Mahler's Guide to

Advanced Ratemaking

CAS Exam 8

prepared by
Howard C. Mahler, FCAS

Study Aid 2018-**8**

Howard Mahler
hmahler@mac.com
www.howardmahler.com/Teaching

Mahler's Guide to Advanced Ratemaking

Information in bold or sections whose title is in bold are more important for passing the exam. Larger bold type indicates it is extremely important. Information presented in italics (including subsections whose titles are in italics) should not be needed to directly answer exam questions and should be skipped on first reading. It is provided to aid the reader's overall understanding of the subject, and to be useful in practical applications.

Solutions to problems are at the end of each section.[1]

[1] Note that problems include both some written by me and some from past exams. The latter are copyright by the Casualty Actuarial Society and are reproduced here solely to aid students in studying for exams. The solutions and comments are solely the responsibility of the author; the CAS bears no responsibility for their accuracy. While some of the comments may seem critical of certain questions, this is intended solely to aid you in studying and in no way is intended as a criticism of the many volunteers who work extremely long and hard to produce quality exams. There are also some past exam questions copyright by the Society of Actuaries.

Past Exam Questions by Section

Sec.		1995 Exam 9	1996 Exam 9	1997 Exam 9	1998 Exam 9
1	Mahler, Shifting Risk Parameters	10, 31	20	44, 45, 46	13, 14, 25
2	Bailey & Simon, Cred. Single Car	6, 30, 32	50	19	26
3	Goldburd, Khare and Tevet, GLMs				
4	ASOP 12: Risk Classification			18	15, 22
5	Robertson, Hazard Group Mapping				
6	Couret & Venter, Class Freq.				
7	Clark, Reinsurance Pricing				
8	Bernegger, Exposure Curves				
9	Grossi & Kunreuther, Catastrophes				
10	Experience Rating	20, 40, 42	4, 27, 28c&d	31a, 32	18, 37b, 38, 39
11	NCCI Experience Rating Plan	16, 41	24, 25	10, 34	17, 20, 36
12	ISO Experience Rating Plan	17	1, 21, 22, 23	9, 33	41
13	Frequency and Loss Distributions				
14	Bahnemann, Distrib. for Actuaries	11, 33, 35	36, 38, 41, 42	13, 36a, 40a	30a, 31, 33, 34
15	Lee Diagrams, Loss Distributions		39	37	29
16	Retrospective Rating	21, 22, 24, 44	29, 31, 32, 34	1, 21, 27	4, 44c, 47
17	NCCI Retro. Rating Plan	46, 47	9	5	2, 5, 42, 46
18	Table M Construction	45	10	22, 23	
19	Table L	25	30, 35		43
20	Lee Diagrams, Retro. Rating	50		4, 26	
21	Limited Table M				
22	Other Loss Sensitive Plans				
23	Pricing LDD Policies				
24	Conclud. Remarks, Indiv. Risk Rat.				

Some questions are based on more than one syllabus reading, particularly on recent exams.
In any case, sometimes it is unclear what is the best section in which to put a question.
In those cases, I have made one of the possible reasonable choices of where to put a question.

Sec.		1999 Exam 9	2000 Exam 9	2001 Exam 9	2002 Exam 9
1	Mahler, Shifting Risk Parameters	48	34	1	
2	Bailey & Simon, Cred. Single Car	1	32	2, 22	47
3	Goldburd, Khare and Tevet, GLMs				
4	ASOP 12: Risk Classification	2, 43b			48
5	Robertson, Hazard Group Mapping				
6	Couret & Venter, Class Freq.				
7	Clark, Reinsurance Pricing				
8	Bernegger, Exposure Curves				
9	Grossi & Kunreuther, Catastrophes				
10	Experience Rating	12, 13, 31	1, 4, 40		
11	NCCI Experience Rating Plan	28	17, 42	25	33
12	ISO Experience Rating Plan	30	2	27	11, 12, 34
13	Frequency and Loss Distributions				
14	Bahnemann, Distrib. for Actuaries	35, 38, 40, 41	39	11, 35, 37c	41, 42
15	Lee Diagrams, Loss Distributions	34, 39	37		43
16	Retrospective Rating	5, 6, 21	5, 6, 44	8, 10, 31, 32	14, 15, 16
17	NCCI Retro. Rating Plan	8, 9, 10, 22, 23, 25		9, 33, 34	35, 40
18	Table M Construction		19, 48	30	36
19	Table L	26	45		38, 39
20	Lee Diagrams, Retro. Rating				17
21	Limited Table M				
22	Other Loss Sensitive Plans				
23	Pricing LDD Policies	42	38		1
24	Conclud. Remarks, Indiv. Risk Rat.				

Sec.		2003 Exam 9	2004 Exam 9	2005 Exam 9	2006 Exam 9
1	Mahler, Shifting Risk Parameters	21	3	2	
2	Bailey & Simon, Cred. Single Car	22	2	3	2
3	Goldburd, Khare and Tevet, GLMs	25			5
4	ASOP 12: Risk Classification		23		
5	Robertson, Hazard Group Mapping			9	
6	Couret & Venter, Class Freq.				
7	Clark, Reinsurance Pricing				
8	Bernegger, Exposure Curves				
9	Grossi & Kunreuther, Catastrophes				
10	Experience Rating	2, 6, 26, 28	15, 16, 39	26	23, 27
11	NCCI Experience Rating Plan	27		24, 27	24
12	ISO Experience Rating Plan	3, 4, 5	14, 41	28	28
13	Frequency and Loss Distributions				
14	Bahnemann, Distrib. for Actuaries	13, 37, 38, 43	5, 6, 19 25, 26	6, 7, 10 23a, 35	6, 8
15	Lee Diagrams, Loss Distributions				
16	Retrospective Rating	7, 32, 33	47	30, 32	32, 35
17	NCCI Retro. Rating Plan	31	18, 20, 45	31	30
18	Table M Construction		43	8	9
19	Table L	30	44		7
20	Lee Diagrams, Retro. Rating	8, 9, 29	4, 17	33	29, 34
21	Limited Table M				
22	Other Loss Sensitive Plans				
23	Pricing LDD Policies	35	46, 48	34, 36	31, 33, 36
24	Conclud. Remarks, Indiv. Risk Rat.				

Sec.		2007 Exam 9	2008 Exam 9	2009 Exam 9	2010 Exam 9
1	Mahler, Shifting Risk Parameters	6			
2	Bailey & Simon, Cred. Single Car	2	5	4	5
3	Goldburd, Khare and Tevet, GLMs	4a	3	3	3
4	ASOP 12: Risk Classification				
5	Robertson, Hazard Group Mapping				
6	Couret & Venter, Class Freq.				
7	Clark, Reinsurance Pricing				
8	Bernegger, Exposure Curves				
9	Grossi & Kunreuther, Catastrophes				
10	Experience Rating	26	23	20	23
11	NCCI Experience Rating Plan	25, 28	25	21	20
12	ISO Experience Rating Plan	27	24	22	21
13	Frequency and Loss Distributions				
14	Bahnemann, Distrib. for Actuaries	7, 8, 10	26, 27	17, 18, 26	17, 26
15	Lee Diagrams, Loss Distributions			24	
16	Retrospective Rating	32, 35		28, 31	27, 29
17	NCCI Retro. Rating Plan		36	30	
18	Table M Construction	30, 34	28		
19	Table L		32, 33	32	
20	Lee Diagrams, Retro. Rating	31	29		25, 31
21	Limited Table M				
22	Other Loss Sensitive Plans				
23	Pricing LDD Policies	33, 36	30, 31	29a	28
24	Conclud. Remarks, Indiv. Risk Rat.			27	24

Sec.		2011 Exam 8	2012 Exam 8	2013 Exam 8	2014 Exam 8
1	Mahler, Shifting Risk Parameters		3		
2	Bailey & Simon, Cred. Single Car	1	6		5
3	Goldburd, Khare and Tevet, GLMs	*3*	2, 4	2	3
4	ASOP 12: Risk Classification				
5	Robertson, Hazard Group Mapping	4	1	4	2
6	Couret & Venter, Class Freq.	2	5	3	1, 4
7	Clark, Reinsurance Pricing	7, 8	7, 10	21, 23, 25	20, 21, 22 23, 25
8	Bernegger, Exposure Curves	9	8	20, 22	
9	Grossi & Kunreuther, Catastrophes	5, 6	9	24	24
10	Experience Rating	15, 16b&c	11, 16a&c	9, 10b	9, 11
11	NCCI Experience Rating Plan	12	13		10
12	ISO Experience Rating Plan	14	14	8	8
13	Frequency and Loss Distributions				
14	Bahnemann, Distrib. for Actuaries	10, 17	15	6	7
15	Lee Diagrams, Loss Distributions	11	22		6
16	Retrospective Rating	20, 25		14	17
17	NCCI Retro. Rating Plan	21	19, 23		
18	Table M Construction			12	13
19	Table L		18	13	
20	Lee Diagrams, Retro. Rating	22	21	15	12, 18
21	Limited Table M				
22	Other Loss Sensitive Plans				
23	Pricing LDD Policies	18, 19	20	16, 19	16, 19
24	Conclud. Remarks, Indiv. Risk Rat.	23			

Added for the 2011 Exam: Bernegger, Robertson, Couret & Venter, Grossi & Kunreuther. Clark Reinsurance Pricing was on Exam 6 prior to 2011.

For the 2016 exam, Goldburd, M.; Khare, A.; and Tevet, D., “Generalized Linear Models for Insurance Rating,” replaced Anderson, D.; Feldblum, S; Modlin, C; Schirmacher, D.; Schirmacher, E.; and Thandi, N., “A Practitioner’s Guide to Generalized Linear Models”

Sec.		2015 Exam 8	2016 Exam 8	2017 Exam 8
1	Mahler, Shifting Risk Parameters	4		
2	Bailey & Simon, Cred. Single Car	1	1	3
3	Goldburd, Khare and Tevet, GLMs	3	4, 5, 6, 7	4, 5, 6
4	ASOP 12: Risk Classification		3	
5	Robertson, Hazard Group Mapping	6	2	2
6	Couret & Venter, Class Freq.	5		
7	Clark, Reinsurance Pricing	21, 23	20	19
8	Bernegger, Exposure Curves	20	21	18
9	Grossi & Kunreuther, Catastrophes	22	18, 19	20
10	Experience Rating	10, 11, 12	11	11
11	NCCI Experience Rating Plan		9, 10	
12	ISO Experience Rating Plan	9		9, 10
13	Frequency and Loss Distributions			
14	Bahnemann, Distrib. for Actuaries	8a		7, 8, 14
15	Lee Diagrams, Loss Distributions	7		12
16	Retrospective Rating	15, 17		13
17	NCCI Retro. Rating Plan	16		15, 17
18	Table M Construction		12	16
19	Table L		14	
20	Lee Diagrams, Retro. Rating		13	
21	Limited Table M			
22	Other Loss Sensitive Plans			1
23	Pricing LDD Policies	13, 14, 18, 19	15, 16	
24	Conclud. Remarks, Indiv. Risk Rat.			

ASOP No. 12 Risk Classification was added to the syllabus for 2017.
It replaced American Academy of Actuaries "Risk Classification Statement of Principles."
Some of the past exam questions on Risk Classification no longer apply.

For the 2017 exam, many previous readings were replaced by:
a CAS Study Note "Individual Risk Rating," by Fisher, McTaggart, Petker, and Pettingell,
and a CAS Monograph "Loss Distributions for Actuaries," by Bahnemann.

Section 1, Mahler, Shifting Risk Parameters[1]

Errata for "An Example of Credibility and Shifting Risk Parameters" by Howard C. Mahler:[2]

Page 286, first sentence:[3]
τ is distributed on the range [-1, 1].
If the actual correlation, $\rho = 0$, then t is symmetrically distributed on the range [-1, 1].

Page 297, fourth line:[4]

$$\text{Cov}[X_i, X_j] = \begin{cases} \ell(\text{lj-il})\, \zeta^2 & i \neq j \\ \zeta^2 + \delta^2 & i = j \end{cases}$$

[1] "An Example of Credibility and Shifting Risk Parameters", by Howard C. Mahler, PCAS 1990.
Candidates will not be tested on the Appendices.
CAS Learning Objective A1.
[2] Not official.
[3] In Appendix B, not on the syllabus.
[4] In Appendix D, not on the syllabus.

Shifting Risk Parameters:

Shifting risk parameters: The parameters defining the risk process for an individual insured are not constant over time. There are (a series of perhaps small) permanent changes to the insured's initial risk process as one looks over several years.[5]

For example, a private passenger automobile insured's risk parameters might shift if a major new road were opened in his locality or if he changed the location to which he commutes to work.

In another example, the private passenger automobile insurance experience of a town relative to the rest of the state, in other words the town's relativity, could shift as that town becomes more densely populated.

In yet another example, the procedures and machines used to manufacture widgets change over time. This could result in changes over time in the expected pure premium and therefore the relativity for the Widget Manufacturing Class for Workers Compensation Insurance.

For insurance situations, risk parameters are never totally constant over decades. However, depending on the length of the time period considered and the particular data, the magnitude of the shifts can be large or small.

If risk parameters shift significantly over time, this will significantly effect the optimal credibility to assign to years of past data in order to predict the future.

The Baseball Paradigm:

In Mahler's "An Example of Credibility and Shifting Risk Parameters," the author evaluates various estimates for baseball teams' future losing percentages using historical losing percentages Mahler discusses the impact of shifting parameters over time in this context.

Mahler combines a substantive actuarial topic, the effect of shifting risk parameters on optimal credibility values, with an excellent baseball analogy.
Mahler seeks optimal credibility values, primarily for experience rating but also for class ratemaking, reserving, and other actuarial topics. Section 11 of the paper explains the covariance structure and provides the formulas for estimating optimal credibility values. But many readers of the paper have trouble digesting the theory. The baseball analogy is an excellent means of explaining the intuition.

This is an analogy with the characteristics needed, but without the problems of insurance data.

---

[5] Taken from page 456 of "Credibility With Shifting Risk Parameters, Risk Heterogeneity, and Parameter Uncertainty," by Howard C. Mahler. PCAS 1998, not on the syllabus.

1. Insurance applications of credibility are complex, since different size risks have different degrees of partial credibility. The baseball teams all play the same number of games; they are the same size, so there is no need for partial credibilities.
2. Insurance is complicated by loss development. There is no loss development in baseball; when the season is over, we know the won-loss record.
3. An insurance portfolio changes over time, as new insureds are added and as old insureds leave. Mahler has the same baseball teams for 60 years.

The analogy of the baseball example to an insurance industry situation:

losing percentage of baseball team. ⇔ loss ratio of an insured (or class).

losing percentage of team compared to average.
⇔ loss ratio of an insured compared to average. ⇔ relativity of a class.

predicting future losing percentage of a team.
⇔ experience rating an insured. ⇔ determining new class relativity.

Advantages of the Baseball Data:[6]

1. Over a very extended period of time there is a constant set of risks (teams).
In insurance, there are generally insureds who leave the data base and new ones that enter.

2. The loss data over this extended period of time are readily available, accurate and final.
In insurance, the loss data are sometimes hard to compile or obtain and are subject to possible reporting errors and loss development.

3. Each of the teams in each year plays roughly the same number of games.
Thus the loss experience is generated by risks of roughly equal "size."
Thus, in this example, one need not consider the dependence of credibility on size of risk.

[6] See Section 3.1 of the paper.

Sampling Error:

The use of credibility mitigates distortions caused by sampling error. Part of sampling error is the inability to get accurate readings because the measuring instruments are too crude. We don't get accurate estimates of incurred losses until years after the accident, because we can not observe future court decisions. Mahler wants to avoid this topic, so that he can focus on shifting risk parameters over time. Therefore, Mahler analyses a data set, baseball won-loss records, that mitigates sampling error problems.

Team Differences:

Mahler demonstrate that baseball losing percentages have the characteristics that are relevant for credibility studies. If all insureds were the same, there would be no use for experience rating. So Mahler shows that the losing percentages of the various teams are not random; there are better teams and worse teams.[7]

One might still argue: "Maybe teams are not the same, but perhaps past performance is a poor predictor of future performance." So Mahler shows that experience in one period has predictive power for other periods. Specifically, Mahler shows that **there is a significant correlation between the results of years close in time. Thus recent years can be usefully employed to predict the future.**[8]

[7] See page 229 and Table 3 of Mahler. The so-called Binomial Test; see 9, 11/98, Q.25.
[8] See page 236 of Mahler.

Chi-Square Test of Whether the Risk Parameters Shift Over Time:

Mahler uses two methods to test whether risk parameters shift over time:
(i) He does chi-square tests.
(ii) He examines the correlations in pairs of years separated by a constant period.

The Chi-Square Test as used by Mahler can be summarized as follows:[9]

- Applied to the data of one team.
- H_0: The expected losing percentage is the same over time for this team.
- Group data into appropriate intervals.
  Mahler groups the 60 years into 5 year non-overlapping intervals.
- Calculate the mean losing percentage for the team over the 60 years.
- Then calculate for each interval: $(A - E)^2/E$,
  where A = actual observation = (5 year mean losing percentage)(5 years)(150 games),
  and E = expected observation = (60 year mean losing percentage)(5 years)(150 games).
- Sum up the contributions for all 12 intervals in order to get the chi-square statistic.
- If the statistic is greater than the critical value for number of intervals - 1 = 11 degrees of freedom, then reject the null hypothesis that parameters do not shift over time.

For each team, Mahler finds that there is less than a 0.2% chance that the different five-year segments were all drawn from the same distribution. Therefore, he rejects the hypothesis that the means are the same over time, in favor of the hypothesis that the parameters shift (at a noticeable amount) over time.

Chi-Square Statistics and p-values[10] [11]

NL1	NL2	NL3	NL4	NL5	NL6	NL7	NL8
107	45	98	35	39	73	114	119
7×10^{-18}	5×10^{-6}	4×10^{-16}	0.025%	5×10^{-5}	5×10^{-11}	3×10^{-19}	3×10^{-20}

AL1	AL2	AL3	AL4	AL5	AL6	AL7	AL8
114	69	34	30	97	162	53	65
3×10^{-19}	2×10^{-10}	0.036%	0.158%	7×10^{-16}	5×10^{-29}	2×10^{-7}	1×10^{-9}

[9] See Table 4 in Mahler. This an application of material covered on preliminary exams.
[10] The values of the Chi-Square statistic are taken from Mahler's Table 4. I have added the probability values. Note that all of the p-values are less than 0.2%.
[11] The teams are identified in footnote 6 on page 229 of the paper by Mahler.
For example, AL5 is the New York Yankees.

Correlations Test of Whether the Risk Parameters Shift Over Time:

Here is a description of Mahler's correlation test, as applied to insurance data.

Suppose we have N similar risks and T years. We denote the manual loss ratio for risk n in year t as $LR_{n,t}$. (We use manual loss ratios, not standard loss ratios.)
For each year t, we have N loss ratios $\{LR_{1,t}, LR_{2,t}, ... , LR_{N,t}\}$.

For the one year differential, we examine the correlation of the T - 1 sets of pairs:
$\{LR_{1,1}, LR_{2,1} ... , LR_{N,1}\}$ with $\{LR_{1,2}, LR_{2,2}, ... , LR_{N,2}\}$
$\{LR_{1,2}, LR_{2,2} ... , LR_{N,2}\}$ with $\{LR_{1,3}, LR_{2,3}, ... , LR_{N,3}\}$
etc.
We take the average correlation for the one year differential.

We do the same for the two year differential, using the correlation of the T - 2 sets of pairs:
$\{LR_{1,1}, LR_{2,1} ... , LR_{N,1}\}$ with $\{LR_{1,3}, LR_{2,3}, ... , LR_{N,3}\}$
$\{LR_{1,2}, LR_{2,2} ... , LR_{N,2}\}$ with $\{LR_{1,4}, LR_{2,4}, ... , LR_{N,4}\}$
etc.
We take the average correlation for the two year differential.

We do the similar calculation for the other differentials in years.[12]

If the risk parameters do not shift over time, the average correlation should not differ significantly between the one year differential, two year differential, and so forth. If the risk parameters shift over time, the average correlation should be highest for the one year differential, second highest for the two year differential, and so forth. The rate at which the correlation drops as the differential widens measures how fast the risk parameters shift over time.[13]

Mahler's results in his Table 5 indicate that the risk parameters are shifting at a high rate for the baseball data examined.

[12] Results are shown in Table 5 in Mahler.
[13] This is discussed further at pages 640 to 642 of "A Markov Chain Model of Shifting Risk Parameters," by Howard C. Mahler, PCAS 1997, not on the syllabus.

Table 5 from the Paper

	Correlations	
Years Separating Data	NL	AL
1	0.651	0.633
2	0.498	0.513
3	0.448	0.438
4	0.386	0.360
5	0.312	0.265
6	0.269	0.228
7	0.221	0.157
8	0.190	0.124

The correlations decline as the separation increases.

Years further apart are less correlated than years closer together.

Data from last year is more valuable to predict the coming year, than data from 5 years ago.

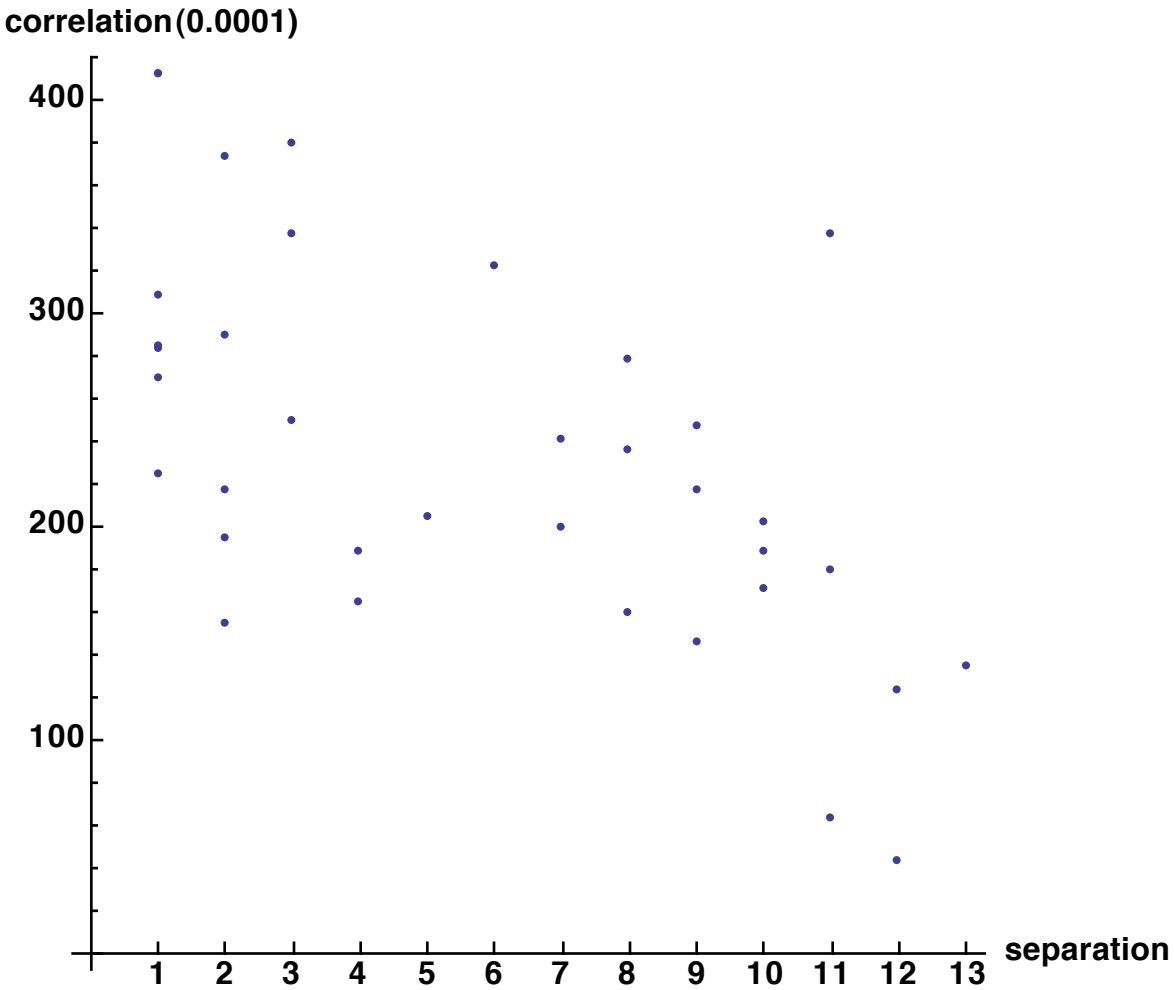
Thus the NCCI Experience Rating Plan, which assuming equal volume of data for each year gives equal weight to each year of data, is an approximation to the theoretically most accurate plan.[14]

[14] There are other complications such as the maturity of the data.
See "Credibility With Shifting Risk Parameters, Risk Heterogeneity, and Parameter Uncertainty," by Howard C. Mahler. PCAS 1998, not on the syllabus.

California Driver Data:[15]

A similar correlations test has been performed on data for drivers in California.
The data show the number of accidents annually in 1961-1963 and 1969-1974, for a sample of drivers licensed from 1961 to 1974. There were 54,165 drivers divided between male and female.

Correlations were computed for pairs of years of data separated by different numbers of years. For example, 1961 and 1962 are separated by one year, while 1961 and 1970 are separated by 9 years. Here is a graph of the results for female drivers:[16] [17]

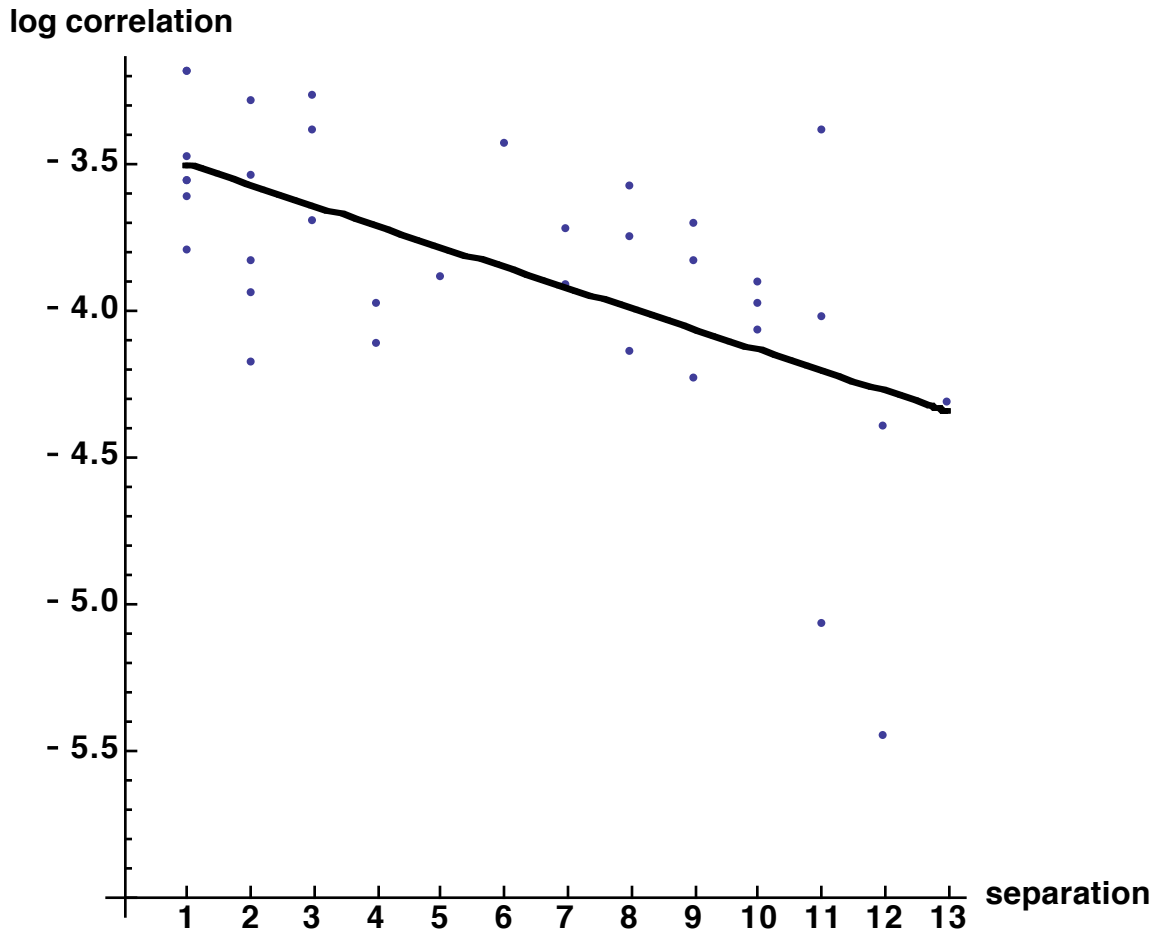


Real insurance data can be messy; the data and thus the correlations between years are subject to significant random fluctuation. However, the correlations do appear to be declining as the separation between years increases.

---

[15] From Exhibit 1 in "The Credibility of a Single Private Passenger Driver," by Howard C. Mahler, PCAS 1991.
[16] Due to the gap in the years of data, some separations have fewer values than one would otherwise expect.
[17] There were two cases where for a separation of one year the correlation was 0.0412.

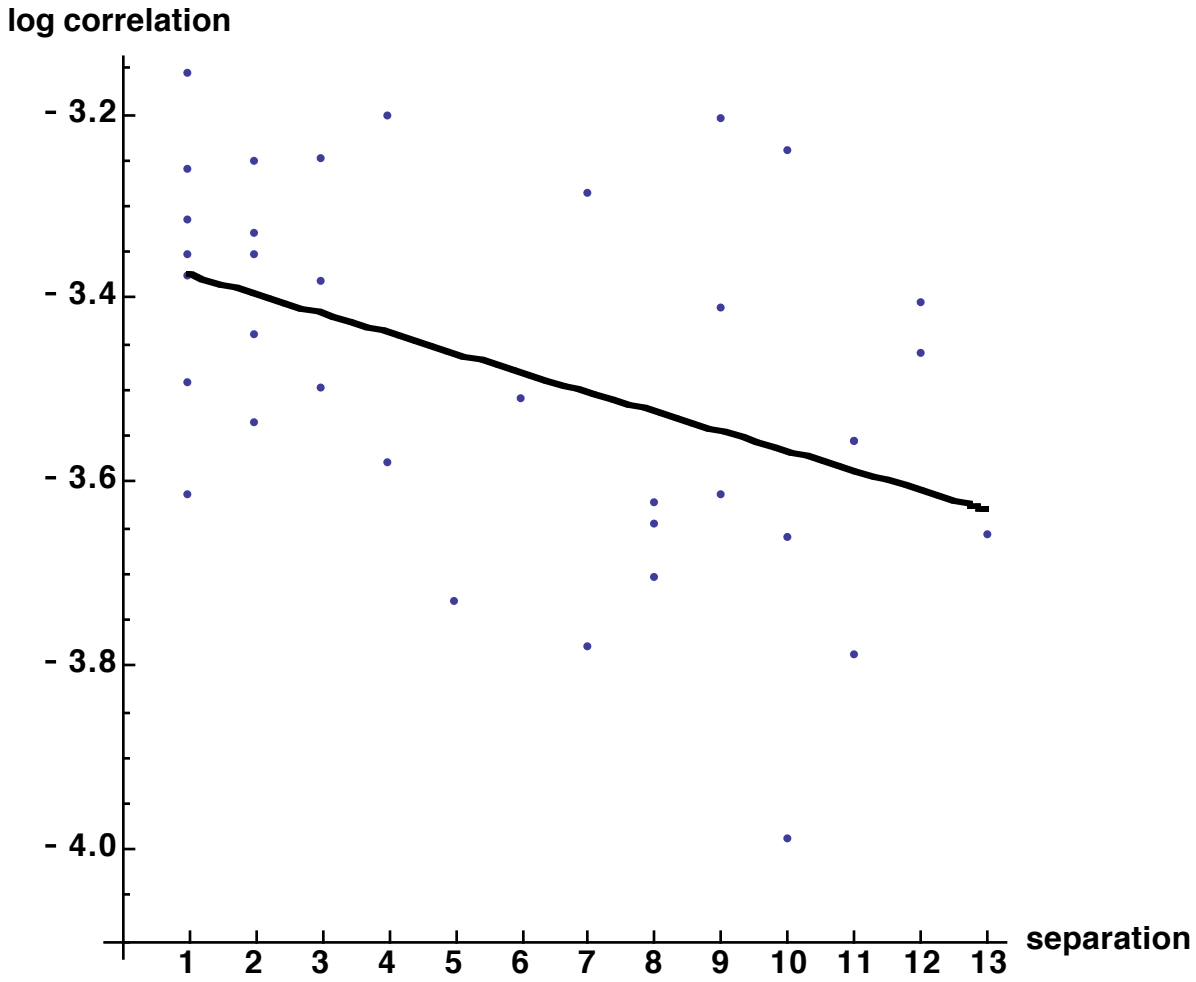
Here is the same data on a log scale. Also shown is the least squares line fit to the logs of the correlations, -3.435 - 0.06999 x.[18]



Thus there is evidence that the correlations are declining with separation and thus that parameters are shifting over time. The least squares line is: ln(corr) = -3.435 - 0.06999 x. $\Leftrightarrow$ Correlation = (0.0322) (0.932^x), where x is the separation in years.[19]
The 0.932 measures the rate at which parameters are shifting; the further this base is from one, the more quickly parameters are shifting.

[18] R^2 = 0.36. The p-value for testing whether the slope is zero is 0.0001; so there is very good evidence that the slope is not zero. A negative slope corresponds to correlations declining with separation.
[19] There is a theoretical reason to expect correlations to follow this type of curve.
See "A Markov Chain Model of Shifting Risk Parameters," by Howard C. Mahler, PCAS 1997.

Here is a similar graph, but for the male drivers. The least squares line fit to the logs of these correlations is: -3.354 - 0.02140 x.[20]



Again there is evidence that the correlations are declining with separation and thus that parameters are shifting over time. The least squares line is: ln(corr) = -3.354 - 0.02140 x.⇔

Correlation = (0.0349) (0.979^x), where x is the separation in years.

The 0.979 measures the rate at which parameters are shifting. For females the similar base was 0.932, indicating parameters are shifting much more quickly for female drivers than for male drivers.[21]
For male drivers, the number of years of separation required for the correlation to decline to half of its original value is: ln(0.5) / ln(0.979) = 33; for females it is: ln(0.5) / ln(0.932) = 10.

[20] R^2 = 0.19. The p-value for testing whether the slope is zero is 0.007; so there is good evidence that the slope is not zero. A negative slope corresponds to correlations declining with separation.
[21] This conclusion based on this one data set should be taken with a grain of salt.

The Effect of this Pattern of Correlations:[22]

The correlation between years that are close together is higher than the correlation between years that are further apart. Therefore, the credibility assigned to more recent years of data should be higher for predicting the future.[23]

Delays in receiving data make estimates of the future less accurate.
Therefore, the optimal credibility decreases with increased delays in receiving the data.
When predicting year 5, it is better to have data for years 2, 3, and 4 than for years 1, 2, and 3.[24]

Up to a given point, using more year of data, with an optimal set of credibilities applied to each year, increases the accuracy of the estimate of the future. However, at a certain point adding more older years of data, no longer increases (measurably) the accuracy of the estimate.[25]

With equal weight to each year, at a certain point adding more older years of data, no longer increases the accuracy of the estimate; instead in this case at some point adding older years of data decreases the accuracy of the estimate.[26]

Estimators:

A credibility weighting formulas (credibility estimator) might be 60% of last year's loss ratio plus 40% of the overall average loss ratio. This estimator has two terms; a simple estimator has a single term, such as the overall mean, last year's experience, or the experience from two years ago.

Mahler's credibility estimators are:
1. A linear combinations of a few simple estimators.
2. Unbiased for the set of teams as a whole.
3. Analogous to experience rating.

---

[22] See for example, 9, 11/01, Q.1.
[23] In this paper, Mahler assumes the different years contain the same volume of data.
[24] Ignoring possible complications such as loss development.
See for example, Sections 7.10, 7.11, 7.12 and 10.10 of "Credibility With Shifting Risk Parameters, Risk Heterogeneity, and Parameter Uncertainty," by Howard C. Mahler. PCAS 1998, not on the syllabus.
See also the 9th and 10th pages of "Workers' Compensation Classification Credibilities", by Howard C. Mahler, Fall 1999 CAS Forum.
[25] See Table 19 in Mahler. The slower the rate of shifting parameters, the longer it takes to reach such a point of diminishing returns.
[26] See Table 19 in Mahler. The slower the rate of shifting parameters, the longer it takes to reach a point where one should stop adding older years.

The Three Criteria:[27]

Mahler discusses the use of three criteria to determine optimal credibilities:
1. Least Squares Error.[28]
2. Small chance of a large error.[29]
3. Meyers/Dorweiler

If the predicted value is E (expected) and the observed value is O, the squared error is $(E - O)^2$. To find the optimal credibility formula, we write SE = squared error = $\Sigma (E - O)^2$ as a function of the credibility Z and we set to zero the partial derivative of the squared error with respect to Z. The estimator (credibility formula) that gives the smallest squared error, on average, is the best. We minimize the expected squared error, not the squared error for a particular estimate.

The chance of a large error is the probability that the absolute value of (E - O)/E is more than a given number k. Small chance of large error chooses the credibility formula that minimizes Prob[$|(E - O)/E| > k$]. The estimator (credibility formula) that gives the smallest number of large errors is the best.

Meyers/Dorweiler is different.[30] Perhaps the optimal experience rating plan uses 3 years of data and 40% credibility, but we use a plan with 6 years of data and 50% credibility. We fear that there may be patterns in the errors, meaning that the underwriter prefers to write either risks with credit modifications or risks with debit modifications. No matter the magnitude of the errors in the experience rating plan, the plan passes the Meyers/Dorweiler test if underwriters are indifferent between credit risks and debit risks.[31]

Meyers/Dorweiler criterion is concerned with the pattern of the errors. Unlike the other two criteria, large errors are not an issue for the Meyers/Dorweiler criterion, as long as there is no pattern relating the errors to the experience modification.[32]

---

[27] See Section 7 of Mahler. Know the three methods, how they work, and any unusual characteristics.
These methods - and particularly the Meyers/Dorweiler method - form the basis of likely exam questions.
[28] The basis of Buhlmann Credibility or greatest accuracy credibility.
[29] The idea behind Classical Credibility.
[30] Taken from Glenn G. Meyers in "An Analysis of Experience Rating", PCAS 1985, based upon the ideas of Paul Dorweiler.
[31] Dorweiler's view is quoted in "Workers Compensation Experience Rating, What Every Actuary Should Know," by Gillam at page 218: "A necessary condition for proper credibility is that the credit risks and debit risks equally reproduce the permissible loss ratio."
See also "Experience Rating - Equity and Predictive Accuracy," by Venter, at page 7: "On a standard premium basis . . . the loss ratios should be less dispersed, and, ideally, all equal for a better working plan." and at page 2: "From the viewpoint of the insurer, after experience rating, all insureds have the same expected profit potential, regardless of their past loss history."
[32] See the example at page 271 of Mahler.

The Meyers/Dorweiler criterion uses Kendall's t (tau), a measure of correlation.[33]
The optimal credibility using the Meyers/Dorweiler criterion has a Kendall's tau of 0.
We measure the correlation of:
1. (actual losing percentage)/(predicted losing percentage), and
2. (predicted losing percentage)/(overall average losing percentage).

Item #2 is analogous to the experience modification.[34]
Item #1 is analogous to the modified loss ratio, the ratio of losses to modified premium.[35] [36]
Thus the Meyers/Dorweiler criterion desires that the correlation between the experience modification and the modified loss ratio be zero.

If this correlation were positive, then debit risks, those with modifications greater than 1, would tend to have larger modified loss ratios. In other words, after applying the experience rating plan, underwriters would on average not want to write debit risks. Credit risks would tend to have smaller modified loss ratios. In other words, after applying the experience rating plan, underwriters would on average want to write credit risks.

If this correlation were negative, then debit risks, would tend to have smaller modified loss ratios. In other words, after applying the experience rating plan, underwriters would on average want to write debit risks. Credit risks would tend to have larger modified loss ratios. In other words, after applying the experience rating plan, underwriters would on average not want to write credit risks.

Unlike the other two criteria, the Meyers/Dorweiler criterion can not be used to distinguish between using different number of years of data. For each value of N, there is a value of Z such that the correlation is zero.[37]

---

[33] The details of computing Kendall's tau are in Appendix B of Mahler, not on the syllabus.
It involves comparing the ranked order of the two vectors.

[34] If for example, the predicted losing percentage is 60%, then the ratio to the average losing percentage is 60%/50% = 1.2. This is similar to an experience modification factor of 1.2; this team (insured) is predicted to be worse than average. Similarly, a predicted losing percentage of 45% corresponds to an experience modification factor of 45%/50% = 0.9; this team (insured) is predicted to be better than average.

[35] Modified premium = (manual premium)(experience modification).
The modified premium is what would be called standard premium in Workers Compensation.

[36] Modified loss ratio = losses/{(manual premium)(experience modification)}.
Losses ⇔ actual losing percentage. manual premium ⇔ overall mean = 50%.
experience modification ⇔ (predicted losing percentage)/(overall average losing percentage).
Therefore Item #1 = (actual losing percentage)/(predicted losing percentage) ⇔
Losses/{(manual premium)(experience modification)} = Modified loss ratio.

[37] See page 249 of Mahler.

Testing an Experience Rating Plan:

There are several ways to test an experience rating plan:

- We examine whether credit risks or debit risks are more profitable. If credit risks are more profitable than debit risks, then the experience rating credibility is too low; we should give credit risks bigger credits. If credit risks are less profitable than debit risks, then the experience rating credibility is too high; we should give credit risks smaller credits.
- The quintiles test is conceptually the same as the credit vs debit above, but it uses five categories of risks, ranked in order of the experience modifications, instead of two.[38]
- *The ratio of variances generalizes the quintiles test: we rank the risks by their modifications into N groups, from lowest mods to highest mods. We determine the average manual and standard loss ratios in each group, and we compute the variance of the average standard loss ratios divided by the variance of the average manual loss ratios. The lower the ratio of the variances, the better the experience rating plan.*
- The Meyers-Dorweiler test uses the Kendall t statistic for the correlation between the actual loss ratio relativities and the indicated loss ratio relativities.
- The minimum squared error test sums the squared errors between the actual loss ratio relativity and the indicated loss ratio relativity; the lower the sum of the squared errors, the better the experience rating plan. *Alternatives to the minimum squared error test are the minimum χ^2 test and the minimum absolute error test.*
- *Let μ be the expected loss ratio for an insured prior to experience rating. Let M = E[μ] over a group of insureds. Let F be the estimator of* m*, in this context the result of using an experience rating plan. Then the efficiency is of F is: 1 - $\frac{E[(m - F)^2]}{E[(m - M)^2]}$.*

  The higher the efficiency, the better the experience rating plan.[39]

---

[38] See "Experience Rating - Equity and Predictive Accuracy," by Gary G. Venter.
According to William R. Gillam at pages 219-220 of "Workers Compensation Experience Rating: What Every Actuary Should Know", "The test statistic for each size group is the variance of the modified ratios divided by the variance of the unmodified ratios. A low test statistic indicates a plan that has eliminated much of the between variance (in risk theoretic terms) or made risks of differing experience more equally desirable."

[39] This is the efficiency test of Glenn G. Meyers in "An Analysis of Experience Rating", PCAS 1985, mentioned at page 220 of "Workers Compensation Experience Rating: What Every Actuary Should Know," by William R. Gillam. Meyers applies efficiency to models where the risk parameters vary between insureds within a group. In such models we are assumed to know the expected loss ratio for each insured, and we see how well the experience rating plan works for a set of data generated from this group.

Rating plans which do well on one test often do well on other tests. But the tests examine different characteristics of the rating plan. Some tests check for bias, often referred to as patterns of errors (credit-debit; quintiles; Meyers-Dorweiler) and some tests check for accuracy (minimum squared error, minimum χ^2, minimum absolute error, ratios test).

A plan is biased if the experience modification helps us select among risks. For example, suppose we gave all risks 10% credibility, but the proper credibility is higher. A risk with a credit modification is overpriced, since the true experience modification would be lower with greater credibility, and a risk with debit modification is underpriced, since the true experience modification would be higher with greater credibility.

In a perfect plan, the loss ratio relativity predicted by the plan would be the expected relativity.[40]

- The plan is unbiased if no value of the experience modification is a predictor of rate redundancy or inadequacy versus other risks after application of the modification, in other words with respect to standard premium.
- The plan is accurate if the difference between the predicted loss ratio and the expected loss ratio is zero; any differences between the predicted loss ratio and the actual loss ratio stem from random loss fluctuations.

We desire an experience rating plan that is as close to unbiased and as accurate as practical.[41]

Whether a risk is a debit risk or a credit risks depends on the plan. It is tempting to presume that the credit risks are the risks with expected loss ratio relativities less than one and the debit risks are the risks with expected loss ratio relativities more than one. This is not correct, since we do not know the expected losses for any risk.

Rather, the credit risks are the risks with experience modifications below one. Whether a risk is a credit risk or a debit risk depends on the plan parameters, such as the expected losses, the state accident limit, the primary-excess split, and the credibilities. For a particular plan, we desire that most of the credit risks actually are better than average, and that most of the debit risks are actually worse than average.

---

[40] No actual experience rating plan is ever perfect.

[41] As mentioned by Venter in "Experience Rating - Equity and Predictive Accuracy," we also want the experience rating plan to provide incentives for the insured to reduce losses.

An Example of Comparing Experience Rating Plans:

Consider five prototypical insureds of similar size. We show the experience modifications predicted by two rating plans, P and Q. We show the subsequently observed loss ratio to manual premium relativities, for the period of time predicted by the experience rating plans.[42] [43]

	Experience Modification		Subsequently Observed
Risk	P	Q	Manual Loss Ratio Relativity
1	0.75	0.86	0.71
2	0.80	0.90	0.79
3	0.91	0.94	0.94
4	1.05	1.02	1.14
5	1.44	1.26	1.42

Both plans P and Q seem to do a reasonable of predicting which risks will be better than average and which risks will be worse than average. Either plan would be better than no experience rating.

Let us look at the loss ratios to standard premium relativities for each plan.
For example, for Risk 1 for Plan P, 0.710/0.750 = 0.947.

Risk	Manual L.R.	Mod for P	Standard L.R. for P	Mod for Q	Standard L.R. for Q
1	0.710	0.750	0.947	0.860	0.826
2	0.790	0.800	0.988	0.900	0.878
3	0.940	0.910	1.033	0.940	1.000
4	1.140	1.050	1.086	1.020	1.118
5	1.420	1.440	0.986	1.260	1.127

Ideally we would like the loss ratios to standard premium to be similar for debit and credit risks; in other words after the application of experience rating all risks should ideally have the same expected loss ratio.[44]

---

[42] Any useful comparison would involve thousands of insureds. We show 5 solely for illustrative purposes.
[43] Analogous to Exhibit 3, Part 2 of "Parameterizing the Workers Compensation Experience Rating Plan," by William R. Gillam, not on the syllabus.
The values shown in Gillam are for risks grouped into quintiles.
The lowest quintile for Plan P, would be the insureds with the 20% lowest modifications using Plan P.
The lowest quintile for Plan Q, would be the insureds with the 20% lowest modifications using Plan Q.
The lowest quintile for Plan P, would be similar to the lowest quintile for Plan Q, but would consist of a somewhat different set of insureds.
[44] For thousands of risks, the observed loss ratio for a quintile would be close to the expected loss ratio.
For a single insured, this need not be the case.

We see that for the best and worst risks, Plan P does a better job of this than Plan Q.
Plan P appears to be more responsive than Plan Q; in other words Plan P assigns a higher credibility to the insureds own experience.[45] [46]

The Meyers/Dorweiler criteria would compute the correlation between the loss ratios to standard premium and the experience modification. We prefer this correlation to be close to zero.

Meyers and Mahler use the Kendall t statistic; however, how to compute that is in Appendix B of Mahler, not on the syllabus.[47] For illustrative purposes we can use the usual sample correlation:
$r = \hat{Cov}[X, Y] / (s_X s_Y) = \Sigma(X_i - \bar{X})(Y_i - \bar{Y}) / \sqrt{\Sigma(X_i - \bar{X})^2 \, \Sigma(Y_i - \bar{Y})^2}$.

We take X_i to be the standard loss ratio relativities, and Y_i to be the experience modification for a given plan.

Exercise: For Plan P, calculate the sample correlation between the loss ratios to standard and the experience modifications.
[Solution: X = (0.947, 0.988, 1.033, 1.086, 0.986).
Y = (0.75, 0.8, 0.91, 1.05, 1.44).
$\bar{X} = 1.008$. $\bar{Y} = 0.99$.
$s_X^2 = \{(0.947 - 1.008)^2 + ... + (0.986 - 1.008)^2\} / (5 - 1) = 0.0028285$. $s_Y^2 = 0.07655$.
$\hat{Cov}[X, Y] = \{(0.947 - 1.008)(0.75 - 0.99) + ... + (0.986 - 1.008)(1.44 - 0.99)\} / (5 - 1) = 0.002805$.
$r = 0.002805 / \sqrt{(0.0028285)\,(0.07655)} = 0.1906$.]

For Plan P, r = 0.19, while for Plan Q, r = 0.83. Thus by this criterion, very similar to the Meyers/ Dorweiler criteria, Plan P is better than Plan Q.

The quintiles test would compare the squared deviations from the mean before and after experience rating.[48] We would like the squared deviations after experience rating to be small.

We group the insureds into five equally sized groups. The first group would contain those insureds with the smallest modifications under the given experience rating plan. The final group would contain those insureds with the largest modifications under the given experience rating plan.[49]

---

[45] There may be other differences between the plans, such as whether and how they split the losses into primary and excess.
[46] Plan R that assigned even more credibility to this size of risk than Plan P, might do worse than Plan P. Higher credibility is not always better!
[47] See pages 300-302 of Meyers, "An Analysis of Experience Rating," PCAS 1985, not on the syllabus.
[48] See "Workers Compensation Experience Rating: What Every Actuary Should Know", by Gillam, not on the syllabus.
[49] In this illustrative example, we only have five insureds, so each quintile has just one insured.

For the manual premium loss ratio relativities, the average is 1.00.
The sum of squared differences from the mean is:
$(0.71 - 1)^2 + (0.79 - 1)^2 + (0.94 - 1)^2 + (1.14 - 1)^2 + (1.42 - 1)^2 = 0.3278$.

For the standard premium loss ratio relativities for Plan P, the average is 1.008.
The sum of squared differences from the mean is:
$(0.947 - 1.008)^2 + (0.988 - 1.008)^2 + (1.033 - 1.008)^2 + (1.086 - 1.008)^2 + (0.986 - 1.008)^2 = 0.0113$.

For the standard premium loss ratio relativities for Plan Q, the average is 0.990.
The sum of squared differences from the mean is:
$(0.826 - 0.990)^2 + (0.878 - 0.990)^2 + (1.000 - 0.990)^2 + (1.118 - 0.990)^2 + (1.127 - 0.990)^2 = 0.0747$.

For Plan P, the quintiles test statistic is: 0.0113/0.3278 = 0.034.
For Plan Q, the quintiles test statistic is: 0.0747/0.3278 = 0.228.

Therefore, based on the quintiles test, Plan P works better than Plan Q.[50]
After the application of the experience modifications from Plan P, the loss ratios to standard vary less among the quintiles of insureds, than they do for Plan Q.

[50] For this size of risk, and for the limited number of risks looked at for illustrative purposes.

Kendall's Tau:[51]

Kendall's tau is a measure of correlation that depends on ranks.
Kendall's Tau is not as sensitive to strong outliers as Pearson's correlation coefficient

In order to compute Kendall's tau, the first step is to order the first elements of the pairs from smallest to largest. Then list the resulting ranks of the second elements.[52]

For example let us take eight pairs of heights of fathers and their adult son:

Father	Son	Rank	Concordant	Discordant
53	56	1	7	0
54	58	2	6	0
57	61	4	4	1
58	60	3	4	0
61	63	6	2	1
62	62	5	2	0
63	65	8	0	1
66	64	7		
Sum			25	3

The number concordant listed in a row is the number of ranks below it in the column that are greater than the given rank. The number discordant listed in a row is the number of ranks below it in the column that are less than the given rank.

Then $\tau = \frac{C - D}{C + D} = \frac{25 - 3}{25 + 3} = 0.7857$.

Note that the denominator, $25 + 3 = 28 = (8)(8 - 1) / 2 = n (n-1)/2$.

[51] See Appendix B of Mahler, not on the syllabus.

[52] One would get the same Kendall's correlation by instead ordering the second elements of the pairs from smallest to largest, and then listing the resulting ranks of the first elements.

When there are no ties, in order to calculate Kendall's tau:[53]

1. Order the first elements of the pairs from smallest to largest.
2. List the resulting ranks of the second elements of the pairs.
3. The number concordant listed in a row is the number of ranks below it in the column that are greater than the given rank. C = sum of concordants.
4. The number discordant listed in a row is the number of ranks below it in the column that are less than the given rank. D = sum of discordants.
5. $\tau = \frac{C - D}{C + D} = \frac{C - D}{n (n-1) / 2}$.

Exercise: You are given six risks of similar size.

Risk	Experience Modification	Subsequent Loss Ratio to Standard Premium
A	1.00	74%
B	0.70	66%
C	1.20	107%
D	1.40	88%
E	0.80	71%
F	0.90	63%

Calculate Kendall's tau between the experience modifications and the subsequent loss ratios to standard premium.

[Solution: Order the risks by the rank of their experience modification.

Risk	Experience Mod.	Loss Ratio to Stand. Prem.	Concordant	Discordant
B	0.70	66%	4	1
E	0.80	71%	3	1
F	0.90	63%	3	0
A	1.00	74%	2	0
C	1.20	107%	0	1
D	1.40	88%		
Sum			12	3

$\tau = \frac{C - D}{C + D} = \frac{12 - 3}{12 + 3} = 0.6$.

Comment: If Kendall's tau were close to zero, that would indicate that an Experience Rating Plan is working well according to the Meyers/Dorweiler criterion. For a practical application, we would look at many more than 6 insureds.]

[53] Things get a little more complicated when there are ties.

Assuming independence, and thus that the actual correlation is zero, Kendall's tau has a mean of zero and variance of: $\frac{2\ (2n+5)}{9n\ (n-1)}$.

Thus one can use Kendall's Rank Correlation Coefficient and the Normal approximation to test the hypothesis that there is no relationship between the two samples.[54]

$Z = \tau \sqrt{\frac{9n\ (n-1)}{2\ (2n+5)}}$.

One can perform the usual two-sided and one-sided tests.

For the heights example, with a sample size of 8, $\tau = 0.7857$.

H_0: The correlation of the joint distribution from which the paired samples were drawn is zero.

H_1: The correlation of the joint distribution from which the paired samples were drawn is positive.

$Z = 0.7857 \sqrt{\frac{(9)(8)(7)}{(2)(21)}} = 2.722.$

Thus for this one-sided test, the probability-value is: $1 - \Phi[2.722] = 0.32\%$.

We reject H_0 at a 0.5% level.

In other words, at a 0.5% significance level we conclude that there is a positive correlation between the heights of fathers and sons; taller fathers tend to have taller sons.

Exercise: For 200 experience rated risks of a similar size, an actuary calculates Kendall's tau between the experience modifications and the subsequent loss ratios to standard premium.

$\tau = -0.03$.

H_0: The correlation between the experience modifications and the subsequent loss ratios to standard premium is zero.

H_1: The correlation between the experience modifications and the subsequent loss ratios to standard premium is not zero.

What is the probability-value of this test?

[Solution: $\text{Var}[\tau] = \frac{2\ (2n+5)}{9n\ (n-1)} = \frac{2\ (405)}{(9)\ (200)\ (199)} = 0.002261$.

Using the Normal approximation, $Z = -0.03 / \sqrt{0.002261} = -0.631$.

For this two-sided test, the probability-value is: $2\ \Phi[-0.63] = 53\%$.

Comment: Since τ is close to zero, by the Meyers-Dorweiler criterion, the experience rating plan is doing a good job of correcting for risk differences.]

---

[54] In practical applications, one would use an exact statistical table for sample sizes of 10 or less.
For example, for a one-sided test with n = 7, one would reject at 5% for $\tau > 11/21$, and reject at 1% for $t > 15/21$.

Geometrically Declining Weights:[55]

Parallel to one of the examples in Mahler, pure premium for a class are projected based on the formula:[56]

E = Z X + (1 - Z) P, where
X = the most recent accident year's pure premium
P = the prior estimate of the most recent accident year
Z = the credibility assigned to the most recent accident year

Exercise: Assume no delay in obtaining data and Z = 20%.
What is the weight given to accident year 2007 data in the estimate of accident year 2009?
[Solution: $P_{2009} = Z X_{2008} + (1 - Z) P_{2008} = Z X_{2008} + (1 - Z) \{Z X_{2007} + (1 - Z) P_{2007}\} =$
$Z X_{2008} + (1 - Z) Z X_{2007} + (1 - Z)^2 P_{2007} = 0.2 X_{2008} + 0.16 X_{2007} + 0.64 P_{2007}$.
The weight given to AY 2007 data is 16%.
Comment: The weight given to AY 2008 data is 20%.
The remaining weight of 64% is given to the prior estimate of the 2007 pure premium; this estimate was based in turn on data from years prior to 2007.]

In general, let the credibility be Z for the latest experience.
If we forecast for year t, and there is no delay, then the weight given to each past year of data is:[57]

Year	Weight
t-1	Z
t-2	$Z (1 - Z)$
t-3	$Z (1 - Z)^2$
t-4	$Z (1 - Z)^3$
t-n	$Z (1 - Z)^{n-1}$

As $Z \to 100\%$, we give full weight to the most recent experience and no weight to older experience. As $Z \to 0\%$, the weights for each year become similar and very small.

---

[55] See page 255 of Mahler. See 9, 11/05, Q.2.
[56] The most recent experience has been developed to ultimate, and has been adjusted for trend and any other changes. The prior estimate has been been adjusted for trend and any other changes. Such complications do not occur with the baseball data.
[57] These weights are from a Geometric Distribution. The weight given to year t-n is f(n-1).
Using the notation in Loss Models, $\beta/(1 + \beta) = 1 - Z$, or $\beta = (1 - Z)/Z = 1/Z - 1$.
This is an example of (single) exponential smoothing.

If risk parameters shift at a faster rate, then all of the past years become a worse predictor of the future. However, more recent experience becomes relatively more useful than older experience to predict the future. For example, as a predictor of 2009, 2007 data is more affected by an increase in the rate of shifting than is 2008 data. Since all of the weight is being applied to some past year of data, the weight to the most recent year of data increases.

Therefore, if the risk parameters shift at a faster rate, then Z increases.
If instead the risk parameters shift at a slower rate, then Z decreases.

Geometrically Declining Weights with Delay:

This form of estimator is similar to pure premium ratemaking, where the credibility weighted pure premium is: Z (the indicated pure premium) + (1 - Z) (the underlying pure premium).
In insurance applications we usually have a delay in getting information.

Exercise: Assume a delay in obtaining data. For example, we have year 2007 data available to predict year 2009, but do not have 2008 data available at that time. Z = 20%. What is the weight given to accident year 2005 losses in the estimate of accident year 2009 losses?
[Solution: $P_{2009} = Z X_{2007} + (1 - Z) P_{2008} = Z X_{2007} + (1 - Z) \{Z X_{2006} + (1 - Z) P_{2007}\} =$
$Z X_{2007} + (1 - Z) Z X_{2006} + (1 - Z)^2 P_{2007} =$
$Z X_{2007} + (1 - Z) Z X_{2006} + (1 - Z)^2 \{Z X_{2005} + (1 - Z) P_{2006}\} =$
$Z X_{2007} + (1 - Z) Z X_{2006} + (1 - Z)^2 Z X_{2005} + (1 - Z)^3 P_{2006}\}.$
The weight given to 2005 losses is: $(1 - Z)^2 Z = (0.8^2)(0.2) = 12.8\%$.]

Least Squares Credibilities:[58]

The least squares credibilities minimize the expected squared error between the estimate and the observation.[59] [60] The least squares credibilities depend on the years used in the estimator as well as the assumed covariance structure.[61] Table 16 shows the resulting credibilities for the covariance structure underlying the years of baseball data, with no delay.[62]

Portion of Table 16 in the Paper

Number of Years	Years Between Data and Estimate				
of Data Used	1	2	3	4	5
1	66.0%	-	-	-	-
2	57.7%	12.6%	-	-	-
3	56.1%	4.8%	13.5%	-	-
4	55.6%	4.6%	11.5%	3.5%	-
5	55.7%	5.1%	11.7%	6.0	-4.4%

Let us interpret this table. Let us assume we are trying to predict 1960.

If we use 1959, Z = 66.0%.
We give the remaining weight of 34.0% to the overall mean relativity of 1.
If instead we use 1958 and 1959, then we weight 1958 12.6% and weight 1959 57.7%.
We give the remaining weight of 29.7% to the overall mean relativity of 1.

If we use 1957, 1958 and 1959, then we weight 1957 13.5%, 1958 4.8% and 1959 56.1%.
We give the remaining weight of 25.6% to the overall mean relativity of 1.
If we use 1956, 1957, 1958 and 1959, then we weight 1956 3.5%, 1957 11.5%, 1958 4.6%, and 1959 55.6%. We give the remaining weight of 24.8% to the overall mean relativity of 1.

We notice that using two years of data, due to shifting risk parameters over time, the more recent year 1959 is given more weight than the more distant year 1958. This follows from the fact that 1959 is more closely correlated with 1960 than is 1958.

The pattern for more years of data gets more complicated. Some of that is due to the specific values for the covariances used here in the paper.[63]

---

[58] See Section 11 of the paper.
[59] I subsequently show how to solve the linear equations in order to solve for the least squares credibilities.
[60] For the given form of linear estimator. So for example, we would specify in advance that we using a linear combination of years 1, 2 and 3 and the overall mean, in order to estimate year 4.
[61] Unlike in Table 9 in the paper, we allow different years of data to be given different weight.
Unlike in Table 9 in the paper, here we work with an assumed covariance structure based on the baseball data, rather than working directly with the baseball data.
[62] I will subsequently discuss the linear equations that are solved for the least squares credibilities.
[63] Subsequently, I show a similar example with a more regular pattern of credibilities.

For three years of data, 1959 is given by far the most weight of 56.1%, but 1957 is given weight of 13.5%, which is more than the 4.8% given to year 1958. This is due to an "edge effect". 1957 is more closely correlated with 1956 and earlier years than is 1958. By giving somewhat more weight to 1957, we in some sense capture some information about years 1956 and prior. Thus we end up getting a better estimate of 1960.

For five years of data, one of the weights is negative. This can happen; there is nothing in the mathematics to prevent it. In some cases, giving negative weight to one year allows one to give more weight to another year and reduce the expected squared error. If one desired, one could constrain each of the weights to be at least zero and no more than one, as one would want in the case of items labeled credibilities.

Table 19 in the paper compares the mean squared errors of different situations.[64]

Portion of Table 19 in the Paper, Using the Credibilities from Table 16

Number of Years of Data Used	Mean Squared Error (0.0001)
1	52
2	51
3	49
4	48
5	48

We note that for example, using 2 years of data is a special case of the using three years of data with one of the credibilities constrained to be zero. Thus as we use more years of data, with varying credibilities by year, the minimum expected squared error declines.[65]

With varying credibilities by year, using more years of data leads to a smaller mean squared error.[66]
Given the number of years of data to be used, we solve for the least squares credibilities, with separate credibilities assigned to each year. Using the most recent two years of data is the same as using three years and setting Z = 0 for the most distant year. We can do at least as well and usually better if we solve for the best credibilities when we use three years of data, rather than setting one of them equal to zero.[67]

When using varying weights by year, including more years of data usually decreases the minimum expected mean squared error, although eventually it stays the same.

---

[64] Subsequently, I will discuss how to calculate the expected mean squared error.
[65] The minimum mean squared error for using 5 years of data is slightly less than that for using 4 years of data, even though in the table they round to the same value. For this particular example, after about 6 years of data one reaches a point of extremely small improvement from using more years of data.
[66] See also page 260 of the paper by Mahler
[67] This is similar to the idea that the loglikelihood for the maximum likelihood Gamma Distribution must be at least as good as the loglikelihood for the maximum likelihood Exponential Distribution, since the Exponential is a special case of the Gamma with $\alpha = 1$.

Rather than separate credibilities by year, instead one could give each year the same weight.

Portion of Table 17 in the Paper

Number of Years of Data Used	Credibility	Z/N
1	66.0%	66.0%
2	70.3%	35.2%
3	72.9%	24.3%
4	73.6%	18.4%
5	72.2%	14.4%

Thus for example, if estimating 1960 using three years data with equal weights, we would give each of 1957, 1958, 1959 weight 24.3%.

Table 19 in the paper compares the mean squared errors of these different situations.[68]

Portion of Table 19 in the Paper

	Mean Squared Error (0.0001)	
Number of Years of Data Used	Differing Credibilities	Equal Weights
1	52	52
2	51	54
3	49	55
4	48	57
5	48	60

Using one year of data, the two cases are identical. Using two or more years of data, having the weights constrained to be equal is a special case of varying weights, and thus can not do as well.

Thus when we use equal weights, the minimum expected squared error is greater than or equal to that using weights that are not necessarily equal. For example, for two years of data 54 > 51.

With equal weights, using more years of data is not a special case of using fewer years of data. Thus the mean squared errors do not necessarily decrease as we increase the number of years used. In fact, **due to shifting risk parameters, when using equal weights, eventually including more years of data increases the minimum expected mean squared error**.[69] In fact, in this case, two years with equal weights does worse than using one year of data.

---

[68] Subsequently, I will discuss how to calculate the mean squared error.

[69] At page 245 of the syllabus reading, discussing the mean squared error criterion when applying equal weights to each year of data: "The results of applying the first criterion are shown in Table 6. Based on most actuarial uses of credibility, an actuary would expect the optimal credibilities to increase as more years of data are used. In this example they do not. In fact, using more than one or two years of data does an inferior job according to this criterion. This result is to be expected, since the parameters shift substantially over time. Thus the use of older data (with equal weight) eventually leads to a worse estimate."

Table 18 in the paper shows credibilities with no weight given to the overall mean, so that the credibilities are constrained to add to one.[70]

Portion of Table 18 in the Paper

Number of Years	Years Between Data and Estimate				
of Data Used	1	2	3	4	5
1	100.0%	-	-	-	-
2	72.6%	27.4%	-	-	-
3	66.1%	10.3%	23.6%	-	-
4	63.5%	9.1%	16.0%	11.4%	-
5	63.1%	8.7%	15.8%	9.5%	2.9%

For example, if using 1957, 1958, and 1959 to estimate 1960, we would give 1957 weight 23.6%, 1958 weight 10.3%, and 1959 weight 66.1%.

Table 19 in the paper also compares the mean squared errors depending on whether there is weight given to the overall mean or not.

Portion of Table 19 in the Paper

	Mean Squared Error (0.0001)	
Number of Years of Data Used	Weight to Overall Mean	No Weight to Overall Mean
1	52	63
2	51	58
3	49	54
4	48	52
5	48	52

Again using fewer years of data is a special case of using more years of data. Thus the minimum mean squared errors decline as more years of data are used, with very limited improvement eventually. Having the weight to the overall mean constrained to be zero is a special case of using the optimal weight on the overall mean, so the mean squared error is at least as big, For example, for two years, $58 > 51$.

[70] These are calculated using equations 11.6 and 11.7, which you are extremely unlikely to be asked about.

Here is a somewhat different set of least squares credibilities, based on modeling the same baseball data via Markov Chains.[71] We allow each year to have a different credibility, give the remaining weight to the overall mean, and have no delay in getting data; thus these credibilities are similar to those shown in Table 16 in the syllabus reading.

Number of Years	Years Between Data and Estimate				
of Data Used	1	2	3	4	5
1	67.0%	-	-	-	-
2	55.1%	17.7%	-	-	-
3	54.3%	15.0%	4.9%	-	-
4	54.2%	14.8%	4.2%	1.4%	-
5	54.2%	14.8%	4.1%	1.2%	0.4%

This model has smoothed out the peculiarities of the covariances of the baseball data that are due to random fluctuation. Thus we see a much more regular pattern of credibilities. More distant years get less credibility than more recent years, declining in a nice pattern. Due to the high rate at which parameters shift in the baseball data, the credibilities for distant years get small quickly. The sum of the credibilities approaches 74.7%.[72] Unlike Table 16, there are no negative credibilities.

[71] Taken from Table 7, in "A Markov Chain Model of Shifting Risk Parameters," by Howard C. Mahler, PCAS 1997, not on the syllabus. In terms of number of games lost, the covariances between years can be approximated by: $Cov[X_i, X_j] = (170)(0.818^{|i-j|}) + 37\ \delta_{ij}$, where δ_{ij} is zero if $i \neq j$ and one if $i = j$.

[72] With shifting risk parameters, the limit of the sum of the credibilities as N approaches infinity is less than 1. The faster the rate of shifting, the smaller is this limit.
For 10 years of data, the least squares credibilities are: 54.2%, 14.8%, 4.1%, 1.2%, 0.3%, 0.1%, 0, 0, 0, 0.

An Example of Solving for the Least Squares Credibility:[73]

The paper uses the following notation:
τ^2 = between variance.
C(k) = covariance for data of the same risk, k years apart = "within covariance"
C(0) = "within variance".

For a data set, you are given: $\tau^2 = 8$, C(0) = 50, C(1) = 20, C(2) = 15, and C(3) = 10.[74]

For two different years: $Cov[X_i, X_j] = \tau^2 + C(|i - j|)$.

For example, $Cov[X_1, X_4] = \tau^2 + C(3) = 8 + 10 = 18$.

For a single year of data: $Cov[X_i, X_i] = Var[X_i] = \tau^2 + C(0) = 8 + 50 = 58$.

Thus the covariance matrix is:

Year 1	58	28	23	18
Year 2	28	58	28	23
Year 3	23	28	58	28
Year 4	18	23	28	58

Use data from year 3 to predict year 4.
Give weight Z to the relativity for year 3, and weight 1 - Z to the over mean relativity of 1.
Then $Z = \frac{Cov[X_3, X_4]}{Var[X_3]} = 28/58 = 48.3\%$.

Instead let us use data from years 2 and 3 to predict year 4.

In other words, we will give weight Z_2 to the relativity for year 2, weight Z_3 to the relativity for year 3, and weight $1 - (Z_2 + Z_3)$ to the overall mean relativity of 1.

Equations 11.3 are the linear equations for the least squares credibility:[75]
$\sum_{j=1}^{N} Z_j \, Cov[X_i, X_j] = Cov[X_i, X_{N+\Delta}]$, where we are predicting year $N + \Delta$, using years 1 to N.

---

[73] Based on recent exams, this is unlikely to be asked.
[74] These are illustrative values. Note that the variance of a single year is more than the covariance between two different years. Also, the covariance between years further apart is less than between years that are closer together. This is the pattern we get with shifting risk parameters over time.
[75] These are the Normal Equations for credibility; see equations 20.25 & 20.26 in Loss Models, not on the syllabus. Note that if all of the covariances are multiplied by the same constant, the credibilities remain the same.

Here N = 2 and $\Delta = 1$, and we get two linear equations in two unknowns:[76]
$58Z_2 + 28Z_3 = 23.$
$28Z_2 + 58Z_3 = 28.$

Solving: $Z_2 = 55/258 = 21.3\%$, and $Z_3 = 49/129 = 38.0\%$. Thus we give weight 21.3% to year 2, weight 38.0% to year 3, and the remaining weight of 40.7% to the overall mean relativity of 1. Due to shifting risk parameters, Year 2 is less correlated with year 4 than is year 3. Year 3 is more useful for predicting year 4 than is year 2; year 3 is given more weight.

Exercise: Assume that a team has a losing percentage of 0.453 in year 2, and a losing percentage of 0.411 in year 3. Predict the losing percentage for this team in year 4.
[Solution: (21.3%)(0.453) + (38.0%)(0.411) + (40.7%)(0.500) = 0.456.
Comment: One could divide everything by the overall mean losing percentage of 0.5 in order to put everything in terms of relativities with respect to average.]

Let us instead assume we give weight Z to the average of the relativities for years 2 and 3, and weight 1 - Z to the overall mean relativity of 1.[77]

$Cov[(X_2 + X_3)/2, X_4] = \{Cov[X_2, X_4] + Cov[X_3, X_4]\} / 2 = (23 + 28)/2 = 25.5.$
$Var[(X_2 + X_3)/2] = \{Var[X_2] + Var[X_3] + 2\, Cov[X_2, X_3]\} / 2^2 = \{58 + 58 + (2)(28)\} / 4 = 43.$

Thus the linear equation for Z, analogous to equation 11.3 is:
$43\, Z = 25.5. \Rightarrow Z = 25.5 / 43 = 59.3\%.$

Exercise: Assume that a team has a losing percentage of 0.453 in year 2, and a losing percentage of 0.411 in year 3. Predict the losing percentage for this team in year 4.
[Solution: (59.3%) (0.453 + 0.411)/2 + (1 - 59.3%) (0.500) = 0.460.
Comment: Differs slightly from the previous estimate using separate credibilities by year.]

The previous separate credibilities were: $Z_2 = 21.3\%$, and $Z_3 = 38.0\%$.
They sum to 59.3%, the same as the single Z applied to the average.[78]

Note that using a single Z, we constrained the weights given to years two and three to be equal. Thus this is a special case of solving for the least squares Z_2 and Z_3. Thus the expected squared error using the best single Z must be greater than or equal to that from using the best Z_2 and Z_3.

---

[76] The coefficients on the lefthand side are the second and third rows and columns of the covariance matrix. The values on the righthand side are the second and third rows of column four, since we are predicting Year 4.
[77] In this paper, every year of data has the same volume of data, so we are giving weight Z/2 to each year.
[78] One can show algebraically, that this will be true in general when using only two years of data.

Repeating the covariance matrix:

Year 1	58	28	23	18
Year 2	28	58	28	23
Year 3	23	28	58	28
Year 4	18	23	28	58

Let us instead use data from years 1 and 2 to predict year 4. There is now a one year delay.
We give weight Z_1 to the relativity for year 1, weight Z_2 to the relativity for year 2,
and weight 1 - (Z_1 + Z_2) to the overall mean relativity of 1.

Equations 11.3 are the linear equations for the least squares credibility:
$\sum_{j=1}^{N} Z_j \text{ Cov}[X_i , X_j] = \text{Cov}[X_i, X_{N+\Delta}]$, where we are predicting year N + Δ, using years 1 to N.

Exercise: Write down the linear equations for the least squares credibilities.
[Solution: Here N = 2 and Δ = 2, and we get two linear equations in two unknowns:
$58Z_1 + 28Z_2 = 18$.
$28Z_1 + 58Z_2 = 23$.
Comment: The coefficients on the lefthand side are the first two rows and columns of the covariance matrix. The righthand side is the first two rows of column four, since we are predicting Year 4.]

Solving these linear equations: Z_1 = 20/129 = 15.5%, and Z_2 = 83/258 = 32.2%.
Thus we give weight 15.5% to the relativity for year 1, 32.2% weight to the relativity for year 2, and the remaining weight of 52.3% to the overall mean relativity of 1.

This compares to the previous case with no delay when Z_2 = 21.3% and Z_3 = 38.0%.
With no delay, we give more weight to the data: 21.3% > 15.5%, and 38.0% > 32.2%.

Due to shifting risk parameters, more distant years are less useful for predicting the future.
Thus with a delay the credibilities are smaller. The bigger the delay, the smaller the credibilities.

Interestingly, with the delay the weight assigned to year 2 is larger than it was without the delay. That is because least squares credibility is a relative concept. The weight assigned to a year of data depends on how good an estimator is each of the other years being used.

Year 3 is a better estimator of year 4 for than is year 2; thus when using year 3 and year 2 this tends to decrease the weight given to year 2. In contrast, year 1 is a worse estimator of year 4 than is year 2; thus when using year 1 and year 2 this tends to increase the weight given to year 2.

Let us instead assume we give weight Z to the average of the relativities for years 1 and 2 and weight 1 - Z to the overall mean relativity of 1.[79]

$\text{Cov}[(X_1 + X_2)/2, X_4] = \{\text{Cov}[X_1, X_4] + \text{Cov}[X_2, X_4]\} / 2 = (18 + 23)/2 = 20.5.$

$\text{Var}[(X_1 + X_2)/2] = \{\text{Var}[X_1] + \text{Var}[X_2] + 2\,\text{Cov}[X_1, X_2]\} / 2^2 = \{58 + 58 + (2)(28)\} / 4 = 43.$

Thus the linear equation for Z, analogous to equation 11.3 is:
$43\,Z = 20.5. \Rightarrow Z = 20.5 / 43 = 47.7\%.$

The previous separate credibilities were: $Z_1 = 15.5\%$, and $Z_2 = 32.2\%$.
They sum to 47.7%, the same as the single Z applied to the average.

Due to shifting risk parameters, Z = 47.7% when using years 1 and 2 to predict year 4 is smaller than Z = 59.3% when instead using years 2 and 3 to predict year 4.

Repeating the covariance matrix:

Year 1	58	28	23	18
Year 2	28	58	28	23
Year 3	23	28	58	28
Year 4	18	23	28	58

Exercise: We will use years 1, 2, and 3 to predict year 4.
Write down the linear equations for the least squares credibilities.
[Solution: We get three linear equations in three unknowns:
$58Z_1 + 28Z_2 + 23Z_3 = 18.$
$28Z_1 + 58Z_2 + 28Z_3 = 23.$
$23Z_1 + 28Z_2 + 58Z_3 = 28.$
Comment: The coefficients on the lefthand side are the first 3 rows and columns of the covariance matrix. The righthand side is the first 3 rows of column four, since we are predicting Year 4.]

Solving these linear equations:
$Z_1 = 170/2191 = 7.8\%$, $Z_2 = 115/626 = 18.4\%$, and $Z_3 = 796/2191 = 36.3\%$.[80]

Thus we give weight 7.8% to the relativity for year 1, 18.4% weight to the relativity for year 2, 36.3% weight to the relativity for year 3, and the remaining weight of 37.5% to the overall mean relativity of 1.

[79] In this paper, every year of data has the same volume of data, so we are giving weight Z/2 to each year.
[80] You will not be asked to solve three linear equations on your exam.

Let us instead assume we give weight Z to the average of the relativities for years 1, 2, and 3 and weight 1 - Z to the overall mean relativity of 1.[81]

$\text{Cov}[(X_1 + X_2 + X_3)/3, X_4] = \{\text{Cov}[X_1, X_4] + \text{Cov}[X_2, X_4) + \text{Cov}[X_3, X_4]\} / 3 =$
$(18 + 23 + 28)/3 = 23.$
$\text{Var}[(X_1 + X_2 + X_3)/3] =$
$\{\text{Var}[X_1] + \text{Var}[X_2] + \text{Var}[X_3] + 2\,\text{Cov}[X_1, X_2] + 2\,\text{Cov}[X_1, X_3] + 2\,\text{Cov}[X_2, X_3]\} / 3^2 =$
$\{58 + 58 + 58 + (2)(28) + (2)(23) + (2)(28)\} / 9 = 332/9.$

Thus the linear equation for Z, analogous to equation 11.3 is:
$(332/9)\, Z = 23. \Rightarrow Z = 207 / 332 = 62.3\%.$

Alternately, we could use equation 11.4:[82]

$$Z = N \frac{N \tau^2 + \sum_{l=1}^{N} C(N+\Delta-i)}{N^2 \tau^2 + \sum_{j=1}^{N}\sum_{l=1}^{N} C(|i-j|)}.$$

With $N = 3$ and $\Delta = 1$:

$$Z = (3) \frac{(3)(8) + C(3) + C(2) + C(1)}{(9)(8) + C(0) + C(1) + C(2) + C(1) + C(0) + C(1) + C(2) + C(1) + C(0)} =$$

$$(3) \frac{24 + 10 + 15 + 20}{72 + 50 + 20 + 15 + 20 + 50 + 20 + 15 + 20 + 50}$$

$= (3)(69) / 332 = 207 / 332 = 62.3\%.$

The separate credibilities were: $Z_1 = 7.8\%$, $Z_2 = 18.4\%$, and $Z_3 = 36.3\%$.
These sum to 62.5%, close to but different than the single credibility of 62.3%.[83]
In general, we expect them to be similar but not identical.

[81] In this paper, every year of data has the same volume of data, so we are giving weight Z/3 to each year.
[82] I would not memorize this equation.
[83] The sum is: 391/626 = 0.62460, while the single Z = 207/332 = 0.62349.

Expected Squared Errors:[84]

The least squares credibilities minimize the expected squared error between the estimate and the observation.[85] The expected squared error is given by equation 11.2:[86]

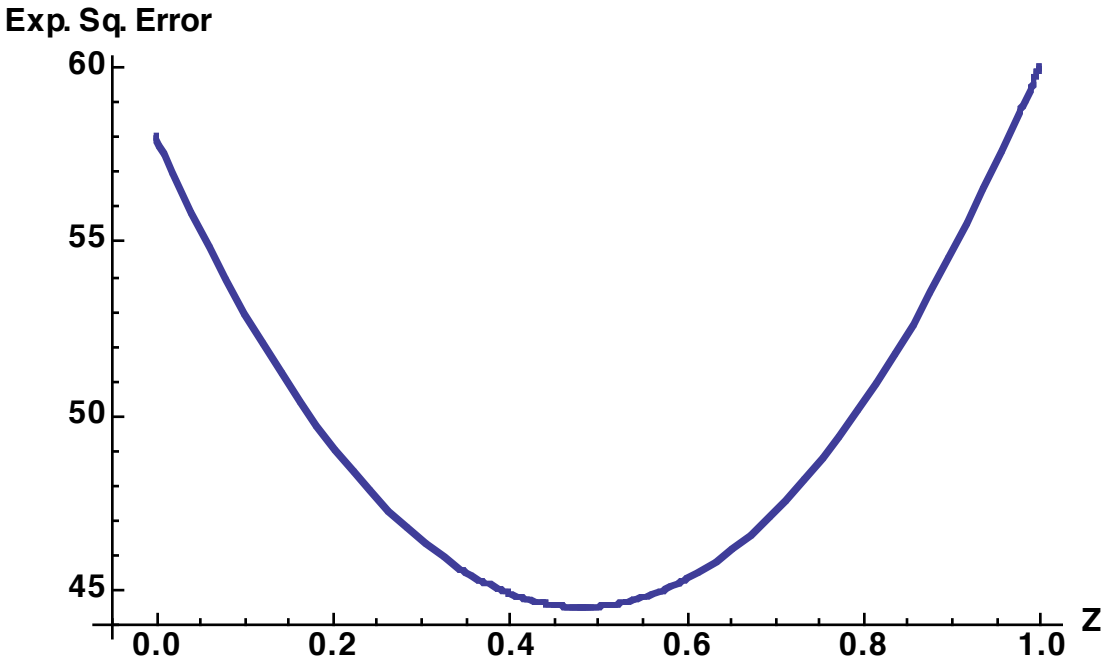
$$V(\vec{Z}) = \sum_{i=1}^{N}\sum_{j=1}^{N} Z_i Z_j \{\tau^2 + C(|i-j|)\} - 2\sum_{i=1}^{N} Z_i \{\tau^2 + C(N+\Delta-i)\} + \tau^2 + C(0).$$

For the previous example, we had: $\tau^2 = 8$, $C(0) = 50$, $C(1) = 20$, $C(2) = 15$, and $C(3) = 10$.

If we are using year 3 to estimate year 4, then $N = 1$ and $\Delta = 1$, and:
$V(Z) = Z^2\{\tau^2 + C(0)\} - 2Z\{\tau^2 + C(1)\} + \tau^2 + C(0) = 58Z^2 - 56Z + 58$.

Setting the derivative of V(Z) equal to zero in order to minimize the expected squared error:
$(2)(58)Z - 56 = 0. \Rightarrow Z = 28/58 = 48.3\%$, matching a previous result.

At $Z = 48.3\%$, the expected squared error is: $(58)(0.483^2) - (56)(0.483) + 58 = 44.48$.
Here is a graph of the expected squared error as a function of Z, a parabola:



[84] Based on recent exams, this is unlikely to be asked about in any detail.
[85] For the given form of linear estimator. So for example, we would specify in advance that we using a linear combination of years 1, 2 and 3 and the overall mean, in order to estimate year 4. We would also need to specify whether we will give each year the same weight or instead apply separate credibilities to each year of data.
[86] I would not memorize this equation.

Exercise: We are instead using year 2 to estimate year 4.
Determine the expected squared error as a function of Z, and find its minimum.
[Solution: $V(Z) = Z^2\{\tau^2 + C(0)\} - 2Z\{\tau^2 + C(2)\} + \tau^2 + C(0) = 58Z^2 - 46Z + 58$.
Setting the derivative of V(Z) equal to zero in order to minimize the expected squared error:
$(2)(58)Z - 46 = 0. \Rightarrow Z = 46/116 = 39.7\%$.
At Z = 39.7%, the expected squared error is: $(58)(0.397^2) - (46)(0.397) + 58 = 48.88$.]

Note that when using year 2 rather than year 3, the credibility is smaller while the expected mean squared error is larger. Specifically, the minimum squared error is now 48.88 compared to 44.48.
Due to shifting parameters over time, year 2 is a worse predictor of year 4 than is year 3, and thus the minimum expected squared error is greater if we use year 2 rather than year 3.

Now let us use data from years 2 and 3 to predict year 4. Then using equation 11.2:
V(Z) =
$Z_2^2\{\tau^2 + C(0)\} + 2Z_2Z_3\{\tau^2 + C(1)\} + Z_3^2\{\tau^2 + C(0)\} - 2Z_2\{\tau^2 + C(2)\} - 2Z_3\{\tau^2 + C(1)\} + \tau^2 + C(0) =$
$58Z_2^2 + 56Z_2Z_3 + 58Z_3^2 - 46Z_2 - 56Z_3 + 58$.

It turns out that Equation 11.2 can be rewritten in matrix form,
Mean Squared Error = $V(Z) = Z^T C Z$.
C is the matrix of covariances for the years of data.
Z is the (column) vector with credibilities in the years used to estimate, -1 in the year being estimated, and zeros in any other years. Z^T is the transpose of Z.

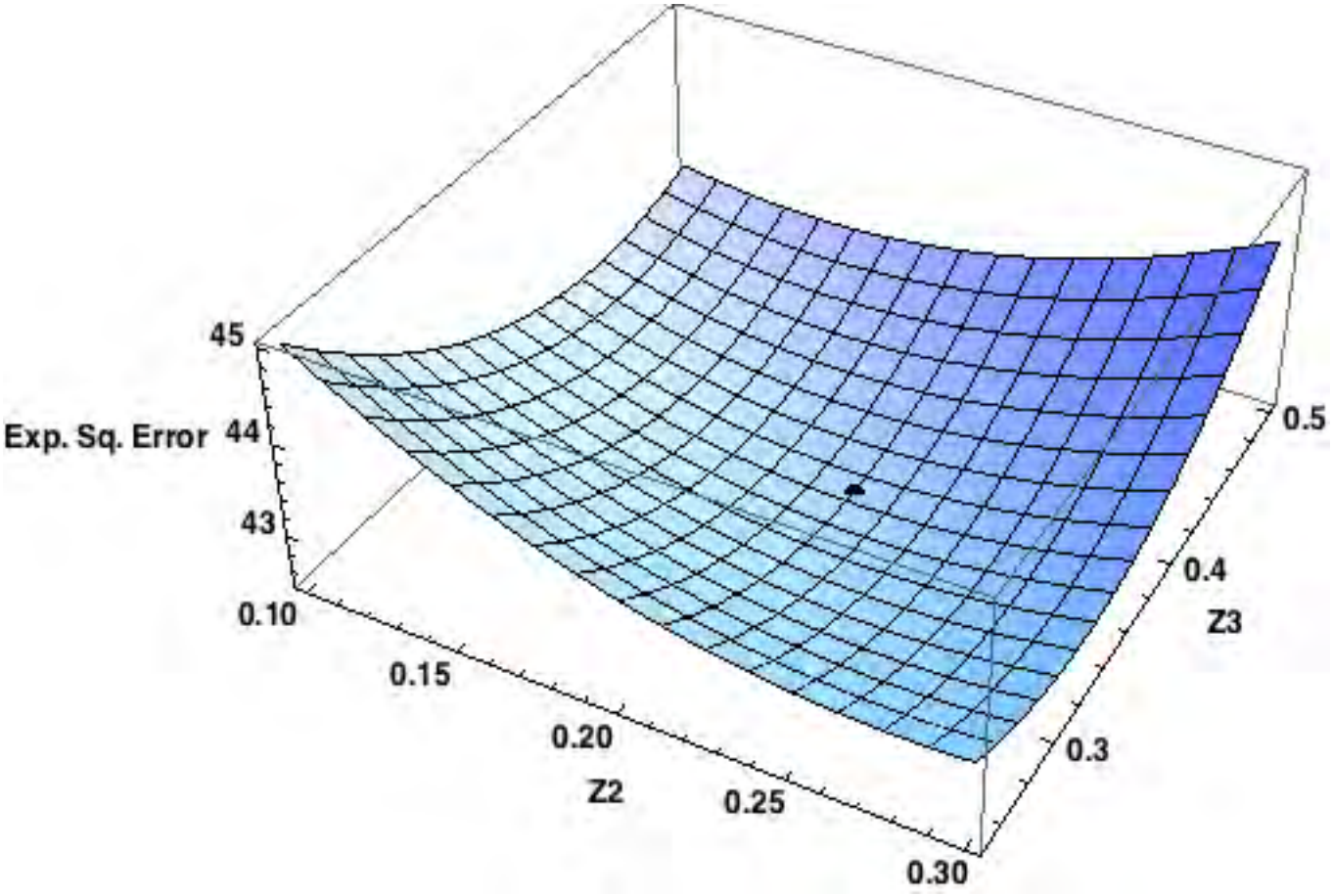
In this case: $V(Z) = (Z_2, Z_3, -1) \begin{pmatrix} 58 & 28 & 23 \\ 28 & 58 & 28 \\ 23 & 28 & 58 \end{pmatrix} \begin{pmatrix} Z_2 \\ Z_3 \\ -1 \end{pmatrix} = (Z_2, Z_3, -1) \begin{pmatrix} 58Z_2 + 28Z_3 - 23 \\ 28Z_2 + 58Z_3 - 28 \\ 23Z_2 + 28Z_3 - 58 \end{pmatrix}$

$= 58Z_2^2 + 56Z_2Z_3 + 58Z_3^2 - 46Z_2 - 56Z_3 + 58$.

Setting the partial derivative with respect to Z_2 equal to zero:
$116 Z_2 + 56 Z_3 = 46$.
Setting the partial derivative with respect to Z_3 equal to zero:
$56 Z_2 + 116 Z_3 = 56$.

These are equivalent to the two equations we got before for this situation, and the solution is: $Z_2 = 21.3\%$ and $Z_3 = 38.0\%$. For these least squares credibilities, the expected squared error is: $(58)(0.213^2) + (56)(0.213)(0.380) + (58)(0.380^2) - (46)(0.213) - (56)(0.380) + 58 = 42.46$.

Here is a graph of the expected squared error as a function of Z_1 and Z_2, with the minimum shown as a dot, (0.213, 0.380, 42.46):



"The optimal credibilities are uniquely determined given the covariance structure. However, there are many other sets of credibilities which produce expected squared errors very close to minimal."[87]

Note that using just year 3 is a special case of using years 2 and 3, with Z_2 constrained to be zero. Thus the minimum expected squared error using just year 3 has to be greater than or equal that from using both years 2 and 3. In this example, 44.48 is greater than 42.46.

As shown previously, when giving the same weight to years 2 and 3, the least squares credibility is Z = 59.3%. In other words, when constrained to be equal, $Z_2 = Z_3 = 59.3\%/2$. Then, V(Z) = $(58)(0.2965^2) + (56)(0.2965^2) + (58)(0.2965^2) - (46)(0.2965) - (56)(0.2965) + 58 = 42.88$.
As it has to be, the minimum expected squared error when the weights are constrained to be equal is greater than that when the weights are allowed to be different; 42.88 > 42.46.

[87] See page 265 of "An Example of Credibility and Shifting Risk Parameters," by Howard C. Mahler.

Half-Life:[88]

One can model shifting risks parameters via a covariance structure between years of data that is of the form: $Cov[X_i, X_j] = a \rho^{|i-j|} + b \delta_{ij}$, where δ_{ij} is zero if $i \neq j$ and one if $i=j$.[89]

$\rho < 1$ measures the speed at which risk parameters shift. The correlations between years decline by a factor of ρ when the separation between those years of data increases by a year.

Define the "half-life" as the length of time for the correlations between years to decline by a factor of one-half: $\rho^{\text{half-life}} = 0.5. \Rightarrow \text{half-life} = \ln(0.5) / \ln(\rho)$.
The half-life is a somewhat more intuitive way to quantify the rate at which parameters shift.

Here are some examples, with the approximate values of ρ and the corresponding half-lives:

Example	ρ	Half-life
Baseball Win/Loss Data by Team	0.82	3.5 years
California P.P. Auto Driving Data[90]	0.95	13.5 years
Workers Compensation Classes[91]	0.94	11.2 years
Workers Comp. Experience Rating[92]	0.82	3.5 years

We see that the rate of shifting in the baseball example in the syllabus reading is much faster than that in the first two insurance examples. While this made the baseball data set a good one to use to develop these ideas and illustrate the results, the effect of shifting risk parameters will be significantly smaller in many applications to insurance data.

---

[88] See "A Markov Chain Model of Shifting Risk Parameters," by Howard C. Mahler, PCAS 1997, not on the syllabus.
[89] For $\rho < 1$, this models shifting risk parameters over time. This is an approximation to the form in "A Markov Chain Model of Shifting Risk Parameters," by Howard C. Mahler, PCAS 1997.
[90] With ρ approximately 0.94 for Female Drives and 0.97 for Male drivers. It is not clear whether this difference between males and females is significant or just due to random fluctuations in the data set.
See "The Credibility of a Single Private Passenger Driver", by Howard C. Mahler, PCAS 1991.
[91] Classification relativities for the Manufacturing Industry in Massachusetts, for classes with expected annual losses between $300,000 and $1 million.
See page 535 of "Credibility With Shifting Risk Parameters, Risk Heterogeneity, and Parameter Uncertainty," by Howard C. Mahler, PCAS 1998.
The rate of shifting risk parameters may be more rapid for smaller classes than for larger classes.
In "Workers' Compensation Classification Credibilities", by Howard C. Mahler, Fall 1999 CAS Forum, the equivalent of $\rho = 0.85$ for the very smallest classes and $\rho = 0.99$ for the very largest classes were selected.
[92] See page 589 of "Credibility With Shifting Risk Parameters, Risk Heterogeneity, and Parameter Uncertainty," by Howard C. Mahler, PCAS 1998. Based on experience rating data for Massachusetts. The selected values differ somewhat between primary and excess and by size of insured.

Conclusions:

Know well the three paragraphs on page 280, the conclusions of the paper:

When shifting parameters over time is an important phenomenon, older years of data should be given substantially less credibility than more recent years of data. The more significant this phenomenon, the more important it is to minimize the delay in receiving the data that is to be used to make the prediction.

Three different criteria were examined that can be used to select the optimal credibility: least squares, limited fluctuation, and Meyers/Dorweiler. In applications, one or more of these three criteria should be useful. While the first two criteria are closely related, the third criterion can give substantially different results than the others.

Generally the mean squared error can be written as a second order polynomial in the credibilities. The coefficients of this polynomial can be written in terms of the covariance structure of the data. This in turn **allows one to obtain linear equation(s) which can be solved for the least squares credibilities in terms of the covariance structure**.

Further Reading and Resources:[93]

The NEAS webpage has an illustrative worksheet showing Mahler's baseball analysis.
Go to the regression analysis VEE course discussion board, and find the thread for student project template for sports won-loss records: http://tempforum.neas-seminars.com/Forum177.aspx

"A Markov Chain Model of Shifting Risk Parameters", by Howard C. Mahler, PCAS 1997.
www.casact.org/pubs/proceed/proceed97/97581.pdf
This 1997 paper expands on "An Example of Credibility and Shifting Risk Parameters."
Pages 629-639 revisit the baseball example.

"Credibility With Shifting Risk Parameters, Risk Heterogeneity, and Parameter Uncertainty," by Howard C. Mahler, PCAS 1998. www.casact.org/pubs/proceed/proceed98/980455.pdf
This 1998 paper expands on the 1997 paper.
Pages 615-617 briefly discuss the baseball example.

"Workers' Compensation Classification Credibilities", by Howard C. Mahler,
Fall 1999 CAS Forum. www.casact.org/pubs/forum/99fforum/99ff433.pdf
This paper applies the ideas of the 1998 paper to a practical example of classification ratemaking.

Stuart A. Klugman, "Credibility with Shifting Risk Parameters," SOA Study Note, 2014.

---

[93] Not on the syllabus.

Problems:

1.1. (1 point) All but which of the following are reasons Mahler uses baseball data to study experience rating?
A. Each baseball team plays the same number of games.
B. The won-loss data are accurate, final, and readily available.
C. The set of teams does not change over the period of time studied.
D. Baseball teams win a similar percentage of games over a decade.
E. All of A, B, C, and D are true.

1.2. (1 point) Which of the following did Mahler conclude regarding differences between teams?
1. A team that had been worse than average over one period of time is more likely to be better than average over the subsequent period of time.
2. Observed differences between teams over six decades are greater than could be attributed to chance alone if teams were inherently equal.
3. The fact that one team's loss is another team's win has a material effect on the distribution of losing percentages in the baseball analogy.

1.3. (1 point) Which of the following among Mahler's conclusions regarding changes in the inherent winning potential of the teams over time?
1. For all the teams, a Chi-Square test showed differences over time that are significant at the 1 % level.
2. Significant correlations exist between a team's results in one year and its results in other years less than ten years before or after.
3. A team's experience in recent years is useful in predicting its experience in the upcoming year.

Use the following information for the next two questions:
We are estimating optimal credibility for experience rating of NBA basketball teams under three versions of the draft rule:
(1) Team with poorest record gets the first draft pick.
(2) Team with best record gets the first draft pick.
(3) The order of draft picks is chosen randomly.
Let $Z(i)$ is the credibility under rule i ($i = 1, 2, 3$).

1.4. (1 point) Rank the experience rating credibility under these three rules.

1.5. (1 point) Which of the following are true?
1. $0 \leq Z(1) \leq 1$
2. $0 \leq Z(2) \leq 1$
3. $0 \leq Z(3) \leq 1$

1.6. (1 point) Which of the following statements are true of Mahler's credibility estimators?
1. They are linear combinations of a few simple estimators.
2. They are unbiased for the set of teams as a whole.
3. They are more analogous to schedule rating than to experience rating.

1.7. (3 points) Eleven insureds have the following relative loss ratios in two consecutive years:

1	2	3	4	5	6	7	8	9	10	11
0.90	0.93	0.96	0.98	0.99	1.00	1.01	1.02	1.04	1.07	1.10
0.98	0.93	1.00	0.90	0.98	1.02	0.99	0.99	1.07	1.10	1.04

Based on the least squares criterion, what is the proper credibility for these insureds?

1.8. (1 point) We are designing an experience rating system which weights the class mean with the unweighted mean of the risk's latest N years of data. Which of the following criteria can be used to select optimal values for the credibility Z and the number of years N?
1. Least squared error
2. Small chance of large error
3. Meyers/Dorweiler

1.9. (1 point) Mahler in "An Example of Credibility and Shifting Risk Parameters," concludes that to predict baseball losing percentages, a reasonable method is to use three years of data with $Z_1 = 10\%$, $Z_2 = 10\%$, $Z_3 = 55\%$, and the remaining weight to the grand mean.
A baseball team had the following record:
2005: won 67 games and lost 95 games.
2006: won 61 games and lost 101 games.
2007: won 66 games and lost 96 games.
Using the above method, in 2008, what is the predicted record for this team for its first 88 games?

1.10. (1 point) Which of the following are true about optimal credibility estimators?
1. When one combines the unweighted N-year average with the grand mean, the estimate continues to improve as N increases.
2. The exact values of the optimal credibility weights $\{Z_i\}$ are not critical as long as one is close to the optimal set.
3. The ideal credibility estimator would reduce the mean squared error between the estimated and observed values to zero.

1.11. (1 point) Let a set of Z_i's be the credibility factors that minimize the expected squared error as determined by the covariance structure. Which of the following are true?
1. The expected squared error is a linear function of the Z_i.
2. The optimal Z_i are all nonnegative.
3. It is necessary to distinguish among three sources of variance: variance between risks (τ^2), the process variance excluding the effect of shifting parameters over time (δ^2), and the portion of the process variance due to shifting parameters over time (ζ^2).

1.12. (1 point) The optimal credibility weights for an experience rating plan depend on the variance between risks (τ^2), the process variance excluding the effect of shifting parameters over time (δ^2), and the portion of the within variance due to shifting parameters over time (ζ^2).
A certain plan uses five years of experience and a two step credibility procedure: the risk's own experience gets credibility Z of 50%, divided between five years of experience are weighted 10%, 15%, 20%, 25%, and 30%.
We are updating the credibility weights, based on new estimates of τ^2, δ^2, and ζ^2.
Which of the following statements are correct?
1. As τ^2 increases, the value of Z decreases.
2. As δ^2 increases, the value of Z decreases.
3. As ζ^2 increases, the weight for year 1 (now 10%) decreases.

1.13. (2 points) The optimal credibility weights depend on the variance between risks (τ^2), the process variance excluding the effect of shifting parameters over time (δ^2), and the portion of the within variance due to shifting parameters over time (ζ^2). Briefly explain how each of these three elements differs between class ratemaking and experience rating.

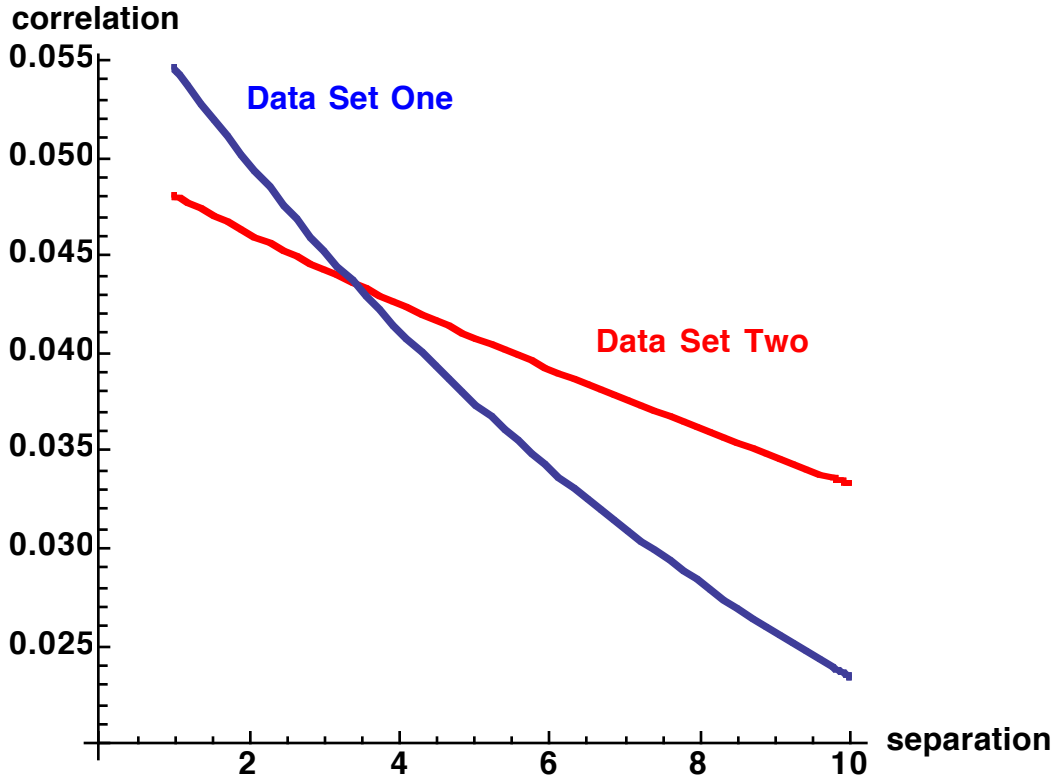
1.14. (3 points) You are analyzing an experience rating plan.
Briefly explain how each of the changes affect the following items:
Between Variance (τ^2), Within Variance ($\delta^2 + \zeta^2$), Effect of Shifting Risk Parameters,
and Credibility (Z).
(a) Change from a no-split experience rating plan to one with a reasonable primary-excess split.
(b) Use 2 years of data instead of 5 in the experience rating plan.
(c) Refine the classification plan.

1.15. (1 point) Which of the following are true of the Meyers/Dorweiler criterion?
1. The criterion assures that debit and credit risks are equally attractive to insurers.
2. As credibility approaches zero, the Kendall t statistic approaches one.
3. An experience rating plan that satisfies the criterion is an acceptable plan.

1.16. (1 point) Using Mahler's terminology,
1. Let $X(\theta, t)$ be the observation for risk θ at time t.
2. Let $\mu(\theta, t)$ be the expected value for risk θ at time t.
3. Let $\mu(\theta)$ be the long-term expected value for risk θ.
4. Let M be the long-term all-risk grand mean.
Which of the above may be directly observed in an insurance pricing situation?

1.17. (1 point) What problems are caused by delays in obtaining data?
1. Delays in obtaining data degrade the performance of experience rating plans.
2. The optimal credibility decreases as the delay increases.
3. If the delay exceeds three years, then for the data set examined in Mahler's "An Example of Credibility and Shifting Risk Parameters," the predictive value of the data is close to zero.

1.18. (1 points) You are looking at two different data sets, each consisting of many years of data. In each case, you calculate correlations between pairs of years of data. You then fit a curve to the correlations as a function of the separation of the years of data. Here is a graph of the results:



What conclusions do you draw and why?

1.19. (1 point) Mahler discusses an estimate of the form F = Z X + (1 - Z) P, where X is the most recent data point, P is the previous estimate, and Z is a selected weight.
Assume that there is no delay in obtaining the data.
If Z = 55%, what weight is given to the data for 2006 in estimating losses for 2009?

1.20. (1 point) A rate indication for 2009 uses weighted experience from 2002 through 2006.
Based on the considerations outlined by Mahler, which of the following statements are true?
1. It is appropriate to assign equal weights to the years of data.
2. It is not appropriate to assign nonzero weight to data from 2002.
3. The traditional weights of 10%, 15%,20%, 25%, 30% perform significantly worse than an optimal set of weights derived using Mahler's techniques.

1.21. (1 point) According to Mahler in "Credibility and Shifting Risk Parameters," which of the following statements are correct?
1. The mean squared error can generally be written as a second order polynomial in the credibilities.
2. The coefficients of this polynomial can be written in terms of the covariance structure of the data.
3. This in turn allows one to obtain quadratic equations which can be solved for the least squares credibilities in terms of the covariance structure.

1.22. (2 points) (For baseball fans) You are updating the study in Mahler's paper using similar baseball data from 1961 to the present.
(a) Mention two complications that would occur that Mahler did not have to deal with.
(b) Would you expect shifting risk parameters to have a bigger effect or smaller effect than in Mahler's study? Why?

1.23. (15 points) In "An Example of Credibility and Shifting Risk Parameters," Mahler uses the following notation:
τ^2 = between variance.
C(k) = covariance for data of the same risk, k years apart = "within covariance"
C(0) = "within variance".
For a data set, you are given $\tau^2 = 5$, C(0) = 50, C(1) = 10, C(2) = 8, C(3) = 6, and C(4) = 4.
One will be using least squares credibility, with the complement of credibility given to the grand mean and varying weights to each year of data.
In each case, determine the optimal credibilities to be assigned to each year of data.
(a) (1 point) Use data for Year 1 to Predict Year 2.
(b) (1 point) Use data for Year 1 to Predict Year 3.
(c) (1 point) Use data for Year 1 to Predict Year 4.
(d) (2 points) Use data for Years 1 and 2 to Predict Year 3.
(e) (2 points) Use data for Years 1 and 2 to Predict Year 4.
(f) (4 points) Use data for Years 1, 2, and 3 to Predict Year 4.
(g) (4 points) Use data for Years 1, 2, and 3 to Predict Year 5.

1.24. (10 points) In the previous question, in parts (d) through (g), instead require that the weight given to each year be the same. Calculate the resulting least squares credibility.

1.25. (2 points)
(a) Define the phenomena of shifting risk parameters.
(b) Pick a line of insurance and give one reason why risk parameters would shift over time for an insured.
Do not discuss something that would likely result in a change in classification or territory.

1.26. (2 points) (For football fans) You are updating the study in Mahler's paper using data from the National Football League from 1961 to the present.
(a) Mention two issues that Mahler's study did not have to deal with.
(b) Would you expect shifting risk parameters to have a bigger effect or smaller effect than in Mahler's study? Why?

1.27. (10 points) In "An Example of Credibility and Shifting Risk Parameters," Mahler uses the following notation:
τ^2 = between variance.
C(k) = covariance for data of the same risk, k years apart = "within covariance"
C(0) = "within variance".
For a data set, you are given $\tau^2 = 10$, C(0) = 30, C(1) = 15, C(2) = 10, C(3) = 6, and C(4) = 3.
One will be using least squares credibility, with the complement of credibility given to the grand mean and varying weights to each year of data.
In each case, determine the optimal credibilities to be assigned to each year of data.
(a) (1 point) Use data for Year 1 to Predict Year 2.
(b) (1 point) Use data for Year 1 to Predict Year 3.
(c) (1 point) Use data for Year 1 to Predict Year 4.
(d) (1 point) Use data for Year 1 to Predict Year 5.
(e) (2 points) Use data for Years 1 and 2 to Predict Year 3.
(f) (2 points) Use data for Years 1 and 2 to Predict Year 4.
(g) (2 points) Use data for Years 1 and 2 to Predict Year 5.

1.28. (10 points) For each of the parts of the previous question, calculate the corresponding minimum expected squared error.

1.29. (1 point) In "An Example of Credibility and Shifting Risk Parameters," one of the techniques used by Mahler is least squares credibility, with the complement of credibility given to the grand mean and varying weights to each year of data.
For an example, Mahler determines the optimal credibilities to be assigned to each year of data and displays them in Table 16.
Which of the following statements is true about these optimal credibilities?
A. They are not negative.
B. More distant years are given less weight than more current years.
C. As more years of data are used, the credibility assigned to the first year of data does not increase.
D. As more years of data are used, the expected squared error does not increase.
E. None of A, B, C, or D.

1.30. (6 points) Use the following information:

- You are using data from years 1 through 5 in order to predict year 6.
- The variance of each year of data is 6.
- The covariance between different years of data is:
  $Cov[X_i, X_j] = 0.9^{|i-j|}$.

In "An Example of Credibility and Shifting Risk Parameters," one of the techniques used by Mahler is least squares credibility, with the complement of credibility given to the grand mean and varying weights to each year of data.
Determine the credibilities to assign to each of the five years of data.
(Use a computer to help you with the computations.)

1.31. (2 points) Three experience rating plans are being compared.
You are trying to evaluate which is optimal.
Each rating plan has been tested on the same five different policies of similar size.
You compare the modification factor for each plan calculated before the policy period to the subsequent experience during the policy period.
The following tables summarize the indicated modifications and policy period experience.

Policy Number	Rating Plan 1 Modification Factor	Rating Plan 2 Modification Factor	Rating Plan 3 Modification Factor		Policy Period Experience
1	0.80	0.87	0.81	0.85	
2	0.90	0.87	0.83	0.85	
3	1.00	1.00	1.00	1.00	
4	1.10	1.03	1.09	1.05	
5	1.20	1.23	1.27		1.25

Which is the preferred plan based on the Meyers/Dorweiler criterion? Why?
Which is the preferred plan based on the least squared error criterion? Why?

1.32. (3 points) You are using N years of data without any delay in order to estimate the next year. The remaining weight will be given to the grand mean.
Allowing the credibilities to differ by year, the following least squares credibilities were determined,
with year 1 being the most recent year.
Also shown is the corresponding minimum mean squared error (0.00001):

Year	N=1	N=2	N=3	N=4	N=5	N=6	N=7	N=8	N=9	N=10
1	3.43%	3.32%	3.23%	3.16%	3.10%	3.05%	3.01%	2.98%	2.95%	2.93%
2		3.11%	3.01%	2.94%	2.87%	2.82%	2.78%	2.74%	2.71%	2.69%
3			2.82%	2.74%	2.67%	2.61%	2.57%	2.53%	2.50%	2.47%
4				2.56%	2.49%	2.43%	2.38%	2.34%	2.30%	2.27%
5					2.33%	2.26%	2.21%	2.16%	2.13%	2.10%
6						2.12%	2.06%	2.01%	1.97%	1.94%
7							1.93%	1.87%	1.83%	1.79%
8								1.75%	1.71%	1.67%
9									1.60%	1.55%
10										1.45%
Total	3.43%	6.43%	8.24%	11.40%	13.46%	15.29%	16.93%	18.38%	19.69%	20.86%
MSE	3835	3832	3829	3826	3824	3822	3821	3820	3819	3818

Note that the values shown in a column may not sum to the total shown due to rounding.

Fully discuss the results shown.

1.33. (2 points)
Compare and contrast the following 3 covariance structures between years of data.

	Y1	Y2	Y3
Year 1	200	200	200
Year 2	200	200	200
Year 3	200	200	200

	Y1	Y2	Y3
Year 1	200	140	140
Year 2	140	200	140
Year 3	140	140	200

	Y1	Y2	Y3
Year 1	200	140	110
Year 2	140	200	140
Year 3	110	140	200

Which one corresponds to a situation of shifting risk parameters over time? Explain why.

1.34. (20 points)
You are using N years of data without any delay in order to estimate the next year.
The remaining weight will be given to the grand mean.
Allow the credibilities to differ by year.
The covariance between different years of data is:
$Cov[X_i, X_j] = (127.5)\ 0.75^{|i-j|} + (42.5)\ 0.965^{|i-j|} + 37\ \delta_{ij}$, where δ_{ij} is zero if $i \neq j$ and one if $i=j$.
With the aid of a computer, for N = 1, 2, 3, ... , 10, in each case determine the least squares credibilities and the corresponding minimum mean squared errors.

1.35. (3 points) Risk parameters are shifting over time.
In order to estimate the next year, you are using N years of data without any delay.
The remaining weight will be given to the grand mean.
In each case, the least squares credibilities have been determined as well as the corresponding minimum expected squared errors.
In one set of calculations, the credibilities were allowed to differ by year.
In a second set of calculations, the credibilities were the same for each year of data used.
The resulting minimum expected squared errors were as follows:

	N=1	N=2	N=3	N=4	N=5	N=6	N=7	N=8	N=9	N=10
Creds. Differ	58.71	58.00	57.60	57.36	57.23	57.14	57.10	57.07	57.05	57.04
Creds. the Same	58.71	58.01	57.63	57.45	57.37	57.36	57.40	57.47	57.55	57.64

Discuss fully these results.

1.36. (2 points) Mahler performs a chi-square test on his baseball data.
(a) (0.5 points) What is the purpose of this test?
(b) (1.5 points) Fully describe how this test is performed.

1.37. (9, 11/95, Q.10) (1 point) Which of the following are conclusions of Mahler in "An Example of Credibility and Shifting Risk Parameters"?
1. When parameter shift is present, the optimal credibility (based on least squares criterion) for the most recent available year of data increases as the delay in receiving the data increases.
2. Older years of data receive greater credibility when parameter shift is present than when it is not.
3. When parameter shift is present, use as many years of data as possible to maximize the accuracy of the prediction.

1.38. (9, 11/95, Q.31) (3 points) List and describe the three (3) evaluation methods used by Mahler in his paper "An Example of Credibility and Shifting Risk Parameters" to arrive at estimates for optimal credibility. Which one does Mahler suggest might give results that disagree with the others, and why might this be?

1.39. (9, 11/96, Q.20) (1 point) According to Mahler's "An Example of Credibility and Shifting Risk Parameters," which of the following are true ?
1. The best that can be done using credibility to combine two estimates is to reduce the mean squared error between the estimated and observed values to 50% of the minimum of the squared errors from either relying solely on the data or ignoring the data.
2. It is desirable to have the correlation between the experience modification and the loss ratio modified by experience modification to be zero.
3. Mahler recommends using as many years of data as there are available.

Note: I have rewritten statement #1 in order to match the current syllabus.

1.*40. (9, 11/97, Q.44)* *(3 points): Using Mahler's "An Example of Credibility and Shifting Risk Parameters," calculate the proportion of the total variance due to parameter shifting for the following scenario:*

- *There are 10 baseball teams.*
- *Each team plays 200 games.*
- *5 teams have true mean losing potential of 0.4.*
- *The other 5 teams have true mean losing potential of 0.6.*
- *The number of losses is Poisson distributed around its true mean losing potential.*
- *The actual number of losses for each team are:*

Team	*Number of Losses*
1	*75*
2	*115*
3	*61*
4	*110*
5	*94*
6	*133*
7	*139*
8	*98*
9	*81*
10	*94*

Show all work.

1.41. (9, 11/97, Q.45) (3 points) A major retailer, the Unlimited, has contracted you to project their loss ratio for general liability. The previous actuary was fired by the Unlimited, because she would rely only on the industry loss ratio to make the projection. The Unlimited has asked you to consider giving half of the credibility weight to the industry loss ratio and half to its own loss ratio from the previous year. The industrywide loss ratio is 65%. Using the least squares criterion, do you agree or disagree with your client? The Unlimited's historical data is as follows:

Policy Year	Loss Ratio
1/1 - 12/31/92	75%
1/1 - 12/31/93	70%
1/1 - 12/31/94	65%
1/1 - 12/31/95	60%
1/1 - 12/31/96	55%

1.42. (9, 11/97, Q.46) (2 points) Mahler's "An Example of Credibility and Shifting Risk Parameters" describes the Meyers/Dorweiler criterion for evaluating methods of assigning credibility to past data in order to predict future performance.

a. (1 point) In utilizing this criterion, what are the two ratios Mahler calculates to evaluate the predictors of baseball losing percentages?

b. (1 point) Within the contest of an experience rating plan, what quantities would be the equivalents to each of the ratios given in part (a)?

1.43. (9, 11/98, Q.13) (1 point) Following the approach described by Mahler in "An Example of Credibility and Shifting Risk Parameters" and given the following data, use exponential smoothing to calculate the expected 1999 loss ratio for the Increasingly Risky Corporation.
Increasingly Risky has produced the following historical loss ratios:

1998	100%
1997	90%
1996	80%
1995	70%
1994 and Prior	60%

Credibility Z = 30%

1.44. (9, 11/98, Q.14) (1 point) In "An Example of Credibility and Shifting Risk Parameters," Mahler discusses the maximum reduction in the mean squared error of an estimate that can be accomplished by using credibility.
You are given the following estimates based upon one year of data:
Mean squared error relying solely on the data = 80.
Mean squared error ignoring the data = 100.
What is the best mean squared error that can be achieved using a linear weighted average of the two estimates?

1.45. (9, 11/98, Q.25) (4 points) For the past 25 years, the Bermuda Captives have battled in the highly competitive Island Sunshine League. Their losses in each individual 100 game season are shown below, in five year intervals. Also shown below are the 25 year average losing percentages for each team in the Island Sunshine League. Each team played 100 games in each of the 25 years.

Bermuda Captives Loss Record	5 Year Subtotal
Seasons 1 - 5	160
Seasons 6 -10	170
Seasons 11 - 15	294
Seasons 16 - 20	330
Seasons 21 - 25	296

Team	25 Year Average Loss %
Bermuda Captives	50.0%
Barbados Bombers	60.0%
Jamaica White Sox	55.0%
Trinidad Hurricanes	45.0%
Cayman Cubs	40.0%

Critical Chi-Square statistic at 95% confidence level: 9.488

In Mahler's paper "An Example of Credibility and Shifting Risk Parameters," the author discusses three tests to perform on the data sets being observed. Use Mahler and the data above to answer the following questions.

a. (0.5 point) Mahler performs a test using the binomial distribution on the data set. What is the purpose of this test?

b. (0.75 point) Perform the binomial test at the 95% confidence level using the standard normal approximation, and give your conclusion of that test with respect to the above data. Show all work.

c. (0.5 point) Mahler performs a chi-square test on the data set. What is the purpose of this test?

d. (0.75 point) Perform the chi-square test described by Mahler at the 95% confidence level, and give your conclusion of that test with respect to the above data. Show all work.

e. (0.5 point) Mahler performs a correlation test on the data set. What is the purpose of this test?

f. (1 point) Describe how one would perform the correlation test on the above data set. What would the likely conclusion be on the above data set?

1.46. (9, 11/99, Q.48) (4 points) In Mahler's "An Example of Credibility and Shifting Parameters," the author gives the following equation:

$$V(Z) = \sum_{i=1}^{N}\sum_{j=1}^{N} Z_i Z_j (\tau^2 + C(|i-j|)) - 2\sum_{i=1}^{N} Z_i (\tau^2 + C(N+\Delta-i)) + \tau^2 + C(0)$$

where Z_1 is the credibility for the earliest year used.

a. (1 point) Define the following terms:

i) V(Z)

ii) τ^2

iii) C(k)

iv) Δ

b. (3 points) The Cayman Island Captives play in the Actuarial Baseball League. Using the following information, predict the Captives' winning percentage in the year 2000, based on least squares credibility as described by Mahler with N = 2 years of data. Show all work.

Year	Winning Percentage
1997	55%
1998	40%
1999	45%
Grand Mean	50%

τ^2	0.1000
C(0)	0.8000
C(1)	0.5000
C(2)	0.3500

1.47. (9, 11/00, Q.34) (2 points) Answer the following based on Mahler's "An Example of Credibility and Shifting Risk Parameters."

a. (1.5 points) Briefly describe three criteria used to compare the performance of credibility methods.

b. (0.5 point) Mahler states that one criterion differs from the other two criteria on a conceptual level. Which criterion is that? Briefly state in what way it differs from the others.

1.48. (9, 11/01, Q.1) (1 point) In Mahler's "An Example of Credibility and Shifting Risk Parameters," the author evaluates various estimates for baseball teams' future losing percentages using historical losing percentages He discusses the impact of shifting parameters over time in this context. According to Mahler, which of the following statements regarding shifting risk parameters is false?

A. The correlation between years that are close together is significantly less than the correlation between years that are further apart.

B. With delays in receiving historical data, the resulting estimates of the future will be less accurate.

C. Based on the least squares criterion, the optimal credibility decreases with increased delays in receiving the data.

D. If the data available to predict the next year, $Year_{x+1}$, included only data from $Year_{x-1}$, there is a significant increase in the squared error as compared to what would result if the data available included $Year_x$.

E. Older years of data should be given substantially less credibility than more recent years of data.

1.49. (9, 11/03, Q.21) (1 point) Briefly describe two methods to test whether risk parameters shift over time.

1.50. (9, 11/04, Q.3) (1 point) Three experience rating plans have been developed and you are trying to evaluate which is optimal. Each rating plan has been tested on four different risks. The following tables summarize the indicated modifications and the resulting errors.

Plan 1

Risk Number	Predicted Modification Factor	Error
1	1.30	40%
2	1.30	40%
3	0.70	30%
4	0.70	30%

Plan 2

Risk Number	Predicted Modification Factor	Error
5	1.30	10%
6	1.30	-10%
7	0.70	-20%
8	0.70	20%

Plan 3

Risk Number	Predicted Modification Factor	Error
9	1.30	4%
10	1.20	2%
11	0.80	-2%
12	0.70	-4%

Which of the following summarizes the preferred plan based on the Meyers/Dorweiler criterion and the least squared error criterion?

	Meyers/Dorweiler Criterion	Least Squared Error Criterion
A.	Plan 1	Plan 2
B.	Plan 1	Plan 3
C.	Plan 2	Plan 1
D.	Plan 2	Plan 3
E.	Plan 3	Plan 3

1.51. (9, 11/05, Q.2) (3 points)
a. (1.5 points) Expected losses for a risk within a class are projected based on the formula:
E = Z X + (1 - Z) P, where
X = the most recent accident year's losses
P = the prior estimate of the most recent accident year
Z = the credibility assigned to the most recent accident year
Assume:

- No delay in obtaining data
- Z = 10%

What is the difference in the weight given to accident year 2001 losses in accident year 2002's estimate and the weight given to accident year 20011 losses in accident year 2005's estimate?
b. (1.5 points) If there are significant shifts in risk parameters that require Z to be reevaluated, will the answer to part a. above increase, decrease, or remain constant. Explain your answer. Assume that there are no changes other than the shifts in risk parameters.

1.52. (9, 11/07, Q.6) (2 points)
The actuary for an insurance company has been asked by senior management to determine whether the company's expected frequency has been shifting over time.
The actuary knows that the company has maintained a constant number of exposures and a uniform mix of business since 1997.
Based on an assumption that expected frequency has remained constant during the period, the actuary has compiled the following data.

Year	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Actual Claims	475	420	460	500	490	525	515	510	540	575
Expected Claims	500	500	500	500	500	500	500	500	500	500

Discuss two methods that the actuary could use to test whether the expected frequency has been shifting over time. Describe any assumptions, calculations, or additional information that would be necessary to completely formulate and carry out each test.

1.53. (8, 11/12, Q.3) (1.75 points) The table below shows property claim frequency by year for the last five years. Assume that claim frequencies are Poisson distributed with a mean of 1.5.

Year	Exposures	Frequency
2011	118	1.5
2010	132	1.7
2009	121	1.3
2008	109	1.6
2007	97	1.3

The critical value for the relevant chi-squared distribution is 9.49.

a. (1.25 points) Calculate the chi-squared test statistic for whether the claim frequency is shifting over time. Interpret the result.

b. (0.5 point) Describe a second method for testing whether the claim frequency is shifting over time.

1.54. (8, 11/15, Q.4) (2.25 points)

An actuary is reviewing an account that has been with the company for over ten years.

Given the following:

- The claim frequency for this account follows a Poisson distribution, with $\lambda = 0.012$
- The recorded frequency for the last five years is as follows:

Year	Exposures	Frequency
2010	9,500	0.011
2011	11,000	0.010
2012	13,000	0.013
2013	10,500	0.012
2014	12,000	0.010

- The critical value for the relevant Chi-squared distribution is 9.49

a. (1.5 points) Use the Chi-squared test to evaluate whether the claim frequency is shifting over time. Include the hypotheses, test statistic, and provide an interpretation of the result.

b. (0.75 points)

Fully describe another method for determining whether claim frequency is shifting over time.

Solutions:

1.1. D.

1.2. Statement 1 is false.
Only statement number 2 is correct.
The losing percentages of the various teams are not random; there are better teams and worse teams (statement 2).
One might still argue: "Maybe teams are not the same, but perhaps past performance is a poor predictor of future performance." So Mahler shows that experience in one period has predictive power for other periods.
Statement 3 is false. Insurance risks are independent; a loss for one insured does not imply anything about losses for other insureds. (Certain lines of business, such as wind losses for homeowners, are exceptions.) Baseball differs from insurance by the constraint on the overall losing percentage: it is always 50%. If there were only two teams, the won-loss percentage of one team tells us the won-loss percentage of the other team, but with enough teams and games played each year, this constraint is not material.

1.3. Statement 1 is true; Mahler shows this for all of the teams (see page 236 of the text).
Statements 2 and 3 are Mahler's conclusions from his data; he says (page 239):
On the other hand, there is a significant correlation between the results of years close in time. Thus recent years can be usefully employed to predict the future.

1.4. D. Under rule #1, a bad record one year is less likely to produce a bad record the following year than under rule #3. Therefore, $Z(1) < Z(3)$.
Under rule #2, a good record one year is more likely to produce a good record the following year than under rule #3. Therefore, $Z(2) > Z(3)$.
Thus, **$Z(1) < Z(3) < Z(2)$**.
Comment: None of these is the rule used by the NBA, but rule #1 is the closest.

1.5. Statement 1: With the current draft rules, where the worst team gets the first draft pick, the optimal credibility may be less than zero. To see this, suppose there were just one player on each team and 100 teams. The team that does worst one year, gets the first draft pick, and it becomes one of the best teams.
For example, a team with a losing percentage of 80% one year, might be expected the next year to have a losing percentage of 40%. Z (80%) + (1 - Z)(50%) = 40%, would imply Z = -1/3, so the credibility is less than zero.
With 12 players on a team (though only five starters) and only two to three dozen teams, the worst team does not necessarily become one of the best from a single draft pick, but it may move up to above average, which also implies negative credibility.
Statement 2: If the better teams get higher draft picks, the teams which are good one year are expected to become even better the next year, and the teams which are poor one year are expected to become even worse the next year.
For example, a team with a losing percentage of 40% one year, might be expected the next year to have a losing percentage of 35%. Z (40%) + (1 - Z)(50%) = 35%, would imply Z = 1.5, so the credibility is greater than one.
Comment: The period of time studied in Mahler's paper was prior to the 1965 introduction of baseball's draft of players.
For a history of the rules for the NBA draft see:
http://www.nba.com/history/draft_evolution.html

1.6. Statement 1: Mahler uses linear combinations of the previous years' loss ratios and the overall average loss ratio.
Statement 2: Each estimator gives an overall 50% losing percentage, so it is unbiased. In the baseball analogy, the overall expected won-loss ratio is 50%. Similarly, the overall mean is 50%, the average of last year's won-loss ratio for all teams is 50%, and so forth.
Rating procedures can be biased for several reasons; we give illustrations:
Trended or developed losses are generally biased, though we may not know size or even the direction of the bias. A rate review may use an 8% loss cost trend; if the trend is higher or lower than 8%, the trended loss ratio is biased.
A loss ratio credibility weighted with loss ratios from other states or other insurers is biased, since other states or insurers may have higher or lower expected loss ratios.
Statement 3: These estimators are like experience rating; they use past experience to predict future experience. An analogy for schedule rating would be to look at the recent draft picks to predict the changes in next year's losing percentages.

1.7. With a credibility of Z and a relative loss ratio the previous year of L_1, the predicted relative loss ratio for the second year is: $Z L_1 + (1 - Z)$.

The squared error is: $\{Z L_1 + (1 - Z) - L_2\}^2 = \{Z (L_1 - 1) - (L_2 - 1)\}^2$.

Taking the sum of the squared errors, and setting the derivative with respect to Z equal to zero:

$\Sigma\ 2 \{Z (L_1 - 1) - (L_2 - 1)\} (L_1 - 1) = 0$.

$\Rightarrow Z = \Sigma (L_1 - 1)(L_2 - 1) / \Sigma (L_1 - 1)^2$.

$\Sigma (L_1 - 1)(L_2 - 1) = (-0.10)(-.02) + (-0.07)(-0.07) + ... + (0.10)(0.04) = 0.0226$.

$\Sigma (L_1 - 1)^2 = (-0.10)^2 + (-0.07)^2 + ... + (0.10)^2 = 0.0340$.

The credibility is 0.0226/ 0.0340 = **66.5%**.

Comment: The mean of each years relative loss ratios is 1, by definition.

The credibility is the slope of the regression line, which is the linear curve of best fit to the data points, using least squares.

1.8. 1 and 2 only. See page 249 of Mahler.

1.9. The teams predicted losing percentage is:

(10%){95/(67 + 95)} + (10%){101/(61 + 101)} + (55%){96/(66 + 96)} + (25%)(50%) = 0.572.

Out of 88 games, this team is expected to lose: (0.572)(88) = 50.3 games.

Therefore, the predicted record is about: **38 wins and 50 losses**.

Comment: This data is for the Tampa Bay Rays of the American League.

Through 7/7/08 inclusive, their record in 2008 was 55 wins and 33 losses.

This is an example of a large prediction error.

It is impossible to avoid some large prediction errors, particularly when using a simple technique based solely on past losing percentages.

Hopefully, such large prediction errors are rare in experience rating.

1.10. Statement 1 uses an unweighted average. If risk parameters shift over time, last year's losses may be a good predictor of next year's losses, but losses from five years ago may be a poor predictor. Using more years of an unweighted average may not improve the estimate. See section 8.3 of Mahler, page 245, second to last paragraph from bottom:
"Based on most actuarial uses of credibility, an actuary would expect the optimal credibilities to increase as more years of data are used. In this example they do not. In fact, using more than one or two years of data does an inferior job according to this criterion."
The use of older data with equal weight eventually leads to a worse outcome.
Table 17 on page 266 of Mahler shows empirical results. With a 1 year unweighted average, the optimal credibility is 66.0%. The optimal credibility increases to a maximum of 73.6% with 4 years of data, and then decreases as the years increase. At 10 years, the credibility has decreased to 66.9%.
Statement 2 says that if we know the general range in which the optimal credibility value lies, it doesn't make much difference what value we pick from that range.
In the past, some actuaries believed it was important to choose the proper full credibility standard, since different credibility values may produce more or less accurate rates. Not so, says Mahler. If the credibility values is near the optimal value, there is little difference in the accuracy of the rates.
Statement 3 is false. The best we can expect is to reduce the mean squared error to about 75% of the lower of the original estimates. See the third paragraph on page 252 of Mahler:
"In the current case, the best that can be done using credibility to combine two estimates is to reduce the mean squared error between the estimated and observed values to 75% of the minimum of the squared errors from either relying solely on the data or ignoring the data."
Comment: If Statement 1 were changed to a weighted N-year average, it be would true.
A weighted average using N years is a special case of the weighted average using N+1 years, since it is an N+1 year average with a weight of 0 for the oldest year. Since the N year average is one instance of an N+1 year average, the optimal N+1 year average must be at least as good as the optimal N year average.
In statement 2, Mahler is not saying that the credibility values do not affect the rate indication. Different credibility values give different indications. Suppose the indicated pure premium is $5.00 per $100 of payroll and the underlying pure premium is $2.00 per $100 of payroll. A credibility of 60% gives a rate of (60%)($5.00) + (40%)($2.00) = $3.80, and a credibility of 40% gives (40%)($5.00) + (60%)($2.00) = $3.20. This is a difference of about 15 to 20% in the rates. Different credibility values give different indicated rates. But the two sets of rates may have about the same expected squared error.
Suppose the true pure premium is $4.00 per $1 00 of payroll. The $3.80 rate has a squared error of $(4.00 - 3.80)^2 = 0.040$ and the $3.20 rate has a squared error of $(4.00 - 3.20)^2 = 0.640$. This is an enormous difference. However, we are speaking about the expected squared error. For a given size, we are choosing between 40% and 60% credibility. Sometimes the 40% credibility gives a higher rate and sometimes the 60% credibility gives a higher rate. Mahler says that the mean squared error won't differ much, as long as both values are close to the optimal value. If the optimal value is between 40% and 60%, both credibilities give about the same expected squared error.
Let's change the scenario. Suppose we don't know the true pure premium. One credibility value gives an indicated pure premium of $3.20, whereas another credibility value gives an indicated pure premium of $3.80. We don't know which estimate is closer to the true pure premium. Sometimes the first estimate is better, and sometimes the second is better.

We ask, over the full distribution of true pure premiums, which estimate is better? Mahler says: As long as the credibility is near the optimal value, there is not much difference.

Suppose a credibility estimator gives a pure premium that just equals the future loss costs. The squared error is zero, which is less than 75% of the minimum. However, Mahler is talking about the expected squared error, not the actual squared error in any particular instance. The actual squared error may be 0% by happenstance.

If the mean squared error is zero, the estimator is right on the mark; predicting future experience perfectly. This never happens, since there is random fluctuation in the losses.

You might recall the 75% figure as follows. If the optimal credibility is close to 0% or to 100%, it doesn't reduce the mean squared error much from the lower of relying entirely on the data or not relying on the data at all. If the optimal credibility is 10%, the difference between 0% and 10% is not great. The largest effect occurs when the optimal credibility is 50%. In that case, we should be using 50% of each estimator instead of 100% of either the experience or the overall mean. The mean squared error is reduced by the square of 50%, or 25%; this is the complement of 75%. $1 - 0.5^2 = 0.75$.

The previous paragraph is obviously not a mathematical derivation of the 75% result; see Appendix E of Mahler's paper, not on the syllabus, for the derivation. It simply shows the intuition.

1.11. None of these statements are true!
Statement 1 should say second order function, not linear. Mahler says on the top of page 264:
"Equation 11.2 shows that the squared error is a second order polynomial in the Z_i. This equation is the fundamental result for analyzing least squares credibility."
In equation 11.2 on page 263, V(Z) is the expected squared error. The second order polynomial comes from the $Z_i Z_j$ product in the first double summation.
Statement 2 is false. Table 16 on page 264 shows the solution of the matrix equations, and many of the credibilities are negative. In footnote 45 of page 265, Mahler says:
"Giving negative weight to some years allows a larger weight to be given to other years. The net effect is to reduce the expected squared error."
This result is counterintuitive. One might think that the negative credibilities stem from random loss fluctuations. But the negative credibilities all show up in the same columns (columns 5, 7, and 8 of Table 16), so there is some systematic effect, though finding an intuitive explanation is difficult.
Statement 3 is false. Mahler divides the variance into three parts, but he does not use this division for least squares credibilities. As he says at the top of page 263:
"It is possible to divide the within variance into two parts. The first part is the process variance excluding the effect of shifting parameters over time. The second part is that portion of the within variance due to shifting parameters over time. While this division may aid our understanding, it is not necessary for the calculation of the least squares credibilities."
Illustration: Suppose we examine a group of 100 large employers with $100 million of payroll apiece. The average employer has five workers' compensation claims a month. If we draw a sample of ten months, with each month taken from a random employer, what is the expected variance? That is, if we look at Employer #23 for January, Employer #41 for February, and so forth, what is the expected variance among the number of monthly claims?
Suppose first that there is no process variance and no between variance. If each employer has 5 claims each month, the variance of the number of claims in the sample is zero.
Suppose there is process variance but no between variance: that is, each employer has a Poisson distribution of claims with a mean of five. The process variance for each employer is 5, and the expected variance of the number of claims in the sample is 5.
Suppose the process variance is zero: that is, each employer has its expected number of claims each month, but the between variance is not zero: the expected claims differ by employer. If the between variance is 8, then the expected variance of the sample is 8.
Suppose the between variance is zero (all employers are identical) and the process variance at any moment in time is zero (the employer always realizes the expected number of claims). If the expected number of claims changes over time, then the variance of the claims in the ten samples is more than zero. This is what is captured by Mahler's ζ^2.
Comment: To prepare for the exam, know equations 11.1, 11.2, 11.3, and 11.4. The derivation of equation 11.2 is in Appendix C of Mahler, not on the syllabus. Question 48 of the Fall 1999 exam asked a question about equation 11.2, which was given to you.

1.12. Statement 1 is false. Individual risk rating is most important (credibility is highest) when classes are heterogeneous; as the between variance increases, the credibility Z increases. For individual risk rating, the between variance is the variance among the risks in the class, not the variance between the classes.

If $\tau^2 = 0$, then the class is perfectly homogeneous; the expected losses of risks within the class do not differ. In this case, the class rate is the proper rate for each risk. The loss history of any risk gives us extra noise, not extra information; the proper experience rating credibility is zero.

If the class is exceedingly heterogeneous, in other words if τ^2 is large, then the class rate tells us little about the proper rate for any insured.

The insured's experience is a combination of noise and information. We use credibility to separate the information from the noise.

Statement 2 is true. Statement 2 deals with the flip side of this issue. Individual risk rating is useful if the risk's experience gives us information about that risk's loss propensities. Suppose that the within variance is zero; i.e., the losses don't change from year to year. If last year the loss costs were $100,000, they are $100,000 this year as well (adjusted for exposure changes and loss cost trend). As the within variance goes to zero, the credibility goes to 100%.

If the within variance is high, the loss history doesn't tell us much about expected losses. As the within variance increases, the credibility decreases. Small insureds have high within variance, so their credibility is low. Large insureds have low within variance, so their credibility is high. The within variance is $\delta^2 + \zeta^2$.

Statement 3: As the variance due to shifting parameters increases, we give less weight to older accident years and more weight to more recent years: as the variance due to shifting parameters increases, the weight given to year 1 (now 10%) decreases.

If instead $\zeta^2 = 0$, then there are no shifting risk parameters, and every year would be equally good for predicting the future.

Comment: The weights sum to 100%, so the weight given to year 5 (now 30%) increases if ζ^2 increases. We can't say anything about the weights for years 2, 3, and 4. Presumably, the weight for year 2 decreases, and the weight for year 4 increases, but we can't say this with certainty.

The within variance would also be called the Expected Value of the Process Variance (EPV). The between variance would be also called the Variance of the Hypothetical Means (VHM).

1.13. The variance between risks (τ^2) is the variance of the true class rates for class ratemaking, or the variance between the individual risk propensities within a class for experience rating. The process variance excluding the effect of shifting parameters over time (δ^2) is the random fluctuation of the class loss costs for class ratemaking, or the random fluctuation of an individual risk's loss costs for experience rating. The portion of the within variance due to shifting parameters over time (ζ^2) is the variance of the average class pure premium stemming from changes in the class risk parameters over time, or the variance in the individual risk's expected pure premium stemming from changes in the insured's attributes over time.

1.14. (a) A rating plan that uses a reasonable primary-excess split has less variance of the individual risk's credibility weighted experience, whereas a no-split rating plan has higher variance. Thus this change in the rating plan doesn't change the variance of the hypothetical means, but it effectively decreases the process variance of the individual risk's experience. This decreases the K value in the credibility formula and increases the credibility. There is no change in the effect of shifting risk parameters.
(b) Using 2 years instead of 5 in the rating plan does not change the variance of the hypothetical means. If we use a weighted average of the years, using only 2 years degrades the rating plan and increases the process variance, so the experience rating credibility decreases.
Using only the most recently available 2 years of data reduces the effects of shifting risk parameters on the experience rating plan. (Which years we use does not change the covariance structure of the entire data set.) If we use an unweighted average of the years of data, using only 2 years may improve the rating plan compared to using 5 years, if the effect of shifting risk parameters is large. (See Table 6 in Mahler.) In that case, the experience rating credibility could be larger for either 2 or 5 years, depending on the details. (See Table 9 in Mahler.)
(c) Refining the classification plan decreases the variance of the hypothetical means, since all risks within any class are more alike. There is no change on the process variance of any individual risk. The K value in the credibility formula increases and credibility decreases. We can rely more on the class rate and less on the individual's experience, since the class rate is now a better estimate of the individual's loss potential than it was before the classification plan was refined.
There is no change in the effect of shifting risk parameters.

1.15. Only statement 1 is correct. At the top of page 285, Mahler writes:
"If an experience rating plan works properly, then after the application of experience rating, an insurer should be equally willing to write debit and credit risks. In other words, the modified loss ratio of expected losses to modified premiums should be the same for debit and credit risks."
Underwriters sometimes say: "We don't want to give this risk such a large debit, we don't want to punish him too much for one accident." This sentence is confused; the confusion stems from common parlance. The experience rating plan is not rewarding or punishing a risk for good or poor experience. Rather, the past experience helps predict future loss costs, and the modified rates are the best estimate of the future loss costs.
Since the standard premium, which includes the debit or credit mod, is the best estimate of future loss costs, insurers should be indifferent between debit and credit risks.
Statement 3 is false. Meyers/Dorweiler solely deals with if there is a pattern in the errors. For any experience rating plan, there is some credibility that satisfies Meyers/Dorweiler. If the plan successfully identifies good and poor risks, the credibility should be high; if the plan can not identify good and poor risks, the credibility should be low. In each case, the proper credibility gives a Kendall tau statistic of zero and satisfies the Meyers-Dorweiler criterion.
The plan may have enormous errors, but if there is no pattern, the ideal credibility satisfies Meyers-Dorweiler. See page 271 of Mahler.
Statement 2 is false. The Kendall τ statistic reflects the correlation in the order of two series. If two series are from uncorrelated distributions, the expected Kendall τ statistic is zero, and the actual correlation is symmetrically distributed on [-1, +1]. The same statements are true for the Kendall t statistic as for the statistical correlation; see page 286 of Mahler's paper in Appendix B, not on the syllabus. If the credibility approaches zero, past experience is not used at all. The modification is one for all risks, and the correlation with the actual loss costs is zero.
The expected value of τ is zero; the actual value in any instance is symmetrically distributed over [-1, +1].

1.16. 1 only.
We do not directly observe expected values; that eliminates choices 2 and 3. Insurance, unlike baseball, has no constraint on the grand mean. We estimate the mean by observing all risks over a long period. However, that is still an estimate subject to random fluctuation.
For insurance situations where we are interested in relativities compared to average, then by definition M = 1, however it is not directly observed.

1.17. 1 and 2 only.
Statement 1 is true. Mahler says on pages 252-253:
“If there is a delay before the data are available for use in experience rating, the resulting estimate of the future will be less accurate.
As the delay increases, the squared error increases significantly.
Statement 2 is correct as well. As Mahler says on page 254:
"The optimal credibility (as determined using the least squares criterion) decreases as the delay increases. Less current information is less valuable for estimating the future.”
Statement 3 is false for the data set examined. The predictive value declines slowly as the delay increases, and it takes many years before it gets close to zero. Table 11 on page 254 shows the figures. The average credibility is about 70% with a 1 year lag between latest data point and future prediction and about 45% with a 4 year lag. Statement 3 might be true for some data set where risk parameters were shifting significantly faster than in the baseball data examined by Mahler.
However, this is extremely unlikely to be the case for insurance data; insurance data tend have parameters that are more stable than in Mahler’s baseball data.
Comment: What is the relation between delays and shifting risk parameters?
Suppose we predict the pure premium for year 6 using 3 years of experience data.
If there is no delay, we use years 3, 4, and 5.
If there is a short delay, we use years 2, 3, and 4.
If there is a long delay, we use years 1, 2, and 3.
If the risk parameters don't shift over time, all three methods should have similar expected squared errors.
If the risk parameters shift over time, then the first method is best, and the last method is worst.

1.18. In both cases, the correlations decline with increasing separation of the years.
This indicates that parameters are shifting over time.
The rate of decline in correlations is swifter for data set one, indicating that parameters are shifting more quickly for data set one than for data set two.

1.19. Let P_{2009} = estimate of 2009. Let X_{2008} = observation for 2008.

$P_{2009} = Z X_{2008} + (1 - Z) P_{2008}$.

Similarly, $P_{2008} = Z X_{2007} + (1 - Z) P_{2007}$.

$P_{2007} = Z X_{2006} + (1 - Z) P_{2006}$.

Therefore, $P_{2009} = Z X_{2008} + (1 - Z) P_{2008} = Z X_{2008} + (1 - Z)Z X_{2007} + (1 - Z)^2 P_{2007}$

$= Z X_{2008} + (1 - Z)Z X_{2007} + (1 - Z)^2 Z X_{2006} + (1 - Z)^3 P_{2006}$.

The coefficient for X_{2006} is $(1 - Z)^2 Z$.

When $Z = 55\%$, $(1 - Z)^2 Z = (1 - 0.55)^2 (0.55) =$ **11.1 %**.

Comment: The weights applied to years of data decline geometrically.
This form of estimator is similar to what is done in pure premium ratemaking.
For pure premium ratemaking, the credibility weighted pure premium is:
Z (the indicated pure premium) + (1 - Z) (the underlying pure premium).
In loss ratio ratemaking, the credibility weighted loss ratio is:
(Z) the experience loss ratio + (1 - Z) (the permissible loss ratio).

1.20. None of 1, 2, or 3 is said by Mahler.
Comment: See Mahler at page 272.
Statement 3 is one of the most practical implications from Mahler's paper: if we know the approximate credibility, a refined figure doesn't make much of a difference. For example, any credibility figure between 40% and 70% might give about the same expected squared error. Actuaries sometimes argue whether the full credibility standard should be 2,500 claims or 3,000 claims. In many cases, it doesn't make much difference.

1.21. Statements 1 and 2 are correct; the mean squared error is the expected squared error. Statement 3 is false. To solve the second order equation, Mahler takes partial derivatives. This creates linear equations, which can be solved for the credibilities.
Comment: See Mahler at page 280.

1.22. (a) (1) Some new teams entered the leagues due to expansion. Mahler had the same 8 teams in each league throughout. We would have varying numbers of teams. For example, in 1969 the Kansas City Royals and Seattle Pilots (now the Milwaukee Brewers) joined the American League. These new teams were worse than average. Thus the existing teams seemed to improve on average between 1968 and 1969.
(2) Some seasons were shortened by strikes. Thus there are some years where a significantly smaller number of games were played.
(3) Leagues were split into divisions, and in recent seasons, teams play teams within their division more frequently. Thus unlike in Mahler's study, teams do not play approximately the same number of games against each other team in their league. If in a given season a certain division is significantly stronger than average, then the teams in that division play opponents who are stronger than average. Therefore, the expected winning percentages for teams in that division would be lower than it would otherwise be if there was a balanced schedule.
(4) Interleague play was introduced recently. While only about 10% of games involve play between the two leagues, this complication was not present in Mahler's Study.
The average winning percentage for a league is no longer 50% each year.
(For example, in 2006 the American League won 154 out of 252 interleague games; 154/252 = 61%. Thus that year, the average winning percentage for the American League was greater than 50%.) Also the expected winning percentage of a team is effected by which teams it is scheduled to play that season. Each season, a team only plays some of the teams in the other league and that varies from year to year.
(b) Since Mahler's study, free agency was introduced. Thus players switch teams more frequently now. Thus I would expect the effect of shifting risk parameters to be greater than in Mahler's study.
Alternately, the difference between the best and the worst teams is usually less than in Mahler's study; there is more parity among the teams. Therefore, there is a smaller region in which the winning percentages can vary from year to year. Thus I would expect the effect of shifting risk parameters to be less than in Mahler's study.
Alternately, since Mahler's study, baseball has instituted a draft. Teams with the worst record get to draft earlier. This will tend to allow bad teams to get better more quickly. Conversely good teams will have a harder time staying good for a long time. Therefore, parameters may shift more quickly than in the era in Mahler's study.
Comment: There are many possible reasonable answers. In part (a) only give two reasons.

1.23. For two different years, $Cov[X_i, X_j] = \tau^2 + C(|i - j|)$.

For example, $Cov[X_1, X_3] = \tau^2 + C(2) = 5 + 8 = 13$.

For a single year of data, $Cov[X_i, X_i] = Var[X_i] = \tau^2 + C(0) = 5 + 50 = 55$.

A covariance matrix is:

Year 1	55	15	13	11	9
Year 2	15	55	15	13	11
Year 3	13	15	55	15	13
Year 4	11	13	15	55	15
Year 5	9	11	13	15	55

$\sum_{j=1}^{N} Z_j \, Cov[X_i, X_j] = Cov[X_i, X_{N+\Delta}]$, where we are predicting year $N + \Delta$, using years 1 to N.

(a) Using data for Year 1 to Predict Year 2, the equation is:

$55Z = 15. \Rightarrow Z = 15/55 =$ **27.3%**.

(b) Using data for Year 1 to Predict Year 3, the equation is:

$55Z = 13. \Rightarrow Z = 13/55 =$ **23.6%**.

(c) Using data for Year 1 to Predict Year 4, the equation is:

$55Z = 11. \Rightarrow Z = 11/55 =$ **20.0%**.

(d) Using data for Years 1 and 2 to Predict Year 3, the equations are:

$55Z_1 + 15Z_2 = 13.$

$15Z_1 + 55Z_2 = 15.$

The coefficients on the lefthand side are the first two rows and the first two columns of the covariance matrix, since we are using data from Years 1 and 2. The values on the righthand side are the first two rows of column three, since we are predicting year 3.

Solving, $\mathbf{Z_1 = 17.5\%}$ and $\mathbf{Z_2 = 22.5\%}$.

(e) Using data for Years 1 and 2 to Predict Year 4, the equations are:

$55Z_1 + 15Z_2 = 11.$

$15Z_1 + 55Z_2 = 13.$

The values on the righthand side are the first two rows of column four, since we are predicting Year 4.

Solving, $\mathbf{Z_1 = 14.6\%}$ and $\mathbf{Z_2 = 19.6\%}$.

(f) Using data for Years 1, 2, and 3 to Predict Year 4, the equations are:

$55Z_1 + 15Z_2 + 13Z_3 = 11.$

$15Z_1 + 55Z_2 + 15Z_3 = 13.$

$13Z_1 + 15Z_2 + 55Z_3 = 15.$

The coefficients on the lefthand side are the first three rows and the first three columns of the covariance matrix, since we are using data from Years 1, 2, and 3. The values on the righthand side are the first three rows of column four, since we are predicting Year 4.

Solving, $\mathbf{Z_1 = 11.0\%}$, $\mathbf{Z_2 = 15.0\%}$, and $\mathbf{Z_3 = 20.6\%}$.

(g) Using data for Years 1, 2, and 3 to Predict Year 4, the equations are:

$55Z_1 + 15Z_2 + 13Z_3 = 9.$

$15Z_1 + 55Z_2 + 15Z_3 = 11.$

$13Z_1 + 15Z_2 + 55Z_3 = 13.$

The values on the righthand side are the first three rows of column five, since we are predicting Year 5.

Solving, $\mathbf{Z_1 = 8.6\%}$, $\mathbf{Z_2 = 12.7}\%$, and $\mathbf{Z_3 = 18.1}\%$.

Comment: Parts f and g are beyond what you should be asked on your exam.

See Equation 11.3 in Mahler.

These linear equations are called the Normal Equations, as discussed in Loss Models.

The notation used in the syllabus paper by Mahler, written in 1988 and published in 1990, is unnecessarily complex. All one really needs is the covariance matrix. The least squares credibilities are determined by the relative sizes of the elements of the covariance matrix.

With no delay in getting data, $\Delta = 1$, similar to Mahler's Table 16:

	Years Between Data and Estimate		
Number of Years of Data Used (N)	1	2	3
1	27.3%		
2	22.5%	17.5%	
3	20.6%	15.0%	11.0%

With a delay in getting data, $\Delta = 2$:

	Years Between Data and Estimate		
Number of Years of Data Used (N)	2	3	4
1	23.6%		
2	19.6%	14.6%	
3	18.1%	12.7%	8.6%

1.24. Use data for Years 1 and 2 to Predict Year 3.
$Cov[(X_1 + X_2)/2, X_3] = \{Cov[X_1, X_3] + Cov[X_2, X_3]\} / 2 = (13 + 15)/2 = 14.$
$Var[(X_1 + X_2)/2] = \{Var[X_1] + Var[X_2] + 2\, Cov[X_1, X_2]\} / 2^2 = \{55 + 55 + (2)(15)\} / 4 = 35.$
Thus the weight given to the average of the years is: $Z = 14/35 =$ **40%**.
(Thus 20% weight is given to each year. When giving different weights we got: $Z_1 = 17.5\%$ and $Z_2 = 22.5\%$. Note that 17.5% + 22.5% = 40%.)
Use data for Years 1 and 2 to Predict Year 4.
$Cov[(X_1 + X_2)/2, X_4] = \{Cov[X_1, X_4] + Cov[X_2, X_4]\} / 2 = (11 + 13)/2 = 12.$
$Var[(X_1 + X_2)/2] = \{Var[X_1] + Var[X_2] + 2\, Cov[X_1, X_2]\} / 2^2 = \{55 + 55 + (2)(15)\} / 4 = 35.$
Thus the weight given to the average of the years is: $Z = 12/35 =$ **34.3%**.
Use data for Years 1, 2 and 3 to Predict Year 4.
$Cov[(X_1 + X_2 + X_3)/3, X_4] = \{Cov[X_1, X_4] + Cov[X_2, X_4] + Cov[X_3, X_4]\} / 3 =$
$(11 + 13 + 15)/3 = 13.$
$Var[(X_1 + X_2 + X_3)/3] =$
$\{Var[X_1] + Var[X_2] + Var[X_3] + 2\, Cov[X_1, X_2] + 2\, Cov[X_1, X_3] + 2\, Cov[X_2, X_3]\} / 3^2 =$
$\{55 + 55 + 55 + (2)(15) + (2)(13) + (2)15)\} / 9 = 251/9.$
Thus the weight given to the average of the years is: $Z = 13 / (251/9) =$ **46.6%**.
Use data for Years 1, 2 and 3 to Predict Year 5.
$Cov[(X_1 + X_2 + X_3)/3, X_5] = \{Cov[X_1, X_5] + Cov[X_2, X_5] + Cov[X_3, X_5]\} / 3 =$
$(9 + 11 + 13)/3 = 11.$
$Var[(X_1 + X_2 + X_3)/3] =$
$\{Var[X_1] + Var[X_2] + Var[X_3] + 2\, Cov[X_1, X_2] + 2\, Cov[X_1, X_3] + 2\, Cov[X_2, X_3]\} / 3^2 =$
$\{55 + 55 + 55 + (2)(15) + (2)(13) + (2)(15)\} / 9 = 251/9.$
Thus the weight given to the average of the years is: $Z = 11 / (251/9) =$ **39.4%**.

Comment: One could use equation 11.4: $Z = N \dfrac{N\,\tau^2 + \sum_{l=1}^{N} C(N+\Delta-i)}{N^2\,\tau^2 + \sum_{j=1}^{N}\sum_{l=1}^{N} C(|i-j|)}$.

Requiring that the weights applied to each year be equal results in a larger minimum mean squared error than allowing the weights to vary.

1.25. (a) Shifting Risk Parameters: The parameters defining the risk process for an individual insured are not constant over time. There are (a series of perhaps small) permanent changes to the individual insured's risk process as one looks over several years.
(b) 1. Private Passenger Automobile Insurance:
A driver's risk parameters might shift if he changed the location to which he commutes to work. (He drivers the same distance, but it is over different type of roads.)
So for example, if he now drives to work over local streets while before he mostly drove on a highway, his expected pure premium changes.
2. Workers Compensation Insurance:
There might be a change in the attitude of management with regard to workplace safety.
If management of the company paid more attention to workplace safety, then the expected pure premium would go down.
3. Homeowners Insurance:
The number of children in the neighborhood changes over time.
(The insured remains in the same house and there are no children living there.)
As the number of neighborhood children increases, there is more chance of a liability claim; the expected pure premium for the liability coverage would increase.
Comment: There are many possible examples. one could give in part (b).
The risk parameters of a Workers Compensation class can shift over time, so that its relativity compared to average for its Industry Group changes over time. This could be due to changes in the manufacturing process, how the work is performed, or the nature of the job.
The automobile experience of a town relative to the rest of the state could shift as that town becomes more densely populated.
The insurance experience of a town relative to the rest of the state could shift as that town undertook an effective campaign against insurance fraud.

1.26. (a) (1) Some new teams entered the leagues due to expansion. Mahler had the same 8 teams in each league throughout. We would have varying numbers of teams. When new and weaker teams enter the league, the existing teams seem to improve compared to average.
(2) The number of games per season was increased over this period of time from 14 to 16. Thus the amount of data varies.
(3) Some seasons were shortened by strikes. Thus there is one year (1982) where a significantly smaller number of games were played.
(4) Unlike in Mahler's study, teams do not play approximately the same number of games against each other team in their league. If in a given season a certain division is significantly stronger than average, then the teams in that division play opponents who are stronger than average. Therefore, the expected winning percentages for teams in that division would be lower than it would otherwise be if there was a balanced schedule.
(5) Each season a team plays at most 16 games, compared to about 150 in Mahler's study. Thus there is much more random fluctuation in the data than in Mahler's Study.
(b) Since the average career of a star football player is shorter than the average career of a star baseball player, I would expect the quality of a team to change more quickly in football. Thus, I would expect shifting risk parameters to have more effect on football data.
Alternately, since there are more players on a football team than a baseball team, the effect on the quality of the team from replacing one player is less than for baseball. I would expect the quality of a team to change less quickly in football. Thus, I would expect shifting risk parameters to have less effect on football data.
Comment: There are many possible reasonable answers. In part (a) only give two reasons. Feel free to make up a similar question to answer based on some other team sport you may prefer, such as basketball, hockey, soccer, etc.

1.27. For two different years, $Cov[X_i, X_j] = \tau^2 + C(|i - j|)$.

For example, $Cov[X_2, X_5] = \tau^2 + C(3) = 10 + 6 = 16$.

For a single year of data, $Cov[X_i, X_i] = Var[X_i] = \tau^2 + C(0) = 10 + 30 = 40$.

A covariance matrix is:

Year 1	40	25	20	16	13
Year 2	25	40	25	20	16
Year 3	20	25	40	25	20
Year 4	16	20	25	40	25
Year 5	13	16	20	25	40

$\sum_{j=1}^{N} Z_j \, Cov[X_i, X_j] = Cov[X_i, X_{N+\Delta}]$, where we are predicting year $N + \Delta$, using years 1 to N.

(a) Using data for Year 1 to Predict Year 2, the equation is: $40Z = 25$.

$\Rightarrow Z = 25/40 =$ **5/8 = 62.5%**.

(b) Using data for Year 1 to Predict Year 3, the equation is: $40Z = 20$.

$\Rightarrow Z = 20/40 =$ **1/2 = 50.0%**.

(c) Using data for Year 1 to Predict Year 4, the equation is: $40Z = 16$.

$\Rightarrow Z = 16/40 =$ **40.0%**.

(d) Using data for Year 1 to Predict Year 5, the equation is: $40Z = 13$.

$\Rightarrow Z = 13/40 =$ **32.5%**.

(e) Using data for Years 1 and 2 to Predict Year 3, the equations are:

$40Z_1 + 25Z_2 = 20$.

$25Z_1 + 40Z_2 = 25$.

The coefficients on the lefthand side are the first two rows and the first two columns of the covariance matrix, since we are using data from Years 1 and 2. The values on the righthand side are the first two rows of column three, since we are predicting year 3.

Solving, $\mathbf{Z_1 = 7/39 = 17.9\%}$ and $\mathbf{Z_2 = 20/39 = 51.3\%}$.

(f) Using data for Years 1 and 2 to Predict Year 4, the equations are:

$40Z_1 + 25Z_2 = 16$.

$25Z_1 + 40Z_2 = 20$.

The values on the righthand side are the first two rows of column four, since we are predicting Year 4.

Solving, $\mathbf{Z_1 = 28/195 = 14.4\%}$ and $\mathbf{Z_2 = 16/39 = 41.0\%}$.

(g) Using data for Years 1 and 2 to Predict Year 5, the equations are:

$40Z_1 + 25Z_2 = 13$. $25Z_1 + 40Z_2 = 16$.

Solving, $\mathbf{Z_1 = 8/65 = 12.3\%}$ and $\mathbf{Z_2 = 21/65 = 32.3\%}$.

Comment: See Equation 11.3 in Mahler.

1.28. (a) $V(Z) = Z^2\{\tau^2 + C(0)\} - 2\,Z\{\tau^2 + C(1)\} + \tau^2 + C(0) =$
$40Z^2 - (2)(25)Z + 40 = (40)(5/8)^2 - (50)(5/8) + 40 =$ **24.375**.
(b) $V(Z) = Z^2\{\tau^2 + C(0)\} - 2\,Z\{\tau^2 + C(2)\} + t^2 + C(0) =$
$40Z^2 - (2)(20)Z + 40 = (40)(1/2)^2 - (40)(1/2) + 40 =$ **30**.
(c) $V(Z) = Z^2\{\tau^2 + C(0)\} - 2\,Z\{\tau^2 + C(3)\} + \tau^2 + C(0) =$
$40Z^2 - (2)(16)Z + 40 = (40)(0.4)^2 - (32)(0.4) + 40 =$ **33.6**.
(d) $V(Z) = Z^2\{\tau^2 + C(0)\} - 2\,Z\{t^2 + C(4)\} + \tau^2 + C(0) =$
$40Z^2 - (2)(13)Z + 40 = (40)(0.325)^2 - (26)(0.325) + 40 =$ **35.775**.
(e) Years 1 and 2 predicting year 3.

$$V(Z) = (Z_1, Z_2, -1, 0, 0) \begin{pmatrix} 40 & 25 & 20 & 16 & 13 \\ 25 & 40 & 25 & 20 & 16 \\ 20 & 25 & 40 & 25 & 20 \\ 16 & 20 & 25 & 40 & 25 \\ 13 & 16 & 20 & 25 & 40 \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \\ -1 \\ 0 \\ 0 \end{pmatrix} =$$

$$(7/39, 20/39, -1, 0, 0) \begin{pmatrix} 40 & 25 & 20 & 16 & 13 \\ 25 & 40 & 25 & 20 & 16 \\ 20 & 25 & 40 & 25 & 20 \\ 16 & 20 & 25 & 40 & 25 \\ 13 & 16 & 20 & 25 & 40 \end{pmatrix} \begin{pmatrix} 7/39 \\ 20/39 \\ -1 \\ 0 \\ 0 \end{pmatrix} =$$

(7/39, 20/39, -1, 0, 0) . (0, 0, -920/39, -463/39, -369/39) = 920/39 = **23.59**.
(f) Years 1 and 2 predicting year 4.

$$V(Z) = (Z_1, Z_2, 0, -1, 0) \begin{pmatrix} 40 & 25 & 20 & 16 & 13 \\ 25 & 40 & 25 & 20 & 16 \\ 20 & 25 & 40 & 25 & 20 \\ 16 & 20 & 25 & 40 & 25 \\ 13 & 16 & 20 & 25 & 40 \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \\ -1 \\ 0 \\ 0 \end{pmatrix} =$$

$$(28/195, 16/39, 0, -1, 0) \begin{pmatrix} 40 & 25 & 20 & 16 & 13 \\ 25 & 40 & 25 & 20 & 16 \\ 20 & 25 & 40 & 25 & 20 \\ 16 & 20 & 25 & 40 & 25 \\ 13 & 16 & 20 & 25 & 40 \end{pmatrix} \begin{pmatrix} 28/195 \\ 16/39 \\ 0 \\ -1 \\ 0 \end{pmatrix} =$$

(28/195, 16/39, 0, -1, 0) . (0, 0, -463/39, -5752/195, -1077/65) = 5752/195 = **29.50**.

(g) Years 1 and 2 predicting year 5.

$$V(Z) = (Z_1, Z_2, 0, 0, -1) \begin{pmatrix} 40 & 25 & 20 & 16 & 13 \\ 25 & 40 & 25 & 20 & 16 \\ 20 & 25 & 40 & 25 & 20 \\ 16 & 20 & 25 & 40 & 25 \\ 13 & 16 & 20 & 25 & 40 \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \\ 0 \\ 0 \\ -1 \end{pmatrix} =$$

$$(8/65, 21/65, 0, 0, -1) \begin{pmatrix} 40 & 25 & 20 & 16 & 13 \\ 25 & 40 & 25 & 20 & 16 \\ 20 & 25 & 40 & 25 & 20 \\ 16 & 20 & 25 & 40 & 25 \\ 13 & 16 & 20 & 25 & 40 \end{pmatrix} \begin{pmatrix} 8/65 \\ 21/65 \\ 0 \\ 0 \\ -1 \end{pmatrix} =$$

(8/65, 21/65, 0, 0, -1) . (0, 0, -123/13, -1077/65, -432/13) = 432/13 = **33.23**.

Comment: The expected squared error is given by equation 11.2:

$$V(\vec{Z}) = \sum_{i=1}^{N}\sum_{j=1}^{N} Z_i Z_j \{\tau^2 + C(|i-j|)\} - 2\sum_{i=1}^{N} Z_i \{\tau^2 + C(N+\Delta-i)\} + \tau^2 + C(0).$$

It turns out that Equation 11.2 can be rewritten in matrix form,
Mean Squared Error = $V(Z) = Z^T C Z$. C is the matrix of covariances for the years of data.
Z is the (column) vector with credibilities in the years used to estimate, -1 in the year being estimated, and zeros in any other years. Z^T is the transpose of Z.
Note that year 1 is a worse predictor of year 3 than it is of year 2. Therefore, the mean square error is larger for predicting year 3 than it is for predicting year 2; 30 > 24.375.
The longer the delay in getting data, the larger the mean squared error.
Using years 1 and 2 to predict year 3 is a better estimator than using just year 2. Therefore, the mean square error is larger using just year 2 than it is using years 1 and 2; 30 > 23.59.

1.29. D. While we would like statement A to be true, with several years of data and a particular set of covariances, the optimal weights can be negative.
While statement B sounds like it should be true, with several years of data and a particular set of covariances, the optimal weight for 1953 data may be more than that for 1954 data.
While statement C sounds like it should be true, the optimal weight assigned to the most recent year of data may be slightly more when using for example 6 years of data rather than 5 years of data.
Statement D is true. See Table 19 in Mahler.
For example using 1952 and 1953 to predict 1954 is a special case of using 1951, 1952, and 1953 to predict 1954, where 1951 is given a weight of zero. Thus the minimum expected squared error from the latter can not be more than that from the former.
Comment: If the covariance matrix was more structured, one could usually avoid negative credibilities. See for example, Tables 4 and 7 in "A Markov Chain Model of Shifting Risk Parameters", by Howard C. Mahler, PCAS 1997.

1.30. $\text{Var}[X] = 6$. $\text{Cov}[X_1, X_2] = 0.9$. $\text{Cov}[X_1, X_3] = 0.9^2 = 0.81$.

$\text{Cov}[X_1, X_4] = 0.9^3 = 0.729$. $\text{Cov}[X_1, X_5] = 0.9^4 = 0.6561$. $\text{Cov}[X_1, X_6] = 0.9^5 = 0.59049$.

The covariance matrix between the years of data is:

$$\begin{pmatrix} 6 & 0.9 & 0.81 & 0.729 & 0.6561 & 0.59049 \\ 0.9 & 6 & 0.9 & 0.81 & 0.729 & 0.6561 \\ 0.81 & 0.9 & 6 & 0.9 & 0.81 & 0.729 \\ 0.729 & 0.81 & 0.9 & 6 & 0.9 & 0.81 \\ 0.6561 & 0.729 & 0.81 & 0.9 & 6 & 0.9 \end{pmatrix}$$

Therefore, the equations for the least squares credibilities (the Normal Equations) are:

$6Z_1 + 0.9Z_2 + 0.81Z_3 + 0.729Z_4 + 0.6561Z_5 = 0.59049$.

$0.9Z_1 + 6Z_2 + 0.9Z_3 + 0.81Z_4 + 0.729Z_5 = 0.6561$.

$0.81Z_1 + 0.9Z_2 + 6Z_3 + 0.9Z_4 + 0.81Z_5 = 0.729$.

$0.729Z_1 + 0.81Z_2 + 0.9Z_3 + 6Z_4 + 0.9Z_5 = 0.81$.

$0.6561Z_1 + 0.729Z_2 + 0.81Z_3 + 0.9Z_4 + 6Z_5 = 0.9$.

Solving: $Z_1 = 5.525\%$, $Z_2 = 6.373\%$, $Z_3 = 7.560\%$, $Z_4 = 9.150\%$, $Z_5 = 11.228\%$.

$5.525\% + 6.373\% + 7.560\% + 9.150\% + 11.228\% = 39.836\%$.

The remaining weight of 60.164% is given to the a priori mean.

Comment: Beyond what you will be asked on your exam.

The older years are less correlated with year 6, the year we wish to estimate, and thus their data is given less weight.

See “A Markov Chain Model of Shifting Risk Parameters,” by Howard C. Mahler, PCAS 1997, not on the syllabus.

1.31. For each set of predictions we calculate the errors: predicted - observed.

Policy Number	Rating Plan 1 Modification Factor	Error
1	0.80	-0.05
2	0.90	+0.05
3	1.00	0
4	1.10	+0.05
5	1.20	-0.05

Policy Number	Rating Plan 2 Modification Factor	Error
1	0.87	+0.02
2	0.87	+0.02
3	1.00	0
4	1.03	-0.02
5	1.23	-0.02

Policy Number	Rating Plan 3 Modification Factor	Error
1	0.81	-0.04
2	0.83	-0.02
3	1.00	0
4	1.09	+0.04
5	1.27	+0.02

Plan 2 has positive errors for debit risks and negative errors for credit risks.
Plan 3 has negative errors for debit risks and positive errors for credit risks.
In both cases, the errors are correlated with the experience modifications.
In the case of Plan 1, the errors have a correlation close to zero with the experience modifications.
Thus **by the Meyers/Dorweiler criterion, we prefer Plan 1**.
Plan 1 has a larger average squared error than plan 3, which has a larger average squared error than plan 2. Thus **by the least squared error criterion we prefer plan 2**.
Comment: Intended as an improvement on the less than completely logical past exam question: 9, 11/04, Q.3.
One would do such testing on thousands of policies rather than just 5.

1.32. This probably is based on a situation with shifting risk parameters.
More distant years are given less weight than more recent years.
For example, if we were using 2003, 2004, and 2005 to predict 2006, we would give 2003 weight 2.82%, 2004 weight 3.01% and 2005 weight 3.23%. (The remaining weight of 90.94% would be given to the overall mean.) This makes sense, since 2003 is less correlated with 2006 than is 2005, and thus is a worse predictor of 2006 than is 2005.
Using fewer years of data is a special case of using more years of data, where some of the credibilities have been constrained to be zero. (The credibilities by year are allowed to differ.) Thus using more years of data does a better job than using fewer years of data. Thus the minimum mean squared errors should decline as we use the least square credibilities for more years of data. This is in fact what we observe. For example, for 4 years of data the minimum mean squared error is 0.03826, while for 5 years of data it is 0.03824. (As more years are added, the MSE continues to improve, but only very slowly. Eventually there will no longer be any significant improvement from adding years.)
The sum of the credibilities increases as the number of years of data increase; we give less weight to the overall mean. The sum of credibilities increases at a decreasing rate. (With shifting risk parameters, as the number of years of data approaches infinity, the sum of credibilities will approach a value less than one. This differs from the Buhlmann Credibility formula, $Z = N / (N+K)$, where the limit is one.)
Comment: Based on an approximation to the model of California Female P. P. Auto Drivers in "A Markov Chain Model of Shifting Risk Parameters", by Howard C. Mahler, PCAS 1997.
The credibilities shown by year are similar to those in Table 4 of that paper.
The results shown in the question were based on: $Cov[X_i, X_j] = (0.0014)\ 0.94^{|i-j|} + 0.037\ \delta_{ij}$, where δ_{ij} is zero if $i \neq j$ and one if $i=j$.
The sum of the credibilities approaches about 32% as the number of years approach infinity. (Similar to Figure 12 in the Markov Chain paper.)
The minimum mean squared error approaches 0.03814 as the number of years approach infinity.
This model has smoothed out the peculiarities of the covariances of the data that are due to random fluctuation. Thus we see a regular pattern of credibilities. More distant years get less credibility than more recent years, declining in a nice pattern. This female driver data has a slower rate of shifting risk parameters than does the baseball data, thus the credibilities for distant years decline more slowly for the driver data than the baseball data.

1.33. The first covariance matrix has all of its elements equal. The variance of a year is the same as the covariance of two different years. Thus all of the years of data are perfectly correlated. Whatever the observed relativity (for a baseball team, or insured, or class) is in one year, it is the same in every other year. This is not a reasonable model for insurance.
In the second covariance matrix, the variance of year of data is 200, while the covariance between different years is 140.
Therefore, the correlation of any two different years of data is: 140/200 = 70%.
The correlation between different years does not depend on how far apart they are.
In the third covariance matrix, the variance of a year of data is 200, while the covariance between consecutive years is 140, and the covariance of year 1 and year 3 is 110.
Therefore, the correlation of consecutive years of data is: 140/200 = 70%, the correlation of years 1 and 3 is: 110/200 = 55%. The correlation between years further apart is less than the correlation of years closer together. This is what we expect with shifting risk parameters over time.
The third matrix corresponds to a situation of shifting risk parameters over time.
Comment: The second matrix is an example of the Buhlmann covariance structure.

1.34. For years 1, 2, and 3, the covariance matrix is: $\begin{pmatrix} 207 & 136.638 & 111.296 \\ 136.638 & 207 & 136.368 \\ 111.296 & 136.368 & 207 \end{pmatrix}$.

Thus if we are using years 1 and 2 to predict year 3, the least squares credibilities satisfy:
$207 Z_1 + 136.638 Z_2 = 111.296$.

$136.638 Z_1 + 207 Z_2 = 136.368$.

Solving $Z_1 = 0.1807$ and $Z_2 = 0.5408$.

Then using equation 11.2, the minimum expected squared error is:
$207 Z_1^2 + 207 Z_2^2 + (2)(136.638) Z_1 Z_2 - (2)(111.296) Z_1 - (2)(136.638) Z_2 + 207 = 112.994$.

With 1 being the most recent year, proceeding in a similar manner we get:

Year	N=1	N=2	N=3	N=4	N=5	N=6	N=7	N=8	N=9	N=10
1	66.0%	54.1%	52.9%	52.7%	52.7%	52.6%	52.6%	52.6%	52.5%	52.5%
2		18.1%	14.7%	14.2%	14.1%	14.1%	14.0%	14.0%	14.0%	14.0%
3			6.3%	4.6%	4.3%	4.2%	4.2%	4.1%	4.1%	4.1%
4				3.2%	2.0%	1.7%	1.6%	1.6%	1.6%	1.6%
5					2.3%	1.3%	1.0%	0.9%	0.9%	0.9%
6						2.0%	1.0%	0.8%	0.7%	0.7%
7							1.8%	0.9%	0.7%	0.6%
8								1.7%	0.8%	0.6%
9									1.5%	0.8%
10										1.4%
Total	66.0%	72.1%	73.9%	74.7%	75.3%	75.8%	76.2%	76.6%	77.0%	77.3%
MSE	116.81	112.99	112.55	112.43	112.38	112.33	112.30	112.27	112.24	112.22

Note that the values shown in a column may not sum to the total shown due to rounding.

Comment: The covariances are on a basis of number of games lost for the baseball data; they are based on a model of the baseball data shown at page 661 of "Credibility With Shifting Risk Parameters, Risk Heterogeneity, and Parameter Uncertainty," by Howard C. Mahler, PCAS 1998. (This is a mixture of two Markov Chains with different rates of shifting of risk parameters.) This model has smoothed out the peculiarities of the covariances of the data that are due to random fluctuation. Thus we see a regular pattern of credibilities. (Contrast here to Table 16 in the syllabus reading.) More distant years get less credibility than more recent years, declining in a nice pattern.

However, there is also an "edge effect"; the most distant year used tends to get more weight since it is correlated with even more distant years. For example, when using five years of data, for example 1951 to 1955, then the most distant year, 1951, contains useful information about years 1950, 1949, 1948, etc. Thus 1951 is given more weight than 1952; 2.3% > 2.0%.

1.35. In the case of credibilities that can differ by year, using three years of data is a special case of using four years of data, with the credibility of the most distant year constrained to be zero. Therefore, the best credibilities for four years of data do at least as well and probably better than the best credibilities for three years of data. As expected, we see the minimum expected squared errors decline as we used more years of data. After a while we reach of point of diminishing improvement. For example, ten years with a MSE of 57.04 is only slightly better than nine years with a MSE of 57.05.
In each case, constraining the credibilities to be the same by year is a special case of allowing the credibilities to differ. Thus allowing the credibilities to vary does at least as well and probably better than requiring the credibilities to be the same by year. In fact, the mean squared errors are smaller for the case where the credibilities differ. (For N =1 the two methods are the same.) For example, for four years of data, 57.36 is better than 57.45.
In the case of credibilities that are the same by year, using fewer years of data is a not a special case of using more years of data. Distant years should be given very little weight, but we are requiring all years to be given the same weight. While adding year of data may be better, with shifting risk parameters, eventually adding years of data will be worse. In this case, the mean squared errors improve through 6 years of data. However, after that the mean squared errors increase. For example, using 7 years of data has a mean squared error of 57.40, worse than the 57.36 for 6 years of data.
Comment: The results shown in this question are based the covariance between different years of data being: $Cov[X_i, X_j] = (10)\ 0.88^{|i-j|} + 50\ \delta_{ij}$, where δ_{ij} is zero if $i \neq j$ and one if $i=j$.
Parameters shift more slowly than in the baseball example.
Given the covariances, one could solve for the least squares credibilities.
For example, using 3 three years of data: $Z_1 = 8.3\%$, $Z_2 = 9.9\%$, and $Z_3 = 12.1\%$.
Using 3 three years of data instead with equal weights, each year is weighted 10.1%.

1.36. (a) The purpose is to test whether risk parameters shift over time.
In other words, determine whether inherent loss potential (L%) is shifting over time for each team.
(b) The test is applied separately to the data of one baseball team.
H_0: The expected losing percentage is the same over time for this team.
Compute this team's losing percentage over the whole experience period (of 60 years).
Then group data for that team into appropriate intervals; Mahler groups the 60 years into 5 year non-overlapping intervals.
Calculate for each interval: $(A - E)^2/E$,
where A = actual observation = (5 year mean losing percentage)(5 years)(150 games),
and E = expected observation = (60 year mean losing percentage)(5 years)(150 games).
Sum up the contributions for all intervals in order to get the chi-square statistic.
Compare to the Chi-Square Distribution with number of degrees of freedom equal to the number of intervals minus one; in the paper Mahler compares to the Chi-Square with 11 degrees of freedom.
If the statistic is greater than the critical value for the appropriate significance level, for example 5%, then for this team we reject the null hypothesis that parameters do not shift over time.
Comment: See Table 4 in the paper. For each of the 16 teams, the p-value was less than 0.2%.

1.37. Statement 1 is backwards. As the delay in receiving data increases, its predictive value decreases and the credibility decreases.
Statement 2 is backwards.
Statement 3 is backwards. If one gives each year equal weight, as the number of years increases, eventually the accuracy will decrease. (If one determines separate optimal credibilities by year, as the number of years increases, eventually the accuracy will no longer increase significantly.)
Comment: The conclusions in this exam question are those of Bizarro-Mahler on a planet opposite of the real world.

1.38. 1. Least squared error.
Minimize the squared error or the mean squared error between the observed and predicted results. Analogous to Buhlmann credibility.
2. Small chance of large errors.
Minimize the probability that the observed results will be more than some chosen % different from the predicted. Analogous to classical credibility.
3. Meyers/Dorweiler.
Minimize, in other words make equal to zero, the correlation between:
$\frac{\text{observed}}{\text{predicted}}$ and $\frac{\text{predicted}}{\text{overall average}}$.
Use some correlation measure; Mahler uses Kendall's statistic, which counts inversions.
Meyers/Dorweiler results differ from the others because it's concerned with patterns rather than sizes of errors.
Comment: The results from the first two methods are usually very similar.

1.39. 1. False; should say 75% rather than 50%. See page 252 (and Appendix E.)
2. True. See page 271 (and Appendix B.)
3. False! See page 277.
Comment: The original statement #1 was false from Appendix E, no longer on the syllabus. "Credibility methods reduce the squared error between the observed value and the estimated/ predicted value to a greater extent than they reduce the squared error between the true mean and the estimated predicted mean."

1.40. For an individual team, the number of games lost is Poisson with mean $n\lambda$.
Therefore, for an individual team, the variance in number of games lost is also $n\lambda$.
The losing percentage is the number of games lost divided by n.
Therefore, the variance in losing percentage is: $n\lambda/n^2 = \lambda/n$.
Thus the expected value of the process variance in losing percentage is:
$(0.5)(0.4/200) + (0.5)(0.6/200) = 0.0025$.
The variance of the hypothetical mean losing percentages is:
$(0.5)(0.4 - 0.5)^2 + (0.5)(0.6 - 0.5)^2 = 0.01$.
The observed variance in losing percentages is:
$\{(75/200 - 0.5)^2 + ... + (94/200 - 0.5)^2\}/10 = 0.0138$.
Therefore, we can back out the amount of variance due to shifting risk parameters as:
$0.0138 - 0.0025 - 0.01 = 0.0013$.
The percentage of the total variance due to shifting risk parameters is: $0.0013/0.0138 =$ **9.4%**.
Comment: See page 297 of Mahler's Appendix D, no longer on the syllabus.
The paper observed 60 years and averaged the observed variances for the individual years.
The estimate from just one year of data is not reliable.
Also the paper assumed a Binomial Model.
There is no need to divide up the variance into three pieces in order to calculate credibilities.
This is something which may help your understanding, but is not necessary.
I would not have done this if I were rewriting the paper today.

1.41. For Z = 50%, the predicted loss ratios are:
For 1993: (65% + 75%)/2 = 70%. For 1994: (65% + 70%)/2 = 67.5%.
For 1993: (65% + 65%)/2 = 65%. For 1994: (65% + 60%)/2 = 62.5%
The total of the squared errors is: $(70 - 70)^2 + (65 - 67.5)^2 + (60 - 65)^2 + (55 - 62.5)^2 = 87.5$.
For Z = 0, the predicted loss ratios are all 65%, and the total of the squared errors is:
$(70 - 65)^2 + (65 - 65)^2 + (60 - 65)^2 + (55 - 65)^2 = 150$.
Since Z = 50% has a lower sum of squared errors than Z = 0, I agree with the client.
Comment: In practical applications one would not apply the least squares criterion to only 5 years of data from one insured. One could apply it to years of data from many similar insureds of similar size in order to determine which value of Z performs well.

1.42. a. Ratio 1 = (Team's actual losing percentage)/(Team's predicted losing percentage).
Ratio 2 = (Team's predicted losing percentage)/(grand mean of 50%).

b. Ratio 1 $\Leftrightarrow$ The loss ratio to modified premium (loss ratio to standard premium).

Ratio 2 $\Leftrightarrow$ The experience modification.
Comment: See Section 7.3 in the paper by Mahler.
Part b of this exam question is from Appendix B, which is no longer on the syllabus.
However, it would be a good idea to know this anyway.

1.43. New estimate = Z Latest year of data + (1 - Z)(prior estimate).
We start with an estimate of 60%.
Estimate of 1996 using data from 1995: (30%)(70%) + (1 - 30%) (60%) = 63%.
Estimate of 1997 using data from 1996: (30%)(80%) + (1 - 30%) (63%) = 68.1%.
Estimate of 1998 using data from 1997: (30%)(90%) + (1 - 30%) (68.1%) = 74.67%.
Estimate of 1999 using data from 1998: (30%)(100%) + (1 - 30%) (74.67%) = **82.269%**.
Comment: See Section 9.1 in the paper by Mahler.

1.44. The best that can be done using credibility to combine two estimates is to reduce the mean squared error between the estimated and observed values to 75% of the minimum of the squared errors from either relying solely on the data or ignoring the data.
(75%)(80) = **60**.
Comment: See Section 8.5 in the paper by Mahler.

1.45. a. To determine whether the data for each team was drawn from the same probability distribution. In other words, to determine whether an "inherent difference" in loss % exists between teams.
b. The variance in losing percentage in 2500 games would be: (0.5)(0.5)/2500 = 0.0001.
standard deviation is: 1%.
If the data for each team was drawn from the same probability distribution, we would expect to see about 95% of the teams results between: 50% ± (2)(1%) = 48% to 52%.
In this case only 1 out of 5 teams is in that range.
(Two of the teams have losing percentages 5 standard deviations from average, while two team have losing percentages 10 standard deviations from average!)
Thus we conclude that the teams differ.
c. The purpose is to test whether risk parameters shift over time. In other words, determine whether inherent loss potential (L%) is shifting over time for each team.
d. The Bermuda Captives have an overall losing percentage of 50%.
The observed number of losses per 5 years for this team is: (5)(100)(50%) = 250.
(For this team this happens to also be the a priori mean.)
Chi-Square statistic is: $(160 - 250)^2/250 + (170 - 250)^2/250 + (294 - 250)^2/250 +$
$(330 - 250)^2/250 + (296 - 250)^2/250 = 99.808$.
(This statistic has: number of groups - 1 = 5 - 1 = 4 degrees of freedom.)
Since 99.808 > 9.488, we reject the null hypothesis at the 95% confidence level (5% significance level). We conclude that the risk parameters shift over time, at least for the Bermuda Captives.
e. The purpose is to test whether risk parameters shift over time.
f. For each year we have a vector of length 5 of losing percentages by team.
For the one year differential, we examine the correlation of the 24 sets of pairs of data separated by one year: year 1 versus year 2, year 2 versus year 3, etc.
Mahler uses Kendall's tau to measure the correlation.
We take the average of these 24 correlations for the one year differential.
We do the same for the two year differential, using the correlation of the 23 sets of pairs of data by two years. We take the average correlation for the two year differential.
We do the similar calculation for the other differentials in years.
If the risk parameters do not shift over time, the average correlation should not differ significantly between the one year differential, two year differential, and so forth. If the risk parameters shift over time, the average correlation should be highest for the one year differential, second highest for the two year differential, and so forth.
Given the results of the Chi-Square Test for the Bermuda Captives, the likely conclusion of this test is that the risk parameters shift over time.
Comment: See Section 4 in the paper by Mahler.

1.46. (a) V(Z) = the expected squared error between the observation and predication.
τ^2 = between variance.
C(k) = covariance for data for the same risk, k years apart = "within covariance."
Δ = the length of time between the latest year of data used and the year being estimated.
If $\Delta = 1$, then there is no delay in receiving information.

(b) $V(Z) = Z_1^2(\tau^2 + C(0)) + Z_2^2(\tau^2 + C(0)) + 2 Z_1Z_2(\tau^2 + C(1)) - 2 Z_1(\tau^2 + C(2))$
$- 2 Z_2(\tau^2 + C(1)) + \tau^2 + C(0)$

$V(Z) = 0.9 Z_1^2 + 0.9 Z_2^2 + 1.2 Z_1Z_2 - 0.9 Z_1 - 1.2 Z_2 + 0.9.$

Setting the derivative of V with respect to Z_1 equal to zero:

$0 = 1.8Z_1 + 1.2Z_2 - 0.9.$

Setting the derivative of V with respect to Z_2 equal to zero:

$0 = 1.8Z_2 + 1.2Z_1 - 1.2.$ Solving, $Z_1 = 10\%$ and $Z_2 = 60\%$.

Therefore, the weight given to the overall mean is: 1 - 10% - 60% = 30%.
Therefore, the estimate for the year 2000 is: (10%)(40%) + (60%)(45%) + (30%)(50%) = **46%**.

Alternately, for two different years, $Cov[X_i , X_j] = \tau^2 + C(|i - j|)$.

For example, $Cov[X_{1998}, X_{2000}] = \tau^2 + C(2) = 0.1000 + 0.3500 = 0.45.$

For a single year of data, $Cov[X_i, X_i] = Var[X_i] = \tau^2 + C(0) = 0.1000 + 0.8000 = 0.9000.$

A covariance matrix is:

1998	0.90	0.60	0.45
1999	0.60	0.90	0.60
2000	0.45	0.60	0.90

$\Sigma Z_i Cov[X_i , X_j] = Cov[X_i, X_{N+\Delta}]$, where we are predicting year $N + \Delta$, using years 1 to N.
Using data for Years 1998 and 1999 to Predict Year 2000, the equations are:
$0.9Z_1 + 0.6Z_2 = 0.45.$
$0.6Z_1 + 0.9Z_2 = 0.60.$
The coefficients on the lefthand side are the first two rows and the first two columns of the covariance matrix, since we are using data from Years 1998 and 1999. The values on the righthand side are the first two rows of column three, since we are predicting year 2000.
Proceed as before.
Comment: See page 263 and Equation 11.3 in Mahler. We give 1999 more weight than 1998. Since N = 2, we do not use the information from 1997. In order to determine a least squares credibility to assign to 1997, we would need to be given C(3).
Mahler works with losing percentages. If one converted the data and the grand mean to losing percentages, the predicted losing percentage in 2000 would be:
(10%)(60%) + (60%)(55%) + (30%)(50%) = 54% = 1 - 46%.

1.47. a. 1. Least squares - minimize the total squared error between actual and predicted result.
2. Small chance of large error - minimize the likelihood that any one actual observation will be a certain % different from the predicted result.
3. Meyers/Dorweiler - minimize the correlation between the ratio of actual/predicted and the predicted/average actual.
b. Meyers/Dorweiler is different from the first two which focus on minimizing prediction error.
In contrast, Meyers/Dorweiler focuses on the pattern of the errors.

1.48. A. The principles for shifting risk parameters are:
Statements A and E. Years that are closer together have a higher correlation than years that are further apart, so credibility should be higher for more recent years. .
Statements B and C. Delays in receiving data make the experience less useful and it should receive less credibility.
The use of the current year of data to help predict next year increases the accuracy of the estimate, so statement D is true.

1.49. Correlation test:

- Group data by pairs based on time lag
- Calculate correlation for each pair
- Calculate the average correlation by time lag
- If the correlation decreases as time lag increases, then risk parameters shift over time.

Chi-Square Test:
Null Hypothesis - H_0: risk parameters do not shift over time

- Group data into appropriate intervals
- Calculate the overall expected value
- Then calculate for each interval, $(A - E)^2/E$,
  where A = actual observation and E = expected observation
- Sum up the contributions for all intervals in order to get the chi-square statistic.
- If the total statistic is greater than the critical value for number of intervals -1
  degrees of freedom, then reject the null hypothesis that parameters do not shift over time.

1.50. D. Under Plans 1 and 3, the risks with higher mods have larger errors.
Under Plan 2, there is no correlation between the mods and the errors; underwriters would be indifferent between writing credit or debit risks.
Therefore, Plan 2 does best under the Meyers/Dorweiler criterion.
Plan 3 has the smallest average squared error, so Plan 3 is preferred under the Least Squared Error Criterion.
Comment: This past exam questions was not really properly put together.
If plans 1 and 2 produce the same modification for the each risk, and they should have the same errors; they should perform the same.
Actual experience rating plans are tested on thousands of risks.
Based on this data, Plan 1 is a bad experience rating plan.

1.51. a. The weight given to accident year 2001 losses in accident year 2002's estimate is Z = 10%. We work out the weight given to accident year 2001 losses in accident year 2005's estimate as follows:

$P_{2005} = Z X_{2004} + (1 - Z) P_{2004} = Z X_{2004} + (1 - Z) \{Z X_{2003} + (1 - Z) P_{2003}\} =$

$Z X_{2004} + (1 - Z) Z X_{2003} + (1 - Z)^2 P_{2003} =$

$Z X_{2004} + (1 - Z) Z X_{2003} + (1 - Z)^2 \{Z X_{2002} + (1 - Z) P_{2002}\} =$

$Z X_{2004} + (1 - Z) Z X_{2003} + (1 - Z)^2 Z X_{2002} + (1 - Z)^3 P_{2002} =$

$Z X_{2004} + (1 - Z) Z X_{2003} + (1 - Z)^2 Z X_{2002} + (1 - Z)^3 \{Z X_{2001} + (1 - Z) P_{2001}\} =$

$Z X_{2004} + (1 - Z) Z X_{2003} + (1 - Z)^2 Z X_{2002} + (1 - Z)^3 Z X_{2001} + (1 - Z)^4 P_{2001}$.

The weight given to X_{2001} is: $(1 - Z)^3 Z$.

When Z = 10%, $(1 - Z)^3 Z = (1 - .1)^3 (.1) = 7.29\%$.

The difference in the weight given to accident year 2001 losses in accident year 2002's estimate and the weight given to accident year 2001 losses in accident year 2005's estimate is:
10% - 7.29% = **2.71%**.

b. If there is a significant shift in risk parameters, then older years of data become much less predictive. Therefore, less weight is given to 2001 losses in the estimate of 2005 than when there was less shifting in risk parameters. This will make the difference in part (a) **increase**.

Comment: See page 255 of Mahler.

1.52. 1) χ^2 (Chi-Square Method).
The test statistic is: S(Actual – Expected)^2 / Expected.
Null Hypothesis: Expected number of claims is the same for each year.
Calculate the test statistic which sums the relative errors (squared)
Compare the test statistic to the critical value (from χ^2 distribution) with n-1 degrees of freedom.
If test statistic > critical value, then reject null and accept alternative, that risk parameters shift over time.
2) Correlation Test
Group data by pair for all possible combinations of time lag.
Calculate the correlation for each possible pair.
If the correlation decreases as the time lag increases, then there is a shifting of risk parameters over time.
Comment: Here is the result of the Chi-Square Test.
You would want the observed and assumed columns to add to the same amount,
thus the expected number of claims should be 501 rather than 500 as shown in the question.
(Using 500 would result in a statistic of 33.80.)

Year	Observed Number	Assumed Number	Chi Square
1997	475	501	1.35
1998	420	501	13.10
1999	460	501	3.36
2000	500	501	0.00
2001	490	501	0.24
2002	525	501	1.15
2003	515	501	0.39
2004	510	501	0.16
2005	540	501	3.04
2006	575	501	10.93
Sum	5,010	5,010	33.71

There are 10 years, and 10 - 1 = 9 degrees of freedom.
For 9 degrees of freedom, the critical value for 1/2% is 23.589.
(Value taken from the Chi-Square Table attached to a preliminary exam.)
Since 33.71 > 23.589, we reject the null hypothesis at 1/2%.
One could group data by interval of a few years (Mahler uses groups of 5 years over a period of 60 years.) He applies the test separately to each of the 16 teams.
The Chi-Square Test is shown by Mahler in his Table 4.
In item 2 of the solution, one would be calculating autocorrelations as per Time Series. See Introductory Times Series with R, by Cowpertwait & Metcalfe, not on the syllabus of this exam.
While that is a similar idea to what is done in the syllabus reading, it is not quite the same.
In the paper, one looks at the correlations of the vector of the losing percentages (each length 8) for 1901 and 1902. Then for 1902 and 1903. Then for 1903 and 1904. etc.
Then we average these results. This is the listed correlation for separation of 1 year.

Here the sample correlation is:

$$r = \frac{\text{estimated covariance of X and Y}}{\sqrt{(\text{estimated standard deviation of X})\ (\text{estimated standard deviation of Y})}}$$

$$= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}.$$

For data separated by one year, the two vectors are:

X = (475, 420, 460, 500, 490, 525, 515, 510, 540). $\bar{X}$ = 492.778.

Y= (420, 460, 500, 490, 525, 515, 510, 540, 575). $\bar{Y}$ = 503.889.

(Which you call X and which you call Y is irrelevant.)

$X - \bar{X}$ = (-17.778, -72.778, -32.778, 7.222, -2.778, 32.222, 22.222, 17.222, 47.222).

$Y - \bar{Y}$ = (-83.889, -43.889, -3.889, -13.889, 21.111, 11.111, 6.111, 36.111, 71.111).

$\sum (X_i - \bar{X})(Y_i - \bar{Y})$ = 9127.8.

$\sum (X_i - \bar{X})^2$ = 10,805.6.

$\sum (Y_i - \bar{Y})^2$ = 16,138.9.

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} = \frac{9127.8}{\sqrt{(10{,}805.6)(16{,}138.9)}} = 0.691.$$

Alternately, one can fit a linear regression between X and Y using the stat functions on a calculator.

The output r is the desired correlation.

Similar to Mahler's Table 5, the autocorrelations for the data in this question are:

Separation	Correlation
1	0.691
2	0.528
3	0.717
4	0.470
5	0.654
6	0.770
7	-0.220

One would need more years of data, in order to draw a reliable conclusion from the correlation test.

The paper has 60 years of data rather the 10 years here.

1.53. a) The expected number of claims in a year are: 1.5 times exposures.
The observed number of claims in a year are: (observed frequency)(exposures).

Year	Observed	Exposures	Expected	((Observed - Expected)^2)/Expected
	Number		Number	
2011	177	118	177.00	0.000
2010	224.4	132	198.00	3.520
2009	157.3	121	181.50	3.227
2008	174.4	109	163.50	0.727
2007	126.1	97	145.50	2.587
Sum	859.2	577	865.50	**10.060**

The Chi-square statistic is 10.060. 10.060 > 9.49, so we reject the null hypothesis.

$\Rightarrow$ The different years are not all drawn from the same Poisson Distribution.

$\Rightarrow$ The parameters are shifting over time.

b) Compute the correlations between different pairs of years of data for individuals.
Then average the correlations for years separated by a given number of years.
If the correlations decline as the separation increases, this indicates that parameters are shifting over time; the quicker the decline the more quickly parameters are shifting.
Comment: We have 5 - 1 = 4 degrees of freedom; the 5% critical value is 9.49.
See Tables 4 and 5 in Mahler.

1.54. (a) H_0: The expected frequency is 1.2% for each year.

H_1: Not H_0.

For 2011 the observed number is: (11,000)(0.010) = 110,
and the expected number is: (11,000)(0.012) = 132.

Contribution is: (Observed - Expected)^2/ Expected = (110 - 132)^2 / 132 = 3.6667

Year	Exposures	Frequency	Observed	Expected	Chi-Square Contribution
2010	9,500	0.011	104.5	114	0.79167
2011	11,000	0.010	110	132	3.6667
2012	13,000	0.013	169	156	1.0833
2013	10,500	0.012	126	126	0
2014	12,000	0.010	120	144	4
					9.54

Since the Chi-Square statistic is 9.54 > 9.49, at the corresponding significance level we reject the null hypothesis. This is evidence that (expected) claim frequency is shifting over time.
(b) For a given risk, compute the correlations between pairs of different years of data.
Average the correlations for all pairs with the same number of years between them.
If these average correlations decline quickly towards zero as the distance between pairs of years increases, then parameters are shifting at a significant rate.
Comment: 9.49 is the 5% critical value for a Chi-Square Distribution with 4 degrees of freedom.

Section 2, Bailey and Simon, Merit Rating[1] [2]

In their classic paper, Bailey and Simon use Merit Rating data to determine the credibility to assign to the experience of a single private passenger car. The most important parts of this concise paper are Tables 2 and 3, and their conclusions.

A key concept is that when using credibility, Z is the discount compared to average given to an insured who is claims-free. This credibility varies by class and the number of years claims-free.

Merit Rating is a very simplified form of Experience Rating. As has been discussed previously, one way to analyze Experience Rating is to compare experience during a prior and subsequent period in order to determine how a plan would have worked in the past.[3]

Bailey-Simon compare a prior three year period to a subsequent one year period for Private Passenger Automobile Insurance in Canada.[4] They compare the subsequent frequency for groups with different numbers of years claims-free.[5] They found that Merit Rating has useful predictive ability beyond that of class and territory.[6]

---

[1] "An Actuarial Note on the Credibility of Experience of a Single Private Passenger Car," by Robert A. Bailey and Leroy J. Simon, PCAS XLVI, 1959, pp. 159-164.
Including discussion of paper: Hazam, W. J., PCAS XLVII, 1960, pp. 150-152.
CAS Learning Objectives A1-A2.

[2] This excellent paper has been on the exam syllabus since shortly after it was written.
Thus it has been on the syllabus for over half a century!

[3] However, their method differs from those discussed by Gillam, Venter, and Mahler in their syllabus readings.

[4] For the prior period they only record how many years a car has been claims-free prior to the "present".

[5] In their much less important Table 4, they look at loss ratios, which involve dollars of loss.

[6] Since the average annual frequency is low, three years of private passenger auto data for a single car contains a lot of noise and relatively little signal; the credibilities are much smaller than those for a large commercial insured.

Merit Rating Plans:[7]

The Canadian Merit Rating Plan in place when Bailey and Simon wrote their paper is relatively simple.[8]

Those who are claim-free for only one year get a discount of 10%, Group Y.
Those who are claim-free for only 2 years get a discount of 20%, Group X.
Those who are claim-free for 3 or more years get a discount of 35%, Group A.[9] [10]
These discounts are off the base rate for those who are not claims-free, Group B.

Group A ⇔ no claims in the 3 year experience period has a claim.

Group X ⇔ the most recent 2 years claims free,
while the earliest year in the 3 year experience period has a claim.
For example, Merit Rating a 1958 policy: 1956 and 1957 claim free, but 1955 has a claim.

Group Y ⇔ the most recent 1 year claims free,
while the second year in the 3 year experience period has a claim.
For example, Merit Rating a 1958 policy: 1957 claim free, but 1956 has a claim.

As stated at the first page of Bailey-Simon:
Earned premiums are converted to a common rate basis by use of the relationship in the rate structure that A: X: Y: B = 65: 80: 90: 100.
Bailey-Simon put premiums on the level that would have been charged for Merit Rating Class B, those who are not claims free. For example, if the actual premiums for Merit Rating Group A were 6.5 million, then on a Group B basis they would be: 6.5 / (1 - 35%) = 10 million.

Currently in many states in the U.S., many insurers apply a simple form of Experience Rating to private passenger automobile insurance, often called Safe Driver Insurance Plans (SDIP).[11] They are usually somewhat more complex than the Merit Rating Plan discussed in Bailey-Simon. The number of moving traffic violations and/or at-fault claims will be determined for each driver over some recent period such as 3 or 5 years.[12] Those drivers with worse records will pay more than average, while those drivers with better records will pay less than average for their insurance.

---

[7] For background. You should not be tested on the details of Merit Rating Plans or Safe Driver Insurance Plans.
[8] See "The Canadian Merit Rating Plan for Individual Automobile Risks," by Herbert E. Wittick, PCAS 1958.
[9] See page 159 of Bailey-Simon.
[10] All operators of a vehicle must be claim-free in order to get the discount; we are only looking at liability claims.
[11] See pages G-6 to G-9 of the ISO Personal Automobile Manual, not on the syllabus of this exam.
[12] For example, in Massachusetts as of 2004, the Safe Driver Insurance Plan (SDIP) uses 5 years of data on minor and major traffic law violations, and minor and major at fault accidents. Thus while this plan is largely based on frequency, there is a small component that depends on severity. Driving under the influence results in a larger surcharge than speeding. There are a number of additional complicated details in how this specific plan works.

The actuarial theory behind such plans is similar to that for the CGL and Workers Compensation Experience Rating Plans. However, the details differ. There is much smaller volume of data generated by a single private passenger automobile. The Canadian Merit Rating plan only uses frequency not severity.[13] Also most SDIPs use moving violations, so that someone who has no losses may still get a surcharge.

Also, the Canadian Merit Rating Plan differs from a plan that just added up the number of claims over the last three years. For example, let us assume the experience period is 1955, 1956, and 1957, and we are rating a 1958 policy.

	Number of Claims by Year		
Car	1955	1956	1957
1	0	0	0
2	1	0	0
3	2	0	0
4	1	1	0
5	0	1	0
6	0	0	1
7	1	1	1

Car 1 is put in Group A and gets the 35% discount.
Cars 2 and 3 are both put in Group X and get a 20% discount.
Cars 4 and 5 are both put in Group Y and get a 10% discount.
Cars 6 and 7 are both put in Group B and get a no discount.

Note that insureds with different numbers of claims over the last three years may be charged the same amounts by the Canadian Merit Rating Plan. Also insureds with the same numbers of claims over the last three years can be charged different amounts by the Canadian Merit Rating Plan.[14]

In order to be put in Group A, the insured and/or principal operator must have been licensed for at least three years. In order to be put in Group X, the insured and/or principal operator must have been licensed for at least two years.

---

[13] As the volume of data declines, so does the optimal accident limit in an Experience Rating Plan; as the accident limit gets very low, the plan approaches a frequency only plan.

[14] While a plan that used the number of claims over three years might have more predictive power, Bailey-Simon is not comparing plans, but just trying to determine the predictive power of the plan then used in Canada.

Claims-Free Discount:

As mentioned, those who are claims-free get a discount; the longer the claims-free period the larger the discount. There are two ways to look at the size of the discounts. First there is the discount from the base rate. In the case of the Canadian Merit Rating Plan, the discount is off of the rate charged Group B, which is higher than average.

For example, from Table 1 in Bailey-Simon, for Class 1, the average discount is:
$$\frac{(35\%)(159{,}108) + (20\%)(7910) + (10\%)(9862)}{194{,}106} = 30.01\%.$$
The average Class 1 premium at Group B rates (in other words prior to any discounts) is: 194,106,000 / 3,325,714 = $58.36.[15] However, after the effect of the Merit Rating discounts, the average rate paid is only: ($58.36)(1 - 0.3001) = $40.85

The advertised discount for three years claims-free is 35%; however, the discount off of the average rate is: 1 - (1 - 35%)/(1 - 30.01%) = 7.1%.

The advertised discount for two years claims-free is 20%; however, the surcharge above the average rate is: (1 - 20%)/(1 - 30.01%) - 1 = 14.3%.[16]

The Group X insureds are claims-free for only two-years; they had a claim three year ago. Thus they are worse than the insureds who have been claims-free for at least 3 years, and Group X pays more than average.

In the context of credibility theory, actuaries are interested in the experience and discounts with respect to average. Bailey-Simon will compute claims-free discounts compared to average for those who have been claims-free at least one year (A +X+Y), claims-free at least two years (A + X), and claims-free at least three years (A).

[15] Remember this is for 1957 and 1958. Also these are Canadian dollars.
[16] Public acceptance of the plan is much better if one advertises discounts off of the base rate, rather than making it obvious that some insureds are paying more than average.

Table 1, Bailey-Simon:

Private Passenger Auto Liability data from Canada (excluding Saskatchewan) from Policy Years 1957 and 1958.[17] The data is divided into five classes.[18] Their analysis will be performed separately on each class.

Within each class are four Groups, based on how long they have been claims-free:

A	3 or more years claims-free
X	2 years claims free
Y	1 year claims free
B	0 years claims-free
A + X	2 or more years claims-free
A + X + Y	1 or more years claims-free

We have exposures (earned car years), premiums (earned premiums at present Group B rates), and claims (number of claims incurred).

Then the number of claims is divided by premiums in $1000, rather than exposures.
For example, for Group A in Class 1: 217,151 / 159,108 = 1.365.

Bailey and Simon "have chosen to calculate Relative Claim Frequency on the basis of premium rather than car years. This avoids the maldistribution created by having higher claim frequency territories produce more X, Y, and B risks and also produce higher territorial premiums." [19]

---

[17] For an individual car, assume we have a policy written during 1958.
Then the Merit rating class (A, X, Y, B) would have been based on 3 past years of data.
(This may be 1955, 1956, and 1957 without a gap in obtaining data for Merit Rating.)
I believe the Class (1, 2, 3, 4, or 5) is the one for the 1958 policy.
For example, some insured who were in Class 1 (Pleasure - no male operator under 25) during 1958 would have been in different classes during 1955, 1956, or 1957.
The data was not scrubbed to remove insureds who switched classes over the relevant period.

[18] Class 1 is Pleasure - no male operator under 25.
Class 2 is Pleasure - Non-principal male operator under 25.
Class 3 is Business use.
Class 4 is Unmarried owner or principal operator under 25.
Class 5 is Married owner or principal operator under 25.

[19] Average premiums by territory within a class will vary due to differences in frequency, differences in severity, as well as to some extent incorrect territory relativities.

The use of premium based frequencies avoids double counting. If instead one used caryears as the denominator of frequency, the credibility calculation would account for both "within territory differences" and "between territory differences". However, territory relativities already account for the between territory differences.[20] [21] Compounding territory relativities with credibility would double count the between territory differences, and therefore the credibility would be overstated. Therefore, claims free drivers would be undercharged while other drivers would be overcharged.[22]

In order to remove this "double counting", we use premium as exposure. An assumption for this is that the premium differences should reflect the true pure premium differences between the territories.[23]

The premiums are put on the basis of Group B, in other words prior to any discounts for Merit Rating.[24] [25] We are removing the effects of any current discounts due to Merit Rating in order to estimate the indicated discounts, rather than estimating a change in the current discounts. As discussed previously, Bailey-Simon will be estimating discounts compared to average.

Then Bailey-Simon divide the premium based frequency for a group by that for the whole class, in order to to get the relative claim frequency.
For example, for Group A in Class 1: 1.365/1.484 = 0.920.

---

[20] Hazam points out in his discussion, we need to assume that the territory relativities are correct and reflect differences in frequency (per caryear) between the territories.

[21] If for some reason territorial rating is not used in spite of differences between territories in frequency, then there would be no double counting resulting from using car years as the denominator of frequency. See 8, 11/15, Q.1. In the absence of territorial rating, the appropriate Merit Rating credibilities are larger than they otherwise would be. In general, the less accurate the class/territory plan and relativities, the more predictive work there is for experience rating to do, and thus the larger the appropriate credibility for experience rating.

[22] Subsequently I have a detailed example illustrating this.

[23] See the discussion by Hazam.

[24] One could use another group as the base, and the relative claim frequencies would be the same.

[25] “Earned premiums are converted to a common rate basis by use of the relationship in the rate structure that A:X:Y:B = 65:80:90:100.” In other words as mentioned previously, those in Merit Rating Class A currently get a 35% discount with respect to Merit Rating Class B, those in Merit Rating Class X currently get a 20% discount with respect to Merit Rating Class B, and those in Merit Rating Class Y currently get a 10% discount with respect to Merit Rating Class B. Thus for example, in order to be put on a Merit Rating Class B level, premium from an insured in Merit Rating Class A would be divided by 0.65.

Here is the calculation for the Class 2 data shown in their Table 1:

Class 2 - Pleasure - Non-principal male operator under 25

Group	Years Claims-Free	Group B Premium	Number of Claims	Freq.	Rel. Freq.
A	3 or more	11,840,000	14,506	1.225	0.932
A+X	2 or more	12,552,000	15,507	1.235	0.940
A+X+Y	1 or more	13,496,000	16,937	1.255	0.955
Total		15,488,000	20,358	1.314	1.000

We need to combine Groups A and X in order to get those who are claims free for 2 years or more:
Claims-free at least 2 years = (3 or more years claims-free) + (2 years claims-free).
A + X + Y is those who are claims free for 1 year or more.[26]

Table 2, Bailey-Simon:

In their very important Table 2, for each class separately, the credibilities for one, two, and three years of data are calculated from the indicated claims-free discount compared to average.
For example, for Class 2, the overall frequency on a premium basis in Table 1 is:
20,358 / 15,488 = 1.314.
The frequency on a premium basis for Group A (3 years claims-free) is: 14,506 / 11,840 = 1.225.
Thus the indicated experience modification for Group A is: 1.225/1.314 = 0.932.
This is the relative claim frequency also shown in Table 1.

Then the claims free discount is: 1 - 0.932 = 6.8%.
This is the estimated credibility for three years of data shown in Table 2 for Class 2.
In general, for a given class and numbers of years or more claims-free:

$$\mathbf{1 - Z = M = \frac{Premium\ Based\ Claim\ Frequency\ for\ Those\ Claims\text{-}Free\ N\ or\ More\ Years}{Overall\ Premium\ Based\ Claim\ Frequency\ for\ the\ Class}}.$$

Calculating in this manner the credibilities for one, two or three years is the most commonly asked exam question on this paper. For Class 2:[27]
The one-year credibility is: 1 - 1.255/1.314 = 1 - 0.955 = 4.5%.
The two-year credibility is: 1 - 1.235/1.314 = 1 - 0.940 = 6.0%.
The three-year credibility is: 1 - 1.225/1.314 = 1 - 0.932 = 6.8%.

---

[26] If they use the letters for the groups, we expect them to tell you their meaning in the question.
A = claims-free 3 or more years. X = claims-free 2 years. Y = claims-free one year. B = not claims-free.

[27] These match what is shown for Class 2 in Table 2 in Bailey-Simon.
These do not match the then current discounts in the Canadian plan, which as discussed were with respect to the base rate rather than with respect to average and were the same regardless of class.
The then current discounts in the Canadian plan preceded the study by Bailey and Simon.

Ratio of Credibility to Frequency:

In addition, in Table 2, for each class Bailey-Simon takes the ratio of the three-year credibility to the frequency.[28] For example for Class 2, the overall exposure based frequency is: 20.358 / 168,998 = 0.120. Then the ratio of the 3-year credibility to frequency is: 0.068 / 0.120 = 0.567.

The credibilities depend on the Expected Value of the Process Variance (EPV) and the Variance of the Hypothetical Means (VHM).[29] If each insured is Poisson, then the EPV is equal to the average frequency for the class. In any case, the EPV should be roughly proportional to the mean frequency.

If the Buhlmann Credibility formula holds, then the three-year credibility is $Z = 3 / (3 + K)$, with $K = EPV / VHM$.[30] [31]

For K big compared to 3, as it is in the situations in Bailey Simon: $Z \cong 3/K = (3)(VHM / EPV)$.

Let m be the overall mean frequency, which is also the mean of the hypothetical mean frequencies.
Assume the EPV is (approximately) proportional to the overall mean frequency: $EPV = c \mu$.
Then the ratio of the credibility to the mean frequency is approximately:
$(3)(VHM / EPV) / \mu = (3/c) VHM / \mu^2$.

Thus the ratio of the credibility to the mean frequency is proportional to the square of the coefficient of variation of the hypothetical means: VHM / μ^2. Thus the smaller this ratio, the smaller the CV of the hypothetical means, and the less variation between the insureds within a class.

Thus the smaller this ratio of credibility to frequency, the more homogeneous the class.

The more homogeneous the class, the less the credibility assigned to the experience of an individual, as experience of an individual that differs from the average would more likely be random than a real difference. To take the extreme case, if all the risks in a class were known to be exactly alike, we would know that any variations in the experience of an individual from average for its class are random, and therefore should be given no credibility.

---

[28] I would prefer using the one-year credibility, since as will be discussed, the one-year credibility is less affected by shifting risk parameters over time than is the three-year credibility.
[29] Subsequently, I have a review of Buhlmann Credibility and some related material.
[30] As will be discussed subsequently, the Buhlmann Credibility formula does not hold for this data.
[31] A car that has been claims-free for at least three years may have many years of data. However if all we know is that it has been claim free for at least three years, then we are looking at the most recent three years of data. N = 3.
Similarly, if we look at all the cars that has been claim free for at least the last two years (combining those that have been claims-free for exactly two years with those who have been claims-free for at leas 3 years), then N = 2.

All other things being equal, more claims means higher credibility. All other things being equal, one car for one year when the mean frequency for the class is 10% has more credibility than when the mean frequency is 5%; approximately, twice as much credibility in the first case than the second, all else being equal. Thus we divide by the mean frequency to adjust for its effect. This leaves the effect of homogeneity, which we are trying to compare between classes.

As shown in Table 2 of Bailey-Simon:

Class	Three-Year Credibility	Claim frequency per car-year	Ratio
1	8.0%	8.7%	0.920
2	6.8%	12.0%	0.567
3	8.0%	14.2%	0.563
4	9.9%	16.2%	0.611
5	5.9%	11.0%	0.536

With the highest ratio of credibility to mean frequency, Class 1 is the least homogeneous, in other words the most heterogeneous.[32] With the lowest ratio of credibility to mean frequency, Class 5 is the most homogeneous, although Classes 2 and 3 are nearly as homogeneous.[33]

"Classes 2, 3, 4 and 5 are more narrowly defined than Class 1, and the fact that the ratios in the last column of Table 2 for these classes are less than the ratio for Class 1 confirms the expectation that there is less variation of individual hazards in those classes. This also illustrates that **credibility for experience rating depends not only on the volume of data in the experience period but also on the amount of variation of individual hazards within the class**."[34]

[32] Class 1 is Pleasure - no male operator under 25.
[33] Class 2 is Pleasure - Non-principal male operator under 25, Class 3 is Business use, Class 4 is Unmarried owner or principal operator under 25, and Class 5 is Married owner or principal operator under 25.
[34] The homogeneity of classes is also discussed in the ASOP 12: Risk Classification.

Table 3, Bailey-Simon:

In their important Table 3, for each class separately, the two-year and three-year credibilities are compared to the one-year credibility.

As shown in Table 2 of Bailey-Simon:

Class	One-Year Credibility	Two-Year Credibility	Three-Year Credibility
1	4.6%	6.8%	8.0%
2	4.5%	6.0%	6.8%
3	5.1%	6.8%	8.0%
4	7.1%	8.5%	9.9%
5	3.8%	5.0%	5.9%

For Class 1, the ratio of the two-year to one-year credibility is: 6.8% / 4.6% = 1.48.

Then as shown in Table 3 of Bailey-Simon:

	Relative Credibility		
Class	One-Year	Two-Year	Three-Year
1	1.00	1.48	1.74
2	1.00	1.33	1.51
3	1.00	1.33	1.57
4	1.00	1.20	1.39
5	1.00	1.32	1.55

These credibilities go up much less than linearly as the number of years of data increase.

Bailey-Simon gives the following possible reasons:[35]

1. Risks entering and leaving the class.

2. An individual insured's chance for an accident changes from time to time within a year and from one year to the next.

3. The risk distribution of individual insureds has a marked skewness reflecting varying degrees of accident proneness.

4. The Buhlmann Credibility formula, $Z = N / (N+K)$, increases somewhat less than linearly with N.

[35] To be discussed in more detail subsequently. The fourth reason is from the discussion by Hazam.

Table 4, Bailey-Simon:

For the class with the most data, Class 1, Bailey-Simon also works with loss ratios rather than frequencies.[36] The denominator is the same premium at Group B rates. The numerator is incurred losses rather than number of claims.

The overall loss ratio is 43.6%.
The loss ratio for Group A (3 or more years claims-free) is 39.7%.
The relative loss ratio is: 39.7% / 43.6% = 0.911.
Thus the three-year credibility is: 1 - 0.911 = 5.5%.

The relative loss ratio for those who are claims free at least 2 years (A + X) is 0.924.
Thus the two-year credibility is: 1 - 0.924 = 7.6%.

The credibilities are:

1 Year	2 Year	3 Year
5.5%	7.6%	8.9%

These are similar to those for Class 1 based on frequency as shown in Table 2, but slightly bigger. This seems to indicate that those who are claims-free also have a lower expected future severity compared to those who are not claims-free.

The relative credibilities are:

1 Year	2 Year	3 Year
1.00	1.38	1.62

This is similar pattern as seen for the credibilities based on frequency. For Class 1, here the credibilities are slightly further from linear than were those based on frequency.

[36] The aggregate losses for an insured are affected by severity as well as frequency, and thus loss ratios are subject to more random fluctuation than are frequencies. Thus an analysis of loss ratios requires more data than a similar analysis of frequencies.

An Alternate Way to Estimate the One-Year Credibility:[37]

Bailey-Simon also backs out a one-year credibility by comparing the observed frequency in the prior year of those who were not claims-free (Merit Rating Group B) to their observed frequency in the subsequent year.

For example, as shown in Table 1, for Class 1 the observed overall frequency per exposure is: 288,019 / 3,325,714 = 0.0866. Assume that the overall frequency is Poisson with mean λ. Then the mean number of claims for those who were not claim free (Group B) is:[38]
$\lambda / (1 - e^{-\lambda}) = 0.0866 / (1 - e^{-0.0866}) = 1.044$.

Thus Group B has a frequency relative to average within Class 1 of: $1 / (1 - e^{-\lambda}) = 1/ (1 - e^{-0.0866}) = 12.05$. However, based on its relative premium based frequency, in Table 1 we have an estimated modification for Group B in Class 1 of: 2.190 / 1.484 = 1.476.
Thus, $1.476 = (12.05) Z + (1)(1 - Z). \Rightarrow Z = (1.476 - 1) / (12.05 - 1) = 4.3\%$.[39] This is similar to the 4.6% one-year credibility for Class 1 shown in Table 2 and based on the claims-free discount.

Let λ = the mean claim frequency (per exposure) for the class.
M = relative premium based frequency for risks with one or more claims in the past year.
Then, $M = Z / (1 - e^{-\lambda}) + (1 - Z)(1). \Rightarrow Z = \frac{M - 1}{1 / (1 - e^{-\lambda}) - 1} = (M - 1) (e^{\lambda} - 1)$.

Here are the similar results for all of the classes:

Class	Mean Freq. Overall	Mean Freq. For Group B	Prior Rel. For Group B	Subseq. Rel. For Group B	One Year Credibility	Table 2 1 year Z
1	8.66%	1.044	12.05	1.476	4.3%	4.6%
2	12.05%	1.061	8.81	1.307	3.9%	4.5%
3	14.24%	1.073	7.53	1.362	5.5%	5.1%
4	16.21%	1.083	6.68	1.247	4.3%	7.1%
5	10.96%	1.056	9.63	1.302	3.5%	3.8%

There is a reasonable match between the credibilities from looking at Group B and those from the claims-free discount, with the exception of Class 4. As will be discussed subsequently, there is an inherent problem with using the claim free discount to estimate credibilities for Class 4, which includes many drivers who have less than three years of driving experience. In any case, these two different techniques are expected to produce similar but somewhat different results, neither of which is equal to the least squares credibility.

[37] See page 160 and Appendix II in Bailey-Simon. See for example, 9, 11/05, Q.3, and 9, 11/09, Q.4a.
[38] See Appendix II in Bailey-Simon, to be discussed subsequently.
[39] Matching the result shown at the bottom of page 160 in Bailey-Simon.

Standard Method	Alternative Method
actual past claim frequency	theoretical past claim frequency
nonparametric	Poisson Distribution
claim frequency to premiums	claim frequency to exposures
claims-free risks	not claims-free risks
1, 2, and 3 year credibilities	one year credibility

Conclusions of Bailey-Simon:[40]

(1) The experience for one car for one year has significant and measurable credibility for experience rating.

(2) In a highly refined private passenger rating classification system which reflects inherent hazard, there would not be much accuracy in an individual risk merit rating plan, but where a wide range of hazard is encompassed within a classification, credibility is much larger.

(3) If we are given one year's experience and add a second year we increase the credibility roughly two-fifths. Given two years' experience, a third year will increase the credibility by one-sixth of its two-year value.

Conclusion number 1 has two parts. Bailey-Simon have demonstrated a practical and simple way to measure this credibility. Also they show in Table 2 that the credibility is big enough to make Merit Rating of Private Passenger Automobile Insurance practical and worthwhile from an actuarial point of view.[41]

Conclusion number 2 follows from general credibility theory applied to experience rating. The more homogeneous a class, the less credibility is given to the experience of an individual insured.

The key idea in conclusion number 3 is that based on their Table 3, the credibilities increase much less than linearly. The specific values are not anywhere near as important.

[40] They were writing more than half a century ago; things that may be obvious today were far from obvious then.
[41] As discussed by Hazam, including moving violations makes Merit Rating more worthwhile to use.

Buhlmann Credibility (Least Squares Credibility), Review:[42]

EPV = Expected Value of the Process Variance = $E_\theta[\ VAR[X \mid \theta]\ ]$.

VHM = Variance of the Hypothetical Means = $VAR_\theta[\ E[X \mid \theta]\ ]$.

Buhlmann Credibility Parameter = $\mathbf{K = \frac{EPV}{VHM}}$,

where the Expected Value of the Process Variance and the Variance of the Hypothetical Means are each calculated for a single observation of the risk process.
One calculates the EPV, VHM, and K prior to knowing the particular observation!

If one is estimating claim frequencies or pure premiums, then N is in exposures.
If one is estimating claim severities, then N is in number of claims.

For N observations, the Buhlmann Credibility Factor is: $\mathbf{Z = \frac{N}{N + K}}$.[43]

Estimate of the future = (Z) (Observation) + (1 - Z) (Prior Mean).

Assumptions:
- (1 - Z) is applied to the prior mean.
- The risk parameters and risk process do not shift over time.
- The expected value of the process variance (EPV) of the sum of N observations increases with N.
- The variance of the hypothetical means (VHM) of the sum of N observations increases with N^2.

For experience rating, we compare the individual relative to its class; the class has a relativity of one, and thus the estimated relativity = Z (observed relativity) + (1 - Z)(1).

Bayes Analysis, Review:[44]

The prior estimate is adjusted to reflect the new information.

Bayes' Theorem: $\mathbf{P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}}$.

$\mathbf{P(Risk\ Type \mid Observation) = \frac{P(Observation \mid Risk\ Type)\, P(Risk\ Type)}{P(Observation)}}$.

---

[42] I do not expect you to be tested directly on any of this other than the use of the formula for Z.
[43] For the situations in Bailey-Simon, K is big compared to N, and thus Z should be approximately proportional to N.
[44] I do not expect you to be tested on any of this.

If $\pi(\theta)$ is the assumed prior distribution of the parameter θ,
then the posterior distribution of θ is proportional to: $\pi(\theta)$ P(Observation | θ).

The posterior distribution of θ is: $\frac{\pi(\theta) \text{ Prob[Observation | } \theta]}{\int \pi(\theta) \text{ Prob[Observation | } \theta] \, d\theta}$

The Bayes estimate is: $\frac{\int (\text{Mean given } \theta) \, \pi(\theta) \text{ Prob[Obs. | } \theta] \, d\theta}{\int \pi(\theta) \text{ Prob[Obs. | } \theta] \, d\theta}$.

Bühlmann Credibility (Least Squares Credibility) is the weighted least squares line fit to the Bayesian estimate. In certain special mathematical situations, such as the Gamma-Poisson or the Beta-Binomial, the Bayesian analysis estimate is equal to that from Bühlmann Credibility (Least Squares Credibility). Due to the greater complexities of its probabilistic nature, Bayesian analysis is not used as commonly in practical applications in insurance as is Bühlmann credibility.[45]

Claims Free Discount Versus Least Squares (Buhlmann) Credibility:

Assume we are using credibility to estimate future frequency.
Then the estimated future frequency for an insured who had no claims is: $Z\ 0 + (1\text{-}Z)\mu = \mu - \mu Z$.
Thus as a percent, the estimated future frequency is Z less than average.
Thus Z is the claim free discount.

Bailey-Simon sets the credibility equal to the indicated claims free discount:

$1 - Z = \frac{\text{observed frequency for those who were claims free}}{\text{overall frequency}}$.[46]

In general, the least squares credibility does not equal this indicated claims free discount. The least squares credibility is the linear estimator that best approximates the Bayes Estimates for all of the possible observations. In contrast, this indicated claims free discount only looks at the observed result for those with no claims. Usually, the claims free discount will be close to the least squares credibility.

One important special case is the Gamma-Poisson.[47] For the
Gamma-Poisson the least squares credibility is equal to the Bayes Estimates; the Bayes Estimates are on a straight line. Thus, in this case, the claim free discount is equal to least squares credibility.[48]

On page 160, Bailey-Simon also backs out a one year credibility by comparing the observed frequency in the prior year of those who were not claim free to their observed frequency in the next year. Again, this will be very similar to but not identical to the least squares credibility.

---

[45] Bayes Analysis is harder to explain to nonactuaries.
[46] They do this separately for those who were claim free for at least a year, at least two years, and at least 3 years.
[47] The Gamma-Poisson is usually a pretty good model for private passenger auto frequencies.
[48] With finite data sets, the two ways to estimate the credibility will differ somewhat.

Review of the Mathematics Behind Experience Rating:

Assume that a insured has had no accidents over the last decade. This provides evidence that he is a safer than average insured; his expected claim frequency is lower than average for his class. Thus for automobile insurance one might give him a "safe driver discount" off of the otherwise applicable rate for his class.

This is an example of experience rating. Generally, experience rating consists of modifying the rate charged to an insured (driver, business, etc.) based on its past experience. While such plans can be somewhat complex in detail, in broad outline they all reward better than expected experience and penalize worse than expected experience. Depending on the particular circumstances more or less weight is put on the insured's observed experience from the recent past.[49]

The new estimate of the insured's frequency or pure premium is a weighted average of that for his classification and the observation. The amount of weight given to the observation is the credibility assigned to the individual insured's data. In general, how much credibility to assign to an individual insured's data should depend on:

1. What is being estimated. Pure Premiums are harder to estimate than frequencies.
Total Limits losses are harder to estimate than basic limits losses.
In a split plan, primary losses are easier to predict than excess losses.[50]

2. The volume of data. All other things being equal, the more data the more credibility is assigned to the observation.[51]

3. The Expected Value of the Process Variance. The more volatile the experience, the less credibility is assigned to it.

4. The variance of the hypothetical means within classes; the more homogeneous the classification the smaller this variance and the less credibility is assigned to the insured's individual experience compared to that for the whole classification.

---

[49] The period of past experience used varies between the different Experience Rating Plans.
[50] For a thorough discussion of whether or not to use severity in addition to frequency, split versus non-split plans, and the choice of accident limits, see "An Analysis of Experience Rating," by Glenn G. Meyers, PCAS 1985, and the discussion by Howard C. Mahler, PCAS 1987.
[51] For example, in Workers' Compensation Insurance the data from a business with $10,000 in Expected Losses would be given much less credibility for Experience Rating than the data from a business with $1 million in Expected Losses.

The more homogeneous the classes, the less variation between the risks within the class, the less credibility assigned an individual's data and the more to the average for the class, when performing experience rating (individual risk rating.) The credibility is a relative measure of the value of the information contained in the observation of the individual versus the information in the class average. The more homogeneous the classes, the more value we place on the class average and the less we place in the individual's experience.

Thus low credibility is neither good nor bad. It merely reflects the relative values of two pieces of information. With a well designed class plan, the less we need to rely on the observations of the individual, compared to a poorly designed class plan. In auto insurance if we classified insureds based on their middle initials, we would expect to give the insureds individual experience a lot of credibility. A poor class plan leads one to rely more on individual experience.

Note that the role of the class in Experience Rating has changed from its role in Classification Ratemaking. In Experience Rating, the class experience receives the complement of credibility not given to the individual's experience. In the case of classification rating, the class experience gets the credibility while the complement of credibility is assigned to the experience of all classes combined. In Experience Rating, the insured is the smaller unit while the class is the larger unit. In Classification Ratemaking, the class is the smaller unit while the state is the larger unit. In both cases, the weight given to the classification's experience is larger the more homogeneous the class. Thus the more homogeneous the classes, the more credibility is given to the experience of each class for Classification Ratemaking. The more homogeneous the class, the less credibility is assigned to the individual's experience and therefore the more weight is given to the class experience for Experience Rating.

Simple models may help one to understand the mathematics behind experience rating.[52] The Gamma-Poisson frequency process is a good model for this purpose. Each insured's frequency is given by a Poisson Process. The mean frequencies of the insureds within a class are distributed via a Gamma Distribution. The variance of this Gamma Distribution quantifies the homogeneity of the class. The smaller the variance of this Gamma, the more homogeneous the class.

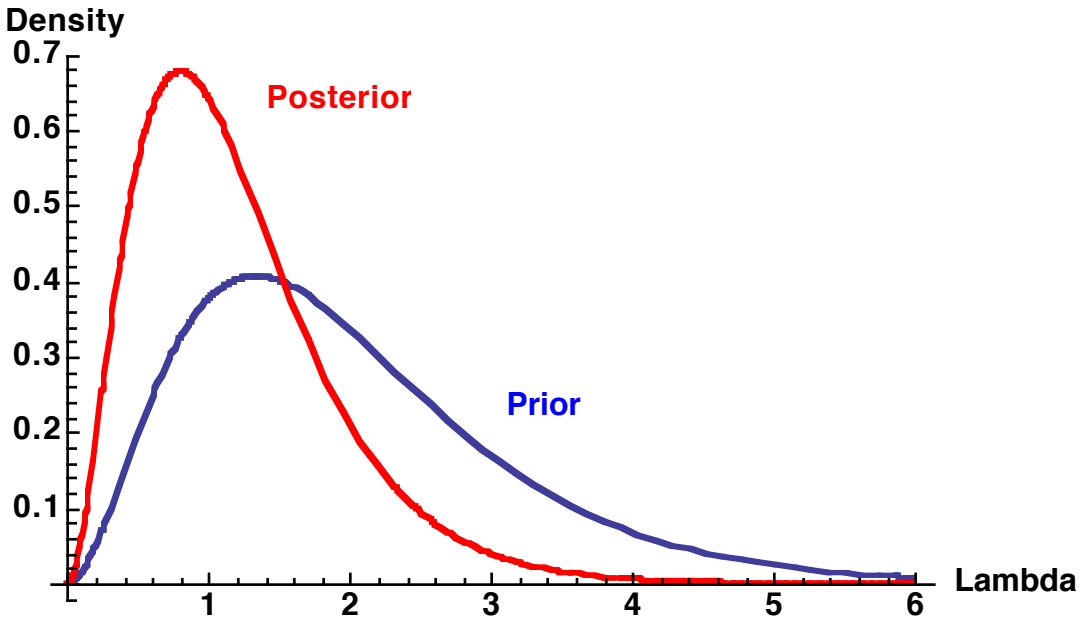
The observed experience of an insured can be used to improve the estimate of that insured's future claim frequency. We assume a priori that the average claim frequencies of the insureds in a class are distributed via a Gamma Distribution with $\alpha = 3$ and $\theta = 2/3$. The average frequency for the class is $(3)(2/3) = 2$.

If we observe no claims in a year, then the posterior distribution of that insured's (unknown) Poisson parameter is a Gamma distribution with $\alpha = 3$ and $\theta = 0.4$, with an average of: $(3)(0.4) = 1.2$.[53]
Thus the observation has lowered our estimate of this insured's future claim frequency.

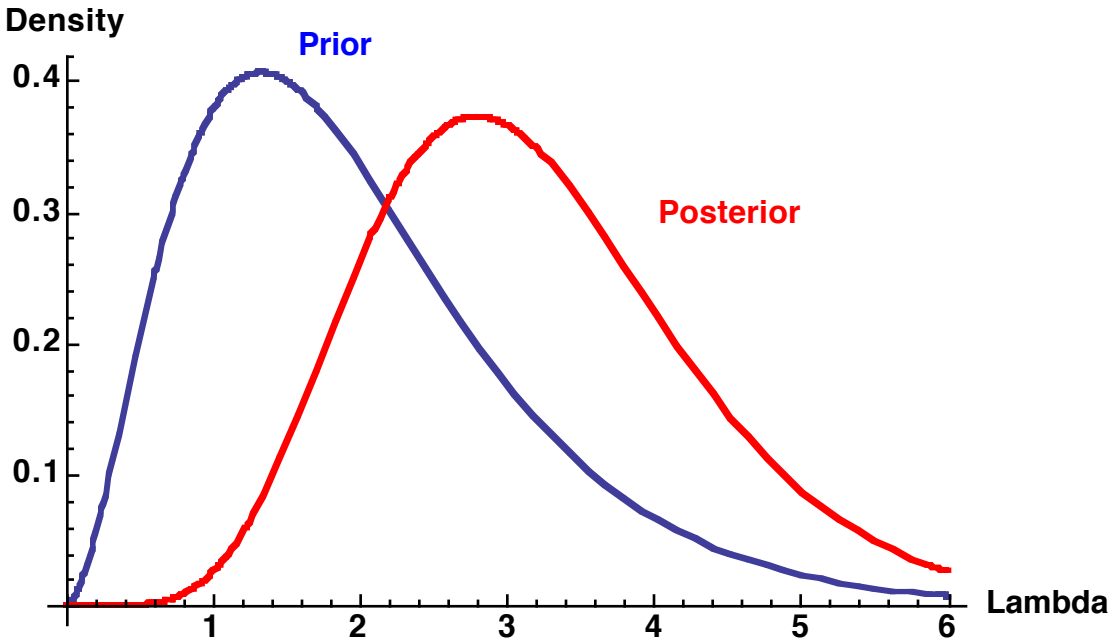
---

[52] See for example, "A Graphical Illustration of Experience Rating Credibilities," by Howard C. Mahler, PCAS 1998.
[53] The posterior alpha is 3 + 0 = 0. The posterior theta = 1/{1+1/(2/3)} = 1/ 2.5 = 0.4.

The prior Gamma with $\alpha = 3$ and $\theta = 2/3$, and the posterior Gamma with $\alpha = 3$ and $\theta = 0.4$, are shown:



If instead we observe 5 claims in a year, then the posterior distribution of that insured's (unknown) Poisson parameter is a Gamma distribution with $\alpha = 8$ and $\theta = 0.4$, with an average of: (8)(0.4) = 3.2.[54] Thus this observation has raised our estimate of this insured's future claim frequency.
The posterior Gamma in the case of this alternate observation is shown below:



[54] The posterior alpha is 3 + 5 = 8. The posterior theta = 1/{1+1/(2/3)} = 1/ 2.5 = 0.4.

Shifting Risk Parameters:

One possible explanation for the credibilities increasing significantly less than linearly provided by Bailey-Simon is: "an individual insured's chance for an accident changes from time to time within a year and from one year to the next."[55] This is the concept of shifting risk parameters as discussed in the syllabus reading by Mahler, "An Example of Credibility and Shifting Risk Parameters."

When parameters shift over time, more distant years are worse predictors than they otherwise would have been. For example, let us assume 1956, 1957 and 1958 are available for predicting 1959. 1956 will be more affected by shifting risk parameters than would be 1958. Due to shifting risk parameters, all of the credibilities will be smaller than they otherwise would be, but the three year credibility (data from 1956 to 1958) is affected more than is the one year credibility (1958 data).

Thus we see a ratio of the three year to the one year credibility that is significantly less than 3. The more rapid the shifting, the larger the effect and thus the smaller this ratio.

A Model with No Shifting Risk Parameters:[56]

Assume there are no territories and we are looking at one class. Insureds do not move in and out of this class. Each insured is Poisson.
There are 100,000 insureds with $\lambda = 10\%$, and 100,000 insureds with $\lambda = 30\%$.

Of those with $\lambda = 10\%$, the expected number claims free for one year is: $100{,}000\ e^{-0.1} = 90{,}484$.
Of those with $\lambda = 30\%$, the expected number claims free for one year is: $100{,}000\ e^{-0.3} = 74{,}082$.
The expected future frequency for those who were claim free for one year is:

$$\frac{(90{,}484)(10\%) + (74{,}082)(30\%)}{90{,}484 + 74{,}082} = 19.00\%.$$

The overall frequency is 20%.
Thus, $1 - Z = 19.00\% / 20\%. \Rightarrow Z = 5.00\%$.

[55] "The fact that the relative credibilities in Table 3 for two and three years are much less than 2.00 and 3.00 is partially caused by risks entering and leaving the class. But it can be fully accounted for only if an individual insured's chance for an accident changes from time to time within a year and from one year to the next, or if the risk distribution of individual insureds has a marked skewness reflecting varying degrees of accident proneness."
[56] Similar to a simple Bayes Analysis question on a preliminary exam.

Exercise: Using the technique in Bailey-Simon, determine the two-year credibility.
[Solution: Of those with $\lambda = 10\%$, the number claims free for 2 years is: $100,000\ e^{-0.2} = 81,873$.
Of those with $\lambda = 30\%$, the number claims free for 2 years is: $100,000\ e^{-0.6} = 54,881$.
The expected future frequency for those who were claims-free for two years is:
$\frac{(81,873)(10\%) + (54,881)(30\%)}{81,873 + 54,881} = 18.026\%$.

Thus, $1 - Z = 18.026\% / 20\%$. $\Rightarrow Z = 9.87\%$.
Comment: Bailey-Simon use data. We have applied their technique to the data we would expect to see if the given model were correct.]

Exercise: Using the technique in Bailey-Simon, determine the three-year credibility.
[Solution: Of those with $\lambda = 10\%$, the number claims free for 3 years is: $100,000\ e^{-0.3} = 74,082$.
Of those with $\lambda = 30\%$, the number claims free for 3 years is: $100,000\ e^{-0.9} = 40,657$.
The expected future frequency for those who were claims-free for three years is:
$\frac{(74,082)(10\%) + (40,657)(30\%)}{74,082 + 40,657} = 17.087\%$.

Thus, $1 - Z = 17.087\% / 20\%$. $\Rightarrow Z = 14.57\%$.]

The ratio of the two-year credibility to the one-year credibility is: 9.87% / 5% = 1.974.
The ratio of the three-year credibility to the one-year credibility is: 14.57% / 5% = 2.914.
Thus these credibilities increase slightly less than linearly, but much closer to linearly than those in Bailey-Simon.[57] This behavior can be explained by the Buhlmann Credibility Formula, $Z = N / (N+K)$.

Exercise: Determine the Buhlmann Credibility Parameter, K, for this model.
[Solution: EPV = (10% + 30%)/2 = 0.2. VHM = $\{(0.1 - 0.2)^2 + (0.3 - 0.2)^2\}/2 = 0.01$.
$K = EPV / VHM = 0.2 / 0.01 = 20$.]

Comparing the Buhlmann (least squares) Credibilities with those from the claims-free discounts:

N	Credibility from Claims-Free	Buhlmann Credibility
1	5.00%	1/(1+20) = 4.76%
2	9.87%	2/(2+20) = 9.09%
3	14.57%	3/(3+20) = 13.04%

As expected the credibilities from the claims-free discounts are similar to those from Buhlmann Credibility, which increase somewhat less than linearly.

[57] In Table 3 of Bailey-Simon for Class 1, the ratios are 1.48 and 1.74.

Bayes Analysis versus Buhlmann Credibility:

For this simple example, let us assume we observe the total number of claims over three years for an individual insured of unknown type.

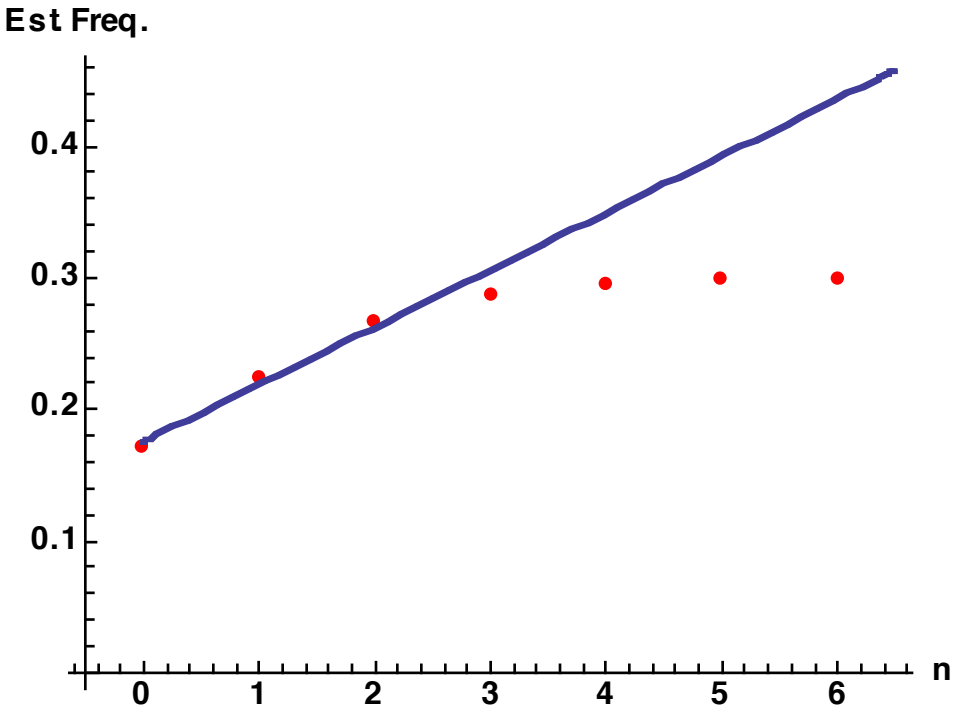
We had previously computed K = 20, Z = 3/23. Thus if we observe n claims in three years, the estimated future annual frequency is: (3/23) n + (20/23)(0.2).

Exercise: Assume we see one claim in three years.
Use Bayes Analysis to estimate the future annual frequency for that insured.
[Solution: Over three years we have a Poison with mean 3λ.
The chances of the observation are: $0.3\, e^{-0.3}$, and $0.9\, e^{-0.9}$.
Since the risk types are equally likely, the posterior probabilities are:

$$\frac{0.3\, e^{-0.3}}{0.3\, e^{-0.3} + 0.9\, e^{-0.9}} = 0.378, \text{ and } \frac{0.9\, e^{-0.9}}{0.3\, e^{-0.3} + 0.9\, e^{-0.9}} = 0.622.$$

Thus the estimated future estimated annual frequency for this insured is:
(0.378)(10%) + (0.622)(30%) = 22.44%.]

Proceeding in a similar manner, we can get the estimate from Bayes Analysis for other possible observations. Here is a graph with the Buhlmann Credibility Estimate as the straight line, and the estimates from Bayes Analysis as the dots, for n = 0, 1, ..., 6:[58]

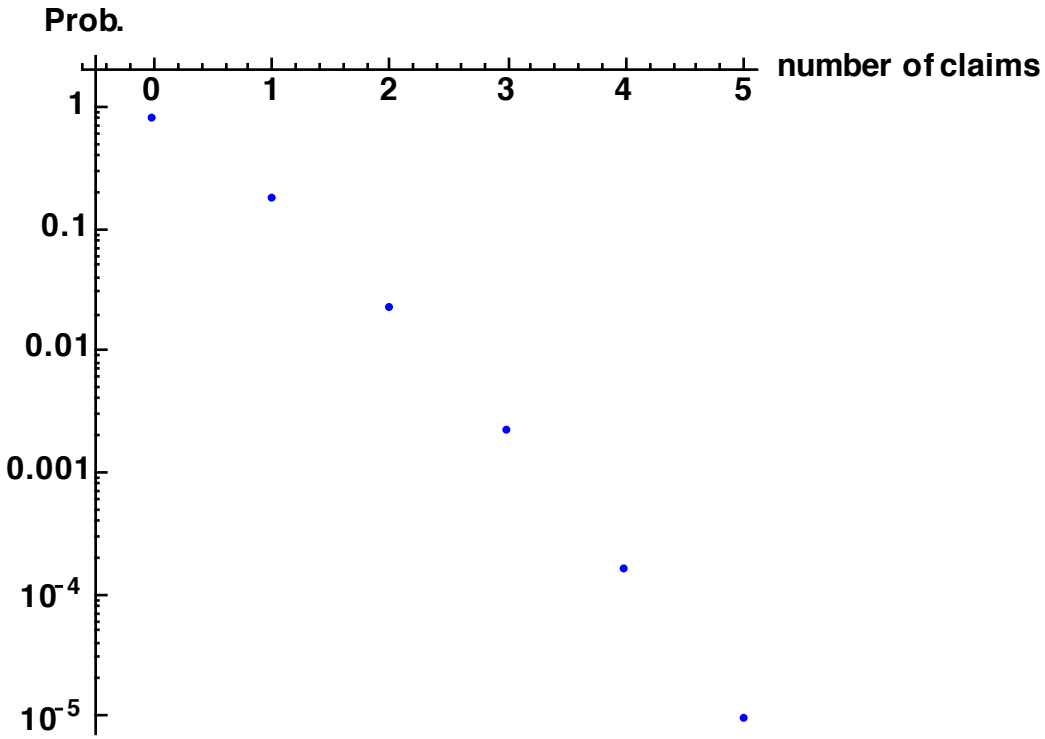


[58] We can observe more than 6 claims.

In general, the line formed by the Buhlmann Credibility estimates is the weighted least squares line to the Bayesian estimates, with the a priori probability of each outcome acting as the weights. The slope of this weighted least squares line to the Bayesian Estimates is the Buhlmann Credibility. Buhlmann Credibility is the Least Squares approximation to the Bayesian Estimates.

Exercise: Assume we see one claim in three years.
Use Bayes Analysis to estimate the probability of seeing two claims next year.
[Solution: From the previous exercise, the posterior probabilities are: 0.378, and 0.622.
Thus the probability that this insured will have 2 claims next year is:
$(0.378)(0.1^2 e^{-0.1} / 2) + (0.622)(0.3^2 e^{-0.3} / 2) = 2.24\%$.
Comment: A question for a preliminary exam, that you do not expect to be asked on this exam.]

For an insured who had one claim in three years, here its distribution of number of claims for the following year, with probability shown on a log scale:



Number of Insureds Claims-Free for Exact Numbers of Years:[59]

For the previous model, there are no territories and we are looking at one class.
Insureds do not move in and out of this class. Each insured is Poisson.
There are 100,000 insureds with $\lambda = 10\%$, and 100,000 insureds with $\lambda = 30\%$.

The expected number of insureds with no years claims-free, in other words who have at least one claim the first year is: $100,000 (1 - e^{-0.1}) + 100,000 (1 - e^{-0.3}) = 35,434$.

Exercise: Determine the expected number of insureds claims-free for exactly one year.
[Solution: The expected number claims-free for at least one year is:
$100,000 e^{-0.1} + 100,000 e^{-0.3} = 164,566$.
The expected number claims-free for at least two years is:
$100,000 e^{-0.2} + 100,000 e^{-0.6} = 136,754$.
Thus the expected number claims-free for exactly one year is: 164,566 - 136,754 = 27,812.]

Exercise: Determine the expected number of insureds claims-free for exactly two years.
[Solution: $100,000 (e^{-0.2} - e^{-0.3}) + 100,000 (e^{-0.6} - e^{-0.9}) = 22,015$.]

Here is a list of the expected number of insureds claims-free for exactly t years:

t = 0	t=1	t=2	t=3	t=4	t=5	t=6
35,435	27,811	22,015	17,587	14.185	11,555	9507

t=7	t=8	t=9	t=10	t=11	t=12	t=13
7899	6627	5611	4791	4124	3574	3118

t=14	t=15	t=16	t=17	t=18	t=19	t=20
2735	2411	2135	1896	1690	1510	1352

more than 20
12,433

[59] See 8, 11/16, Q.1.

A Model with Shifting Risk Parameters:[60]

Alter the previous model so that each year an insured of a given type has a 20% chance of switching to the other type.[61] Thus an insured who has $\lambda = 10\%$ this year, has an expected frequency next year of: (80%)(10%) + (20%)(30%) = 14%. An insured who has $\lambda = 30\%$ this year, has an expected frequency next year of: (80%)(30%) + (20%)(10%) = 26%.

Of those with $\lambda = 10\%$, the number claims-free for one year is: $100,000\ e^{-0.1} = 90,484$.
Of those with $\lambda = 30\%$, the number claims-free for one year is: $100,000\ e^{-0.3} = 74,082$.
Thus the expected future frequency for those who were claims-free for one year is:
$\frac{(90,484)(14\%) + (74,082)(26\%)}{90,484 + 74,082} = 19.402\%$.

The overall frequency is 20%.
Thus, $1 - Z = 19.402\% / 20\%. \Rightarrow Z = 2.99\%$.[62]

Of those with $\lambda = 10\%$ in the first year who were claims-free, the next year (0.8)(90,484) = 72,387 of them have $\lambda = 10\%$, while (0.2)(90,484) = 18,097 of them have $\lambda = 30\%$.
Of those with $\lambda = 30\%$ in the first year who were claims-free, the next year (0.2)(74,082) = 14,816 of them have $\lambda = 10\%$, while (0.8)(74,082) = 59,266 of them have $\lambda = 30\%$.
In summary, of those who were claims-free the first year, during the second year 72,387 + 14,816 = 87,203 will have $\lambda = 10\%$, while 18,097 + 59,266 = 77,363 will have $\lambda = 30\%$.

Of those who were claims-free in year one and with $\lambda = 10\%$ in year two, the number claims-free in year two is: $87,203\ e^{-0.1} = 78,905$.
Of those who were claims-free in year one and with $\lambda = 30\%$ in year two, the number claims-free in year two is: $77,363\ e^{-0.3} = 57,312$.
Thus the expected future frequency for those who were claims-free for two years is:
$\frac{(78,905)(14\%) + (57,312)(26\%)}{78,905 + 57,312} = 19.049\%$.

Thus, $1 - Z = 19.049\% / 20\%. \Rightarrow Z = 4.76\%$.

[60] You are very unlikely to be asked a numerical question requiring you to work with such a model on your exam.
[61] This is relatively fast rate of shifting risk parameters over time. This is a extremely simplified and unrealistic version of the models in "A Markov Chain Model of Shifting Risk Parameters", by Howard Mahler, PCAS 1997.
[62] Due to shifting risk parameters over time, the one-year credibility has declined from 5.00% to 2.99%.

Exercise: Determine the credibility for three years claims-free.
[Solution: Of those with $\lambda = 10\%$ in the 2nd year who were claims-free, the next year $(0.8)(78,905) = 63,124$ of them have $\lambda = 10\%$, and $(0.2)(78,905) = 15,781$ have $\lambda = 30\%$.
Of those with $\lambda = 30\%$ in the 2nd year who were claims-free, the next year $(0.2)(57,312) = 11,462$ of them have $\lambda = 10\%$, while $(0.8)(57,312) = 45,850$ of them have $\lambda = 30\%$.
In summary, of those who were claims-free for 2 years, during the third year $63,124 + 11,462 = 74,586$ will have $\lambda = 10\%$, while $15,781 + 45,850 = 61,631$ will have $\lambda = 30\%$.
Of those who were claims-free in years one and two with $\lambda = 10\%$ in year three, the number claims-free in year three is: $74,586\ e^{-0.1} = 67,488$.
Of those who were claims-free in year one and with $\lambda = 30\%$ in year two, the number claims-free in year two is: $61,631\ e^{-0.3} = 45,657$.
Thus the expected future frequency for those who were claims-free for three years is:
$\frac{(67,488)(14\%) + (45,657)(26\%)}{67,488 + 45,657} = 18.842\%$.
Thus, $1 - Z = 18.842\% / 20\%$. $\Rightarrow Z = 5.79\%$.]

Comparing credibilities from the claims-free discounts with and without shifting risk parameters:

N	No Shifting	Ratio to One-Year	With Shifting	Ratio to One-Year
1	5.00%		2.99%	
2	9.87%	1.97	4.76%	1.59
3	14.57%	2.91	5.79%	1.94

With shifting risk parameters, the credibilities increase much less than linearly. This is similar to the pattern in Bailey-Simon.[63]

[63] In Table 3 of Bailey-Simon for Class 1, the ratios are 1.48 and 1.74.

Number of Insureds Claims-Free for Exact Numbers of Years, Shifting Risk Parameters.[64]

The expected number of insureds with no years claims-free, in other words who have at least one claim the first year is: $100,000 (1 - e^{-0.1}) + 100,000 (1 - e^{-0.3}) = 35,434$,
the same as without shifting risk parameters.

Of the original 200,000 insureds, over two years there are 4 groups:
80,000 with $\lambda = 10\%$ in both years, 20,000 with $\lambda = 10\%$ the first year and 30% the second year,
80,000 with $\lambda = 30\%$ in both years, 20,000 with $\lambda = 30\%$ the first year and 10% the second year.

Thus the expected number claims-free for exactly one year is:
$80,000 (e^{-0.1} - e^{-0.2}) + 20,000 (e^{-0.1} - e^{-0.4}) + 80,000 (e^{-0.3} - e^{-0.6}) + 20,000 (e^{-0.3} - e^{-0.4})$
= 28,349. This compares to 27,811 without shifting risk parameters.

This type of calculation quickly gets very tedious, so instead I simulated this situation. Here is comparison between no shifting and shifting risk parameters of the numbers claims-free:

	No Shifting[65]	With Shifting[66]		No Shifting	With Shifting
t=0	35,435	35,623	t=11	4124	4354
t=1	27,811	28,543	t=12	3574	3711
t=2	22,015	22,698	t=13	3118	3009
t=3	17,587	19,179	t=14	2735	2578
t=4	14.185	15,662	t=15	2411	2133
t=5	11,555	12,991	t=16	2135	1772
t=6	9507	10,770	t=17	1896	1473
t=7	7899	9208	t=18	1690	1220
t=8	6627	7429	t=19	1510	1015
t=9	5611	6290	t≥20	13,785	5162
t=10	4791	5180			

The two patterns are similar. However, in the the case of shifting risk parameters fewer insureds are claims-free for very long periods of time than when risk parameters are not shifting.[67]
It is not at all clear to me how one could use the information solely of the observed numbers of insureds claims-free for exactly t years in order to determine whether or not risk parameters are shifting and if so how quickly.

---

[64] See 8, 11/16, Q.1.
[65] Expected numbers calculated previously.
[66] From simulation.
Note that the first two simulated numbers differ somewhat from the expected numbers calculated previously.
[67] In this case, the two types of insureds have very different mean claim frequencies, and the rate at which parameters shift is large, in order to make the effects easier to spot.

Risks Entering and Leaving a Class:

One possible explanation for the credibilities increasing significantly less than linearly provided by Bailey-Simon is: "risks entering and leaving the class."[68]

Let us assume that cars are frequently moving from one class to another.[69] For example, let us assume it is common for a car that is pleasure use one year to be business use the next year, or vice-versa. So in the Bailey-Simon data it is common to move from Class 1 to Class 3 or vice-versa.

For example, let us assume one car was pleasure use in 1956 to 1959, while another car was business use in 1956 and 1957, but pleasure use in 1958 and 1959. Then the three years of data 1956-1958 will be worse at predicting 1959 in the latter case than the former case. When we combine a whole bunch of data consisting of both situations, the credibility of three years of data for predicting the future will be lower than if all the cars had remained in the same class.

In this example, the data for 1958 is an equally good predictor of 1959 for both cars. However, there is another car, for which the class would be different in 1958 and 1959. So again, when we combine a whole bunch of data consisting of different situations, the credibility of one year of data for predicting the future will be lower than if all the cars had remained in the same class.

However, for a given car, its class in 1956 is more likely to be different than that in 1959, than is 1958 to be different than 1959. Thus the average effect on the credibility of more distant years is greater than that on more recent years. Thus the credibility of three years of data is more affected by shifting of classes than is the credibility of one year of data. Thus the credibilities go up less than linearly. The more frequently on average the classes of cars shift, the more the effect on the credibilities, and the lower is the ratio of the three year credibility to the one year credibility.

The effect of shifting classes is mathematically the same as shifting risk parameters. However, often in theoretical work on credibility, the term "shifting risk parameters" is restricted to those cases where there has been no change in the classifications and territories used for rating.

---

[68] "The fact that the relative credibilities in Table 3 for two and three years are much less than 2.00 and 3.00 is partially caused by risks entering and leaving the class. But it can be fully accounted for only if an individual insured's chance for an accident changes from time to time within a year and from one year to the next, or if the risk distribution of individual insureds has a marked skewness reflecting varying degrees of accident proneness."

[69] Frequently could be once every five years on average.
This can equally well be moving from one territory to another.

Marked Skewness Reflecting Varying Degrees of Accident Proneness:

One possible explanation for the credibilities increasing significantly less than linearly provided by Bailey-Simon is: "if the risk distribution of individual insureds has a marked skewness reflecting varying degrees of accident proneness."[70] While they provide no further explanation, I believe they are referring to the Gamma-Poisson, which was starting to be discussed about that time with respect modeling Merit Rating.[71] [72] Unfortunately, I do not believe that this is a possible cause of the observed behavior of the credibilities.[73]

The overall mean frequency (for a class) is observed, and thus constrained in any model of the Canadian data in Bailey-Simon. Similarly, we can determine the credibility applied to one year of data; in fact Bailey-Simon shows two complementary ways to do so. From this credibility one can back out the Buhlmann Credibility Parameter K. There is then a unique Gamma-Poisson model (for each class.)

In the absence of shifting risk parameters or insureds entering and leaving classes, we would have (for each class) a Gamma-Poisson model. The claim free discounts come from Bayes Analysis, which for the Gamma-Poisson is the same as the least squares (Buhlmann) credibility. $Z = N / (N + K)$. From the magnitude of the credibilities for one year, K must be relatively big. Therefore, the credibilities are approximately linear in N.

I do not see how having a Gamma-Poisson or some other model changes this, since K is backed out of the data, and does not depend on which particular model is used.

---

[70] "The fact that the relative credibilities in Table 3 for two and three years are much less than 2.00 and 3.00 is partially caused by risks entering and leaving the class. But it can be fully accounted for only if an individual insured's chance for an accident changes from time to time within a year and from one year to the next, or if the risk distribution of individual insureds has a marked skewness reflecting varying degrees of accident proneness."

[71] See for example, "Automobile Merit Rating and Inverse Probabilities," by Lester B. Dropkin, PCAS 1960. Bailey and Simon were each very involved in the literature on this and related subjects at this time.

[72] Each insured has a Poisson frequency with mean l. However, across the class l varies via a Gamma Distribution.

[73] To be fair Bailey and Simon were each pioneers in the development of credibility theory, and did not have the benefit we have of the many developments since they wrote their classical paper. By the way, Robert's father Arthur Bailey developed and published the mathematics of what would later be called "Buhlmann Credibility," about 15 years before Buhlmann published.

Appendix I:

In their Appendix I, Bailey-Simon demonstrate why one would expect the credibility to increase approximately linearly with the number of years of data, given certain assumptions.[74] They set up a discrete risk type model, the type of model which should be familiar from earlier exams.

Each insured has a Poisson frequency. For each insured their mean is the same every year.[75]

Percent of Insureds	Poisson Parameter (mean annual frequency λ)
40%	5%
40%	10%
20%	20%

Then the a priori mean frequency is: (40%)(5%) + (40%)(10%) + (20%)(20%) = 10%.

Assume an insured picked at random is claim free for one year, let us use Bayes Analysis to estimate that insured's future annual frequency.[76]

Percent of Insureds	λ	Chance of Observation	Posterior Chance of Risk Type
40%	5%	$e^{-0.05}$	$0.4e^{-0.05} / (0.4e^{-0.05} + 0.4\ e^{-0.1} + 0.2\ e^{-0.2}) = 41.989\%$
40%	10%	$e^{-0.1}$	$0.4e^{-0.10} / (0.4e^{-0.05} + 0.4\ e^{-0.1} + 0.2\ e^{-0.2}) = 39.941\%$
20%	20%	$e^{-0.2}$	$0.2e^{-0.20} / (0.4e^{-0.05} + 0.4\ e^{-0.1} + 0.2\ e^{-0.2}) = 18.070\%$

Thus the estimated future frequency for this insured is:
(41.989%)(5%) + (39.941%)(10%) + (18.070%)(20%) = 9.707%.[77]

This is lower than the 10% overall a priori frequency. Since the λ for each insured remains the same, and the proportion of risks of each type remains the same, the expected overall future annual frequency is also 10%.

Thus the modification for one year claims free is: 9.707/10 = 0.9707.
The credibility for one year claim free is: Z = 1 - 0.9707 = 2.93%.[78]

---

[74] A key conclusion of their paper is that the credibilities increase much less than linearly.
[75] No shifting risk parameters.
[76] I do not expect you to be asked to do Bayes Analysis on this exam.
[77] Matches the 0.09707 claim frequency after one year claim free shown in Bailey-Simon. What they have done is mathematically the same as Bayes Analysis, just assuming for convenience a total of 250,000 insureds.
[78] Matches the result shown in Bailey-Simon.

Exercise: An insured is picked at random and has two years claims free.
Use Bayes Analysis to estimate this insured's future annual claim frequency.
[Solution: The chance of the observation is Exp[-2λ].

	A Priori	Poisson	Chance of	Probability	Posterior	
Type	Probability	Parameter	Observation	Weights	Probability	Mean
A	40%	5%	90.484%	0.36193	43.951%	5%
B	40%	10%	81.873%	0.32749	39.769%	10%
C	20%	20%	67.032%	0.13406	16.280%	20%
Sum	100%	10%		0.82349	100.000%	**9.430%**

Comment: Matches the result shown in Bailey-Simon for t = 2.]

Then for two years claim free: 1 - Z = 9.430%/10%. ⇒ Z = 5.70%.

Exercise: Using the technique in Bailey-Simon, determine the credibility for 3 years claims free.
[Solution: The chance of the observation is Exp[-3λ].

	A Priori	Poisson	Chance of	Probability	Posterior	
Type	Probability	Parameter	Observation	Weights	Probability	Mean
A	40%	5%	86.071%	0.34428	45.882%	5%
B	40%	10%	74.082%	0.29633	39.491%	10%
C	20%	20%	54.881%	0.10976	14.628%	20%
Sum	100%	10%		0.75037	100.000%	**9.169%**

Then for three years claim free: 1 - Z = 9.169%/10%. ⇒ Z = 8.31%.
Comment: Matches the result shown in Bailey-Simon for t = 3.]

The three credibilities are: 2.93%, 5.70%, and 8.31%.
The ratio of the two year to the one year credibility is: 5.70%/2.93% = 1.945.
The ratio of the three year to the one year credibility is: 8.31%/2.93% = 2.836.

While these credibilities increase somewhat less than linearly, it is much closer to linear than the results Bailey-Simon get for the Canadian data, as shown in their Table 3. In Table 3, for example, for Class 1 the ratio of the two year to the one year credibility is 1.48, while the ratio of the three year to the one year credibility is only 1.74.

One could instead apply Buhlmann Credibility to their simple model in Appendix I.[79]

The process variance for each type is I, so the Expected Value of the Process Variance is:[80]
(40%)(5%) + (40%)(10%) + (20%)(20%) = 10%.

[79] I do not expect you to be asked to do a Buhlmann Credibility problem on this exam.
[80] When mixing Poissons, the EPV is equal to the overall mean.

The first moment of the hypothetical means is the a priori overall mean:
(40%)(5%) + (40%)(10%) + (20%)(20%) = 0.1.
The second moment of the hypothetical means is:
(40%)(5%^{2}) + (40%)(10%^{2}) + (20%)(20%^{2}) = 0.013.
Therefore, the Variance of the Hypothetical Means is: 0.013 - 0.1^2 = 0.003.

The Buhlmann Credibility Parameter is: K = EPV / VHM = 0.1 / 0.003 = 33.33.

Thus the Buhlmann (least squares) Credibility for one year of data is: $\frac{1}{1 + 33.33}$ = 2.91%.

The Buhlmann Credibility for two years of data is: $\frac{2}{2 + 33.33}$ = 5.66%.

The Buhlmann Credibility for three years of data is: $\frac{3}{3 + 33.33}$ = 8.26%.

Again these credibilities increase somewhat less than linearly.

As pointed out in the discussion by Hazam, in general $Z = \frac{N}{N + K}$ increases less than linearly; however, for K large compared to N this formula is not that far from linear.

We note that while the Buhlmann Credibilities are close to the claim free credits, they are not the same. For example, 8.26% ≠ 8.31%. Except in special mathematical cases where they are equal, the two types of credibilities will be close but not the same.

Appendix II:

Assume that the overall frequency is Poisson with mean λ.[81] The portion of insureds with no claims in a year is $e^{-\lambda}$. Then the portion of insureds with at least one claim in a year is: $1 - e^{-\lambda}$. Let x be the mean number of claims had by such insureds. Then since the overall mean is λ, we must have:
$\lambda = (0)(e^{-\lambda}) + x(1- e^{-\lambda}). \Rightarrow x = \lambda/(1- e^{-\lambda})$.[82]

For example, as shown in Table 1, for Class 1 the observed overall frequency (per exposure) is: 288,019 / 3,325,714 = 0.0866. Thus we assume that those insureds who were not claim free during the most recent year, had an average number of claims of approximately:
$0.0866 / (1- e^{-0.0866}) = 1.044$.[83]

Note that in Appendix I, the model is instead a mixture of Poissons. For the example shown there, the overall frequency is 10%. Also the percentage claims free is:
226,544 / 250,000 = 0.9062.

Let x be the mean number of claims had by insureds who had at least one claim.
Then since the overall mean is 10%, we must have:
$0.10 = (0)(0.9062) + (x)(1 - 0.9062). \Rightarrow x = 1.066$.

For the overall mean of 10%, and the technique Bailey-Simon uses, one would instead estimate the mean number of claims for those who have at least one claim to be instead:
$0.1 / (1 - e^{-0.1}) = 1.051$.

Thus the simple model in Appendix II would produce slightly different results than the more complicated model in Appendix I. For the limited purpose for which it is used by Bailey-Simon, the simpler method is okay.

---

[81] A better model would be a mixture of Poissons as per Appendix I.

[82] For small λ, this is approximately: $\lambda / (\lambda - \lambda^2/2) = 2/(2-\lambda)$.

[83] Matching the result in Bailey-Simon.

The Discussion by William J. Hazam:[84]

The areas discussed are: use of premium based rather than exposure based frequencies, the Buhlmann Credibility formula, and the use of convictions for moving traffic violations.

As discussed, Bailey-Simon divide claims by premiums at the Group B rate, in order to get frequencies to compare.[85] This avoids double-counting. Hazam points out:
"that a premium base eliminates maldistribution only if (1) high frequency territories are also high premium territories and (2) if territorial differentials are proper."[86]

While most is due to frequency, some of the variation in premiums by territory is due to differences in severity.[87] Nevertheless, using premiums in the denominator is an improvement.[88]

When Bailey-Simon was written, all expenses were treated as variable. Currently, some expenses are treated as fixed. This would raise another issue with the use of premiums in the denominator .

The Buhlmann Credibility formula says Z = N / (N+K).[89] For large K, the credibility increases only slightly less than linearly. While this does not explain the behavior observed by Bailey-Simon, it is one reason why the credibilities would go up less than linearly.

Given the one year credibilities in Table 2 of Bailey-Simon, we can back out a Buhlmann Credibility Parameter. For example for Class 1, 1/(1+K) = 4.6%. Thus K = 20.7. We can then use this K to calculate 2-year and 3-year credibilities.

Class	One-Year Cred.	K	Two-Year Cred.	Three Year Cred.
1	4.6%	20.7	8.8%	12.7%
2	4.5%	21.2	8.6%	12.4%
3	5.1%	18.6	9.7%	13.9%
4	7.1%	13.1	13.2%	18.6%
5	3.8%	25.3	7.3%	10.6%

[84] This 3 page discussion of Bailey-Simon is also on the syllabus.
[85] "The authors have chosen to calculate Relative Claim Frequency on the basis of premium rather than car years. This avoids the maldistribution created by having higher claim frequency territories produce more X, Y, and B risks and also produce higher territorial premiums."
[86] In other words, if all expenses are treated as variable, then the expected loss ratios by territory should be equal.
[87] After adjusting for difference in the average class rating factor, most of the difference in average pure premiums between territories for Private Passenger Automobile is due to difference in average frequency. Some is due to difference in average severity. Based on my work on Massachusetts Private Passenger Automobile, I estimate that somewhere around 1/5 of the difference is due to severity while the remaining 4/5 is due to frequency.
[88] "However, premium, although not perfect, is an improvement over exposure as a base for this type of study. The fact that either or both of these inherent assumptions may not always exist does not detract from the qualitative nature of the conclusions but may alter somewhat the basic relative frequencies of Table 1 and the consequent values in Tables 2 and 3."
[89] Hazam's review was written before Buhlmann published his papers. This formula goes back to the 1918 PCAS.

We can see that these two-year and three-year credibilities are a poor match to those in Table 2 of Bailey-Simon.

	Two-Year Credibilities		Three-Year Credibilities	
Class	Buhlmann Formula	Table 2	Buhlmann Formula	Table 2
1	8.8%	6.8%	12.7%	8.0%
2	8.6%	6.0%	12.4%	6.8%
3	9.7%	6.8%	13.9%	8.0%
4	13.2%	8.5%	18.6%	9.9%
5	7.3%	5.0%	10.6%	5.9%

Due to shifting risk parameters and other possible causes mentioned by Bailey-Simon, the Buhlmann Credibility formula is not a good model for the credibilities for different numbers of years shown in Table 2 of Bailey-Simon.

Finally, Hazam mentions that many Merit Rating plans in the U.S. use moving traffic violations in addition to claims.[90] The addition of this useful information allows one to better distinguish between insureds within the same class, and therefore justifies larger credits and larger surcharges than when using just claims history.[91]

The amount of credibility depends as well on how refined the class plan is. The more homogeneous the classes, the less need there is for Merit Rating, and the smaller the credibility assigned to the data of an individual insured.

In any case, an actuary should use appropriate caution about extending the results on one set of data to other somewhat different situations.

---

[90] "It may be surmised from this approach to the Canadian results that, in a balanced merit rating plan, there is not enough credibility by class to warrant the magnitude of credits now being offered by many U. S. plans. We must remember, however, that these results are based strictly on claim frequencies, not claim frequencies plus convictions frequencies. Adding convictions no doubt helps substantiate larger credits but it is dubious that it will support current merit rating differentials, if the Canadian experience is at all indicative of what we might expect in this country."

[91] I looked extensively at such data for Massachusetts Private Passenger Automobile Insurance when I was involved in the redesign of the mandatory SDIP in the early 1980s. It was clear that for example someone who had recently been convicted of speeding had a higher expected future claim frequency than an otherwise similar driver who had not.

The Impact of Different Territories, and Why We Use Premiums in the Denominator:

Let us take an extremely simple model. There are two territories with equal exposures, and no classes. Each insured is Poisson, and l does not vary over time. In Territory 1, half of the insureds have $\lambda = 2\%$ and the other half have $\lambda = 8\%$.[92] In Territory 2, half of the insureds have $\lambda = 6\%$ and the other half have $\lambda = 14\%$. The average severity for all insureds is $10,000.

The overall frequency is 7.5%. The overall pure premium is $750.
Territory 1 has a mean frequency of 5%, while Territory 2 has a mean frequency of 10%.
Thus Territory 1 has a pure premium of $500, while Territory 1 has a pure premium of $1000.
Assuming no fixed expenses, we charge Territory 2 twice as much on average as Territory 1.

Let us assume we give a percentage credit to those who are claim free for at least three years.
Let us see what happens if we calculate the three-year credibility using exposures (rather than base class premiums) in the denominator. For convenience, assume 400,000 insureds in total.[93]

Type	Number who are 3 years claims-free	Number who are not 3 years claims-free
$\lambda = 2\%$	$(100,000)(e^{-0.06}) = 94,176$	5,824
$\lambda = 8\%$	$(100,000)(e^{-0.24}) = 78,663$	21,337
$\lambda = 6\%$	$(100,000)(e^{-0.18}) = 83,527$	16,473
$\lambda = 14\%$	$(100,000)(e^{-0.42}) = 65,705$	34,295

In total, there are 322,072 claims-free and 77,929 who are not.
The average future annual (exposure based) frequency for those who were claims free is:

$$\frac{(2\%)(94,176) + (8\%)(78,663) + (6\%)(83,527) + (14\%)(65,705)}{322,072} = 6.742\%.$$

The average future annual (exposure based) frequency overall is 7.5%.
Thus, $1 - Z = 6.742/7.5. \Rightarrow Z = 10.1\%. \Rightarrow$ A 10.1% discount from average.

We wish to charge Territory 1 $500 on average.
Thus we wish to charge those who are claims-free: (0.899)($500) = $449.5.
Let the base rate be x.
There are claims-free 94,176 + 78,663 = 172,839, and not claims free: 5824 + 21,337 = 27,161.
$27,161\ x + (172,839)($449.5) = (200,000)($500). \Rightarrow x = \$821.36.$

[92] There is no way to distinguish the two types.
I have chosen the means to be very different for illustrative purposes.
[93] For an insured with lambda = 8%, the three years frequency is Poisson with lambda = 24%.
For simplicity assume each insured has been licensed for at least three years and no insured switches territories.

We wish to charge Territory 2 $1000 on average.
Thus we wish to charge those who are claims-free: (0.899)($1000) = $899.
Let the base rate be y.
There are claims-free 83,527 + 65,705 = 149,232, and not claims free:
16,473 + 34,295 = 50,768.
50,768 y + (149,232)($899) = (200,000)($1000). $\Rightarrow$ y = $1296.89.

For those claims-free in Territory 1, the expected pure premium is:
($10,000) {(2%)(94,176) + (8%)(78,663)} / 172,839 = $473.07.
For those not claims-free in Territory 1, the expected pure premium is:
($10,000) {(2%)(5824) + (8%)(21,337)} / 27,161 = $671.34.

For those claims-free in Territory 2, the expected pure premium is:
($10,000) {(6%)(83,527) + (14%)(65,705)} / 149,232 = $952.23.
For those not claims-free in Territory 2, the expected pure premium is:
($10,000) {(6%)(16,473) + (14%)(34,295)} / 50,768 = $1140.42.

Let us compare the amount charged to the expected pure premiums:

Territory	Claims-free	Expected Pure Premium	Premium Charged
1	Yes	$473.07	$449.50
1	No	$671.34	$821.36
1	All	$500	$500
2	Yes	$952.23	$899.00
2	No	$1140.42	$1296.89
2	All	$1000	$1000

Using the exposure based frequencies to determine the claims-free credibility and discount, the pure premiums by cell do not match well to the premiums charged. Let us see what happens if instead we use premium based frequencies, as per Bailey-Simon.

There are 322,072 claims-free. The expected number of claims next year for these insureds is:
(2%)(94,176) + (8%)(78,663) + (6%)(83,527) + (14%)(65,705) = 22,387.

Assume that the current base rate for Territory 2 is twice that of Territory 1, 1000 versus 500.[94]
Then the annual premium at base rates next year for these insureds is:
(500)(94,176) + (500)(78,663) + (1000)(83,527) + (1000)(65,705) = 235,651,500.

The premium based frequency (per $1000) for those who were claims-free is:
22,387 / 235,651.5 = 0.09500.

The average premium based frequency overall is 7.5%/0.75 = 0.10000.
Thus, 1 - Z = 0.09500/0.10000. $\Rightarrow$ Z = 5%.

[94] All that is important is that the ratio is two to one, so that the current territory relativity is correct.

Taking into account the mix of the claims-free insureds by territory has resulted in this case in a much smaller credibility of 5% rather than 10.1%.

We wish to charge Territory 1 $500 on average.
Thus we wish to charge those who are claims-free: (0.95)($500) = $475.
Let the base rate be x.
There are claims-free 94,176 + 78,663 = 172,839, and not claims free: 5824 + 21,337 = 27,161.
27,161 x + (172,839)($475) = (200,000)($500). $\Rightarrow$ x = $659.09.

We wish to charge Territory 2 $1000 on average.
Thus we wish to charge those who are claims-free: (0.95)($1000) = $950.
Let the base rate be y.
There are claims-free 83,527 + 65,705 = 149,232, and not claims free:
16,473 + 34,295 = 50,768.
50,768 y + (149,232)($950) = (200,000)($1000). $\Rightarrow$ y = $1146.97.

Now the comparison of the amount charged to the expected pure premiums is:

Territory	Claims-free	Expected Pure Premium	Premium Charged
1	Yes	$473.07	$475
1	No	$671.34	$659.09
1	All	$500	$500
2	Yes	$952.23	$950
2	No	$1140.42	$1146.97
2	All	$1000	$1000

As expected from Bailey-Simon, the premium based frequencies do a much better job of estimating appropriate claim-free discounts than do the exposure based frequencies.
The remaining discrepancy comes from having a single discount for both territories.

If we look at a single territory, then it will not matter whether we use premiums or exposures in the denominator. In Territory 1, the (exposure based) frequency for those who are claim free is:
$\frac{(2\%)(94,176) + (8\%)(78,663)}{94,176 + 78,663}$ = 4.7307%. The overall frequency in Territory 1 is 5%.

Thus, 1 - Z = 4.7307%/5%. $\Rightarrow$ Z = 5.385%.

In Territory 2, the (exposure based) frequency for those who are claim free is:
$\frac{(6\%)(83,527) + (14\%)(65,705)}{83,527 + 65,705}$ = 9.522%. The overall frequency in Territory 2 is 10%.

Thus, 1 - Z = 9.522%/10%. $\Rightarrow$ Z = 4.777%.

An Example Using Driver Data [95]

Here is example of the results of study similar to that in Bailey-Simon, which was done at the same time. However, the data was on drivers rather than cars.[96]

Some cars are driven by more than one driver, while some drivers drive more than one car. Also a much larger percent of licensed drivers do not drive during a year or drive only a minimal number of miles, compared to the percent of insured cars that are not driven or are only driven a minimal number of miles. So it makes some difference in the results whether one analyzes cars or drivers.

Drivers were grouped by the number of traffic violations they had over a three year period.[97] Then the number of accidents over a three year period by the drivers in the different groups was compared.[98] As expected, those drivers with more violations had a higher mean frequency.

Number of Violations	Number of Drivers	Mean Number of Accidents	Variance of Number of Accidents
0	55,757	0.087	0.096
1	20.613	0.194	0.207
2	8,753	0.274	0.299
3	4,320	0.354	0.395
4	2,297	0.426	0.501
5 or more	3,195	0.553	0.610
Total	94,935	0.163	0.193

Also the variance of the number of accidents within each group was computed.
If we assume that for each driver the number of accidents is Poisson distributed with mean λ, and the lambdas vary via Gamma Distribution, then the mixed distribution is Negative Binomial. As discussed on a preliminary exam, if the Gamma has parameters α and θ, then the Negative Binomial has parameters $r = \alpha$ and $\beta = \theta$.

It was found that in total, the number of accident data was fit well by a Negative Binomial. One can fit via the method of moments a Negative Binomial, to the total and to each group above. Set $r\beta$ = mean, and $r\beta(1+\beta)$ = variance. The results are shown below.

---

[95] Taken from "Some Considerations on Automobile Rating Systems Utilizing Individual Driving Records," by Lester B. Dropkin, PCAS 1959, not on the syllabus. See also the discussion by Robert A. Bailey in PCAS 1960. See also "Merit Rating in Private Passenger Automobile Liability Insurance and the California Driver Record Study," by Frank Harwayne, PCAS 1959.

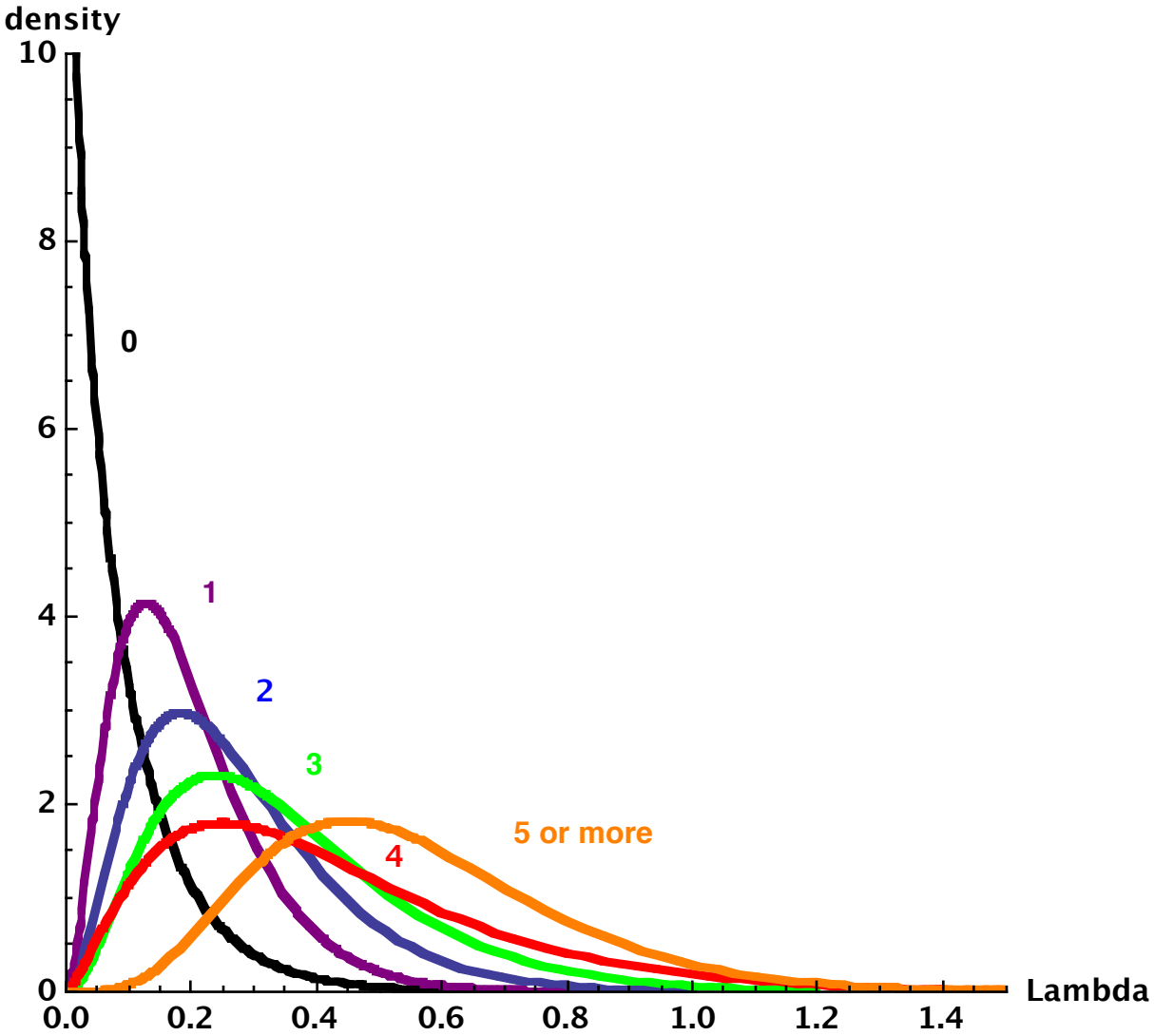
[96] Also the data was from California rather than Canada as in Bailey-Simon.

[97] Bailey-Simon instead looked at the number years a car had been claims-free.

[98] I believe it was over the same three year period as the violations.

Number of Violations	Fitted r	Fitted β
0	0.84	0.103
1	2.90	0.067
2	3.01	0.091
3	3.05	0.116
4	2.42	0.176
5 or more	5.37	0.103
Total	0.89	0.184

Then one can infer the parameters of the Gamma: r = α and $\beta = \theta$.
Here is a graph of the Gamma Distributions for the different groups of number of violations:[99]



[99] The group with no violations includes those licensed drivers who did not drive or drove only a minimal number of miles; this probably explains why its Gamma Distribution has a mode of zero. (Alpha is less than one.)

While each violations group is more homogeneous than the overall set of drivers, there is still lots of variation in expected mean frequency between drivers within a group.

One way to measure the homogeneity of each violations group is via the coefficient of variation (CV), the ratio of the standard deviation to the mean. For the Gamma Distribution the CV is $1/\sqrt{\alpha}$.

Number of Violations	Fitted alpha	CV
0	0.84	1.09
1	2.90	0.59
2	3.01	0.57
3	3.05	0.57
4	2.42	0.64
5 or more	5.37	0.43
Total	0.89	1.06

Based on the this measure, the group of those drivers with no violations is significantly more heterogeneous than the other groupings.[100]

Also as we would expect there is lots of overlap between the different groups. Here are the 10th and 90th percentile of the distributions of lambdas for the different groups.[101]

Number of Violations	Tenth Percentile	Ninetieth Percentile
0	0.006	0.209
1	0.070	0.347
2	0.101	0.486
3	0.131	0.625
4	0.134	0.793
5 or more	0.278	0.872
Total	0.014	0.388

Based on the Gamma Distribution inferred for all of the drivers in total, three years of accident data would be given 15.5% credibility for predicting future accident frequency.[102] This compares to three year credibilities in Bailey-Simon ranging from 5.9% to 9.9%.

---

[100] As mentioned before, the no violations group is mixture of those who did not drive a significant number of miles and those did, making it more heterogeneous.

[101] Recall that these are three year accident frequencies.

[102] The overall Gamma Distribution of three year mean frequencies has $\theta = \beta = 0.184$.
The Buhlmann Credibility parameter is $K = 1/\theta = 5.435$.
However, here we have treated three years of data as one draw from the risk process.
$Z = 1 / (1 + 5.435) = 15.5\%$.

However, these credibilities are not comparable because of a number of reasons including:
1. Here we are looking at drivers rather than cars as in Bailey-Simon.
2. Different overall mean annual frequencies.
3. The 15.5% credibility would be in the absence of any classifications or territories, while in Bailey-Simon cars were divided between five classifications.[103]
4. Here we have the Buhlmann Credibility based on a Gamma-Poisson model, while in Bailey-Simon the credibility was based on the indicated claims-free discount.

Assuming, one divided the drivers into classes based on their number of violations, one could infer the credibility of three years of accident data from the Gamma fit to each violation group:[104]

Number of Violations	Fitted θ	Three Year Credibility
0	0.103	9.3%
1	0.067	6.3%
2	0.091	8.3%
3	0.116	10.4%
4	0.176	15.0%
5 or more	0.103	9.3%

Based on the above, we might give about 8% credibility to three years of accident data from drivers, if drivers were classified solely based on the number of violations they had over the last three years.

There is not enough information to infer what the credibility assigned to three years of either accident or violation data should be if there were a reasonable set of classifications and territories. The appropriate credibility given to the individual's experience would be less with a class and territory plan than in the absence of one. The better the class plan, the less credibility should be to assigned to the individual's experience in individual risk rating.

There is not enough information to infer what credibility should be assigned to three years of accident and violation data combined.[105] However, more credibility would be assigned than would be assigned to either the accident or violation data separately. Again the appropriate credibility given to the individual's experience would be less with a class and territory plan than in the absence of one.

---

[103] The better the class plan, the lower the credibility given to the experience of the individual.
[104] There are only 2300 to 4300 drivers in each of the last 3 categories, so there is considerable random fluctuation.
[105] One could just add the number of violations and accidents. Instead one could assign different numbers of "points" to different types of moving violations, and different numbers of points to different severities of at-fault and single vehicle accidents, as is done in some Safe Driver Insurance Plans.

Another Example, California Female Private Passenger Auto Drivers:[106]

Here is another example similar to that in Bailey-Simon. The data and analysis are different. Specifically, the data was on drivers rather than cars, tracked drivers over many more years than three, and the analysis was similar to that in Mahler's syllabus reading on shifting risk parameters.[107]

Number of Years	Years Between Data and Estimate					
of Data Used	1 (Most recent)	2	3	4	5	Total
1	3.2% -	-	-	-	-	3.2%
2	3.1%	2.9%	-	-	-	6.0%
3	3.1%	2.8%	2.6%	-	-	8.5%
4	3.0%	2.7%	2.5%	2.3%	-	10.5%
5	3.0%	2.7%	2.4%	2.2%	2.1%	12.4%

Due to shifting risk parameters, the credibilities given to more distant years are less than those given to more recent years.

Also note that the total of the credibilities goes up more slowly than linearly with the number of years of data used. This is the same important pattern noted by Bailey and Simon for their data, although the increase is much closer to linear here than in Bailey-Simon.[108] Nevertheless, the increase in credibilities is less than would be expected from the Buhlmann Credibility formula: $Z = N / (N+K)$.

---

[106] Taken from Table 4, in "A Markov Chain Model of Shifting Risk Parameters," by Howard C. Mahler, PCAS 1997, not on the syllabus. These are least squares credibilities for no delay in receiving data. They were solved for in a manner parallel to "An Example of Credibility and Shifting Risk Parameters," by Howard C. Mahler.

[107] A fraction of drivers licensed in a state will not drive during a year, at least in that state. Some cars will be driven frequently by different drivers during a year. In any case, modeling licensed drivers is somewhat different than modeling the experience of insured cars.

[108] The credibilities depend on among other things the data used. Unlike the Bailey-Simon data from Canada, this data is not divided into classes and only includes female drivers. (There is another similar data set with male drivers.)

Fitting a Model of Shifting Risk Parameters to the Bailey-Simon Credibilities:

Assume that the covariance structure between years of data is of the form:[109]
$Cov[X_i, X_j] = a\, \rho^{|i-j|} + b\, \delta_{ij}$, where δ_{ij} is zero if $i \neq j$ and one if $i=j$.
Since one can multiply the covariances by any constant and not change the least squares credibilities, for convenience let us take $b = 1$, so that $Cov[X_i, X_j] = a\, \rho^{|i-j|} + \delta_{ij}$.[110]

Then the covariance matrix:

Year 1	$1+a$	$a\rho$	$a\rho^2$	$a\rho^3$
Year 2	$a\rho$	$1+a$	$a\rho$	$a\rho^2$
Year 3	$a\rho^2$	$a\rho$	$1+a$	$a\rho$
Year 4	$a\rho^3$	$a\rho^2$	$a\rho$	$1+a$

Then applying credibility Z to the average of N years of data with no delay:[111]

$$Z = N \frac{\sum_{l=1}^{N} Cov[X_i, X_{N+1}]}{\sum_{j=1}^{N}\sum_{l=1}^{N} Cov[X_i, X_j]}.$$

For one year of data: $Z = Cov[X_1, X_2] / Var[X_1] = a\rho / (1+a)$.

For two years of data:

$$Z = 2 \frac{Cov[X_1, X_3] + Cov[X_2, X_3]}{Var[X_1] + Cov[X_1, X_2] + Cov[X_2, X_1] + Var[X_2]} = 2 \frac{a\rho + a\rho^2}{2 + 2a + 2a\rho}.$$

For three years of data:[112]

$$Z = 3 \frac{Cov[X_1, X_4] + Cov[X_2, X_4] + Cov[X_3, X_4]}{3Var[X] + 2Cov[X_1, X_2] + 2Cov[X_1, X_3] + 2Cov[X_2, X_3]} = 3 \frac{a\rho + a\rho^2 + a\rho^3}{3 + 3a + 4a\rho + 2a\rho^2}.$$

---

[109] For $\rho < 1$, this models shifting risk parameters over time. This is an approximation to the form in "A Markov Chain Model of Shifting Risk Parameters," by Howard C. Mahler, PCAS 1997.
[110] Taking $b = 1$, then $1/a$ is similar to the Buhlmann Credibility Parameter K.
[111] Mathematically equivalent to equation 11.4 in "An Example of Credibility and Shifting Risk Parameters," by Howard C. Mahler.
[112] Note that if $\rho = 1$, in other words there are no shifting risk parameters, and replacing $1/a$ by K, then $Z = (3)(3a) / (3 + 9a) = 3/(3+K)$, the usual Buhlmann Credibility formula.

For example, for Class 1 in the Bailey-Simon data, the credibilities are:[113]

One Year	Two Year	Three Year
4.6%	6.0%	8.0%

Setting these credibilities for one and two years equal to the previous formulas, we get two equations in two unknowns:

$a\rho / (1+a) = 0.046$.

$$2 \frac{a\rho + a\rho^2}{2 + 2a + 2a\rho} = 0.060.$$

Solving (with the aid of a computer): $a = 0.09195$, and $\rho = 0.5463$.

Plugging these values into the previous equation for the credibility for 3 years, we get:

$$3 \frac{a\rho + a\rho^2 + a\rho^3}{3 + 3a + 4a\rho + 2a\rho^2} = 7.9\%.$$

This is a reasonable match to the 8.0% in Bailey-Simon.

Proceeding in a similar manner for the other classes, we get:[114]

Class	Fitted a	Fitted ρ	Fitted 3 Year Credibility	Bailey-Simon 3 Year Cred.
1	0.09195	0.5463	7.9%	8.0%
2	0.12919	0.3933	6.5%	6.8%
3	0.09195	0.5463	7.4%	8.0%
4	0.33620	0.2822	8.7%	9.9%
5	0.11593	0.3658	5.4%	5.9%

There is a good match for Class 1, a fair match for Classes 2, 3, and 5, but a poor match for Class 4. This is due to an inherent problem in the use of Class 4 in the claims-free analysis of Bailey-Simon, which applies to a lesser extent, to Class 5.[115]

The definitions of the classes are given in the Bailey-Simon paper. Class 1 is Pleasure-No Male Operator under 25. Class 2 is Pleasure-Non-principal Male Operator under 25. Class 3 is Business Use. Class 4 is Unmarried Owner or Principal Operator under 25. Class 5 is Married Owner or Principal Operator under 25.

---

[113] See Table 2 in Bailey-Simon.

[114] Similar to Table 2.1 in Howard Mahler's Discussion of "An Analysis of Experience Rating" by Glenn Meyers, PCAS 1987, not on the syllabus.

[115] See Howard Mahler's Discussion of "An Analysis of Experience Rating," not on the syllabus.

The key point is that one cannot have three clean years of experience unless one has been licensed for at least three years. Class 4 includes many drivers who have less than three years of driving experience. Those risks with one year of experience go into Merit Rating Class Y (clean for one year) if they are clean, and Merit Rating Class B (clean for less than one year) if they are not.

Both Merit Rating Class A (clean for three years) and Merit Rating Class X (clean for two years) contain no risks with only one year of experience. We expect drivers with only one year of experience to be worse than the average for Class 4. Thus Merit Rating Class A (clean for three years) for driving Class 4, will have a lower frequency than the average for driving Class 4, merely because all of its drivers have at least three years of experience. Thus when we compare it to the remainder of driving Class 4, the resulting Bailey-Simon credibility for three years of data is overstated. The same is true to a lesser extent for the Bailey-Simon credibility for two years of data.

Note that in the fitted model, r is the rate at which the correlations decline as we increase the years of separation. For Class 1 in Bailey-Simon $\rho = 0.55$, which compares to an approximate value of $\rho = 0.95$ for the California Driver Data.[116]

Thus this would indicate that parameters are shifting much more quickly for the Canadian data in Bailey-Simon than the California Data. I find this unlikely, and suspect that something else explains the behavior in Bailey-Simon's data in addition to shifting risk parameters.[117]

Using the model fit to the Bailey-Simon credibilities for Class 1, the least squares credibilities with no delay are by year:[118]

Number of Years	Years Between Data and Estimate					
of Data Used	1 (Most recent)	2	3	4	5	Total[119]
1	4.6%	-	-	-	-	4.6%
2	4.5%	2.3%	-	-	-	6.8%
3	4.5%	2.3%	1.2%	-	-	7.9%
4	4.5%	2.2%	1.1%	0.6%	-	8.4%
5	4.5%	2.2%	1.1%	0.6%	0.3%	8.7%

The total credibilities increase much less than linearly.

---

[116] With ρ approximately 0.94 for Female Drives and 0.97 for Male drivers. It is not clear that this difference between males and females is significant or just due to random fluctuations in the data set.

[117] In "An Analysis of Experience Rating," Glenn Meyers suggest parameter uncertainty is affecting the credibilities. Bailey-Simon mentions insureds switching classes and "the risk distribution of individual insureds has a marked skewness reflecting varying degrees of accident proneness" in addition to shifting risk parameters.

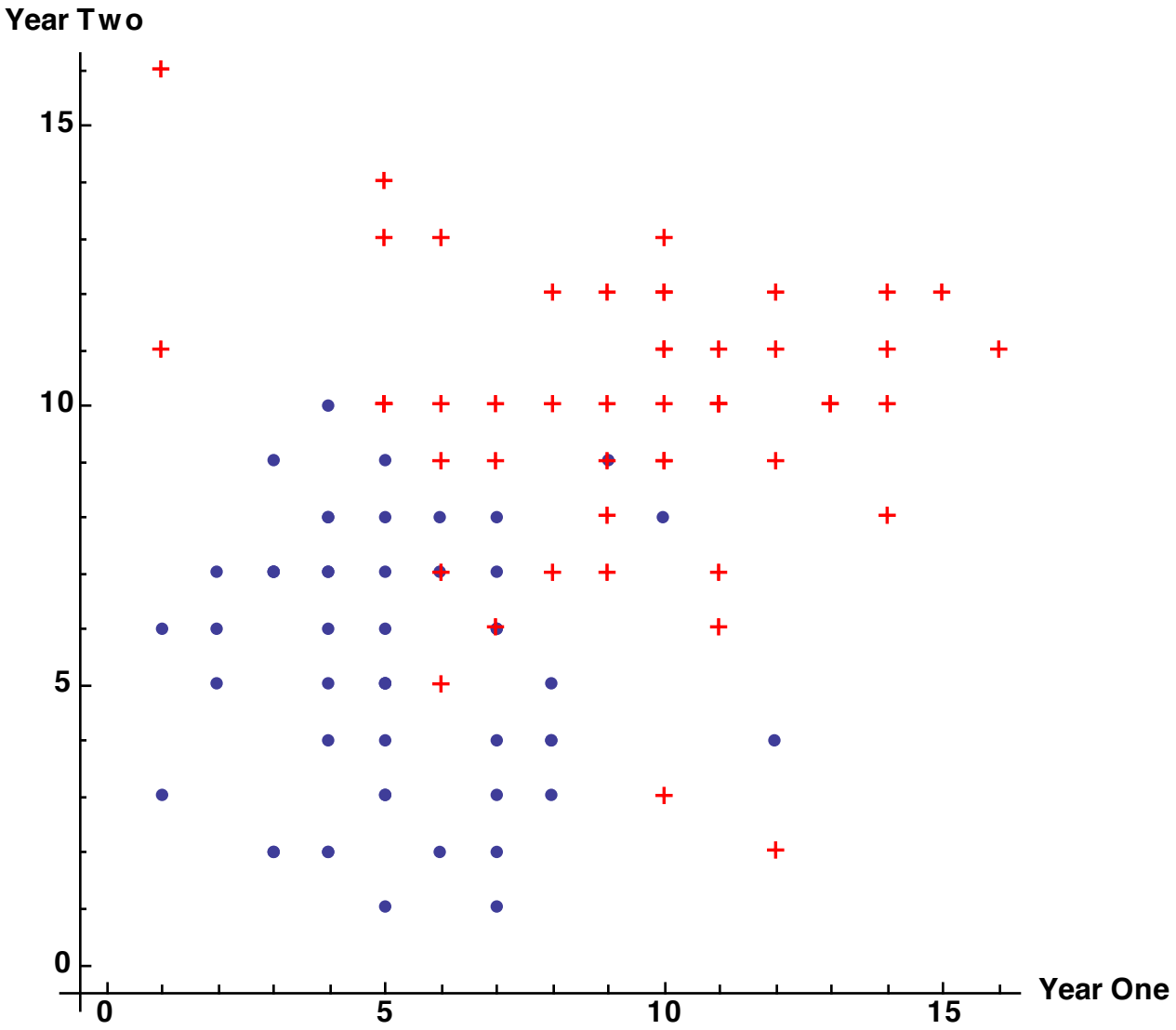
[118] Fit as per the method in "An Example of Credibility and Shifting Risk Parameters," by Howard C. Mahler.

[119] Total may differ from the sum of displayed values due to rounding of the displayed values. For ten years the sum of the credibilities is 8.94%; the total approaches a limit of 8.95%.

A Graphical Illustration:[120]

Assume two type of insureds equally likely: Poisson with $\lambda = 5$ and Poisson with $\lambda = 10$.[121]

Let us simulate two years of frequency data from 50 insureds of each type.
Those with $\lambda = 5$ are shown as blue dots, while those with $\lambda = 10$ are shown as red pluses:

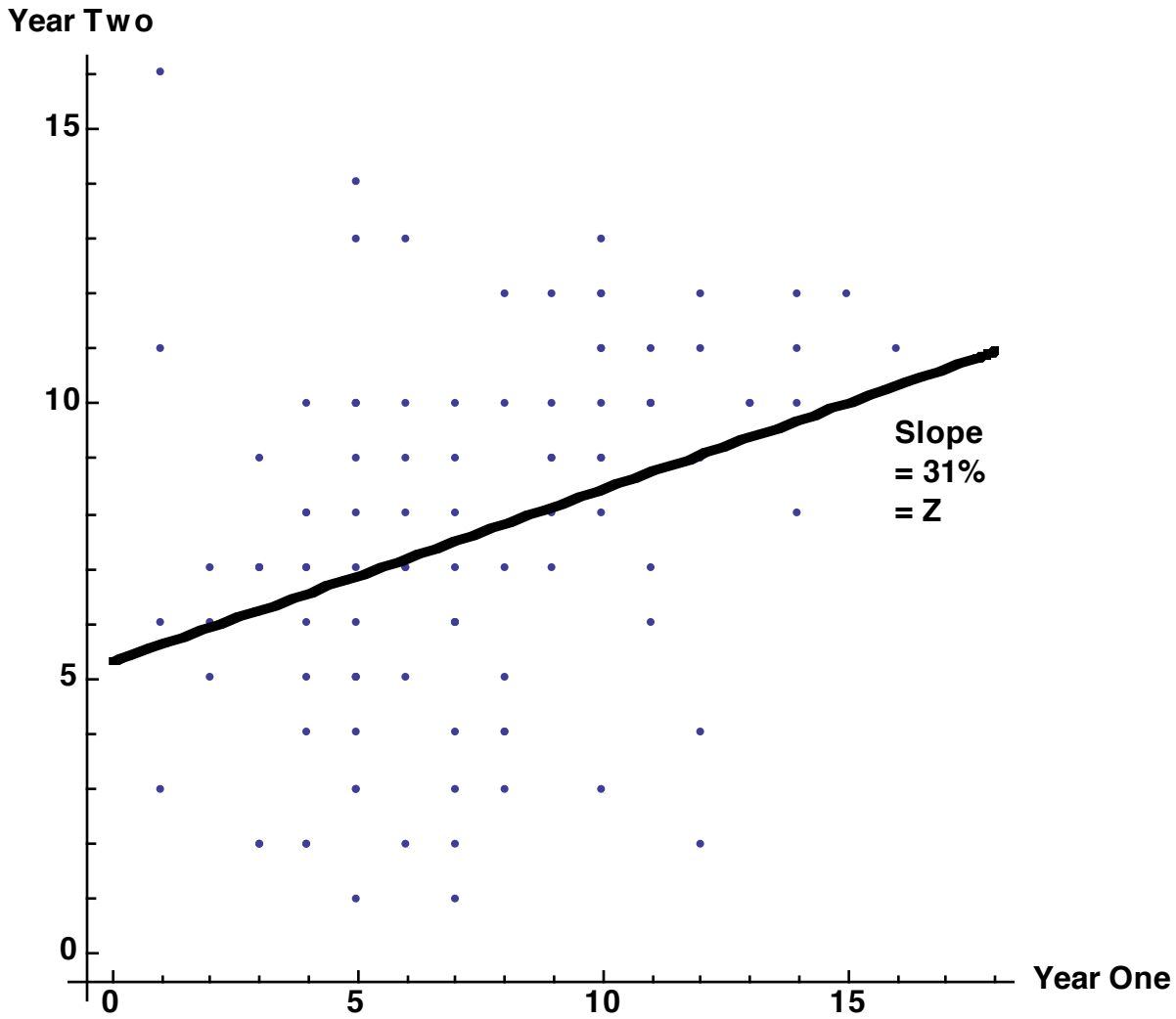


As expected, those with more claims than average in year 1 are more likely to have more claims than average in year 2. In other words, we can use past experience to predict future experience for an insured. This is the idea behind merit rating.

[120] See for example, "A Graphical Illustration of Experience Rating Credibilities," by Howard C. Mahler, PCAS 1998.
[121] The expected frequencies were chosen to be so large so that things would show up well in the graphs. Clearly this is not a model of Private Passenger Automobile Insurance.

In an insurance application of experience rating, we are assuming there is no way to distinguish the two types, other than through their past experience.

Here is a graph of the same data, without identifying the types of insureds:



A least squares line was fit to this data.[122] The slope of this fitted line is an estimate of the credibility of one year of data, in this case 31%.[123]

[122] You should not be asked to fit a regression on your exam.
[123] Due to limited data, this estimate of Z from data is subject to random fluctuation.

Important Ideas, Bailey-Simon:

Using claim frequency relative to premium instead of relative to exposures avoids distortions from maldistribution of merit rating classes between territory.
They use frequencies relative to average for those claims-free for various periods of time in order to estimate credibilities.

The alternative method of estimating a one-year credibility, compares frequencies relative to premiums vs. exposures for the group that is not claims-free.
Assuming Poisson frequency, the mean number of claims for those who were not claim free is: $\lambda / (1 - e^{-\lambda})$. Let λ = the mean claim frequency (per exposure) for the class.
M = relative premium based frequency for risks with one or more claims in the past year.

Then, $M = Z / (1 - e^{-\lambda}) + (1 - Z)(1). \Rightarrow Z = \frac{M - 1}{1 / (1 - e^{-\lambda}) - 1} = (M - 1)(e^{\lambda} - 1)$.

The ratio of three year to one year credibility is much lower than three due to:
1. Marked skewness of the distribution of accident proneness.
2. Shifting risk parameters, which Mahler discusses in more detail.
3. Movement of insureds in and out of classes.
4. The nonlinearity of the credibility formula.

Merit rating credibility varies with claim frequency. A higher frequency is like a longer experience period for a Poisson distribution. Drivers with higher expected claim frequency have higher merit rating credibility, all else being equal.
The ratio of merit rating credibility to claim frequency varies by class. Homogeneous classes have higher class credibility and lower merit rating credibility. Merit rating extract the information after class rating has done its work. **A higher ratio of merit rating credibility to claim frequency in a class indicates greater heterogeneity of the drivers in that class**. As the class plan is more refined, classes are more homogeneous and the credibility of each risk declines.

The Three Conclusions of Bailey-Simon:
(1) The experience for one car for one year has significant and measurable credibility for experience rating.
(2) In a highly refined private passenger rating classification system which reflects inherent hazard, there would not be much accuracy in an individual risk merit rating plan, but where a wide range of hazard is encompassed within a classification, credibility is much larger.
(3) If we are given one year's experience and add a second year we increase the credibility roughly two-fifths. Given two years' experience, a third year will increase the credibility by one-sixth of its two-year value.

Problems:

2.1. (1 point) Which of the following are conclusions reached by Bailey and Simon in their paper?

1. The experience for one car year has significant and measurable credibility for experience rating.
2. Merit rating adds a significant degree of accuracy to a private passenger rating system in which the classification system is highly refined, but it is of dubious value where a wide range of hazard is encompassed within a class.
3. If we are given one year's experience and add a second year, we increase the credibility roughly two-thirds.

2.2. (5 points) You are given the following data on the Adult Drivers Class for P.P. Auto Liability. Shown is the number of years they were without accident prior to 2010, the number of claims they had during 2010, and their loss cost premium during 2010 prior to the effects of Merit Rating:

Years since last accident	Premium ($ million)	Claims
5+	1520	134,200
4	70	8,900
3	80	10,400
2	90	12,500
1	100	14,400
0	140	19,600
Total	2000	200,000

a. (1 point) What is the credibility of 5 or more accident-free years of experience?
b. (1 point) What is the credibility of 4 or more accident-free years of experience?
c. (1 point) What is the credibility of 3 or more accident-free years of experience?
d. (1 point) What is the credibility of 2 or more accident-free years of experience?
e. (1 point) What is the credibility of 1 or more accident-free years of experience?

2.3. (2 points) Compare and contrast the Canadian Merit Rating Plan and the NCCI Experience Rating Plan, with respect to frequency and severity.

2.4. (1 point) Within a certain class and territory, you are given the following information for private passenger automobile insurance:

- Drivers with no claims in one year are expected to have 0.05 claims the next year.
- Drivers with 1 claim in one year are expected to have 0.12 claims the next year.

Determine the credibility of a single year of experience of a single private passenger car.

2.5. (2 points) You are an actuary at an insurer which writes private passenger automobile insurance. Alf Nadler is a critic of the insurance industry. Alf asks why for private passenger automobile insurance you use driver characteristics such as sex, age, marital status, principal place of garaging, and credit score, which are not socially acceptable, not controllable by the driver, and have no clear relation to future accidents. Alf proposes to the state legislature that insurers instead be required to use past accident history, which is socially acceptable, controllable by the insured, and has a clear relation to the expected future accidents. You are helping your company respond to Alf's proposal. What are some actuarial points you think your company's representative should make?

2.6. (1 point) Why do Bailey and Simon calculate claim frequency based on premiums rather than on car years when determining the credibility of claim-free experience?
A. Reliable data in terms of car years was not available at the level of detail required.
B. Premium as an exposure base adjusts for inflation from one year to another.
C. Because the same manual rates apply to each merit rating class, there was no material difference between the two exposure bases.
D. Premium as the denominator avoids distortion caused by variation in claim severity by territory.
E. Premium as the exposure base avoids distortion caused by variation in claim frequency by territory.

2.7. (1 point) The average pure premium in a territory for a class of private passenger automobile cars is $500 during 2012. You look at those cars within that class and territory that had no claims during 2011. The average pure premium for these cars during 2012 is $470. How much credibility would you give to a single private passenger car?

2.8. (2 points) Bailey and Simon present two methods of estimating a credibility for one year of data from a single private passenger car both based on data for the number of claims. Compare and contrast these two methods.

2.9. (1 point) If Bailey and Simon used claim frequency relative to car years instead of premium, their estimates of merit rating credibility would be:

	Cars with at least One Year Claim Free	Cars with No Claim Free Years
A.	understated	overstated
B.	overstated	understated
C	understated	understated
D.	overstated	overstated
E.	unbiased	unbiased

2.10. (1 point) You are examining experience for private passenger automobile liability for 2 classes. Class 1 and 2 are similar, except class 1 has a mean frequency of 4%, while class 2 has a mean frequency of 12%.
Compare and discuss the Merit Rating credibilities of a single car from Class 1 for three years and a single car from Class 2 for one year.

2.11. (4.5 points) Based on Bailey and Simon's paper "An Actuarial Note on the Credibility of Experience of a Single Private Passenger Car" and the information given below, calculate the credibilities that can be assigned to the experience of a single private passenger car from each of the following two groups:
a. (1.5 points) The group of risks that have been claim free for one (1) or more years.
b. (1.5 points) The group of risks that have been claim free for no (0) years.
c. (1.5 points) Discuss why the techniques in parts (a) and (b) usually give different estimates of the credibility of one year of data.

Group	Number of Years Claim Free	Earned Car Years	Earned Premium at Present B Rates	Number of Claims Incurred
A	3 or more	185,000	225,000,000	18,200
X	2	12,000	15,000,000	1,400
Y	1	15,000	20,000,000	2,200
B	0	28,000	40,000,000	5,200
Total		240,000	300,000,000	27,000

2.12. (1 point) You are examining experience for private passenger automobile liability for 4 classes: retired drivers, young unmarried males, business use, and all others.
For which class would you expect to find the highest credibility for one year from a single car, relative to claim frequency? Briefly explain why.

2.13. (3 points) Les N. DeRisk is an actuary who is studying personal auto liability insurance for drivers aged 30 to 55.
Les assumes that personal auto claims are independent events; a claim in one week does not affect the likelihood of claims in other weeks.
Les correlates claims in years X and X+1 and finds that:
• Drivers with more claims in Year X are more likely to have claims in Year X+1.
• Across a large number of drivers, the correlation of the number of claims in Year X+1 and Year X for individual drivers is 10%.
a. (1 point) Does this correlation negate the assumption that claims are independent?
b. (1 point) How should Les test the assumption that claims are independent?
c. (0.5 point) Does the 10% correlation imply a 10% merit rating credibility for one year of data?
d. (0.5 point) How should Les infer the merit rating credibility?

2.14. (3 points) An insurance company has a private passenger auto book of business. There is the following claims experience for Class 1 in State X:

Territory	Earned Premium at Present Rates Prior to Merit Rating	Earned Car Years	Number of Claims
A	$15,000,000	20,000	800
B	$25,000,000	28,000	1250
C	$30,000,000	30,000	1300
D	$25,000,000	23,000	1100
E	$20,000,000	17,000	800
Total	$115,000,000	118,000	5250

You will be trying to determine the credibility of a single private passenger car for Class 1 in State X, by comparing the experience of those who are claims-free for various periods of time to the experience of all cars in Class 1 in State X.
Which ratio would be more appropriate to use in this analysis:

$$\frac{\text{Number of Claims}}{\text{Number of Earned Car Years}} \text{ or } \frac{\text{Number of Claims}}{\text{Dollars of Earned Premiums}}?$$

Justify your selection.
Is there some other ratio that you would use instead of these two?

2.15. (2 points) Using the procedures and formulas from Bailey and Simon's paper "An Actuarial Note on the Credibility of Experience of a Single Private Passenger Car," determine which of the current classes exhibits less variation of individual hazards than the others.
Use the data shown below:

	Claim Frequency per $1,000 Earned Premium	Earned Premium per Earned Car Year	Credibility of 3 years of Data from a Single Car
Class 1	0.263	$300	5.8%
Class 2	0.369	$400	9.3%
Class 3	0.311	$350	8.1%

Assume that the earned premiums are adjusted to a common current rate level.
Show all work.

2.16. (4.5 points) Use the following information for private passenger automobile insurance in the province of Manaberta:

- There are two territories with the same number of car years in each.

Territory	Average Premium	Average Frequency Per Car Year	Average Severity
1	400	10%	2400
2	500	8%	3750

- For those cars that are claims free for at least the last 3 years:

Territory	Car Years	Premium	Subsequent Year Number of Claims
1	100,000	38 million	9000
2	110,000	53 million	8100

In each case, determine the credibility for three years of data.

(a) (0.5 point) Combining the data for the two territories, and using premiums as the denominator of "claim frequency".

(b) (0.5 point) Combining the data for the two territories, and using car years as the denominator of claim frequency.

(c) (1 point) For each territory separately, and using premiums as the denominator of "claim frequency".

(d) (1 point) For each territory separately, and using car years as the denominator of claim frequency.

(e) (1.5 points) Discuss the differences in the results in the previous parts.

2.17. (2 points)

You are given the following private passenger automobile results for the state of Fremont.

Using the techniques from Bailey and Simon's "An Actuarial Note on the Credibility of a Single Private Passenger Car," answer the questions below:

Class	Claim Frequency per Car Year	One-year Credibility		Three-year Credibility
1	0.07	0.05	0.10	
2	0.08	0.09		0.17
3	0.09	0.08		0.17

a. (1 point) For which class do its insured have more stable expected claim frequencies over the three year period?

Assume that there is no change in the exposures in each class during the three years and that the risk distribution in each class is not markedly skewed. Explain your answer.

b. (1 point) Which class has less variation in expected claim frequency between individual risks within its class? Explain your answer.

2.18. (2 points) For a specific class, the following data shows the experience of a merit rating plan.

Merit Rating	Number of Accident-Free Years	Earned Premium at Present B Rates	Number of Incurred Claims
A	3 or More	$2400 million	12,000
X	2	$200 million	1200
Y	1	$220 million	1400
B	0	$380 million	2600
	Total	$3200 million	17,200

The base rate (for Merit Rating B) is $800 per exposure for this class.
Calculate the appropriate premium for an exposure that is accident free for one or more years.

2.19. (1963, CAS Fellowship Exam IV, part b, Q.9)
In “An Actuarial Note on the Credibility of Experience of a Single Private Passenger Car”, relative claim frequency was calculated on the basis of premium rather than car years.
(a) Why was this approach taken?
(b) What are the assumptions underlying this approach?

2.20. (9, 11/88, Q.11a) (1 point) The 1986 policy year collision experience of a sample of 100,000 cars, each of which had been insured for at least the preceding three years, was tabulated as follows:

Merit Rating Class Number of Years Claims-Free Prior to 1986 Policy Year	Policy Year 1986 Exposure (Car-Years)	Policy Year 1986 Number of Claims
3 or more	71,000	7,800
2	9,000	1,400
1	10,000	1,600
0	10,000	1,700
Total	100,000	12,500

Use the method of Bailey and Simon in their paper "An Actuarial Note on the Credibility of Experience of a Single Private Passenger Car" to estimate the credibility of the experience of one car for one year.

2.21. (9, 11/88, Q.12) (1 point) You are the actuary for the XYZ Insurance Company. Currently, you are considering implementing an experience rating program for your private passenger automobile insureds based on each insured's experience. Your analysis shows that, while an insured's past claim frequency is very credible in predicting future claim frequency, an insured's past loss ratio is not very credible in predicting the future loss ratio. Based on Bailey and Simon's paper "An Actuarial Note on the Credibility of Experience of a Single Private Passenger Car", list two potential nonrandom causes of this phenomenon.

2.22. (9, 11/94, Q.31) (2 points) Based on the methodology and notation used by Bailey and Simon in "An Actuarial Note on the Credibility of Experience of a Single Private Passenger Car," and the table below calculate the credibility for category B risks (i.e., risks whose number of claims-free years equals zero) for a one-year experience period. (You can assume that the Poisson distribution reasonably approximates the distribution of observed claim counts among the risks from all merit rating groups combined.) Show all of your work.

Merit Rating (Number of Accident-Free Years)	Earned Car Years	Earned Premium at Present Category B Rates	Number of Claims Incurred
A(3+)	3,005,000	195,400,000	260,000
X(2)	148,000	10,700,000	18,000
Y(1)	184,000	13,200,000	25,000
B(0)	330,000	23,000,000	46,000
Total	3,667,000	242,300,000	349,000

2.23. (2 points) In the previous question, 9, 11/94, Q.31, assume instead that the Geometric distribution reasonably approximates the distribution of observed claim counts among the risks from all merit rating groups combined. Calculate the credibility for category B risks.

2.24. (9, 11/95, Q.6) (1 point) According to Hazam's discussion of Bailey and Simon's paper "An Actuarial Note on the Credibility of Experience of a Single Private Passenger Car," which of the following are true?

1. For a study like that presented by Bailey and Simon, the use of premium as a base is an improvement over the use of exposure as a base.
2. Using a premium base eliminates the maldistribution only if high frequency territories are also high premium territories and if territorial differentials are proper.
3. Bailey and Simon's statement "the credibilities for experience periods of one, two, and three years would be expected to vary approximately in proportion to the number of years" holds largely true only for low credibilities.

Comment: I have rewritten this past exam question.

2.25. (9, 11/95, Q.30) (3 points) Based on Bailey and Simon's paper "An Actuarial Note on the Credibility of Experience of a Single Private Passenger Car" and the information given below, calculate the credibilities that can be assigned to the experience of a single private passenger car from each of the following two groups:
a. (1.5 points) The group of risks that have been claim free for two (2) or more years.
b. (1.5 points) The group of risks that have been claim free for no (0) years.
Show all work.

Group	Number of Years Claim Free	Earned Car Years	Earned Premium at Present D Rates	Number of Claims Incurred
A	3 or more	650,000	390,000,000	54,250
B	2	200,000	120,000,000	21,000
C	1	75,000	45,000,000	10,125
D	0	75,000	45,000,000	14,625
Total		1,000,000	600,000,000	100,000

2.26. (9, 11/95, Q.32) (3 points) You have been retained as a consulting actuary for Hirisk Auto Insurance Company. The company has asked for you to determine if any of the three classifications in use is possibly in need of further refinement. The only data available are shown below:

	Claim Frequency Per $1,000 Earned Premium
Class A Total	1.625
Class B Total	1.750
Class C Total	2.212

Only Risks with 3 or More Years Loss Free	Earned Premium Per Earned Car Year	Credibility of a Single Risk
Class A	$150	0.082
Class B	$148	0.046
Class C	$190	0.079

Using the procedures and formulas from Bailey and Simon's paper "An Actuarial Note on the Credibility of Experience of a Single Private Passenger Car," determine whether one or more of the current classes exhibit(s) more variation of individual hazards than do(es) the other(s). Assume that the earned premiums are adjusted to a common current rate level. Show all work.

2.27. (9, 11/96, Q.50) (2 points)
You are given the following private passenger automobile results for a hypothetical state.
Using the techniques from Bailey and Simon's "An Actuarial Note on the Credibility of a Single Private Passenger Car," answer the questions below:

Class	Description
A	Pleasure Class - Unmarried Male Operator under age 25
B	Pleasure Class - Unmarried Female Operator under age 25
C	Pleasure Class - Operator over age 55

Class	1995 Claim Frequency	1995 One-year Credibility	1993-1995 Three-year Credibility
A	0.12	0.18	0.36
B	0.10	0.08	0.22
C	0.08	0.16	0.48

a. (1 point) Which class has a more stable claim frequency over the three year period?
Assume that there is no change in the exposures in each class during the three years and that the risk distribution in each class is not markedly skewed. Explain your answer.
b. (1 point) Which class has less variability in claim frequency within its class?
Explain your answer.

2.28. (9, 11/97, Q.19) (1 point) According to Bailey and Simon's "An Actuarial Note on the Credibility of Experience of a Single Private Passenger Car," which of the following are true?
1. Relative claim frequency is calculated on a premium basis to avoid biases due to the fact that exposure based frequency varies by territory.
2. Credibility for experience rating depends only on the volume of data in the experience period.
3. The experience for one car for one year has significant and measurable credibility for experience rating.

2.29. (9, 11/98, Q.26) (3 points) Based on Bailey and Simon's "An Actuarial Note on the Credibility of Experience of a Single Private Passenger Car," answer the following questions.
a. (2 points) Using the information below, calculate the number of claim incurred for Group C. Show all work.

Group	Number of Years Claim Free	Earned Car Years	Earned Premium at Present Group D Rates	Number of Claims Incurred
A	3 or more	700,000	$420,000	62,376
B	2	175,000	$105,000	15,955
C	1	100,000	$60,000	?????
D	0	25,000	$15,000	?????
Totals		1,000,000	$600,000	98,000

Credibility for the group of risks with one or more claim-free years (Z) = 0.086
b. (0.5 point) What conclusion do the authors reach with respect to merit rating using one year's worth of experience?
c. (0.5 point) In a highly refined private passenger rating classification system, what relative credibilities would the authors conclude should be assigned to the experience of an individual risk compared to the experience of a class?

2.30. (9, 11/99, Q.1) (1 point) In Bailey and Simon's "An Actuarial Note on the Credibility of Experience of a Single Private Passenger Car," the authors state that under certain conditions, the credibilities associated with experience periods of one, two, and three accident-free years for insureds within a given class would be expected to vary approximately in proportion to the number of years. Which of the following are reasons why this would not be true?
1. Changes in an individual insured's chance for an accident within a year.
2. Skewness in the risk distribution of individual insureds.
3. The impact of risks entering and leaving the class.

2.31. (9, 11/00, Q.32) (3 points)
Based on Bailey and Simon's "An Actuarial Note on the Credibility of Experience of a Single Private Passenger Car" and the table below, answer the following.

Private Passenger Automobile Liability - Non-Farmers
Class 3 - Business Use

Merit Rating	Earned Car Years	Earned Premium at Present B Rates	Number of Claims Incurred	Claim Frequency per $1,000 of Premium	Relative Claim Frequency
A	247,424	$25,846,000	31,964	1.237	0.920
X	15,868	$1,783,000	2,695	1.511	1.123
Y	20,369	$2,281,000	3,546	1.555	1.156
B	37,666	$4,129,000	7,565	1.832	1.362
Total	321,327	$34,039,000	45,770	1.345	1.000

where: Class A - Three or more years claim free
Class X - Two years claim free
Class Y - One year claim free
Class B - Zero years claim free

a. (1.5 points) Calculate the credibilities for a single private passenger car for one year, two years, and three years. Show all work.
b. (0.5 point) Briefly describe the relationship that Bailey and Simon expect between the three credibilities from part (a).
c. (1 point) Do the credibilities calculated in part (a) follow the relationship described in part (b)? Briefly explain why or why not.

2.32. (9, 11/01, Q.2) (1 point) According to Bailey and Simon's "An Actuarial Note on the Credibility of Experience of a Single Private Passenger Car," which of the following is false?
A. The experience for one car for one year has significant and measurable credibility for experience rating.
B. Credibility for experience rating depends on the variation of individual hazards within the class.
C. In a highly refined private passenger rating classification system that reflects inherent hazard, there would not be much accuracy in an individual risk merit rating plan.
D. In experience rating, an increase in the volume of data in the experience period increases the reliability of the indication in proportion to the square root of the volume.
E. None of A, B, C, or D are false.

2.33. (9, 11/01, Q.22) (2.5 points) Use Bailey and Simon's "An Actuarial Note on the Credibility of Experience of a Single Private Passenger Car," and Hazam's discussion to answer the following questions.
a. (1.5 points) Using the information below, calculate the credibility for 1-year and 2-year claim free periods for Class 1. Show all work.

Number of Years Claim Free	Earned Premium at Present Rates	Number of Claims Incurred	Earned Car Years
2 or more	$5,000,000	7,000	15,000
1	$7,000,000	10,000	12,250
0	$1,000,000	2,000	400
Total	$13,000,000	19,000	27,650

b. (0.5 point) What exposure base do the authors use? Explain why.
c. (0.5 point) According to Hazam, what two conditions must be met to use the exposure base described in part (b)?

2.34. (9, 11/02, Q.47) (2 points)
a. (1.5 points)
Given the following data, calculate the credibilities for 1-year and 2-year claim free periods.

A represents 3 or more years since the most recent accident.
X represents 2 years since the most recent accident.
Y represents 1 year since the most recent accident.
B represents 0 years since the most recent accident.

	Earned Car Years	Earned Premium at Present Class B Rates	Number of Claims
A	50,000	$5,500,000	5,000
X	6,500	$682,500	1,000
Y	5,000	$535,000	850
B	4,500	$490,500	900
TOTAL	66,000	$7,208,000	7,750

b. (0.5 point)
Give two possible reasons that the 2-year credibility is less than 2 times the 1-year credibility.

2.35. (9, 11/03, Q.22) (3 points) You are given the following data:

Class	Years since last accident	Actual Earned Premium at Present B Rates	Earned Car Years	Number of Claims
A	3+	375,000	2,500	200
X	2	15,000	100	12
Y	1	22,500	150	20
B	0	37,500	250	38

Assume that the same rate is charged to all insureds within a class and there have been no rate changes in or since the experience period.

a. (1 point) What is the credibility of 3 or more accident-free years of experience?

b. (1 point) What is the credibility of 1 or more accident-free years of experience?

c. (1 point) Give two possible reasons why the answer in part (a) is not 3 times the answer in part b.

2.36. (9, 11/04, Q.2) (1 point) Given the following information:

Class	Number of Years Since Most Recent Accident	Earned Car Years	Earned Premium at Present B Rates	Number of Claims
A	3 or more	10,000	$1,000,000	1000
X	2	7,000	$770,000	1155
Y	1	5,000	$625,000	1250
B	0	2,000	$400,000	1000
Total		24,500	$2,795,000	4405

Calculate the credibility of one or more accident-free years of experience.

2.37. (9, 11/05, Q.3) (3 points)

a. (2 points) Given the following information:

N = the number of drivers in the population

m = the mean claim frequency for all drivers

Mod = the credibility weighted modification factor for risks with one or more claims in the past year

Derive the formula for the credibility assigned to the experience of drivers with one or more claims in the past year. Assume that claim frequency follows a Poisson distribution.

b. (1 point) If there is a switch from a less refined class plan to a highly refined class plan, describe the likely change in the credibility assigned to an individual risk.

2.38. (9, 11/06, Q.2) (4 points)
(3 points) Given the following information about an automobile insurance portfolio:

Group	Number of Accident-Free Years	Earned Premium at Present Group D Rates	Number of Claims Incurred
A	3	$25,000,000	40,000
B	2	$8,000,000	15,000
C	1	$13,000,000	25,000
D	0	$8,000,000	30,000

Calculate the credibility of a single car for each of the following: one-year, two-year, and three-year accident-free periods.
b. (1 point) In performing the analysis in part (a) above, would using car years instead of earned premium as an exposure base be more preferable? Explain why or why not.

2.39. (9, 11/07, Q.2) (3.5 points)
a. (2 points) The following data were compiled from the ABC automobile insurance portfolio:

Group	Number of Accident-Free Years	Earned Premium at Present Group D Rates	Number of Claims Incurred
A	3 or more	$ 100,000,000	120,000
B	2	$ 10,000,000	25,000
C	1	$ 17,000,000	44,000
D	0	$ 10,000,000	36,000

Calculate the credibility of a single car for each of the following ranges of accident-free years:
I, 1 or more
ii. 2 or more
iii. 3 or more
b. (1 point)
The following table provides the single car credibility for the XYZ automobile insurance portfolio:

Accident-Free Years	Single Car Credibility
1 or More	0.06
2 or More	0.10
3 or More	0.14

Discuss two conclusions than can be drawn from the different credibility results of the ABC and XYZ portfolios.
c. (0.5 point) Explain why analysis of two portfolios with different classification plans could assign different values to the credibility of the experience of a single car.
Note: I have rewritten part (b) of this past exam question.

2.40. (9, 11/08, Q.5) (2 points)
A liability insurer collects the following data for a particular class of private passenger auto risks:

Accident-Free Years	Earned Exposures	Incurred Losses ($)
2 or more	2,500	1,000,000
1	500	500,000
0	1,000	2,500,000
Total	4.000	4,000,000

Assume the following:

- The base rate is $1,250 per exposure.
- An experience rating factor is the only factor applied to the base rate.

a. (1 point) Calculate the credibility of an exposure that is accident-free for 1 or more years.
b. (1 point) Calculate the premium for an exposure that is accident-free for 2 or more years.

2.41. (9, 11/09, Q.4) (3.5 points) The following information can be used to calculate the credibility assigned to the experience of a single private passenger car:

Group	Years Since Last Accident	Earned Car Years	Earned Premium at Present B Rates	Number of Claims
A	3 or more	650,000	400,000,000	50,000
X	2	230,000	150,000,000	20,000
Y	1	100,000	75,000,000	12,000
B	0	M	45,000,000	18,000
Total		980,000 + M	670,000,000	100,000

Assume claim counts follow a Poisson distribution.
a. (2.5 points) Calculate M, the earned car years for Group B, given that the credibility for an insured that has had no claim-free years is equal to 0.167.
b. (1 point)
Calculate the credibility for the group of risks that have been claim-free for two or more years.

2.42. (9, 11/10, Q.5) (1 point) An insurance company has a private passenger auto book of business with the following claims experience:

Group	Number of Accident-Free Years	Earned Premium at Present Group D Rates	Number of Claims Incurred
A	3 or more	60,000,000	45,000
B	2	15,000,000	15,000
C	1	20,000,000	29,300
D	0	15,000,000	18,700
		100,000,000	108,000

Calculate the credibility of a single car for a driver with one or more accident-free years.

2.43. (8, 11/11, Q.1) (3 points) An insurance company is using a merit rating plan for drivers in two states. State X has the following claims experience:

Group	Number of Accident-Free Years	Earned Premium at Present Group D Rates	Number of Claims Incurred
A	3 or more	$500,000	240
B	2	$150,000	125
C	1	$200,000	190
D	None	$300,000	300

State Y has the following relative claim frequencies for accident-free experience:

Number of Accident-Free Years	Relative Claim Frequencies to Total
3 or more	0.70
2 or more	0.77
1 or more	0.84

Assuming that no new risks enter or leave either state, use relative credibility to explain which state has more variation in an individual insured's probability of an accident.

2.44. (8, 11/12, Q.6) (2.5 points) An insurance company has a private passenger auto book of business with the following claims experience:

Territory	Years Since Last Accident	Earned Premium at Present Rates for Two Years Since Last Accident	Earned Car Years	Number of Claims	Incurred Loss
1	0	$15,000,000	15,000	5,000	$9,000,000
1	1	$125,000,000	125,000	41,000	$75,000,000
1	2+	$230,000,000	230,000	76,000	$138,000,000
2	0	$25,000,000	25,000	7,000	$16,000,000
2	1	$310,000,000	300,000	84,000	$187,000,000
2	2+	$550,000,000	535,000	147,000	$328,000,000
3	0	$10,000,000	10,000	4,000	$7,000,000
3	1	$80,000,000	100,000	35,000	$43,000,000
3	2+	$160,000,000	170,000	60,000	$100,000,00

Choose an appropriate exposure base for calculating credibility. Justify the selection.

2.45. (8, 11/14, Q.5) (2.5 points)
The following data shows the experience of a merit rating plan for a specific state.

Number of Accident-Free Years	Earned Car Years	Earned Premium ($000)		Number of Incurred Claims
3 or More	250,000	250,000		1,200
2	300,000	100,000	625	
1	25,000	100,000	750	
0	12,000	150,000	1,500	
Total	587,000	600,000		4,075

The base rate is $1,000 per exposure. No other rating variables are applicable.
a. (0.5 point) The typical exposure base used to develop the merit rating plan is earned premium. Briefly discuss two assumptions in selecting this exposure base.
b. (1.5 points) Calculate the ratio of credibility for an exposure with two or more years accident-free experience to one or more years accident-free experience.
c. (0.5 point) Calculate the premium for an exposure that is accident free for two or more years.

2.46. (8, 11/15, Q.1) (2.5 points)
An actuary is evaluating a merit rating plan for private passenger cars.
Given the following:

Number of Accident-Free Years	Earned Car Years	Number of Claims Incurred
2 or More	500,000	20,000
1	200,000	15,000
0	100,000	9,000
Total	800,000	44,000

- Frequency varies by territory.
- State law prohibits reflecting territory differences in rating.
- Annual claims for an individual driver follow a Poisson distribution.
- Claim cost distributions are similar across all drivers.

a. (0.5 point) Identify one potential issue with the exposure base used. Briefly explain whether or not earned premium would be a better choice for the exposure base.
b. (1.0 point) Calculate the credibility of one driver with one or more year's accident-free experience.
c. (1.0 point) Calculate the credibility of one driver with 0 Accident-Free years.

2.47. (8, 11/16, Q.1) (2.75 points) A group of insureds have different expected claim frequencies. The number of insureds claim-free for the past t years is as follows:

Expected Claim Frequency	t=0	t=1	t=2	t=3
0.05	50,000	47,500	45,000	44,000
0.10	50,000	45,000	43,000	36,000
0.20	25,000	20,500	16,500	14,000
Total	125,000	113,000	104,500	94,000

Determine whether the variation of an individual insured's chance for an accident changes over time.

2.48. (8, 11/17, Q.3) (1.5 points)
The following data shows the experience of a merit rating plan for private passenger vehicles. The merit rating plan uses multiple rating variables, including territory.

Number of Accident-Free Years	Earned Car Years (000s)	Earned Premium ($000s)	Number of Incurred Claims
5 or More	250	500,000	15,000
3 and 4	100	90,000	13,500
1 and 2	80	60,000	8,000
0	70	50,000	10,500
Total	500	700,000	47,000

Territory	Frequency	Average Premium
A	0.05	1,500
B	0.10	2,000
C	0.15	1,250

a. (0.75 point) Recommend and justify an exposure base for this merit rating plan.
b. (0.75 point) Calculate the relative credibility of an exposure that has been three or more years accident-free using the exposure base from part (a) above.

Solutions:

2.1. Statement 1 is conclusion #1 of Bailey-Simon.
Statement 2 is backwards from conclusion #2 of Bailey-Simon.
Conclusion #3 of Bailey-Simon states that the credibility increases roughly by (only) two-fifths.
Only statement #1 is true.

2.2. The overall claim frequency on a premium basis is: 200,000/2000 = 100.
(a) Claim frequency on a premium basis for 5 or more years claim free: 134,200/1520 = 88.289.
1 - Z = 88.289 / 100. $\Rightarrow$ Z = **11.7%**.
(b) Claim frequency on a premium basis for 4 or more years claim free:
(134,200 + 8900) / (1520 + 70) = 90.
1 - Z = 90 / 100. $\Rightarrow$ Z = **10.0%**.
(c) Claim frequency on a premium basis for 3 or more years claim free:
(134,200 + 8900 + 10,400) / (1520 + 70 + 80) = 91.916.
1 - Z = 91.916 / 100. $\Rightarrow$ Z = **8.1%**.
(d) Claim frequency on a premium basis for 2 or more years claim free:
(134,200 + 8900 + 10,400 + 12,500) / (1520 + 70 + 80 + 90) = 94.318.
1 - Z = 94.318 / 100. $\Rightarrow$ Z = **5.7%**.
(e) Claim frequency on a premium basis for 1 or more years claim free:
(134,200 + 8900 + 10,400 + 12,500 + 14,400) / (1520 + 70 + 80 + 90 + 100) = 96.989.
1 - Z = 96.989 / 100. $\Rightarrow$ Z = **3.0%**.
Comment: In part (b) those who have no claims in a 4 year window are:
those 4 years claims free plus those claims free for 5 or more years.
Different merit rating plans will have a different experience period.
Presumably this data was collected from a situation where the experience period was 5 years.
While all of these insureds are in the Adult Driver class, they may have different vehicle usage, different territories, etc. The premiums are prior to the impacts of any discounts for Merit Rating, and thus are analogous to the premium at current Group B rates in Bailey-Simon. Also they are loss costs premiums so as to get at the expected effects on frequency of the other rating factors without being distorted by fixed expenses. We could instead take premiums less expense fees.

2.3. The Merit Rating Plan results are not affected at all by severity. Claim frequency is only used in so far as the number of years claims-free. Using the total number of claims in the last three years for an individual driver would produce somewhat different results.
In the NCCI Experience Rating Plan, each claim is divided into primary and excess losses, with a split point of $5000. (Initially each loss is limited by the State Accident Limit.) The primary losses are more affected by frequency than severity, while the excess losses are more affected by severity than frequency. Primary credibilities are larger than excess credibilities. Therefore, the experience modification is more affected by frequency than severity. The ratio of excess credibility to primary credibility increases as the size of insured increases. Thus the modifications of larger insureds are more sensitive to severity than are those of smaller insureds.
Comment: Currently, some Merit Rating Plans (SDIPs) will reflect severity to a limited extent. For example, an at-fault claim of size more than $2000 might be assigned "4 points," while an at-fault claim of size less than $2000 but at least $500 is only assigned "3 points." (This is what is done in the Massachusetts SDIP.)
As one reduced the size of the split point in an experience rating plan, the primary losses became more and more like frequency; for a split point of $1, the primary losses would be the number claims. The optimal credibilities depend on the split point. Theoretically, the optimal split point depends on the size of the insured; smaller risks have a smaller optimal split point. Put another way, for smaller risks there is less predictive value to severity than for larger risks, all other things being equal. See for example, Howard C. Mahler's discussion of "An Analysis of Experience Rating" by Glenn G. Meyers, PCAS 1987, not on the syllabus.
A single private passenger car generates the same amount of data as the very smallest Workers Compensation insureds. For example, a car might have expected annual liability losses of $1000.
A Workers Compensation insured with only $1000 in expected annual losses is too small to qualify for Experience Rating. (See the Table on page 16 of the plan. For example, a state might require $4000 in average annual premium, which correspond to about $3000 in expected annual losses.)
Such very small Workers Compensation insureds have too little data to fit into the Experience Rating Plan. However, predicting the future losses of such very small risks would benefit from a simple Merit Plan which just used frequency in a limited manner similar to the Canadian Merit Plan.

2.4. Let Y be the expected claim frequency for the average risk and Z be the one-year credibility for a single car. We have two equations:
Z x (0 claims) + (1 - Z) x (Y claims) = 0.05 claims.
Z x (1 claim) + (1 - Z) x (Y claims) = 0.12 claims.
Solving, Z = (0.12 - 0.05) / 1 = **7%**.
Comment: A similar idea to what Bailey and Simon do, but somewhat different.

2.5. Predictive accuracy is an important part of allowing any private insurance system to operate;
it allows rates to be not unfairly discriminatory.
The driver characteristics we use all have been shown to have significant value in predicting the future losses of insureds. In addition, we use the past experience of an insured in our Safe Driver Insurance Plan in order to improve that prediction.
However, the past experience of a single private passenger car has a lot of randomness. Thus, there is not enough volume of data from a single private passenger car to by itself get an accurate prediction of future experience. (According to Bailey-Simon, the number of claims for one car over three years has about 6% to 10% credibility. The credibility would be higher in the absence of any class plan. Nevertheless, in the context of one car having 5 times or more the expected losses of another car in the same state, this is relatively small.) Past experience of a single private passenger car, including moving violations, is a useful supplement to a well designed, refined class plan. However, past experience can not replace the class plan.
Comment: Whether certain risk characteristics are socially acceptable is a matter of opinion, and not something actuarial. While controllability is desirable it is not necessary for a classification variable.

2.6. E. Low rated territories will have lower expected frequencies and thus more insureds who are claims-free for 3 or more years. Using premium based relative frequencies adjusts for this approximately.

2.7. \$470 = (0)Z + (1-Z)(\$500). Z = 1 - 470/500 = **6%**.
Comment: A similar idea to what Bailey and Simon do, but somewhat different.

2.8. In both methods one divides the data for a single class into Merit Rating Groups based on how many years the car has been claims-free.
In the first method, one compares the subsequent premium based frequency for those who are claims free for at least one year to the overall. This ratio is 1 - Z.
The second method instead compares those who are not claims free (Group B) to average.
One gets a modification M for Group B as in the first method.
Using a Poisson assumption and the average exposures based frequency for the class, one determines the average exposure based frequency for those in Group B for the experience period.
Using this together with M, one can back out a credibility for Group B:

$$Z = \frac{M - 1}{(\text{Group B experience period frequency relative to average}) - 1}.$$

Note that the first method uses neither exposures nor an assumption of the distributional form of the frequency.
Comment: The first method is also applied to estimate two and three year credibilities.
M = relative premium based frequency for risks with one or more claims in the past year.
Let λ = the mean claim frequency (per exposure) for the class.
Group B has a frequency relative to average within its class of: $1 / (1 - e^{-\lambda})$.

2.9. D. The claims-free group contains more insureds from low rated-territories, which makes their future exposure based frequency better than it otherwise would be; the estimated credibilities from comparing to average are too big. The not claims-free group contains more insureds from high
rated-territories, which makes their future exposure based frequency worse than it otherwise would be; the estimated credibilities are again too big.

2.10. The volume of data is the same in each case; (3)(4%) = 12%. However, shifting risk parameters make the more distant years of data less valuable for predicting the future. Therefore, I would expect the one year for a driver from Class 2 to have more credibility than three years of data for a driver from Class 1.

2.11. a. The overall premium based frequency is: 27,000 / 300 = 90.
The premium based frequency for those claims-free for 1 or more years (A + X + Y) is:
(18,200 + 1400 + 2200) / (225 + 15 + 20) = 83.85.
1 - Z = 83.85 / 90. $\Rightarrow$ Z = **6.8%**.
b. The premium based frequency for those claims-free for 0 years (B) is: 5200 / 40 = 130.
Thus the modification for Group B is:
Future Relative Claim Frequency = 130 / 90 = 1.444.
Overall frequency per exposure is: 27,000 / 240,000 = 0.1125.
Given the Poisson assumption, the relative observed frequency for those who had at least one claim is: $1 / (1 - e^{-\lambda}) = 1 / (1 - e^{-0.1125}) = 9.398$.
Thus we must have: 1.444 = Z 9.398 + (1 - Z) 1.
$\Rightarrow$ Z = (1.444 - 1) / (9.398 - 1) = **5.3%**.
c. As always with finite data sets we have random fluctuation.
In addition, each technique makes assumptions and approximations. The premium based frequencies only approximately adjust for the maldistribution of the Groups by territory. In part (b) we had to make use of a Poisson assumption.
However, more fundamentally, we are measuring two somewhat different things. In part (a) we are attempting to back out the weight that would have done best in predicting the future experience of those insureds who had no claims this year (A + X + Y). In part (b), we are attempting to back out the weight that would have done best in predicting the future experience of those insureds who had at least one claims this year (B).
The Bayes Analysis estimates for different groups, those with 0 claims, those with 1 claim, those with 2 claims, etc. usually do <u>not</u> lie upon a straight line. (Only in special cases such as the Gamma-Poisson, are the Bayes estimates along a straight line, and thus Buhlmann Credibility equals Bayes Analysis.) Thus the optimal weight to use in each of these situations would be different.
<u>Comment</u>: The Buhlmann credibility is the slope of the weighted least squares line fit to the Bayes Estimates as function of the observations. Thus we would expect the estimates in parts (a) and (b) to differ from each other as well as the Buhlmann credibility.

2.12. The most heterogeneous class would have the highest credibility for a car. I would expect this to be the "all other" class, since it contains many different types of drivers with different potentials for loss.
Comment: The less homogeneous a class is, the more rely on the experience of an individual car within that class.

2.13. a. For any driver, claims may be independent.
However, the correlation compares claim rates of different drivers.
Some drivers have high claim frequency, with high expected rates in both years.
Some drivers have low claim frequency, with low expected rates in both years.
A high correlation implies that:
• Drivers are heterogeneous with stable risk parameters.
• Good drivers usually stay good for a second year; bad drivers usually stay bad for a second year.
A high correlation from year to year does not mean claims are not independent.
b. The actuary should examine the serial correlation in each driver's claim history.
If a given driver has a claim in Year X but not in Year Y, does claim frequency tend to be greater in Year X+1 than in Year Y+1?
Under independence, the answer should be no.
c. Merit rating is applied after class rating.
The 10% correlation implies that class rating plus merit rating has a 10% credibility.
In personal auto, class and territory rating separates drivers into relatively homogeneous classes.
Class and territory rating gets much of the credibility, leaving much less than 10% for merit rating.
d. The actuary should examine the correlation within classes.
Bailey-Simon examines a single class at a time; compare the future (premium based) frequency of those who were claims-free to that of the overall class.

2.14. Bailey-Simon uses $\frac{\text{Number of Car Years}}{\text{Dollars of Earned Premiums}}$, in order to adjust for the maldistribution that would result from low frequency territories having a larger portion of insureds who are claims-free.

It would be better to use premiums, provided the high rated territories have higher frequency and provided the territory relativities are correct.

Territory	Average Rate	Relative to Avg.	Frequency per Car-Year	Relative to Avg.
A	$750	0.769	4.00%	0.899
B	$893	0.916	4.46%	1.002
C	$1000	1.026	4.33%	0.973
D	$1086	1.114	4.78%	1.074
E	$1176	1.206	4.70%	1.056
Total	$975	1.000	4.45%	1.000

There is a tendency for the higher rated territories to have higher frequencies.
However, the relative average rates have a much wider spread than the relative average frequencies. Thus the average premiums largely reflect differences in severity and/or reflect incorrect territory relativities in the current rates.

Using for each subgroup (0 years claims-free, 1 year claims-free, 2 years claims free, etc.) $\frac{\text{Number of Claims}}{\text{Dollars of Earned Premiums}}$ would adjust for the differences in frequency by territory, but would significantly over-adjust due to whatever is causing the wider differences in average premium.

Using $\frac{\text{Number of Claims}}{\text{Number of Earned Car Years}}$ would <u>not</u> adjust for the differences in frequency by territory.

In this case, the other reasons for differences in average premiums seem to have a bigger effect than differences in frequency. Thus on balance I would prefer to use $\frac{\text{Number of Claims}}{\text{Number of Earned Car Years}}$ rather than $\frac{\text{Number of Claims}}{\text{Dollars of Earned Premiums}}$. We want to adjust for the different mixes of territory for the subgroups, due to the different frequencies by territory. If possible, it would probably be better to use for each subgroup (0 years claims-free, 1 year claims-free, 2 years claims free, etc.):

$$\frac{\text{Number of Claims}}{\sum_{\text{territories}} \text{(car years for subgroup in territory) (frequency within territory relative to whole class)}}.$$

The relative frequencies for the territories within Class 1 are: 0.899, 1.002, 0.973, 1.074, 1.056. Assume that the subgroup that is claim free for at least 3 years has exposures within Class 1 by territory: 17,700, 24,500, 26,300, 19,900, 14,800. Then the above denominator would be: (0.889)(17,700) + (1.002)(24,500) + (0.973)(26,300) + (1.074)(19,900) + (1.056)(14,800) = 102,876. This is less than the sum of exposures for this subgroup of 103,200, reflecting the somewhat higher proportion of low frequency territories in this subgroup than in all of Class 1.

<u>Comment</u>: Similar to 8, 11/12, Q.6.

2.15. For each class, we get the frequency per exposure by multiplying the frequency per $ premium times the premium per exposure.
For example, for Class 1: (0.000263)(300) = 7.89%.
Then take the ratio of the 3-year credibility to this frequency, as per Table 2 in Bailey-Simon.
For example, for Class 1: 5.8% / 7.890% = 0.7351.

Class	Cred.	Class Freq. per Prem.	Prem. per Expos.	Freq. per Expos.	Cred. / Freq.
1	5.8%	0.000263	300	7.890%	**0.7351**
2	9.3%	0.000369	400	14.760%	**0.6301**
3	8.1%	0.000311	350	10.885%	**0.7441**

A more homogeneous class will have a ratio of credibility for experience rating to frequency that is lower.
Thus Class 2 is more homogeneous than Classes 1 and 3;
Class 2 exhibits less variation of individual hazards than do the others.
Comment: Similar to 9, 11/95, Q.32.

2.16. (a) Assume each territory has x exposures.
Then total number of claims is: x 10% + x 8% = x 18%.
Total premium is: 400 x + 500 x = 900 x.
Overall premium based frequency is: 18% / 900 = 0.0002.
Claims-free premium based frequency is: 17,100 / 91 million = 0.000188.
Z = 1 - 0.000188 / 0.0002 = **6.0%**.
(b) Overall frequency is: (1/2)(10%) + (1/2)(8%) = 9%.
Claim free frequency is: (9000 + 8100) / (100,000 + 110,000) = 8.14%.
Z = 1 - 8.14% / 9% = **9.6%**.
(c) Overall frequency Territory 1 is: 10% / 400 = 0.00025.
Claims-free frequency Territory 1 is: 9000 / 38 million = 0.000237.
Territory 1 credibility is: 1 - 0.000237 / 0.00025 = **5.2%**.
Overall frequency Territory 2 is: 8% / 500 = 0.00016.
Claims-free frequency Territory 2 is: 8100 / 53 million = 0.000153.
Territory 2 credibility is: 1 - 0.000153 / 0.00016 = **4.4%**.
(d) Overall frequency Territory 1 is 10%.
Claim free frequency Territory 1 is: 9000 / 100,000 = 9%.
Territory 1 credibility is: 1 - 9%/10% = **10%**.
Overall frequency Territory 2 is 8%.
Claim free frequency Territory 2 is: 8100 / 110,000 = 7.36%.
Territory 2 credibility is: 1 - 7.36%/8% = **8.0%**.
(e) The pure premiums are: (10%)(2400) = 240, and (8%)(3750) = 300.
The ratio of pure premiums to average premium are: 240/400 = 60%, and 300/500 = 60%.
Thus the territory relativities appear to be correct.
However, the higher rated territory has the lower frequency.
Thus in part (a), using premiums in the denominator is not a good idea; it would not be adjusting for the differences in frequency between the territories.
Therefore, the result in part (b) is preferable to that in part (a).
There are difference in the classes within a territory in average premiums and frequencies.
If the higher rated classes are also higher frequency, then the results in part (c) using premiums in the denominator would be preferable to those in part (d) using car years in the denominator.
It makes sense that three years of data from the higher frequency territory 1 would have a larger credibility than three years of data from the lower frequency territory 2.
(However, in practical applications of a Safe Driver Insurance Plan, one would probably give the same credibility to a car year of data from any class and territory.)
Comment: The data in this question is not arranged in exactly the same way as in Bailey-Simon.
I do not have an opinion as to whether the results in part (b) or part (c) are preferable;
I would need to investigate further as to why they differ.
There are probably other reasonable answers to part (e).

2.17. (a) Bailey & Simon give 3 reasons why the credibilities increase less than linearly with number of year of data. The question has eliminated two of these reasons; the one that is left is shifting risk parameters. The faster parameters shift over time, the greater the effect of lowering the ratio of 3-year to 1-year credibility.
The ratios of three year to one year credibilities are for the given classes: 2, 1.9, and 2.1.
Thus Class 2 has been most affected by shifting risk parameters over time and Class 3 the least. Thus, the insureds in Class **3** have more stable expected claim frequencies from year to year.
(b) Less variation in individual hazard within its class is a smaller Variance of the Hypothetical Means. Such a class would have a smaller credibility all else being equal. However, a higher mean frequency would also produce a higher credibility, all else being equal.
Compare the one-year credibility to the mean frequency, the ratios are: 0.7, 1.1, and 0.9.
Thus Class **1** has less variability in expected claim frequency within its class.
As per Table 2 in Bailey-Simon, comparing the three-year credibility to the mean frequency, the ratios are: 1.4, 2.1, and 1.9. (I would prefer to use the one-year credibilities, which are less affected by shifting risk parameters.) A lower ratio indicates that lower relative credibility is assigned, meaning a more homogeneous class.
Thus Class **1** has less variability in expected claim frequency within its class.
Comment: Similar to 9, 11/96, Q.50. Part b is similar to 8, 11/11, Q.1.
See Table 2 in Bailey-Simon.

2.18. The indicated rate compared to average for those who are one or more years claims free is: $\frac{(12000 + 1200 + 1400) / (2400 + 200 + 220)}{17{,}200/3200} = 5.1773 / 5.375 = 0.9632.$

The indicated rate compared to average for those who are not claims free is:
$\frac{2600/380}{17{,}200/3200} = 6.8421 / 5.375 = 1.2729.$

Thus the appropriate premium for an exposure that is accident free for one or more years is:
(0.9632/1.2729) ($800) = **$605.36**.
Alternately, (5.1773/6.8421) ($800) = **$605.35**.
Comment: Similar to 8, 11/14, Q.5.

2.19. (a) "Earned premiums are converted to a common rate basis by use of the relationship in the rate structure that A: X : Y :B = 65 : 80 : 90 : 100.
The authors have chosen to calculate Relative Claim Frequency on the basis of premium rather than car years. This avoids the maldistribution created by having higher claim frequency territories produce more X, Y, and B risks and also produce higher territorial premiums."
In other words, Bailey and Simon were concerned about the inherent correlation of exposures between Merit Rating Groups and territories.
We would expect that Group B (not claims free) would have a larger percentage of exposures in territories with higher than average frequencies than would Group A (claims-free for at least three years). However, we are already charging insureds in those territories more than average. If we did not adjust for that here by dividing by premiums rather than exposures, we would be double counting. This adjustment removes the impact of things that are already included in the rate structure via territory relativities.
(b) We are assuming that the territory relativities underlying the current rates are a reasonably accurate reflection of differences in frequency between territories. We are assuming that little if any of the difference in territory rates are due to differences in average severity. Similarly, we are assuming that the effect of any other classification factors other than Merit Rating that underlay the current rates accurately reflect differences in frequency and do not reflect differences in severity.

2.20. The overall frequency is: 12,500 / 100,000 = 0.125.
The frequency for those who are claims-free for at least a year is: 10,800 / 90,000 = 0.120.
Their relative frequency is: 0.120 / 0.125 = 0.96.

$1 - Z = 0.96. \Rightarrow Z =$ **4.0%**.

Alternately, the subsequent frequency for those who are not claims-free is: 1700/10,000 = 0.17.
Assuming a Poisson frequency, the average number of claims for those who were not claims-free is:

$\lambda / (1 - e^{-\lambda}) = 0.125 / (1 - e^{-0.125}) = 1.0638.$

$Z\ 1.0638 + (1 - Z)(0.125) = 0.170. \Rightarrow Z =$ **4.8%**.

Comment: Bailey-Simon uses premium based frequency. The first method is the intended solution.
Let, M = relative premium based frequency for risks with one or more claims in the past year.
Then, $Z = (M - 1) (e^{\lambda} - 1) = (0.17/0.125 - 1) (e^{0.125} - 1) = 4.8\%$.

2.21. 1. Loss ratios have premiums rather than exposures in the denominator. Premiums reflect class and territory differentials, which could account for most of the variance between the loss potential of individual insureds.
2. Severity is systematically opposite to frequency.
Comment: in the second response, we could for example have a model with two types:

Type	Mean Frequency	Mean Severity	Mean Pure Premium
1	5%	$10,000	$500
2	10%	$5,000	$500

2.22. Overall claim frequency: 349,000 / 3,667,000 = 0.0952.
Assuming Poisson, the average number of claims for Group B is:
$\lambda / (1 - e^{-\lambda}) = 0.0952 / (1 - e^{-0.0952}) = 1.0484$.
Relative frequency for Group B is: 1.0484 / 0.0952 = 11.01.
The overall premium based frequency is: 349,000 / 242,300 = 1.440.
The premium based frequency for Group B is: 46,000 / 23,000 = 2.
$\Rightarrow$ Modification for Group B is: 2/1.440 = 1.389.
Thus, 1.389 = Z 11.01 + (1-Z) 1. $\Rightarrow$ Z = **3.9%**.

2.23. Take β equal to the overall mean of 0.0952.
The probability of no claims is: $1/(1+\beta) = 1/1.0952 = 0.9131$.
Let the average number of claims for Group B be x.
$(0)(0.9131) + x(1 - 0.9131) = 0.0952. \Rightarrow x = 1.0955$.
Relative frequency for Group B is: 1.0955 / 0.0952 = 11.51.
$\Rightarrow$ Modification for Group B is 1.389.
Thus, 1.389 = Z 11.51 + (1-Z) 1. $\Rightarrow$ Z = **3.7%**.
Comment: The credibility depends only slightly on the Poisson versus Geometric assumption.

2.24. All three statements are true.

2.25. a) The overall premium based frequency is: 100,000/ 600,000 = 1/6.
The premium based frequency for those claims-free for 2 or more years (A+B) is:
(54,250 + 21,000) / (390,000 + 120,000) = 0.1475.
1 - Z = 0.1475 / (1/6). $\Rightarrow$ Z = **11.5%**.
b) The premium based frequency for those claims-free for 0 years (D) is:
14,625 / 45,000 = 0.325.
Thus the modification for Group D is:
Future Relative Claim Frequency = (0.325) / (1/6) = 1.95.
Overall frequency per exposure is: 100,000 / 1,000,000 = 0.1.
Given the Poisson assumption, the relative observed frequency for those who had at least one claim is: $1 / (1 - e^{-\lambda}) = 1 / (1 - e^{-0.1}) = 10.51$.
Thus we must have: 1.95 = Z 10.51 + (1 - Z) 1.
$\Rightarrow$ Z = (1.95 - 1) / (10.51 - 1) = **10.0%**.
Comment: In part B, $Z = \frac{\text{(Future Relative Frequency) - 1}}{\text{(Past Relative Frequency) - 1}}$.

2.26. We are not given the average premium for each class.
I will estimate that the average premium for each class is approximately such that:
(average premium for class) (1 - Z) = (average premium for 3-years claims free and in class).
Thus the average premium for Class A is: 150 / (1 - 0.082) = 163.40.
For each class, we get the frequency per exposure by multiplying the frequency per $ premium times the premium per exposure.
For example, for Class A: (0.001625)(163.4) = 26.55%.
Then take the ratio of the 3-year credibility to this frequency, as per Table 2 in Bailey-Simon.
For example, for Class A: 8.2% / 26.55% = 0.3088.

Class	Z	Class Freq. per Prem.	Claims-Free Prem. per Expo.	Class Prem. per Expo.	Freq. per Expos.	Z / Freq.
A	8.2%	0.001625	$150	$163.40	26.55%	**0.3088**
B	4.6%	0.001750	$148	$155.14	27.15%	**0.1694**
C	7.9%	0.002212	$190	$206.30	45.63%	**0.1731**

A more homogeneous class will have a ratio of credibility for experience rating to frequency that is lower.
Thus Class A is more heterogeneous than Classes B and C;
Class A exhibits more variation of individual hazards than do the others.

2.27. (a) Bailey & Simon give 3 reasons why the credibilities increase less than linearly with number of year of data. The question has eliminated two of these reasons; the one that is left is shifting risk parameters. The faster parameters shift over time, the greater the effect of lowering the ratio of 3-year to 1-year credibility.
The ratios of three year to one year credibilities are for the given classes: 2, 2.75, and 3.
Thus Class A has been most affected by shifting risk parameters over time and Class C the least. Thus, assuming that the exam question meant in which class do the insureds have more stable expected claim frequencies from year to year, that is Class **C**.
(b) Less variation in individual hazard within its class is a smaller Variance of the Hypothetical Means. Such a class would have a smaller credibility all else being equal. However, a higher mean frequency would also produce a higher credibility, all else being equal.
Compare the one-year credibility to the mean frequency, the ratios are: 1.5, 0.8, and 2.
Thus Class **B** has less variability in claim frequency within its class.
As per Table 2 in Bailey-Simon, comparing the three-year credibility to the mean frequency, the ratios are: 3, 2.2, and 6. (I would prefer to use the one-year credibilities, which are less affected by shifting risk parameters.) A lower ratio indicates that lower relative credibility is assigned, meaning a more homogeneous class. Thus Class **B** has less variability in claim frequency within its class.
Alternately, assume the one-year credibility is $1/(1+K)$. $\Rightarrow K = 1/Z - 1$.
Also assume that the Expected Value of the Process Variance is equal to the mean.
(EPV = mean if each insured has a Poisson frequency.
For comparison purposes we need only assume it is proportional.)

Class	1995 Claim Frequency	1995 One-year Credibility	K = EPV/VHM	VHM
A	0.12	0.18	1/0.18 - 1 = 4.56	0.12/4.56 = 0.0263
B	0.10	0.08	11.5	0.0087
C	0.08	0.16	5.25	0.0152

Class B has smallest ratio of VHM / (mean freq.)^2. Thus Class **B** has less variability in claim frequency within its class, as measured by the square of the coefficient of variation.
Comment: Part b would have been better if it had been worded: “Which class has less variation in expected claim frequency between individual risks within its class?”

2.28. Statement #1 is true. We would expect that Class B (not claims free) would have a larger percentage of exposures in territories with higher than average frequencies than would Class A (claims-free for at least three years). To avoid double counting effects that are already reflected in the territory relativities, we divide by premiums at base class rates; a class with a higher than the average percentage of exposures in a high frequency territory will also have a higher than average base class premium.
2. Statement #2 is false. We would also be interested in the homogeneity of a class. To the extent that the insureds in a class are more similar, the credibility for experience rating (individual risk rating) is smaller.
3. Statement #3 is true. See conclusion #1 of the paper.
Comment: As discussed on a preliminary exam, for Buhlmann credibility we would be interested in the EPV, VHM, and volume of data. In this context, the variance of hypothetical means measures how different the insureds are within a class, the expected value of the process variance measures how much random fluctuation there is in the data, and the volume of data is the number of years from an individual car.

2.29. a) Let x be the number of claims for Group C.
The frequency on a premium basis for one or more claim free years is:
$$\frac{62{,}376 + 15{,}955 + x}{420{,}000 + 105{,}000 + 60{,}000} = \frac{78{,}333 + x}{585{,}000}.$$
The overall frequency on a premium basis is: 98,000 / 600,000.
We have: $1 - Z = 1 - 0.086 = M = \frac{\text{frequency for at least one year claims free}}{\text{overall frequency}}$. $\Rightarrow$
$0.914 = \frac{78{,}333 + x}{585{,}000}$ / (98,000 / 600,000). $\Rightarrow$ x = **9000**.
b) "The experience for one car for one year has significant and measurable credibility for experience rating."
c) "In a highly refined private passenger rating classification system which reflects inherent hazard, there would not be much accuracy in an individual risk merit rating plan, but where a wide range of hazard is encompassed within a classification, credibility is much larger."
If the class system is highly refined and each class is homogeneous (not much variation in hazard), then the majority of the credibility (weight) should be assigned to the class experience rather than the individual risk experience.
Comment: See the first and second conclusions of the paper.

2.30. All three statements are true.
"The fact that the relative credibilities in Table 3 for two and three years are much less than 2.00 and 3.00 is partially caused by risks entering and leaving the class. But it can be fully accounted for only if an individual insured's chance for an accident changes from time to time within a year and from one year to the next, or if the risk distribution of individual insureds has a marked skewness reflecting varying degrees of accident proneness."

2.31. a. For three years: 1 - Z = 0.920. ⇒ Z = **8.0%**.
For two or more years claim free, claim frequency is: (31,964 + 2695) / (25,846 + 1783) = 1.254.
1 - Z = 1.254 / 1.345. ⇒ Z = **6.8%**.
For one or more years claim free (A + X + Y), claim frequency is:
(31,964 + 2695 + 3546) / (25,846 + 1783 + 2281) = 1.277.
1 - Z = 1.277 / 1.345. ⇒ Z = **5.1%**.
b. If the chance of accident for an individual risk remains constant and no risks enter or leave, then the credibility should vary approximately in proportion to the number of experience years.
c. Comparing the credibilities for one year and two years: 6.8/5.1 = 1.33 ≠ 2.
Comparing the credibilities for two years and three years: 8.0/6.8 = 1.18 ≠ 1.5.
The credibilities do not follow the expected pattern. An individual insured's chance for an accident changes over time and/or risks may be entering or leaving.
Comment: Conclusion #3 of the paper: "If we are given one year's experience and add a second year we increase the credibility roughly two-fifths. Given two years' experience, a third year will increase the credibility by one-sixth of its two-year value."
In part (a) we could use the alternate method to get a one year credibility.
The premium based frequency for those claims-free for 0 years is given as 1.362.
Overall frequency per exposure is: 45,770 / 321,327 = 0.1424.
Given the Poisson assumption, the relative observed frequency for those who had at least one claim is: $1 / (1 - e^{-\lambda}) = 1 / (1 - e^{-0.1424}) = 7.534$.
Thus we must have: 1.362 = Z 7.534 + (1 - Z) 1.
⇒ Z = (1.362 - 1) / (7.534 - 1) = **5.5%**.
A somewhat different answer than using the other method.

2.32. D. Statement #1 is conclusion #1 from the paper and thus true.
Statement #2 is true. See page 160 of the paper: "This also illustrates that credibility for experience rating depends not only on the volume of data in the experience period but also on the amount of variation of individual hazards within the class."
Statement #3 is conclusion #2 from the paper and thus true.
While Statement #4 could be true, as per the square root rule from Classical Credibility, this is not what Bailey & Simon find for their particular data.
Comment: Conclusion #3 of the paper: "If we are given one year's experience and add a second year we increase the credibility roughly two-fifths. Given two years' experience, a third year will increase the credibility by one-sixth of its two-year value."
If it followed the square root rule, then the ratio of the credibilities for 3 years and 2 years would be: $\sqrt{3/2} = 1.225$ rather than 7/6 = 1.167.

2.33. a) total frequency is: 19,000/13,000 = 1.462.
frequency for those 1 or more claim free is: (10,000 + 7,000) / (7,000 + 5,000) = 1.417.
⇒ One year credibility is: 1 - 1.417/1.462 = **3.1%**.
frequency for those 2 or more claim free is: 7,000/5,000 = 1.4.
⇒ Two year credibility is: 1 - 1.4/1.462 = **4.2%**.
b) The authors use earned premium as their exposure base to avoid the maldistribution caused when lower frequency territories produce a larger percentage of risks that are claims-free than higher frequency territories.
c) 1. Higher-frequency territories must also be higher-premium territories.
2. The territorial differentials in the current rates must be proper.
Comment: The given premiums should be prior to the effects of Merit Rating.

2.34. (a) Overall the claim frequency on a premium basis is: 7750 / 7208 = 1.0752.
For two or more years claim free (A + X), claim frequency is:
(5000 + 1000) / (5500 + 682.5) = 0.9705.
1 - Z = 0.9705 / 1.0752. ⇒ Z = **9.7%**.
For one or more years claim free (A + X + Y), claim frequency is:
(5000 + 1000 + 850) / (5500 + 682.5 + 535) = 1.0197.
1 - Z = 1.0197 / 1.0752. ⇒ Z = **5.2%**.
(b) 1. Individual insured's chance for an accident changes from time to time within a year or from one year to the next.
2. Insureds are entering or leaving the class.
3. Individuals' accident propensities in a class vary and are markedly skewed.
4. The Buhlmann Credibility formula is less than linear.

2.35. (a) Overall the claim frequency is:
(200 + 12 + 20 + 38) / (2500 + 100 + 150 + 250) = 0.09.
For three or more years claim free (A), claim frequency is: 200/2500 = 0.08.
1 - Z = 0.08 / 0.09. ⇒ Z = **11.1%**.
For one or more years claim free (A + X + Y), claim frequency is:
(200 + 12 + 20) / (2500 + 100 + 150) = 0.08436.
1 - Z = 0.08436 / 0.09. ⇒ Z = **6.3%**.
(c) 1. Individual insured's chance for an accident changes from time to time within a year or from one year to the next.
2. Insureds are entering or leaving the class.
3. The risk distribution of individual insureds has a marked skewness reflecting varying degrees of accident proneness.
Comment: The volume of data in this question is way less than used by Bailey-Simon.
If every insured within a class is charged the same rate, then we can use the usual exposure based frequencies rather than the premium based frequencies used by Bailey-Simon. It makes no difference in the result, since consistent with the statement that every insured within a class is charged the same rate, each of the premiums at current class B rates are 150 times the exposures.

2.36. Overall the claim frequency on a premium basis is: 4405 / 2795 = 1.5760.
For one or more years claim free (A + X + Y), claim frequency is:
(1000 + 1155 + 1250) / (1000 + 770 + 625) = 1.4217.
1 - Z = 1.4217 / 1.5760. $\Rightarrow$ Z = **9.8%**.

2.37. (a) Let Group B be those drivers with at least one claim last year.
Let x be the average number of claims per insured for Group B.
For a Poisson with mean μ, $f(0) = e^{-\mu}$.
Therefore, $\mu = (0)(e^{-\mu}) + (x)(1 - e^{-\mu})$. $\Rightarrow x = \mu / (1 - e^{-\mu})$.
Thus the relativity of Group B compared to average is: $x/m = 1/ (1 - e^{-\mu})$.
Then we have that the credibility weighted modification factor for risks in Group B is:
$M = Z / (1 - e^{-\mu}) + (1 - Z)(1)$.

$$\Rightarrow Z = \frac{M - 1}{1 / (1 - e^{-\mu}) - 1} = \mathbf{(M - 1)\ (e^{\mu} - 1)}.$$

(b) Credibility for an individual risk is lowered when the class plan is highly refined, because it is more difficult to identify differences in the loss potential for a particular risk from the average risk in the class. In other words, the Variance of Hypothetical Means within a class is less, so that the Buhlmann Credibility Parameter K is larger, and Z is less.
Put another way, if the class plan is more refined, it does a better job of estimating the expected pure premium, and there is less need to rely upon the experience of an individual insured. The relative value of the information in the data from the individual has declined, and Z is less.
If a class plan were to get so refined that each class was homogeneous, in other words if every insured in the class had the same expected pure premium, there would be no need for merit rating (experience rating) and Z for the experience of the individual would be zero.
Comment: Part (a) tests the alternate technique at page 160 of Bailey-Simon based on looking at the relativity for those with at least one claim. For Class 1 in Bailey-Simon, from their Table 1, μ = 288,019 / 3,325,714 = 0.0866, and M = 1.476. Thus, $Z = (1.476 - 1) (e^{0.0866} - 1) = 0.043$.

2.38. (a) The overall claim frequency on a premium basis is: 110,000/54 = 2037.
Claim frequency on a premium basis for 3 or more claim free: 40,000 / 25 = 1600.
1 - Z = 1600 / 2037. ⇒ Z = **21.5%**.
Claim frequency on a premium basis for 2 or more claim free:
(40,000 + 15,000) / (25 + 8) = 1667.
1 - Z = 1667 / 2037. ⇒ Z = **18.2%**.
Claim frequency on a premium basis for 1 or more claim free:
(40,000 + 15,000 + 25,000) / (25 + 8 + 13) = 1739.
1 - Z = 1739 / 2037. ⇒ Z = **14.6%**.
(b) Using car years is not preferable to using earned premiums. Using earned premiums adjusts for the mix of business by territory; it adjusts for the effect of territories with higher than average expected frequencies by dividing by their higher than average premiums. Using cars years would not adjust for this, and thus we would be double counting the effect of territories via territorial rating factors and the experience of the insured via Merit Rating.
Specifically, a higher expected frequency territory would have a lower than average proportion of Group A and a higher than average proportion of Group D. Thus Group A would have a higher than average proportion of risks from lower rated territories. Thus the future experience of Group A would look better compared to the overall average than it otherwise would. We want to use Z in Merit Rating to adjust the estimated future frequency for an insured compared to its territory and class, not compared to the overall average. Thus, here what we want to do is compare the experience of group A to the average frequency in its mix of territories rather than the overall average. This is approximated by using earned premium in the denominator which adjusts for the expected frequency for the mix of territories in each Group.

2.39. (a) The overall claim frequency on a premium basis is: 225,000/137 = 1642.
Claim frequency on a premium basis for 3 or more claim free: 120,000/100 = 1200.
1 - Z = 1200 / 1642. ⇒ Z = **26.9%**.
Claim frequency on a premium basis for 2 or more claim free:
(120,000 + 25,000) / (100 + 10) = 1318.
1 - Z = 1318 / 1642. ⇒ Z = **19.7%**.
Claim frequency on a premium basis for 1 or more claim free:
(120,000 + 25,000 + 44,000) / (100 + 10 + 17) = 1488.
1 - Z = 1488 / 1642. ⇒ Z = **9.4%**.
(b) 1. The credibilities are smaller for XYZ than ABC. This is probably due to a more refined classification system for XYZ than ABC. This could also be due to a much lower mean frequency for ABC, so that one year from ABC contains less useful information than from XYZ. For XYZ the ratio of the three year credibility to the one year credibility is: 14/6 = 2.33, while for ABC it is 26.9/9.4 = 2.86. Since for XYZ the credibilities are further from increasing linearly, there are probably more rapidly shifting risk parameters over time for XYZ than for ABC. This could instead or also be due to XYZ having more risks entering and leaving classes than for ABC.
(c) If one portfolio has a more refined class plan then the credibility assigned to the experience of a single car would be lower relative to the other portfolio which has a less refined plan.
Comment: The two-year credibility of 19.7% is more than twice the one-year credibility of 9.4%. Rather, we would expect the two-year credibility to be less than twice the one-year credibility.

2.40. (a) The overall pure premiums is: 4,000,000/4000 = 1000.
Pure premium for 1 or more years claims-free:
(1,000,000 + 500,000) / (2500 + 500) = 500.
1 - Z = 500 / 1000. $\Rightarrow$ Z = **50.0%**.
(b) Pure premium for 2 or more years claims-free: 1,000,000/2500 = 400.
The overall pure premium is $1000.
Thus the premium for an exposure that is accident-free for 2 or more years is:
($1250)(400/1000) = **$500**.
Alternately, for a risk that is accident-free for 2 or more years:
1 - Z = 400/1000. $\Rightarrow$ Z = 60.0%.
There are no losses during the two years, so that the mod is: (0)(0.6) + (1)(1 - 0.6) = 0.4.
Premium is: (0.4)($1250) = **$500**.
Comment: Bailey-Simon work with frequencies rather than pure premiums. All other things being equal, the credibility of one year for estimating future pure premiums is usually less than that for estimating frequencies. (It is easier to estimate future frequencies than pure premiums.)
The given earned exposures and incurred losses are for a subsequent year.
Given the assumptions, we are fine using exposures as the denominator of frequency.
We could instead use in the denominator 1250 times the exposures, making no difference in the estimated credibilities.
In part (b) we have assumed either that there are no fixed expenses, or that there is a separate expense fee which is not adjusted for Merit Rating and which we ignore.
In part (b) we might have an insured who is claim free in 2006 and 2007 and we are using these two years of experience to predict 2008; we give a claim-free discount of 60%.
The CAS sample solutions to part (b) make no sense to me.

2.41. (a) The overall (exposure based) frequency is m = 100,000 / (980,000 + M).
Assuming Poisson frequency, the mean number of claims for those in Group B is: $m / (1 - e^{-m})$.
The relative frequency for Group B is: $1 / (1 - e^{-m})$.
The premium based frequency for Group B is: 18,000/45,000,000.
The overall premium based frequency is: 100,000/670,000,000.
Therefore, the modification for Group B is: (18/45) / (100/670) = 2.68.
Thus we must have: $2.68 = (0.167)\{1 / (1 - e^{-m})\} + (1 - 0.167)(1). \Rightarrow$
$1 - e^{-m} = 0.0904168. \Rightarrow 0.947688 = m = 100,000 / (980,000 + M). \Rightarrow$ M = **75,198**.
(b) Premium based frequency for those who are claim-free for two or more years:
(50,000 + 20,000) / (400 + 150) = 127.27.
Premium based frequency overall: 100,000/670 = 149.25.
1 - Z = 127.27/149.25. $\Rightarrow$ Z = **14.7%**.
Comment: Part (a) tests the alternate technique at page 160 of Bailey-Simon based on looking at the relativity for those with at least one claim.
In part (a), I found it confusing that they used the letter M for the missing number of exposures.

2.42. Premium based frequency for those who are claims-free for one or more years:
(45,000 + 15,000 + 29,300) / (60 + 15 + 20) = 940.
Premium based frequency overall: 108,000/100 = 1080.
1 - Z = 940/1080. $\Rightarrow$ Z = **13.0%**.

2.43. For State X we have the total claim frequency is: 855 / 1150 = 0.7435.

Number of Accident-Free Years	Relative Claim Frequencies to Total
3 or more	(240/500) / 0.7435 = 0.6456
2 or more	(365/650) / 0.7435 = 0.7553
1 or more	(555/850) / 0.7435 = 0.8782

In state X the ratio of three year to one year credibility is: (1 - 0.6456) / (1 - 0.8782) = 2.91.
In state Y the ratio of three year to one year credibility is: (1 - 0.70) / (1 - 0.84) = 1.875.
State Y credibilities go up much less than linearly, and thus state Y is more affected by shifting risk parameters.
State Y is more variation (over time) in an individual insured's probability of an accident.

2.44. It would be better to use premiums, provided the high rated territories have higher frequency and provided the territory relativities are correct.
The average rates are:
Territory 1: (15 + 125 + 230) / (15 + 125 + 230) = $1000.
Territory 2: (25 + 310 + 550) / (25 + 300 + 535) = $1029.
Territory 3: (10 + 80 + 160) / (10 + 100 + 170) = $893.
(There is not a large spread of rates, but Territory 3 is the lowest rated.)
The average frequencies are:
Territory 1: (5 + 41 + 76) / (15 + 125 + 230) = 0.330.
Territory 2: (7 + 84 + 147) / (25 + 300 + 535) = 0.277.
Territory 3: (4 + 35 + 60) / (10 + 100 + 170) = 0.354.
While Territory 3 is the lowest rated, it has the highest frequency.
So using premiums as the denominator of frequency would not adjust for a maldistribution.
Thus, I would use **car-years** as the denominator of frequency in determining the credibility of a single private car using the general type of technique in Bailey and Simon.
Comment: When I read this question, it was very unclear to me what they were trying to get at. If this happens to you on your exam, skip the question and come back later if you have time. It would have helped me if they had said "choose an appropriate denominator to divide into the number of claims to use in determining the credibility of a single private car using the general type of technique in Bailey and Simon." In my opinion, this was far from one of their better questions.
In Bailey-Simon, I would consider years as the exposure base for credibility; the more years of data for a car, the more credibility.
Bailey and Simon use premium as the denominator to eliminate maldistribution due to high frequency territories having a high territorial relativity and a lower number of accident free risks. The purpose is to adjust for the mix of territories by subgroup (0 years claims-free, 1 year claims-free, 2 years claims free, etc.); we are concerned about the different relative claim frequencies by territory.
Hazam says that using premium as the denominator works only when:
High frequency territories are also high premium territories, and territorial relativities are proper. (However, he does not say that when this is not the case we should use car-years as the denominator.)
We can check whether the territory current relativities are correct. The current loss ratios are:
Territory 1: (9 + 75 + 138) / (15 + 125 + 230) = 60%.
Territory 2: (16 + 187 + 328) / (25 + 310 + 550) = 60%.
Territory 3: (7 + 43 + 100) / (10 + 80 + 160) = 60%.
Thus the current territory relativities appear to be correct.
The average rates by years since last accident are for Territory 1 all $1000.
In Territory 2, the average rate for those with zero years claims-free is $1000, while for 2 years claims-free it is $1028.
This is not the pattern we expect. We would assume that those who are claims-free for two years are on average in lower rated classes than those who have zero years claims free.

2.45. (a) I will assume we are analyzing data separately for each class, as per Bailey and Simon.
(Here there do not seem to be any classes; "No other rating variables are applicable.")
Also I will assume as per Bailey and Simon that the premiums have been put on the current "Class B" rate level, in other words on the Merit Rating level of those with no years claims free; we need to remove the current impact of the Merit Rating Plan.
We assume that the current territory relativities are correct, and that differences in territory relativities are due to differences in expected frequency (per caryear) rather than expected severity.
According to the review by Hazam: "a premium base eliminates maldistribution only if (1) high frequency territories are also high premium territories and (2) if territorial differentials are proper."
(b) Overall, frequency with respect to premium ($ million) is: 4075/600 = 6.792.
For two or more years claims free, frequency with respect to premium ($ million) is:
(1200 + 625) / (250 + 100) = 5.214.
Thus for two or more years claims free, Z = 1 - 5.214/6.792 = 23.2%.
For one or more years claims free, frequency with respect to premium ($ million) is:
(1200 + 625 + 750) / (250 + 100 + 100) = 5.722.
Thus for one or more years claims free, Z = 1 - 5.722/6.792 = 15.8%.
The ratio of these two credibilities is: 23.2% / 15.8% = **1.47**.
(c) Assume that the base rate is to be applied to an exposure which has zero years claim free.
For exposures who are zero years claims free, frequency with respect to premium ($ million) is:
1500/150 = 10.
Thus we should charge an exposure that is accident free for two or more years:
(1000)(5.214/10) = **$521**.
Alternately, compared to average, we should give an exposure that is accident free for two or more years a discount of 23.2%.
Compared to average those with zero years claims free they should get a surcharge of:
10/6.792 - 1 = 47.2%.
Thus we should charge an exposure that is accident free for two or more years:
(1000/1.472)(1 - 23.2%) = $522.
Alternately, assuming that the base rate is the average rate, then we should charge an exposure that is accident free for two or more years: (1000) (1 - 23.2%) = **$768**.
Comment: In part (c), the examiners seem unaware that the base rate is for Merit Rating Class B, those who are zero years claims free. Rather they seem to assume that the base rate is the average rate, which is not how it is done in the real world. Bailey and Simon put all of their premiums on a Class B level; in other words they treat Merit Rating Class B as the base class. In any case, the calculated mods are with respect to average.
The credibilities determined are unrealistically big.
The given data is very unusual and unrealistic, including the average premiums:

Number of Accident-Free Years	Earned Car Years	Earned Premium ($000)	Average Premiums
3 or More	250,000	250,000	$1000
2	300,000	100,000	$333
1	25,000	100,000	$4000
0	12,000	150,000	$12,500
Total	587,000	600,000	$1022

2.46. (a) Assume as in Bailey-Simon that this is data for one class.
Using car years may create maldistribution because some territories have higher frequency. Using car years as the denominator of frequency, the credibility calculation would account for both "within territory differences" and "between territory differences". However, usually territory relativities already account for the between territory differences. We want Merit Rating to account for differences between cars not already accounted for by the class/territory relativities. Therefore using car years as the exposure base would double count territory differences, which usually would result in the credibility estimated for Merit Rating being too large.
However, since in this case state law prohibits reflecting territory differences in rating, using earned premium as the exposure base (dividing number of claims by earned premium) should work just as well as using earned exposures. Here using car years is appropriate due to the lack of territory differences in rating. Due to the rates not reflecting frequency differences between territory, the appropriate credibilities for Merit Rating are larger than they otherwise would be. Alternately, premium may still be a stronger exposure base if nonterritorial factors are captured correctly, thereby reducing the maldistribution that exists using car years.
(b) Overall frequency is: 44/800 = 0.055.
Frequency of those with one or more years accident-free is:
(20 + 15) / (500 + 200) = 0.050.
$Z = 1 - 0.05/0.055 =$ **9.09%**.
(c) Frequency of those with no years accident-free is: 9/100 = 9%.
$9\%/5.5\% = M = Z / (1 - e^{-0.055}) + (1 - Z)(1). \Rightarrow 17.69Z = 0.6364. \Rightarrow Z =$ **3.60%**.
Comment: For part (c) we are using the alternate method discussed at page 160 in Bailey-Simon.
It uses the Poisson assumption. Let λ = the mean claim frequency (per exposure) for the class.
M = relative premium based frequency for risks with one or more claims in the past year.
Then, $M = Z / (1 - e^{-\lambda}) + (1 - Z)(1). \Rightarrow Z = \dfrac{M - 1}{1 / (1 - e^{-\lambda}) - 1} = (M - 1)(e^{\lambda} - 1)$.

The estimated credibilities in parts (b) and (c) are both for one year of data, and we would expect them to be more similar than they are here.
Bailey and Simon “have chosen to calculate Relative Claim Frequency on the basis of premium rather than car years. This avoids the maldistribution created by having higher claim frequency territories produce more X, Y, and B risks and also produce higher territorial premiums.”

2.47. Here is the best solution I could come up with, expanding on the ideas in Appendix II of Bailey-Simon.

Let us assume that each insured is Poisson with mean λ, with the lambdas varying across the portfolio. Assume that over several years each insured has a constant expected frequency λ.

Then the probability of being claim free for zero years is: $1 - e^{-\lambda}$.

The probability of being claim free for at least one year is: $e^{-\lambda}$.

The probability of being claim free for at least two years is: $e^{-2\lambda}$.

Thus the probability of being claim free for exactly one year is: $e^{-\lambda} - e^{-2\lambda}$.

The probability of being claim free for at least two years is: $e^{-3\lambda}$.

Thus the probability of being claim free for exactly two years is: $e^{-2\lambda} - e^{-3\lambda}$.

Similarly, the probability of being claim free for exactly three years is: $e^{-3\lambda} - e^{-4\lambda}$.

Then for a subset of insureds with the same lambda:

$$\frac{\text{expected number claim free for exactly one year}}{\text{expected number not claim free}} = (e^{-\lambda} - e^{-2\lambda}) / (1 - e^{-\lambda}) = e^{-\lambda}.$$

$$\frac{\text{expected number claim free for exactly two years}}{\text{expected number claim free for exactly one year}} = (e^{-2\lambda} - e^{-3\lambda}) / (e^{-\lambda} - e^{-2\lambda}) = e^{-\lambda}.$$

$$\frac{\text{expected number claim free for exactly three years}}{\text{expected number claim free for exactly two years}} = (e^{-3\lambda} - e^{-4\lambda}) / (e^{-2\lambda} - e^{-3\lambda}) = e^{-\lambda}.$$

Thus within each of the given rows, if the assumptions are correct, we would expect these observed ratios to be close to equal. (Ignore the issue of how would one know the expected claim frequencies for the different rows of insureds.)

For the first row, the observed ratios are: 47,500/50,000 = 0.95, 45,000/47,500 = 0.947, and 44,000/45,000 = 0.978. The last ratio is dissimilar from the other two.

For the second row, the observed ratios are: 45,000/50,000 = 0.90, 43,000/45,000 = 0.956, and 36,000/43,000 = 0.837. These ratios are not similar to each other!

For the third row, the observed ratios are: 20,500/25,000 = 0.82, 16,500/20,500 = 0.805, and 14,000/16,500 = 0.849. These ratios are dissimilar from each other.

We do not see what we would expect; therefore something is wrong with the assumptions.

One or more of the following are true: individuals risk parameters are shifting over time, the frequency process is not Poisson, or insureds are entering and leaving the data base over the period of time studied.

Here is a sample solution from the CAS Examiner's Report that attempts to apply the ideas from Bailey-Simon (but fails to do so correctly, see my comments below):
Total insureds: 125,000 + 113,000 + 104,5000 + 94,000 = 436,500.
Insureds claims free for at least one year (in fact for those claims free for exactly 1, 2 or 3 years): 113,000 + 104,5000 + 94,000 = 311,500.
Insureds claims free for at least two years (in fact for those claims free for exactly 2 or 3 years): 104,5000 + 94,000 = 198,500.
Insureds claims free for at least three years (in fact for those claims free for exactly 3 years): 94,000.
Total expected claims: (186,500)(0.05) + (174,000)(0.10) + (76,000)(0.20) = 41,925.
Expected claims for those claims free exactly one year:
(47,500)(0.05) + (45,000)(0.10) + (20,500)(0.20) = 10,975.
Expected claims for those claims free exactly two years:
(45,000)(0.05) + (43,000)(0.10) + (16,500)(0.20) = 9,850.
Expected claims for those claims free exactly three years:
(44,000)(0.05) + (36,000)(0.10) + (14,000)(0.20) = 8,600.
Expected claims for insureds claims free for at least one year (in fact for those claims free for exactly 1, 2 or 3 years): 10,975 + 9,850 + 8,600 = 29,425.
Expected claims for insureds claims free for at least two years (in fact for those claims free for exactly 2 or 3 years): 9,850 + 8,600 = 18,450.
Expected claims for insureds claims free for at least three years (in fact for those claims free for exactly 3 years): 8,600.
Then for example, the expected frequency for those claims free for at least three years (in fact for those claims free for exactly 3 years): 8600/94,000 = 0.0915. Then, 0.0915/0.0960 = 0.9525.

n	# Claim free n or more years	Expected Claims	Expected Frequency	Relative Exp. Freq.	"Credibility"
3	94,000	8,600	0.0915	0.9525	0.0475
2	198,500	18,450	0.0929	0.9677	0.0323
1	311,500	29,425	0.0945	0.9835	0.0165
Total	436,500	41,925	0.0960	1	

For example, the "credibility" for three years is: 1 - 0.9525 = 0.0475.
If the variation of an insured's chance for an accident is not changing over time, then $\frac{\text{3 year credibility}}{\text{1 year credibility}}$ will be approximately equal to 3, and $\frac{\text{2 year credibility}}{\text{1 year credibility}}$ will be approximately equal to 2.
$\frac{\text{3 year credibility}}{\text{1 year credibility}} = 0.0475 / 0.0165 = 2.88$. $\frac{\text{2 year credibility}}{\text{1 year credibility}} = 0.0323 / 0.0165 = 1.96$.
The ratios are approximately 3 and 2, and therefore the chance for an accident is stable.

Here is a second sample solution from the CAS Examiner's Report that is a parody of the calculations in Bailey-Simon, demonstrating a lack of understanding of the ideas in Bailey-Simon.
Expected claims at t = 0 (actually for those not claims free):
(50,500)(0.05) + (50,000)(0.10) + (25,000)(0.20) = 12,500.
Expected claims at t = 1 (actually for those claims free exactly one year):
(47,500)(0.05) + (45,000)(0.10) + (20,500)(0.20) = 10,975.
Expected claims at t = 2 (actually for those claims free exactly two years):
(45,000)(0.05) + (43,000)(0.10) + (16,500)(0.20) = 9,850.
Expected claims at t = 3 (actually for those claims free exactly three years):
(44,000)(0.05) + (36,000)(0.10) + (14,000)(0.20) = 8,600.
Then the "frequency at t = 0": 12,000/125,000 = 0.1000.
The "frequency at t = 1": 10,975/113,000 = 0.09712.
The "frequency at t = 2": 9850/104,500 = 0.09426.
The "frequency at t = 3": 8600/94,000 = 0.09149.
The "frequency at t = 1 relative to t = 0": 0.09712/0.1000 = 0.9712.
$\Rightarrow$ One year "credibility": 1 - 0.9712 = 2.88%.
The "frequency at t = 2 relative to t = 0": 0.09426/0.1000 = 0.9426.
$\Rightarrow$ Two year "credibility": 1 - 0.9426 = 5.74%.
The "frequency at t = 3 relative to t = 0": 0.09149/0.1000 = 0.9149.
$\Rightarrow$ Three year "credibility": 1 - 0.9149 = 8.51%.
(Note this is <u>not</u> how Bailey-Simon calculates credibilities. Within a rating class, they compare for example the observed subsequent (premium based) frequency for those who are claims free for 2 years or more, to the overall observed subsequent (premium based) frequency.
Then Bailey-Simon are backing out the credibility for 2 years of data based on an observed credit appropriate for 2 or more years claims free.)
If the variation of an insured's chance for an accident is <u>not</u> changing over time, then $\frac{\text{3 year credibility}}{\text{1 year credibility}}$ will be approximately equal to 3, and $\frac{\text{2 year credibility}}{\text{1 year credibility}}$ will be approximately equal to 2.
$\frac{\text{3 year credibility}}{\text{1 year credibility}}$ = 8.51% / 2.88% = 2.95. $\frac{\text{2 year credibility}}{\text{1 year credibility}}$ = 5.74% / 2.88% = 1.99.
The ratios are approximately 3 and 2, and therefore the chance for an accident is stable.

Here is a third sample solution from the CAS Examiner's Report that is a parody of the calculations in the paper by Mahler, demonstrating a lack of understanding of the ideas in that paper.
Determine the percent of the insureds in each column that are in each of the three rows.

t=0	t=1	t=2	t=3
50/125 = 40%	47.5/113 = 42.03%.	45/104.5 = 43.06%	44/94 = 46.81%
50/125 = 40%	45/113 = 39.82%	43/104.5 = 41.15%	36/94 = 38.30%
25/125 = 20%	20.5/113 = 18.14%	16.5/104.5 = 15.79%	14/94 = 14.89%

Now calculate the correlations between the various columns:

"lag 1"	t=0 vs t=1: 0.9965	t=1 vs t=2: 0.9998	t=2 vs t=3: 0.9806	AVG: 0.9923.
"lag 2"	t=0 vs t=2: 0.9980	t=1 vs t=3: 0.9845		AVG: 0.9913
"lag 3"	t=0 vs t=3: 0.9663			AVG: 0.9663

Since the correlations are decreasing with lag, this indicates that parameters are shifting over time.
Here is a fourth sample solution from the CAS Examiner's Report that is another parody of the calculations in the paper by Mahler, demonstrating a lack of understanding of the ideas in that paper.
Determine the expected claims for each entry in the rows and columns.

t=0	t=1	t=2	t=3
(0.05)(50,000) = 2500	2375	2250	2200
(0.10)(50,000) = 5000	4500	4300	3600
(0.20)(25,000) = 5000	4100	3300	2800

Then compute the correlations between the different columns.
"lag 1" r(0,1) = 0.9842, r(1,2) = 0.9456, r(2,3) = 0.9954. Average = 0.9750.
"lag 2" r(0,2) = 0.8730, r(1,3) = 0.9909. Average = 0.8914.
"lag 3" r(0,3) = 0.8220. Average = 0.8220.
Downward trending average correlation as lag increases. ⇒ Risk parameters are shifting.
Here is a fifth sample solution from the CAS Examiner's Report that is another parody of the calculations in the paper by Mahler, demonstrating a lack of understanding of the ideas in that paper.
Determine the ratios of the number of insureds in adjacent columns in each of the three rows.
For the first row, the observed ratios are: 47,500/50,000 = 0.95, 45,000/47,500 = 0.9474, and 44,000/45,000 = 0.9778.
For the second row, the observed ratios are: 45,000/50,000 = 0.90, 43,000/45,000 = 0.9556, and 36,000/43,000 = 0.8372.
For the third row, the observed ratios are: 20,500/25,000 = 0.82, 16,500/20,500 = 0.8049, and 14,000/16,500 = 0.8485.
Then take the correlations between these sets of ratios:
corr[{0.95, 0.90, 0.82}, {0.9474, 0.9556, 0.8049}] = 0.9049,
corr[{0.9474, 0.9556, 0.8049}, {0.9778, 0.8372, 0.8485}] = 0.3920.
corr[[{0.95, 0.90, 0.82}, {0.9778, 0.8372, 0.8485}]] = 0.748.
Average of correlations for "lag 1": (0.9049 + 0.3920)/2 = 0.6485.
Average of correlations for "lag 2": 0.748.
These correlations are not declining with increase in lags.
Thus there is no evidence that parameters are shifting over time.

Comment: This question does not follow any of the syllabus readings. Although this question bears a similarity to ideas in Bailey-Simon and to the shifting risk parameters paper by Mahler, the information needed to properly apply the ideas in those syllabus readings is not provided.
In my opinion, this is a terrible exam question, which demonstrates the lack of understanding of this material by its writer. I suspect those with a better understanding of this material did worse in attempting to somehow answer this exam question. For study purposes, I think this question has negative educational value. Of course, you might want to know how to mechanically reproduce one of the sample solutions in case this exact same form of question is repeated. Therefore, I have given the sample answers from the CAS Examiner's Report.
My commentary on the question and sample solutions follows.
How would one know the expected claim frequencies for the different subsets of insureds?
If an insured's individual chance of an accident changes over time, what could it mean to be in one of the given rows? If an insured's individual chance of an accident changes over time, the insureds in a given row can not have the same expected claim frequency over several years.
Although we are not shown the information, aren't there insureds who are claims-free for exactly four years, exactly five years, etc.? Thus, we do not know for example how many insureds were claims-free for 3 or more years.
In the first sample solution I showed, expected claims are calculated "at time t", by multiplying the number of insureds by the expected claim frequency. What does this mean? Yes if we have 50,000 insureds with an expected claim frequency of 0.05 then we would expect 2500 claims. However, these 50,000 insureds in the first column were not claim free, so they each had at least one claim. Perhaps this means we would expect 2500 claims the following year from these insureds; however; this would ignore the fact that those in a given (heterogeneous) group who are not claim free have higher than average expected future claim frequency compared to the group (the idea behind using credibility) and also that insureds claim propensity may change over time.
Rather as per Bailey-Simon, what we want to know is for a class of insureds the subsequent actual experience of those who were not claim-free, those who were claim free for at least one year, those who were claim free for at least two years, etc. Here we not given this vital information. The solution compares "expected" frequencies rather than as it should observed actual subsequent frequencies.
There is a comparison of the data for those claims free for exactly 1 to 3 years, those claims free for exactly 2 or 3 years, and those who are claims free for exactly 3. The correct comparisons would be between those claims free for at least one year, those who are claims free for at least 2 years, and those who were claims free for at least 3 years; we do not have that information.
Having performed a bunch of arithmetic, "credibilities" supposedly for one, two and three years of data are determined, which are not in fact credibilities in any meaningful sense. However, the conclusion drawn from these "credibilities" is correct. If risk parameters were shifting significantly over time, then the credibilities for one, two, and three years should increase significantly less than linearly.
In the paper by Mahler, the correlations are between different years of actual experience for a set of individual risks. After doing some arithmetic, the sample solutions compute correlations. However, these are not the type of correlations one would use to answer the question of whether we have shifting risk parameters.

The third sample solution works with correlations of the percent of insureds who are claims-free for exactly t years. There is no reason to assume that if risk parameters are constant, that this type of correlation will be independent of the differences in t. If these types of correlations decline as the difference in t (which is not the lag between different years of data) increases, this does not demonstrate that parameters are shifting.
The fourth sample solution and fifth sample solutions are also invalid.
Partial credit was also given for a Chi-Square approach, which is not shown. The Examiner's Report does not explain how one would know the expected number of insureds claims-free for exactly t years, to compare to the actual number. Nor does the Examiner's Report explain exactly how this has any relation to whether or not risk parameters shift.

2.48. (a) The use of premiums as the exposure base (as Bailey-Simon did) would make sense if the high rated territories are the high frequency territories. However, this is not the case here; territory C with the highest frequency has the lowest average premium.
(Different average severities seem to be responsible for a significant amount of the variation in premiums between territories.)
Thus I will use **earned car years as the exposure base**.
Note that in order to use premium as the exposure base to correct for maldistribution, one would also require that the territory differentials are properly priced; there is no way to determine whether or not that is the case here.

(b) Number of Accident-Free Years	Car Years (000s)	Number of Claims	Frequency	Relative Freq.
3 or More	350	28,500	0.0814	0.866 = 814/940
1 or more	430	36,500	0.0849	0.903 = 849/940
Total	500	47,000	0.0940	1.000

Three year credibility is: 1 - 0.866 = 13.4%.
One year credibility is: 1 - 0.903 = 9.7%.
Three year credibility relative to the one year credibility: 13.4% / 9.7% = **1.38**.
Alternately, one can estimate the credibility for one year of data from the experience of those who were not claim free. The frequency per car year for those who are not claim free is: 10,500 / 70,000 = 0.1500. Relative frequency is: 0.1500/0.0940 = 1.596.
Assume a Poisson frequency with mean equal to the overall mean: $\lambda = 0.0940$.
Then the average frequency for those who are not claim free is: $\lambda / (1 - e^{-\lambda})$.
Thus the relative frequency of those who are not claim free is: $1 / (1 - e^{-\lambda}) = 1 / (1 - e^{-0.094})$.
$\Rightarrow 1.596 = M = Z/ (1 - e^{-0.094}) + (1-Z)(1)$. $\Rightarrow$ credibility for one year of data = Z = 5.9%.

Three year credibility relative to the one year credibility: 13.4% / 5.9% = **2.27**.

Comment: For part (b), see Tables 1 and 3 in Bailey-Simon. In Bailey-Simon, the premiums have been adjusted to remove the effect of any discounts from the (current) Merit Rating Plan.
In part (a), the CAS allowed arguing that "while the frequencies do not appear to be in-line with premiums by territory, that premium may still be a better choice as it addresses some maldistribution and should be still used as the exposure base."
In that case, in part (b), one should have gotten:

Number of Accident-Free Years	Premium ($million)	Number of Claims	Frequency	Relative Freq.
3 or More	590	28,500	48.31	0.720 = 48.31/67.14
1 or more	650	36,500	56.15	0.836 = 56.15/67.14
Total	700	47,000	67,14	1.000

Three year credibility is: 1 - 0.720 = 28.0%.
One year credibility is: 1 - 0.839 = 16.4%.
Three year credibility relative to the one year credibility: 28.0% / 16.4% = 1.71.
The merit rating plan uses the number of years an insured is claims free.
The merit rating plan does not "use multiple rating variables, including territory." Rather the rating plan upon which merit rating is superimposed, uses multiple rating variables, including territory. These other rating variables should be controlled for. This is why Bailey and Simon apply this technique to data from each class separately.
Part (b) is unclear; it should have said "three year credibility relative to the one year credibility."

Section 3, Generalized Linear Models, Goldburd, Khare, and Tevet[1]

Generalized Linear Models are widely used by actuaries in ratemaking, loss reserving, etc.

GLMs can be thought of as a generalization of multiple linear regressions.
However, **the distribution of random errors need not be Normal**.
Common distributions for the errors are:
Normal, Poisson, Gamma, Binomial, Negative Binomial, Inverse Gaussian, and Tweedie.

Also **there is a link function that connects the linear combination of variables and the thing to be modeled.**
Common link functions are: identity, inverse, logarithmic, logit, and inverse square.
In a linear model, the link function is equal to the identity function.
In a multiplicative model, the link function is logarithmic; this is analogous to an Exponential regression.

Generalized Linear Models are fit via maximum likelihood.

Our goal in modeling is to find the right balance where we pick up as much of the systematic effects (called the signal) as possible and as little of the randomness in the data (called the noise).

Based on the syllabus reading, I do not expect you to be asked to fit a model. Rather you should concentrate on how to set up a GLM, choose between different models, and how to interpret computer output.

Therefore, do not get bogged down in the mathematical details of some of the examples I give, which are provided for those who find that concrete examples help them to learn the material.

This CAS Study Note also discusses some things that apply to most modeling and actuarial work, rather than just to GLMs.

---

[1] "Generalized Linear Models for Insurance Rating", by Mark Goldburd, Anand Khare, and Dan Tevet, CAS monograph series number 5, added to the syllabus for 2016.
Only Sections 1 to 8 are on the syllabus.
Section 9 "Variations on the Generalized Linear Models" is not on the syllabus.

Types of Variables:

Variables can be continuous: size of loss, height, weight, Body Mass Index (BMI), etc.

Variables can be discrete: number of children, number of claims in the last three years, etc.

Variables can be categorical; there are a discrete number of categories.
The different possible values that a categorical variable can take on are called its levels.

In the case of nominal variables, the categories do not have a natural order.
For example, type of vehicle: sedan, SUV, truck, van.

Sometimes however, the categories have a natural order; such variables are called ordinal.
For example injuries may be categorized as: minor, serious, catastrophic, and fatal.
This also occurs when a continuous variable is grouped into categories.

Additive and Multiplicative Models:

When one uses the identity function, the model is additive:
$\mu = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$.
This is analogous to a linear regression.

For example, $\mu = 100 + 5x_1 - 3x_2$.
Each increase of 1 in x_1 results in an increase of 5 in m.
Each increase of 1 in x_2 results in an decrease of 3 in m.

When one uses the log link function, the model is multiplicative:
$\ln[\mu] = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p. \Leftrightarrow m = \exp[\beta_0 + \beta_1 x_1 + ... + \beta_p x_p]$.
This is analogous to an exponential regression.

For example, $\mu = \exp[5 + 0.2x_1 - 0.1x_2]$.
Each increase of 1 in x_1 results in μ being multiplied by $e^{0.2} = 1.221$.
Each increase of 1 in x_2 results in μ being multiplied by $e^{-0.1} = 0.905$.

Advantages of Multiplicative Rating Structures:[2]

1. **A multiplicative plan guarantees positive premium**.
Having additive terms in a model can result in negative premiums, which doesn't make sense; you may have to implement clunky patches like minimum premium rules.

2. **A multiplicative model has more intuitive appeal**.

"It doesn't make much sense to say that having a violation should increase your auto premium by $500, regardless of whether your base premium is $1,000 or $10,000. Rather it makes more sense to say that the surcharge for having a violation is 10%."[3] [4]

"For these and other reasons, **log link models, which produce multiplicative structures, are usually the most natural model for insurance risk**."

Nevertheless, sometimes a multiplicative model (model using a log link function) does not do a good job of modeling the data, while a different link function does a better job. This is an empirical issue. Most factors in insurance rating algorithms are multiplicative, however it is not uncommon to also have additive elements as well.[5]

Even if one uses a log link function, when interaction terms are included in a model, the structure of the model will no longer have all of the nice features of a multiplicative model.
For example: $\mu = \exp[5 + 0.2x_1 - 0.1x_2 + 0.03x_1x_2]$.
Now the effect of a change in x_1 depends on the value of x_2, while the effect of a change in x_2 depends on the value of x_1. For example, in private passenger auto insurance, the effect on expected pure premiums of gender varies by age.

Also keep in mind that for a binary or binomial target variable, for example whether or not a policy is renewed, a logit link function is commonly used as will be discussed.

---

[2] See page 5 of Goldburd, Khare, and Tevet.
[3] This is an empirical question. For example, a more complicated surcharge such as $100 plus 10% of base premium might be a better prediction of the extra future expected costs.
[4] It would be extremely unusual to pay $10,000 or more as a base premium for private passenger automobile. Perhaps they are referring to commercial automobile. In any case, the $10,000 is just for illustrative purposes.
[5] Chapter 2 of "Basic Ratemaking" by Werner and Modlin has some examples.

Other Uses of GLMs:[6]

While GLMs are commonly used for classification ratemaking, the benefits of GLMs are not restricted to the application of pricing.
The following are a few of the other applications for which insurance companies are using GLMs:

- Practitioners are using GLMs to reduce a variety of risk variables into one score. This has obvious application in regards to creating underwriting tiers, credit scores, fire protection scores, vehicle symbols, etc.

- Many companies have begun to perform elasticity modeling. By building elasticity models for new and renewal business, companies can predict the impact of various actions on market share. A few companies are already linking the profitability and elasticity models to find the optimal pricing decision.

- Claims handlers are starting to see the advantages of GLMs and are using them to help set more accurate reserves and to provide early identification of claims that may be fraudulent or are most likely to end up in a lawsuit.

- Competitive analysis units are using GLMs to reverse-engineer competitors' rates given a large sample of rating quotes.

---

[6] Quoted from “GLM Basic Modeling: Avoiding Common Pitfalls,” by Geoff Werner and Serhat Guven, CAS Forum Winter 2007, not on the syllabus.

Common Link Functions:

$g(\mu) = \sum \beta_i x_i. \Leftrightarrow \mu = g^{-1}(\sum \beta_i x_i).$
The x_i are the predictor or explanatory variables.
The β_i are the coefficients, which are to be fit.
$\beta x = \sum \beta_i x_i$, is the linear predictor.
g is the link function, whose form needs to be specified.

The link function must satisfy the condition that it be differentiable and monotonic (either strictly increasing or strictly decreasing). Common link functions to use include:

Identity	$g(\mu) = \mu$	$g^{-1}(y) = y$	$\mu = \beta x$
Log	$g(\mu) = \ln(\mu)$	$g^{-1}(y) = e^y$	$\mu = e^{\beta x}$
Logit	$g(\mu) = \ln[\mu/(1 - \mu)]$	$g^{-1}(y) = \frac{e^y}{e^y + 1}$	$\mu = \frac{e^{\beta x}}{e^{\beta x} + 1}$
Reciprocal	$g(\mu) = 1/\mu$	$g^{-1}(y) = 1/y$	$\mu = 1 / (\beta x)$

With more than one variable, the use of the log link function results in a familiar multiplicative model for classification relativities.
One can also use other powers as a link function, such as $g(\mu) = 1/\mu^2$ or $g(\mu) = \sqrt{\mu}$.

Let p be the probability of policy renewal. Then $0 < p < 1$.
Thus, $0 < p / (1 - p) < \infty$.
Applying the logit link function, $-\infty < \ln[p / (1 - p)] < \infty$.
So we have converted the domain from 0 to 1 to a range of minus infinity to infinity.

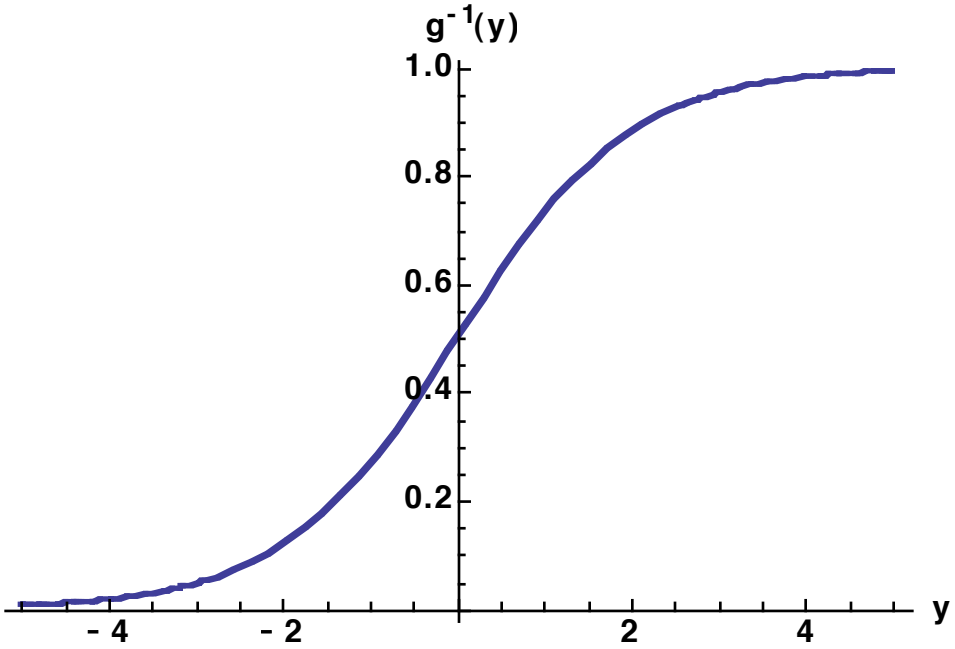
The inverse of the logit link function, $\frac{e^y}{e^y + 1}$, converts the interval from minus infinity to infinity to the interval from zero to one, which would be appropriate for probabilities.[7]

Exercise: $m = \frac{e^{\beta x}}{e^{\beta x} + 1}$. Determine μ for $\beta x = -2$, $\beta x = 0$, and $\beta x = 2$.

[Solution: $e^{-2} / (e^{-2} + 1) = 0.119$. $e^0 / (e^0 + 1) = 0.5$. $e^2 / (e^2 + 1) = 0.881$.
Comment: These all make sense as probabilities.]

[7] Note that $F(x) = e^x / (e^x + 1)$, $-\infty < x < \infty$ is the logistic function.

Here is a graph of the inverse of the logit link function:



It is common to pick the form of the variable X, to be a member of an exponential family.
In that case, there are corresponding "canonical link functions".[8]

Distribution Form	*Canonical Link Function*
Normal[9]	*Identity*
Poisson[10]	*Log: ln(μ)*
Gamma[11]	*Reciprocal: 1/μ*
Binomial[12]	*Logit: ln[μ/(1 -μ)]*
Inverse Gaussian	$1/\mu^2$

Using the canonical link function makes the estimate from the GLM unbiased.

---

[8] While these choices result in some nice mathematica properties, they are not required.
[9] For example, ordinary linear regression.
[10] Could be used to model claim frequencies or claim counts.
[11] Could be used to model claim severities. In that case, one could use the log link function, ln(μ).
[12] Could be used to model probability of policy renewal.
The use of the logit link function with the Binomial or special case Bernoulli is the idea behind logistic regression.

Structure of Generalized Linear Models:

One can state the assumptions of a Generalized Linear Model as:

1. **Random component: Each component of Y, the target variable is independent and is from one of the exponential family of distributions.**[13]

2. **Systematic component:**
   The p explanatory variables are combined to give the linear predictor X b**.**

3. **Link function: The relationship between the random and systematic components is specified via a link function, g, that is differentiable and monotonic such that:**
   $\mathbf{E[Y] = \mu = g^{-1}(X \beta). \Leftrightarrow X \beta = g(\mu).}$

The target variable, also called the dependent variable,Y, is the thing being modeled; it may be: frequency, severity, pure premiums, loss ratios, or something like the probability of policy renewal.

The predictor variables, also called response variables or independent variables, x's, the things being used as inputs to the model, can be things like: age, gender, amount of insurance, etc.

The linear predictor has an intercept β_0 plus p slopes: $\eta = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$.

$\eta = g(\mu)$.

Several different models may be fit to the same data, with one or more of the above features differing. Then the models would be compared using the output diagnostics, in order to determine the best model to use for the purpose.[14]

---

[13] Y is a vector; "each component" of Y refers to the elements of that vector.

[14] Similar diagnostics are available as for a multiple linear regression.

A One Dimensional Example of Generalized Linear Models:[15]

Let us assume a set of three observations: (1, 1), (2, 2), (3, 9).

The predictor variable x takes on the values 1, 2, and 3 for the observations.[16]
The target variable Y takes on the values 1, 2 and 9 for the observations.

In a generalized linear model, Y will have some distributional form. The mean of the distribution will vary with x. However, any other parameters will be constant.

For now let us assume the identity link function, $g(\mu) = \mu$, so that $\mu = \sum \beta_i x_i = \beta_0 + \beta_1 x$.[17]
Thus for now we are fitting a straight line. In general, the identity link function leads to a linear model.

Assume that Y is Poisson, with mean μ.[18]
$\mu = \beta_0 + \beta_1 x$.

For the Poisson Distribution as per Loss Models, $f(y) = e^{-\lambda} \lambda^y / y!$.
$\ln f(y) = -\lambda + y \ln(\lambda) - \ln(y!) = -\mu + y \ln(\mu) - \ln(y!)$.

The loglikelihood is the sum of the contributions from the three observations:
$-(\beta_0 + \beta_1) - (\beta_0 + 2\beta_1) - (\beta_0 + 3\beta_1) + \ln(\beta_0 + \beta_1) + 2\ln(\beta_0 + 2\beta_1) + 9\ln(\beta_0 + 3\beta_1)$
$- \ln(1) - \ln(2) - \ln(9!)$.

To maximize the loglikelihood, we set its partial derivatives equal to zero.
Setting the partial derivative with respect to β_0 equal to zero:
$0 = -3 + 1/(\beta_0 + \beta_1) + 2/(\beta_0 + 2\beta_1) + 9/(\beta_0 + 3\beta_1)$.
Setting the partial derivative with respect to β_1 equal to zero:
$0 = -6 + 1/(\beta_0 + \beta_1) + 4/(\beta_0 + 2\beta_1) + 27/(\beta_0 + 3\beta_1)$.

Solving these two equations in two unknowns: $\beta_0 = -12/5 = -2.4$ and $\beta_1 = 16/5 = 3.2$.[19]
$\mu = -2.4 + 3.2x$. For $x = 1$, $\mu = 0.8$. For $x = 2$, $\mu = 4.0$. For $x = 3$, $\mu = 7.2$.[20]

---

[15] I do not expect you to have to go into this level of detail on your exam.
See page 15 of "A Practitioners Guide to Generalized Linear Models," by Duncan Anderson, Sholom Feldblum, Claudine Modlin, Dora Schirmacher, Ernesto Schirmacher and Neeza Thandi, in the 2004 CAS Discussion Paper Program, not on the syllabus of this exam.

[16] It is not clear in this example whether x can take on values other than 1, 2 and 3. These may be the only possible values, or they might be the three values for which we happen to have had an observation. In practical applications, when x is discrete, we would expect to have many observations for each value of x.

[17] I have treated x_0 as the constant 1 and x_1 as the predictor variable x.

[18] In the case of a Poisson, there are no additional parameters beyond the mean.

[19] I used a computer to solve these two equations. One can confirm that these values satisfy these equations.

[20] This differs from what would be obtained if one assumed Y was Normal rather than Poisson.

This model should be interpreted as follows. For a given value of x, Y is Poisson Distributed with mean = -2.4 + 3.2x. For example, for x = 3, the mean = 7.2. However, due to random fluctuation, for x = 3 we will observe values of Y varying around the expected value of 7.2.[21]
If we make a very large number of observations of individuals with x = 3, then we expect to observe a Poisson Distribution of outcomes with mean 7.2.

As discussed, another important decision is the choice of the link function.
In this example, let us maintain the assumption of a Poisson Distribution, but instead of the identity link function let us use the log link function.

$\ln(\mu) = \sum\beta_i x_i = \beta_0 + \beta_1 x. \Rightarrow \mu = \exp[\sum\beta_i x_i] = \exp[\beta_0 + \beta_1 x].$

$f(y) = e^{-\lambda} \lambda^y / y!.$
$\ln f(y) = -\lambda + y\ln(\lambda) - \ln(y!) = -\mu + y\ln(\mu) - \ln(y!) = -\exp[\beta_0 + \beta_1 x] + y(\beta_0 + \beta_1 x) - \ln(y!).$

The loglikelihood is the sum of the contributions from the three observations:
$-\exp[\beta_0 + \beta_1] - \exp[\beta_0 + 2\beta_1] - \exp[\beta_0 + 3\beta_1] + \beta_0 + \beta_1 + 2(\beta_0 + 2\beta_1) + 9(\beta_0 + 3\beta_1)$
$- \ln(1) - \ln(2) - \ln(9!).$

To maximize the loglikelihood, we set its partial derivatives equal to zero.
Setting the partial derivative with respect to β_0 equal to zero:

$0 = -\exp[\beta_0 + \beta_1] - \exp[\beta_0 + 2\beta_1] - \exp[\beta_0 + 3\beta_1] + 12.$

Setting the partial derivative with respect to β_1 equal to zero:

$0 = -\exp[\beta_0 + \beta_1] - 2\exp[\beta_0 + 2\beta_1] - 3\exp[\beta_0 + 3\beta_1] + 32.$

Thus we have two equations in two unknowns:
$\exp[\beta_0 + \beta_1]\{1 + \exp[\beta_1] + \exp[2\beta_1]\} = 12.$

$\exp[\beta_0 + \beta_1]\{1 + 2\exp[\beta_1] + 3\exp[2\beta_1]\} = 32.$

Dividing the second equation by the first equation:

$$\frac{1 + 2\exp[\beta_1] + 3\exp[2\beta_1]}{1 + \exp[\beta_1] + \exp[2\beta_1]} = 8/3.$$

$\Rightarrow \exp[2\beta_1] - 2\exp[\beta_1] - 5 = 0.$

[21] For the Poisson Distribution, the variance is equal to the mean.

Letting $v = \exp[\beta_1]$, this equation is: $v^2 - 2v - 5 = 0$, with positive solution $v = 1 + \sqrt{6} = 3.4495$.

$\exp[\beta_1] = 3.4495. \Rightarrow \beta_1 = 1.238.$

$\Rightarrow \exp[\beta_0] = 12/\{\exp[\beta_1] + \exp[2\beta_1] + \exp[3\beta_1]\} = 12/\{3.4495 + 3.4495^2 + 3.4495^3\} = 0.2128.$

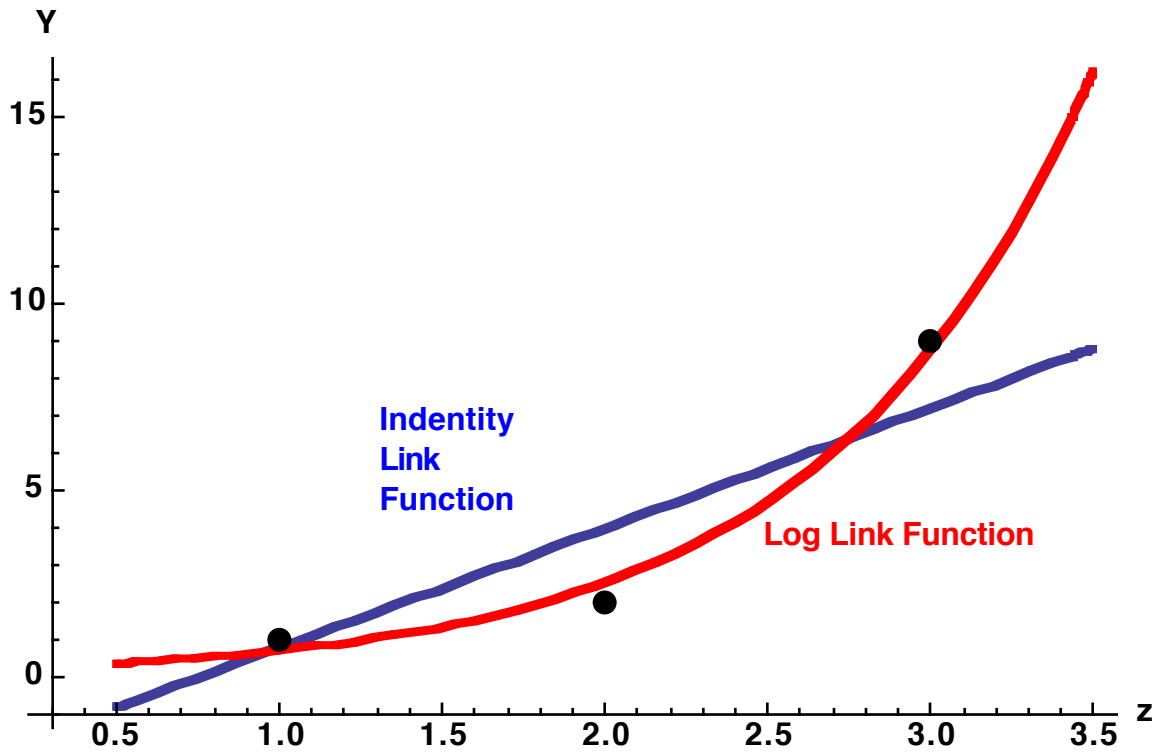
$\Rightarrow \beta_0 = -1.547.$

$\mu = \exp[\beta_0 + \beta_1 x] = \exp[\beta_0] \exp[\beta_1]^x = (0.2128)(3.4495^x).$

For x = 1, $\mu = 0.734$. For x = 2, $\mu = 2.532$. For x = 3, $\mu = 8.735$.
This differs from the result obtained previously when using the identity link function:

x	Observed	Poisson, Identity Link	Poisson, Log Link Function
1	1	0.8	0.734
2	2	4.0	2.532
3	9	7.2	8.735

Here is the same information in the form of a graph, with the data shown as dots:



In general, the choice of a link function makes a difference.
Using the log link function we got an exponential model rather than a linear model. With more explanatory variables, the log link function gives a multiplicative rather than an additive model.

Exponential Families:

Linear Exponential Families include:
Bernoulli, Binomial (m fixed), Poisson, Geometric, Negative Binomial (r fixed), Exponential, Gamma (α fixed), Normal (σ fixed), Inverse Gaussian (θ fixed), and the Tweedie Distribution.

Confusingly, when working on GLMs, "Exponential Family" means "Linear Exponential Family."[22] This is how the syllabus reading refers to them, and thus from now on I will do the same.

Exponential Families have two parameters, μ the mean, and ϕ the dispersion parameter.
The dispersion parameter is related to the variance. In a GLM ϕ is fixed across the observations and is treated as a nuisance parameter, in the same way that σ is treated in multiple regression.

It turns out that the relationship between the mean and variance uniquely identifies which linear exponential family we have.
Var[Y] = ϕ V(μ), where the form of V(μ) depends on which exponential family we have.

If the variance does not depend on the mean, then we have a Normal Distribution.
If the variance is proportional to the square of the mean, then we have a Gamma Distribution.
If the variance is proportional to the cube of the mean, then we have a Inverse Gaussian Distribution.
If the variance is proportional to the mean and we have a discrete distribution, then we have a Poisson Distribution.

For the Gamma Distribution, $f(y) = (y/\theta)^\alpha \exp[-y/\theta] / (y\,\Gamma[\alpha])$. $E[Y] = \alpha\theta$. $Var[Y] = \alpha\theta^2$.
If used in a GLM, then we are assuming that we have a Gamma Distribution with α fixed.
Then, Variance = $\alpha\theta^2 = (\alpha\theta)^2/\alpha = (\text{mean})^2 / \alpha$.
Thus for the Gamma Distribution (with α fixed) the variance is proportional to the square of the mean.
For the Gamma Distribution: $V(\mu) = \mu^2$ and $\phi = 1/\alpha$.

For the following members of the exponential family of distributions, where m is their mean, their variance is proportional to μ^p:
- **Normal distribution, p = 0.**
- **Poisson distribution, p = 1.**
- **Gamma distribution, p = 2.**
- **Tweedie distribution, 1 < p < 2.**
- **Inverse Gaussian distribution, p = 3.**

[22] Linear Exponential families are defined via the form of their density; however, this definition is not on the syllabus of this exam.

The syllabus reading gives a list of $V(\mu)$ for different exponential families.[23]
The syllabus reading does not go into detail on how to relate the parameterization of exponential families using μ and ϕ to that which you may already be familiar from for example Loss Models.
However, in order to make things a little more concrete here is a table.

Distribution	μ	ϕ	$V(\mu)$
Normal	μ	σ^2	1
Poisson[24]	λ	1	μ
Gamma[25]	$\alpha\theta$	$1/\alpha$	μ^2
Inverse Gaussian[26]	μ	$1/\theta$	μ^3
Negative Binomial[27]	β/κ	1	$\mu(1 + \kappa\mu)$
Binomial[28]	mq	1	μ (1 - μ/m)
Tweedie[29]			μ^p

---

[23] See Table 1 in Goldburd, Khare, and Tevet.

[24] As discussed subsequently, for the overdispersed Poisson $\phi > 1$.

[25] As per Loss Models, with mean = $\alpha\theta$ and variance = $\alpha\theta^2$, with α fixed.

[26] As per Loss Models, with mean = μ and variance = μ^3/θ, with θ fixed.

[27] Where $\kappa = 1/r$, fixed. κ is called the overdispersion parameter.
As per Loss Models, the Negative Binomial has mean = r β and variance = r β (1+β).

[28] As per Loss Models with m fixed, with mean = mq and variance = mq(1-q). m = 1 is a Bernoulli.
Goldburd, Khare, and Tevet give $V(\mu)$ for the case where m = 1.

[29] To be discussed subsequently,

Gamma Distribution:

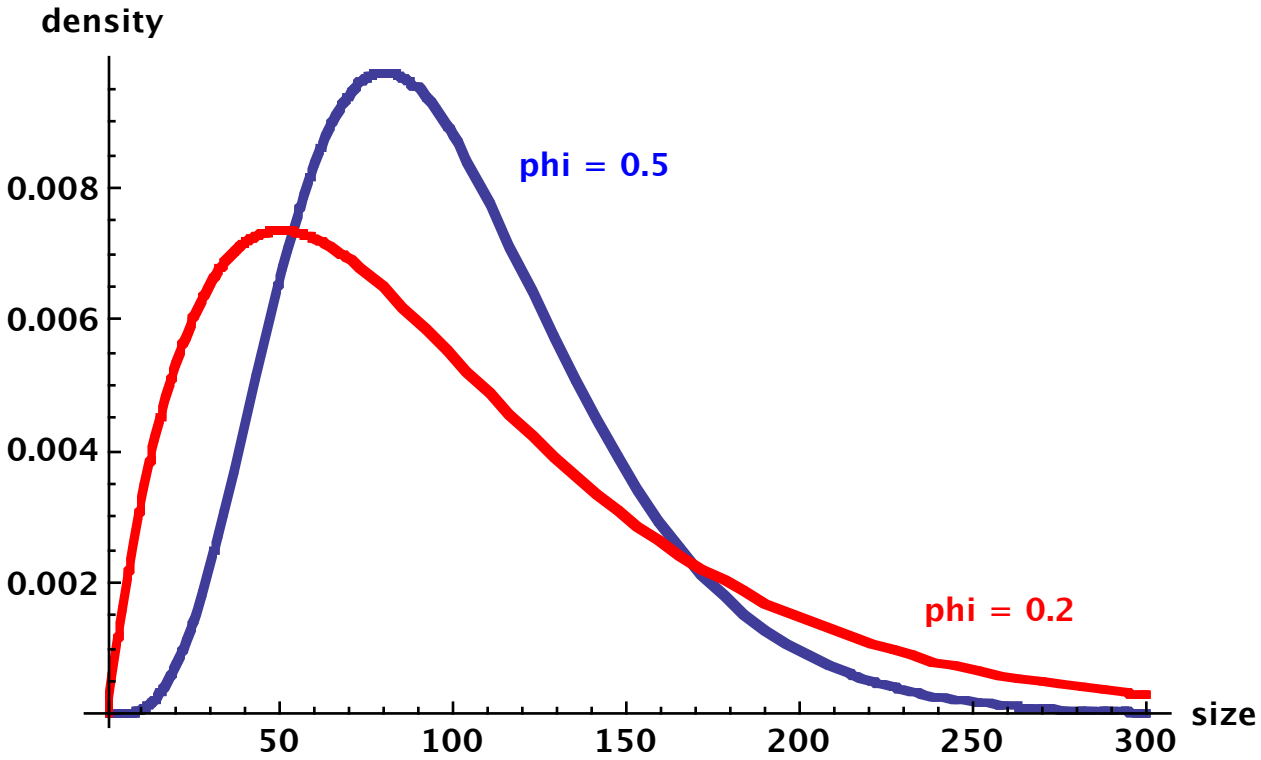
$f(x) = (x/\theta)^a \exp[-x/\theta] / (x\,\Gamma[\alpha])$, $x > 0$.[30]

Mean = $\alpha\theta$. Variance = $\alpha\theta^2$. CV = $1/\sqrt{\alpha}$

$\phi = 1/\alpha$. $V(\mu) = \mu^2$.

The Gamma Distribution is commonly used to model severity.

Here are graphs of the densities of Gamma Distributions with $\mu = 100$ and $\phi = 1/5$ or $1/2$:[31]



The Gamma Distribution has support from 0 to infinity. The Gamma Distribution is right-skewed (has positive skewness), with a sharp peak and a long tail to the right.

Exercise: Determine the variance for a Gamma Distribution with $\mu = 20$ and $\phi = 1/4$.
[Solution: Variance = $\phi\, V(\mu) = \phi\, \mu^2 = (1/4)(20^2) = 100$.
Comment: The coefficient of variation is: $\sqrt{100}/20 = 1/2 = \sqrt{\phi}$.]

[30] Parameterized as per Loss Models, not on the syllabus of this exam.
I do not expect you to need to know the density.
[31] The first has $\alpha = 5$ and $\Theta = 20$, while the second has $\alpha = 2$ and $\theta = 50$.

Inverse Gaussian Distribution:[32]

As per Loss Models: $f(x) = \sqrt{\frac{\theta}{2\pi}} \frac{\exp\left[-\frac{\theta\left(\frac{x}{\mu} - 1\right)^2}{2x}\right]}{x^{1.5}}$, $x > 0$.[33]

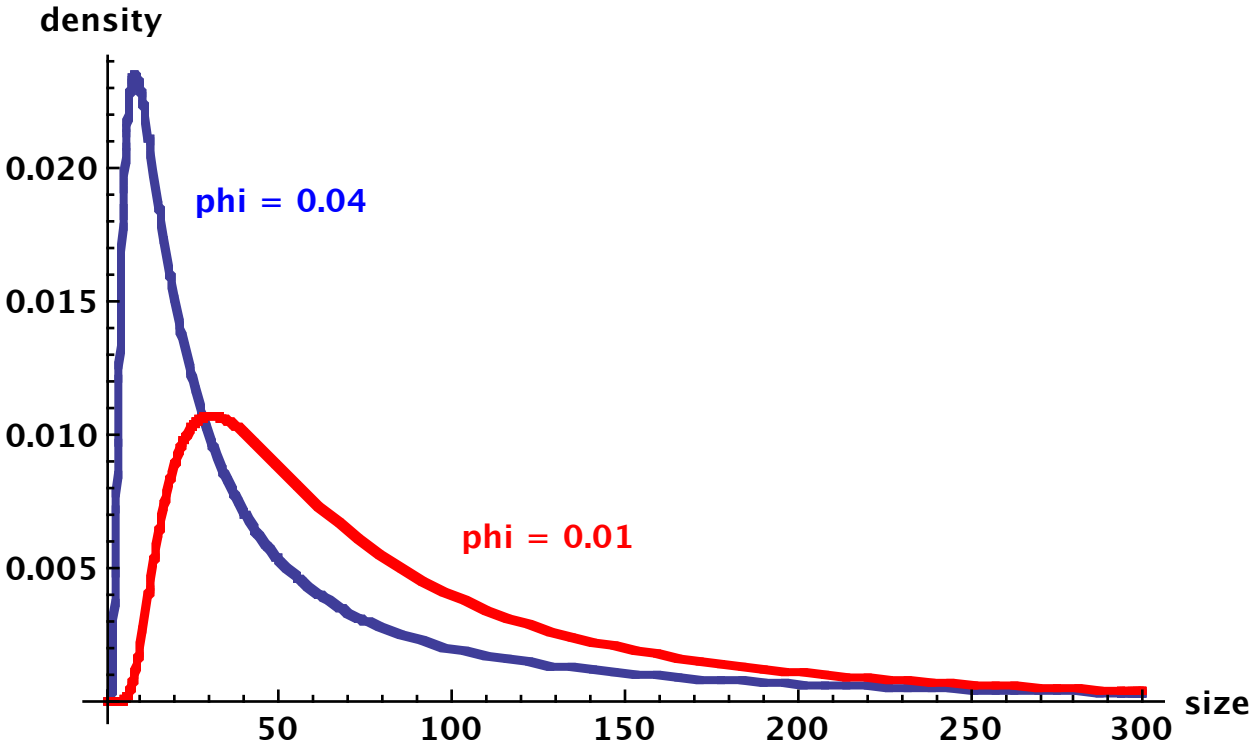
Mean = μ. Variance = μ^3 / θ.

$\phi = 1/\theta$. $V(\mu) = \mu^3$.

The Inverse Gaussian Distribution can be used to model severity. The Inverse Gaussian Distribution is appropriate when the severity has a larger skewness than for a Gamma.

Exercise: Determine the variance for an Inverse Gamma Distribution with $\mu = 20$ and $\phi = 1/5$.
[Solution: Variance = $\phi V(\mu) = \phi \mu^3 = (1/5)(20^3) = 1600$.]

Graphs of the densities of Inverse Gaussian Distributions with $\mu = 100$ and $\phi = 0.04$ or 0.01:[34]



[32] While the Gaussian (normal) describes a Brownian Motion's level at a fixed time, the Inverse Gaussian describes the distribution of the time a Brownian Motion with positive drift takes to reach a fixed positive level.
The cumulant generating function is the natural log of the moment generating function.
The cumulant generating function of an Inverse Gaussian is the inverse function of that of a Gaussian (Normal).

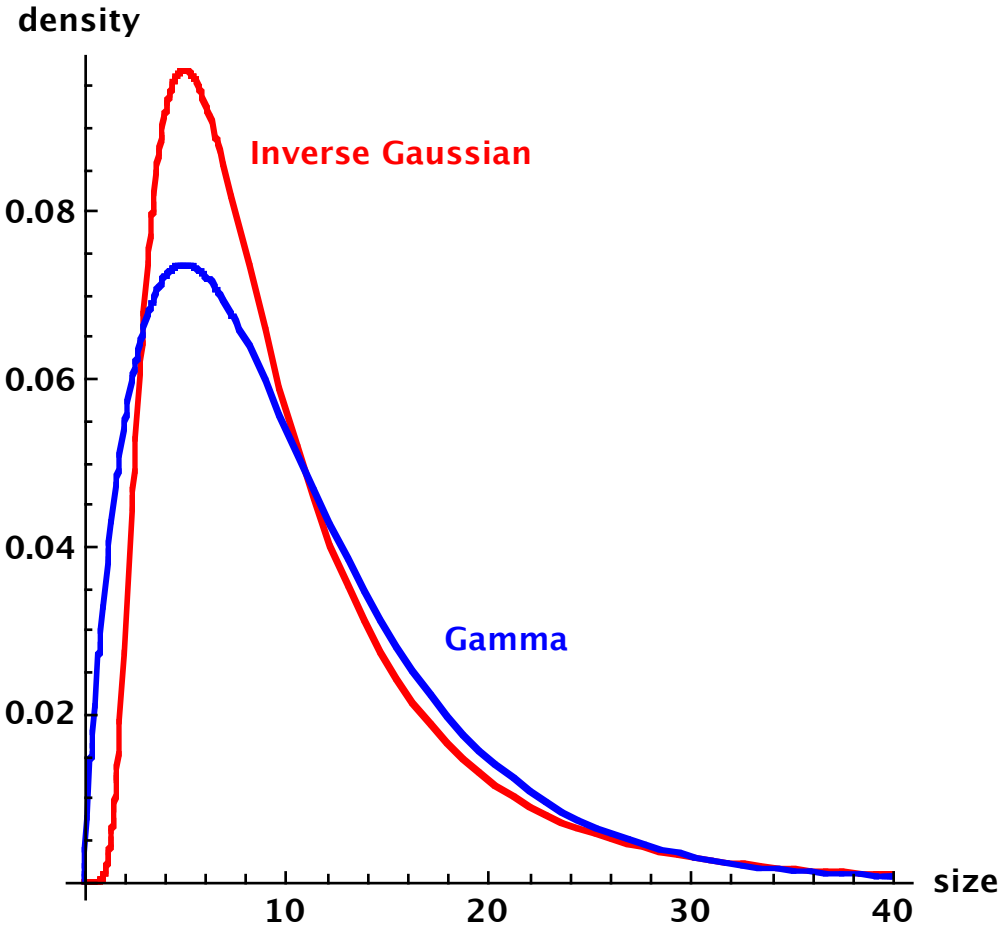
[33] I do not expect you to need to know the density.

[34] The first has $\theta = 25$, while the second has $\theta = 100$.

For the Gamma the variance is proportional to the square of the mean,
while for the Inverse Gaussian the variance is proportional to the cube of the mean.
The Inverse Gaussian and Gamma are similar, but the Inverse Gaussian has larger skewness and a higher peak.[35]

For example, a Gamma Distribution with $\mu = 10$ and $\alpha = 2$ has mean = 10, and variance = $10^2/2 = 50$. An Inverse Gaussian Distribution with $\mu = 10$ and $\phi = 20$ has mean = 10, and variance = $10^3/20 = 50$. Thus these two distributions have the same mean and variance.

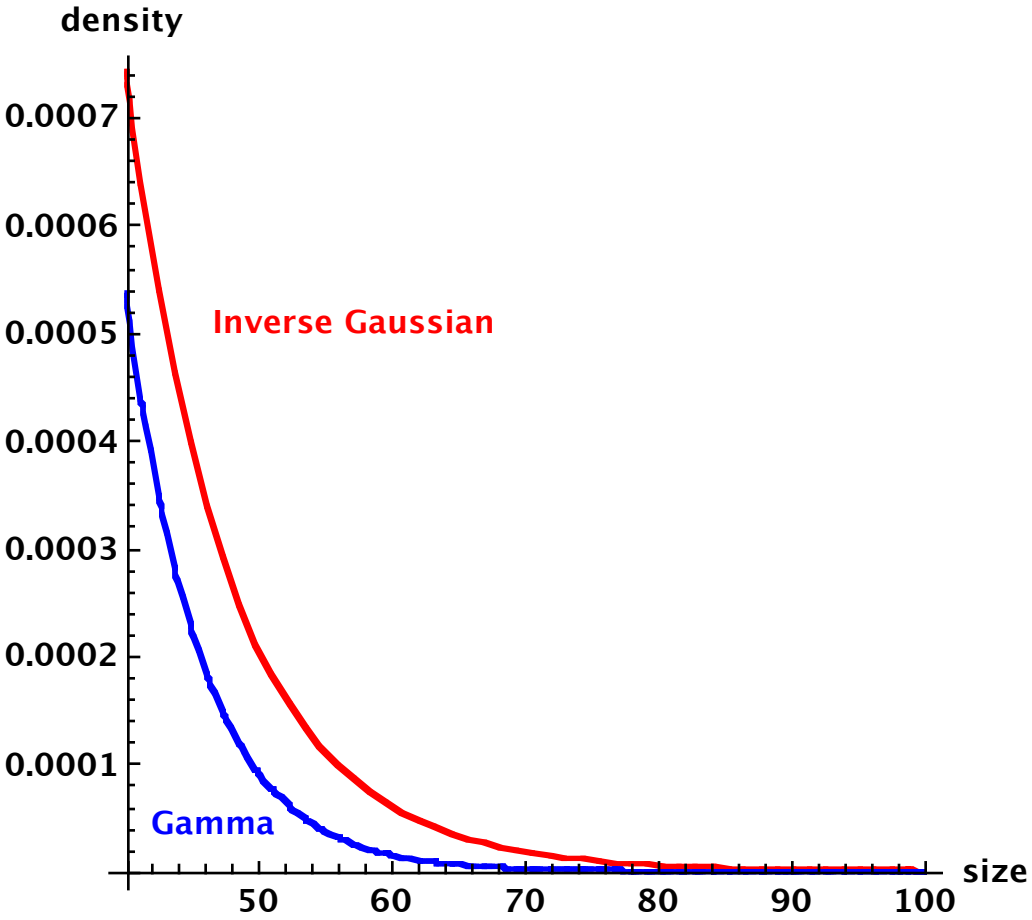
Here is a graph comparing these two densities:



The Inverse Gaussian Distribution has a higher peak than the Gamma Distribution.

[35] The skewness for the Gamma distribution is always twice times the coefficient of variation.
The skewness for the Inverse Gaussian distribution is always three times the coefficient of variation.

The Inverse Gaussian has more probability in the extreme righthand tail. With the aid of a computer, for this Gamma Distribution the survival function at 40 is S(40) = 0.30%, while for this Inverse Gaussian Distribution, S(40) = 0.58%.



A Two Dimensional Example of Generalized Linear Models:[36]

Let us assume we have two types of drivers, male and female, and two territories, urban and rural. Then there are a total of four combinations of gender and territory.
We assume an equal number of claims in each of the four combinations.

Let us assume that we have the following observed severities:

		Urban	Rural
Male		800	500
Female		400	200

Let us assume the following generalized linear model:
Gamma Function
Reciprocal link function[37]

Define male and rural as the base level, which introduces a constant term.
Then the constant, β_0, applies to all observations.
Let $X_1 = 1$ if female and 0 if male.[38]
Let $X_2 = 1$ if urban and 0 if rural.[39]

$$1/\mu = \sum \beta_i x_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ . \Rightarrow \mu = \frac{1}{\beta_0 + \beta_1 x_1 + \beta_2 x_2} \ .$$

Therefore, the modeled means are:

		Urban	Rural
Male		$1/(\beta 0+\beta 2)$	$1/\beta 0$
Female		$1/(\beta 0+\beta 1+\beta 2)$	$1/(\beta 0+\beta 1)$

For the Gamma Distribution as per Loss Models, $f(y) = (y/\theta)^\alpha \exp[-y/\theta] / (y\,\Gamma[\alpha])$.

$$\ln f(y) = (\alpha-1)\ln(y) - y/\theta - \alpha\ln(\theta) - \ln[\Gamma[\alpha]] = (\alpha-1)\ln(y) - y/(\mu/\alpha) - \alpha\ln(\mu/\alpha) - \ln[\Gamma[\alpha]]$$
$$= (\alpha-1)\ln(y) - \alpha y/\mu - \alpha\ln(\mu) + \alpha\ln(a) - \ln[\Gamma[\alpha]]$$
$$= (\alpha-1)\ln(y) - \alpha y(\beta_0 + \beta_1 x_1 + \beta_2 x_2) + \alpha\ln(\beta_0 + \beta_1 x_1 + \beta_2 x_2) + \alpha\ln(\alpha) - \ln[\Gamma[\alpha]].$$

---

[36] I do not expect you to have to go into this level of detail on your exam.
See page 24 and Appendix F of "A Practitioners Guide to Generalized Linear Models," by Anderson, et. al.
[37] One could instead use the log link function, and obtain somewhat different results.
[38] Since we have taken male as the base level, the covariate has to involve not male.
[39] Since we have taken rural as the base level, the covariate has to involve not rural.

The loglikelihood is the sum of the contributions from the four observations:
$(\alpha-1)\{\ln(800) + \ln(400) + \ln(500) + \ln(200)\}$
$- \alpha\{800(\beta_0 + \beta_2) + 400(\beta_0 + \beta_1 + \beta_2) + 500\beta_0 + 200(\beta_0 + \beta_1)\}$
$+ \alpha\{\ln(\beta_0 + \beta_2) + \ln(\beta_0 + \beta_1 + \beta_2) + \ln(\beta_0) + \ln(\beta_0 + \beta_1)\} + 4\alpha\ln(\alpha) - 4\ln[\Gamma(\alpha)]$.

To maximize the loglikelihood, we set its partial derivatives equal to zero.
Setting the partial derivative with respect to β_0 equal to zero:

$0 = -\alpha(800 + 400 + 500 + 200) + \alpha\{1/(\beta_0 + \beta_2) + 1/(\beta_0 + \beta_1 + \beta_2) + 1/\beta_0 + 1/(\beta_0 + \beta_1)\}$.

$\Rightarrow 1/(\beta_0 + \beta_2) + 1/(\beta_0 + \beta_1 + \beta_2) + 1/\beta_0 + 1/(\beta_0 + \beta_1) = 1900$.

Setting the partial derivative with respect to β_1 equal to zero:

$0 = -\alpha(400 + 200) + \alpha\{1/(\beta_0 + \beta_1 + \beta_2) + 1/(\beta_0 + \beta_1)\}$. $\Rightarrow 1/(\beta_0 + \beta_1 + \beta_2) + 1/(\beta_0 + \beta_1) = 600$.

Setting the partial derivative with respect to β_2 equal to zero:

$0 = -\alpha(800 + 400) + \alpha\{1/(\beta_0 + \beta_2) + 1/(\beta_0 + \beta_1 + \beta_2)\}$. $\Rightarrow 1/(\beta_0 + \beta_2) + 1/(\beta_0 + \beta_1 + \beta_2) = 1200$.

Solving these three equations in three unknowns:[40]
$\beta_0 = 0.00223811$, $\beta_1 = 0.00171142$, and $\beta_2 = -0.00106605$.

$$\mu = \frac{1}{0.00223811 + 0.00171142x_1 - 0.00106605x_2}.$$

For Male and Urban: $x_1 = 0$, $x_2 = 1$, and $\mu = 1 / (0.00223811 - 0.00106605) = 853.20$.
For Female and Urban: $x_1 = 1$, $x_2 = 1$,
and $\mu = 1 / (0.00223811 + 0.00171142 - 0.00106605) = 346.80$.
For Male and Rural: $x_1 = 0$, $x_2 = 0$, and $\mu = 1/0.00223811 = 446.81$.
For Female and Rural: $x_1 = 1$, $x_2 = 0$, and $\mu = 1/(0.00223811 + 0.00171142) = 253.20$.

The fitted severities by cell are:[41]

		Urban	Rural	Average
Male		853.20	446.81	650.01
Female		346.80	253.20	300.00
Average		600.00	350.01	475.00

[40] I used a computer to solve these three equations.
There is no need to solve for α in order to calculate the fitted pure premiums by cell.
However, using a computer, the maximum likelihood alpha is 45.6.
[41] The averages were computed assuming the same number of claims by cell.

This compares to the observed severities by cell:

		Urban	Rural	Average
Male		800	500	650
Female		400	200	300
Average		600	350	475

Notice how the averages for male, female, urban, and rural are equal for the fitted and observed. The overall experience of each class and territory has been reproduced by the model.

In general, the estimates will be in balance as they were here, when one uses the canonical link function; the canonical link function for the Gamma is the reciprocal link function.[42]

Exercise: For the Urban territory, what the relativity of male compared to female indicated by the GLM?
[Solution: 853.20/346.80 = 2.460.]

Exercise: For the Rural territory, what the relativity of male compared to female indicated by the GLM?
[Solution: 446.81/253.20 = 1.765.]

The relativities are different in the different territories. In general for a particular GLM, the relativities for one predictor variable can depend on the level(s) of the other predictor variable(s).

Here we have used the reciprocal link function. If instead the log link function had been used, the model would have been multiplicative, and the indicated multiplicative relativities would not have depended on territory. If instead the identity link function had been used, the model would have been additive, and the indicated additive relativities would not have depended on territory.

We could instead change the definitions of the covariates, and have a model without an intercept:
$x_1 = 1$ if male.
$x_2 = 1$ if female.
$x_3 = 1$ if urban and $x_3 = 0$ if rural.

Then $1/\mu = \sum \beta_i x_i = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3. \Rightarrow \mu = \frac{1}{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}$.

[42] See "A Systematic Relationship Between Minimum Bias and Generalized Linear Models", by Stephen J. Mildenhall, PCAS 1999, not on the syllabus.

Therefore, the modeled means are:

		Urban	Rural
Male		$1/(\beta 1+\beta 3)$	$1/\beta 1$
Female		$1/(\beta 2+\beta 3)$	$1/\beta 2$

For the Gamma Distribution as per Loss Models, $f(y) = (y/\theta)^\alpha \exp[-y/\theta] / (y\,\Gamma[\alpha])$..

$\ln f(y) = (\alpha-1)\ln(y) - y/\theta - \alpha\ln(\theta) - \ln[\Gamma[\alpha]] = (\alpha-1)\ln(y) - y/(\mu/\alpha) - \alpha\ln(\mu/\alpha) - \ln[\Gamma[\alpha]]$

$= (\alpha-1)\ln(y) - \alpha y/\mu - \alpha\ln(\mu) + \alpha\ln(a) - \ln[\Gamma[\alpha]]$

$= (\alpha-1)\ln(y) - \alpha y\,(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3) + \alpha\ln(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3) + \alpha\ln(\alpha) - \ln[\Gamma[\alpha]]$.

The loglikelihood is the sum of the contributions from the four observations:

$(\alpha-1)\{\ln(800) + \ln(400) + \ln(500) + \ln(200)\}$

$- \alpha\{800(\beta_1 + \beta_3) + 400(\beta_2 + \beta_3) + 500\beta_1 + 200\beta_2\}$

$+ \alpha\{\ln(\beta_1 + \beta_3) + \ln(\beta_2 + \beta_3) + \ln(\beta_1) + \ln(\beta_2)\} + 4\,\alpha\ln(\alpha) - 4\,\ln[\Gamma[\alpha]]$.

To maximize the loglikelihood, we set its partial derivatives equal to zero.

Setting the partial derivative with respect to β_1 equal to zero:

$0 = -\alpha(800 + 500) + \alpha\{1/(\beta_1 + \beta_3) + 1/\beta_1\}. \Rightarrow 1/(\beta_1 + \beta_3) + 1/\beta_1 = 1300.$

Setting the partial derivative with respect to β_2 equal to zero:

$0 = -\alpha(400 + 200) + \alpha\{1/(\beta_2 + \beta_3) + 1/\beta_2\}. \Rightarrow 1/(\beta_2 + \beta_3) + 1/\beta_2 = 600.$

Setting the partial derivative with respect to β_3 equal to zero:

$0 = -\alpha(800 + 400) + \alpha\{1/(\beta_1 + \beta_3) + 1/(\beta_2 + \beta_3)\}. \Rightarrow 1/(\beta_1 + \beta_3) + 1/(\beta_2 + \beta_3) = 1200.$

Solving these three equations in three unknowns:[43]

$\beta_1 = 0.00223811$, $\beta_2 = 0.00394952$, and $\beta_3 = -0.00106605$.

$$\mu = \frac{1}{0.00223811x_1 + 0.00394952x_2 - 0.00106605x_3}.$$

For Male and Urban: $x_1 = 1$, $x_2 = 0$, $x_3 = 1$, and $\mu = 1 / (0.00223811 - 0.00106605) = 853.20$.

For Female and Urban: $x_1 = 0$, $x_2 = 1$, $x_3 = 1$, and $\mu = 1 / (0.00394952 - 0.00106605) = 346.80$.

For Male and Rural: $x_1 = 1$, $x_2 = 0$, $x_3 = 0$, and $\mu = 1/0.00223811 = 446.81$.

For Female and Rural: $x_1 = 0$, $x_2 = 1$, $x_3 = 0$, and $\mu = 1/0.00394952 = 253.20$.

The modeled means are the same as in the other version of the model with a base level.

[43] I used a computer to solve these three equations.

Instead fit an Inverse Gaussian with the inverse square link function to this same data.[44] [45]

For the Inverse Gaussian: $f(x) = \sqrt{\frac{\theta}{2\pi}} \frac{\exp\left[-\frac{\theta\left(\frac{x}{\mu} - 1\right)^2}{2x}\right]}{x^{1.5}}$, mean = μ, variance = μ^3 / θ.

Ignoring terms that do not involve μ,

$\ln f(x) = -\frac{\theta\left(\frac{x}{\mu} - 1\right)^2}{2x} = -\frac{\theta}{2x}\left(\frac{x^2}{\mu^2} - 2\frac{x}{\mu} + 1\right) = -\frac{\theta x}{2\mu^2} + \frac{\theta}{\mu} - \frac{\theta}{2x}$.

Use $x_1 = 1$ if male.
$x_2 = 1$ if female.
$x_3 = 1$ if urban and $x_3 = 0$ if rural.

Using the squared reciprocal link function: $1/\mu^2 = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.
Thus ignoring terms that do not include μ, the loglikelihood is:
$\frac{-\theta}{2}\{800(\beta_1 + \beta_3) + 500(\beta_1) + 400(\beta_2 + \beta_3) + 200(\beta_2)\} + \theta\{\sqrt{\beta_1 + \beta_3} + \sqrt{\beta_1} + + \sqrt{\beta_2 + \beta_3}\}$.

Setting the partial derivative with respect to β_1 equal to zero:
$0 = \frac{-\theta}{2}\{800 + 500\} + \frac{\theta}{2}\{1/\sqrt{\beta_1 + \beta_3} + 1/\sqrt{\beta_1}\}. \Rightarrow 1300 = 1/\sqrt{\beta_2 + \beta_3} + 1/\sqrt{\beta_2}$.
Setting the partial derivative with respect to β_2 equal to zero:
$0 = \frac{-\theta}{2}\{400 + 200\} + \frac{\theta}{2}\{1/\sqrt{\beta_2 + \beta_3} + 1/\sqrt{\beta_2}\}. \Rightarrow 600 = 1/\sqrt{\beta_2 + \beta_3} + 1/\sqrt{\beta_2}$.
Setting the partial derivative with respect to β_3 equal to zero:
$0 = \frac{-\theta}{2}\{800 + 400\} + \frac{\theta}{2}\{1/\sqrt{\beta_1 + \beta_3} + 1/\sqrt{\beta_2 + \beta_3}\}. \Rightarrow 1200 = 1/\sqrt{\beta_1 + \beta_3} + 1/\sqrt{\beta_2 + \beta_3}$.

Solving these three equations in three unknowns:[46]
$\beta_1 = 0.0000054693$, $\beta_2 = 0.0000134722$, and $\beta_3 = -0.00000415544$.
$1/\mu^2 = 0.0000054693x_1 + 0.0000134722x_2 - 0.00000415544x_3$.

---

[44] Again assuming equal claims per cell.
[45] While the inverse square is the canonical link function for the Inverse Gaussian, one could use a different link function.
[46] I used a computer to solve these three equations.
There is no need to solve for θ in order to calculate the fitted pure premiums by cell.

For Male and Urban: $x_1 = 1$, $x_2 = 0$, $x_3 = 1$, and
$\mu = 1 / \sqrt{0.0000054693 - 0.00000415544} = 872.42$.
For Female and Urban: $x_1 = 0$, $x_2 = 1$, $x_3 = 1$, and
$\mu = 1 / \sqrt{0.0000134722 - 0.00000415544} = 327.62$.
For Male and Rural: $x_1 = 1$, $x_2 = 0$, $x_3 = 0$, and $\mu = 1 / \sqrt{0.0000054693} = 427.60$.
For Female and Rural: $x_1 = 0$, $x_2 = 1$, $x_3 = 0$, and $\mu = 1 / \sqrt{0.0000134722} = 272.45$.

The fitted severities by cell differ from the previous model and are as follows:[47]

		Urban	Rural	Average
Male		872.42	427.60	650.01
Female		327.62	272.45	300.04
Average		600.02	350.03	475.02

This compares to the observed severities by cell:

		Urban	Rural	Average
Male		800	500	650
Female		400	200	300
Average		600	350	475

Notice how subject to rounding, again the averages for male, female, urban, and rural are equal for the fitted and observed. The overall experience of each class and territory has been reproduced by the model.

In general, the estimates will be in balance as they were here, when one uses the canonical link function; the canonical link function for the Inverse Gaussian is the inverse square link function.[48]
When the weights differ by cell, this balance involves weighted averages.

[47] The averages were computed assuming the same number of claims by cell.
[48] See "A Systematic Relationship Between Minimum Bias and Generalized Linear Models," by Stephen Mildenhall, PCAS 1999, not on the syllabus.

Design Matrix:

As is the case for multiple regression, it is common in GLMs to work with a design matrix.
Each row of the design matrix corresponds to one observation in the data.[49]
Each column of the design matrix corresponds to a covariate in the model.
If there is an intercept or constant term in the model, then the first column refers to it;
the first column of the design matrix will then consist of all ones.

A one dimensional example, with one covariate plus an intercept, was discussed previously:
Three observations: (1, 1), (2, 2), (3, 9).
$Y = \beta_0 + \beta_1 X$.

Then the design matrix is: $\begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}$.

Since the intercept applies to each observation, the first column is all ones.
The second column contains the observed values of the only covariate X.

Note that the design matrix depends on the observations and the definitions of the covariates.
The design matrix does not depend on the link function or the distributional form of the errors.

The response vector would contain the observed values of Y: $\begin{pmatrix} 1 \\ 2 \\ 9 \end{pmatrix}$.

The vector of parameters is: $\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$.

If one used the identity link function, then this model can be written as: $E[Y] = X \beta$,
where X is the design matrix and β is the vector of parameters.
If instead one used the log link function, then this model can be written as: $E[Y] = \exp[X \beta]$.

In general, with a link function g, a GLM can be written as: $\mathbf{E[Y] = g^{-1}[X \beta]}$.

With more covariates, things get a little more complicated. There is not a unique way to define the covariates. The important thing is to have the design matrix be consistent with the chosen definitions of the covariates.

[49] When we have more than one exposure or claim in a cell, a row may correspond to several observations grouped.

A two dimensional model was previously discussed:

		Urban	Rural
Male		800	500
Female		400	200

Usually on your exam, one would define a base level, which introduces a constant term.
For example, as before we could define male/rural as the base level.[50]
Then the constant, β_0, would apply to all observations.

Let $X_1 = 1$ if female and 0 if male.[51]

Let $X_2 = 1$ if urban and 0 if rural.[52]

Then with link function g, the GLM is: $g(E[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.

If we order the observations as follows, then the design matrix is:

$$\begin{pmatrix} \text{Male/Urban} \\ \text{Male/Rural} \\ \text{Female/Urban} \\ \text{Female/Rural} \end{pmatrix} \Leftrightarrow \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

The first column of ones corresponds to the constant term which applies to all observations.
The first row of the design matrix corresponds to male/urban: $X_1 = 0$, $X_2 = 1$.
The second row corresponds to male/rural: $X_1 = 0$, $X_2 = 0$.
The third row corresponds to female/urban: $X_1 = 1$, $X_2 = 1$.
The last row corresponds to female/rural: $X_1 = 1$, $X_2 = 0$.

On your exam the model is likely to be defined with a base level.

---

[50] One could define any of the four combinations as the base level.
[51] Since male is the base level, the covariate has to involve not male.
[52] Since rural is the base level, the covariate has to involve not rural.

Nevertheless, one could instead define:
$X_1 = 1$ if male. (0 if female)
$X_2 = 1$ if female. (0 if male)
$X_3 = 1$ if urban and 0 if rural.
Then with link function g, the GLM is: $g(E[Y]) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.

Then if we order the observations as before, then the design matrix is:[53]

$$\begin{pmatrix} \text{Male/Urban} \\ \text{Male/Rural} \\ \text{Female/Urban} \\ \text{Female/Rural} \end{pmatrix} \Leftrightarrow \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

The first row corresponds to male/urban: $X_1 = 1$, $X_2 = 0$, and $X_3 = 1$.
The second row corresponds to male/rural: $X_1 = 1$, $X_2 = 0$, and $X_3 = 0$.
The third row corresponds to female/urban: $X_1 = 0$, $X_2 = 1$, and $X_3 = 1$.
The last row corresponds to female/rural: $X_1 = 0$, $X_2 = 1$, and $X_3 = 0$.

The response vector would contain the observed values of Y,
in the same order as the rows of the design matrix:

$$\begin{pmatrix} \text{Male/Urban} \\ \text{Male/Rural} \\ \text{Female/Urban} \\ \text{Female/Rural} \end{pmatrix} \Leftrightarrow \begin{pmatrix} 800 \\ 500 \\ 400 \\ 200 \end{pmatrix}.$$

The vector of parameters is: $\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$.

[53] One can put the observations in any order, as long as one is consistent throughout.

This definition of covariates is not unique. For example instead define:
$X_1 = 1$ if urban. (0 if rural)
$X_2 = 1$ if rural. (0 if urban)
$X_3 = 1$ if female and 0 if male.

Exercise: For these definitions, what are the design matrix and the response vector?
[Solution: If we order the observations as before, then the design matrix is:

$$\begin{pmatrix} \text{Male/Urban} \\ \text{Male/Rural} \\ \text{Female/Urban} \\ \text{Female/Rural} \end{pmatrix} \Leftrightarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

The response vector would contain the observed values of Y,
in the same order as the rows of the design matrix:

$$\begin{pmatrix} \text{Male/Urban} \\ \text{Male/Rural} \\ \text{Female/Urban} \\ \text{Female/Rural} \end{pmatrix} \Leftrightarrow \begin{pmatrix} 800 \\ 500 \\ 400 \\ 200 \end{pmatrix}.$$

Comment: While the design matrix is different than before, this version of the model is just as valid as the previous ones, as long as everything is handled consistently.
The first row of the design matrix corresponds to male/urban: $X_1 = 1$, $X_2 = 0$, and $X_3 = 0$.
The second row corresponds to male/rural: $X_1 = 0$, $X_2 = 1$, and $X_3 = 0$.
The third row corresponds to female/urban: $X_1 = 1$, $X_2 = 0$, and $X_3 = 1$.
The last row corresponds to female/rural: $X_1 = 0$, $X_2 = 1$, and $X_3 = 1$.]

Generalized Linear Models, An Example of Adding Dimensions:

Assume we have a one-dimensional model with two territories: Urban and Rural.
While there are several different ways to set up this model, let us define:
Urban is the base level, β_0 is the intercept, $X_1 = 1$ if Rural.

Let us now add another dimension, gender: Male or Female.

We can either let Female/Urban be the base level and $X_2 = 1$ if Male,
or let Male/Urban be the base level and $X_2 = 1$ if Female.
In either case, we add only one more variable to the model we had for one dimension.

We could now add another dimension such as age: Young, Senior, Other. Regardless of which model we had for two dimensions, we would add two more variables to include age. Age has three levels, and in order to add it to our model we need to add 3 - 1 = 2 variables to the model.

Assume our model for two dimensions had:
Female/Rural as the base level, β_0 is the intercept, with $X_1 = 1$ if Urban, $X_2 = 1$ if Male.
Then for example we could take:
Female/Rural/Other as the base level, β_0 is the intercept, with $X_3 = 1$ if Young and $X_4 = 1$ if Senior.

If the model has a base level and corresponding constant term, then each categorical variable introduces a number of covariates equal to the number of its levels minus 1.

In this example, the number of covariates is: (constant term) + (2-1) + (2-1) + (3-1) = 5.

In practical applications, it is important to choose the base level of each category to be one with lots of data. If the chosen base level has little data, then the standard errors of the coefficients will be larger than if one had chosen a base level with lots of data.[54]

[54] See Figure 2 in Goldburd, Khare, and Tevet. Even when the base level has lots of data, the standard errors of coefficients corresponding to levels with little data will be wider than those levels with more data.

For example, assume instead our model for two dimensions had:
No base level with $X_1 = 1$ if Urban, $X_2 = 1$ if Rural, and $X_3 = 1$ if Male.
Then for example we could take: $X_4 = 1$ if Young and $X_5 = 1$ if Senior.

Without a base level and corresponding constant term, then one and only one of the categorical variables has a number of covariates equal to the number of its levels.
Each of the other categorical variables introduces a number of covariates equal to the number of its levels minus one.

For this example, without a base level, territory has a number of covariates equal to its number of levels, while gender and age each have a number of covariates equal to their number of levels minus one. The total number of covariates is: 2 + (2-1) + (3-1) = 5, the same as before.

Design Matrices, Continuous Variables:

We have looked at discrete categorical variables such as territory. GLMs can also use continuous variables such as amount of insurance and time living at current residence.[55] WIth continuous variables, determining the design matrix is somewhat different than it is with discrete variables.

Let us assume we are modeling pure premiums for homeowners and observe five policies:

Policy	Amount of Insurance ($000)	Time at Residence	Pure Premium ($000)
1	100	3	0
2	130	11	30
3	180	0	0
4	250	7	80
5	400	16	0

If X_1 = Amount of Insurance and X_2 = Time at Residence,
then the design matrix and response vector are:

$$X = \begin{pmatrix} 100 & 3 \\ 130 & 11 \\ 180 & 0 \\ 250 & 7 \\ 400 & 16 \end{pmatrix} \quad Y = \begin{pmatrix} 0 \\ 30 \\ 0 \\ 80 \\ 0 \end{pmatrix}.$$

The GLM is: $g(E[Y]) = \beta X$.

[55] In some cases, the model will perform better if such continuous variables are grouped into several categories.

Poisson Distribution:

$f(x) = e^{-\lambda} \lambda^x / x!$, $x = 0, 1, 2, ...$
Mean = λ. Variance = λ.
$\phi = 1$. $V(\mu) = \mu$.

The Poisson Distribution is commonly used to model frequency.

Overdispersion:

$Var[Y_i] = \phi E[Y_i]$. Since for the Poisson $\phi = 1$, the variance is equal the mean.

When the variance is greater than the mean, one could use a Negative Binomial Distribution, which has a variance greater than its mean.[56]

We can instead use an overdispersed Poisson with $\phi > 1$.
$Var[Y_i] = \phi E[Y_i]$. For $\phi > 1$, variance is greater than the mean.
While this does not correspond to the likelihood of any exponential family, otherwise the GLM mathematics works.[57] [58]

Using an overdispersed Poisson (ODP), we get the same estimated betas as for the usual Poisson regression.[59]
However, the standard errors of all of the estimated parameters are multiplied by $\sqrt{\phi}$. [60]

Although not mentioned in the syllabus readings, the usual estimator of the dispersion parameter ϕ is: $\hat{\phi} = \frac{1}{n - p} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\mu_i}$.

---

[56] One way the variance can be greater than the mean is if frequency is Poisson for each insured, but the means of the Poissons vary between insureds. If the Poisson means follow a Gamma Distribution, then the mixed distribution is a Negative Binomial Distribution.
[57] This is called using a quasi-likelihood, although the syllabus reading dos not use that term.
[58] Often using a Negative Binomial Distribution or an overdispersed Poisson approach to fit a GLM will produce similar results.
[59] This is the same reason we can fit the betas in a Normal regression without fitting σ.
[60] The variance of the estimated parameter is multiplied by ϕ.

Negative Binomial Distribution:

$$f(x) = \frac{\Gamma(x+r)}{x!\,\Gamma(r)} \frac{\beta^x}{(1+\beta)^{x+r}} = \frac{\Gamma(x+1/\kappa)}{x!\,\Gamma(1/\kappa)} \frac{(\kappa\mu)^x}{(1+\kappa\mu)^{x+r}}, \quad x = 0, 1, 2, \ldots$$

Mean = $r\beta = \beta/\kappa$. Variance = $r\beta(1+\beta) = (\beta/\kappa)(1+\beta)$.

$\phi = 1$. V($\mu) = \mu(1 + \kappa\mu)$.

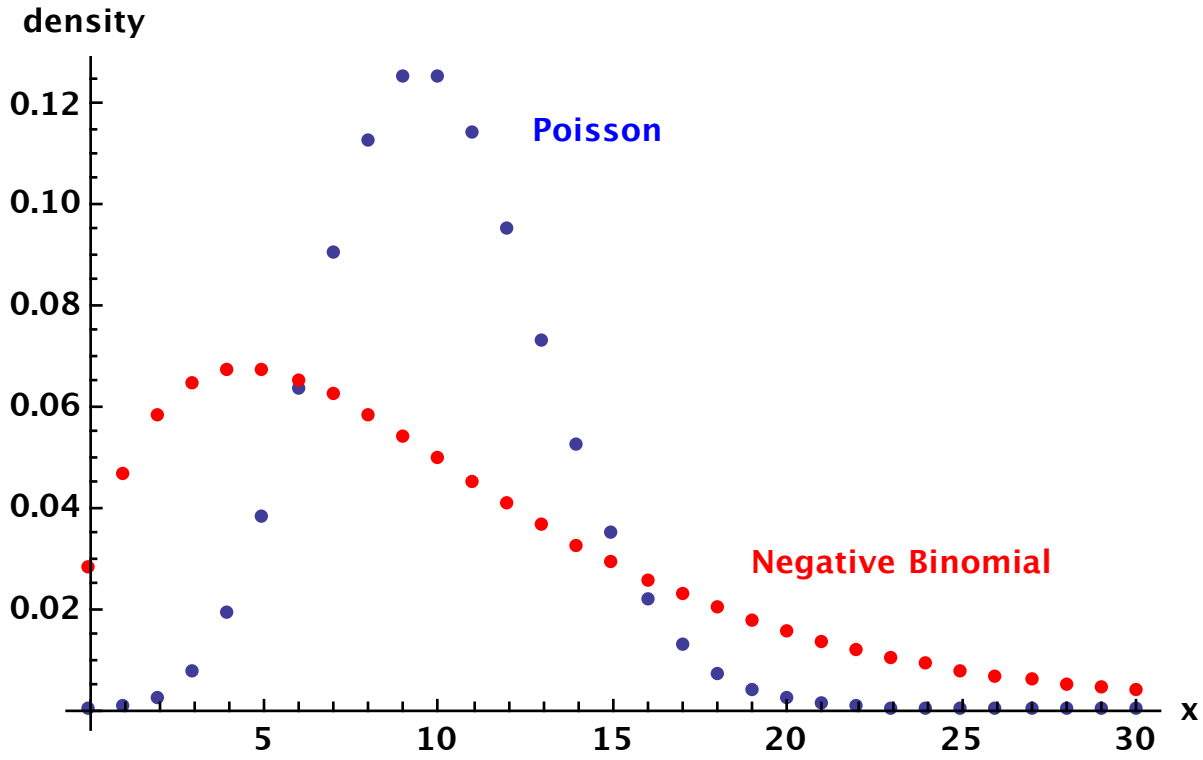
$\kappa = 1/r$ is called the overdispersion parameter.

As κ approaches zero while keeping the mean constant, the Negative Binomial Distribution approaches a Poisson Distribution.[61]

The Negative Binomial Distribution has its variance greater than its mean.
One way a Negative Binomial Distribution arises is as a Gamma mixture of Poisson Distributions.

The Negative Binomial Distribution is used to model frequency.

Here is a graph comparing a the densities of a Poisson with mean 5,
and a Negative Binomial with mean 5 and $\kappa = 1/2$ ($r = 2$):



[61] This is mathematically equivalent to letting β approach zero while keeping the mean constant.

One Dimensional Poisson Example with Exposures:

Exposures are a measure of how much insurance protection has been provided. Car years are an example. If one insures three cars each for two years, that is 6 car years of exposure.

Assume the same three observations: (1, 1), (2, 2), (3, 9).
However, let us assume 2, 3, and 4 exposures respectively.

Let us again fit a GLM using a Poisson with a log link function.
$\lambda_i = \exp[\beta_0 + x\beta_1]$.

We assume that Y_i is Poisson, with mean $n_i \lambda_i$,
where n_i is the number of exposures for observation i.
For example, the third observation is Poisson with mean: $4 \exp[\beta_0 + 3\beta_1]$.

For the Poisson Distribution, $\ln f(y) = -\lambda + y\ln(\lambda) - \ln(y!)$.
Thus the contribution to the loglikelihood from the third observation is:
$-4 \exp[\beta_0 + 3\beta_1] + 9 \{\ln 4 + (\beta_0 + 2\beta_1)\} - \ln[9!]$.

The loglikelihood is the sum of the contributions from the three observations:
$-2 \exp[\beta_0 + \beta_1] - 3 \exp[\beta_0 + 2\beta_1] - 4 \exp[\beta_0 + 3\beta_1] + (\beta_0 + \beta_1) + 2(\beta_0 + 2\beta_1) + 9(\beta_0 + 3\beta_1)$
$+ \ln[2] + 2 \ln[3] + 9 \ln[4] - \ln(1) - \ln(2) - \ln(9!)$.

To maximize the loglikelihood, we set its partial derivatives equal to zero.
Setting the partial derivative with respect to β_0 equal to zero:
$0 = -2 \exp[\beta_0 + \beta_1] - 3 \exp[\beta_0 + 2\beta_1] - 4 \exp[\beta_0 + 3\beta_1] + 12$.
Setting the partial derivative with respect to β_1 equal to zero:
$0 = -2 \exp[\beta_0 + \beta_1] - 6 \exp[\beta_0 + 2\beta_1]) - 12 \exp[\beta_0 + 3\beta_1] + 32$.

Solving these two equations in two unknowns: $\beta_0 = -1.97234$ and $\beta_1 = 0.91629$.[62]
$\mu_i = n_i \exp[-1.97234 + 0.91629 x_i]$.

For x = 1, $\mu = 2 \exp[-1.97234 + 0.91629] = 0.696$.
For x = 2, $\mu = 3 \exp[-1.97234 + (2)(0.91629)] = 2.609$.
For x = 3, $\mu = 4 \exp[-1.97234 + (3)(0.91629)] = 8.696$.

[62] I used a computer to solve these two equations. One can confirm that these values satisfy these equations.

Offsets, Poisson Model with Log Link Function:

When fitting a Poisson Distribution with a log link function, it is common to state the model with an offset term which is ln[exposures].
Offset terms are used to adjust for group size or differing time periods of observation.

With the log link function: $\lambda_i = \exp[\eta_i]$.

We assume that Y_i is Poisson, with mean $n_i \lambda_i$,
where n_i is the number of exposures for observation i.

$\mu_i = n_i \lambda_i = n_i \exp[\eta_i]$. $\Leftrightarrow$ **$\ln[\mu_i] = \ln[n_i] + \eta_i$.**

Thus we have rewritten the usual equation relating the mean to the linear predictor, $\eta = X\beta$, with an additional term, **$\ln[n_i]$ which is called the offset. Note that the offset involves a vector of known amounts, the number of exposures corresponding to each observation**.

In the previous example: $\ln[\mu_i] = \ln[n_i] + \beta_0 + \beta_1 x_i$. $\Leftrightarrow$ $\mu_i = n_i \exp[\beta_0 + \beta_1 x_i]$.
Thus the use of an offset term will produce an equivalent model and the same result as obtained previously.

Computer software to fit GLMs will have an option to include an offset term.

Offsets, When Updating Only Part of the Rating Plan:[63]

Assume for example, one is updating other parts of the rating algorithm, but is leaving the deductible credits the same.[64] The current deductibles and credits are a follows:

$500	Base
$1000	8% credit
$2500	14% credit

Then in a GLM for pure premium using a log link function:

$$\mu = \exp[X\beta]\, f_D,$$

where Xb is the linear predictor (not taking into account deductible),
and f_D is the appropriate deductible factor of: 1, 0.92, or 0.86.

$$\ln[\mu] = X\beta + \ln[f_D] = X\beta + \text{offset}.$$

This is mathematical the same as the use of an offset in the case of a Poisson frequency. However, there the offset was ln[exposures] while here the offset is ln[1 - deductible credit].

If an observation is from a policy with a $500 deductible, then the offset is ln[1] = 0.
If an observation is from a policy with a $1000 deductible, then the offset is ln[1 - 0.08] = -0.0834.
If an observation is from a policy with a $2500 deductible, then the offset is ln[1 - 0.14] = -0.1508.

The expected pure premium for a policy with a $2500 deductible is lower than that of a similar policy with a $500 deductible. If the mix of deductibles varies by the other classification variables, then we know that completely ignoring deductibles would lead to distorted estimates of the effects of the other classification variables. The use of the offset term takes into account deductible; however, we are assuming the effects of deductibles are known based on the current credits and that there is no (significant) interaction of effects between deductible amount and other classification variables.

In general, **an offset factor is a vector of known amounts which adjusts for known effects not otherwise included in the GLM**.

As another example, one could take the current territories and territory relativities as givens, and include an offset term in a GLM of ln[territory relativity].

---

[63] See Section 2.6 of Goldburd, Khare, and Tevet.
[64] Either you will update them at some later date, or the deductible credits will be determined by some technique other than by using a GLM.

Prior Weights:[65]

When observing numbers of claims, the volume of data is numbers of exposures. When observing sizes of claims, the volume of data is numbers of claims.[66] When a given observation is based on more data we give it more weight.

Let us return to the example with two types of drivers, male and female, and two territories, urban and rural. Before we assumed an equal number of claims in each of the four combinations.
Instead let us assume that the Urban/Male combination has twice the volume of the others; in other words Urban/Male has twice as many claims as each of other the other combinations.

Let us assume that we have the same observed average severities:

		Urban	Rural
Male		800	500
Female		400	200

Let us again assume the following generalized linear model:
Gamma Function
Reciprocal link function
$x_1 = 1$ if male.
$x_2 = 1$ if female.
$x_3 = 1$ if urban and $x_3 = 0$ if rural.

Then $1/\mu = \sum \beta_i x_i = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. $\Rightarrow \mu = \dfrac{1}{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}$.

Therefore, the modeled means are:

		Urban	Rural
Male		$1/(\beta 1 + \beta 3)$	$1/\beta 1$
Female		$1/(\beta 2 + \beta 3)$	$1/b2$

For the Gamma Distribution as per Loss Models, $f(y) = (y/\theta)^{\alpha} \exp[-y/\theta] / (y\, \Gamma[\alpha])$.

$\ln f(y) = (\alpha-1)\ln(y) - y/\theta - \alpha\ln(\theta) - \ln[\Gamma[\alpha]] = (\alpha-1)\ln(y) - y/(\mu/\alpha) - \alpha\ln(\mu/\alpha) - \ln[\Gamma[\alpha]]$

$= (\alpha-1)\ln(y) - \alpha y/\mu - \alpha\ln(\mu) + \alpha\ln(a) - \ln[\Gamma[\alpha]]$

$= (\alpha-1)\ln(y) - \alpha y(\beta_0 + \beta_1 x_1 + \beta_2 x_2) + \alpha\ln(\beta_0 + \beta_1 x_1 + \beta_2 x_2) + \alpha\ln(\alpha) - \ln[\Gamma[\alpha]]$.

---

[65] See Section 2.5 of Goldburd, Khare, and Tevet
[66] In Buhlmann Credibility, N is number of exposures when estimating frequency or pure premiums, but N is number of claims when estimating severity.

Since it now has twice the number of claims, we multiply the contribution from Urban/Male by two.

The loglikelihood is the sum of the contributions from the four combinations:
$(\alpha-1)\{2 \ln(800) + \ln(400) + \ln(500) + \ln(200)\}$
$- \alpha\{(2)(800)(\beta_1 + \beta_3) + 400(\beta_2 + \beta_3) + 500\beta_1 + 200\beta_2\}$
$+ \alpha\{2\ln(\beta_1 + \beta_3) + \ln(\beta_2 + \beta_3) + \ln(\beta_1) + \ln(\beta_2)\}\} + 5\alpha\ln(\alpha) - 5 \, l \, \ln[\Gamma[\alpha]]$.

To maximize the loglikelihood, we set its partial derivatives equal to zero.
Setting the partial derivative with respect to β_1 equal to zero:

$0 = -\alpha(1600 + 500) + \alpha\{2/(\beta_1 + \beta_3) + 1/\beta_1\}. \Rightarrow 2/(\beta_1 + \beta_3) + 1/\beta_1 = 2100.$

Setting the partial derivative with respect to β_2 equal to zero:

$0 = -\alpha(400 + 200) + \alpha\{1/(\beta_2 + \beta_3) + 1/\beta_2\}. \Rightarrow 1/(\beta_2 + \beta_3) + 1/b_2 = 600.$

Setting the partial derivative with respect to β_3 equal to zero:

$0 = -\alpha(1600 + 400) + \alpha\{2/(\beta_1 + \beta_3) + 1/(\beta_2 + \beta_3)\}. \Rightarrow 2/(\beta_1 + \beta_3) + 1/(\beta_2 + \beta_3) = 2000.$

Solving these three equations in three unknowns:[67]
$\beta_1 = 0.00224451$, $\beta_2 = 0.00392976$, and $\beta_3 = -0.00103566$.
$\mu = 1 / (0.00224451x_1 + 0.00392976x_2 - 0.00103566x_3)$.

For Male and Urban: $x_1 = 1$, $x_2 = 0$, $x_3 = 1$, and $\mu = 1 / (0.00224451 - 0.00103566) = 827.23$.
For Female and Urban: $x_1 = 0$, $x_2 = 1$, $x_3 = 1$, and $\mu = 1 / (0.00392976 - 0.00103566) = 345.53$.
For Male and Rural: $x_1 = 1$, $x_2 = 0$, $x_3 = 0$, and $\mu = 1/0.00224451 = 445.53$.
For Female and Rural: $x_1 = 0$, $x_2 = 1$, $x_3 = 0$, and $\mu = 1/0.00392976 = 254.47$.

The fitted severities by cell are:

		Urban	Rural
Male		827.23	445.30
Female		345.53	254.47

Which differ from those obtained previously when we had equal weights.

[67] I used a computer to solve these three equations.
There is no need to solve for α in order to calculate the fitted pure premiums by cell.

Let us examine what I did in a little more detail.

My contribution to the loglikelihood for Male/Urban was:
$2 \ln f(800) = 2\{(\alpha-1)\ln(800) - 800/\theta - \alpha\ln(\theta) - \ln[\Gamma(\alpha)]\}$
$= 2\{(\alpha-1)\ln(800) - 800\alpha/\mu - \alpha\ln(\mu/\alpha) - \ln[\Gamma(\alpha)]\}$.
This is the same as assuming two claims each of size 800 were observed.

If instead, we had two claims, one of size 600 and one of size 1000, averaging to the same 800, then the contribution to the loglikelihood for Male/Urban would be:
$\ln f(600) + \ln f(1000) =$
$(\alpha-1)\ln(600) - 600\alpha/\mu - \alpha\ln(\mu/\alpha) - \ln[\Gamma(\alpha)]] + (\alpha-1)\ln(1000) - 10000\alpha/\mu - \alpha\ln(\mu/\alpha) - \ln[\Gamma(\alpha)]]$
$= (\alpha-1)\{\ln(600) + \ln(1000)\} - 1600\alpha/m - 2\alpha\ln(\mu/\alpha) - 2\ln[\Gamma(\alpha)]]$.

This differs from before by some constant times $\alpha - 1$. However, this does not affect the fitted maximum likelihood parameters; when we take a partial derivative with respect to β_i these terms will drop out.

If we only use the fact that Urban/Male has two claims summing to 1600, then we can use the fact that the sum of two identically distributed Gammas has twice the alpha.[68] The mean will also be twice as big, so that $\theta = \mu/\alpha$ would remain the same. Thus the contribution to the loglikelihood for Male/Urban would be the log density of this Gamma with 2α at 1600:
$(2\alpha-1)\ln(1600) - 1600\alpha/\mu - 2\alpha\ln(\mu/\alpha) - \ln[\Gamma(\alpha)]]$.

Again, this differs from before by terms that involve constants and alpha. However, this does not affect the fitted maximum likelihood parameters; when we take a partial derivative with respect to β_i these terms will drop out.

The members of exponential families each have this nice property that the maximum likelihood fit only depends on the average and not the individual values.[69]

In general, **when modeling severity, let the weights w_i be the number of claims**.

So for example, if an observation is the average size of 10 claims, then the variance will be 1/10 of that for an observation of the size of a single claim.

For example, for the Poisson, $f(x) = \lambda^x e^{-\lambda} / x!$. $\ln f(x) = x \ln(\lambda) - \lambda - \ln(x!)$.
If we have two (independent) exposures each with mean frequency x, then we can multiply the contribution to the loglikelihood by two: $2x \ln(\lambda) - 2\lambda - 2\ln(x!)$.

If we have two (independent) exposures each with Poissons with mean λ, then the number of claims is Poisson with mean 2λ.

---

[68] The other exponential families share the property that when one adds up independent, identical copies one gets anther member of the same family.
[69] The mean is a sufficient statistic.

Then with a sum of 2x, and an average frequency of x, the log density is:
$2x \ln(2\lambda) - 2\lambda - \ln(2x!) = 2x \ln(\lambda) - 2\lambda - 2x \ln(2) - \ln(2x!)$.

Except for constants and terms involving x, this is the same loglikelihood as before.
Thus we would get the same maximum likelihood fit.
Thus when modeling claim frequencies, one can weight by the number of exposures.

When modeling claim frequency or pure premiums, let the weights be exposures.

When a weight is specified, the assumed variance for (the mean of) observation i is inversely proportional to the weight:[70] $\mathbf{Var[Y_i] = \phi\, V[\mu_i] / \omega_i}$.

[70] This is our usual assumption that the variance of an average is inversely proportional to the number of items being averaged.

A Three Dimensional Example of a GLM:[71]

Here is a three dimensional example for private passenger automobile insurance claim frequency, with: age of driver, territory, and vehicle class.[72] It is a multiplicative model, in other words a GLM with a log link function.[73]

There are 9 levels for driver age, 8 territories, and 5 classes of vehicle. An intercept term is used. Therefore, since each of the three factors is a categorical variable, each has one less parameter than its number of levels. In addition to the intercept term, there are 8 driver age parameters, 7 territory parameters, and 4 vehicle class parameters.

Choose age group 40-49, territory C, and vehicle class A, as the base levels.[74] [75]
Let b_1 correspond to the intercept term, and assign the other parameters as follows:

Age of driver		Territory		Vehicle class	
Factor level	Parameter	Factor level	Parameter	Factor level	Parameter
17-21	β_2	A	β_{10}	A	
22-24	β_3	B	β_{11}	B	β_{17}
25-29	β_4	C		C	β_{18}
30-34	β_5	D	β_{12}	D	β_{19}
35-39	β_6	E	β_{13}	E	β_{20}
40-49		F	β_{14}		
50-59	β_7	G	β_{15}		
60-69	β_8	H	β_{16}		
70+	β_9				

The total number of cells is: (9)(8)(5) = 360.
So the design matrix would have 360 rows, assuming that there are no cells lacking data.

---

[71] See pages 31 to 32 of "A Practitioner's Guide to Generalized Linear Models," by Duncan Anderson; Sholom Feldblum; Claudine Modlin; Doris Schirmacher; Ernesto Schirmacher; and Neeza Thandi, (Third Edition), CAS Study Note, February 2007. Not on the syllabus of this exam.
[72] Presumably, there would be another GLM fit to severity.
[73] We are not told what distributional form is assumed, but it is probably Poisson.
We are not given any details of the fitting or any diagnostics.
[74] One could make another set of choices and should get the same fitted frequencies.
[75] The standard errors of the fitted parameters are smaller if one chooses as the base level the one with the most exposures.

For example, the first row of the design matrix is probably for age 17-21, Territory A, and Class A, with ones in column 1, 2, and 10:[76]
1 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0.

The last row of the design matrix is probably for age 70+, Territory H, and Class E, with ones in column 1, 9, 16, and 20:
1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 1.

Exercise: For age 35-39, Territory F, and Class C, what does the corresponding row of the design matrix look like?
[Solution: Ones in columns 1, 6, 14, and 18:
1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0.]

The fitted parameters are an intercept term of 0.1412 and:[77] [78]

Age of driver		Territory		Vehicle class	
Factor level	Multiplier	Factor level	Multiplier	Factor level	Multiplier
17-21	1.6477	A	0.9407	A	1.0000
22-24	1.5228	B	0.9567	B	0.9595
25-29	1.5408	C	1.0000	C	1.0325
30-34	1.2465	D	0.9505	D	0.9764
35-39	1.2273	E	1.0975	E	1.1002
40-49	1.0000	F	1.1295		
50-59	0.8244	G	1.1451		
60-69	0.9871	H	1.4529		
70+	0.9466				

The estimated frequency for a 40-49 year old driver, from Territory C and Vehicle Class A, is 0.1412; the estimate for the base levels is the intercept term.[79]
For example, a 22-24 year old driver, from Territory G and Vehicle Class D would have an estimated frequency of: (1.5228)(1.1451)(0.9764)(0.1412) = 0.2404.

Exercise: What is the estimated frequency for a 30-34 year old driver, from Territory B and Vehicle Class E?
[Solution: (1.2465)(0.9567)(1.1002)(0.1412) = 0.1853.]

---

[76] How you arrange the rows of the design matrix does not affect the result, as long as everything is done consistently.

[77] For example, the fitted value of β_2 is ln(1.6447).

The multipliers for the base levels are one by definition.

[78] This is presumably illustrative rather than the output of a GLM fit in a practical application.

[79] In order to estimate the overall average frequency, one would need the distribution of exposures by cell.

Tweedie Distribution:[80]

Another (linear) exponential family is the Tweedie Distribution.
The Tweedie Distribution has mean μ and its **variance is proportional to μ^p, for $1 < p < 2$**.[81]

The Tweedie Distribution is used to model pure premiums (losses divided by exposures) or loss ratios; there is a point mass of probability at zero corresponding to no loss.
The Tweedie Distribution is mathematically a special case of a Compound Poisson Distribution.

When the Tweedie is used in GLMs, p and ϕ are constant across all observations.

When using the Tweedie distribution, it turns out that an increase in pure premium is made up of both an increase in frequency and an increase in severity.[82] Even if this assumption does not hold in an given application, the Tweedie GLM can still produce very useful and well fitting models of pure premium.

Details of the Tweedie Distribution:

It is a Poisson frequency with a Gamma severity, with parameters of the Poisson and Gamma:[83]
$\lambda = \frac{\mu^{2-p}}{\phi\ (2 - p)}$, $\alpha = \frac{2 - p}{p - 1}$, and $\theta = \phi\ (p-1)\ \mu^{p-1}$.

Exercise: Verify the mean and variance of the Tweedie as a Compound Poisson.
[Solution: Mean $= \lambda\alpha\theta = \frac{\mu^{2-p}}{\phi\ (2 - p)} \frac{2 - p}{p - 1} \phi\ (p-1)\ \mu^{p-1} = \mu$.

Variance = λ (2nd moment of Gamma) $= \lambda\ \alpha(\alpha+1)\theta^2$
$= \frac{\mu^{2-p}}{\phi\ (2 - p)} \frac{2 - p}{p - 1} \frac{1}{p - 1} \{\phi\ (p-1)\ \mu^{p-1}\}^2 = \phi\ \mu^p$.]

Exercise: What is the point mass at zero of the Tweedie as a Compound Poisson.
[Solution: This corresponds to the Poisson in the Compound Poisson being zero.
This has probability $e^{-\lambda}$.]

---

[80] See Section 2.7.3 of Goldburd, Khare, and Tevet.
[81] For insurance modeling, p is typically between 1.5 and 1.8.
In some software packages, one can specify the Tweedie distribution, which will in turn cause the software package to find the best value for the power parameter, p, when solving for the parameters (betas) in the linear equation.
[82] This is far from obvious. Why this is the case is discussed subsequently.
[83] For use in a GLM, ϕ and α (and thus p) are fixed for all observations.

$\alpha = \frac{2 - p}{p - 1}$. $\Rightarrow p = \frac{\alpha+2}{\alpha+1}$.

As alpha, the shape parameter of the Gamma, approaches infinity, p approaches 1, and the Tweedie approaches a Poisson. For p near one, the CV of the Gamma is small, and most of the randomness is due to the Poisson frequency.[84] As alpha approaches zero, p approaches 2, and the Tweedie approaches a Gamma.

Several of the other exponential family distributions are in fact special cases of Tweedie, dependent on the value of p:
• A Tweedie with p = 0 is a Normal distribution.
• A Tweedie with p = 1 is a Poisson distribution.
• A Tweedie with p = 2 is a Gamma distribution.
• A Tweedie with p = 3 is an inverse Gaussian distribution.

The mean of the Tweedie is: $\mu = \lambda\alpha\theta$. Also it turns out that: $\phi = \frac{\lambda^{1-p} (\alpha\theta)^{2-p}}{2 - p}$.

For a Compound Poisson with Gamma severity, we have Prob[X = 0] = $e^{-\lambda}$, and for x > 0:[85]

$$f(x) = \sum_{n=1}^{\infty} \frac{e^{-\lambda} \lambda^n}{n!} \frac{e^{-x/\theta} x^{n\alpha-1}}{\Gamma[n\alpha] \theta^{n\alpha}} = \exp[-x/\theta - \lambda] \sum_{n=1}^{\infty} \frac{(\lambda / \theta^\alpha)^n x^{n\alpha-1}}{n! \Gamma[n\alpha]}.$$

$\theta^\alpha = \phi^{(2-p)/(p-1)} (p-1)^{(2-p)/(p-1)} \mu^{2-p}$. We had: $\lambda = \frac{\mu^{2-p}}{\phi (2 - p)}$. Thus λ/θ^α does <u>not</u> depend on μ.

Thus the above sum does <u>not</u> depend on μ.

For a given GLM using the Tweedie, ϕ and $1 < p < 2$ are fixed. $\Rightarrow \alpha = \frac{2 - p}{p - 1}$ is fixed.

If μ increases, then $\lambda = \frac{\mu^{2-p}}{\phi (2 - p)}$ and $\theta = \phi (p-1) \mu^{p-1}$ each also increase.

Thus if the mean increases, then both mean frequency = λ, and mean severity = $\alpha\theta$ increase.

---

[84] For the Gamma Distribution, the coefficient of variation is $1/\sqrt{\alpha}$.

[85] Using the fact that the sum of n independent, identically distributed Gammas is another Gamma Distribution with parameters $n\alpha$ and θ.

An Example of a Tweedie Distribution:

Exercise: Take $\mu = 10$, $p = 1.5$, and $\phi = 4$. Determine the parameters of the Poisson and Gamma.

[Solution: $\lambda = \frac{\mu^{2-p}}{\phi\,(2-p)} = \frac{10^{0.5}}{(4)\,(2-1.5)} = 1.581$. $\alpha = \frac{2-p}{p-1} = (2 - 1.5) / (1.5 - 1) = 1$.

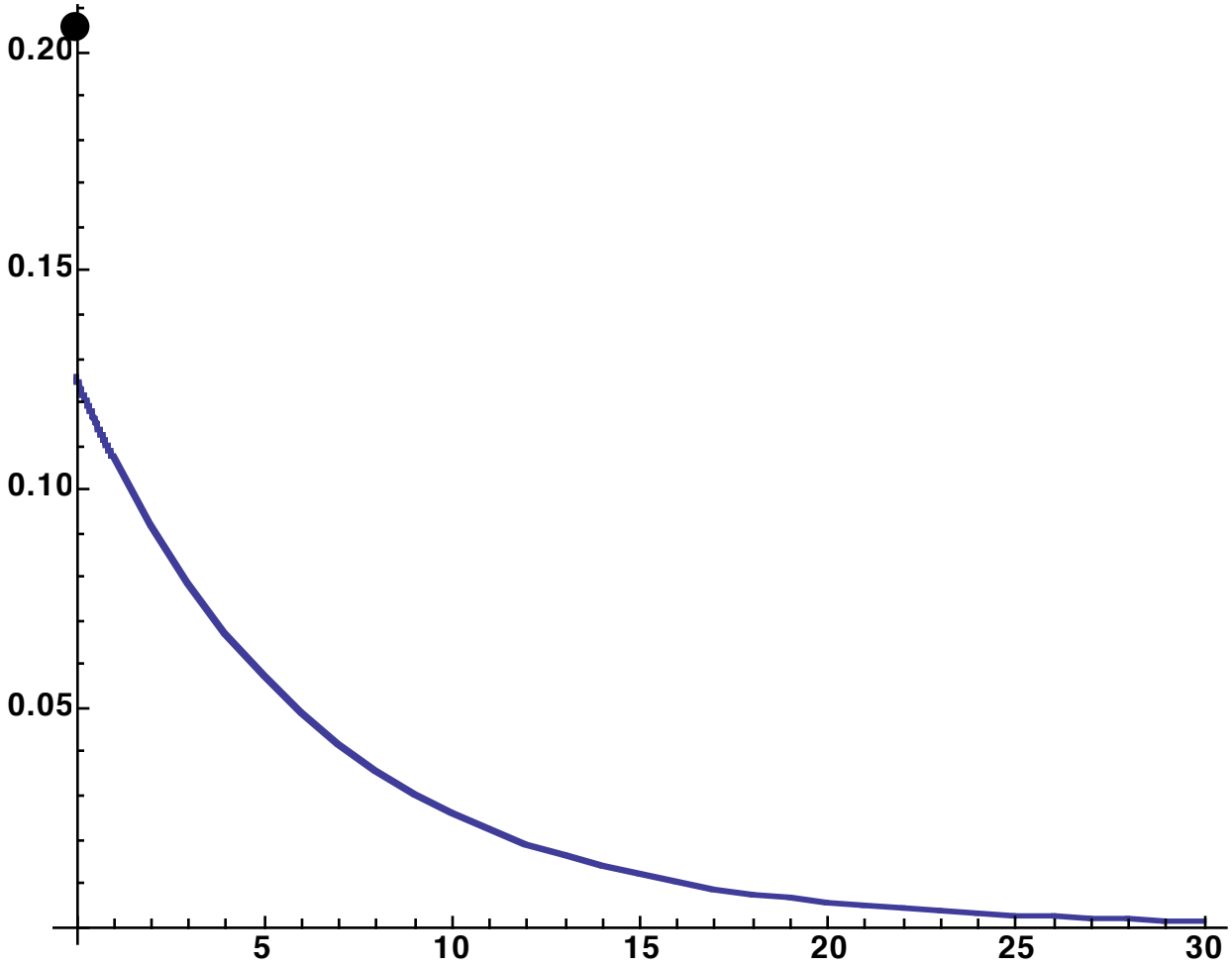
$\theta = f\phi\,(p-1)\,\mu^{p-1} = (4)(1.5 - 1)\,10^{(1.5-1)} = 6.325$.

Comment: The severity piece of the Compound Poisson is an Exponential with mean 6.325.
The mean of the Compound Poisson is: (1.581) (6.325) = 10 = μ.
The variance of the Compound Poisson is:
(mean of Poisson) (second moment of the Exponential) = (1.581) $\{(2)(6.325^2)\}$ = 126.5 = $(4)(10^{1.5}) = \phi\,\mu^p$.]

The density at zero of the Poisson is: $e^{-1.581} = 20.58\%$.
Thus there is a point mass of probability of 20.58% at zero.

Using a computer, this Tweedie has density at one of 0.1072.
This Tweedie has density at ten of 0.0258.
This Tweedie has density at twenty five of 0.0024.

Here is graph of the density of this Tweedie Distribution,
including the point mass of probability 20.58% at zero:[86] [87]



[86] For example, at 0.01 the density of the Tweedie is 0.1254.
[87] See Figure 5 in Goldburd, Khare, and Tevet.

Standard Errors and Confidence Intervals for Fitted Parameters:

A standard error is the standard deviation of an estimated coefficient.[88] Computer software for fitting GLMs will output the fitted coefficients and the corresponding standard errors.[89]

For GLMs for large samples, the Maximum Likelihood estimator is approximately multivariate Normal and asymptotically unbiased. Thus in GLM output, it is common to graph the fitted parameters and also bands plus or minus two standard errors.[90]

For example we might have fitted coefficients of:

$\hat{\beta}_0 = 223$, $\hat{\beta}_1 = 1.95$, and $\hat{\beta}_2 = -1.07$.

With corresponding standard errors of: 30.3, 0.607, and 0.632.

An approximate 95% confidence interval for β_0 is:

$223 \pm (1.960)(30.3) = (164, 282)$.

95% confidence interval for β_i is: $\hat{\beta}_i \pm 1.96$ (standard error of β_i).

Exercise: Determine an approximate 95% confidence interval for β_1,

[Solution: $1.95 \pm (1.960)(0.607) = (0.76, 3.14)$.]

Exercise: Determine an approximate 95% confidence interval for β_2,

[Solution: $-1.07 \pm (1.960)(0.632) = (-2.31, 0.17)$.]

A standard error of 30.3 for $\hat{\beta}_0$ can be thought of as follows: if one simulated similar sized data sets many times and fit GLMs, the estimated intercepts would have a variance of 30.3^2.
A smaller standard error gives us more confidence in the estimate of the corresponding coefficient.

Larger data sets will produce smaller standard errors than otherwise smaller data sets;
the standard errors go down approximately as the square root of the sample size.
The larger the estimated dispersion parameter ϕ, the more randomness there is in the data, and thus the larger the standard error; the standard error goes up as $\sqrt{\phi}$.

---

[88] This is similar to the standard error of a regression.
[89] Most software will also output the covariance matrix. The variances are along the diagonal.
The standard errors are the square roots of the variances.
[90] See Figure 2 in Goldburd, Khare, and Tevet.

One can perform hypothesis tests. For example, we can test $\beta_1 = 0$ versus $\beta_1 \neq 0$.

The probability value of this two-sided test is: $2 \{1 - \Phi[1.95/0.607]\} = 2 \{1 - \Phi[3.21]\} = 0.1\%$.[91]

p-value = Prob[test statistic takes on a value equal to its calculated value or
a value less in agreement with H_0 (in the direction of H_1) | H_0].

For a p-value sufficiently small, we can reject the null hypothesis in favor of the alternative hypothesis that the slope is non-zero. In this case, with a p-value of 0.1% we reject the hypothesis that $\beta_1 = 0$.

Exercise: Test $\beta_2 = 0$ versus $\beta_2 \neq 0$.

[Solution: p-value = $2 \Phi[-1.07/0.632] = 2 \Phi[-1.69] = 9.1\%$.
Therefore, we reject the null hypothesis at 10% and do not reject the null hypothesis at 5%.
Comment: Since zero was not in the 95% confidence interval for b_2,
we reject the null hypothesis at 5%.
Note that "not reject" is the correct statistical language, although actuaries sometimes say "accept".]

At the 10% significance level we can reject the hypothesis that $\beta_2 = 0$. However, at the 5% significance level there is insufficient evidence to reject the hypothesis that $\beta_2 = 0$.

We can perform two-sided tests: $\beta_2 = 0$ versus $\beta_2 \neq 0$.

We can also perform one-sided tests: $\beta_2 = 0$ versus $\beta_2 > 0$, or $\beta_2 = 0$ versus $\beta_2 < 0$.

[91] A table of the Normal Distribution will not be attached to your exam.

Using a p-value of 5%:

"A common statistical rule of thumb is to reject the null hypothesis where the p-value is 0.05 or lower. However, while this value may seem small, note that it allows for a 1-in-20 chance of a variable being accepted as significant when it is not. Since in a typical insurance modeling project we are testing many variables, this threshold may be too high to protect against the possibility of spurious effects making it into the model."[92]

For example, if we are testing the potential usefulness of 60 possible predictor variables, then if we use a p-value of 5%, even if none of the variables actually predict the outcome, on average three of these 60 variables will be selected as significant.

I performed a simulation experiment. I simulated 500 random observations from each of 60 independent normally distributed predictor variables. Then I simulated 500 observations from a normally distributed response variable.[93]

However, the response variable was independent of the predictor variables. In other words, none of the 60 predictor variables was actually useful for predicting the response variable. Then I fit a multiple regression to this data.[94]

The p-values of the 60 fitted slopes, were from smallest to largest:
0.005, 0.009, 0.020, 0.095, 0.109, 0.121, 0.148, 0.159, 0.177, 0.181, 0.196, 0.206, 0.253, 0.275, 0.331, 0.333, 0.387, 0.421, 0.423, 0.455, 0.494, 0.495, 0.495, 0.513, 0.521, 0.522, 0.545, 0.549, 0.562, 0.591, 0.593, 0.610, 0.614, 0.618, 0.629, 0.637, 0.645, 0.649, 0.653, 0.676, 0.684, 0.707, 0.707, 0.758, 0.778, 0.778, 0.790, 0.806, 0.825, 0.861, 0.886, 0.894, 0.894, 0.916, 0.941, 0.952, 0.980, 0.982, 0.987, 0.993.

We note that even though none of the 60 potential predictor variables is useful, three of the slopes are significant at the 5% level.[95] This illustrates the difficulty of relying on p-values when one starts with a large number of potential predictor variables. In such situations, it is very important to test any selected model on a separate holdout data set, as has been discussed.[96]

---

[92] Quoted from Section 2.3.2 of Goldburd, Khare, and Tevet.
[93] A Normal Distribution was used for simplicity.
[94] This is a special case of a GLM, with a Normal response using the identity link function.
[95] While we would expect (5%)(60) = 3 significant slopes, the fact that in this simulation it is exactly three is a coincidence.
[96] One can instead use k-fold validation, as discussed previously.

Log Link Function and Continuous Variables:[97]

As will be discussed, taking the log of continuous variables provides more variety of behaviors.
⇒ One is more likely to find one that fits your data.

Assume we are using the log link function.
For example: $\mu = \exp[\beta_0 + \beta_1 x_1 + \beta_2 x_2]$.

Then $\mu = \exp[\beta_0 + \beta_2 x_2] \exp[\beta_1]^{x_1}$.
Thus the multiplicative relativity for x_1 is $\exp[\beta_1]^{x_1}$.

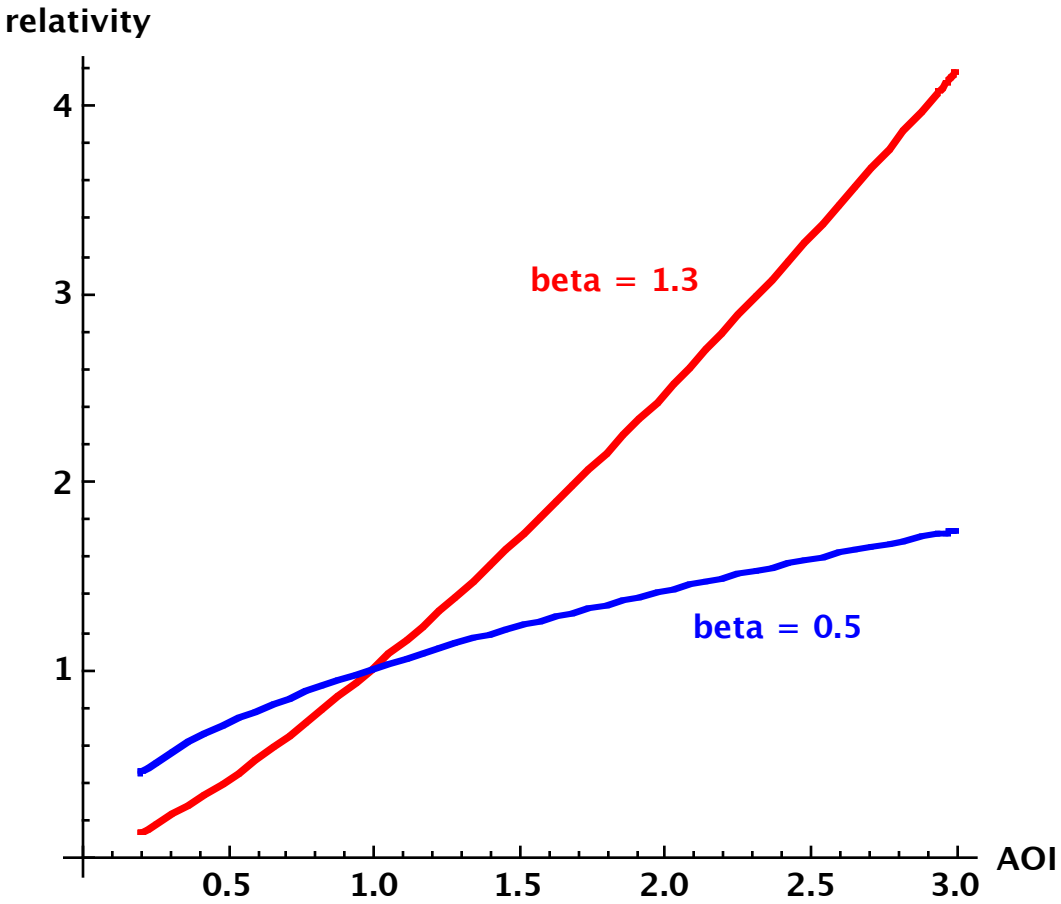
Assume x_1 is a continuous variable such as amount of insurance.[98] [99]
For example, if $\beta_1 = 0.5$, then $\exp[\beta_1]^{x_1} = 1.649^{AOI}$.
If instead $\beta_1 = 1.1$, then $\exp[\beta_1]^{x_1} = 3.004^{AOI}$.
Both of these curves have the same form, exponential growth: c^x, where c is some constant.

What if instead of using x_1 as the predictor variable, we used $\ln[x_1]$?
$m = \exp[\beta_0 + \beta_1 \ln[x_1] + \beta_2 x_2] = \exp[\beta_0 + \beta_2 x_2] \, x_1^{\beta_1}$.
Now the multiplicative relativity for amount of insurance is AOI^{β_1}.

For example, if $\beta_1 = 0.5$, then the multiplicative relativity is $AOI^{0.5}$.
If instead $\beta_1 = 1.3$, then the multiplicative relativity is $AOI^{1.3}$.
These are significantly different behaviors.[100]

---

[97] See Section 2.4.1 of Goldburd, Khare, and Tevet.
[98] We have not grouped the variable into levels.
[99] Usually the model will be easier to interpret if for example we used AOI / $100,000.
While this will be easier to interpret, it produces a mathematically equivalent model to using AOI.
[100] See Figure 1 in Goldburd, Khare, and Tevet.



This variety of behaviors makes it more likely to find a model that fits the data.
⇒ **The authors recommend that when using the log link function in a GLM, you log your continuous predictor variables**.[101]

[101] "This allows the scale of the predictors to match the scale of the entity they are linearly predicting, which in the case of a log link is the log of the mean of the outcome."
This is an empirical question. There will be cases where not taking the log of a continuous predictor variable will result in a GLM that better fits the data; for example, this may be the case when the continuous predictor is year.

Logistic Regression:[102]

A variable can be categorical; there are a discrete number of categories, but the labels attached to them may have no significance. Variables can be binary; this is a special case of categorical variable with only two categories, which can be thought of as either 0 or 1. Examples include: whether a policyholder renews its policy, whether a newly opened claim will exceed $10,000, whether a newly opened claim will lead to a subrogation opportunity, whether a newly opened claim is fraudulent, etc.

When the response variable is binary we use the Bernoulli Distribution, the Binomial with m = 1. In that case, the probability of the event is μ and the probability of not having the event is $1-\mu$. The ratio **$\mu/(1-\mu)$ is called the odds**.

The most common link function to use in this case is the logit, the log of the odds:[103]
$g(m) = \ln[\mu/(1-\mu)]$. $\Leftrightarrow \mu = \exp[x'b] / \{1 + \exp[x'b]\}$.

One can group similar observations in which case one has a Binomial with parameters m_i and q_i,where m_i is the number of observations in the given group.

A GLM with the Bernoulli or Binomial Distribution using the logit link function is called a Logistic Regression.

Example of Logistic Regression:[104]

Fit a logistic regressions to data on whether or not a vehicle had a claim.
If x is the vehicle value in units of $10,000, the model is:
$\ln[\mu/(1-\mu)] = \beta_0 + \beta_1 x + \beta_2 x^2$, with $\hat{\beta}_0 = -2.893$, $\hat{\beta}_1 = 0.220$, $\hat{\beta}_2 = -0.026$.

For a vehicle worth $30,000, $x\beta = -2.893 + (0.220)(3) + (-0.026)(3^2) = -2.467$.
Thus the expected probability of a claim for a vehicle worth $30,000 is:
$e^{-2.467} / (1 + e^{-2.467}) = 7.8\%$.

Exercise: Determine the expected probability of a claim for a vehicle worth $70,000.
[Solution: $x\beta = -2.893 + (0.220)(7) + (-0.026)(7^2) = -2.627$. $e^{-2.627} / (1 + e^{-2.627}) = 6.7\%$.]

---

[102] See Section 2.8 of Goldburd, Khare, and Tevet.
[103] The logit is the canonical link function for the Binomial Distribution, including the special case the Bernoulli.
[104] See Section 7.3 of Generalized Linear Models for Insurance Data, by de Jong and Heller, not on the syllabus.

One can also fit a model, using instead a categorical version of vehicle value such as 6 groups: less than 25,000, 25K to 50K, 50K to 75K, 75K to 100K, 100K to 125K, more than 125,000.
With the first group as the base level, the fitted model had:
$\hat{\beta}_0 = -2.648$, $\hat{\beta}_1 = 0.174$, $\hat{\beta}_2 = 0.102$, $\hat{\beta}_3 = -0.571$, $\hat{\beta}_4 = -0.397$, $\hat{\beta}_5 = -0.818$.

Thus a vehicle of value less than $25,000 has an expected probability of a claim of:
exp[-2.648] / (1 + exp[-2.648]) = 6.61%.

A vehicle of value $25,000 to $50,000 has an expected probability of a claim of:
exp[-2.648 + 0.174] / (1 + exp[-2.648 + 0.174]) = 7.77%.

A vehicle of value greater than $125,000 has an expected probability of a claim of:
exp[-2.648 - 0.818] / (1 + exp[-2.648 - 0.818]) = 3.03%.

The odds for a vehicle of value less than $25,000, the base level is:
6.61%/(1 - 6.61%) = 0.0708 = exp[-2.648] = exp[β_0].

The odds for a vehicle of value 25,000 to $50,000 is:
7.77%/(1 - 7.77%) = 0.0842 = exp[-2.648]exp[0.174] = exp[β_0]exp[β_1].

Thus the odds for the second level are those for the first base level times exp[β_1]. The odds for the second level are higher than those for the base level by a factor of exp[0.174] = 1.190. The odds for the last level are lower than those for the base level by a factor of exp[-0.818] = 0.441.

Grouping Data:

When one has binary variables, one can group the data into the possible combinations.
For example, with vehicle insurance data using driver's age (6 groups), area (6 territories), vehicle body (13 types), and vehicle value (6 groups), there are (6)(6)(13)(6) = 2808 cells.
Only some of these cells contain data.

For example, assume that driver age group 1, Area A, Hatchback, of value less than $25,000 in value has 554 polices with 47 claims.
We would take this as a random draw from a Binomial with m = 554.
In general, for a cell with n_i policies, we would assume the number of claims follows B(n_i, q_i).

We get the same fitted parameters and standard errors using either individual or grouped data, although the test statistics will differ.

Correlation Among Predictors:[105]

When the correlation between two predictor variables is large (in absolute value), the GLM will be unstable. The standard errors of the corresponding coefficients can be large and small changes in the data can produce large changes in the coefficients.

For example, years of education of the father and mother are likely to be highly positively correlated.
Including both in a model may produce problems.[106]

Software may not catch the presence of highly correlated variables and try to fit the model anyway. Due to the extreme correlation, the model will be highly unstable; the fitting procedure may fail to converge, and even if the model run is successful the estimated coefficients will be nonsensical.

When you start with a very long list of possible predictors to use in a GLM, it is common for some pairs of predictors to be highly correlated. Thus one should check the correlations of pairs of proposed predictor variables with each other.

If potential problems are found, one can:
1. Remove one or more predictors from the model.[107]
2. Use techniques that combine predictors in order to reduce the dimension, such as Principal Component Analysis and Factor Analysis.[108]

"Determining accurate estimates of relativities in the presence of correlated rating variables is a primary strength of GLMs versus univariate analyses; unlike univariate methods, the GLM will be able to sort out each variable's unique effect on the outcome, as distinct from the effect of any other variable that may correlate with it, thereby ensuring that no information is double-counted."

---

[105] See Section 2.9 of Goldburd, Khare, and Tevet.
[106] One could instead include an average of these two variables.
[107] While simple, this may lead us to lose valuable information.
[108] You are not responsible for any details.
I discuss Principal Component Analysis in my section on the paper by Robertson.
When a set of variables are highly correlated, either positively or negatively, the first principal component or the first two principal components capture most of the variation in the original variables.
The first principle component is a linear combination of the original variables.

Multicollinearity:

Multicollinearity is a similar situation which also leads to potential problems.
Multicollinearity occurs when two or more predictors in a model are strongly predictive of another one of the predicator variables.[109]

As discussed, we are concerned when pairs of variables are highly correlated. However, even in situations where pairs of variables are not highly correlated, problems can occur when looking at three or more predictor variables in combination.

For example, an insurer uses among others the following policyholder characteristics: age, years of education, and income. The first two characteristics would help to predict the final characteristic. Depending on how close this relationship was for this insurer's data, this could create a problem with the output of a GLM due to multicollinearity.

A high degree of multicollinearity, usually leads to unreliable estimates of the parameters. The estimation equations are ill-conditioned.

A useful statistic for detecting multicollinearity is the variance inflation factor (VIF).
If one or more of the VIFs is large, that is an indication of multicollinearity.
A common statistical rule of thumb is that a VIF greater than 10 is considered high, indicating possible problems from multicollinearity.

You will not be asked to compute VIF.[110] [111] Most software packages give VIF as an output.

Aliasing:

Where two predictors are perfectly correlated, they are said to be aliased, and the GLM will not have a unique solution. Equivalently, aliasing can be defined as a linear dependency among the columns of the design matrix X.

Intrinsic aliasing is a linear dependency between covariates due to the definition.

For example, if you have only three territories, then knowing an insured is not in territory one or territory two, implies they are in territory three. Such intrinsic aliasing is common with categorical variables; every insured must be in one and only one of the categories.

---

[109] Let X be the design matrix and X' be its transpose. In the case of regression, this is often described as X'X being an ill-conditioned matrix; one can also say the data is ill-conditioned.
In this case, the determinant of X'X will be very small.

[110] "The VIF for any predictor is a measure of how much the (squared) standard error for the predictor is increased due to the presence of collinearity with other predictors. It is determined by running a linear model for each of the predictors using all the other predictors as inputs, and measuring the predictive power of those models."

[111] In the case of regression, regress the i^{th} independent variable against all of the other independent variables, and let R_i^2 be the coefficient of determination of this regression.

Then the Variance Inflation Factor is: $VIF_i = 1/(1 - R_i^2)$.

Initially we have three covariates for the three territories and corresponding coefficients: β_1, β_2, and β_3. Ignoring any other factors, the linear predictor is: $\eta = X_1 \beta_1 + X_2 \beta_2 + X_3 \beta_3$.
However, $X_1 + X_2 + X_3 = 1$, so we can eliminate any one of three variables from the model.
For example, $\eta = X_1 \beta_1 + X_2 \beta_2 + (1 - X_1 - X_2) \beta_3 = X_1 (\beta_1 - \beta_3) + X_2 (\beta_2 - \beta_3) + \beta_3$.
Thus one can eliminate X_3 from the model, and include an intercept term if it does not already exist.

The fitted values will be the same regardless of which level is eliminated.
Selecting as the base level for each factor the one with the most exposure is helpful, since this minimizes the standard errors associated with other parameter estimates.

Exercise: Age of driver has only three levels: Youth, Adult, and Senior.
Demonstrate how aliasing can be used to exclude a level from the age variable.
[Solution: We have that $1 = X_{youth} + X_{adult} + X_{senior}$, and thus $X_{adult} = 1 - X_{youth} - X_{senior}$.
Therefore, we7 can eliminate β_{adult} from the model and include an intercept term if it does not already exist.
Comment: One could have eliminated any of the levels.
The adult level, which has the most exposures, would be a good choice for a base level.
The intercept term would now corresponds to the adult base level; there is no separate parameter for adult.
We would still have a parameter for Youth and a parameter for Senior.]

In general, **when we have a categorical variable with N levels, the model should have N-1 parameters in addition to an intercept term**. The chosen base level, which is often the one with the most exposures, is associated with the intercept term and will not have a separate associated parameter.

As another example of intrinsic aliasing, age of vehicle would alias with model year, since if you know one you can determine the other.

Extrinsic aliasing is a linear dependency between covariates that arises due to the particular values in the observed data rather than inherent properties of the covariates themselves.[112]

For example, if all sports cars in a data base just happen to be red cars and vice-versa.

Most software will detect aliasing and automatically drop one of those predictors from the model.

[112] Goldburd, Khare, and Tevet, do not distinguish between intrinsic and extrinsic aliasing.

Limitations of GLMs:[113]

1. GLMs assign full credibility to the data.[114]
2. GLMs assume that the randomness of outcomes are uncorrelated.[115]

As has been discussed on the exam on Basic Ratemaking, when estimating classification relativities by older techniques, an actuary uses credibility. The estimated relativities of classes with less data are given less than full weight.

However, using GLMs the estimated relativities are given full weight.

In fact, for a GLM with just one categorical predictor variable, the estimates will just be the observed average for each level. An actuary would not use the observed average for a small class (or the ratio of its observed average to the observed average for the base level) as a reasonable estimate of the future.

It should be noted, that for a class with little data, the standard errors of the fitted coefficient will be large. Thus we may not reject a value of zero for the coefficient of that small class. In a multiplicative model this would imply a relativity of one. Alternately, we could combine the small class with another class. However, neither of these alternatives is as flexible as giving the observed relativity for this small class some positive weight less than one.

In a regression, we assume that the random components, in other words the errors, ε_i, are uncorrelated.[116] Similarly, in a GLM we assume that the random components are uncorrelated.[117] [118]
This assumption can be violated.

For example, the data set may include several years of data from a single policyholder, which appear as separate records. The outcomes of a single policyholder are correlated.
Another example, in the case of wind losses, the outcomes for policyholders in the same area will be correlated.[119]

If there are large correlations of random components, then the GLM would pick up too much random noise, and produce sub-optimal predictions and overoptimistic measures of statistical significance.

---

[113] See Section 2.10 of Goldburd, Khare, and Tevet.
[114] Section 9 of Goldburd, Khare, and Tevet, not on the syllabus, discusses two ways to incorporate something similar to credibility: generalized linear mixed models and elastic net GLMs.
[115] Goldburd, Khare, and Tevet, mention two methods that account for correlation in the data: generalized linear mixed model, and generalized estimating equations.
[116] This assumption is often violated when dealing with time series.
[117] We assume that the systematic components are correlated.
For example, drivers in the same class and territory are assumed to have similar expected pure premiums.
[118] The random component is the portion of the outcomes driven by causes not in our model.
[119] I am thinking about wind losses from other than catastrophes; catastrophes would not be modeled using GLMs.

The Model-Building Process:[120]

The authors discuss how actuaries build models; much of the material is not specific to GLMs. They give a list of steps or components:[121]
• Setting of objectives and goals
• Communicating with key stakeholders
• Collecting and processing the necessary data for the analysis
• Conducting exploratory data analysis
• Specifying the form of the predictive model
• Evaluating the model output
• Validating the model
• Translating the model results into a product
• Maintaining the model
• Rebuilding the model

Setting Goals and Objectives:

• Determine the goals.
• Determine appropriate data to collect.
• Determine the time frame.
• What are key risks and how can they be mitigated?
• Who will work on the project; do they have the necessary knowledge and expertise?

Communication with Key Stakeholders:

• Legal and regulatory compliance
• Information Technology (IT) Department
• Underwriters
• Agents

Collecting and Processing Data:[122]

• Time-consuming.
• Data is messy.
• Often an iterative process.
• **The data should also be split into at least two subsets, so that the model can be tested on data that was not used to build it**.
• Formulate a strategy for validating the model.

Any analysis performed by an actuary is no better than the quality of the data that goes into that analysis![123]

---

[120] See Section 3 of Goldburd, Khare, and Tevet.
[121] As always, such lists are somewhat arbitrary. Many actuaries do not require such lists to do their jobs. Another possible step is to read the literature to see what has been done in similar situations in the past.
[122] For more detail, see Section 4 of Goldburd, Khare, and Tevet.
[123] Garbage in, garbage out.

Conducting Exploratory Data Analysis (EDA):

Spend some time to better understand the nature of the data and the relationships between the target and explanatory variables. Helpful EDA plots include:

- Plotting each response variable versus the target variable to see what (if any) relationship exists. For continuous variables, such plots may help inform decisions on variable transformations.
- Plotting continuous response variables versus each other, to see the correlation between them.[124]

Specifying Model Form:[125]

- What type of predictive model works best?
- What is the target variable, and which response variables should be included?
- Should transformations be applied to the target variable or to any of the response variables?
- Which link function should be used?

Evaluating Model Output:[126]

- Assessing the overall fit of the model.
- Identifying areas in which the model fit can be improved.
- Analyzing the significance of each predictor variable, and removing or transforming variables accordingly.
- Comparing the lift of a newly constructed model over the existing model or rating structure.

Model Validation:[127]

- Assessing fit with plots of actual vs. predicted on holdout data.
- Measuring lift.
- For Logistic Regression, use Receiver Operating Characteristic (ROC) Curves.

Translating the Model into a Product:

For GLMs, often the desired result is a rating plan.

- The product should be clear and understandable.
- Are there other rating factors included in the rating plan that were not part of the GLM? Then it is important to understand the potential relationship between these additional variable(s) and other variables that were included in the model. Judgmental adjustments may be needed.

---

[124] Recall that a high correlation, either positive or negative, between pairs of predictor variables may lead to problems with the fitted GLM.

[125] For more detail, see Section 5 of Goldburd, Khare, and Tevet.

[126] For more detail, see Section 6 and 7 of Goldburd, Khare, and Tevet.

[127] For more detail, see Section 7 of Goldburd, Khare, and Tevet.

Maintaining and Rebuilding the Model:

Models should be periodically rebuilt in order to maximize their predictive accuracy, but in the interim it may be beneficial to merely refresh the existing model using newer data. In other words, more frequently one would update the classification relativities without updating the rating algorithm or classification definitions. Less frequently, one would do a more complete update, investigating changing the classification definitions, the predictor variables used, and/or the rating algorithm.

In a somewhat different context, perhaps every 2 years one would update ELPPFs using the latest data but the existing grouping of classifications into hazard groups. Perhaps every 10 or 15 years one would update the grouping of classifications into hazard groups.[128]

Data Preparation and Considerations:[129]

Much of this is not unique to GLMs.
Data preparation is time consuming.[130]
Correcting one data error might help you discover another.

- Combining Policy and Claim Data.
- Modifying the Data.
- Splitting the Data.

Ratemaking Data:

Data is used by actuaries for many purposes including ratemaking.
For classification and territory ratemaking, more detailed data on exposures, premiums, losses, and ALAE is used, broken down by class and territory.
Ratemaking data is usually aggregated into calendar years, accidents years, and/or policy years.

---

[128] See the syllabus reading by Robertson.
[129] See Section 4 of Goldburd, Khare, and Tevet.
[130] At a large insurer, much of this work would have been routinely done by someone other than the actuary working on a specific GLM project. The actuary is responsible for determining whether it is reasonable to rely on the data supplied by others. See for example, Actuarial Standard of Practice 23 on Data Quality, not on the syllabus.

Combining Policy and Claim Data:

An insurer's data is often contained in a policy data base with exposures and premiums, and a separate claims data base with losses and alae.[131] These data bases have to be combined in a manner useful to the actuary.

Issues discussed by the authors:
- Are there timing considerations with respect to the way these databases are updated that might render some of the data unusable?
- Is there a unique key that can be used to match the two databases to each other in such a way that each claim record has exactly one matching policy record?
- What level of detail should the data sets be aggregated to before merging?
- Are there fields that can be safely discarded?
- Are there fields that should be in the database but aren't?[132]

Finding and Correcting Errors in the Data:[133]

Any dataset of sufficient size is likely to have errors.

- Check for duplicate records.
- Check categorical fields against available documentation.
- Check numerical fields for unreasonable values.[134]
- Decide how to handle each error or missing value that is discovered.

---

[131] See for example Chapter 3 of "Basic Ratemaking" by Werner and Modlin, on the syllabus of Exam 5.
[132] In which case, the actuary may initiate the process to start collecting this additional information. There are many pieces of information currently collected by insurers and rating bureaus that were not collected 50 years ago.
[133] When I worked at a rating bureau, a good percentage of my time was spent on this. We developed many systematic ways to detect errors. More than one group of people would be looking at the data from somewhat different points of view. Large errors were easy to find, but smaller errors required more diligence to find. Unfortunately, one can never find all of the errors.
[134] For example, an insurer reported to the rating bureau that an employer had as much payroll as the entire state. This error was quickly spotted and when pointed out to the insurer was quickly corrected.

Splitting the Data into Subsets:[135]

For modeling purposes one should split the data into either two or three parts.
This can be done either at random or based on time for example policy year.

The simpler approach is to **split the data into a training set and test (holdout) set.**[136]
For example, the training set could be 2/3 of the data while the test set is the remaining 1/3.

One develops the model on the training set. Then once one has come up with a final model or a few candidates for a final model, **one would test performance on the test set of data, which was not used in developing the model**.[137]

The model was developed to fit well to the training set. In doing so, we are concerned that the model may be picking up peculiarities of the training set. If the model does a good job of predicting for the test set, which was not used in developing the model, then it is likely to also work well at predicting the future.[138]

Reasons to split the data into a training set and a test set:
• Attempting to test the performance of any model on the same set of data on which the model was built will produce overoptimistic results. The model-fitting process optimizes the parameters to best fit the data used to train it. Using the training data to compare our model to a model built on different data would give our model an unfair advantage.
• As we increase the complexity of the model, the fit to the training data will always get better. Thus the performance on the training data can not be used to compare models of different complexity. On the other hand, for data the model fitting process has not seen, eventually increased complexity will worsen the performance of the model.[139] Thus the performance on the test data can be used to compare models of different complexity.

The split of data can be performed either by randomly allocating records between the training and test sets, or by splitting on the basis of a time variable.[140] The latter approach has the advantage in that the model validation is performed “out of time” as well as out of sample, giving us a more accurate view into how the model will perform on unseen years.

---

[135] See Section 4.3 of Goldburd, Khare, and Tevet.
[136] This is done in the syllabus reading by Couret and Venter.
[137] Such testing will be discussed subsequently.
[138] We are interested in how the GLM will perform at predicting the response variable on some future set of data rather than on the set of past data with which we are currently working.
[139] See Figure 7 in Goldburd, Khare, and Tevet.
[140] One could split by month or by calendar/accident year.
As in Couret and Venter one could select either the even or odd years of data as the training set and the other as the test set, in order to be neutral with respect to trend and maturity.

"Out-of-time validation is especially important when modeling perils driven by common events that affect multiple policyholders at once. An example of this is the wind peril, for which a single storm will cause many incurred losses in the same area. If random sampling is used for the split, losses related to the same event will be present in both sets of data, and so the test set will not be true unseen data, since the model has already seen those events in the training set. This will result in overoptimistic validation results. Choosing a test set that covers different time periods than the training set will minimize such overlap and allow for better measures of how the model will perform on the completely unknown future."[141]

The actuary should wait as long as possible in the process to use the test set. Once you start comparing to the test set, if you go back and change the form of the model, the usefulness of the test set for further comparisons has been diminished.

Thus sometimes, one uses the more complicated approach of splitting the data in three subsets: **a training set, validation set, and test (holdout) set**.[142]
For example, the split might be 40%, 30%, 30%.

As before, one develops the model on the training set. Then once one has come up with a good model or several good models, one would test performance on the validation set of data, which was not used in developing the model(s). If any changes in the form of the model are indicated, one goes back and works again with the training set. This iteration continues until the actuary is satisfied.

Then one would test performance on the test set of data, which was not used so far.

In either the simpler or more complicated case, **once a final form of the model has been decided upon, one should go back and use all of the available data to fit the parameters of the GLM**.

---

[141] Quoting from Section 4.3.1 of Goldburd, Khare, and Tevet.
See 8 11/17, Q.4c.
[142] Hopefully the total amount of data available is big enough to allow this.

Underfitting and Overfitting:

A model may be either overfit or underfit. Think of fitting a polynomial to 20 points. A straight line with no intercept, in other words a model with one parameter, will probably not do a good job of fitting the points. A fitted 19^{th} degree polynomial, in other words a model with 20 parameters, will pass through all of the points.

However, actuaries are using a model to predict the behavior in the future. The one parameter model will probably not do a very good job, since it ignored some of the information in the data. It is underfit. The 20 parameter model will not do a good job of predicting, since it picked up all of the random fluctuation (noise) in the data. It is overfit.

A model should be made as simple as possible, but not simpler.

Underfit. ⇔ Too few Parameters. ⇔ Does not use enough of the useful information.
⇔ Does not capture enough of the signal.

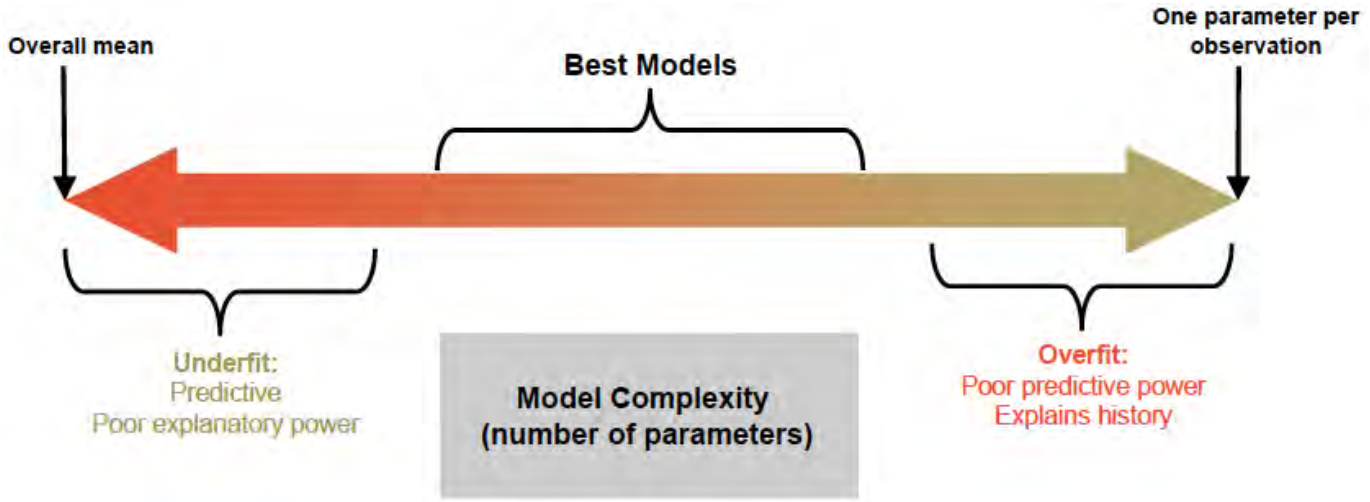
Overfit. ⇔ Too many Parameters. ⇔ Reflects too much of the noise.

We wish to avoid both underfitting and overfitting a model.

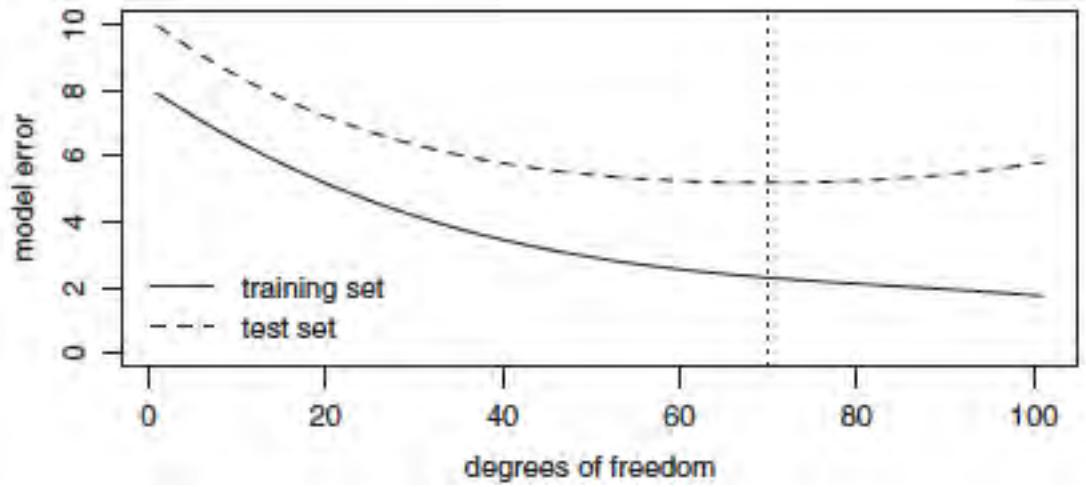
Think of fitting loss distributions. We would not use the most complicated model possible.[143]
We would only add parameters to the extent they were statistically significant.[144]
In a particular situation, it might be that an Exponential Distribution (one parameter) is an underfit model, a Transformed Gamma Distribution (3 parameters) is an overfit model, while a Gamma Distribution (2 parameters) is just right.

[143] Recall that a mixture of two or more distributions can have a lot of parameters.
[144] Think of the Likelihood Ratio Test or the Schwarz Bayesian Criterion.

In order to produce a sensible model that explains recent historical experience and is likely to be predictive of future experience, one needs to avoid both too little and too much complexity:[145]



Each added parameter adds a degree of freedom to the model. This can be due to the addition of a new predictor variable, the addition of a polynomial term, the addition of an interaction term, etc.
Each added degree of freedom makes the model more complex.
Our goal in modeling is to find the right balance where we pick up as much of the signal as possible with minimal noise. This is illustrated in Figure 7 of the syllabus reading:



As we add more parameters, we get a model that fits the training set better. However, when we compare such a model fitted to the training data to the test data, there is a point past which added parameters reduce the fit to the test data. The right balance is indicated by the vertical dotted line, at about 70 degrees of freedom in this case.[146]

---

[145] Taken from "GLM II, Basic Modeling Strategy", presented by Lenard Shuichi Llaguno, FCAS, at the 2012 CAS Ratemaking and Product Management Seminar.
[146] Here the authors use degrees of freedom to refer to the number of parameters in the fitted model. In for example the F-test, many authors instead define the degrees of freedom as number of observations minus number of fitted parameters for the fitted model.

Cross Validation:[147] [148]

Cross Validation is another technique for data splitting, although it is often of limited usefulness for actuarial work.

Split the data into for example 10 groups. Each group is called a fold. For each fold:
• Train the model using the other folds.[149]
• Test the model using the given fold.

Cross validation has the advantage of using all of the data (at some point) to estimate the mean squared error, rather than only using the portion of the data in the holdout set to do so. Thus cross validation should produce a better estimate of the MSE.

In the case of 10-fold cross validation, fit model form A on the data for the first 9 folds. Then compute the mean squared error (MSE) of this fitted model used to make predictions to the data in the remaining tenth fold.

Now fit model form A on the data for the folds other than the ninth. Then compute the mean squared error (MSE) of this fitted model used to make predictions to the data in the remaining ninth fold.

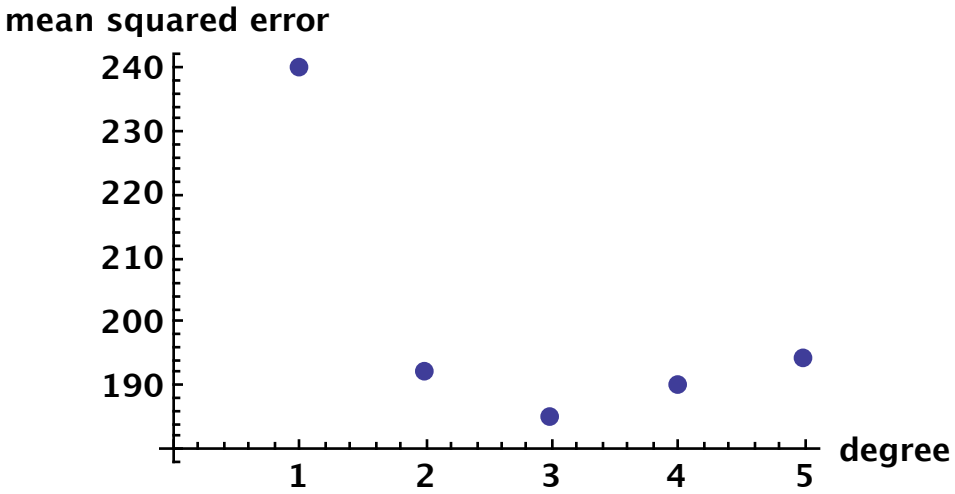
We would continue in this manner and then average these ten mean squared errors. This would be the estimated test MSE for model form A. We could then determine the MSE of several other model forms in a similar manner.[150] The form of model with the lowest test MSE would be best.

For example, we might compare polynomial models with different number of powers of a predictor variable.

---

[147] See Section 4.3.4 of Goldburd, Khare, and Tevet.
[148] See also, An Introduction to Statistical Learning with Applications in R, by James, Witten, Hastie, and Tibshirani, not on the syllabus of this exam. They also discuss how to apply cross validation to other modeling techniques such as ridge regression and the lasso.
[149] According to the authors, this training procedure has to include all of the steps of the model building, including the variable selection and transformation; these steps usually include significant amounts of actuarial judgement.
[150] For example, Model A and Model B might use different sets of predictor variables.

One has fit similar GLMs on a set of data, where one of the predictors enters using polynomials of different degrees. The test MSEs were estimated using ten-fold cross-validation:



The model using the third degree model seems to perform best.[151]

Limitations of Cross-Validation for Actuarial Work:

Cross-validation can be useful for deciding how many polynomial terms to include.[152] However, cross validation is often of limited usefulness for most insurance modeling applications.

The actuary usually applies a great deal of care and judgment in selecting the variables to be included in the model. If using cross validation, this actuarial judgement should be applied separately to each of the data sets created by leaving out one fold. This is not really practical. Thus, using cross validation in place of a holdout set is only really appropriate where a purely automated variable selection process is used.[153]

For most actuarial modeling, the use of a holdout set is preferred to the use of cross validation. The final model valuation should always be done using a distinct set of data held out until the end.

---

[151] Due to the data being assigned to the 10 folds at random, if one performed cross validation again, one would get somewhat different estimates of the test MSEs. Therefore, in practical applications one would perform cross validation several times and compare the results.

[152] This is an example of evaluating a "tuning parameter" of the model.

[153] This is the opinion of the authors of the syllabus reading, who have plenty of experience using GLMs for actuarial modeling.

An Example of k-Fold Cross-Validation:[154]

Eight observations of three independent variables and one dependent variable:

X_1	X_2	X_3	Y
-2	1	-4	6
1	-1	0	8
3	4	4	33
6	-4	8	14
11	0	12	40
15	8	16	118
17	-8	20	2
20	-6	24	61

I will perform 4-fold cross-validation, so that each fold contains 8/4 = 2 observations.
We need to divide the original data into 4 random subsets; the estimated test MSE will depend to some extent on this random subdivision. My four folds will be: (1, 7), (2, 4), (3, 5), (6, 8).

If we leave out the first and seventh observations, and fit a regression model to the remaining six observations, the fitted parameters are:

$\hat{\beta}_0 = 3.78881$, $\hat{\beta}_1 = 5.10444$, $\hat{\beta}_2 = 5.17811$, $\hat{\beta}_3 = -0.621247$.

We now plug into this fitted model the values of the predictors for the first observation:
(5.10444)(-2) + (5.17811)(1) + (-0.621247)(-4) = 1.24303.
We now plug into this fitted model the values of the predictors for the seventh observation:

$\hat{Y} = 3.78881 + (5.10444)(17) + (5.17811)(-8) + (-0.621247)(20) = 36.7145$.
The mean squared difference between the observed values and these predicted values is:
$MSE_1 = \{(6 - 1.24303)^2 + (36.7145 - 2)^2\} / 2 = 613.863$.

Similarly, we would now instead leave out the 2nd and 4th observations.
We continue in this manner, and the four mean squared errors are:
613.863, 231.863, 697.906, 1458.9.
The average of these four values is the 4-fold cross-validation estimate of the test MSE:
750.633.

I used R to perform this same process five separate times and the estimated test MSEs were:[155]
449.2197, 1249.365, 616.1268, 680.8828, 754.928.
With only 8 observations, we see considerable variation in these estimates.

[154] Solely in order to give a simple concrete example; you are not responsible for any details.
[155] Using the R function cv.glm. Each time a different set of random folds is used.

Selection of Model Form:[156]

"Selecting the form of a predictive model is an iterative process, and is often more of an art than a science."

Important decisions on the form of a GLM include:
• Choosing the target variable.
• Choosing a distribution for the target variable.
• Choosing the predictor variables.
• Whether to apply transformations to the predictor variables.
• Grouping categorical variables.
• Whether to include interactions.

[156] See Section 5 of Goldburd, Khare, and Tevet.

Frequency/Severity versus Pure Premium:[157]

An actuary could build two separate models: one for frequency and one for severity.[158]
Alternately the actuary could build a single model for pure premium. If there is time, an actuary could do both and compare the results.

Advantages of the frequency/severity approach over pure premium modeling:
• Provides the actuary with more insight.
• Each of frequency and severity is more stable than pure premium.[159]

Disadvantages of pure premium modeling versus the frequency/severity approach:
• Some interesting effects may go unnoticed.
• Pure premium modeling can lead to underfitting or overfitting.
• The Tweedie distribution used to model pure premium contains the implicit assumption that an increase in pure premiums is made up of an increase in both frequency and severity.[160]

For example, urban driving tends to lead to a higher frequency of accidents (per mile driven) than rural driving. However, urban driving tends to lead to a lower severity of accidents than rural driving.
These two separate effects could be masked in a pure premium model. In any case, with just a pure premium model, the actuary would not get this interesting and perhaps important insight.

While territory would show up as significant in a frequency model, when testing it in a pure premium model the high variance in severity may overwhelm this effect, rendering the territory statistically insignificant.[161] Thus, a useful predictive variable will be excluded from the model, leading to underfitting.

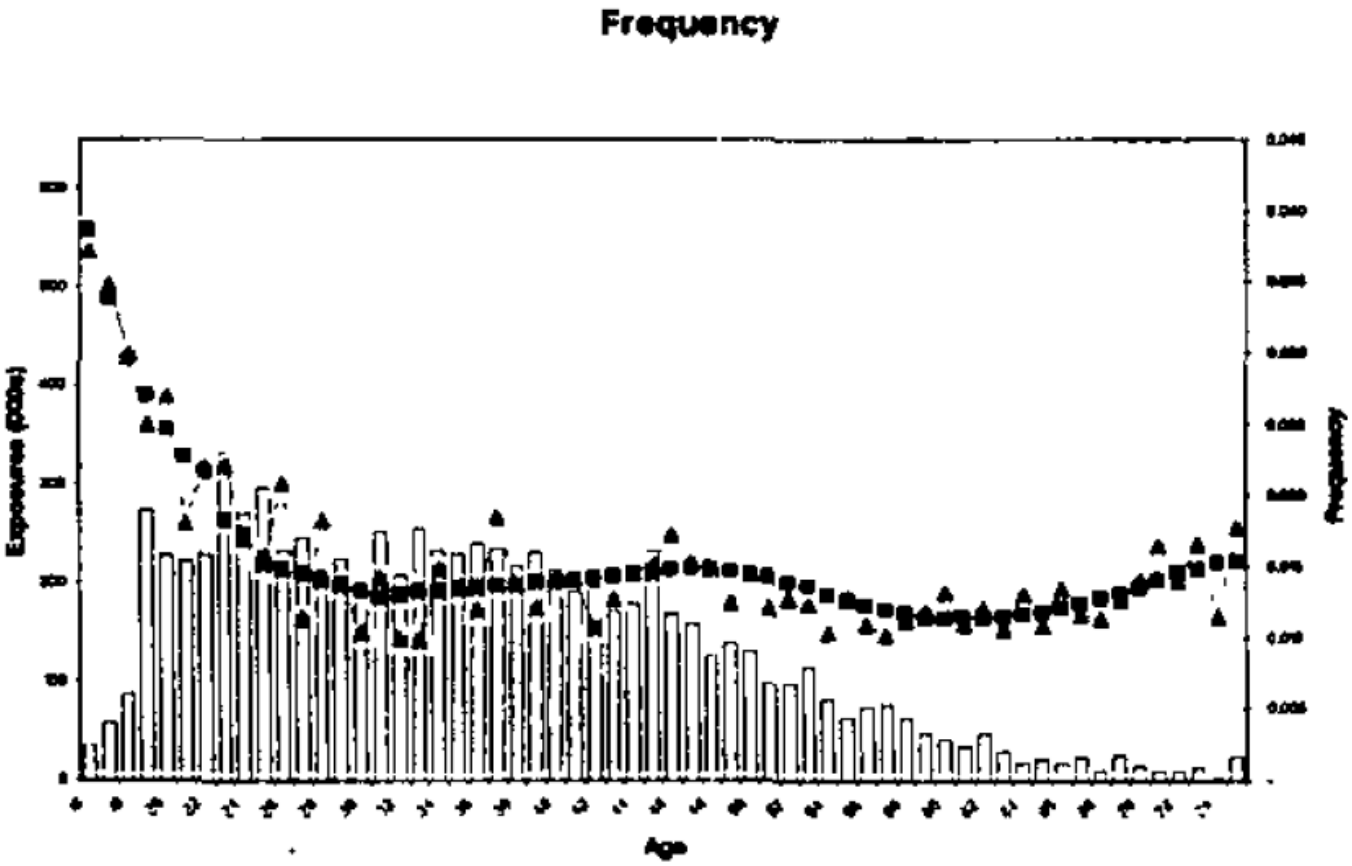
Assume that a predictor variable has a significant effect on frequency and no effect on severity. If that variable is included in a pure premium model, then the fitted GLM will pick up any effect of severity in the training data even if it is just noise. The corresponding parameter will be overfit.

For frequency and severity, a priori expected patterns help the actuary to produce a better model. To the extent that the historical pattern is erratic, the actuary will be able to use appropriate techniques and knowledge about insurance to build a model that captures the signal in the data.

---

[157] See Section 5.1.1 of Goldburd, Khare, and Tevet.
[158] If the log link function is used for both, then the pure premium (multiplicative) relativities will be the product of the separate frequency and severity relativities.
[159] Recall that the standard for full credibility for pure premium is the sum of those for frequency and severity.
[160] The authors assume that one would use the Tweedie Distribution to model pure premiums.
[161] While this could happen in general, in the example I have chosen it is unlikely to do so.

For example, when modeling auto collision frequency, the actuary may expect the frequency by age to decrease from youthful to adult and increase again for the most mature drivers.[162] The following figure compares the historical frequencies (triangles) and modeled frequencies (squares) by age.[163]



The modeled frequencies follow the general pattern expected.

[162] These are frequencies per car year. Most senior citizens have higher expected frequencies per mile driven. However, their average number of miles driven per year is lower.

[163] Figure 5, from "GLM Basic Modeling: Avoiding Common Pitfalls," by Geoff Werner and Serhat Guven, CAS Forum Winter 2007, not on the syllabus.

Modeling Loss Ratios:

If the goal of the project is to identify deficiencies in the existing rating plan, loss ratio may be an appropriate target variable for the GLM.[164] However, there are disadvantages to modeling loss ratios rather than pure premiums or frequency/severity.

Theoretical and practical disadvantages to loss ratio modeling:[165]

- One needs to put premiums on-level at a granular level; difficult and time consuming.
  One has to put on the current rate level individual policies;
  overall on-level factors will not do.
- There is no generally accepted error distribution.[166]
- Difficult to distinguish noise from pattern, compared to modeling frequency/severity.
- If changes are made to the rates, then models cannot be reused from the last review.

[164] See Section 5.1 of Goldburd, Khare, and Tevet.
[165] Taken from "GLM II: Basic Modeling Strategy, " by Claudine Modlin,
CAS Predictive Modeling Seminar, October 2008.
[166] However, as discussed previously, one could use the Tweedie Distribution.

Policies with Multiple Coverages and Perils:[167]

A Businessowners package policy includes building, business personal property, and liability coverage.[168] Each of those coverages should be modeled separately.

In addition, one may models each peril individually.[169] For the Businessowners building model, one may wish to create separate models for: fire and lightning, wind and hail, and all other perils.[170]

One way to combine separate models by peril in order to get a model for all perils:
1. Use the separate models by peril to generate predictions of expected loss due to each peril for some set of exposure data.[171]
2. Add the peril predictions together to form a combined loss cost for each record.
3. Run a model on that data, using the combined loss cost calculated in Step 2 as the target, and the union of all the individual model predictors as the predictor variables.

Transforming the Target Variable:[172]

Sometimes it is useful to transform the target variable. Among the possible transformations:
- Cap large losses for purposes of modeling pure premium or severity.[173]
- Remove catastrophe losses.
- Losses may need to be developed.[174]
- Losses and/or exposures may need to be trended.
- Premium may need to be put on level.[175]

Year could be included in the model, which should pick up any effects on the target variable related to time, such as trend, loss development, and rate changes.

---

[167] See Section 5.1.2 of Goldburd, Khare, and Tevet.
[168] Similar ideas would apply to Homeowners Insurance.
[169] Or group of perils.
[170] Wind and hail should be divided between catastrophe and non-catastrophes; with catastrophes modeled separately as discussed in the syllabus reading by Grossi and Kunreuther.
[171] The data used for this procedure should reflect the expected mix going forward, and so using only the most recent year may be ideal. Since the target data fed into this new model is extremely stable, this procedure doesn't require a whole lot of data.
[172] See Section 5.1.3 of Goldburd, Khare, and Tevet, which discusses familiar things done in ratemaking.
[173] Ideally the level chosen for the cap should capture most of the signal and eliminate most of the noise. This is similar in concept to choosing a reasonable accident limit to use in an Experience Rating Plan.
[174] Either to ultimate or to a common level of maturity.
For a severity model, the development factor should reflect only expected future development on known claims. Since larger claims take on average longer to report, this may not address the whole issue.
For some lines of insurance, one may be better off not using more recent but less mature data in the model.
For a pure premium or loss ratio model, the development factor should include the effect of pure IBNR claims as well.
[175] Premium would be used in a loss ratio model.

Choosing the Distribution for the Target Variable:[176] [177]

If modeling claim frequency, the distribution is likely to be either Poisson or Negative Binomial.[178]

If modeling a binary response, then the Bernoulli or Binomial Distributions are used.

If modeling claim severity, common choices for the distribution are Gamma and Inverse Gaussian.

If modeling pure premiums, the Tweedie Distribution is a common choice.

Selection of Predictor Variables:[179]

Sometimes the actuary is just updating the parameters a model using newer data. Other times, the actuary will do a full review of all aspects of a model, including which predictor variables to include.

One would like a predictor variable to have a statistical significant effect on the target variable. Statistical tests can be performed. One would like a small probability value for the null hypothesis that the corresponding parameter is zero.

There is no magic cutoff, although a p-value of 5% or less is often used.[180] However, if the p-value is 5%, that means that there is 1/20 chance we are including a predictor variable in the model when we should not. If there is large set of possible predictor variables that are tested for inclusion in the model, this can lead to problems.[181]

In addition to statistical significance, the actuary must take into account practical considerations.[182] For example:
• Will it be cost effective?
• Actuarial standards of practice.
• Regulatory and legal requirements.
• Can the IT (Information Technology) department easily implement the change?

---

[176] See Section 5.2 of Goldburd, Khare, and Tevet.
[177] Analysis of the deviance residuals, to be discussed subsequently, can help the actuary to choose.
[178] Recall that one can also use an overdispersed Poisson.
[179] See Section 5.3 of Goldburd, Khare, and Tevet.
[180] For a further discussion of p-values see the following subsection, the ASA statement not on the syllabus.
[181] There are automated variable selection algorithms, which are not on the syllabus.
[182] See ASOP 12: Risk Classification.

ASA Statement on Statistical Significance and P-values:[183] [184]

Introduction

Increased quantification of scientific research and a proliferation of large, complex datasets in recent years have expanded the scope of applications of statistical methods. This has created new avenues for scientific progress, but it also brings concerns about conclusions drawn from research data. The validity of scientific conclusions, including their reproducibility, depends on more than the statistical methods themselves. Appropriately chosen techniques, properly conducted analyses and correct interpretation of statistical results also play a key role in ensuring that conclusions are sound and that uncertainty surrounding them is represented properly.

Underpinning many published scientific conclusions is the concept of "statistical significance," typically assessed with an index called the p-value. While the p-value can be a useful statistical measure, it is commonly misused and misinterpreted. This has led to some scientific journals discouraging the use of p-values, and some scientists and statisticians recommending their abandonment, with some arguments essentially unchanged since p-values were first introduced.

In this context, the American Statistical Association (ASA) believes that the scientific community could benefit from a formal statement clarifying several widely agreed upon principles underlying the proper use and interpretation of the p-value. The issues touched on here affect not only research, but research funding, journal practices, career advancement, scientific education, public policy, journalism, and law. This statement does not seek to resolve all the issues relating to sound statistical practice, nor to settle foundational controversies. Rather, the statement articulates in non-technical terms a few select principles that could improve the conduct or interpretation of quantitative science, according to widespread consensus in the statistical community.

What is a p-value?

Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.

---

[183] February 5, 2016.
Edited by Ronald L. Wasserstein, Executive Director
on behalf of the American Statistical Association Board of Directors
[184] Not on the syllabus.

Principles

1. P-values can indicate how incompatible the data are with a specified statistical model.

A p-value provides one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data. The most common context is a model, constructed under a set of assumptions, together with a so-called "null hypothesis." Often the null hypothesis postulates the absence of an effect, such as no difference between two groups, or the absence of a relationship between a factor and an outcome. The smaller the p-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the p-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.

2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

Researchers often wish to turn a p-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The p-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.

3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

Practices that reduce data analysis or scientific inference to mechanical "bright-line" rules (such as "$p < 0.05$") for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision-making. A conclusion does not immediately become "true" on one side of the divide and "false" on the other. Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis. Pragmatic considerations often require binary, "yes-no" decisions, but this does not mean that p-values alone can ensure that a decision is correct or incorrect. The widespread use of "statistical significance" (generally interpreted as "$p \leq 0.05$") as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.

4. Proper inference requires full reporting and transparency.

P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain p-values (typically those passing a significance threshold) renders the reported p-values essentially uninterpretable. Cherry-picking promising findings, also known by such terms as data dredging, significance chasing, significance questing, selective inference and "p-hacking," leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided. One need not formally carry out multiple statistical tests for this problem to arise: Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis. Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted and all p-values computed. Valid scientific conclusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including p-values) were selected for reporting.

5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

Statistical significance is not equivalent to scientific, human, or economic significance. Smaller p-values do not necessarily imply the presence of larger or more important effects, and larger p-values do not imply a lack of importance or even lack of effect. Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive p-values if the sample size is small or measurements are imprecise. Similarly, identical estimated effects will have different p-values if the precision of the estimates differs.

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Researchers should recognize that a p-value without context or other evidence provides limited information. For example, a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis. Likewise, a relatively large p-value does not imply evidence in favor of the null hypothesis; many other hypotheses may be equally or more consistent with the observed data. For these reasons, data analysis should not end with the calculation of a p-value when other approaches are appropriate and feasible.

Other approaches

In view of the prevalent misuses of and misconceptions concerning p-values, some statisticians prefer to supplement or even replace p-values with other approaches. These include methods that emphasize estimation over testing, such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors; and other approaches such as decision-theoretic modeling and false discovery rates. All these measures and approaches rely on further assumptions, but they may more directly address the size of an effect (and its associated uncertainty) or whether the hypothesis is correct.

Conclusion

Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning.

Transformation of Predictor Variables:[185]

In many cases, a variable will need to be transformed in some way such that the resulting GLM is a better fit to the data. We have already discussed how with a log link function it often make sense to take the log of a continuous variable. Partial Residual Plots are one way for the actuary to detect whether a transforming a predictor variable is indicated.

Partial Residual Plots:[186]

Concentrate on one of the explanatory variables X_j.

Then the partial residuals are: r_i = (ordinary residual) $g'(\mu_i) + x_{ij} \hat{\beta}_j$. [187]

In a Partial Residual Plot, we plot the partial residuals versus the variable of interest.

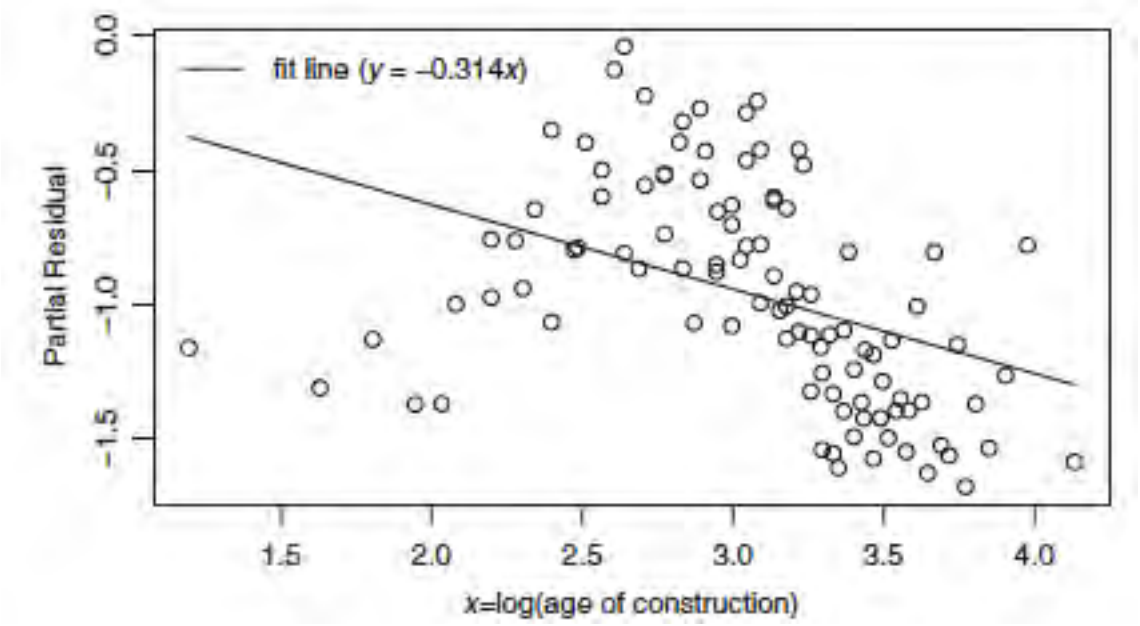
If there seems to be curvature rather than linearity in the plot, that would indicate a departure from linearity between the explanatory variable of interest and $g(\mu)$, adjusting for the effects of the other independent variables.

For a log link, $g'(\mu) = 1/\mu$, so that:

$$r_i = \frac{y_i - \mu_i}{\mu_i} + \hat{\beta}_j x_{ij}.$$

[185] See Section 5.4 of Goldburd, Khare, and Tevet.
[186] See Section 5.4.1 of Goldburd, Khare, and Tevet.
[187] For the identity link ratio, this matches what one would have for multiple regressions.

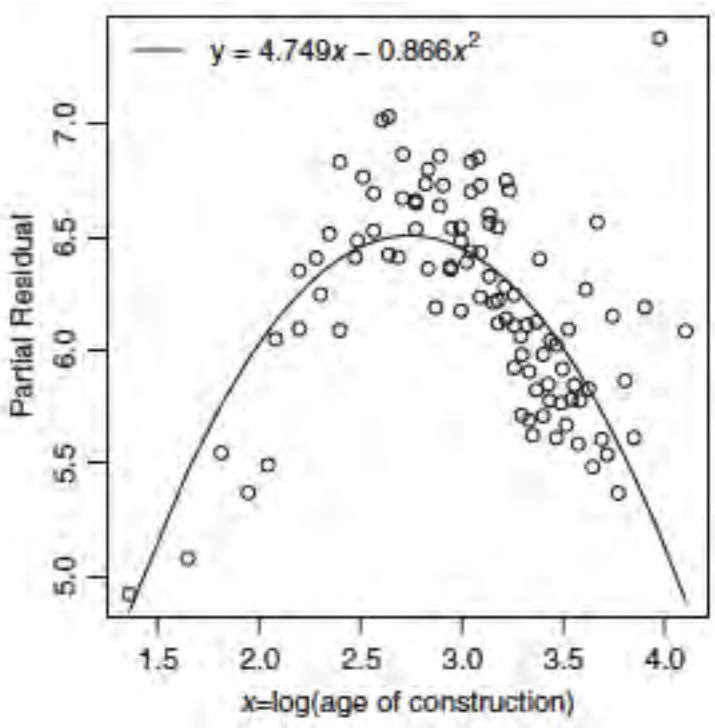
For example, assume a GLM where the fitted coefficient on ln[age of building] is -0.314.
Assume the following graph of the partial residuals:[188]



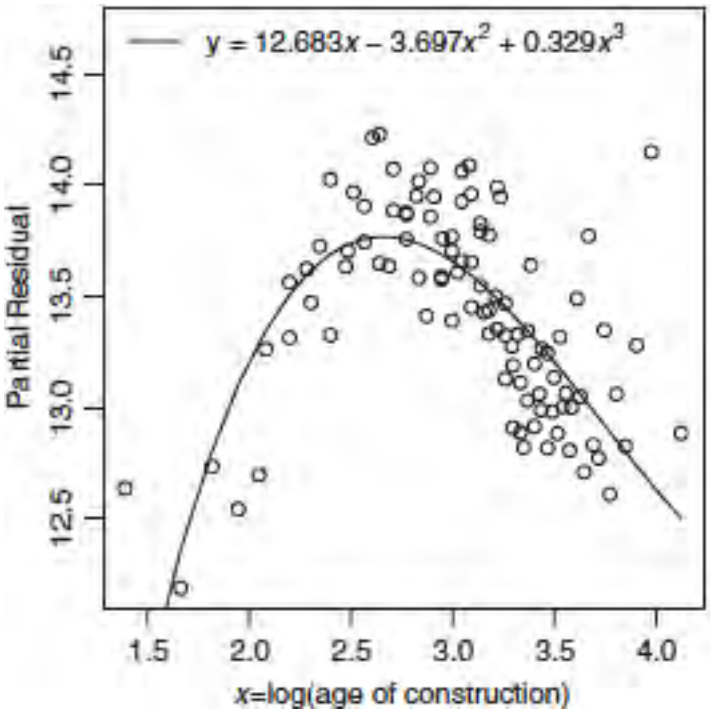
The linear estimate of the GLM, -0.314*x,* is superimposed over the plot of the partial residuals.
The points are missing the line in a systematic way, indicating that this model can be improved.
The model is overpredicting for risks where log building age is less than 2.5, underpredicts between 2.5 and 3.25, and once again overpredicts for older buildings.

A new GLM was fit, including both ln[age of building] and its square.
The following page has a graph of the partial residuals.[189]

[188] Figure 8 taken from Generalized Linear Models for Insurance Rating, by Goldburd, Khare and Tevet.
[189] Figure 10 taken from Generalized Linear Models for Insurance Rating, by Goldburd, Khare and Tevet.



We see that adding the square of the logged building age improves the model.
The following is a graph of the partial residuals when the cube is also added:



The model with the cube of the logged age of building seems to do even better.

A Cherry Tree Example:

We are given the height, diameter, and volume of 31 black cherry trees:[190]

Diameters are: 83, 86, 88, 105, 107, 108, 110, 110, 111, 112, 113, 114, 114, 117, 120, 129, 129, 133, 137, 138, 140, 142, 145, 160, 163, 173, 175, 179, 180, 180, 206.

Heights are: 70, 65, 63, 72, 81, 83, 66, 75, 80, 75, 79, 76, 76, 69, 75, 74, 85, 86, 71, 64, 78, 80, 74, 72, 77, 81, 82, 80, 80, 80, 87.

Volumes are: 103, 103, 102, 164, 188, 197, 156, 182, 226, 199, 242, 210, 214, 213, 191, 222, 338, 274, 257, 249, 345, 317, 363, 383, 426, 554, 557, 583, 515, 510, 770.

I took X_1 = ln[diameter], X_2 = ln[height], and Y = volume.
A GLM was fit using a Gamma Distribution and a log link function.

The fitted parameters were: $\hat{\beta}_0$ = -8.94859, $\hat{\beta}_1$ = 1.98041, $\hat{\beta}_2$ = 1.13288.

$\hat{y}$ = exp[-8.94859 + 1.9804 ln[diameter] + 1.13288 ln[height]]

= 0.00012992 diameter$^{1.9804}$ height$^{1.13288}$.

The covariance matrix is: $\begin{pmatrix} 0.556725 & 0.00760542 & -0.13715 \\ 0.00760542 & 0.00545975 & -0.00788943 \\ -0.13715 & -0.00788943 & 0.0405552 \end{pmatrix}$.

Exercise: Based on geometry, it would make sense for $\beta_1 = 2$. Test whether $\beta_1 = 2$.

[Solution: (1.98041 - 2) / $\sqrt{0.00545975}$ = -0.265. p-value is: 2 Φ[-0.265] = 79.1%.]

Exercise: Based on geometry, it would make sense for $\beta_2 = 1$. Test whether $\beta_2 = 1$.

[Solution: (1.13288 - 1) / $\sqrt{0.0405552}$ = 0.660. p-value is: 2 {1- Φ[0.660]} = 50.9%.]

---

[190] The diameter is measured at 4'6" above the ground.
Data from a study by Ryan, Joiner, and Ryan.

The first predicted volume is: exp[-8.94859 + 1.9804 ln[83] + 1.13288 ln[70]] = 101.04.
Thus the first residual is: 103 - 101.04 = 1.96.

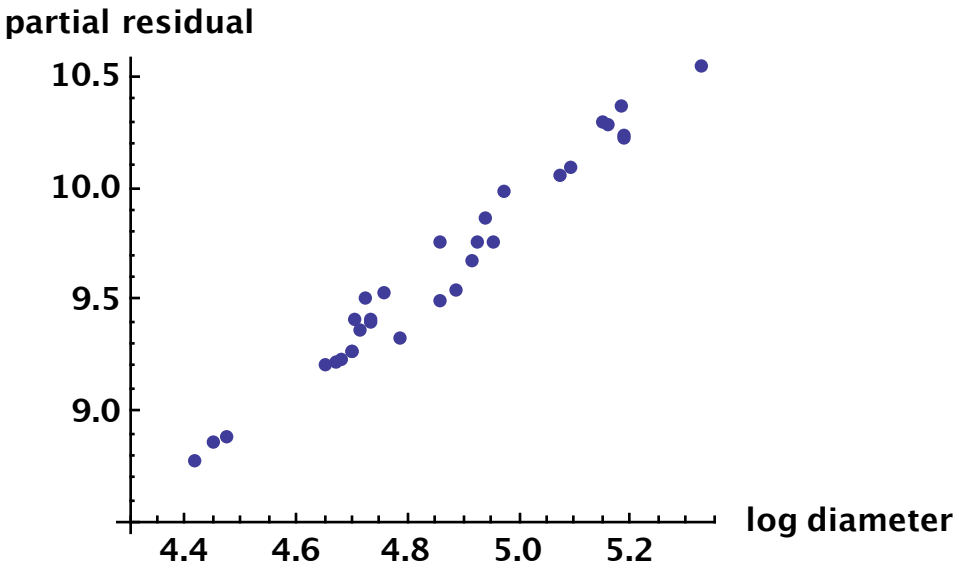
The residuals are: 1.96, 3.32, 1.31, -2.19, -9.14, -9.43, -9.12, -8.85, 16.96, 1.22, 28.50, 2.06, 6.06, 16.79, -35.74, -35.71, 36.48, -50.59, -20.02, -0.86, 23.33, -23.46, 38.15, 0.29, -2.43, 43.48, 27.42, 44.45, -29.52, -34.52, -12.21.

For this example, $g(\mu) = \ln(\mu)$. Thus $g'(\mu) = 1/\mu$. $\Rightarrow r_i = \frac{y_i - \mu_i}{\mu_i} + \hat{\beta}_j x_{ij}$.

Thus for ln[diameter], the partial residuals are: $(y_i - \hat{y}_i) / \hat{y}_i + \ln[(\text{diameter})_i]$ 1.98041.

The first partial residual is: (103 - 101.04)/101.04 + ln[83](1.980401) = 8.77.

The partial residuals for the ln[diameter] are: 8.77, 8.85, 8.88, 9.2, 9.21, 9.23, 9.25, 9.26, 9.41, 9.35, 9.50, 9.39, 9.41, 9.52, 9.32, 9.49, 9.75, 9.53, 9.67, 9.75, 9.86, 9.75, 9.97, 10.05, 10.08, 10.29, 10.28, 10.36, 10.23, 10.22, 10.54.
Here is a graph of these partial residuals versus ln[diameter]:



A departure from linearity is not evident.[191]

[191] If there seems to be curvature rather than linearity in the plot, that would indicate a departure from linearity between the independent variable of interest and g(μ), adjusting for the effects of the other independent variables.

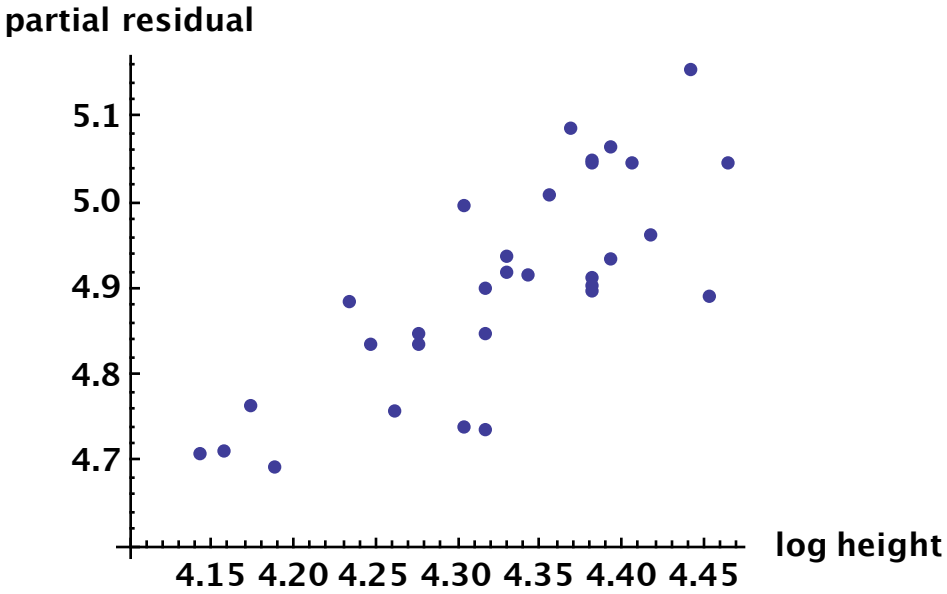
Exercise: For ln[height] what is the first partial residual?

[Solution: The partial residuals are: $(y_i - \hat{y}_i) / \hat{y}_i + \ln[(height)_i]$ 1.13288.

The first partial residual is: (103 - 101.04)/101.04 + ln[70] (1.13288) = 4.83.

The partial residuals for the ln[height] are: 4.83, 4.76, 4.71, 4.83, 4.93, 4.96, 4.69, 4.84, 5.05, 4.90, 5.08, 4.92, 4.94, 4.88, 4.73, 4.74, 5.15, 4.89, 4.76, 4.71, 5.01, 4.90, 4.99, 4.85, 4.92, 5.06, 5.04, 5.05, 4.91, 4.90, 5.04.

Here is a graph of these partial residuals versus ln[height]:



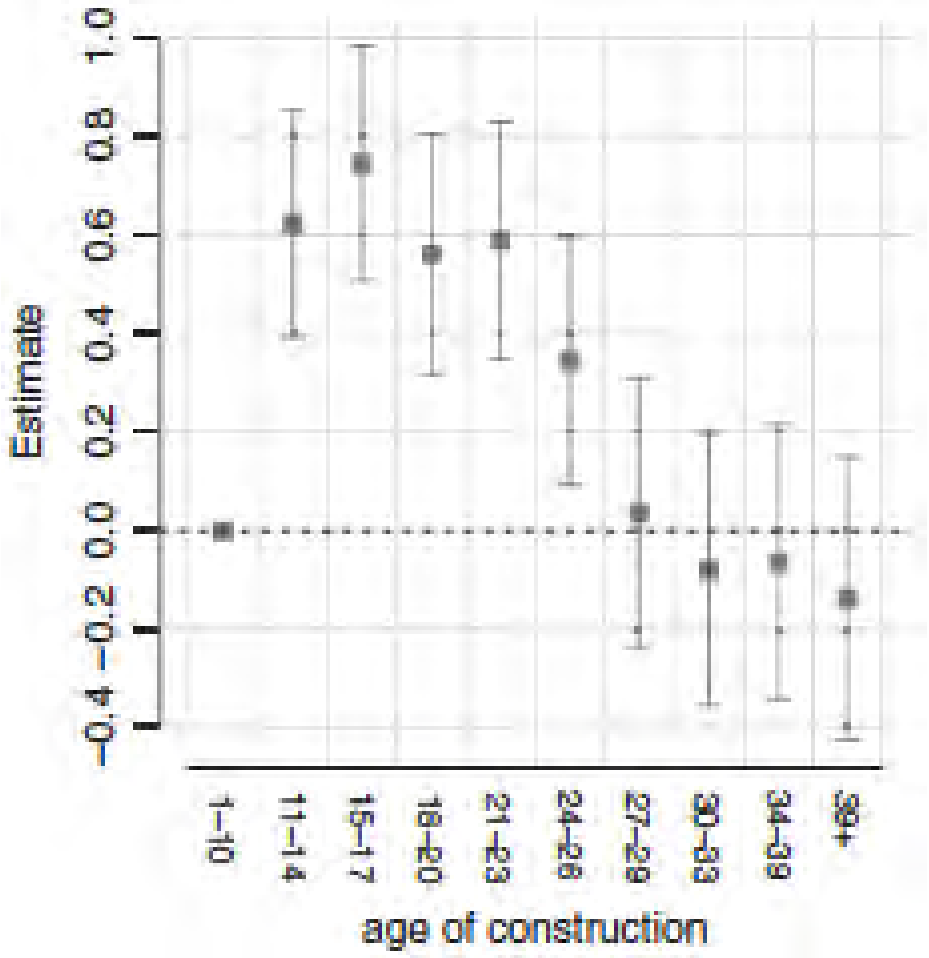
A departure from linearity is not evident.

Binning Continuous Predictors:[192]

If there is nonlinearity, one possible fix for a continuous variable is to group it into intervals.

For example, rather than treat age of construction as a continuous variable, one can group it into several categories. We have converted a continuous variable into a categorical variable.

For their example, the authors group age of construction into ten bins.[193]
Figure 9 in the syllabus reading shows the resulting model:



"The model picked up a shape similar to that seen in the points of the partial residual plot. Average severity rises for buildings older than ten years, reaching a peak at the 15-to-17 year range, then gradually declining."

---

[192] See Section 5.4.2 of Goldburd, Khare, and Tevet.
[193] The bins were chosen so that they each have roughly the same amount of data.
While having bins with roughly equal amounts of data has advantages, it is not a necessity.

Disadvantages of binning (grouping) continuous variables:

1. Adds parameters to the model.[194]
2. Continuity in the estimates is not guaranteed.
   There is no guarantee that the pattern among intervals makes sense.[195] [196]
3. Variation within intervals is ignored.
   For example, it may be that the relativity for age of construction less than 5 years may be significantly different than that for 6 to 10 years. However, if we use an interval consisting of less than 10, our model can not pick up any such difference.

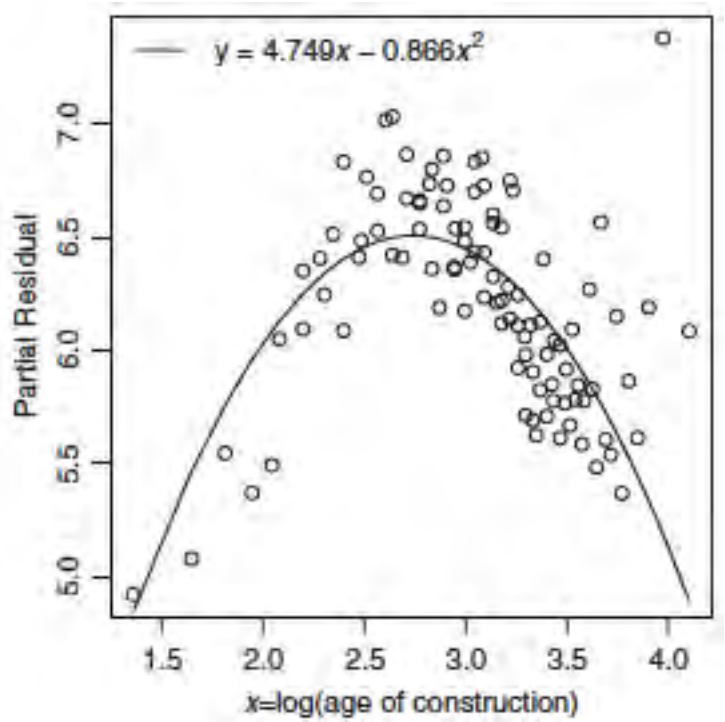
Adding Polynomial Terms:[197]

Rather than a model that uses $\beta_0 + \beta_1 x_1 + ...$,

one can use $\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + ...$, or $\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + ...$

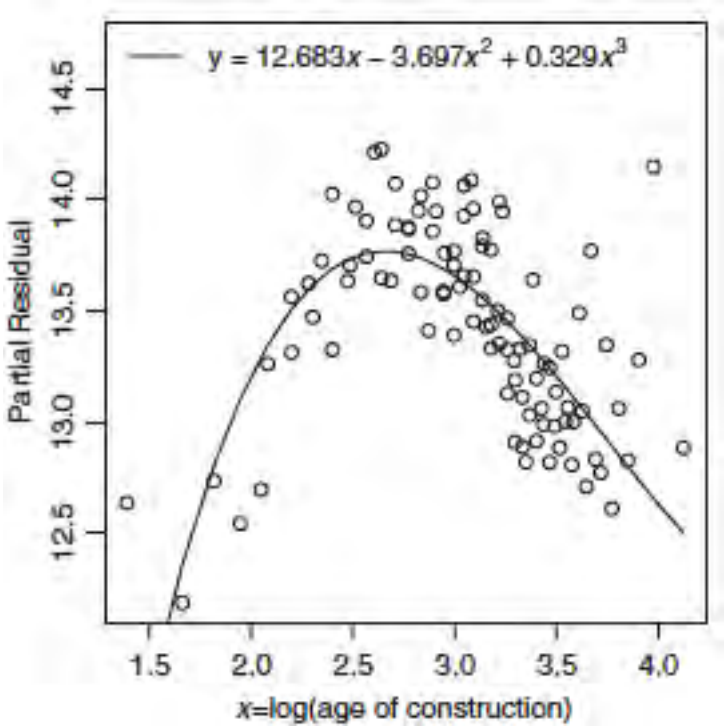
The more polynomial terms that are included, the more flexibility, at the cost of greater complexity.

The authors added the square of the logged building age to their model. Here is the resulting plot of partial residuals with the curve formed by both building age terms superimposed:[198]



[194] By the principle of parsimony, we wish to avoid adding unnecessary parameters to the model.
[195] For example, in the previous graph, the estimate for 21-23 years does not follow the general pattern.
[196] One may be able to alleviate this problem by applying some smoothing process to the estimates from the model. Alternately, one could group together two or more intervals.
[197] See Section 5.4.23 of Goldburd, Khare, and Tevet.
[198] See Figure 10 in Goldburd, Khare, and Tevet.

Then the authors added the cube of the logged building age to their model. Here is the resulting plot of partial residuals with the curve formed by both building age terms superimposed:[199]



"This perhaps yields a better fit, as the points seem to indicate that the declining severity as building age increases does taper off toward the higher end of the scale."

Unfortunately, it is hard to interpret these models that include powers of the logged building age.

Using Piecewise Linear Functions:[200]

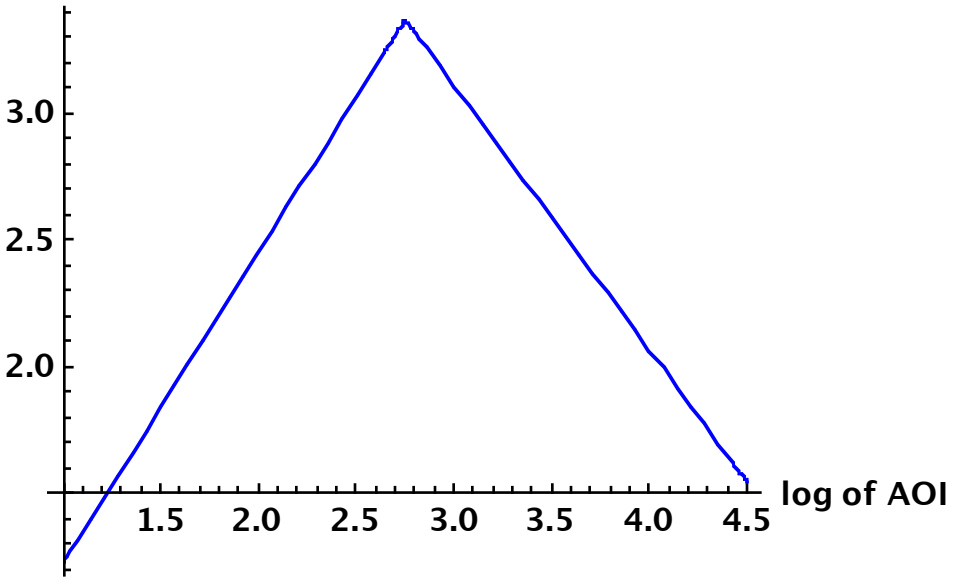
Let X_+ be X if $X \geq 0$ and 0 if $X < 0$.
Then a **hinge function** is: $\max[0, X - c)] = (X - c)_+$, for some constant c.
The constant c would be called the breakpoint.
Hinge functions can be used to create piecewise linear functions which can be used in GLMs.

For example, let X = ln[AOI]. Then a usual linear estimator is: $\beta_0 - 0.314\ x + ...$
Using instead a hinge function: $\beta_0 + 1.225\ x - 2.269\ (x - 2.75)_+ + ...$[201]

---

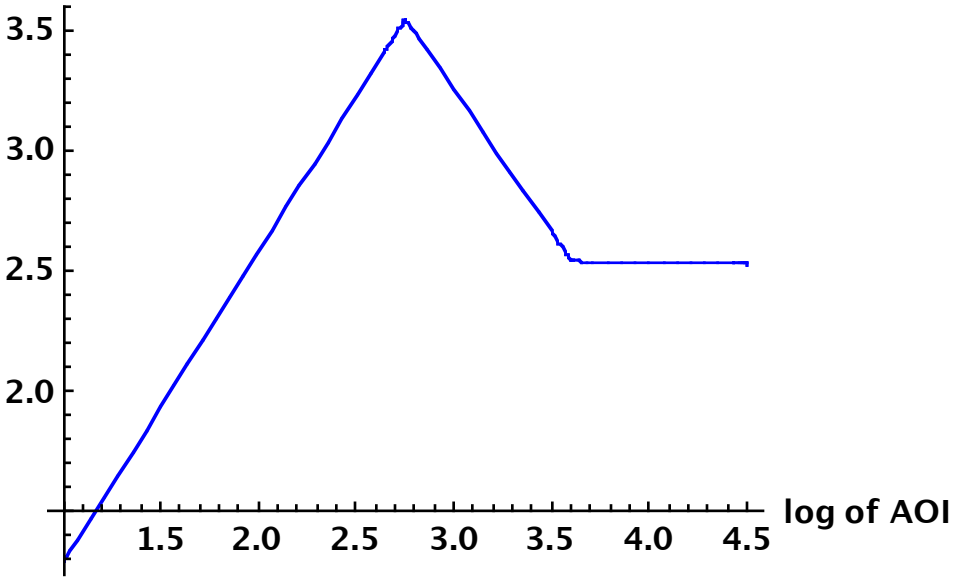
[199] See Figure 10 in Goldburd, Khare, and Tevet.
[200] See Section 5.4.4 of Goldburd, Khare, and Tevet.
[201] See Table 6 in Goldburd, Khare, and Tevet, "adding a breakpoint at 2.75."

Here is a graph of the broken line that results from including the hinge function:



For ln[AOI] < 2.75,we have slope 1.225, while for ln[AOI] > 2.75 we have a slope of:
1.225 - 2.269 = -1.044.

Instead we can use two hinge functions:
$\beta_0 + 1.289\, x - 2.472\, (x - 2..75)_+ + 1.170\, (x - 3.60)_+ + ...$[202]
Here is a graph of the broken line that results from including two hinge functions:

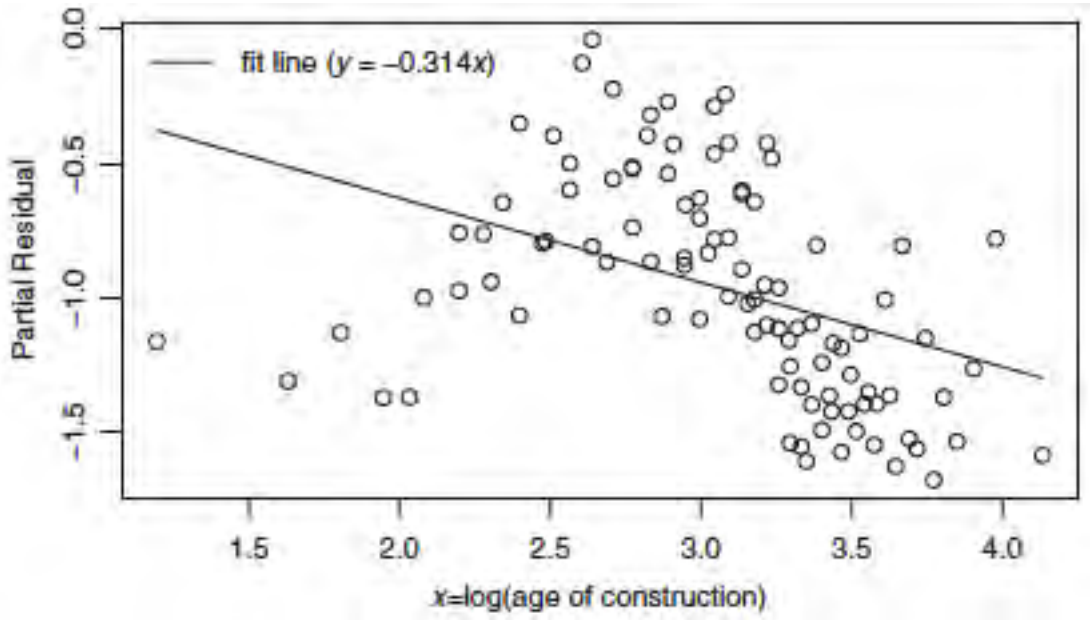


For ln[AOI] < 2.75,we have slope 1.289,
for 3.60 > ln[AOI] > 2.75 the slope is: 1.289 - 2.472 = -1.183,
while for 3.60 > ln[AOI] > 3.60 the slope is: 1.289 - 2.472 + 1.170 = -0.013.

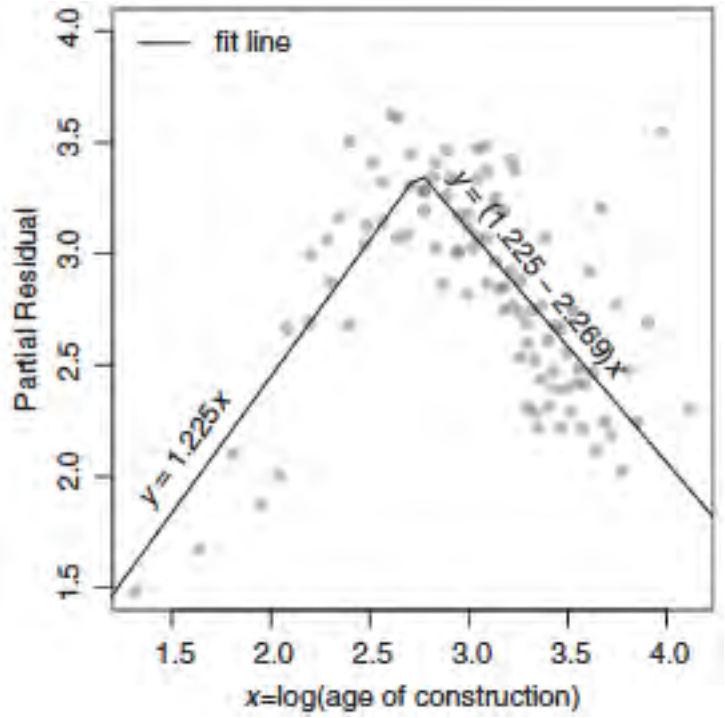
---

[202] See Table 7 in Goldburd, Khare, and Tevet, “adding an additional breakpoint at 3.6.”

Here is a graph of the partial residuals for the straight line:[203]



Here is a graph of the partial residuals for the broken line that results from using one hinge function:[204]

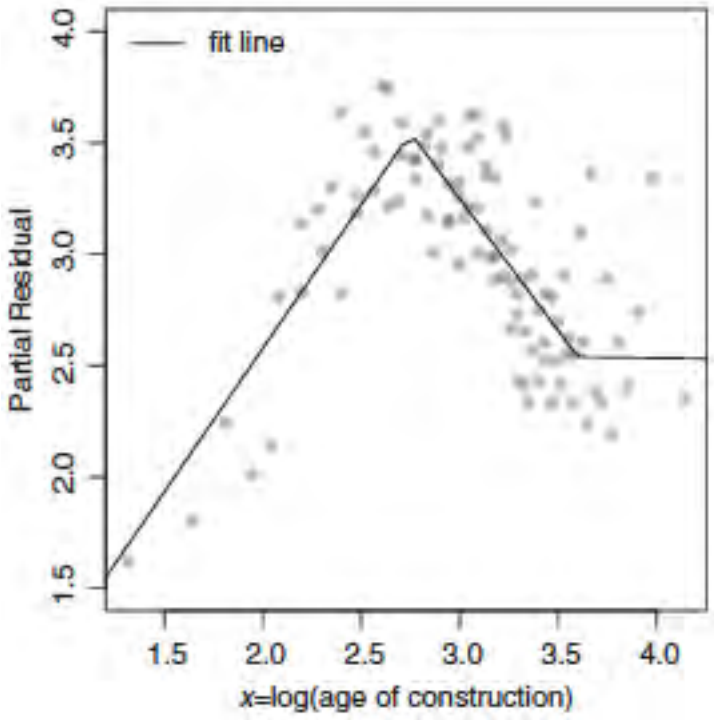


The model using the broken line does a better job of fitting the authors' data than the model that uses the straight line.

---

[203] See Figure 8 in Goldburd, Khare, and Tevet.
[204] See Figure 11 in Goldburd, Khare, and Tevet.

Here is a graph of the partial residuals for the broken line using two hinge functions:[205]



The model using two hinge functions may do a somewhat better job of fitting the authors' data than the model that uses one hinge function. With limited data it is hard to tell.[206]

Hinge functions provide more flexibility at the cost of greater complexity.

[205] See Figure 11 in Goldburd, Khare, and Tevet.
[206] "As this leveling-off effect comports with our intuition, we may decide to keep the third hinge function term in the model."

Grouping Categorical Variables:[207]

Some predictor variables are ordinal; they are discrete with several categories with a natural order. Sometimes it is useful for modeling purposes to group such predictor variables into fewer categories.[208] This is particularly useful when there are many categories.[209]

For example, workers compensation claims are categorized as: medical only, temporary total, minor permanent partial, major permanent partial, permanent total, and fatal. For some purposes it might be useful to group the first three categories into nonserious and the last three categories into serious.

One can start with a model without grouping. Statistical tests can determine whether the coefficients of adjacent levels are significantly different. Then one can group adjacent levels with similar fitted coefficients. Now run a new model using these groupings, and iterate the procedure. One needs to balance the competing priorities of: predictive power, parsimony, and avoiding overfitting.

Interactions:[210]

If x_1 and x_2 are predictor variables, then we can include an interaction term: x_1x_2.
Then the model would be: $g(\mu) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2 + ...$
This provides more flexibility at the cost of complexity.[211]

For example let x_1 be gender and x_2 be age. Then if we include an interaction term the effect of age depends on gender, and the effect of gender depends on age.

The syllabus reading gives an example with building occupancy class and sprinkler status.[212] Models are fit both with and without an interaction term.[213] The model with interactions is:

$$\mu = (\text{mean for base}) \exp[0.2303\, x_1 + 0.4588\, x_2 + 0.0701\, x_3 - 0.2895\, x_4 - 0.2847 x_1x_4 - 0.0244\, x_2x_4 - 0.2622\, x_3x_4],$$

where $x_1 = 1$ if occupancy class 2, $x_2 = 1$ if occupancy class 3, $x_3 = 1$ if occupancy class 4, $x_4 = 1$ if sprinklered, and occupancy class 1 without sprinklers is the base.

---

[207] See Section 5.5 in Goldburd, Khare, and Tevet.
[208] This is analogous to Robertson grouping classes into Hazard Groups.
[209] The syllabus reading uses the example of driver age, which can be thought of as either continuous or discrete. In the case of age, there may not be any clear breakpoints to use for grouping; actuarial judgement may be needed.
[210] See Section 5.6 in Goldburd, Khare, and Tevet.
[211] One would only include the interaction term if its coefficient were significantly different from zero.
[212] This is a commercial building claims frequency model using a Poisson with a log link function.
[213] See Tables 8 and 9 in Goldburd, Khare, and Tevet.
While two of the interaction terms are significantly different from zero, the remaining one is not.
They show an intercept which only makes sense if there are other predictor variables in the model.

For a non-sprinklered building in occupancy class 2, the multiplicative relativity to the base is: exp[0.2303] = 1.259.

For a sprinklered building in occupancy class 2, the multiplicative relativity to the base is: exp[0.2303 - 0.2895 - 0.2847] = 0.709.

Exercise: For a non-sprinklered building in occupancy class 4, determine the multiplicative relativity to the base.
[Solution: exp[0.0701] = 1.073.]

Exercise: For a sprinklered building in occupancy class 4, determine the multiplicative relativity to the base.
[Solution: exp[0.0701 - 0.2985 + 0.2622] = 1.044.
Comment: For occupancy class 4, the effect of sprinklers is small, while for occupancy class 2, the effect of sprinklers is large.]

The syllabus reading also shows another fitted model, with occupancy class, sprinklered, ln[AOI/200,000], plus an interaction term between sprinklered and ln[AOI/200,000]:[214]

μ = (mean for base) exp[0.2919 x_1 + 0.3510 x_2 + 0.0370 x_3 - 0.5153 x_4 + 0.4239 x_5
- 0.1032 $x_4 x_5$],

where x_1 = 1 if occupancy class 2, x_2 = 1 if occupancy class 3, x_3 = 1 if occupancy class 4, x_4 = 1 if sprinklered, x_5 = ln[AOI/200,000],
and AOI = 200,000 in occupancy class 1 without sprinklers is the base.

For a non-sprinklered building in occupancy class 2 with AOI = 500,000, the multiplicative relativity to the base is: exp[0.2919 + 0.4239 ln[2.5]] = 1.975.

Exercise: For a sprinklered building in occupancy class 2 with AOI = 500,000,
determine the multiplicative relativity to the base.
[Solution: exp[0.2919 - 0.5153 + 0.4239 ln[2.5] - 0.1032 ln[2.5]] = 1.073.]

For range of sizes of AOI for buildings that are insured, the expected frequency increases at a slower rate with AOI for sprinklered buildings than for non-sprinklered buildings.

[214] See Table 12 in Goldburd, Khare, and Tevet. They show an intercept of -3.771, which implies an expected frequency for the base level of: exp[-3.771] = 2.3%.

Loglikelihood:[215]

The loglikelihood is the sum of the contributions of the ln[density] at each of the observations. All other things being equal, a larger loglikelihood indicates a better fit. However, the principle of parsimony means that we should not add additional parameters to a model unless it significantly increases the loglikelihood.

The saturated model has as many parameters as the number of observations.
Each fitted value equals the observed value.
The saturated model has the largest possible likelihood, of models of a given form.
The minimal model has only one parameter, the intercept.[216]
The minimal model has the smallest possible likelihood, of models of a given form.[217]

Deviance:[218]

The deviance is twice the difference between the maximum loglikelihood for the saturated model (with as many parameters as data points) and the maximum loglikelihood for the model of interest.

D = Deviance = 2 {(loglikelihood for the saturated model) - (loglikelihood for the fitted model)}.
The smaller the deviance, the better the fit of the GLM to the data.[219]
Maximizing the loglikelihood is equivalent to minimizing the deviance.

By definition, the deviance of the saturated model is zero. Even though the saturated model fits the data perfectly we would not use it to predict the future, since the saturated model is overfit; the saturated model picks up too much of the randomness in the data (called the noise).

The minimal model has the largest possible deviance while the saturated model has the smallest possible deviance of zero. The deviance of a fitted model will lie between those two extremes.

We will be comparing the deviance of models with the same distributional form, same dispersion parameter, and link function, that have been fit to the same data.[220]

---

[215] See Section 6.1.1 in Goldburd, Khare, and Tevet.
They do not give any details on the form of the deviance for the different distributions.
See for example, An Introduction to Generalized Linear Models by Dobson and Barnett.
[216] Also sometimes called the null model.
[217] In the context of GLMs we would be comparing models with the same distributional form and link function, that have been fit to the same data.
[218] See Section 6.1.2 in Goldburd, Khare, and Tevet.
[219] Subsequently we will discuss how to test whether an improvement in deviance is statistically significant.
[220] If a variable has missing values for some records, the default behavior of most model fitting software is to toss out those records when fitting the model. In that case, the resulting measures of fit are no longer comparable, since the second model was fit with fewer records than the first.

Nested Models and the F-Test:[221] [222]

We can use the F-Test to compare two nested models, in other words when one model is a special case of the other. The bigger (more complex) model always has a smaller (better) deviance than the smaller (simpler) model. The question is whether the deviance of the bigger model is significantly better than that of the smaller model (special case).

Assume that we have two nested models.
Then the test statistic (asymptotically) follows an F-Distribution with numbers of degrees of freedom equal to: ν_1 = the difference in number of parameters, and

ν_2 = number of observations minus number of fitted parameters for the smaller model.[223]

The test statistic is: $\frac{(D_S - D_B) / (\text{number of added parameters})}{\hat{\phi}_S} \sim F_{df_S - df_B, df_S}$.[224]

D_S = deviance for the smaller (simpler) model.

D_B = deviance for the bigger (more complex) model.

df_S = number of degrees of freedom for the smaller (simpler) model.

= number of observations minus number of fitted parameters for the simpler model.

df_B = number of degrees of freedom for the bigger (more complex) model

= number of observations minus number of fitted parameters for the more complex model.

number of added parameters = $df_S - df_B$.

$\hat{\phi}_S$ = estimated dispersion parameter for the smaller (simpler) model.[225] [226] [227]

---

[221] See Section 6.2.1 in Goldburd, Khare, and Tevet.

[222] This F-Test is analogous to that used to test slopes in multiple regression.

[223] A Table of the F-Distribution is not attached to your exam, although they could give some values in a question. An F-Distribution is the ratio of two independent Chi-Square Distributions, with each Chi-Square divided by its number of degrees of freedom. ν_1 = the number of degrees of freedom of the Chi-Square in the numerator.

ν_2 = the number of degrees of freedom of the Chi-Square in the denominator.

If $\nu_1 = 1$, then the F-Distribution is related to the t-distribution.

Prob[F-Distribution with 1 and n degrees of freedom > c^2] =
Prob[absolute value of t-distribution with n degrees of freedom > c].
Thus if the difference in the number of parameters is one, then this test reduces to a t-test.

[224] I have seen instead in the denominator the estimated dispersion parameter of the more complex model. In that case, the degrees of freedom associated with the denominator are those of the more complex model. See "A Practitioners Guide to Generalized Linear Models," by Duncan Anderson, Sholom Feldblum, Claudine Modlin, Dora Schirmacher, Ernesto Schirmacher and Neeza Thandi, in the 2004 CAS Discussion Paper Program.

[225] The syllabus reading does not discuss how to estimate the dispersion parameter. One way to estimate the dispersion parameter in a model is as the ratio of the deviance to the number of degrees of freedom of the model.

[226] There is no requirement that the estimated dispersion parameters of the two models be equal.

[227] For cases where the dispersion parameter is one, such as for a Poisson or Negative Binomial Distribution, an actuary would normally use instead the likelihood ratio test, not discussed in the syllabus reading. See "A Practitioners Guide to Generalized Linear Models," by Duncan Anderson, Sholom Feldblum, Claudine Modlin, Dora Schirmacher, Ernesto Schirmacher and Neeza Thandi, in the 2004 CAS Discussion Paper Program.

If the F-Statistic is sufficiently big, then reject the null hypothesis that the data is from the smaller model in favor of the alternate hypothesis that the data is from the bigger model.[228]

Exercise: A GLM using a Gamma Distribution has been fit for modeling expenditures upon admission to a hospital. There are 150 observations. It uses 25 variables.
It uses 4 categories of self-rated physical health: poor, fair, good, and very good.
The deviance is 35.1.
An otherwise similar GLM excluding self-rated physical health has a deviance of 38.4.
The estimated dispersion parameter for this simpler model is 0.3.
Discuss how you would determine whether physical health is a useful variable for this model.
[Solution: The more complex model has 25 variables, and 150 - 25 = 125 degrees of freedom.
In order to incorporate physical health, avoiding aliasing, we need 4 - 1 = 3 variables.
Thus the simpler model has 22 variables, and 150 - 22 = 128 degrees of freedom.
The difference in degrees of freedom is: 128 - 125 = 3 = number of additional variables.

Test statistic is: $\frac{D_S - D_B}{(\text{number of added parameters})\,\hat{\phi}_S} = \frac{38.4 - 35.1}{(3)\,(0.3)} = 3.67.$

We compare the test statistic to an F-distribution with 3 and 128 degrees of freedom.
The null hypothesis is to use the simpler model, the one without physical health
The alternate hypothesis is to use the more complex model.
We reject the null hypothesis if the test statistic is sufficiently big.
Comment: The syllabus reading gives a similar example.]

Using a computer, the p-value of this test is 1.4%.[229]
Thus at a 2.5% significance level we would reject the simpler model in favor of the more complex model. At a 1% significance level we would not reject the simpler model.

If we had used a 2.5% significance level, we would have decided to use physical health.
We had used four levels of physical health: poor, fair, good, and very good.
The next step would be to see how many of these levels are useful. For example, does it significantly improve model performance to separate good from very good?

---

[228] The F-Distribution with ν_1 and $\nu_2 > 2$ degrees of freedom has a mean of $\nu_2/(\nu_2 - 1)$. For ν_2 large this mean is approximately 1. We reject the null hypothesis if the F-Statistic is significantly greater than 1.

[229] The 2.5% critical value is 3.220, while the 1% critical value is 3.938.
In other words, for the F-Distribution with 3 and 129 degrees of freedom, the survival function at 3.220 is 2.5%.

AIC and BIC:[230]

AIC and BIC are each methods of comparing models.
In each case, a smaller value is better.
These penalized measures of fit are particularly useful for comparing models that are not nested.

The Akaike Information Criterion (AIC) is used to compare a bunch of models all fit via maximum likelihood to the same data.[231] The model with the smallest AIC is preferred. For a particular model:
AIC = (-2) (maximum loglikelihood) + (number of parameters)(2).

The number of parameters fitted via maximum likelihood are the betas (slopes plus if applicable an intercept).[232]

Since the deviance = (2) (saturated max. loglikelihood - maximum loglikelihood for model),
we can compare between the models: Deviance + (number of parameters)(2).[233]

Assume for example, assume we have three Generalized Linear Models fit to the same data:

Model #	Number of Parameters	Deviance	Deviance + (number of parameters)(2)
1	4	888.7	896.7
2	5	886.2	896.2
3	6	884.4	896.4

We prefer Model #2, since it has the smallest AIC.[234]

The Bayesian Information Criterion (BIC) can also be used to compare a bunch of models all fit via maximum likelihood to the same data.[235] The model with the smallest BIC is preferred.
For a particular model:
BIC = (-2) (max. loglikelihood) + (number of parameters) ln(number of data points).[236]

Since the deviance = (2) (saturated max. loglikelihood - maximum likelihood for model),
we can compare between the models:
Deviance + (number of parameters) ln(number of data points).[237]

---

[230] See Section 6.2.2 in Goldburd, Khare, and Tevet.
[231] Thus AIC can be applied to Generalized Linear Models.
[232] If a dispersion parameter is fit via maximum likelihood, then the number of parameters in the above formula for AIC is one more. However, if one is using AIC to compare models, it does not matter, as long as one is consistent, since the only difference is to add the same constant to each AIC.
[233] The maximum likelihood for the saturated model is the same in each case.
[234] In each case, the AIC is: Deviance + (number of parameters)(2) - (2)(loglikelihood for the saturated model).
[235] Thus BIC can be applied to Generalized Linear Models.
[236] The GLM monograph uses ln and log interchangeably to both mean the natural log.
[237] The maximum likelihood for the saturated model is the same in each case.

Assume that we have three Generalized Linear Models fit to the same data set of size 20:

Model #	Number of Parameters	Deviance	Deviance + (number parameters) ln(20)
1	4	888.7	900.7
2	5	886.2	901.2
3	6	884.4	902.4

We prefer Model #1, since it has the smallest BIC.[238]
We note that in this case, using AIC or BIC would result in different conclusions.

BIC is mathematically equivalent to the Schwarz Bayesian Criterion.[239] Using the Schwarz Bayesian Criterion, one adjusts the loglikelihoods by subtracting in each case the penalty: (number of fitted parameters) ln(number of data points) / 2.
One then compares these penalized loglikelihoods directly; larger is better.
For a model, when BIC is smaller this penalized loglikelihood is bigger and vice-versa.

"As most insurance models are fit on very large datasets, the penalty for additional parameters imposed by BIC tends to be much larger than the penalty for additional parameters imposed by AIC. In practical terms, the authors have found that **AIC tends to produce more reasonable results**. **Relying too heavily on BIC may result in the exclusion of predictive variables from your model**."

[238] In each case, the BIC is: Deviance + (number of parameters)ln[20] - (2)(loglikelihood for the saturated model).
[239] See for example Loss Models, not on the syllabus of this exam.

A Communicable Disease Example:[240]

Assume we have the following reported occurrences of a communicable disease in two areas:

Number in Area A	Number in Area B	Month
8	9	2
8	12	4
10	9	6
11	14	8
14	15	10
17	19	12
13	20	14
15	21	16
17	25	18
15	23	20

Let $X_1 = 0$ if Region A and 1 if Region B.
Let $X_2 = \ln[\text{month}]$.
Fit a GLM with a Poisson using a log link function.
$\mu = \text{Exp}[\beta_0 + \beta_1 X_1 + \beta_2 X_2]$.

The fitted parameters are: $\beta_0 = 1.54894$, $\beta_1 = 0.265964$, $\beta_2 = 0.435105$.

The covariance matrix is: $\begin{pmatrix} 0.0618301 & -0.00781226 & -0.0226385 \\ -0.00781226 & 0.0138001 & -6.28837 \times 10^{-18} \\ -0.0226385 & -6.28837 \times 10^{-18} & 0.00948766 \end{pmatrix}$.

Therefore, approximate 95% confidence intervals for the parameters are:
$1.54894 \pm 1.960 \sqrt{0.0618301} = (1.06, 2.04)$,
$0.265964 \pm 1.960 \sqrt{0.0138001} = (0.04, 0.50)$,
$0.435105 \pm 1.960 \sqrt{0.00948766} = (0.24, 0.63)$.

The loglikelihood is: -47.0892.
The Deviance is: 4.45650.

[240] Adapted from Section 18.4 of Applied Regression Analysis by Draper and Smith, not on the syllabus.

In order to test whether $\beta_1 = 0$, the test statistic is:

$\hat{\beta}_1$ / StdDev[$\hat{\beta}_1$] = 0.265964 / $\sqrt{0.0138001}$ = 2.264.

The probability value of a two-sided test is: 2{1 - Φ[2.264]} = 2.4%.[241]

Exercise: Test whether $\beta_2 = 0$.

[Solution: $\hat{\beta}_2$ / StdDev[$\hat{\beta}_2$] = 0.435105 / $\sqrt{0.00948766}$ = 4.467.

The probability value of a two-sided test is: 2{1 - Φ[4.467]} = 0%.

Comment: Using a computer, the p-value is 8 x 10^{-6}.]

Exercise: Test whether $\beta_0 = 2$.

[Solution: ($\hat{\beta}_0$ - 2) / StdDev[$\hat{\beta}_0$] = (1.54894 - 2) / $\sqrt{0.0618301}$ = -1.814.

The probability value of a two-sided test is: 2 Φ[-1.814] = 7.0%.]

Now fit an otherwise similar GLM ignoring region, in other words without the dummy variable X_1.

The fitted parameters are: $\beta_0 = 1.69074$, $\beta_2 = 0.435105$.

The covariance matrix is: $\begin{pmatrix} 0.0574127 & -0.0226404 \\ -0.0226404 & 0.00948839 \end{pmatrix}$.

Therefore, approximate 95% confidence intervals for the parameters are:

β_0: 1.69074 ± 1.960 $\sqrt{0.0574127}$ = (1.22, 2.16),

β_2: 0.435105 ± 1.960 $\sqrt{0.00948839}$ = (0.24, 0.63).

The loglikelihood is: -49.6747.
The Deviance is: 9.62755.

For the model including region, the loglikelihood is -47.0892.
There are 20 data points and this model has 3 fitted betas.
AIC = (-2)(-47.0892) + (3)(2) = 100.178.
BIC = (-2)(-47.0892) + 3 ln(20) = 103.166.

For the simpler model excluding region, the loglikelihood is -49.6747.
This model has only 2 fitted betas.
AIC = (-2)(-49.6747) + (2)(2) = 103.349.
BIC = (-2)(-49.6747) + 2 ln(20) = 105.341.

[241] There is not a Normal Distribution Table attached to your exam.

The first more complicated model has the smaller AIC and thus is preferred on this basis.
The more complicated model has the smaller BIC and thus is also preferred on this basis.

The first model has a Deviance of 4.45650, while the second simpler model has a Deviance of 9.62755. Equivalently, we can use these Deviances.

For the first model, Deviance + (number of parameters)(2) = 4.45650 + (3)(2) = 10.45650.
For the second model, Deviance + (number of parameters)(2) = 9.62755 + (2)(2) = 13.62755.
Since 10.45650 < 13.62755, the first more complicated model is preferred on this basis.[242]

For the first model, Deviance + (number of parameters) ln(sample size) = 4.45650 + 3 ln(20) = 13.444.
For the second model, Deviance + (number of parameters) ln(sample size) = 9.62755 + 2 ln(20) = 15.619.
Since 113.444 < 15.619, the first more complicated model is also preferred on this basis.[243]

Deviance Residuals:[244]

The (ordinary) residuals are the difference between the observed and fitted values.
Other types of residuals are useful when working with GLMs, including Deviance Residuals.[245]
Deviance Residuals provide a more general quantification of the conformity of a case to the model specification.

Deviance Residuals are based on the form of the deviance for the particular distribution. Since the syllabus reading does not discuss these forms, you are not responsible for them on this exam.

The square of the deviance residual is the corresponding term in the sum that is the Deviance.[246]

We take the sign of the deviance residual as the same as that of the (ordinary) residual $y_i - \hat{\mu}_i$.

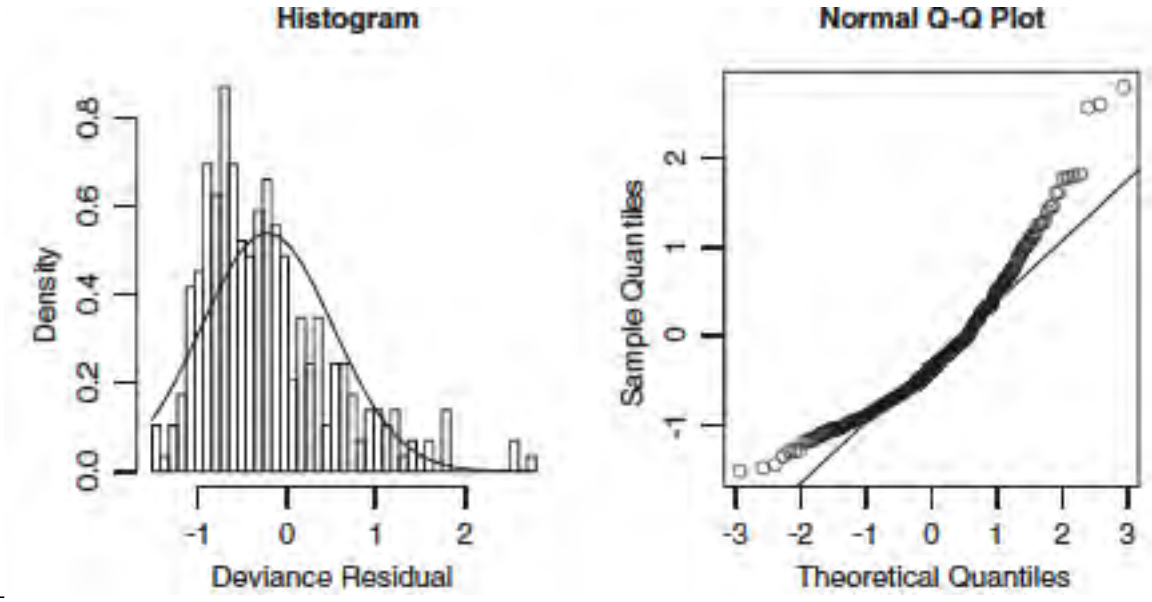
---

[242] This is equivalent to comparing AICs.
[243] This is equivalent to comparing BICs.
[244] See Section 6.3 in Goldburd, Khare, and Tevet.
[245] Pearson Residuals and Anscombe Residuals are also used, but these are not on the syllabus.
See for example Generalized Linear Models by McCullagh and Nelder,
Generalized Linear Models for Insurance Data by de Jong and Heller,
and An Introduction to Generalized Linear Models by Dobson and Barnett.
[246] The syllabus reading incorrectly states that the deviance residual itself rather than its square is the corresponding term in the sum that is the deviance.

"We can think of **the deviance residual as the residual adjusted for the shape of the assumed GLM distribution**, **such that its distribution will be approximately Normal** if the assumed GLM distribution is correct."

If the fitted model is appropriate, then we expect:
• **The deviance residuals should follow no predictable pattern.**[247]
• **The deviance residuals should be Normally distributed, with constant variance.**[248]

The syllabus reading shows an example of how to determine whether the deviance residuals are Normal. In the first case, a model was fit with a Gamma Distribution:[249]



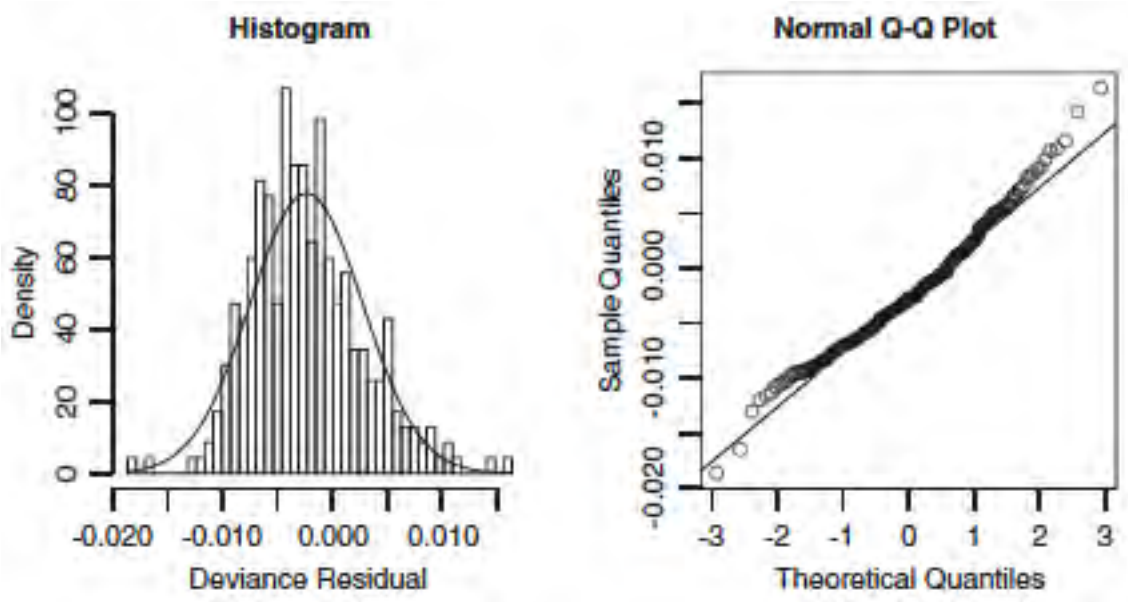
In the histogram, the deviance residuals do not seem close to the best fit Normal.[250]
In the Normal Q-Q plot, the deviance residuals are not near the comparison straight line.[251]
We conclude that the deviance residuals are not Normal and therefore the Gamma Distribution is probably not a good choice to model this data.

In the histogram, the deviance residuals are skewed to the right. Thus an Inverse Gaussian Distribution with greater skewness than a Gamma Distribution, might be better for modeling this data.

---

[247] If we discover a pattern in the deviance residuals then we can probably improve our model to pick this pattern up.
[248] The property of constant variance is called homoscedasticity.
Homoscedasticity is more closely followed for standardized deviance residuals, not on the syllabus.
If the model is correct, standardized residuals should (approximately) follow a Standard Normal Distribution.
See an An Introduction to Generalized Linear Models by Dobson and Barnett.
[249] See Figure 15 in Goldburd, Khare, and Tevet.
[250] See for example Loss Models, not on the syllabus of this exam.
[251] See for example Loss Models, not on the syllabus of this exam.

Here is similar graphs for a model that was fit with a Inverse Gaussian Distribution:[252]



In the histogram, the deviance residuals are much closer to the best fit Normal than before. In the Normal Q-Q plot, the deviance residuals are much nearer to the comparison straight line than before.

We conclude that the deviance residuals are closer to Normal, and therefore the Inverse Gaussian Distribution is probably a better choice to model this data than the Gamma Distribution.

Deviance Residuals for Discrete Distributions:

For discrete distributions such as Poisson or Negative Binomial, or distributions that have a point mass such as the Tweedie, the deviance residuals will likely not follow a Normal Distribution.[253] This makes deviance residuals less useful for assessing the appropriateness of such distributions, when each record is for a single risk.[254]

Fortunately, for data sets where one record may represent the average frequency for a large number of risks, deviance residuals are more useful than when each record is for a single risk.[255]

---

[252] See Figure 16 in Goldburd, Khare, and Tevet.

[253] This is because the deviance residuals do not adjust for the discreteness; the large numbers of records having the same target values cause the residuals to be clustered together in tight groups.

[254] One possible solution is to use randomized quantile residuals, which add random jitter to the discrete points so that they wind up more smoothly spread over the distribution.

[255] The target variable will take on a larger number of distinct values, effectively smoothing out the resulting distribution causing it to lose much of its discrete property and approach a continuous distribution.

Review, Histograms:

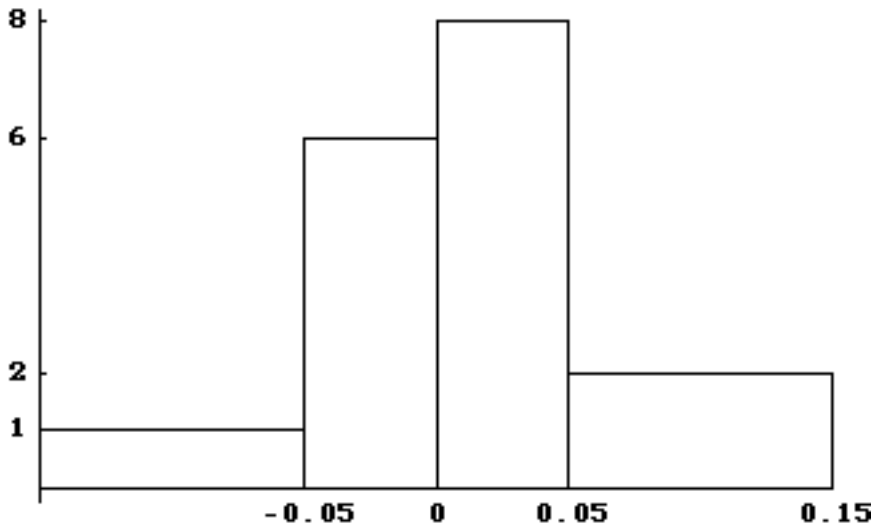
A histogram is an approximate graph of the probability density function.
First we need to group the data into intervals.
The height of each rectangle = $\frac{\text{\# values in the interval}}{\text{(total \# values) (width of interval)}}$.

For example, let us assume we observe 100 values and group them into four intervals:
Number that are between -0.15 and -0.05: 10
Number between -0.05 and 0: 30
Number between 0 and 0.05: 40
Number between 0.05 and 0.15: 20

The first interval has width 0.1. The probability in the first interval is: 10/100.
We want the area of the first rectangle to be equal to the probability in the first interval.
(0.1)(height) = 10/100. $\Rightarrow$ Height = (10/100) / (0.1) = 1.

Similarly, the height of the second rectangle is: (30/100) / (0.05) = 6.
The height of the third rectangle is: (40/100) / (0.05) = 8.
The height of the fourth rectangle is: (20/100) / (0.10) = 2.

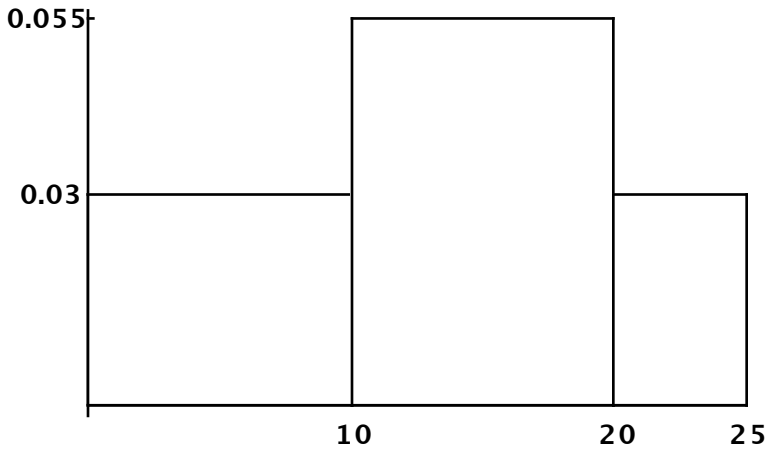
The histogram of these 100 values:



The sum of the areas of the rectangles is: (0.1)(1) + (0.05)(6) + (0.05)(8) + (0.1)(2) = 1.
In general the area under a histogram should sum to one, just as for the graph of a probability density function.

Exercise: Draw a histogram of the following grouped data: 0 -10: 6, 10-20: 11, 20-25: 3.

[Solution: The heights are: $\frac{6}{(20)(10)} = 0.03$, $\frac{11}{(20)(10)} = 0.055$, and $\frac{3}{(20)(5)} = 0.03$.



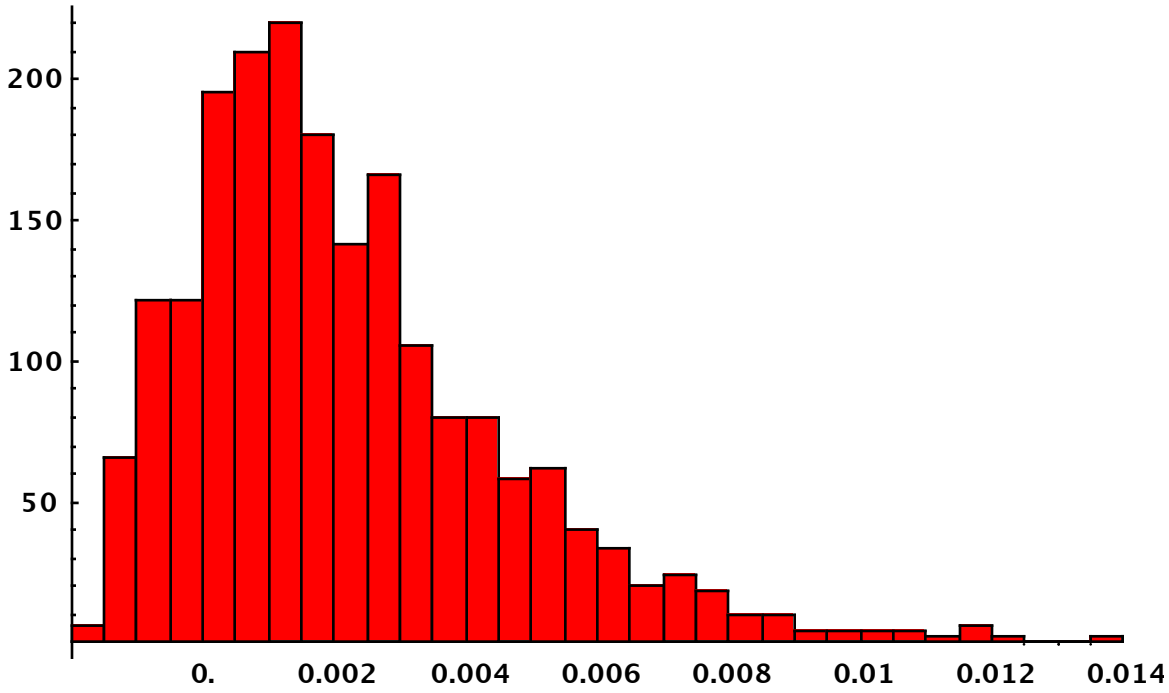
Comment: The sum of the areas of the rectangles is: (10)(0.03) + (10)(0.055) + (5)(0.03) = 1. With more data, we would get a better idea of the probability density function from which this data was drawn.]

Creating a histogram and comparing the histogram to a graph of a Normal Distribution is one way to determine whether the items of interest appear to be Normally distributed.

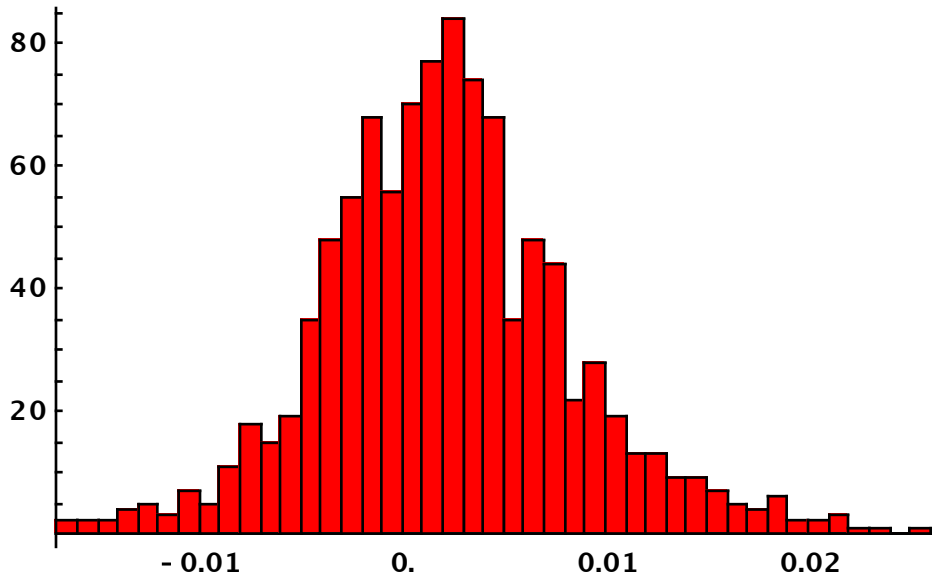
First we would want the histogram to look roughly symmetric, since the Normal Distribution is symmetric around its mean.[256]

[256] If the values are from a Normal Distribution, then one would expect the skewness of the observed values to be close to zero. In addition, since a Normal Distribution has a kurtosis of 3, if the values are from a Normal Distribution, then one would expect the kurtosis of the observed values to be close to 3.

The following histogram is not symmetric, and thus not likely to be a sample from a Normal Distribution:[257]

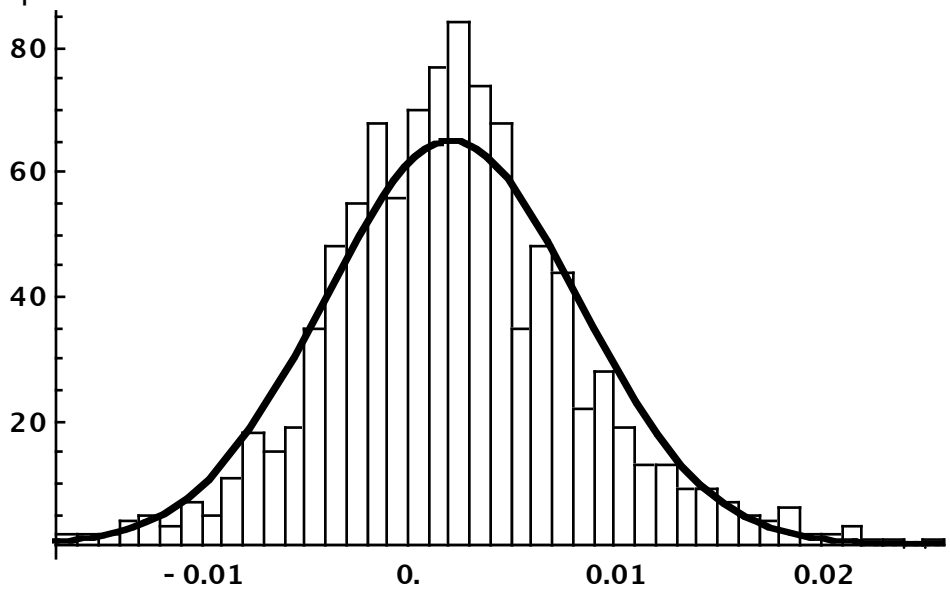


The following histogram looks approximately symmetric:



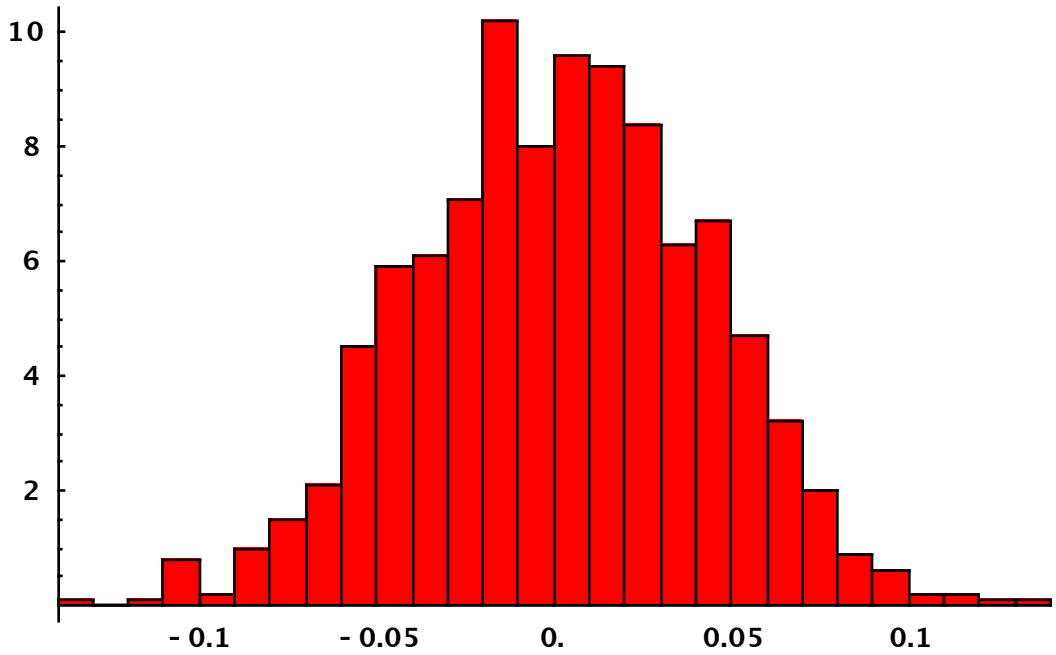
[257] *This histogram was based on 1000 data points simulated from a shifted Gamma Distribution.*

However, one can superimpose upon it a Normal Distribution with parameters $\mu = \overline{X}$ and σ = sample variance:



The histogram of the data seems to be more highly peaked than the Normal and may have heavier tails.[258] *This data has a larger kurtosis than a Normal; the graph displays leptokurtosis.*[259]

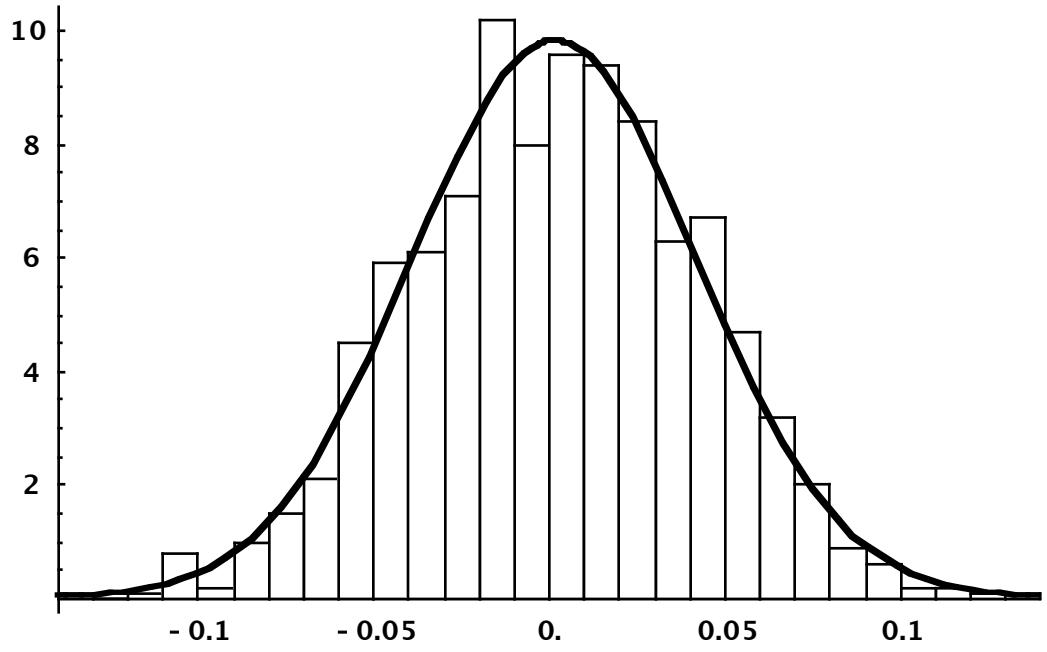
The following histogram, is based on a random sample of size 1000 from a Normal Distribution:



[258] Heavier tails means more probability in both the lefthand and righthand tails.

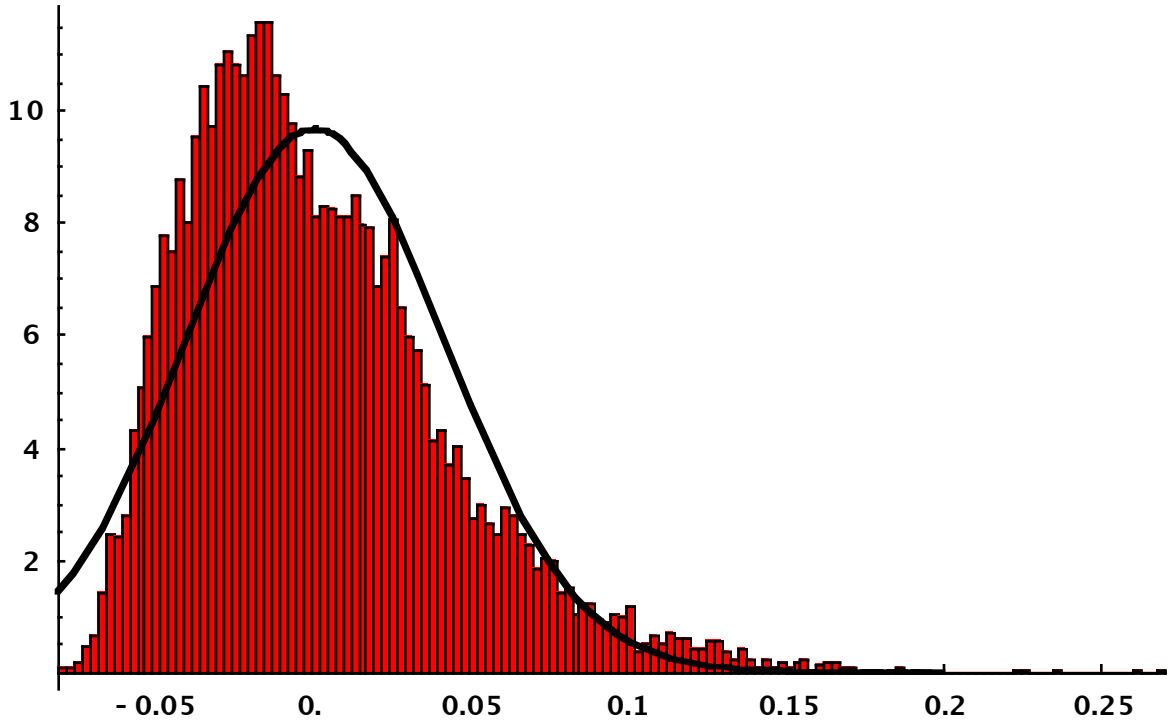
[259] *Kurtosis = 4th central moment / square of the variance. All Normal Distributions have a kurtosis of 3, so one would want the kurtosis of the data to also be close to 3. For the data that generated this histogram the kurtosis is 3.85, indicating somewhat heavier tails than a Normal Distribution.*

I superimposed upon the above histogram a Normal Distribution, with parameters $\mu = \overline{X}$ and σ = sample variance:



As with any finite sample, while the match between the data and a fitted Normal Distribution seems reasonable, it is far from perfect.

Next I simulated 10,000 random draws from a Gamma Distribution (with $\alpha = 4$), and then subtracted a constant.[260] I then compared a histogram of the data to the probability density function of a Normal Distribution with parameters based on the sample mean and sample variance of the data:



The curve of the Normal Distribution is a poor match to the data represented by the histogram.[261]
Even if we did not know the data was simulated from another distribution, we would conclude that this data was not drawn from a Normal Distribution.

Review, Q-Q Plots:

A Q-Q plot or quantile-quantile plot is a graphical technique which can be used to either compare a data set and a distribution or compare two data sets. Q-Q plots are most commonly used as a visual test of whether data is appears to be from a Normal Distribution. These are sometimes called Normal Q-Q Plots.

The 95th percentile is also referred to as $Q_{0.95}$, the 95% quantile.

For a distribution, the quantile Q_α is such for $F(Q_\alpha) = \alpha$. In other words, $Q_\alpha = F^{-1}(\alpha)$.

For example, $Q_{0.95}$ for a Standard Normal Distribution is 1.645, since $\Phi[1.645] = 0.95$.

---

[260] The key idea here is that the Gamma is some distribution different than the Normal Distribution.
[261] Since this Normal Distribution has the same mean and variance as the data, we would expect it to be a good match to the data, provided the data were drawn from a Normal Distribution.

In order to see whether data is drawn from some member of a Distribution Family, which has a scale and/or location parameter, we can create a Q-Q Plot for a standard member of that family F.

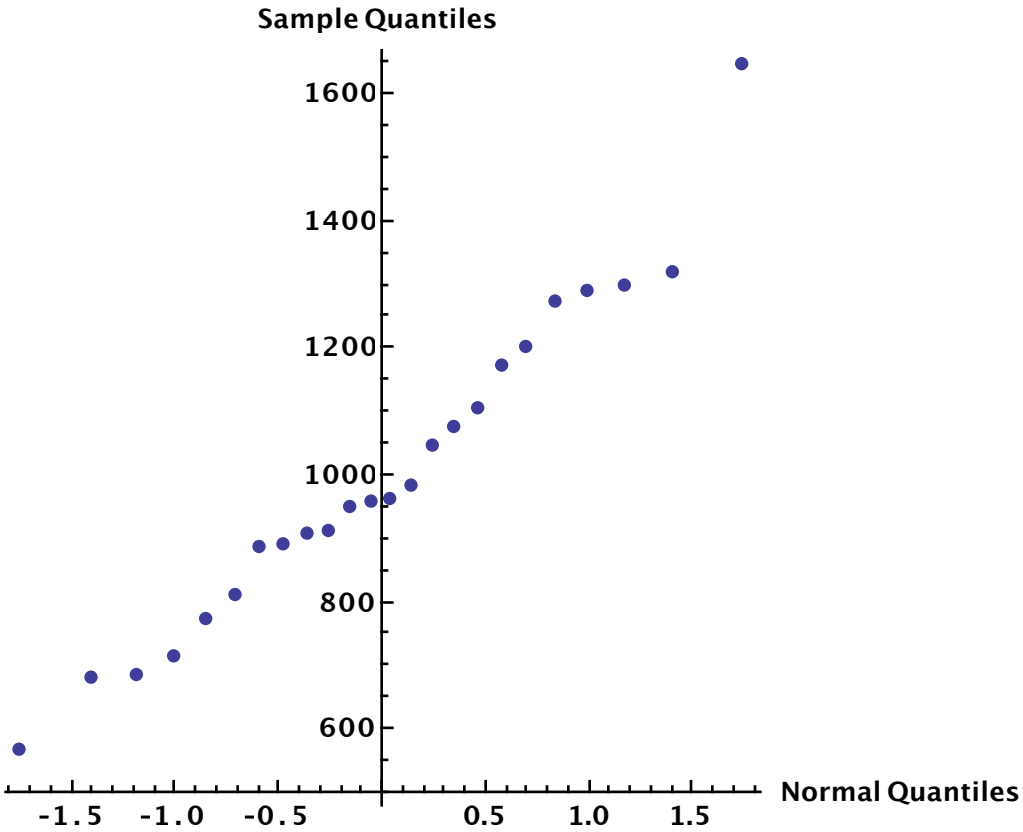
- Grade the n data points from smallest to largest.
- For i = 1 to n, plot the points: $(F^{-1}[\frac{i}{n+1}])$, $x_{(i)})$.

If the data is drawn from the given distribution family, then we expect the plotted points to lie close to some straight line.

Take the following 24 data point arranged from smallest to largest:
565, 678, 681, 713, 769, 809, 883, 890, 906, 909, 946, 956, 961, 983, 1046, 1073, 1103, 1171, 1198, 1269, 1286, 1296, 1316, 1643.
For the Standard Normal, $Q_{1/25} = Q_{0.04} = -1.751$.
Thus the first plotted point in a Normal Q-Q Plot is: (-1.751, 565).

Exercise: What is the second plotted point?
[Solution: $Q_{2/25} = Q_{0.08} = -1.405$. Thus the second plotted point is: (-1.405, 678).]

Here is the resulting Normal Q-Q Plot:

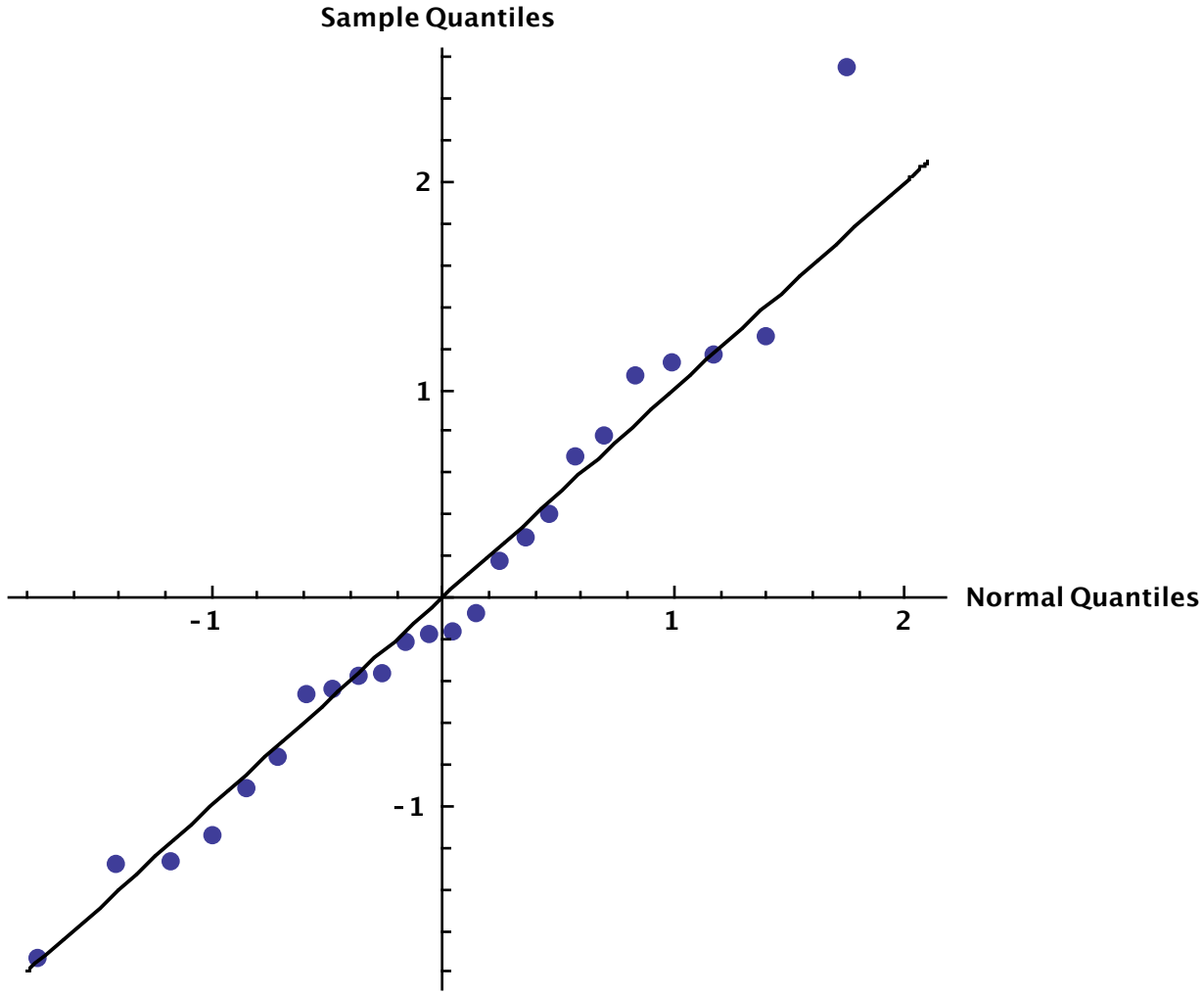


Other than the final point, the plotted points seem approximately linear, and thus this data could very well be from a single Normal Distribution.[262]

One could standardize each data point prior to constructing the Q-Q plot.
The data has a sample mean of 1002.08, and a sample variance of 63,387.9.
Thus we would subtract 1002.08 from each data point and divide by $\sqrt{63,387.9}$.

For example, (565 - 1002.08) / $\sqrt{63,387.9}$ = -1.736.

Here is the Q-Q Plot, using the standardized data, including the comparison line x = y:[263]



Again, other than the final point, the plotted points are close to the 45 degree comparison line, and thus this data could very well be from a single Normal Distribution.

---

[262] With small data sets it is hard to draw a definitive conclusion.
There is no specific numerical test we would apply to the Q-Q plot.

[263] Having standardized the data, when we compare to the Standard Normal Distribution, we expect the plotted points to be close to the 45 degree comparison line x = y.

Form of the Deviance Residual:[264]

The form of the deviance residual depends on the distribution and thus the form of the deviance.

Distribution	Square of the Deviance Residual
Normal	$\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$
Poisson	$2 \sum_{i=1}^{n} \{y_i \ln[y_i / \hat{\lambda}_i] - (y_i - \hat{\lambda}_i)\}$
Binomial	$2 \sum_{i=1}^{n} \{y_i \ln[\frac{y_i}{\hat{y}_i}] + (m_i - y_i) \ln[\frac{m_i - y_i}{m_i - \hat{y}_i}]\}$
Gamma	$2 \alpha \sum_{i=1}^{n} \{-\ln[y_i / \hat{y}_i] + (y_i - \hat{y}_i) / \hat{y}_i\}$
Inverse Gaussian	$\theta \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i^2 \, y_i}$
Negative Binomial	$2 \sum_{i=1}^{n} \{y_i \ln[y_i / \hat{y}_i] - (y_i + r) \ln[\frac{y_i + r}{\hat{y}_i + r}]\}$

Exercise: For a GLM using a Gamma Distribution, the first observed value is 800 with corresponding fitted value of 853.20. The maximum likelihood fitted parameter $\alpha = 45.6$
What is the corresponding deviance residual?
[Solution: $d_1^2 = (2)(45.6) \{-\ln[800/853.20] + (800 - 853.20)/853.20\} = 0.1850$.
Since 800 - 853.20 is negative, we take the deviance residual as negative.
$d_1 = -\sqrt{0.1850} = -0.430$.
Comment: This is for the two-dimensional example I discussed previously, using a reciprocal link function.]

[264] Not on the syllabus of this exam.

Communicable Disease Example Continued:

For the Poisson Distribution, the deviance is:

$$D = 2 \sum_{i=1}^{n} \{y_i \ln[y_i / \hat{\lambda}_i] - (y_i - \hat{\lambda}_i)\}.$$

Then the square of the deviance residual is the corresponding term in the above sum:

$$d_i^2 = 2 \{y_i \ln[y_i / \hat{\lambda}_i] - (y_i - \hat{\lambda}_i)\}.$$

For example, for the Communicable Disease Example which uses a Poisson Distribution, the first observed count is 8 with corresponding fitted value 6.3632.

Thus $d_1^2 = 2 \{8 \ln[8 / 6.3632] - (8 - 6.3632)\} = 0.3889$.

Since the first ordinary residual is positive, $d_1 = \sqrt{0.3889} = 0.6236$.

Exercise: For this example, the third observed count is 10 with corresponding fitted value 10.263.

Determine the corresponding deviance residual.

[Solution: $d_3^2 = 2 \{10 \ln[10 / 10.263] - (10 - 10.263)\} = 0.006798$

Since $10 - 10.263 < 0$, we take the deviance residual as negative.

$d_3 = -\sqrt{0.006798} = -0.0824$.]

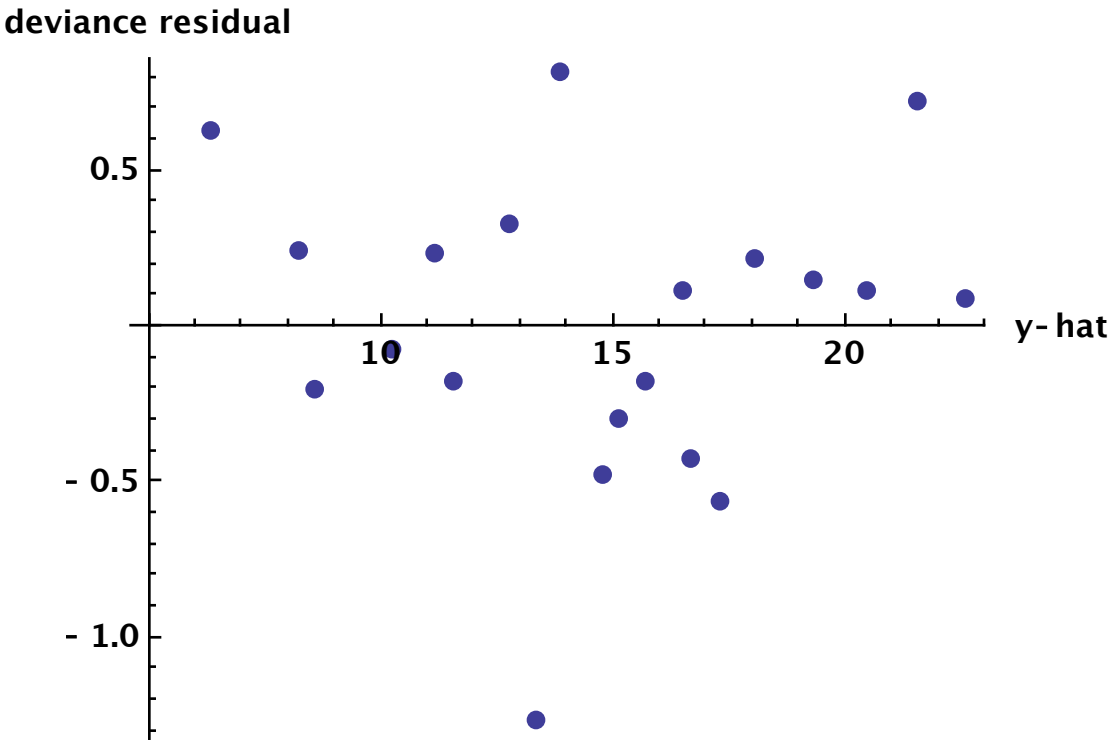
For this example, the deviance residuals are: 0.6237, -0.2081, -0.0824, -0.1869, 0.3254, 0.8099, -0.4876, -0.1845, 0.1094, -0.5728, 0.2390, 0.2289, -1.2763, -0.3058, -0.4288, 0.2090, 0.1448, 0.1061, 0.7144, 0.0819.

If the model is correct, then asymptotically the deviance is Chi-Square with n - p degrees of freedom, where n is the number of observations and p is the number of fitted parameters. Thus the expected value of the deviance is n - p. Therefore, we expect each term of the sum, d_i^2, to contribute about $(n - p)/n \cong 1$. Thus we expect $|d_i| \cong 1$.

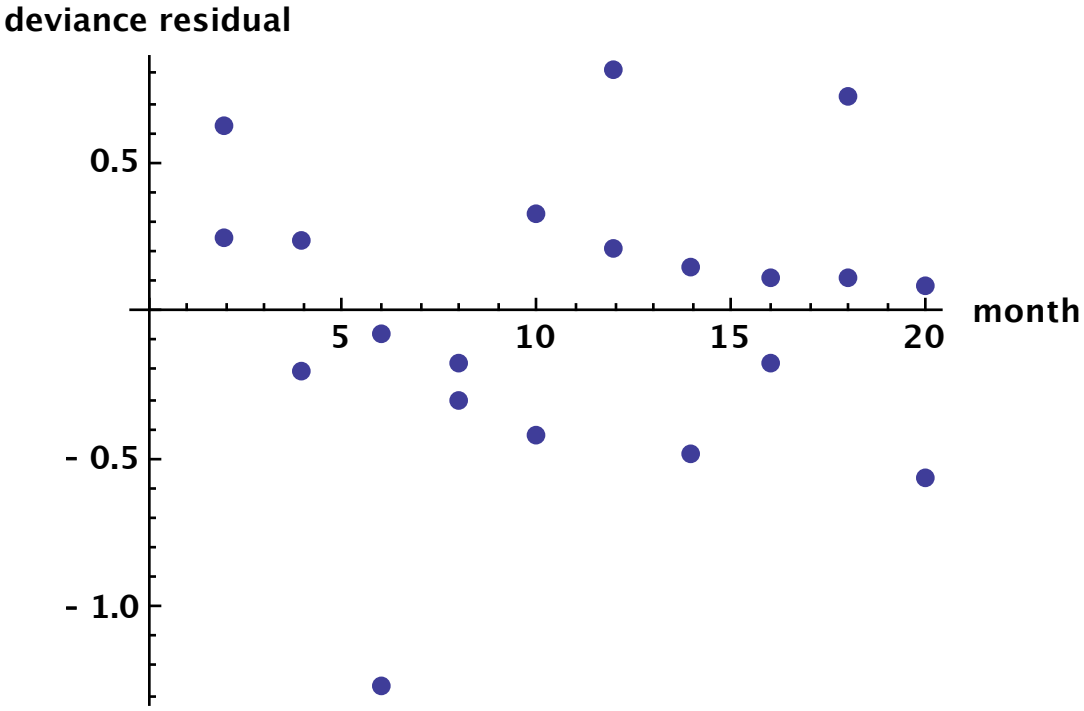
Thus, $|d_i|$ much bigger than one, indicates that observation i is contributing to a lack of fit.

In this example, the largest absolute value is 1.2763, not much bigger than one.

Here is a graph of the deviance residuals versus the fitted values:

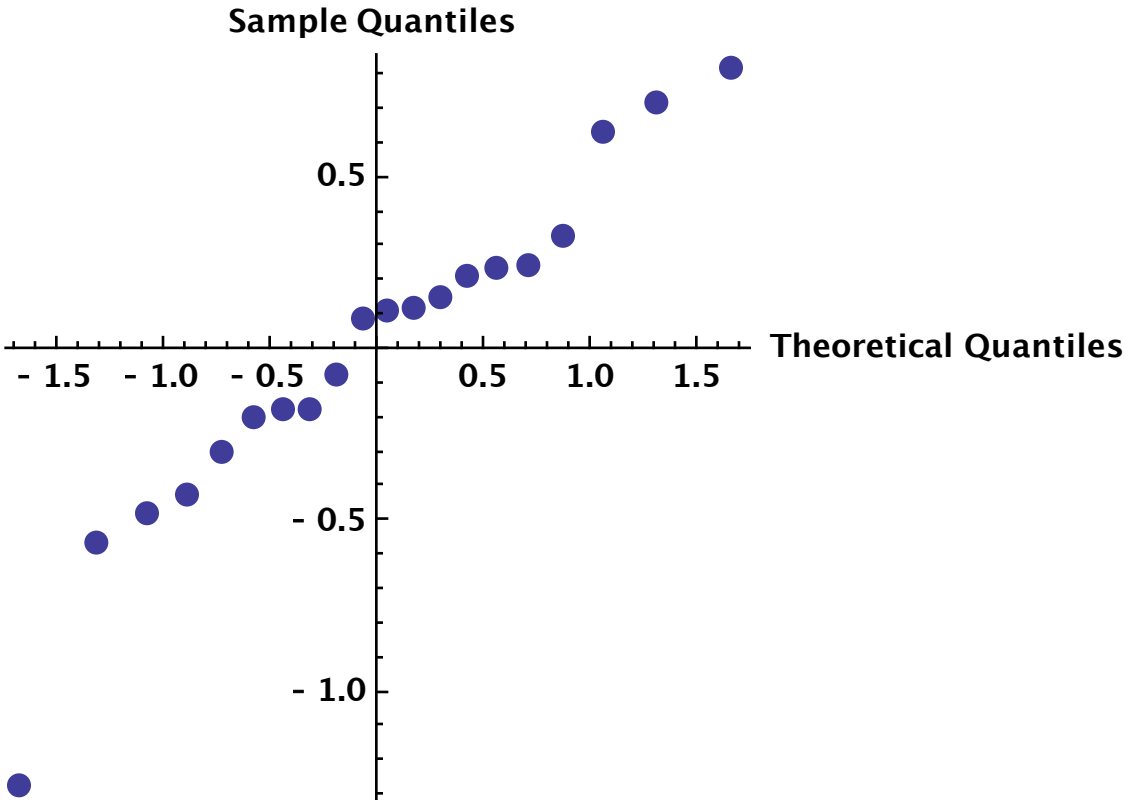


Here is a graph of the deviance residuals versus month:



In neither case do I observe an obvious pattern.

Here is a Normal Q-Q plot of the deviance residuals:



Other than the first point, the points seem to be approximately along a straight line, thus this data could be from a Normal. However, there is too little data to make a definite conclusion.[265]

[265] It turns out the standardized residuals do not seem to follow a Standard Normal. It turns out that for a Gamma Distribution rather than a Poisson Distribution, the standardized residuals seem to follow a Standard Normal. Thus a Gamma Distribution seems to be a better model for this data.

Assessing Model Stability:[266]

The actuary would like the GLM to be stable; in other words, the predictions of the model should not be overly sensitive to small changes in the data.

An observation is influential if it has a large effect on the fitted model. An outlier is an observation such that the corresponding fitted value is far from the observation.

An influential observation is such that its removal from the data set causes a significant change to our modeled results. An observation is influential when one or more of its predictor values are far from its mean and the observation is an outlier.

A common measure of influence is Cook's distance.[267] **The larger the value of Cook's distance, the more influential the observation.**[268]

The actuary should rerun the model excluding the most influential points to see their impact on the results. If this causes large changes in some of the parameter estimates, the actuary should consider for example whether to give these influential observations less weight.

Cross-validation, as discussed previously, can also be used to assess the stability of a GLM. For example, we can divide the data into ten parts. By combining these parts, we can create ten different subsets each of which contains 90% of the total data. We then fit the model to each of these ten subsets.

The results of the models fit to these different subsets of the data ideally should be similar. The amount by which these results vary is a measure of the stability of the model.

Bootstrapping via simulation can also be used to assess the stability of a GLM.[269] The original data is randomly sampled with replacement to create a new set of data of the same size. One then fits the GLM to this new set of data. By repeating this procedure many times one can estimate the distribution of the parameter estimates of the GLM; we can estimate the mean, variance, confidence intervals, etc. "Many modelers prefer bootstrapped confidence intervals to the estimated confidence intervals produced by statistical software in GLM output."

---

[266] See Section 6.4 of Goldburd, Khare, and Tevet.
[267] The syllabus reading gives no details on how Cook's Distance is calculated.
Computer software to fit GLMs will usual include Cook's Distance as one of the possible outputs.
[268] Values of Cook's Distance greater than unity may require further investigation.
[269] See An Introduction to Statistical Learning with Applications in R, by James, Witten, Hastie, and Tibshirani, not on the syllabus of this exam.

Scoring Models:[270]

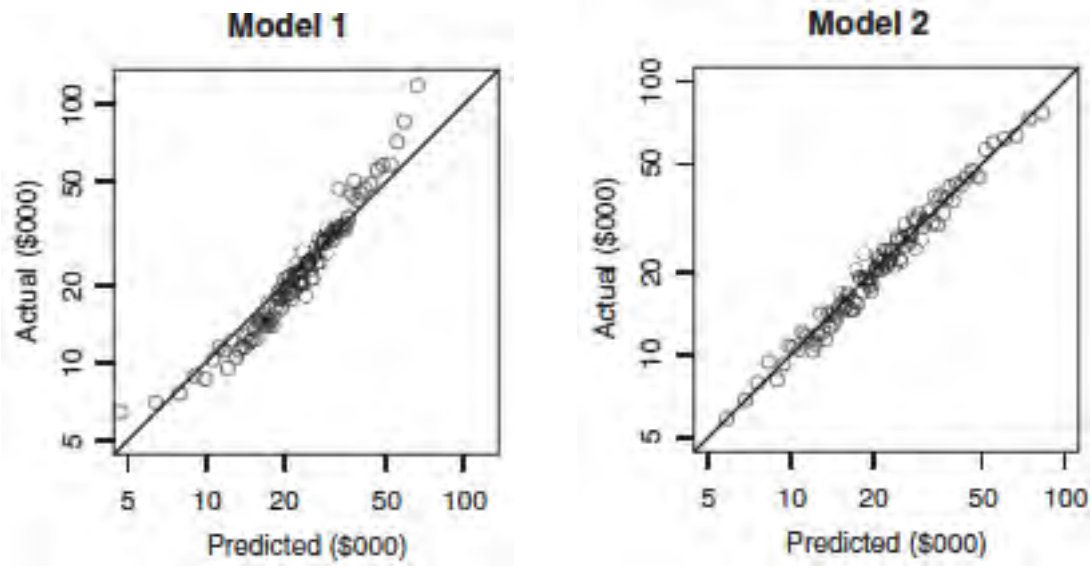
We have a rating plan or rating plans. We may not know what model if any that the plan(s) came from.[271] We wish to evaluate a rating plan or compare two rating plans.

Methods that are discussed: Plots of Actual vs. Predicted, Simple Quantile Plots, Double Lift Charts, Loss Ratio Charts, the Gini Index, and ROC Curves.

In order for these techniques to be used, one only needs a database of historical observations plus the predictions from each of the competing models. The process of assigning predictions to individual records is called scoring.

Assessing Fit with Plots of Actual versus Predicted:[272]

Create a plot of the actual target variable (on the y-axis) versus the predicted target variable (on the x-axis) for each model. If a model fits well, then the actual and predicted target variables should follow each other closely. Here are two examples:[273]



Model 2 fits the data better than Model 1, as there is a much closer agreement between the actual and predicted target variables for Model 2 than there is for Model 1.

These plots should not use data that was used to fit or train the models.
It is common to group the data, for example into percentiles.
Often one will plot the graph on a log scale, as in the above examples.

---

[270] See Section 7 of Goldburd, Khare, and Tevet.
[271] One or more of the rating plans may be proprietary.
[272] See Section 7.1 of Goldburd, Khare, and Tevet.
[273] See Figure 17 of Goldburd, Khare, and Tevet.

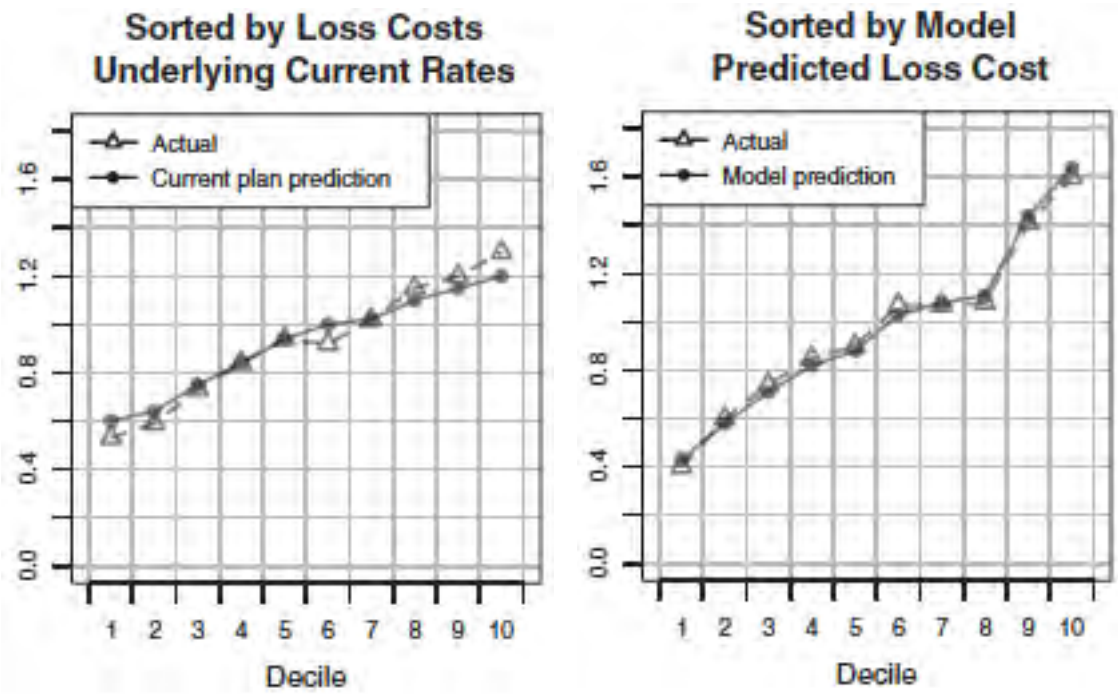
Measuring Model Lift:[274]

Lift refers to a model's ability to prevent adverse selection, **measuring the approximate "economic value" of the model.** Economic value is produced by comparative advantage in avoidance of adverse selection; thus model lift is a relative concept, comparing two or more competing models, or a model and the current plan. Lift measures a model's ability to charge each insured an actuarially fair rate, thereby minimizing the potential for adverse selection. Model lift should always be measured on holdout data, in other words not using data used to fit or build the model.

Simple Quantile Plots:[275]

To create a quantile plot of a model.
1. Sort the dataset based on the model predicted loss cost from smallest to largest.
2. Group the data into quantiles with equal volumes of exposures.[276]
3. Within each group, calculate the average predicted pure premium based on the model, and the average actual pure premium.
4. Plot for each group, the actual pure premium and the predicted pure premium.

One can create separate quantile plots for two models, for example the current rating plan and a proposed rating plan and compare them:[277]

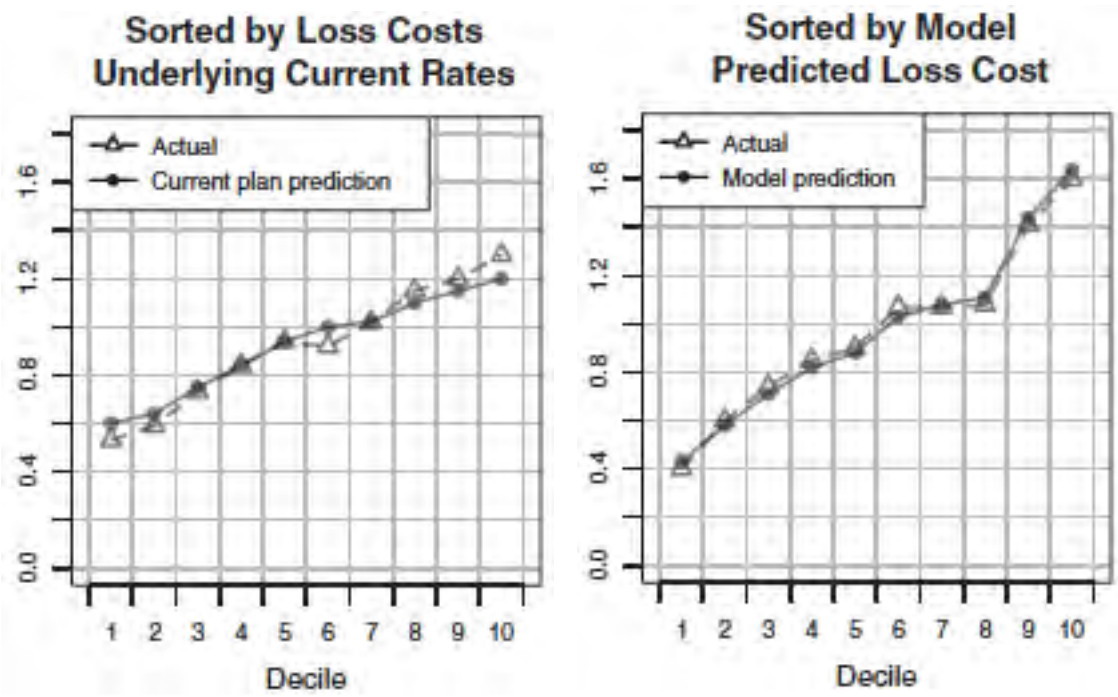


[274] See Section 7.2 of Goldburd, Khare, and Tevet. Lift differs from goodness of fit measures.
[275] See Section 7.2.1 of Goldburd, Khare, and Tevet. ß
[276] For example:.quintiles (5 buckets), deciles (10 buckets), or vigintiles (20 buckets).
[277] See Figure 18 in Goldburd, Khare, and Tevet.

To compare the models use the following 3 criteria:
1. **Predictive accuracy.**
2. **Monotonicity.** The actual pure premium should increase.[278]
3. **Vertical distance between the actuals in the first and last quantiles.**

"A large difference (also called "lift") between the actual pure premium in the quantiles with the smallest and largest predicted loss costs indicates that the model is able to maximally distinguish the best and worst risks."

The previous set of graphs can be used to compare the current and proposed model.



1. Predictive accuracy: the proposed model does a better job of predicting.
2. Monotonicity: the current plan has a reversal in the 6th decile, whereas the proposed model does better with no significant reversals.
3. Vertical distance between the first and last quantiles: The spread of actual loss costs for the current plan is 0.55 to 1.30. The spread of the proposed model is 0.40 to 1.60, which is larger and thus better.

Thus, by all three criteria, the proposed plan outperforms the current one.

[278] Although small reversals are okay.

Double Lift Charts:[279]

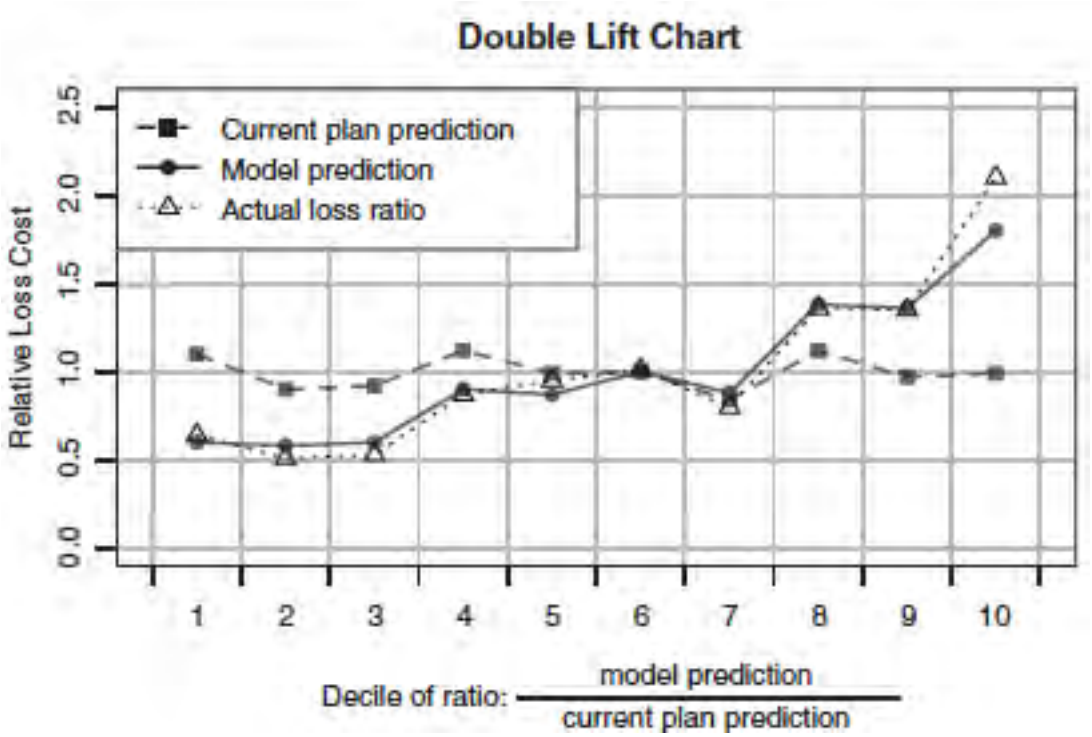
A double lift chart directly compares two models A and B.

To create a double lift chart:

1. For each observation, calculate Sort Ratio = $\frac{\text{Model A Predicted Loss Cost}}{\text{Model B Predicted Loss Cost}}$.[280]
2. Sort the dataset based on the Sort Ratio, from smallest to largest.
3. Group the data.[281]
4. For each group, calculate the pure premiums: predicted by Model A, predicted by Model B, and actual. Then divide the group average by the overall average.
5. For each group, plot the three relativities calculated in the step 4.

The first group contains those risks which Model A thinks are best relative to Model B, while the last group contains those risks which Model B thinks are best relative to Model A. The first and last groups contain those risks on which Models A and B disagree the most in percentage terms.

The "winning" model is the one that more closely matches the actual pure premiums.
Here is an example of a double lift chart, comparing a current and proposed plan:[282]



The proposed model more accurately predicts actual pure premium by decile than does the current rating plan. This is particularly clear when looking at the extreme groups on either end.

---

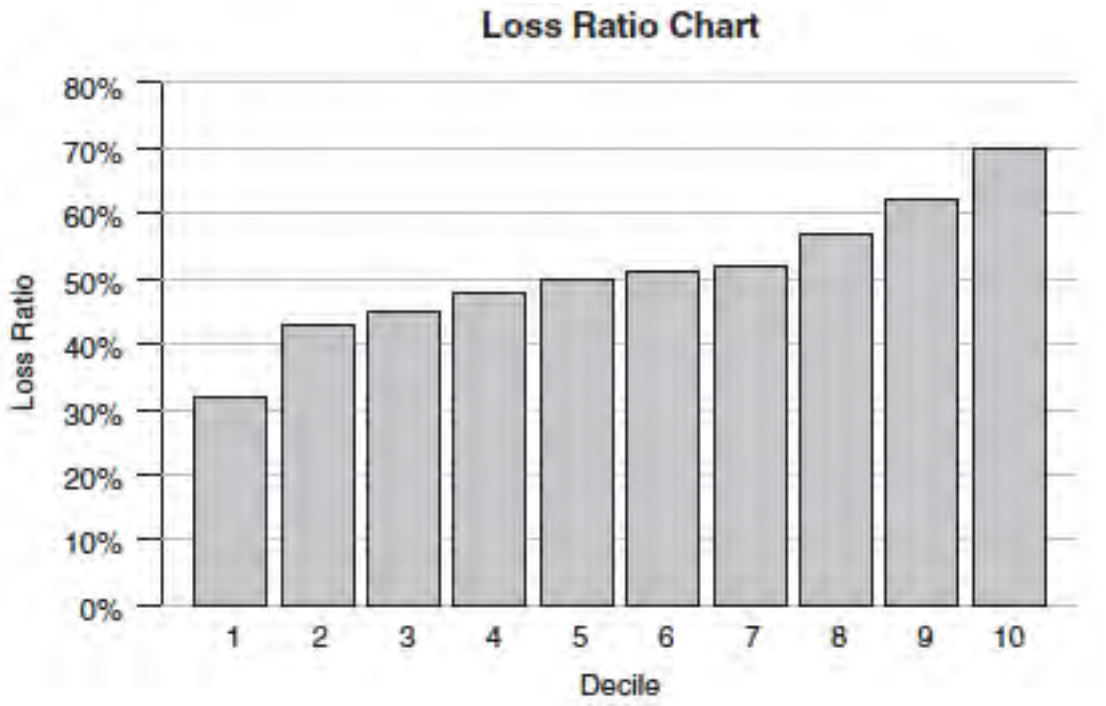
[279] See Section 7.2.2 of Goldburd, Khare, and Tevet.
[280] Thus a low sort ratio means that model A predicts a lower loss cost than does model B.
[281] For example into 5 buckets (quintiles) or 10 buckets (deciles).
[282] See Figure 19 in Goldburd, Khare, and Tevet.

"As an alternate representation of a double lift chart, one can plot two curves: the percent error for the model predictions and the percent error for the current loss costs, where percent error is calculated as $\frac{\text{Predicted Loss Cost}}{\text{Actual Loss Cost}} - 1$. In this case, the winning model is the one with the flatter line centered at y = 0, indicating that its predictions more closely match actual pure premium."

Loss Ratio Charts:[283]

A loss ratio chart is similar to a simple quantile chart, except one works with loss ratios (with respect to the premiums for the current plan) rather than pure premiums.
To create a loss ratio chart:
1. Sort the dataset based on the model prediction.
2. Group the data into quantiles with equal volumes of exposures.
3. Within each group, calculate the actual loss ratio.

Here is an example:[284]



The proposed model is able to segment the data into lower and higher loss ratio buckets, indicating that the proposed model is better than the current model. "The advantage of loss ratio charts over quantile plots and double lift charts is that they are simple to understand and explain."

[283] See Section 7.2.3 of Goldburd, Khare, and Tevet.
[284] See Figure 20 in Goldburd, Khare, and Tevet.

Lorenz Curves:[285]

The Lorenz Curve is used to define the Gini Index, to be discussed subsequently.

Assume that the incomes in a country follow a distribution function F(x).[286]
Then F(x) is the percentage of people with incomes less than x.

The income earned by such people is: $\int_0^x t\, f(t)\, dt = E[X \wedge x] - x\, S(x) = \int_0^x S(t)\, dt$.

The percentage of total income earned by such people is:

$$\frac{\int_0^x y\, f(y)\, dy}{E[X]} = \frac{E[X \wedge x] - x\, S(x)}{E[X]} .$$

Define $G(x) = \frac{\int_0^x y\, f(y)\, dy}{E[X]} = \frac{E[X \wedge x] - x\, S(x)}{E[X]}$.[287]

For example, assume an Exponential Distribution.
Then $F(x) = 1 - e^{-x/\theta}$.

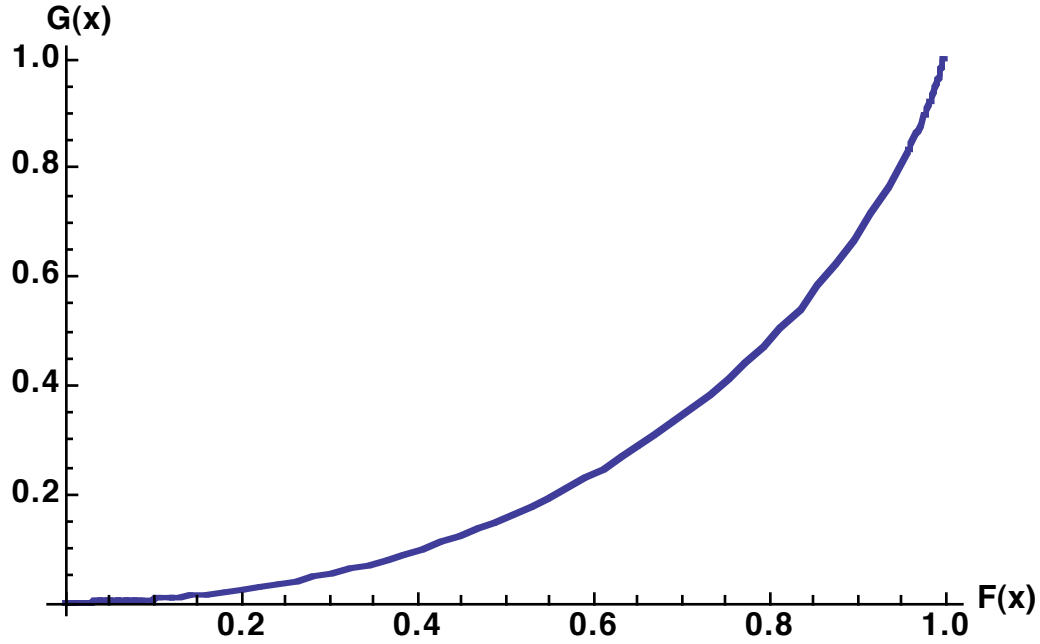
$$G(x) = \frac{E[X \wedge x] - x\, S(x)}{E[X]} = \frac{\theta\, (1 - e^{-x/\theta}) - x\, e^{-x/\theta}}{\theta} = 1 - e^{-x/\theta} - (x/\theta)\, e^{-x/\theta}.$$

Let $t = F(x) = 1 - e^{-x/\theta}$. Therefore, $x/\theta = -\ln(1 - t)$.[288]
Then, $G(t) = t - \{-\ln(1-t)\}\, (1-t) = t + (1-t)\, \ln(1-t)$.

---

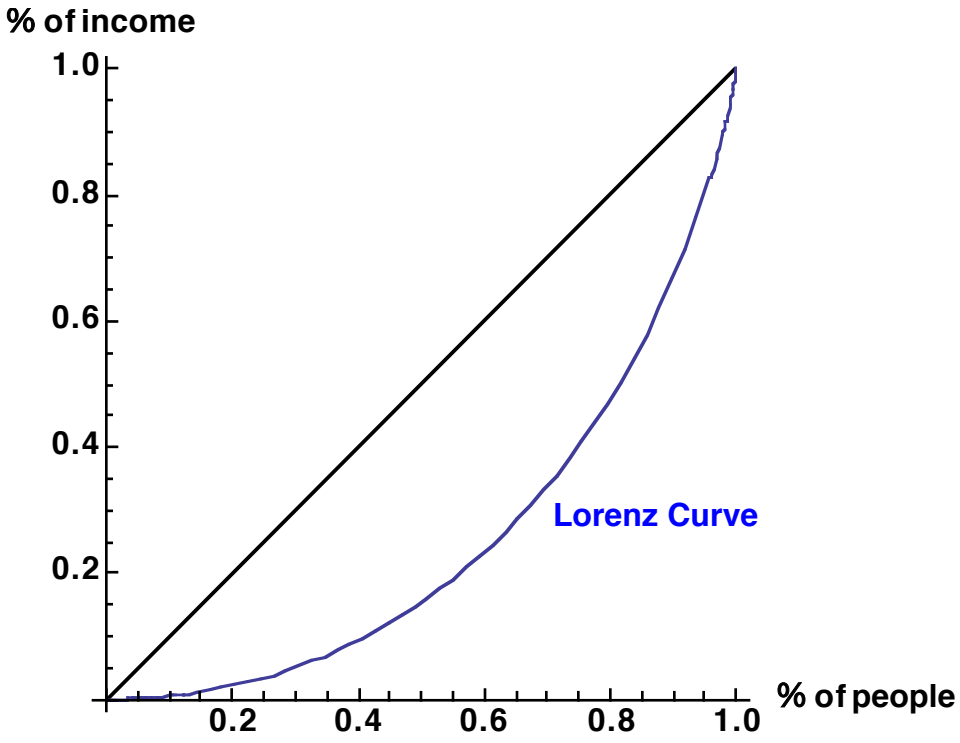
[285] You should not be responsible for any details of the mathematics of Lorenz curves.
[286] Of course, the mathematics applies regardless of what is being modeled.
The distribution of incomes is just the most common context.
[287] This is not standard notation. I have just used G to have some notation.
[288] This is just the VaR formula for the Exponential Distribution.

Then we can graph G as a function of F:



This curve is referred to as the Lorenz curve or the concentration curve.

Since F(0) = 0 = G(0) and F(∞) = 1 = G(∞), the Lorenz curve passes through the points (0, 0) and (1, 1). Usually one would also include in the graph the 45° reference line connecting (0, 0) and (1, 1), called the line of equality, as shown below:



$$G(t) = G[F(x))] = \frac{\int_0^x y\, f(y)\, dy}{E[X]}.$$

$$\frac{dG}{dt} = \frac{dG}{dx} / \frac{dF}{dx} = \frac{x\, f(x)}{E[X]} / f(x) = \frac{x}{E[X]} > 0.$$

$$\frac{d^2G}{dt^2} = \frac{1}{E[X]} \frac{dx}{dx} / \frac{dF}{dx} = \frac{1}{E[X]\, f(x)} > 0.$$

Thus, in the above graph, as well as in general, the Lorenz curve is increasing and concave up. The Lorenz curve is below the 45° reference line, except at the endpoints when they are equal.

The vertical distance between the Lorenz curve and the 45° comparison line is: F - G.

Thus, this vertical distance is a maximum when: $0 = \frac{dF}{dF} - \frac{dG}{dF}$.

$$\Rightarrow \frac{dG}{dF} = 1. \Rightarrow \frac{x}{E[X]} = 1. \Rightarrow x = E[X].$$

Thus the vertical distance between the Lorenz curve and the line of equality is a maximum at the mean income.

Exercise: If incomes follow an Exponential Distribution, what is this maximum vertical distance between the Lorenz curve and the line of equality?

[Solution: The maximum occurs when $x = \theta$.

$F(x) = 1 - e^{-x/\theta}$. From previously, $G(x) = 1 - e^{-x/\theta} - (x/\theta) e^{-x/\theta}$.

$F - G = (x/\theta) e^{-x/\theta}$. At $x = \theta$, this is: $e^{-1} = 0.3679$.]

Exercise: Determine the form of the Lorenz Curve, if the distribution of incomes follows a Shifted Pareto Distribution, with $\alpha > 1$.[289]

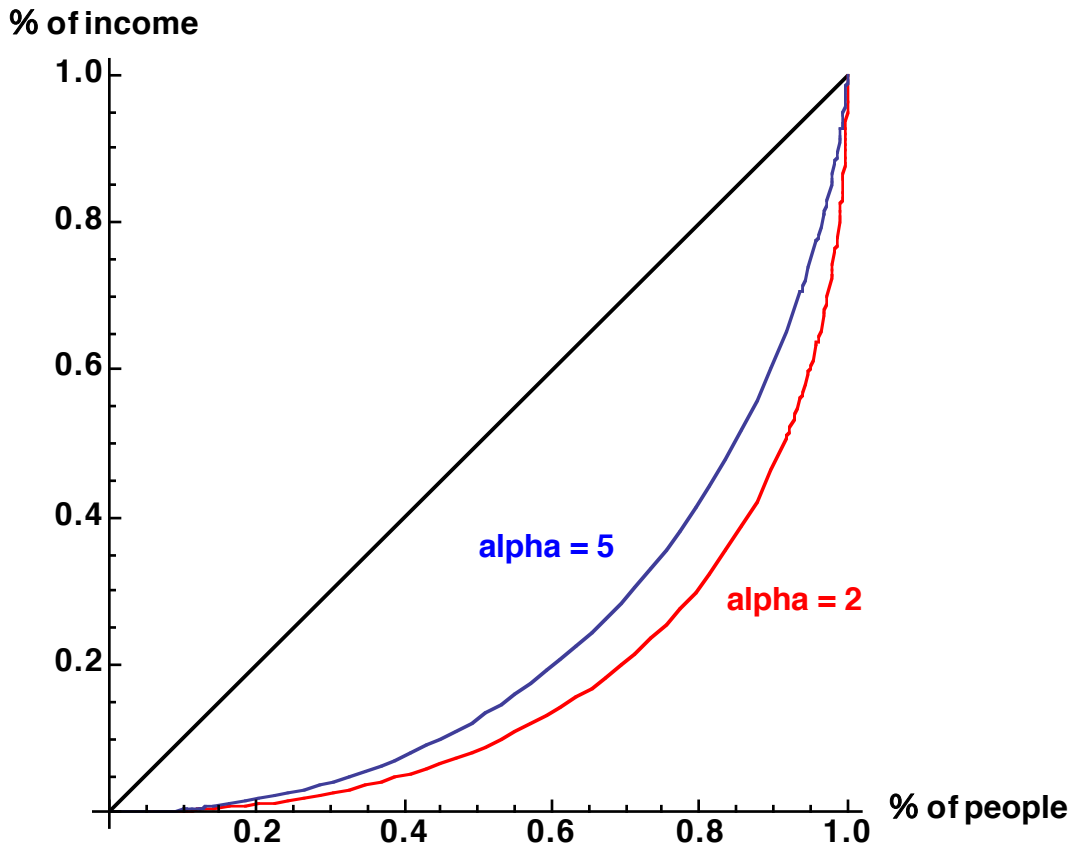
[Solution: $F(x) = 1 - \left(\frac{\theta}{\theta+x}\right)^{\alpha}$, $x > 0$. $E[X] = \frac{\theta}{\alpha - 1}$. $E[X \wedge x] = \frac{\theta}{\alpha - 1}\left\{1 - \left(\frac{\theta}{\theta+x}\right)^{\alpha - 1}\right\}$.

$$G(x) = \frac{E[X \wedge x] - x\,S(x)}{E[X]} = \frac{\frac{\theta}{\alpha-1}\left\{1 - \left(\frac{\theta}{\theta+x}\right)^{\alpha-1}\right\} - x\,S(x)}{\theta / (\alpha-1)} = 1 - \left(\frac{\theta}{\theta+x}\right)^{\alpha-1} - (\alpha-1)\frac{x}{\theta}\,S(x).$$

Let $t = F(x) = 1 - \left(\frac{\theta}{\theta+x}\right)^{\alpha}$. $\Rightarrow \left(\frac{\theta}{\theta+x}\right)^{\alpha} = S(x) = 1 - t$. Also, $x/\theta = (1 - t)^{-1/\alpha} - 1$.[290]

Therefore, $G(t) = 1 - (1 - t)^{(\alpha-1)/\alpha} - (\alpha-1)\{(1 - t)^{-1/\alpha} - 1\}(1 - t) = t + \alpha - t\alpha - \alpha\,(1-t)^{1-1/\alpha}$, $0 \le t \le 1$.
Comment: $G(0) = \alpha - \alpha = 0$. $G(1) = 1 + \alpha - \alpha - 0 = 1$.]

Here is graph comparing the Lorenz curves for Shifted Pareto Distributions with $\alpha = 2$ and $\alpha = 5$:

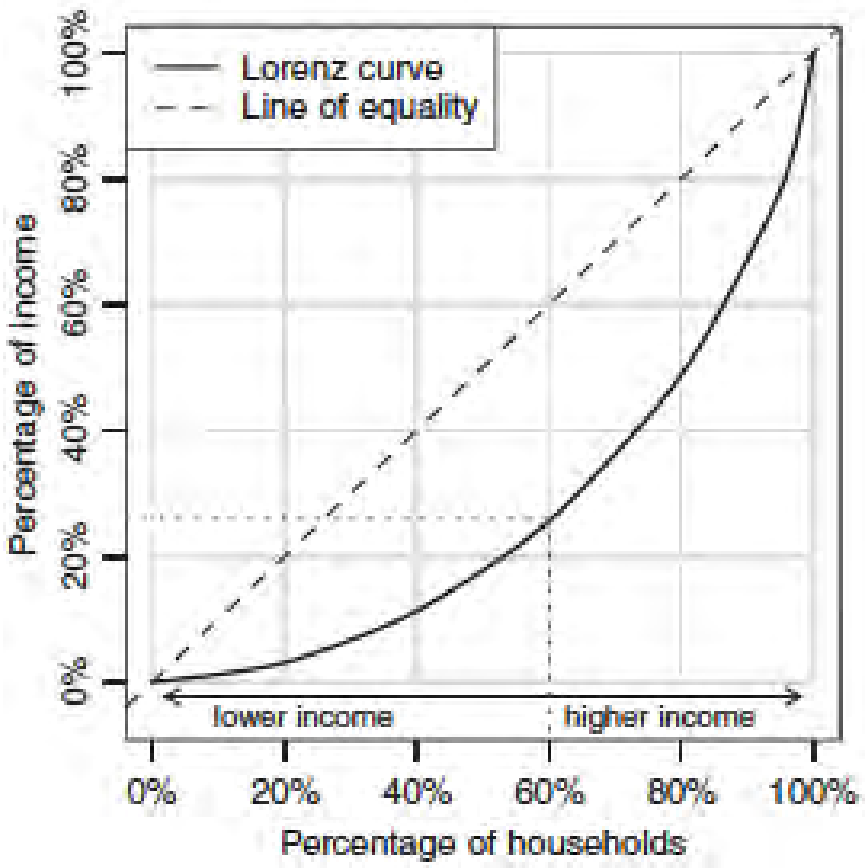


[289] What Bahnemann calls a Shifted Pareto, Loss Models simply calls a Pareto.
[290] This is just the VaR (value at risk) formula for the Shifted Pareto Distribution.

The Shifted Pareto with $\alpha = 2$ has a heavier righthand tail than the Shifted Pareto with $\alpha = 5$. If incomes follow a Shifted Pareto with $\alpha = 2$, then there are more extremely high incomes compared to the mean, than if incomes follow a Shifted Pareto with $\alpha = 5$. In other words, if $\alpha = 2$, then income is more concentrated in the high income individuals than if $\alpha = 5$.[291]

The Lorenz curve for $\alpha = 2$ is below that for $\alpha = 5$. In general, the lower curve corresponds to a higher concentration of income. In other words, a higher concentration of income corresponds to a smaller area under the Lorenz curve. Equivalently, a higher concentration of income corresponds to a larger area between the Lorenz curve and the 45° reference line.

Here is a Lorenz Curve for United States 2014 Household Income:[292]



The Gini index is calculated as twice the area between the Lorenz curve and the line of equality. In this case, the Gini index is 48.0%.

[291] An Exponential Distribution has a lighter righthand tail than either Shifted Pareto. Thus if income followed an Exponential, it would less concentrated than if it followed any Shifted Pareto.

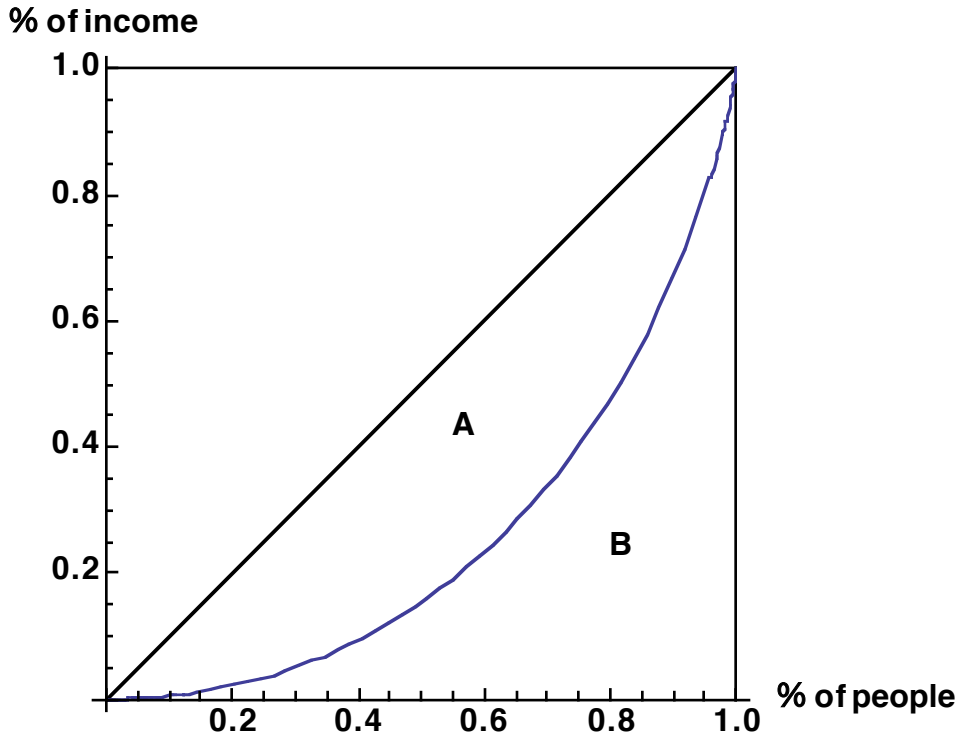
[292] See Figure 21 of Goldburd, Khare, and Tevet.

Gini Index:[293]

The Gini Index comes up for example in economics, when looking at the distribution of incomes. A subsequent section will discuss how the Gini index can be used to evaluate a rating plan.

The Gini index is a measure of inequality. For example if all of the individuals in a group have the same income, then the Gini index is zero. As incomes of the individuals in a group became more and more unequal, the Gini index would increase towards a value of 1. The Gini index has found application in many different fields of study.

As discussed, for incomes, the Lorenz curve would graph percent of people versus percent of income. This correspondence between areas on the graph of the Lorenz curve the concentration of income is the idea behind the Gini index. Let us label the areas in the graph of a Lorenz Curve:



$$\text{Gini Index} = \frac{\text{Area A}}{\text{Area A + Area B}}.$$

However, Area A + Area B add up to a triangle with area 1/2.

$$\text{Therefore, Gini Index} = \frac{\text{Area A}}{\text{Area A + Area B}} = 2A$$

= twice the area between the Lorenz Curve and the line of equality = 1 - 2B.

---

[293] See Section 7.2.4 of Goldburd, Khare, and Tevet.
Also called the Gini Coefficient or coefficient of concentration.

Gini Index for Specific Distributions:[294]

For the Exponential Distribution, the Lorenz curve was: $G(t) = t + (1-t) \ln(1-t)$.

Thus, Area B = area under Lorenz curve = $\int_0^1 t + (1-t) \ln(1-t)\, dt = 1/2 + \int_0^1 s \ln(s)\, ds$.

Applying integration by parts,

$\int_0^1 s \ln(s)\, ds = (s^2/2) \ln(s) \Big]_{s=0}^{s=1} - \int_0^1 (s^2/2)(1/s)\, ds = 0 - 1/4 = -1/4.$

Thus Area B = 1/2 - 1/4 = 1/4.

Therefore, for the Exponential Distribution, the Gini Index is: 1 - (2)(1/4) = 1/2.

For the Uniform Distribution, the Gini Index is: 1/3.

For the Shifted Pareto Distribution, the Gini Index is: $1 / (2\alpha - 1)$, $\alpha > 1$.

We note that the Uniform with the lightest righthand tail of the three has the smallest Gini index, while the Shifted Pareto with the heaviest righthand tail of the three has the largest Gini index. Among Shifted Pareto Distributions, the smaller alpha, the heavier the righthand tail, and the larger the Gini index.[295]
The more concentrated the income is among the higher earners, the larger the Gini index.

For the Classical (Single Parameter) Pareto Distribution, the Gini Index is: $1 / (2\alpha - 1)$, $\alpha > 1$.

For the LogNormal Distribution, the Gini Index is: $2\Phi[\sigma/\sqrt{2}] - 1$.

For the Gamma Distribution, the Gini Index is: $1 - 2\, \beta(\alpha+1, \alpha; 1/2)$.

---

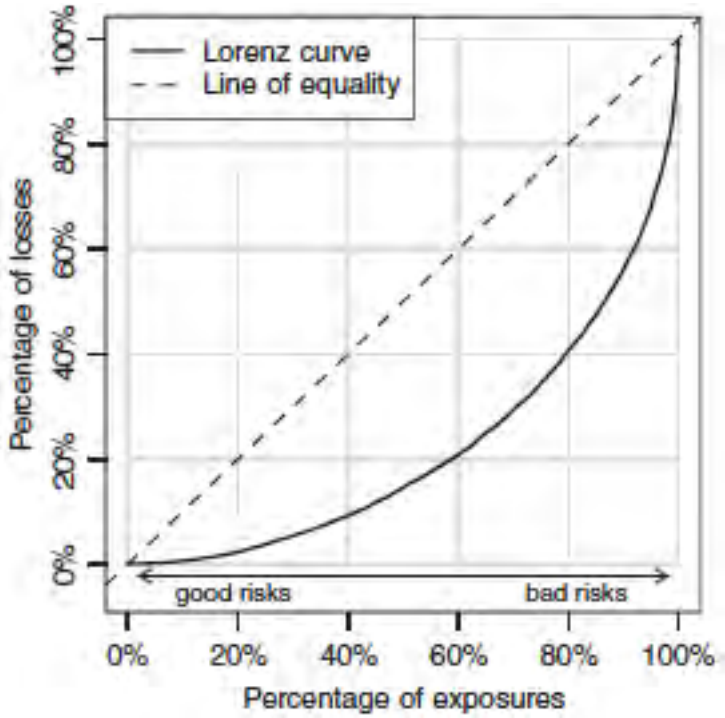
[294] Not on the syllabus.
[295] As alpha approaches one, the Gini coefficient approaches one.

Gini Index and Rating Plans:[296]

The Gini index can also be used to measure the lift of an insurance rating plan by quantifying its ability to segment the population into the best and worst risks. Assume we have a rating plan. Ideally we would want the model to identify those insureds with higher expected pure premiums.

The Lorenz curve for the rating plan is determined as follows:
1. Sort the dataset based on the model predicted loss cost.[297]
2. On the x-axis, plot the cumulative percentage of exposures.
3. On the y-axis, plot the cumulative percentage of actual losses.
Draw a 45-degree line connecting (0, 0) and (1, 1), called the line of equality.

Here is an example:[298]



This model identified 60% of exposures which contribute only 20% of the total losses. **The Gini index is twice the area between the Lorenz curve and the line of equality**, in this case 56.1%. **The higher the Gini index, the better the model is at identifying risk differences**.[299]

---

[296] See Section 7.2.4 of Goldburd, Khare, and Tevet.
[297] This should be done on a dataset not used to develop the rating plan.
[298] See Figure 21 of Goldburd, Khare, and Tevet.
[299] "Note that a Gini index does not quantify the profitability of a particular rating plan, but it does quantify the ability of the rating plan to differentiate the best and worst risks. Assuming that an insurer has pricing and/or underwriting flexibility, this will lead to increased profitability."

An Example of the Gini Index and an Insurance Rating Plan:[300]

We have four classes each with an equal number of exposures, and the result of fitting a GLM.[301]
We have already sorted the classes according to the pure premiums predicted by the GLM.[302]

Class	Predicted Pure Premium
1	100
2	200
3	300
4	400

Ignoring here any misestimating of the overall rate level, the observed pure premiums would differ from the predicted pure premiums for two reasons: [303] [304]
1. Imperfection of the GLM, in other words modeling error.
2. Random fluctuation, in other words process variance.[305]

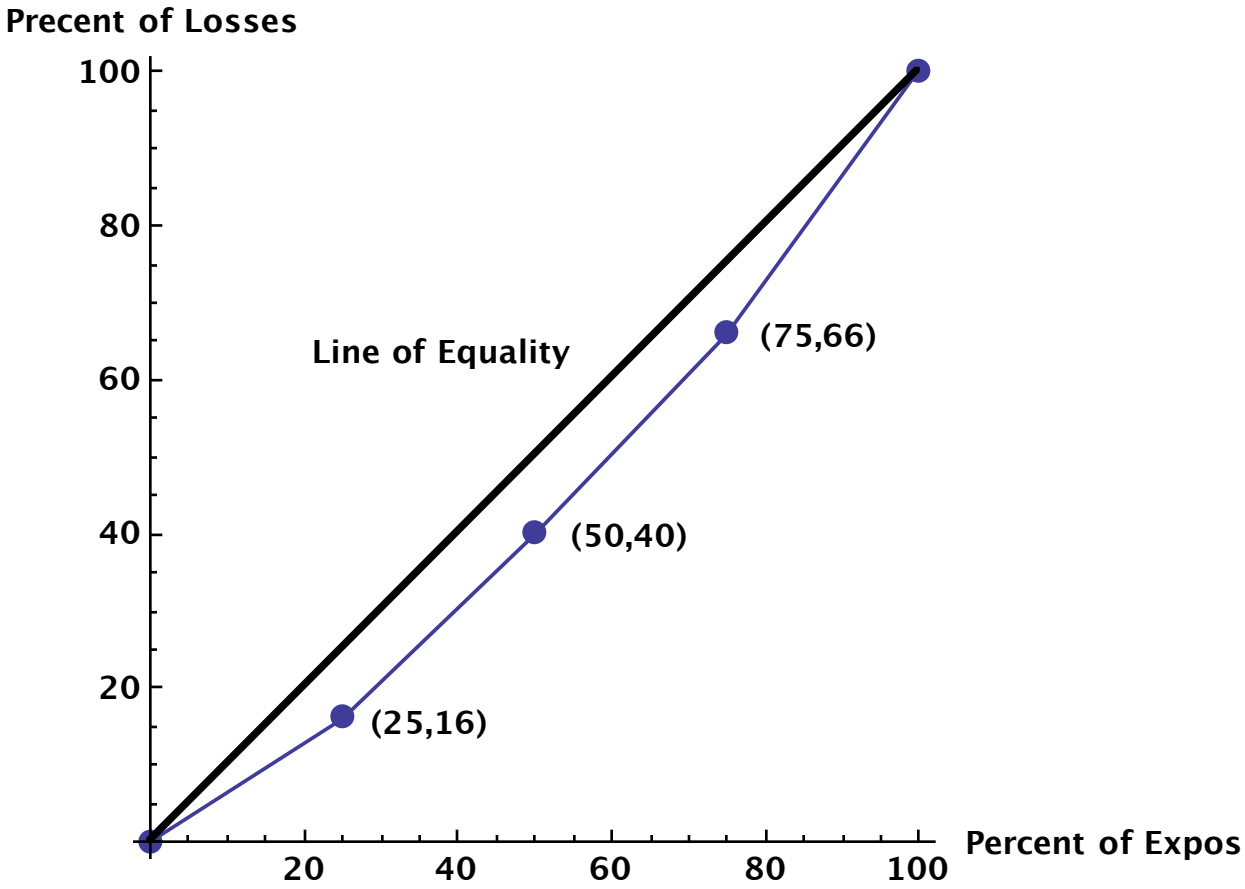
Let us assume the following Actual Pure Premiums:[306]

Class	Actual P.P.	Cumulative Losses[307]	% of Losses	% Expos
1	160	160	16%	25%
2	240	400	40%	50%
3	260	660	66%	75%
4	340	1000	100%	100%

Thus for the Lorenz curve we plot the points: (0, 0), (25, 16), (50, 40), (75, 66), (100, 100).

---

[300] See 8, 11/16, Q.5.
[301] I have chosen a one-dimensional example with only four levels solely for illustrative simplicity. Most GLMs would include more than one risk characteristic, and some characteristics would have more than four levels.
Also the exposures for each level would usually not all be equal.
[302] In a practical application we would have hundreds if not thousands of different cells consisting of risks with all of the same characteristics and thus the same predicted pure premium.
[303] We are using the GLM to predict class relativities rather than the overall rate level.
In some cases, the GLM output will automatically balance to the observed.
[304] Each class is not perfectly homogenous; it may be possible to refine the given classes to produce more homogeneous classes. Of course, if the classes are made too small, we would have issues with credibility.
[305] The more data in a class, the less subject to random fluctuation would be the average observed pure premium for that class.
[306] These observed pure premiums are from a dataset similar to the one to which the GLM was fit.
[307] Assuming solely for simplicity one exposure per class.



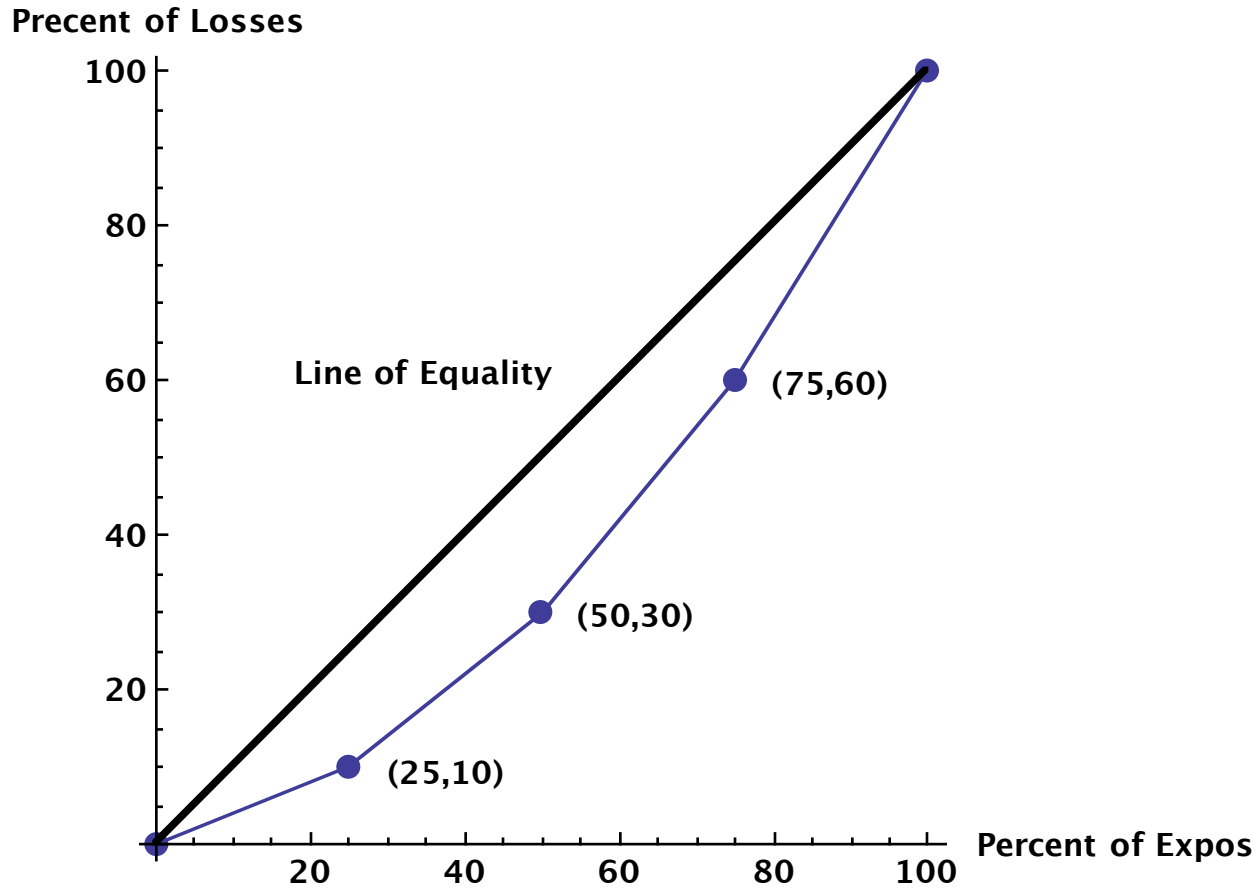
It is possible to calculate the area between the above Lorenz Curve and the Line of Equality, by dividing the area in triangles.[308] [309] This area turns out to be 0.07.[310]
Thus the Gini Index is twice that or 14%.[311]

[308] You will not be asked to do so on your exam!
[309] The six triangles I used were: {(0,0), (25,16), (25,25)}, {{25,16}, {25,25}, {50,40)}, ...
One can calculate the area of a triangle from the length of the sides via Heron's formula, not on the syllabus.
[310] Remembering that for example the value shown as 25 is actually 25% = 0.25.
[311] The higher the Gini index, the better the model is at identifying risk differences.
A more complicated model is likely to do better than this very simple class plan.

Solely for illustrative purposes, let us investigate the Gini Index if instead the actual pure premiums exactly matched the predicted pure premiums for each class.[312]

Class	Actual P.P.	Cumulative Losses[313]	% of Losses	% Expos
1	100	100	10%	25%
2	200	300	30%	50%
3	300	600	60%	75%
4	400	1000	100%	100%

Thus for the Lorenz curve we plot the points: (0, 0), (25, 10), (50, 30), (75, 60), (100, 100).



The area between the above Lorenz Curve and the Line of Equality, turns out to be 0.125. Thus the Gini Index is twice that or 25%, higher than previously.

[312] While this is will not occur in practice, this is the best possible result for this simple plan with only four classes.
[313] Assuming solely for simplicity one exposure per class.

Understanding & Validating a Model:[314]

Model Lift
How well does the model differentiate between best and worst risks?
Does the model help prevent adverse selection?
Is the model better than the current rating plan?

Simple Quantile plots:
Illustrate how well the model helps prevent adverse selection.
Double lift charts:
Compare competing models or compare new model against current rating plan.
Gini Index:
Summarizes model lift into one number.
Loss ratio charts:
Puts lift in a context most people in the insurance industry can understand.

Goodness of Fit
What kind of model statistics are available, and how do you interpret them?
What kind of residual plots should you consider, and how do you interpret them?
What are some considerations regarding actual versus predicted plots?

Internal Stability
How well does the model perform on other data?
How will the model perform over time?
How reliable are the model's parameter estimates?

[314] "And The Winner Is…? How to Pick a Better Model," 2015 CAS RPM Seminar, by Hernan L. Medina.

ROC Curves:[315]

Receiver Operating Characteristic (ROC) Curves can be used to compare models that use the Bernoulli or Binomial Distribution.[316]

The first step is to pick a threshold. For example, if the discrimination threshold were 8%, then we look at all cells with the fitted probability of an event > 8%, in other words $q_i > 8\%$.[317] Then we count up the number of times there was an event when an event was predicted. For example, there might be 3740 such true positives. Assume that there 4625 total events. Then the "sensitivity" is the ratio: 3740/4625 = 0.81.

In general, **above a given threshold, the sensitivity is the portion of the time that an event was predicted by the model out of all the times there is an event** = $\frac{\textbf{true positives}}{\textbf{total times there is an event}}$. Sensitivity is the rate of true positives.[318]
All other things being equal, higher sensitivity is good.

Then we look at all cells with the fitted probability of an event ≤ 8%, in other words $q_i \le 8\%$.
For example, there might be 54,196 such policies without an event. Assume there are a total of 63,232 policies without an event. Then the "specificity" is the ratio: 54,196/63,232 = 0.85.

Below a given threshold, the specificity is the portion of the time that an event was not predicted by the model out of all of the times these is <u>not</u> an event = $\frac{\textbf{true negatives}}{\textbf{total times there is not an event}}$.[319] All other things being equal, higher specificity is good.
Specificity is the rate of true negatives. 1 - specificity is the rate of false positives.

For this example, for a threshold of 8%, we can display the information in a **confusion matrix**:[320]

Discrimination Threshold: 8%

	Predicted		
Actual	Event	No Event	Total
Event	3740	884	4625
No Event	9036	54,196	63,232
Total	12,776	55,080	67,856

---

[315] See Section 7.3.1 in Goldburd, Khare, and Tevet.
[316] ROC analysis was originally developed during World War II for the analysis of radar images.
[317] The event could be a claim, a policy renewal, etc.
[318] If one has a model to predict the probability of a claim being fraudulent, then for a given threshold: Sensitivity = (correct predictions of fraud) / (total number of fraudulent claims).
[319] If one has a model to predict the probability of a claim being fraudulent, then for a given threshold: Specificity = (correct predictions of no fraud) / (total number of non-fraudulent claims).
[320] See Table 13 in Goldburd, Khare, and Tevet.

The general form of a confusion matrix:

	Predicted	
Actual	Event	No Event
Event	true positive	false negative
No Event	false positive	true negative

A confusion matrix is similar to a table from hypothesis testing, where the null hypothesis is no event:[321]

Decision	Reject H_0	Do not reject H_0
H_1 is True	Correct	Type II Error
H_0 is True	Type I Error	Correct

The false negatives are analogous to making a Type II Error.
The false positives are analogous to making a Type I Error.

	Predicted		
Actual	Event	No Event	Total
Event	3740	884	4625
No Event	9036	54,196	63,232
Total	12,776	55,080	67,856

For the 8% threshold, the specificity was: $\frac{\text{true negatives}}{\text{total times there is not an event}} = \frac{54,196}{63,232} = 85\%.$

1 - specificity = $\frac{\text{false positives}}{\text{total times there is not an event}} = \frac{9036}{63,232} = 15\%.$

1 - specificity is analogous to:
chance of making a Type Error I = significance level of a statistical test.

For the 8% threshold, the sensitivity was: $\frac{\text{true positives}}{\text{total times there is an event}} = \frac{3740}{4625} = 81\%.$

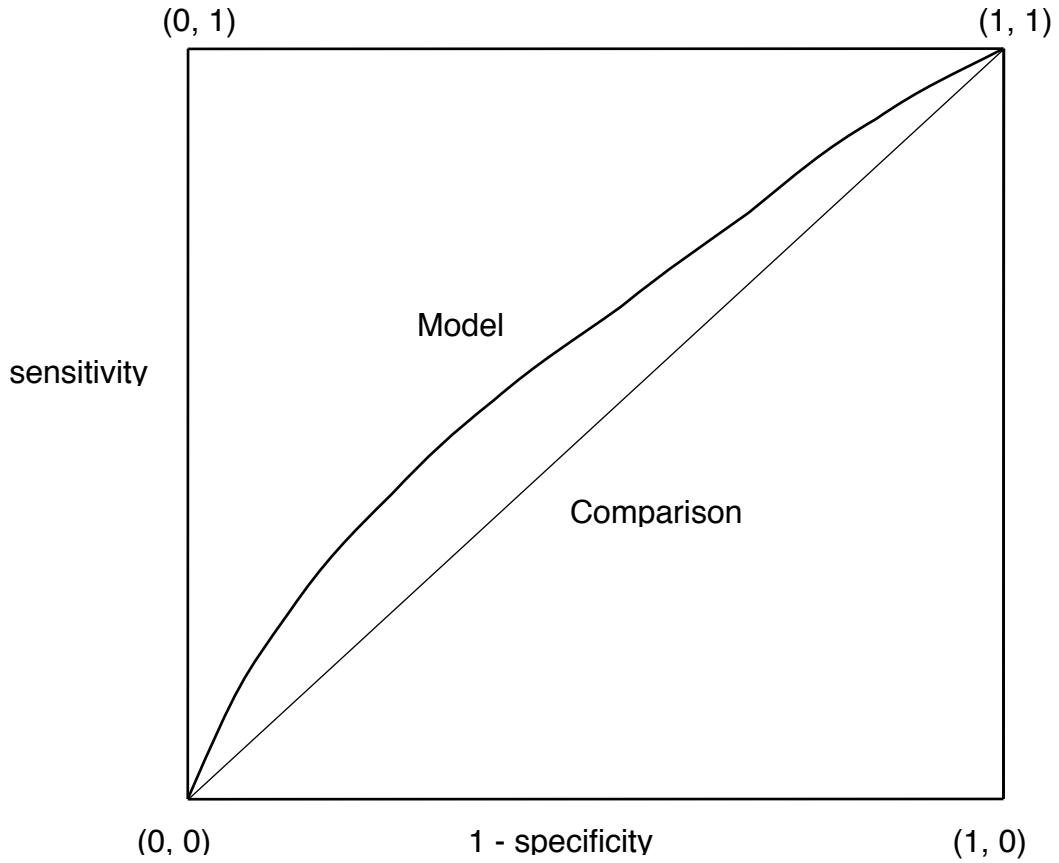
Sensitivity is analogous to: 1 - chance of making a Type Error II = power of a statistical test.

In the ROC Curve we plot the point: (1 - 0.85, 0.81) = (0.15, 0.81).

In general, **the ROC curve consists of plotting for various thresholds:**
(1 - specificity , sensitivity).
In addition, there is a 45% comparison line, the line of equality, from (0, 0) to (1, 1).

[321] While the analogy to hypothesis testing may help your understanding, it should not be tested on your exam.

Here is an example of an ROC curve:[322]



A perfect model would be at (0, 1) in the upper lefthand corner; sensitivity = 1 and specificity = 1. The closer the model curve gets to the upper lefthand corner the better.

The comparison line (line of equality) indicates a model with sensitivity = 1 - specificity, which can be achieved by just flipping a coin to decide your prediction. Thus such models have no predictive value. The closer the model curve gets to the 45 degree comparison line (line of equality), the worse the model.

The comparison line has area 1/2 below it. The larger the area under the model curve, the better it is.

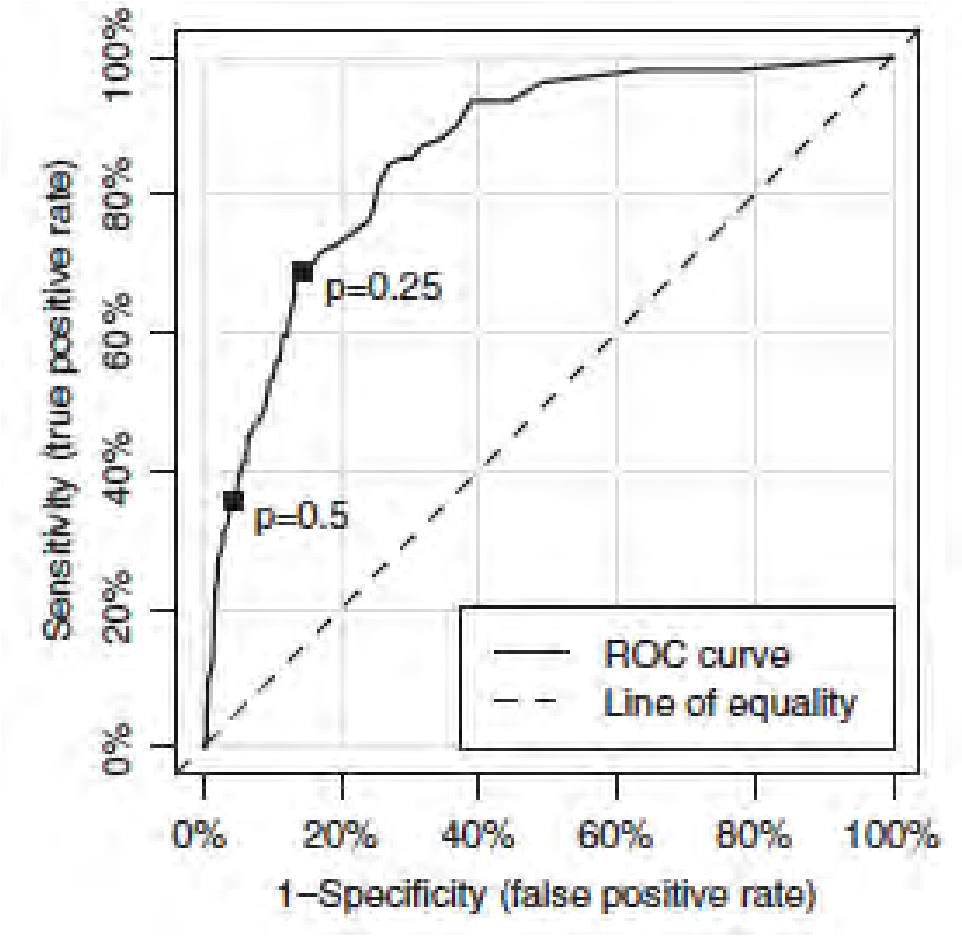
AUROC is the area under the ROC curve; the larger AUROC the better the model.[323]

---

[322] I just drew my diagram; it does not come from an actual fitted GLM.
Similar to Figure 22 in Goldburd, Khare, and Tevet.

[323] The AUROC is equal to: (0.5) (normalized Gini) + 0.5, where the normalized Gini is the ratio of the model's Gini index to the Gini index of the hypothetical "perfect" model (where each record's prediction equals its actual value). Note that the prefect model will not have a Gini index of one; it's Gini index depends on the homogeneity of the risks and the randomness of the loss process.

For an example of modeling fraud on claims, the syllabus reading has the following ROC Curve:[324]



This ROC has an area under ROC (AUROC) of 0.857.[325]

We can see how as one changes the threshold from 0.5 to 0.25, the sensitivity increases, but at the cost a lower specificity. In other words, the rate of true positives increases at the cost of also increasing the rate of false positives.[326]

The selection of the discrimination threshold involves a trade-off: a lower threshold will result in more true positives and fewer false negatives than a higher threshold, but at the cost of more false positives and fewer true negatives.

---

[324] See Figure 22 in Goldburd, Khare, and Tevet.

[325] The perfect model would have an AUROC of 1.

[326] "The ROC curve allows us to select a threshold we are comfortable with after weighing the benefits of true positives against the cost of false positives. Different thresholds may be chosen for different claim conditions; for example, we may choose a lower threshold for a large claim where the cost of undetected fraud is higher. Determination of the optimal threshold is typically a business decision that is out of the scope of the modeling phase."

A Medical Example of ROC:[327]

Let us assume we have a medical test for a disease which results in a numerical score. The lower the score on this test the more likely that the individual has this disease.[328] Assume the following data:

Score on Medical Test	Number with Disease		Number without Disease
5 or less	18		1
5.1 to 7	7	17	
7.1 to 9	4	36	
9 or more	3		39
Total	32		93

We can pick a threshold to use with this test; if the test score is less than or equal to the chosen threshold this indicates that the individual has the disease.

For example, assume a threshold of 5. Then 18 individuals are correctly identified as diseased, and 1 is incorrectly identified as diseased. There are 18 true positives. There is one false positive. 14 individuals who are diseased are incorrectly identified as being without disease. There are 14 false negatives. 92 individuals who are not diseased are correctly identified as being without disease.

We can think of sensitivity as the rate of true positives of a medical test for a disease as a portion of
positives. The rate of true positives out of all diseased is: 18/32 = 0.56.[329]

We can think of specificity as the rate of individuals that the test indicates do not have the disease out of those without the disease. The rate of negatives out of those without the disease: 92/93 = 0.99. One minus the specificity, 1%, is the rate of false positives out of those without the disease.[330]

The confusion matrix is:
Discrimination Threshold: 5

	Predicted		
Actual	Disease	No Disease	Total
Disease	18	1	19
No Disease	14	92	106
Total	32	93	125

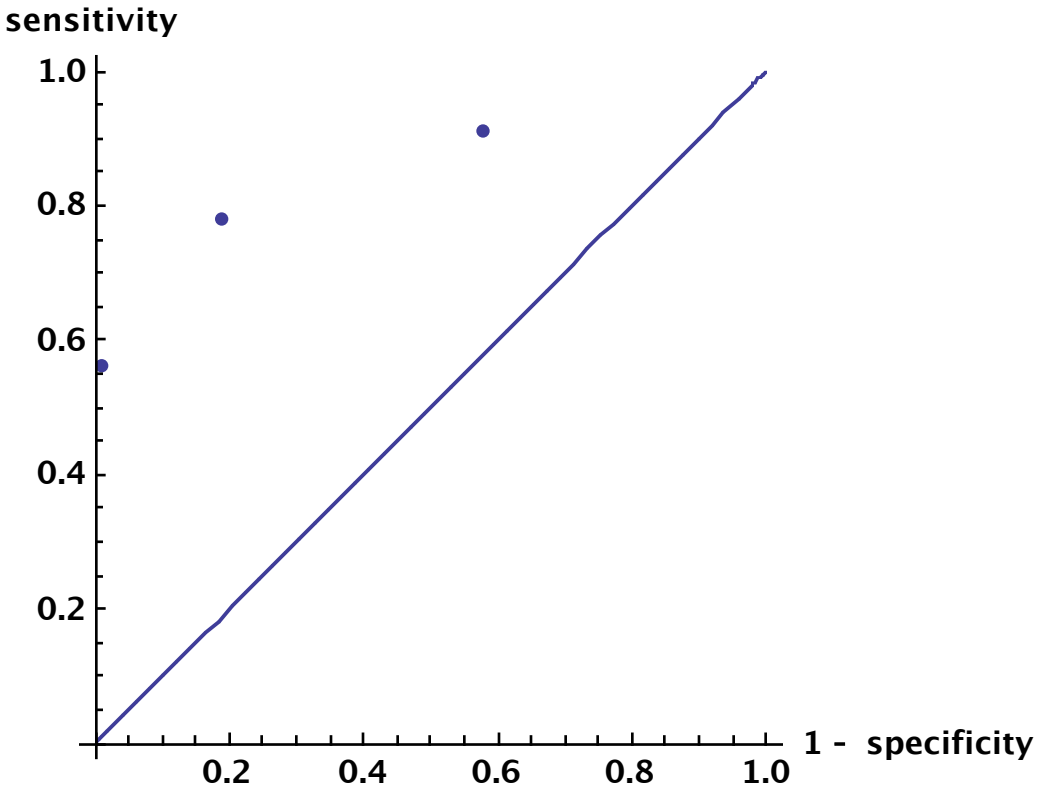
[327] http://gim.unmc.edu/dxtests/ROC1.htm
[328] While a low test score indicates the presence of the disease in this example, it could have been the reverse.
[329] Sensitivity is analogous to the probability of rejecting the null hypothesis (healthy) when it is false, which is the power of the test.
[330] One minus specificity is analogous to the probability of rejecting the null hypothesis (healthy) when it is true, which is the significance level of the test.

Exercise: What are the sensitivity and specificity if one instead uses a threshold of 7?
[Solution: 25 people have positive tests out of 32 with the disease. ⇒ sensitivity is: 25/32 = 0.78.
75 people have negative tests out of 93 who are healthy. ⇒ specificity is: 75/93 = 0.81.
Comment: With a higher threshold the sensitivity is higher but the specificity is lower.
There is a tradeoff between a high sensitivity and a high specificity.]

Exercise: What are the sensitivity and specificity if one instead uses a threshold of 9?
[Solution: 29 people have positive tests, out of 32 with the disease. ⇒ sensitivity is: 29/32 = 0.91.
39 people have negative tests out of 93 who are healthy. ⇒ specificity is: 39/93 = 0.42.]

Threshold	Sensitivity	Specificity	1 - Specificity
5	0.56	0.99	0.01
7	0.78	0.81	0.19
9	0.91	0.42	0.58

The corresponding ROC curve, where I have not connected the dots:[331]



[331] The area under the curve measures discrimination, that is, the ability of the test to correctly classify those with and without the disease.

Other Topics:[332]

The syllabus reading discusses three additional topics:
- Why you probably should not model coverage options with GLMs.
- Why territories are not a good fit for the GLM framework.
- Ensembling.

Coverage Options:[333]

Insureds can choose coverage options such as deductible amount or limit of liability.[334]
There are corresponding deductible credits or increased limits factors.[335]
You probably should not model the rating factors for coverage options with GLMs.

For example, a GLM might indicate that one should charge more for a higher deductible. There may be something systematic about insureds with higher deductibles that may make them a worse risk relative to others in their class.[336] In which case, the coefficients estimated by the GLM are reflecting some of this increased risk due to antiselection effects.

To the extent that the factor indicated by the GLM differs from the pure effect on loss potential, it will affect the way insureds choose coverage options in the future. Thus, the selection dynamic will change and the past results would not be expected to be replicated for new policies.

Thus factors for coverage options should be estimated outside the GLM, using traditional actuarial techniques.[337] The resulting factors should then be included in the GLM as an offset, as has been discussed previously.

---

[332] See Section 8 of Goldburd, Khare, and Tevet.
[333] See Section 8.1 of Goldburd, Khare, and Tevet.
[334] These can be distinguished from characteristics of the insured.
[335] In general, the insured should pay less for less coverage and more for more coverage.
[336] "The choice of high deductible may be the result of a high risk appetite on the part of an insured, which would manifest in other areas as well. Alternately, the underwriter, recognizing an insured as a higher risk, may have required the policy to be written at a higher deductible."
[337] This is the recommendation of Goldburd, Khare, and Tevet.
Even if the final factors for coverage options are not estimated within the GLM, I think the results of including coverage options in a GLM may reveal something interesting and potentially important to the actuary.

Territory Modeling:[338]

Territories are not a good fit for the GLM framework.

There may have hundreds of territories, which requires many levels in the GLM. Therefore, the authors recommend the use of other techniques, such as spatial smoothing, to model territories.[339]

One should include the territory relativities produced by the separate model as an offset in the GLM used to determine classification relativities. Similarly, one should include classification relativities produced by the GLM as an offset in the separate model used to determine territory relativities.[340]
Ideally this should be an iterative process.[341]

Ensembling:[342]

Two (or more) teams model the same item; they build separate models working independently. The models are evaluated and found to be approximately equal in quality.

Combining the answers from both models is likely to perform better than either individually.[343] **A model that combines information from two or more models is called an ensemble model**.

A simple means of ensembling is to average the separate model predictions.[344] "Predictive models each have their strengths and weaknesses. Averaged together, they can balance each other out, and the gain in performance can be significant."

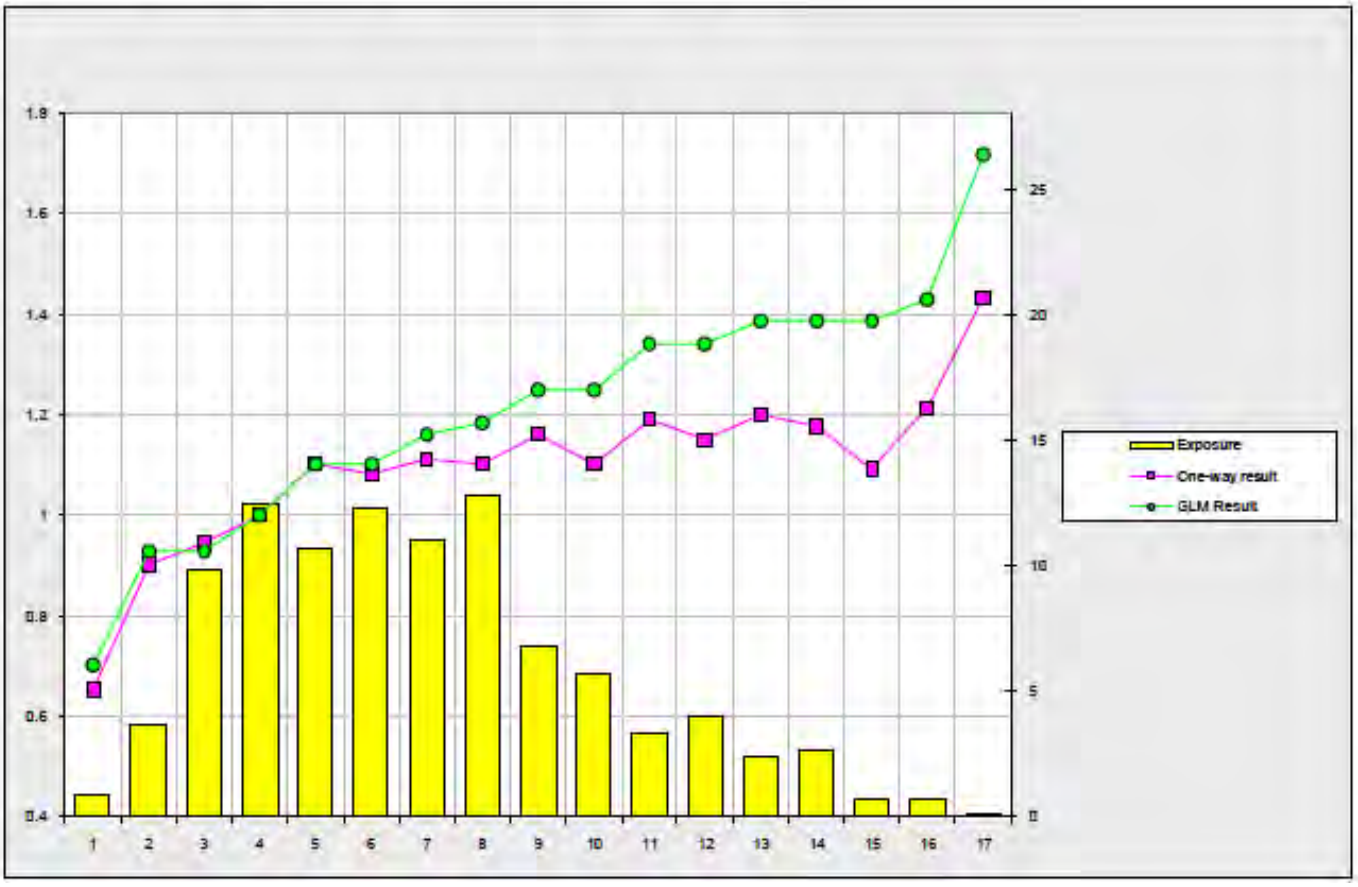
---

[338] See Section 8.2 of Goldburd, Khare, and Tevet.
I believe the authors are discussing determining territory relativities rather than constructing territories from smaller geographical units such as zipcode. However, they may be discussing doing both together.
[339] The authors do not discuss any details of spatial smoothing or other techniques.
[340] In determining territories one should adjust the pure premiums for a zipcode by its average class rating factor. Chapter 11 of Basic Ratemaking by Werner and Modlin have a discussion of determining territories.
[341] If they being updated at the same time, both models should be run, using the other as an offset, until they reach an acceptable level of convergence.
[342] See Section 8.3 of Goldburd, Khare, and Tevet.
[343] Of course it is costly to have two teams build two separate models.
"Done properly, though, ensembles can be quite powerful; if resources permit, it may be worth it."
[344] The authors do not discuss more complicated methods of ensembling.

Examples of GLM Output:[345]

Figure 10.1 shows private passenger automobile collision frequency by symbol.[346] [347] Rectangles represent the volume of exposures. The circles represent a fitted GLM, which includes many more variables than just symbol. So for example, symbol 10 is predicted to have a frequency about 25% higher than symbol 4, with all the other variables being considered.[348]

10.1 Effect of Vehicle Symbol on Automobile Collision Frequency



The squares represent the estimates of a univariate model that only includes symbol. We note that these relativities are significantly different than those from the GLM.

---

[345] Taken from Chapter 10 and Appendix F of Basic Ratemaking, on the syllabus of Exam 5.

[346] Automobiles have been assigned "symbols" for physical damage coverage for many decades.
All automobiles of a particular make and model have the same symbol.
Each symbol represents a group of vehicles that have been combined based on common characteristics (e.g., weight, number of cylinders, horsepower, cost).
See for example, http://www.iso.com/Products/VINMASTER/Physical-Damage-Rating-Symbols.html
The higher the symbol, the higher the expected pure premium, all else being equal.

[347] Everything is shown relative to symbol 4; symbol 4 has a (multiplicative) relativity of 1. Symbol 4, one of the symbols with a lot of exposures, has been chosen as the base symbol. Choosing as the base level one with lots of exposures makes the denominator of the relativity more stable.

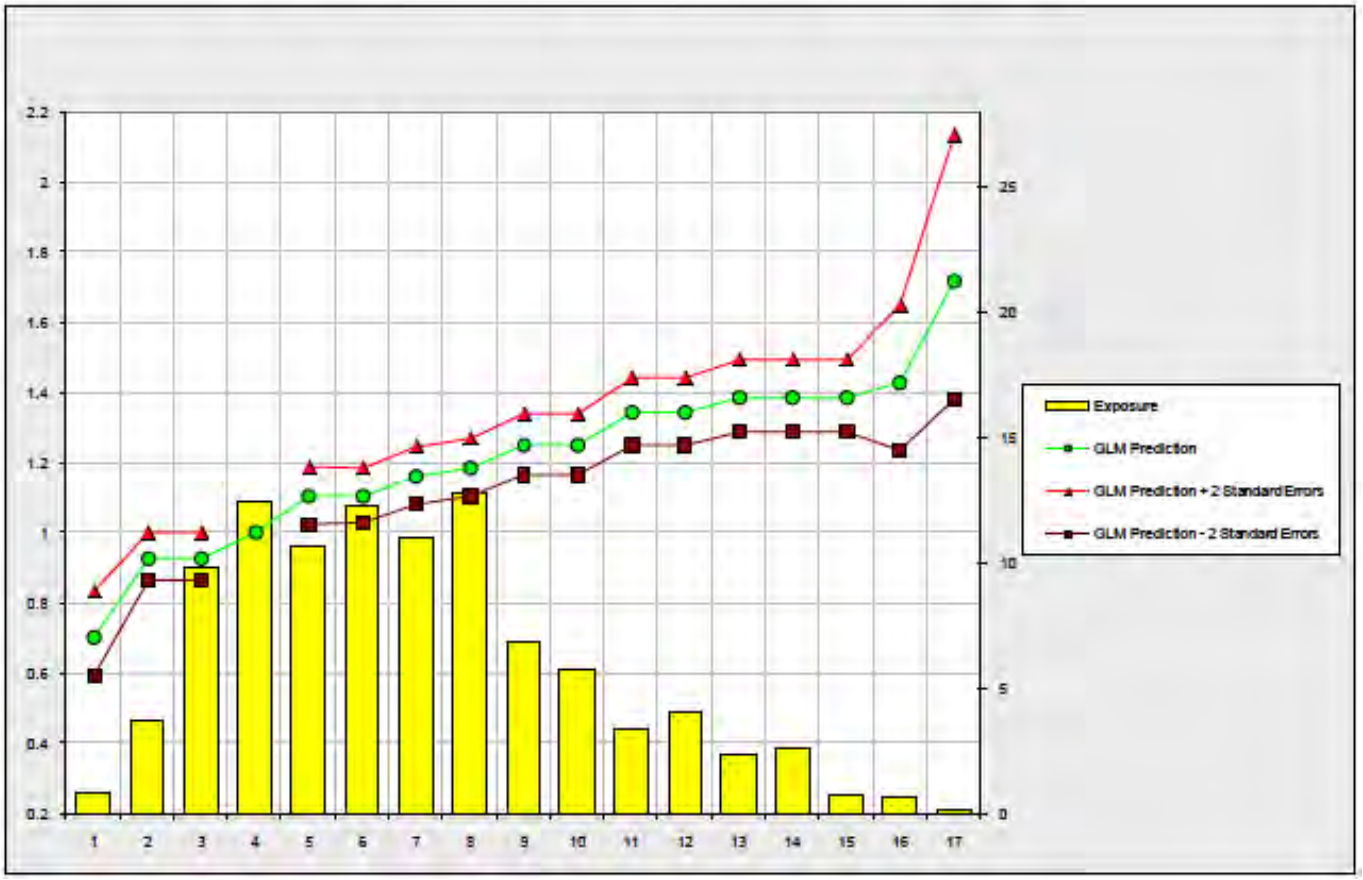
[348] This estimate depends somewhat on which other variables are included in the model.

The difference between the univariate estimates and those from the GLM are probably due to the correlation of symbol with another variable in the model. As discussed previously, univariate analyses can be distorted by such effects.

The GLM results of one variable such as symbol are only meaningful if the results for all other variables are considered at the same time. The indicated relativity of 1.25 for symbol 10 discussed previously will not be valid if variables are removed or added to the model. In other words, the indicated relativities for vehicle symbol are dependent on the other relativities being considered. Also the relativities for one variable usually depends somewhat on the levels of the other variables. The 1.25 relativity for symbol 10 is presumably with all the other variables at their base levels.

Figure 10.2 graphs the fitted relativities and ± 2 standard errors for this GLM.[349] [350]
It is an example of one common diagnostic.

10.2 Standard Errors for Effect of Vehicle Symbol on Automobile Collision Frequency

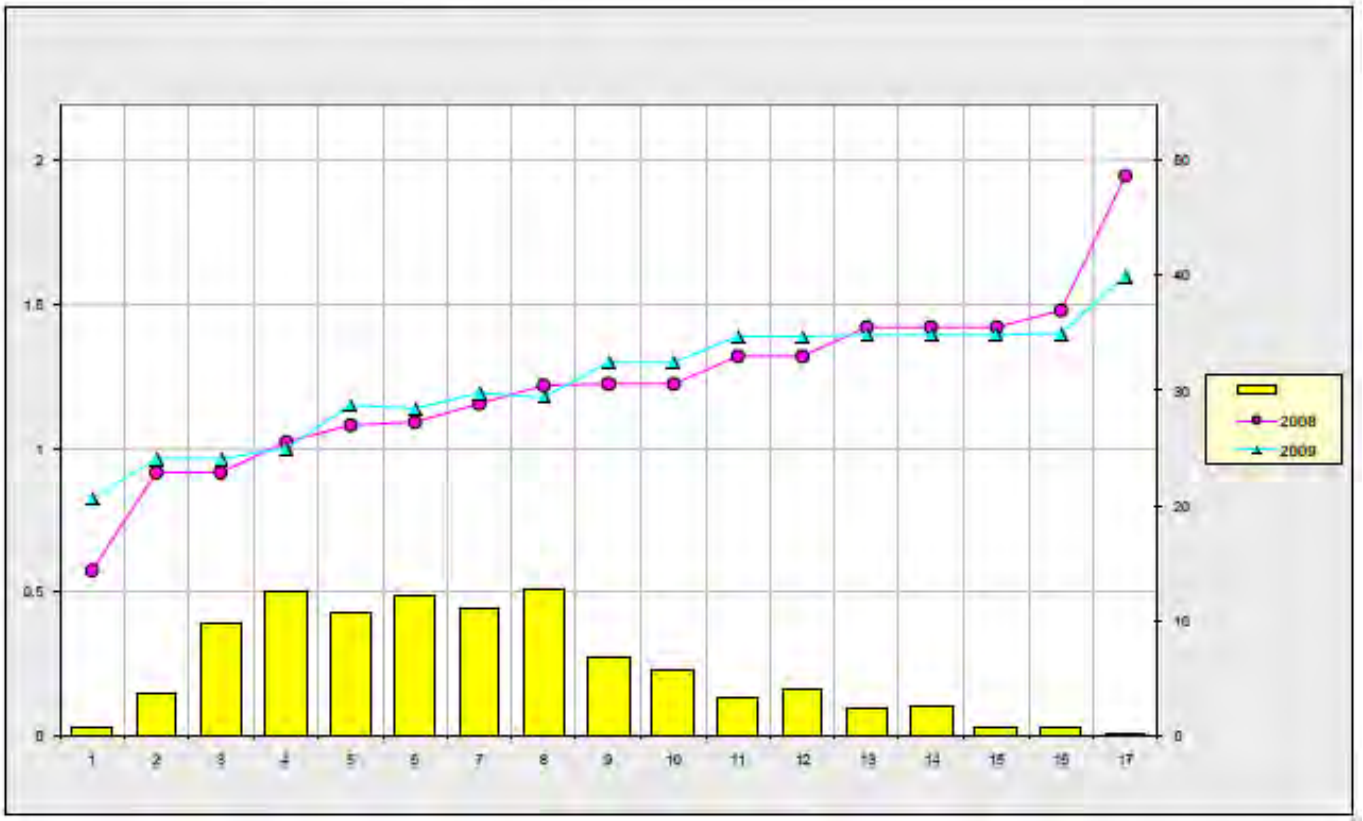


[349] This is presumably for all the other variables in the model at some base level.
[350] As discussed previously, we would expect the actual parameter to be within plus or minus two standard errors of the fitted parameter about 95% of the time.

The relativities go up with symbol. The error bars are relatively narrow, although they do get wider for the last few symbols, where there is not much data. Symbol seems to have a systematic effect on claim frequency.

Figure 10.3 shows the model fit to two separate years of data.[351]
We are interested in whether the model results are consistent based on the different years.

10.3 Consistency of Time for Vehicle Symbol



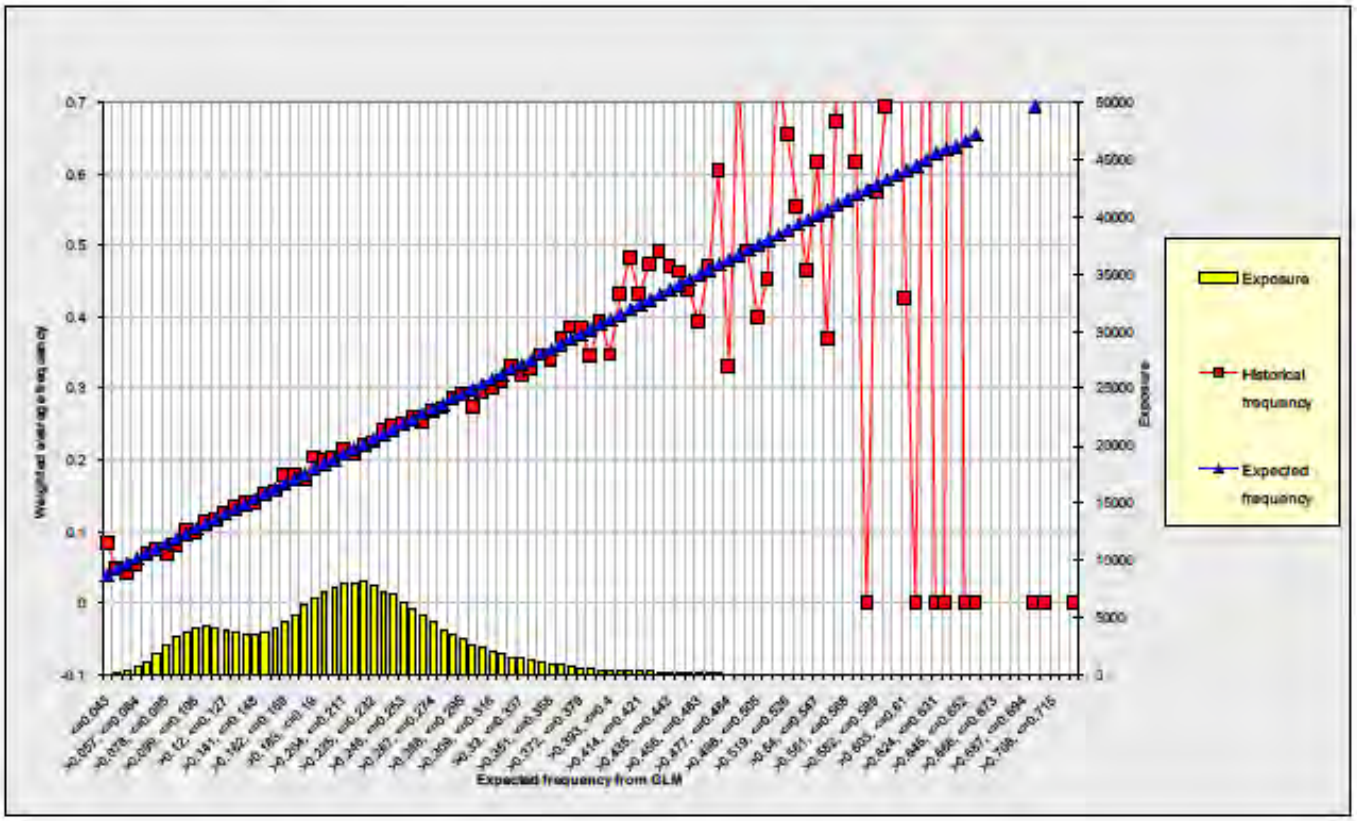
When one splits the original data into separate years like this, each model is based on less data than the original model, so we expect some more random fluctuation. In this case, the results are consistent between the two years, with the exception of symbol 17 where there is very little data. Again the model has been validated.

A key idea is that of a hold-out data set. We intentionally set aside a random portion of the original data, and do not use it to develop and calibrate the model. Then we see how well the model performs at predicting on this hold-out data set.[352]

[351] One could instead split the data into random subsets.
[352] Couret and Venter in their syllabus reading use a hold-out data to test their multidimensional credibility model.

Figure 10.4 compares the fitted model (triangles) to the observed frequency for the hold-out data set (squares). For each exposure in the hold-out data set, the model is used to calculate the expected frequency. Then the exposures in the hold out data set are grouped into intervals by expected frequency.[353] [354] For each of these groups we calculate the observed (historical) frequency.

10.4 Model Validation



In this example, the models performs well. The predicted matches the historical, until we get to the high predicted frequency groups, where there is very little volume, and thus lots of random fluctuation in the historical frequencies.

If in Figure 10.4, we had seen a bad match between the model and historical frequencies, then this might have indicated a model that was either underfit or overfit.
As discussed previously, the actuary wants to avoid both underfitting and overfitting models.

[353] Thus, we see the modeled frequencies (triangles) increase smoothly from left to right.
[354] For example, one group contains all exposures with expected frequencies > 14.1% and ≤ 14.8%.

For homeowners insurance it would be common to construct a GLM for each major peril for frequency and severity separately.[355] [356]

The first example models the frequency of claims for water damage on homeowners insurance. The GLM contains many variables, but here we concentrate on the effect of prior claim history.

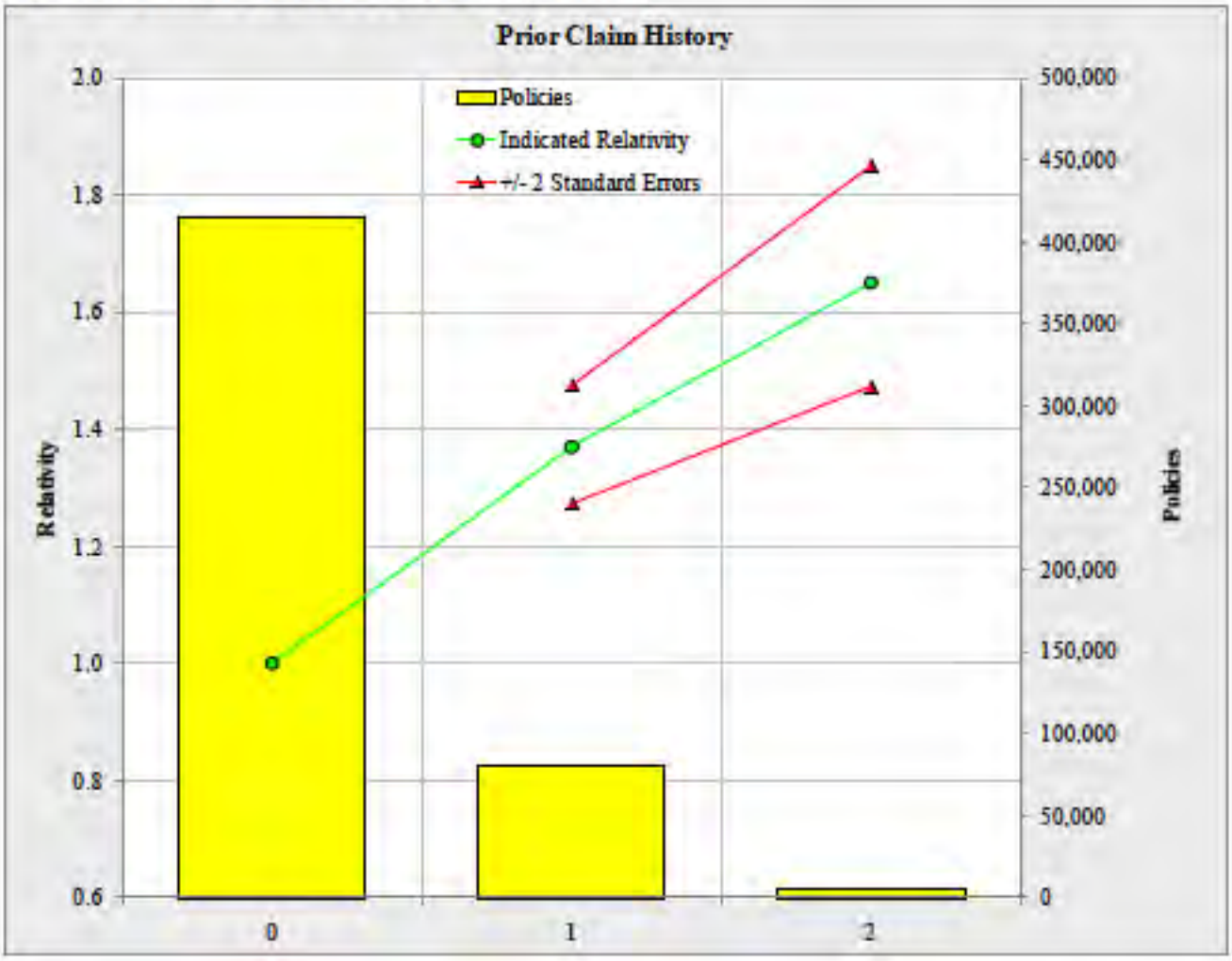
Policies are divided by the number of claims for some unspecified past experience period.[357] Each policy had either 0, 1, or 2 claims.[358] I assume that: each policy covers one home, and that renters and condominium policies are not included.[359]

Figure F.1 shows the fitted model and standard errors.[360] A standard error is the (estimate of) the standard deviation of the underlying errors for the model. If the errors were Normal, which they do not have to be for a GLM, then plus or minus 2 standard errors would cover about 95% probability.

---

[355] Severity has more random fluctuation than frequency, so it is usually harder to model.
[356] Perils would include: Fire, Theft, Wind, Vandalism, Water Damage, Liability, etc.
[357] I assume these are past claims for all perils. It is unclear what period of time is covered. Figure F.2 would lead one to believe that in Figure F.1 we are looking at four years of experience combined, 2011 to 2014.
The labeling of the years is just for illustrative purposes.
[358] While there may have been very few policies with more than 2 claims, they are not shown.
[359] None of these details are essential for interpreting the Figures and validating the model.
[360] This is presumably for all other variables in the model at some base level. Similar to Figure 10.2.

F.1 Main Effect Test for Prior Claim History



First, the fitted model makes sense. Those insureds with more claims in the past are predicted to have a higher expected frequency going forward. Compared to those with no claims, those with 1 prior claim are modeled to have a frequency relativity of about 1.37, in other words 37% more future expected claims from water damage than those with no past claims.

Besides the graph of the model relativities, also shown are ± 2 standard errors.[361] Those with 1 prior claim have a frequency relativity of between about 1.26 and 1.48. This is a relatively tight band, suggesting that it is OK to use prior claim history in the model. Not surprisingly, with a much smaller volume of data, the error bars for those with 2 claims are wider.

[361] As discussed previously, we would expect the actual parameter to be within plus or minus two standard errors of the fitted parameter about 95% of the time.

In general, we want a model that makes sense, and with relatively narrow error bars.

Figure F.2 breaks things down by policy year.[362]

F.2 Consistency Test for Prior Claim History

The models based on a single year each have a larger variance than the model based on all four years combined in Exhibit F.1. However, the lines each slope upwards with similar slopes; the pattern seems consistent over time.

In general, test the consistency of the model by comparing the results on separate subsets of the data base, such as separate years. In general, the actuary should use judgment to check the reasonableness of the results. In this case, it seems reasonable that more past claims would lead to a higher future expected frequency.[363]

---

[362] For the policies from each year, we use the same length of experience period as was used for the previous Exhibit F.1.

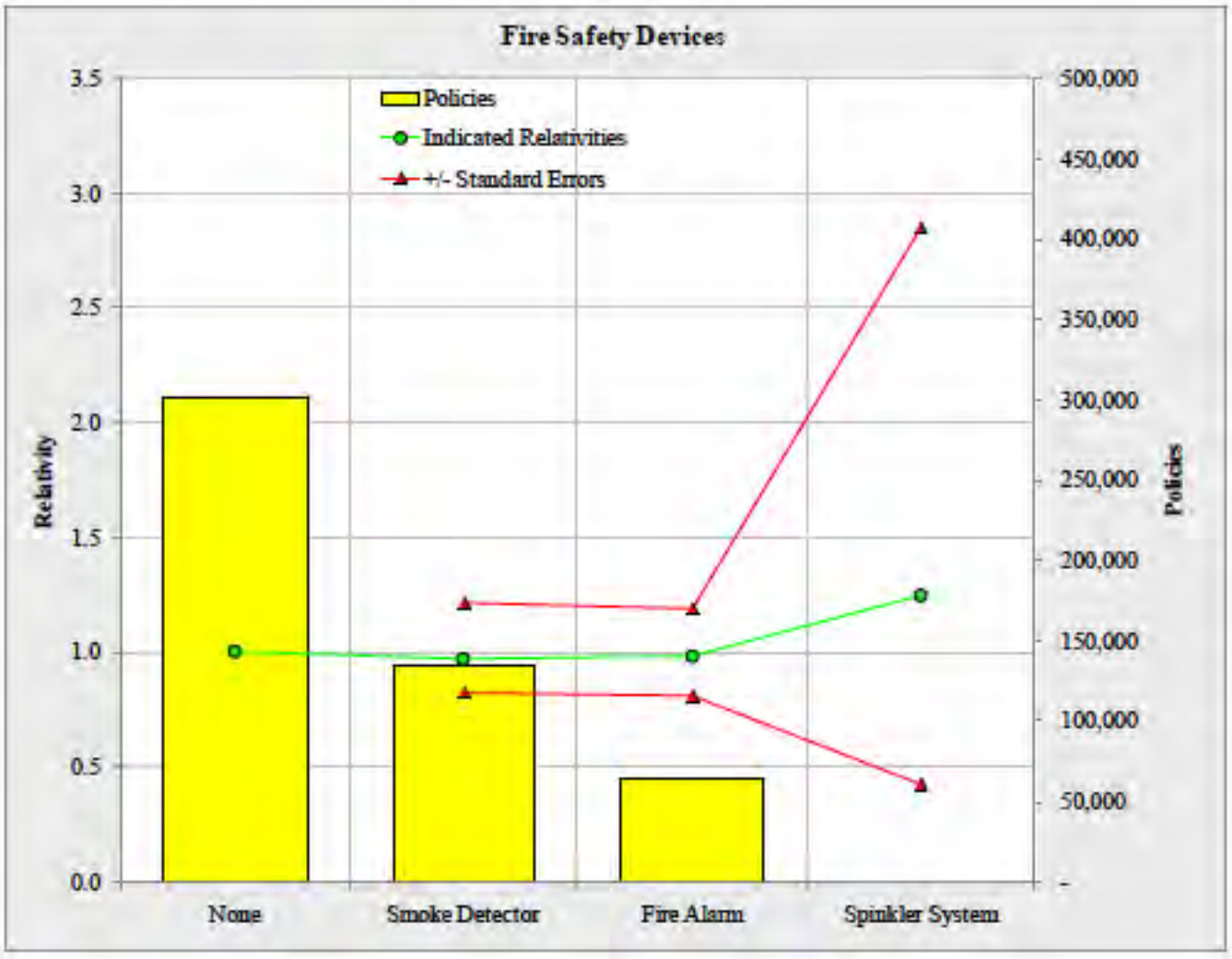
[363] This is the key reason why experience rating is worthwhile.

The second example, is similar to the first, except here we are attempting to predict the frequency of wind losses for homeowners. Again, even though there are many variables in the GLM, we are concentrating on just one, fire safety devices.

First, while we would expect fire safety devices to affect expected fire losses, most actuaries would not expect fire safety devices to significantly affect expected wind losses. In this case, the model does not seem reasonable based on judgement.

Figure F.3 is similar to Figure F.1 from the previous example.
Predicted wind frequency relativities are graphed versus four levels of fire safety device: None, Smoke Detector, Fire Alarm, and Sprinkler System.[364]

F.3 Main Effect Test for Fire Safety Device

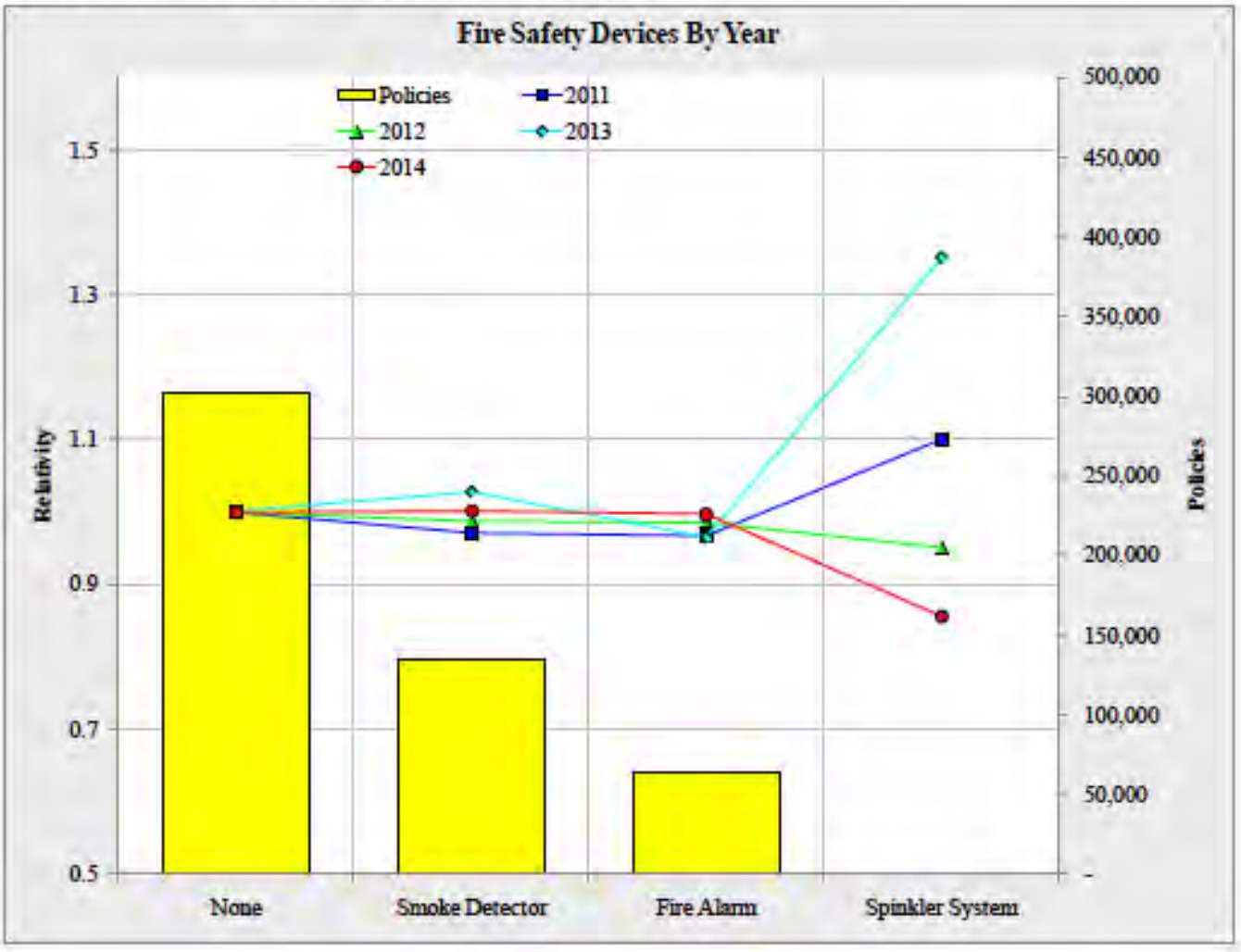


[364] Extremely few homes have a sprinkler system.

The indicated relativities are close to one, except for sprinklers. However, since there is so little data for sprinklers, its standard error is huge. We can conclude that the sprinkler relativity is very likely to be between about 0.45 and 2.9; in other words, this model tells us nothing useful about the relativity for sprinklers. The errors bars on the other relativities are consistent with a relativity of one. We conclude that fire safety devices have no predictive value in this model for frequency of wind losses.

Figure F.4 is similar to Figure F.2 from the previous example.
Again the results are shown for the model run on the data of each of four separate policy years.

F.4 Consistency Test for Fire Safety Device Claim



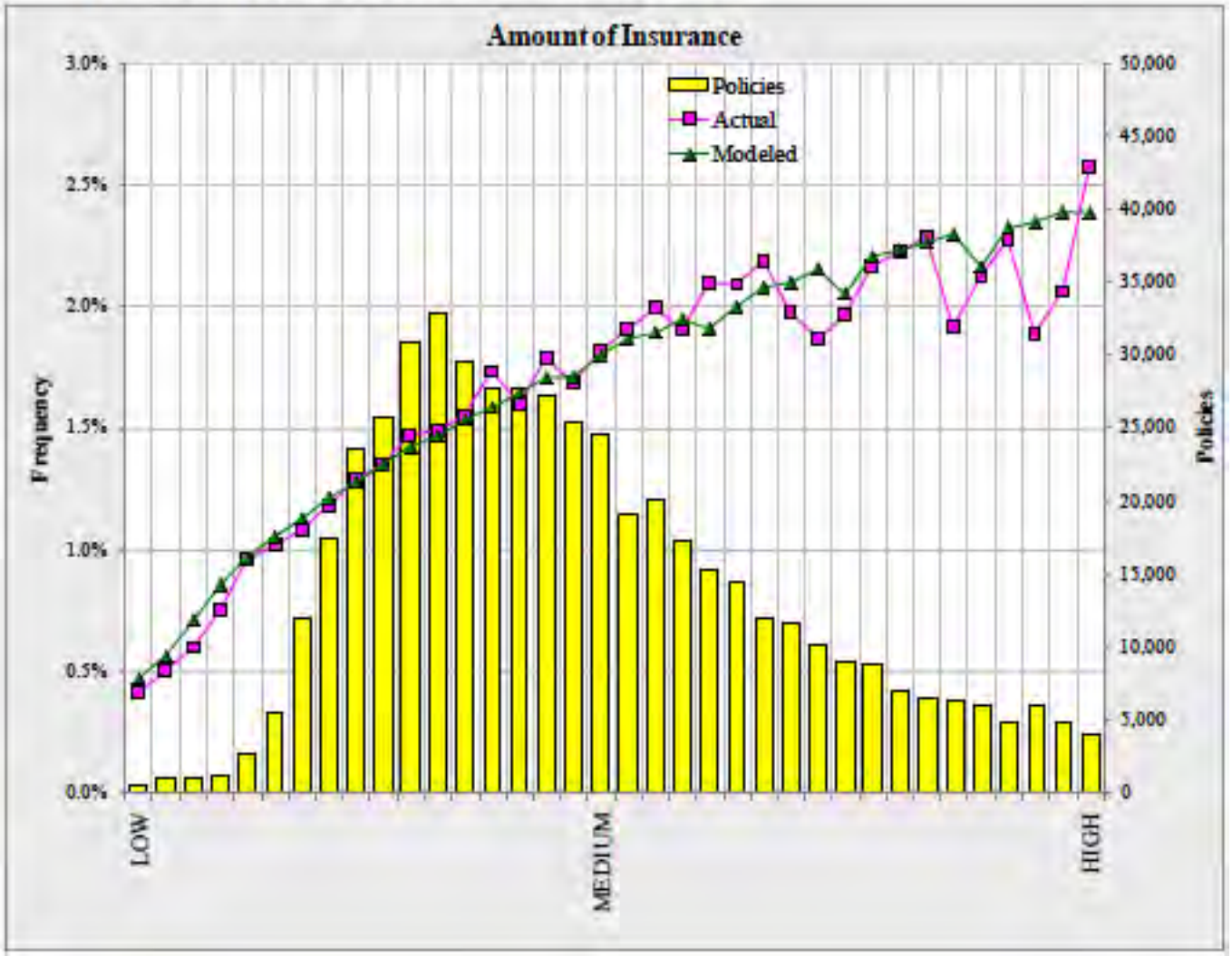
For sprinklers, there is not a consistent pattern across years; the prediction for sprinkles is very volatile due to the small amount of data. From the consistency test we again conclude that the indicated relatively for sprinklers is unreliable, and the model is consistent with a relativity of one for the other fire safety devices.

Next, rather than concentrating on one variable or one peril, we look at output to help us evaluate the performance of the overall model. We are still looking at a model for homeowners insurance.

A key idea is that of a hold-out data set. We intentionally set aside a random portion of the original data, and do not use it to develop and calibrate the model. Then we see how well the model performs at predicting on this hold-out data set. In general, the actuary should test the performance of a GLM on a hold-out data set.

Figure F.5 shows the results of the overall frequency model.

F.5 Actual Results v Modeled Results for AOI

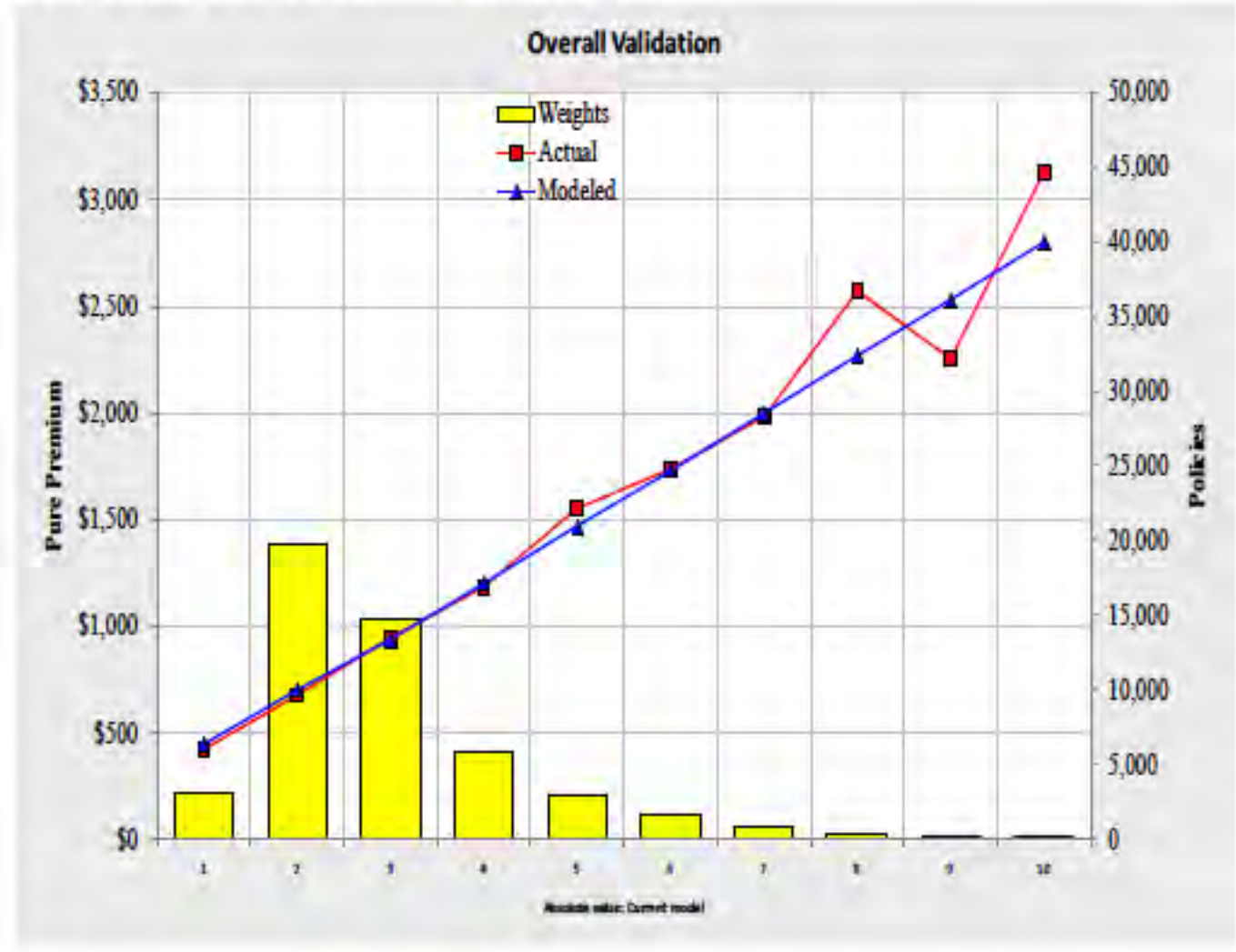


For the hold-out data set, as a function of amount of insurance, the modeled frequency is compared to the actual frequency.[365] We would like a close match between predicted and actual. However, we have a limited amount of policies, particularly for low and high amounts of insurance.

Overall, the match between the model and actual is good. However, the model appears to be underpredicting frequency somewhat for medium sized amounts of insurance and overpredicting frequency somewhat for high amounts of insurance. For extremely low amounts of insurance there is little data and a lot of volatility; however, the graph suggests that the model may be overpredicting for extremely low amounts of insurance.[366]

Figure F.6 displays another way to validate the overall model. This time we compare modeled and actual pure premiums for the hold-out set. We order the hold-out data by modeled pure premium and group it into 10 groups.

F.6 Actual Results v Modeled Results



[365] Remember, the GLM was developed and calibrated without this hold-out data.
[366] "Essentially, all models are wrong, but some are useful," George Box.

There appears to be a reasonable match between actual and predicted. While they differ for the highest pure premium groups, there is too little data in those groups to draw a definitive conclusion.

In general, in the case of a graph like Figure F.6, the actuary should pay particular attention to the extremes on both ends, since they are usually harder for the model to predict.

A More Realistic and Complex Example:

Consider the following data on claim severity for personal auto insurance:[367]

Observation	Age Group	Vehicle-Use	Severity	Claim Count
1	17–20	Pleasure	250.48	21
2	17–20	Drive to Work < 10 miles	274.78	40
3	17–20	Drive to Work > 10 miles	244.52	23
4	17–20	Business	797.80	5
5	21–24	Pleasure	213.71	63
6	21–24	Drive to Work < 10 miles	298.60	171
7	21–24	Drive to Work > 10 miles	298.13	92
8	21–24	Business	362.23	44
9	25–29	Pleasure	250.57	140
10	25–29	Drive to Work < 10 miles	248.56	343
11	25–29	Drive to Work > 10 miles	297.90	318
12	25–29	Business	342.31	129
13	30–34	Pleasure	229.09	123
14	30–34	Drive to Work < 10 miles	228.48	448
15	30–34	Drive to Work > 10 miles	293.87	361
16	30–34	Business	367.46	169
17	35–39	Pleasure	153.62	151
18	35–39	Drive to Work < 10 miles	201.67	479
19	35–39	Drive to Work > 10 miles	238.21	381
20	35–39	Business	256.21	166
21	40–49	Pleasure	208.59	245
22	40–49	Drive to Work < 10 miles	202.80	970
23	40–49	Drive to Work > 10 miles	236.06	719
24	40–49	Business	352.49	304
25	50–59	Pleasure	207.57	266
26	50–59	Drive to Work < 10 miles	202.67	859
27	50–59	Drive to Work > 10 miles	253.63	504
28	50–59	Business	340.56	162
29	60+	Pleasure	192.00	260
30	60+	Drive to Work < 10 miles	196.33	578
31	60+	Drive to Work > 10 miles	259.79	312
32	60+	Business	342.58	96

[367] Data taken from Exhibit 1 of "A Systematic Relationship Between Minimum Bias and Generalized Linear Models," by Stephen J. Mildenhall, PCAS 1999, not on the syllabus.

There are 8 age categories and 4 vehicle use types.
Thus there are a large number of ways to set up a GLM.
I will make age 40-49 and drive to work less than 10 miles as the base levels.

I will use the following definitions of variables:
X_0 corresponds to the base levels.
X_1 is one if 17-20 years old and zero otherwise.
X_2 is one if 21-24 years old and zero otherwise.
X_3 is one if 25-29 years old and zero otherwise.
X_4 is one if 30-34 years old and zero otherwise.
X_5 is one if 35-39 years old and zero otherwise.
X_6 is one if 50-59 years old and zero otherwise.
X_7 is one if 60+ years old and zero otherwise.
X_8 is one if Pleasure Use and zero otherwise.
X_9 is one if Drive to Work > 10 and zero otherwise.
X_{10} is one if Business Use and zero otherwise.

A Gamma Distribution with an identity link function was fit to these data:[368]

Parameter	Fitted Value	Standard Error		p-Value
β_0	203.522	6.54517	0	
β_1	62.9056	37.0291		8.9%
β_2	66.1851	19.4111		0
β_3	46.1676	12.5584	0	
β_4	33.2979	11.3777		0.3%
β_5	-15.289	9.57527		11.0%
β_6	3.57547	8.79087		68.4%
β_7	-1.84956	9.5907		84.7%
β_8	-8.63574	8.22596		29.4%
β_9	45.1086	7.43089	0	
β_{10}	122.802	13.4003	0	

[368] The fitted severities are: 257.79, 266.43, 311.54, 389.23, 261.07, 269.70, 314.82, 392.51, 241.05, 249.69, 294.80, 372.49, 228.19, 236.82, 281.93, 359.62, 179.60, 188.23, 233.34, 311.04, 194.89, 203.52, 248.63, 326.32, 198.46, 207.10, 252.21, 329.90, 193.04, 201.67, 246.78, 324.47.

Based on their large p-values, β_5, β_6, β_7, and β_8 are not significantly different than zero.

Let us test a model in which we eliminate the corresponding variables.
The reduced model will have:
Age 35-39 combined with 40-49.
Age 50-60 combined with 60+.
Pleasure use combined with Drive to Work < 10 miles.
Another GLM with Gamma Distribution with an identity link function was fit to these data.[369] [370]

The deviance for the original model with more variables is 31.2438 [371]
The deviance for the new model with less variables is 37.0310.

We have two nested models. GLM 1 is a special case of GLM 2.
Then the test statistic (asymptotically) follows an F-Distribution with numbers of degrees of freedom equal to: ν_1 = the difference in number of parameters = 3,

and ν_2 = number of degrees of freedom for the more simpler model
= (number of observations) - (number of parameters) = 32 - 7 = 25.

$\hat{\phi}_S$ = estimated dispersion parameter for the smaller (simpler) model
$= D_S / \nu_S = 37.0310/25 = 1.481$.[372]

The test statistic is: $\frac{D_S - D_B}{(\text{number of added parameters})\,\hat{\phi}_S} = \frac{(37.0310 - 31.2438) / 3}{1.481} = 1.303$.

Using a computer, the p-value is 29.5%.
Thus we do <u>not</u> reject the null hypothesis of using the simpler model with fewer parameters.[373]

---

[369] The fitted parameters are: 196.36, 67.41, 71.78, 50.88, 38.42, 6.13, 47.01, 125.74.
[370] The fitted severities are: 263.77, 310.77, 389.51, 268.14, 315.15, 393.88, 247.24, 294.248, 372.98, 234.78, 281.79, 360.52, 196.36, 243.37, 322.10, 202.49, 202.49, 249.49, 328.23.
[371] A computer was used to fit both models and to calculate the deviances.
[372] The syllabus reading does not discuss how to estimate ϕ; this is one way.
[373] One could now compare additional models with different subsets of the original variables.
One could also fit models using different distributional forms and/or link functions.

Example of Homeowners Rating Factors Used in the United Kingdom:[374]

Personal lines rates in the United Kingdom have long been based on GLMs.
One important aspect to using GLMs is to find relevant variables.
Here is a list of some rating variables that are used for Homeowners Insurance.

Postal code (so geodemographic and geophysical factors can be derived)[375]
Amount of insurance
Number of rooms / bedrooms
Wall type
Roof type
State of repair
Extensions
Ownership status (rent/own)
Occupancy in day
Neighborhood watch scheme
Approved locks, alarms, smoke detectors
Deductibles
Endorsements purchased (e.g. riders for jewelry, oriental rugs)
How long held insurance / when last claimed

Policyholder details:
• Age
• Sex
• Marital status
• Number of children
• Occupation
• Residency
• Criminal convictions
• Claims in past 2 or past 5 years

Smokers present in house
Non-family members sharing house
Length of time living at property
Use (principal residence / secondary residence / business / rented)
Coverage selected (buildings/contents/both)
Source of business (e.g. agent, internet, etc.)

---

[374] "Homeowners Modeling" by Claudine Modlin, presentation at the 2006 CAS Seminar on Predictive Modeling.
[375] Geodemographics are the average characteristics in an area. Examples are: population density, length of homeownership, average age of residents, and average family income. Geophysical factors can include soil type, and weather data such as the maximum wind speed, the average rainfall, and the average snowfall.

Homeowners Perils:

There can be advantages to modeling the different homeowners perils separately.[376]
One can either model pure premium or separately model frequency and severity.

Some variables may have different effects on different perils. For example, increased population density may be related to an increased frequency for theft claims while being related to a decreased frequency of fire claims.

Some variables may have a significant effect on one peril but not another. For example, more children in the house may be related to an increased frequency of liability while being unrelated to the frequency for wind.

Here is an example of data by peril for the United States.

Peril	Frequency(in percent)	Median Claim Amount
Fire	0.310	4,152
Lightning	0.527	899
Wind	1.226	1,315
Hail	0.491	4,484
Water-Weather Related	0.776	1,481
Water-NonWeather[377]	1.332	2,167
Liability	0.187	1,000
Other	0.464	875
Theft-Vandalism	0.812	1,119
Total	5.889	1,661

The percent of losses expected by peril varies considerably by geographical location. For example, the expected percent from wind (from hurricanes and other storms) is higher than average on the coast of Florida. For example, the expected percent from theft is higher than average in the center of a large city.

Recently, homeowners insurers have begun to implement rating plans that have separate base rates for each major peril covered and the individual rating variable relativities are applied to the applicable base rate (e.g., burglar alarm discount applies to the theft base rate only).

---

[376] See for example, "Predictive Modeling of Multi-Peril Homeowners Insurance," by Edward W. Frees, Glenn Meyers, and A. David Cummins, in Variance Volume 6 / Issue 1. They show that the perils are not independent.
[377] For example, water from the bursting of a pipe.

Problems:

3.1. (1.5 points) Five Generalized Linear Models have been fit to the same set of 50 observations.

Model	Number of Fitted Parameters	Deviance
A	6	335.8
B	8	331.9
C	10	325.2
D	12	321.4
E	14	317.0

Which model has the best AIC (Akaike Information Criterion)?

3.2. (0.5 points) Briefly discuss how to pick the base level of a categorical variable.

3.3. (1 point) When a log link is used, it is usually appropriate to take the natural logs of continuous predictors before including them in the model, rather than placing them in the model in their original forms. Discuss why.

3.4. (1.5 points) Fully discuss the use of weights in GLMs.

3.5. (0.5 points) Briefly discuss a primary strength of GLMs versus univariate analyses.

3.6. (0.5 points) A continuous predictor x_1 has a coefficient of $\beta_1 = 0.4$ in a logistic model.
For a unit increase in x_1, what is the estimated change in the odds?

3.7. (1 point) Compare and contrast the Poisson and the Negative Binomial Distributions.

3.8. (0.5 points) With respect to GLMs, briefly discuss aliasing.

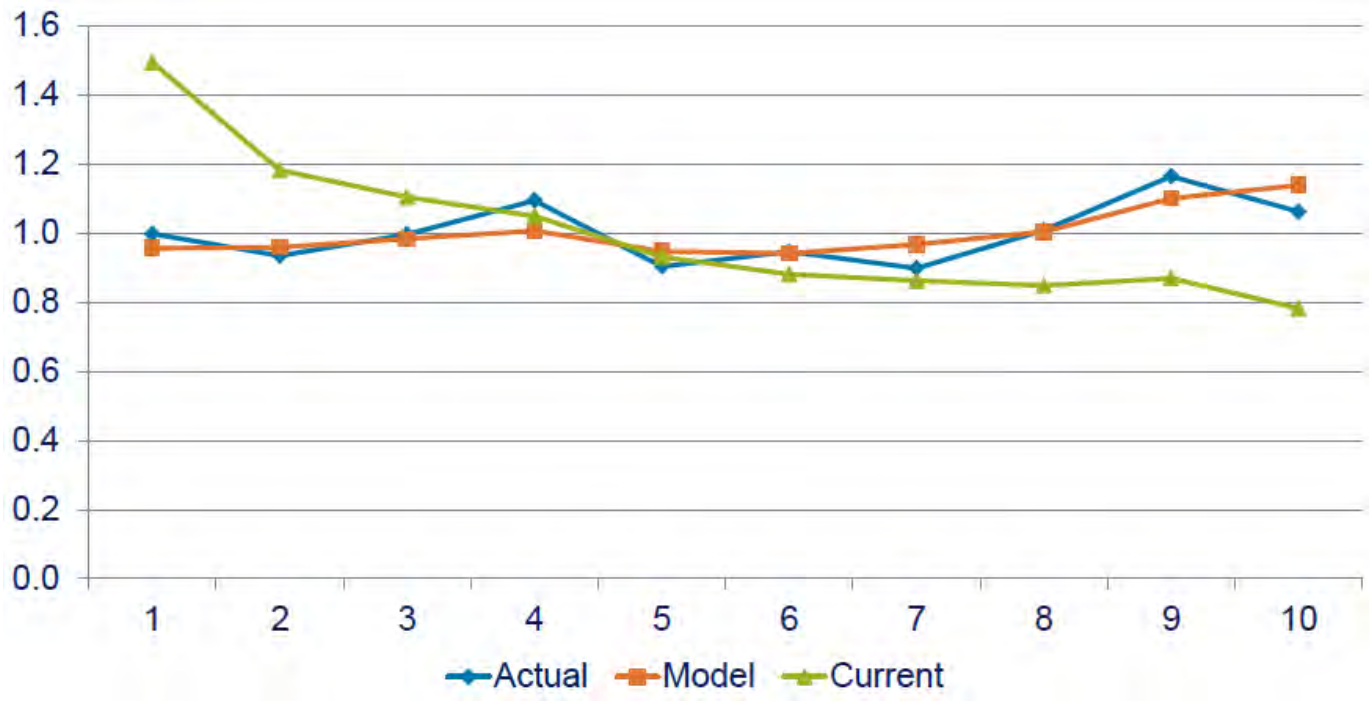
3.9. (0.5 points) List two limitations of GLMs.

3.10. (1 point) One possible fix for nonlinearity in a continuous variable is not to model it as continuous at all; rather, a new categorical variable is created where levels are defined as intervals over the range of the original variable. Briefly discuss two drawbacks to this approach.

3.11. (1.5. points) A GLM has been fit using a Poisson Distribution with $\hat{\beta}_1 = 0.02085$ with standard error 0.00120.
Using instead an overdispersed Poisson the estimate of ϕ is 7.9435.
For this second model, determine a 95% confidence interval for β_1.

3.12. (1 point) Discuss the Tweedie Distribution.

3.13. (1 point) You are given a double lift chart, sorted by ratio of the model prediction over the current plan prediction. Discuss the lift of the proposed model compared to the current plan.



3.14. (1 point) The flexibility afforded by the ability to use a link function is a good thing because it gives us more options in specifying a model, thereby providing greater opportunity to construct a model that best reflects reality. However, when using GLMs to produce insurance rating plans, an added benefit is obtained when the link function is specified to be the natural log function. Briefly discuss this added benefit.

3.15. (1 point) A logistic regression has been fit to some data. For a certain threshold:

		Predicted Claims		
		No	Yes	Total
Actual	No	6000	2000	8000
Claim	Yes	300	700	1000
	Total	6300	2700	9000

What point would be plotted in the ROC curve?

3.16. (2 points) List and briefly discuss four components of a predictive modeling project.

3.17. (1.5 points)
(a) (0.5 points) Define the partial residuals.
(b) (1 point) Discuss partial residual plots.

3.18. (0.5 points) Briefly contrast the following two GLMs:
$\mu = \exp[\beta_0 + \beta_1 X_1 + \beta_2 X_2]$.
$\mu = \exp[\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2]$.

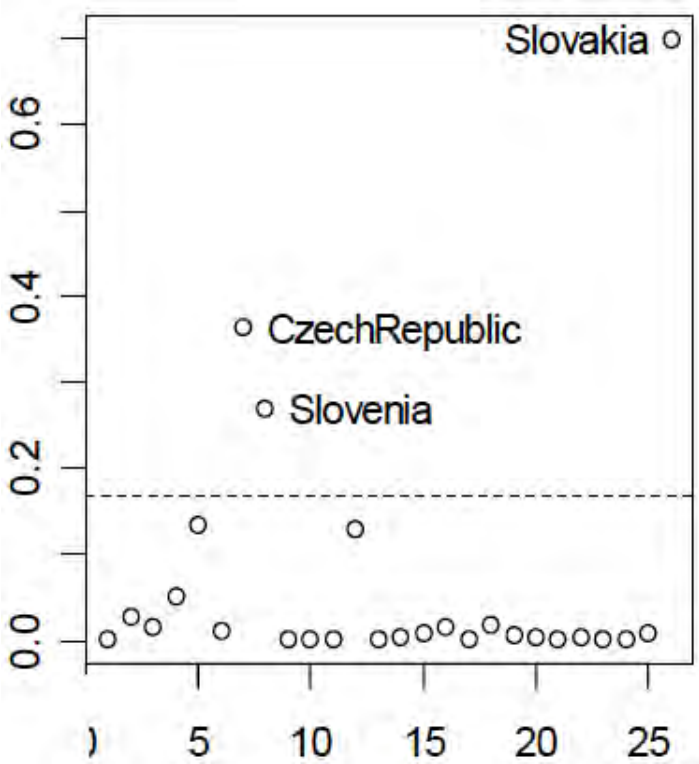
3.19. (1 point) Any data set of sufficient size is likely to have errors.
Briefly discuss two of the steps that should always be taken to attempt to catch and remedy some of the more common errors that can occur.

3.20. (0.5 points) List two types of Exploratory Data Analysis (EDA) plots and their purposes.

3.21. (1 point) Discuss some reasons to use frequency and severity models rather than a pure premium model.

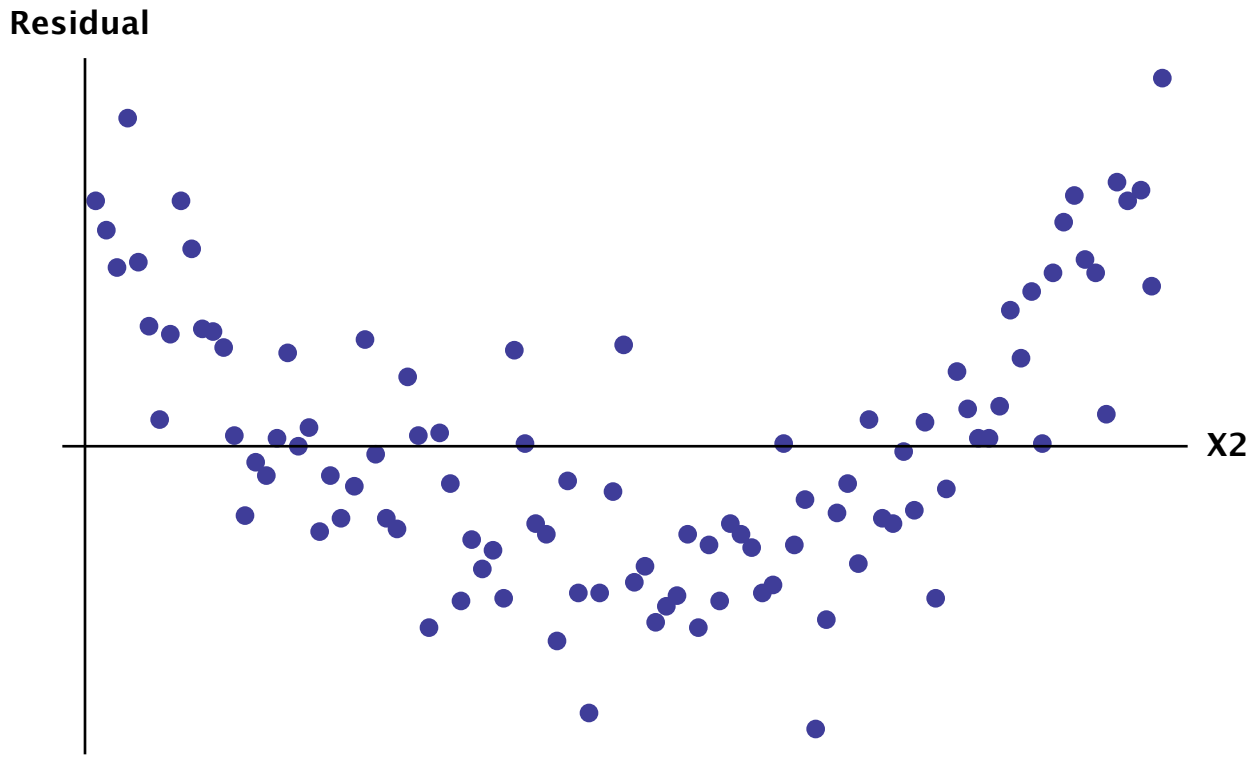
3.22. (1.5 points) Fully discuss the use of an offset term in GLMs.

3.23. (0.5 points) Discuss the following graph of Cook's Distance for 26 observations:



3.24. (1 point) Define the saturated and the null models, and discuss them with respect to deviance.

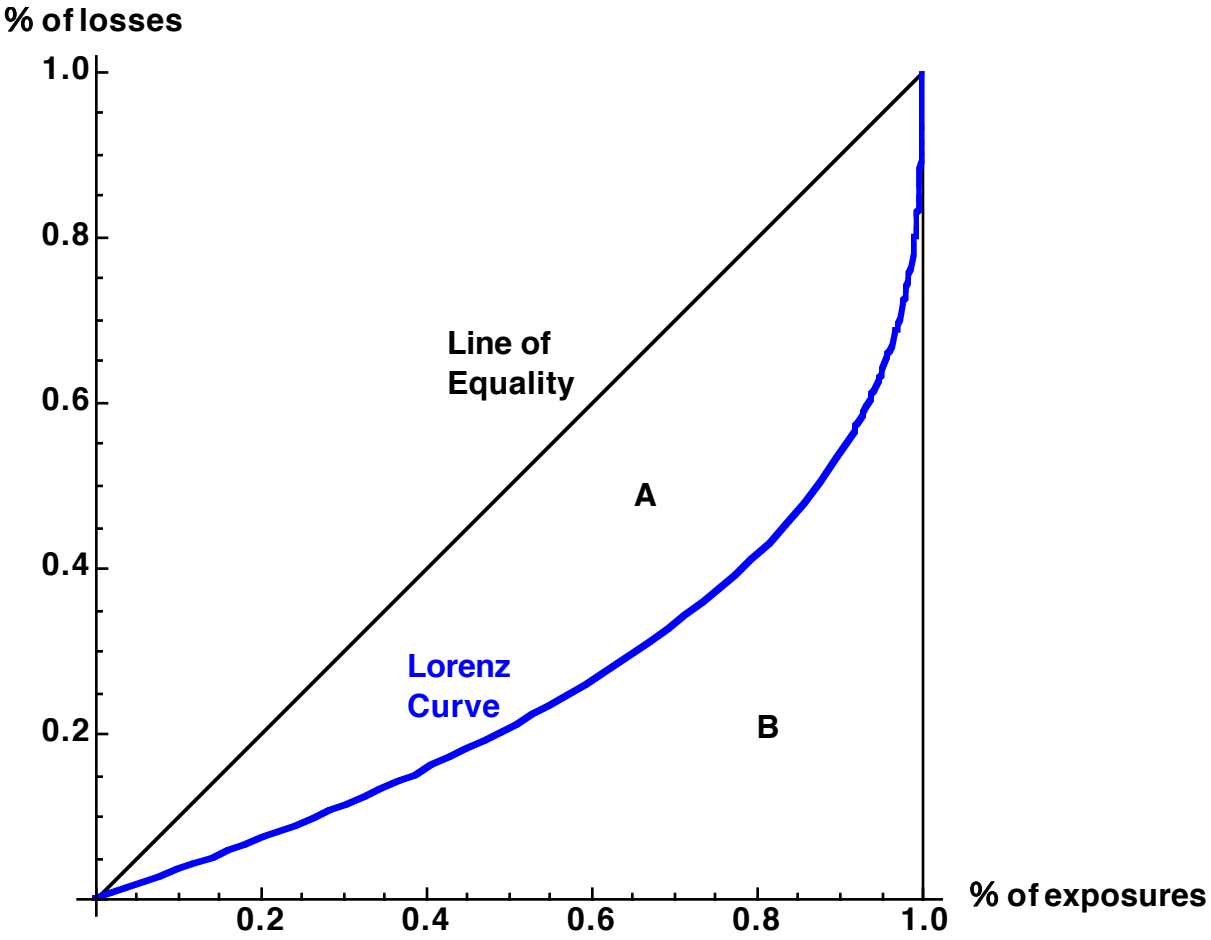
3.25. (1 point) Briefly comment on the following plot of deviance residuals of a model as a function of a predictor variable X_2:



3.26. (2 points) A GLM using a Tweedie Distribution and a log link function is being used to model pure premiums of private passenger automobile property damage liability insurance.
There are 100,000 observations.
10 parameters including an intercept were fit.
The deviance is 233,183.65, and the estimated dispersion parameter is 2.371.
Credit score as a categorical variable is added to the model, with a total of 6 categories.
The deviance for this more complex model is 233,134.37.
Discuss how you would use an F-Test to determine whether credit score should be added to this model.

3.27. (2 points) The following 5 returns on a stock price are observed:
-0.154, 0.239, -0.064, -0.328, 0.195.
Construct the corresponding Normal Q-Q Plot.

3.28. (0.5 points) Areas have been labeled in the following graph of a Lorenz Curve. Determine the Gini index.



3.29. (0.5 points) With respect to GLMs, briefly discuss pricing coverage options such as deductibles or increased limits.

3.30. (0.5 points) Give an example of a hinge function.

3.31. (0.5 points) Five logistic regressions have been fit to the same data.
ROC curves have been drawn for each model.

Model	Number of Parameters	AUROC
A	1	0.58
B	2	0.66
C	3	0.73
D	4	0.79
E	5	0.75

Which model is preferred?

3.32. (1 point) For a GLM, the estimated mean for an individual is 35, with variance 5. Determine a 95% confidence interval for the estimated mean.

3.33. (1.5 points)
Five different Generalized Linear Models, have been fit to the same set of 400 observations.

Model	Number of Fitted Parameters	LogLikelihood
A	3	-730.18
B	4	-726.24
C	5	-723.56
D	6	-721.02
E	7	-717.50

Which model has the best BIC (Bayesian Information Criterion)?

Use the following information for the following four questions:

- There is data on commercial building insurance claims frequency.
- A Poisson GLM was fit using the log link function.
- A categorical predictor used is building occupancy class, coded 1 through 4, with 1 being the base class.
- A binary predictor used is sprinklered status, with 1 being yes and 0 being no.
- A continuous predictor used is: ln[amount of insurance / 200,000] = ln[AOI / 200,000].
- The fitted intercept is $\beta_0 = -3.8$.
- The fitted parameters for building occupancy classes 2, 3, and 4 are: $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.1$.
- The fitted parameter for sprinklers is: $\beta_4 = -0.5$.
- The fitted parameter for ln[AOI / 200,000] is: $\beta_5 = 0.4$.
- An interaction term between sprinkler status and ln[AOI / 200,000] is included in the model; the fitted parameter is: $\beta_6 = -0.1$.

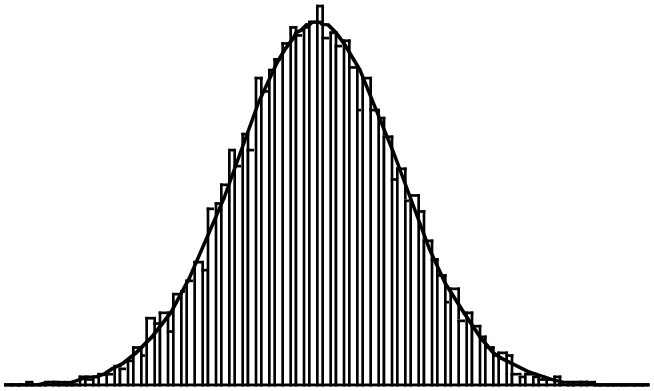
3.34. (1 point) Determine the fitted frequency for a $100,000 building in occupancy class 1 without sprinklers.

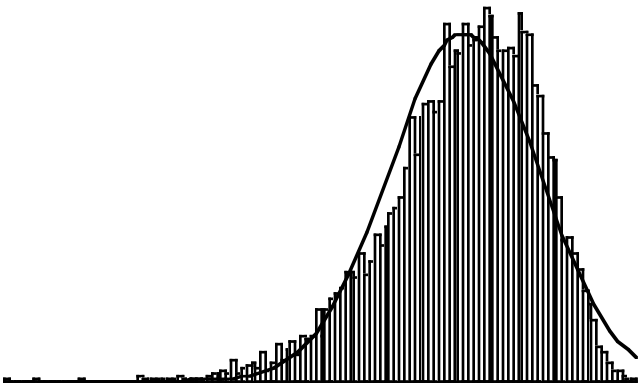
3.35. (1 point) Determine the fitted frequency for a $250,000 building in occupancy class 2 with sprinklers.

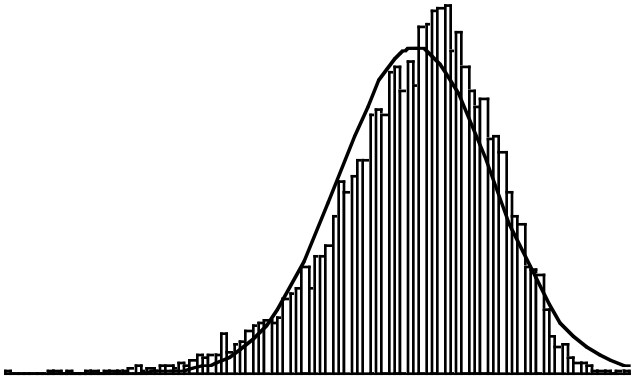
3.36. (1 point) Determine the fitted frequency for a $300,000 building in occupancy class 3 without sprinklers.

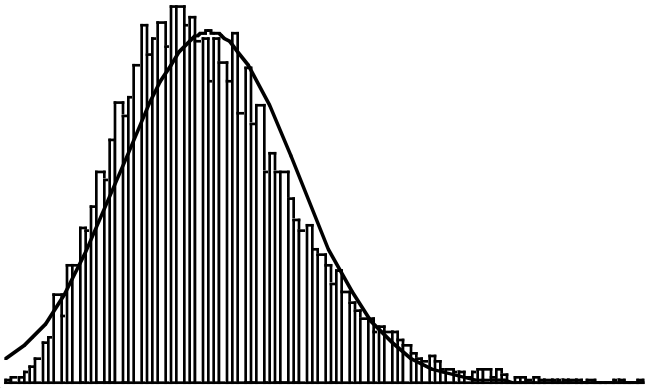
3.37. (1 point) Determine the fitted frequency for a $600,000 building in occupancy class 4 with sprinklers.

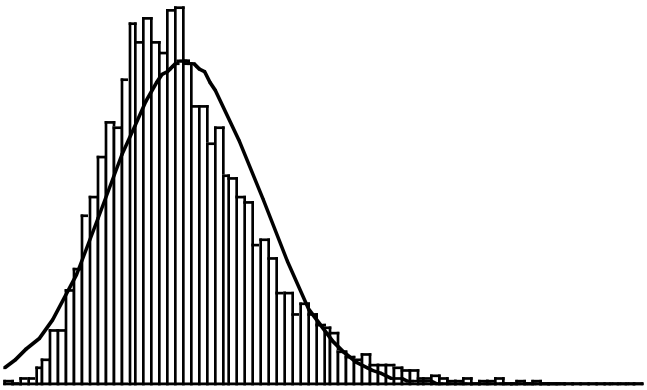
3.38. (1 point) The following are histograms of deviance residuals for GLMs.
Which of the following histograms represents the best model?

A. 

B. 

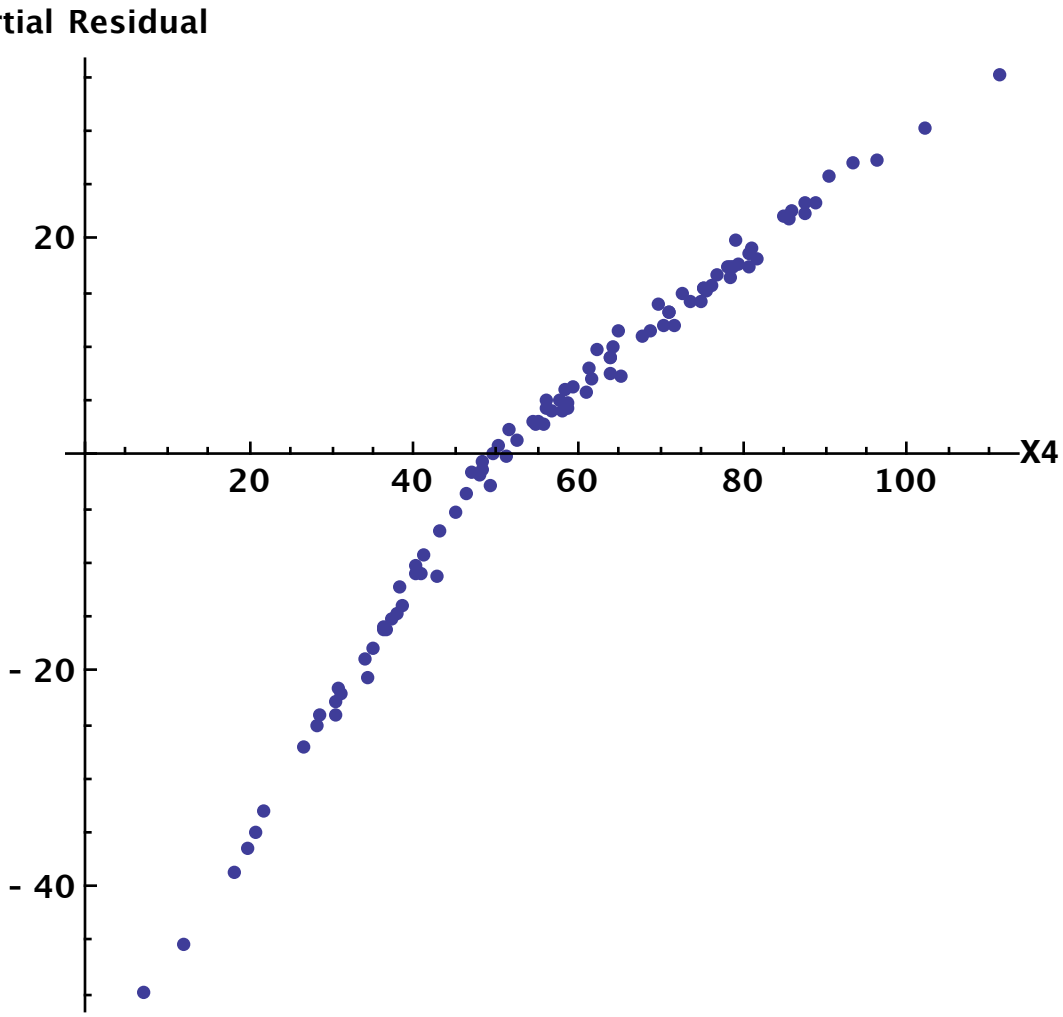
C. 

D. 

E. 

3.39. (2 points) You are constructing a Generalized Linear Model.
(a) (0.5 point) If the model is additive, what link function would you use?
(b) (0.5 point) If the model is multiplicative, what link function would you use?
(c) (0.5 point) If the variance is proportional to the mean, what distribution would you use?
(d) (0.5 point) If the standard deviation is proportional to the mean, what distribution would you use?

3.40. (1 point) For a GLM, here is a partial residual plot for the predictor variable X_4:



Briefly discuss the meaning of this plot.
If necessary, what is a possible solution?

3.41. (1.5 points) With respect to GLMs, discuss the training, validation, and test sets.

3.42. (2 points) Exponential families have a relationship between their mean and variance: $V(Y_i) = \phi\, V(\mu_i) / \omega_i$, where $V(\mu)$ is the variance function.
List different exponential families and their variance functions.

3.43. (6 points) You are given the following 20 breaking strengths of wires:
500, 750, 940, 960, 1100, 1130, 1150, 1170, 1190, 1240, 1260, 1350, 1400, 1450, 1490, 1520, 1550, 1580, 1850, 2000.
With the aid of a computer, construct a Normal Q-Q Plot.

3.44. (5 points) You have the following data on reported occurrences of a communicable disease in two areas of the country at 2 month intervals:

Months	Area A	Area B
2	8	14
4	8	19
6	10	16
8	11	21
10	14	23
12	17	27
14	13	28
16	15	29
18	17	33
20	15	31

Let X_1 = ln(months). Let X_2 = 0 for Area A and 1 for Area B.

Assume the number of occurrences Y_i are Poisson variables with means μ_i, and

$\ln(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$.

Set up the equations to be solved in order to fit this model via maximum likelihood.

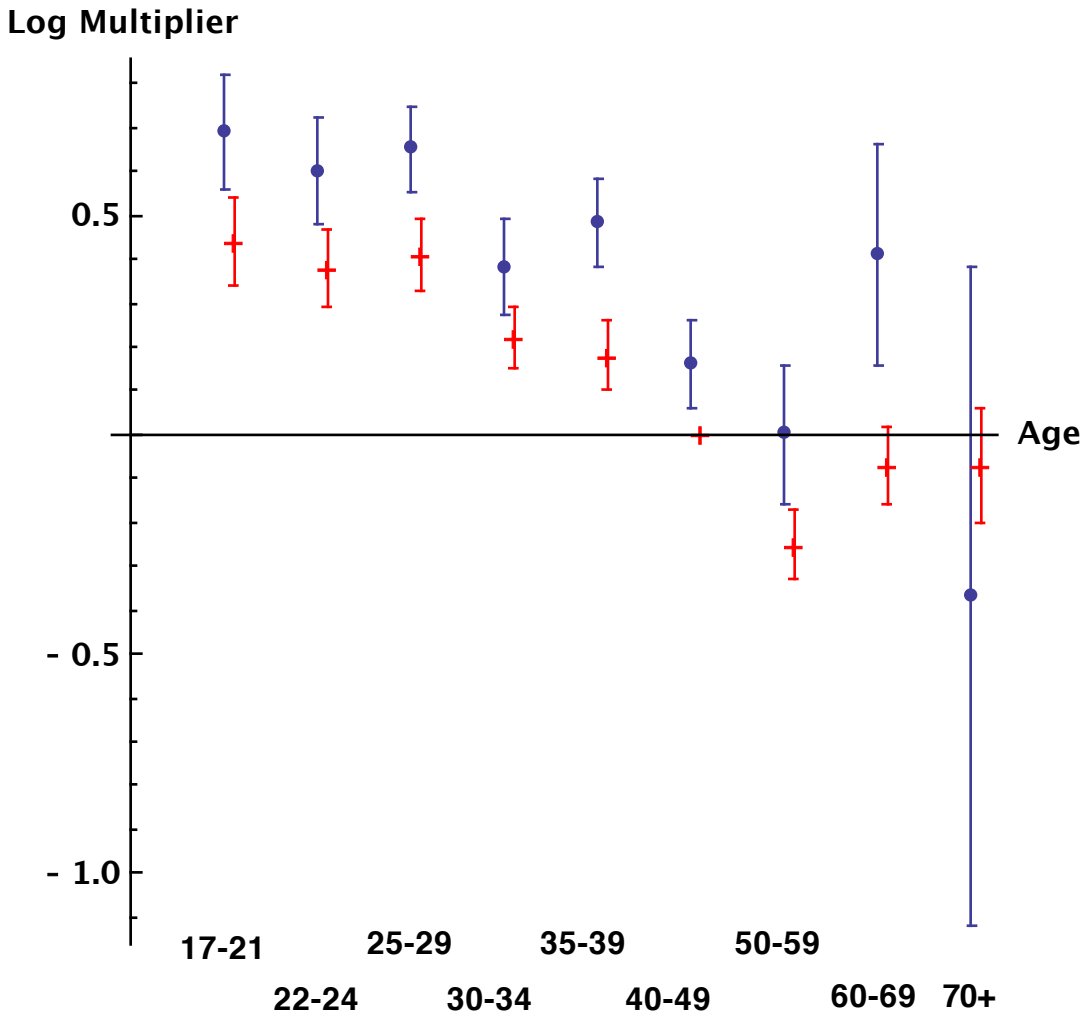
3.45. (1 point) Which of the following statements are true with respect to Generalized Linear Models?

1. Errors are assumed to be Normally Distributed.
2. The link function defines the relationship between the expected response variable and the linear combination of the predictor variables.
3. The use of a log link function assumes the rating variables relate multiplicatively to one another.

3.46. (1.5 points) Generalized Linear Models with a overdispersed Poisson error structure and a log link function have been fit in order to model claim frequency for Homeowners Insurance.
The models use many variables. The homes have been split into four age categories.
A model that uses age has a deviance of 3306.9.
An otherwise similar model that does not use age has a deviance of 3320.2,
and an estimated dispersion parameter of 1.83.
The null hypothesis is to use the model that does not include age.
The alternative hypothesis is to use the model that does include age.
Calculate the F-test statistic.
Discuss how you would perform the test.

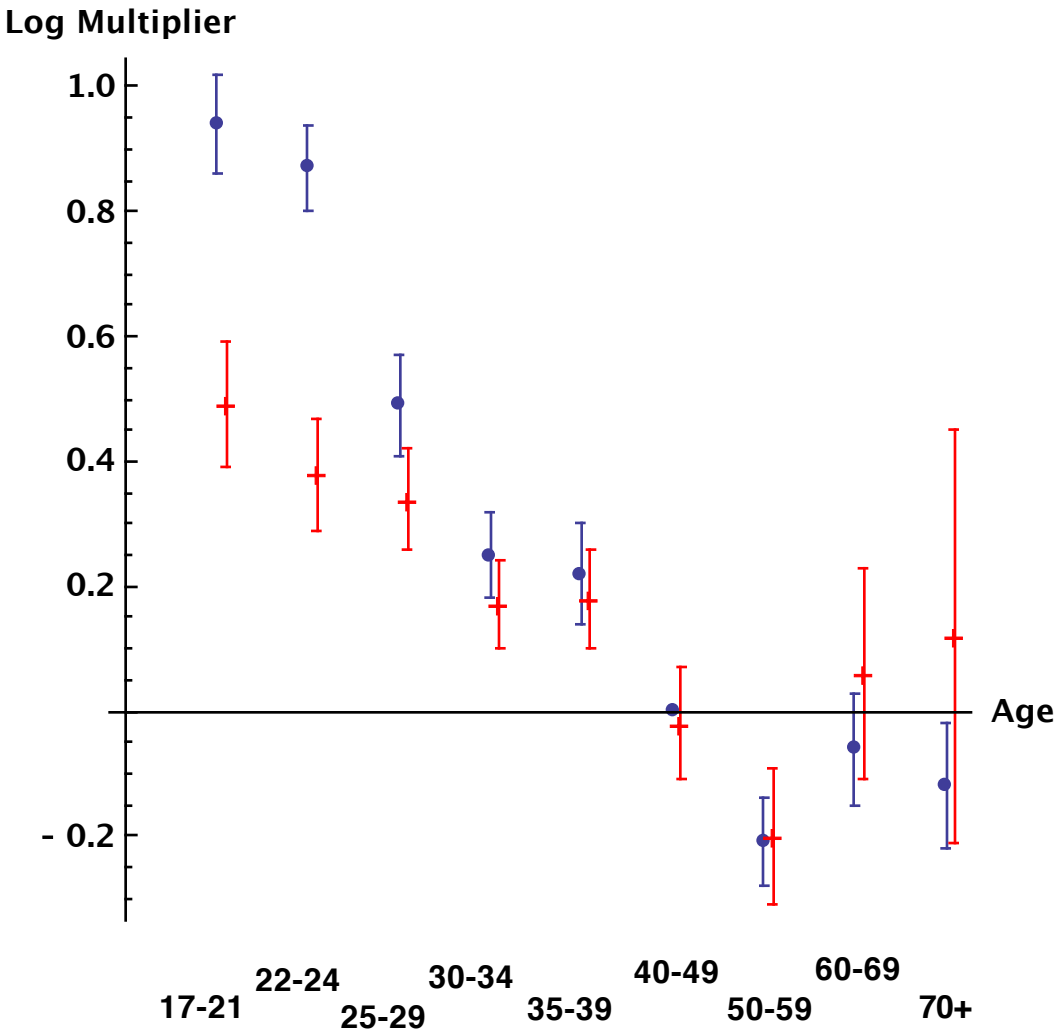
3.47. (1 point) Discuss model lift.

3.48. (1.5 points) The following graph displays the modeled log of the frequency relativity by age for two different frequency of premium payment: yearly in red pluses, and four times a year in blue dots. Also approximate 95% confidence intervals are shown for each case.



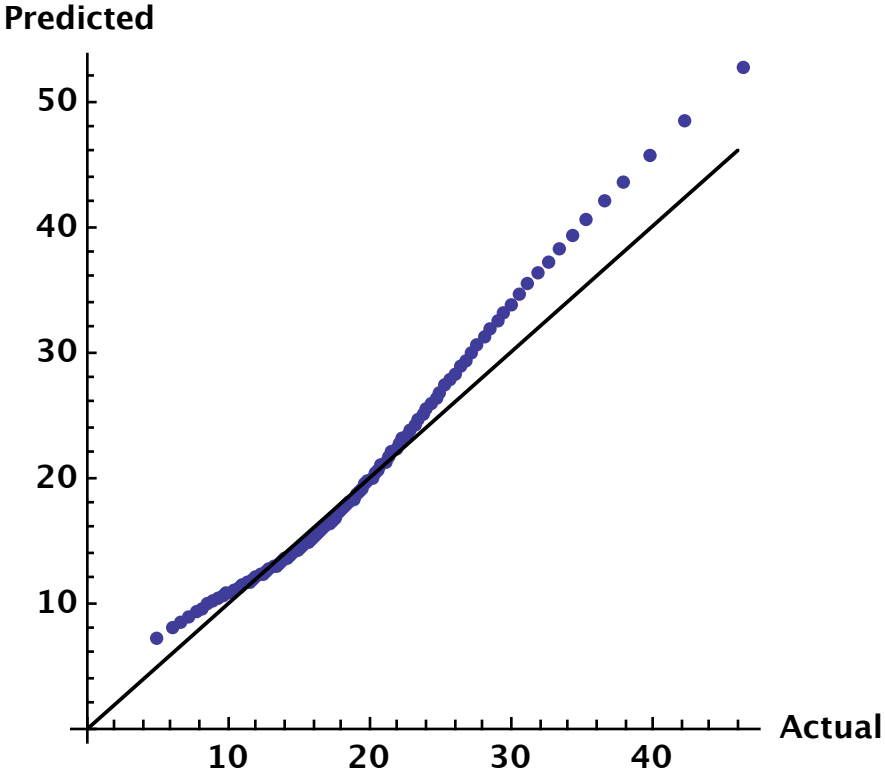
Question continued on the next page.

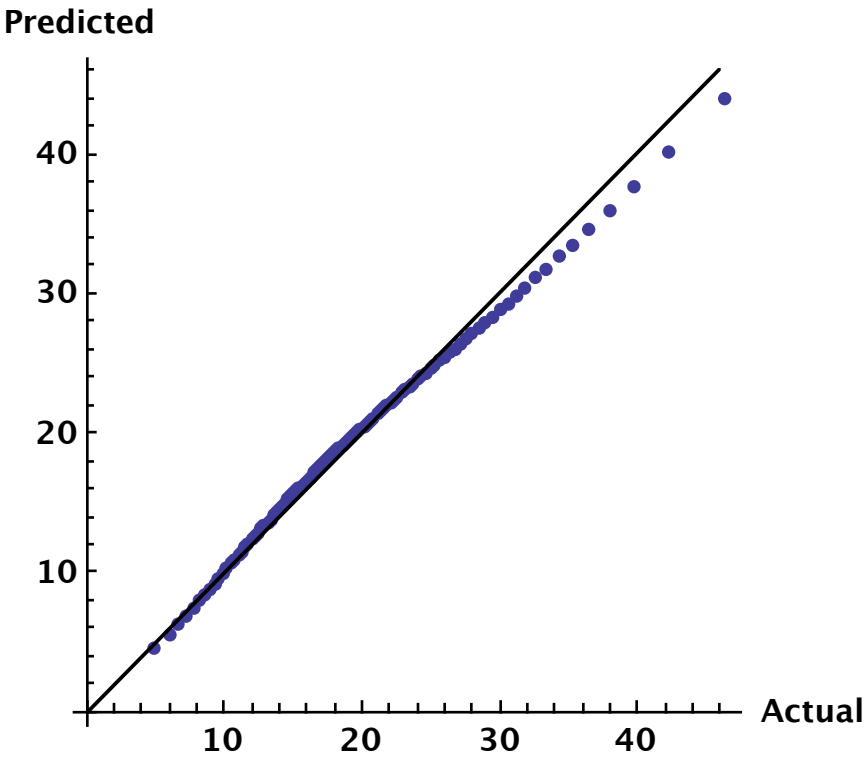
The following similar graph displays the modeled log of the frequency relativity by age for males in blue dots and females in red pluses.
Also approximate 95% confidence intervals are shown for each case.



Briefly compare and contrast the interaction of age of driver and payment frequency with the interaction of age of driver and gender.

3.49. (0.5 points) For two GLMs you are given the following graphs based on holdout data:



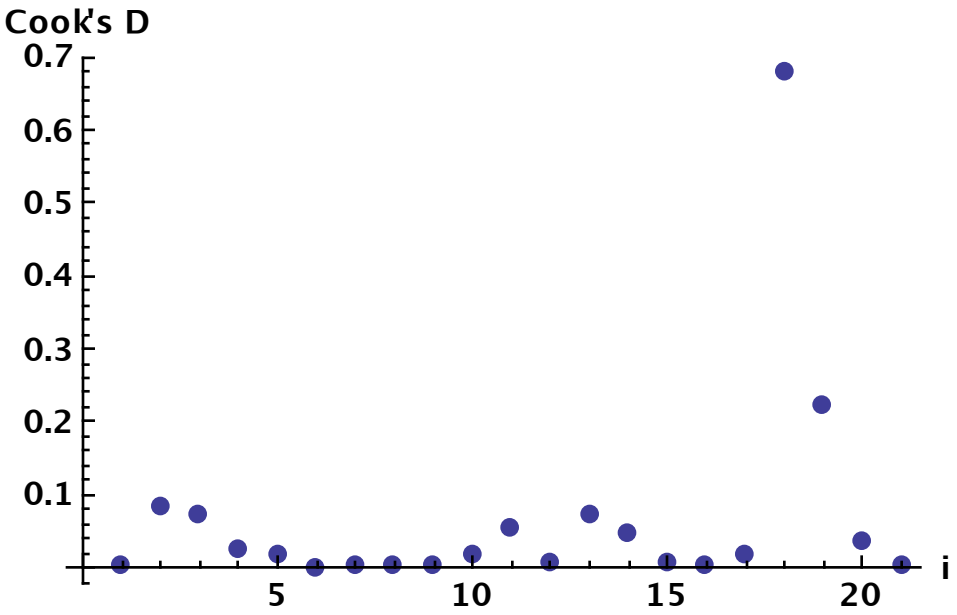


Which model do you prefer and why?

3.50. (2 points) There are three age groups of cars: A, B, C.
There are also three size categories of cars: small, medium, large.
Specify the following structural components of a generalized linear model.
i. Design matrix
ii. Vector of model parameters

3.51. (2 points) Briefly discuss, compare and contrast under-fitting and over-fitting a model.

3.52. (0.5 points) Discuss the following graph of Cook's Distance for 21 observations:



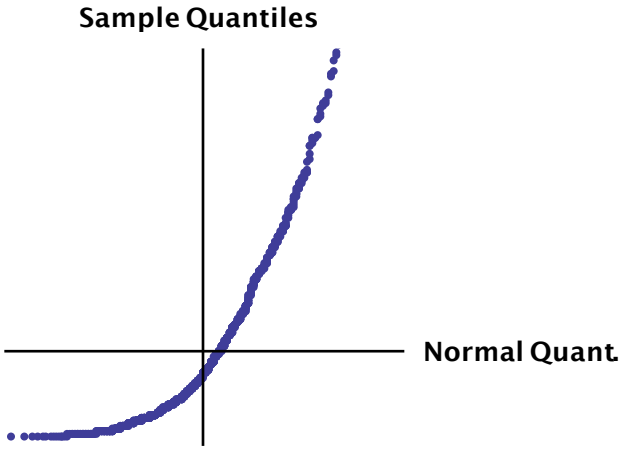
3.53. (2 points)
Use the following information on two Generalized Linear Models fit to the same 100 data points:

Number of Fitted Parameters	Loglikelihood
6	-321.06
7	-319.83

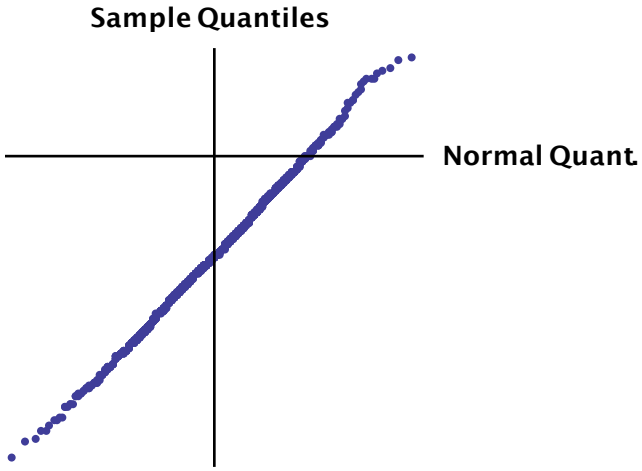
(a) Based on AIC (Akaike Information Criterion), which model is preferred?
(b) Based on BIC (Bayesian Information Criterion), which model is preferred?

3.54. (1 point) Which of the following Normal Q-Q Plots is most likely to be of data drawn from a Normal Distribution?

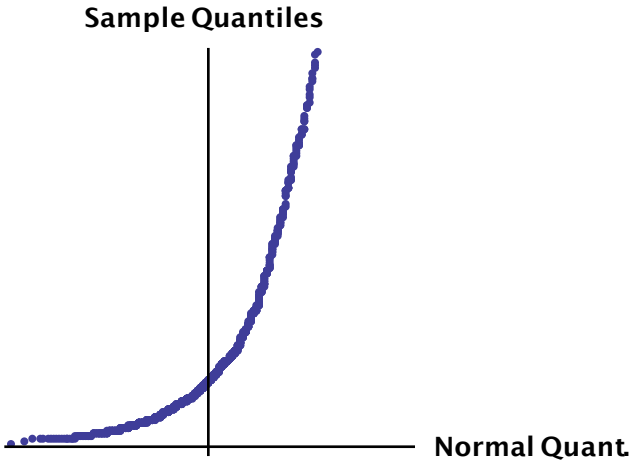
A.



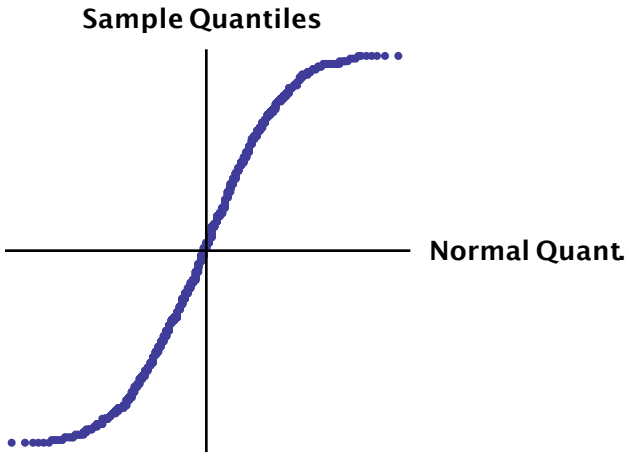
B.



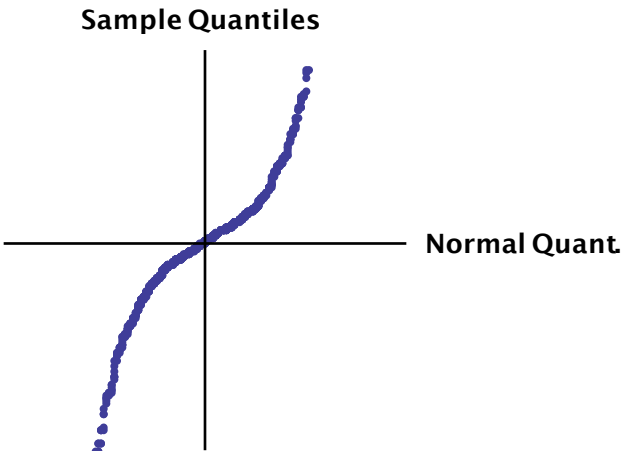
C.



D.



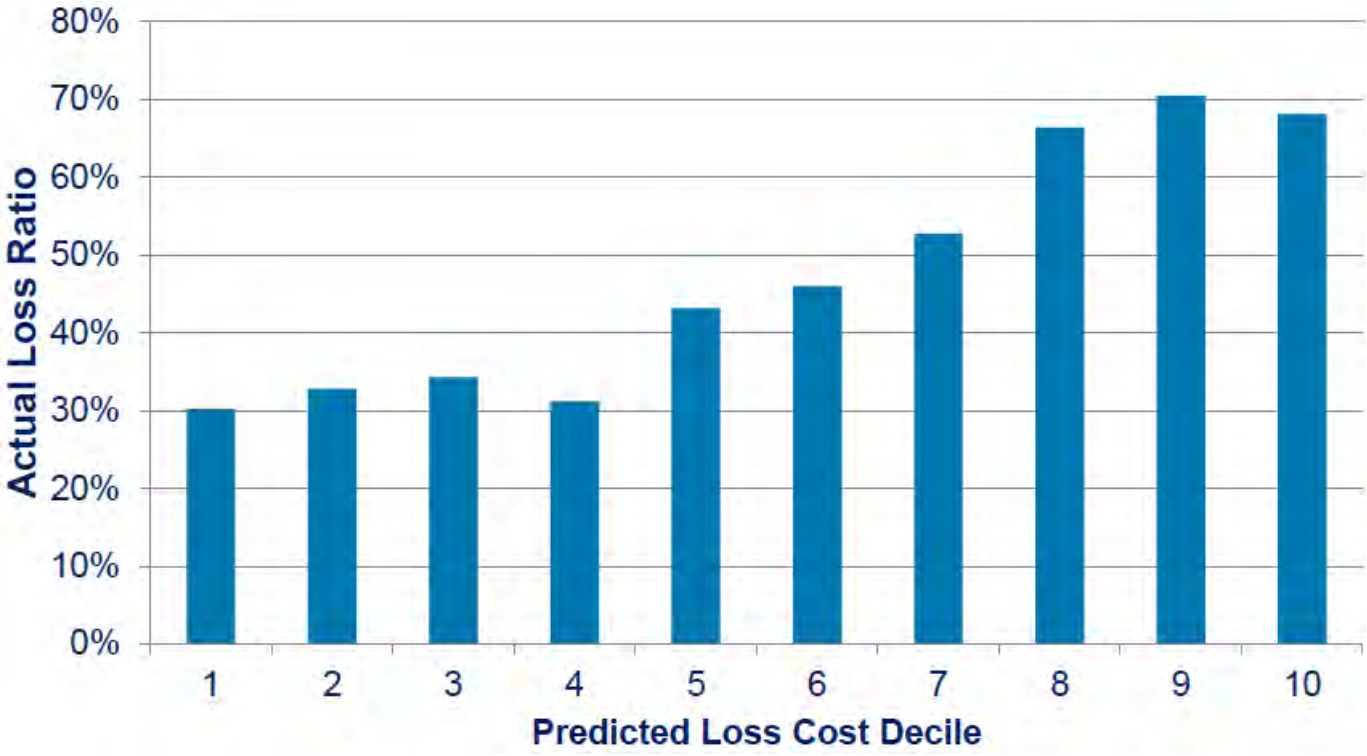
E.



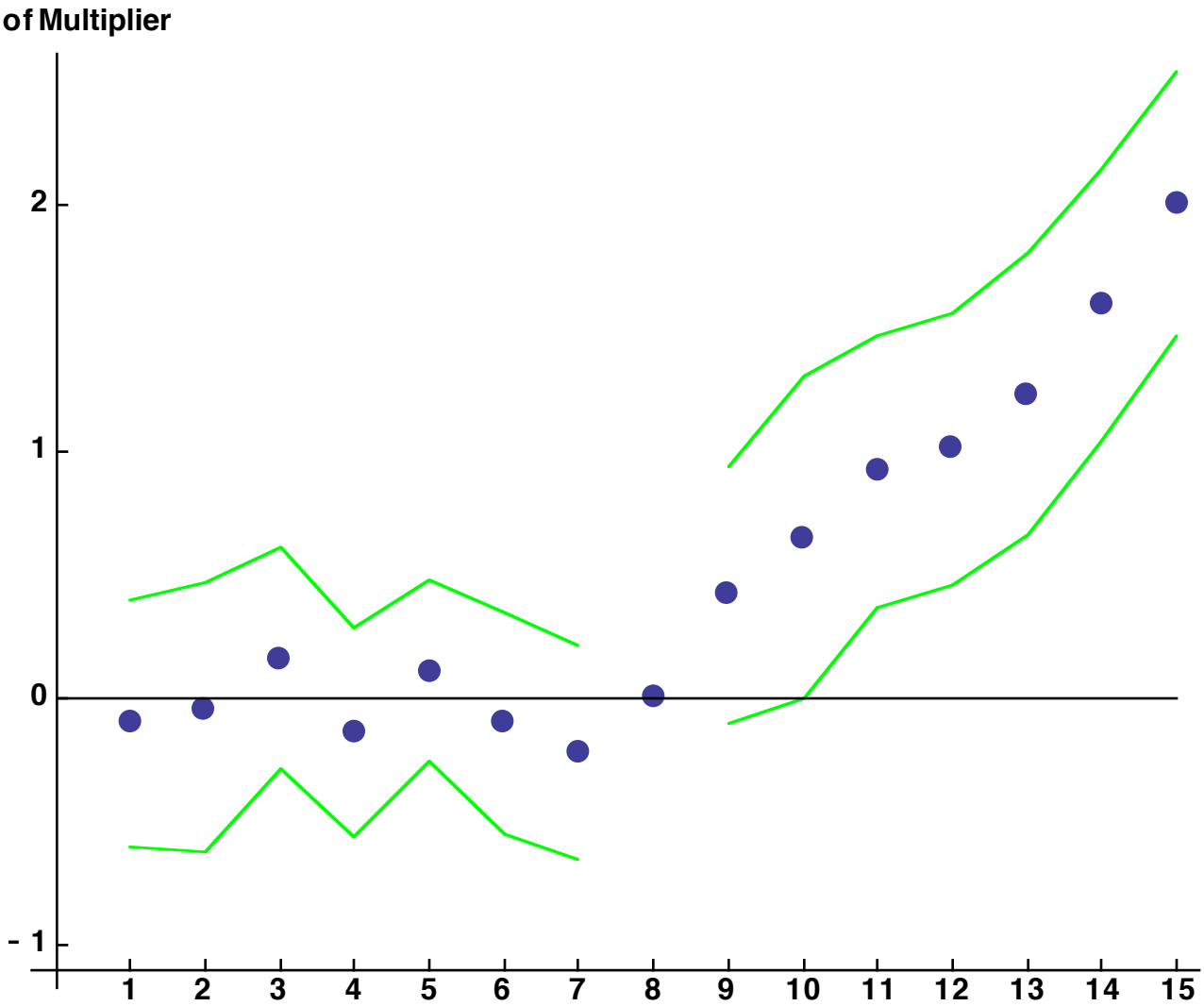
3.55. (2.5 points) For each of the following situations, give the typical generalized linear model form. State the distributional form of the error and link function typically used.
(a) Claim Frequencies.
(b) Claim Counts.
(c) Average Claim Sizes
(d) Probability of Policy Renewal
(e) Pure Premiums

3.56. (0.5 points) You are comparing two rating plans.
The first has a Gini Index of 0.48, while the second has a Gini Index of 0.55.
Which rating plan is preferred?

3.57. (1 point) You are given the following loss ratio chart for a proposed rating plan.
Discuss the lift of the proposed plan compared to the current plan.



3.58. (1 point) Below is a graph of a GLM fit to data, showing the natural log of the fitted multiplicative factors for levels of a variable. (8 is the base level.)
Also shown are approximate 95% confidence intervals.
Briefly discuss what this graph tells the actuary about the fitted model.



3.59. (5 points) The observed claim frequencies for urban vs rural and male vs female drivers are:

Claim frequency	Urban	Rural
Male	0.200	0.100
Female	0.125	0.050

There are equal exposures in each of the four cells.
We will fit a GLM using a Poisson Distribution.
(a) (2.5 points) For an additive model, determine the maximum likelihood equations to be solved.
(b) (2.5 points) For an multiplicative model, determine the maximum likelihood equations to be solved.

3.60. (1 point) A logistic regression has been fit to some data. For a certain threshold:

		Predicted Claims		
		No	Yes	Total
Actual	No	40,000	10,000	50,000
Claim	Yes	1200	1800	3000
	Total	41,200	11,800	53,000

What point would be plotted in the ROC curve?

Use the following information for the next two questions:

X:	1	5	10	25
Y:	5	15	50	100

Y_1, Y_2, Y_3, Y_4 are independently Normally distributed with means $\mu_i = \beta X_i$, i = 1, 2, 3, 4, and common variance σ^2.

3.61. (2 points) Determine $\hat{\beta}$ via maximum likelihood.

3.62. (3 points) Estimate the standard deviation of $\hat{\beta}$.

3.63. (1.5 points) A GLM is used to model claim size.
You are given the following information about the model:

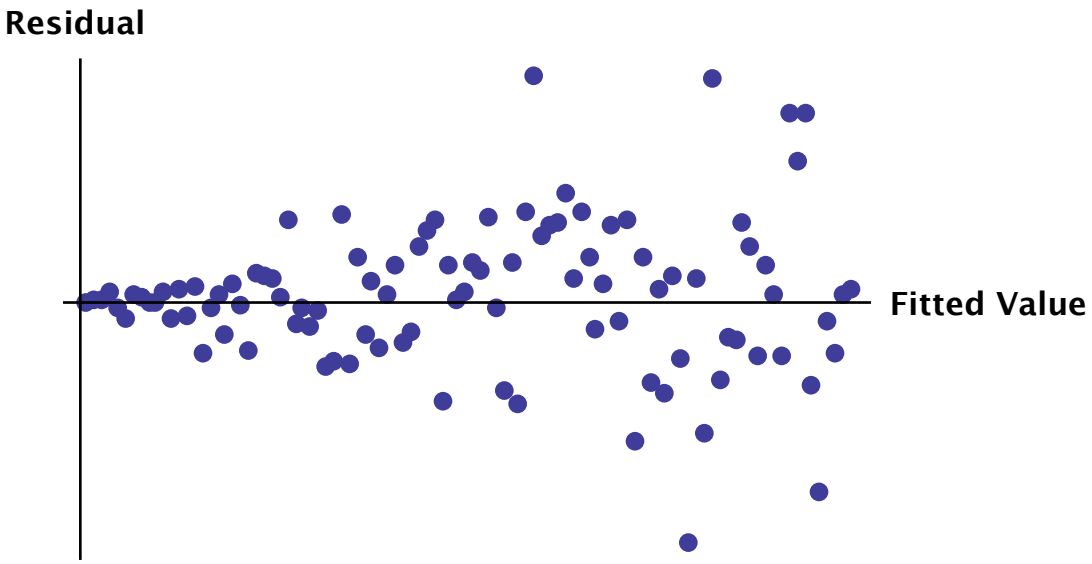
- Claim size follows an Inverse Gaussian distribution.
- Log is the selected link function.
- The dispersion parameter is estimated to be 0.00510.
- Territory and gender are used in the model.
- Selected Model Output:

Variable	$\hat{\beta}$
Intercept	8.03
Territory D	0.18
Gender - Male	0.22

Calculate the standard deviation of the predicted claim size for a male in Territory D.

3.64. (2 points) List four ways that an actuary can analyze the appropriateness of a Generalized Linear Model.

3.65. (1 point) Briefly comment on the following plot of deviance residuals of a model as a function of the fitted values:



3.66. (1 points) You have fit a Generalized Linear Model using an exponential family.
What is the deviance?

3.67. (1 point) A GLM has been fit with a log link function.
Age is used, grouped into categories.
Gender is used.
There are categories of Use of Vehicle.
Territories are used.
The expected pure premium for the base is $207.
For the age group 24-26 the coefficient is 0.43.
For Male the coefficient is 0.22.
For Pleasure Use (No Driving to Work) the coefficient is -0.32.
For Territory H the coefficient is 0.36.
Determine the expected pure premium for a male, 24-26 years old, Pleasure Use, in Territory H.

3.68. (1 point) Define and briefly discuss ensemble models.

3.69. (2 points) A GLM using a Gamma Distribution and a log link function is being used to model severity of personal injury claims. There are 25,000 observations.
3 parameters were fit: an intercept, time until settlement, and whether there is legal representation.
The deviance is 24,359. The estimated dispersion parameter is 1.22.
A variable is added to the model, equal to the product of the time until settlement and the legal representation variable. (This is an interaction variable.)
The deviance is now 24,352.
Determine whether this additional variable should be added to this model.
You may use the following:
If X follows an F-Distribution with 1 and n degrees of freedom,
then $\sqrt{X}$ follows a t-distribution with n degrees of freedom.
For n large, a t-distribution is approximately a Standard Normal Distribution.
Selected percentiles of the Standard Normal Distribution:

	Values of z for selected values of Pr(Z < z)						
z	0.842	1.036	1.282	1.645	1.960	2.326	2.576
Pr(Z < z)	0.800	0.850	0.900	0.950	0.975	0.990	0.995

3.70. (1.5 points) Fully discuss model stability and some ways to assess it.

3.71. (8 points) You are given 19 data points:
258, 636, 652, 814, 833, 860, 895, 937, 950, 1009,
1020, 1059, 1103, 1113, 1127, 1139, 1246, 1335, 1770.
You wish to compare this data to a Normal Distribution with $\mu = 1000$ and $\sigma = 300$.
With the aid of a computer, draw a Q-Q plot.

3.72. (4 points) For private passenger automobile liability claim frequency, you use three factors: gender, age of driver, and territory.
There are 4 levels for driver age, and 3 territories.
A GLM with a log link function is fit.
An intercept term is used.
Let β_1 correspond to the intercept term, β_2 correspond to male,
and assign the other parameters as follows:

Age of driver			Territory	
Factor level	Parameter		Factor level	Parameter
17-21	β_3		A	β_6
22-29	β_4	B		
30-59			C	β_7
60+	β_5			

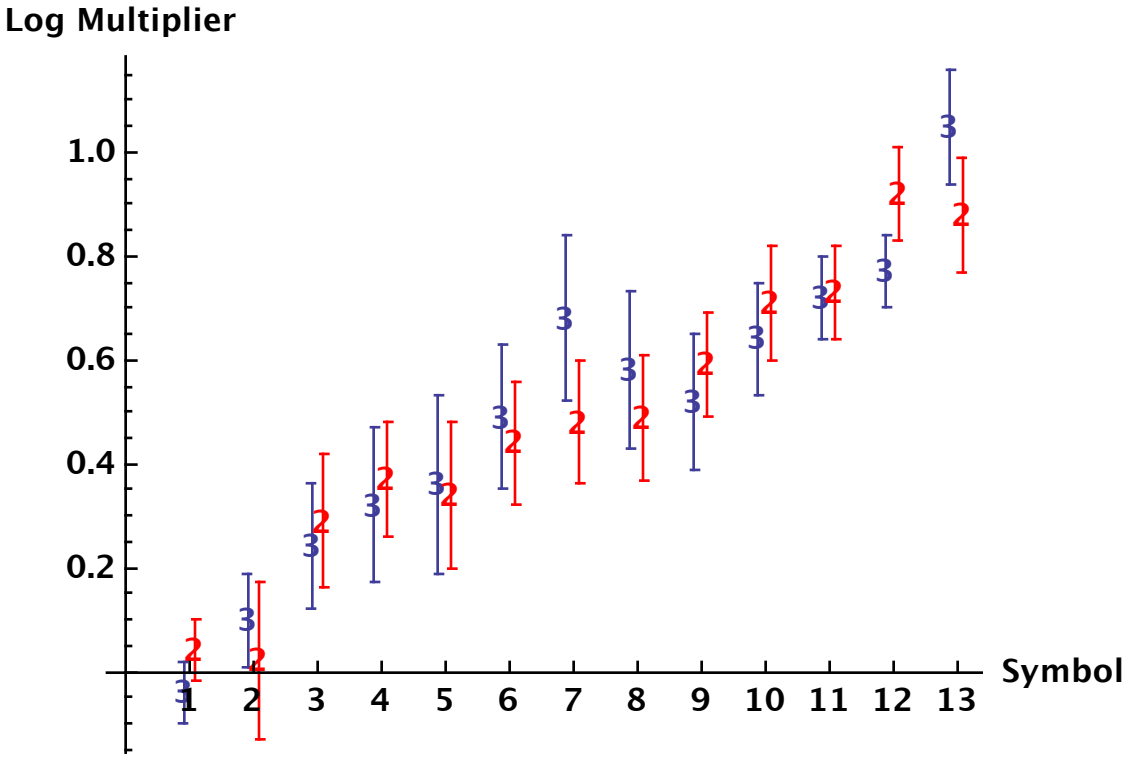
(a) (3 points) What is the design matrix?
(b) (0.5 point) In terms of the fitted parameters, what is the estimated frequency for a 30-59 year old female driver in Territory B?
(c) (0.5 point) In terms of the fitted parameters, what is the estimated frequency for a 22-29 year old male driver in Territory C?

3.73. (1.5 points) Five Generalized Linear Models have been fit to the same set of 200 observations.

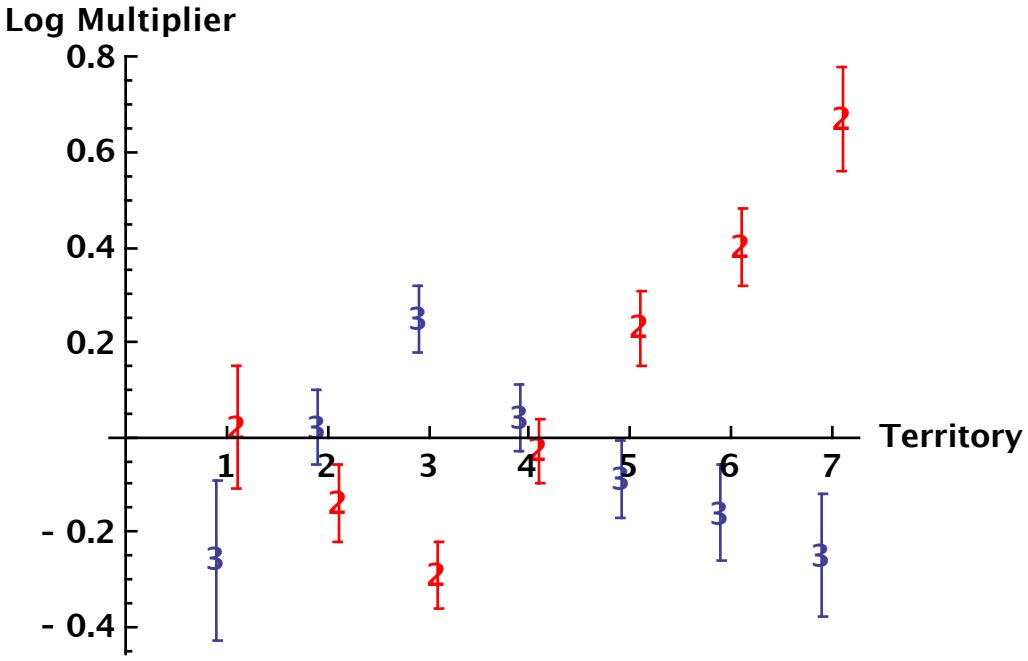
Model	Number of Fitted Parameters	LogLikelihood
A	3	-359.17
B	4	-357.84
C	5	-356.42
D	6	-354.63
E	7	-353.85

Which model has the best AIC (Akaike Information Criterion)?

3.74. (1.5 points) The following graph displays the modeled log of the relativity by vehicle symbol, for a base level of the other predictor variables in a GLM, for two separate years of data.
Approximate 95% confidence intervals are shown.



Here is a second similar graph for a different model, by Territory:



Briefly compare and contrast what the two graphs tell the actuary about each model.

3.75. (1.5 points) Before embarking on a GLM modeling project, it is important to understand the correlation structure among the predictors.
Discuss why this is important and what actions may be indicated.

3.76. (1 point) Multiplicative models are the most common type of rating structure used for pricing insurance, due to a number of advantages they have over other structures.
Briefly discuss two advantages of a multiplicative rating structure.

3.77. (1.5 points) A GLM using a Gamma Distribution has been fit for modeling severity of medical malpractice claims. There are 1000 observations.
50 parameters were fit, including an intercept.
It uses gender and 6 categories of age of claimant.
The deviance is 1120.3.
An otherwise similar GLM excluding gender and age of claimant has a deviance of 1128.1, and an estimated dispersion parameter of 0.395.
Discuss how you would use an F-Test to determine whether age and gender should be used in this model.

3.78. (1.5 point) Briefly discuss limitations on the use of the loglikelihood and deviance to compare the fit of two GLMs.

3.79. (1 point) An insurer sells “Disgrace Insurance” which covers a business against the possibility that their celebrity spokesperson may engage in disgraceful behavior or expressions. You are putting together Generalized Linear Models (GLMs) to try to develop a rating algorithm. Assuming you have plenty of good data, list some variables you would include in your testing of possible GLMs.

3.80. (1 point) Compare and contrast the Gamma and the Inverse Gaussian Distributions.

3.81. (1.5 points) An actuary has historical information relating to personal loan default rates.
A logistic model (GLM with a logit link function) was used to estimate the probability of default for a given customer.
The two variables determined to be significant were the size of loan in thousands of dollars and the credit score of the customer.
β_0 corresponds to the intercept term, β_1 corresponds to size of loan, and
and β_2 corresponds to credit score
The parameter estimates were determined to be as follows:

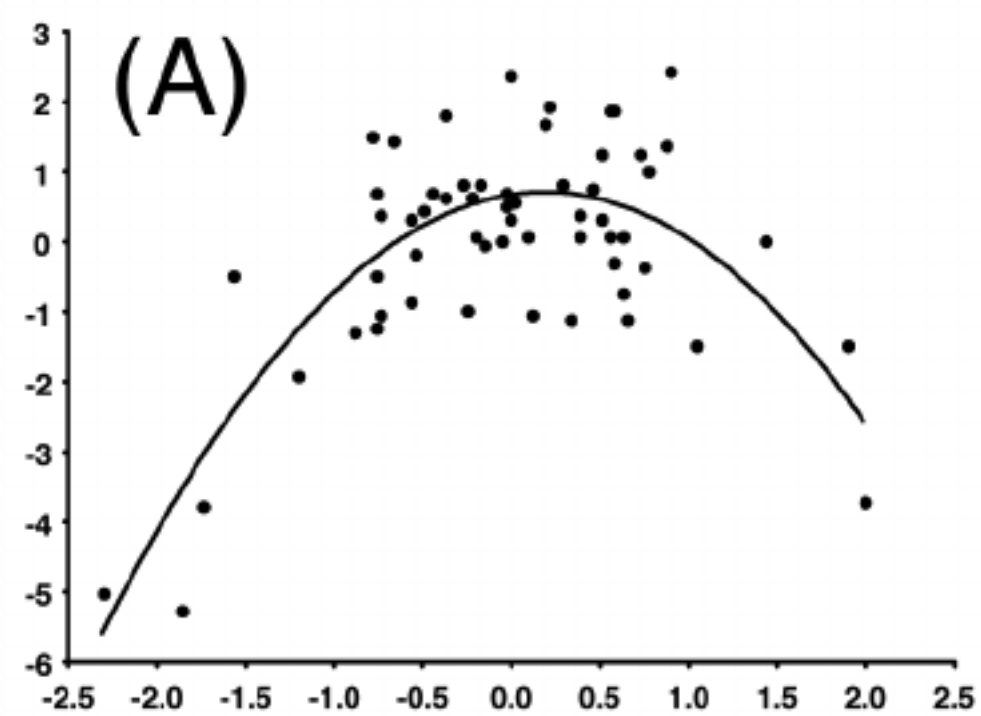
β_0 9.5

β_1 0.01

β_2 -0.02

a. (0.75 point) Calculate the estimated default rate for a customer who has credit score of 670 and took out a loan for $180,000.
b. (0.75 point) Calculate the estimated default rate for a customer who has credit score of 760 and took out a loan for $100,000.

3.82. (1 point) For a GLM, here is a partial residual plot for the predictor variable X_1:



Briefly discuss the conclusion from this plot.
If necessary, what is a possible solution?

3.83. (6 points) We model average claim severity by type and horsepower of the car:
• Type: Sedan or SUV
• Horsepower: Low, Medium, or High
We observe an equal number of vehicles of each of the six possible types,
and the observed average claim severities are:

	Sedan	SUV
Low Horsepower	800	1,500
Medium Horsepower	900	1,700
High Horsepower	1,100	2,000

We will fit a GLM using a Gamma Distribution.
(a) (3 points) For an additive model, determine the maximum likelihood equations to be solved.
(b) (3 points) For an multiplicative model, determine the maximum likelihood equations to be solved.

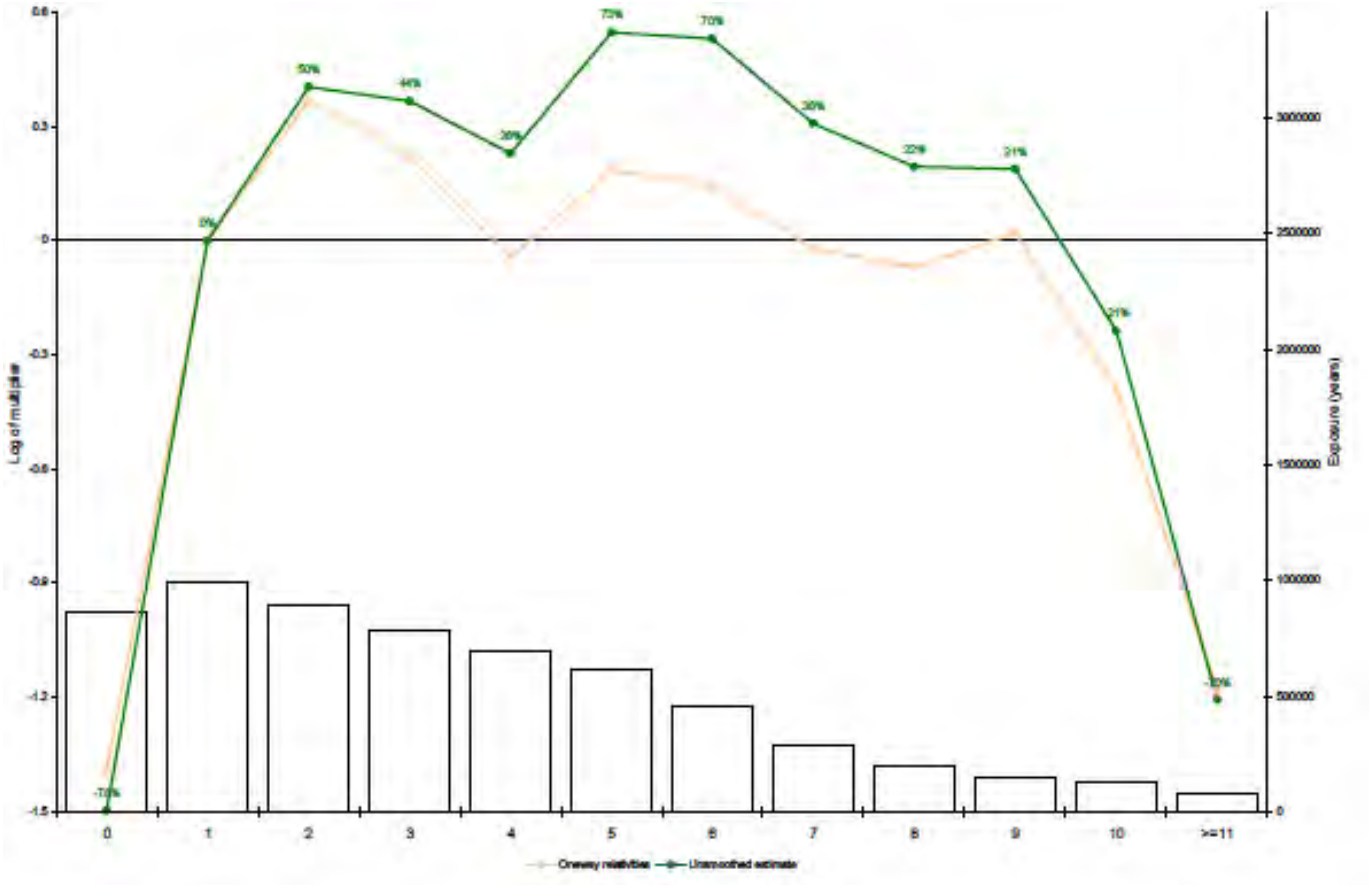
3.84. (2 points) A GLM using an Inverse Gaussian Distribution and an inverse link function is being used to model severity of private passenger automobile property damage liability claims.
There are 2000 observations.
14 parameters including an intercept were fit.
The deviance is 1848.5, and the estimated dispersion parameter is 0.93.
A categorical variable is added to the model based on vehicle type, with a total of 10 categories.
The deviance for this more complex model is 1833.0.
Discuss how you would use an F-Test to determine whether vehicle type should be added to this model at the 5% significance level.

3.85. (1.5 points) Five Generalized Linear Models have been fit to the same set of 250 observations.

Model	Number of Fitted Parameters	Deviance
A	6	1679.1
B	8	1666.4
C	10	1655.9
D	12	1646.2
E	14	1634.5

Which model has the best BIC (Bayesian Information Criterion)?

3.86. (1 point)
A Generalized Linear Model was fit to data on lapse rates for life insurance policies.
Three predictor variables were included in the GLM:
calendar year of exposure, policy duration, and product class.
The graph below displays logs of the relativities by policy duration.
For each band, the black bars at bottom show exposure, quantified on the righthand axis.
The GLM results are in green, and are relative to the base level for policy duration.
The yellow line (lighter line) is what would have been generated by a 'one-way' analysis: i.e., considering just policy duration, without any other factors.



Briefly discuss a likely reason why the green and yellow lines differ.

3.87. (0.5 points) A continuous predictor x_2 has a coefficient of $\beta_2 = -0.3$ in a logistic model.
For a unit increase in x_2, what is the estimated change in the odds?

3.88. (1 point) We are fitting a GLM to private passenger automobile liability pure premiums.
Female drivers age 31 to 59 in a rural territory may have observed pure premiums higher or lower than their fitted values.
Unmarried male drivers age 17 to 21 in an urban territory may have observed pure premiums higher or lower than their fitted values.
Contrast the effect on fitting the GLM of the modeling errors from these two groups.

3.89. (1 point) A logistic regression has been fit to some data. For a certain threshold:

		Predicted Fraud		
		No	Yes	Total
Actual	No	70,000	10,000	80,000
Fraud	Yes	3000	2000	5000
	Total	73,000	12,000	85,000

What point would be plotted in the ROC curve?

3.90. (1 point) How would the standard error help to analyze the results of fitting a Generalized Linear Model (GLM)?

3.91. (1 point) For a rating plan, briefly discuss how to construct a Lorenz Curve and compute the Gini Index.

3.92. (4 points) Assume a set of three observations:
For z = 1, we observe 4. For z = 2, we observe 7. For z = 3, we observe 8.
Fit to these observations a Generalized Linear Model with a Poisson Distribution and a log link function. In other words, assume that each observation is a Poisson random variable, with mean λ and $\ln(\lambda) = \beta_0 + \beta_1 z$.

3.93. (1 point) In addition to statistical significance, give other considerations for variable selection.

3.94. (3.5 points) A personal auto class system has three class dimensions:
• Sex: Male vs female
• Age: Youthful vs adult vs retired
• Territory: Urban vs suburban vs rural
An actuary sets rate relativities from the experience of 20,000 cars.
• Urban is the base level in the territory dimension.
• Adult is the base level in the age dimension.
• Male is the base level in the sex dimension.
a. (0.5 point) How many elements does the vector of covariates have in a multiplicative model?
b. (0.5 point) How many elements does the vector of covariates have in an additive model?
c. (1 point) Specify each element of the vector of parameters, with $\beta_0 \Leftrightarrow$ the base class.
d. (0.5 point) How many columns does the design matrix have?
e. (0.5 point) How many rows does the design matrix have if each record is analyzed separately?
f. (0.5 point) For grouped data, how many rows does the design matrix have?

3.95. (2 points) Answer the following with respect to deviance residuals of a GLM.
(a) (0.5 points) Define the deviance residual.
(b) (0.5 points) Give an intuitive interpretation of deviance residuals.
(c) (1 point) Discuss how deviance residuals can be used to check the fit of a model.

3.96. (4 points) You have the following data on the renewal of homeowners insurance policies with the ABC Insurance Company:

Number of Years Insured	Number of Policies	Number of Policies Renewed
1	1000	900
2	900	820
3	800	740
4	700	660
5	600	580

Let X = number of years insured with ABC Insurance Company.
A Generalized Linear Model using a Binomial Distribution with a logit link function will be fit to this data, including an intercept term.
Determine the equations to be solved in order to fit this model via maximum likelihood.

3.97. (0.5 points) The variance of a distribution from the exponential family can be expressed using the following formula: $\text{Var}(y_i) = \frac{\phi \, V(\mu_i)}{\omega_i}$.

Define the parameters ϕ and ω_i in the formula above.

Use the following information for the next five questions:

X	2	5	8	9
Y	10	6	11	13

Y_1, Y_2, Y_3, Y_4 are independently Normally distributed with means $\mu_i = \beta_0 + \beta_1 X_i$, i = 1, 2, 3, 4, and common variance σ^2.

3.98. (2 points) Determine $\hat{\beta}_1$ via maximum likelihood.

3.99. (2 points) Determine $\hat{\beta}_0$ via maximum likelihood.

3.100. (2 points) Determine $\hat{\sigma}$ via maximum likelihood.

3.101. (3 points) Estimate the standard deviation of $\hat{\beta}_1$.

3.102. (3 points) Estimate the standard deviation of $\hat{\beta}_0$.

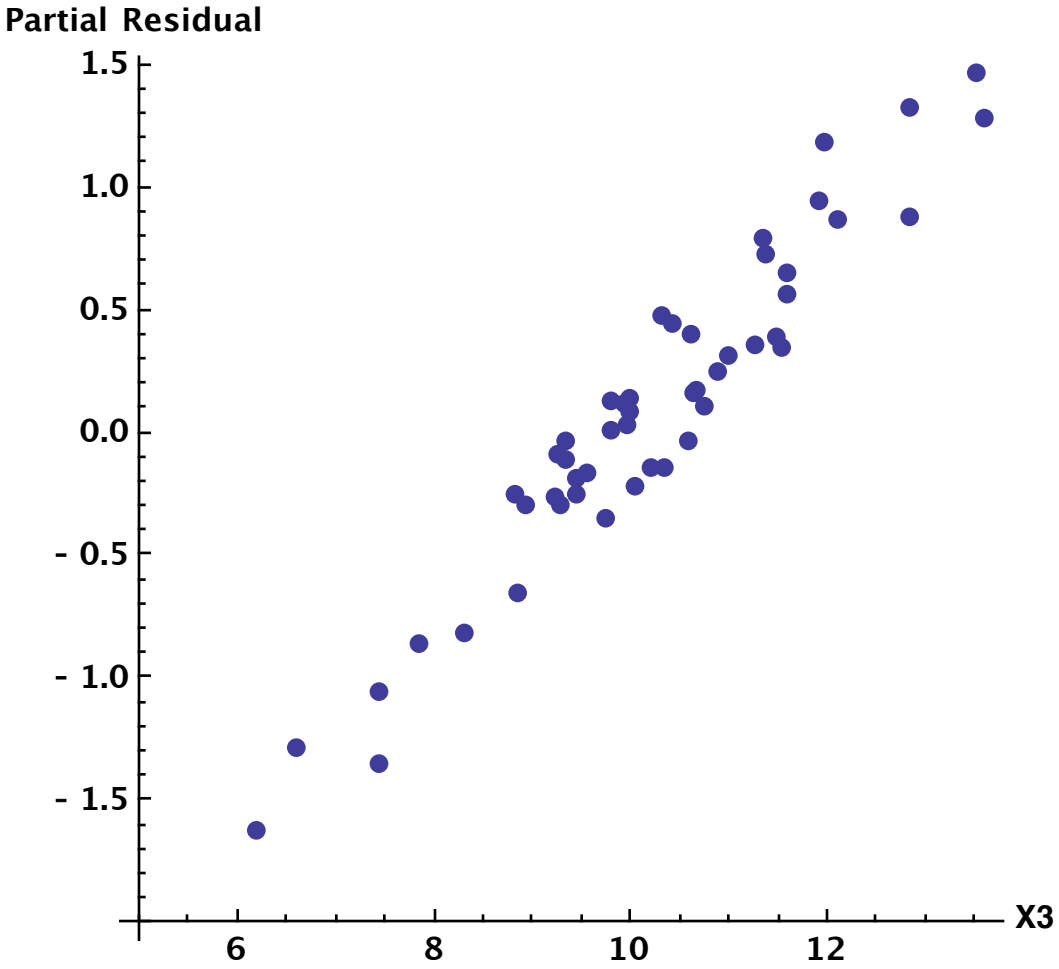
3.103. (1 point) Five Generalized Linear Models have been fit to the same set of observations. Each model uses the same number of parameters.
Which of these models is preferred?

Model	Deviance
A	3609.5
B	3611.0
C	3606.3
D	3602.1
E	3605.8

3.104. (1 point) Discuss the overdispersed Poisson Distribution.

3.105. (1 point) A common statistical rule of thumb is to reject the null hypothesis where the p-value is 0.05 or lower. Is this appropriate for a typical insurance modeling project?
Why or why not?

3.106. (1 point) For a GLM, here is a partial residual plot for the predictor variable X_3:



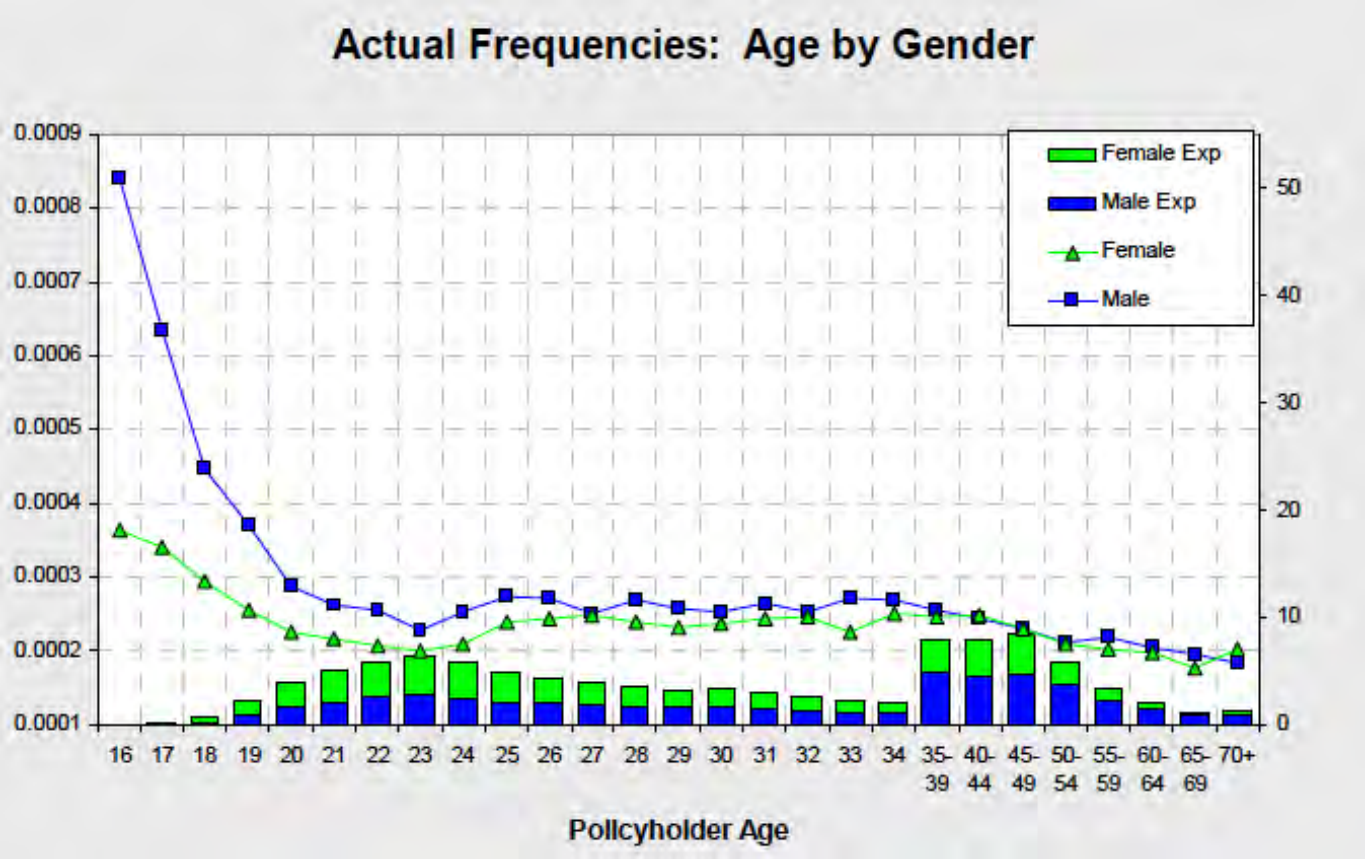
Briefly discuss the meaning of this plot.
If necessary, what is a possible solution?

3.107. (0.5 points) With respect to GLMs, briefly discuss multicollinearity.

3.108. (1 point) An actuary is determining the rates by class and territory.
With respect to GLMs, briefly discuss determining territory relativities.

3.109. (1 point)
Define a holdout sample of data, and briefly discuss how it can be used in GLM validation.

3.110. (1 point) The following graph shows claim frequency for private passenger automobile insurance by gender and age. (The rectangles represent the number of exposures.)



Briefly discuss the implications for modeling frequency via a Generalized Linear Model.

3.111. (2 points) Using Generalized Linear Models, an actuary Edward Conners has developed a policy renewal model for private passenger automobile insurance written by the Some States Insurance Company. There are two predictor variables:

z_1 = the number of years the insured has been with Some States.

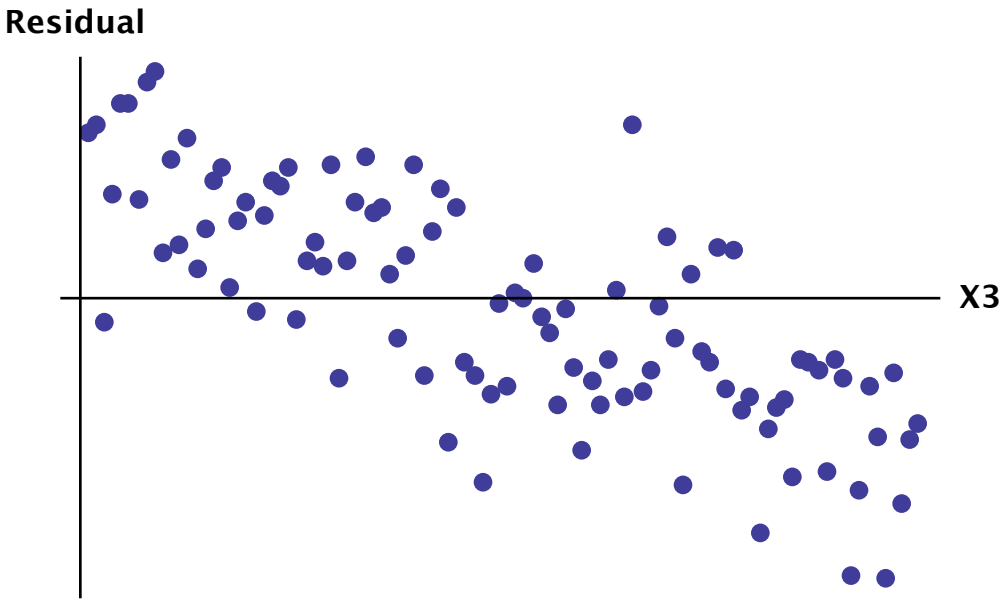
z_2 = the age of the principal operator of the vehicle.

The predicted probability of policy renewal is: $\frac{\text{Exp}[0.6 + 0.05\, z_1 + 0.02\, z_2]}{1 + \text{Exp}[0.6 + 0.05\, z_1 + 0.02\, z_2]}$.

(a) For a principal operator who is 30 years old, what is the multiplicative relativity of 1 year with Some States compared to 10 years with Some States?

(b) For a principal operator who is 50 years old, what is the multiplicative relativity of 1 year with Some States compared to 10 years with Some States?

3.112. (1 point) Briefly comment on the following plot of deviance residuals of a model as a function of a predictor variable X_3:



3.113. (6 points) You are given the following information on the labor force participation of 10 married women between the ages of 25 and 35:

Child of Age 6 or Less	Years of Education	Participating in the Labor Force
No	12	Yes
No	14	No
No	15	Yes
No	16	No
No	17	Yes
Yes	10	No
Yes	11	No
Yes	13	Yes
Yes	15	No
Yes	16	Yes

A Generalized Linear Model using a Binomial Distribution with a logit link function will be fit to this data, including an intercept term.

a. (1 point) What are the design matrix and the response vector?

b. (5 points) Determine the equations to be solved in order to fit this model via maximum likelihood.

3.114. (1 point) Les N. DeRisk is an actuary. Les has scrubbed and adjusted the data he will be using for classification ratemaking for a certain line of insurance.
Les will run a Generalized Linear Model. List 3 things Les has to specify.

3.115. (1.5 points) You are given two simple quantile plots, one sorted by the current plan and one sorted by a proposed plan.
Discuss the lift of the proposed plan compared to the current plan.

3.116. (0.5 point) Give an example of a situation where a GLM with a Binomial distribution and logit link function would be used.

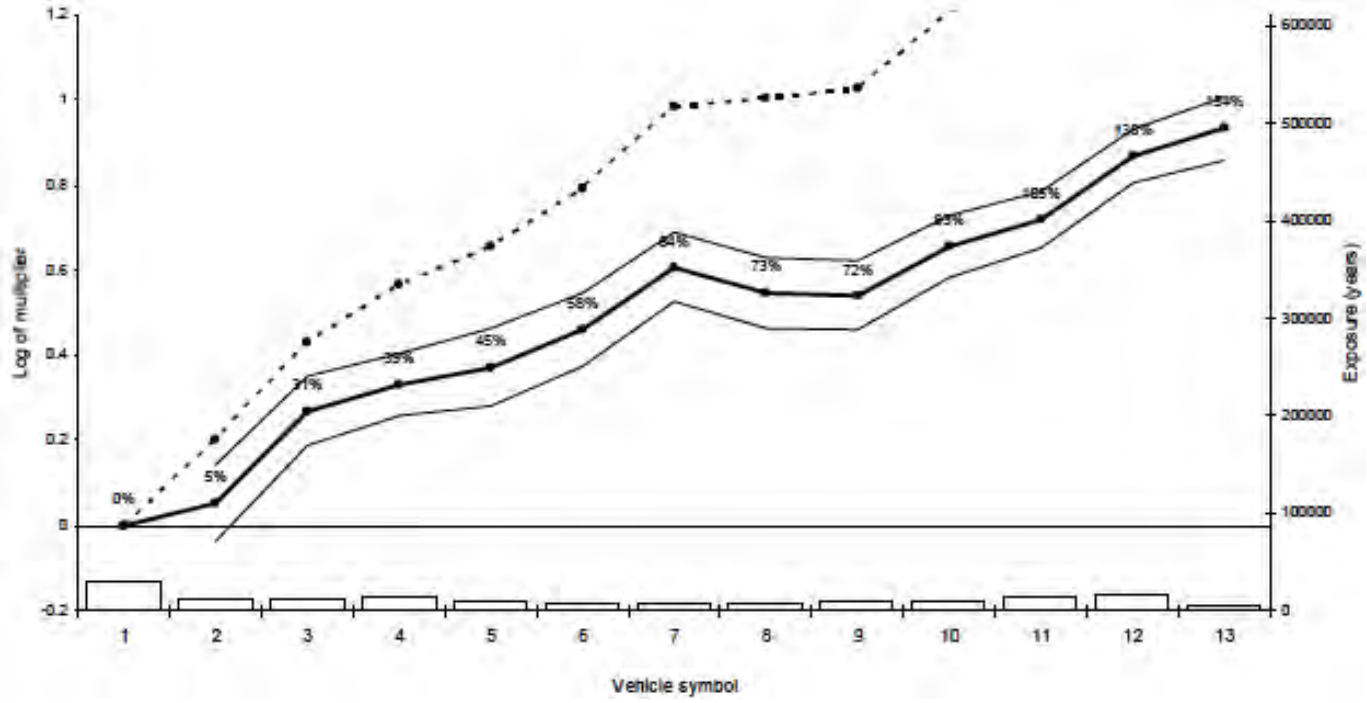
Use the following information for the next two questions:

- A GLM using a Gamma Distribution and a log link function has been fit for modeling severity of auto claims.
- The explanatory variables are: x_1 driver age, and x_2 marital status where 1 = married.
- The fitted coefficients are: $\beta_0 = 8.80$, $\beta_1 = -0.03$, $\beta_2 = -0.15$.
- The estimated $\phi = 0.3$.

3.117. (1 point) Determine the estimated mean severity for a 30 year old married driver.

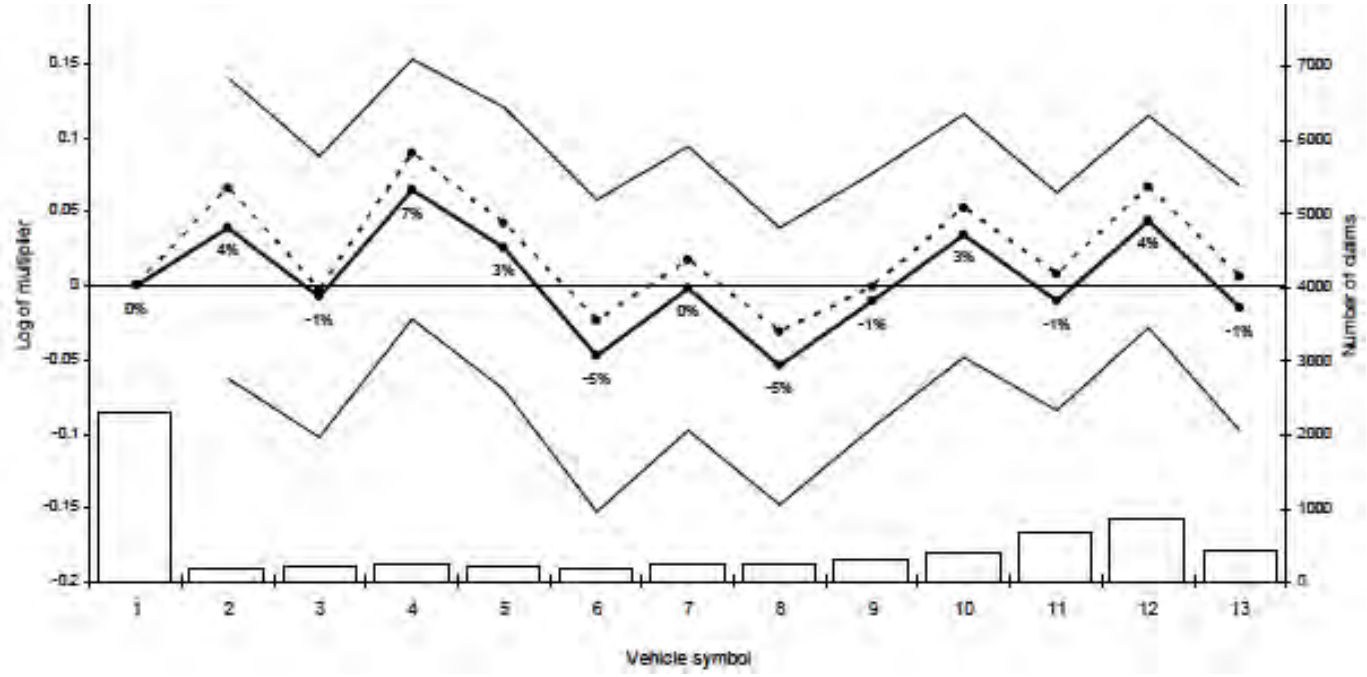
3.118. (1 point) Determine the estimated variance of severity for a 40 year old unmarried driver.

3.119. (2 points) The following graph displays the modeled log of the relativity by vehicle symbol, for a base level of the other predictor variables in a GLM.
The bold line shows the fitted parameter estimates.
Lines indicates two standard errors on either side of the parameter estimate.
The dotted line show the relativities implied by a simple one-way analysis.
The distribution of exposure for all business considered is also shown as a bar chart at the bottom.



Question continued on the next page.

Here is a second similar graph for a different model.



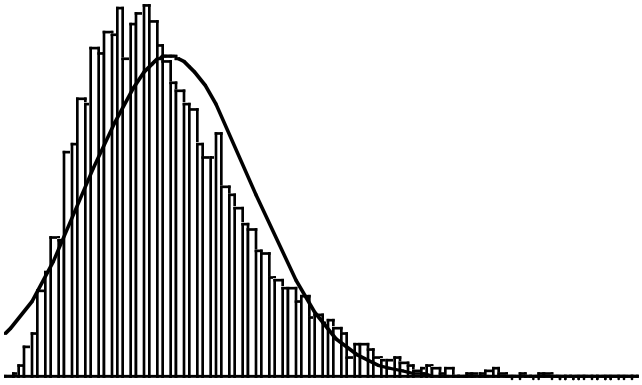
Briefly compare and contrast what the two graphs tell the actuary about each model.

3.120. (1 point) For a line of insurance, an actuary fits separate GLMs to different perils. Discuss one way to combine separate models by peril in order to get a model for all perils.

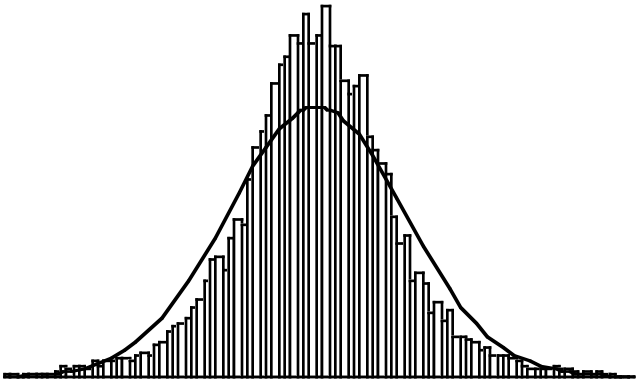
3.121. (2 points) Claim counts for private passenger automobile insurance are Poisson. The mean frequency, m, depends on age and gender.
Briefly discuss and contrast the following two models, where x is age.
(a) (1 point) $\log(\mu) = \alpha_i + \beta x$, where α_1 and α_2 depend on gender.
(b) (1 point) $\log(\mu) = \alpha_i + \beta_i x$, where α_1, α_2, β_1, and β_2 depend on gender.

3.122. (1 point) The following a histograms of deviance residuals for GLMs.
Which of the following histograms represents the best model?

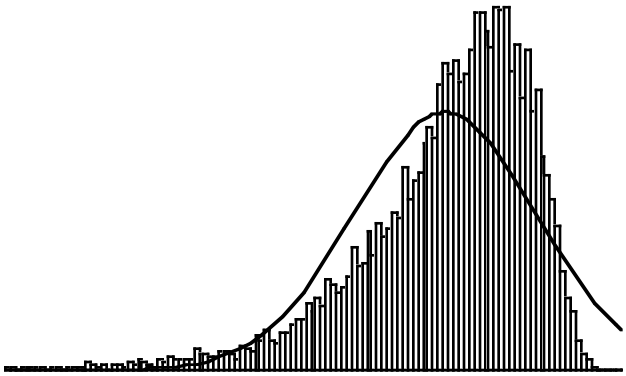
A.



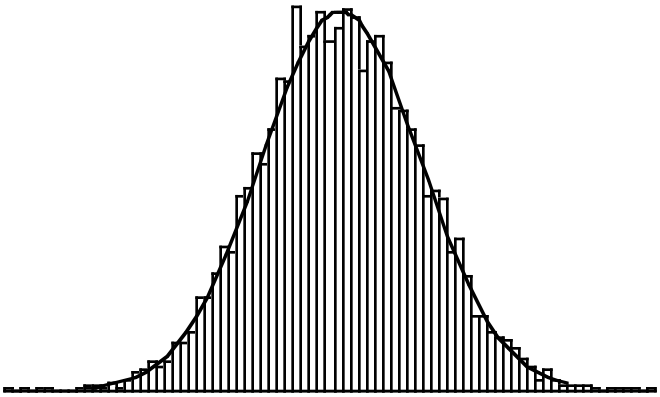
B.



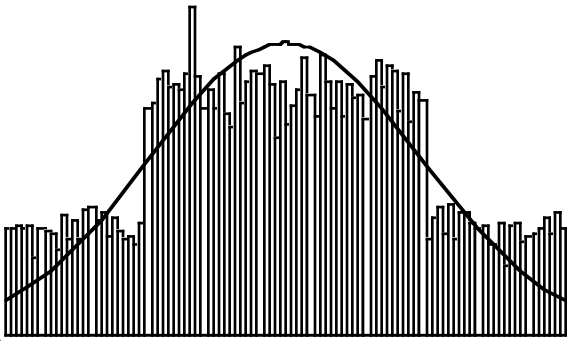
C.



D.



E.



3.123. (1 point) Geoff Linus Modlin is an actuary using Generalized Linear Models (GLMs) to determine classification rates for private passenger automobile insurance.
Geoff notices that the relativity for drivers aged 19 is different between two GLMs based on the same data. Briefly discuss why that can be the case.

3.124. (1 point) You observe 36 monthly returns on a stock.
The 9th value from smallest to largest is 0.004.
What is the corresponding point in the Normal Q-Q Plot?

3.125. (1.5 points) With respect to GLMs, fully discuss variance inflation factors (VIF).

3.126. (1.5 points) Dollar Bill Bradley, an actuary at the Knickerbocker Insurance Company, has fit a Generalized Linear Model with a overdispersed Poisson error structure and a log link function in order to model claim frequency for automobile liability insurance.
His model has a deviance of 2196.1 and estimated dispersion parameter of 2.09.
Bill now introduces into the model an additional categorical variable with five categories:
1. Insured has homeowners insurance with Knickerbocker.
2. Insured has homeowners insurance with another insurer.
3. Insured has renters insurance with Knickerbocker.
4. Insured has renters insurance with another insurer.
5. Other
With this additional variable, the model has a deviance of 2179.3.
The null hypothesis is to use the simpler model.
The alternative hypothesis is to use the more complicated model.
Determine the F-test statistic and discuss how you would perform the statistical test.

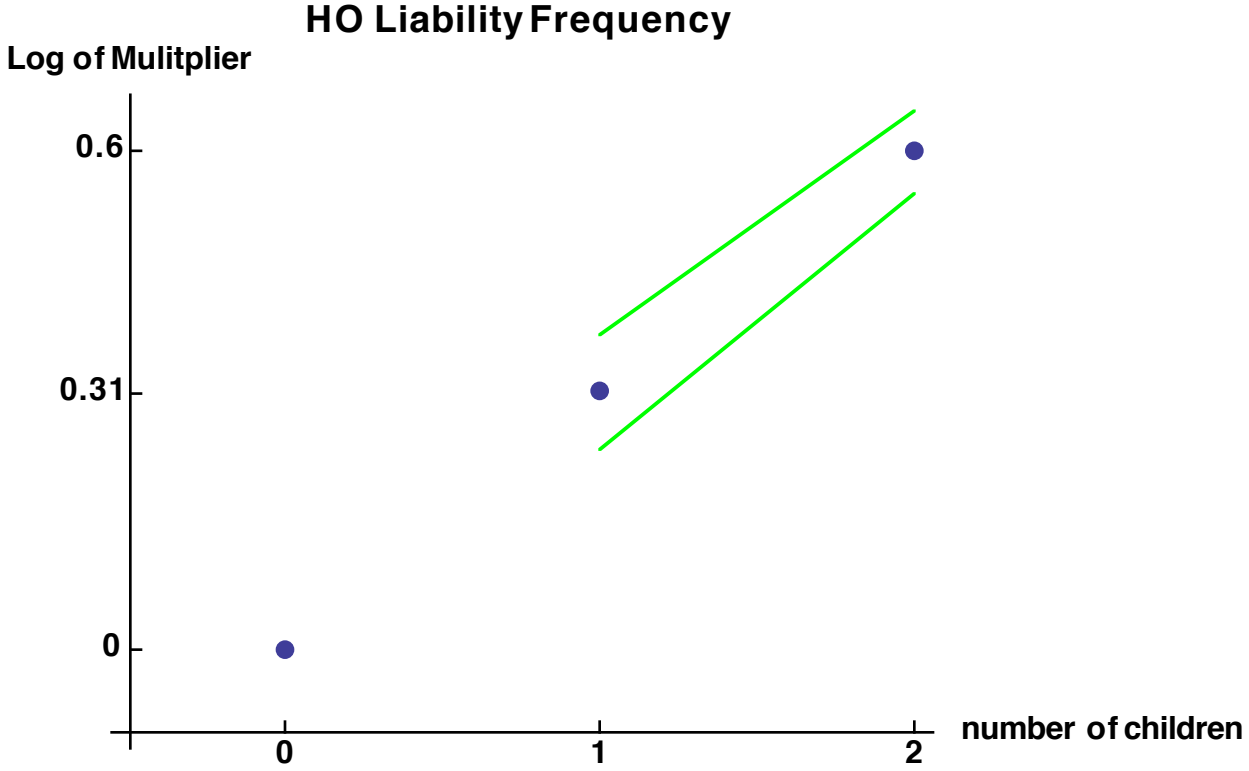
3.127. (1 point) Discuss cross validation as used with GLMs.

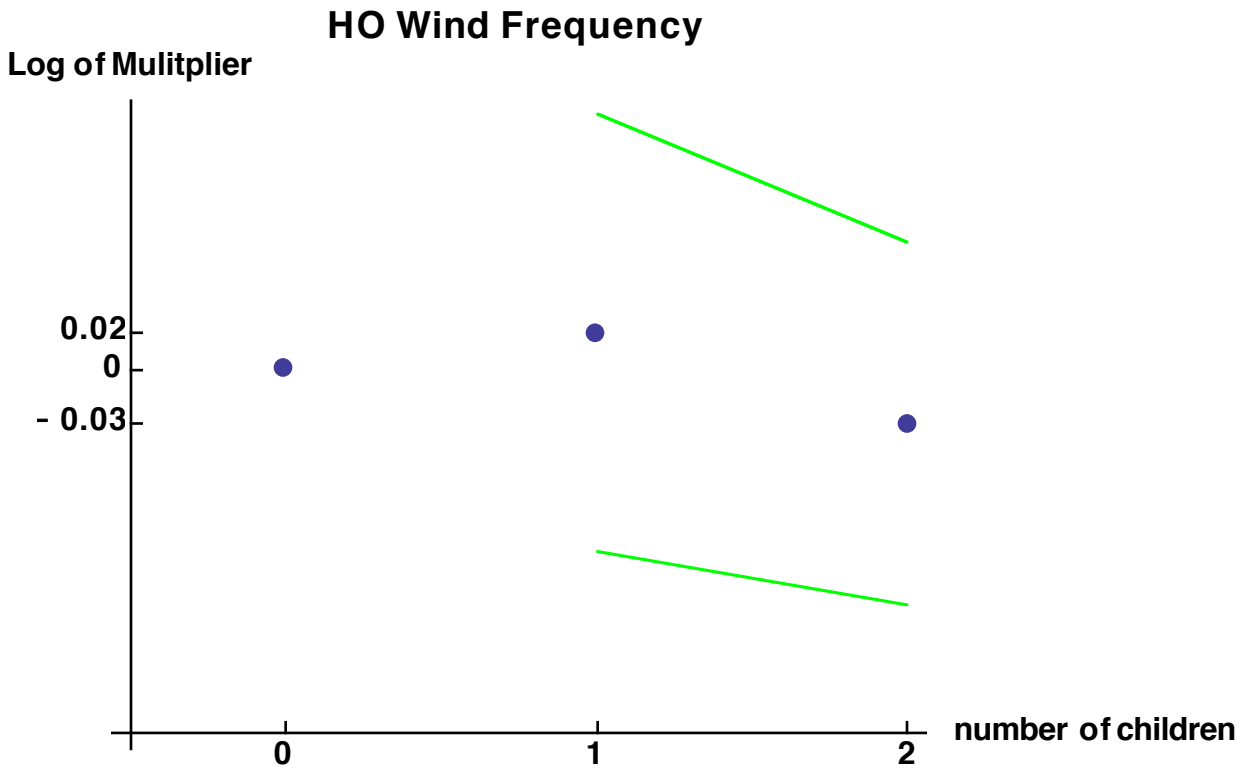
3.128. (1 point) A GLM has been fit in order to predict blood pressure of individuals.

Variable	Coefficient	VIF
Constant	-12.87	
Age	0.7033	1.76
Weight	0.9699	10.42
Body Surface Area	3.780	6.33
Duration of Hypertension	0.0684	1.24
Basal Pulse	-0.0845	4.41
Stress Index	0.00341	1.83

Briefly discuss this output.

3.129. (2 points) Below are graphs of GLMs fit to Homeowners frequency data, showing the natural log of the fitted multiplicative factors for one or two children in the house relative to none. Also shown are approximate 95% confidence intervals.
Briefly compare and contrast what the two graphs tell the actuary about each model.





3.130. (2.5 points) An actuary at a private passenger auto insurance company wishes to use a generalized linear model to create an auto severity model using the data below.

	Dollars of Loss	
Gender	Territory A	Territory B
Male	700,000	500,000
Female	400,000	300,000

	Number of Claims	
Gender	Territory A	Territory B
Male	800	700
Female	600	500

The model will include three parameters: β_1, β_2, β_3, where β_1 is the average severity for males, β_2 is the average severity for Territory A, and β_3 is an intercept.

Assuming $\beta_3 = 570.356$, solve a generalized linear model with a normal error structure and identity link function for β_1.

3.131. (1.5 points) Five Generalized Linear Models have been fit to the same set of 60 observations.

Model	Number of Fitted Parameters	LogLikelihood
A	2	-220.18
B	3	-217.40
C	4	-214.92
D	5	-213.25
E	6	-211.03

Which model has the best BIC (Bayesian Information Criterion)?

3.132. (1 point) You fit a GLM using year as one of the predictor variables.
The values of year in your data are: 2010, 2011, 2012, 2013, and 2014.
You pick 2012 as the base level.
Applying statistical tests you determine that the coefficients for 2011 and 2014 are not significant.
Discuss what would you do.

3.133. (3 points) You are given the following wage distribution table:

Ratio to SAWW	Cumulative Portion of Workers	Cumulative Portion of Wages
0.10	0.18%	0.01%
0.20	0.93%	0.13%
0.30	3.53%	0.79%
0.40	6.85%	1.96%
0.50	11.33%	4.00%
0.60	18.49%	7.98%
0.70	28.57%	14.56%
0.80	40.05%	23.13%
0.90	48.99%	30.75%
1.00	57.47%	38.80%
1.10	64.98%	46.69%
1.20	71.14%	53.76%
1.30	76.34%	60.25%
1.40	80.99%	66.51%
1.50	85.33%	72.80%
1.75	92.86%	84.92%
2.00	96.91%	92.48%
2.25	98.73%	93.41%
2.50	99.28%	94.41%
3.00	99.66%	95.79%
4.00	99.87%	97.28%
5.00	99.93%	98.05%
6.00	99.96%	98.52%
7.00	99.97%	98.84%

With the aid of a computer, draw the corresponding Lorenz curve.

3.134. (2 points) Assume there are two models, Model A and Model B, both of which produce an estimate of the expected loss cost (pure premium) for each policyholder.
Discuss using Simple Quantile Plots to compare the two models A and B.
How are Simple Quantile Plots created?
How would one determine the winning model?

3.135. (1.5 points) A logistic model was built to predict the probability of a claim being fraudulent.
(a) Briefly define the discrimination threshold.
(b) Briefly discuss the selection of what discrimination threshold to use.

3.136. (4 points) An actuary is considering using a generalized linear model to estimate the expected frequency of a recently introduced insurance product.
Given the following assumptions:
- The expected frequency for a risk is assumed to vary by territory and gender.
- A log link function is used.
- A Poisson error structure is used.
- β_0 is the intercept.
- β_1 is the effect of gender = Female.
- β_2 is the effect of Territory = B.

Number of Claims

Gender	Territory A	Territory B
Male	1200	1100
Female	800	900

Number of Exposures

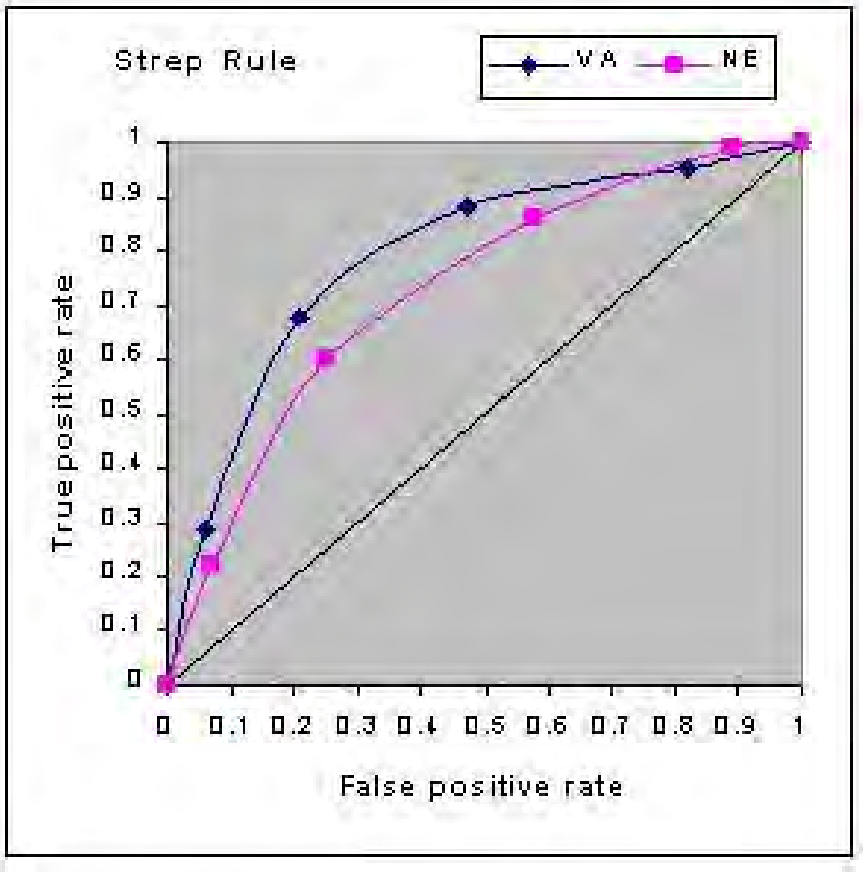
Gender	Territory A	Territory B
Male	24,000	15,000
Female	20,000	13,000

Given that β_0 = -3.0300, determine the expected frequency of a female risk in Territory B.

3.137. (1 point) Briefly discuss interaction in GLMs and give an example of an interaction term.

3.138. (2 points) A GLM has been used to develop an insurance rating plan.
There are only two classes A and B, with equal numbers of exposures.
The predicted pure premium for Class A is less than that for class B
(a) Determine the Gini Index if the actual losses for the two classes are equal.
(b) Determine the Gini Index if the actual losses for Class A are 0 and for Class B are positive.

3.139. (0.5 points) The following ROC curves are for two medical tests for strep throat:



Which test do you prefer and why?

3.140. (1.5 points)

A GLM has been fit using a Poisson Distribution with $\hat{\beta}_1 = 5.624$ with standard error 0.1978.

Using instead an overdispersed Poisson the estimate of ϕ is 3.071.

For this second model, determine a 95% confidence interval for β_1.

3.141. (1.75 points) An analyst has fit several different variations of a GLM to a large dataset in order to predict pure premiums.

For each model variation listed below, draw a simple quintile plot based on the training data.

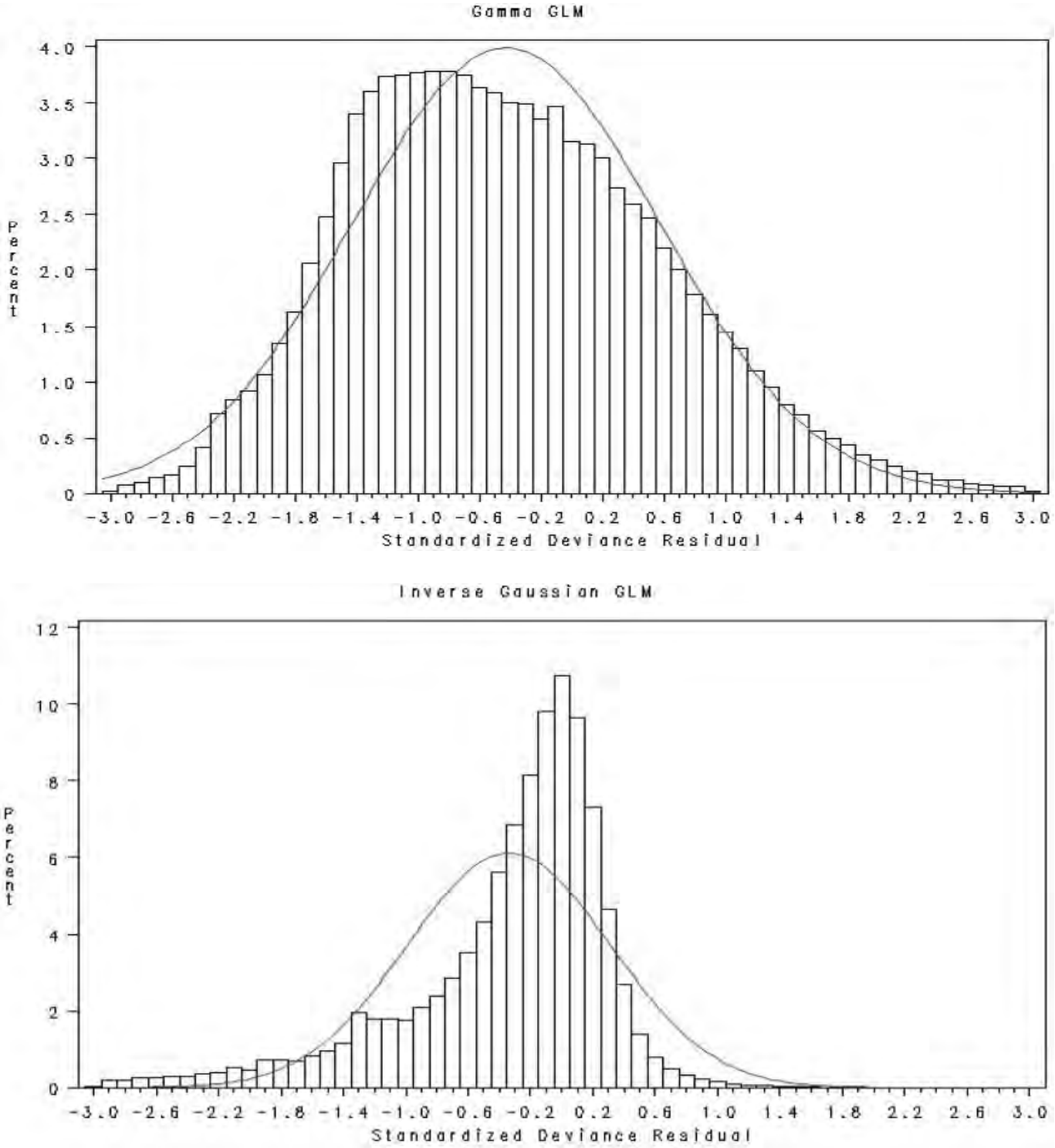
Label the axes and identify each data series.

i. A saturated model

ii. A null model

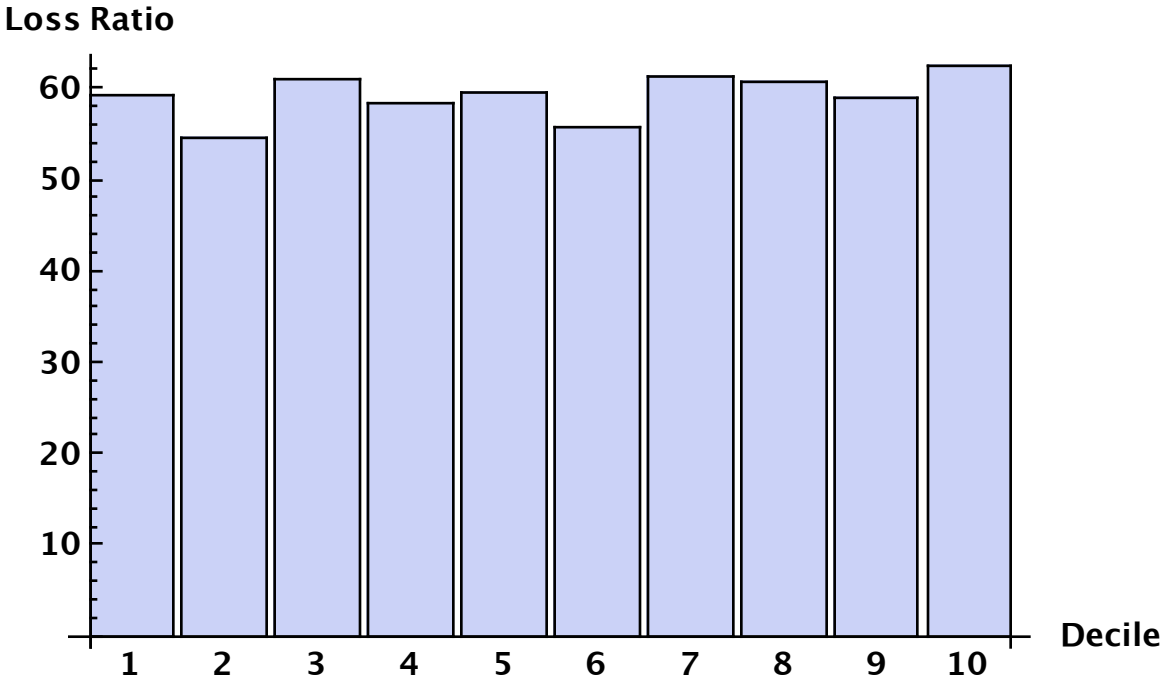
iii. A model that could be used in practice

3.142. (1 point) Otherwise similar GLMs have been fit, one using a Gamma Distribution and the other using an Inverse Gaussian Distribution. Based on the following histograms of standardized deviance residuals which model do you prefer and why.



3.143. (1 point) The following loss ratio chart for a proposed rating plan was created by:
1. Sorting the dataset based on the model prediction.
2. Bucketing the data into deciles, such that each decile has approximately the same volume of exposures.
3. Within each bucket, calculate the actual loss ratio (under the current plan) for risks within that bucket.

Discuss the lift of the proposed plan compared to the current plan.



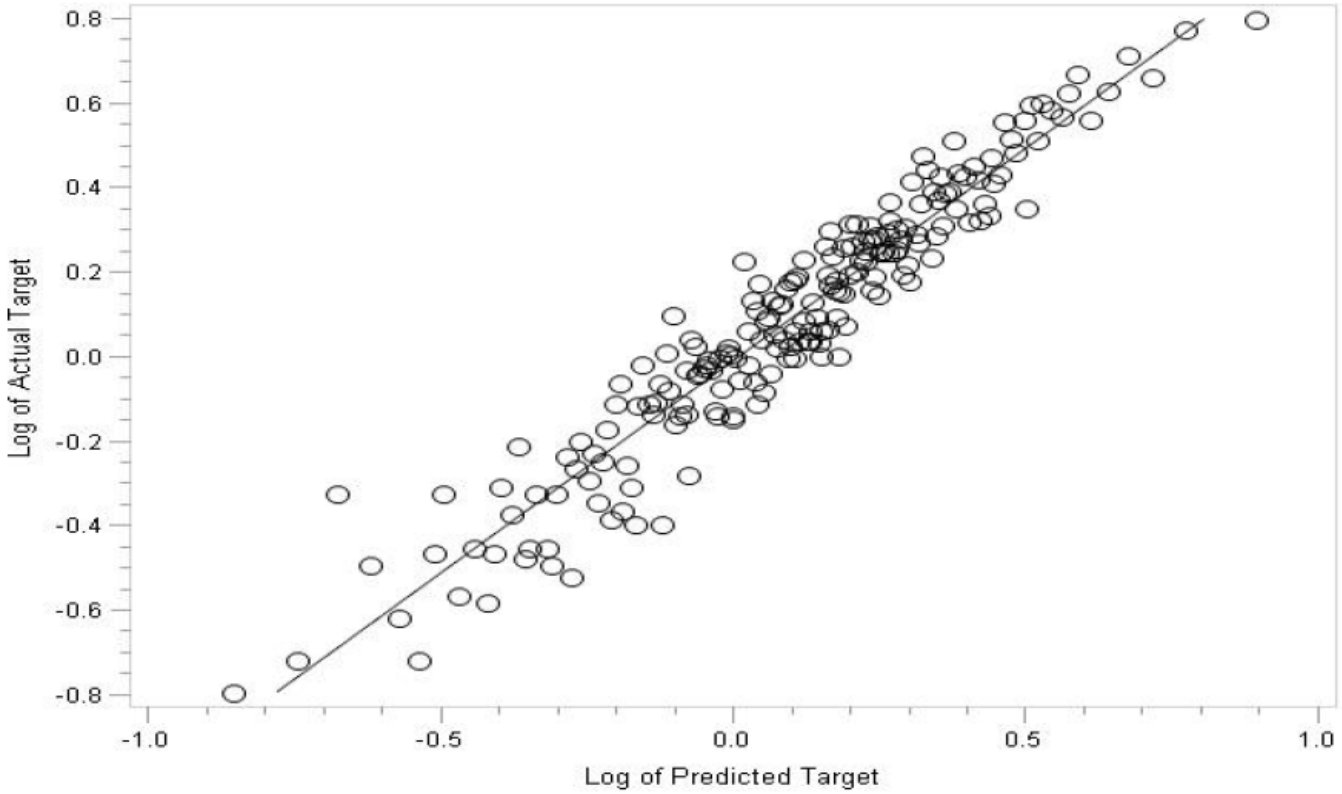
Use the following information for the next two questions:
Three Generalized Linear Models have been fit to the same set of 5000 observations.

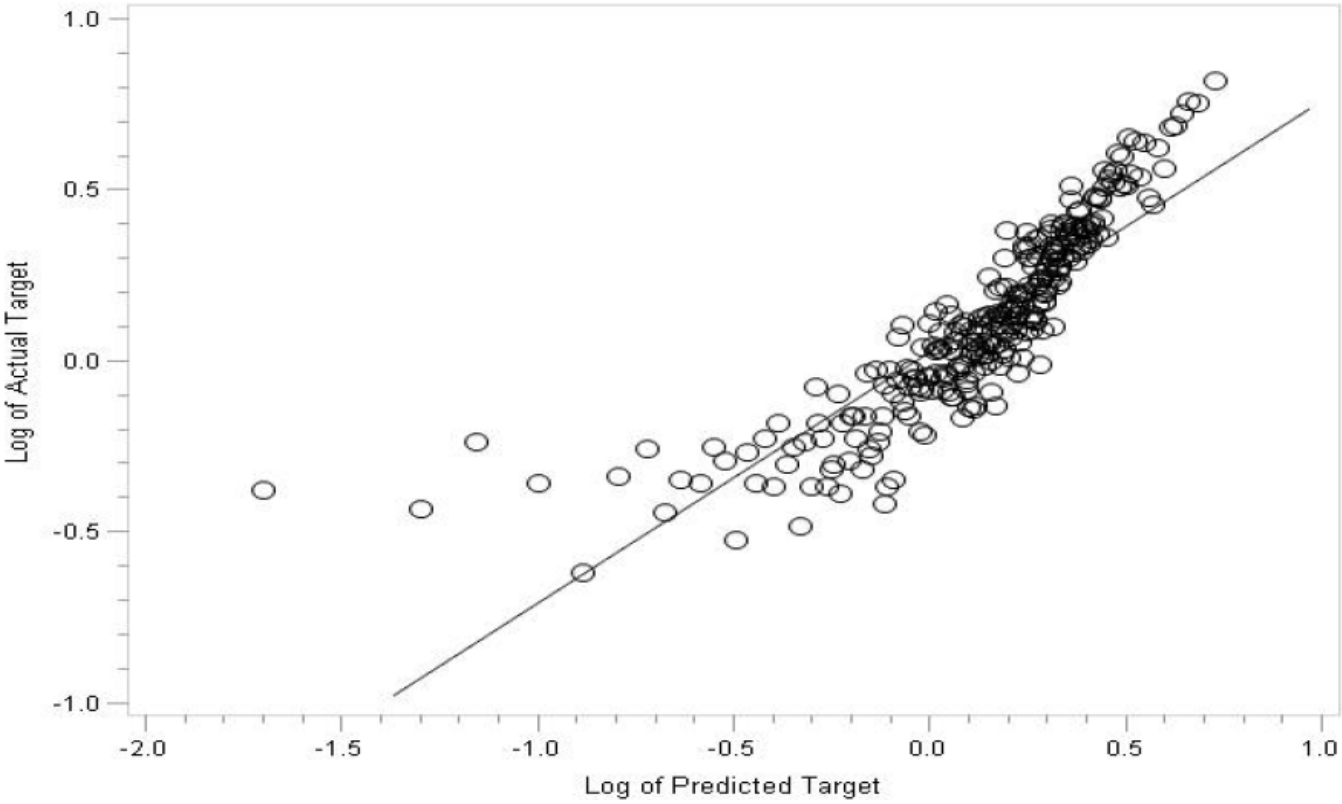
Model	Number of Fitted Parameters	LogLikelihood
A	5	-9844.16
B	10	-9822.48
C	15	-9815.70

3.144. (1 point) Which model has the best AIC (Akaike Information Criterion)?

3.145. (1 point) Which model has the best BIC (Bayesian Information Criterion)?

3.146. (1 point) Below are plots of Actual vs. Predicted for two different GLMs.





Which model do you prefer and why.

3.147. (4 points) A GLM has been used to develop an insurance rating plan.
The results are given below:

Risk	Exposures	Model Predicted Pure Premium	Actual Pure Premium
1	3	7000	6000
2	7	1000	4000
3	8	4000	2000
4	11	5000	8000
5	12	3000	1000
6	16	6000	8000
7	19	8000	6000
8	24	2000	4000

Plot the Lorenz curve for this rating plan.
Label each axis and the coordinates of each point on the curve.

3.148. (2 points) You are given a GLM of collision claim size with the following potential explanatory variables only:
• Vehicle price, which is a continuous variable modeled with a second order polynomial
• Vehicle Age which is a categorical variable with 8 levels
• Average driver age, which is a continuous variable modeled with a first order polynomial
• Number of drivers, which is a categorical variable with three levels
• Gender, which is a categorical variable with two levels
• There is only one interaction in the model, which is between gender and average driver age.
Determine the number of parameters in this model.

3.149. (1.5 points) Discuss how to construct a double lift chart.

3.150. (1.5 points)
Generalized Linear Models have been fit both with and without a certain predictor variable.

Model	With	Without
Deviance	8,901.4414	8,905.6226
Degrees of Freedom	18,169	18,175
Scale Parameter	0.4327	0.4523

The null hypothesis is to use the simpler model.
The alternative hypothesis is to use the more complicated model.
Calculate the F-test statistic.
Discuss how you would perform the test.

3.151. (2 points) You are given the following GLM output:

Response variable		Pure Premium
Response distribution		Gamma
Link		log
Estimated alpha		2.2
Parameter	df	$\hat{\beta}$
Intercept	1	5.07
Risk Group	2	
Group 1	0	0.00
Group 2	1	0.21
Group 3	1	0.48
Vehicle Symbol	1	
Symbol 1	1	-0.36
Symbol 2	0	0.00

Calculate the variance of the pure premium for an insured in Risk Group 3 with Vehicle Symbol 1.

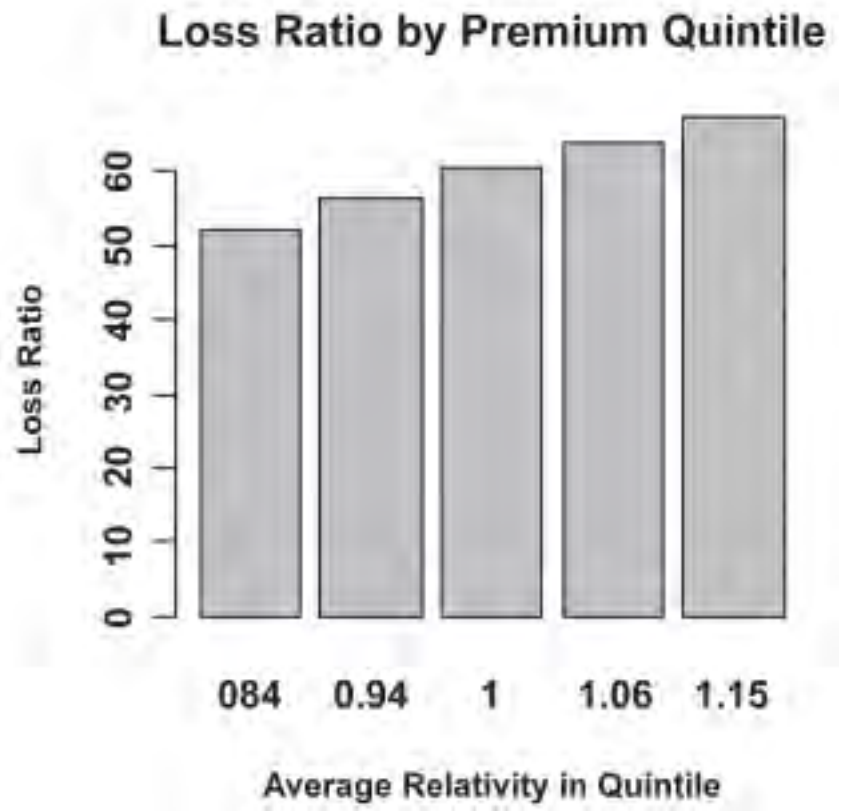
3.152. (1 point) Two GLMs with somewhat different sets of variables have been fit to the same data. Model 1 has a Gini index of 0.16, while Model 2 has a Gini index of 0.12.
Briefly discuss which rating plan has better lift.

3.153. (1 point) Discuss how to construct a loss ratio chart.

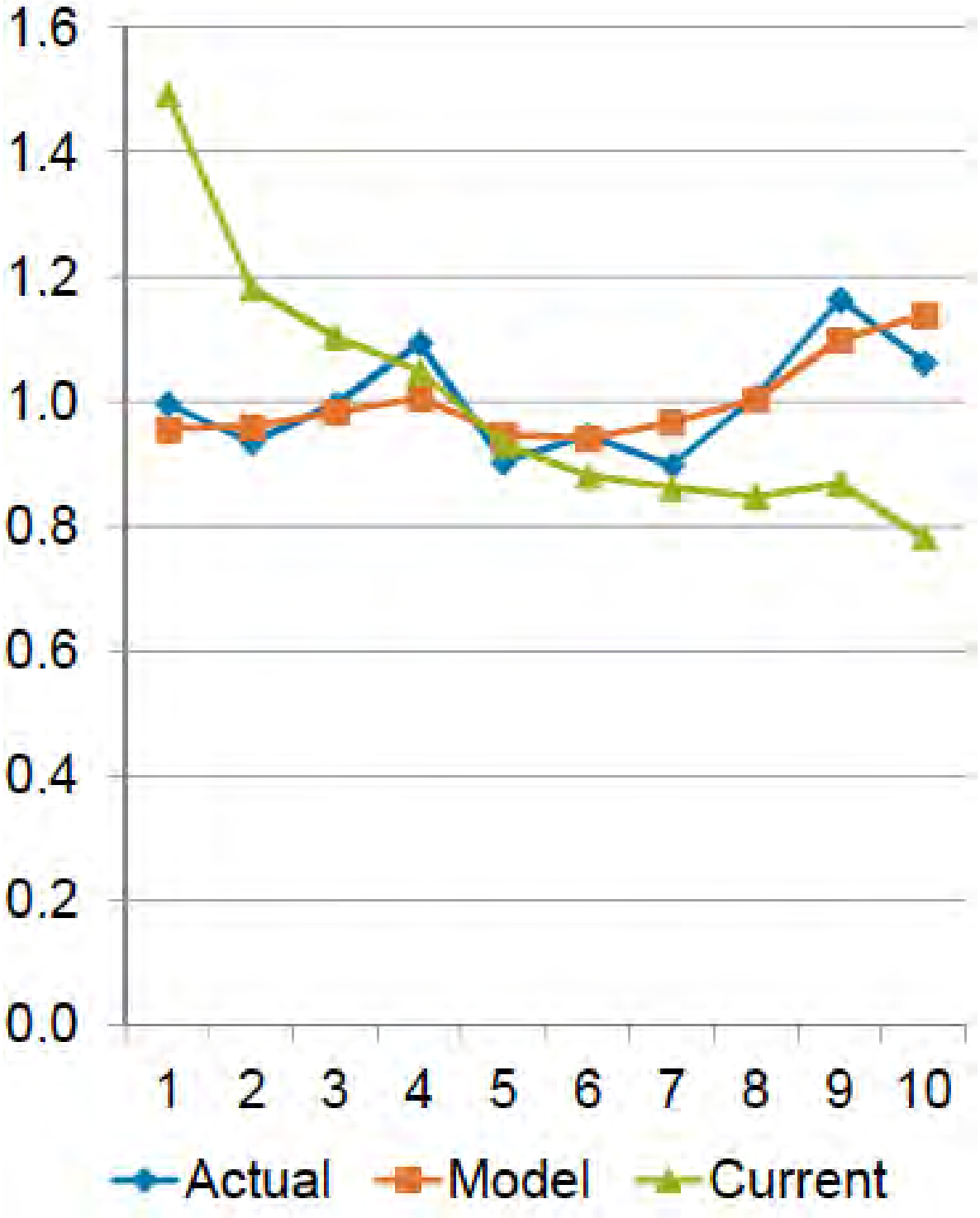
3.154. (1 point) An actuary fits a GLM to a large amount of data on pure premiums for private passenger automobile insurance. The model includes driver age.
The actuary wants to test adding a new variable, number of years claims-free:
0, 1, 2, 3, 4 or more.
The new variable will only be used for drivers at least 25 years old.
The actuary fits an otherwise similar model that includes number of years claims-free to the same data. The effect of driver age in the second model is significantly less than in the first model.
Briefly discuss why this may have occurred.

3.155. (1 point) The following loss ratio chart for a proposed rating plan was created by:
1. Sorting the dataset based on the model prediction.
2. Bucketing the data into quintiles, such that each quintile has approximately the same volume of exposures.
3. Within each bucket, calculate the actual loss ratio (under the current plan) for risks within that bucket.

Discuss the lift of the proposed plan compared to the current plan.

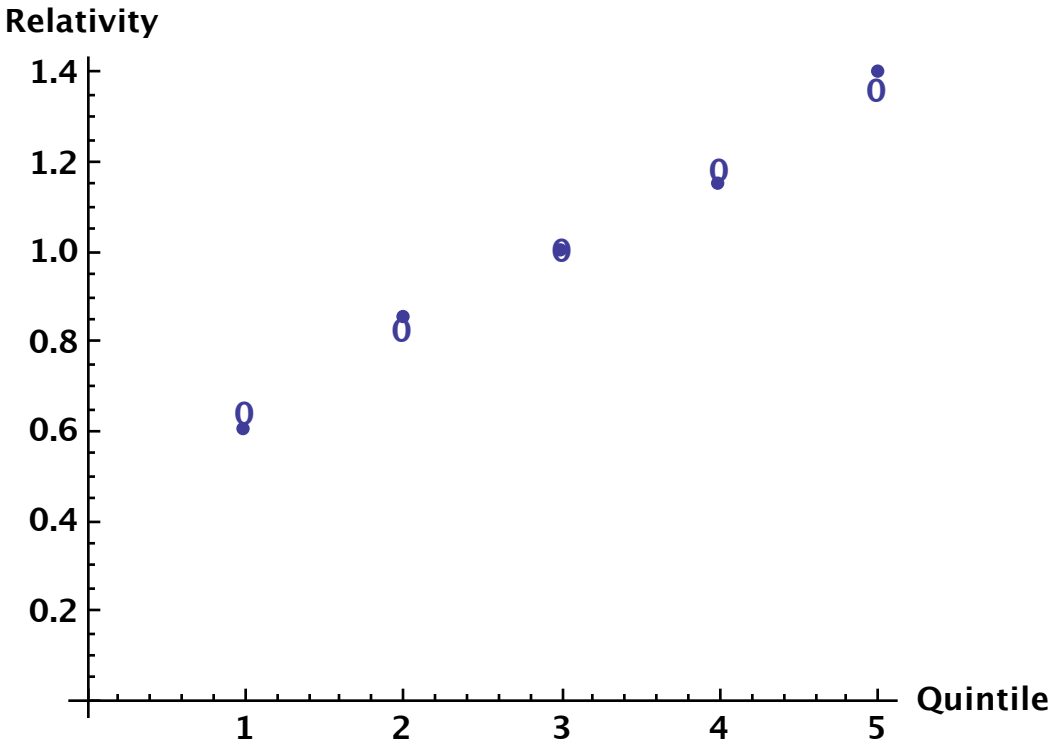


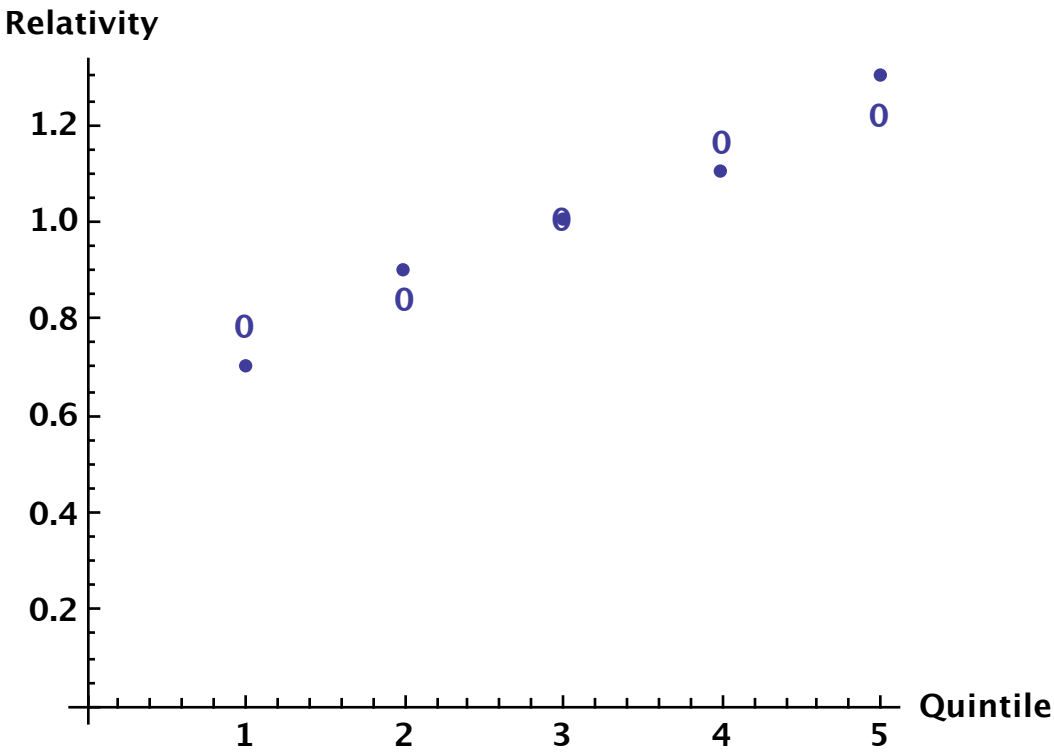
3.156. (1 point) You are given the following double lift chart:



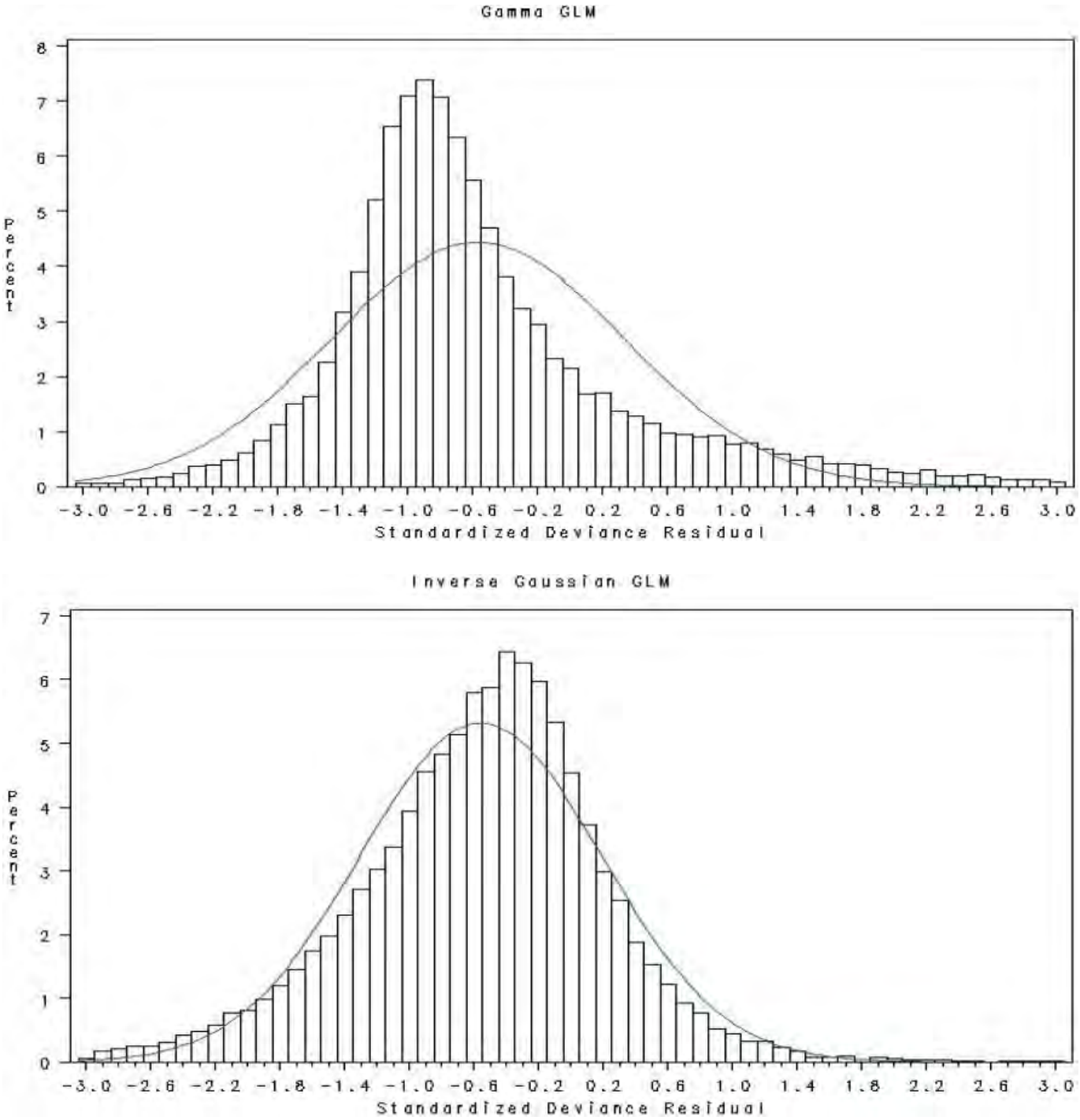
Briefly discuss what conclusion you draw and and why.

3.157. (1 point) Below are shown two simple quintile plots, the first for Plan A and the second for Plan B. In each case, the model plan predictions are shown by dots and the actual by o.
Which plan is preferable and why?





3.158. (1 point) Otherwise similar GLMs have been fit, one using a Gamma Distribution and the other using an Inverse Gaussian Distribution. Based on the following histograms of standardized deviance residuals which model do you prefer and why.



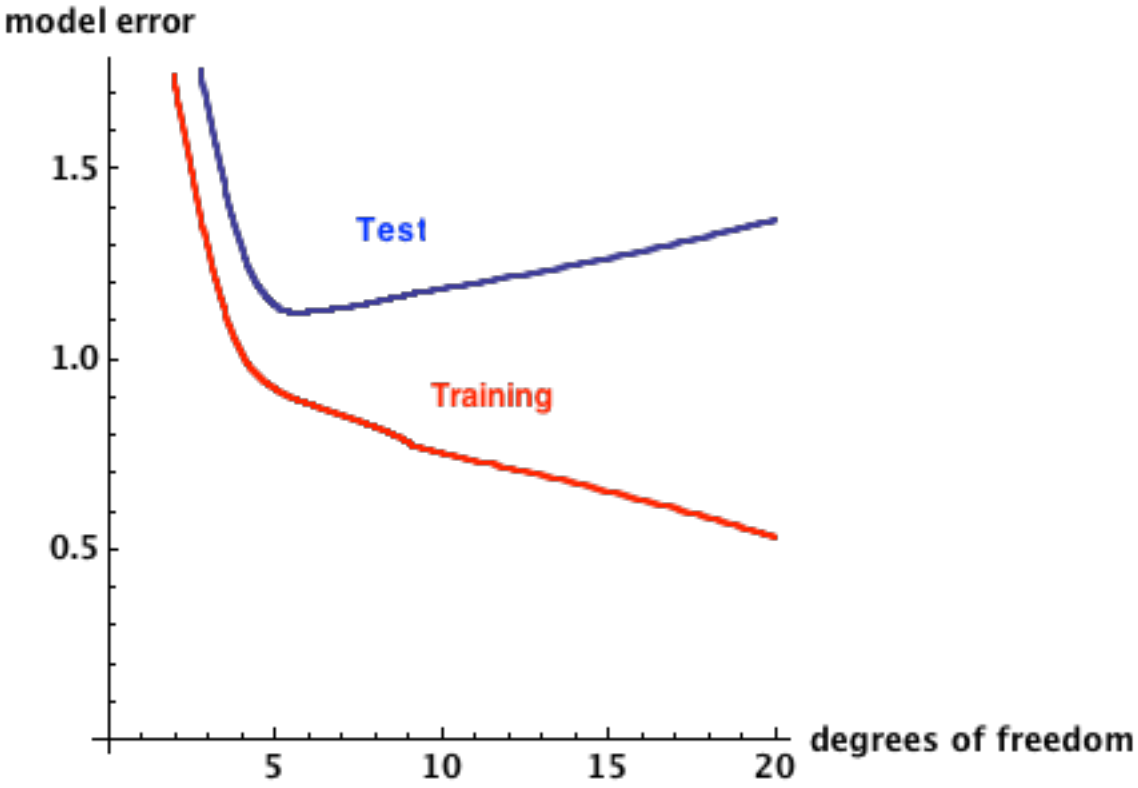
3.159. (3 points) A logistic model was built to predict the probability of a claim being fraudulent. Consider the predicted probabilities for the 15 claims below to be a representative sample of the total model.

Claim Number	Actual Fraud Indicator	Predicted Probability of Fraud
1	N	37%
2	N	46%
3	N	23%
4	N	13%
5	Y	89%
6	N	5%
7	Y	21%
8	N	74%
9	Y	75%
10	Y	69%
11	N	57%
12	Y	54%
13	N	53%
14	N	83%
15	N	49%

a. (1.5 point) Construct confusion matrices for discrimination thresholds of 0.30 and 0.60.

b. (1.5 points) Plot the Receiver Operating Characteristic (ROC) curve with the discrimination thresholds of 0.30 and 0.60.
Label each axis and the coordinates and discrimination threshold of each point on the curve.

3.160. (0.75 points) An actuary has split data into training and test groups for a model. The chart below shows the relationship between model performance and model complexity. Model performance is represented by model error and model complexity is represented by degrees of freedom.



Briefly discuss the optimal balance of complexity and performance.

3.161. (9, 11/03, Q.25) (2 points)
a. (1 point) Explain why one-way analysis of risk classification relativities can produce indicated relativities that are inaccurate and inconsistent with the data.
b. (1 point) Describe an approach to calculating risk classification relativities that would reduce the error produced by a one-way analysis.

3.162. (9, 11/06, Q.5) (4 points)
a. (3 points) Compare the random component, the systematic component, and the link functions of a linear model to those of a generalized linear model.
b. (1 point) Describe two reasons why the assumptions underlying linear models are difficult to guarantee in application.

3.163. (9, 11/07, Q.4a) (1 point) There are a variety of methods available to a ratemaking actuary when determining classification rates.
Compare the Generalized Linear Model to the Classical Linear Model with respect to the following:

i. The distribution of the response variable.
ii. The relationship between the mean and variance of the response variable.

3.164. (9, 11/08, Q.3) (2 points) When using a Generalized Linear Model one of the concerns of which the practitioner must be aware is the presence of aliasing within the model.
a. (1 point) Discuss the two types of aliasing and provide an example of how each can arise in a model.
b. (1 point) An actuary is using a Generalized Linear Model to determine possible interactions between pure premiums. While reviewing the model, the actuary observes the following pure premiums for liability coverage:

	Liability Pure Premium		
	Vehicle Size		
Territory	Small	Medium	Large
North	100	150	250
South	80	110	290
East	90	170	200
West	180	260	540

Assuming equal exposure distribution across all combinations of territory and vehicle size, demonstrate how aliasing can be used to exclude a level from either the territory or the vehicle size variable.

3.165. (2 points) Use the information in the previous question, 9, 11/08, Q.3.
Take North and Medium as the base levels.
Specify the following structural components of a generalized linear model:
Definition of Variables, Design matrix, Vector of responses, Vector of model parameters.

3.166. (9, 11/09, Q.3) *(3 points) Consider a simple private passenger auto classification system that has two rating variables: territory (urban or rural) and gender (male or female).*
The observed average claim severities are:

Gender	*Urban*	*Rural*
Male	*$400*	*$250*
Female	*$200*	*$100*

Y, the response variable, is the average claim severity. Male (x1), Female (x2), Urban (x3) and Rural (x4) are the 4 covariates. A uniquely defined model is:

$Y = \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + e.$

a. (2 points) Using the classical linear model, derive the equations to solve for the parameters β_1, β_2 and β_3 using the sum of squared errors. (Do NOT solve the equations.)

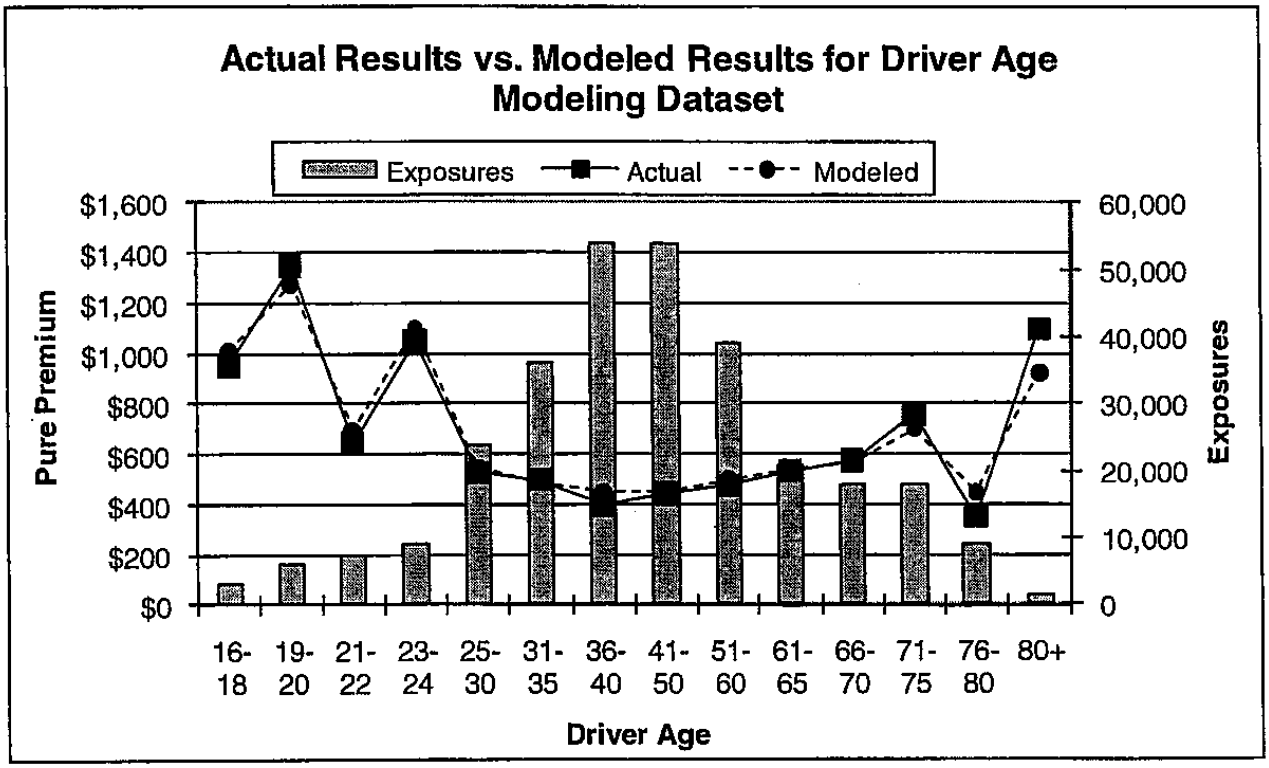
b. (1 point) Briefly describe two underlying assumptions of the classical linear model. Explain why the model may not be able to guarantee these assumptions.

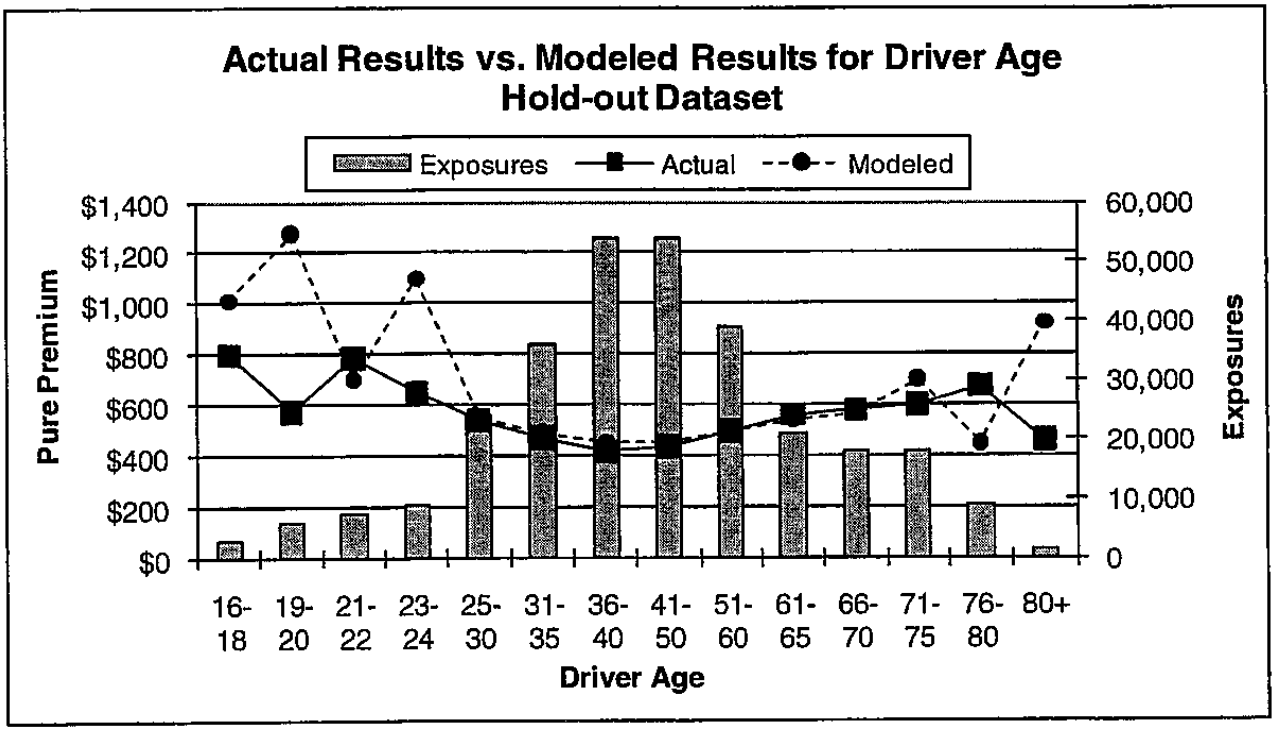
3.167. *(9 points) Use the information in the previous question, 9, 11/09, Q.3.*
As per the exam question, use the following variables: Male (X_1), Female (X_2), Urban (X_3).

a. (2 points) Specify the following structural components of a generalized linear model: Design matrix, Vector of responses, Vector of model parameters.
b. (2 points) Determine the equations that would need to be solved in order to fit the model. Assume a Gamma Distribution and the identity link function.
Assume equal exposures for each cell.
c. (2 points) Determine the equations that would need to be solved in order to fit the model. Assume a Gamma Distribution and the inverse link function.
Assume equal exposures for each cell.
d. (3 points) Determine the equations that would need to be solved in order to fit the model. Assume a Inverse Gaussian Distribution and the squared inverse link function.
Assume equal exposures for each cell.

For the Inverse Gaussian : $f(x) = \sqrt{\frac{\theta}{2\pi}} \frac{\exp\left[-\frac{\theta\left(\frac{x}{\mu} - 1\right)^2}{2x}\right]}{x^{1.5}}$, *mean* $= \mu$, *variance* $= \mu^3 / \theta$.

3.168. (5, 5/10, Q.36) (1 point)
Company XYZ applied generalized linear modeling to its personal auto data. Graphs of the actual and modeled pure premiums by the driver groupings were produced by the analysis. The first graph is a plot of the values using the modeling dataset. The second graph is a plot of the values using a hold-out dataset. The modeling dataset and the hold-out dataset have the same number of exposures. Explain whether or not the model appears to be appropriate.





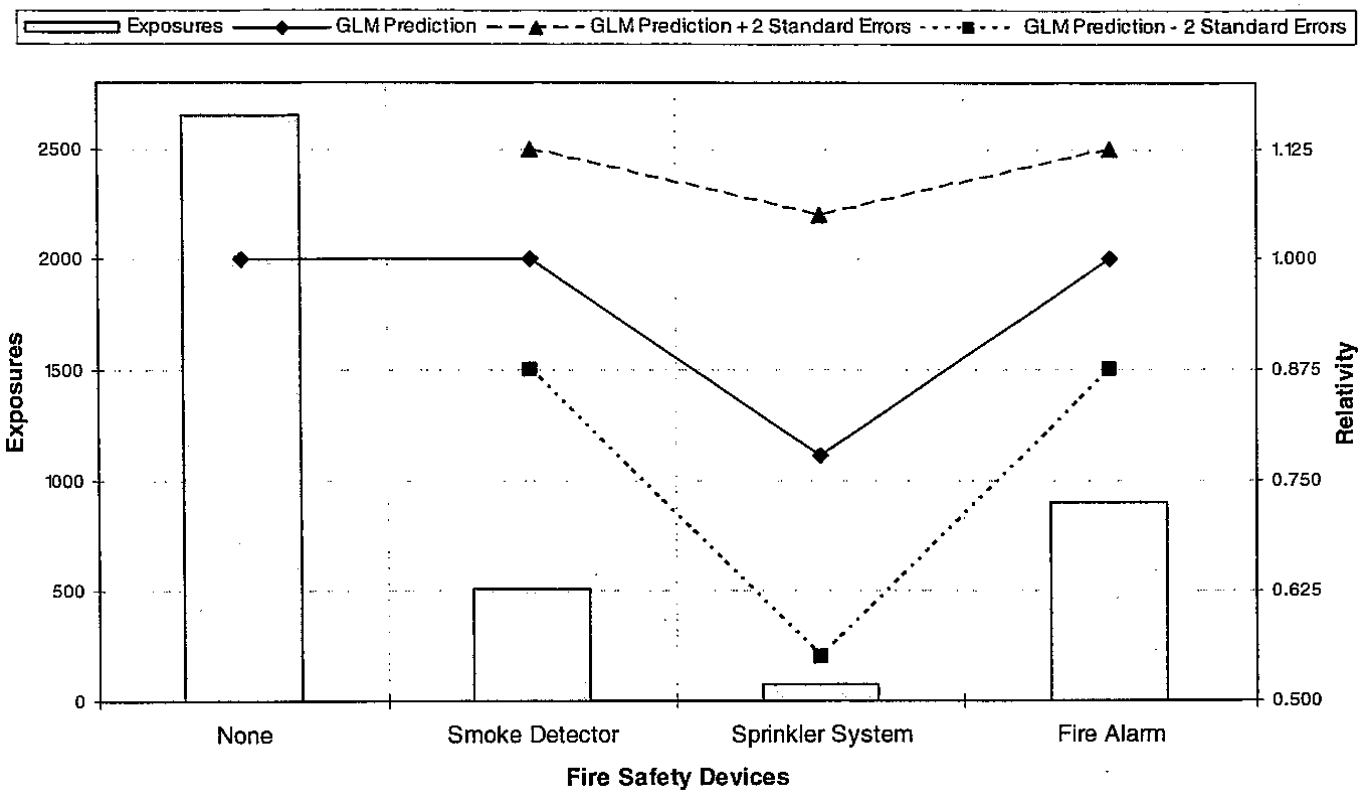
3.169. (9, 11/10, Q.3) (3.5 points)
The following chart represents claim frequencies for a commercial auto book of business:

Claim Frequencies (1,000 Vehicle-Years)

	Private Passenger	Light Truck	Medium Truck
Territory A	10	12	20
Territory B	5	10	18

a. (2 points) Complete the first step in solving a generalized linear model by specifying the design matrix and vector of beta parameters.
b. (0.5 point) For each of the Poisson and gamma error structures, describe the relationship between the variance and the expected value and how these relationships differ.
c. (1 point) Once the link function and error structure have been selected, describe the process to determine the final beta parameters.

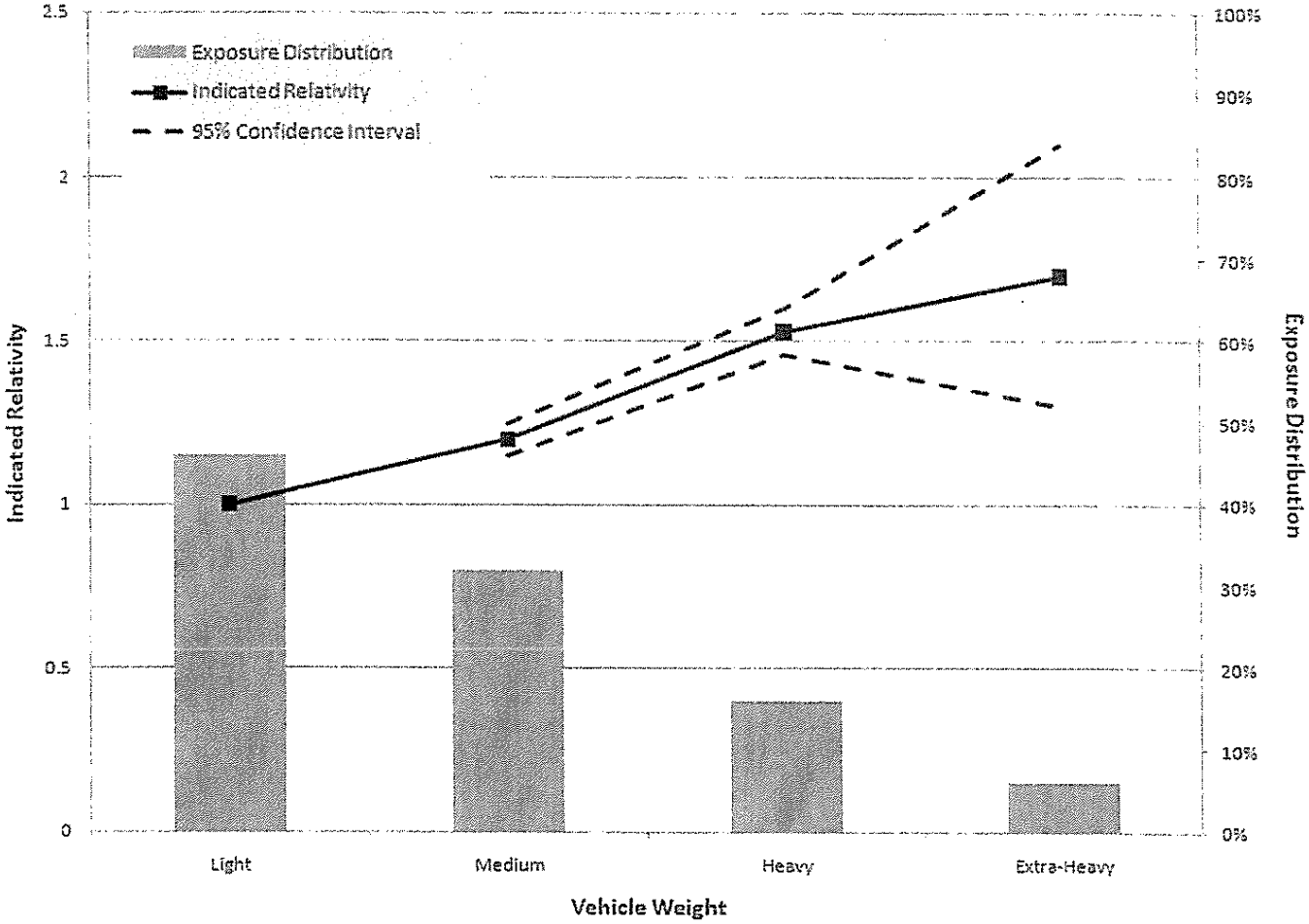
3.170. (5, 5/11, Q.13) (1 point)
A company applied generalized linear modeling to its homeowners data. A graph of indicated relativities and their standard errors for a fire safety device rating variable is shown below. Evaluate the effectiveness of the variable in the model.



Note that the legend reads: Exposures, GLM Prediction,
GLM Prediction + 2 Standard Deviations, GLM Prediction - 2 Standard Deviations.

3.171. (8, 11/11, Q.3) *(1.5 points) An actuary is considering performing a one-way analysis to provide pricing guidance for an insurance company's personal auto book of business.*
a. (0.5 point) Briefly describe two shortcomings associated with one-way analyses.
b. (1 point) Provide an example of how each shortcoming in part a. above may arise.

3.172. (5, 5/12, Q.11) (1.5 points) An insurer uses several rating variables, including vehicle weight, to determine premium charges for commercial automobiles. Your manager has requested a review of the vehicle weight rating relativities. The following diagnostic chart displays the results for vehicle weight from a generalized linear model. (Light, Medium, Heavy, and Extra Heavy.)



Company management plans to expand its commercial auto marketshare with an emphasis on writing more businesses that operate with extra-heavy weight vehicles. Management wants to charge the same rates for both heavy and extra-heavy weight vehicles.
Based on the model results, provide your recommendation to management and explain the considerations supporting your position. Include a discussion of any potential risks associated with it.

3.173. (8, 11/12, Q.2) (2.25 points) A private passenger auto insurance company orders a report whenever it writes a policy, showing what other insurance the policyholder has purchased. The following table shows claim frequencies (per 100 earned car-years) for bodily injury liability coverage, split by whether the policyholder has a homeowners policy and whether the policyholder had a prior auto policy:

	Homeowners Policy	
Prior Auto Policy	Yes	No
Yes	3	5
No	8	12

The table does not include the experience of policyholders with missing data.
a. (1.25 points) Specify the following structural components of a generalized linear model that estimates frequencies for this book of business.
i. Error distribution
ii. Link function
iii. Vector of responses
iv. Vector of model parameters
v. Design matrix
b. (1 point) Describe how the missing data may cause problems for the company in developing the model, and suggest a solution.

3.174. (8, 11/12, Q.4) (1.75 points) An actuary has historical information relating to customer retention. A logistic model was used to estimate the probability of renewal for a given customer. The two variables determined to be significant were the size of rate change and number of phone calls the insured made to the company. The parameter estimates were determined to be as follows:

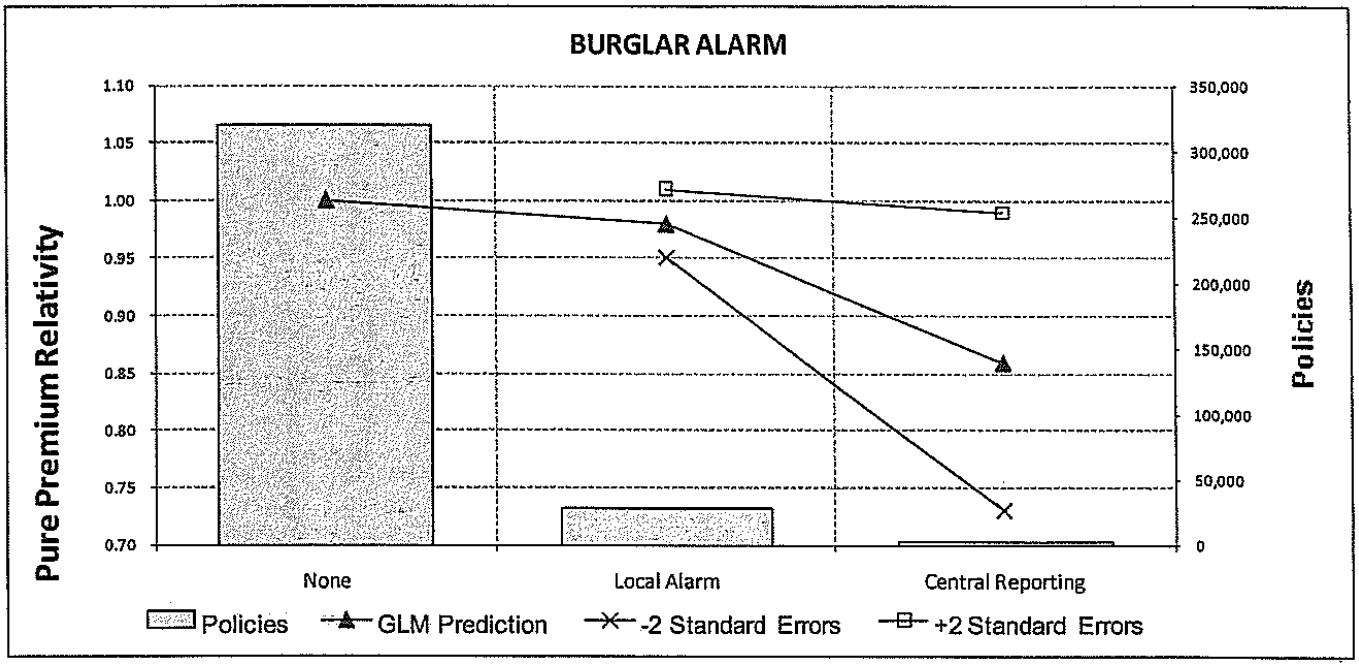
Rate Change	Parameter Estimate
Decrease to 3.9% increase	0.3323
4.0% to 6.9% increase	0
Increase of 7.0% or more	-0.4172

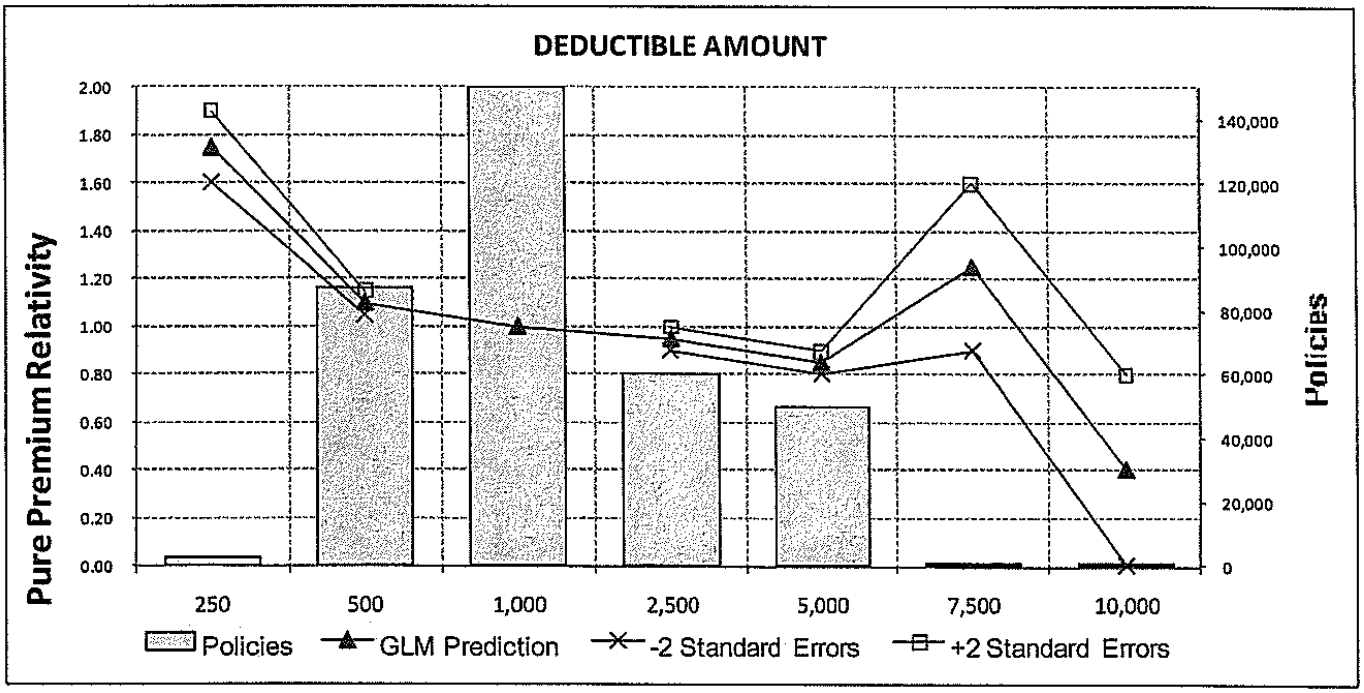
Number of Phone Calls in Past Year	Parameter Estimate
0	0
1	-0.2128
2+	-0.4239

Intercept Term	1.793

a. (0.75 point) Calculate the renewal probability for a customer who has a 7% rate increase and called the company twice in the past year.
b. (1 point) The company needs policyholder retention to be above 78% to maintain growth and expense ratio goals. A possible strategy is to add the number of phone calls to the classification plan and use the model results to determine the rate increase.
Construct an argument either in favor of or against the strategy above, describing two reasons for that position.

3.175. (5, 5/13, Q.12) (3 points) An insurer is planning to revise burglar alarm and deductible rating plan factors for its Homeowners program. Given the following generalized linear model output:





Question is continued on the next page.

(question continued)

Burglar Alarm	GLM Prediction	-2 Standard Errors	+2 Standard Errors	Policies
None	1.00			320,000
Local Alarm	0.98	0.950	1.010	27,500
Central Reporting	0.86	0.730	0.990	2,500

Deductible	GLM Prediction	-2 Standard Errors	+2 Standard Errors	Policies
$250	1.75	1.60	1.90	2,700
$500	1.10	1.05	1.15	87,000
$1,000	1.00			150,000
$2,500	0.95	0.90	1.00	60,000
$5,000	0.85	0.80	0.90	50,100
$7,500	1.25	0.90	1.60	150
$10,000	0.40	0.00	0.80	50

Propose revised burglar alarm and deductible rating plan factors.
Document the relevant analysis and rationale to support the proposal.

3.176. (5, 11/13, Q.11) (2.25 points) Given the following information:

	Policies with a $100,000 Limit		Policies with a $250,000 Limit		Policies with a $500,000 Limit	
Size of Loss	Claims	Losses	Claims	Losses	Claims	Losses
$X \leq \$100,000$	100	$8,000,000	35	$1,800,000	35	$1,800,000
$\$100,000 < X \leq \$250,000$			40	$7,400,000	25	$3,900,000
$\$250,000 < X \leq \$500,000$					15	$5,200,000

Limit	Indicated factor (pure premium generalized linear model analysis)
$100,000	1.00
$250,000	0.95
$500,000	1.15

For the $250,000 policy limit:
a. (1.25 points) Calculate the indicated increased limits factor, assuming a basic limit of $100,000.
b. (0.5 point) Explain the difference between the indicated increased limits factor calculated in part a. above and the generalized linear model results.
c. (0.5 point) Select an increased limit factor and briefly explain the rationale for the selection.

3.177. (8, 11/13, Q.2) (3.5 points) An actuary at a private passenger auto insurance company wishes to use a generalized linear model to create an auto frequency model using the data below.

	Number of Claims	
Gender	Territory A	Territory B
Male	700	600
Female	400	420

	Number of Exposures	
Gender	Territory A	Territory B
Male	1,400	1,000
Female	1,000	1,200

The model will include three parameters: β_1, β_2, β_3, where β_1 is the average frequency for males, β_2 is the average frequency for Territory A, and β_3 is an intercept.

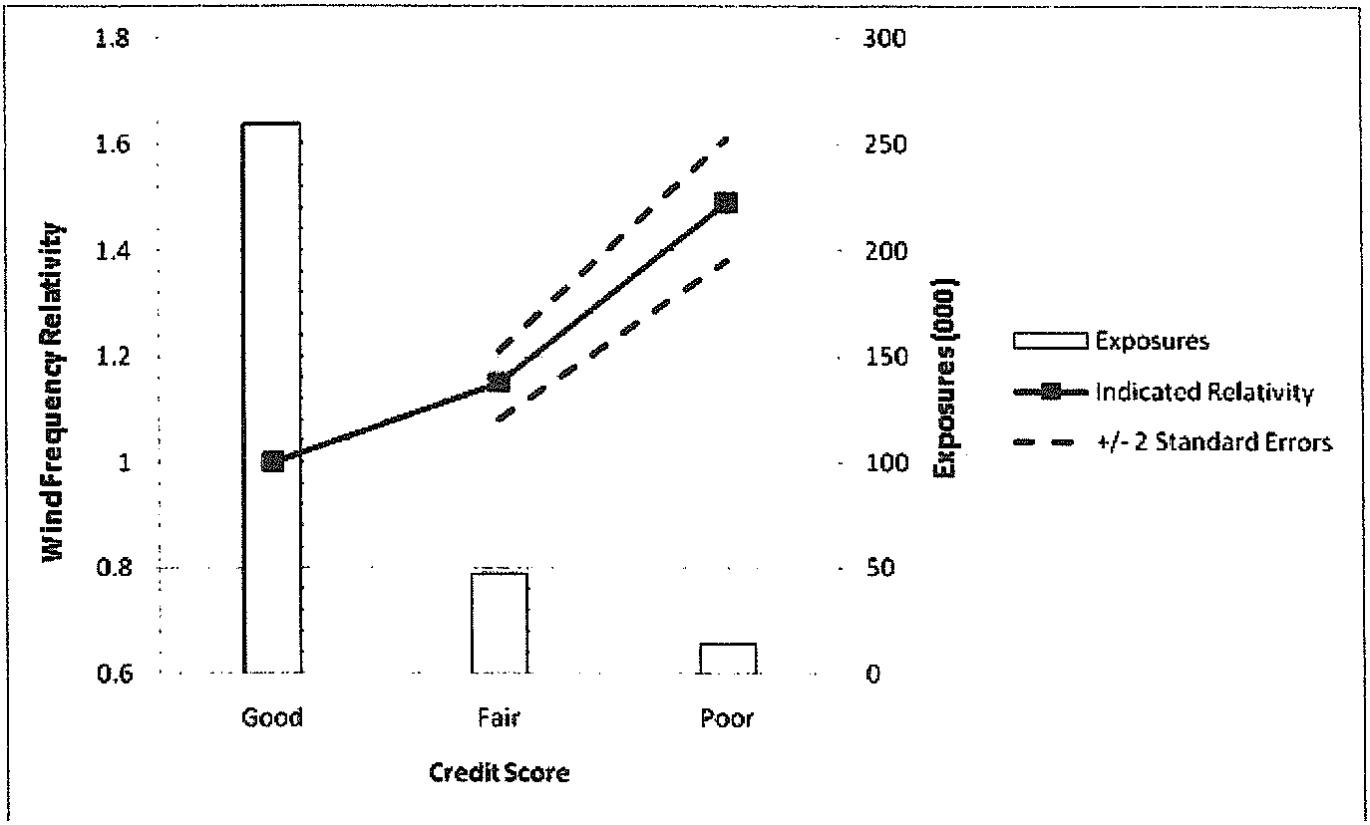
a. (0.5 point) Define the design matrix [X].

b. (0.25 point) Define the vector of responses [Y].

c. (2.25 points) Assuming $\beta_3 = 0.35$, solve a generalized linear model with a normal error structure and identity link function for β_1.

d. (0.5 point) The actuary determines that the analysis results would be improved by assuming a Poisson error structure with a log link function.
Identify two reasons this structure may better suit this data.

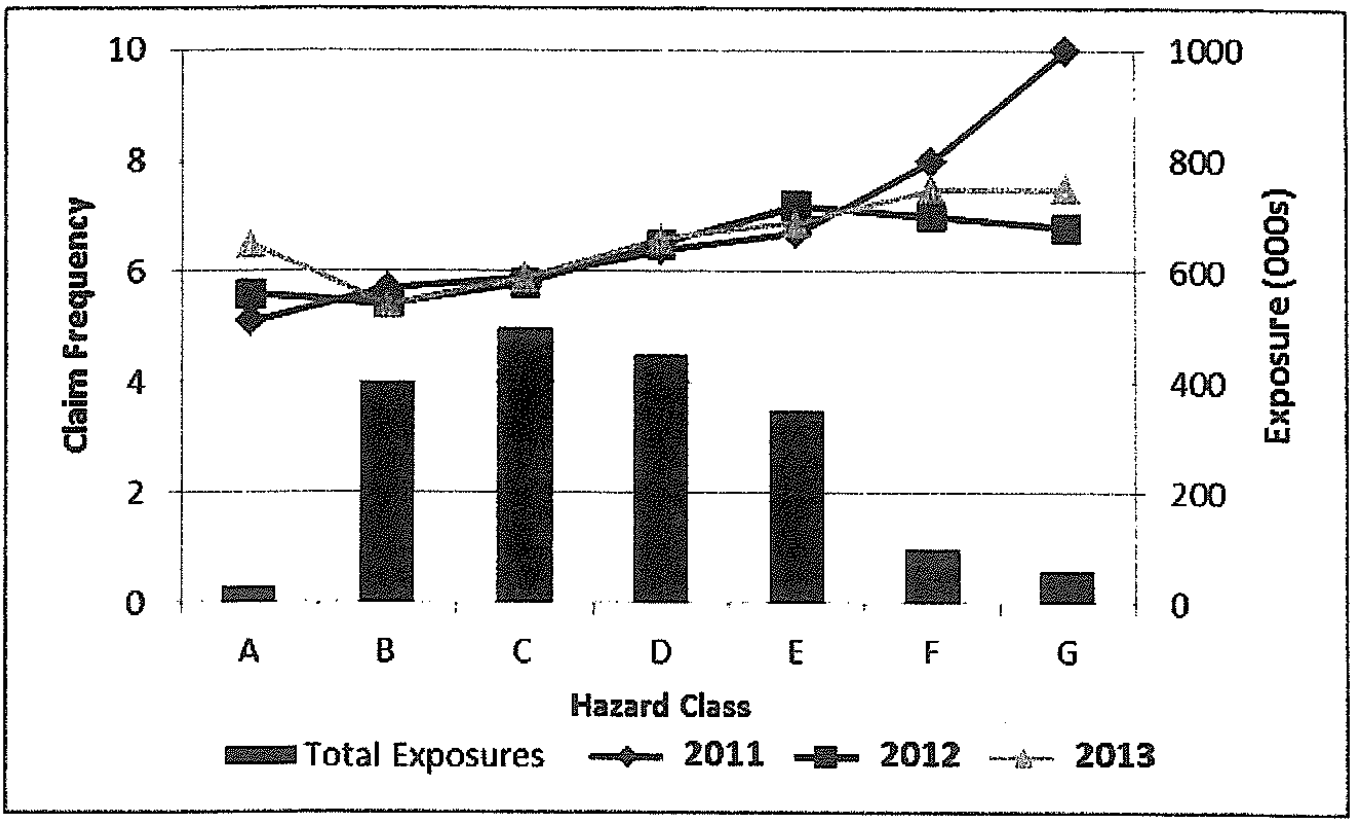
3.178. (5, 5/14, Q.9) (2 points)
An insurer is considering using credit score to further segment its homeowners book of business. The insurer has developed a generalized linear model to evaluate different variables' contribution to expected frequency of wind claims.
The following diagnostic chart displays the results of a countrywide analysis performed on one year of data from a generalized linear model:



(The solid line is the indicated relativity, while the dashed lines are ± 2 standard errors.)

Using the generalized linear model output, as well as other considerations, justify whether the insurer should add credit score to the homeowners rating plan for the wind peril.

3.179. (5, 11/14, Q.10a) (0.75 points)
An actuary performed an analysis of a products liability class plan using a Generalized Linear Model (GLM) for the first time on this book of business. The insureds are categorized by hazard classes A through G. The following graph shows claim frequency and exposure data by hazard class.



Fully evaluate the predictive value of hazard class based on the information provided above.

3.180. (8, 11/14, Q.3) (2 points) The random component of a generalized linear model must come from the exponential family of distributions. The variance of a distribution from the exponential family can be expressed using the following formula: $\text{Var}(Y_i) = \frac{\phi\, V(\mu_i)}{\omega_i}$

a. (0.5 point) Define the parameters ϕ and ω_i in the formula above.

b. (1 point) For each of the data sets below, identify the error distribution that should be used to model the data. Briefly explain why that error distribution is appropriate.

i. Severity

ii, Policy Renewal Retention

c. (0.5 point) For each of the error distributions in part b. above, provide an example of how w_i should be assigned for the type of data being modeled.

3.181. (CAS S Sample Exam 2015, Q.4) (2 points)
An actuary wants to estimate the probability of a home insurance policy having a claim by using a logistic regression model. He has the following pieces of information from 1,000 historical policies:
• Cost of the home, in $000s
• Age of the home, in years
• Whether or not there was a claim on the policy

The actuary is considering a number of different model specifications. Below are the models he is considering along with the calculated deviance of each model:

Model #	Included Variables	Deviance
1	Intercept + Cost	1085.0
2	Intercept + Cost + Age	1084.8
3	Intercept + Cost + (Cost * Age)	1083.0
4	Intercept + Cost + $Cost^2$ + $Cost^3$	1081.9
5	Intercept + Cost + $Cost^2$ + $Cost^3$ + $Cost^4$	1081.6

Determine the optimal model using the Bayesian Information Criterion.

3.182. (2 points) In the previous question, determine the optimal model using instead the Akaike Information Criterion.

3.183. (CAS S, 11/15, Q.32) (2 points)
A GLM is used to model claim size. You are given the following information about the model:
• Claim size follows a Gamma distribution.
• Log is the selected link function.
• The dispersion parameter f is estimated to be 2.
• Model Output:

Variable	$\hat{\beta}$
(Intercept)	2.32
Location - Urban	0.00
Location - Rural	-0.64
Gender - Female	0.00
Gender - Male	0.76

Calculate the variance of the predicted claim size for a rural male.

3.184. (CAS S, 11/15, Q.33) (2 points)
You are given the following output from a GLM to estimate the probability of a claim:
- Distribution selected is Binomial.
- Link selected is Logit.

Parameter	β
Intercept	-1.485
Vehicle Body	
Coupe	-0.881
Roadster	-1.047
Sedan	-1.175
Station wagon	-1.083
Truck	-1.118
Utility	-1.330
Driver's Gender	
Male	-0.025
Area	
B	0.094
C	0.037
D	-0.101

Calculate the estimated probability of a claim for:
- Driver Gender: Female
- Vehicle Body: Sedan
- Area: D

3.185. (CAS S, 11/15, Q.34) (1 point)
You are given the following information for a model of vehicle claim counts by policy:
- The response distribution is Poisson and the model has a log link function.
- The model uses two categorical explanatory variables: Number of Youthful Drivers and Number of Adult Drivers.
- The parameters of the model are given:

Parameter	Degrees of Freedom	$\hat{\beta}$
Intercept	1	-2.663
Number of Youthful Drivers		
0		
1	1	0.132
Number of Adult Drivers		
1		
2	1	-0.031

Calculate the predicted claim count for a policy with one adult driver and one youthful driver.

3.186. (CAS S, 11/15, Q.35) (2 points)
You are given a GLM of liability claim size with the following potential explanatory variables only:
• Vehicle price, which is a continuous variable modeled with a third order polynomial
• Average driver age, which is a continuous variable modeled with a first order polynomial
• Number of drivers, which is a categorical variable with four levels
• Gender, which is a categorical variable with two levels
• There is only one interaction in the model, which is between gender and average driver age.
Determine the maximum number of parameters in this model.

3.187. (CAS S, 11/15, Q.36) (2 points) You are given the following information for two potential logistic models used to predict the occurrence of a claim:
• Model 1: (AIC = 262.68)

Parameter	$\hat{\beta}$
(Intercept)	-3.264
Vehicle Value ($000s)	0.212
Gender-Female	0.000
Gender-Male	0.727

• Model 2: (AIC = 263.39)

Parameter	$\hat{\beta}$
(Intercept)	-2.894
Gender-Female	0.000
Gender-Male	0.727

• AIC is used to select the most appropriate model.
Calculate the probability of a claim for a male policyholder with a vehicle valued $12,000 by using the selected model.

3.188. (CAS S, 11/15, Q.38) (2 points)
You are testing the addition of a new categorical variable into an existing GLM.
You are given the following information:
• The change in model deviance after adding the new variable is -53.
• The change in AIC after adding the new variable is -47.
• The change in BIC after adding the new variable is -32.
• Prior to adding the new variable, the model had 15 parameters.
Calculate the number of observations in the model.

3.189. (8, 11/15, Q.3) (2.5 points) An actuary is considering using a generalized linear model to estimate the expected frequency of a recently introduced insurance product.
Given the following assumptions:
- The expected frequency for a risk is assumed to vary by state and gender.
- A log link function is used.
- A Poisson error structure is used.
- The likelihood function of a Poisson is

$$l(y; \mu) = \sum \ln f(y_i; \mu_i) = \sum \{-\mu_i + y_i \ln[\mu_i] - \ln[y_i!]\}$$

- β_1 is the effect of gender = Male.
- β_2 is the effect of gender = Female.
- β_3 is the effect of State = State A.

	Claim Frequency	
	State A	State B
Male	0.0920	0.0267
Female	0.1500	0.0500

Given that $\beta_3 = 1.149$, determine the expected frequency of a male risk in State A.

3.190. (CAS S, 5/16, Q.29) (2 points) You are given the following information for a fitted GLM:

Response variable		Occurrence of Accidents
Response distribution		Binomial
Link		Logit

Parameter	df		$\hat{\beta}$
Intercept	1	x	
Driver's Age	2		
1	1		0.288
2	1		0.064
3	0		0
Area	2		
A	1		-0.036
B	1		0.053
C	0		0
Vehicle Body	2		
Bus	1		1.136
Other	1		-0.371
Sedan	0		0

The probability of a driver in age group 2, from area C and with vehicle body type Other, having an accident is 0.22.
Calculate the odds ratio of the driver in age group 3, from area C and with vehicle body type Sedan having an accident.

3.191. (CAS S, 5/16, Q.30) (2 points) You are given the following information for a fitted GLM:

Response variable	Occurrence of Accidents
Response distribution	Binomial
Link	Logit

Parameter		df	$\hat{\beta}$	se
Intercept		1	-2.358	0.048
Area		2		
	Suburban	0	0.000	
	Urban	1	0.905	0.062
	Rural	1	-1.129	0.151

Calculate the modeled probability of an Urban driver having an accident.

3.192. (CAS S, 5/16, Q.31) (2 points) You are given the following information for a fitted GLM:

Response variable	Claim size	
Response distribution	Gamma	
Link	Log	
Estimated alpha	1	
Parameter	df	$\hat{\beta}$
Intercept	1	2.100
Zone	4	
1	1	7.678
2	1	4.227
3	1	1.336
4	0	0.000
5	1	1.734
Vehicle Class	6	
Convertible	1	1.200
Coupe	1	1.300
Sedan	0	0.000
Truck	1	1.406
Minivan	1	1.875
Station wagon	1	2.000
Utility	1	2.500
Driver Age	2	
Youth	1	2.000
Middle age	0	0.000
Old	1	1.800

Calculate the predicted claim size for an observation from Zone 3,
with Vehicle Class Truck and Driver Age Old.

3.193. (CAS S, 5/16, Q.32) (2 points) You are given the following information for a fitted GLM:

Response variable	Claim size	
Response distribution	Gamma	
Link	Log	
Estimated f	1	
Parameter	df	$\hat{\beta}$
Intercept	1	2.100
Zone	4	
1	1	7.678
2	1	4.227
3	1	1.336
4	0	0.000
5	1	1.734
Vehicle Class	6	
Convertible	1	1.200
Coupe.	1	1.300
Sedan	0	0.000
Truck	1	1.406
Minivan	1	1.875
Station wagon	1	2.000
Utility	1	2.500
Driver Age	2	
Youth	1	2.000
Middle age	0	0.000
Old	1	1.800

Calculate the variance of a claim size for an observation from Zone 4,
with Vehicle Class Sedan and Driver Age Middle age

3.194. (CAS S, 5/16, Q.33) (2 points)
You are given the following information for a GLM of customer retention:

Response variable	Retention	
Response distribution	Binomial	
Link	Logit	
Parameter	df	$\hat{\beta}$
Intercept	1	1.530
Number of Drivers	1	
1	0	0.000
>1	1	0.735
Last Rate Change	2	
<0%	0	0.000
0%-10%	1	-0.031
>10%	1	·0.372

Calculate the probability of retention for a policy with 3 drivers and a prior rate change of 5%.

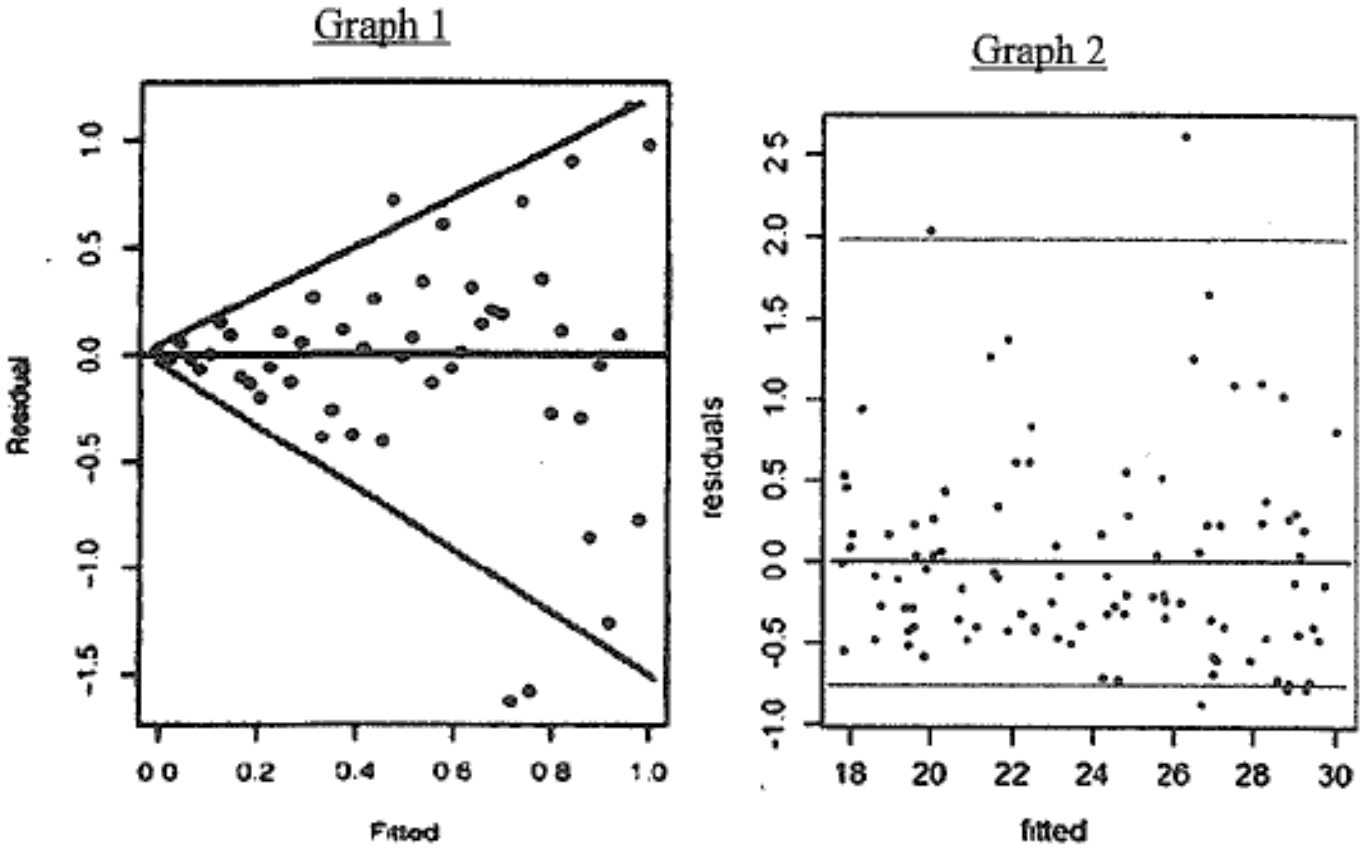
3.195. (CAS S, 5/16, Q.35) (2 points) You are given the following information about three candidates for a Poisson frequency GLM on a group of condominium policies:

Model	Variables in the Model	DF	Log Likelihood	AIC	BIC
1	Risk Class	5	-47,704	95,418	95,473.61182
2	Risk Class + Region		-47,495		
3	Risk Class + Region + Claim Indicator	10	-47,365	94,750	

- Insureds are from one of five Risk Class: A, B, C, D, E
- Condominium policies are located in several regions
- Claim Indicator is either Yes or No
- All models are built on the same data

Calculate the absolute difference between the AIC and the BIC for Model 2.

3.196. (CAS S, 5/16, Q.36) (2 points) You are given the following two graphs comparing the fitted values to the residuals of two different linear models:



Determine which of the following statements are true.

I. Graph 1 indicates the data is homoscedastic
II. . Graph 1 indicates the data is heteroskedastic (a lack of homoscedasticity)
III. Graph 2 indicates the data is non-normal

3.197. (CAS S, 5/16, Q.37) (2 points)

Determine which of the following GLM selection considerations is true.

A. The model with the largest AIC is always the best model in model selection process.
B. The model with the largest BIC is always the best model in model selection process.
C. The model with the largest deviance is always the best model in model selection process.
D. Other things equal, when the number of observations > 1000, AIC penalizes more for the number of parameters used in the model than BIC.
E. Other things equal, when number of observations > 1000, BIC penalizes more for the number of parameters used in the model than AIC.

3.198. (CAS S, 5/16, Q.38) (2 points) You are testing the addition of a new categorical variable into an existing GLM, and are given the following information:
- A is the change in AIC and B is the change in BIC after adding the new variable.
- B > A + 25
- There are 1500 observations in the model.

Calculate the minimum possible number of levels in the new categorical variable.

3.199. (CAS S, 5/16, Q.41) (1 point) A Poisson regression model with log link is used to estimate the number of diabetes deaths. The parameter estimates for the model are:

Response variable	Number of Diabetes Deaths
Response distribution	Poisson
Link	Log

Parameter	df	$\hat{\beta}$	p-value
Intercept	1	-15.000	<0.0001
Gender: Female	1	-1.200	<0.0001
Gender: Male	1	0.000	
Age	1	0.150	<0.0001
Age^2	1	0.004	<0.0001
Age ×Gender: Female	1	0.012	<0.0001
Age ×Gender: Male	0	0.000	

Calculate the expected number of deaths for a population of 100,000 females age 25.

3.200. (CAS 8, 11/16, Q.4) (3 points) An actuary is conducting a generalized linear model (GLM) analysis on historical personal automobile data in order to develop a rating plan.

a. (1.5 points)
Argue against the following factors being included as predictors in the actuary's GLM analysis:
i. Limit of liability.
ii. Number of coverage changes during the current policy period.
iii. ZIP code of the garaging location of the automobile.

b. (1 point) The actuary is modeling pure premium with a log-link function and a Tweedie error distribution ($1 < p < 2$). Provide two arguments against the inclusion of deductible as a predictor in the actuary's GLM analysis.

c. (0.5 point) Other than including deductible as a predictor in the GLM, describe how to determine deductible relativities and how such relativities can be incorporated in a GLM.

3.201. (CAS 8, 11/16, Q.5) (2.25 points)
A GLM has been used to develop an insurance rating plan.
The results are given below:

Risk	Model Predicted Loss	Actual Loss
1	2,000	2,050
2	500	220
3	1,500	1,480
4	800	850
5	200	400

a. (1.75 points) Plot the Lorenz curve for this rating plan.
Label each axis and the coordinates of each point on the curve.
b. (0.5 point) Briefly describe how the Gini index is calculated and what the Gini index measures when applied to an insurance rating program. Do not calculate the Gini index.

3.202. (CAS 8, 11/16, Q.6) (2.5 points) An actuary has constructed a three-variable Tweedie GLM with a log-link function to estimate loss ratios for commercial property new business.
The actuary wants to create a second model for renewal business that will include all of the variables from the new business model, plus a variable for the prior year claim count.
The actuary requires that the coefficients of the variables: Average Building Age, log(Manual Premium), and Location Count, are consistent between the new and renewal models.
The fitted new business model parameters are as follows:

Variable	Name	Estimate
	intercept	0.910
Average Building Age (Years)	age	0.013
log(Manual Premium)	logprem	-0.187
Location Count	loccnt	0.062

a. (0.75 point) Calculate the modeled loss ratio for a new business policy with a manual premium of $25,000, an average building age of four years, and having eight locations.
b. (0.75 point) Briefly describe how to produce the renewal business model, and specify the resulting equation for the renewal business modeled loss ratio.
c. (1 point) Identify and briefly describe two techniques that the actuary can use to assess the stability of the new variable in the renewal business model.

3.203. (CAS 8, 11/16, Q.7) (1.5 points) A company is considering modifying its rating plan to include factors by age group. Below are statistics for the base model and for the new model.

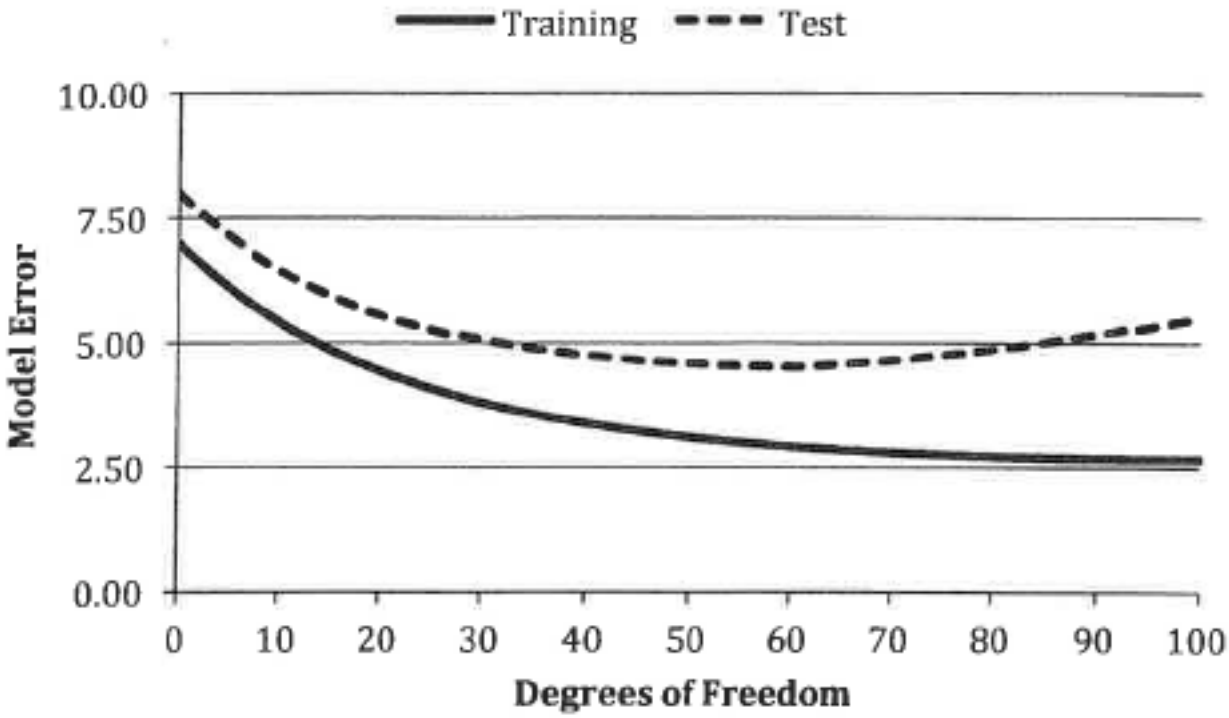
Statistic	Base Model	New Model
Loglikelihood	-750	-737.5
Deviance	500	475
Parameters	10	15
Data points	1,000,000	1,000,000

a. (1 point) Calculate the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for both models.

b. (0.25 point) Explain whether AIC or BIC is a more reliable test statistic as an indicator of whether to adopt the new model.

c. (0.25 point) Recommend and briefly justify whether to adopt the new model.

3.204. (CAS 8, 11/17, Q.4) (1.75 points) An actuary has split data into training and test groups for a model. The chart below shows the relationship between model performance and model complexity. Model performance is represented by model error and model complexity is represented by degrees of freedom.



a. (0.5 point) Briefly describe two reasons for splitting modeling data into training and test groups.

b. (0.75 point) Briefly describe whether each of the following model iterations has an optimal balance of complexity and performance.
   i. Model iteration 1: 10 degrees of freedom
   ii. Model iteration 2: 60 degrees of freedom
   iii. Model iteration 3: 100 degrees of freedom

c. (0.5 points) Identify and briefly describe one situation where it is an advantage to split the data by time rather than by random assignment.

3.205. (CAS 8, 11/17, Q.5) (1.75 points) An analyst has fit several different variations of a logistic GLM to a dataset containing 1,000 records of fraudulent claims and 9,000 records of legitimate claims.
For each model variation listed below, draw a quintile plot based on the training data.
Label the axes and identify each data series.
i. A saturated model
ii. A null model
iii. A model that could be used in practice

3.206. (CAS 8, 11/17, Q.6) (3.5 points) A logistic model was built to predict the probability of a claim being fraudulent. Consider the predicted probabilities for the 10 claims below to be a representative sample of the total model.

Claim Number	Actual Fraud Indicator	Predicted Probability of Fraud
1	Y	11%
2	N	23%
3	N	15%
4	N	70%
5	Y	91%
6	Y	30%
7	N	11%
8	Y	75%
9	N	58%
10	N	27%

a. (1 point) Construct confusion matrices for discrimination thresholds of 0.50 and 0.25.
b. (1.5 points) Plot the Receiver Operating Characteristic (ROC) curve with the discrimination thresholds of 0.50 and 0.25.
Label each axis and the coordinates and discrimination threshold of each point on the curve.
c. (0.5 point) Describe an advantage and a disadvantage of selecting a discrimination threshold of 0.25 instead of 0.50.
d. (0.5 point) Describe whether a discrimination threshold of 0.25 or 0.50 is more appropriate for a line of business with low frequency and high severity.

Solutions:

3.1. Ignoring the loglikelihood of the saturated model, which is a constant,
AIC = Deviance + (number of parameters)(2).
For example, AIC = 335.6 + (6)(2) = 347.6.

Model	Number of Parameters	Deviance	AIC
A	6	335.60	347.60
B	8	331.90	347.90
C	10	325.20	345.20
D	12	321.40	345.40
E	14	317.00	**345.00**

Since AIC is smallest for model E, model E is preferred.

3.2. When using categorical variables, it is important to set the base level to be one with populous data, so that our measures of significance will be most accurate.
By choosing the base level to be one with lots of data, the estimates of the coefficients for the non-base levels are more stable.

3.3. This allows the scale of the predictors to match the scale of the entity they are linearly predicting, which in the case of a log link is the log of the mean of the outcome.
When a logged continuous predictor is placed in a log link model, the resulting coefficient becomes a power transform of the original variable. The coefficient b_1 becomes an exponent applied to the original variable x_1.
Including continuous predictors in their logged form allows a log link GLM flexibility in fitting the appropriate response curve. On the other hand, if the variable x is not logged, the response curve for any positive coefficient will always have the same basic shape: exponential growth, that is, increasing at an increasing rate.

3.4. Frequently, the dataset going into a GLM will include records that represent the averages of the outcomes of groups of similar risks rather than the outcomes of individual risks.
In such instances, it is intuitive that records that represent a greater number of risks should carry more weight in the estimation of the model coefficients, as their outcome values are based on more data. GLMs accommodate that by allowing the user to include a weight variable, which specifies the weight given to each record in the estimation process.
The weight is the number of exposures for frequency or pure premium models.
For severity models, the weight is the number of claims.
The weight variable, usually denoted ω, formally works its way into the math of GLMs as a modification to the assumed variance: $\text{Var}[y_i] = \frac{\phi \, V(\mu_i)}{\omega_i}$.

3.5. Determining accurate estimates of relativities in the presence of moderately correlated rating variables is a primary strength of GLMs versus univariate analyses. Unlike univariate methods, the GLM will be able to sort out each variable's unique effect on the outcome, as distinct from the effect of any other variable that may correlate with it, thereby ensuring that no information is double-counted.

3.6. exp[0.4] - 1 = 49.2%.
Comment: See page 25 of Goldburd, Khare, and Tevet.
For a logistic model: Odds = μ / (1 - μ).

3.7. Both are discrete distributions used to model frequency.
Both have support from zero to infinity. Both have $\phi = 1$.
The Negative Binomial Distribution has an additional parameter $k > 0$, called the overdispersion parameter.
The Poisson Distribution has variance function $V(\mu) = \mu$, while the Negative Binomial Distribution has variance function $V(\mu) = \mu(1 + \kappa\mu)$. Thus the Negative Binomial Distribution has a variance greater than its mean, while the Poisson has a variance equal to its mean.
The Negative Binomial Distribution has a heavier righthand tail than the Poisson Distribution.
Comment: One way a Negative Binomial Distribution can arise is as a Gamma mixture of Poissons.

3.8. Where two predictors are perfectly correlated, they are said to be aliased, and the GLM will not have a unique solution.

3.9. 1. GLMs assign full credibility to the data.
2. GLMs assume that the randomness of outcomes are uncorrelated.

3.10. "**Continuity in the Estimates is Not Guaranteed.** Allowing each interval to move freely may not always be a good thing. The ordinal property of the levels of the binned variable have no meaning in the GLM; there is no way to force the GLM to have the estimates behave in any continuous fashion, and each estimate is derived independently of the others. Therefore, there is a risk that some estimates will be inconsistent with others due to random noise."
Variation within Intervals is Ignored. Since each bin is assigned a single estimate, the GLM ignores any variation that may exist within the bins.
Comment: See Section 5.4.2 of Goldburd, Khare, and Tevet.

3.11. The fitted parameter(s) are the same, while the standard errors are multiplied by $\sqrt{7.9435}$.

The standard error of $\hat{\beta}_1$ is: $0.00120\sqrt{7.9435} = 0.00338$.

95% confidence interval for β_1: 0.02085 ± (1.96) (0.00338) = **0.02085 ± 0.00662**.

Comment: One could instead use: 0.02085 ± (2) (0.00338) = 0.02085 ± 0.00676.

3.12. The Tweedie Distribution is an (linear) exponential family,
used for modeling pure premiums.
Besides the usual parameters μ and ϕ, the Tweedie Distribution has a power parameter p.
The variance function for Tweedie is $V(\mu) = \mu^p$. For use in GLMs we usually take $1 < p < 2$.
The Tweedie Distribution can be represented as a compound Poisson with a Gamma severity.
One rather interesting characteristic of the Tweedie distribution is that several of the other exponential family distributions are in fact special cases of Tweedie, dependent on the value of p.

3.13. It is clear that the proposed model more accurately predicts actual pure premium by decile than does the current rating plan. Specifically, consider the first decile. It contains the risks that the model thinks are best relative to the current plan. As it turns out, the model is correct.
Similarly, in the 10th decile, the model more accurately predicts pure premium than does the current plan.
Comment: Graph taken from "Introduction to Predictive Modeling Using GLMs A Practitioner's Viewpoint," a presentation by Dan Tevet and Anand Khare.

3.14. The use of a log link results in the linear predictor, which begins as a series of additive terms, transforming into a series of multiplicative factors when deriving the model prediction. Multiplicative models are the most common type of rating structure used for pricing insurance, due to a number of advantages they have over other structures.

3.15. The sensitivity is: $\frac{\text{true positives}}{\text{total times there is an event}} = 700 / 1000 = 0.70.$

The specificity is: $\frac{\text{true negatives}}{\text{total times there is not an event}} = 6000 / 8000 = 0.75.$

For this threshold, we graph the point: (1 - specificity , sensitivity) = **(0.25, 0.70)**.

3.16. 1. Setting of objectives and goals.
Determine the goals. Determine appropriate data to collect. Determine the time frame.
What are key risks and how can they be mitigated?
Who will work on the project; do they have the necessary knowledge and expertise?

2. Communicating with key stakeholders.
Legal and regulatory compliance. Information. Technology (IT) Department.
Underwriters. Agents.

3. Collecting and processing the necessary data for the analysis.
Time-consuming. Data is messy. Often an iterative process. The data should also be split into at least two subsets, so that the model can be tested on data that was not used to build it.
Formulate a strategy for validating the model.

4. Conducting exploratory data analysis (EDA).
Spend some time to better understand the nature of the data and the relationships between the target and explanatory variables. Helpful EDA plots include:
Plotting each response variable versus the target variable to see what (if any) relationship exists.
Plotting continuous response variables versus each other, to see the correlation between them.

5. Specifying the form of the predictive model.
What type of predictive model works best?
What is the target variable, and which response variables should be included?
Should transformations be applied to the target variable or to any of the response variables?
Which link function should be used?

6. Evaluating the model output.
Assessing the overall fit of the model.
Identifying areas in which the model fit can be improved.
Analyzing the significance of each predictor variable, and removing or transforming variables accordingly.
Comparing the lift of a newly constructed model over the existing model or rating structure.

7. Validating the model.
Assessing fit with plots of actual vs. predicted on holdout data. Measuring lift.
For Logistic Regression, use Receiver Operating Characteristic (ROC) Curves.

8. Translating the model results into a product.
For GLMs, often the desired result is a rating plan.
The product should be clear and understandable.
Are there other rating factors included in the rating plan that were not part of the GLM?

9. Maintaining the model.
Models should be periodically rebuilt in order to maximize their predictive accuracy, but in the interim it may be beneficial to merely refresh the existing model using newer data.

10. Rebuilding the model.
More frequently one would update the classification relativities without updating the rating algorithm or classification definitions. Less frequently, one would do a more complete update, investigating changing the classification definitions, the predictor variables used, and/or the rating algorithm.

Comment: See Section 3 of Goldburd, Khare, and Tevet.

3.17. (a) Concentrate on one of the explanatory variables X_j.

The partial residuals are: (ordinary residual) $g'(\hat{\mu}_i) + x_{ij} \hat{\beta}_j$.

(b) In a Partial Residual Plot, we plot the partial residuals versus the variable of interest.
If there seems to be curvature rather than linearity in the plot, that would indicate a departure from linearity between the explanatory variable of interest and $g(\mu)$, adjusting for the effects of the other independent variables.

3.18. The second model includes an interaction term.
In the second model, the effect of X_1 depends on the level of X_2 and vice-versa.
In contrast, for the first model, the effects of X_1 and X_2 are independent.

3.19. "**Check for duplicate records.** If there are any records that are exactly identical, this likely represents an error of some sort. This check should be done prior to aggregation and combination of policy and claim data."
"**Cross-check categorical fields against available documentation.** If data base documentation indicates that a roof can be of type A, B, or C, but there are records where the roof type is coded as D, this must be investigated. Are these transcription errors, or is the documentation out of date?"
"**Check numerical fields for unreasonable values.** For every numerical field, there are ranges of values that can safely be dismissed as unreasonable, and ranges that might require further investigation. A record detailing an auto policy covering a truck with an original cost (new) of $30 can safely be called an error. But if that original cost is $5,000, investigation may be needed."
Comment: Quoted from Section 4.2 of Goldburd, Khare, and Tevet.
"**Decide how to handle each error or missing value that is discovered.** The solution to duplicate records is easy, delete the duplicates. But fields with unreasonable or impossible values that cannot be corrected may be more difficult to handle."

3.20. 1. Plot each response variable versus the target variable, to see what if any relationship exists.
2. Plot continuous response variables versus each other, to see the correlation between them.

3.21. Advantages of the frequency/severity approach over pure premium modeling:
• Provides the actuary with more insight.
• Each of frequency and severity is more stable than pure premium.
Disadvantages of pure premium modeling versus the frequency/severity approach:
• Some interesting effects may go unnoticed.
• Pure premium modeling can lead to underfitting or overfitting.
• The Tweedie distribution used to model pure premium contains the implicit assumption that an increase in pure premiums is made up of an increase in both frequency and severity.

3.22. An offset is used with a Poisson Distribution and a log link function, and there are exposures associated with each observation. The offset term is ln(exposure) = $\ln(n_i)$.

Then the model is: $\ln(Y_i) = \ln(n_i) + \eta_i \Leftrightarrow Y_i = n_i \exp[\eta_i]$.

In general, an offset factor is a vector of known amounts which adjusts for known effects not otherwise included in the GLM. For example, one could take the current territories and territory relativities as givens, and include an offset term in a GLM of ln[territory relativity].

3.23. The observation for Slovakia has by far the biggest Cook's Distance, and is thus the most influential. The observations for the Czech Republic and Slovenia are less influential than Slovakia, but more influential than the others.

3.24. The saturated model has an equal number of predictors as there are records in the dataset. Since the saturated model predicts each record perfectly it is the theoretical best a model can possibly do.

The null model has only an intercept and no predictors. The null model produces the same prediction for every record: the grand mean.

The deviance for the saturated model is zero, while the deviance of the null model can be thought of as the total deviance inherent in the data. The deviance for your model will lie between those two extremes.

3.25. The deviance residuals seem to be on average positive for small and large values of X_2, while being on average negative for middle values of X_2. Such a pattern is not good. This indicates that one should investigate other possible forms of the model, for example, a model including a term involving X_2^2.

3.26. Adding credit score adds 6 - 1 = 5 parameters to the model.

Test statistic is:

$$F = \frac{D_S - D_B}{(\text{number of added parameters})\,\hat{\phi}_S} = \frac{(233{,}183.65 - 233{,}134.37) / 5}{2.371} = 4.157.$$

The number of degrees of freedom in the numerator is 5.

The number of degrees of freedom in the denominator is:

number of observations minus the number of parameters in the smaller model

= 100,000 - 10 = 99,990.

We compare the test statistic to an F-distribution with 5 and 99,990 degrees of freedom.

The null hypothesis is to use the simpler model.

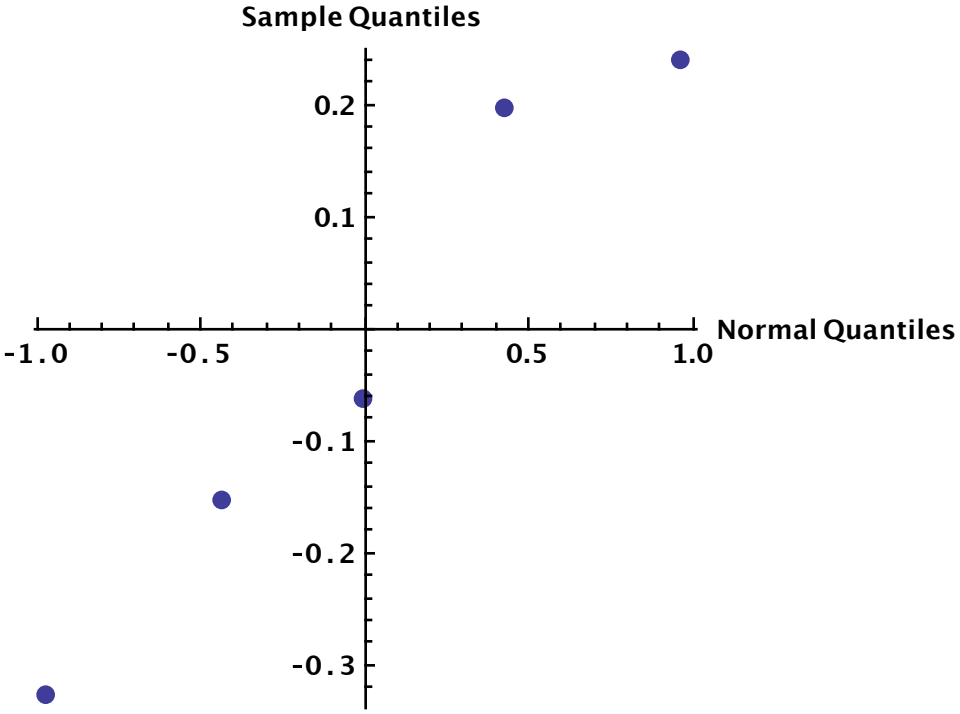
The alternate hypothesis is to use the more complex model including credit score.

We reject the null hypothesis when the F-Statistic is big.

Comment: Using a computer, the p-value of this test is 0.09%.

Thus one would use the more complex model including credit score rather than the simpler model.

3.27. Arranged from smallest to largest: -0.328, -0.154, -0.064, 0.195, 0.239.
Plot $(Q_{i/6}, x_{(i)})$.
$Q_{1/6} = -0.967$, since $\Phi[-0.967] = 1/6$. $Q_{2/6} = -0.431$. $Q_{3/6} = 0$. $Q_{4/6} = 0.431$. $Q_{5/6} = 0.967$.
Thus the five plotted points are:
(-0.967, -0.328), (-0.431, -0.154), (0, -0.064), (0.431, 0.195), (0.967, 0.239).
Here are the 5 points plotted:



There is too little data to decide whether or not these stock price returns are Normally distributed.

3.28. Gini index = **2A**.

Comment: Gini index = $\frac{\text{Area A}}{\text{Area A + Area B}}$.

However, Area A + Area B add up to a triangle with area 1/2.

Therefore, Gini index = $\frac{\text{Area A}}{\text{Area A + Area B}} = 2\text{A}$

= twice the area between the Lorenz Curve and the line of equality = 1 - 2B.

3.29. Factors for coverage options should be estimated outside the GLM, using traditional actuarial techniques. The resulting factors should then be included in the GLM as an offset.

3.30. Min[X - c, 0], where c is some constant and X is a variable.
For example, $\text{Min}[X_2 - 13, 0]$ is a hinge function.

3.31. Model D is preferred. Bigger Area Under ROC Curve (AUROC) is better.

3.32. $35 \pm 1.96 \sqrt{5} =$ **(30.62, 39.38)**.

3.33. BIC = (-2) (maximum loglikelihood) + (number of parameters)ln[400].
For example, BIC = (-2)(-730.18) + 3 ln[400] = 1478.33.

Model	Number of Parameters	Loglikelihood	BIC
A	3	-730.18	1478.33
B	4	-726.24	**1476.45**
C	5	-723.56	1477.08
D	6	-721.02	1477.99
E	7	-717.50	1476.94

Since BIC is smallest for model B, model B is preferred.

3.34. exp[-3.8 + (0.4)ln[1/2]] = **1.7%**.
Comment: Loosely based on Table 12 in Generalized Linear Models for Insurance Rating, by Goldburd, Khare and Tevet.

3.35. exp[-3.8 + 0.3 - 0.5 + (0.4) ln[2.5/2] - (0.1) ln[2.5/2]] = **2.0%**.

3.36. exp[-3.8 + 0.5 + (0.4)ln[3/2]] = **4.3%**.

3.37. exp[-3.8 + 0.1 - 0.5 + (0.4)ln[6/2] - (0.1) ln[6/2]] = **2.1%**.

3.38. Histogram A most closely matches the Normal Distribution.

3.39. a) Identity link function.
b) Log link function.
c) Poisson Distribution.
d) For the variance proportional to the square of the mean, use the Gamma Distribution.

3.40. The partial residual plot is not linear; thus, we should do something to improve the model. Since the slope seems to change somewhere around 50 or 60, we could use a hinge function: Min[0, X_4 - 50] or Min[0, X_4 - 60].
Comment: In general, we could instead group the variable, or add polynomial terms to the model.

3.41. We can divide the original data into three sets.
We fit GLMs to the training data, until we have one or more good candidate models.
Then we see how these models perform on the validation set.
Based on what we find out, we can go back and fit some other GLMs to the training data.
The validation set is used to refine the models during the building process.
The test set (holdout data) is held out until the end.
We compare the performance of models on the test set to pick a final model to use.

3.42.

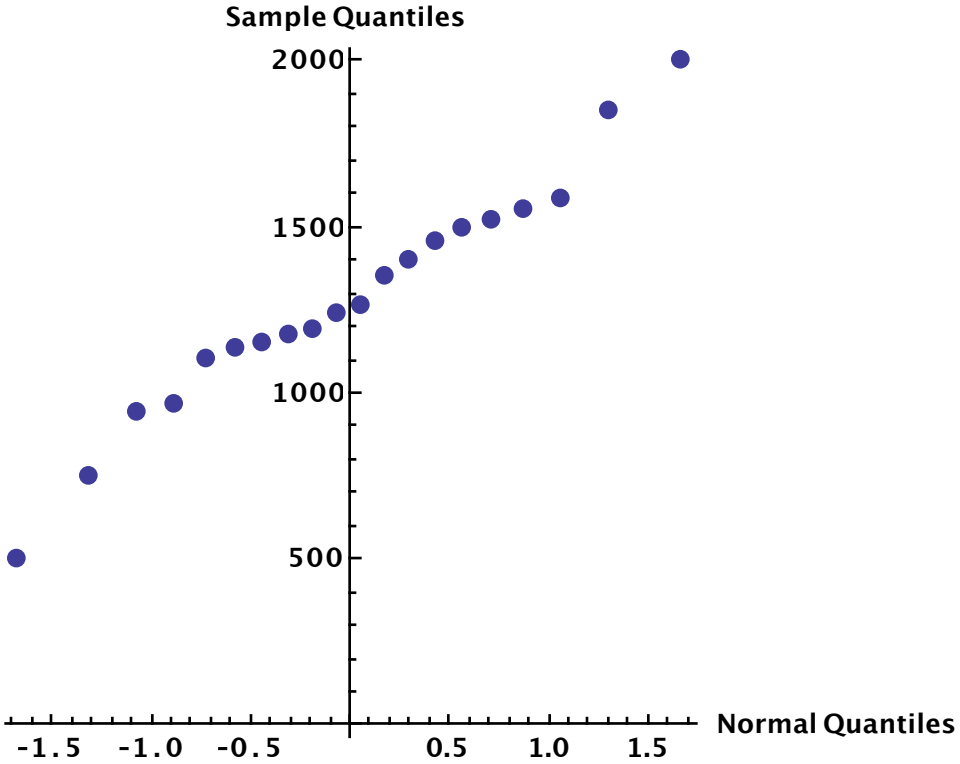
Distribution	V(μ)
Normal	$\mu^0 = 1$
Poisson	$\mu^1 = \mu$
Gamma	μ^2
Binomial (one trial)	$\mu\ (1-\mu)$
Inverse Gaussian	μ^3
Tweedie	μ^p, $p < 0$, $1 < p < 2$, or $p > 2$.

Alternately, for the Binomial Distribution, $V(\mu) = \mu\ (1 - \mu/m)$.

3.43. $Q_{1/21} = -1.668$, since $\Phi[-1.668] = 1/21$.

Thus the first plotted point is: (-1.668 , 500).

The Q-Q Plot:



3.44. For a Poisson, $f(n) = e^{-\lambda}\lambda^n/n!$.

$\ln f(n) = -\lambda + n\ln\lambda - \ln(n!) = -\exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}] + n_i(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}) - \ln(n_i!)$.

loglikelihood $= -\sum\exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}] + \sum Y_i(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})$ + constants.

Setting the partial derivatives of the loglikelihood with respect to β_0, β_1, and β_2 equal to zero:

$0 = -\sum\exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}] + \sum Y_i$.

$0 = -\sum X_{1i}\exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}] + \sum Y_i X_{1i}$.

$0 = -\sum X_{2i}\exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}] + \sum Y_i X_{2i}$.

$\sum Y_i = 8 + 8 + 10 + ... + 33 + 31 = 369$.

$\sum Y_i X_{1i} = 8\ln(2) + 8\ln(4) + 10\ln(6) + ... + 33\ln(18) + 31\ln(20) = 872.856$.

$\sum Y_i X_{2i} = 14 + 19 + ... + 33 + 31 = 241$.

$\exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}] = \exp[\beta_0]\exp[\beta_1 X_{1i}]\exp[\beta_2 X_{2i}] = \exp[\beta_0]\exp[X_{1i}]^{\beta_1}\exp[\beta_2 X_{2i}]$.

The first equation becomes:

$\exp[\beta_0]\{2^{\beta_1} + 4^{\beta_1} + ... + 20^{\beta_1} + 2^{\beta_1}\exp[\beta_2] + 4^{\beta_1}\exp[\beta_2] + 20^{\beta_1}\exp[\beta_2]\} = 369. \Rightarrow$

$\exp[\beta_0](1 + \exp[\beta_2])\{2^{\beta_1} + 4^{\beta_1} + 6^{\beta_1} + ... + 20^{\beta_1}\} = 369$.

The second equation becomes:

$\exp[\beta_0](1 + \exp[\beta_2])\{\ln(2)2^{\beta_1} + \ln(4)4^{\beta_1} + \ln(6)6^{\beta_1} + ... + \ln(20)20^{\beta_1}\} = 872.856$.

The third equation becomes:

$\exp[\beta_0]\exp[\beta_2]\{2^{\beta_1} + 4^{\beta_1} + 6^{\beta_1} + ... + 20^{\beta_1}\} = 241$.

Comment: Well beyond what you should be asked on your exam!

A Poisson variable with a logarithmic link function.

Dividing the 1st and 3rd equations:

$(1 + \exp[\beta_2])/\exp[\beta_2] = 369/241. \Rightarrow \beta_2 = \ln(241/148) = 0.6328$.

Using a computer, the fitted parameters are: $\beta_0 = 1.684$, $\beta_1 = 0.3784$, $\beta_2 = 0.6328$.

One can verify that these values satisfy the three equations.

Example taken from Applied Regression Analysis by Draper and Smith.

3.45. While one may assume that the errors are Normally Distributed, in a GLM one could assume a different distribution of errors, such as Gamma or Poisson.

Thus Statement #1 is not true.

Statements #2 and #3 are true.

3.46. With four age categories, we add 4 -1 = 3 parameters.

Test statistic is: $F = \frac{D_S - D_B}{(\text{number of added parameters})\, \hat{\phi}_S} = \frac{3320.2 - 3306.9}{(3)\ (1.83)} = 2.42.$

The number of degrees of freedom in the numerator is 3.
The number of degrees of freedom in the denominator is:
number of observations minus the number of parameters in the smaller model.
We compare the test statistic to the appropriate F-distribution.
We reject the null hypothesis if the test statistic is sufficiently big.

3.47. "Broadly speaking, model lift is the economic value of a model. The phrase "economic value" doesn't necessarily mean the profit that an insurer can expect to earn as a result of implementing a model, but rather it refers to a model's ability to prevent adverse selection. The lift measures ... attempt to visually demonstrate or quantify a model's ability to charge each insured an actuarially fair rate, thereby minimizing the potential for adverse selection.
Model lift is a relative concept, so it requires two or more competing models. That is, it doesn't generally make sense to talk about the lift of a specific model, but rather the lift of one model over another.
In order to prevent overfitting, model lift should always be measured on holdout data."
Comment: Quoted from Section 7.2 of Goldburd, Khare, and Tevet.

3.48. The effects of age and gender interact strongly. For example, the relationship between male and female relativities is very different for young drivers than it is for middle-aged drivers. In contrast, the effects of frequency of payment and age do not appear to interact significantly; there seems to be approximately the same relationship for each age group.
Comment: The graphs are adapted from "A Practitioner's Guide to Generalized Linear Models," by Duncan Anderson, Sholom Feldblum, Claudine Modlin, Doris Schirmacher, Ernesto Schirmacher, and Neeza Thandi.

3.49. The second model is preferred since the predictions are closer to the actual than in Model 1.
Comment: See Figure 17 in Goldburd, Khare, and Tevet.

3.50. Let $X_1 = 1$ if age group A, and 0 otherwise.
$X_2 = 1$ if age group B, and 0 otherwise.
$X_3 = 1$ if small, and 0 otherwise.
$X_4 = 1$ if medium, and 0 otherwise.
$X_5 = 1$ if large, and 0 otherwise.
Then the design matrix is:

$$\begin{pmatrix} \text{A/small} \\ \text{A/medium} \\ \text{A/large} \\ \text{B/small} \\ \text{B/medium} \\ \text{B/large} \\ \text{C/small} \\ \text{C/medium} \\ \text{C/large} \end{pmatrix} \Leftrightarrow \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

For example, the first row corresponds to age group A and small:
$X_1 = 1$, $X_2 = 0$, $X_3 = 1$, $X_4 = 0$, and $X_5 = 0$.
The last row corresponds to age group C and large: $X_1 = 0$, $X_2 = 0$, $X_3 = 0$, $X_4 = 0$, and $X_5 = 1$.

The vector of parameters is: $\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}$.

Alternately, define medium and age group C as the base level.
Then the constant, b_0, would apply to all observations.
Let $X_1 = 1$ if age group A, and 0 otherwise.
$X_2 = 1$ if age group B, and 0 otherwise.
$X_3 = 1$ if small, and 0 otherwise.
$X_4 = 1$ if large, and 0 otherwise.
Then the design matrix is:

$$\begin{pmatrix} \text{A/small} \\ \text{A/medium} \\ \text{A/large} \\ \text{B/small} \\ \text{B/medium} \\ \text{B/large} \\ \text{C/small} \\ \text{C/medium} \\ \text{C/large} \end{pmatrix} \Leftrightarrow \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The first column of ones corresponds to the constant term which applies to all observations.
For example, the first row corresponds to age group A and small:
$X_0 = 1$, $X_1 = 1$, $X_2 = 0$, $X_3 = 1$, $X_4 = 0$.
The last row corresponds to age group C and large: $X_0 = 1$, $X_1 = 0$, $X_2 = 0$, $X_3 = 0$, $X_4 = 1$.

The vector of parameters is: $\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$.

Comment: There is no unique answer. I have given two out of the many possible answers.
There are 3 age categories and 3 size categories, so we need to have either 3 + 3 - 1 = 5 covariates, or 4 covariates and a constant term.
The data would be arranged in a grid such as:

	Small	Medium	Large
Age A	???	???	???
Age B	???	???	???
Age C	???	???	???

The response vector would have 9 rows and one column, containing the observations in the same order as the rows of the design matrix.

3.51. "If the modeler retains variables in the model that reflect a non-systematic effect on the response variable (i.e., noise) or over-specifies the model with high order polynomials, the result is over-fitting. Such a model will replicate the historical data very well (including the noise) but is not going to predict future outcomes reliably (as the future experience will most likely not have the same noise).
Conversely, if the model is missing important statistical effects (the extreme being a model that contains no explanatory variables and fits to the overall mean), the result is under-fitting. This model will predict future outcomes (e.g., in the extreme case mentioned above, the future mean) reliably but hardly help the modeler explain what is driving the result."
"Considerable disparity between actual and expected results on the hold-out sample may indicate that the model is over or under-fitting."

Underfit. ⇔ Too few Parameters. ⇔ Does not use enough of the useful information.

Overfit. ⇔ Too many Parameters. ⇔ Reflects too much of the noise.
In general, the actuary wants to avoid both underfitting and overfitting models.
Comment: See page 182 of Basic Ratemaking, on Exam 5.

3.52. The 18th observation has by far the biggest Cook's Distance, and is thus the most influential.

3.53. (a) AIC = (-2) (maximum loglikelihood) + (number of parameters)(2).
For the first model, AIC = (-2)(-321.06) + (6)(2) = 654.12.
For the second model, AIC = (-2)(-319.83) + (7)(2) = 653.66.
Since AIC is smaller for the second model, the second model is preferred.
(b) BIC = (-2) (maximum loglikelihood)) + (number of parameters) ln(number of data points).
For the first model, BIC = (-2)(-321.06) + (6) ln[100] = 669.75.
For the second model, BIC = (-2)(-319.83) + (7) ln[100] = 671.90.
Since BIC is smaller for the first model, the first model is preferred.
Comment: An example where using AIC and BIC lead to different conclusions.

3.54. Graph B is closest to a straight line.
Comment: If the data was drawn from a Normal Distribution with $m \neq 0$, then we would expect the plotted points to be close to a straight line, but not a straight line through the origin.

3.55. (a) Poisson with log link function.
(b) Poisson or Negative Binomial with log link function.
(c) Gamma with log link function.
(d) Binomial with logit link function.
(e) Tweedie with log link function.
Comment: Claim frequency is claim count per exposure. If each insured has the same number of exposures, then a model of claim counts and claim frequency are mathematically equivalent.

3.56. Larger Gini index is better, all else being equal. The second rating plan is preferred
Comment: The higher the Gini index, the better the model is at identifying risk differences.

3.57. If the current rating plan were perfect, then all risks should have the same loss ratio. The fact that the proposed model is able to segment the data into lower and higher loss ratio buckets is a strong indicator that it is outperforming the current rating plan.
Comment: Graph taken from "Introduction to Predictive Modeling Using GLMs A Practitioner's Viewpoint," a presentation by Dan Tevet and Anand Khare.

3.58. For levels 1 to 7 of the variable, the log of the multiplier is not significantly different than zero; in other words the relativity is not significantly different from one. Also for levels 1 to 7, there is no consistent pattern. Thus perhaps, levels 1 to 8 of this variable should be grouped into one level for purposes of the model; this would be treated as the new base.
In contrast, for levels 9 to 15 there is pattern of increasing relativities. For levels 11 to 15 the relativities are significantly different from one. Given the pattern, one could also use the indicated relativities for levels 9 and 10.
Comment: As always, more testing may lead to a different conclusion. For example, it would be interesting to compare the results for different years of data to see if they are consistent.
For example, the levels of the variable could be groups of annual income.

3.59. There are many ways to define the variables.
Let us define $X_1 = 1$ if male and zero otherwise.
$X_2 = 1$ if female and zero otherwise.
$X_3 = 1$ if urban and zero otherwise.
For the Poisson, $f(x) = \lambda^x e^{-\lambda} / x!$. $\ln f(x) = x \ln(\lambda) - \lambda - \ln(x!) = x \ln(\mu) - \mu$ - constants.
(a) We use an identity link function. The estimated means are:

	Urban	Rural
Male	$\beta_1 + \beta_3$	β_1
Female	$\beta_2 + \beta_3$	β_2

Ignoring constants, the loglikelihood is:
$0.2 \ln(\beta_1 + \beta_3) - (\beta_1 + \beta_3) + 0.1 \ln(\beta_1) - (\beta_1) + 0.125 \ln(\beta_2 + \beta_3) - (\beta_2 + \beta_3) + 0.05 \ln(\beta_2) - (\beta_2)$.
Setting the partial derivative with respect to β_1 equal to zero: $0.2/(\beta_1 + \beta_3) + 0.1/\beta_1 = 2$.
Setting the partial derivative with respect to β_2 equal to zero: $0.125/(\beta_2 + \beta_3) + 0.05/\beta_2 = 2$.
Setting the partial derivative with respect to β_3 equal to zero: $0.2/(\beta_1 + \beta_3) + 0.125/(\beta_2 + \beta_3) = 2$.
(b) We use an log link function. The estimated means are:

	Urban	Rural
Male	$\exp[\beta_1 + \beta_3]$	$\exp[\beta_1]$
Female	$\exp[\beta_2 + \beta_3]$	$\exp[\beta_2]$

Ignoring constants, the loglikelihood is:
$0.2(\beta_1 + \beta_3) - \exp[\beta_1 + \beta_3] + 0.1\beta_1 - \exp[\beta_1] + 0.125(\beta_2 + \beta_3) - \exp[\beta_2 + \beta_3] + 0.05\,\beta_2 - \exp[\beta_2]$.
Setting the partial derivative with respect to β_1 equal to zero: $\exp[\beta_1 + \beta_3] + \exp[\beta_1] = 0.3$.
Setting the partial derivative with respect to β_2 equal to zero: $\exp[\beta_2 + \beta_3] + \exp[\beta_2] = 0.175$.
Setting the partial derivative with respect to β_3 equal to zero: $\exp[\beta_1 + \beta_3] + \exp[\beta_2 + \beta_3] = 0.325$.
Comment: Using a computer, the fitted parameters in part (a) are:
$\beta_1 = 0.105556$, $\beta_2 = 0.047500$, $\beta_3 = 0.084444$.
The fitted frequencies are: 0.1900, 0.1056, 0.1319, 0.0475.
Using a computer, the fitted parameters in part (b) are:
$\beta_1 = -2.35665$, $\beta_2 = -2.89565$, $\beta_3 = 0.77319$.
The fitted frequencies are: 0.2053, 0.0947, 0.1197, 0.0553.

3.60. The sensitivity is: $\frac{\text{true positives}}{\text{total times there is an event}} = 1800/3000 = 0.6$.

The specificity is: $\frac{\text{true negatives}}{\text{total times there is not an event}} = 40{,}000/50{,}000 = 0.8$.

For this threshold, we graph the point: (1 - specificity , sensitivity) = **(0.2, 0.6)**.

3.61. $f(y) = \exp[-(y - \mu)^2/(2\sigma^2)]/\{\sigma\sqrt{2\pi}\}$. $\ln f(Y_i) = -(Y_i - \beta X_i)^2/(2\sigma^2) - \ln(\sigma) - \ln(2\pi)/2$.

Loglikelihood is: $-\sum(Y_i - \beta X_i)^2/(2\sigma^2) - n \ln(\sigma) - n \ln(2\pi)/2$.

Set the partial derivative of the loglikelihood with respect to b equal to zero:

$0 = \sum X_i(Y_i - \beta X_i)/\sigma^2. \Rightarrow \sum X_i Y_i = \beta \sum X_i^2. \Rightarrow \hat{\beta} = \sum X_i Y_i / \sum X_i^2 = 3080/751 =$ **4.10**.

Comment: Matches the linear regression model with no intercept, $\hat{\beta} = \sum X_i Y_i / \sum X_i^2$.

3.62. Set the partial derivative of the loglikelihood with respect to s equal to zero:

$0 = \sum(Y_i - \beta X_i)^2/\sigma^3 - n/\sigma. \Rightarrow \sigma^2 = \sum(Y_i - \beta X_i)^2/n =$

$$\frac{\{5 - (1)(4.1)\}^2 + \{15 - (5)(4.1)\}^2 + \{50 - (10)(4.1)\}^2 + \{100 - (25)(4.1)\}^2}{4} = 29.58.$$

$\hat{\beta} = \sum X_i Y_i / \sum X_i^2$. $\text{Var}[\hat{\beta}] = \text{Var}[\sum X_i Y_i / \sum X_i^2] = \sum \text{Var}[X_i Y_i / \sum X_i^2] = \sum X_i^2 \text{Var}[Y_i]/ (\sum X_i^2)^2 =$

$\sum X_i^2 \sigma^2/ (\sum X_i^2)^2 = \sigma^2/\sum X_i^2 = 29.58/751 = 0.0394$.

$\text{StdDev}[\hat{\beta}] = \sqrt{0.0394} =$ **0.198.**

Comment: In the linear regression version of this same example, one would estimate the variance of the regression as: $\sigma^2 = \sum \hat{\varepsilon}_i^2 / (N - 1) =$

$$\frac{\{5 - (1)(4.1)\}^2 + \{15 - (5)(4.1)\}^2 + \{50 - (10)(4.1)\}^2 + \{100 - (25)(4.1)\}^2}{4 - 1} = 39.4.$$ This is an unbiased estimate of σ^2, which is not equal to that from maximum likelihood which is biased.

3.63. Estimated mean severity for a male in Territory D is: $\exp[8.03 + 0.18 + 0.22] = 4583$.

For the Inverse Gaussian Distribution, $\text{Var}[Y] = \phi\mu^3 = (0.00510)(4583^3) = 490{,}930{,}199$.

$\text{StdDev}[Y] = \sqrt{490{,}930{,}199} =$ **22,157**.

3.64. 1. Actuarial judgement. Does the model make sense; is the model reasonable.
2. Statistical Tests such as the F-Test.
3. Graph the modeled relativities plus or minus two standard errors.
   We would like the range between plus and minus two standard errors to be relatively narrow.
4. Check the consistency of the model run on different years of data.
5. Check the predictive accuracy of the model on a hold-out data set.

Comment: There are other possible answers.

3.65. The variance of the residuals appears to increasing with the fitted values, indicating heteroscedasticity (a lack of homoscedasticity.) This is not good, and one should try to refine the current model.

3.66. The deviance is equal to twice the difference between the maximum achievable loglikelihood (i.e., the loglikelihood where the fitted value is equal to the observation for every record) and the loglikelihood of the model.
Alternately, the deviance is equal to twice the difference between the loglikelihood of the saturated model and the loglikelihood of the fitted model.

3.67. $\ln[\mu/207] = 0.43 + 0.22 - 0.32 + 0.36 = 0.69$.
$\mu = 207 \exp[0.69] =$ **$413**.
Comment: This is a multiplicative model with four categorical variables.

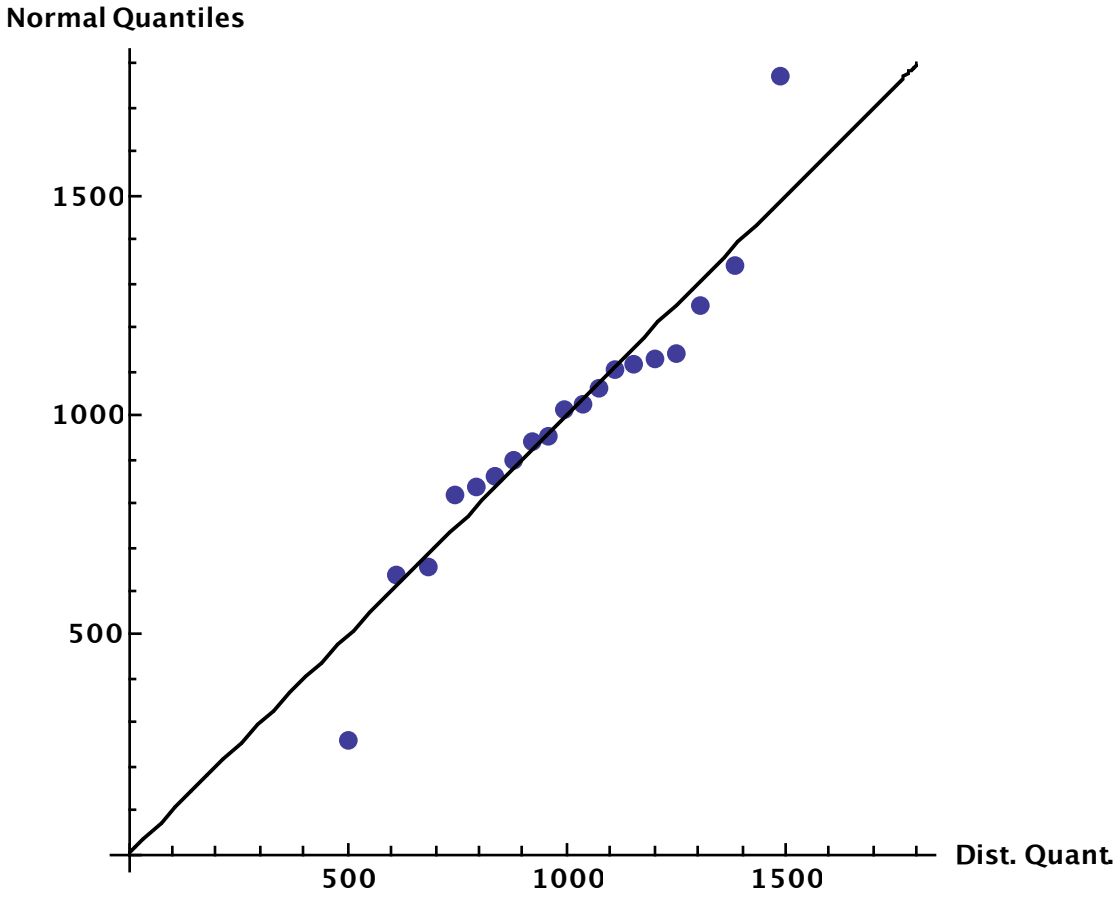
3.68. A model that combines information from two or more models is called an ensemble model.
Two (or more) teams model the same item; they build separate models working independently.
Combining the answers from both models is likely to perform better than either individually.
A simple means of ensembling is to average the separate model predictions.

3.69. Test statistic is: $F = \frac{D_S - D_B}{(\text{number of added parameters}) \hat{\phi}_S} = \frac{(24{,}359 - 24{,}352) / 1}{1.22} = 5.738$.

The number of degrees of freedom in the numerator is 1.
The number of degrees of freedom in the denominator is:
number of observations minus the number of parameters in the smaller model
= 20,000 - 4 = 19,996.
This is equivalent to a two sided t-test at $\sqrt{5.738} = 2.395$, with 19,996 degrees of freedom.
Using the Normal approximation, the p-value is: (2) (1 - Φ[2.395]).
Since 2.326 < 2.395 < 2.576, the two-sided p-value is between 2% and 1%.
Thus at a 2% significance level we should use the more complex model with the added variable, but at a 1% significance level we should use the simpler model without the additional variable.
Comment: Using a computer, the p-value of this test is 1.66%.
The null hypothesis is to use the simpler model. The alternate hypothesis is to use the more complex model. We reject the null hypothesis if the test statistic is sufficiently big.

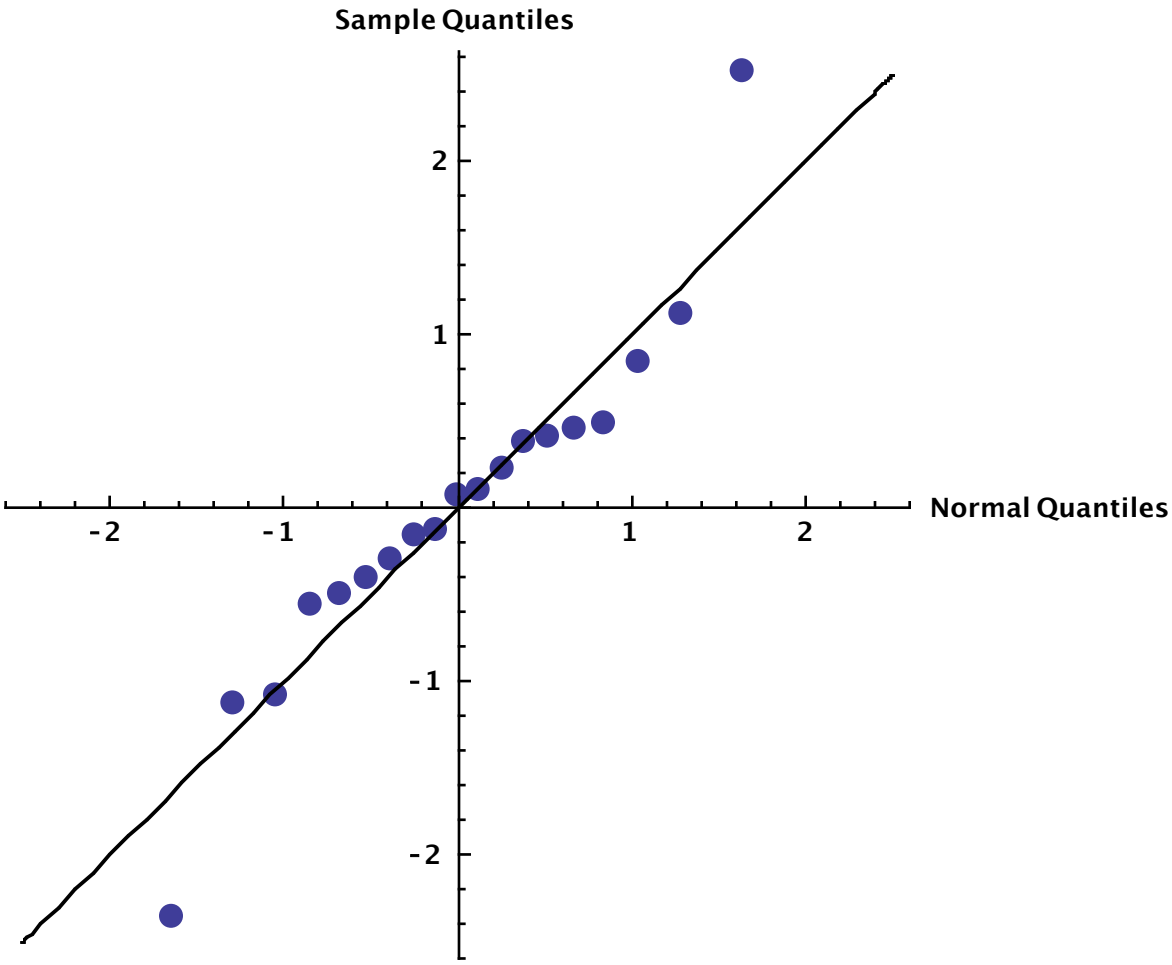
3.70. The actuary would like the GLM to be stable; in other words, the predictions of the model should not be overly sensitive to small changes in the data.
An observation is influential if it has a large effect on the fitted model.
The larger the value of Cook's distance, the more influential the observation.
The actuary should rerun the model excluding the most influential points to see their impact on the results. If this causes large changes in some of the parameter estimates, the actuary should consider for example whether to give these influential observations less weight.
Cross-validation can also be used to assess the stability of a GLM. A single model can be run on the set of folds. The results of the models fit to these different subsets of the data ideally should be similar. The amount by which these results vary is a measure of the stability of the model.
Bootstrapping via simulation can also be used to assess the stability of a GLM. The original data is randomly sampled with replacement to create a new set of data of the same size. One then fits the GLM to this new set of data. By repeating this procedure many times one can estimate the distribution of the parameter estimates of the GLM; we can estimate the mean, variance, confidence intervals, etc.

3.71. $\Phi[-1.645] = 1/20$. Thus for the given Normal, $Q_{0.05} = 1000 - (1.645)(300) = 506.5$.
The 19 plotted points are: (506.5, 258), (615.5, 636), (689.1, 652), (747.5, 814), (797.7, 833), (842.7, 860), (884.4, 895), (924.0, 937), (962.3, 950), (1000.0, 1009), (1037.7, 1020), (1076.0, 1059), (1115.6, 1103), (1157.3, 1113), (1202.3, 1127), (1252.5, 1139), (1310.9, 1246), (1384.5, 1335), (1493.5, 1770).
The resulting Q-Q plot:



Alternately, one could standardize the data, by subtracting the sample mean of 987.158 and dividing by the square root of the sample variance of 96,057.8.
For example, (258 - 987.158) / $\sqrt{96{,}057.8}$ = -2.353.

Then one compares the standardized data to the quantiles of the Standard Normal Distribution.
(-1.645, -2.353), (-1.282, -1.133), (-1.036, -1.081), (-0.842, -0.559), (-0.674, -0.497), (-0.524, -0.410), (-0.385, -0.297), (-0.253, -0.162), (-0.126, -0.120), (0, 0.070), (0.126, 0.106), (0.253, 0.232), (0.385, 0.374), (0.524, 0.406), (0.674, 0.451), (0.842, 0.490), (1.036, 0.835), (1.282, 1.122), (1.645, 2.526).
The resulting Q-Q plot:



Comment: With the exception of the first and last plotted points, the points stay close to the 45 degree comparison line, indicating that this data may be normally distributed.

3.72. a) The total number of cells is: (2)(4)(3) = 24. So the design matrix would have 24 rows. Each row has a one in the first column; the intercept term applies to all insureds.
For example, the first row has one in columns 3 and 6 corresponding to age 17-21 and Territory A.

1	0	1	0	0	1	0	F 17-21 A
1	0	0	1	0	1	0	F 22-29 A
1	0	0	0	0	1	0	F 30-59 A
1	0	0	0	1	1	0	F 60+ A
1	1	1	0	0	1	0	M 17-21 A
1	1	0	1	0	1	0	M 22-29 A
1	1	0	0	0	1	0	M 30-59 A
1	1	0	0	1	1	0	M 60+ A
1	0	1	0	0	0	0	F 17-21 B
1	0	0	1	0	0	0	F 22-29 B
1	0	0	0	0	0	0	F 30-59 B
1	0	0	0	1	0	0	F 60+ B
1	1	1	0	0	0	0	M 17-21 B
1	1	0	1	0	0	0	M 22-29 B
1	1	0	0	0	0	0	M 30-59 B
1	1	0	0	1	0	0	M 60+ B
1	0	1	0	0	0	1	F 17-21 C
1	0	0	1	0	0	1	F 22-29 C
1	0	0	0	0	0	1	F 30-59 C
1	0	0	0	1	0	1	F 60+ C
1	1	1	0	0	0	1	M 17-21 C
1	1	0	1	0	0	1	M 22-29 C
1	1	0	0	0	0	1	M 30-59 C
1	1	0	0	1	0	1	M 60+ C

b) 30-59 year old female driver in Territory B is the base. Estimated frequency is $\exp[\hat{\beta}_1]$.

c) For 22-29 year old male driver in Territory C, the estimated frequency is:
$\exp[\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_4 + \hat{\beta}_7]$.

Comment: One can arrange the rows of the design matrix differently, as long as everything is consistent. Since there is an intercept term, and since each of the factors is a categorical variable, each has one less parameter than its number of levels.
We have chosen 30-59 year old female driver in Territory B as the base; some other choice could have been made.

3.73. AIC = (-2) (maximum loglikelihood) + (number of parameters)(2).
For example, AIC = (-2)(-359.17) + (3)(2) = 724.34.

Model	Number of Parameters	Loglikelihood	AIC
A	3	-359.17	724.34
B	4	-357.84	723.68
C	5	-356.42	722.84
D	6	-354.63	**721.26**
E	7	-353.85	721.70

Since AIC is smallest for model D, model D is preferred.

3.74. In the first graph, the relativities indicated by the separate years are similar to each other. Also the relativities for each year display a similar pattern of increase with vehicle symbol, which makes sense. Vehicle symbol appears to be a significant factor for the first model; it is likely to be a good predictor of future experience.
In the second graph, the relativities indicated by separate years are not consistent. Territory does not appear to be a significant factor for the second model.
Comment: The graphs are adapted from ones showing more information in Sections 2.40-2.41 of "A Practitioner's Guide to Generalized Linear Models,"by Duncan Anderson, Sholom Feldblum, Claudine Modlin, Doris Schirmacher, Ernesto Schirmacher, and Neeza Thandi, not on the syllabus.

3.75. If two predictors are highly correlated (have a correlation coefficient close to plus or minus one) coefficients may behave erratically. Furthermore, the standard errors associated with those coefficients will be large, and small perturbations in the data may swing the coefficient estimates wildly. Such instability in a model should be avoided. As such it is important to look out for instances of high correlation prior to modeling, by examining two-way correlation tables.
Where high correlation is detected, means of dealing with this include the following:
• For any group of correlated predictors, remove all but one from the model.
• Preprocess the data using dimensionality reduction techniques such as principal component analysis.
Multicollinearity: A more subtle potential problem may exist where two or more predictors in a model may be strongly predictive of a third, a situation known as multicollinearity. The same instability problems as above may result. A useful statistic for detecting multicollinearity is the variance inflation factor (VIF), which can be output by most statistical packages. A common statistical rule of thumb is that a VIF greater than 10 is considered high.
Aliasing: Where two predictors are perfectly correlated, they are said to be aliased, and the GLM will not have a unique solution. Where they are nearly perfectly correlated, the model will be highly unstable; the fitting procedure may fail to converge, and even if the model run is successful the
estimated coefficients will be nonsensical. Such problems can be avoided by looking out for and properly handling correlations among predictors, as discussed above.
Comment: See Section 2.9 of Goldburd, Khare, and Tevet.
Not necessary to say all of the above rather than some of the above.

3.76. 1. They are simple and practical to implement.
2. Having additive terms in a model can result in negative premiums, which doesn't make sense. With a multiplicative plan you guarantee positive premium without having to implement clunky patches like minimum premium rules.
3. A multiplicative model has more intuitive appeal. It doesn't make much sense to say that having a violation should increase your auto premium by $500, regardless of whether your base premium is $1,000 or $10,000.
Rather it makes more sense to say that the surcharge for having a violation is 10%.
Comment: For these and other reasons, log link models, which produce multiplicative structures, are usually the most natural model for insurance risk.
"As for the link function, it is usually the case that the desirability of a multiplicative rating plan trumps all other considerations, so the log link is almost always used. One notable exception is where the target variable is binary (i.e., occurrence or non-occurrence of an event), for which a special link function (logistic) must be used."

3.77. In order to incorporate age, avoiding aliasing, we need 6 - 1 = 5 variables.
In order to incorporate gender, we would need one more variable for a total of 6.
So getting rid of age and gender would produce a model with 6 fewer parameters.

Test statistic is: $F = \frac{D_S - D_B}{(\text{number of added parameters})\,\hat{\phi}_S} = \frac{(1128.1 - 1120.3) / 6}{0.395} = 3.291.$

The number of degrees of freedom in the numerator is 6.
The number of degrees of freedom in the denominator is:
number of observations minus the number of parameters in the smaller model
= 1000 - 44 = 956.
We compare the test statistic to an F-distribution with 6 and 956 degrees of freedom.
The null hypothesis is to use the simpler model, the one without age and gender.
The alternate hypothesis is to use the more complex model.
We reject the null hypothesis if the test statistic is sufficiently big.
Comment: Using a computer, the p-value of this test is 3.3%.

3.78. Firstly, when comparing two models using log-likelihood or deviance, the comparison is valid only if the data sets used to fit the two models are exactly identical. If a new variable has missing values for some records, the default behavior of most model fitting software is to toss out those records when fitting the model. In that case, the resulting measures of fit are no longer comparable, since the second model was fit with fewer records than the first.
For any comparisons of models that use deviance it is also necessary that the assumed distribution and the dispersion parameter (basically, everything other than the coefficients) must be identical as well.
Comment: See Section 6.1.3 of Goldburd, Khare, and Tevet.

3.79. Age of spokesperson, gender of spokesperson, marital status of the spokesperson, time he has been a spokesperson, type of celebrity (actor, singer, athlete, etc.), criminal record of the spokesperson, past drug/alcohol abuse of the spokesperson, etc.
Comment: There are other reasonable answers.
This is often sold as death, disability, and disgrace insurance.

3.80. Both are continuous distributions used to model severity. Both are right-skewed, with a sharp peak and a long tail to the right, and a lower bound at zero.
The Gamma Distribution has variance function $V(\mu) = \mu^2$, while the Inverse Gaussian Distribution has variance function $V(\mu) = \mu^3$.
The Inverse Gaussian Distribution has a sharper peak and a wider tail than the Gamma Distribution.
Therefore, the Inverse Gaussian Distribution is appropriate for situations where the skewness of the severity curve is more extreme.
Comment: The skewness for the Gamma distribution is always twice times the coefficient of variation, while the skewness for the Inverse Gaussian distribution is always three times the coefficient of variation.

3.81. a) 9.5 + (0.01)(180) + (-0.02)(670) = -2.1.
Using the inverse of the logit link function, the probability of default is:
$\frac{\exp(-2.1)}{1 + \exp(-2.1)}$ = **10.9%**.

b) 9.5 + (0.01)(100) + (-0.02)(760) = -4.7.
Probability of default is: $\frac{\exp(-4.7)}{1 + \exp(-4.7)}$ = **0.9%**.

Comment: Similar to 8, 11/12, Q.4a. Not intended as a realistic model.

3.82. The partial residual plot is not linear; thus, we should do something to improve the model.
We could group the variable X_1, converting it into a categorical variable.
We could add polynomial terms such X_1^2 to the model.
We could use piecewise linear functions such as: Min[0, X_1 + 1] and Min[0, X_1 - 1].

3.83. There are many ways to define the variables.
Let us define $X_1 = 1$ if low horsepower and zero otherwise.
$X_2 = 1$ if medium horsepower and zero otherwise.
$X_3 = 1$ if high horsepower and zero otherwise.
$X_4 = 1$ if sedan and zero otherwise.

For the Gamma Distribution, $f(y) = \theta^{-\alpha} y^{\alpha-1} e^{-y/\theta} / \Gamma(\alpha)$.
$\ln f(y) = (\alpha-1)\ln(y) - y/\theta - \alpha\ln(\theta) - \ln[\Gamma(\alpha)] = (\alpha-1)\ln(y) - y/(\mu/\alpha) - \alpha\ln(\mu/\alpha) - \ln[\Gamma(\alpha)]$
$= (\alpha-1)\ln(y) - \alpha y/\mu - \alpha\ln(\mu) + \alpha\ln(\alpha) - \ln[\Gamma(\alpha)]$.
a) With the identity link function: $\mu = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$.
Ignoring terms that do not involve the betas, the loglikelihood is:
$-\alpha 800/(\beta_1 + \beta_4) - \alpha \ln(\beta_1 + \beta_4) - \alpha 900/(\beta_2 + \beta_4) - \alpha \ln(\beta_2 + \beta_4) - \alpha 1100/(\beta_3 + \beta_4) - \alpha \ln(\beta_3 + \beta_4)$
$- \alpha 1500/(\beta_1) - \alpha \ln(\beta_1) - \alpha 1700/(\beta_2) - \alpha \ln(\beta_2) - \alpha 2000/(\beta_3) - \alpha \ln(\beta_3)$.
Setting the partial derivative with respect to β_1 equal to zero:
$800/(\beta_1 + \beta_4)^2 + 1500/\beta_1^2 = 1/(\beta_1 + \beta_4) + 1/\beta_1$.
Setting the partial derivative with respect to β_2 equal to zero:
$900/(\beta_2 + \beta_4)^2 + 1700/\beta_2^2 = 1/(\beta_2 + \beta_4) + 1/\beta_2$.
Setting the partial derivative with respect to β_3 equal to zero:
$1100/(\beta_3 + \beta_4)^2 + 2000/\beta_3^2 = 1/(\beta_3 + \beta_4) + 1/\beta_3$.
Setting the partial derivative with respect to β_4 equal to zero:
$800/(\beta_1 + \beta_4)^2 + 900/(\beta_2 + \beta_4)^2 + 1100/(\beta_3 + \beta_4)^2 = 1/(\beta_1 + \beta_4) + 1/(\beta_2 + \beta_4) + 1/(\beta_3 + \beta_4)$.

(b) We use a log link function. $\mu = \exp[\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4]$.
Ignoring terms that do not involve the betas, the loglikelihood is:
$-\alpha 800\exp[-\beta_1 - \beta_4] - \alpha(\beta_1+\beta_4) - \alpha 900\exp[-\beta_2 - \beta_4] - \alpha(\beta_2+\beta_4) - \alpha 1100\exp[-\beta_3 + \beta_4] - \alpha(\beta_3+\beta_4)$
$- \alpha 1500\exp[-\beta_1] - a(\beta_1) - \alpha 1700\exp[-\beta_2] - \alpha(\beta_2) - \alpha 2000\exp[-\beta_3] - \alpha(\beta_3)$.
Setting the partial derivative with respect to β_1 equal to zero:
$800\exp[-\beta_1 - \beta_4] + 1500\exp[-\beta_1] = 2$.
Setting the partial derivative with respect to β_2 equal to zero:
$900\exp[-\beta_2 - \beta_4] + 1700\exp[-\beta_2] = 2$.
Setting the partial derivative with respect to β_3 equal to zero:
$1100\exp[-\beta_3 - \beta_4] + 2000\exp[-\beta_3] = 2$.
Setting the partial derivative with respect to β_4 equal to zero:
$800\exp[-\beta_1 - \beta_4] + 900\exp[-\beta_2 - \beta_4] + 1100\exp[-\beta_3 - \beta_4] = 3$.
Comment: Using a computer, the fitted parameters in part (a) are:
$\beta_1 = 1567.71$, $\beta_2 = 1688.03$, $\beta_3 = 1914.41$, $\beta_4 = -784.60$.
The fitted severities are: 783.11, 903.43, 1129.81, 1567.71, 1688.03, 1914.41.
Using a computer, the fitted parameters in part (b) are:
$\beta_1 = 7.30933$, $\beta_2 = 7.43082$, $\beta_3 = 7.61246$, $\beta_4 = -0.620811$.
The fitted severities are: 803.13, 906.88, 1087.51, 1494.17, 1687.20, 2023.24.

3.84. Adding vehicle type adds 10 - 1 = 9 parameters to the model.

Test statistic is: $F = \dfrac{D_S - D_B}{(\text{number of added parameters})\,\hat{\phi}_S} = \dfrac{(1848.5 - 1833.0) / 9}{0.93} = 1.852.$

The number of degrees of freedom in the numerator is 9.
The number of degrees of freedom in the denominator is:
number of observations minus the number of parameters in the smaller model
= 2000 - 14 = 1986.
We compare the test statistic to an F-distribution with 9 and 1986 degrees of freedom.
The null hypothesis is to use the simpler model, the one without vehicle type.
The alternate hypothesis is to use the more complex model.
We reject the null hypothesis at 5% if the test statistic is bigger than the 5% critical value, which is where the F-distribution is 95%.
Comment: Using a computer, the p-value of this test is 5.5%.
Thus we would reject the null hypothesis at 5%.
If we reduced the number of vehicle type categories by combining some of the 10 categories we used, it might turn out that now we should use vehicle type at the 5% significance level.

3.85. Ignoring the loglikelihood of the saturated model, which is a constant,
BIC = Deviance + (number of parameters) ln[250].
For example, BIC = 1679.1 + 6 ln[250] = 1712.23.

Model	Number of Parameters	Deviance	BIC
A	6	1679.10	1712.23
B	8	1666.40	**1710.57**
C	10	1655.90	1711.11
D	12	1646.20	1712.46
E	14	1634.50	1711.80

Since BIC is smallest for model B, model B is preferred.

3.86. The difference between the yellow univariate line and the green GLM line, which better represents the underlying reality, arises from correlation between policy duration shown in the graph and the two other factors in the model.
Comment: One does not have to understand the life insurance details in order to answer the question asked.

3.87. exp[-0.3] - 1 = -25.9%.
Comment: See page 25 of Goldburd, Khare, and Tevet.
For a logistic model: Odds = μ / (1 - μ).

3.88. Female drivers age 31 to 59 in a rural territory have lower (process) variances than unmarried male drivers age 17 to 21 in an urban territory.
Therefore, the fitted model shifts to agree more closely with the observed values for the first group compared to the second group.
A GLM is more concerned with differences between observed and fitted where the (process) variances in observations are smaller. A GLM is less concerned with differences between observed and fitted where the variances in observations are larger.

3.89. The sensitivity is: 2000/5000 = 0.40.
The specificity is: 70,000 / 80,000 = 0.875.
For this threshold, we graph the point: (1 - specificity , sensitivity) = **(0.125, 0.40)**.

3.90. "Two standard errors from the parameter estimates are akin to a 95% confidence interval. This means the GLM parameter estimate is a point estimate, and the standard errors show the range in which the modeler can be 95% confident the true answer lies within."
Comment: See page 179 of Basic Ratemaking, on Exam 5.

3.91. The Lorenz curve for the rating plan is determined as follows:
1. Sort the dataset based on the model predicted loss cost.
2. On the x-axis, plot the cumulative percentage of exposures.
3. On the y-axis, plot the cumulative percentage of losses.
Draw a 45-degree line connecting (0, 0) and (1, 1), called the line of equality.
The Gini index is twice the area between the Lorenz curve and the line of equality.

3.92. $\ln(\lambda) = \beta_0 + \beta_1 z. \Rightarrow \lambda = \exp[\beta_0 + \beta_1 z]$.

For the Poisson Distribution: $f(y) = e^{-\lambda} \lambda^y / y!$.

$\ln f(y) = -\lambda + y\ln(\lambda) - \ln(y!) = -\exp[\beta_0 + \beta_1 z] + y(\beta_0 + \beta_1 z) - \ln(y!)$.

The loglikelihood is the sum of the contributions from the three observations:

$-\exp[\beta_0 + \beta_1] - \exp[\beta_0 + 2\beta_1] - \exp[\beta_0 + 3\beta_1] + 4(\beta_0 + \beta_1) + 7(\beta_0 + 2\beta_1) + 8(\beta_0 + 3\beta_1)$
$- \ln(4!) - \ln(7!) - \ln(8!)$.

To maximize the loglikelihood, we set its partial derivatives equal to zero.
Setting the partial derivative with respect to β_0 equal to zero:

$0 = -\exp[\beta_0 + \beta_1] - \exp[\beta_0 + 2\beta_1] - \exp[\beta_0 + 3\beta_1] + 19$.

Setting the partial derivative with respect to b_1 equal to zero:

$0 = -\exp[\beta_0 + \beta_1] - 2\exp[\beta_0 + 2\beta_1] - 3\exp[\beta_0 + 3\beta_1] + 42$.

Thus we have two equations in two unknowns:

$\exp[\beta_0 + \beta_1]\{1 + \exp[\beta_1] + \exp[2\beta_1]\} = 19$.

$\exp[\beta_0 + \beta_1]\{1 + 2\exp[\beta_1] + 3\exp[2\beta_1]\} = 42$.

Dividing the second equation by the first equation:

$\{1 + 2\exp[\beta_1] + 3\exp[2\beta_1]\}/\{1 + \exp[\beta_1] + \exp[2\beta_1]\} = 42/19$.

$\Rightarrow 19 + 38\exp[\beta_1] + 57\exp[2\beta_1] = 42 + 42\exp[\beta_1] + 42\exp[2\beta_1]$.

$\Rightarrow 15\exp[2\beta_1] - 4\exp[\beta_1] - 23 = 0$.

Letting $v = \exp[\beta_1]$, this equation is: $15v^2 - 4v - 23 = 0$, with positive solution:

$v = (4 + \sqrt{1396})/30 = 1.3788$.

$\exp[\beta_1] = 1.3788. \Rightarrow \beta_1 = 0.3212$.

$\Rightarrow \exp[\beta_0] = 19/\{\exp[\beta_1] + \exp[2\beta_1] + \exp[3\beta_1]\} = 19/\{1.3788 + 1.3788^2 + 1.3788^3\} = 3.2197$.

$\Rightarrow \beta_0 = 1.1693$.

$\lambda = \exp[\beta_0 + \beta_1 z] = \exp[\beta_0] \exp[\beta_1]^z = (3.2197)(1.3788^z)$.

For $z = 1$, $\lambda = 4.439$. For $z = 2$, $\lambda = 6.121$. For $z = 3$, $\lambda = 8.440$.
Comment: An ordinary linear regression fit to these same observations turns out to be:
$y = 2.333 + 2x$, with fitted values: 4.333, 6.333, and 8.333.

3.93. Examples include:
• Will it be cost-effective to collect the value of this variable when writing new and renewal business?
• Does inclusion of this variable in a rating plan conform to actuarial standards of practice and regulatory requirements?
• Can the electronic quotation system be easily modified to handle the inclusion of this variable in the rating formula?

3.94. a. We would have one parameter for gender, two parameters for age, and two parameters for territory. In addition we would have a parameter related to the base level.
A total of **6** parameters.
(2-1) + (3-1) + (3-1) + 1 = 6.
Sex Age Terr. Base
b. A total of **6** parameters. The link function does not affect the number of parameters.
c. β_0 is the intercept term that applies to all insureds.

β_1 corresponds to Female.

β_2 corresponds to Youthful.

β_3 corresponds to Retired.

β_4 corresponds to Suburban.

β_5 corresponds to Rural.

(There are many other possible orders for the parameters.)
d. With 6 parameters, the design matrix has **6** columns.
e. With 20,000 cars, the design matrix has **20,000** rows.
f. The number combinations are: (2)(3)(3) = 18. Thus the design matrix has **18** rows.
(I have assumed that none of these cells is empty.
I have assumed that there are no records with missing classification information.)

3.95. (a) The deviance residual for any given record is defined as that record's contribution to the deviance, adjusted for the sign of actual minus predicted; the deviance residual is taken to be negative where actual is less expected, and positive where actual is more than expected.
(Actually the deviance residual is the square root of the contribution to the deviance.)
(b) Intuitively, we can think of the deviance residual as the residual adjusted for the shape of the assumed GLM distribution, such that its distribution will be approximately normal if the assumed GLM distribution is correct.
(c) In a well-fit model, we expect deviance residuals to follow no predictable pattern, and be normally distributed, with constant variance.
One could plot the deviance residuals versus the fitted values or versus an important predictor variable, in order to see whether there is a pattern.
We can check for the normality of the deviance residuals via either a histogram or q-q plot.

3.96. $p/(1-p) = \exp[\beta_0 + \beta_1 X]$. $\Rightarrow 1/p - 1 = \exp[-\beta_0 - \beta_1 X]$. $\Rightarrow p = 1/(1 + \exp[-\beta_0 - \beta_1 X])$.

$\Rightarrow 1 - p = \exp[-\beta_0 - \beta_1 X] / (1 + \exp[-\beta_0 - \beta_1 X]) = 1 / (1 + \exp[\beta_0 + \beta_1 X])$.

For a Binomial with parameters m and p, $f(n) = p^n (1-p)^{m-n} m! / \{(n!)(m-n)!\}$.

$\ln f(n) = n \ln p + (m-n)\ln(1-p) + \ln(m!) - \ln(n!) - \ln[(m-n)!] = n \ln[p/(1-p)] + m \ln(1-p) + \text{constants} = n(\beta_0 + \beta_1 X) - m \ln[(1 + \exp[\beta_0 + \beta_1 X])] + \text{constants}$.

loglikelihood $= \sum n_i(\beta_0 + \beta_1 X_i) - \sum m_i \ln[(1 + \exp[\beta_0 + \beta_1 X_i])] + \text{constants}$.

Setting the partial derivatives of the loglikelihood with respect to β_0 and β_1 equal to zero:

$0 = \sum n_i - \sum m_i \exp[\beta_0 + \beta_1 X_i]/(1 + \exp[\beta_0 + \beta_1 X_i])$.

$0 = \sum n_i X_i - \sum m_i X_i \exp[\beta_0 + \beta_1 X_i]/(1 + \exp[\beta_0 + \beta_1 X_i])$.

$\sum n_i = 900 + 820 + 740 + 660 + 580 = 3700$.

$\sum n_i X_i = (1)(900) + (2)(820) + (3)(740) + (4)(660) + (5)(580) = 10{,}300$.

The first equation becomes:

$3700 = 1000/(1 + \exp[-\beta_0 - \beta_1]) + 900/(1 + \exp[-\beta_0 - 2\beta_1]) + 800/(1 + \exp[-\beta_0 - 3\beta_1])$
$+ 700/(1 + \exp[-\beta_0 - 4\beta_1]) + 600/(1 + \exp[-\beta_0 - 5\beta_1])$.

The second equation becomes:

$10300 = 1000/(1 + \exp[-\beta_0 - \beta_1]) + 1800/(1 + \exp[-\beta_0 - 2\beta_1]) + 2400/(1 + \exp[-\beta_0 - 3\beta_1])$
$+ 2800/(1 + \exp[-\beta_0 - 4\beta_1]) + 3000/(1 + \exp[-\beta_0 - 5\beta_1])$.

Comment: An example of a Logistic Regression.

Using a computer, the maximum likelihood fit is: $\beta_0 = 1.88543$ and $\beta_1 = 0.245509$.

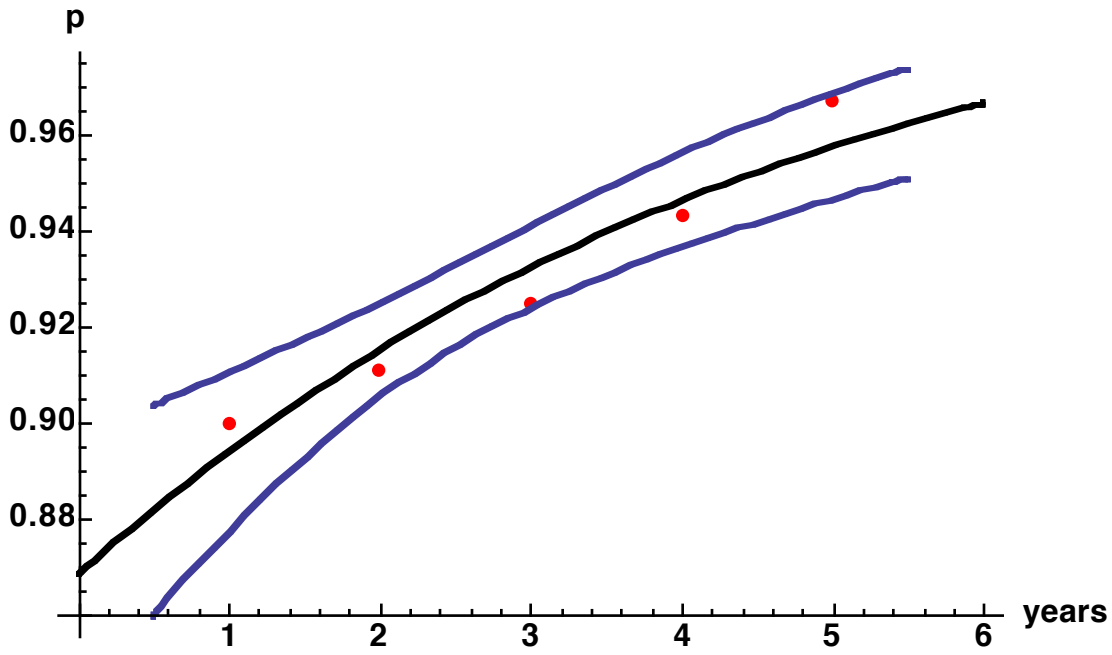
The covariance matrix of the fitted parameters is:

	β_0	β_1
β_0	0.0154836	-0.00501396
β_1	-0.00501396	0.00212092

Thus the standard error of β_0 is: $\sqrt{0.0154836} = 0.1244$,

and the standard error of β_1 is: $\sqrt{0.00212092} = 0.04605$.

Here is a graph of the data, the fitted curve, and 95% confidence intervals:



3.97. ϕ is the dispersion parameter, which scales the variance.
ω_i is a (prior) weight, representing the amount of data we have for observation i; the variance is inversely proportional to the volume of data.

3.98. & 3.99. $f(y) = \exp[-(y - \mu)^2/(2\sigma^2)] / \{\sigma\sqrt{2\pi}\}$.

$\ln f(Y_i) = -(Y_i - \beta_0 - \beta_1 X_i)^2/(2\sigma^2) - \ln(\sigma) - \ln(2\pi)/2$.

Loglikelihood is: $-\sum(Y_i - \beta_0 - \beta_1 X_i)^2/(2\sigma^2) - n \ln(\sigma) - n \ln(2\pi)/2$.

Set the partial derivative of the loglikelihood with respect to β_0 equal to zero:

$0 = \sum(Y_i - \beta_0 - \beta_1 X_i)/\sigma^2. \Rightarrow \sum Y_i = n\beta_0 + \beta_1 \sum X_i. \Rightarrow \beta_0 = \bar{Y} - \beta_1 \bar{X}$.

Set the partial derivative of the loglikelihood with respect to β_1 equal to zero:

$0 = \sum X_i(Y_i - \beta_0 - \beta_1 X_i)/\sigma^2. \Rightarrow \sum X_i Y_i = \beta_0 \sum X_i + \beta_1 \sum X_i^2. \Rightarrow \sum X_i Y_i = (\bar{Y} - \beta_1 \bar{X}) \sum X_i + \beta_1 \sum X_i^2$.

$\Rightarrow \hat{\beta}_1 = \{\sum X_i Y_i - \bar{Y} \sum X_i\} / \{\sum X_i^2 - \bar{X} \sum X_i\} = \{255 - (10)(24)\} / \{174 - (6)(24)\} = 15/30 = \mathbf{0.5}$.

$\Rightarrow \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 10 - (0.5)(6) = \mathbf{7}$.

Comment: Matches the linear regression model with an intercept.
For example, in deviations form:
$\bar{X} = 24/4 = 6$. $x = X - \bar{X} = -4, -1, 2, 3$. $\bar{Y} = 40/4 = 10$. $y = Y - \bar{Y} = 0, -4, 1, 3$.

$\hat{\beta} = \sum x_i y_i / \sum x_i^2 = 15/30 = 0.5$. $\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} = 10 - (0.5)(6) = 7$.

3.100. Set the partial derivative of the loglikelihood with respect to σ equal to zero:
$0 = \sum(Y_i - \beta_0 - \beta_1 X_i)^2/\sigma^3 - n/\sigma$. $\Rightarrow \sigma^2 = \sum(Y_i - \beta_0 - \beta_1 X_i)^2/n = \sum(Y_i - 7 - (0.5)X_i)^2/4 =$

$$\frac{\{10 - 7 - (0.5)(2)\}^2 + \{6 - 7 - (0.5)(5)\}^2 + \{11 - 7 - (0.5)(8)\}^2 + \{13 - 7 - (0.5)(9)\}^2}{4} = 18.5/4 = 4.625.$$

$\Rightarrow \hat{\sigma} = \sqrt{4.625} =$ **2.15**.

3.101. & 3.102. Let $x = X - \bar{X} = -4, -1, 2, 3$, and $y = Y - \bar{Y} = 0, -4, 1, 3$.
Then, $\sum X_i Y_i - \bar{Y}\sum X_i = \sum X_j Y_j - \sum Y_j \sum X_i/n = \sum Y_j(X_j - \bar{X}) = \sum Y_j x_j$.
Also, $\sum X_i^2 - \bar{X}\sum X_i = \sum X_i(X_i - \bar{X}) = \sum X_i x_i = \sum (X_i - \bar{X})x_i + \sum \bar{X} x_i = \sum x_i^2 + \bar{X} \sum x_i = \sum x_i^2 + \bar{X}(0)$
$= \sum x_i^2$.
$\hat{\beta}_1 = \{\sum X_i Y_i - \bar{Y}\sum X_i\} / \{\sum X_i^2 - \bar{X}\sum X_i\} = \sum Y_i x_i / \sum x_i^2$.
$\text{Var}[\hat{\beta}_1] = \text{Var}[\sum Y_i x_i / \sum x_i^2] = \sum \text{Var}[Y_i x_i]/\{\sum x_i^2\}^2 = \sum x_i^2 \text{Var}[Y_i] / \{\sum x_i^2\}^2 = \sigma^2 \sum x_i^2/\{\sum x_i^2\}^2 = \sigma^2/\sum x_i^2$
$= 4.625/30 = 0.1542$. $\text{StdDev}[\hat{\beta}_1] = \sqrt{0.1542} =$ **0.393**.
$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = (Y_1 + Y_2 + Y_3 + Y_4)/4 - (\sum Y_i x_i / \sum x_i^2)(6) =$
$(Y_1 + Y_2 + Y_3 + Y_4)/4 - (-4Y_1 - Y_2 + 2Y_3 + 3Y_4)(6/30) = 1.05Y_1 + 0.45Y_2 - 0.15Y_3 - 0.35Y_4$.
Recalling that the Y_i are independent and each have variance σ^2:
$\text{Var}[\hat{\beta}_0] = \sigma^2(1.05^2 + 0.45^2 + 0.15^2 + 0.35^2) = 1.45\sigma^2 = (1.45)(4.625) = 6.706$.
$\text{StdDev}[\hat{\beta}_0] = \sqrt{6.706} =$ **2.59**.
Comment: Beyond what you should be asked on your exam.

One can show in general that $\text{Var}[\hat{\beta}] = \sigma^2 / \sum x_i^2$ and $\text{Var}[\hat{\alpha}] = \sigma^2 \sum X_i^2 / (N\sum x_i^2)$.
While the maximum likelihood results are similar, they do not match linear regression:

$\hat{Y} = \hat{\alpha} + \hat{\beta} X = 8, 9.5, 11, 11.5$. $\hat{\varepsilon} = Y - \hat{Y} = 2, -3.5, 0, 1.5$. $\text{ESS} = \sum \hat{\varepsilon}_i^2 = 18.5$.

$s^2 = \text{ESS} / (N - 2) = 18.5 / (4 - 2) = 9.25$.
$\text{Var}[\hat{\beta}] = s^2 / \sum x_i^2 = 9.25/30 = 0.3083$. $s_{\hat{\beta}} = \sqrt{0.3083} = 0.555$.

$\text{Var}[\hat{\alpha}] = s^2 \sum X_i^2 / (N\sum x_i^2) = (9.25)(174) / ((4)(30)) = 13.41$. $s_{\hat{\alpha}} = \sqrt{13.41} = 3.66$.

3.103. Since they have all been fit to the same data and have the same number of parameters, the model with the smallest Deviance is best.
Comment: If we were to apply either AIC or BIC, in this case the ranks of the models would be the same as that of their deviances.

3.104. When the variance is greater than the mean we can use an overdispersed Poisson with $\phi > 1$.
$\text{Var}[Y_i] = \phi\, E[Y_i]$. For $\phi > 1$, variance is greater than the mean. While this does not correspond to the likelihood of any exponential family, otherwise the GLM mathematics works.
Using an overdispersed Poisson (ODP), we get the same estimated betas as for the usual Poisson regression. However, the standard errors of all of the estimated parameters are multiplied by $\sqrt{\phi}$.

Comment: When the variance is greater than the mean, one could use a Negative Binomial Distribution, which has a variance greater than its mean.
Often the results of using an overdispersed Poisson and a Negative Binomial will be similar.

3.105. While a 5% probability value may seem small, it allows for a 1-in-20 chance of a variable being accepted as significant when it is not. Since in a typical insurance modeling project we are testing many variables, this threshold may be too high to protect against the possibility of spurious effects making it into the model.
For example, if we are testing the potential usefulness of 40 possible predictor variables, then if we use a p-value of 5%, even if none of the variables actually predict the outcome, on average two of these 40 variables will be selected as significant.
Comment: See page 9 of "Generalized Linear Models for Insurance Rating".
"Spurious correlations exist when the historical correlation between two variables is random or coincidental. In these cases, one variable cannot reliably be used to inform a projection of the other variable going forward. For example, over the past year the number of California Department of Insurance rate regulation actuaries has increased, as has California average rainfall. Unfortunately, however, we cannot expect to influence future California rainfall by hiring additional actuaries."
Quoted from "Predictive Analytics: Regulatory Review" by Rachel Hemphill
in the AAA Casualty Quarterly, Summer 2017.

3.106. The partial residual plot seems linear; thus, no action is indicated.

3.107. A potential problem may exist where two or more predictors in a model may be strongly predictive of a third, a situation known as multicollinearity. Instability problems may result, since the information contained in the third variable is also present in the model in the form of the combination of the other two variables. However, the variable may not be highly correlated with either of the other two predictors individually, and so this effect will not show up in a correlation matrix, making it more difficult to detect.

3.108. Territories are not a good fit for the GLM framework.
One should include the territory relativities produced by the separate model as an offset in the GLM used to determine classification relativities. Similarly, one should include classification relativities produced by the GLM as an offset in the separate model used to determine territory relativities.

3.109. A hold-out sample is data that was not used in the development of the model so that it could be used to test the effectiveness of the model. (This could either be a random sample of the original data, or an additional year of data.) One compares the expected outcome of the model with results on the hold-out sample. The extent to which the model results track closely to results on the hold-out sample for a large part of the portfolio is an indication of how well the model validates.

3.110. You can use age groups, but probably want to group fewer ages together for the younger ages. (Unfortunately, the volume of data is smaller for the very youngest ages, so there is a trade-off between homogeneity and credibility.) For ages above about 25, the affect of gender is relatively small and similar by age. In contrast, for younger ages the affect of gender is large and differs by age. Thus a simple multiplicative model with a single relativity for male compared to female will not work. One would need to have a gender relativity that varied by age. (This may be possible to accomplish this by having an interaction term in the GLM.)

3.111. (a) For $z_1 = 1$ and $z_2 = 30$, renewal probability is:

$$\frac{\text{Exp}[0.6 + (0.05)(1) + (0.02)(30)]}{1 + \text{Exp}[0.6 + (0.05)(1) + (0.02)(30)]} = 0.7773.$$

For $z_1 = 10$ and $z_2 = 30$, renewal probability is:

$$\frac{\text{Exp}[0.6 + (0.05)(10) + (0.02)(30)]}{1 + \text{Exp}[0.6 + (0.05)(10) + (0.02)(30)]} = 0.8455.$$

0.7773 / 0.8455 = **0.919**.
(b) For $z_1 = 1$ and $z_2 = 50$, renewal probability is:

$$\frac{\text{Exp}[0.6 + (0.05)(1) + (0.02)(50)]}{1 + \text{Exp}[0.6 + (0.05)(1) + (0.02)(50)]} = 0.8389.$$

For $z_1 = 10$ and $z_2 = 50$, renewal probability is:

$$\frac{\text{Exp}[0.6 + (0.05)(10) + (0.02)(50)]}{1 + \text{Exp}[0.6 + (0.05)(10) + (0.02)(50)]} = 0.8909.$$

0.8389 / 0.8909 = **0.942**.
Comment: Not intended as a realistic model of policy renewal.
In general for a particular GLM, the relativities for one predictor variable can depend on the level (s) of the other predictor variable(s).
This model was based on the logit link function. If instead the log link function had been used, the model would have been multiplicative, and the indicated multiplicative relativities would not have depended on the other predictor variable. If instead the identity link function had been used, the model would have been additive, and the indicated additive relativities would not have depended on the other predictor variable.

3.112. The deviance residuals seem to decrease on average with X_3.
The lack of independence of the deviance residuals and X_3 is not good.
One should investigate refining the model.

3.113. Let X_0 correspond to the constant term.
Let X_1 be 1 if there is child. Let X_2 be the years of education.

a. $X = \begin{pmatrix} 1 & 0 & 12 \\ 1 & 0 & 14 \\ 1 & 0 & 15 \\ 1 & 0 & 16 \\ 1 & 0 & 17 \\ 1 & 1 & 10 \\ 1 & 1 & 11 \\ 1 & 1 & 13 \\ 1 & 1 & 15 \\ 1 & 1 & 16 \end{pmatrix}$ $Y = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$

b. $p/(1-p) = \exp[\beta_0 + \beta_1 X_1 + \beta_2 X_2]. \Rightarrow 1/p - 1 = \exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2)].$

$\Rightarrow p = 1/ (1 + \exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2)]).$

$\Rightarrow 1 - p = \exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2)] / (1 + \exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2)])$

$= 1 / (1 + \exp[\beta_0 + \beta_1 X_1 + \beta_2 X_2]).$

For a Bernoulli (yes/no) with parameter p, $f(y) = p^y (1-p)^{1-y}$.
$\ln f(y) = y \ln p + (1-y)\ln(1-p) = y \ln[p/(1-p)] + \ln(1-p) =$
$y(\beta_0 + \beta_1 X_1 + \beta_2 X_2) - \ln[1 + \exp[\beta_0 + \beta_1 X_1 + \beta_2 X_2]].$

loglikelihood $= \sum y_i(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}) - \sum \ln[1 + \exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{i2}]].$
Setting the partial derivatives of the loglikelihood with respect to β_0, β_1, and β_2 equal to zero:

$0 = \sum y_i - \sum \exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}]/(1 + \exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}]).$

$0 = \sum y_i X_{1i} - \sum X_{1i} \exp[\beta_0 + \beta_1 X_i]/(1 + \exp[\beta_0 + \beta_1 X_{1i} + \beta b_2 X_{2i}]).$

$0 = \sum y_i X_{2i} - \sum X_{2i} \exp[\beta_0 + \beta_1 X_i]/(1 + \exp[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}]).$

$\sum y_i = 1 + 0 + 1 + 0 + 1 + 0 + 0 + 1 + 0 + 1 = 5.$

$\sum y_i X_{1i} = 2.$

$\sum y_i X_{2i} = 12 + 15 + 17 + 13 + 16 = 73.$

The first equation becomes:

$$5 = 1/(1 + \exp[-\beta_0 - 12\beta_2]) + 1/(1 + \exp[-\beta_0 - 14\beta_2]) + 1/(1 + \exp[-\beta_0 - 15\beta_2])$$
$$+ 1/(1 + \exp[-\beta_0 - 16\beta_2]) + 1/(1 + \exp[-\beta_0 - 17\beta_2]) + 1/(1 + \exp[-\beta_0 - \beta_1 - 10\beta_2])$$
$$+ 1/(1 + \exp[-\beta_0 - \beta_1 - 11\beta_2]) + 1/(1 + \exp[-\beta_0 - \beta_1 - 13\beta_2]) + 1/(1 + \exp[-\beta_0 - \beta_1 - 15\beta_2])$$
$$+ 1/(1 + \exp[-\beta_0 - \beta_1 - 16\beta_2]).$$

The second equation becomes:

$$2 = 1/(1 + \exp[-\beta_0 - \beta_1 - 10\beta_2]) + 1/(1 + \exp[-\beta_0 - \beta_1 - 11\beta_2]) + 1/(1 + \exp[-\beta_0 - \beta_1 - 13\beta_2])$$
$$+ 1/(1 + \exp[-\beta_0 - \beta_1 - 15\beta_2]) + 1/(1 + \exp[-\beta_0 - \beta_1 - 16\beta_2]).$$

The third equation becomes:

$$73 = 12/(1 + \exp[-\beta_0 - 12\beta_2]) + 14/(1 + \exp[-\beta_0 - 14\beta_2]) + 15/(1 + \exp[-\beta_0 - 15\beta_2])$$
$$+ 16/(1 + \exp[-\beta_0 - 16\beta_2]) + 17/(1 + \exp[-\beta_0 - 17\beta_2]) + 10/(1 + \exp[-\beta_0 - \beta_1 - 10\beta_2])$$
$$+ 11/(1 + \exp[-\beta_0 - \beta_1 - 11\beta_2]) + 13/(1 + \exp[-\beta_0 - \beta_1 - 13\beta_2])$$
$$+ 15/(1 + \exp[-\beta_0 - \beta_1 - 15\beta_2]) + 16/(1 + \exp[-\beta_0 - \beta_1 - 16\beta_2]).$$

Comment: In a practical application, one would have at least several hundred data points.
Using a computer, the fitted parameters are:
$\beta_0 = -3.65238$, $\beta_1 = -0.373673$, $\beta_2 = 0.275467$.
The fitted probabilities of workplace participation are:
0.4142, 0.5509, 0.6177, 0.6803, 0.7370, 0.2190, 0.2697, 0.3906, 0.5265, 0.5942.
For example, with a child and 10 years of education, the estimated probability of participating in the workforce is:

$$\frac{\exp[-3.65238 - (1)(0.373673) + (10)(0.275467)]}{1 + \exp[-3.65238 - (1)(0.373673) + (10)(0.275467)]} = \frac{\exp[-1.271383]}{1 + \exp[-1.271383]} = 21.90\%.$$

3.114. 1. The variables to be considered.
2. The distributional form of the errors.
3. The link function.
4. Whether he is modeling frequency, severity, or pure premium.
5. Whether he will be modeling all of the perils together or he will be modeling one of the major perils separately.

Comment: There are probably other good answers.

3.115. 1. Predictive accuracy: for the right panel graph, the plotted loss costs correspond more closely between the two lines than for the left panel graph, indicating that the proposed model seems to predict actual loss costs better than the current rating plan does.
2. Monotonicity: the current plan has a reversal in the 6th decile, whereas the model has no significant reversals.
3. Vertical distance between the first and last quantiles: the spread of actual loss costs for the current plan is about 0.6 to 1.2, which is not very much. The spread of the proposed model is larger.
Thus, by all three metrics, the new plan outperforms the current one.
Comment: Graphs taken from "Introduction to Predictive Modeling Using GLMs A Practitioner's Viewpoint," a presentation by Dan Tevet and Anand Khare.

3.116. Modeling personal auto probability of policy renewal.
Modeling fraud on claims.
Comment: Many other possible answers.

3.117. exp[8.8 + (-0.03)(30) - 0.15] = **2322**.

3.118. mean = exp[8.8 + (-0.03)(40)] = 1998. Variance = ϕ mean^2 = (0.3)(1998^2) = **1,197,601**

3.119. Approximately 95% of the time the actual relativity should be within the bands two standard errors on either side of the parameter estimate.
In the first graph the bands are relatively narrow. Also the relativities display an increase with vehicle symbol, which makes sense. Vehicle symbol appears to be a significant factor for the first model.
In the second graph, the bands are wide. Also the relativities display no consistent pattern with vehicle symbol. Vehicle symbol does not appear to be a significant factor for the second model. There are no parameter estimates more than two standard errors from zero.
In other words, the results are consistent with a multiplicative relativity of one for all symbols.
Comment: The graphs are taken from "A Practitioner's Guide to Generalized Linear Models," by Duncan Anderson, Sholom Feldblum, Claudine Modlin, Doris Schirmacher, Ernesto Schirmacher, and Neeza Thandi. We note that in the first graph the one-way (univariate analysis) comes up with different relativities than the GLM, presumably because vehicle symbol is correlated with other significant predictor variables in the GLM. The bottom righthand of the original of the first graph shows a p-value of 0%, indicating that vehicle symbol is significant. The original of the second graph shows a p-value of 52.5% indicating that vehicle symbol is not significant in this second model.

3.120. One way to combine separate models by peril in order to get a model for all perils:
1. Use the separate models by peril to generate predictions of expected loss due to each peril for some set of exposure data.
2. Add the peril predictions together to form a combined loss cost for each record.
3. Run a model on that data, using the combined loss cost calculated in Step 2 as the target, and the union of all the individual model predictors as the predictor variables.

3.121. (a) $\mu = \text{Exp}[\alpha_i + \beta x] = \text{Exp}[\alpha_i]\ \text{Exp}[\beta x]$.
This is a multiplicative model, with relativities for gender and relativities for age.
The age relativities are the same for males and females.
If $\beta < 0$, then the relative frequencies decline exponentially with age.
(b) $\mu = \text{Exp}[\alpha_i + \beta_i x] = \text{Exp}[\alpha_i]\ \text{Exp}[\beta_i x]$.
Similar to the previous model, except now the age relativities differ by gender.
For example, the relativity for age 20 relative to age 30 is:
$\text{Exp}[20\beta_i] / \text{Exp}[30\beta_i] = \text{Exp}[-10\beta_i]$, which differs by gender.
(If $\beta_1 = \beta_2$, then this reduces to the previous model.)
Comment: Even for $\beta_i < 0$, this is not a realistic model of expected claim frequencies by driver age. Instead one would group the ages into for example, 17-20, 21-24, etc., and treat the age groups as categorical variables.

3.122. D. Histogram D most closely matches the Normal Distribution.

3.123. The results of a GLM depend on the choice of link functions. So perhaps the two models have different link functions. The results of a GLM depend on the choice of predictor variables. So perhaps the two models have different sets of predictor variables other than driver age.
The results of a GLM depend on the choice of the assumed distributional form of the errors. So perhaps the two models have different distributional forms of their errors.
Comment: Usually the actuary analyzes the relativities for driver age assuming all of the other predictor variables in the GLM are at the base level. If one varies the levels of the other predictor variables in the GLM, then relativities between driver ages will also usually vary.

3.124. Plot $(\Phi^{-1}[i/37)], x_{(i)})$.
$Q_{9/37} = Q_{0.243} = -0.696$, since $\Phi[-0.696] = 0.243$.
Thus the plotted point is: **(-0.696, 0.004)**.

3.125. A useful statistic for detecting multicollinearity is the variance inflation factor (VIF). The VIF for any predictor is a measure of how much the squared standard error for the predictor is increased due to the presence of collinearity with other predictors. It is determined by running a linear model for each of the predictors using all the other predictors as inputs, and measuring the predictive power of those models.
A common statistical rule of thumb is that a VIF greater than 10 is considered high. However, where large VIFs are indicated, it is important to look deeper into the collinearity structure in order to make an informed decision about how best to handle it in the model.

3.126. The new categorical variable has five categories, so adds 4 degrees of freedom.

Test statistic is: $F = \frac{D_S - D_B}{(\text{number of added parameters})\ \hat{\phi}_S} = \frac{(2196.1 - 2179.3) / 4}{2.09} = 2.010.$

The number of degrees of freedom in the numerator is 4.
The number of degrees of freedom in the denominator is:
number of observations minus the number of parameters in the smaller model.
We compare the test statistic to an F-distribution.
We reject the null hypothesis if the test statistic is big.

3.127. Cross Validation is another technique for data splitting.
Split the data into for example 10 groups. Each group is called a fold. For each fold:
• Train the model using the other folds.
• Test the model using the given fold.
Several models can be compared by running the procedure for each of them on the same set of folds and comparing their relative performances for each fold.
However, cross validation is often of limited usefulness for most insurance modeling applications.
Using cross validation in place of a holdout set is only appropriate where a purely automated variable selection process is used.
Comment: See Section 4.3.4 of Goldburd, Khare, and Tevet.
Purely automated variable selection processes should be used with appropriate caution.

3.128. A common statistical rule of thumb is that a VIF greater than 10 is considered high.
Thus, there is probably multicollinearity related to Weight; two or more predictors in the model are probably strongly predictive of Weight. This may cause instability problems with the model.
This situation should be investigated further.
It may help to either remove Weight from the model or to preprocess the data using dimensionality reduction techniques such as principal components analysis.
Comment: The VIF of 6.33 for Body Surface Area may also warrant some investigation.

3.129. In the first graph for liability losses, the number of children seems to have a significant impact on frequency. The 95% confidence intervals do not include a log of the multiplier of 0; in other words the multiplier is significantly different from one. Also while one child increases the frequency compared to none, two children also increase the frequency compared to one. It seems as if the number of children in the household is a useful variable for modeling liability frequency for Homeowners.
In the second graph for wind losses, the number of children seems to have a insignificant impact on frequency. The 95% confidence intervals do include a log of the multiplier of 0; in other words the multiplier is not significantly different from one. Also while one child increases the frequency compared to none, two children decreases the frequency compared to one. The number of children in the household is not a useful variable for modeling wind frequency for Homeowners.
Comment: There is no logical relationship between the number of children and wind losses.
A child (or any relative) who lives in the house is covered for any liability claim he or she causes. Also having children in the house may lead to more neighborhood children coming on your property with the potential for liability claims if they are injured on your property. Thus there is some logical relationship between the number of children in the household and the frequency of liability claims for Homeowners.
Presumably, the liability relativity for three children would be higher than for two children. (Three children was not shown in the graph in order to keep things simple.)
One would want to apply statistical tests to see if the number of children in the household is a useful variable for modeling liability frequency. Also one would want to check the consistency over time of the indicated relativities.

3.130. With a Normal error function and an identity link function, this is the same as a multiple regression. The squared error is:

$$800 (\beta_1 + \beta_2 + \beta_3 - 700{,}000/800)^2 + 600 (\beta_2 + \beta_3 - 400{,}000/600)^2$$
$$+ 700 (\beta_1 + \beta_3 - 500{,}000/700)^2 + 500 (\beta_3 - 300{,}000/500)^2.$$

We are given that $\beta_3 = 570.356$, thus the squared error is:

$800 (\beta_1 + \beta_2 - 304.644)^2 + 600 (\beta_2 - 96.311)^2 + 700 (\beta_1 - 143.930)^2 + 500 (-29.644)^2$.

Setting the partial derivative with respect to β_1 equal to zero:

$0 = 1600 (\beta_1 + \beta_2 - 304.644) + 1400 (\beta_1 - 143.930). \Rightarrow 3000 \beta_1 + 1600 \beta_2 = 688{,}932.$

Setting the partial derivative with respect to β_2 equal to zero:

$0 = 1600 (\beta_1 + \beta_2 - 304.644) + 1200 (\beta_2 - 96.311). \Rightarrow 1600 \beta_1 + 2800 \beta_2 = 603{,}004.$

$\Rightarrow \beta_2 = (603{,}004 - 1600 \beta_1) / 2800.$

Plugging back into the first equation: $3000 \beta_1 + 1600 (603{,}004 - 1600 \beta_1) / 2800 = 688{,}932.$

$$\Rightarrow \beta_1 = \frac{(2800)(688{,}932) - (1600)(603{,}004)}{(3000)(2800) - (1600)(1600)} = 964{,}203{,}200 / 5{,}840{,}00 = \mathbf{165.103}.$$

$\Rightarrow \beta_2 = 121.014.$

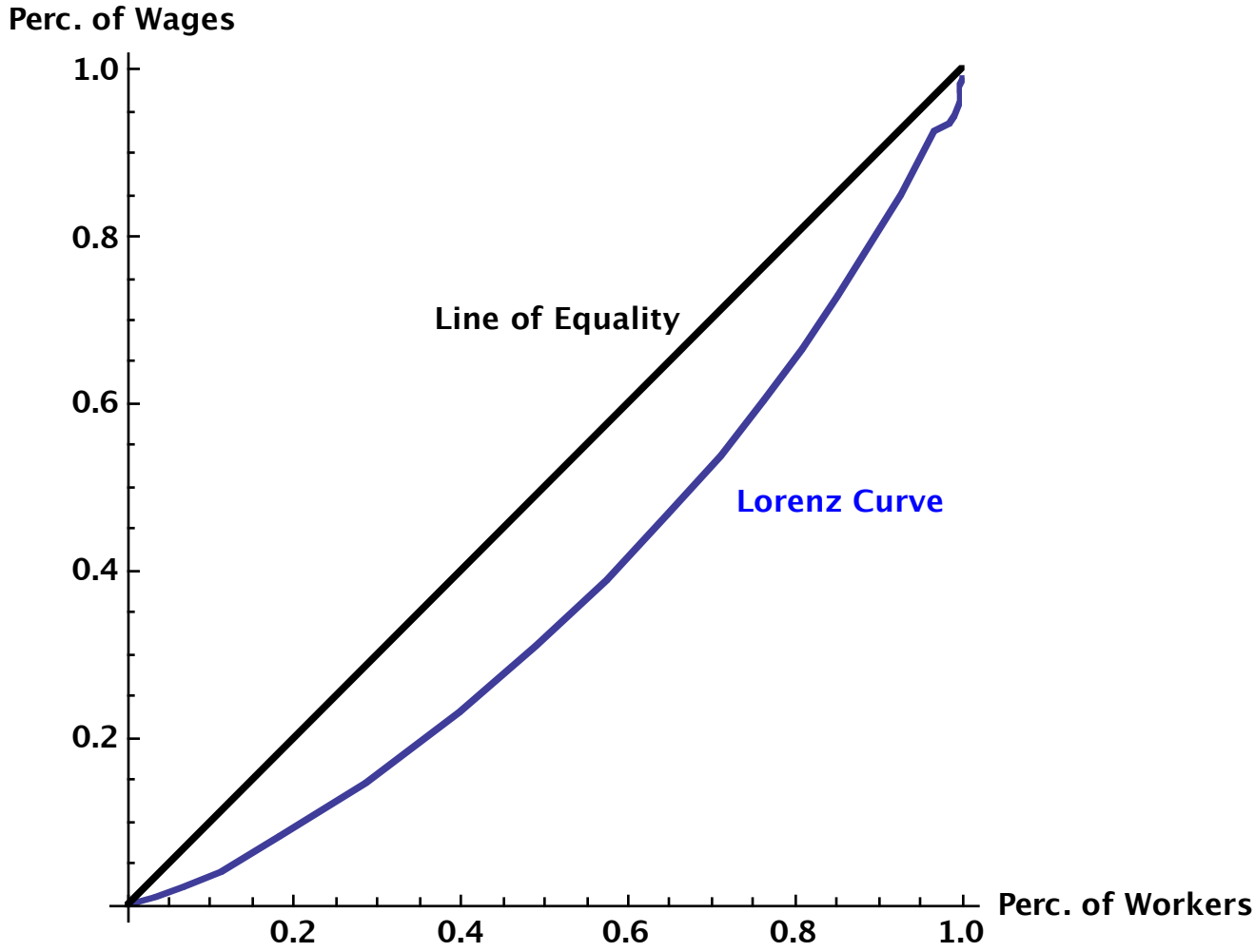
3.131. BIC = (-2) (maximum loglikelihood) + (number of parameters)ln[60].
For example, BIC = (-2)(-220.18) + 2 ln[60] = 448.55.

Model	Number of Parameters	Loglikelihood	BIC
A	2	-220.18	448.55
B	3	-217.40	447.08
C	4	-214.92	**446.22**
D	5	-213.25	446.97
E	6	-211.03	454.81

Since BIC is smallest for model C, model C is preferred.

3.132. One should also perform a statistical test to compare a model with year to a simpler model without year.
Before excluding year as a variable, it would be better to first try a model where you group the years into fewer categories, for example: 2010-2011, 2012, 2013-2014.
(We may not have enough data from each year in order to be statistically confident of separate coefficients by year.)
Then if after fitting the new model the revised coefficients for years are still not significant, one could exclude year from the model.
Comment: The actuary would want to determine whether the pattern between years of the fitted coefficients makes any sense to him given his knowledge of the situation being modeled.
Statistical tests are important, but just one tool. Actuarial judgement is also important.

3.133. The third plotted point is: (0.0353, 0.0079).
The last plotted point is: (0.9997, 0.9884).



Comment: In my graph, I have had the computer join the plotted points.
Information was taken from the 1998 Massachusetts Wage Distribution Table.
The Gini index is twice the area between the Lorenz Curve and the Line of Equality.

3.134. Simple quantile plots are created via the following steps:
1. Sort the dataset based on the Model A predicted loss cost (from smallest to largest).
2. Bucket the data into quantiles, such that each quantile has the same volume of exposures. Common choices are quintiles (5 buckets), deciles (10 buckets), or vigintiles (20 buckets).
3. Within each bucket, calculate the average predicted pure premium (predicted loss per unit of exposure) based on the Model A predicted loss cost, and calculate the average actual pure premium.
4. Plot, for each quantile, the actual pure premium and the pure premium predicted by Model A.
5. Repeat steps 1 through 4 using the Model B predicted loss costs.
There are now two quantile plot; one for Model A and one for Model B.
6. Compare the two quantile plots to determine which model provides better lift.
In order to determine the "winning" model, consider the following 3 criteria:
1. **Predictive accuracy.** How well each model is able to predict the actual pure premium in each quantile.
2. **Monotonicity.** By definition, the predicted pure premium will monotonically increase as the quantile increases, but the actual pure premium should also increase (though small reversals are okay).
3. **Vertical distance between the first and last quantiles.** The first quantile contains the risks that the model believes will have the best experience, and the last quantile contains the risks that the model believes will have the worst experience. A large difference (also called "lift") between the actual pure premium in the quantiles with the smallest and largest predicted loss costs indicates that the model is able to maximally distinguish the best and worst risks.
Comment: See Section 7.2.1 of Goldburd, Khare, and Tevet.

3.135. (a) For example, using a discrimination threshold of 25%, one would be predicting fraud for any claim for which the GLM says the probability of fraud is greater than 25%.
Alternately, choose a specific probability level, called the discrimination threshold, above which we will investigate the claim for fraud and below which we will not. This determination may be thought of as the model's "prediction" in a binary (i.e., fraud/no fraud) sense.
(b) Using a lower threshold would detect more of the fraudulent claims, at the cost of also having to investigate more claims which turned out not to be fraudulent. Using a higher threshold would detect fewer of the fraudulent claims, but we would have to investigate fewer claims which turned out not to be fraudulent.
Alternately, there is trade-off: a lower threshold results in a higher sensitivity (true positive rate), while a higher threshold results in a higher specificity (and thus a lower false positive rate).
Comment: Similar to 8, 11/17, Q.6d.
See pages 75 to 77 of Generalized Linear Models for Insurance Rating.
"The selection of a discrimination threshold involves a trade-off: a lower threshold will result in more true positives and fewer false negatives than a higher threshold, but at the cost of more false positives and fewer true negatives."

3.136. The mean modeled claim counts are:

	Terr. A	Terr. B
Male	24,000 exp[β_0]	15,000 exp[$\beta_0 + \beta_2$]
Female	20,000 exp[$\beta_0 + \beta_1$]	13,000 exp[$\beta_0 + \beta_1 + \beta_2$]

The likelihood function of a Poisson is : $\sum \ln f(y_i; \mu_i) = \sum \{-\mu_i + y_i \ln[\mu_i] - \ln[y_i!]\}$

The loglikelihood ignoring terms that do not depend on the betas is:

$-24{,}000 \exp[\beta_0] + 1200 (\beta_0) - 20{,}000 \exp[\beta_0 + \beta_1] + 800 (\beta_0 + \beta_1)$

$- 15{,}000 \exp[\beta_0 + \beta_2] + 1100 (\beta_0 + \beta_2) - 13{,}000 \exp[\beta_0 + \beta_1 + \beta_2] + 900 (\beta_0 + \beta_1 + \beta_2)$.

Setting the partial derivative of the loglikelihood with respect to β_1 equal to zero:

$- 20{,}000 \exp[\beta_0 + \beta_1] + 800 - 13{,}000 \exp[\beta_0 + \beta_1 + \beta_2] + 900 = 0$.

Given $\beta_0 = -3.0300$: $1700 = 966.3 \exp[\beta_1] + 628.1 \exp[\beta_1] \exp[\beta_2]$.

Setting the partial derivative of the loglikelihood with respect to β_2 equal to zero:

$- 15{,}000 \exp[\beta_0 + \beta_2] + 1100 - 13{,}000 \exp[\beta_0 + \beta_1 + \beta_2] + 900 = 0$.

$\Rightarrow 2000 = 724.7 \exp[\beta_2] + 628.1 \exp[\beta_1] \exp[\beta_2]$.

Subtracting two equations: $300 = 724.7 \exp[\beta_2] - 966.3 \exp[\beta_1]$.

$\Rightarrow \exp[\beta_2] = 0.4140 + 1.3334 \exp[\beta_1]$.

$\Rightarrow 1700 = 966.3 \exp[\beta_1] + 628.1 \exp[\beta_1] (0.4140 + 1.3334 \exp[\beta_1])$.

Let $x = \exp[\beta_1]$. $\Rightarrow 1700 = 966.3 x + 628.1 x (0.4140 + 1.3334 x)$.

$\Rightarrow 837.5x^2 + 1226.3 x - 1700 = 0$.

$\Rightarrow x = \dfrac{-1226.3 \pm \sqrt{1226.3^2 - (4)(837.5)(-1700)}}{(2)(837.5)} = 0.8697$, taking the positive root.

$\Rightarrow \beta_1 = \ln(0.8697) = -0.1396$.

$\Rightarrow \exp[\beta_2] = 0.4140 + 1.3334 \exp[\beta_1] = 0.4140 + (1.3334)(0.8697) = 1.5737$.

$\Rightarrow \beta_2 = \ln(1.5737) = 0.4534$.

Expected frequency of a female risk in Territory B is:

$\exp[\beta_0 + \beta_1 + \beta_2] = \exp[-3.0300 - 0.1396 + 0.4534] =$ **6.61%**.

Comment: Similar to 8, 11/15, Q.3.

Using a computer, without being given β_0, the maximum likelihood fit is:

$\hat{\beta}_0 = -3.02999$, $\hat{\beta}_1 = -0.139599$, $\hat{\beta}_2 = 0.453335$.

The mean modeled frequencies are:

	Territory A	Territory B
Male	exp[-3.02999] = 4.83%	exp[-3.02999 + 0.453335] = 7.60%
Female	exp[-3.02999 - 0.139599] = 4.20%	exp[-3.02999 - 0.139599 + 0.453335] = 6.61%

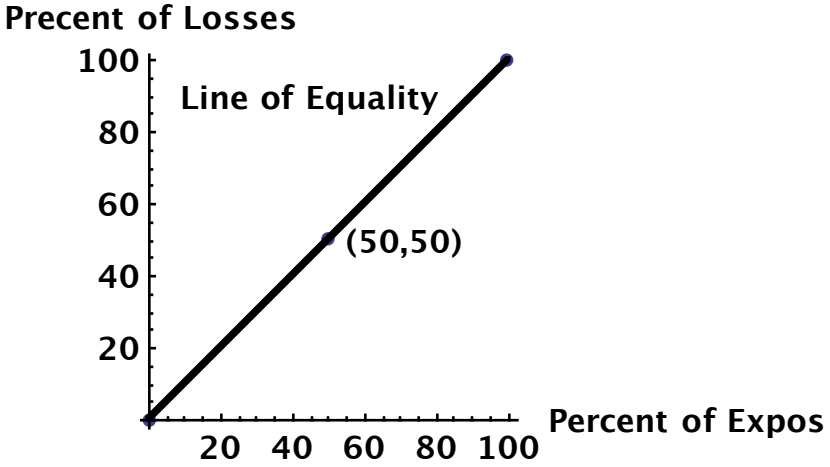
3.137. When the effect of one predictor depends on the level of another predictor, and vice-versa, such an relationship is called an interaction.
An example of an interaction term: X_1X_2.

In this example, $g(\mu) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3 X_1X_2 + ...$

The effect of X_1 depends on the level of X_2 and vice-versa.

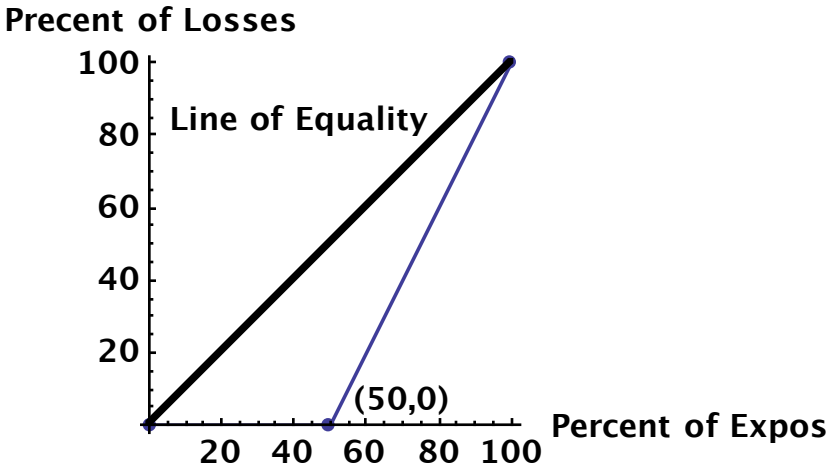
Comment: See Section 5.6 of Goldburd, Khare, and Tevet.
The actuary can use the GLM significance statistics in order to determine whether the inclusion of an interaction significantly improves the model.

3.138. (a) The percent of losses for A are 50%.. So the Lorenz Curve has the point (50, 50).



The Lorenz curve is equal to the line of equality, and thus the area between them is zero. The Gini Index is twice that, or **zero**.

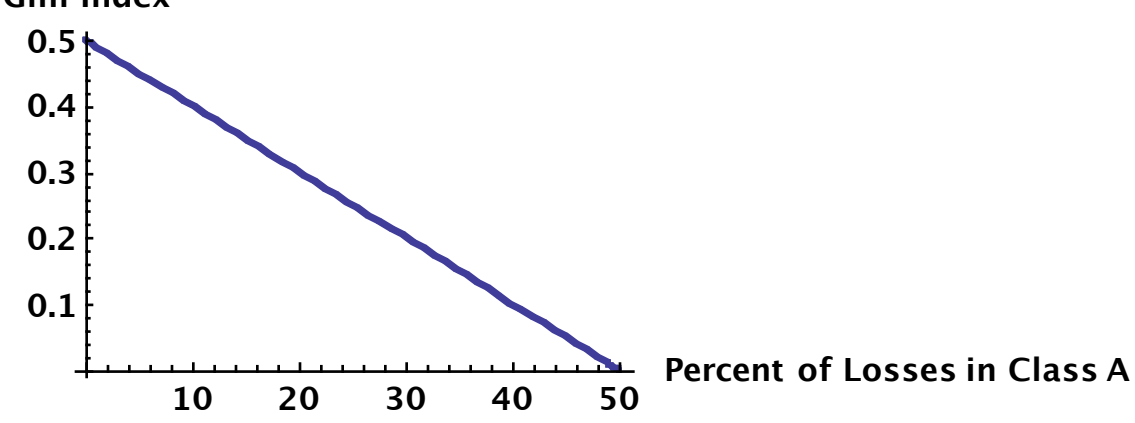
(b) The percent of losses for A are 0%. So the Lorenz Curve has the point (50, 0).



The region between the Lorenz curve and the line of equality is a triangle of base 50% and height 100%, and thus area: (1/2)(50%)(100%) = 0.25. The Gini Index is twice that, or **50%**.

Comment: We looked at the two extreme cases, which will not occur in practice.

Here is a graph of the Gini Index versus the percent of total actual losses in Class A:



3.139. The curve corresponding to the text labeled VA has more area under it, so it is better than the test labeled NE.

3.140. The fitted parameter(s) are the same, while the standard errors are multiplied by $\sqrt{3.071}$.

The standard error of $\hat{\beta}_1$ is: $0.1978 \sqrt{3.071} = 0.3466$.

95% confidence interval for β_1: 5.624 ± (1.96) (0.3466) = **5.624 ± 0.679**.

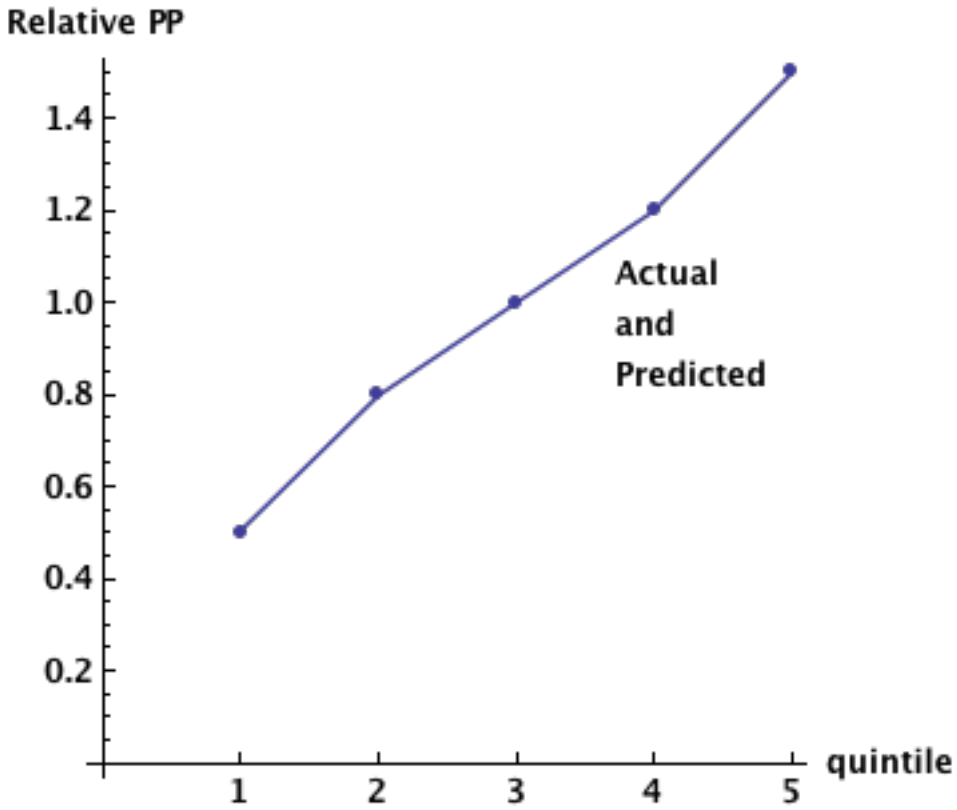
Comment: One could instead use: 5.624 ± (2) (0.3466) = 5.624 ± 0.693.

3.141. A simple quintile plot is a simple quantile plot with 5 buckets.

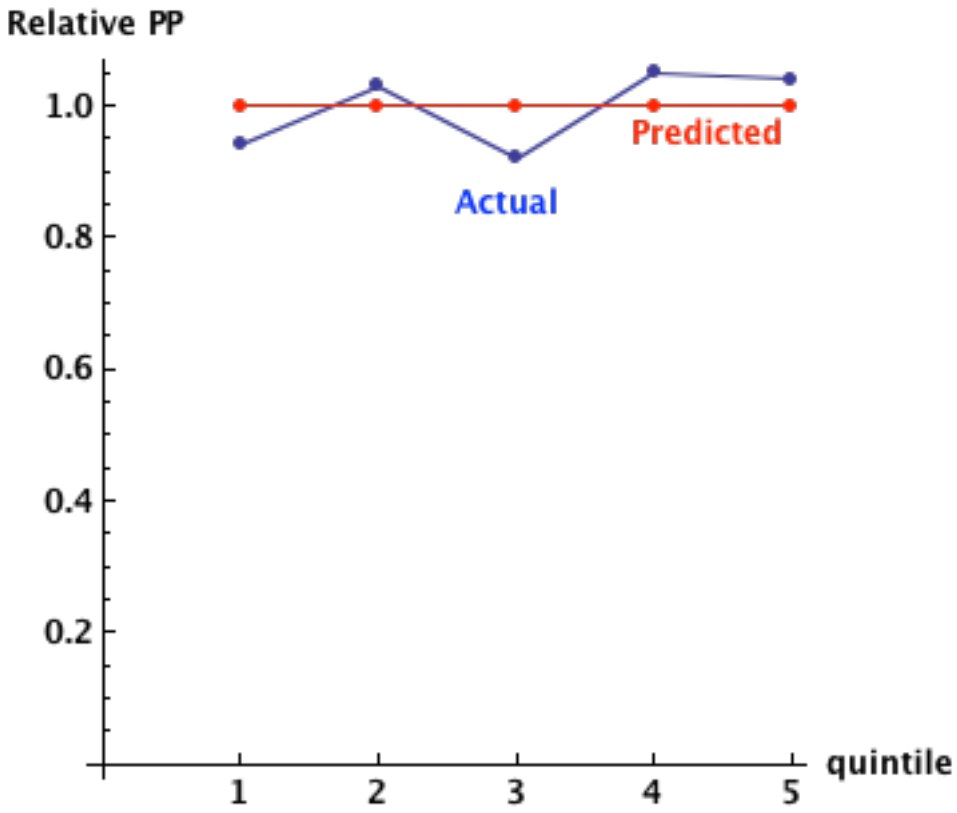
- Sort the dataset based on the model predicted pure premium from smallest to largest.
- Group the data into 5 buckets with equal volume.
- Within each group, calculate the average predicted pure premium based on the model, and the average actual pure premium.
- Plot for each group, the actual pure premium and the predicted pure premium.

Since we are not given the overall average pure premium, I will plot the pure premiums relative to average.

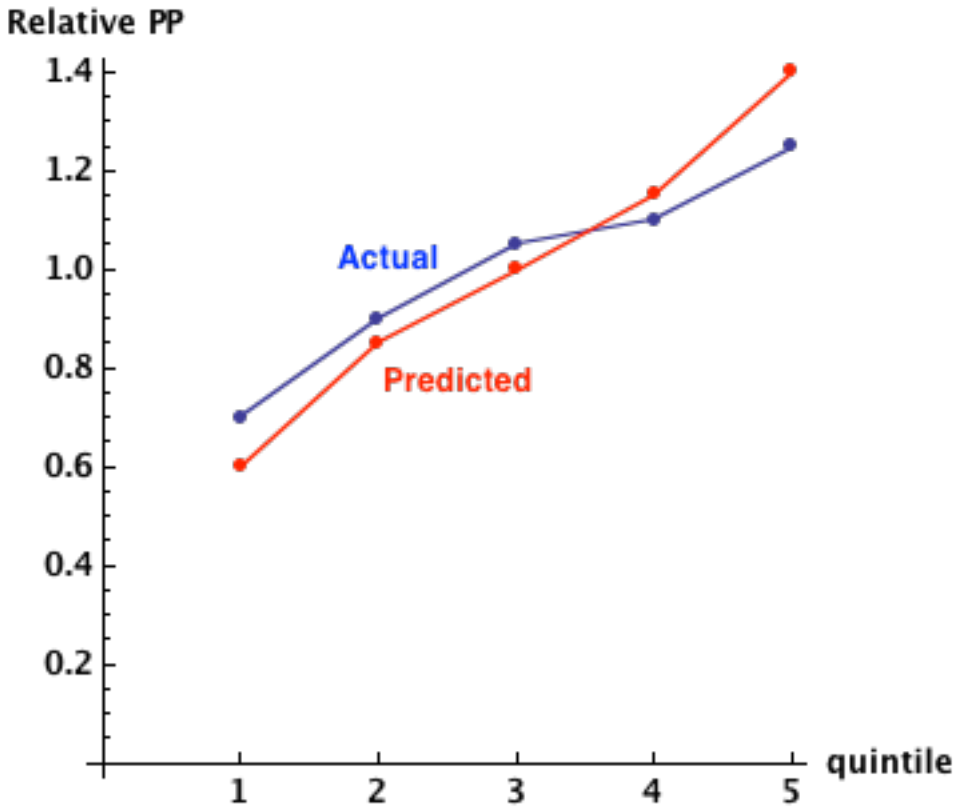
The saturated model has as many predictors as data points. Thus for the saturated model, the predictions exactly match the observations for each record.
The simple quintile plot:



The null model, has no predictors, only an intercept. Thus for the null model the prediction is the same for every record: the grand mean.
Since every risk has the same prediction, one would assign them to buckets at random.
Thus all of the actuals by quintile should be close to the grand mean, with small differences due to the randomness of assignments. The simple quintile plot:



"A model that could be used in practice", would have the actuals increase monotonically, have good but not perfect predictive accuracy, and a reasonably large vertical distance between the actuals in the first and last quintiles. A simple quintile plot:



Comment: Similar to 8, 11/07, Q. 5.
There are many possible examples of the last plot.
Since the records are ordered by predicted values, the records in each bucket change for each graph. Thus, actuals are not the same for each graph.
Quintile plots are sorted by predicted values from smallest to largest value. Thus the predicted values must be monotonically increasing (or in the case of the null model equal). Actuals need not be monotonically increasing, although that is desirable.
In every graph, the average of the actuals should be the grand mean.
In the final plot, the average of the predicteds should be close to if not equal to the grand mean; the GLM may have a small bias.
In the final plot, the predicted and actuals for the final quintile should each be less than in the saturated model. In the final plot, the predicted and actuals for the final quintile should each be more than in the null model.

3.142. I prefer the Gamma model, since the standardized deviance residuals are much closer to being Normally Distributed.

3.143. Since the proposed model is not able to segment the data into lower and higher loss ratio buckets, the proposed model is not significantly outperforming the current rating plan.
Comment: See Section 7.2.3 of Goldburd, Khare, and Tevet.

3.144. AIC = (-2) (maximum loglikelihood) + (number of parameters)(2).
For example, AIC = (-2)(-9844.16) + (5)(2) = 19,698.32.

Model	Number of Parameters	Loglikelihood	AIC
A	5	-9844.16	19,698.32
B	10	-9822.48	19,664.96
C	15	-9815.70	19,661.40

Since AIC is smallest for model C, model C is preferred.

3.145. BIC = (-2) (maximum loglikelihood) + (number of parameters) ln[number of data points].
For example, BIC = (-2)(-9844.16) + 5 ln[5000] = 19730.91.

Model	Number of Parameters	Loglikelihood	BIC
A	5	-9844.16	19,730.91
B	10	-9822.48	19,730.13
C	15	-9815.70	19,759.16

Since BIC is smallest for model B, model B is preferred.
Comment: Similar to 8, 11/16, Q.7.
See Section 6.2.2 in Goldburd, Khare, and Tevet.
Most actuarial GLMs are fit to many more than 5000 data points.
"As most insurance models are fit on very large datasets, the penalty for additional parameters imposed by BIC tends to be much larger than the penalty for additional parameters imposed by AIC. In practical terms, the authors have found that AIC tends to produce more reasonable results. Relying too heavily on BIC may result in the exclusion of predictive variables from your model."

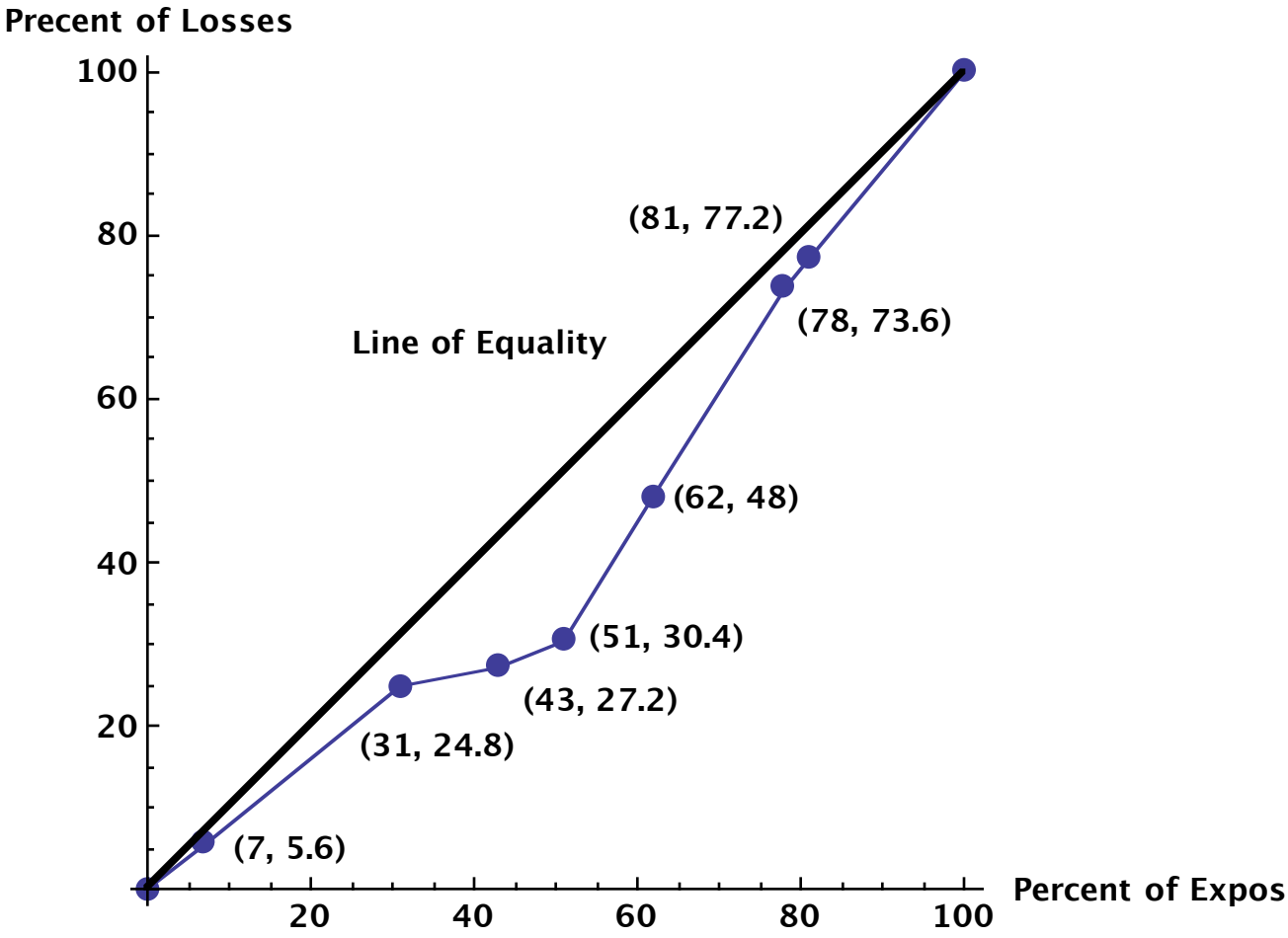
3.146. The first model does a better job of fitting the data and is thus preferred.

3.147. Sort the risks from best to worst based on the model predicted pure premium.

Risk	Model P.P.	Exposures	Cumulative Exposures	% of Exposures
2	1000	7	7	7%
8	2000	24	31	31%
5	3000	12	43	43%
3	4000	8	51	51%
4	5000	11	62	62%
6	6000	16	78	78%
1	7000	3	81	81%
7	8000	19	100	100%

Risk	Exposures	Actual P.P.	Actual Losses	Cumulative Losses	% of Losses
2	7	4000	28,000	28,000	5.6%
8	24	4000	96,000	124,000	24.8%
5	12	1000	12,000	136,000	27.2%
3	8	2000	16,000	152,000	30.4%
4	11	8000	88,000	240,000	48.0%
6	16	8000	128,000	368,000	73.6%
1	3	6000	18,000	386,000	77.2%
7	19	6000	114,000	500,000	100.0%

On the x-axis, plot the cumulative percentage of exposures.
On the y-axis, plot the cumulative percentage of actual losses.
The plotted points are: (0, 0), (7%, 5.6%), (31%, 24.8%), ... , (81%, 77.2%), (100%, 100%).



Comment: Similar to 8, 11/16, 5a.
The Gini index is twice the area between the Lorenz Curve and the line of equality.
The higher the Gini Index, the better the rating plan is at identifying risk differences.

3.148.

Variable	Number of Parameters
Vehicle Price	3
Vehicle Age	8 - 1 = 7
Driver age	2 - 1 = 1
Number of drivers	3 - 1 = 2
Gender	2 - 1 = 1
Interaction Gender & Driver Age	1

Number of parameter is: 3 + 7 + 1 + 2 + 1 + 1 = **15**.
Comment: Similar to CAS S, 11/15, Q.35.
A model with only Vehicle Price would involve: $\beta_0 + \beta_1 (vp) + \beta_2 (vp)^2$.
The interaction of gender and driver age only uses one parameter since each of gender and driver age only use one parameter.

3.149. A double lift chart compares the current rating plan to a proposed model.
Sort data by ratio of model prediction to current premium.
Subdivide sorted data into quantiles with equal exposure.
For each quantile calculate average actual loss cost, average model predicted loss cost and the average loss cost underlying the current manual premium .
Index the quantile averages to the overall averages.
Plot the results.
Comment: The "winning" model is the one that more closely matches the actual pure premiums.

3.150. The difference in degrees of freedom is: 18,175 - 18,169 = 6; we add 6 parameters.

Test statistic is: $F = \frac{D_S - D_B}{(\text{number of added parameters})\ \hat{\phi}_S} = \frac{8{,}905.6226 - 8{,}901.4414}{(6)\ (0.4523)} = 1.541.$

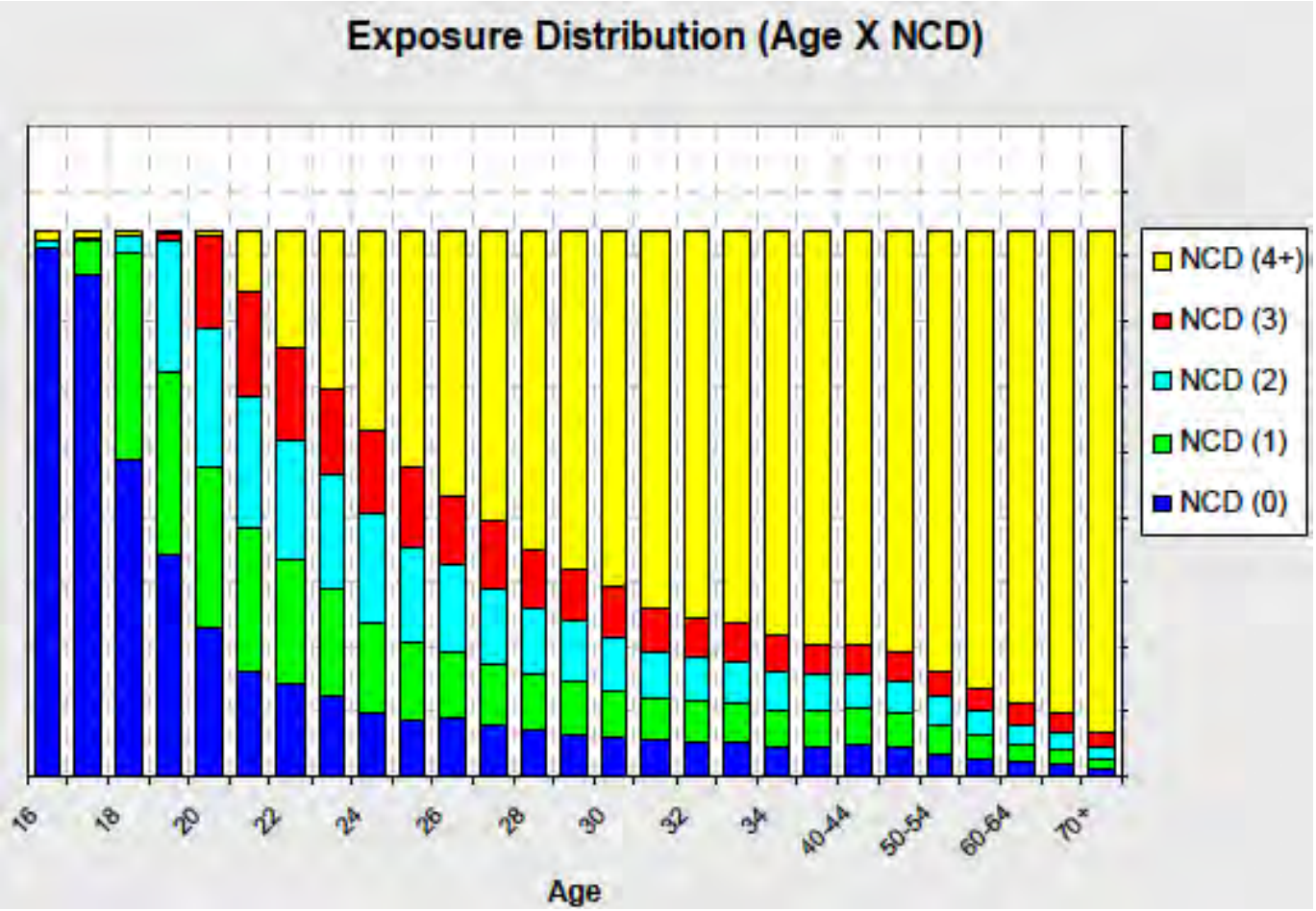
The number of degrees of freedom in the numerator is 6.
The number of degrees of freedom in the denominator is:
number of degrees of freedom for the simpler model = 18,175.
We compare the test statistic to the appropriate F-distribution.
We reject the null hypothesis if the test statistic is sufficiently big.
Comment: Using a computer, the p-value is 16.0%. Thus at for example a 10% significance level, we do not reject the null hypothesis to use the simpler model.

3.151. mean = exp[5.07 + 0.48 - 0.36] = 179.5.
Variance = mean^2 / α = 179.5^2 / 2.2 = **14,646**.
Comment: Similar to CAS S, 5/16, Q.32.

3.152. The Gini index can be used to measure the lift of an insurance rating plan by quantifying its ability to segment the population into the best and worst risks.
The larger the Gini index, the better job the rating plan does of segmenting.
Thus the rating plan used in Model 1 has more lift than the rating plan used in Model 2.

3.153. One works with loss ratios with respect to the premiums for the current plan.
To create a loss ratio chart:
1. Sort the dataset based on the model prediction.
2. Group the data into quantiles with equal volumes of exposures.
3. Within each group, calculate the actual loss ratio.
Comment: If the proposed model is able to segment the data into lower and higher loss ratio buckets, then the proposed model is better than the current model.

3.154. Driver age and number of years claims-free are positively correlated. Older drivers are likely to be claims-free for more years than younger drivers. Thus in order to avoid double counting effects, the GLM lessens the effect of each variable somewhat compared to a model that just used one of the two variables.
Comment: A graph of number of years claims-free versus driver age:



Graph taken from "GLM II: Basic Modeling Strategy," by Claudine Modlin, CAS Predictive Modeling Seminar, October 2008.
If two variables are very highly correlated, which is not the case here, then the GLM will have trouble converging and the parameter estimates may be unreliable.

3.155. If the current rating plan were perfect, then all risks should have the same loss ratio. The fact that the proposed model is able to segment the data into lower and higher loss ratio buckets is a strong indicator that it is outperforming the current rating plan.
Comment: Graph taken from "Goodness of Fit vs. Goodness of Lift," by Glenn Meyers and David Cummings, August 2009 Actuarial Review.
If one insurer were to use the current rating plan, while another insurer were to use the proposed rating plan, the second insurer should be able to attract better risks from the first insurer. The first insurer who continued to use the current plan would be subject to adverse selection.

3.156. The "winning" model is the one that more closely matches the actual pure premiums. The proposed model does a much better job than the current rating plan; thus the proposed model is preferred.

3.157. 1. Model A does a better job of matching the actual than does Model B. Thus based on the criterion of predictive accuracy I prefer Model A.
Both models satisfy the criterion of monotonicity; the actuals increase with quintile.
Model A has a larger vertical distance between the actuals for the first and last quintiles than does Model B. Thus based on this criterion I prefer Model A.
Thus overall I prefer Model A to Model B.
Comment: In order to determine the "winning" model, consider the following 3 criteria:
1. Predictive accuracy. How well each model is able to predict the actual pure premium in each quantile.
2. Monotonicity. By definition, the predicted pure premium will monotonically increase as the quantile increases, but the actual pure premium should also increase (though small reversals are okay).
3. Vertical distance between the first and last quantiles. The first quantile contains the risks that the model believes will have the best experience, and the last quantile contains the risks that the model believes will have the worst experience. A large difference (also called "lift") between the actual pure premium in the quantiles with the smallest and largest predicted loss costs indicates that the model is able to maximally distinguish the best and worst risks.

3.158. I prefer the Inverse Gaussian model, since the standardized deviance residuals are much closer to being Normally Distributed.

3.159.

		30% Threshold		60% Threshold	
Claim #	Fraud	Predict.		Predict.	
1	N	Y	False Pos.	N	True Neg.
2	N	Y	False Pos.	N	True Neg.
3	N	N	True Neg.	N	True Neg.
4	N	N	True Neg.	N	True Neg.
5	Y	Y	True Pos.	Y	True Pos.
6	N	N	True Neg.	N	True Neg.
7	Y	N	False Neg.	N	False Neg.
8	N	Y	False Pos.	Y	False Pos.
9	Y	Y	True Pos.	Y	True Pos.
10	Y	Y	True Pos.	Y	True Pos.
11	N	Y	False Pos.	N	True Neg.
12	Y	Y	True Pos.	N	False Neg.
13	N	Y	False Pos.	N	True Neg.
14	N	Y	False Pos.	Y	False Pos.
15	N	Y	False Pos.	N	True Neg.

(a)

30% Threshold

	Predicted		
Actual	Fraud	No Fraud	Total
Fraud	true pos.: 4	false neg.: 1	5
No Fraud	false pos.: 7	true neg.: 3	10
Total	11	4	15

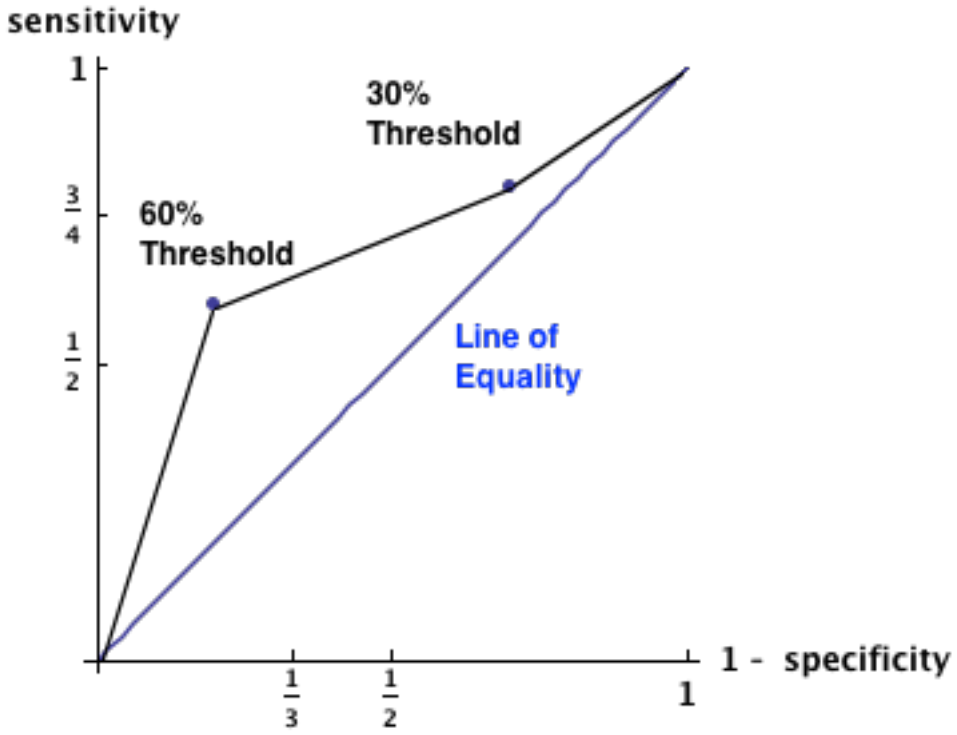
60% Threshold

	Predicted		
Actual	Fraud	No Fraud	Total
Fraud	true pos.: 3	false neg.: 2	5
No Fraud	false pos.: 2	true neg.: 8	10
Total	5	10	15

(b) Sensitivity = $\frac{\text{True Positives}}{\text{Total Number of Events}} = \frac{\text{Correct Predictions of Fraud}}{\text{Total Number of Fraudulent Claims}}$.

Specificity = $\frac{\text{True Negatives}}{\text{Total Number of Non-Events}} = \frac{\text{Correct Predictons of No Fraud}}{\text{Total Number of Nonfraudulent Claims}}$.

30% threshold: sensitivity = 4/5, and specificity = 3/10. Graph (1 - 3/10, 4/5).
60% threshold: sensitivity = 3/5, and specificity = 8/10 = 4/5. Graph (1 - 4/5, 3/5).
The ROC Curve, plus the 45-degree comparison line:



Comment: Similar to 8, 11/07, Q. 6a&b.

3.160. A model with approximately 6 degrees of freedom has the right balance, since it has the smallest test MSE.
A model with approximately 2 degrees of freedom is too simple, it has a larger test MSE.
A model with approximately 20 degrees of freedom is too complex, it has a larger test MSE; while the training MSE is smaller that is due to this model being overfit.
Comment: Similar to 8, 11/17, Q 4b.
See Figure 7 in Generalized Linear Models for Insurance Rating.
We are interested in how the GLM will perform at predicting the response variable on some future set of data rather than on the set of past data with which we are currently working.
Our goal in modeling is to find the right balance where we pick up as much of the signal as possible with minimal noise, represented in this case by model with about 6 degrees of freedom.

3.161. One-way or univariate analysis does not accurately take into account the effect of other rating variables. It does not consider exposure correlations with other rating variables.

3.162. a. Linear Model:
• Random Component: Each component of Y is independent and normally distributed. Their means may differ, but they have common variance.
• Systematic Component: The covariates are combined to produce the linear predictor $\eta = X\beta$.
• Link Function: The relationship between the random component and the systematic component is specified with the identity link function: $E(Y) = \mu = \eta$.
(if g is the identity link function, $g^{-1}(\eta) = \eta$.)
Generalized Linear Model:
• Random Component: Each component of Y is independent and a member of an exponential family. (While the Normal is one possibility, there are others.)
• Systematic Component: The covariates are combined to produce the linear predictor $\eta = X\beta$.
• Link Function: The relationship between the random component and the systematic component is
specified with the link function, which is differentiable and monotonic such that:
$E(Y) = \mu = g^{-1}(\eta)$. (While the identity link function is one possibility, there are others.)
b) 1) The assumption of normality with common variance is often not true.
2) Sometimes the response variable may be restricted to be positive, but normality with the identity link function violates this.

3.163. i. Classical Linear Model: Response variable is normally distributed.
Generalized Linear Model: Response variable is from the exponential family.
ii) Classical Linear Model: The variance is constant but the mean is allowed to vary.
Generalized Linear Model: The variance is a function of the mean (exponential family).

3.164. a. Intrinsic aliasing is a linear dependency between covariates due to the definition. For example, if we have only black, red and blue cars, the red cars can be determined from [total cars] - black - blue. As another example, age of vehicle would alias with model year, since if you know one you can determine the other.
Extrinsic aliasing is a linear dependency between covariates that arises due to the nature of the data
rather than inherent properties of the covariates themselves. For example, if in the data all cars with unknown color also have an unknown number of doors, and vice-versa.
b. We have that [all cars] - large cars - medium cars = small cars, so we can say that $X_{small} = 1 - X_{large} - X_{medium}$.
If we do not have a base level, then we could have two size variables such as Large and Medium, plus all four territories.
We have that [all cars] - North - South - West = East, so we can say that $X_{East} = 1 - X_{North} - X_{South} - X_{West}$.
If we do not have a base level, then we could have three territory variables such as North, South, and West, plus all three sizes.
Alternately, we can eliminate β_{small} and β_{East} from the model and include an intercept term; Small / East would be the base level. Intercept plus 2 size and 3 territory variables.
Comment: The current syllabus reading does not distinguish between intrinsic and extrinsic aliasing.
In part (b) we should end up with 6 variables in total.
If we have an intercept term, we would have in addition three territory levels and two size levels.
Aliasing occurs when there is a linear dependency among the observed covariates. Equivalently, aliasing can be defined as a linear dependency among the columns of the design matrix X.
Near aliasing is a common problem and occurs when two or more factors contain levels that are almost, but not quite, perfectly correlated. This same problem comes up when performing multiple linear regressions.

3.165. There are many possible ways to set this up.
Taking North and Medium as the base levels as instructed.
Let X_1 correspond to the intercept term. It is one for all cells.
Let X_2 correspond to South. $X_2 = 1$ if South and 0 otherwise.
Let X_3 correspond to East. $X_3 = 1$ if East and 0 otherwise.
Let X_4 correspond to West. $X_4 = 1$ if West and 0 otherwise.
Let X_5 correspond to Small. $X_5 = 1$ if Small and 0 otherwise.
Let X_6 correspond to Large. $X_6 = 1$ if Large and 0 otherwise.
Then the design matrix, X, and response vector Y are:

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{matrix} \text{North Small} \\ \text{North Medium} \\ \text{North Large} \\ \text{South Small} \\ \text{South Medium} \\ \text{South Large} \\ \text{East Small} \\ \text{East Medium} \\ \text{East Large} \\ \text{West Small} \\ \text{West Medium} \\ \text{West Large} \end{matrix} \quad Y = \begin{pmatrix} 100 \\ 150 \\ 250 \\ 80 \\ 110 \\ 290 \\ 90 \\ 170 \\ 200 \\ 180 \\ 260 \\ 540 \end{pmatrix}$$

The vector of parameters is:

$$\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix}$$

Comment: While one could fit a Poisson to pure premiums, and treat the result as a discrete approximation, it is more common to fit a Tweedie Distribution.
Using a computer the fitted parameters are:
$\beta_1 = 4.95978$, $\beta_2 = -0.040822$, $\beta_3 = -0.0833816$, $\beta_4 = 0.672944$, $\beta_5 = -0.427444$, $\beta_6 = 0.617924$.
We have a multiplicative model with relativities:
South: exp[-0.040822] = 0.960, East: exp[-0.0833816] = 0.920, West: exp[0.672944] = 1.960,
Small: exp[-0.427444] = 0.652, Large: exp[0.617924] = 1.855.
For example, the fitted value for South and Small is:
$\exp[\beta_1 + \beta_2 + \beta_5] = \exp[4.95978 - 0.040822 - 0.427444] = 89.26$.

The fitted values are:

	Vehicle Size			
Territory	Small	Medium	Large	Total
North	92.98	142.56	264.46	500
South	89.26	136.86	253.88	480
East	85.54	131.16	243.31	460.01
West	182.23	279.42	518.35	980
Total	450.01	690	1280	2420.01

Subject to rounding, the totals for the fitted match those for the data.
In general, the estimates will be in balance as they were here, when one uses the canonical link function; the canonical link function for the Poisson is the log link function.
See “A Systematic Relationship Between Minimum Bias and Generalized Linear Models,” by Stephen Mildenhall, PCAS 1999.

3.166. a) $Y = \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + e$.

$Y_1 = 400 = \beta_1 + 0 + \beta_3 + e_1$.

$Y_2 = 250 = \beta_1 + 0 + 0 + e_2$.

$Y_3 = 200 = 0 + \beta_2 + \beta_3 + e_3$.

$Y_4 = 100 = 0 + \beta_2 + 0 + e_4$.

Sum of Squared Errors $= e_1^2 + e_2^2 + e_3^2 + e_4^2$

$= (400 - \beta_1 - \beta_3)^2 + (250 - \beta_1)^2 + (200 - \beta_2 - \beta_3)^2 + (100 - \beta_2)^2$.

Set equal to zero the partial derivatives with respect to betas:

$2(400 - \beta_1 - \beta_3)(-1) + 2(250 - \beta_1)(-1) = 0. \Rightarrow 2\beta_1 + \beta_3 = 650.$

$2(200 - \beta_2 - \beta_3)(-1) + 2(100 - \beta_2)(-1) = 0. \Rightarrow 2\beta_2 + \beta_3 = 300.$

$2(400 - \beta_1 - \beta_3)(-1) + 2(200 - \beta_2 - \beta_3)(-1) = 0. \Rightarrow \beta_1 + \beta_2 + 2\beta_3 = 600.$

Solve for the betas.

b) i. constant variance. However, the variance is often a function of the mean.

ii. The components of the response variable are normally distributed.

For example, the response variable may be restricted to non-negative values, violating normality.

iii. Additivity of effects. Many factors in reality have multiplicative effects.

Comment: GLMs relax all of the three assumptions in part (b).

In the additive model in the question, we are taking Rural as the base; we have three categorical variables that each can take on the values zero or one, although when X1 = 1 we must have X2 = 0 and vice-versa.

The solution to the three equations is: $\beta_1 = 525/2$, $\beta_2 = 175/2$, and $\beta_3 = 125$.

The resulting estimates are:

Gender	Urban	Rural
Male	525/2 + 125 = 387.5	525/2 = 262.5
Female	175/2 + 125 = 212.5	175/2 = 87.5

The corresponding minimum sum of squared errors is:

$(400 - 387.5)^2 + (250 - 262.5)^2 + (200 - 212.5)^2 + (100 - 87.5)^2 = 625.$

3.167. a. As per the exam question, take Male (X_1), Female (X_2), Urban (X_3).
Then the design matrix, X, and response vector Y are:

$$X = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix} \quad \begin{matrix} \text{Male Urban} \\ \text{Male Rural} \\ \text{Female Urban} \\ \text{Female Rural} \end{matrix} \quad Y = \begin{pmatrix} 400 \\ 250 \\ 200 \\ 100 \end{pmatrix}.$$

The vector of parameters is: $\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$.

b. For the Gamma Distribution, $f(y) = \theta^{-\alpha} y^{\alpha-1} e^{-y/\theta} / \Gamma(\alpha)$.

$\ln f(y) = (\alpha-1)\ln(y) - y/\theta - \alpha\ln(\theta) - \ln[\Gamma(\alpha)] = (\alpha-1)\ln(y) - y/(\mu/\alpha) - \alpha\ln(\mu/\alpha) - \ln[\Gamma(\alpha)]$

$= (\alpha-1)\ln(y) - \alpha y/\mu - \alpha\ln(\mu) + \alpha\ln(\alpha) - \ln[\Gamma(\alpha)]$.

With the identity link function: $\mu = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.

Thus the loglikelihood is:

$(\alpha-1)\ln(400) - \alpha 400/(\beta_1 + \beta_3) - \alpha n(\beta_1 + \beta_3) + \alpha\ln(\alpha) - \ln[\Gamma(\alpha)] +$

$(\alpha-1)\ln(250) - \alpha 250/(\beta_1) - \alpha\ln(\beta_1) + \alpha\ln(\alpha) - \ln[\Gamma(\alpha)] +$

$(\alpha-1)\ln(200) - \alpha 200/(\beta_2 + \beta_3) - \alpha\ln(\beta_2 + \beta_3) + \alpha\ln(\alpha) - \ln[\Gamma(\alpha)] +$

$(\alpha-1)\ln(100) - \alpha 100/(\beta_2) - \alpha\ln(\beta_2) + \alpha\ln(\alpha) - \ln[\Gamma(\alpha)]$.

Setting the partial derivative with respect to α_1 equal to zero:

$0 = \alpha 400/(\beta_1 + \beta_3)^2 - \alpha/(\beta_1 + \beta_3) + \alpha 250/(\beta_1)^2 - \alpha/(\beta_1). \Rightarrow$

$400/(\beta_1 + \beta_3)^2 + 250/\beta_1^2 = 1/(\beta_1 + \beta_3) + 1/\beta_1$.

Setting the partial derivative with respect to β_2 equal to zero:

$0 = \alpha 200/(\beta_2 + \beta_3)^2 - \alpha/(\beta_2 + \beta_3) + \alpha 100/(\beta_2)^2 - \alpha/(\beta_2). \Rightarrow$

$200/(\beta_2 + \beta_3)^2 + 100/\beta_2^2 = 1/(\beta_2 + \beta_3) + 1/\beta_2$.

Setting the partial derivative with respect to β_3 equal to zero:

$0 = \alpha 400/(\beta_1 + \beta_3)^2 - \alpha/(\beta_1 + \beta_3) + \alpha 200/(\beta_2 + \beta_3)^2 - \alpha/(\beta_2 + \beta_3). \Rightarrow$

$400/(\beta_1 + \beta_3)^2 + 200/(\beta_2 + \beta_3)^2 = 1/(\beta_1 + \beta_3) + 1/(\beta_2 + \beta_3)$.

c. For the Gamma Distribution, $f(y) = \theta^{-\alpha} y^{\alpha-1} e^{-y/\theta} / \Gamma(\alpha)$.

$\ln f(y) = (\alpha-1)\ln(y) - y/\theta - \alpha\ln(\theta) - \ln[\Gamma(\alpha)] = (\alpha-1)\ln(y) - y/(\mu/\alpha) - \alpha\ln(\mu/\alpha) - \ln[\Gamma(\alpha)]$

$= (\alpha-1)\ln(y) - \alpha y/\mu - \alpha\ln(\mu) + \alpha\ln(\alpha) - \ln[\Gamma(\alpha)]$.

With the inverse link function: $1/\mu = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.

Thus the loglikelihood is:

$(\alpha-1)\ln(400) - \alpha 400(\beta_1 + \beta_3) + \alpha\ln(\beta_1 + \beta_3) + \alpha\ln(\alpha) - \ln[\Gamma(\alpha)] +$

$(\alpha-1)\ln(250) - \alpha 250(\beta_1) + \alpha\ln(\beta_1) + \alpha\ln(\alpha) - \ln[\Gamma(\alpha)] +$

$(\alpha-1)\ln(200) - \alpha 200(\beta_2 + \beta_3) + \alpha\ln(\beta_2 + \beta_3) + \alpha\ln(\alpha) - \ln[\Gamma(\alpha)] +$

$(\alpha-1)\ln(100) - \alpha 100(\beta_2) + \alpha\ln(\beta_2) + \alpha\ln(\alpha) - \ln[\Gamma(\alpha)]$.

Setting the partial derivative with respect to β_1 equal to zero:

$0 = -\alpha 400 + \alpha/(\beta_1 + \beta_3) - \alpha 250 + \alpha/(\beta_1)$. $\Rightarrow 650 = 1/(\beta_1 + \beta_3) + 1/\beta_1$.

Setting the partial derivative with respect to β_2 equal to zero:

$0 = -\alpha 200 + \alpha/(\beta_2 + \beta_3) - \alpha 100 + \alpha/(\beta_2)$. $\Rightarrow 300 = 1/(\beta_2 + \beta_3) + 1/\beta_2$.

Setting the partial derivative with respect to β_3 equal to zero:

$0 = -\alpha 400 + \alpha/(\beta_1 + \beta_3) - \alpha 200 - \alpha/(\beta_2 + \beta_3)$. $\Rightarrow 600 = 1/(\beta_1 + \beta_3) + 1/(\beta_2 + \beta_3)$.

(d) $f(x) = \sqrt{\frac{\theta}{2\pi}} \frac{\exp\left[-\frac{\theta\left(\frac{x}{\mu} - 1\right)^2}{2x}\right]}{x^{1.5}}$.

Ignoring terms that do not involve μ, $\ln f(x) = -\frac{\theta\left(\frac{x}{\mu} - 1\right)^2}{2x} = -\frac{\theta}{2x}\left(\frac{x^2}{\mu^2} - 2\frac{x}{\mu} + 1\right)$

$= -\frac{\theta x}{2\mu^2} + \frac{\theta}{\mu} - \frac{\theta}{2x}$.

Using the squared reciprocal link function: $1/\mu^2 = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.

Thus ignoring terms that do not include μ, the loglikelihood is:

$\frac{-\theta}{2}\{400(\beta_1 + \beta_3) + 250(\beta_1) + 200(\beta_2 + \beta_3) + 100(\beta_2)\} + \theta\{\sqrt{\beta_1 + \beta_3} + \sqrt{\beta_1} + \sqrt{\beta_2 + \beta_3} + \sqrt{\beta_2}\}$.

Setting the partial derivative with respect to β_1 equal to zero:

$0 = \frac{-\theta}{2}\{400 + 250\} + \frac{\theta}{2}\{1/\sqrt{\beta_1 + \beta_3} + 1/\sqrt{\beta_1}\}. \Rightarrow 650 = 1/\sqrt{\beta_1 + \beta_3} + 1/\sqrt{\beta_1}$.

Setting the partial derivative with respect to β_2 equal to zero:

$0 = \frac{-\theta}{2}\{200 + 100\} + \frac{\theta}{2}\{1/\sqrt{\beta_2 + \beta_3} + 1/\sqrt{\beta_2}\}. \Rightarrow 300 = 1/\sqrt{\beta_2 + \beta_3} + 1/\sqrt{\beta_2}$.

Setting the partial derivative with respect to β_3 equal to zero:

$0 = \frac{-\theta}{2}\{400 + 200\} + \frac{\theta}{2}\{1/\sqrt{\beta_1 + \beta_3} + 1/\sqrt{\beta_2 + \beta_3}\}. \Rightarrow 600 = 1/\sqrt{\beta_1 + \beta_3} + 1/\sqrt{\beta_2 + \beta_3}$.

Comment: Using a computer, the fitted parameters in part b are:

$\beta_1 = 263.236$, $\beta_2 = 98.160$, $\beta_3 = 110.129$.

For example, the fitted value for Female and Urban is: 98.160 + 110.129 = 208.29.

The fitted values in part b are:

Gender	Urban	Rural
Male	373.36	263.24
Female	208.29	98.16

Using a computer, the fitted parameters in part c are:

$\beta_1 = 0.00447623$, $\beta_2 = 0.00789904$, $\beta_3 = -0.0021321$.

For example, the fitted value for Female and Urban is: 1/ (0.00789904 - 0.0021321) = 173.40.

The fitted values in part c are:

Gender	Urban	Rural	Total
Male	426.60	223.40	650
Female	173.40	126.60	300
Total	600	350	950

The totals for the fitted match those for the data.
These were the equations that needed to be solved for this model in part c.
In general, the estimates will be in balance as they were here, when one uses the canonical link function; the canonical link function for the Gamma is the reciprocal link function.
See "A Systematic Relationship Between Minimum Bias and Generalized Linear Models," by Stephen Mildenhall, PCAS 1999.
Note that when the weights differ by cell, this balance involves weighted averages.
Using a computer, the fitted parameters in part d are:
$\beta_1 = 0.0000218789$, $\beta_2 = 0.000053899$, $\beta_3 = -0.0000166235$.

For example, the fitted value for Female and Urban is:
$1 / \sqrt{0.000053899 - 0.0000166235} = 163.79$.

The fitted values in part d are:

Gender	Urban	Rural	Total
Male	436.21	213.79	650
Female	163.79	136.21	300
Total	600	350	950

Since the canonical link function for the Inverse Gaussian is the squared reciprocal link function, again the totals for the fitted match those for the data.

3.168.
The model appears to be appropriate for the ages nearer the center, from about 25 to 70.
The model does not appear to be appropriate for either the younger ages below 25 or the older ages above 70.
The model appears to be over fitted. In the age ranges without many exposures, the model is picking up the random fluctuations in the data, in other words the noise. In these age ranges, the model matches the data to which it was fit but fails to match the holdout dataset to which it was not fit.
If the model had been picking up useful information about a pattern in these age ranges, in other words had been picking up a signal, then it should have matched to some extent the holdout data set as well.
Comment: See Exhibit F.5 in Appendix F of Basic Ratemaking, on Exam 5.

3.169. a. Let β_1 represent territory A. Let β_2 represent territory B.

Let β_3 represent private passenger. Let β_4 represent light trucks.

(These are not the only choices. We have chosen medium trucks as the base level.)

Design matrix = X = $\begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$.

The first row of (1, 0, 1, 0) corresponds to the cell for Territory A (β_1) and private passenger (β_3).

The second row of (1, 0, 0, 1) corresponds to the cell for Territory A (β_1) and light truck (β_4).

The third row of (1, 0, 0, 0) corresponds to the cell for Territory A (β_1) and medium truck.

(There are other ways to arrange the design matrix.)

The corresponding vector of betas (parameters) is:

$(\beta_1 + \beta_3, \beta_1 + \beta_4, \beta_1, \beta_2 + \beta_3, \beta_2 + \beta_4, \beta_2)$.

b. For a poisson error structure variance is a function of the expected value, while under the gamma error structure the variance of an observation is a function of its mean squared.

c. Determine the form of the density for the chosen error structure (distribution of errors.)

Using this density and the chosen link function take a product of the chances of the observations; this is the likelihood as a function of the parameters.

Maximize the log of the likelihood function by setting the partial derivatives with respect to each of the parameters equal to zero.

Solve the resulting system of equations for the fitted parameters.

Compute the predicted values.

Comment: In part (a) one could have instead for example taken:

Let β_1 be an intercept. Let β_2 represent territory A.

Let β_3 represent light trucks. Let β_4 represent medium trucks.

In that case, we have taken Territory B / Heavy Trucks as the base level.

Instead, other combinations of territory and truck weight could have been chosen as the base level.

If we use an intercept term, then we can have only one coefficient for territory, and two coefficients for vehicle type. Including the intercept, we still have a total of four coefficients in our model.

In more complicated situations one would not be able to solve the equations for the parameters in closed form. Fortunately, there are commercial packages of computer software specifically designed to solve and analyze GLMs.

3.170. There are relatively wide error bars around the estimates of the model. For example, for the smoke detector class, the relativity could be between about 0.88 and 1.13, or a 12% credit to a 13% debit. The error bars are even wider for the Sprinkler System class, which is not surprising given the smaller number of exposures. A different relativity for Sprinkler System than the others although indicated is not justified by the model. Intuition would lead one to expect all of these devices to lower expected fire losses somewhat. However, these results of the model are not sufficient to justify giving any credit, let alone quantifying the size of such a credit. Based on the data analyzed, this is not an effective variable in the model. One would probably need significantly more exposures than was used here in order to properly analyze this whole situation.
Comment: A much shorter answer should suffice for full credit.
See Exhibit F.3 in Appendix F of Basic Ratemaking on Exam 5, which models frequency of wind losses. We are not told any details about the graph in this exam question. Is it modeling frequency or pure premiums? Is the graph in this exam question modeling a particular peril or all losses?
A good example of where a fancy model can not make up for having too small a volume of data. One would need to observe more homes and/or more years.

3.171. One-way analysis doesn't consider:
1. Correlations between rating variables, in other words correlations of exposures by cell.
For example, young people drive older cars more often. Worse loss ratios for older cars can be partially driven by the larger proportion of youthful drivers.
For example, age may be correlated with territory if a greater proportion of senior citizens live in certain parts of a state. The relative loss ratios for such territories will be better due to the higher proportion of drivers who are senior citizens.
2. Interdependencies among rating variables.
For example, the rate differentials between male and female drivers vary by age.
For example, young drivers who have expensive cars may be poor risks, but old drivers who have expensive cars may be good risks.

3.172. The indicated rate for extra-heavy vehicles is 10 to 15 percent higher than that for heavy vehicles. However, the relativity for extra-heavy vehicles is based on few exposures and thus the 95% confidence interval is wide. The 95% confidence interval stretches from about a rate 10% lower to about a rate 40% higher than heavy vehicles. Thus the proposal by management is well within the range indicated by the generalized linear model based on the available data.
On the other hand, based on the pattern for light, medium, and heavy vehicles, the insurance cost increases significantly with weight. Thus it is logical that extra-heavy vehicles would cost more than heavy vehicles.
Thus I would recommend that we charge the indicated relativity.
If we expand the number of extra-heavy vehicles written, we will get more data to better estimate an appropriate relativity in the future.
There is a risk that if we write a lot of new extra-heavy vehicles, they will be on average poorer risks or at least have different risk characteristics.
If management's proposal were adopted, and if the rates for heavy vehicles are lower than costs, we may attract a significant volume of new business, but lose money. If the combined rate for heavy and extra heavy vehicles is higher than costs for heavy vehicles, then we risk adverse selection.
We would be able to write lots of underpriced extra-heavy vehicles and lose a lot of our current heavy vehicles which will be overpriced.
It would be useful to see what competitors are charging for extra-heavy vehicles versus heavy vehicles. If the data used in the GLM is from one state, it would be useful to get information from other states. It would also be useful to investigate the interaction of vehicle weight with the other rating variables in more detail.
Comment: There is no one right answer. Given the limited information available in an exam question, one has to make some assumptions and do the best one can to answer in a sensible manner.

3.173. a) In the absence of any other information, I would choose a Poisson error function which is commonly used for frequency.
The frequencies look like they might follow a multiplicative model; the ratios of the columns look kind of similar and the ratios of the rows look kind of similar.
(In contrast, the differences in the columns look kind of different and the differences in the rows look kind of different. Thus I would not choose an additive model and the identity link function.)
Therefore, I will use a log link function corresponding to a multiplicative model.
$g(x) = \ln(x)$. $g^{-1}(x) = e^x$.
One needs to pick a base level.
It is likely that Yes/Yes has the most exposures, so I will pick that as the base level.
The vector of model parameters:
Let β_0 be the intercept term, which is a parameter which applies to all observations.
Let β_1 correspond to no for homeowners.
Let β_2 correspond to no for auto policy.

Then the response vector would be: $\begin{pmatrix} \text{Yes HO / Yes Auto} \\ \text{No HO / Yes Auto} \\ \text{Yes HO / No Auto} \\ \text{No HO / No Auto} \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \\ 8 \\ 12 \end{pmatrix}$.

The design matrix would be: $\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}$. $X_1 \Leftrightarrow \begin{cases} \text{No H.O.} = 1 \\ \text{Otherwise} = 0 \end{cases}$. $X_2 \Leftrightarrow \begin{cases} \text{No Auto} = 1 \\ \text{Otherwise} = 0 \end{cases}$.

Alternately, the vector of model parameters: Let β_1 correspond to yes for auto.
Let β_2 correspond to no for auto. Let β_3 correspond to yes for homeowners.

Then the response vector would be: $\begin{pmatrix} \text{Yes HO / Yes Auto} \\ \text{No HO / Yes Auto} \\ \text{Yes HO / No Auto} \\ \text{No HO / No Auto} \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \\ 8 \\ 12 \end{pmatrix}$.

The design matrix would be: $\begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$.

$X_1 \Leftrightarrow \begin{cases} \text{Yes Auto} = 1 \\ \text{Otherwise} = 0 \end{cases}$. $X_2 \Leftrightarrow \begin{cases} \text{No Auto} = 1 \\ \text{Otherwise} = 0 \end{cases}$. $X_3 \Leftrightarrow \begin{cases} \text{Yes H.O.} = 1 \\ \text{Otherwise} = 0 \end{cases}$.

b) Missing data can lead to aliasing. For the missing data, if in most cases both whether it had an auto and homeowner policy is missing, then there is the potential problem of aliasing or near aliasing. No data auto and no data homeowners would be either perfectly or highly correlated. With aliasing the model parameters make no sense. Near aliasing creates problems with convergence of the model.
A solution would be to exclude these missing data records from modeling.
Another solution is to eliminate the unknown level from one of the factors so there are no linear dependencies. In other words, one further covariate needs to be removed; this could either be the “unknown” auto covariate or the “unknown” homeowners covariate.
Comment: One can make other choices in part (a) for the vector of parameters and get full credit provided the design matrix is consistent.

3.174. a) Increase of 7.0% ⇔ -0.4172. 2 phone calls ⇔ -0.4239.
-0.4172 - 0.4239 + 1.793 = 0.9519.
Using the logistic model, probability of renewal is: exp(0.9519) / {1 + exp(0.9519)} = **72.15%**.
b) I would not use this strategy:
1. There is no reasonable connection between insurance losses and the number of times an insured calls an insurer. Charging those who call the insurer more would not be acceptable to the public.
Insurance regulators are very unlikely to allow the use of this rating variable.
While causality is not required, this is an extreme case of the opposite of causality.
2. The proposed variable is easily manipulated by the insured.
3. The proposed variable lacks constancy; the number of phone calls from an insured is likely to change from year to year.
4. We do not know why the rate of renewal decreases with number of phone calls made by the insured to the insurer. Could this be because when they call, insureds get impolite or incompetent service? In that case, a better strategy would be to improve the insurer's service so that they do not lose so many customers.
5. The number of phone calls is likely related to other variables which are more directly related to renewal probability, such as moving or age. The actuary should go back and try to find variables that are the underlying reasons for the model results.
6. If many of these calls are from insureds who are making claims, then perhaps the lower renewal is due to poor claims service. It would be a better strategy to improve claims service.
7. By raising the rate of insureds who made phone calls, you are making their future renewal rate even lower. The insurer is likely to lose a lot of insureds if it followed this strategy.
Alternately, I would use this strategy:
1. When pricing based on the lifetime of a policy, it makes sense to take into account the expected renewal rate. Those with a lower expected renewal rate, such as those who make several phone calls to the insurer, should be charged more, all else being equal.
2. Those who call the insurer more often are likely to be reporting a claim. Those with claims in the past, have a higher expected future claim frequency. So it makes sense to charge those with more calls more, since their future average claim frequency is higher than average.
3. Those who call more often in the past are more likely to call more often in the future, resulting in higher expense for the insurer.
4. Those who call the insurer more often are more likely to make a small claim when they suffer a small loss, and thus have higher expected future loses.
Comment: I found it much, much easier in this case, to argue against using the strategy.
(My reasons in favor other than the first, have nothing to do with the given model.)
On your exam, pick whichever side of the argument allows you to quickly come up with two good reasons.
Without diagnostics there is no way to check the statistical significance of the modeled result.
Some of the extra phone calls may be from insureds who got big increases and are calling to complain or to see if this insurer will match a quote from another insurer. Thus the two variables in the model may be correlated.

3.175. Burglar Alarm:
Based solely on the GLM, there is little evidence to support a discount; there is relatively little data for the non-base classes, particularly for central reporting.
There are wide confidence intervals for both Local Alarm and Central Reporting groups. The Local Alarm standard errors suggest it is not significantly different than the None category; the confidence interval encompass a relativity of one. Central reporting has very few exposures and large standard errors.
I would recommend this variable not be used; in other words, 1.00 factor for all groups.
Alternately, based solely on the GLM, there is little evidence to support a discount.
On the other hand, it is logical that a local burglar alarm will reduce theft losses.
It is logical that a central reporting burglar alarm would be more effective at reducing theft losses than a local alarm.
However, theft losses are only one of many perils covered by Homeowners.
We are given no information on what portion of the expected losses are due to theft; this varies by geographical location.
Based on the logic and the limited statistical support from the GLM, small discounts make sense.
I judgmentally select 0.98 for local alarm and 0.96 for central reporting.
Deductible:
Based solely on the GLM, due to the small amount of data, there is little evidence to support a discount for the $7500 and $10,000 deductibles. Also the discount for a $7500 should be smaller than for a $10,000 deductible; the GLM fitted relativities indicate the opposite.
There is somewhat more data for the $250 deductible, but the error bars are relatively wide.
On the other hand, we know that expected losses paid are more for a lower deductible and are lower for a higher deductible.
For the $7500 and $10,000 deductible, based on the difference between the indicated relativities for 2500 and 5000, I will judgmentally select relativities of 0.78 and 0.74.
(For evenly spaced deductibles, the difference in Loss Elimination Ratios gets smaller as the deductible increases, in the absence of either favorable or adverse selection.)
For the $250 deductible, based on the difference between the $500 and the $1000 relativities, a relatively of something like 1.20 might make sense. The 1.75 prediction from the GLM might be due to adverse selection. So I will judgmentally select a relativity of 1.30.
For the other deductibles, I will use the GLM output.
Thus my selected relativities are: 1.30, 1.10, 1.00, 0.95, 0.85, 0.78, 0.74.
Comment: There are many possible reasonable answers. Additional information besides the GLM output would be very helpful, for example competitor's rates.

3.176. (a) Limited expected value at $100,000 is:

$$\frac{8M + 1.8M + (100K)(40) + 1.8M + (100K)(40)}{100 + 35 + 40 + 35 + 25 + 15} = \$78,400.$$

The 40 large claims on policies with $250,000 limit contribute to the layer from $100,000 to $250,000: 7,400,000 - (40)(100,000).
The 15 largest claims on policies with $500,000 limit contribute to the layer from $100,000 to $250,000: (15)(250,000 - 100,000).
Estimate of the difference between the limited expected values at $100,000 and $250,000 is:

$$\frac{7.4M - (40)(100K) + 3.9M - (25)(100K) + (150K)(15)}{35 + 40 + 35 + 25 + 15} = \$47,000.$$

Indicated $250,000 ILF: 1 + $47,000 / $78,400 = **1.599**.

(b) The GLM considers the interaction of limit purchased with other characteristics such as class and territory. It is estimating the ratio of expected pure premiums of otherwise similar insureds (same class and territory) who buy different limits of coverage. Based on the model output, it seems as if those who choose to buy $250,000 limits are better risks than those who choose to buy basic limits of $100,000; there is favorable selection. Presumably the expected frequency of those who buy $250,000 limits is significantly lower than that of similar insureds who buy basic limits. Thus even though those who buy $250,000 limits are getting more coverage, the GLM estimates that their expected pure premium is lower than that of those who choose to buy basic limits. A GLM can sometimes produce counter intuitive results, such as lower increased limits factors for higher limits.
It should be noted that many class/territory/limit purchased cells will have little data.
Therefore, some of the GLM results may be due to random fluctuation (noise rather than signal.)
(It would have been useful to have the standard errors associated with the GLM output.)
In contrast, the calculation in part (a) implicitly assumes that the expected frequency does not vary by limit purchased. Also it does not consider the mix of classes and territories by limit purchased.

(c) Using an ILF of 0.95 for $250,000 would result in charging less for more coverage.
Soon anybody who would otherwise have bought basic limits would instead buy $250,000 limits; in addition we would attract insureds who currently buy basic limits from other insurers. These insureds would get more coverage for less premium, and therefore our premiums for $250,000 limits will be inadequate. (If we used the 0.95 ILF for $250,000, the favorable selection that the GLM said is in the data with respect to purchasers of $250,000 limits would vanish.)
Therefore, I will not use the ILF indicated by the GLM.
On the other hand, the GLM output leads me to believe that the indicated ILF from part (a) is too high. One could select something in between such as 1.30.
(It would be very helpful to have more information on the GLM output including but not limited to standard errors. It would be helpful to have more information such as the current ILFs, competitors ILFs, etc.)
Alternately, one can base ones selection on the 1.15 GLM indicated for $500,000.
The $250,000 ILF should be greater than one and less than the $500,000 ILF.
Linearly interpolating one would get: 1 + (0.15)(250 - 100) / (500 - 100) = 1.056.
However, ILFs should decrease at a decreasing rate, so I will select 1.10.

Comment: There are other reasonable selections you could make in part (c).
If one does not use the output of the GLM to set ILFs, one needs to go back and make sure the class and territory relativities from a GLM will work well with the ILFs actually selected.
The estimate of the difference between the limited expected values at $250,000 and $500,000 is: $\frac{5.2M - (15)(250K)}{35 + 25 + 15} = \$19,333.$
Therefore, in part (a) the Indicated $500,000 ILF would be:
1 + ($47,000 + $19,333) / $78,400 = 1.846.

3.177. (a) The first column refers to β_1 whether or not we have a male,
the second column refers to β_2 whether or not we are in Territory A,
the third column refers to β_3 the intercept, and thus is all ones.

$$X = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{matrix} \text{A, Male} \\ \text{A, Female} \\ \text{B, Male} \\ \text{B, Female} \end{matrix} \quad \text{or } X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{matrix} \text{A, Male} \\ \text{B, Male} \\ \text{A, Female} \\ \text{B, Female} \end{matrix}$$

$$\text{(b) } Y = \begin{pmatrix} 700/1400 \\ 400/1000 \\ 600/1000 \\ 420/1200 \end{pmatrix} = \begin{pmatrix} 0.50 \\ 0.40 \\ 0.60 \\ 0.35 \end{pmatrix} \begin{matrix} \text{A, Male} \\ \text{A, Female} \\ \text{B, Male} \\ \text{B, Female} \end{matrix} \quad \text{or } Y = \begin{pmatrix} 0.50 \\ 0.60 \\ 0.40 \\ 0.35 \end{pmatrix} \begin{matrix} \text{A, Male} \\ \text{B, Male} \\ \text{A, Female} \\ \text{B, Female} \end{matrix}$$

(c) With a Normal error function and an identity link function, this is the same as a multiple regression.
Assuming $\beta_3 = 0.35$, then the squared error is:

$$1400 (\beta_1 + \beta_2 + 0.35 - 0.5)^2 + 1000 (\beta_2 + 0.35 - 0.4)^2$$
$$+ 1000 (\beta_1 + 0.35 - 0.6)^2 + 1200 (0.35 - 0.35)^2 =$$

$1400 (\beta_1 + \beta_2 - 0.15)^2 + 1000 (\beta_2 - 0.05)^2 + 1000 (\beta_1 - 0.25)^2$.

Setting the partial derivative with respect to β_1 equal to zero:

$0 = 2800(\beta_1 + \beta_2 - 0.15) + 2000(\beta_1 - 0.25). \Rightarrow 4800 \beta_1 + 2800 \beta_2 = 920.$

Setting the partial derivative with respect to β_2 equal to zero:

$0 = 2800(\beta_1 + \beta_2 - 0.15) + 2000(\beta_2 - 0.05). \Rightarrow 2800 \beta_1 + 4800 \beta_2 = 520.$

$\Rightarrow \beta_2 = (520 - 2800 \beta_1) / 4800.$

Plugging back into the first equation: $4800 \beta_1 + 2800 (520 - 2800 \beta_1) / 4800 = 920.$

$\Rightarrow \beta_1 =$ **0.1947**. $\Rightarrow \beta_2 = -0.0052.$

Alternately, without taking into account exposures by cell, the squared error is:
$(\beta_1 + \beta_2 + 0.35 - 0.5)^2 + (\beta_2 + 0.35 - 0.4)^2 + (\beta_1 + 0.35 - 0.6)^2 + (0.35 - 0.35)^2 =$
$(\beta_1 + \beta_2 - 0.15)^2 + (\beta_2 - 0.05)^2 + (\beta_1 - 0.25)^2$.
Setting the partial derivative with respect to β_1 equal to zero:
$0 = 2(\beta_1 + \beta_2 - 0.15) + 2(\beta_1 - 0.25). \Rightarrow 4\beta_1 + 2\beta_2 = 0.8.$
Setting the partial derivative with respect to β_2 equal to zero:
$0 = 2(\beta_1 + \beta_2 - 0.15) + 2(\beta_2 - 0.05). \Rightarrow 2\beta_1 + 4\beta_2 = 0.4.$
$\Rightarrow \beta_2 = (0.4 - 2\beta_1) / 4 = 0.1 - 0.5\beta_1.$
Plugging back into the first equation: $4\beta_1 + 2(0.1 - 0.5\beta_1) = 0.8.$
$\Rightarrow \beta_1 =$ **0.2**. $\Rightarrow \beta_2 =$ **0**.

Alternately, in either case one can fit via maximum likelihood and get the same result as by minimizing the squared errors.

For the Normal Distribution, $f(x) = \frac{\exp[-\frac{(x-\mu)^2}{2\sigma^2}]}{\sigma\sqrt{2\pi}}$. $\ln[f(x)] = -\frac{(x-\mu)^2}{2\sigma^2} - \ln[\sigma] - \ln[2\pi]/2.$

Without taking info account exposures by cell, the loglikelihood is:
$-(\beta_1 + \beta_2 + 0.35 - 0.5)^2 / (2\sigma^2) - (\beta_1 + 0.35 - 0.4)^2 / (2\sigma^2) - (\beta_2 + 0.35 - 0.6)^2 / (2\sigma^2)$
$- (0.35 - 0.35)^2 / (2\sigma^2) - 4 \ln[\sigma] - \ln[2\pi]/2.$
Setting the partial derivative of the loglikelihood with respect to β_1 equal to zero:
$0 = -(\beta_1 + \beta_2 - 0.15) / \sigma^2 - (\beta_1 - 0.25) / \sigma^2. \Rightarrow 2\beta_1 + \beta_2 = 0.4.$
Setting the partial derivative with respect to β_2 equal to zero:
$0 = -(\beta_1 + \beta_2 - 0.15) / \sigma^2 - (\beta_2 - 0.05) / \sigma^2. \Rightarrow \beta_1 + 2\beta_2 = 0.2.$
Solving two equations in two unknowns: $\beta_1 =$ **0.2**, and $\beta_2 =$ **0**.
Setting the partial derivative with respect to σ equal to zero:
$0 = (\beta_1 + \beta_2 - 0.15)^2 / \sigma^3 - (\beta_1 - 0.05)^2 / \sigma^3 - (\beta_2 - 0.25)^2 / \sigma^3 - 4 / \sigma.$
$\Rightarrow \sigma^2 = \{(0.2 + 0 - 0.15)^2 + (0.2 + 0 - 0.05)^2 + (0.2 + 0 - 0.25)^2\} / 4 = 0.006875.$

(d) 1. The Normal Distribution allows negative values, while the Poisson Distribution does not. Since claim frequencies are never negative, the Poisson error structure is preferable here.

2. The Normal error structure assumes that the process variances of the frequency in the four cells are equal. In contrast, a Poisson error structure assumes that the process variances of the frequency in the four cells are equal to their means. I would expect that those cells with higher expected frequencies would have higher process variances than those with lower expected frequencies, and thus would prefer the Poisson error structure to the Normal.

3. The log link function would assume multiplicative relativities, while the identity link function assumes additive relativities. If the relationship is approximately multiplicative, then the log link function would do a better job than the identity link function.

Comment: In parts (a) and (b), the order in which one lists the rows is arbitrary;
it would be a good idea to label what you did.
In part (c), if one includes the exposures in the sum of squared errors, that is equivalent to using exposures as the prior weights in the GLM, or using exposures in an offset term.
Including the exposures is equivalent to doing a weighted multiple regression.
The observed frequencies are:

Gender	Territory A	Territory B	Difference
Male	0.50	0.60	0.10
Female	0.40	0.35	-0.05
Difference	-0.10	-0.25	

The differences between territories are not similar for the two genders.
The differences between genders are not similar for the two territories.

Gender	Territory A	Territory B	Ratio
Male	0.50	0.60	1.2
Female	0.40	0.35	0.875
Ratio	0.80	0.583	

The ratios between territories are not similar for the two genders.
The ratios between genders are not similar for the two territories.
Thus perhaps neither an additive nor a multiplicative relationship is appropriate.

3.178. Based solely on the given output of the GLM, it makes sense to add credit score to the homeowners rating plan for the wind peril. The ± 2 standard deviation bands around the indicated relativities for fair and poor each do not contain one; the indicated frequency relativities are statistically significantly different than 1. However, I would also want to see an analysis of pure premiums.
Countrywide there are only about 45,000 exposures in the fair category and 15,000 exposures for the poor category. This raises concerns about the credibility of the data from those classes. (The vast majority of the exposures are in the good category. Perhaps some other breakdown of scores into categories would be better.)
I question whether there is casual relationship between credit scores and claim frequency from wind.
This is a countrywide study. Could it be that the average credit scores may vary by state, with those states with higher average wind losses also having lower average credit scores?
(On the other hand, perhaps those with poorer credit scores are less likely to properly maintain the roof of their house, leading to some wind claims that would not have been otherwise made. It is important during hurricanes that the roof remain intact and attached to the home.)
Overall, I would recommend that the variable not be added (at this time) based on the lack of causality and the lack of reliable relativities due to the small volume of data for only one year. More data as well as more analysis is needed.
Comment: One should not spend much time commenting on the general issue of using credit scores in rating insurance. I think it should be sufficient to discuss the issue of causality, whether or not there is a logical connection between credit scores and wind losses.
On the general issue of using credit scores in rating insurance:
1. Assuming the insurer writes a reasonable amount of business and credit scores are grouped into intervals that are not tiny, there should be enough data in each rating group to measure costs with sufficient accuracy. The criterion of credibility is fulfilled.
2. Insureds with similar premiums after the use of credit scores have a range of expected costs, just as with any other rating variable. However, the use of rating scores decreases this variation and thus improves the homogeneity.
3. Studies have shown that credit scores are correlated with insurance costs.
Credit scores have been used for several years and the relationship to costs has been reasonably stable over time. Thus the criterion of statistical significance is fulfilled.
4. There are errors in credit reports. Individuals can get copies of their credit reports and try to get the credit bureau to correct any errors. However, the information in the credit report are not subject to manipulation or lying by the insured. The criterion of verifiability is fulfilled.
5. There is considerable expense in obtaining credit reports and turning them into credit scores to use for rating insurance. Either the insurer will incur that cost or pay someone else to do this work. In either case the criterion of low administrative costs is not fulfilled.
6. One can construct credit scores for use in rating insurance using objective definitions, with little ambiguity. Class definitions based on ranges of credit scores can be mutually exclusive and exhaustive. There should not be much administrative error, as the credit scores can be calculated by computer. The criterion of objectivity can be fulfilled.
7. Since when they apply for a home mortgage or a car loan, their credit reports are examined, it is not an issue when these same reports are used for insurance. The criterion of privacy is fulfilled.
8. Both high and low income insureds have good and bad credit reports. The effect of using credit scores should not be correlated with income. The criterion of affordability is fulfilled.

9. The items recorded in a credit report, such as a late payment of a bill, are not responsible for differences in insurance costs. The criterion of causality is not fulfilled.
10. An insured can modify his behavior in order to improve his credit report in the future. The criterion of controllability is fulfilled.

3.179. (a) The indicated frequencies differ significantly by hazard group. (We are not given information in order to determine whether these differences are statistically significant.) Indicated relativities generally increase with Hazard Group, with the exceptions of Hazard Groups A and G which have much less data than the others. The separate indications for the three years are consistent, with the exceptions of Hazard Groups A and G which have much less data than the others. Therefore, I conclude that hazard group is useful for predicting expected frequency.

3.180. (a) ϕ is the scale or dispersion parameter, which scales the variance.
ω_i is a (prior) weight, representing the amount of data we have for observation i; the variance is inversely proportional to the volume of data.
(b) i. Gamma Distribution is most commonly used to model the error structure for severity; it works well in many situations based on diagnostics.
The Gamma is continuous with support from zero to infinity.
The gamma distribution also has an intuitively attractive property for modeling claim amounts since it is invariant to measures of currency. In other words measuring severities in dollars and measuring severities in cents will yield the same results using a gamma multiplicative GLM. (This is not true of some other distributions such as Poisson, but would be for the Inverse Gaussian.)
For the Gamma: $V(\mu_i) = \mu_i^2$.
ii. For policy renewal a Bernoulli or Binomial is used, since policy renewal is a yes/no process.
For the Bernoulli: $V(\mu_i) = \mu_i (1 - \mu_i)$.
For the Binomial representing μ trials (μ policies): $V(\mu_i) = \mu_i (1 - \mu_i) / m$.
(c) 1. For severity, ω_i would be the number of claims, the measure of how much data we have.
2. For policy renewal, if using the Bernoulli, ω_i would be the number of policies.
If using the Binomial, $\omega_i = 1$.

3.181. Smaller Bayesian Information Criterion is better.
BIC = -2 (maximum loglikelihood) + p ln(n),
where n = 1000 is the sample size and p is the number of parameters.
Since the deviance = (2) (saturated max. loglikelihood - maximum likelihood for model),
we can compare between the models: Deviance + p ln(n) = Deviance + p ln(1000).
(The maximum likelihood for the saturated model is the same in each case.)

Model #	p	Deviance	Deviance + p ln(1000)
1	2	1085.0	1098.82
2	3	1084.8	1105.52
3	3	1083.0	1103.72
4	4	1081.9	1109.53
5	5	1081.6	1116.14

The smallest Deviance + p ln(n) is for **Model 1**.

3.182. Smaller Akaike Information Criterion is better.
AIC = -2 (maximum loglikelihood) + (number of parameters)(2).
Since the deviance = (2) (saturated max. loglikelihood - maximum likelihood for model),
we can compare between the models:
Deviance + p 2 = Deviance + (number of parameters)(2).
(The maximum likelihood for the saturated model is the same in each case.)

Model #	Number of Parameters	Deviance	Deviance + (number of parameters)(2)
1	2	1085.0	1089.0
2	3	1084.8	1090.8
3	3	1083.0	1089.0
4	4	1081.9	1089.9
5	5	1081.6	1091.6

The smallest AIC is a tie between **Model 1 and Model 3**.

3.183. Estimated mean severity for a rural male is: exp[2.32 - 0.64 + 0.76] = 11.473.
For the Gamma Distribution, Var[Y] = $\phi\mu^2$ = (2) (11.473^2) = **263.3**.

3.184. exp[βx] = exp[-1.485 + 0 - 1.175 - 0.101] = $e^{-2.761}$ = 0.06323.
For the logit link function: $\mu = \frac{e^{\beta x}}{e^{\beta x} + 1}$ = 0.06323 / (0.06323 + 1) = **5.95%**.

3.185. μ = exp[-2.633 + 0.132 + 0] = **0.07957**.

3.186.

Variable	Number of Parameters
Vehicle Price	4
Driver age	2 - 1 = 1
Number of drivers	4 - 1 = 3
Gender	2 - 1 = 1
Interaction Gender & Driver Age	1

Maximum number of parameter is: 4 + 1 + 3 + 1 + 1 = **10**.

Comment: A model with only Vehicle Price would involve: $\beta_0 + \beta_1 (vp) + \beta_2 (vp)^2 + \beta_3 (vp)^3$.

The interaction of gender and driver age only uses one parameter since each of gender and driver age only use one parameter.

3.187. Smaller AIC is better, so we prefer Model 1.

$\exp[\beta x] = \exp[-3.264 + (12)(0.212) + 0.727] = e^{0.007} = 1.007$.

For the logit link function: $\mu = \frac{e^{\beta x}}{e^{\beta x} + 1} = 1.007 / (1.007 + 1) =$ **50.2%**.

3.188. Let x be the number of additional parameters for the new model.

Let ℓ_1 be the loglikelihood for the original model, and ℓ_2 be the loglikelihood for the model including the new variable.

Deviance = (2) (saturated max. loglikelihood - maximum likelihood for model).

Thus the change in model deviance is: $-2(\ell_2 - \ell_1) = -53$.

AIC = (-2) (maximum loglikelihood) + (number of parameters)(2).

Thus the change in AIC is: $(-2)(\ell_2 - \ell_1) + 2x = -53 + 2x = -47. \Rightarrow x = 3$.

BIC = (-2) (maximum loglikelihood) + (number of parameters) ln(number of data points).

Thus the change in BIC is: $(-2)(\ell_2 - \ell_1) + x \ln(n) = -53 + 3 \ln(n) = -32. \Rightarrow n = e^7 =$ **1097**.

3.189. We have to assume equal exposures in each of the four cells.
The mean modeled frequencies are:

	State A	State B
Male	$\exp[\beta_1 + \beta_3]$	$\exp[\beta_1]$
Female	$\exp[\beta_2 + \beta_3]$	$\exp[\beta_2]$

The loglikelihood ignoring terms that do not depend on the betas is:
$-\exp[\beta_1 + \beta_3] + 0.0920 (\beta_1 + \beta_3) - \exp[\beta_2 + \beta_3] + 0.1500 (\beta_2 + \beta_3)$
$- \exp[\beta_1] + 0.0267 \beta_1 - \exp[\beta_2] + 0.0500 \beta_2$.
Setting the partial derivative of the loglikelihood with respect to β_1 equal to zero:
$-\exp[\beta_1 + \beta_3] + 0.0920 - \exp[\beta_1] + 0.0267 = 0$.
Given $\beta_3 = 1.149$: $-\exp[\beta_1] e^{1.149} + 0.0920 - \exp[\beta_1] + 0.0267 = 0$.
$\Rightarrow \exp[\beta_1] = (0.0920 + 0.0267) / (1 + e^{1.149}) = 0.02857$.
$\Rightarrow \exp[\beta_1 + \beta_3] = 0.02857 e^{1.149} =$ **0.0901** = expected frequency of a male risk in State A.
Comment: Similar to 8, 11/13, Q.2c.
What the exam questions calls "the likelihood function" is the loglikelihood function.
$\hat{\beta}_1 = \ln(0.02857) = -3.555$.
Setting the partial derivative of the loglikelihood with respect to β_2 equal to zero:
$-\exp[\beta_2 + \beta_3] + 0.1500 - \exp[\beta_2] + 0.0500$.
Given $\beta_3 = 1.149$: $-\exp[\beta_2] e^{1.149} + 0.1500 - \exp[\beta_2] + 0.0500 = 0$.
$\Rightarrow \exp[\beta_2] = (0.1500 + 0.0500) / (1 + e^{1.149}) = 0.04813$. $\Rightarrow \hat{\beta}_2 = -3.034$.
Using a computer, without being given β_3, the maximum likelihood fit is:
$\hat{\beta}_1 = -3.5555$, $\hat{\beta}_2 = -3.0338$, and $\hat{\beta}_3 = 1.1490$.
The mean modeled frequencies are:

	State A	State B
Male	exp[-3.5555 + 1.1490] = 9.01%	exp[-3.5555] = 2.86%
Female	exp[-3.0338 + 1.1490] = 15.19%	exp[-3.0338] = 4.81%

3.190. In order to solve for the unknown intercept, we use the given probability of accident for a driver in age group 2, from area C and with vehicle body type Other.
0.22 = exp[x + 0.064 - 0.371] / {exp[x + 0.064 - 0.371] + 1}.
$\Rightarrow$ exp[x + 0.064 - 0.371] = 0.22 / (1 - 0.22) = 0.28205.
$\Rightarrow$ x + 0.064 - 0.371 = ln[0.28205] = -1.2657. $\Rightarrow$ x = -0.9587.
Thus for a driver in age group 3, from area C and with vehicle body type Sedan, the odds (ratio) is: π / (1 - π) = exp[-0.9857 + 0 + 0 + 0] = **0.3834**.
Comment: The probability of having an accident for a driver in age group 3, from area C and with vehicle body type Sedan is: 0.3834 / (1 + 0.3834) = 0.277.
Note that 0.277 / (1 - 0.277) = 0.383.

3.191. $\frac{\text{Exp}[-2.358 + 0.905]}{1 + \text{Exp}[-2.358 + 0.905]}$ = **0.190**.

3.192. exp[2.100 + 1.336 + 1.406 + 1.800] = **766.63**.

3.193. For an observation from Zone 4, with Vehicle Class Sedan and Driver Age Middle age, the mean is: exp[2.1] = 8.166.
For the Gamma Distribution the variance is: $\phi \mu^2$ = (1) (8.166^2) = **66.7**.

3.194. $\frac{\exp[1.530 + 0.735 - 0.031]}{1 + \exp[1.530 + 0.735 - 0.031]}$ = **90.33%**.

3.195. Since Model 2 has one fewer parameter than model 3,
model 2 has 9 degrees of freedom.
AIC = (-2) (maximum loglikelihood) + (number of parameters)(2).
BIC = (-2) (maximum loglikelihood)) + (number of parameters) ln(number of data points).
Therefore from Model 1: 95,473.61182 = (-2)(-47,704) + (5) ln(n). $\Rightarrow$ n = 500,000.
For Model 2, AIC = -47,495 + (9)(2).
For Model 2, BIC = -47,495 + (9) ln(500,000).
The absolute difference between the AIC and the BIC for Model 2 is:
l (9) ln(500,000) - 18 l = **100.1**.

3.196. Graph one shows an increasing variance with fitted value.
Homoscedasticity would be constant variance, so statement 1 is false; statement 2 is true.
The residuals in Graph 2 are not symmetric around zero; there are more extreme positive values than there are extreme negative values. This indicates that the residuals are not normally distributed.
Statement 3 is true.
Comment: In Graph 2 it is not clear the meaning of the horizontal lines.
A Normal Q-Q Plot would have been much more useful than Graph 2.

3.197. The model with the smallest AIC is usually the best model in model selection process, all other things being equal. Statement A is not true.
The model with the smallest BIC is usually the best model in model selection process, all other things being equal. Statement B is not true.
The model with the smallest deviance is usually the best model in model selection process, all other things being equal. Statement C is not true.
AIC = (-2) (maximum loglikelihood) + (number of parameters)(2).
BIC = (-2) (maximum loglikelihood)) + (number of parameters) ln(number of data points).
The penalty for AIC is (number of parameters)(2).
The penalty for BIC is (number of parameters) ln(number of data points).
So the penalties are equal for: 2 = ln(number data Points). $\Rightarrow$ number of data points = $e^2 = 7.4$.
Thus, other things equal, when number of observations ≥ 8, BIC penalizes more for the number of parameters used in the model than AIC. **Thus statement E is true**.
Comment: Since statements D and E are opposites, it is likely that one of them is true.

3.198. Change in AIC is: (2) (number of parameters added).
Change in BIC is: ln(1500) (number of parameters added).
We want: ln(1500) (number of parameters added) > (2) (number of parameters added) + 25.
$\Rightarrow$ Number of parameters added > 4.7. $\Rightarrow$ Number of added parameters is at least 5.
$\Rightarrow$ Minimum possible number of levels in the new categorical variable is: 5 + 1 = **6**.

3.199. 100,000 exp[-15 - 1.2 + (0.15)(25) + (0.004)(25^2) + (0.012)(25)]
= 100,000 $e^{-9.65}$ = **6.44** deaths.

3.200. i. Where the variable in question relates to a policy option selected by the insured, having its factor reflect anything other than the excess losses due to higher limit is not a good idea. One can get counterintuitive results such as charging less for more coverage.
Even if the indicated result is not counterintuitive, to the extent that the factor differs from the pure effect on loss potential, it will affect the way insureds choose coverage options in the future. Thus, the selection dynamic will change and the past results would not be expected to replicate for new policies. For this reason it is recommended that factors for coverage options such as increased limit factors be estimated outside the GLM, using traditional actuarial techniques. (The resulting factors should then be included in the GLM as an offset.)

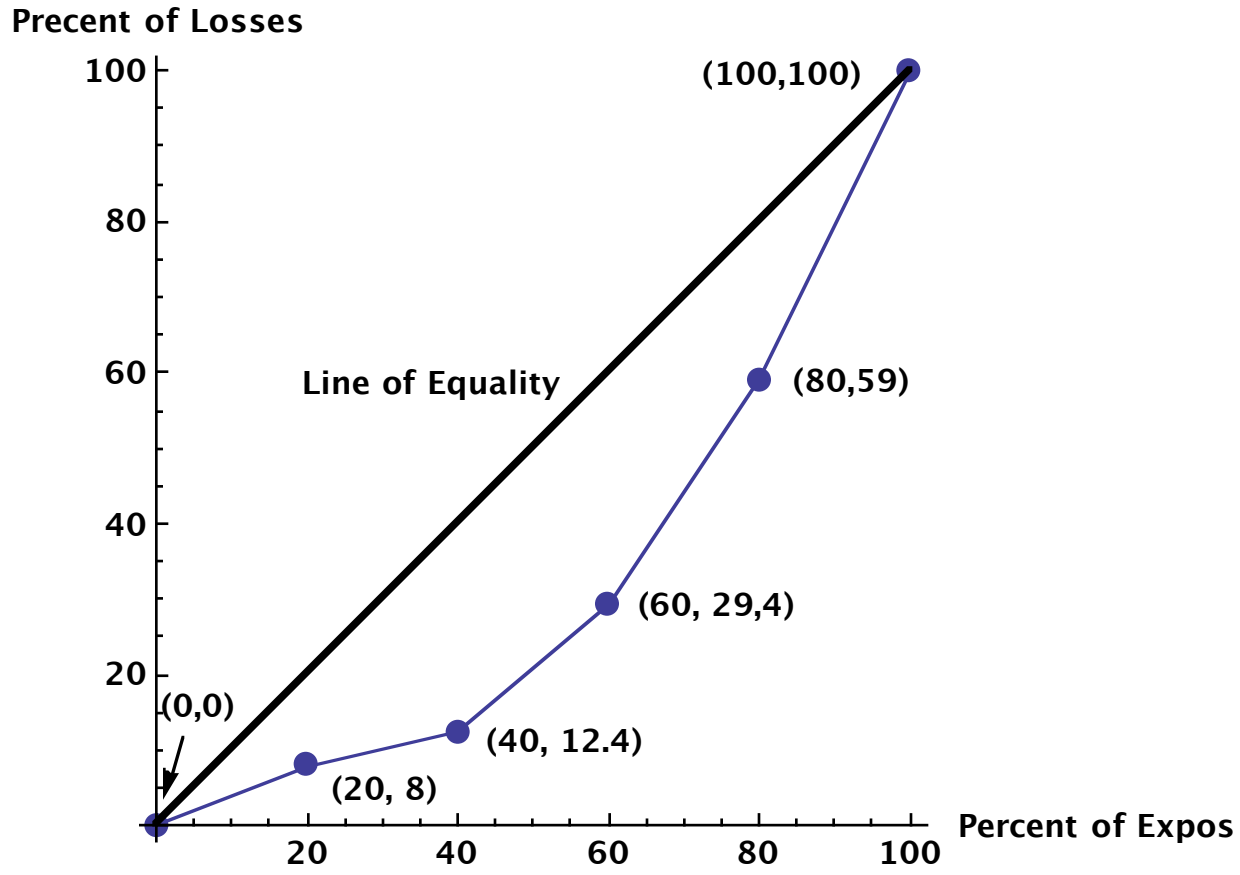
ii. I assume what is intended is that the number of coverage changes during the current policy period will be used to help rate the policy during its next policy period. (We are not given any information on whether the number of coverage changes in a policy period is related to the insurance costs the following period compared to otherwise similar insureds.)
The number of changes during a given policy period is not a good classification variable.
It is something that is likely to be zero for many policy periods, and vary somewhat randomly over time. If those with more coverage changes are changed more it is unlikely to be acceptable to insurance regulators and the public. If those with more coverage changes are changed more, then it will give insureds less incentive to make necessary coverage changes during a policy period; some of these coverage changes would have resulted in additional premiums for the insurer.
Alternately, the information will not be available for new business since we are building a GLM for the prospective period.
Alternately, the number of coverage changes is likely to change from what it is in the current policy period and thereafter year by year.
iii. Territories are not a good fit for the GLM framework. You may have thousands of zipcodes to consider and aggregating them to a manageable level will cause you to lose a great deal of important signal. If one does not aggregate the large number of zipcodes, then there are too many parameters which can lead to overfitting.
Using a spatial smoothing technique would be a more appropriate technique; one would then include the value determined for ZIP code as an offset term in the GLM.

(b) 1. One can get counterintuitive results such as charging more for less coverage.
2. Even if the indicated result is not counterintuitive, to the extent that the factor differs from the pure effect on loss potential, it will affect the way insureds choose coverage options in the future. Thus, the selection dynamic will change and the past results would not be expected to replicate for new policies.
3. Deductibles should lower frequency (small losses below deductible not reported) but usually increase severity (since claims that do get reported are higher average cost). This violates the assumption for the Tweedie Distribution, that a lower pure premium is due to both a lower frequency and a lower severity.
(c) One can calculate deductible relativities from loss elimination ratios.
Deductible Relativity = (1 - LER for chosen deductible) / (1 - LER for Base Deductible).
Loss elimination ratios can be estimated from size of loss data.
Loss elimination ratio = (Limited Expected Value at Deducible Amount) / Mean.
In the GLM, one would then include an offset of ln[deductible relativity].
Comment: While the average size of non-zero payment, equal to the mean residual life, usually increases as the size of deductible increases, this is not always the case.
Deductible factors may produce higher relativities at higher deductibles due to factors other than pure losses elimination:
1. Insureds at high loss potential and high premiums may be more likely to elect high deductibles in order to reduce their premium.
2. Underwriters may force high deductibles on riskier insureds.

3.201. a. Sort the risks from best to worst based on the model predicted loss.

Risk	Model Predicted Loss	Actual Loss	Cumulative Losses	% of Losses
5	200	400	400	8.0%
2	500	220	620	12.4%
4	800	850	1470	29.4%
3	1,500	1,480	2950	59.0%
1	2,000	2,050	5000	100.0%
Total		5000		

On the x-axis, plot the cumulative percentage of exposures.
I will assume that each risk has the same number of exposures.
On the y-axis, plot the cumulative percentage of actual losses.



b. The Gini index is twice the area between the Lorenz Curve and the line of equality.
The higher the Gini Index, the better the rating plan is at identifying risk differences, in other words the rating plan has more lift.
"The Gini index can also be used to measure the lift of an insurance rating plan by quantifying its ability to segment the population into the best and worst risks."
Comment: See Section 7.2.4 including Figure 21 in Goldburd, Khare, and Tevet.
Usually, one would be working with thousands of risks.

3.202. (a) $\exp[0.910 + (3)(0.013) + \ln[25{,}000]\ (-0.187) + (8)(0.062)] = e^{-0.4357} =$ **64.7%.**
(b) One could take the coefficients of the new business model as a given, other than b_0, which will be re-estimated.
Let the prior year claim count be x for renewal business.
Then the renewal business model is:
$\mu = \exp[\beta_0 + 0.013\ age + (-0.187)\ logprem + 0.62\ locont + \beta_4\ x]$.
We would fit the model via maximum likelihood to the data for renewal business, taking into account the form of density for the Tweedie Distribution.
Alternately, one can fit a single model to the data for new and renewal business.
Let the prior year claim count be x for renewal business
Let D = 0 if new business and 1 if renewal business.
Then the combined model is: $\mu = \exp[\beta_0 + \beta_1\ age + \beta_2\ logprem + \beta_3\ locont + D\ \beta_4 + D\ \beta_5\ x]$.
We would fit the model via maximum likelihood to the combined data, taking into account the form of density for the Tweedie Distribution.
(c) 1. Time-consistency. One can fit the model to the data for separate years and compare the coefficients. If the fitted coefficients are similar, that indicates stability over time.
Alternately, one could introduces dummy variables into the model for the various years of data.
For example, if we have data from 2012, 2013 and 2104,
then we could take 2012 = base year, $x_5 = 1$ if 2013, $x_6 = 1$ if 2014.
Then test whether the coefficients of these variables are significantly different from zero. If one or more of the fitted coefficients are significantly different than zero, that indicates instability over time.
2. Bootstrapping. Create multiple datasets from the initial dataset by sampling with replacement. Run the model on each sampled set. Assess stability of estimates of coefficients by comparing the results from each run.
3. Cross-Validation. Split the data into k parts and run the model on the (k-1) parts, then validate the result on the remaining part. Compare how similar the estimates are from the k iterations to assess variable stability.
4. Validation on Holdout Dataset. Split the data into two subsets, training and holdout. Determine the best model on the training set. Ideally, this model should fit well the holdout data.
5. Cook's Distance. Sort the observations based on their Cook's Distance value (higher distance = more influence on the model.) Remove one or more of the most influential observations and rerun the model on this new set of data to see the effect on estimated parameters.

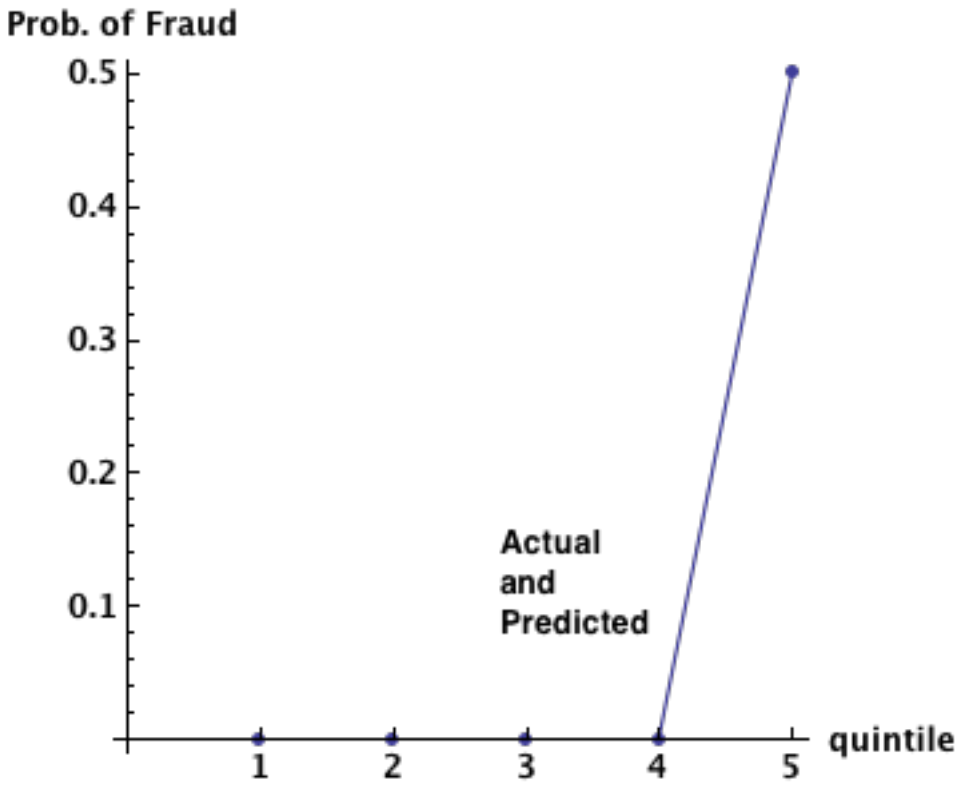
3.203. (a) For the base model:
AIC = (-2)(-750) + (2)(10) = **1520**.
BIC = (-2)(-750) + 10 ln[1,000,000] = **1638.2**.
For the new model:
AIC = (-2)(-737.5) + (2)(15) = **1505**.
BIC = (-2)(-737.5) + 15 ln[1,000,000] = **1682.2**.
(b) AIC is preferable. As here, most actuarial models involve a lot of data points. Therefore, the penalty for more parameters is very large for the BIC. Using BIC will tend to result in too simple models. In contrast, AIC does not depend on the number of data points.
(c) Based on part (b), I will rely on AIC.
Smaller AIC is better, so I will recommend the new model.
Comment: See Section 6.2.2 in Goldburd, Khare, and Tevet.
If one instead relied on BIC, the base model would be preferred.
Deviance = 2 (loglikelihood of saturated model - loglikelihood of model).
Thus equivalently to using AIC, one could compare models using: Deviance + 2p.
For the base model, Deviance + 2p = 500 + (2)(10) = 520.
For the new model, Deviance + 2p = 475 + (2)(15) = 505.
Since 505 < 530, we prefer the new model based on this criterion.
Equivalently to using BIC, one could compare models using: Deviance + p ln[N]..
For the base model, Deviance + 2p = 500 + 10 ln[1 million] = 638.16.
For the new model, Deviance + 2p = 475 + 15 ln[1 million] = 682.23.
Since 638.16 < 682.23, we prefer the base model based on this criterion.

3.204. (a) 1. Attempting to test the performance of any model on the same set of data on which the model was built will produce overoptimistic results. Using the training data to compare this model to any model built on different data would give our model an unfair advantage.
2. As we increase the complexity of the model, the fit to the training data will always get better. In contrast, for data the model fitting process has not seen, additional complexity may not improve the performance of a model; as the model gets more complex its performance on the holdout data (test data) will eventually get worse, as shown in the figure in this question.
(b) Model 2 has the right balance, since it has the smallest test MSE.
Model 1 is too simple (fewer degrees of freedom than Model 2), while model 3 is too complex (more degrees of freedom than Model 2).
(c) "Out-of-time validation is especially important when modeling perils driven by common events that affect multiple policyholders at once. An example of this is the wind peril, for which a single storm will cause many incurred losses in the same area. If random sampling is used for the split, losses related to the same event will be present in both sets of data, and so the test set will not be true unseen data, since the model has already seen those events in the training set. This will result in overoptimistic validation results. Choosing a test set that covers different time periods than the training set will minimize such overlap and allow for better measures of how the model will perform on the completely unknown future."
Alternately, as in Couret and Venter, one may select either the even or odd years of data as the training set and the other as the holdout set, in order to be neutral with respect to trend and maturity.
Comment: See Section 4.3 of Generalized Linear Models for Insurance Rating.
The figure shown is very similar to Figure 7 in Generalized Linear Models for Insurance Rating. We are interested in how the GLM will perform at predicting the response variable on some future set of data rather than on the set of past data with which we are currently working.
Our goal in modeling is to find the right balance where we pick up as much of the signal as possible with minimal noise, represented in this case by Model 2.

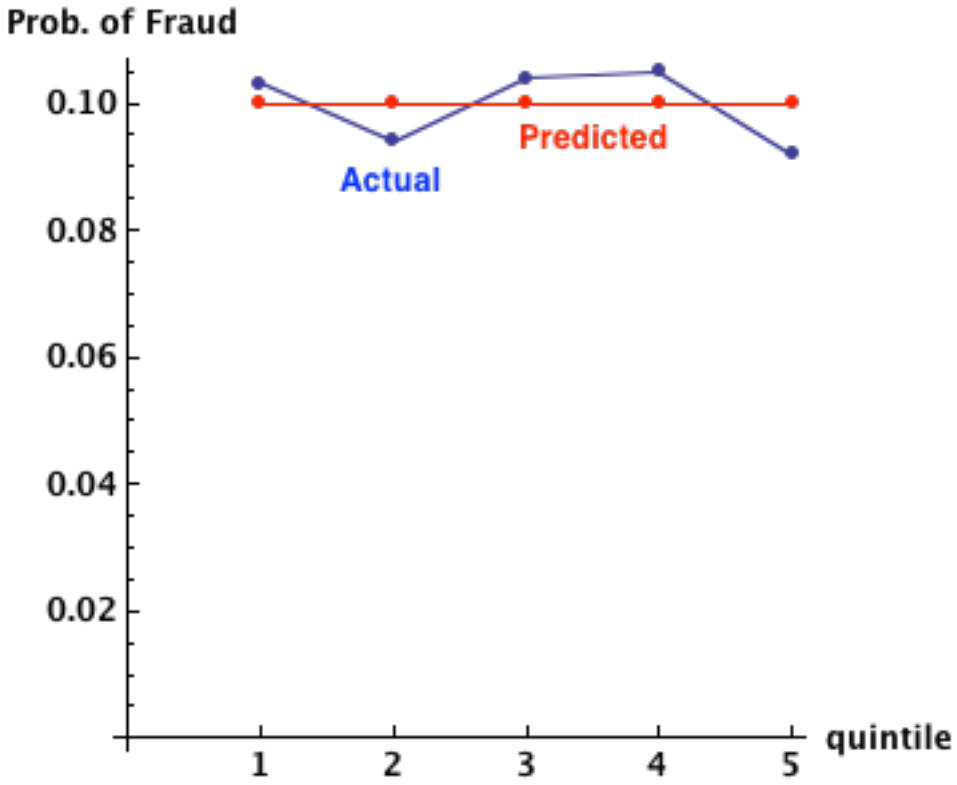
3.205. A simple quintile plot is a simple quantile plot with 5 buckets.

- Sort the dataset based on the model predicted fraud rate from smallest to largest.
- Group the data into 5 buckets with equal volume. (In this case 2000 claims in each.)
- Within each group, calculate the average predicted fraud rate based on the model, and the average actual fraud rate.
- Plot for each group, the actual fraud rate and the predicted fraud rate.

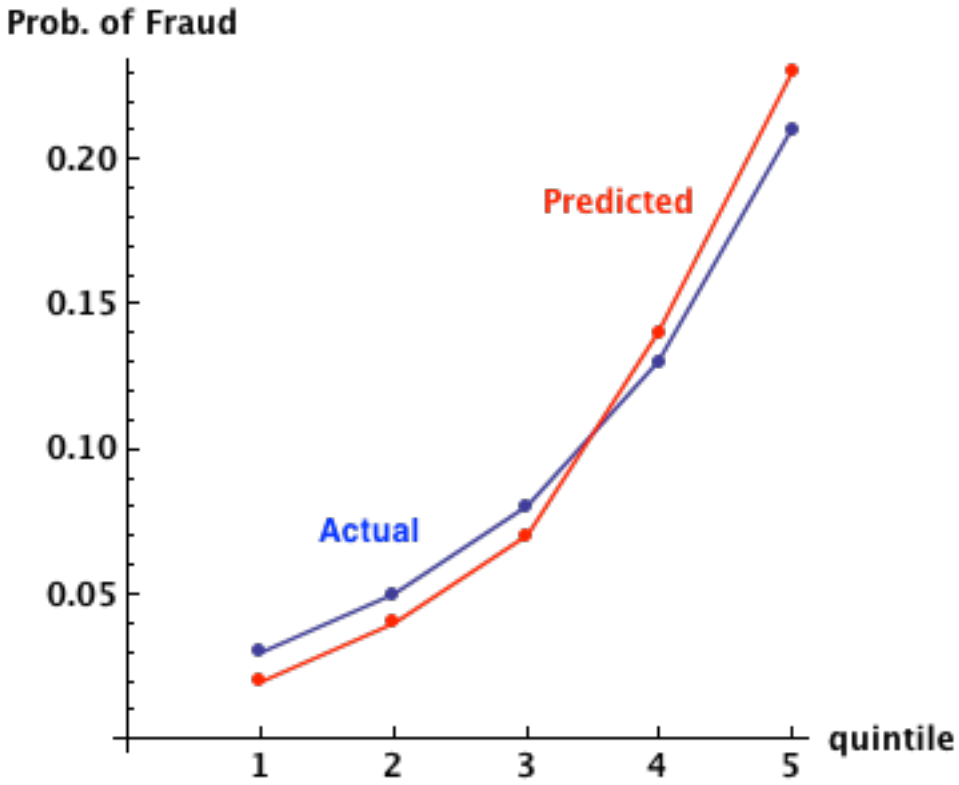
The saturated model has as many predictors as data points. Thus for the saturated model, the predictions exactly match the observations for each claim. In this case, 1000 of the claims involve fraud, and would all be placed in the last quintile. Thus the last quintile would consist of 1000 claims with fraud and 1000 claims without fraud.
The simple quintile plot:



The null model, has no predictors, only an intercept. Thus for the null model the prediction is the same for every record: the grand mean.
In this case, the overall probability of fraud is: 1,000/10,000 = 10%.
Since every risk has the same prediction, one would assign them to buckets at random.
Thus all of the actuals by quintile should be close to the grand mean, with small differences due to the randomness of assignments. The simple quintile plot:



"A model that could be used in practice", would have the actuals increase monotonically, have good but not perfect predictive accuracy, and a reasonably large vertical distance between the actuals in the first and last quintiles. A simple quintile plot:



Comment: See Section 7.2.1 and page 59 of "GLMs for Insurance Rating."
Combines separate ideas in the syllabus reading.
There are many possible examples of the last plot.
Since the records are ordered by predicted values, the records in each bucket change for each graph. Thus, actuals are not the same for each graph.
Quintile plots are sorted by predicted values from smallest to largest value. Thus the predicted values must be monotonically increasing (or in the case of the null model equal). Actuals need not be monotonically increasing, although that is desirable.
In every graph, the average of the actuals should be the grand mean of 10%.
In the final plot, the average of the predicteds should be close to if not equal to 10%; the GLM may have a small bias.
In the final plot, the predicted and actuals for the final quintile should each be less than the 50% in the saturated model. In the final plot, the predicted and actuals for the final quintile should each be more than the 10% in the null model.

3.206.

Claim #	Fraud	25% Threshold Predict.		50% Threshold Predict.	
1	Y	N	False Neg.	N	False Neg.
2	N	N	True Neg.	N	True Neg
3	N	N	True Neg.	N	True Neg.
4	N	Y	False Pos.	Y	False Pos.
5	Y	Y	True Pos.	Y	True Pos.
6	Y	Y	True Pos.	N	False Neg.
7	N	N	True Neg.	N	True Neg.
8	Y	Y	True Pos.	Y	True Pos.
9	N	Y	False Pos.	Y	False Pos.
10	N	Y	False Pos.	N	True Neg.

(a)

25% Threshold

	Predicted		
Actual	Fraud	No Fraud	Total
Fraud	true pos.: 3	false neg.: 1	4
No Fraud	false pos.: 3	true neg.: 3	6
Total	6	4	10

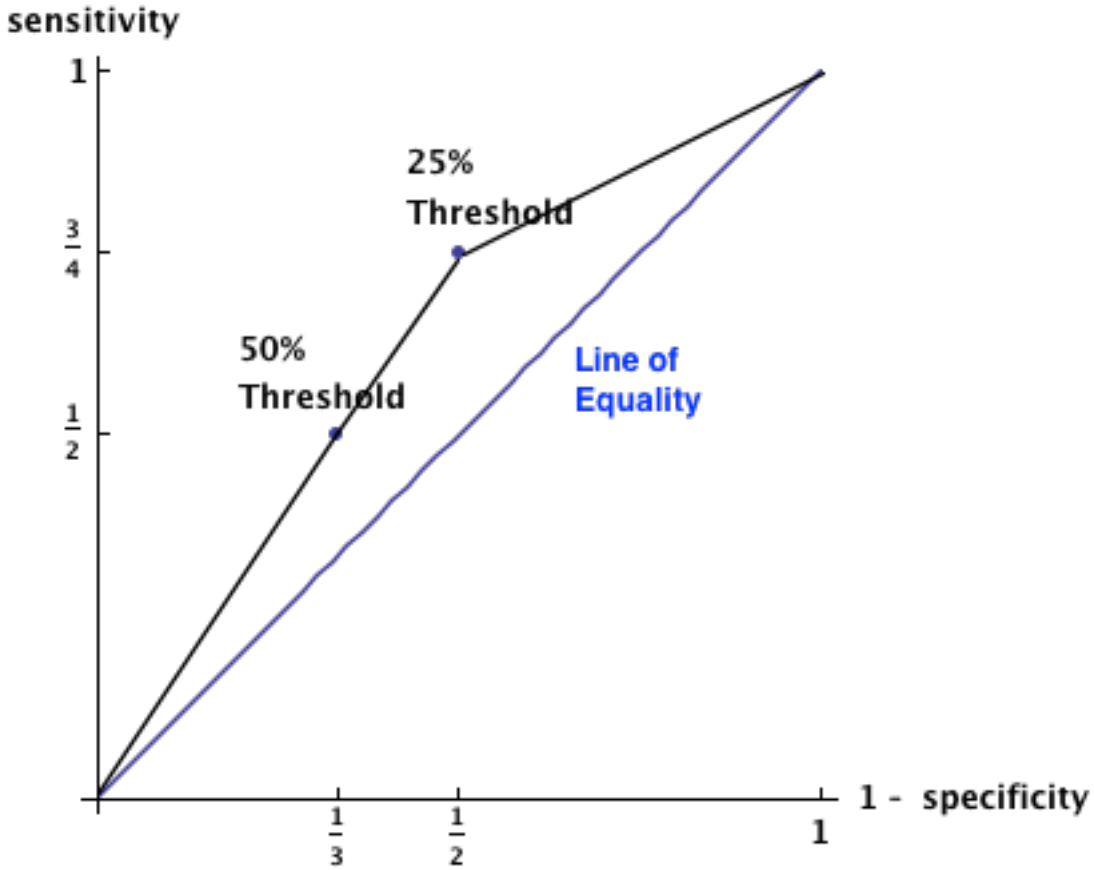
50% Threshold

	Predicted		
Actual	Fraud	No Fraud	Total
Fraud	true pos.: 2	false neg.: 2	4
No Fraud	false pos.: 2	true neg.: 4	6
Total	6	4	10

(b) Sensitivity = $\frac{\text{True Positives}}{\text{Total Number of Events}} = \frac{\text{Correct Predictions of Fraud}}{\text{Total Number of Fraudulent Claims}}$.

Specificity = $\frac{\text{True Negatives}}{\text{Total Number of Non-Events}} = \frac{\text{Correct Predictons of No Fraud}}{\text{Total Number of Nonfraudulent Claims}}$.

25% threshold: sensitivity = 3/4, and specificity = 3/6 = 1/2. Graph (1 - 1/2, 3/4).
50% threshold: sensitivity = 2/4 = 1/2, and specificity = 4/6 = 2/3. Graph (1 - 2/3, 1/2).
The ROC Curve, plus the 45-degree comparison line:



(c) Using a 25% threshold results in more predictions of fraud than using a 50% threshold. Therefore, the 25% threshold has greater sensitivity, more true positives, which is good; however, this is at the cost of lower specificity, more false positives, which is bad.
Alternately, Advantage: You will catch more actual fraud claims because you will have a higher true positive rate. Disadvantage: You will have a higher false positive rate as well, which means you will waste resources to review claims that are not fraudulent.

(d) There are few claims, but they are large. Thus we are very willing to spend money investigating claims for possible fraud; we do not want to miss any true positives and are willing to live with false positives. Therefore, we would prefer the lower threshold of 25%, which has greater sensitivity.
Alternately, a threshold of 0.25 is more appropriate. The high severity makes the cost of not investigating a fraudulent claim very high. The low frequency means that the number of additional claims that will need to be investigated is not very large. The cost of investigating these few additional claims is far less than the cost of potentially missing a few fraudulent claims at a higher discrimination threshold.
Comment: See Table 13 and Figure 22 in "GLMs for Insurance Rating."
According to the CAS Examiner's Report, in part (a) one was required to show a table similar to the one I have, showing the origin of the true positives, false positives, true negatives, and false negatives.