ACTEX

LEARN TODAY. LEAD TOMORROW.

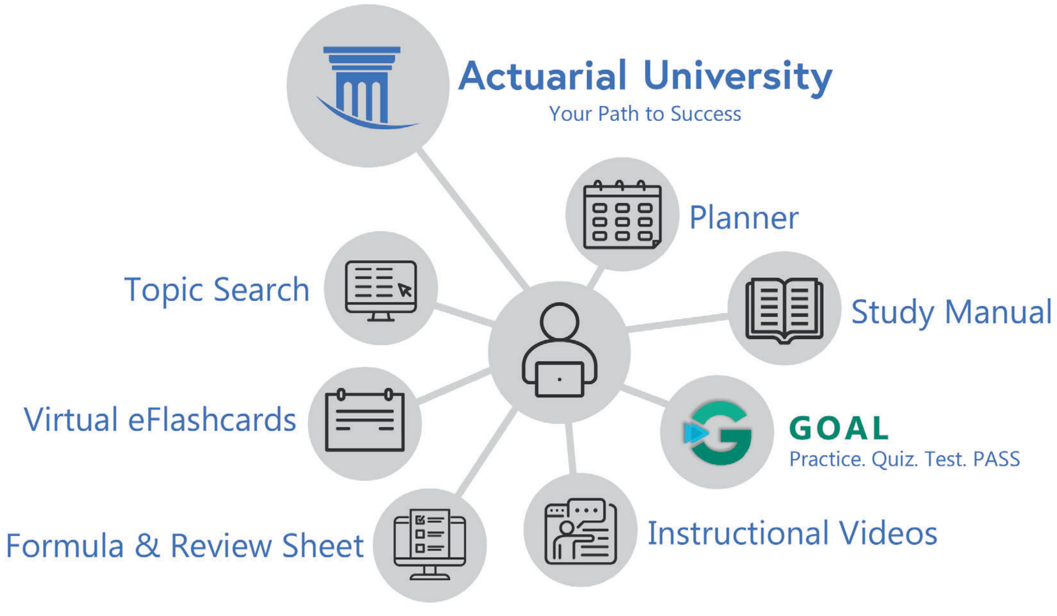
EXAM SRM

Study Manual for SOA Exam SRM

8th Edition

by
Runhuan Feng, Ph.D., FSA, CERA
Daniël Linders, Ph.D.
Ambrose Lo, Ph.D., FSA, CERA

Your Integrated Study Manual Program Includes:



Actuarial University
Your Path to Success

Planner

Topic Search

Study Manual

Virtual eFlashcards

GOAL
Practice. Quiz. Test. PASS

Formula & Review Sheet

Instructional Videos

A
C
T
E
X

Study Manual for SOA Exam SRM

8th Edition

by
Runhuan Feng, Ph.D., FSA, CERA
Daniël Linders, Ph.D.
Ambrose Lo, Ph.D., FSA, CERA

Your Integrated Study Manual Program Includes:

ACTEX Learning | Learn Today. Lead Tomorrow.

ACTEX Learning
Learn Today. Lead Tomorrow.

Actuarial & Financial Risk Resource Materials
Since 1972

Welcome to Actuarial University

Actuarial University is a reimagined platform built around a more simplified way to study. It combines all the products you use to study into one interactive learning center.

You can find integrated topics using this network icon.

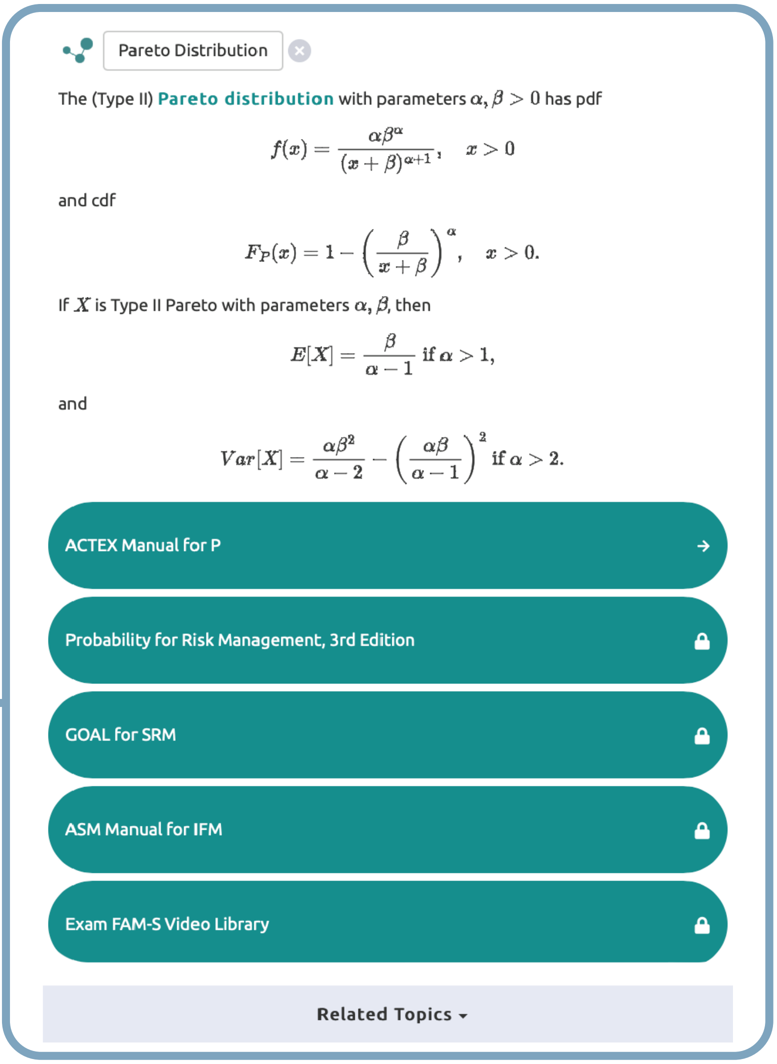
When this icon appears, it will be next to an important topic in the manual. Click the **link** in your digital manual, or search the underlined topic in your print manual.

1. Login to: www.actuarialuniversity.com

2. Locate the **Topic Search** on your exam dashboard and enter the word or phrase into the search field, selecting the best match.

3. A topic "**Hub**" will display a list of integrated products that offer more ways to study the material.

4. Here is an example of the topic **Pareto Distribution:**



The (Type II) **Pareto distribution** with parameters $\alpha, \beta > 0$ has pdf

$$f(x) = \frac{\alpha\beta^\alpha}{(x+\beta)^{\alpha+1}}, \quad x > 0$$

and cdf

$$F_P(x) = 1 - \left(\frac{\beta}{x+\beta}\right)^\alpha, \quad x > 0.$$

If X is Type II Pareto with parameters α, β, then

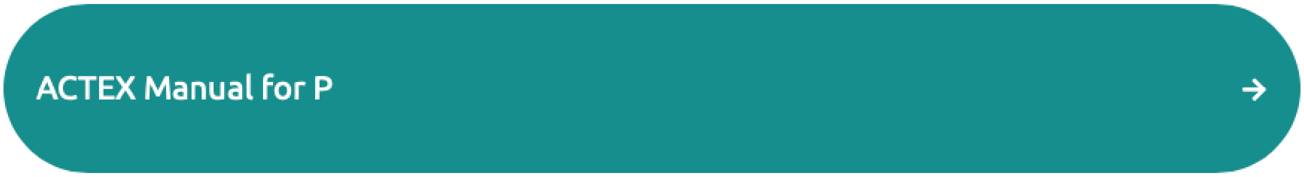
$$E[X] = \frac{\beta}{\alpha - 1} \text{ if } \alpha > 1,$$

and

$$Var[X] = \frac{\alpha\beta^2}{\alpha - 2} - \left(\frac{\alpha\beta}{\alpha - 1}\right)^2 \text{ if } \alpha > 2.$$

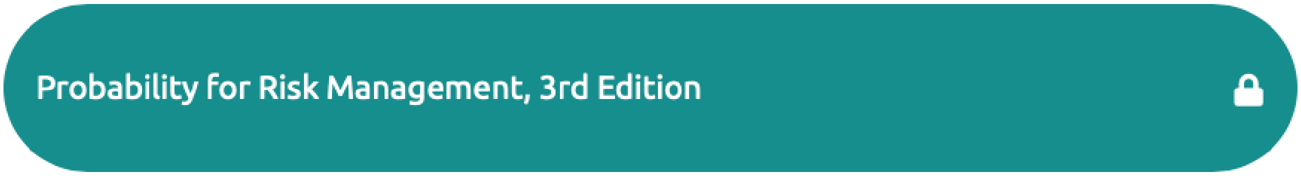
Within the **Hub** there will be unlocked and locked products.

Unlocked Products are the products that you own.



ACTEX Manual for P →

Locked Products are products that you do not own, and are available for purchase.

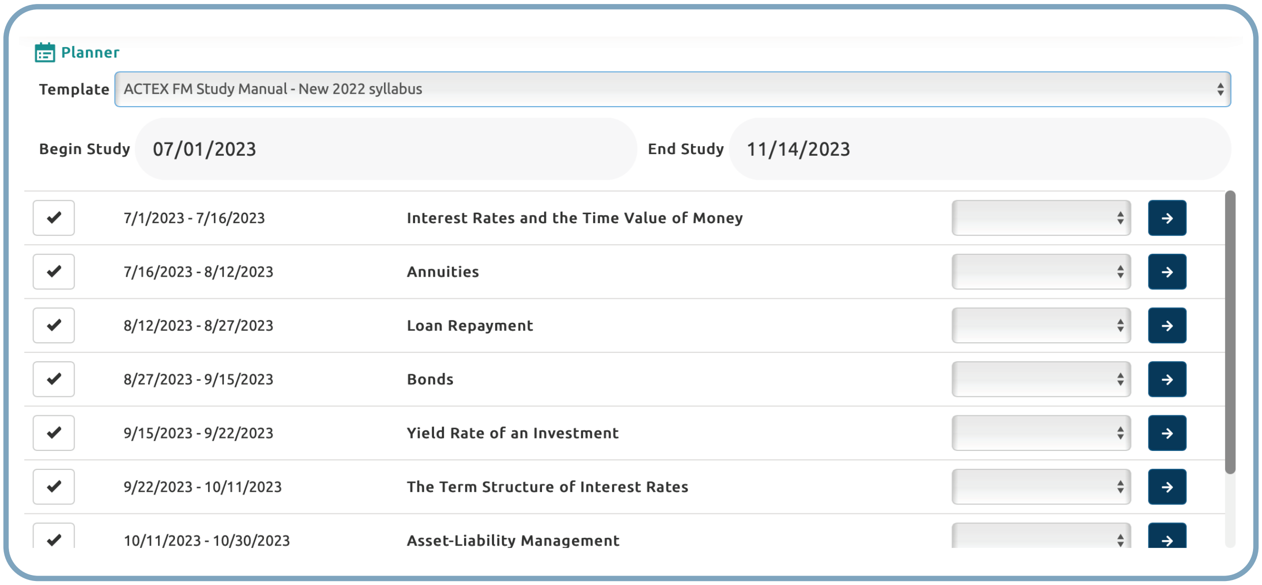


Probability for Risk Management, 3rd Edition

Many of Actuarial University's features are already unlocked with your study program, including:

Instructional Videos*	**Planner**
Topic Search	**Formula & Review Sheet**

Make your study session more efficient with our Planner!



Planner

Template: ACTEX FM Study Manual - New 2022 syllabus

Begin Study: 07/01/2023 End Study: 11/14/2023

✓	7/1/2023 - 7/16/2023	Interest Rates and the Time Value of Money
✓	7/16/2023 - 8/12/2023	Annuities
✓	8/12/2023 - 8/27/2023	Loan Repayment
✓	8/27/2023 - 9/15/2023	Bonds
✓	9/15/2023 - 9/22/2023	Yield Rate of an Investment
✓	9/22/2023 - 10/11/2023	The Term Structure of Interest Rates
✓	10/11/2023 - 10/30/2023	Asset-Liability Management

Available standalone, or included with the Study Manual Program Video Bundle

GOAL

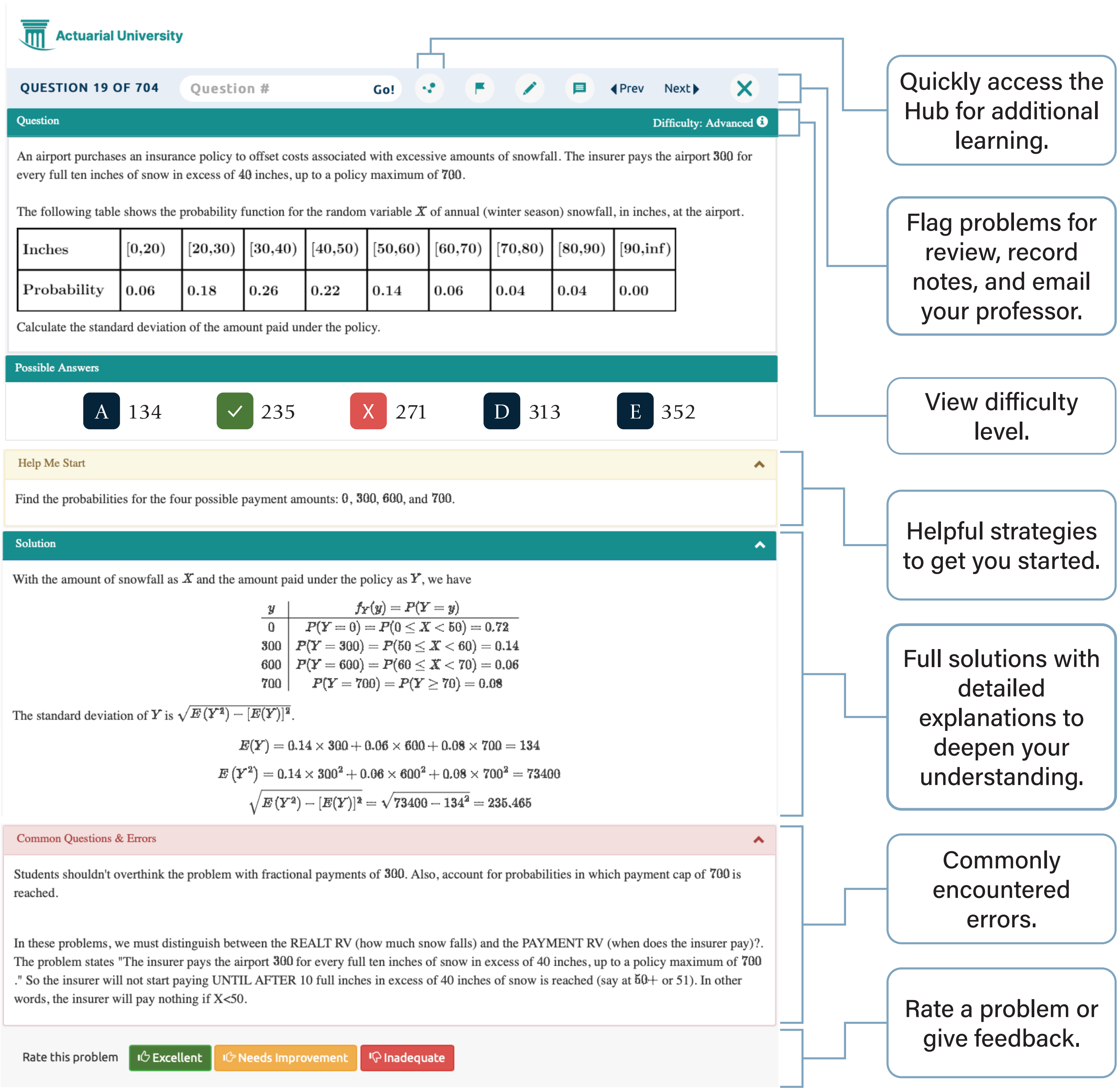
Practice. Quiz. Test. Pass!

- 16,000+ Exam-Style Problems
- Detailed Solutions
- Adaptive Quizzes
- 3 Learning Modes
- 3 Difficulty Modes

Free with your ACTEX or ASM Interactive Study Manual

Available for P, FM, FAM, FAM-L, FAM-S, ALTAM, ASTAM, MAS-I, MAS-II, CAS 5, CAS 6U & CAS 6C

Prepare for your exam confidently with GOAL custom Practice Sessions, Quizzes, & Simulated Exams



Actuarial University

QUESTION 19 OF 704 | Question # | Go! | ◂Prev | Next▸ | ✕

Question — Difficulty: Advanced

An airport purchases an insurance policy to offset costs associated with excessive amounts of snowfall. The insurer pays the airport 300 for every full ten inches of snow in excess of 40 inches, up to a policy maximum of 700.

The following table shows the probability function for the random variable X of annual (winter season) snowfall, in inches, at the airport.

Inches	[0,20)	[20,30)	[30,40)	[40,50)	[50,60)	[60,70)	[70,80)	[80,90)	[90,inf)
Probability	0.06	0.18	0.26	0.22	0.14	0.06	0.04	0.04	0.00

Calculate the standard deviation of the amount paid under the policy.

Possible Answers

A 134 ✓ 235 X 271 D 313 E 352

Help Me Start

Find the probabilities for the four possible payment amounts: 0, 300, 600, and 700.

Solution

With the amount of snowfall as X and the amount paid under the policy as Y, we have

$$\begin{array}{c|c} y & f_Y(y)=P(Y=y) \\ \hline 0 & P(Y=0)=P(0\le X<50)=0.72 \\ 300 & P(Y=300)=P(50\le X<60)=0.14 \\ 600 & P(Y=600)=P(60\le X<70)=0.06 \\ 700 & P(Y=700)=P(Y\ge 70)=0.08 \end{array}$$

The standard deviation of Y is $\sqrt{E(Y^2)-[E(Y)]^2}$.

$$E(Y)=0.14\times 300+0.06\times 600+0.08\times 700=134$$

$$E\left(Y^2\right)=0.14\times 300^2+0.06\times 600^2+0.08\times 700^2=73400$$

$$\sqrt{E\left(Y^2\right)-[E(Y)]^2}=\sqrt{73400-134^2}=235.465$$

Common Questions & Errors

Students shouldn't overthink the problem with fractional payments of 300. Also, account for probabilities in which payment cap of 700 is reached.

In these problems, we must distinguish between the REALT RV (how much snow falls) and the PAYMENT RV (when does the insurer pay)?. The problem states "The insurer pays the airport 300 for every full ten inches of snow in excess of 40 inches, up to a policy maximum of 700 ." So the insurer will not start paying UNTIL AFTER 10 full inches in excess of 40 inches of snow is reached (say at 50+ or 51). In other words, the insurer will pay nothing if X<50.

Rate this problem | Excellent | Needs Improvement | Inadequate

Quickly access the Hub for additional learning.

Flag problems for review, record notes, and email your professor.

View difficulty level.

Helpful strategies to get you started.

Full solutions with detailed explanations to deepen your understanding.

Commonly encountered errors.

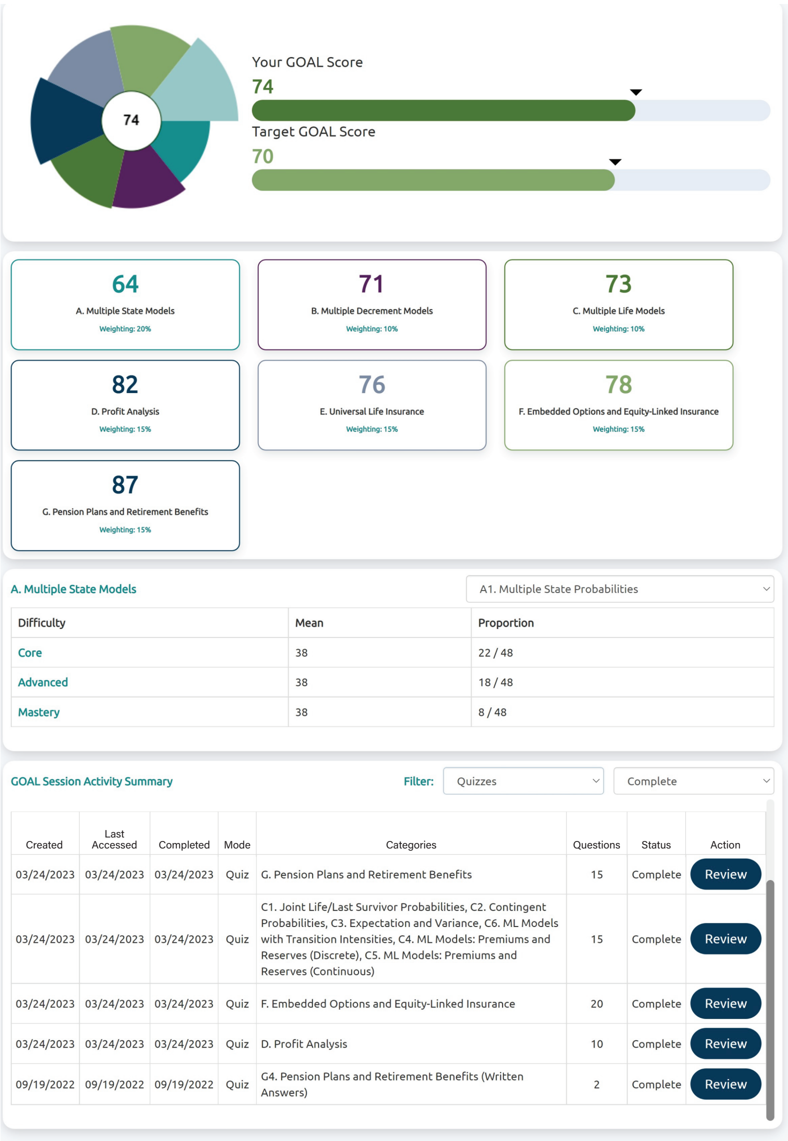
Rate a problem or give feedback.

Track your exam readiness with GOAL Score!

Available for P, FM, FAM, FAM-L, FAM-S, ALTAM, ASTAM, MAS-I, and MAS-II

GOAL Score tracks your performance through GOAL Practice Sessions, Quizzes, and Exams, resulting in an aggregate weighted score that gauges your exam preparedness.

By measuring both your performance, and the consistency of your performance, **GOAL Score** produces a reliable number that will give you confidence in your preparation before you sit for your exam.



Your GOAL Score
74

Target GOAL Score
70

64 A. Multiple State Models Weighting: 20%	71 B. Multiple Decrement Models Weighting: 10%	73 C. Multiple Life Models Weighting: 10%
82 D. Profit Analysis Weighting: 15%	76 E. Universal Life Insurance Weighting: 15%	78 F. Embedded Options and Equity-Linked Insurance Weighting: 15%
87 G. Pension Plans and Retirement Benefits Weighting: 15%		

A. Multiple State Models — A1. Multiple State Probabilities

Difficulty	Mean	Proportion
Core	38	22 / 48
Advanced	38	18 / 48
Mastery	38	8 / 48

GOAL Session Activity Summary — Filter: Quizzes, Complete

Created	Last Accessed	Completed	Mode	Categories	Questions	Status	Action
03/24/2023	03/24/2023	03/24/2023	Quiz	G. Pension Plans and Retirement Benefits	15	Complete	Review
03/24/2023	03/24/2023	03/24/2023	Quiz	C1. Joint Life/Last Survivor Probabilities, C2. Contingent Probabilities, C3. Expectation and Variance, C6. ML Models with Transition Intensities, C4. ML Models: Premiums and Reserves (Discrete), C5. ML Models: Premiums and Reserves (Continuous)	15	Complete	Review
03/24/2023	03/24/2023	03/24/2023	Quiz	F. Embedded Options and Equity-Linked Insurance	20	Complete	Review
03/24/2023	03/24/2023	03/24/2023	Quiz	D. Profit Analysis	10	Complete	Review
09/19/2022	09/19/2022	09/19/2022	Quiz	G4. Pension Plans and Retirement Benefits (Written Answers)	2	Complete	Review

If your GOAL Score is a 70 or higher, you are well-prepared to sit for your exam!

See key areas where you can improve.

Detailed performance tracking.

Quickly return to previous sessions.

Contents

Preface

⚠ **NOTE TO STUDENTS** ⚠

Please read this preface carefully. It contains very important information that will help you navigate this manual and Exam SRM smoothly! 👍

1 About Exam SRM

In Fall 2018, the Society of Actuaries (SOA) added a considerable amount of material on predictive analytics to its Associateship curriculum in view of the growing relevance of this discipline to actuarial work. The most significant changes were the introduction of two inter-related exams:

- Exam SRM (Statistics for Risk Modeling)
- Exam PA (Predictive Analytics)

In January 2022, the ATPA (Advanced Topics in Predictive Analytics) Assessment was also added. This study manual prepares you adequately for Exam SRM, but also paves the way for Exams PA and ATPA, which are largely a sequel to SRM. (See page xviii for further discussion.)

Exam Theme: What is SRM Like?

At a high level, Exams SRM, PA, and ATPA all share the same theme of working with *models*—more specifically, constructing predictive models from data, interpreting the output of these models, evaluating their performance, selecting the best model according to certain criteria, and applying the selected model to make predictions for the future. The whole process involves a sequence of complex and inter-related decisions that do not lend themselves to a multiple-choice exam format, which can only elicit a simple response. In Exams PA and ATPA, you will accomplish these modeling tasks using a computer equipped with Microsoft Word W, Microsoft Excel X, and/or R, and prepare your responses in a written-answer format.

As a precursor, Exam SRM is a traditional multiple-choice exam that serves to provide you with the foundational knowledge behind the modeling process and get you up to speed. You will learn the general tools available for constructing and evaluating predictive models (e.g., training/test set split, cross-validation), and the technical details of specific types of models and techniques (e.g., linear models, generalized linear models, regression-based time series models, decision trees, principal components analysis, and clustering). Despite the title of the exam, what you learn in SRM are widely applicable tools and techniques that can be used not only for "Risk Modeling," but also in many other contexts, even non-actuarial ones. Multiple-choice questions ☰ work here because the objective of this exam is to ensure that candidates are familiar with the basic predictive analytic concepts, at a rather high level, before setting foot in the PA and ATPA arena and seeing things in action. In brief, SRM gives you the conceptual groundwork for PA and ATPA.

Exam Administrations

More specifically, Exam SRM is a 3.5-hour computer-based exam consisting of 35 multiple-choice questions. In 2024 (and likely thereafter), it will be delivered via computer-based testing (CBT) in January, May, and September. Specifics of each testing window (e.g., exam dates, registration deadlines) can be found at:

https://www.soa.org/education/exam-req/exam-day-info/exam-schedules/.

The latest syllabus of Exam SRM, available from

https://www.soa.org/education/exam-req/edu-exam-srm-detail/study/,

is very broad in scope, covering miscellaneous topics in linear regression models, generalized linear models, time series analysis, and statistical learning techniques, many of which are contemporary topics introduced to the ASA curriculum for the first time. The following table shows the five main topics of the syllabus along with their approximate weights and where they are covered in this manual:

Topic	Range of Weight	Relevant Chapters of ACTEX SRM Manual
1. Basics of Statistical Learning	5–10%	Chapter 4
2. Linear Models	40–50% (heavy!)	Chapters 1–5
3. Time Series Models	10–15%	Chapters 6–7
4. Decision Trees	20–25%	Chapter 8
5. Unsupervised Learning Techniques	10–15%	Chapters 9–10

Note that effective from the May 2023 sitting, the weight assigned to Topic 4 has increased quite substantially from 10–15% to 20–25%, so be sure to study Chapter 8 carefully!

As a rough estimate, you will need about **THREE months** of intensive study 📖 to master the material in this exam. There are A LOT to learn and absorb. (In fact, the real exam, even with 35 questions, will likely only test a small subset of the whole exam syllabus.) Don't worry about studying too hard—what you learn in Exam SRM will continue to be useful for Exams PA and ATPA!

Mathematical Prerequisites

The first word of SRM is "Statistics" and, not surprisingly, we will do a lot of statistics in this exam, so it is assumed that you have taken a calculus-based mathematical statistics course (e.g., the one you use to fulfill your VEE Mathematical Statistics requirement) and are no stranger to concepts like t-, F-, and chi-square distributions, point estimators (maximum likelihood estimators in particular), confidence intervals, and hypothesis tests, which will be used quite heavily in Chapters 1, 2, and

5 of this manual. There will also be limited instances (mostly in Chapters 2, 3, and 9) in which you will perform some simple matrix multiplication, which you should have learned in your linear algebra class. Prior exposure to the R programming language, which is used in one of the syllabus textbooks, is not required, however. According to the exam syllabus,

> "[the] ability to solve problems using the R programming language will not be assumed. However, questions may present (self-explanatory) R output for interpretation."

Exam Style

As of September 2023, the SOA has released a total of 67 sample questions, which can be accessed from[i]

https://www.soa.org/Files/Edu/2018/exam-srm-sample-questions.pdf.

According to students' comments, these sample questions are quite indicative of the style and level of difficulty of the exam questions you will see on the real exam, and all of them have been included in the relevant sections of this manual. Judging by these sample questions, we can infer that most SRM exam questions fall into two categories:

Type 1. *Simple computational questions given a small raw dataset or summarized model output (roughly* 1/3 *of the exam)*

In some exam questions (e.g., Sample Questions 1, 3, 4, 9, 11, 15, 17, 18, 19, 23, 24, 27, 28, 33, 35, 44, 45-48, 51, 54, 55, 57-59, 62, 63, 66, 67), you will be asked to do some simple calculations in one of two scenarios:

- You may be given a small dataset, e.g., one with not more than 10 observations. While almost all predictive analytic techniques in the exam syllabus require computers to implement, the small size of the dataset makes it possible to perform at least part of the analysis.

- You may also be given some summarized model output such as tables of parameter estimates and/or graphical output 📈. Then you are asked to perform some simple tasks like interpreting the results of the model, conducting a hypothesis test, making point/interval estimations/predictions, and assessing the goodness of fit of the model, all of which require only pen-and-paper calculations 🖩.

You may ask:

> Why should the SOA make these unrealistic exam questions? Aren't we all using computer 💻 to do the work in real life?

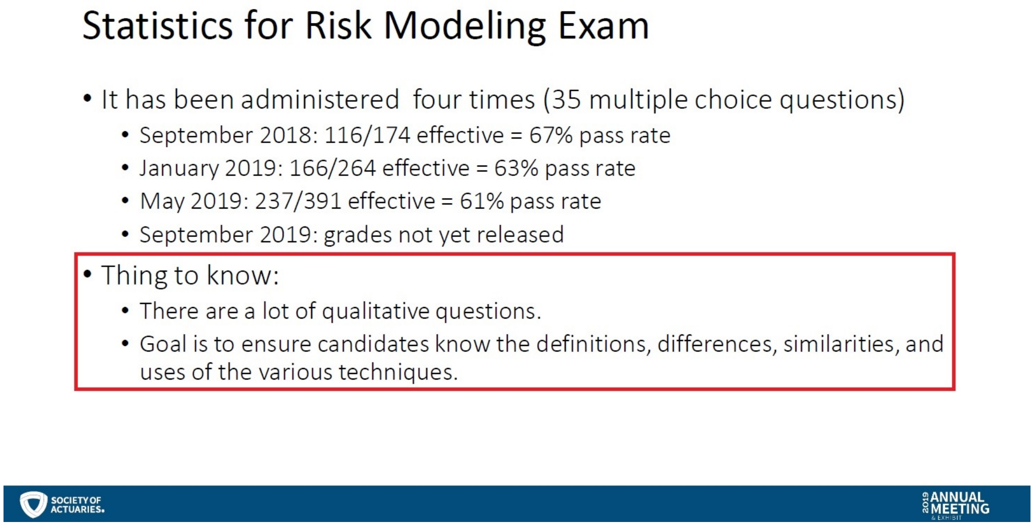
Although you probably will not have the chance to perform hand calculations in the workplace, these quantitative questions encourage you to understand the mechanics of the statistical methodology being tested—you need to know what happens in a particular step of the modeling process, which formulas to use, and what the model output means—and are instructive from an educational point of view.

[i] The SOA deleted Questions 17, 28, 47, and 65 because they test concepts "no longer on the syllabus." There have been no changes in the syllabus readings, which form the backbone of the exam, so it is unclear why these questions were deleted.

Type 2. [IMPORTANT ⚠] *Conceptual/True-or-false questions (roughly 2/3 of the exam)*

A distinguishing characteristic of Exam SRM compared to other multiple-choice ASA-level exams is that most of the questions in this exam are **conceptual** (a.k.a. **qualitative**, **true-or-false**) in nature, testing the uses, motivations, considerations, pros and cons, do's and don'ts of different predictive models, and their similarities and differences. As the SOA publicly admitted in the 2019 Annual Meeting & Exhibit,

"there are a lot of qualitative questions [in Exam SRM]."



Statistics for Risk Modeling Exam

- It has been administered four times (35 multiple choice questions)
  - September 2018: 116/174 effective = 67% pass rate
  - January 2019: 166/264 effective = 63% pass rate
  - May 2019: 237/391 effective = 61% pass rate
  - September 2019: grades not yet released
- Thing to know:
  - There are a lot of qualitative questions.
  - Goal is to ensure candidates know the definitions, differences, similarities, and uses of the various techniques.

Sample Questions 2, 5, 6, 7, 8, 10, 12, 13, 14, 16, 20, 21, 22, 25, 26, 29-32, 34, 36-43, 49, 50, 52, 53, 56, 60-61, 64, and 65 all belong to this type of questions. You are typically given three statements (I, II, and III) and asked to pick the correct one(s). The five answer choices often take a symmetric structure:

TYPICAL FORM OF CONCEPTUAL SRM QUESTIONS

Determine which of the following statements about *[...a particular statistical concept/method...]* is/are true.

I. *[blah blah blah...💬]*

II. *[blah blah blah...💬]*

III. *[blah blah blah...💬]*

(A) I only (B) II only

(C) III only (D) I, II, and III

(E) The correct answer is not given by (A), (B), (C), or (D).

or

(A) None (B) I and II only

(C) I and III only (D) II and III only

(E) The correct answer is not given by (A), (B), (C), or (D).

Do not be under the impression that these conceptual questions must be easy. They can test the obscure ins and outs of different predictive analytic techniques, some of which are mentioned only in one or two lines in the syllabus readings. At times, they can also be quite vague or controversial: Rather than an absolute "yes" or "no," the statement is more a matter of extent. Sadly, if you get any of Statements I, II, or III incorrect, you will likely be led to an incorrect final answer. By the way, Answer (E), which says that (A) to (D) are all wrong, occasionally turns out to be the right answer—it is not a filler!

Historical Pass Rates and Pass Marks

The table below shows the number of sitting candidates, number of passing candidates, pass rates, and pass marks for Exam SRM since it was offered in September 2018.

Sitting	# Candidates	# Passing Candidates	Pass Rate	Pass Mark
September 2023	(To be posted on the SOA webpage 🔗)			
May 2023	1820	1352	74.3%	60%
January 2023	1151	847	73.6%	60%
September 2022	1241	939	75.7%	60%
May 2022	1004	768	76.5%	65%
January 2022	794	625	78.7%	65%
September 2021	810	592	73.1%	65%
May 2021	843	660	78.3%	65%
January 2021	754	566	75.1%	65%
September 2020	814	627	77.0%	63%
May 2020	278	205	73.7%	63%
January 2020	587	372	63.4%	67%
September 2019	554	411	74.2%	60%
May 2019	410	237	57.8%	67%
January 2019	274	166	60.6%	67%
September 2018	181	116	64.1%	70%

Perhaps to your astonishment, the pass rates of Exam SRM have been anomalously high, typically in the 60-75% range, compared to only 40-50% for typical ASA-level exams. The pass marks in the three most recent sittings are all 60%, which means that candidates need to get about **21 out of 35** questions[ii] correct to earn a pass.

[ii] According to the exam syllabus, one or two questions in the CBT environment may be pilot questions that are included to judge their effectiveness for future exams, but they will not be graded.

Syllabus Texts

Exam SRM has two required textbooks:

1. *An Introduction to Statistical Learning: With Applications in R*, by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2021, second edition, available online thanks to the mercy of the authors (⬆ click to access the book freely...and legally!)

   Referred to as **ISLR** in this study manual and used by statistical learning courses all over the world, this highly popular book covers both classical and modern predictive analytic techniques in and beyond the linear regression framework. Although written by four distinguished statisticians, this book is designed for non-statisticians and de-emphasizes technical details (formulas and proofs in particular). In fact, one of the greatest selling points of the book is to facilitate the implementation of the statistical learning techniques introduced in the book using R by a wide range of audience.

   On Internet forums (e.g., Reddit, Discord), many users recommend reading ISLR as an important way to prepare for SRM. While the book is available online and well-written in general, it is rather text-heavy (there are many long paragraphs!) and has hardly any exam-type problems. With this self-contained study manual, reading ISLR is completely optional. Of course, if you still have time left after finishing this manual in its entirety, then there is no harm in taking a look at ISLR as a further way to consolidate your understanding.

2. *Regression Modeling with Actuarial and Financial Applications*, by Edward W. Frees, 2010

   Referred to as **Frees** in the sequel and written by a retired professor at the University of Wisconsin–Madison, this is an ambitious textbook that deals with a wide range of topics in regression analysis with a rather traditional treatment. It tries to cover a lot of ground, but the explanations are not always clear or adequate. Unlike ISLR, Frees is, ironically, not a free book. If you happen to have a copy of Frees and intend to read it (not necessary at all!), do remember to check out the errata list (⬅ click to access). It is LONG!

Among the five topics in the exam syllabus, Frees covers Topics 2, 3, and part of Topic 1, while ISLR covers Topics 4, 5, and most of Topic 1. These two texts duplicate to a certain extent when it comes to the chapters on linear regression models. In this study manual, we have streamlined the material in both texts to result in a coherent exposition without unnecessary repetition. As far as possible, we have followed the notation in the two texts because exam questions can freely use the symbols and shorthand in the syllabus readings (e.g., F, RSS, TSS) without further definitions.

Exam Tables ⊞

In the real exam, you will have access to three statistical tables, namely, the standard normal distribution, t-distribution, and chi-square distribution tables. They are available for download from

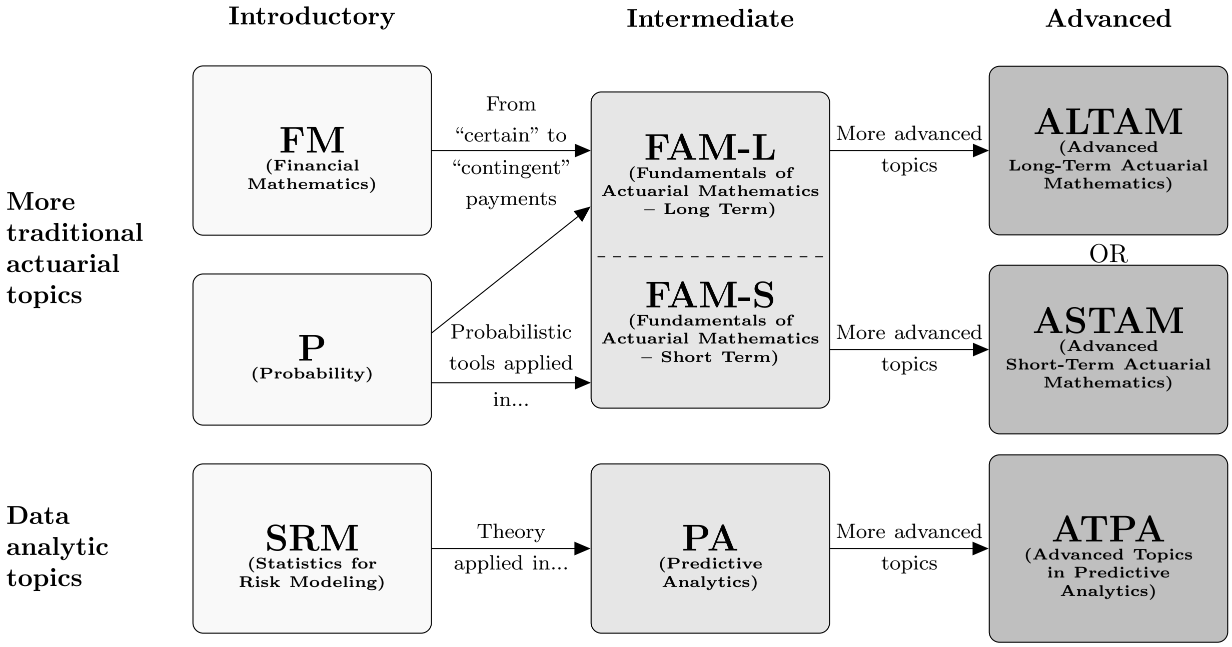
⬇ https://www.soa.org/Files/Edu/2018/exam-srm-tables.pdf

and will be used quite in Chapters 1, 2, 5, and 7 of this study manual. It is a good idea to print out a copy of these tables 🖶 and learn how to locate the relevant entries as you work out the examples and problems in this manual.

Predictive Analytics Trio: SRM, PA, and ATPA

As we noted earlier, Exam SRM is an important stepping stone to Exams PA and ATPA. The flowchart below shows how these three exams (and other ASA exams for your information) are related. While there is no set order in which the exams should be taken, students typically attempt exams from left to right, or from introductory, intermediate to advanced. In the case of the predictive analytics trio, that means taking SRM, PA, and ATPA, in this order.

Flowchart of ASA Exams Effective from 2022


Introductory
Intermediate
Advanced
More traditional actuarial topics
Data analytic topics
FM (Financial Mathematics)
From "certain" to "contingent" payments
FAM-L (Fundamentals of Actuarial Mathematics – Long Term)
FAM-S (Fundamentals of Actuarial Mathematics – Short Term)
P (Probability)
Probabilistic tools applied in...
More advanced topics
ALTAM (Advanced Long-Term Actuarial Mathematics)
OR
ASTAM (Advanced Short-Term Actuarial Mathematics)
More advanced topics
SRM (Statistics for Risk Modeling)
Theory applied in...
PA (Predictive Analytics)
More advanced topics
ATPA (Advanced Topics in Predictive Analytics)

SRM vs. PA. Exam PA is a 3.5-hour computer-based exam offered twice a year, in April and in October. Unlike SRM, PA is not a multiple-choice exam ⚠. Instead, you will be given a business problem broken down into a series of well-defined tasks and asked to write your responses in Microsoft Word addressing those tasks.

In essence, Exams SRM and PA are about the same subject, but test it differently. While Exam SRM emphasizes the theory underlying different predictive analytic techniques, Exam PA will have you apply the theory you learned in Exam SRM to real data by means of computer-based implementations using R. Some additional topics and practical considerations are also presented. After taking Exam PA, you will see the predictive models you learned in Exam SRM in action and gain a much more thorough understanding.

Note that ACTEX Learning has released a separate study manual for Exam PA written by the main author (Ambrose Lo) of this manual. To learn more, please check out:

https://www.actexlearning.com/exams/pa.

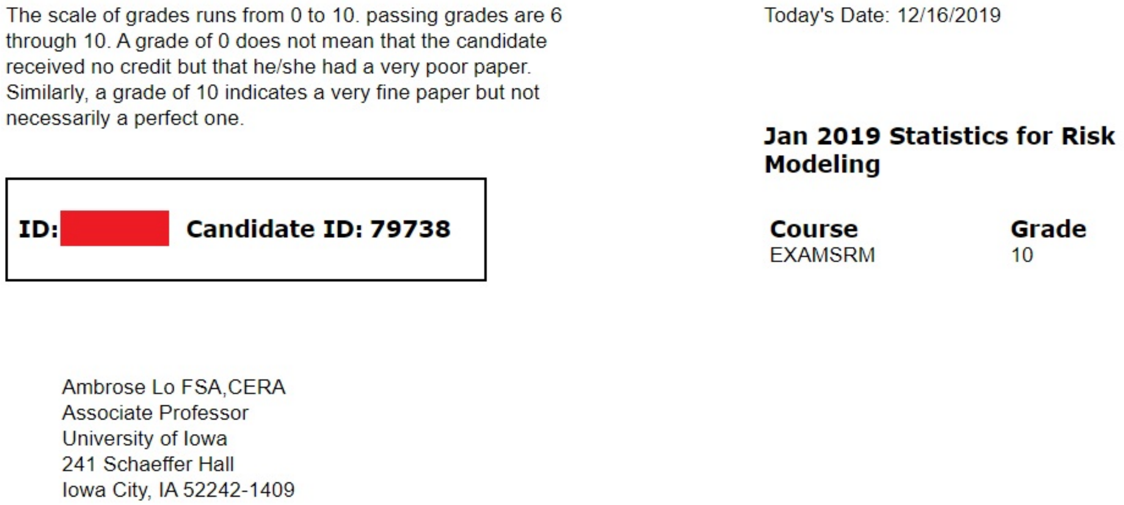
PA vs. ATPA. Exam ATPA is a 96-hour take-home computer-based assessment (rather than a proctored exam, so Exam ATPA is also called the "ATPA Assessment"). It tests additional data and modeling concepts on the basis of those in Exams SRM and PA, and consists of inter-related and more open-ended tasks than those in PA. As a result, ATPA is preferably taken after passing SRM and PA. In contrast to PA, which only requires some basic knowledge of R programming, proficiency with R is critical to success in ATPA. During the 96-hour window, you will spend most of your time dealing with various data issues, constructing and evaluating more advanced predictive models than those covered in SRM and PA, and finally turning your results into a written report. Make sure that you have set aside enough free time in your schedule 📅 for the next 4 days before you start the assessment. You may need more than a day just to clean the data and get it in good shape in R. You will be busy!

2 About this Study Manual

What is Special about This Study Manual?

We fully understand that you have an acutely limited amount of study time and that the SRM exam syllabus is enormous. With this in mind, the overriding objective of this study manual is to help you grasp the material in Exam SRM as effectively and efficiently as possible, so that you will pass the exam on your first try easily and go on to Exams PA and ATPA confidently. Here are some unique features of this manual to make this possible:

- Usually coaches don't play 😄, but the main author of this manual (Ambrose Lo) took the initiative to write the SRM exam in January 2019 to experience first-hand what the real exam was like, despite having been an FSA since 2013 (and technically free from exams thereafter!). He made this decision in the belief that braving the exam myself is the best way to ensure that this manual is indeed effective for exam preparation. (If the manual is useful, then at the minimum the author himself can do well, right?) Drawing upon his "real battle experience," we can assure you that this manual is written from an exam taker's perspective and the examples/problems are designed in a way that best prepares you for SRM.



The scale of grades runs from 0 to 10. passing grades are 6 through 10. A grade of 0 does not mean that the candidate received no credit but that he/she had a very poor paper. Similarly, a grade of 10 indicates a very fine paper but not necessarily a perfect one.

Today's Date: 12/16/2019

ID: [redacted] **Candidate ID: 79738**

Jan 2019 Statistics for Risk Modeling

Course	**Grade**
EXAMSRM	10

Ambrose Lo FSA,CERA
Associate Professor
University of Iowa
241 Schaeffer Hall
Iowa City, IA 52242-1409

- Each chapter starts by explicitly stating which learning objectives and outcomes of the SRM exam syllabus we are going to cover, to assure you that we are on track and hitting the right target.

- The explanations in each chapter are thorough, but exam-focused and integrated with carefully chosen past exam/sample questions for illustration, so that you will learn the syllabus material effectively and efficiently. Throughout, we strive to keep you motivated by showing you how different concepts are typically tested, how different formulas are used, and where the exam focus lies in each section. As you read, you will develop a solid understanding of the concepts in SRM and know how to study for the exam.

- Formulas and results of utmost importance are $\boxed{\text{boxed}}$ for easy identification and numbered (in the (X.X.X) format) for later references. Mnemonics and shortcuts are emphasized, so are highlights of important exam items and common mistakes committed by students.

- While the focus of this study manual is on exam preparation, we take every opportunity to explain the intuitive meaning and mathematical structure of various formulas in the syllabus. The interpretations and insights we provide will foster a genuine understanding of the syllabus material and reduce the need for slavish memorization. It is the authors' belief and personal experience that a solid understanding of the underlying concepts is always conducive to achieving good exam results.

- To succeed in any actuarial exam, we can't overemphasize the importance of practicing a wide variety of exam-type problems to sharpen your understanding and develop proficiency. This study manual embraces this learning by doing approach and features more than 250 in-text examples and 400 end-of-section/chapter problems, which are either taken/adapted from relevant SOA/CAS past exams or original, all with step-by-step solutions, to consolidate your understanding and give you a sense of what you can expect to see in the real exam. Many of the original problems are of the true-or-false type, which, according to students' comments, has figured prominently in recent SRM exams.

  To maximize the effectiveness and efficiency of your learning, we have marked the most representative and instructive practice problems in each section with an asterisk (*). These selected problems, which are generally about 50% of the whole set of problems, will add most value to your learning. Here is our suggestion:

  > Read the main text of each section, including *all* of the in-text examples, and work out the asterisked end-of-section practice problems. Then go on to the next section or chapter. Repeat this process until you finish all of the ten chapters in the core of the manual.

  This should be a good learning strategy for developing a thorough understanding of the syllabus material, and a level of proficiency and confidence necessary for exam taking, while avoiding burn-out. Of course, if there is time left after you finish the entire manual (including the practice exams), it would be great if you work out some of the non-asterisked practice problems as well, especially those related to your weak spots.

- Although this study manual is self-contained in the sense that studying the manual carefully is already sufficient to pass the exam, relevant chapters and sections of the two syllabus texts are referenced at the beginning of each chapter of the manual, for the benefit of students who like to read more. (Remember that ISLR is freely available online.)

- Six (6) original full-length practice exams designed to mimic the real SRM exam in terms of style and difficulty conclude this study manual and give you a holistic review of the syllabus material. Detailed illustrative solutions are provided.

Contact Us 🤝

If you encounter problems with your learning, we stand ready to help.

- For **technical** issues (e.g., not able to access your manual online, extending your digital license, upgrading your product, exercising the Pass Guarantee), please email Customer Service at `support@actexlearning.com`. ✉
- Questions related to **specific contents** of this manual, including potential errors (typographical or otherwise), can be directed to the main author (Ambrose Lo) by emailing `amblo201011@gmail.com`. ✉

Acknowledgments

We would like to thank Dr. Michelle A. Larson for sharing with us many pre-2000 SOA/CAS exam papers. Despite their seniority and the use of different syllabus texts, these hard-earned old exam papers, of which the SOA and CAS own the sole copyright, have proved invaluable in illustrating a number of less commonly tested exam topics in the current syllabus. Ambrose Lo is also grateful to students at The University of Iowa in his SRM courses STAT:4560 (Statistics for Risk Modeling I) in Fall 2019-2022 and STAT:4561 (Statistics for Risk Modeling II) in Spring 2023, and his VEE Applied Statistics course STAT:4510 (Regression, Time Series, and Forecasting) in Fall 2016 and Fall 2017 for class testing earlier versions of this study manual.

About the Authors

Runhuan Feng, PhD, FSA, CERA, is a professor and the Director of Actuarial Science Program at the University of Illinois at Urbana–Champaign. He obtained his PhD in Actuarial Science from the University of Waterloo, Canada. He is a Helen Corley Petit Professorial Scholar and the State Farm Companies Foundation Scholar in Actuarial Science. Prior to joining Illinois, he held a tenure-track position at the University of Wisconsin-Milwaukee. Runhuan has published extensively on stochastic analytics in risk theory and quantitative risk management. Over the recent years, he has dedicated himself to developing computational methods for managing market innovations in areas of investment combined insurance and retirement planning. He has authored several research monographs including *An Introduction to Computational Risk Management of Equity-Linked Insurance*.

Daniël Linders, PhD, is an assistant professor at the University of Illinois at Urbana-Champaign. At the University of Leuven, Belgium, he obtained an M.S. degree in Mathematics, an Advanced M.S. degree in Actuarial Science and a PhD in Business Economics. Before joining the University of Illinois, he was a postdoctoral researcher at the University of Amsterdam, The Netherlands and the Technical University in Munich, Germany. He is a member of the Belgian Institute of Actuaries and has the Certificate in Quantitative Finance from the CQF Institute. Daniël Linders has wide teaching experience. He taught various courses on Predictive Analytics, Life Contingencies, Pension Financing and Risk Measurement. He is currently teaching at the University of Illinois and is guest lecturer at the University of Leuven and the ISM-Adonaí, Benin.

Ambrose Lo, PhD, FSA, CERA, was formerly Associate Professor of Actuarial Science with tenure at the Department of Statistics and Actuarial Science, The University of Iowa. He earned his B.S. in Actuarial Science (first class honors) and PhD in Actuarial Science from The University of Hong Kong in 2010 and 2014, respectively, and attained his Fellowship of the Society of Actuaries (FSA) in 2013. He joined The University of Iowa as Assistant Professor of Actuarial Science in August 2014, and was tenured and promoted to Associate Professor in July 2019. His research interests lie in dependence structures, quantitative risk management as well as optimal (re)insurance. His research papers have been published in top-tier actuarial journals, such as *ASTIN Bulletin: The Journal of the International Actuarial Association*, *Insurance: Mathematics and Economics*, and *Scandinavian Actuarial Journal*.

Besides dedicating himself to actuarial research, Ambrose attaches equal importance to teaching and education, through which he nurtures the next generation of actuaries and serves the actuarial profession. He has taught courses on financial derivatives, mathematical finance, life contingencies, and statistics for risk modeling. He has (co)authored the *ACTEX Study Manuals for Exams MAS-I, MAS-II, PA, and SRM*, a *Study Manual for Exam FAM*, and the textbook *Derivative Pricing: A Problem-Based Primer* (2018) published by Chapman & Hall/CRC Press. Although helping students pass actuarial exams is an important goal of his teaching, inculcating students with a thorough understanding of the subject and concrete problem-solving skills is always his top priority. In recognition of his exemplary teaching, Ambrose has received a number of awards and honors ever since he was a graduate student, including the 2012 Excellent Teaching Assistant Award from the Faculty of Science, The University of Hong Kong, public recognition in the *Daily Iowan* as a faculty member "making a positive difference in students' lives during their time at The University of Iowa" for eight years in a row (2016 to 2023), and the 2019-2020 Collegiate Teaching Award from the College of Liberal Arts and Sciences, The University of Iowa.

Part I

Regression Models

Chapter 2

Multiple Linear Regression

*****FROM THE SRM EXAM SYLLABUS*****

2. Topic: Linear Models (40-50%)

Learning Objectives

The Candidate will understand key concepts concerning generalized linear models.

Learning Outcomes

The Candidate will be able to:

a) Compare model assumptions for ordinary least squares and generalized linear models.

c) Apply the business context of a problem to interpret parameters.

e) Select an appropriate model, including:

  - Variable transformations and interactions
  - t and F tests

f) Calculate and interpret predicted values, and confidence and prediction intervals.

OPTIONAL SRM SYLLABUS READINGS

- Frees, Chapter 3 and Section 6.1
- ISLR, Section 3.2, Subsections 3.3.1-3.3.2

Chapter overview: This chapter extends the discussion in Chapter 1 to the case of more than one predictor. Such a statistical model linearly relating a response variable to "multiple" predictors is aptly called a *multiple linear regression* (MLR) model (or sometimes simply a *linear model*), which is a considerable generalization of an SLR model. Capitalizing on the information brought by a collection of predictors, MLR opens the door to many more questions of practical interest that can be explored and answered in a statistical framework. Examples include:

(i) Are certain predictors useful for explaining the variability of the response variable?

(ii) How should we form the regression function? Is there any interaction between some of the predictors in explaining the response variable?

(iii) Given several competing MLR models, which one is the best? Under what criterion?

We begin in Section 2.1 with the usual fitting, inference, and prediction issues, which are natural extensions of the results in Chapter 1 with mostly notational adjustments. Subtle and unique aspects of MLR unfold in Sections 2.2 to 2.4. Section 2.2 presents a notion of correlation that controls for other predictors not of primary interest and more accurately reflects the linear relationship between two variables. Section 2.3 introduces techniques for quantitatively representing different types of predictors and their interaction in an MLR model. Section 2.4 concludes this chapter with a generalization of the F-test we first learned in connection with the ANOVA table. Such a generalized F-test allows us to test for the significance of a subset of predictors and provides statisticians with considerable flexibility to investigate a wide variety of questions.

⚠ EXAM NOTE ⚠

Just like Chapter 1, there are usually **3 to 5 questions** set on this chapter in a typical SRM exam.

2.1 From SLR to MLR: Fundamental Results

OPTIONAL SRM SYLLABUS READINGS

- Frees, Sections 3.1 to Subsection 3.4.2, Subsection 5.5.4, and Section 6.1
- ISLR, Section 3.2

Model equation. An MLR model is a natural extension of an SLR model in that we employ more than one predictor to gain a better understanding of the behavior of the response variable. The primary interest here is how the predictors operate *together* to influence the response. Mathematically, the model equation of a generic MLR model is expanded to

$$\boxed{y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}_{\text{expanded regression function}} + \varepsilon,}$$

where:

- p is the number of predictors ($p = 1$ in SLR).

  (**Note:** Frees denotes the number of predictors by k while ISLR uses p. In this study manual, we follow ISLR's usage, which is more popular in the statistical learning community. In most cases, k and p can be used interchangeably. The only exception is Section 4.3 of this manual.)

- β_0 is the intercept.
- β_j is the regression coefficient attached to the jth predictor, for $j = 1, \ldots, p$.
- ε is again the random error term.

We assume that a sample of n observations from the model is available in the form of $\{(y_i, x_{i1}, \ldots, x_{ip})\}_{i=1}^n$. The equation governing the ith observation is

$$\boxed{y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad \text{where } \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2), \quad i = 1, \ldots, n.}$$

For notational consistency, in the remainder of this manual the subscript i is often used to index the observations (from 1 to n) and the subscript j is used to index the predictors (from 1 to p), so we may write the model equation compactly as $y_i = \sum_{j=0}^p \beta_j x_{ij}$, with $x_{i0} := 1$ corresponding to the intercept. The dataset can be displayed in a rectangular form as in Table 2.1, where observations are shown across the rows of the table and the corresponding predictor variable values are shown across the columns (in fact, rectangular datasets are very common in data science). The same assumptions concerning the predictors and the random errors as in SLR on page 6 are in force for an MLR model.

Observation	Response Variable	Predictors			
	y	x_1	x_2	$\cdots$	x_p
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

Table 2.1: Typical data structure of an MLR model.

Model fitting. To develop results for MLR, it is often convenient to recast the equation of an MLR model compactly in terms of matrix notation:

$$\boxed{\mathbf{y} = \mathbf{X}\beta + \varepsilon,} \tag{2.1.1}$$

or, on a component-wise basis,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

where: (In this study manual, we use **boldface** to denote vectors and matrices.)

- $\mathbf{y}$ is the $n \times 1$ *response vector*
- $\mathbf{X}$ is the matrix (sometimes known as the *data matrix* or *design matrix*) containing values of the predictors, with the first column of 1's corresponding to the intercept
- $\boldsymbol{\beta}$ is the vector of $p+1$ regression coefficients or parameters, which are to be estimated and inference is to be made
- $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of random errors

To estimate $\boldsymbol{\beta}$, we apply the same least squares techniques we used in Chapter 1 and minimize the sum of squares function

$$\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

where the prime " $'$ " denotes the transpose of a matrix, over $\boldsymbol{\beta}$. By matrix calculus, we solve the normal equations $\frac{\partial}{\partial \boldsymbol{\beta}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$ ($\mathbf{0}$ is a vector of zeros) to get the least squares estimator (LSE) of $\boldsymbol{\beta}$. The LSE need not be unique, but usually[i] it is, in which case it takes the following vector form:

$$\boxed{\hat{\boldsymbol{\beta}} := \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.} \tag{2.1.2}$$

Two points about this formidable vector formula deserve attention:

- Unlike the case of SLR, where we have closed-form algebraic formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ (recall (1.1.4)), (2.1.2) is all we have for the LSEs in an MLR model. In other words, we have a (magnificent!) formula for the entire vector of LSEs, but not separate algebraic formulas for the individual LSEs.

- To apply (2.1.2), we have to invert the $(p+1)\times(p+1)$ matrix $\mathbf{X}'\mathbf{X}$, which is hard to perform by pen-and-paper calculations unless $p = 1$ (i.e., SLR).[ii]

There are several ways a multiple-choice exam question can test (2.1.2):

Type 1. In many old SOA exam questions, $(\mathbf{X}'\mathbf{X})^{-1}$ is directly given to aid your computations. You will need to compute $\mathbf{X}'\mathbf{y} = \begin{pmatrix} \sum_{i=1}^n y_i & \sum_{i=1}^n x_{i1}y_i & \cdots & \sum_{i=1}^n x_{ip}y_i \end{pmatrix}'$, then multiply it by $(\mathbf{X}'\mathbf{X})^{-1}$ on the left.

[i] The LSE is unique as long as the matrix $\mathbf{X}'\mathbf{X}$ is invertible.

[ii] *(If you are interested)* When $p = 1$,

$$\begin{aligned}
\mathbf{X}'\mathbf{X} &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{pmatrix}, \\
(\mathbf{X}'\mathbf{X})^{-1} &= \frac{1}{n\sum_{i=1}^n x_i^2 - (n\bar{x})^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} = \frac{1}{nS_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}, \\
\mathbf{X}'\mathbf{y} &= \begin{pmatrix} n\bar{y} \\ \sum_{i=1}^n x_i y_i \end{pmatrix},
\end{aligned}$$

and (2.1.2) reduces to

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y}) = \frac{1}{nS_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} \begin{pmatrix} n\bar{y} \\ \sum_{i=1}^n x_i y_i \end{pmatrix} = \begin{pmatrix} \bar{y} - (S_{xy}/S_{xx})\bar{x} \\ S_{xy}/S_{xx} \end{pmatrix},$$

which is (1.1.4).

Example 2.1.1. (SOA Course 120 Study Note 120-82-94 Question 11: Given $(\mathbf{X}'\mathbf{X})^{-1}$) An automobile insurance company wants to use gender ($x_1 = 0$ if female, 1 if male) and traffic penalty points (x_2) to predict the number of claims (y). The observed values of these variables for a sample of six motorists are given by:

Motorist	1	2	3	4	5	6
x_1	0	0	0	1	1	1
x_2	0	1	2	0	1	2
y	1	0	2	1	3	5

You are to use the following model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \qquad i = 1, 2, \ldots, 6$$

You have determined:

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{12}\begin{pmatrix} 7 & -4 & -3 \\ -4 & 8 & 0 \\ -3 & 0 & 3 \end{pmatrix}$$

Determine $\hat{\beta}_2$.

(A) -0.25 (B) 0.25 (C) 1.25

(D) 2.00 (E) 4.25

Solution. The general formula for $\hat{\boldsymbol{\beta}}$ is $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, where

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \overset{(p=2)}{=} \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \sum_{i=1}^n x_{i2}y_i \end{pmatrix} = \begin{pmatrix} 12 \\ 9 \\ 17 \end{pmatrix}.$$

By (2.1.2), (irrelevant entries are marked by *)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \frac{1}{12}\begin{pmatrix} * & * & * \\ * & * & * \\ -3 & 0 & 3 \end{pmatrix}\begin{pmatrix} 12 \\ 9 \\ 17 \end{pmatrix} = \begin{pmatrix} * \\ * \\ \boxed{1.25} \end{pmatrix}.$$ **(Answer: (C))** □

Type 2. Another type of questions centers on a no-intercept MLR model with $p = 2$ predictors. In this case, the matrix $\mathbf{X}'\mathbf{X}$ is of dimension 2×2 and easy to invert using the matrix formula

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix},$$

provided that $ad - bc \neq 0$. In words, you simply interchange the two diagonal entries (a and d), put a negative sign on the two off-diagonal entries (b and c), and divide the transformed matrix by the determinant $ad - bc$.

Example 2.1.2. (SOA Part 4 May 1983 Question 10: No-intercept MLR model) Advertising expenditures and sales for the last 5 quarters have been as follows:

Quarter	Advertising	Sales
1	1	4
2	1	5
3	2	6
4	2	7
5	4	8

In quarter 3, a new product was introduced that would influence sales in quarters 3, 4, and 5.

The following model is established:

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where y = sales, x_1 = advertising expenditures, x_2 is a variable that is 1 when the new product is available and 0 otherwise, and ε is an error component.

Find the least squares estimate of β_1.

(A) 25/14 (B) 29/16 (C) 33/16

(D) 29/14 (E) 33/14

Solution. The design matrix is $\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 2 & 1 \\ 2 & 1 \\ 4 & 1 \end{pmatrix}$, so

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & 1 & 2 & 2 & 4 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 2 & 1 \\ 2 & 1 \\ 4 & 1 \end{pmatrix} = \begin{pmatrix} 26 & 8 \\ 8 & 3 \end{pmatrix} \quad \Rightarrow \quad (\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{14} \begin{pmatrix} 3 & -8 \\ -8 & 26 \end{pmatrix}.$$

The LSE of β_1 is the first component of

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \frac{1}{14} \begin{pmatrix} 3 & -8 \\ -8 & 26 \end{pmatrix} \begin{pmatrix} 67 \\ 21 \end{pmatrix} = \begin{pmatrix} \boxed{33/14} \\ * \end{pmatrix}. \quad \textbf{(Answer: (E))}$$ □

Type 3. If you do need to compute $(\mathbf{X}'\mathbf{X})^{-1}$ in an exam, chances are that $\mathbf{X}'\mathbf{X}$ is diagonal. Simply invert the diagonal entries to obtain the matrix inverse. Yes, the dataset needs to be carefully manipulated for this to happen!

Example 2.1.3. (SOA Course 4 Fall 2001 Question 13: LSE as a weighted average of response values) You fit the following model to four observations:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad i = 1, 2, 3, 4$$

You are given:

i	x_{1i}	x_{2i}
1	-3	-1
2	-1	3
3	1	-3
4	3	1

The least squares estimator of β_2 is expressed as $\hat{\beta}_2 = \sum_{i=1}^{4} w_i y_i$.

Determine (w_1, w_2, w_3, w_4).

(A) $(-0.15, -0.05, 0.05, 0.15)$

(B) $(-0.05, 0.15, -0.15, 0.05)$

(C) $(-0.05, 0.05, -0.15, 0.15)$

(D) $(-0.3, -0.1, 0.1, 0.3)$

(E) $(-0.1, 0.3, -0.3, 0.1)$

Solution. With $\mathbf{X} = \begin{pmatrix} 1 & -3 & -1 \\ 1 & -1 & 3 \\ 1 & 1 & -3 \\ 1 & 3 & 1 \end{pmatrix}$, we have $\mathbf{X}'\mathbf{X} = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 20 \end{pmatrix}$, which is diagonal, and thus $(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 1/20 & 0 \\ 0 & 0 & 1/20 \end{pmatrix}$. Then

$$\begin{aligned} \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} &= \begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 1/20 & 0 \\ 0 & 0 & 1/20 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ -3 & -1 & 1 & 3 \\ -1 & 3 & -3 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} \\ &= \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ -3/20 & -1/20 & 1/20 & 3/20 \\ -1/20 & 3/20 & -3/20 & 1/20 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}, \end{aligned}$$

i.e., $\boxed{\hat{\beta}_2 = -0.05y_1 + 0.15y_2 - 0.15y_3 + 0.05y_4}$. **(Answer: (B))** □

Remark. Every LSE must be a linear combination of the response values. See page 114 for more details.

Example 2.1.4. (SOA Course 120 November 1990 Question 19: Another diagonal $\mathbf{X}'\mathbf{X}$) You are performing a multiple regression of the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

You have obtained the following data:

y	x_1	x_2
1	−1	−1
2	1	−1
3	−1	1
4	1	1

Determine $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2$.

(A) 3.5 (B) 4.0 (C) 4.5

(D) 5.0 (E) 5.5

Solution. Since $\mathbf{X}'\mathbf{X} = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}$ and $\mathbf{X}'\mathbf{y} = \begin{pmatrix} 10 \\ 2 \\ 4 \end{pmatrix}$, we have

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/4 \end{pmatrix} \begin{pmatrix} 10 \\ 2 \\ 4 \end{pmatrix} = \begin{pmatrix} 2.5 \\ 0.5 \\ 1 \end{pmatrix}.$$

In particular, $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 = 2.5 + 0.5 + 1 = \boxed{4}$. **(Answer: (B))** □

Type 4. *(Most likely in Exam SRM)* You are directly provided with the vector of LSEs $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 & \hat{\beta}_1 & \cdots & \hat{\beta}_p \end{pmatrix}'$ or other summarized model output, based on which you will do some further analysis. You will see more examples of this sort in the later part of this chapter.

Having found the LSEs from the observed data, we can, as in SLR, compute the *fitted values*

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}, \quad i = 1, 2, \ldots, n,$$

which, for various values of the x_{ij}'s, determine the *fitted regression plane* (as opposed to a line in SLR), and the *residuals* $e_i = y_i - \hat{y}_i$. For $i = 1, 2, \ldots, n$, the ith residual is the difference between the observed and fitted response values for the ith observation.

Example 2.1.5. (CAS Exam MAS-I Fall 2018 Question 35: Given a bunch of matrices!) You are fitting a linear regression model of the form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}; \quad \varepsilon_i \sim \mathrm{N}(0, \sigma^2)$$

and are given the following values used in this model:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 & 9 \\ 1 & 1 & 1 & 15 \\ 1 & 1 & 1 & 8 \\ 1 & 1 & 0 & 7 \\ 1 & 1 & 0 & 6 \\ 1 & 0 & 0 & 6 \end{bmatrix}; \quad \mathbf{y} = \begin{bmatrix} 19 \\ 32 \\ 19 \\ 17 \\ 13 \\ 15 \end{bmatrix}; \quad \mathbf{X}'\mathbf{X} = \begin{bmatrix} 6 & 4 & 3 & 51 \\ 4 & 4 & 2 & 36 \\ 3 & 2 & 3 & 32 \\ 51 & 36 & 32 & 491 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 1.75 & -0.20 & 0.54 & -0.20 \\ -0.20 & 0.84 & 0.25 & -0.06 \\ 0.54 & 0.25 & 1.38 & -0.16 \\ -0.20 & -0.06 & -0.16 & 0.04 \end{bmatrix}; \quad (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} 2.335 \\ 0.297 \\ -0.196 \\ 1.968 \end{bmatrix}$$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \begin{bmatrix} 0.684 & 0.070 & 0.247 & -0.171 & -0.146 & 0.316 \\ 0.070 & 0.975 & -0.044 & 0.108 & -0.038 & -0.070 \\ 0.247 & -0.044 & 0.797 & 0.063 & 0.184 & -0.247 \\ -0.171 & 0.108 & 0.063 & 0.418 & 0.411 & 0.171 \\ -0.146 & -0.038 & 0.184 & 0.411 & 0.443 & 0.146 \\ 0.316 & -0.070 & -0.247 & 0.171 & 0.146 & 0.684 \end{bmatrix}$$

Calculate the residual for the 5th observation.

(A) Less than −1

(B) At least −1, but less than 0

(C) At least 0, but less than 1

(D) At least 1, but less than 2

(E) At least 2

Comments: On first encounter, this question with so many matrices seems intimidating. However, all you have to do is extract the entries relevant to the 5th observation and perform some simple calculations.

Solution. Recall that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is the vector of LSEs and the 5th row of the design matrix $\mathbf{X}$ carries the predictor variable values for the 5th observation. With $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (2.335, 0.297, -0.196, 1.968)$ and $(x_{50}, x_{51}, x_{52}, x_{53}) = (1, 1, 0, 6)$, the fitted value of the 5th observation is

$$\hat{y}_5 = \hat{\beta}_0 + \hat{\beta}_1(1) + \hat{\beta}_2(0) + \hat{\beta}_3(6) = 2.335 + 0.297 + 1.968(6) = 14.44.$$

Therefore, the residual for the 5th observation is $e_5 = y_5 - \hat{y}_5 = 13 - 14.44 = \boxed{-1.44}$. **(Answer: (A))** □

Remark. (i) As in an SLR model, residuals for an MLR model satisfy zero-to-sum constraints. With p predictors here, we have $p+1$ sum-to-zero constraints: $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} x_{ij}e_i = 0$ for all $j = 1, \ldots, p$.

(ii) The matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the hat matrix and will be of use in Subsection 3.2.1.

Example 2.1.6. (CAS Exam MAS-I Spring 2019 Question 32: Can you sense something unusual?) You are fitting the following linear regression model with an intercept:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}; \quad \varepsilon_i \sim \mathrm{N}(0, \sigma^2)$$

and are given the following values used in this model:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 & 9 \\ 1 & 1 & 1 & 15 \\ 1 & 1 & 1 & 8 \\ 0 & 1 & 1 & 7 \\ 0 & 1 & 1 & 6 \\ 0 & 0 & 1 & 6 \end{bmatrix}; \quad \mathbf{y} = \begin{bmatrix} 21 \\ 32 \\ 19 \\ 17 \\ 13 \\ 15 \end{bmatrix}; \quad \mathbf{X}'\mathbf{X} = \begin{bmatrix} 3 & 2 & 3 & 32 \\ 2 & 4 & 4 & 36 \\ 3 & 4 & 6 & 51 \\ 32 & 36 & 51 & 491 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 1.38 & 0.25 & 0.54 & -0.16 \\ 0.25 & 0.84 & -0.20 & -0.06 \\ 0.54 & -0.20 & 1.75 & -0.20 \\ -0.16 & -0.16 & -0.20 & 0.04 \end{bmatrix}$$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \begin{bmatrix} 0.684 & 0.070 & 0.247 & -0.171 & -0.146 & 0.316 \\ 0.070 & 0.975 & -0.044 & 0.108 & -0.038 & -0.070 \\ 0.247 & -0.044 & 0.797 & 0.063 & 0.184 & -0.247 \\ -0.171 & 0.108 & 0.063 & 0.418 & 0.411 & 0.171 \\ -0.146 & -0.038 & 0.184 & 0.411 & 0.443 & 0.146 \\ 0.316 & -0.070 & -0.247 & 0.171 & 0.146 & 0.684 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} 0.297 \\ -0.032 \\ 3.943 \\ 1.854 \end{bmatrix}; \quad \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} 20.93 \\ 32.03 \\ 19.04 \\ 16.89 \\ 15.04 \\ 15.07 \end{bmatrix}; \quad \sigma^2 = 0.012657$$

Calculate the modeled estimate of the intercept parameter.

(A) Less than 0

(B) At least 0, but less than 1

(C) At least 1, but less than 2

(D) At least 2, but less than 3

(E) At least 3

Comments: Watch out! The given design matrix is unusual in some way!

Solution. Note that the third column of the design matrix $\mathbf{X}$ consists of all 1's and corresponds to the intercept, so the LSE of the intercept should be the third (not the first!) entry of the vector of LSEs, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. The estimated intercept is $\hat{\beta}_0 = \boxed{3.943}$. **(Answer: (E))** □

Remark. (i) If you take the first entry of $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ as the answer, then you will be led to Answer B, which is incorrect.

(ii) This MAS-I exam question has drawn a lot of criticism from various candidates on online actuarial forums. Many decided to write to the CAS in an attempt to invalidate this question (unfortunately, to no avail). My opinion is that in practice, it is extremely rare that one would use a column other than the first to represent the intercept. The only motivation to do so is to create confusion and trick candidates in an exam!

Interpretations of regression coefficients. How do we interpret the coefficients in the MLR setting? If x_j is a continuous variable, then we can interpret $\beta_j = \partial\mathbb{E}[y]/\partial x_j$ as the *expected* change in y (also called the *expected effect* on y) per *unit change* in x_j, *holding all other predictors fixed.* The "everything else fixed" assumption is part of the definition when computing partial derivatives, as you learned in a multi-variable calculus class.

Sometimes the response variables and/or predictors are measured on logarithmic scale. In these cases, the parameter interpretation will differ somewhat:

- If the response variable is $\ln y$, i.e., the model is $\ln y = \beta_0 + \cdots + \beta_j x_j + \cdots + \varepsilon$, then

$$\beta_j = \frac{\partial \ln y}{\partial x_j} \overset{\text{(chain rule)}}{=} \frac{\partial y/\partial x_j}{y},$$

  which is the change in y for a small change in x *as a proportion of* y. For example, if $\ln y = \beta_0 + \cdots + 0.2x_j + \cdots + \varepsilon$, then as x_j increases by 0.1, we expect $\ln y$ to increase by $0.2(0.1) = 0.02$ and, upon exponentiation, y to increase by $\mathrm{e}^{0.02} - 1 = 2.02\%$ in proportion. (This is close to, but not exactly the same as $\beta_j \times$ change in $x_j = 0.2(0.1) = 2\%$ because the change in x_j is not infinitesimally small.)

- If the response variable is $\ln y$ and one of the predictors is $\ln x_j$, i.e., the model is $\ln y = \beta_0 + \cdots + \beta_j \ln x_j + \cdots + \varepsilon$, then

$$\beta_j = \frac{\partial \ln y}{\partial \ln x_j} = \frac{\partial y/y}{\partial x_j/x_j},$$

which is the ratio of the *percentage* change in y to the *percentage* change in x. This is known as *elasticity*, which is a concept emanating from economics.[iii]

As you will see shortly, many of the model fitting, statistical inference, and prediction concepts in SLR can carry over to MLR without substantial differences, just that the degrees of freedom of some probabilistic quantities need to be updated to reflect the increase in the number of predictors (from 1 to p).

ANOVA table. In addition to the structure of the model equation and the definition of fitted values and residuals, many other results established for SLR carry over directly to an MLR model, with the exception of some cosmetic notational differences owing to the presence of an increased number of predictors. For example, the ANOVA identity

$$\text{TSS} = \text{RSS} + \text{Reg SS}$$

continues to hold true even in the multiple linear regression setting. The ANOVA table possesses the same structure as that in SLR (see Section 1.2), except that the *df* column needs to be updated to reflect the fact that there are now p predictors:

Source	Sum of Squares	*df*	Mean Square	F
Regression	Reg SS	p	Reg SS$/p$	$\dfrac{\text{Reg SS}/p}{\text{RSS}/[n-(p+1)]}$
Error	RSS	$n-(p+1)$	$s^2 = \text{RSS}/[n-(p+1)]$	
Total	TSS	$n-1$		

(**Note:** $p+1$ is the total number of regression coefficients in the model, *including the intercept*.)

Here, the MSE s^2 can be shown to be an unbiased estimator for the error variance σ^2, as in the SLR framework, but the F-statistic is now for testing whether the p predictors are *collectively* useful for explaining the response variable:

$$\boxed{\text{H}_0: \underbrace{\beta_1 = \beta_2 = \cdots = \beta_p = 0}_{\text{intercept-only model}} \quad \text{vs.} \quad \text{H}_a: \underbrace{\text{at least one } \beta_j \text{ is non-zero}}_{\text{MLR model}}.}$$

Under H_0, the F-statistic[iv] is expected to take a value close to 1.[v] Under H_a, the F-statistic is expected to be greater than 1. Note that the proper interpretation of the result of the F-test in the MLR setting is:

> If H_0 is rejected, then we have strong evidence that *at least one* of the p predictors is an important predictor for the response variable. However, we do *not know which of these predictors are really useful!*

[iii] You probably have seen the concept of option elasticity in Exam IFM.

[iv] If you are interested in knowing, the F-statistic follows an $F_{p,n-(p+1)}$ distribution under $\text{H}_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$.

[v] Page 76 of ISLR provides a heuristic justification of why the F-statistic is expected to be close to 1 under H_0 and much larger than 1 under H_a. The reason is that we always have $\mathbb{E}[\text{RSS}/(n-p-1)] = \sigma^2$, no matter whether H_0 or H_a is true, and $\mathbb{E}[\text{Reg SS}/p] \begin{cases} = \sigma^2, & \text{under H}_0 \\ > \sigma^2, & \text{under H}_a \end{cases}$.

Example 2.1.7. (CAS Exam MAS-I Fall 2018 Question 32: Calculation of F-statistic given sums of squares) An actuary uses a multiple regression model to estimate money spent on kitchen equipment using income, education, and savings. He uses 20 observations to perform the analysis and obtains the following output:

Coefficient	Estimate	Standard Error	t-value
Intercept	0.15085	0.73776	0.20447
Income	0.26528	0.10127	2.61953
Education	6.64357	2.01212	3.30178
Savings	7.31450	2.73977	2.66975

	Sum of
	Squares
Regression	2.65376
Total	7.62956

He wants to test the following hypothesis:

- $\text{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0$
- H_1 : At least one of $\beta_1, \beta_2, \beta_3 \neq 0$

Calculate the value of the F-statistic used in this test.

(A) Less than 1

(B) At least 1, but less than 3

(C) At least 3, but less than 5

(D) At least 5

(E) The answer cannot be computed from the information given.

Solution. There are three predictors, so $p = 3$. Given the values of Reg SS and TSS, the value of the F-statistic is

$$F = \frac{\text{Reg SS}/3}{\text{RSS}/(n-p-1)} = \frac{2.65376/3}{(7.62956 - 2.65376)/(20-3-1)} = \boxed{2.8444}. \quad \textbf{(Answer: (B))}$$

□

Remark. The given coefficient estimates, standard errors, and t-values are all redundant.

Coefficient of determination. Based on the ANOVA table, the coefficient of determination R^2 is again defined by

$$\boxed{R^2 = \frac{\text{Reg SS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}}$$

and measures the proportion of the variation of the response variable (about its mean) that can be explained by the MLR model. In the MLR framework, R^2 is no longer the square of the sample correlation between y and individual predictors (we don't have a single x anymore!); however, it continues to be the *square* of the sample correlation between the observed y and the fitted value $\hat{y}$,[vi] i.e.,

$$\boxed{R^2 = \text{corr}(y, \hat{y})^2.}$$

Frees suggests referring to $R = \sqrt{R^2}$ (the positive square root of R^2) as the *multiple correlation coefficient*, which can be interpreted as the correlation between the response and the best linear combination of the explanatory variables (the fitted values).

Example 2.1.8. (SOA Exam SRM Sample Question 24: Going from F-statistic to R^2) Sarah performs a regression of the return on a mutual fund (y) on four predictors plus an intercept. She uses monthly returns over 105 months.

Her software calculates the F-statistic for the regression as $F = 20.0$, but then it quits working before it calculates the value of R^2. While she waits on hold with the help desk, she tries to calculate R^2 from the F-statistic.

Determine which of the following statements about the attempted calculation is true.

(A) There is insufficient information, but it could be calculated if she had the value of the residual sum of squares (RSS).

(B) There is insufficient information, but it could be calculated if she had the value of the total sum of squares (TSS) and RSS.

(C) $R^2 = 0.44$

(D) $R^2 = 0.56$

(E) $R^2 = 0.80$

Solution. We are given in the first paragraph that $n = 105$. To relate the F-statistic and R^2, we divide the numerator and denominator of F by TSS to get

$$F = \frac{(\text{Reg SS})/p}{\text{RSS}/(n-p-1)} = \frac{R^2/p}{(1-R^2)/(n-p-1)} = \frac{R^2/4}{(1-R^2)/(105-4-1)} = 20,$$

which gives $R^2 = 4/9 \approx \boxed{0.44}$. **(Answer: (C))** □

[vi] The fact that $R^2 = \text{Corr}(y, \hat{y})^2$ is also true for SLR models. After all, SLR is a special case of MLR.

Remark. The general formula relating the F-statistic and R^2 in a p-predictor MLR model is

$$F = \frac{n-p-1}{p} \times \frac{R^2}{1-R^2}, \tag{2.1.3}$$

which generalizes (1.2.5) on page 41 (which is for $p = 1$).

Example 2.1.9. (What can we say given a "large" F-statistic?) Following Example 2.1.8, determine which of the following statements is true.

(A) At least one of the four predictors is related to the response variable.

(B) All of the four predictors are related to the response variable.

(C) At least one of the four predictors is not related to the response variable.

(D) All of the four predictors are not related to the response variable.

(E) None of (A), (B), (C), or (D) are true.

Solution. If $\text{H}_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ is true, then the expected value of the F-statistic is close to one. As its observed value of 20 is much larger than 1, we can deduce that H_0 is not true, which means that at least one of $\beta_1, \beta_2, \beta_3$, and β_4 is non-zero. This in turn implies that at least one of the four predictors is (linearly) associated with the response variable. **(Answer: (A))** □

Example 2.1.10. (CAS Exam MAS-I Spring 2018 Question 33: Going from R^2 to F-statistic) Consider a multiple regression model with an intercept, 3 independent variables and 13 observations. The value of $R^2 = 0.838547$.

Calculate the value of the F-statistic used to test the hypothesis $\text{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

(A) Less than 5

(B) At least 5, but less than 10

(C) At least 10, but less than 15

(D) At least 15, but less than 20

(E) At least 20

Solution. In terms of R^2, the F-statistic is

$$F = \frac{n-p-1}{p}\left(\frac{R^2}{1-R^2}\right) = \frac{13-3-1}{3}\left(\frac{0.838547}{1-0.838547}\right) = \boxed{15.5813}. \quad \textbf{(Answer: (D))}$$

□

Example 2.1.11. **(CAS Exam MAS-I Spring 2019 Question 30: Calculating $\boldsymbol{R^2}$ from $\boldsymbol{y_i}$'s and $\boldsymbol{\hat{y}_i}$'s)** You are given the following information about a linear model:

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

Observed Y's	Estimated Y's
2.441	1.827
3.627	3.816
5.126	5.806
7.266	7.796
10.570	9.785

- Residual Sum of Squares = 1.772

Calculate the R^2 of this model.

(A) Less than 0.6

(B) At least 0.6, but less than 0.7

(C) At least 0.7, but less than 0.8

(D) At least 0.8, but less than 0.9

(E) At least 0.9

Solution 1. The total sum of squares is

$$\text{TSS} = \sum_{i=1}^{5}(y_i - \bar{y})^2 = (2.441 - 5.806)^2 + \cdots + (10.570 - 5.806)^2 = 41.3610.$$

The R^2 of the model is

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{1.772}{41.3610} = \boxed{0.9572}. \quad \textbf{(Answer: (E))}$$

□

Solution 2 (Shorter and preferred!). Inputting $\{(y_i, \hat{y}_i)\}_{i=1}^{5}$ into a financial calculator and reading the sample correlation coefficient, we directly find $R^2 = r^2 = 0.978341^2 = \boxed{0.9572}$. **(Answer: (E))** □

Remark. (i) The given RSS can be computed as

$$\sum_{i=1}^{5}(y_i - \hat{y}_i)^2 = (2.441 - 1.827)^2 + \cdots + (10.570 - 9.785)^2 = 1.772.$$

As Solution 2 shows, the value of RSS is not required for calculating R^2.

(ii) If you use Solution 2, don't forget to square the sample correlation coefficient!

(iii) The form of the linear model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, plays no role in calculating R^2. It plays a role when calculating the adjusted R^2 (to be introduced in Subsection 4.3.1); we need to know how many predictors there are.

Distribution of LSEs. Under the assumption of i.i.d. normal errors, the response observations $y_1, \ldots, y_n$ are independent (but not identically distributed) normal random variables. Then as in Subsection 1.3.1, the vector of LSEs $\hat{\boldsymbol{\beta}} = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}$, as a non-random linear transformation of the response vector $\mathbf{y}$, is also normally distributed—this time a *multivariate* normal distribution with $p+1$ components. However, closed-form algebraic formulas for the standard errors of the individual LSEs are not easily available and are best represented in matrix terms. Symbolically, we have

$$\hat{\boldsymbol{\beta}} \sim \underset{\substack{\text{multivariate}\\ \text{normal dist.}}}{\mathrm{N}_{p+1}} \Big(\underset{\substack{\text{mean}\\ \text{vector}}}{\boldsymbol{\beta}} \, , \underset{\substack{\text{variance-covariance}\\ \text{matrix}}}{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}} \Big),$$

where:

- N_{p+1} denotes a $(p+1)$-dimensional multivariate normal distribution.
- The *mean vector* of the multivariate normal distribution is the parameter vector $\boldsymbol{\beta}$, i.e., $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$,[vii] or component-wise, $\mathbb{E}[\hat{\beta}_j] = \beta_j$ for all $j = 0, 1, \ldots, p$. In the language of mathematical statistics, we say that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.
- The *variance-covariance matrix* hosting the variances of and covariances between the LSEs is

  $$\boxed{\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.} \qquad \textbf{(2.1.4)}$$

  This is a $(p+1) \times (p+1)$ matrix whose general form is

  $$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} \mathrm{Var}(\hat{\beta}_0) & \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \cdots & \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_p) \\ \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \mathrm{Var}(\hat{\beta}_1) & \cdots & \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_p) & \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_p) & \cdots & \mathrm{Var}(\hat{\beta}_p) \end{pmatrix},$$

  where the diagonal entries provide the variances of the LSEs, and the off-diagonal entries provide the covariances between the LSEs. For example, $\mathrm{Var}(\hat{\beta}_1)$ is the 2nd diagonal entry and $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1)$ is the $(1,2)$-th entry (or $(2,1)$-th entry as the matrix is symmetric) of the variance-covariance matrix. The entries of the matrix depend on the unknown random error variance σ^2, so they are not computable in general. As in SLR, we can replace σ^2 by the MSE s^2 to get the *estimated* variances and covariances, and the *estimated* variance-covariance matrix is

  $$\boxed{\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = s^2(\mathbf{X}'\mathbf{X})^{-1}.} \qquad \textbf{(2.1.5)}$$

  In the special case of SLR, this matrix becomes a 2×2 matrix and its diagonal entries are given in (1.3.2):

  $$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} \widehat{\mathrm{Var}}(\hat{\beta}_0) & \widehat{\mathrm{Cov}}(\hat{\beta}_0, \hat{\beta}_1) \\ \widehat{\mathrm{Cov}}(\hat{\beta}_0, \hat{\beta}_1) & \widehat{\mathrm{Var}}(\hat{\beta}_1) \end{pmatrix} = \begin{pmatrix} s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) & * \\ * & \frac{s^2}{S_{xx}} \end{pmatrix}.$$

[vii] Here is a simple proof: Because expectation is linear and the design matrix $\mathbf{X}$ consists of non-random elements, we can take it outside the expectation operator, leading to

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbb{E}[\underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\text{non-random}}\mathbf{y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underbrace{\mathbb{E}[\mathbf{y}]}_{\mathbf{X}\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} = \boldsymbol{\beta}.$$

To derive (2.1.4), we make use of the following formula for the variance-covariance matrix of a non-random linear transformation of a random vector:

$$\text{Var}(\mathbf{AZ}) = \mathbf{A}\underbrace{\text{Var}(\mathbf{Z})}_{\text{cov. matrix}}\mathbf{A}', \tag{2.1.6}$$

where $\mathbf{Z}$ is a random vector, and $\mathbf{A}$ is a non-random matrix such that the matrix product $\mathbf{AZ}$ is well-defined. This result is the multidimensional generalization of the familiar univariate result $\text{Var}(aZ) = a^2\text{Var}(Z) = a \cdot \text{Var}(Z) \cdot a$ for any random variable Z and any real scalar a. Applying (2.1.6) with $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{Z} = \mathbf{y}$, we get[viii]

$$\text{Var}(\hat{\boldsymbol{\beta}}) = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\underbrace{\text{Var}(\mathbf{y})}_{\sigma^2\mathbf{I}_n}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' = \sigma^2\underbrace{[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{X}}_{\text{cancel}}(\mathbf{X}'\mathbf{X})^{-1}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

An exam question will not test (2.1.6), but learning it can help you understand where $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ comes from and some otherwise difficult results later.

The above distributional results cast some light on the optimality of the LSEs. It can be shown that within the class of *linear unbiased* estimators (i.e., estimators that are unbiased and can be expressed as linear combinations of the response values), LSEs are the *minimum variance*[ix] *unbiased estimator* of the parameter vector $\boldsymbol{\beta}$. This result is known as the *Gauss–Markov theorem*, which is true even when the random errors are not normally distributed.

Example 2.1.12. (SOA Course 4 Fall 2001 Question 35: What can we say about an alternative estimator?) You observe n independent observations from a process whose true model is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

You are given:

(i) $z_i = x_i^2$, for $i = 1, 2, \ldots, n$

(ii) $b_1^* = \dfrac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})(x_i - \bar{x})}$

Which of the following is true?

(A) b_1^* is a non-linear estimator of β_1.

(B) b_1^* is a heteroscedasticity-consistent estimator (HCE) of β_1.

(C) b_1^* is a linear biased estimator of β_1.

(D) b_1^* is a linear unbiased estimator of β_1, but not the best linear unbiased estimator (BLUE) of β_1.

[viii] Recall from your linear algebra class that $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$ for any two matrices $\mathbf{A}$ and $\mathbf{B}$ such that $\mathbf{AB}$ is well-defined.

[ix] (If you are interested) In a multivariate framework, the fact that $\hat{\boldsymbol{\beta}}$ has the "minimum variance" means that if $\hat{\boldsymbol{\beta}}'$ is another estimator, then the difference of the two variance-covariance matrices, $\text{Var}(\hat{\boldsymbol{\beta}}') - \text{Var}(\hat{\boldsymbol{\beta}})$, is a non-negative definite matrix.

(E) b_1^* is the best linear unbiased estimator (BLUE) of β_1.

Solution. Note that b_1^* is a linear estimator because

$$b_1^* = \frac{\sum(z_i-\bar{z})(y_i-\bar{y})}{\sum(z_i-\bar{z})(x_i-\bar{x})} = \frac{\sum(z_i-\bar{z})y_i}{\sum(z_i-\bar{z})(x_i-\bar{x})} - \frac{\bar{y}\overbrace{\sum(z_i-\bar{z})}^{0}}{\sum(z_i-\bar{z})(x_i-\bar{x})} = \frac{\sum(z_i-\bar{z})y_i}{\sum(z_i-\bar{z})(x_i-\bar{x})},$$

which is a linear combination of the y_i's. It is also unbiased because

$$\mathbb{E}[b_1^*] = \frac{\sum(z_i-\bar{z})(\mathbb{E}[y_i]-\mathbb{E}[\bar{y}])}{\sum(z_i-\bar{z})(x_i-\bar{x})} = \frac{\sum(z_i-\bar{z})[\beta_1(x_i-\bar{x})]}{\sum(z_i-\bar{z})(x_i-\bar{x})} = \beta_1.$$

However, b_1^* is not the LSE $\hat{\beta}_1 = S_{xy}/S_{xx}$, so b_1^* is not the best linear unbiased estimator of β_1. **(Answer: (D))** □

Remark. Heteroscedasticity-consistent estimators are concerned with estimating variances when the random errors have unequal variances; see page 241 for details.

Example 2.1.13. (SOA Course 4 Fall 2003 Question 36: Standard error of a linear combination of LSEs) For the model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$, you are given:

(i) $n = 15$

(ii) $(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 13.66 & -0.33 & 2.05 & -6.31 \\ -0.33 & 0.03 & 0.11 & 0.00 \\ 2.05 & 0.11 & 2.14 & -2.52 \\ -6.31 & 0.00 & -2.52 & 4.32 \end{pmatrix}$

(iii) RSS $= 282.82$

Calculate the standard error of $\hat{\beta}_2 - \hat{\beta}_1$.

(A) 6.4 (B) 6.8 (C) 7.1

(D) 7.5 (E) 7.8

Comments: In many exam questions testing the use of (2.1.5), you are typically given the matrix $(\mathbf{X}'\mathbf{X})^{-1}$. You will have to extract and multiply the appropriate entries of this matrix by the MSE s^2 to get the desired estimated variances and covariances, as this example illustrates.

Solution. Let's begin by breaking down the estimated variance of $\hat{\beta}_2 - \hat{\beta}_1$ into

$$\widehat{\mathrm{Var}}(\hat{\beta}_2 - \hat{\beta}_1) = \widehat{\mathrm{Var}}(\hat{\beta}_2) \underset{(\text{not } -!)}{+} \widehat{\mathrm{Var}}(\hat{\beta}_1) - 2\widehat{\mathrm{Cov}}(\hat{\beta}_1, \hat{\beta}_2).$$

Practice Exam 1

1. For a simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for $i = 1, \ldots, 15$, you are given:

(i) $\sum_{i=1}^{15} x_i = 73$ and $\sum_{i=1}^{15}(x_i - \bar{x})^2 = 125.7333$

(ii) $\sum_{i=1}^{15} y_i = 1815$ and $\sum_{i=1}^{15}(y_i - \bar{y})^2 = 61032$

(iii) The sample correlation coefficient between x and y is 0.9873.

Calculate the least squares estimate of β_0.

○ (A) 15
○ (B) 18
○ (C) 20
○ (D) 22
○ (E) 25

2. For a simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for $i = 1, \ldots, 20$, you are given:

(i) $\sum_{i=1}^{20}(x_i - \bar{x})^2 = 1,000$

(ii) $\sum_{i=1}^{20}(y_i - \bar{y})^2 = 640$

(iii) The least squares estimate of β_1 is -0.75.

Calculate the value of the F-statistic for testing the significance of x.

○ (A) 131
○ (B) 132
○ (C) 133
○ (D) 134
○ (E) 135

3. For a simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ fitted to 20 observations, you are given:

(i) The least squares estimate of β_1 is 4.5.

(ii) The sample standard deviation of the explanatory variable is 5.

(iii) The residual standard error is 50.

You use a t-test to test $\text{H}_0 : \beta_1 = 0$ against $\text{H}_a : \beta_1 \neq 0$.

Determine which of the following statements concerning the result of the test is correct.

○ (A) Do not reject H_0 at the 0.100 significance level.

○ (B) Reject H_0 at the 0.100 significance level, but not at the 0.050 significance level.

○ (C) Reject H_0 at the 0.050 significance level, but not at the 0.025 significance level.

○ (D) Reject H_0 at the 0.025 significance level, but not at the 0.010 significance level.

○ (E) Reject H_0 at the 0.010 significance level.

4. You fit the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ to 10 observed values (x_i, y_i).

You are given:

$$\begin{aligned}
\sum (y_i - \hat{y}_i)^2 &= 2.79 \\
\sum (x_i - \bar{x})^2 &= 180 \\
\sum (y_i - \bar{y})^2 &= 152.40 \\
\bar{x} &= 6 \\
\bar{y} &= 7.78
\end{aligned}$$

Determine the width of the symmetric 95% prediction interval for y_* when $x_* = 8$.

○ (A) 1

○ (B) 2

○ (C) 3

○ (D) 4

○ (E) 5

5. For the multiple linear regression model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ with $i = 1, \ldots, 15$, you are given:

(i) $$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 1.00 & 0.25 & 0.25 \\ 0.25 & 0.50 & -0.25 \\ 0.25 & -0.25 & 2.00 \end{pmatrix}$$

(ii) $\hat{\beta}_0 = 10$ and $\hat{\beta}_1 = 12$

(iii) The 98% symmetric confidence interval for β_2 is $(9.638, 20.362)$.

Calculate the 99% symmetric prediction interval for y_* observed at $x_{*1} = 1.5$ and $x_{*2} = 4.5$.

○ (A) $(79, 112)$

○ (B) $(75, 116)$

○ (C) $(71, 120)$

○ (D) $(67, 124)$

○ (E) $(63, 128)$

6. Consider the multiple linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. Let

$$\begin{aligned} R^2 &= \text{coefficient of determination of the above model,} \\ r_j &= \text{correlation coefficient between } y \text{ and } x_j, \text{ for } j = 1, 2, \\ r_{12} &= \text{correlation coefficient between } x_1 \text{ and } x_2. \end{aligned}$$

Determine which of the following inequalities is always correct.

○ (A) $R^2 \le r_{12}^2$

○ (B) $r_{12}^2 \le R^2$

○ (C) $R^2 \le r_1^2$

○ (D) $r_1^2 \le R^2$

○ (E) None of the above

7. Interviews were conducted with 15 street vendors to study their annual incomes. Data were collected on annual income (y), age (x_1) and the number of hours worked per day (x_2). The following multiple linear regression model is suggested for the data:

$$\text{Model (1)}: \quad y = \beta_0 + \beta_1 x_1 + \gamma_1 x_1^2 + \beta_2 x_2 + \varepsilon,$$

for some unknown parameters $\beta_0, \beta_1, \gamma_1, \beta_2$.

A number of alternative models are proposed in place of model (1) as follows:

(2) $y = \beta_0 + \beta_1 x_1 + \gamma_1 x_1^2 + \varepsilon$

(3) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

(4) $y = \beta_0 + \beta_1 x_1 + \varepsilon$

(5) $y = \beta_0 + \beta_2 x_2 + \varepsilon$

The following summarizes the residual sum of squares (RSS) obtained by fitting the above models:

Model	(1)	(2)	(3)	(4)	(5)
RSS	2,250,956	2,549,146	3,600,196	8,017,930	4,508,761

Calculate the F-statistic for testing for the significance of age.

○ (A) 2.2

○ (B) 3.3

○ (C) 4.4

○ (D) 5.5

○ (E) 6.6

8. For a heteroscedastic simple linear regression model, you are given:

(i) $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, for $i = 1, 2, \ldots, 5$

(ii) $\text{Var}(\varepsilon_i) \propto x_i^2$

(iii)

i	x_i	y_i
1	1	1
2	2	-2
3	4	5
4	9	-10
5	16	25

Calculate the weighted least squares estimate of β_1.

○ (A) 0.15
○ (B) 0.18
○ (C) 0.20
○ (D) 0.22
○ (E) 0.25

9. Using ordinary least squares, Steve has fitted a simple linear regression model to predict an individual's weight from the individual's height. It turns out that some of the individuals in the study are members of the same family and so have been exposed to the same environmental factors.

Determine which of the following statements about Steve's model is/are true.

I. The estimated standard errors of the coefficient estimates are lower than they should be.

II. Confidence and prediction intervals are narrower than they should be.

III. He may be led to erroneously conclude that height is a statistically significant predictor of weight when it is not.

○ (A) I only
○ (B) II only
○ (C) III only
○ (D) I, II, and III
○ (E) The correct answer is not given by (A), (B), (C), or (D).

10. Consider the following formula (all symbols carry their usual meaning):

$$\mathbb{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \operatorname{Var}\left(\hat{f}(x_0)\right) + \left[\operatorname{Bias}\left(\hat{f}(x_0)\right)\right]^2 + \operatorname{Var}(\varepsilon).$$

Determine which of the following statements about this formula is/are true.

I. It is for a quantitative response variable.

II. It refers to the squared discrepancy between the response variable of a previously unseen observation and the prediction of a model fitted to a fixed training set, averaged over a large number of test observations (x_0, y_0).

III. $\operatorname{Var}\left(\hat{f}(x_0)\right) + \left[\operatorname{Bias}\left(\hat{f}(x_0)\right)\right]^2$ is known as the irreducible error.

○ (A) I only

○ (B) II only

○ (C) III only

○ (D) I, II, and III

○ (E) The correct answer is not given by (A), (B), (C), or (D).

11. Determine which of the following statements about different resampling methods is/are true.

I. The validation set approach is a special case of k-fold cross-validation (CV).

II. LOOCV has lower bias than k-fold CV when $k < n$.

III. k-fold CV is generally less computationally expensive than LOOCV when $k < n$.

○ (A) I and II only

○ (B) I and III only

○ (C) II and III only

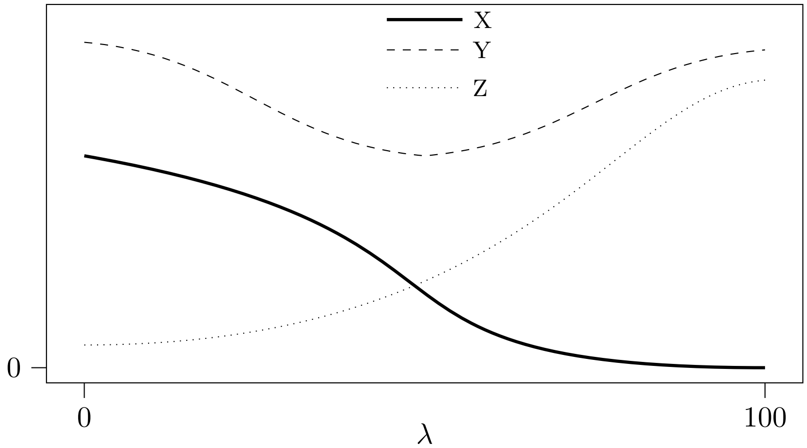
○ (D) I, II, and III

○ (E) The correct answer is not given by (A), (B), (C), or (D)

12. You are estimating the coefficients of a linear regression model by minimizing the sum:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2.$$

From this model, you have produced the following plot of various statistics as a function of the parameter, λ:



Determine which of the following sets of quantities best matches the three curves.

		X	Y	Z
○	(A)	Squared bias	Test MSE	Training MSE
○	(B)	Squared bias	Variance	Training MSE
○	(C)	Variance	Squared bias	Training MSE
○	(D)	Squared bias	Test MSE	Variance
○	(E)	Variance	Test MSE	Squared bias

13. A multiple linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ is fitted, leading to the following table of parameter estimates:

Variable	Estimate	Standard Error
Intercept	0.6	0.15
x_1	−0.2	0.30
x_2	0.5	0.65
x_3	−1.4	0.45

Determine which of the following variables will be eliminated in the first step of the backward selection procedure.

○ (A) Intercept

○ (B) x_1

○ (C) x_2

○ (D) x_3

○ (E) None should be dropped from the model

14. You are given the following information about a GLM:

- The model uses four categorical explanatory variables:

  (a) x_1 is a categorical variables with three levels.

  (b) x_2, x_3 are categorical variables with two levels.

  (c) x_4 is a categorical variable with six levels.

- The model also uses a continuous explanatory variable x_5 modeled with a first order polynomial.
- There is only one interaction in the model, which is between x_1 and x_5.

Determine the maximum number of parameters in this model.

○ (A) Less than 13

○ (B) 13

○ (C) 14

○ (D) 15

○ (E) At least 16

15. You are given the following GLM output:

Response variable	Pure Premium
Response distribution	Gamma
Link	Log
Scale parameter	1.1

Parameter	df	$\hat{\beta}$
Intercept	1	4.78
Risk Group	2	
Group 1	0	0.00
Group 2	1	-0.20
Group 3	1	-0.35
Vehicle Symbol	1	
Symbol 1	0	0.00
Symbol 2	1	0.42

Calculate the estimated variance of the pure premium for an insured in Risk Group 3 with Vehicle Symbol 1.

○ (A) 84

○ (B) 92

○ (C) 7044

○ (D) 7749

○ (E) 591253

16. You are given the following table for model selection:

Model	Negative Loglikelihood	Number of Parameters	AIC
Intercept + Age	A	5	435
Intercept + Vehicle Body	196	11	414
Intercept + Age + Vehicle Value	196	X	446
Intercept + Age + Vehicle Body + Vehicle Value	B	Y	500

Calculate B.

○ (A) 211

○ (B) 212

○ (C) 213

○ (D) 214

○ (E) 215

17. Determine which of the following statements about GLMs with normal responses and identity link function is/are true.

I. A large deviance indicates a poor fit for a model.

II. The deviance reduces to the residual sum of squares.

III. The deviance residual is the same as the Pearson residual.

○ (A) I only

○ (B) II only

○ (C) III only

○ (D) I, II, and III

○ (E) The correct answer is not given by (A), (B), (C), or (D).

18. You are given the following information for a logistic regression model to estimate the probability of a claim for a portfolio of independent policies:

(i) The model uses two explanatory variables:

(a) Age group, which is treated as a continuous explanatory variable taking values of 1, 2 and 3, modeled with a second order polynomial

(b) Sex, which is a categorical explanatory variable with two levels

(ii) Parameter estimates:

Parameter	$\hat{\beta}$
Intercept	-1.1155
Sex	
Female	0.0000
Male	-0.4192
Age group	1.2167
$(\text{Age group})^2$	-0.5412

(iii) A policy is predicted to have a claim if the fitted probability of a claim is greater than 0.25.

Determine which of the following policies is/are predicted to have claims.

Policy	Sex	Age Group
I	Male	1
II	Male	2
III	Female	3

○ (A) I only

○ (B) II only

○ (C) III only

○ (D) I, II, and III

○ (E) The correct answer is not given by (A), (B), (C), or (D).

19. You are given:

- $y_1, y_2, \ldots, y_n$ are independent Poisson random variables with respective means μ_i for $i = 1, 2, \ldots, n$.
- A Poisson GLM was fitted to the data with an identity link function:

  $$\mu_i = \beta_0 + \beta_1 x_i$$

  where x_i refers to the value of the explanatory variable of the ith observation.
- Analysis of the data produced the following output:

x_i	y_i	$\hat{\mu}_i$	$y_i \log(y_i/\hat{\mu}_i)$
-1	2	?	??
-1	3	?	??
0	6	7.45163	-1.30004
0	7	7.45163	-0.43766
0	8	7.45163	0.56807
0	9	7.45163	1.69913
1	10	12.38693	-2.14057
1	12	12.38693	-0.38082
1	15	12.38693	2.87112

Calculate the deviance of the model.

○ (A) 0.4
○ (B) 0.9
○ (C) 1.4
○ (D) 1.9
○ (E) There is not enough information to determine the answer.

20. You are given the following sample of size 6 from a time series:

1 1.5 1.6 1.4 1.5 1.7

Calculate the sample lag-3 autocorrelation.

○ (A) -0.25
○ (B) -0.04
○ (C) -0.03
○ (D) 0.21
○ (E) 0.25

21. You are given:

(i) The random walk model

$$y_t = y_0 + c_1 + c_2 + \cdots + c_t$$

where c_t, $t = 1, 2, \ldots, 8$ denote observations from a Gaussian white noise process.

(ii) The following eight observed values of y_t:

t	1	2	3	4	5	6	7	8
y_t	2	−1	4	7	11	13	17	16

(iii) $y_0 = 0$

(iv) The 7-step ahead forecast of y_{15}, $\hat{y}_{15}$, is determined based on the observed value of y_8.

Determine the 95% symmetric prediction interval for y_{15}.

○ (A) $(20, 40)$

○ (B) $(18, 42)$

○ (C) $(16, 44)$

○ (D) $(14, 46)$

○ (E) $(12, 48)$

22. You are performing out-of-sample validation for exponential smoothed forecasts with $w = 0.8$ and $\hat{s}_0 = 25$. The validation sample is:

t	y_t
1	20
2	30
3	60
4	40
5	15

Calculate the mean absolute percentage error.

○ (A) 30

○ (B) 35

○ (C) 40

○ (D) 45

○ (E) 50

23. For a stationary first-order autoregressive process $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$, you are given:

(i) Based on the observed series $\{y_{2001}, y_{2002}, \ldots, y_{2018}\}$, the estimated parameters are

$$\hat{\beta}_0 = 0.75, \qquad \hat{\beta}_1 = -0.6, \qquad s^2 = 0.5.$$

(iii) $y_{2018} = 10$

Determine the 95% forecast interval for y_{2020}.

○ (A) $(2.5, 5.3)$

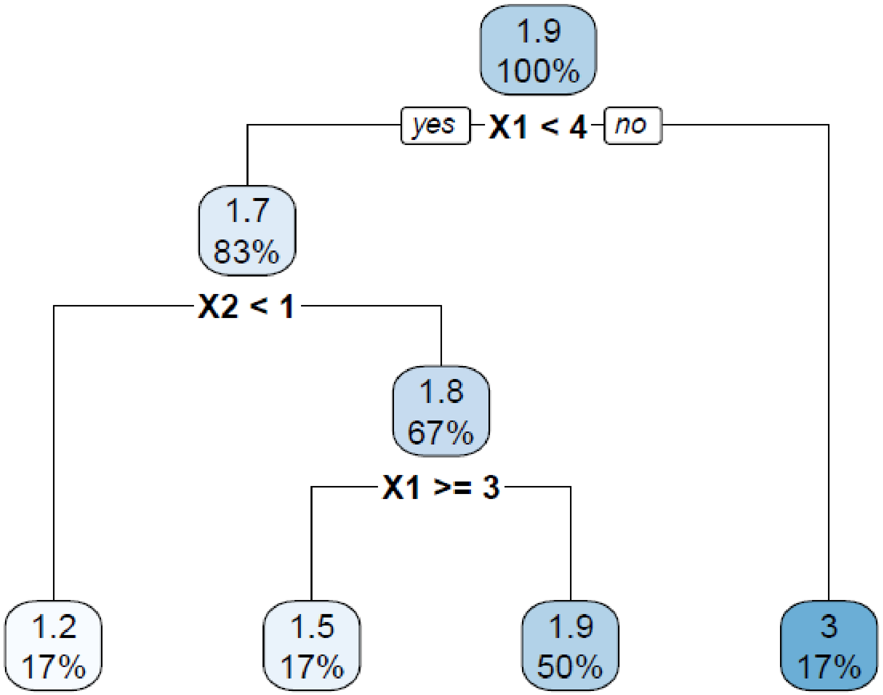
○ (B) $(2.4, 5.4)$

○ (C) $(2.3, 5.5)$

○ (D) $(2.2, 5.6)$

○ (E) $(2.1, 5.7)$

24. You are given the following regression tree using X_1 and X_2 as numeric predictors:



In each node, the first number is the response mean of the training observations in that node.

Determine which of the following statements is/are true.

I. The predicted value when $X_1 = 5$ and $X_2 = 2$ is 3.

II. The predicted value when $X_1 = 3$ and $X_2 = 0$ is 1.2.

III. The predicted value when $X_1 = 2$ and $X_2 = 3$ is.1.5

○ (A) None

○ (B) I and II only

○ (C) I and III only

○ (D) II and III only

○ (E) The correct answer is not given by (A), (B), (C), or (D).

25. Determine which of the following statements about recursive binary splitting for decision trees is/are true.

I. It is a greedy algorithm.

II. It is a top-down approach.

III. In making each split, one of the predictors is randomly chosen as the split variable.

○ (A) I only
○ (B) II only
○ (C) III only
○ (D) I, II, and III
○ (E) The correct answer is not given by (A), (B), (C), or (D).

26. Determine which of the following statements about cost complexity pruning for decision trees is/are true.

I. A tree split is made so long as the decrease in node impurity due to that split exceeds some threshold.

II. It has the effect of increasing the variance compared to an unpruned tree.

III. It is also known as strongest link pruning.

○ (A) None
○ (B) I and II only
○ (C) I and III only
○ (D) II and III only
○ (E) The correct answer is not given by (A), (B), (C), or (D).

27. Consider a classification tree with a binary response variable.

Determine which of the following statements is/are true.

I. A pure node is characterized by a classification error rate close to zero.

II. A pure node is characterized by a Gini index close to zero.

III. The classification error is always bounded from above by the Gini index.

○ (A) I only
○ (B) II only
○ (C) III only
○ (D) I, II, and III
○ (E) The correct answer is not given by (A), (B), (C), or (D).

28. Determine which of the following statements about the advantages and disadvantages of decision trees relative to generalized linear models is/are true.

I. Decision trees can be displayed graphically.

II. Decision trees can capture interactions between variables without the use of additional features.

III. Decision trees tend to be non-robust in the sense that a small change in the data can cause a large change in the final estimated tree.

○ (A) I only

○ (B) II only

○ (C) III only

○ (D) I, II, and III

○ (E) The correct answer is not given by (A), (B), (C), or (D).

29. You apply bagging with $B = 10$ to predict the probability that an insurance policy will lapse. Applying a classification tree to each bootstrapped sample, we obtain the following 10 estimates of the probability of lapse at the same set of predictor variable values:

$$0.1, \quad 0.15, \quad 0.15, \quad 0.25, \quad 0.55, \quad 0.6, \quad 0.6, \quad 0.65, \quad 0.7, \quad 0.75.$$

Determine which of the following combinations is correct.

	Average probability	Overall predicted class determined
	of lapse	by the majority vote approach
○ (A)	0.45	Lapse
○ (B)	0.45	Non-lapse
○ (C)	0.55	Lapse
○ (D)	0.55	Non-lapse
○ (E)	None of (A), (B), (C), or (D).	

30. Determine which of the following statements about bagging and random forests is/are true.

I. Bagging can only be used for regression problems while random forests can be used for both regression and classification problems.

II. Only bagging (but not random forests) involves randomly selecting predictors when making a split.

III. Only the test error of a bagged model (but not that of a random forest) can be estimated by out-of-bag estimation.

○ (A) I only
○ (B) II only
○ (C) III only
○ (D) I, II, and III
○ (E) The correct answer is not given by (A), (B), (C), or (D).

31. Determine which of the following statements about principal components analysis is/are true.

I. Scaling the variables has no effect on the results of principal components analysis.

II. Each principal component loading vector is unique.

III. If the number of principal components is one less than the number of observations, then the representation of observed data in terms of principal components is exact.

○ (A) I only
○ (B) II only
○ (C) III only
○ (D) I, II, and III
○ (E) The correct answer is not given by (A), (B), (C), or (D).

32. Determine for which of the following statistical learning methods the variables are typically standardized.

I. Shrinkage methods

II. Principal components analysis

III. K-means clustering

○ (A) None

○ (B) I and II only

○ (C) I and III only

○ (D) II and III only

○ (E) The correct answer is not given by (A), (B), (C), or (D).

33. Determine which of the following statements about K-means clustering and hierarchical clustering is/are true.

I. If the number of clusters is known a priori, then K-means clustering is always preferred over hierarchical clustering.

II. If hierarchical clustering is applied to n observations and n clusters are desired, then each cluster contains only one observation.

III. In each step of the K-means clustering algorithm, only one observation can move to another cluster.

○ (A) I only

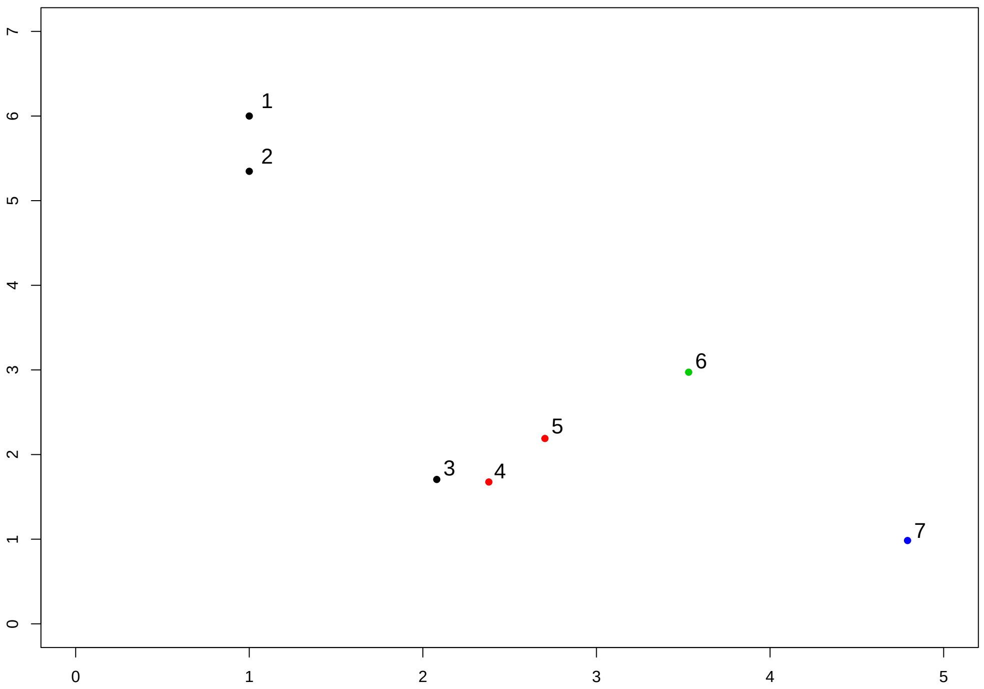
○ (B) II only

○ (C) III only

○ (D) I, II, and III

○ (E) The correct answer is not given by (A), (B), (C), or (D).

34. Consider seven two-dimensional data points numbered 1 through 7 shown in the following scatterplot:



Hierarchical clustering with single linkage is used to determine the clusters. If four clusters are desired, then they are given by

$$C_1 = \{1, 2\}, \quad C_2 = \{3, 4, 5\}, \quad C_3 = \{6\}, \quad C_4 = \{7\}.$$

Determine which two clusters will be merged if three clusters are desired.

○ (A) C_1 and C_2

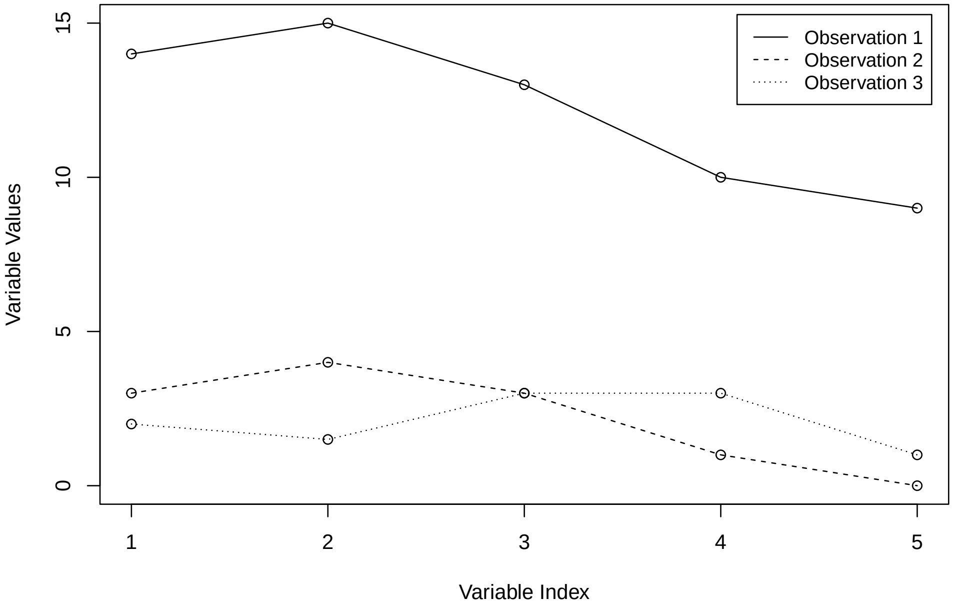
○ (B) C_1 and C_3

○ (C) C_1 and C_4

○ (D) C_2 and C_3

○ (E) C_2 and C_4

35. The following figure shows three observations, each with five variables.



Determine which of the following statements about the three observations is/are true.

I. Observations 1 and 2 are the closest in terms of Euclidean distance.

II. Observations 2 and 3 are the closest in terms of Euclidean distance.

III. Observations 1 and 2 are the closest in terms of correlation-based distance.

○ (A) None

○ (B) I and II only

○ (C) I and III only

○ (D) II and III only

○ (E) The correct answer is not given by (A), (B), (C), or (D).

Auto-Grade My Exam (Click Here)

****END OF PRACTICE EXAM 1****

1. (LSEs of β_0 and β_1 in SLR)

Solution. By (1.1.6), the LSE of β_1 is

$$\hat{\beta}_1 = r \times \frac{s_y}{s_x} = r \times \sqrt{\frac{\sum(y_i - \bar{y})^2/(n-1)}{\sum(x_i - \bar{x})^2/(n-1)}} = 0.9873 \times \sqrt{\frac{61032}{125.7333}} = 21.7522.$$

Then by (1.1.4), the LSE of β_0 is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1815}{15} - 21.7522 \times \frac{73}{15} = \boxed{15.14}. \quad \textbf{(Answer: (A))}$$

□

2. (Calculation of F-statistic from summarized data)

Solution 1 (Better). By (1.1.6),

$$-0.75 = \hat{\beta}_1 = r \times \frac{s_y}{s_x} = r \times \sqrt{\frac{640}{1,000}},$$

which gives $r = -0.9375$. Then $R^2 = r^2 = 0.878906$. (**Note:** Don't forget to square. ⚠) By (1.2.5), the F-statistic equals

$$F = (n-2)\left(\frac{R^2}{1-R^2}\right) = (20-2)\left(\frac{0.878906}{1-0.878906}\right) = \boxed{130.6452}. \quad \textbf{(Answer: (A))}$$

□

Solution 2. Alternatively, we can use the formula Reg SS $= \hat{\beta}_1^2 S_{xx} = (-0.75)^2(1,000) = 562.5$. Then by definition, the F-statistic equals

$$F = \frac{\text{Reg SS}/1}{\text{RSS}/(n-2)} = \frac{562.5}{(640-562.5)/(20-2)} = \boxed{130.6452}. \quad \textbf{(Answer: (A))}$$

□

3. (Result of a t-test for various α)

Solution. By (1.3.3), the standard error of $\hat{\beta}_1$ is

$$\text{SE}(\hat{\beta}_1) = \sqrt{\frac{s^2}{S_{xx}}} = \frac{50}{5\sqrt{19}} = 2.294157.$$

The t-statistic for $\text{H}_0 : \beta_1 = 0$ is

$$t(\hat{\beta}_1) = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)} = \frac{4.5 - 0}{2.294157} = 1.9615,$$

which is between $t_{18,0.050} = 1.7341$ and $t_{18,0.025} = 2.1009$. The critical region of the two-sided test is $\{|t(\hat{\beta}_1)| > t_{18,\alpha/2}\}$, where α is the size of the test. Thus $\text{H}_0 : \beta_1 = 0$ is rejected at $\alpha = 2(0.050) = 0.10$, but not at $\alpha = 2(0.025) = 0.05$. **(Answer: (B))** □

Remark. (i) Equivalently, the p-value of the test is between 0.05 and 0.10.

(ii) Answer (C) is for the one-sided alternative $\text{H}_a : \beta_1 > 0$.

Farewell Message

At long last, you have reached the end of your Exam SRM preparation. Although SRM is generally regarded as a relatively easy exam (well, by those who have passed it!), the amount of exam material is still enormous and the syllabus topics are by no means trivial. We are truly grateful to all of you for using this study manual. It is our sincere hope that you have found this manual useful for your exam preparation, and that the manual has increased your interest in predictive analytics, at least slightly.

Before the exam

- Visit the main author's personal website from time to time for any latest announcements or updates on this study manual:

  https://sites.google.com/site/ambroseloyp/publications/SRM.

- Attempt at least 2 to 3 of the six practice exams. Don't be too frustrated if you don't score well. Instead, use them to identify the syllabus topics in which you need more "training."

During the exam

- *(Important!)* Keep the "6 minutes per question" rule in mind and don't deviate from it too much. Don't spend a disproportionate amount of time (say, more than 10 minutes) on a single question. Remember: Each question, however difficult, carries the same weight.

- *(Very important!!)* Don't feel too bad if you get stumped in a question. The real exam likely has a small number of unfamiliar, more difficult questions to distinguish the very best candidates from the less outstanding, but passing candidates. Doing not so well in one or two questions is not enough to strip you of a pass. Remember: All we need is a pass, or Grade 6.

- According to the exam syllabus:

  > "There is no set requirement for the distribution of correct answers for the multiple-choice preliminary examinations. It is possible that a particular answer choice could appear many times on an examination or not at all. Candidates are advised to answer each question to the best of their ability, independently from how they have answered other questions on the exam."

  so there is no need to count how many A's, B's, C's, D's, and E's you have chosen in other questions that you are sure about, and pick the less common letters in the difficult questions.

After the exam

Unfortunately, instant pass/fail is still not available for Exam SRM yet (it has been 5 years since SRM was introduced! 😠). Instead, the pass list will be posted on

https://www.soa.org/education/general-info/exam-results/edu-exam-results-detail/

on Friday (9 am CST) approximately 8 weeks after each testing window ends. It will be quite a LOOOOONG wait! In the meantime, take a short break, relax as much as possible, and think about when to take Exams PA and ATPA.

Last but not least...

We Wish You The Best of Luck in Your SRM Exam
and Look Forward to Seeing You Again
in our PA Study Manual![vi] 👍

Runhuan Feng
Daniël Linders
Ambrose Lo

♢ ♣ ♡ ♠ ☺ **END** ☺ ♠ ♡ ♣ ♢

[vi] https://www.actexlearning.com/exams/pa

ACTEX Learning

Advanced Topics in Predictive Analytics - Video Course

Johnny Li, PhD, FSA

Designed to help you more easily climb the steep learning curve of the new SOA ATPA Exam. In this course, leading Professor of Predictive Analytics Johnny Li will provide you with detailed explanations of the three models outlined in Topic 3 of the ATPA syllabus. The strong technical knowledge developed through this course will enable students to write the ATPA exam more confidently, and increase the likelihood of passing the exam.

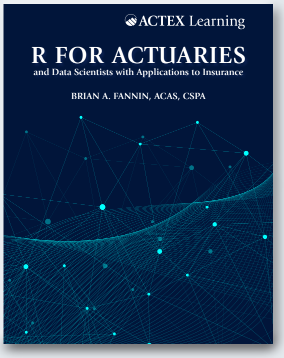
- **Lecture Slides** so you can follow along with the instructor and take notes.
- **Access to Instructor** during the duration of the course.
- **Access to Discussion Forum** to ask questions and see what other students are saying.
- **15+ Videos (6 Hours!)**, meticulously organized into three modules, helping you achieve the learning objectives of this course with ease.
- **3 End-of-Module Assessments** to evaluate your understanding of the material.
- **Certificate of Completion**

R for Actuaries and Data Scientists with Applications to Insurance

Brian Fannin, ACAS, CSPA

Written in a light, conversational style, this book will show you how to install and get up to speed with R in no time. It will also give you an overview of the key modeling techniques in modern data science including generalized linear models, decision trees, and random forests, and illustrates the use of these techniques with real datasets from insurance.

Engaging and at times funny, this book will be valuable for both newcomers to R and experienced practitioners who would like a better understanding of how R can be applied in insurance.



ACTEX Learning

R FOR ACTUARIES

and Data Scientists with Applications to Insurance

BRIAN A. FANNIN, ACAS, CSPA

ACTEX Study Manual for SOA Exam PA + Videos

Ambrose Lo, PhD, FSA, CERA

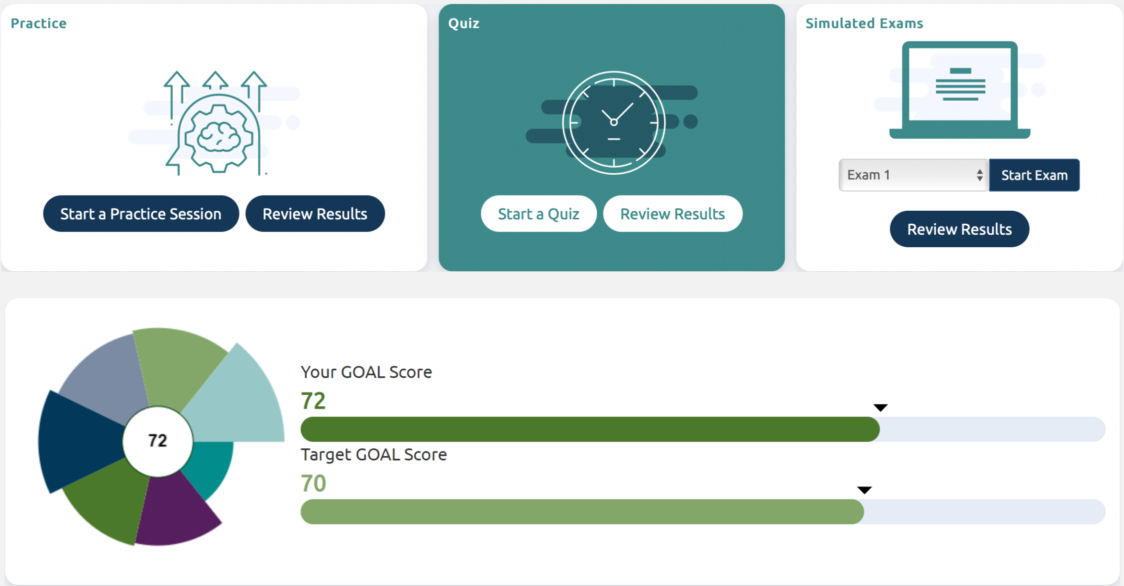
The ACTEX Study manual for Exam PA takes an integrated approach to learning Predictive Analytics with a three-component structure: A crash course in R covering the elements of R programming that are particularly germane to this exam, the theory and practice of predictive modeling techniques, and the discussions on past PA exams.

Included with the manual and videos:

- **110+ In-text Exercises**
- **2 Full-Length Sample Projects**
- **83 End-of-chapter Conceptual Review Questions**
- **Full Syllabus Coverage**
- **All datasets, R scripts, & R markdown files** used are available virtually
- **Over 50 instructional videos** to go along with the study manual. In these videos, Professor Lo will walk you through the fundamental concepts in Predictive Analytics, with a strong emphasis on key test items in Exam PA. Videos are not sold separately.

What's Your GOAL?

We'll help you break it down.



Your GOAL Score
72

Target GOAL Score
70

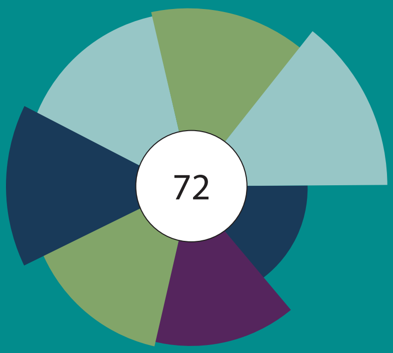
Customizable Exam Prep

- Three Learning Modes:
  - Practice
  - Quiz
  - Simulated Exams
- Three Difficulty Levels
  - Core
  - Advanced
  - Mastery

Practice. Predict. Pass.

Using GOAL Score to gauge your readiness

- Measure how prepared you are to pass your exam with a tool that suits any study approach.
- Flag problem areas for later, receive tips when you need them
- Analyze your strengths and weaknesses by category, topic, level of difficulty, performance on exercises, and consistency of your performance.
- A score of 70 or higher indicates you're ready for your exam, giving you a clear milestone in your studies.



Available for Exams P, FM, FAM, FAM-L, FAM-S, ALTAM, ASTAM, SRM, MAS-I, MAS-II, CAS 5, CAS 6 US & CAS 6 CAN