

ACTEX Learning

Study Manual for Exam PA

10th Edition

Ambrose Lo, PhD, FSA, CERA



An SOA Exam

 **ACTEX Learning**

**Study Manual for
Exam PA**

10th Edition

Ambrose Lo, PhD, FSA, CERA



Actuarial & Financial Risk Resource Materials
Since 1972

Copyright © 2023, ACTEX Learning, a division of ArchiMedia Advantage Inc.

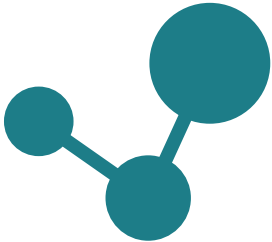
Printed in the United States of America.

No portion of this ACTEX Study Manual may be reproduced or transmitted in any part or by any means without the permission of the publisher.



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**:

 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

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

- ACTEX Manual for P →
- Probability for Risk Management, 3rd Edition 🔒
- GOAL for SRM 🔒
- ASM Manual for IFM 🔒
- Exam FAM-S Video Library 🔒


Related Topics ▾

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

Unlocked Products are the products that you own.

ACTEX Manual for P 

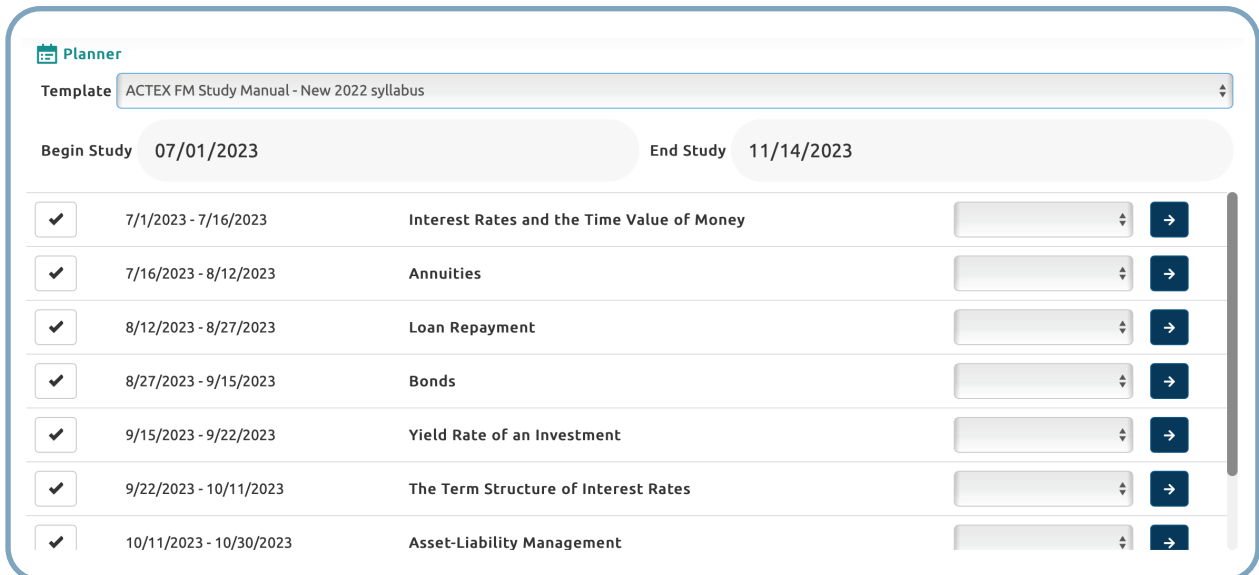
Locked Products are products that you do not own, and are available for purchase.

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

Checkmark	Period	Topic	Dropdown	Arrow
✓	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*

Contents

Preface		xi
1	About Exam PA	xii
2	About this Study Manual	xix
I	A Crash Course in R	1
Chapter 1	Basics of R Programming	3
1.1	Getting Started in R	5
1.1.1	Basic Infrastructure	5
1.1.2	Data Types	12
1.2	Data Structures	16
1.2.1	Vectors	16
1.2.2	Matrices	23
1.2.3	Data Frames	27
1.2.4	Lists	33
1.2.5	Sidebar: Functions	36
1.3	Basic Data Management	40
1.4	for Loops	58
1.5	End-of-Chapter Practice Problems	65
Chapter 2	Data Exploration and Visualization	75
2.1	Making ggplots	76
2.1.1	Basic Features	76
2.1.2	Customizing Your Plots	90
2.2	Data Exploration	92
2.2.1	Univariate Data Exploration	93
2.2.2	Bivariate Data Exploration	107
2.3	End-of-Chapter Practice Problems	119
II	Theory of and Case Studies in Predictive Analytics	129
Chapter 3	Linear Models	131
3.1	A Primer on Predictive Analytics	133
3.1.1	Basic Terminology	134

3.1.2	The Model Building Process	139
3.1.3	Bias-Variance Trade-off	166
3.1.4	Feature Generation and Selection	181
3.2	Linear Models: Conceptual Foundations	192
3.2.1	Model Formulation	192
3.2.2	Model Evaluation and Validation	195
3.2.3	Feature Generation	208
3.2.4	Feature Selection	233
3.2.5	Regularization	241
3.3	Case Study 1: Fitting Linear Models in R	249
3.3.1	Exploratory Data Analysis	251
3.3.2	Simple Linear Regression	257
3.3.3	Multiple Linear Regression	264
3.3.4	Evaluation of Linear Models	276
3.4	Case Study 2: Feature Selection and Regularization	280
3.4.1	Preparatory Work	280
3.4.2	Model Construction and Feature Selection	295
3.4.3	Model Validation	314
3.4.4	Regularization	318
	Conceptual Review Questions for Chapter 3	330
Chapter 4	Generalized Linear Models	335
4.1	Conceptual Foundations of GLMs	336
4.1.1	Selection of Target Distributions and Link Functions	339
4.1.2	Weights and Offsets	351
4.1.3	Fitting and Assessing the Performance of a GLM	356
4.1.4	Performance Metrics for Classifiers	371
4.2	Case Study 1: GLMs for Continuous Target Variables	387
4.2.1	Data Preparation	387
4.2.2	Model Construction and Evaluation	389
4.2.3	Model Validation and Interpretation	397
4.3	Case Study 2: GLMs for Binary Target Variables	401
4.3.1	Data Exploration and Preparation	402
4.3.2	Model Construction and Selection	415
4.3.3	Interpretation of Model Results	431
4.4	Case Study 3: GLMs for Count and Aggregate Loss Variables	436
4.4.1	Data Exploration and Preparation	436
4.4.2	Model Construction and Evaluation	446
4.4.3	Predictions	458
	Conceptual Review Questions for Chapter 4	463
Chapter 5	Tree-Based Models	467
5.1	Conceptual Foundations of Decision Trees	468
5.1.1	Single Decision Trees	468
5.1.2	Ensemble Tree Model I: Random Forests	501
5.1.3	Ensemble Tree Model II: Boosting	507

5.2	Mini-Case Study: A Toy Decision Tree	514
5.2.1	Basic Functions and Arguments	515
5.2.2	Pruning a Decision Tree	521
5.3	Extended Case Study: Classification Trees	528
5.3.1	Problem Set-up and Preparatory Steps	528
5.3.2	Construction and Evaluation of Single Classification Trees	541
5.3.3	Construction and Evaluation of Ensemble Trees	563
	Conceptual Review Questions for Chapter 5	585
Chapter 6	Unsupervised Learning Techniques	589
6.1	Principal Components Analysis	591
6.1.1	Conceptual Foundations	591
6.1.2	Additional PCA Issues	597
6.1.3	A Simple Case Study	605
6.2	Cluster Analysis	629
6.2.1	K -means Clustering	632
6.2.2	Hierarchical Clustering	641
6.2.3	Practical Issues in Clustering	651
6.2.4	A Simple Case Study	653
	Conceptual Review Questions for Chapter 6	670
III	Final Preparation	675
Chapter 7	Discussions on Past PA Exams	677
7.1	October 2023 Exam PA	680
7.2	April 2023 Exam PA	680
7.3	October 2022 Exam PA	698
7.3.1	October 11 Exam	699
7.3.2	October 12 Exam	716
7.4	April 2022 Exam PA	724
7.4.1	April 12 Exam	724
7.4.2	April 14 Exam	739
7.5	December 2021 Exam PA	748
7.5.1	December 13 Exam	749
7.5.2	December 14 Exam	764
7.6	June 2021 Exam PA	776
7.6.1	June 21 Exam	776
7.6.2	June 22 Exam	791
7.7	December 2020 Exam PA	803
7.7.1	December 7 Exam	803
7.7.2	December 8 Exam	814
7.8	June 2020 Exam PA	829
7.8.1	June 16 and 19 Exams	829
7.8.2	June 17 and 18 Exams	839
7.9	December 2019 Exam PA	850

7.10	June 2019 Exam PA	862
Chapter 8	Practice Exams	871
	Practice Exam 1 Project Statement	875
	Practice Exam 1 Suggested Solutions	890

Preface

⚠ NOTE TO STUDENTS ⚠

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

Why this Study Manual?

“The PA modules are so difficult to follow.”

“The PA modules make things unnecessarily complicated and are riddled with errors.”

“I feel that the PA modules don’t cover enough ground for me to handle the exam well. I have to supplement my learning with external resources.”

“I hate having to alternate among the PA modules, the R Markdown files, the required textbooks, and online readings.”

“There is a lack of useful study resources for Exam PA in the market.”

These are some of the most common comments PA exam candidates who studied for the exam solely using the Society of Actuaries (SOA)’s e-learning modules have voiced on Internet forums, e.g., the old Actuarial Outpost, Reddit 🗨, Discord 🗨. These “complaints” and the importance of passing this exam to earn the Associateship of the Society of Actuaries (ASA) designation in today’s exam curriculum have motivated me to develop a completely new Exam PA study manual with the goal of streamlining, synthesizing, and augmenting the materials in the PA e-learning modules in a coherent and exam-oriented format. With this manual, you will have in your possession a reliable learning resource that hosts all of the useful materials in a single place and shows you how to prepare for this exam effectively and efficiently. There is no longer a need to alternate among the e-learning modules, the suggested textbooks in the syllabus, R markdown files, and additional online readings. Starting from the very basics and adopting a case study approach, we will learn fundamental concepts in predictive analytics, implement predictive models in R (a powerful programming language) step by step in concrete settings, make some fancy and informative graphs 📊, understand what the output in R means, and write your responses to the liking of PA exam graders. No prior knowledge in R or the SRM exam material is assumed.

1 About Exam PA

Exam Administrations

Exam PA (Predictive Analytics) is a 3.5-hour computer-based exam offered for the first time in December 2018 by the SOA. There are two sittings each year, one in April and one in October, and each testing window lasts for four days.ⁱ In April 2024, the exam will be delivered via computer-based testing (CBT) in a Prometric exam center on **April 16-19**. The registration deadline is March 12. You can check out the exam’s official homepage for more information:



<https://www.soa.org/education/exam-req/edu-exam-pa-detail/>.

Once you have registered for the exam and paid the (exorbitant! \$\$\$) \$1,125 exam fee, you will receive access to the PA e-learning modules until the end of the month in which the exam is administered (April 30 for the April sitting, October 31 for the October sitting). According to the exam homepage and syllabus, these modules

“provide support designed to enhance candidates’ knowledge from the SRM Exam learning objectives and readings”

and

“guidance regarding knowledge and approaches that will be expected in the exam.”

There are a total of 5 modules, plus an additional module that provides an introduction to R. (Well, the modules are not easy to read, and with this study manual, it is not really necessary to go over the modules. ☺)

What is Exam PA Like and How to Study for It?




Typically one of the last exams students take before attaining their ASA designation, Exam PA is the first of its kind in the history of actuarial exams that heavily integrates predictive modeling, R programming, and written communication in a fully proctored setting, and this new exam style calls for a completely different assessment and learning approach.

Exam format. In Exam PA, you will be asked to perform a *data-driven* analysis of a business problem,ⁱⁱ using general tools for constructing and evaluating predictive models (e.g., training/test set split, cross-validation), and specific types of models and techniques (e.g., generalized linear models, decision trees, principal components analysis, and clustering). Such an analysis does not lend itself to the multiple-choice format of many other preliminary exams you have taken, which can only elicit a simple response. Instead, Exam PA is a computer-based **written-answer** 🗨️ exam consisting of a series of well-defined and independent tasks (usually 7 to 9 tasks), most of which are further broken down into one or more subtasks that require reasonably short answers (you need not write long reports!). The whole exam carries a total of **70 points**, with the points for

ⁱOver the four days in each sitting, there are usually two different sets of exam papers (some tasks are common to both sets of papers), but the four days may be shuffled randomly, e.g., day 1 may be paired with day 2, day 3, or day 4.

ⁱⁱThe business problem is not necessarily actuarial in focus. Even if it is actuarially related, there is no expectation that candidates have specific product or practice-area knowledge.

each task and subtask shown at the beginning of the (sub)task *in italics*. As the exam lasts for 3.5 hours, or 210 minutes, on average you should spend 3 minutes per exam point. A 10-point task, for example, should translate to approximately 30 minutes of work. If you have worked on that task for 50 minutes, then you know that it is time to move on.

Wondering where to put your written answers? The project statement, available in Microsoft Word  format, includes designated spaces labelled “ANSWER:” for you to type  your written responses to different exam subtasks. At the end of the exam, you will upload  the entire Word file for grading. You will be assessed on both the technical accuracy of your answers as well as the clarity of your thought process. Unlike other ASA exams, questions in Exam PA tend to be more open-ended and often there is not a unique best answer, as is true of predictive modeling in practice. To score high, you are expected to justify the decisions you make carefully and adequately, based on the business problem and your prior knowledge of predictive analytics.

Typical exam questions. To give you a first taste of the exam, here are some representative tasks taken from the two latest released exams.

- **Type 1: Conceptual subtasks**

Each exam has quite a number of subtasks that require you to *describe* or *explain* the predictive analytic concepts covered in the syllabus. Here are some examples:

▷ April 2023 exam, Task 5 (a):

(3 points) Compare and contrast single decision tree and tree-based ensemble models.

▷ April 2023 exam, Task 8 (c):


(2 points) Describe the process of searching for the optimal value of the hyperparameter lambda in a lasso regression.


▷ October 11, 2022 exam, Task 3 (c):

(2 points) Contrast best subset and stepwise selection for selecting predictors.

▷ October 11, 2022 exam, Task 8 (a):

(3 points) Describe the cost-complexity pruning algorithm and what purpose it serves.

These descriptive subtasks are good ways for the SOA to test your conceptual understanding of predictive analytics. You can secure these easy exam points simply by studying this manual (in particular, the conceptual foundations sections) carefully and practicing explaining different concepts. These subtasks also mean that there are definitions and descriptions you have to memorize  in advance as part of your exam preparation.

- **Type 2: Examining graphs  and output**

In the majority of the exam tasks, you will examine some externally generated graphs and output, and provide explanations, critiques, or recommendations (e.g., which model is the best, in what sense?).

▷ April 2023 exam, Task 1 (b):

(2 points) Explain, using the graph above, why the **Daytype** variable is statistically significant while the **DayofWeek** variable is not.

▷ April 2023 exam, Task 1 (c):

(2 points) Recommend two modeling enhancements that your assistant could explore based on Graph 3 above.

▷ April 2023 exam, Task 5 (c):

(2 points) Determine if this tree shows an interaction between month and year. If there is an interaction, describe it. If not, explain why there is no interaction.

▷ October 11, 2022 exam, Task 7 (b):

(2 points) Recommend the number of clusters to use and justify your recommendation.

▷ October 12, 2022 exam, Task 8 (c):

(2 points) Recommend a tree to present to the client. Justify your recommendation based on the applicability to the business problem.

These interpretational tasks are more demanding (and interesting!) than the descriptive tasks above, because you will have to formulate your responses based on the given output coupled with your prior knowledge in predictive analytics. It is not enough to memorize.

- **Type 3: Simple calculations **

There are also some subtasks where you are asked to reproduce or calculate certain model quantities by hand.

▷ April 2023 exam, Task 8 (f):

(2 points) You are provided with the confusion matrix produced by the lasso model with a positive response cutoff threshold of 0.5.

Calculate sensitivity and specificity. Show all work.

▷ October 11, 2022 exam, Task 1 (d):

(2 points) Calculate the residual for the predicted time to resolution using the values in the following table for a single observation. Show both the formula(s) used (with values substituted for variables) and the final value to two decimal places.

As you can see, the shift of exam focus from working out multiple-choice problems efficiently to crafting computer-aided written responses makes Exam PA a completely different (and hopefully more enjoyable and practical!) learning experience compared with all other ASA exams you have taken. To study for this exam effectively:

It is very important to spend time *understanding* the subject, at least at a conceptual level, and learning how to *communicate* your thoughts precisely and concisely. (Having taught in a [CAE university](#) for about 10 years and graded hundreds of mock exams submitted by past PA students, I can say written communication is an area in which actuarial science students leave much to be desired. 🙄) Unlike other ASA exams, you can't expect to do well just by drilling mechanical practice problems mindlessly. Instead, make an effort to understand, describe, and explain things, which is far from trivial!

New Exam Format Effective from April 2023

Ever since Exam PA was introduced in December 2018, its format and style have undergone significant changes. The latest revamp took place in the April 2023 sitting, effective from which the exam time has reduced remarkably from 5.25 hours to 3.5 hours. Perhaps the more striking change is:

Starting with the April 2023 administration, **R and RStudio (a convenient platform to implement R) will not be available on the exam.**

How will this “big” change affect the exam and our preparation? My answer, which is confirmed by the April 2023 and October 2023 exams, is:

You will learn the material and prepare for the exam in essentially the same way, perhaps paying less attention to code syntax.

Even when R and RStudio were available on the exam from December 2018 to October 2022, Exam PA was never designed as a coding exam. Candidates did have to know *some* R, but only to the extent that they understood what the code (contained in a separate R markdown file generously provided by the SOA) was doing and knew how to make minor adjustments if necessary. The focus of the exam has always been on *understanding* and *interpretation*, reflected by the abundance of past exam questions belonging to Types 1 and 2 above. With R and RStudio no longer available, perhaps the only major difference is that the code and output relevant to the exam tasks will be provided directly in the project statement; you need not take the trouble to run the code in RStudio or see the R output. The emphasis on conceptual understanding and interpretation is likely to remain (or will even be greater).

There is one change I do expect to see in the new exam format:

There may be more tasks of Type 3, where you have to do some simple analysis and manual calculations based on the R code and R output given.

With all the useful code and output given in the project statement, you may be asked to explain what a certain number in the output means or how it is calculated, or to use the output to do

some simple arithmetic calculations. This is a good way for the SOA to test your understanding. You can't just rely on R to do everything!

Historical Pass Rates %

The following table shows the number of sitting candidates, number of passing candidates, and pass rates for Exam PA since it was offered in December 2018:

(For written-answer exams, including PA, the SOA does not release pass marks, i.e., the actual score you have to get to pass the exam. Yes, the grading is very much a black-box process. 🙄)

Sitting	# Candidates	# Passing Candidates	Pass Rate
October 2023	(To be posted on the SOA webpage 📄 on December 8)		
April 2023	2193	1552	70.8% (highest ever!)
October 2022	1554	1005	64.7%
April 2022	1171	773	66.0%
December 2021	1922	1321	68.7%
June 2021	1691	1055	62.4%
December 2020	1954	1228	62.8%
June 2020	1389	812	58.5%
December 2019	2048	1098	53.6% (I took this exam! 😊)
June 2019	1282	642	50.1%
December 2018	1042	524	50.3%

The pass rates, which are usually in the 50-65% range,ⁱⁱⁱ are higher than those of other ASA-level exams, which are typically 40-50%. Meanwhile, about 40% of the candidates failed every time, even among those who have reached this far in the ASA journey. I have heard of candidates who have failed PA twice or thrice (☹), so the exam is neither a beast nor a breeze! Worst of all, the exam is offered only twice a year, so in the unfortunate event that you fail, you will need to wait another six months, which adds a lot to your travel time to ASA.

ⁱⁱⁱThey became noticeably higher after COVID-19 broke out possibly due to the [temporary refund policy](#), which allowed students to withdraw 14 days before the exam started. This policy is likely to end soon.

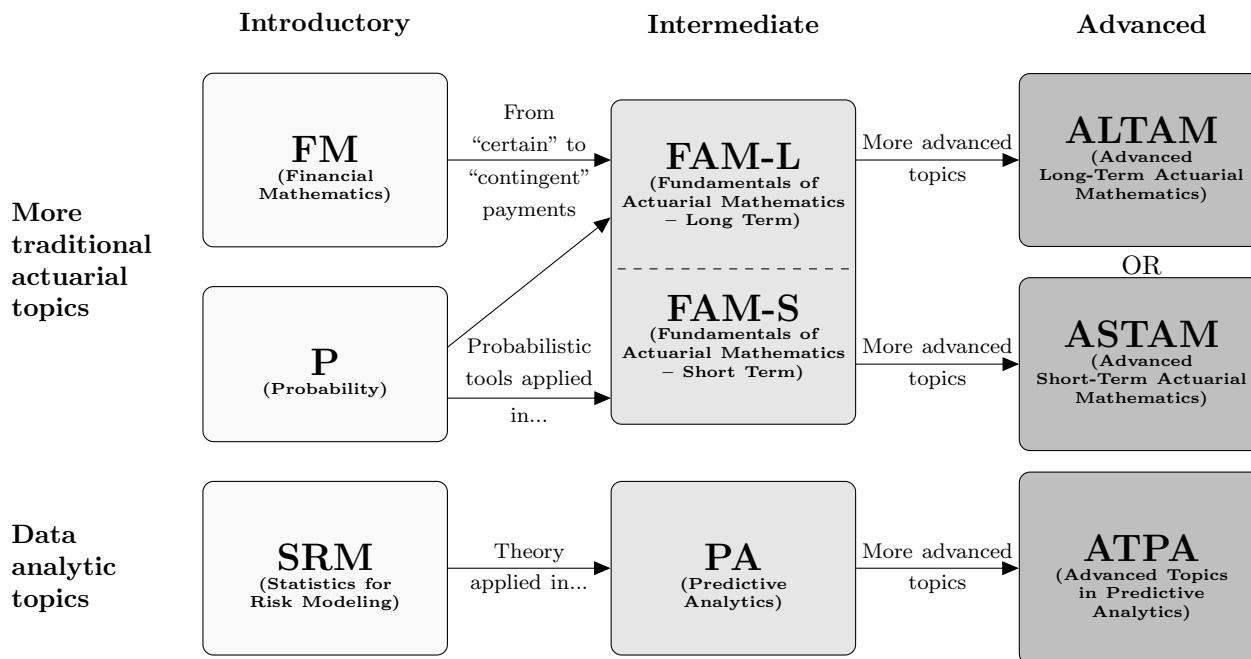
Predictive Analytics Trio 🌀: SRM, PA, and ATPA

Since 2018, the SOA has redesigned the ASA curriculum to reflect more contemporary and powerful predictive analytic methods that have proved useful in actuarial practice. In the current curriculum, there are a total of 3 exams (or assessments) with a heavy focus on predictive analytics:

- SRM (Statistics for Risk Modeling)
- PA
- ATPA (Advanced Topics in Predictive Analytics)

The flowchart below shows how these 3 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



SRM vs. PA. From December 2018 to June 2021, Exam SRM was a formal prerequisite for Exam PA. Although this prerequisite is no longer in place, knowledge of the SRM materials is still assumed. As the PA exam syllabus says,

“Exam PA assumes knowledge of probability, mathematical statistics, and selected analytical techniques as covered in Exam P (Probability), VEE Mathematical Statistics, and Exam SRM (Statistics for Risk Modeling),”

so it is reasonable to prepare for PA at the same time as or shortly after taking SRM, e.g., taking SRM in early September and PA in mid-October, or SRM in early January and PA in mid-April.

In essence, Exams SRM and PA share the same theme of working with *models*, but test it differently. As a precursor, Exam SRM is a traditional multiple-choice exam that serves to provide you with the foundational knowledge behind the modeling process. The emphasis is on the underlying theory, including the uses, motivations, mechanics, pros and cons, do's and don'ts of, and similarities and differences between different predictive analytic techniques. As a natural continuation, Exam PA will have you apply the theory you learned in Exam SRM to a business problem and see first hand how things play out. Although SRM is an important stepping stone to PA and the two exams have a rather big overlap, I would still recommend spending about **2 months** 📅 studying for PA intensively, even if you have taken SRM. Here are the reasons:

- (*Different skills tested*) Even though you will not apply mathematical formulas or do calculations by hand as often as in SRM, you will need time to gain hands-on experience with fitting and interpreting predictive models in R, and need practice on communicating your thoughts in writing. The written-answer format of Exam PA means that the SOA can test the material of SRM in greater breadth and depth, and assess your higher-level thinking, e.g., can you describe a certain concept or explain why something is true? You have to know how things work, at least at a conceptual level, and organize your thoughts in words.
- (*Scope*) There are some additional concepts (e.g., exploratory data analysis, elastic nets, performance metrics for classifiers, elbow method for K -means clustering) and practical considerations that are tested in PA, but not seen in SRM. You do have to hit the books (or this manual)! 📖

PA vs. ATPA. Introduced in January 2022, the ATPA Assessment is a 96-hour take-home computer-based assessment (rather than a proctored exam) that 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. Unlike 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 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. In my experience, you may need more than a day just to clean the data and get it in good shape in R before building any models. You will be busy!

2 About this Study Manual

What is Special about This Study Manual?

I fully understand that you have an acutely limited amount of study time and that Exam PA, as a written-answer exam with a new format effective from April 2023, is not easy to prepare for. With this in mind, the overriding objective of this study manual is to help you develop a conceptual understanding of and hands-on experience with the materials of Exam PA as effectively and efficiently as possible, so that you will pass the exam on your first try easily, go on to ATPA confidently, and get your ASA ASAP. Here are some unique features of this manual to make this possible.

Feature 1: The Coach DID Play!

Usually coaches don't play 😊, but as a study manual author, I took the initiative to write the **December 2019 Exam PA** and the **February-April 2023 ATPA Assessment** to experience first-hand what the real exams were like, despite having been an FSA since 2013 (and technically free from exams thereafter!). I made this decision in the belief that braving the exam myself is the best way to ensure that this manual is indeed useful for exam preparation. (If the manual is useful, then at the minimum the author himself can do well, right?) I am thrilled that with the help of my own manual, I received Grade 10 in Exam PA and passed ATPA, both on the first try.



Shopping Cart Section My Account

You are here: [My Account](#) » [My Transcripts](#) » [Grade Slip](#)

Grade Slip

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: 1/24/2022

Dec 2019 Predictive Analytics

Course	Grade
EXAMPA	10

ID: [REDACTED] Candidate ID: 67666

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

[Printer Friendly Version](#)



Online Services

Shopping Cart Section My Account

You are here: [My Account](#) » [My Transcripts](#) » [Grade Slip](#)

Grade Slip

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: 7/1/2023

Jun 2023 Advanced Topics in Predictive Analytics

ID:	
-----	--

Course
ATPA

Grade
11

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

If you use this PA study manual, you can rest assured that it is written from an exam taker's perspective by a professional instructor who has experienced the "pain" of (AT)PA candidates and truly understands their needs. Drawing upon his "real battle experience" and firm grasp of the exam topics, the author will go to great lengths to help you prepare for this challenging exam in the best possible way. You are in capable hands.

Feature 2: Three-part Structure

To maximize your learning effectiveness and efficiency, I have divided this study manual into three parts:

- **Part I: A Crash Course in R**

The first part of the manual is a crash course in R covering the elements of R programming that are particularly germane to Exam PA to get you up to speed. They include the basics of R programming and data visualization using the `ggplot2` package, covered respectively in Chapters 1 and 2 of the manual. At the completion of this part, you will be equipped with the fundamental R programming skills necessary for constructing predictive models and making some simple but informative graphs in the rest of this manual.

- **Part II: Theory of and Case Studies in Predictive Analytics**




Armed with R basics, you will learn the theory of different types of predictive analytic techniques illustrated by a series of case studies in the second part (Chapters 3 to 6), also

the linchpin, of this manual. Each chapter in this part follows the same arrangement:

- ▷ *Theory*: Each chapter begins with a conceptual foundations section describing the mechanics of various predictive analytic techniques, including linear models (Chapter 3), generalized linear models (Chapter 4), decision trees (Chapter 5), and principal components and cluster analyses (Chapter 6).
- ▷ *Practice*: After learning the ins and outs, pros and cons, and do's and don'ts of these techniques, we will turn to their practical implementations and gain some hands-on experience through a number of task-based case studies using R. Do read these case studies carefully as they illustrate a wide range of skills necessary for tackling various types of tasks in Exam PA, ranging from data pre-processing, data exploration, model construction, model evaluation, and model selection.


• Part III: Final Preparation

Last but not least, the third part concludes this manual with the following resources:

- ▷ *Chapter 7*: My commentary on the SOA's past PA exams, which reflect the SOA's expectations of PA candidates and are quite indicative of future exams
- ▷ *Chapter 8*: Two original full-length practice exams updated for the new exam format and designed to mimic the real PA exam in terms of style and difficulty, with detailed illustrative solutions provided
- ▷ A downloadable  and printable  cheat sheet (available on [Actuarial University](https://www.actuarialuniversity.com) as a separate file) that provides a “helicopter”  view of the entire PA exam, and is useful for both regular review and last-minute exam preparation

To access the cheat sheet, please log in to [Actuarial University](https://www.actuarialuniversity.com):

www.actuarialuniversity.com.

Then click the Actuarial University logo  at the top of the page, and scroll down to the “Helpful Links” section at the bottom. There you will find a green button that says “Printable SOA and CAS Formula and Review Sheets,” which will bring you to a set of printable review sheets, the PA cheat sheet being one of them.

After completing Part III, you will be ready to take (and pass!) the April 2024 PA exam.

Other Features

This manual throughout is also characterized by the following features that make your learning as smooth as possible:

- Each chapter in Parts I and II starts by explicitly stating which learning objectives and outcomes of the PA exam syllabus we are going to cover, to assure you that we are on track and hitting the right target.
- Objects in R are shown in `typewriter` font and code chunks with output in gray boxes for aesthetic reasons.

```
...LOTS OF R CODE HERE...
...LOTS OF R CODE HERE...
...LOTS OF R CODE HERE...
```

Formulas, functions, and commands that are of great importance are boxed to aid identification and retention.

- Important exam items and common mistakes committed by students are highlighted by boxes that look like:

⚠ EXAM NOTE ⚠

Be sure to pay special attention to boxes like this!

- The main text of this manual is interspersed with more than 110 exercises, all with complete solutions, to assess your understanding regularly. Some of these exercises are based on recent SOA and CAS exams, but many are original. (If you have used the *ACTEX Study Manual for Exam SRM*, you may have seen some of these past exam questions in some form, but I have rewritten many of them in the language and style of Exam PA. There is also no harm in giving them a second look!) These examples are instrumental in illustrating a number of conceptual items that can be tested in Exam PA.
- Each chapter in Part II concludes with a number of conceptual review questions designed to help you look back on the most important conceptual issues in that chapter. Solutions to these questions can be found in the main text, indicated by marginal labels such as the one on the left.

Q3.1


Supplementary Files



This study manual comes with a number of supplementary files (e.g., R Markdown files with completely reproducible R code, datasets, cheat sheet, and files to be released) that can be downloaded from [Actuarial University](#). All users of the manual (either the printed or digital version) will receive by email a keycode that provides electronic access to all supplementary files shortly after their order is placed. If you can't retrieve that email (be sure to check your junk/spam folders), please reach out to support@actexlearning.com for assistance.

It is a good idea (but not absolutely essential, given the new exam format) to run the R Markdown files as you work through this manual, making sure that your output agrees with what is shown here. This is especially important if you have ordered a printed copy of this study manual—run the code to see the beautiful colors! ☺

⚠ NOTE ⚠

Commentary on the not-yet-released October 2023 PA exam and Practice Exam 2 will be available on [Actuarial University](#) shortly after the SOA posts the exam with solutions [online](#). 

Two Add-ons

If you have purchased this manual and are interested in upgrading your manual to include any of the following add-ons, please email Customer Service at support@actexlearning.com.

Instructional videos. 🎥 Instructional videos (<https://www.actexlearning.com/exams/pa/exam-pa-study-manual>) accompanying the core of this manual (Parts I and II, or Chapters 1 to 6) are available for purchase as an add-on. In these videos, I (Ambrose) will walk you through the fundamental concepts in predictive analytics and the construction of predictive models in R step by step, with a strong emphasis on key test items in Exam PA. With the aid of visuals, these videos aim to make the materials in the manual as accessible as possible and will add substantial value to your learning.

When it comes to learning strategies, some students find it useful to watch the videos to get the “big picture,” then read the manual to learn the details. Alternatively, you may first read the manual, then watch the videos to consolidate your understanding. Both modes of learning are fine and which one is better depends entirely on your preferences.

Graded mock exam. 📝 In addition to the two practice exams in Chapter 8 of this manual, we offer a separate mock exam (<https://www.actexlearning.com/exams/pa/exam-pa-mock-exam>), with completely different questions, and an optional 1:1 live feedback session.

A common “complaint” against Exam PA is that its written and somewhat open-ended exam format makes it difficult for students to evaluate their work even after reading the SOA’s model solutions to past exams, e.g., if you write this, how many points can you expect to get? How to improve your answers? This is precisely why we create this mock exam with grading service, which provides a valuable opportunity for you to assess your overall understanding of the PA exam syllabus and, more importantly, have your work graded from a critical eye 👁, and receive **personalized feedback** 🗨 (not generated by AI in any way!). You will work on the mock exam under simulated exam conditions and submit your solutions to us. Having taken PA in the past and now teaching for PA, we (Ambrose and his team) will then grade your work from start to finish, with a score out of 70, and offer specific (and critical!) feedback that will help you enhance the quality of your write-up and improve your performance on the real exam.

For the April 2024 sitting, the graded mock exam (currently available for pre-order) is expected to be released on *Actuarial University* in February and the last day of submission is April 2 (Tuesday). Within 2 weeks of your submission, you will receive by email:

1. Your graded mock exam with personalized feedback
2. An access key, which, once activated, will grant you access to the detailed illustrative solutions to the mock exam on Actuarial University

Announcements

As time goes by, I will post news and announcements (e.g., new files becoming available) about this study manual and Exam PA on my personal web page:

<https://sites.google.com/site/ambroseloyp/publications/PA>

An errata list is also maintained. I would greatly appreciate it if you could bring any potential errors, typographical or otherwise, to my attention via email (see below) so that they can be fixed in a future edition of the manual.

Contact Us

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

- ✉ • For **technical issues** (e.g., not able to access, download, or print supplementary files from *Actuarial University*, extending your digital license, upgrading your product, exercising the Pass Guarantee), please email Customer Service at support@actexlearning.com. The list of FAQs available on <https://www.actuarialuniversity.com/help/faq> may also be useful.
- ✉ • Questions related to **specific contents** of this manual, including potential errors (typographical or otherwise), can be directed to me (Ambrose) by emailing amblo201011@gmail.com. Please note:

▷ Remember to check out the errata list on my personal web page. It may happen that the errors you discover have already been addressed.

▷ Instead of saying

“You mention (somewhere) in your manual that...,”

it would be great to quote the specific page(s) of the manual your questions are about. This will provide a concrete context and make our discussion much more fruitful.

▷ (*Less important in the new exam format*) If you experience issues with R, e.g., your code can't run and you keep seeing weird error messages, please provide the version of R (not RStudio!) you are using and a screenshot of the error messages.

NOTE

- For a faster turnaround, it would be greatly appreciated if you could reach out to the appropriate email address. 😊
- I will strive to get back to you ASAP. ↩ Please check your spam folder if you don't hear back from me within 2-3 days.

Acknowledgments

I am grateful to Mr. Tony Pistilli for proofreading an early version of this study manual and many past students for taking the time to send me comments and suggestions, which have improved the quality of the manual in no small measure. All errors that remain are solely mine.

About the Author

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 is the (co)author of 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.

Chapter 2

Data Exploration and Visualization

*****FROM THE PA EXAM SYLLABUS*****

2. Topic: Data Exploration and Visualization (20-30%)

Learning Objectives

The Candidate will be able to work with various data types, understand principles of data design, and construct a variety of common visualizations for exploring data.

Learning Outcomes

The Candidate will be able to:

- d) Apply the key principles of constructing graphs.
- e) Apply univariate data exploration techniques.
- f) Apply bivariate data exploration techniques.

Chapter overview: An integral part of any predictive analytic exercise is the use of graphical displays to investigate the characteristics of the variables of interest, on their own and in relation to one another, and to visualize the results of the predictive models constructed. In this regard, one of the key strengths of R as a programming language is that it offers versatile graphing capabilities, both in the base installation and with add-on packages. With a minimal amount of code, we can produce a wide variety of high-quality graphs. In Exam PA, you will be asked to take advantage of R’s graphing capabilities and make sense of different types of graphical displays [\[1\]](#). Instead of using R’s base graphical platform, you will make graphs using the `ggplot2` package,ⁱ which may be new to you even if you have used R before. Compared to R’s base graphics system, `ggplot2` involves vastly different syntax based on the so-called “grammar of graphics” (in fact, “gg” stands for “grammar of graphics”) and lends itself to producing sophisticated graphs that

ⁱThe `ggplot2` package is developed by Hadley Wickham, who is also a core developer of RStudio. Earlier the package was called `ggplot`, but substantial changes were made later, so the name of the package was upgraded to `ggplot2`.

would be cumbersome to create using base R graphics.

Synthesizing the material in the first four chapters of the book *Data Visualization: A Practical Introduction* (which is listed in the exam syllabus), this chapter presents some of the important graphical functions in the `ggplot2` package most relevant to Exam PA. These functions can be used to construct different types of graphs such as scatterplots, histograms, boxplots, and bar charts, which will all be illustrated in the context of a real insurance dataset. In Section 2.1, we will learn the basic structure of a ggplot, make some simple but informative plots, and learn how to tweak the appearance of a ggplot. Section 2.2 draws upon the data visualization techniques covered in Section 2.1 to perform exploratory data analysis, which is the use of graphs and summary statistics to uncover patterns and relationships in a set of data, and generate hypotheses which can be answered quantitatively in a predictive model at a later stage.

2.1 Making ggplots

2.1.1 Basic Features

Let's begin by installing (make sure to install a package the first time you use it!) and loading the `ggplot2` package.

```
# CHUNK 1
# Uncomment the next line the first time you use ggplot2
#install.packages("ggplot2")
library(ggplot2)
```

With the last command, we can use all the functions in the `ggplot2` package until the end of the current R session.


Skeleton. In its simplest form, a ggplot consists of two parts: The core `ggplot()` function (not `ggplot2()`!) and a chain of additional functions pasted together using the plus (+) sign defining the exact type of plot to be made.

1. *ggplot() function:* The `ggplot()` function initializes the plot, defines the source of data using the `data` argument (almost always a **data frame** in Exam PA), and, most importantly, specifies what variables in the data are “mapped” to visual elements in the plot by the `mapping` argument. Mappings in a ggplot are specified using the `aes()` function, with `aes` standing for “aesthetics.” They determine the role different variables play in the plot. The variables may, for instance, correspond to visual elements such as the x- or y-variables, color, size, and shape, specified by the `x`, `y`, `color`, `size`, and `shape` aesthetics, respectively.
2. *Geom functions:* Subsequent to the `ggplot()` function, we put in *geometric objects*, or *geoms* for short, which include points, lines, bars, histograms, boxplots, and many other possibilities, by means of one or more *geom functions*. Placed layer by layer, these geoms determine what kind of plot is to be drawn and modify its visual characteristics, taking the data and aesthetic mappings specified in the `ggplot()` function as inputs.

Here is the generic structure of a ggplot: (The uppercase letters are placeholders.)

```
ggplot(data = DATA, mapping = aes(AESTHETIC_1 = VARIABLE_1,
                                   AESTHETIC_2 = VARIABLE_2,
                                   ...)) +
  geom_TYPE(...) +
  geom_TYPE(...) +
  OTHER_FUNCTIONS +
  ...
```

Don't worry if the ideas above seem puzzling at this stage. It is commonly acknowledged that the learning curve of ggplots is steep, much more so than R's base graphics system, but taking some time to learn how to make ggplots will pay dividends not only in Exam PA, but also in your real work. You will gain a much better understanding of how a ggplot works after going through the example plots in this chapter and in the rest of this study manual.

Case study: Personal injury insurance dataset. To illustrate data visualization and exploration techniques, in this chapter we will look at a personal injury insurance dataset.ⁱⁱ  This dataset contains the information of 22,036 settled personal injury insurance claims. These claims were reported during the period from July 1989 to the end of 1999, with claims settled with zero payment excluded. The variables in the dataset are described in Table 2.1.



Variable	Description
amt	settled claim amount (continuous numeric variable)
inj	injury code, with seven levels: 1 (no injury), 2, 3, 4, 5, 6 (fatal ) , 9 (not recorded)
legrep	legal representation (0 = no, 1 = yes)
op_time	operational time (a standardized amount of time elapsed between the time when the injury was reported and the time when the claim was settled)

Table 2.1: Data dictionary for the personal injury (`persinj`) insurance claims dataset.

In Section 4.2, we will build a model to predict the size of personal injury insurance claims using other variables in the dataset. For now, we will perform data exploration of the variables in the dataset. The insights we gain here will go a long way towards constructing a good predictive model.

ⁱⁱThis dataset is a pre-processed version of the `persinj.csv` file that accompanies the textbook *Generalized Linear Models for Insurance Data* (2008), by de Jong and Heller.

- To get started, let's run CHUNK 2 to load the external CSV file containing the dataset into R as a data frame called `persinj` (meaning “personal injury”) using the `read.csv()` function, which takes the name of the CSV file  supplied as a character string as an argument. In this preliminary section, we will take out a subset of 50 observations from the `persinj` data, called `persinj50`, and explain the main characteristics of a ggplot using these 50 observations. This will help us appreciate the different types of visual effects that can be produced on a ggplot more easily. In Section 2.2, we will return to the full dataset and learn why we use a certain plot for a certain purpose.

ONE MORE REMINDER!

Please read page xxii of the preface of this manual about how to access the Rmd files as well as datasets that go with this manual.

```
# CHUNK 2
persinj <- read.csv("persinj.csv")
# Take out a subset of 50 observations from the full data
persinj50 <- persinj[seq(1, nrow(persinj), length = 50), ]
```

- **First encounter with ggplots.** As our first example, let's make a *scatterplot* for the two numeric variables in the `persinj50` data, `amt` and `op_time`. The plot, produced by the code in CHUNK 3, is given in Figure 2.1.1. The code obeys the two-part structure discussed earlier:

- *ggplot() function:* The first line of the code makes it clear that we are using the `persinj50` data, where the variables `op_time` and `amt` are mapped to the variables on the x-axis and y-axis through the `x` and `y` aesthetics, respectively. There is no need to name the variables as `persinj50$op_time` or `persinj50$amt` as the data source is already specified in the `data` argument.
- *Geom:* Given these mappings, we use `geom_point()` to make a scatterplot of `amt` (the y-variable) against `op_time` (the x-variable). The plot comprises 50 points ● (hence the name of the function) corresponding to the 50 paired values of the two variables and allows us to see the two variables in comparison with each other.

Later, we will fine-tune this plot in different ways to capture different sorts of information.

```
# CHUNK 3
ggplot(data = persin50, mapping = aes(x = op_time, y = amt)) +
  geom_point()
```

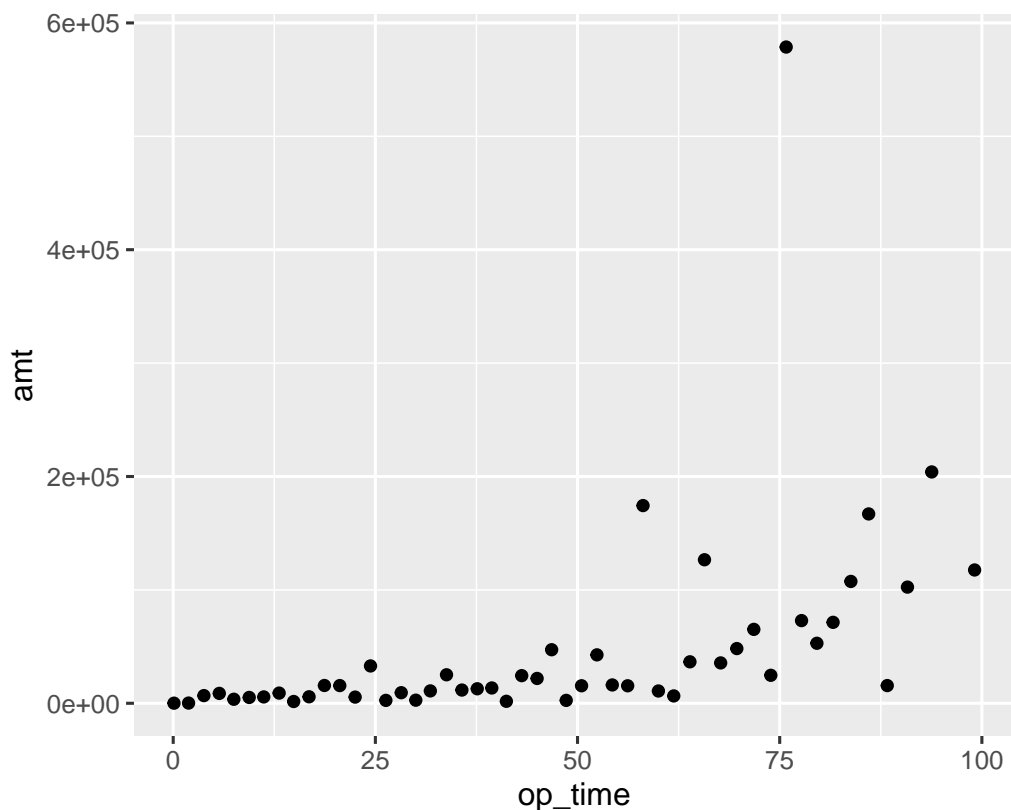


Figure 2.1.1: A basic scatterplot of `amt` against `op_time` in the `persin50` dataset.

Using aesthetics the right way: The essence of aesthetic mappings. One of the most common ways to modify the appearance of a plot is to color the observations in order to produce a more impressive visual effect. To color the data points say, in `blue`, you may be tempted to make use of the `color`ⁱⁱⁱ aesthetic and simply insert `color = "blue"` as an additional argument to the `aes()` function, as in CHUNK 4. Doing so will produce unexpected and undesirable results as shown in Figure 2.1.2. To your astonishment, all of the data points are colored in `red` instead of `blue` and there is a legend saying “blue.” What has gone awry here?

ⁱⁱⁱBoth the American spelling (`color`) and British spelling (`colour`) are accepted.

```
# CHUNK 4
# It is OK to suppress the names of the data and mapping arguments
# so long as they are supplied in order
ggplot(persinj50, aes(x = op_time, y = amt, color = "blue")) +
  geom_point()
```

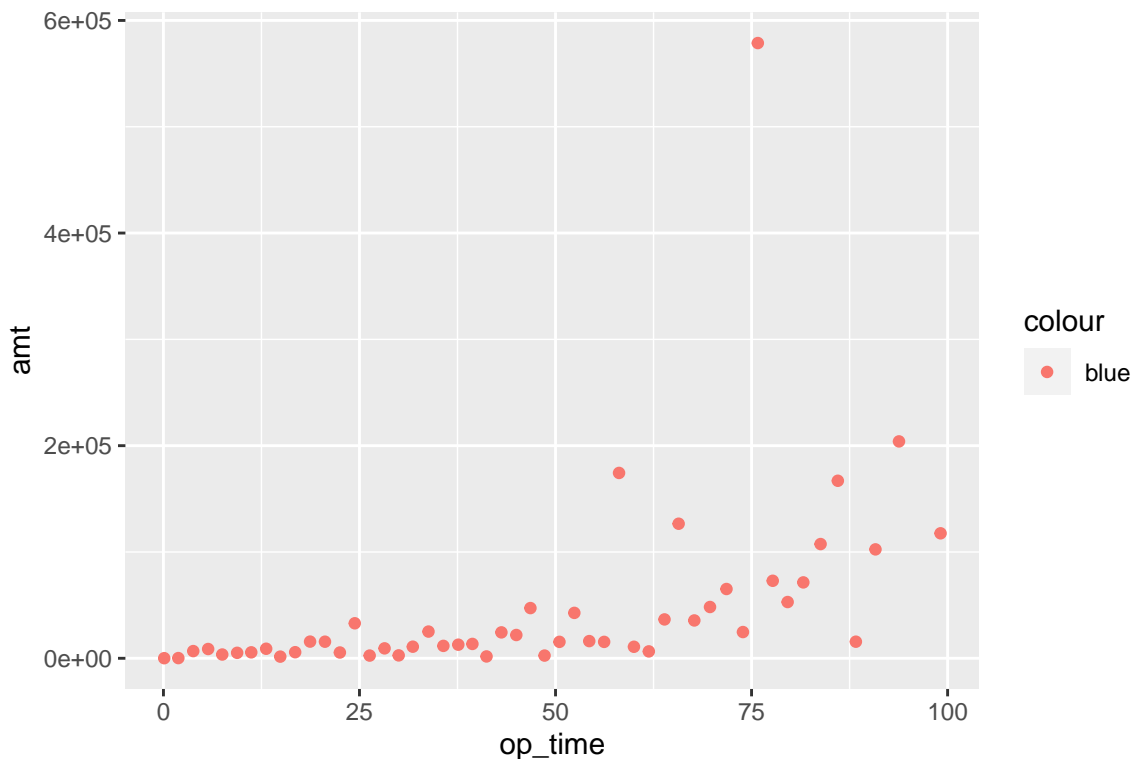


Figure 2.1.2: A version of Figure 2.1.1 with all of the points inadvertently colored in red.

Bear in mind that an aesthetic is a mapping between variables in our data and visual properties of the graph. The use of `color = "blue"` instructs the `aes()` function to map the `color` aesthetic to a variable named "blue" in the `persinj50` dataset. There is no such variable in our data, but the `aes()` function will do its best by treating "blue" as if it were a variable. The effect is the creation of a new character variable behind the scenes taking one and only one value, "blue". As all observations in the data share the same "blue" value, all of them will be mapped to the same color. In `ggplot2`, the default first-category hue is red (not blue!). This explains why every point in the scatterplot becomes red in color.

To do the coloring the right way, we should realize that making all the points blue in color does *not involve any mapping* between variables in our data and the `color` aesthetic. After all, all the observations are colored in blue and are not distinguished on the basis of color. As a result, we should not put `color = "blue"` inside the `aes()` function. It should instead be placed inside the `geom_point()` function to modify the color of the plotted points. Figure 2.1.3 shows the desired scatterplot using the code in CHUNK 5. To our liking, all of the points are colored in blue.


```
# CHUNK 5
ggplot(persinj50, aes(x = op_time, y = amt)) +
  geom_point(color = "blue")
```

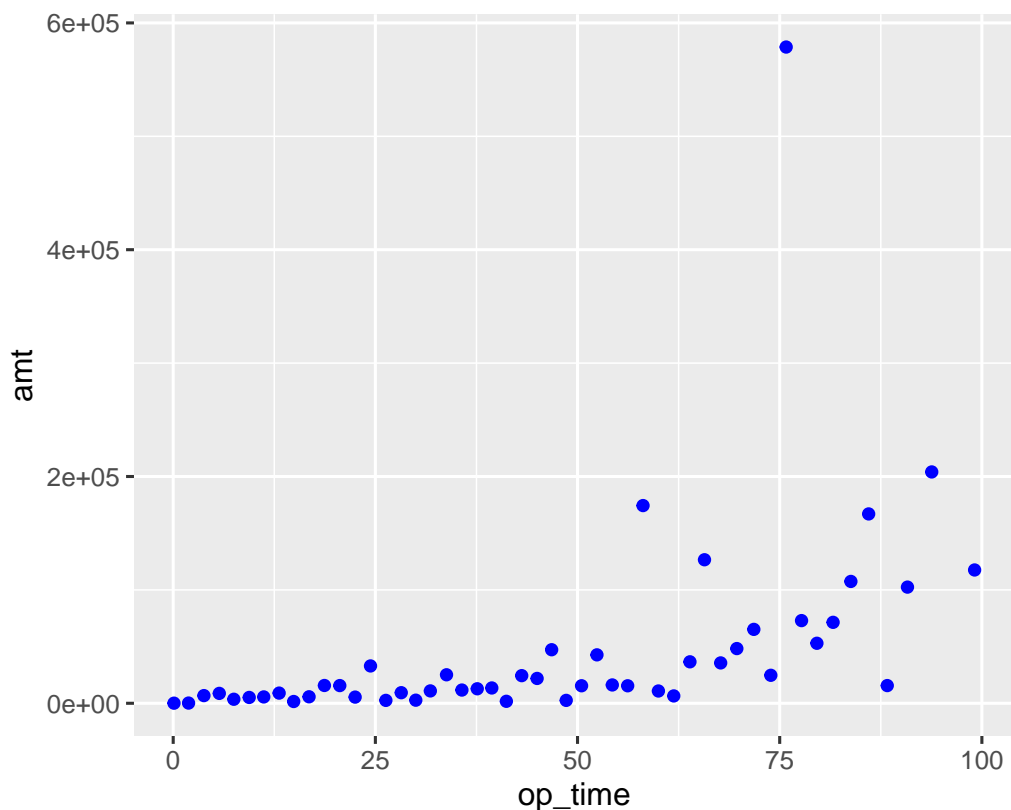


Figure 2.1.3: A version of Figure 2.1.1 with all of the points correctly colored in blue.

Figures 2.1.2 and 2.1.3 highlight a subtle but extremely important mentality when making ggplots:

The `aes()` function is reserved for mappings between aesthetics and *variables*.

To set a property that affects how a plot looks but does not involve mapping variables to aesthetic elements, we should do it outside the `aes()` function—in the geom functions. To put it another way, the aesthetics determine *what* relationships we want to see in the plot whereas the geoms determine *how* we want to see the relationships.

Now let's see an example of using the `color` aesthetic correctly. In the `persinj50` data, the `legrep` variable is a binary variable equal to 1 for injuries with legal representation and 0 for those without. To color the different injuries according to the presence of legal representation, we map the `color` aesthetic to `legrep` treated as a factor (recall that factors are discussed on page 21). The resulting scatterplot, generated by the code in CHUNK 6, is given in Figure 2.1.4, where injuries without legal representation (`legrep = 0`) are displayed in red whereas those with legal representation (`legrep = 1`) are displayed in teal. A legend is produced accordingly. Notice that there is a genuine mapping between the `legrep` variable and `color`, with `legrep`

= 0 mapped to the **red** color and **legrep** = 1 mapped to the **teal** color. In other words, the observations are differentiated on the basis of the **legrep** variable by color. (The **color** aesthetic does not say what colors are used to discriminate injuries on the basis of **legrep**, though.)

CHUNK 6

```
ggplot(persinj50, aes(x = op_time, y = amt, color = factor(legrep))) +
  geom_point()
```

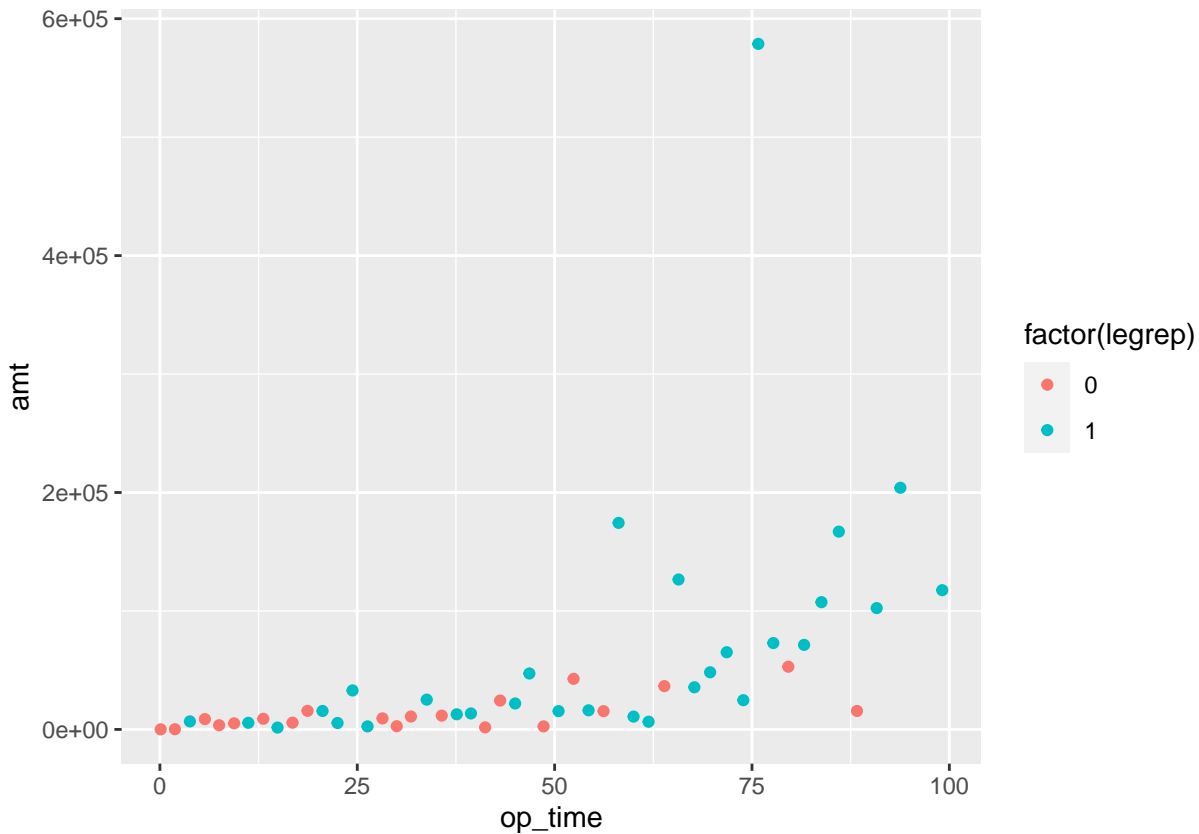


Figure 2.1.4: A version of Figure 2.1.1 with the observations distinguished by **legrep**.

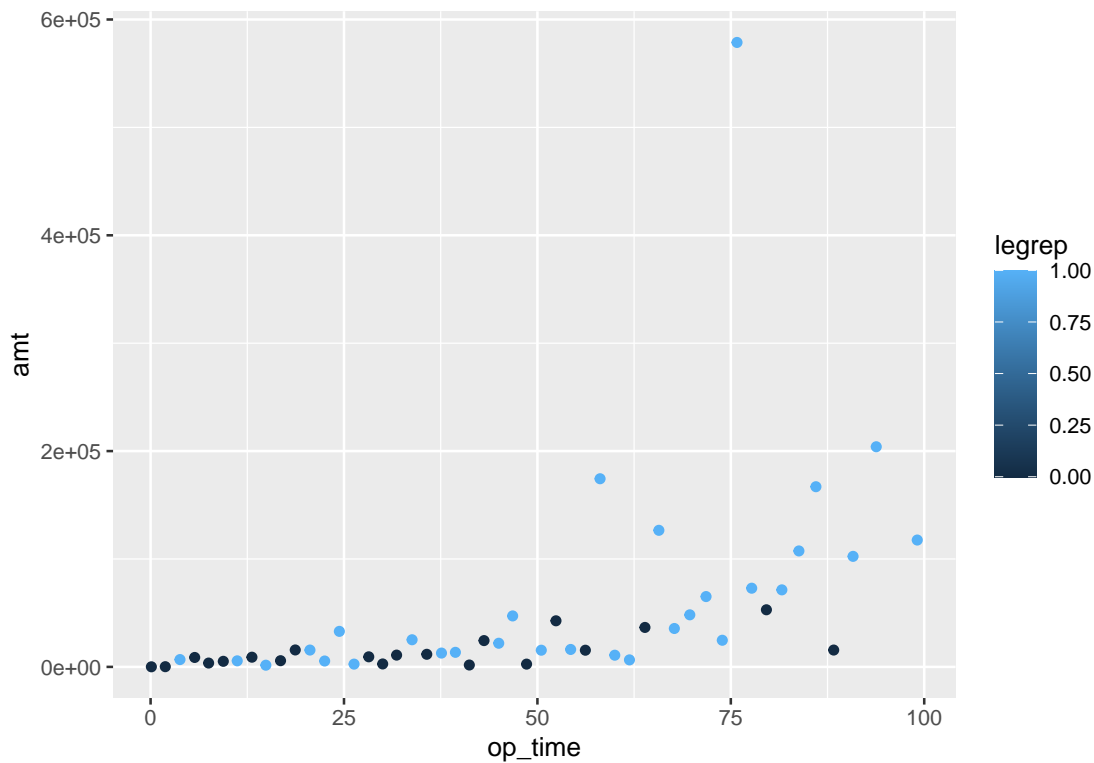
Exercise 2.1.1. 🧠 (Why do we need to convert **legrep** to a factor?) To see why the conversion of **legrep** to a factor variable is needed, run the following code in CHUNK 7:

CHUNK 7

```
ggplot(persinj50, aes(x = op_time, y = amt, color = legrep)) +
  geom_point()
```

What do you notice? Why is the coloring done in the way you observe?

Solution. Running the code produces the following scatterplot with a color gradient from 0 to 1:



This is because the `legrep` variable in the `persinj50` dataset is treated by design as a continuous numeric variable even though the two levels, 0 and 1, are merely class labels that do not convey any sense of numeric order. R implicitly allows for values between 0 and 1 for the `legrep` variable and therefore uses the color gradient to differentiate the observations by color (though you can observe that there are only two colors in the plot, corresponding to the two extremes in the color gradient). This explains the point made in Subsection 1.2.1 that whether to treat a categorical variable as a factor can affect how the resulting graphical output looks, sometimes materially. \square

Some important geoms for Exam PA. Besides `geom_point()`, there are a number of geoms that are important for Exam PA. They are listed below along with their commonly used arguments which affect how the plot looks. (No need to memorize the table entries!)

Geom	Type of Object Produced	Frequently Used Arguments
<code>geom_bar()</code>	Bar chart	<code>fill</code> , <code>alpha</code>
<code>geom_boxplot()</code>	Boxplot	<code>fill</code> , <code>alpha</code>
<code>geom_histogram()</code>	Histogram	<code>fill</code> , <code>alpha</code> , <code>bins</code>
<code>geom_point()</code>	Scatterplot	<code>color</code> , <code>alpha</code> , <code>shape</code> , <code>size</code>
<code>geom_smooth()</code>	Smoothed curve	<code>color</code> , <code>fill</code> , <code>method</code> , <code>se</code>

The names of these geoms are pretty self-explanatory. For example, `geom_smooth()`, as its name suggests, produces a “smoothed” curve and, by default, produces a ribbon around the curve showing the standard error bands. It is typically used in conjunction with `geom_point()`. The smoothed curve can be generated by different statistical methods, such as a linear fit (by setting `method = "lm"`). The default is the use of nonparametric smoothing methods (`method = "gam"` or `method = "loess"`), which are beyond the syllabus of Exam PA. To switch off the standard error bands, you can set `se = FALSE`.

We will illustrate the use of `geom_bar()`, `geom_boxplot()`, and `geom_histogram()` in Section 2.2. For now, let’s continue with the scatterplots we have just produced and make them more fancy and informative. In Figure 2.1.5, we plot the 50 observations in the `persinj50` dataset using large points (`size = 2`) and a small amount of transparency (`alpha = 0.5`), classify them according to whether they have legal representation or not, and fit a separate smoothed curve to each kind of injuries via the `geom_smooth()` function. The commands are collected in CHUNK 8.

Note that:

- For each of the two types of injuries, the smoothed curve and the standard error ribbon are indicated by the same color (red for those without legal representation and teal for those with legal representation), which is appealing from an aesthetic perspective. The consistent coloring is achieved by mapping both the `color` aesthetic and `fill` aesthetic (which controls filled areas of bars, polygons and, in this case, the interior of standard error bands) to the `legrep` variable treated as a factor variable.
- If you omit `fill = factor(legrep)` (try this in R!), then the two smoothed curves will still be colored according to the presence of legal representation due to the `color` aesthetic, but without the `fill` aesthetic, the `geom_smooth()` function will shade the two standard error ribbons by its default color, which is gray.
- The `alpha` argument controls the transparency of the plotted objects on a scale from 0 (fully transparent) to 1 (opaque); the default value is 1. The transparency of the objects increases by decreasing `alpha`; the lower the value of `alpha`, the more transparent the points. In the limit when `alpha` is exactly zero, the objects become completely invisible. The `alpha` argument is particularly useful when there is a lot of overlapping among the data points. By setting `alpha` to an intermediate value such as 0.5, we make it easy to see where most of the observations cluster.

```
# CHUNK 8
ggplot(persinj50, aes(x = op_time, y = amt,
                     color = factor(legrep), fill = factor(legrep))) +
  geom_point(size = 2, alpha = 0.5) +
  geom_smooth()
```

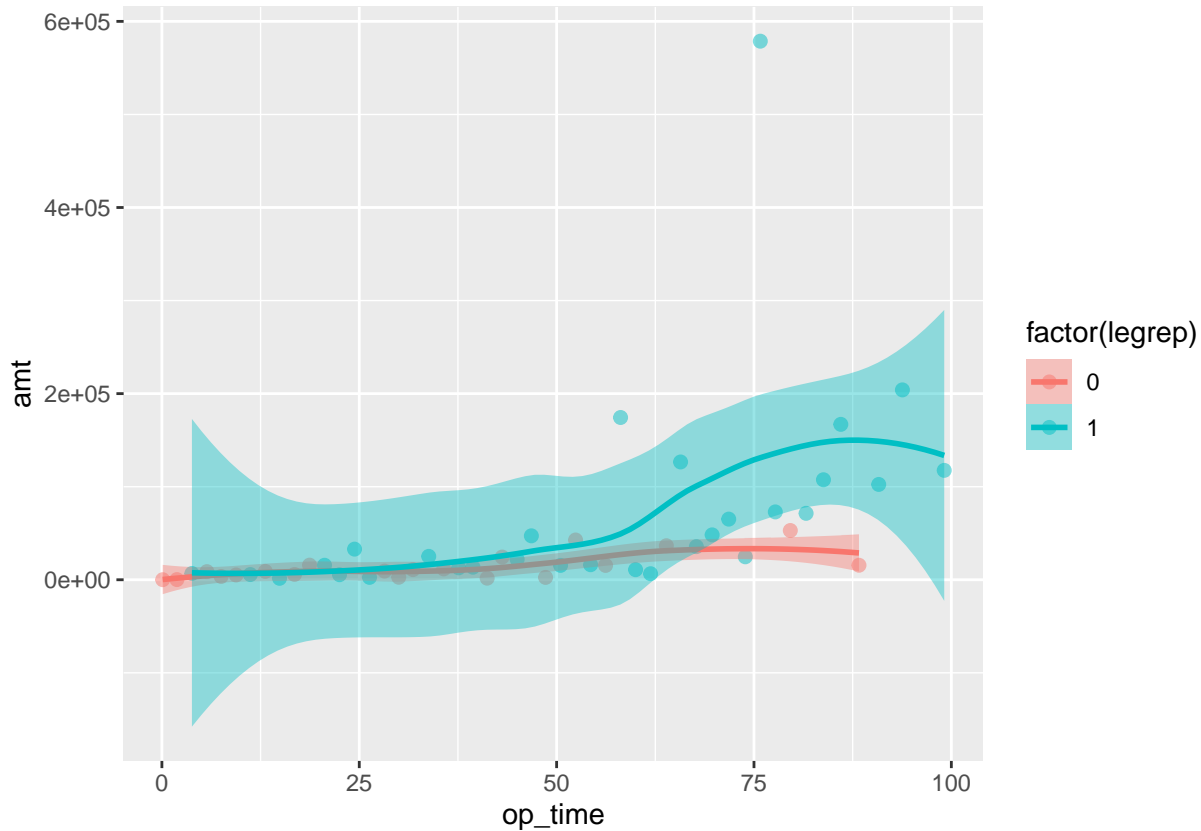


Figure 2.1.5: A version of Figure 2.1.4 with the data points enlarged and the standard error bands added.

Thanks to Figure 2.1.5, we can see that the claim amount of the two types of injuries behaves quite differently with respect to `op_time`, with those for legal representation being much more sensitive to changes in `op_time`. In Section 4.2, we will use a predictive model to quantify the difference between the two forms of behavior formally. Figure 2.1.5, produced by `ggplot2`, allows us to discover such a phenomenon in the first place and is a very useful starting point for such an investigation.

Geom-specific aesthetics. What if you want to make just one smoothed curve applied to all 50 observations in the `persinj50` dataset while still having them colored according to the presence of legal representation? We can do so by specifying different aesthetic mappings for different geoms as in CHUNK 9.

```
# CHUNK 9
```

```
ggplot(persinj50, aes(x = op_time, y = amt)) +
  geom_point(aes(color = factor(legrep)), size = 2, alpha = 0.5) +
  geom_smooth()
```

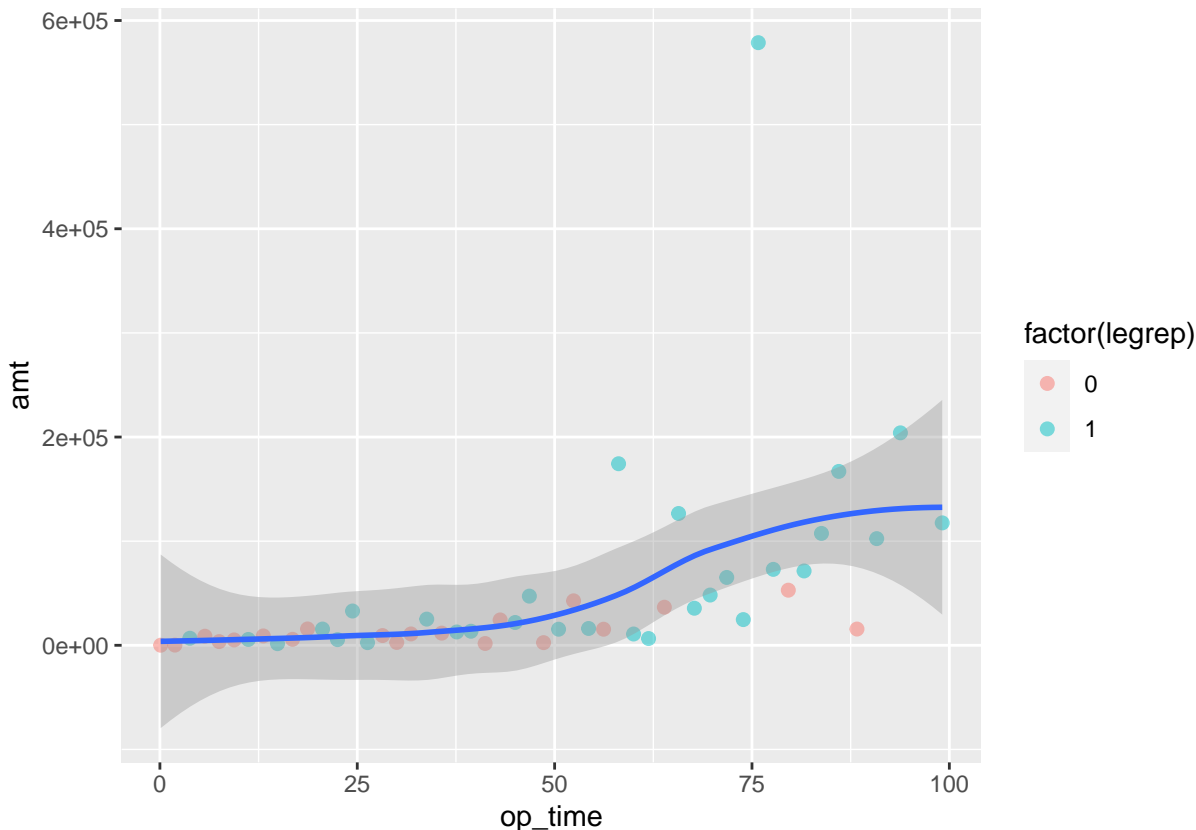



Figure 2.1.6: A variation of Figure 2.1.5 with a single standard error band.

Notice that the `aes()` function in the `ggplot()` call only has the `x` and `y` aesthetics; the `color` aesthetic is moved to the `geom_point()` function. As a result, the 50 points will be distinguished in color by the `legrep` variable. However, there is no such mapping in the `geom_smooth()` function, so a single smoothed curve (colored in blue by default) fitted to all of the 50 observations surrounded by two standard error bands (colored in gray by default) will be produced as shown in Figure 2.1.6. In general, aesthetic mappings common to most, if not all, geoms can be specified in the initial `ggplot()` call. These mappings will be inherited by all geoms. If needed, you can then put in additional aesthetics that apply only to a particular geom to override the default aesthetics.

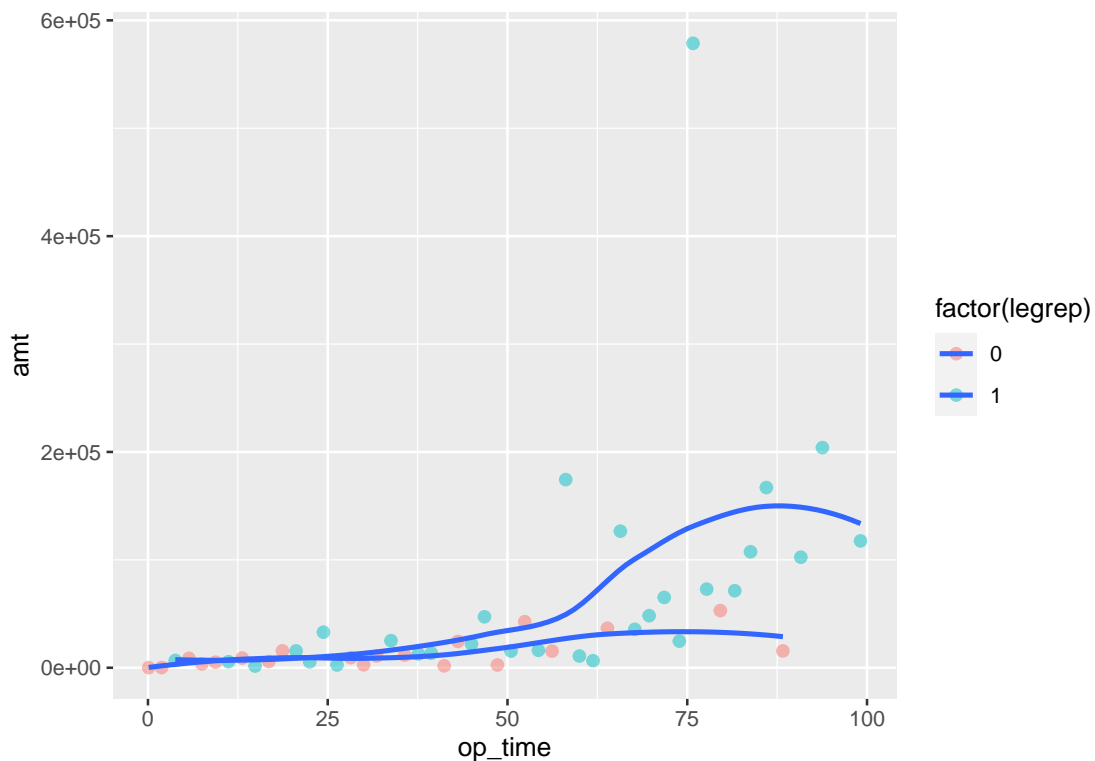
Exercise 2.1.2.  (**Variations of CHUNK 9**) Consider the following variations of CHUNK 9. Think about what kind of plots will be produced. Then run the code and see what happens.

```
# CHUNK 10
ggplot(persinj50, aes(x = op_time, y = amt, fill = factor(legrep))) +
  geom_point(aes(color = factor(legrep)), size = 2, alpha = 0.5) +
  geom_smooth(se = FALSE)
```

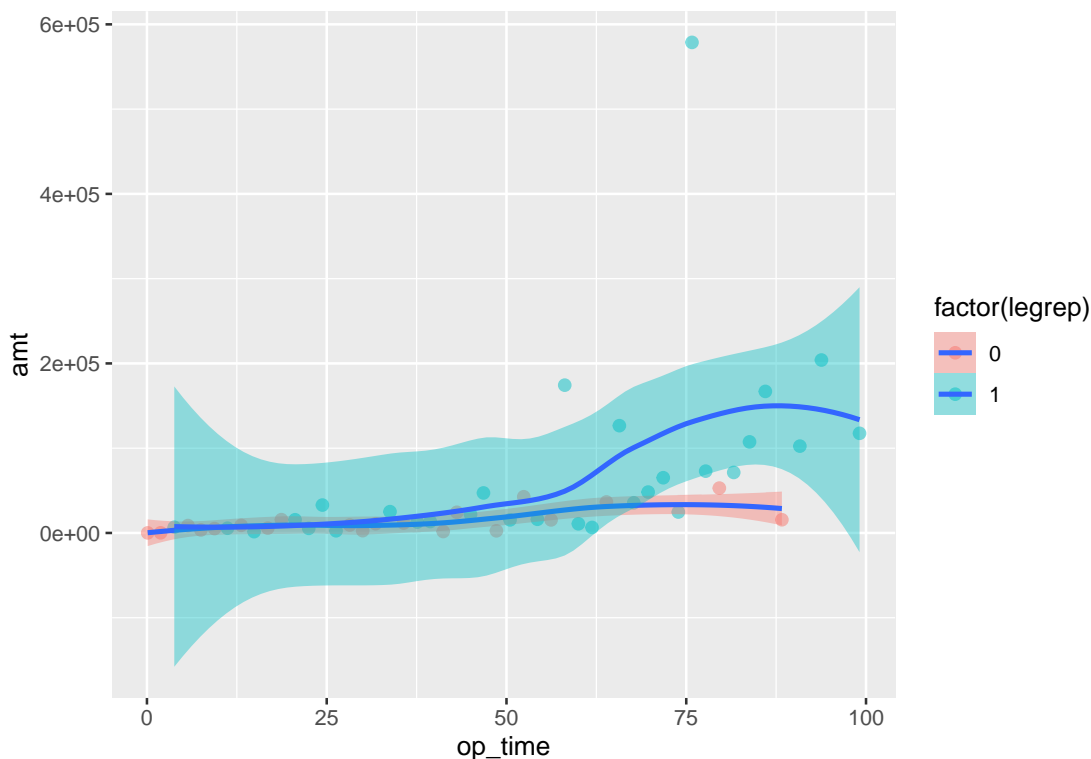
```
# CHUNK 11
ggplot(persinj50, aes(x = op_time, y = amt)) +
  geom_point(aes(color = factor(legrep)), size = 2, alpha = 0.5) +
  geom_smooth(aes(fill = factor(legrep)))
```

Solution. Let's look at the two chunks separately.

- *CHUNK 10:* Compared to CHUNK 9, the code in CHUNK 10 has the `fill` aesthetic added to the initial `ggplot()` call and the option `se = FALSE` added to the `geom_smooth()` function. The `fill` aesthetic has no effect on the `geom_point()` function, but it does affect the `geom_smooth()` function, which, by default, produces the standard error bands that are filled in color according to the `fill` aesthetic. Even though these bands are switched off due to the option `se = FALSE`, separate smoothed curves are still fitted to the two groups of injuries. Without the `color` aesthetic, however, the two curves share the same color, which is [blue](#).



- *CHUNK 11*: Compared to *CHUNK 9*, the code in *CHUNK 11* has the `fill` aesthetic added to the `geom_smooth()` function. As a result, separate smoothed curves are fitted to the two groups of injuries with the standard error bands filled in color according to `legrep`. As the `color` aesthetic is absent in `geom_smooth()`, the two smoothed curves still share the same color (blue).



Remark. This example once again illustrates the subtlety of ggplots. A seemingly minor change in your code can lead to substantially different output. □

Faceting. In Figures 2.1.4 to 2.1.6, we grouped the 50 observations in the `persinj50` data according to the presence of legal representation. In `ggplot2`, grouping is achieved by mapping one or more categorical variables in the data to visual elements like `color`, `shape`, `fill`, `size`, and `linetype`, as we did earlier.

- We now consider *faceting*, which is another useful way to categorize our data into distinct groups. While grouping showcases two or more groups of observations in a single plot, faceting displays the observations in *separate* plots (known as a “small multiple” plot) produced for each value of the faceting variable placed side-by-side, usually on the same scale, to facilitate comparison. In `ggplot2`, faceting is accomplished by the `facet_wrap()` function or the `facet_grid()` function, depending on how many faceting variables there are. The `facet_wrap()` function is often used when there is only one faceting variable. Its generic syntax is


```
facet_wrap(~ FACET_VAR, ncol = N),
```

where the first argument specifies, following the tilde character (`~`), the faceting variable by means of R’s formula syntax (more details on formulas in R will be given in Chapter 3). The second argument, which is optional, determines the number of columns used to display the facets. The code in CHUNK 12, for example, produces the faceted scatterplot in Figure 2.1.7.

```
# CHUNK 12
```

```
ggplot(persinj50, aes(x = op_time, y = amt)) +
  geom_point(size = 2, alpha = 0.5) +
  geom_smooth() +
  facet_wrap(~ legrep) # Try to add scales = "free" to see what happens
```

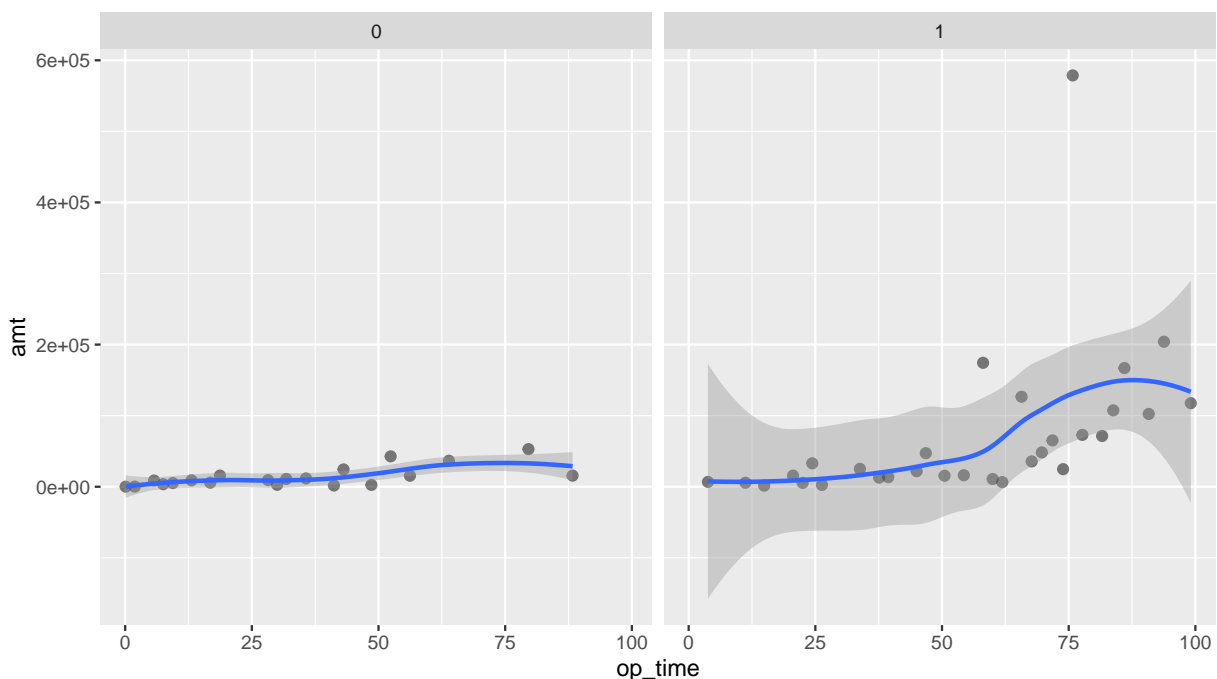


Figure 2.1.7: Scatterplots of `amt` against `op_time` faceted by `legrep` in the `persinj50` dataset.


The two scatterplots are laid out in order according to the two values of `legrep`. A label is displayed at the top of each facet (0 and 1) for easy identification. By default, the two plots are designed to share the same scale of the vertical axis. Adding the option `scales = "free"`, which is used in some past PA exams (e.g., December 2020 exam), to the `facet_wrap()` function will relax this constraint and “free” the scales (try it!). Regardless, Figure 2.1.7 is once again a manifestation that the claim amount of injuries with legal representation behaves differently as a function of `op_time` from those without.

It is possible to do faceting when there are two faceting variables. In this case, we can do a cross-classification using the `facet_grid()` function, which produces a two-dimensional “grid” of plots. Its syntax is similar to that of `facet_wrap()`:

```
facet_grid(FACET_VAR_1 ~ FACET_VAR_2, ncol = N)
```

2.1.2 Customizing Your Plots

We learned the basic structure of a ggplot in the previous subsection. We now look at how to customize a ggplot in terms of the appearance and range of coordinate axes, and how to add and modify cosmetic enhancements such as legends, titles, and subtitles. You will see that it is easy to fine-tune a ggplot to suit your needs.

Axes. Many datasets have outlier values which, when included in the plots, may distort the scaling of the axes  in such a way to obscure the big picture in the data. In other cases, you may want to zoom in and focus on observations lying in a certain range of values, which you can achieve by tweaking the coordinate axes. In `ggplot2`, you can adjust the range of values of the coordinate axes by using the `xlim` and `ylim` arguments of the `coord_cartesian()` function. The two arguments are set to a two-element numeric vector indicating the desired lower and upper limits of the x-axis and y-axis. Data points outside the limits are thrown away.

Another way to display the data points more effectively is to adjust the scale, for example, from a linear scale to a log scale. This is especially useful when dealing with highly skewed variables, as we will see in the next section. The function `scale_x_log10()` (do not miss the parentheses!) converts the scale of the x-axis of a ggplot to a log 10 basis and re-positions the data points accordingly. When this function is applied, the points 10, 100, 1000, 10000 will be shown as consecutive numbers on the x-axis because their log 10 counterparts, $\log_{10} 10 = 1$, $\log_{10} 100 = 2$, $\log_{10} 1000 = 3$, and $\log_{10} 10000 = 4$ are consecutive. As you can expect, the function `scale_y_log10()` performs the same operation on the y-axis.

Titles, subtitles, and captions. In some cases, you want to give more fancy names for the axis labels. The `labs()` function allows us to set the label for the x-axis, y-axis, and the text for the title, subtitle, and caption of a ggplot using the `x`, `y`, `title`, `subtitle`, and `caption` arguments, respectively, with the desired label or text supplied as a character string.^{iv}

⚠ EXAM NOTE ⚠

In quite a few past PA exams, you are asked to comment on the strengths and weaknesses of a given plot. The lack of (useful) axis labels and titles, while minor in most cases, is often a weakness considered by the SOA. Try to keep this in mind for future exams.

To illustrate the use of the cosmetic enhancements above, run CHUNK 13 to produce an enhanced version of the scatterplot in CHUNK 8 (Figure 2.1.5) with labels for the x-axis, y-axis, and the title added, and with the y-axis restricted to the range between $-200,000$ and $300,000$. The resulting scatterplot is shown in Figure 2.1.8. The restriction of the y-axis has the effect

^{iv}If you want to set just the label for the x-axis, y-axis, or the text for the title, you can use the `xlab()`, `ylab()`, and `ggtitle()` functions, respectively.

of excluding the outlier whose claim amount is close to 600,000, way more than other claim amounts, and helping us focus on the main trend in the data much more easily.

```
# CHUNK 13
ggplot(persinj50, aes(x = op_time,
                     y = amt,
                     color = factor(legrep),
                     fill = factor(legrep))) +
  geom_point(size = 2, alpha = 0.5) +
  geom_smooth() +
  labs(title = "Personal Injury Dataset",
       x = "Operational Time",
       y = "Claim Amount") +
  coord_cartesian(ylim = c(-200000, 300000))
```

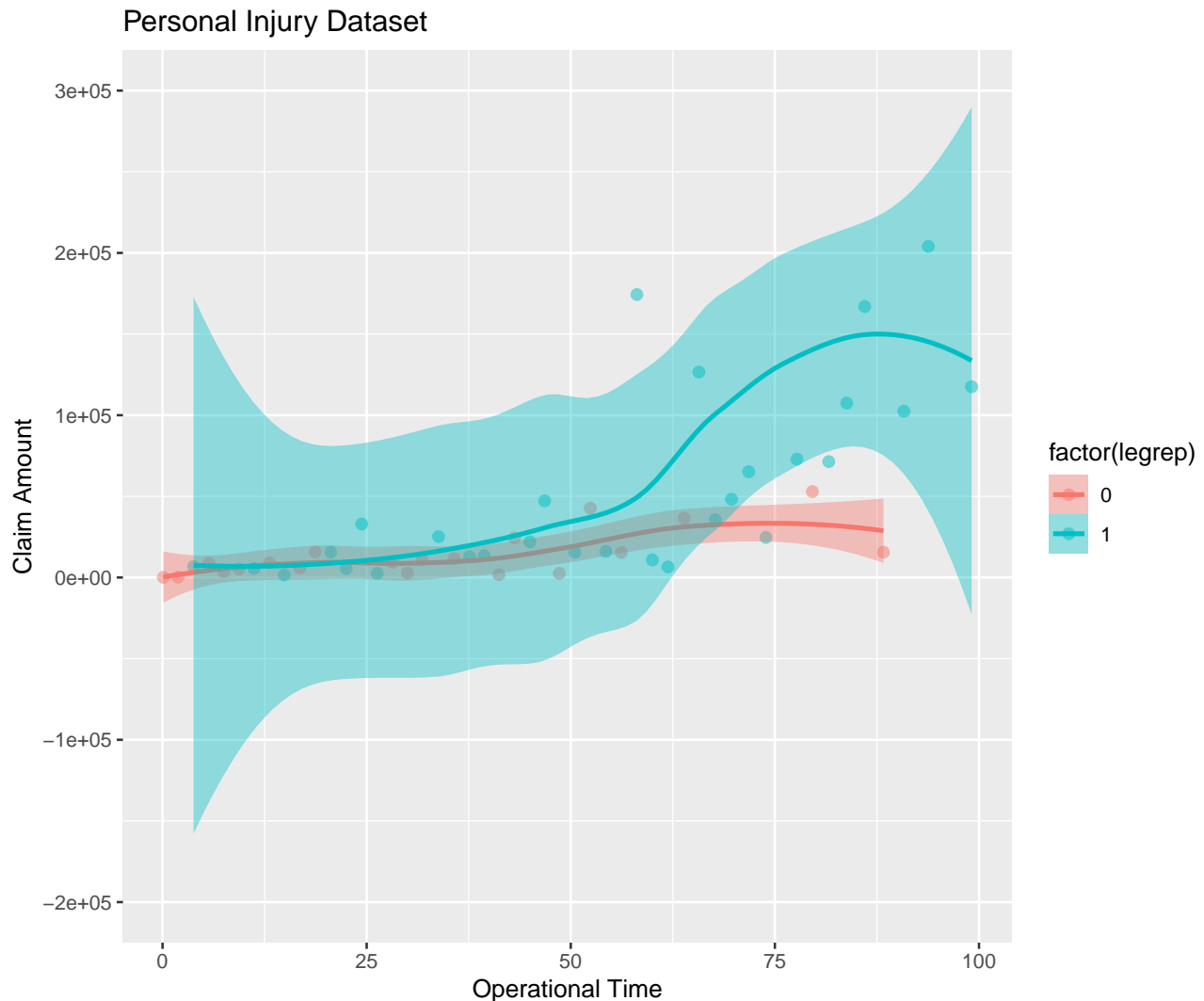


Figure 2.1.8: A version of Figure 2.1.5 with labels for the x-axis, y-axis, and the title added, and with the y-axis restricted to the range between $-200,000$ and $300,000$.

2.2 Data Exploration

- Now that we have learned how to make some simple ggplots, in this section we will apply our data visualization techniques to perform *exploratory data analysis* (EDA), which is an indispensable part of any predictive modeling exercise and is tested in almost every PA exam. The following tasks from the three most recent PA exams will convince you of how important EDA is.

April 2023 PA Exam: Task 1

Your assistant creates two graphs and wants to choose the graph that provides the more easily understood visualization of the relative number of boardings at each stop.

- (a) (2 points) State which graph your assistant should use and explain why this graph is better than the alternative.

October 12, 2022 PA Exam: Task 3

Your boss wants to understand how the distribution of `Time.to.resolution` for requests to pick up sofas differs between departments. There are three departments that pick up sofas: Blight, Grounds, and Sanitation.

- (a) (3 points) Create a single plot of `Time.to.resolution` by department for the provided sofa data.
- (b) (2 points) Describe the observed differences by department.

April 12, 2022 PA Exam: Task 4

Your boss, B, has asked you to use data visualization techniques to better understand the distributions of response time or its components by station.

- (a) (3 points) Describe strengths and weaknesses of the graph above, which was created by your assistant to depict **travel.time**.
- (b) (4 points) Create an informative boxplot of **response.time** by **station** that B can include in a report to the city manager. Include a horizontal line at 360 seconds. Paste the code used to create the graph and the image of the graph below.
- (c) (2 points) Compare the outliers in travel time and response time between the assistant's chart in part (a) and the chart you produced in (b) and describe what is surprising.

EDA is an integral part of predictive modeling because it serves two important purposes:

- *Data validation*: It allows us to perform commonsense checks and identify nonsensical data values (e.g., a negative value for age, which is impossible), which are potential data errors that may lead to unreasonable model results and should be fixed. It also reveals the possible existence of outliers that merit further considerations. After anomalous and inappropriate data values have been removed, the data becomes ready for analysis.
- *Characteristics of variables*: It also helps us understand the key characteristics of the variables in the data. Such an understanding may suggest useful ways to pre-process the variables to improve the prediction performance and interpretability of the models we will construct, and, most importantly, decide on an appropriate type of predictive model that is likely to meet our business needs.

Typically, EDA is accomplished by a *combination* of two kinds of tools:

1. *Descriptive statistics* (a.k.a. summary statistics) that quickly summarize different distributional properties of the variable(s) of interest

Examples: Mean, variance, mode, correlation, table of frequency counts

2. *Graphical displays* (a.k.a. visual displays) that allow us to get a quick impression of the overall distribution of the variable(s) of interest

Graphs are often more informative than a table of summary statistics and sometimes can reveal information that would be missed otherwise, e.g., the presence of outliers.

Examples: Histograms, boxplots, bar charts, and their variants

In this section, we will return to the full `persin` data (with 22,036 observations) and use it to illustrate the creation and interpretation of descriptive statistics and graphical displays.

2.2.1 Univariate Data Exploration

Let's begin with *univariate* data exploration—exploration that sheds light on the distribution of only one variable at a time. The specific statistics and graphical tools will depend on whether the variables you are analyzing are numeric or categorical (precise definitions of numeric and categorical variables will be given on page 137). Both types of variable are part of the dataset of a typical PA exam.

Numeric Variables

Descriptive statistics. Statistical summaries are mainly used to reveal two aspects of the distribution of a **numeric variable**:

- *Central tendency*: The central tendency of a numeric variable, whether it be continuous or discrete, is often quantified by its **mean** and **median**. These two metrics capture, in a loose sense, the typical “size” of the variable and can be readily produced in R by applying the `summary()` function to the variable of interest.

- *Dispersion:* Common measures of dispersion include **variance**, **standard deviation**, and **inter-quartile range** (defined as the difference between the 75% quantile and the 25% quantile of a variable), all of which measure in a way how spread out the values of the numeric variable are over its range.

Graphical displays. To visualize the distribution of numeric variables, histograms and boxplots are convenient graphical aids.

- *Histograms:* **Histograms** divide the observations into several equally spaced bins (or buckets) and provide a visual summary of the count or relative frequency in each bin. Looking at a histogram, we can learn about the overall shape of the distribution of a numeric variable and where most observations lie. Figure 2.2.1 (a) shows a prototypical histogram.

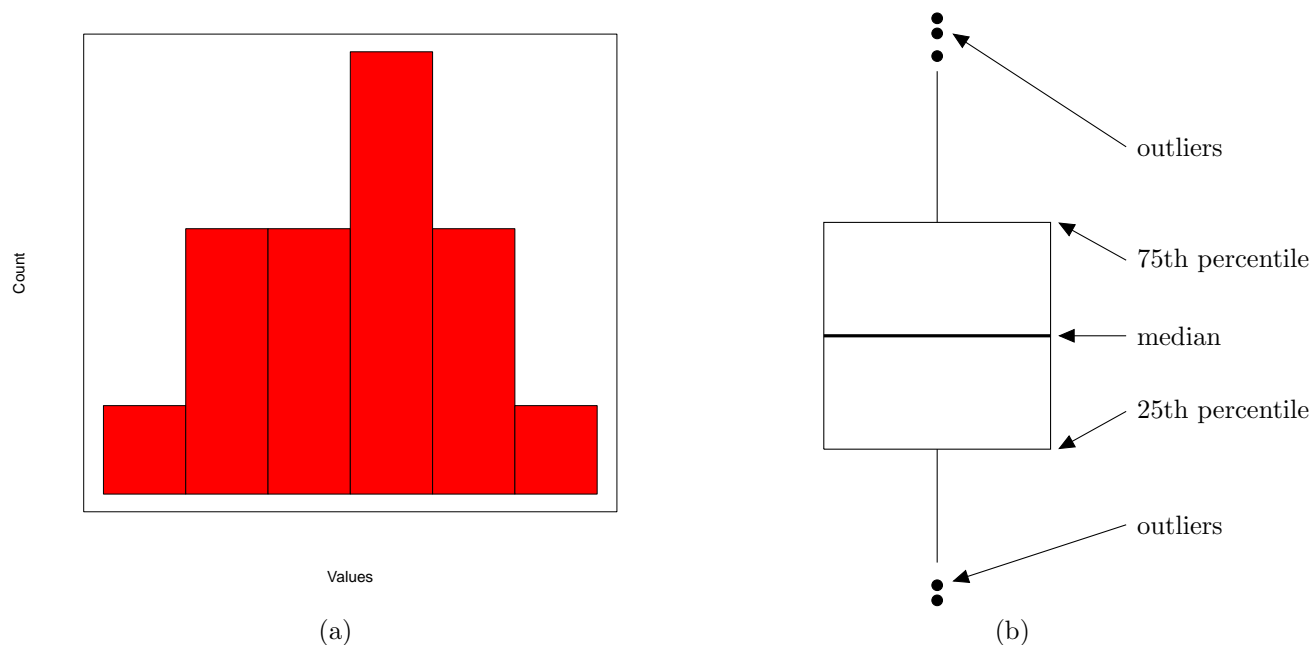


Figure 2.2.1: Prototypical histogram (left) and boxplot (right).

- *Boxplots:* **Boxplots**, a.k.a. *box plots* and *box-and-whiskers plots*, visualize the distribution of a numeric variable by placing its 25% quantile, the median, the 75% quantile in a “box,” with the rest of the data points constituting the “**whiskers**.” The amount of spacing between different parts of a boxplot reflects the degree of dispersion and skewness of the variable’s distribution. “**Outliers**,” defined here as data points that are above or below 1.5 times the inter-quartile range from either edge of the box, are shown as large dotted points. See Figure 2.2.1 (b) for a prototypical boxplot.

Although boxplots do not directly show the actual shape of the variable’s distribution, they offer a useful graphical summary of the key numeric statistics and allow for a visual

Chapter 8

Practice Exams

Introduction

Having been well *trained* on the core of this study manual (Chapters 1 to 6) and past PA exam projects (Chapter 7), you need to be exposed to unseen *tests* to avoid overfitting and identify areas in which you need more *training*. To this end, please make good use of the two substantially updated practice exams in this chapter, one on classification and one on regression. Each exam comes with the following resources:

1. A project statement describing a business problem, a data dictionary, and a series of tasks you have to complete

According to the PA exam syllabus,

“A hardcopy of the problem statements will *not* be available at Prometric testing centers. The statements will be available for the entirety of the exam on-screen.”

This is rather unfortunate because when I took the exam in December 2019, having a printed statement to look at helped quite a lot.

2. (*Available for download on Actuarial University as a separate file*) A Microsoft Word document with spaces labeled as “**ANSWER:**” for you to write your responses to each specific subtask when you practice

On the real exam, this Word document and the project statement are the same file. In other words, you will enter your responses directly in the project statement, similar to FSA written-answer exams. This is the only file you will submit for grading.

3. Detailed illustrative solutions with sample responses and related learning outcomes from the PA exam syllabus identifiedⁱ

ⁱNot all subtasks conveniently fit into the learning outcomes in the syllabus, but they still lie within the general scope of Exam PA.

⚠ NOTE ⚠

In addition to Practice Exams 1 and 2, we have introduced a graded mock exam product with completely different questions, which is available for separate purchase. Please refer to page xxiii of the preface or check out <https://www.actexlearning.com/exams/pa/exam-pa-mock-exam> for more details.

What are these two practice exams like?

Designed taking the new exam format effective from April 2023 and the style of recent PA exams into account, these two practice exams give you a holistic review of the entire PA exam syllabus and have the following characteristics:

- They consist of **7 to 9 tasks**, with a total of **70 points**. Some tasks are longer and some shorter. Almost all tasks are further broken down into a few subtasks. Exam points are provided for each subtask, so you have some idea of how much you should write. As I mentioned in the preface of this manual, you should spend about 3 minutes per exam point.
- Following the exam format effective from the December 2021 sitting, different tasks are mutually independent and can be answered in any order (unless you have a special preference, you may simply start with Task 1). Even if you struggle in a certain task, you can proceed to the next task and start anew. You will not make data preparation or modeling decisions that affect the rest of the project. There are also no tasks about comparing the performance of models constructed in different tasks. (You may have to rank models and select the best model *within the same task*, however.)
- Like recent exams, there are a large number of conceptual or descriptive tasks testing your prior understanding of predictive analytic concepts (look for the verbs “Describe” and “Explain” in the question prompt). You can complete these tasks without looking at any R code or output, or referring to the business problem.
- (*New!*) Starting from the April 2023 sitting, R and RStudio will not be available on the exam, but as the PA exam syllabus says,

“all code and output relevant to the tasks will be provided as part of the exam materials.”

The two practice exams embrace this new format. In quite a few tasks (e.g., Tasks 1, 2, 3, 6, and 8 of Practice Exam 1), you are given some R output and asked to use the output to answer the questions. Sometimes R code is also provided (e.g., Task 5 of Practice Exam 1). You are expected to know what the code does to address some subtasks adequately. This is what Chapters 1 and 2, and the R-based case studies in the manual are for.







- They strike a good balance between easy items testing topics regularly featured in past exams and harder, more unfamiliar items. As comprehensive as this study manual is, each PA exam will likely have a small number of unfamiliar tasks designed to identify the candidates that thrive on new, unseen exam tasks and are not overfitted to past exams.

The harder items in the practice exams are in a similar vein.

(In fact, I am not surprised if members of the PA exam committee have access to this manual and deliberately test obscure things I did not discuss at length! 🤔)

How to use these practice exams?

To make the most of these practice exams, please:

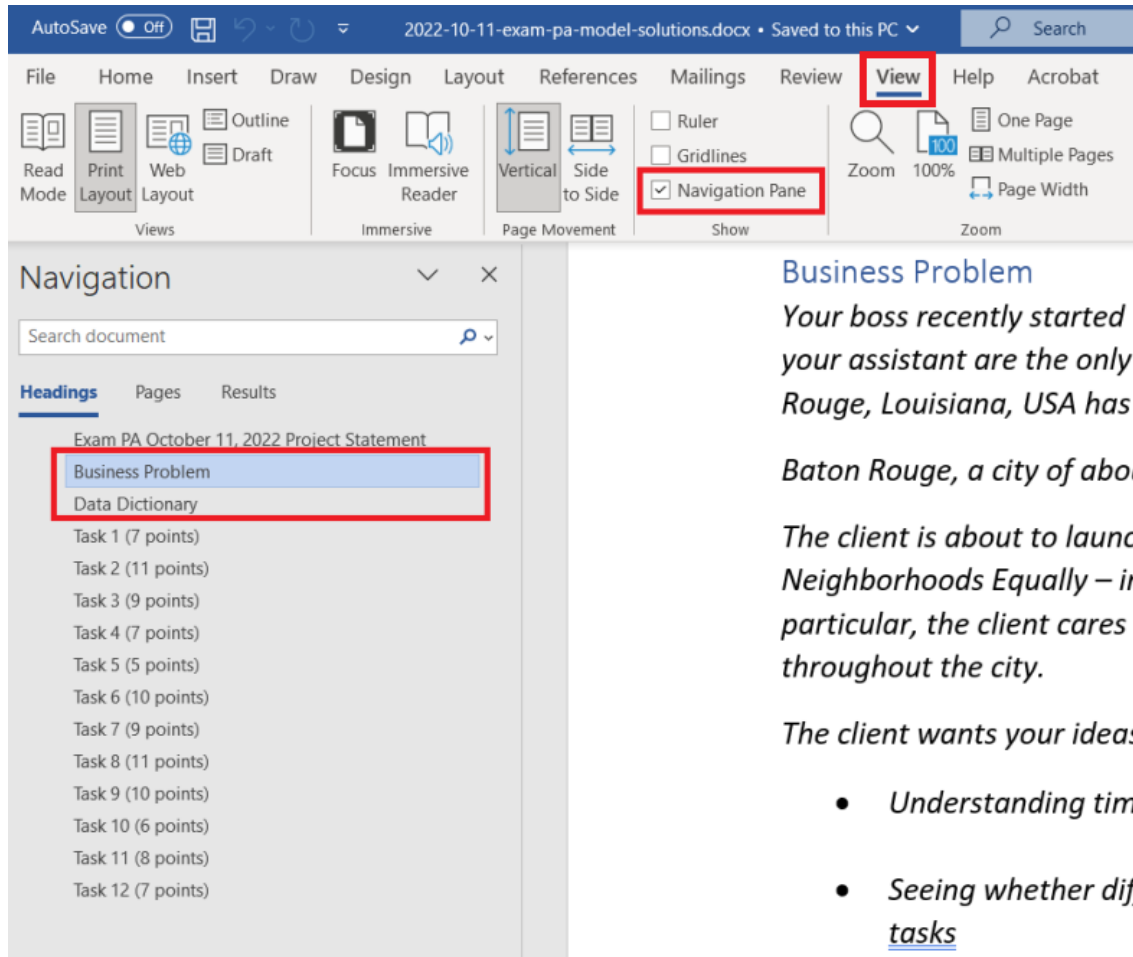
- Attempt them only when you have finished reading the study manual and studied recent PA exams (at least the October 2023 and April 2023 exams). Working on the practice exams when you are not fully ready defeats their purpose.
- Set aside exactly 3 hours 30 minutes and work on each exam in a simulated exam environment detached from distractions. Put away your manual, notes, and phone—no Facebook , Instagram , Twitter , or Snapchat  for 3.5 hours! You can only have your calculator  with you (you may also use Excel , which is available on the exam, to do calculations if you prefer). When you are finished, compare your responses with my suggested solutions and see how well you have done.
- Be sure to read the task statement and the Business Problem section carefully. Almost always, the Business Problem section has something useful for answering a few subtasks, and a seemingly minor point mentioned there can make a huge difference.
- Budget your time wisely. Don't spend a disproportionate amount of time on a single subtask, no matter how difficult it seems. As I mentioned in the preface, you should spend 3 minutes per exam point on average.

NOTE

- Don't feel too frustrated if you find these two exams hard and long—they are probably (a bit) harder than the real exam! It is better to see something more difficult when you practice than to be defeated on the real exam, right? 😊
- Practice Exam 2 will be posted on www.actuarialuniversity.com after the SOA releases the exam paper and solutions online

A Note on Navigation

During the exam, you may want to scroll back to the Business Problem and Data Dictionary at the beginning of the Microsoft Word file, then continue to work on different tasks. To navigate back and forth efficiently, press **Ctrl+F**, or click **View > Navigation Pane**.



This may save you some time and trouble on the exam, where every second counts!

ACTEX PA Manual Practice Exam 1 Project Statement

IMPORTANT NOTICE – THIS IS THE PROJECT STATEMENT OF THE FIRST PRACTICE EXAM. IF YOU ARE NOT READY FOR IT YET, LEAVE IMMEDIATELY AND RETURN LATER.

General Information for Candidates

This examination has 9 tasks numbered 1 through 9 with a total of 70 points. The points for each task are indicated at the beginning of the task, and the points for subtasks are shown with each subtask.

Each task pertains to the business problem and data dictionary described below. Additional information on the business problem may be included in specific tasks—where additional information is provided, including variations in the target variable, it applies only to that task and not to other tasks. For this exam there is no data file or .Rmd file provided. Neither R nor RStudio are available or required.

The responses to each specific subtask should be written after the subtask and the answer label, which is typically ANSWER, in the separate Word document.ⁱⁱ Each subtask will be graded individually, so be sure any work that addresses a given subtask is done in the space provided for that subtask. Some subtasks have multiple labels for answers where multiple items are asked for—each answer label should have an answer after it.

Each task will be graded on the quality of your thought process (as documented in your submission), conclusions, and quality of the presentation. The answer should be confined to the question as set. No response to any task needs to be written as a formal report. Unless a subtask specifies otherwise, the audience for the responses is the examination grading team and technical language can be used.

Prior to uploading your Word file, it should be saved and renamed with your five-digit candidate number in the file name. If any part of your exam was answered in French, also include “French” in the file name. Please keep the exam date as part of the file name.

The Word file that contains your answers must be uploaded before the five-minute upload period time expires.

ⁱⁱAs mentioned before, you will write your responses directly in the project statement on the real exam.

Business Problem

You work at ABC, a large actuarial consulting firm, and have been asked to assist School Wiz, a group dedicated to providing remedial education to troubled students. School Wiz has heard about you the legend, because you are among the very few who got Grade 10 in Exam PA, and wants to explore using your services to advance their business goals. They have collected preliminary dataⁱⁱⁱ of past high school students. They would like to be able to identify which of the incoming high school students have a high tendency to fail, before they enter their high school year. These students will receive remedial services in time.

School Wiz has determined that out of the three grade variables in the data, G1, G2, and G3, they would like you to just focus on building predictive models based on G3. A student who receives a grade of 10 or more will pass. Your goal is to use the available data to construct two models that will predict if a student will pass (rather than the overall grade). One model should be GLM-based and one should be tree-based.

School Wiz has provided the following data dictionary.

ⁱⁱⁱThis practice exam is based on the setting of the Student Success sample project (available from pages 8 and 9 of the [June 2021 PA exam syllabus](#)) and turns it into a much more useful task-based project consistent with the current exam format. The dataset for this sample project in turn is adapted from the Student Performance Data Set contributed by Paulo Cortez to the UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.

Data Dictionary

Name	Description	Variable Values
sex	Student's sex	Binary: F (female) or M (male)
age	Student's age	Integer from 15 to 22
Medu	Mother's education	Integer from 0 (none) to 4 (higher education)
Fedu	Father's education	Integer from 0 (none) to 4 (higher education)
Mjob	Mother's job	Factor: at_home, health (health care related), other, services (civil services, administrative or police), teacher
Fjob	Father's job	Same levels as Mjob
studytime	Weekly study time	Integer from 1 (very short) to 4 (very long)
failures	Number of past class failures	Integer from 0 to 3
schoolsup	Extra educational support	Binary: yes or no
famsup	Extra family supplement	Binary: yes or no
paid	Extra paid classes	Binary: yes or no
activities	Extra-curricular activities	Binary: yes or no
internet	Internet access at home	Binary: yes or no
romantic	Has a romantic relationship	Binary: yes or no
famrel	Quality of family relationships	Integer from 1 (very bad) to 5 (excellent)
freetime	Free time after school	Integer from 1 (very low) to 5 (very high)
goout	Going out with friends	Integer from 1 (very low) to 5 (very high)
Dalc	Weekday alcohol consumption	Integer from 1 (very low) to 5 (very high)
Walc	Weekend alcohol consumption	Integer from 1 (very low) to 5 (very high)
absences	Number of absences in high school year	Integer from 0 to 75
G1	First trimester grade in high school year	Integer from 0 to 20
G2	Second trimester grade in high school year	Integer from 0 to 20
G3	Third trimester grade in high school year	Integer from 0 to 20
pass	Pass indicator	0 if a student fails and 1 if a student passes

Task 1 (7 points)

The following is the correlation matrix for **G1**, **G2**, and **G3**:

	G1	G2	G3
G1	1.0000000	0.8821056	0.8301591
G2	0.8821056	1.0000000	0.9151279
G3	0.8301591	0.9151279	1.0000000

- (a) (2 points) Describe one strength and one weakness of a correlation matrix as a bivariate data exploration tool.

ANSWER:

- (b) (2 points) Based on the correlation matrix, explain why basing pass or fail entirely on **G3**, as requested by School Wiz, may be a sensible decision.

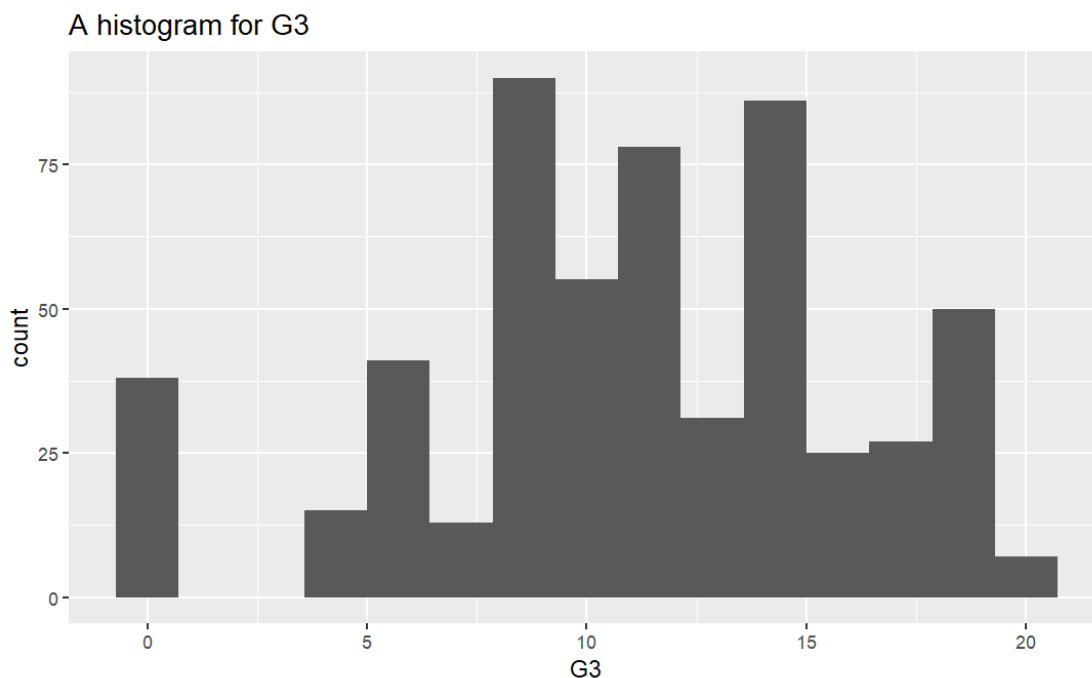
ANSWER:

- (c) (3 points) Describe how principal components analysis can provide an alternative method for determining whether a student will pass or not based on all of **G1**, **G2**, and **G3**.

ANSWER:

Task 2 (5 points)

An alternative to modeling **pass** is to treat **G3** as the target variable and model it directly to determine pass or fail. To explore this alternative, your assistant has produced the following histogram for **G3**:



(a) (2 points) Describe the distributional characteristics of **G3**.

ANSWER:

(b) (3 points) Discuss one advantage and one disadvantage of modeling **G3** as the target variable over modeling **pass** for School Wiz from a GLM perspective.

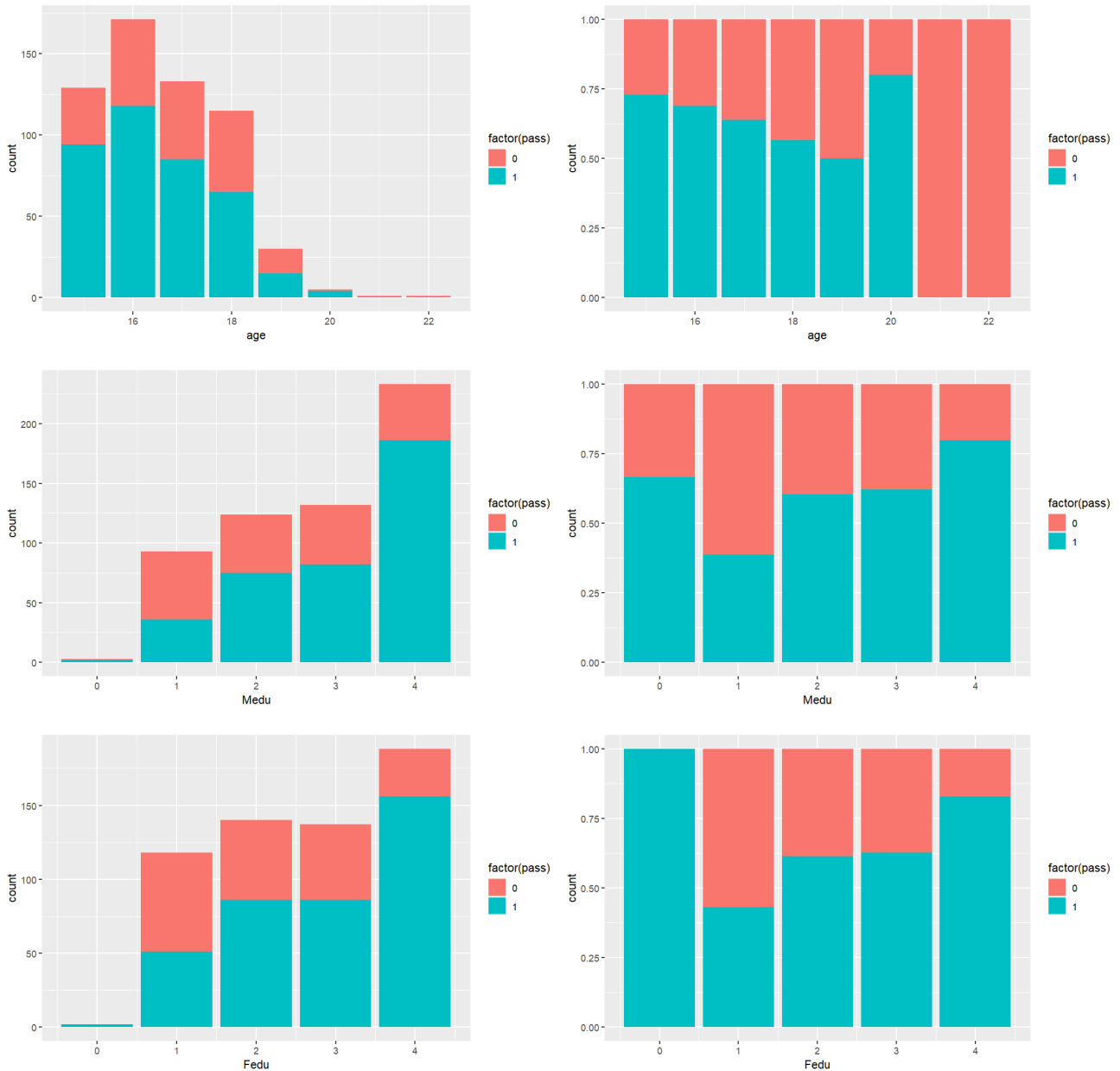
ANSWER:

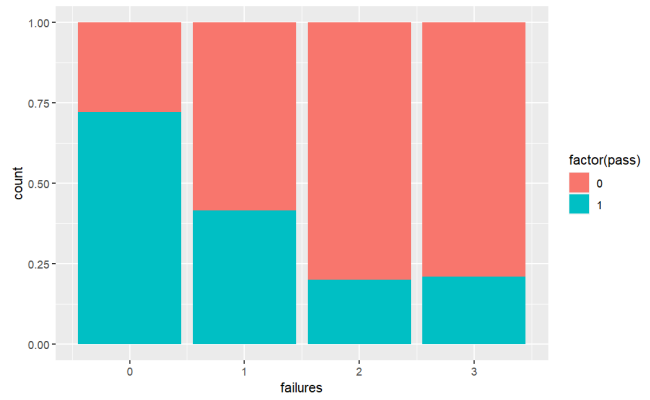
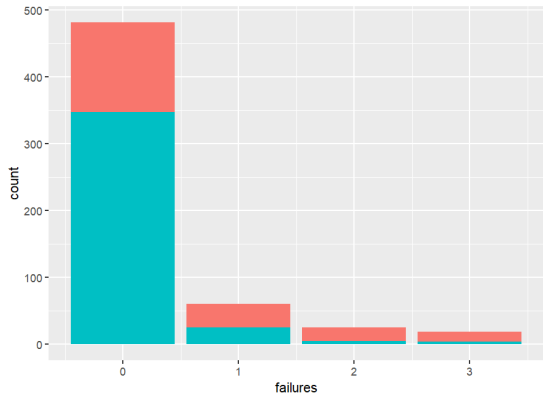
Task 3 (7 points)

You have asked your assistant to investigate the variables in the data.

(a) (2 points) Explain the problem with using **absences** for predicting **pass**.

Your assistant has conducted exploratory data analysis for **pass**. Here is part of the output:





(b) (3 points) Describe two anomalies of the data revealed by the output above.

ANSWER:

(c) (2 points) Identify and explain which variable above appears to be the most important predictor of **pass**.

ANSWER:

Task 4 (5 points)

Your assistant suggested creating a new variable that flags any previous class failures and including this flag variable in your models, in addition to variables that already exist in the data. The value of the variable is 1 if **failures** is higher than or equal to 1, and 0 if **failures** is 0. Your assistant thinks that this variable may be a useful feature for predicting **pass**.

- (a) (3 points) Explain the modeling impacts, if any, of adding the new flag variable when running a GLM.

ANSWER:

- (b) (2 points) Explain the modeling impacts, if any, of adding the new flag variable when running a decision tree.

ANSWER:

Task 5 (9 points)

Your assistant has provided the following R code to perform a certain cluster analysis.

```
data.hc <- data.all[, c("Medu", "Fedu")]
```

```
data.hc$Medu <- scale(data.hc$Medu)
```

```
data.hc$Fedu <- scale(data.hc$Fedu)
```

```
hc <- hclust(dist(data.hc))
```

(a) (2 points) Explain how cluster analysis can be used to develop features for a predictive model.

ANSWER:

(b) (3 points) Explain what kind of cluster analysis is performed by your assistant.

ANSWER:

(c) (2 points) Explain what the **scale()** function in your assistant's code does and why it is important.

ANSWER:

In retrospect, your assistant thinks that the code should have included a random seed so that the same output will be obtained every time the code is run. He apologizes for this omission.

(d) (2 points) Critique your assistant's statement.

ANSWER:

Task 6 (12 points)

Having read the *ACTEX Study Manual for Exam PA*, your assistant knows that accuracy, sensitivity, specificity, and AUC are commonly used performance metrics for a classifier.

(a) (3 points) Define accuracy, sensitivity, specificity, and AUC for a general classifier.

ANSWER:

Accuracy:

Sensitivity:

Specificity:

AUC:

(b) (4 points) Describe how accuracy, sensitivity, specificity, and AUC vary with the cutoff of a classifier.

ANSWER:

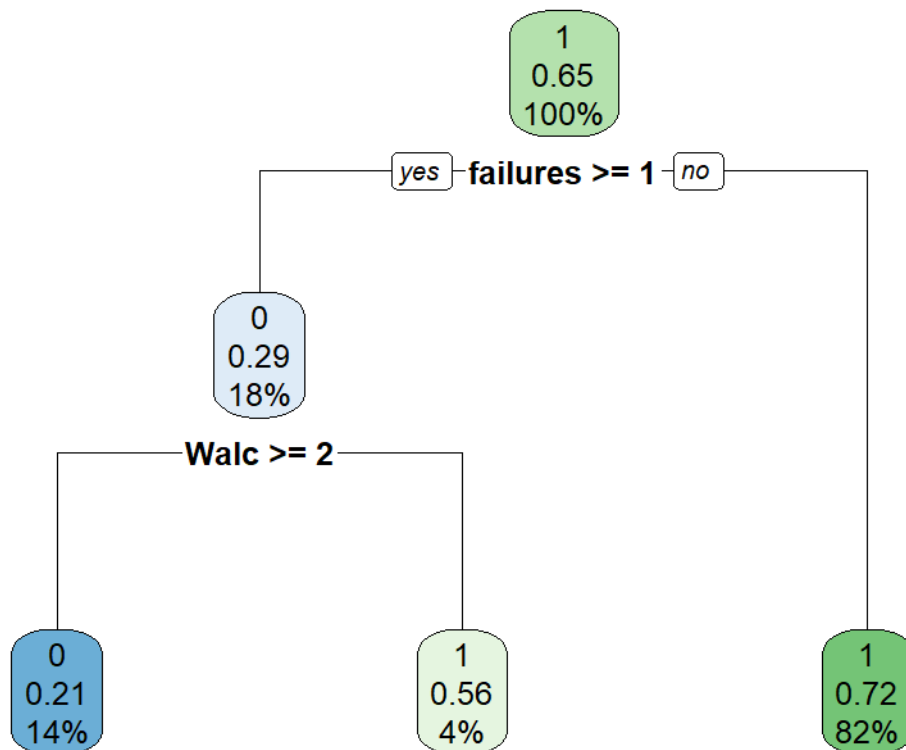
Accuracy:

Sensitivity:

Specificity:

AUC:

Your assistant has fitted a classification tree for **pass** on a subset of the data containing 390 observations, resulting in the following tree.



(c) (5 points) Fill in the following confusion matrix for the classification tree based on a cutoff of 0.5. Show your work.

Prediction	Reference	
	0	1
0	?	?
1	?	?

Task 7 (11 points)

Your assistant has provided code to split the data into the training (70%) and test sets (30%).

- (a) (2 points) Describe the trade-off involved when selecting the percentages of data in the training and test sets.

ANSWER:

Your assistant has also set up code for fitting a regularized logistic regression model for **pass** on the training set.

- (b) (3 points) Explain why **lambda** and **alpha** in an elastic net are hyperparameters and how these two parameters affect an elastic net.

ANSWER:

Why lambda and alpha are hyperparameters:

Lambda:

Alpha:

- (c) (2 points) Describe the significance of using alpha equal to 1 in this business problem.

ANSWER:

The following shows the coefficient estimates of the variables selected in the resulting model:

	s0
(Intercept)	0.49996348
Medu	0.23399344
Fedu	0.09891187
Mjobother	-0.14728207
failures	-0.68078016
famsupyes	-0.51504749
goout	-0.07769020
Walc	-0.02494307

(d) (4 points) Interpret the estimates of the intercept, and the coefficients for the categorical variable and the numeric variable with the most significant impact on **pass**.

ANSWER:

Intercept:

Coefficient for categorical variable:

Coefficient for numeric variable:

Task 8 (7 points)

Your assistant has run a boosted classification tree for **pass** on the training set.

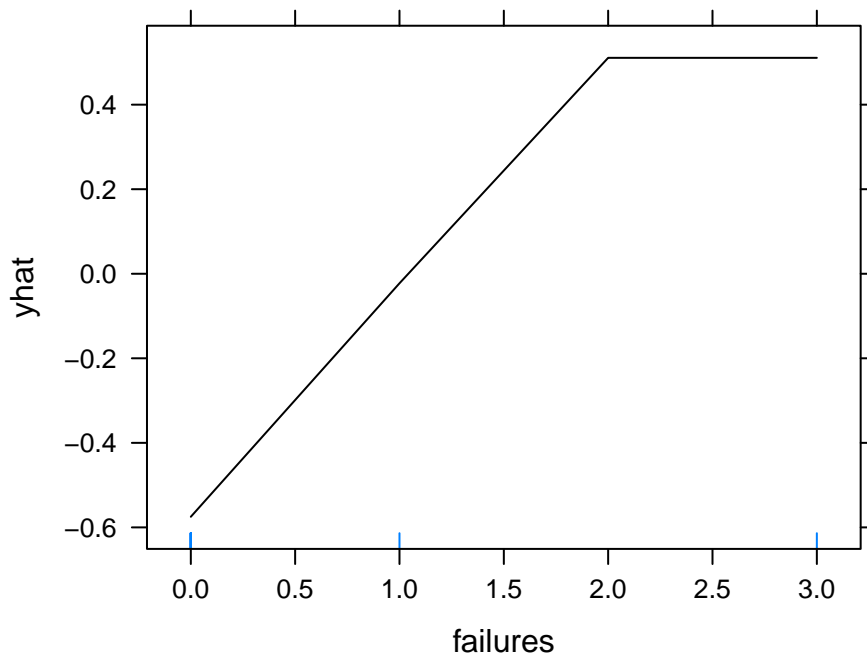
- (a) (3 points) Explain the role played by the **eta** and **nrounds** parameters in a boosted tree, and the considerations for selecting these two parameters.

ANSWER:

eta:

nrounds:

The partial dependence plot for **failures** is provided below.



- (b) (2 points) Provide an interpretation of the plot above.

ANSWER:

- (c) (2 points) Describe the limitation of a partial dependence plot with respect to the interaction between variables.

ANSWER:

Task 9 (7 points)

To help School Wiz put the prediction performance of the models you have constructed or will construct in perspective, your assistant suggests fitting an intercept-only GLM for **pass** on the training set.

- (a) (2 points) Explain how the intercept-only GLM can be used to assess the prediction performance of other models.

ANSWER:

- (b) (3 points) Describe the characteristics of the prediction produced by this model and its ROC curve.

ANSWER:

- (c) (2 points) Determine the test AUC of this model.

ANSWER:

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

Practice Exam 1 Suggested Solutions

▲ NOTE ▲

The following apply to both Practice Exams 1 and 2:

- Each practice exam has a fairly comprehensive coverage of the topics in the PA exam syllabus, ranging from the business problem, data exploration, data preparation, to modeling issues concerning GLMs and decision trees. Apart from conceptual and descriptive items, I made a deliberate attempt to include some subtasks that test your understanding of basic R code and hand calculation based on some R output. These subtasks may figure more prominently in the new exam format.
- The following “suggested” solutions are mainly for illustration purposes. Even though they are likely to be more detailed than what you can write in 3.5 hours, they are by no means perfect. Feel free to augment and refine my responses as you see fit. In many cases, there is a range of fully satisfactory approaches and there may be valid alternatives not discussed.
- Particularly important points in the solutions are underlined for easy identification. While these points (or phrases with similar meaning) can definitely enrich your responses, there is **no expectation that you cover all of them**. As the [Guide to SOA Exams](#) (check out this file if you haven't!) says,
“...candidates do not always need to cover every possible aspect of the solution to receive full points...”
- Some commentary and exam-taking strategies for specific subtasks are shown in *italics*. They are not part of the solutions.

Task 1 – Justify using G3 to determine pass or fail (7 points)

Ambrose’s comments: This is an unseen, but not-so-demanding task specific to this business problem. It is about why using only one grade variable to determine pass or fail makes sense (though it may not be the optimal decision) with reference to the strong correlations among the three grade variables. A closely related topic is PCA.

Relevant PA exam learning outcomes:

- 2b) Identify the types of variables and terminology used in predictive modeling.
- 2f) Apply bivariate data exploration techniques.
- 3b) Apply principal components analysis to transform data.