

# ACTEX Learning

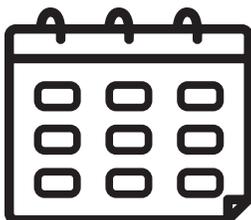
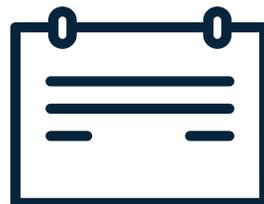
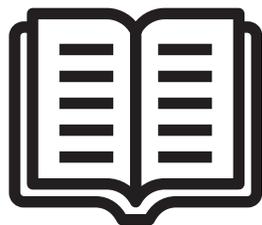
## Exam SRM Study Guide

8<sup>th</sup> Edition 2<sup>nd</sup> Printing

Runhuan Feng, PhD, FSA, CERA

Daniël Linders, PhD

Ambrose Lo, PhD, FSA, CERA



An SOA Exam



 **ACTEX Learning**

**Exam SRM**

**Study Guide**

**8<sup>th</sup> Edition 2<sup>nd</sup> Printing**

**Runhuan Feng, PhD, FSA, CERA**

**Daniël Linders, PhD**

**Ambrose Lo, PhD, FSA, CERA**



*Actuarial & Financial Risk Resource Materials*  
**Since 1972**

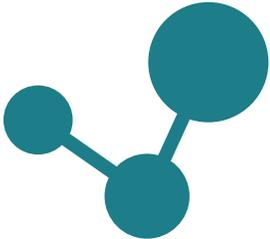
Copyright © 2024, ACTEX Learning, a division of ArchiMedia Advantage Inc.

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](http://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  $\alpha, \beta$ , 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.



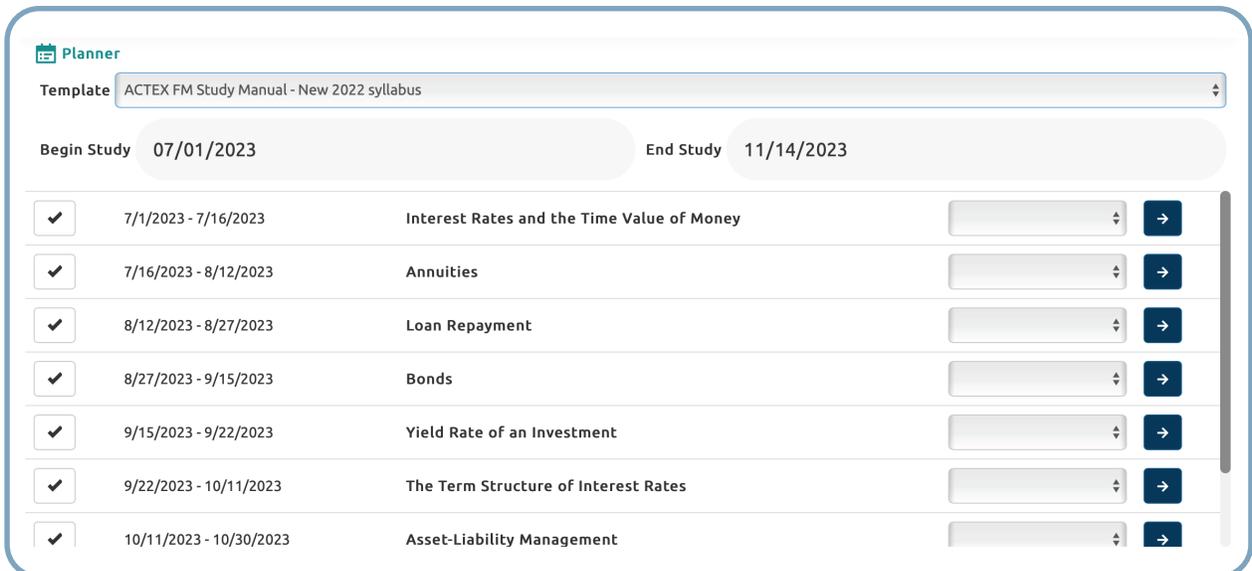
**Locked Products** are products that you do not own, and are available for purchase.



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

<b>Instructional Videos*</b>	<b>Planner</b>
<b>Topic Search</b>	<b>Formula &amp; Review Sheet</b>

**Make your study session more efficient with our Planner!**



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



## Practice. Quiz. Test. Pass!

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

Free with your  
ACTEX or ASM  
Interactive Study  
Manual

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

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

Actuarial University

QUESTION 19 OF 704    Question #    Go!    ⌂    🚩    ✎    📧    ⏪ Prev    Next ⏩    ✕

Question Difficulty: Advanced ⓘ

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

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

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

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

Possible Answers

A 134
✓ 235
✗ 271
D 313
E 352

Help Me Start

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

Solution

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

$y$	$f_Y(y) = P(Y = y)$
0	$P(Y = 0) = P(0 \leq X < 50) = 0.72$
300	$P(Y = 300) = P(50 \leq X < 60) = 0.14$
600	$P(Y = 600) = P(60 \leq X < 70) = 0.06$
700	$P(Y = 700) = P(X \geq 70) = 0.08$

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

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

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

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

Common Questions & Errors

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

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

Rate this problem
👍 Excellent
👎 Needs Improvement
👎 Inadequate

Quickly access the Hub for additional learning.

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

View difficulty level.

Helpful strategies to get you started.

Full solutions with detailed explanations to deepen your understanding.

Commonly encountered errors.

Rate a problem or give feedback.

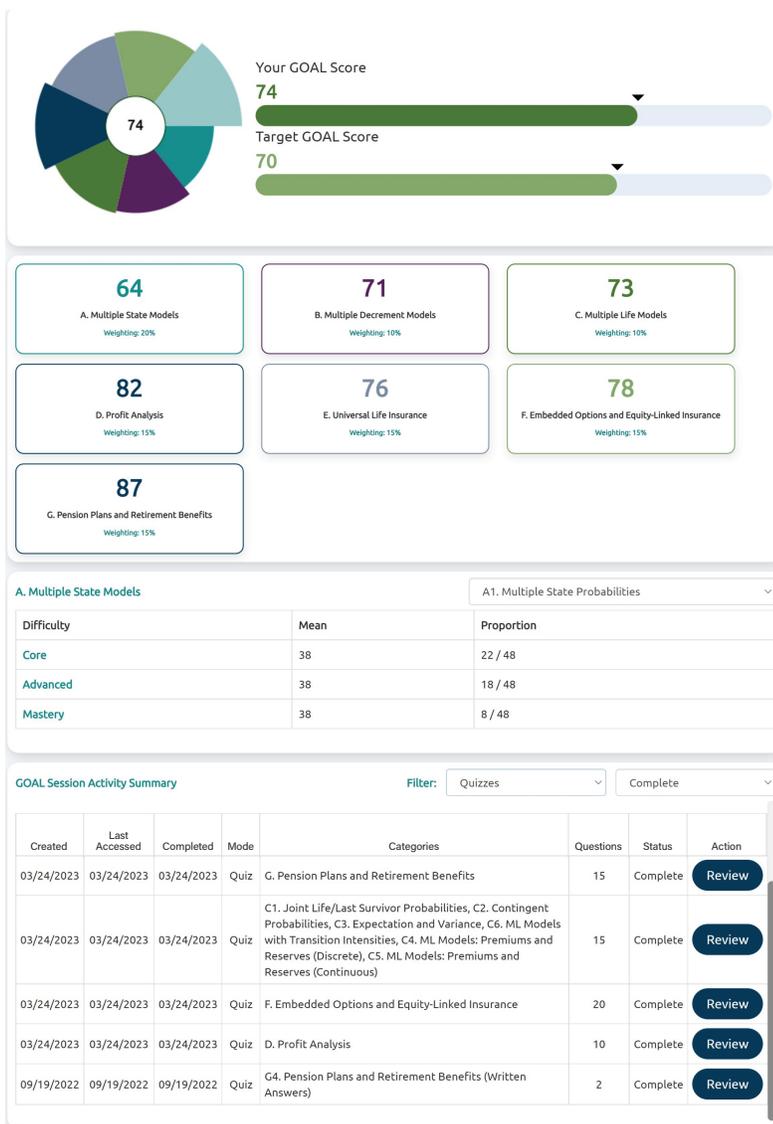


# Track your exam readiness with GOAL Score!

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

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

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



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

See key areas where you can improve.

Detailed performance tracking.

Quickly return to previous sessions.

# Contents

<b>Preface</b>	<b>xiii</b>
P.1 About Exam SRM . . . . .	xiii
P.2 About this Study Manual . . . . .	xxi
<b>I Regression Models</b>	<b>1</b>
<b>Chapter 1 Simple Linear Regression</b>	<b>3</b>
1.1 Model Formulation and Fitting . . . . .	5
1.1.1 What Does an SLR Model Look Like? . . . . .	5
1.1.2 Model Fitting by the Least Squares Method . . . . .	7
Practice Problems for Section 1.1 . . . . .	19
1.2 Assessing the Goodness of Fit of an SLR Model . . . . .	32
1.2.1 Partitioning the Sum of Squares . . . . .	32
1.2.2 F-test . . . . .	40
Practice Problems for Section 1.2 . . . . .	44
1.3 Statistical Inference about Regression Coefficients . . . . .	54
1.3.1 Sampling Distributions of LSEs . . . . .	54
1.3.2 Hypothesis Tests and Confidence Intervals . . . . .	57
Practice Problems for Section 1.3 . . . . .	64
1.4 Prediction . . . . .	81
Practice Problems for Section 1.4 . . . . .	86
<b>Chapter 2 Multiple Linear Regression</b>	<b>97</b>
2.1 From SLR to MLR: Fundamental Results . . . . .	99
Practice Problems for Section 2.1 . . . . .	125
2.2 Partial Correlations . . . . .	151
Practice Problems for Section 2.2 . . . . .	159
2.3 Model Construction . . . . .	164
2.3.1 Types of Predictors . . . . .	164
2.3.2 Relaxing the Linear Assumption . . . . .	173
2.3.3 Relaxing the Additive Assumption . . . . .	175
Practice Problems for Section 2.3 . . . . .	187
2.4 Generalized F-test . . . . .	202
Practice Problems for Section 2.4 . . . . .	210
<b>Chapter 3 Regression Diagnostics</b>	<b>223</b>
3.1 Residual Analysis . . . . .	224

3.1.1	Identification of Outliers . . . . .	224
3.1.2	Detection of Missed Relationships . . . . .	226
3.2	Influential Points . . . . .	228
3.2.1	High-Leverage Points . . . . .	228
3.2.2	Cook's Distance . . . . .	231
3.3	Heteroscedasticity . . . . .	236
3.3.1	Detection of Heteroscedasticity . . . . .	236
3.3.2	Solutions to Heteroscedasticity . . . . .	239
3.4	Collinearity . . . . .	244
3.5	End-of-Chapter Practice Problems . . . . .	256
<b>Chapter 4 Linear Models from a Statistical Learning Perspective</b>		<b>283</b>
4.1	A Primer on Statistical Learning . . . . .	285
4.1.1	Fundamental Concepts . . . . .	285
4.1.2	Assessing Model Accuracy . . . . .	290
4.1.3	Case Study 1: Linear Models . . . . .	298
4.1.4	Case Study 2: $K$ -Nearest Neighbors . . . . .	302
	Practice Problems for Section 4.1 . . . . .	312
4.2	Resampling Methods . . . . .	328
4.2.1	Validation Set Approach . . . . .	329
4.2.2	Cross-Validation . . . . .	331
	Practice Problems for Section 4.2 . . . . .	342
4.3	Variable Selection . . . . .	350
4.3.1	Model Comparison Statistics . . . . .	350
4.3.2	Best Subset Selection . . . . .	359
4.3.3	Automatic Variable Selection Procedures . . . . .	362
	Practice Problems for Section 4.3 . . . . .	374
4.4	Shrinkage Methods . . . . .	390
4.4.1	Shrinkage Method 1: Ridge Regression . . . . .	390
4.4.2	Shrinkage Method 2: The Lasso . . . . .	395
4.4.3	Epilogue: Working in High-Dimensional Settings . . . . .	405
	Practice Problems for Section 4.4 . . . . .	407
<b>Chapter 5 Generalized Linear Models</b>		<b>421</b>
5.1	GLMs: General Theory . . . . .	423
5.1.1	Model Formulation . . . . .	423
5.1.2	Estimation of Parameters . . . . .	434
5.1.3	Assessing Model Fit . . . . .	438
5.1.4	Comparisons between Different GLMs . . . . .	447
	Practice Problems for Section 5.1 . . . . .	455
5.2	GLM Case Study 1: Categorical Response Variables . . . . .	489
5.2.1	Binary Response Variables . . . . .	489
5.2.2	Nominal and Ordinal Regression . . . . .	502
	Practice Problems for Section 5.2 . . . . .	508
5.3	GLM Case Study 2: Count Response Variables . . . . .	525
5.3.1	Poisson Regression . . . . .	525
5.3.2	Overdispersion . . . . .	532
	Practice Problems for Section 5.3 . . . . .	538

<b>II</b>	<b>Regression-Based Time Series Models</b>	<b>553</b>
<b>Chapter 6</b>	<b>Fundamentals of Time Series Analysis</b>	<b>555</b>
6.1	Fundamental Components of a Time Series . . . . .	556
6.1.1	Modeling Trends in Time . . . . .	559
6.1.2	Modeling Seasonal Effects . . . . .	561
6.1.3	Some Qualitative Discussion . . . . .	566
	Practice Problems for Section 6.1 . . . . .	568
6.2	Two Primitive Time Series Models . . . . .	570
6.2.1	White Noise . . . . .	570
6.2.2	Random Walks . . . . .	576
6.2.3	Filtering to Achieve Stationarity . . . . .	583
6.2.4	Unit Root Test . . . . .	586
	Practice Problems for Section 6.2 . . . . .	590
<b>Chapter 7</b>	<b>Time Series Forecasting</b>	<b>603</b>
7.1	Smoothing . . . . .	603
7.1.1	Moving Averages . . . . .	604
7.1.2	Exponential Smoothing . . . . .	607
	Practice Problems for Section 7.1 . . . . .	615
7.2	Autoregressive Models . . . . .	626
7.2.1	Definition and Model Properties . . . . .	626
7.2.2	Parameter Estimation and Diagnostics . . . . .	632
7.2.3	Forecasting . . . . .	636
	Practice Problems for Section 7.2 . . . . .	643
7.3	Forecasting Volatility: ARCH and GARCH Models . . . . .	657
7.3.1	ARCH Models . . . . .	657
7.3.2	GARCH Models . . . . .	658
	Practice Problems for Section 7.3 . . . . .	661
7.4	Forecast Evaluation . . . . .	666
	Practice Problems for Section 7.4 . . . . .	669
<b>III</b>	<b>Statistical Learning Methods Beyond Regression</b>	<b>673</b>
<b>Chapter 8</b>	<b>Decision Trees</b>	<b>675</b>
8.1	Single Decision Trees . . . . .	676
8.1.1	What does a Decision Tree Look Like? . . . . .	676
8.1.2	Constructing and Pruning Decision Trees . . . . .	683
8.1.3	Classification Trees . . . . .	690
8.1.4	Decision Trees vs. Linear Regression Models . . . . .	701
	Practice Problems for Section 8.1 . . . . .	709
8.2	Ensemble Trees . . . . .	732
8.2.1	Bagging . . . . .	732
8.2.2	Random Forests . . . . .	741
8.2.3	Boosting . . . . .	743
	Practice Problems for Section 8.2 . . . . .	749
<b>Chapter 9</b>	<b>Principal Components Analysis</b>	<b>761</b>
9.1	PCA: Fundamental Ideas . . . . .	762

9.1.1	Interpretation 1: Maximal Variance Directions . . . . .	765
9.1.2	Interpretation 2: Closest Hyperplanes . . . . .	771
9.1.3	Additional PCA Issues . . . . .	773
	Practice Problems for Section 9.1 . . . . .	782
9.2	PCA: Applications to Supervised Learning . . . . .	796
9.2.1	Principal Components Regression . . . . .	796
9.2.2	Partial Least Squares . . . . .	799
	Practice Problems for Section 9.2 . . . . .	803
<b>Chapter 10 Cluster Analysis</b>		<b>807</b>
10.1	Introduction . . . . .	808
10.2	$K$ -means Clustering . . . . .	810
	Practice Problems for Section 10.2 . . . . .	820
10.3	Hierarchical Clustering . . . . .	827
10.3.1	Practical Issues in Clustering . . . . .	837
	Practice Problems for Section 10.3 . . . . .	840
<b>IV Practice Examinations</b>		<b>855</b>
<b>Practice Exam 1</b>		<b>859</b>
	Solutions to Practice Exam 1 . . . . .	880
<b>Practice Exam 2</b>		<b>891</b>
	Solutions to Practice Exam 2 . . . . .	910
<b>Practice Exam 3</b>		<b>923</b>
	Solutions to Practice Exam 3 . . . . .	940
<b>Practice Exam 4</b>		<b>951</b>
	Solutions to Practice Exam 4 . . . . .	968
<b>Practice Exam 5</b>		<b>981</b>
	Solutions to Practice Exam 5 . . . . .	1000
<b>Practice Exam 6</b>		<b>1011</b>
	Solutions to Practice Exam 6 . . . . .	1028
<b>Farewell Message</b>		<b>1037</b>

# Preface

## ⚠ NOTE TO STUDENTS ⚠

Please read this preface carefully 📖, even if it looks long. It contains **VERY** important information that will help you navigate this study manual smoothly ✂ and ease your learning.

## P.1 About Exam SRM

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

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

In January 2022, the ATPA (Advanced Topics in Predictive Analytics) Assessment was also added. This study manual not only prepares you adequately for Exam SRM, but also paves the way ⚠ for Exams PA and ATPA, which are largely a sequel to SRM. (See page xx for further discussion.)

### Exam Administrations

Exam SRM is a 3.5-hour computer-based exam consisting of 35 multiple-choice questions. In 2024 and 2025 (and likely thereafter), it will be delivered via computer-based testing (CBT) 🖥 in January, May, and September. You can find the specifics of each testing window (e.g., exam dates, registration deadlines) at

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

When the registration window is open, you will register online at

<https://www.soa.org/education/exam-req/registration/edu-registration/>,

receive an email confirmation letter ✉ from the SOA containing your candidate ID, then schedule an appointment at Prometric (<https://www.prometric.com/soa>).

## Exam Theme: What is SRM Like?

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

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

The latest syllabus of Exam SRM, available from

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

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

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

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

As a rough estimate, you will need about **THREE months** of intensive study  to master the material in this exam. There are A LOT to learn and absorb. (In fact, the real exam, even with 35 questions, will likely only test a small subset of the whole exam syllabus.) Don’t worry about

studying too hard—what you learn in Exam SRM will continue to be useful for Exams PA and ATPA.

## Exam Style

As of July 2024, the SOA has released a total of 72 sample questions,<sup>1</sup> which can be accessed from

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

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

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

In some exam questions (e.g., Sample Questions 1, 3, 4, 9, 11, 15, 17, 18, 19, 23, 24, 27, 28, 33, 35, 44, 45-48, 51, 54, 55, 57-59, 62, 63, 66, 67-70, 72), you will be asked to do some simple calculations  in one of two scenarios:

- Case 1.* You may be given a small dataset, e.g., one with not more than 10 observations. While almost all predictive analytic techniques in the exam syllabus require computers to implement, the small size of the dataset makes it possible to perform at least part of the analysis.
- Case 2.* You may also be given some summarized model output such as tables of parameter estimates and/or graphical output.  Then you are asked to perform some simple tasks like interpreting the results of the model, conducting a hypothesis test, making point/interval estimations/predictions, and assessing the goodness of fit of the model, all of which require only pen-and-paper calculations.

You may ask:

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

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

---

<sup>1</sup>The SOA deleted Questions 17, 28, 47, and 65 because they test concepts “no longer on the syllabus.” There have been no changes in the syllabus readings, which form the backbone of the exam, so it is unclear why these questions were deleted.

- [Important ] **Type 2: Conceptual/True-or-false questions (roughly 2/3 of the exam)**

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

*“there are a lot of qualitative questions [in Exam SRM].”*

## Statistics for Risk Modeling Exam

- It has been administered four times (35 multiple choice questions)
  - September 2018: 116/174 effective = 67% pass rate
  - January 2019: 166/264 effective = 63% pass rate
  - May 2019: 237/391 effective = 61% pass rate
  - September 2019: grades not yet released

- Thing to know:
  - There are a lot of qualitative questions.
  - Goal is to ensure candidates know the definitions, differences, similarities, and uses of the various techniques.



Sample Questions 2, 5, 6, 7, 8, 10, 12, 13, 14, 16, 20, 21, 22, 25, 26, 29-32, 34, 36-43, 49, 50, 52, 53, 56, 60-61, 64, 65, and 71 all belong to this type of questions. You are typically given three statements (I, II, and III) and asked to pick the correct one(s). The five answer choices often take a symmetric structure shown overleaf:



## Historical Pass Rates and Pass Marks %

The table overleaf shows the number of sitting candidates, number of passing candidates, pass rates, and pass marks (= the actual percentage score you have to get to pass the exam) for Exam SRM since it was offered in September 2018.

Sitting	# Candidates	# Passing Candidates	Pass Rate	Pass Mark
September 2024	(To be posted on the <a href="#">SOA webpage</a>  )			
May 2024	1664	1154	69.4%	63%
January 2024	1264	809	64.0%	63%
September 2023	1351	912	67.5%	60%
May 2023	1820	1352	74.3%	60%
January 2023	1151	847	73.6%	60%
September 2022	1241	939	75.7%	60%
May 2022	1004	768	76.5%	65%
January 2022	794	625	78.7%	65%
September 2021	810	592	73.1%	65%
May 2021	843	660	78.3%	65%
January 2021	754	566	75.1%	65%
September 2020	814	627	77.0%	63%
May 2020	278	205	73.7%	63%
January 2020	587	372	63.4%	67%
September 2019	554	411	74.2%	60%
May 2019	410	237	57.8%	67%
January 2019	274	166	60.6%	67%
September 2018	181	116	64.1%	70%

Perhaps to your astonishment, the pass rates of Exam SRM have been anomalously high, typically in the 60-75% range, compared to only 40-50% for typical ASA-level exams. The pass marks in the most recent sittings are 60-63%, which means that candidates need to get about **22 (out of 35)** questions<sup>2</sup> correct ✓ to earn a pass.

<sup>2</sup>According to the exam syllabus, one or two questions in the CBT environment may be pilot questions that are included to judge their effectiveness for future exams, but they will not be graded.

## Syllabus Texts

Exam SRM has two required textbooks:

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

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

On Internet forums (e.g., Reddit , Discord ) , many users recommend reading ISLR as an important way to prepare for SRM. While the book is available online and well-written in general, it is rather text-heavy (there are many long paragraphs!) and has hardly any exam-type problems. After all, it is not designed for exam preparation. With this self-contained study manual, reading ISLR is completely optional. Of course, if you still have time left after finishing this manual in its entirety, then there is no harm in taking a look at ISLR as a further way to consolidate your understanding.

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

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

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

## Exam Tables

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



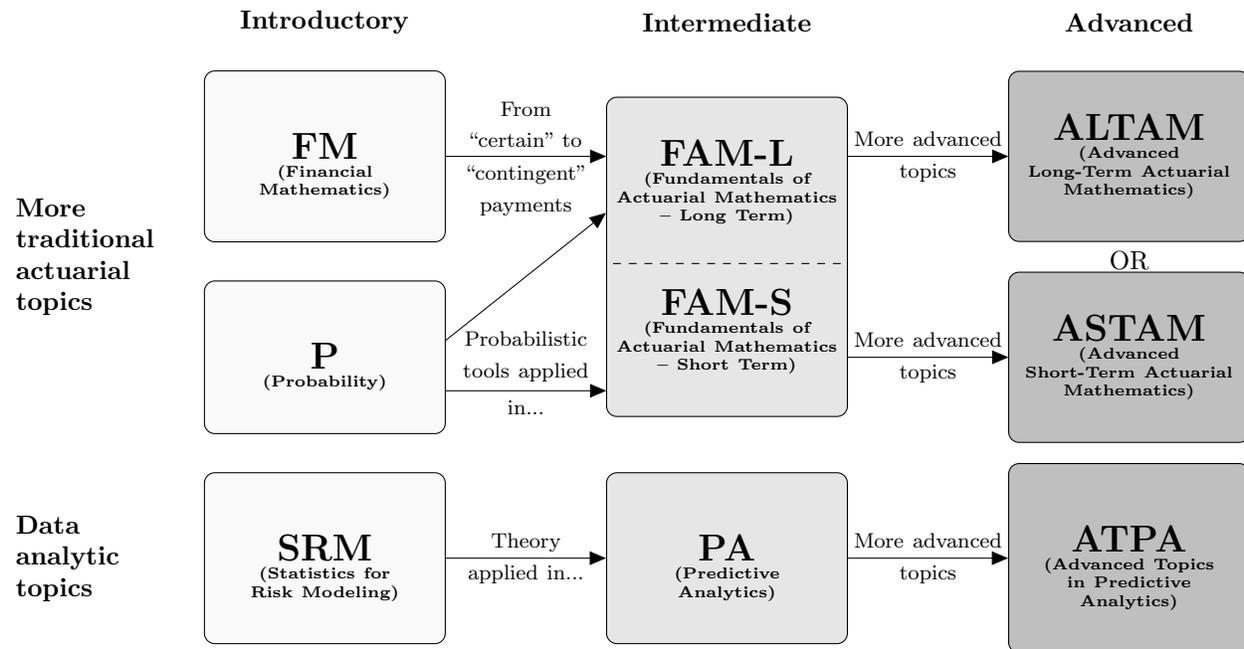
<https://www.soa.org/Files/Edu/2018/exam-srm-tables.pdf>

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

## Predictive Analytics Trio : SRM, PA, and ATPA

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

### Flowchart of ASA Exams Effective from 2022



Let's talk about how SRM, PA, and ATPA relate to one another.

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

In essence, Exams SRM and PA are about the same subject, but test it differently. While Exam SRM emphasizes the theory underlying different predictive analytic techniques, Exam PA will have you apply the theory you learned in Exam SRM to a concrete setting and see first hand how things play out. Some additional topics and practical considerations are also presented. After taking Exam PA, you will see the predictive models you learned in Exam SRM in action and gain a much more thorough understanding.

**PA vs. ATPA.** Exam ATPA is a 96-hour take-home computer-based assessment (rather than a proctored exam, so Exam ATPA is also called the “ATPA Assessment”). It tests additional data and modeling concepts on the basis of those in Exams SRM and PA, and consists of more involved and open-ended tasks than those in PA. As a result, ATPA is preferably taken after passing SRM and PA.

Although ATPA is a take-home assessment and 4 days seem a lot of time, you would be wise not to underestimate the amount of time and effort necessary to master the topics that can be tested, and the workload and pressure that the assessment can create. 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 doing coding 🖥️ and writing! 📝

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

<https://www.actexlearning.com/exams/pa>,

<https://www.actexlearning.com/exams/atpa/exam-atpa-study-manual>.

## P.2 About this Study Manual

### What is Special about This Study Manual?

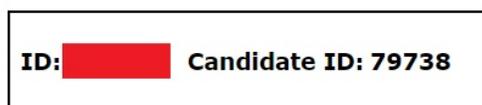
We fully understand that you have an acutely limited amount of study time and that the SRM exam syllabus is enormous. With this in mind, the overriding objective of this study manual is to help you grasp the material in Exam SRM as effectively and efficiently as possible, so that you will pass the exam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

### Feature 1: The Coach DID Play!

Usually coaches don't play 😊, but the main author of this manual (Ambrose) took the initiative to write the SRM exam in January 2019 to experience first-hand what the real exam was like, despite having been an FSA since 2013 (and technically free from exams thereafter!). He made this decision in the belief that *teaching* an exam and *taking* an exam are rather different activities, and braving the exam himself is the best way to ensure that this manual is indeed effective for exam preparation. If the manual is useful, then at the minimum the author himself can do well, right?

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

Today's Date: 12/16/2019



### Jan 2019 Statistics for Risk Modeling

Course	Grade
EXAMSRM	10

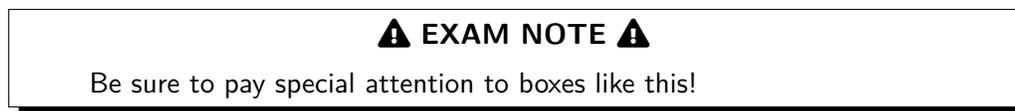
Ambrose Lo FSA, CERA  
Associate Professor  
University of Iowa  
241 Schaeffer Hall  
Iowa City, IA 52242-1409

If you use this SRM 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 SRM candidates and truly understands their needs. Drawing upon their "real battle experience" and firm grasp of the exam topics, the authors 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: Exam-focused Content

- (*Let's start with the syllabus!*) Each chapter starts by explicitly stating which learning objectives and outcomes of the SRM exam syllabus we are going to cover, to assure you that we are on track and hitting the right target.
- (*In-text explanations*) The explanations in each chapter are thorough, but exam-focused and learning-oriented. Besides having a coherent narrative flow that shows the connections ↔ between different exam topics, this manual strives to keep you motivated by showing how the concepts are typically tested, how the formulas are used, and where the exam focus lies in each section. Instead of showing unnecessary mathematical proofs that add little value to exam preparation, we grasp the chance to explain the intuitive meaning and mathematical structure of various formulas in the syllabus, to help you better remember and apply these formulas on the exam.
- (*In-text illustrative examples*) The main text of the manual is interspersed with carefully chosen past SOA/CAS exam questions, with full bibliographic details given (name of exam, year of examination, question number). Complementing the in-text explanations, these illustrative examples are meant to show you how the concepts you have just learned are usually tested and are an essential part of your reading.

- (*Boxed formulas and exam notes*) Formulas and results of utmost importance are boxed for easy identification and numbered (in the (X.X.X) format) for later references. Important exam items and common mistakes committed by students are highlighted by boxes that look like:



- (*End-of-section/chapter problems*) To succeed in any actuarial exam, we can't overemphasize the importance of practicing a wide variety of exam-type problems to sharpen your understanding and develop proficiency. This study manual embraces this learning by doing approach and features more than 400 end-of-section/chapter problems. Designed to reinforce what you have learned in the main text of the manual and provide additional opportunities, these practice problems are either additional problems taken/adapted from relevant SOA/CAS past exams, or are original problems intended to illustrate less commonly tested items, all with step-by-step solutions and many with problem-solving remarks. Many of the original problems are of the true-or-false type, which, according to students' comments, has figured prominently in recent SRM exams.

**⚠** To maximize the effectiveness and efficiency of your learning, we have marked the most representative and instructive practice problems in each section with an **asterisk (\*)**. These selected problems, which are generally not more than 50% of the whole set of problems, span different themes and will add most value to your learning. Here is the study approach we recommend:

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

For computational examples and problems, be sure to get a paper, do the math   , and try to “replicate” our solutions (we mean it!). For any actuarial exams, keep in mind that it is not enough to be able to do problems; you need to do problems *accurately* and *efficiently*.

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

- (*Optional syllabus readings*) Although this study manual is self-contained in the sense that studying the manual carefully is already sufficient to pass the exam, relevant chapters and sections of the two SRM syllabus texts  are referenced at the beginning of each chapter of the manual, for the benefit of students who like to read more. (Remember that ISLR is freely available online.)
- (*Practice exams*) Six (6) original full-length practice exams designed to mimic the real SRM exam in terms of coverage, style, and difficulty conclude this study manual. These practice exams give you a holistic review of the syllabus material and assess your readiness to take and pass the real exam. Detailed illustrative solutions are provided for each exam.

## Contact Us

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

- For **technical** issues  (e.g., not able to access your manual online, extending your digital license, upgrading your product, exercising the Pass Guarantee), please email ACTEX Learning's Customer Service at



`support@actexlearning.com`

The list of FAQs on <https://www.actuarialuniversity.com/help/faq> may also be useful.

- For questions related to **specific contents** of this manual and Exam SRM, including potential errors (typographical or otherwise), please feel free to raise them in the SRM community on [ACTEX's Discord channel](#), which provides a convenient platform for you to network with other SRM students, and we will strive to respond to your questions ASAP.



## Acknowledgments

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

## About the Authors

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

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

**Ambrose Lo**, PhD, FSA, CERA, is the author of several study manuals for professional actuarial examinations and an Adjunct Associate Professor at the Department of Statistics and Actuarial Science, the University of Hong Kong (HKU). He earned his BSc in Actuarial Science (first class honors) and PhD in Actuarial Science from HKU in 2010 and 2014, respectively, and attained his Fellowship of the Society of Actuaries (FSA) in 2013. He joined the Department of Statistics and Actuarial Science, the University of Iowa (UI) as Assistant Professor of Actuarial Science in August 2014, and was promoted to Associate Professor with tenure 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*. He left the UI and returned to Hong Kong in July 2023.

Besides dedicating himself to actuarial research, Ambrose attaches equal (if not more!) importance to teaching and education, through which he nurtures the next generation of actuaries and serves the actuarial profession. He has taught courses on a wide range of actuarial science topics, such as financial derivatives, mathematics of finance, life contingencies, and statistics for risk modeling. He is also the (co)author of the *ACTEX Study Manuals for Exams ATPA, 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 logical reasoning is always his top priority. In recognition of his outstanding 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, HKU, public recognition in the *Daily Iowan* as a faculty member “making a positive difference in students’ lives during their time at UI” for nine years in a row (2016 to 2024), and the 2019-2020 Collegiate Teaching Award from the UI College of Liberal Arts and Sciences.



## Chapter 2

# Multiple Linear Regression

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

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

#### **Learning Objectives**

The Candidate will understand key concepts concerning generalized linear models.

#### **Learning Outcomes**

The Candidate will be able to:

- a) Compare model assumptions for ordinary least squares and generalized linear models.
- c) Apply the business context of a problem to interpret parameters.
- e) Select an appropriate model, including:
  - Variable transformations and interactions
  - t and F tests
- f) Calculate and interpret predicted values, and confidence and prediction intervals.

#### **OPTIONAL SRM SYLLABUS READINGS**

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

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

- (i) Are certain predictors useful for explaining the variability of the response variable?
- (ii) How should we form the regression function? Is there any interaction between some of the predictors in explaining the response variable?
- (iii) Given several competing MLR models, which one is the best? Under what criterion?

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

**⚠ EXAM NOTE ⚠**

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

## 2.1 From SLR to MLR: Fundamental Results

### OPTIONAL SRM SYLLABUS READINGS

- Frees, Sections 3.1 to Subsection 3.4.2, Subsection 5.5.4, and Section 6.1
- [ISLR, Section 3.2](#)

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

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

where:

- $p$  is the number of predictors ( $p = 1$  in SLR).  
(**Note:** Frees denotes the number of predictors by  $k$  while ISLR uses  $p$ . In this study manual, we follow ISLR's usage, which is more popular in the statistical learning community. In most cases,  $k$  and  $p$  can be used interchangeably. The only exception is Section 4.3 of this manual.)
- $\beta_0$  is the [intercept](#).
- $\beta_j$  is the [regression coefficient](#) attached to the  $j$ th predictor, for  $j = 1, \dots, p$ .
- $\varepsilon$  is again the [random error term](#).

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

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

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

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

Table 2.1: Typical data structure of an MLR model.

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (2.1.1)$$

or, on a component-wise basis,

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

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

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

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

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

where the prime “ ’ ” denotes the transpose of a matrix, over  $\beta$ . By matrix calculus, we solve the normal equations  $\frac{\partial}{\partial \beta}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}$  ( $\mathbf{0}$  is a vector of zeros) to get the **least squares estimator (LSE)** of  $\beta$ . The LSE need not be unique, but usually<sup>1</sup> it is, in which case it takes the following vector form:

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

Two points about this formidable vector formula deserve attention:

- Unlike the case of SLR, where we have closed-form algebraic formulas for  $\hat{\beta}_0$  and  $\hat{\beta}_1$  (recall (1.1.4)), (2.1.2) is all we have for the LSEs in an MLR model. In other words, we have a (magnificent!) formula for the entire vector of LSEs, but not separate algebraic formulas for the individual LSEs.
- To apply (2.1.2), we have to invert the  $(p+1) \times (p+1)$  matrix  $\mathbf{X}'\mathbf{X}$ , which is hard to perform by pen-and-paper calculations unless  $p = 1$  (i.e., SLR).<sup>2</sup>

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

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

<sup>1</sup>The LSE is unique as long as the matrix  $\mathbf{X}'\mathbf{X}$  is invertible.

<sup>2</sup>(If you are interested) When  $p = 1$ ,

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

and (2.1.2) reduces to

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

which is (1.1.4).



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

**Example 2.1.2.**  (SOA Part 4 May 1983 Question 10: No-intercept MLR model)

Advertising expenditures and sales for the last 5 quarters have been as follows:

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

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

The following model is established:

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

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

Find the least squares estimate of  $\beta_1$ .

- (A) 25/14                      (B) 29/16                      (C) 33/16  
 (D) 29/14                      (E) 33/14

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

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

The LSE of  $\beta_1$  is the first component of

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

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

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

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

You are given:

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

The least squares estimator of  $\beta_2$  is expressed as  $\hat{\beta}_2 = \sum_{i=1}^4 w_i y_i$ .

Determine  $(w_1, w_2, w_3, w_4)$ .

- (A)  $(-0.15, -0.05, 0.05, 0.15)$                       (B)  $(-0.05, 0.15, -0.15, 0.05)$   
 (C)  $(-0.05, 0.05, -0.15, 0.15)$                       (D)  $(-0.3, -0.1, 0.1, 0.3)$   
 (E)  $(-0.1, 0.3, -0.3, 0.1)$

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

thus  $(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 1/20 & 0 \\ 0 & 0 & 1/20 \end{pmatrix}$ . Then

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





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

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

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

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

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

and are given the following values used in this model:

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

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

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

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



which is the ratio of the *percentage* change in  $y$  to the *percentage* change in  $x$ . This is known as *elasticity*, which is a concept emanating from economics.<sup>3</sup>

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

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

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

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

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

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

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

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

Under  $\text{H}_0$ , the F-statistic<sup>4</sup> is expected to take a value close to 1.<sup>5</sup> Under  $\text{H}_a$ , the F-statistic is expected to be greater than 1. Note that the proper interpretation of the result of the F-test in the MLR setting is:

If  $\text{H}_0$  is rejected, then we have strong evidence that *at least one* of the  $p$  predictors is an important predictor for the response variable. However, we do *not know which of these predictors are really useful!*

<sup>3</sup>You probably have seen the concept of option elasticity in Exam IFM.

<sup>4</sup>If you are interested in knowing, the F-statistic follows an  $F_{p, n-(p+1)}$  distribution under  $\text{H}_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$ .

<sup>5</sup>Page 76 of ISLR provides a heuristic justification of why the F-statistic is expected to be close to 1 under  $\text{H}_0$  and much larger than 1 under  $\text{H}_a$ . The reason is that we always have  $\mathbb{E}[\text{RSS}/(n - p - 1)] = \sigma^2$ , no matter whether  $\text{H}_0$  or  $\text{H}_a$  is true, and  $\mathbb{E}[\text{Reg SS}/p] \begin{cases} = \sigma^2, & \text{under } \text{H}_0 \\ > \sigma^2, & \text{under } \text{H}_a \end{cases}$ .



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

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

and measures the proportion of the variation of the response variable (about its mean) that can be explained by the MLR model. In the MLR framework,  $R^2$  is no longer the square of the sample correlation between  $y$  and individual predictors (we don't have a single  $x$  anymore!); however, it continues to be the *square* of the sample correlation between the observed  $y$  and the fitted value  $\hat{y}$ ,<sup>6</sup> i.e.,

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

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

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

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

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

- (A) There is insufficient information, but it could be calculated if she had the value of the residual sum of squares (RSS).
- (B) There is insufficient information, but it could be calculated if she had the value of the total sum of squares (TSS) and RSS.
- (C)  $R^2 = 0.44$
- (D)  $R^2 = 0.56$
- (E)  $R^2 = 0.80$

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

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

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

<sup>6</sup>The fact that  $R^2 = \text{Corr}(y, \hat{y})^2$  is also true for SLR models. After all, SLR is a special case of MLR.

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

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

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

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

- (A) At least one of the four predictors is related to the response variable.
- (B) All of the four predictors are related to the response variable.
- (C) At least one of the four predictors is not related to the response variable.
- (D) All of the four predictors are not related to the response variable.
- (E) None of (A), (B), (C), or (D) are true.

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

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

Calculate the value of the F-statistic used to test the hypothesis  $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ .

- (A) Less than 5
- (B) At least 5, but less than 10
- (C) At least 10, but less than 15
- (D) At least 15, but less than 20
- (E) At least 20

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

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

□



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



$$\hat{\beta} \sim N_{p+1} \left( \beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \right),$$

multivariate normal dist.
mean vector
variance-covariance matrix

where:

- $N_{p+1}$  denotes a  $(p + 1)$ -dimensional multivariate normal distribution.
- The *mean vector* of the multivariate normal distribution is the parameter vector  $\beta$ , i.e.,  $\mathbb{E}[\hat{\beta}] = \beta$ ,<sup>7</sup> or component-wise,  $\mathbb{E}[\hat{\beta}_j] = \beta_j$  for all  $j = 0, 1, \dots, p$ . In the language of mathematical statistics, we say that  $\hat{\beta}$  is an unbiased estimator of  $\beta$ .
- The *variance-covariance matrix* hosting the variances of and covariances between the LSEs is

$$\text{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \quad (2.1.4)$$

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

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

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

$$\widehat{\text{Var}}(\hat{\beta}) = s^2(\mathbf{X}'\mathbf{X})^{-1}. \quad (2.1.5)$$

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

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

<sup>7</sup>Here is a simple proof: Because expectation is linear and the design matrix  $\mathbf{X}$  consists of non-random elements, we can take it outside the expectation operator, leading to

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

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

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

where  $\mathbf{Z}$  is a random vector, and  $\mathbf{A}$  is a non-random matrix such that the matrix product  $\mathbf{AZ}$  is well-defined. This result is the multidimensional generalization of the familiar univariate result  $\text{Var}(aZ) = a^2\text{Var}(Z) = a \cdot \text{Var}(Z) \cdot a$  for any random variable  $Z$  and any real scalar  $a$ . Applying (2.1.6) with  $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  and  $\mathbf{Z} = \mathbf{y}$ , we get<sup>8</sup>

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

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

The above distributional results cast some light on the optimality of the LSEs. It can be shown that within the class of *linear unbiased* estimators (i.e., estimators that are unbiased and can be expressed as linear combinations of the response values), LSEs are the *minimum variance*<sup>9</sup> *unbiased estimator* of the parameter vector  $\boldsymbol{\beta}$ . This result is known as the *Gauss–Markov theorem*, which is true even when the random errors are not normally distributed.

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

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

You are given:

- (i)  $z_i = x_i^2$ , for  $i = 1, 2, \dots, n$
- (ii)  $b_1^* = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})(x_i - \bar{x})}$

Which of the following is true?

- (A)  $b_1^*$  is a non-linear estimator of  $\beta_1$ .
- (B)  $b_1^*$  is a heteroscedasticity-consistent estimator (HCE) of  $\beta_1$ .
- (C)  $b_1^*$  is a linear biased estimator of  $\beta_1$ .
- (D)  $b_1^*$  is a linear unbiased estimator of  $\beta_1$ , but not the best linear unbiased estimator (BLUE) of  $\beta_1$ .

<sup>8</sup>Recall from your linear algebra class that  $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$  for any two matrices  $\mathbf{A}$  and  $\mathbf{B}$  such that  $\mathbf{AB}$  is well-defined.

<sup>9</sup>(If you are interested) In a multivariate framework, the fact that  $\hat{\boldsymbol{\beta}}$  has the “minimum variance” means that if  $\hat{\boldsymbol{\beta}}'$  is another estimator, then the difference of the two variance-covariance matrices,  $\text{Var}(\hat{\boldsymbol{\beta}}') - \text{Var}(\hat{\boldsymbol{\beta}})$ , is a non-negative definite matrix.

(E)  $b_1^*$  is the best linear unbiased estimator (BLUE) of  $\beta_1$ .

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

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

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

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

However,  $b_1^*$  is not the LSE  $\hat{\beta}_1 = S_{xy}/S_{xx}$ , so  $b_1^*$  is not the best linear unbiased estimator of  $\beta_1$ .  
(Answer: (D))  $\square$

*Remark.* Heteroscedasticity-consistent estimators are concerned with estimating variances when the random errors have unequal variances; see page 243 for details.

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

(i)  $n = 15$

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

(iii)  $\text{RSS} = 282.82$

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

(A) 6.4

(B) 6.8

(C) 7.1

(D) 7.5

(E) 7.8

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

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

$$\widehat{\text{Var}}(\hat{\beta}_2 - \hat{\beta}_1) = \widehat{\text{Var}}(\hat{\beta}_2) + \widehat{\text{Var}}(\hat{\beta}_1) - 2\widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2).$$

(not -!)

Using (2.1.5) and extracting the highlighted entries

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{cccc} & 0 & 1 & 2 & 3 \\ 0 & 13.66 & -0.33 & 2.05 & -6.31 \\ 1 & -0.33 & \boxed{0.03} & \boxed{0.11} & 0.00 \\ 2 & 2.05 & \boxed{0.11} & \boxed{2.14} & -2.52 \\ 3 & -6.31 & 0.00 & -2.52 & 4.32 \end{array},$$

we get the estimated variances and covariances we need:

$$\widehat{\text{Var}}(\hat{\beta}_2) = s^2(2.14), \quad \widehat{\text{Var}}(\hat{\beta}_1) = s^2(0.03), \quad \widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2) = s^2(0.11).$$

To find the MSE, we use (i) and (iii) to get  $s^2 = 282.82/(15 - 4) = 25.710909$ , so

$$\widehat{\text{Var}}(\hat{\beta}_2 - \hat{\beta}_1) = 25.710909[2.14 + 0.03 - 2(0.11)] = 50.1363.$$

Finally, the standard error of  $\hat{\beta}_2 - \hat{\beta}_1$  is the square root of the estimated variance, or  $\sqrt{50.1363} = \boxed{7.0807}$ . (**Answer: (C)**)  $\square$

*Remark.* Failure to take into account the estimated covariance between  $\hat{\beta}_1$  and  $\hat{\beta}_2$  leads to Answer (D), which is incorrect.

**Hypothesis tests and confidence intervals for regression coefficients.** The construction of hypothesis tests and confidence intervals for a *single* regression coefficient is in principle no different than the SLR case, except that the  $t_{n-2}$  distribution and  $t_{n-2, \alpha/2}$  upper quantile are replaced by  $t_{n-p-1}$  and  $t_{n-p-1, \alpha/2}$ , respectively, due to the presence of  $p$  predictors. Specifically:

$H_0 : \beta_j = d$ . The **t-statistic** for testing the null hypothesis  $H_0 : \beta_j = d$ , where  $d$  is a user-specified value, takes the familiar form

$$t(\hat{\beta}_j) = \frac{\hat{\beta}_j - d}{\text{SE}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - d}{\sqrt{s^2 \times (j+1)\text{th diagonal element of } (\mathbf{X}'\mathbf{X})^{-1}}}.$$

This hypothesis allows us to examine the individual effects of the  $j$ th predictor. Under the null hypothesis, the t-statistic follows a  $t_{n-p-1}$  distribution. With this fact, Table 1.3, which presents the decision-making procedure for the t-test in the SLR framework, can be revised as in Table 2.2.

$H_a$	Reject $H_0$ in favor of $H_a$ if...	p-value ( $t$ is the observed value of $t(\hat{\beta}_j)$ )
$\beta_j \neq d$	$ t(\hat{\beta}_j)  > t_{n-p-1, \alpha/2}$	$\mathbb{P}( t_{n-p-1}  >  t ) = 2\mathbb{P}(t_{n-p-1} >  t )$
$\beta_j > d$	$t(\hat{\beta}_j) > t_{n-p-1, \alpha}$	$\mathbb{P}(t_{n-p-1} > t)$
$\beta_j < d$	$t(\hat{\beta}_j) < -t_{n-p-1, \alpha}$	$\mathbb{P}(t_{n-p-1} < t)$

Table 2.2: Decision-making procedures for testing  $H_0 : \beta_j = d$  against various alternative hypotheses by means of a t-test in the multiple linear regression setting.

- C.I. Based on the fact that  $(\hat{\beta}_j - \beta_j)/SE(\hat{\beta}_j) \sim t_{n-p-1}$ , a  $100(1 - \alpha)\%$  confidence interval for  $\beta_j$  is given by

$$\begin{aligned} & \hat{\beta}_j \pm t_{n-p-1, \alpha/2} \times SE(\hat{\beta}_j) \\ = & \boxed{\hat{\beta}_j \pm t_{n-p-1, \alpha/2} \times \sqrt{s^2 \times (j+1)\text{th diagonal element of } (\mathbf{X}'\mathbf{X})^{-1}}. \end{aligned}$$

This interval provides a range of reliability for the true value of  $\beta_j$ .

**Example 2.1.14.** •• (SOA Exam SRM Sample Question 27: Dropping insignificant variables) Trevor is modeling monthly incurred dental claims. Trevor has 48 monthly claims observations and three potential predictors:

- Number of weekdays in the month
- Number of weekend days in the month
- Average number of insured members during the month

Trevor obtained the following results from a linear regression:

	Coefficient	Standard Error	t Stat	p-value
Intercept	-45,765,767.76	20,441,816.55	-2.24	0.0303
Number of weekdays	513,280.76	233,143.23	2.20	0.0330
Number of weekend days	280,148.46	483,001.55	0.58	0.5649
Average number of members	38.64	6.42	6.01	0.0000

Determine which of the following variables should be dropped, using a 5% significance level.

- I. Intercept
  - II. Number of weekdays
  - III. Number of weekend days
  - IV. Number of members
- (A) I only                      (B) II only                      (C) III only  
 (D) IV only                      (E) None should be dropped from the model

*Solution.* At the 5% significance level, a variable should be dropped (i.e., the corresponding  $\beta_j$  is zero) if its associated p-value for testing  $H_0 : \beta_j = 0$  against  $H_a : \beta_j \neq 0$  is larger than 5%. Only the number of weekend days satisfies this criterion, so it is the only variable to be dropped. (**Answer: (C)**) □



p-value), but  $x_j$  does not provide much additional explanatory power *when other predictors have been included in the model*. If this happens, it is often the case that  $x_j$  is *correlated* with predictors that are actually related to the response variable. In this sense,  $x_j$  is serving as a *surrogate* for those predictors and inherits their explanatory power.

It is also possible, but less common, that predictors, taken individually, have minimal effect on the response, but become highly significant when taken collectively. These variables, which increase the importance of other predictors, are called *suppressor variables* by Frees.

- *Individual t-tests vs. whole model F-test*: We now have two ways to assess the importance of the set of predictors:

- (1) Testing each of the following  $p$  null hypotheses separately by means of the t-test:

$$H_0^{(1)} : \beta_1 = 0, \quad H_0^{(2)} : \beta_2 = 0, \quad \dots, \quad H_0^{(p)} : \beta_p = 0.$$

- (2) Testing the following “global” null hypothesis by means of the overall F-test:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0.$$

Which way do we prefer? Suppose that the true model is the i.i.d. model  $y = \beta_0 + \varepsilon$ . That is,  $H_0, H_0^{(1)}, \dots, H_0^{(p)}$  are all correct. For a fixed significance level, say  $\alpha = 0.05$ , each of  $H_0$  and the  $H_0^{(j)}$ 's will be erroneously rejected with a probability of 5%. For  $p = 100$ , we expect to see a significant result in about 5 of the 100  $H_0^{(j)}$ 's. If the results in the 100 hypothesis tests *are* independent, then the probability of correctly accepting all of the 100  $H_0^{(j)}$ 's is  $0.95^{100} \approx 0$  (i.e., we are almost guaranteed to conclude that some predictors are related to the response when they are not), but the probability of correctly accepting  $H_0$  is 0.95. The F-test has the substantial advantage of taking into account the number of predictors, regardless of which the test only has a 5% significance level.

**Example 2.1.16.**  (Based on ISLR’s Advertising dataset) For four quantitative variables, sales, TV, radio, and newspaper, you are given:

- The coefficient estimates for the simple linear regression model of sales on newspaper:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	12.3514	0.6214	19.9	<2e-16 ***
newspaper	0.0547	0.0166	3.3	0.0011 **

- The coefficient estimates for the multiple linear regression model of sales on TV, radio, and newspaper:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.93889	0.31191	9.42	<2e-16 ***
TV	0.04576	0.00139	32.81	<2e-16 ***
radio	0.18853	0.00861	21.89	<2e-16 ***
newspaper	-0.00104	0.00587	-0.18	0.86

(iii) The correlation matrix for the four variables:

	TV	radio	newspaper	sales
TV	1.000	0.055	0.057	0.78
radio	0.055	1.000	0.354	0.58
newspaper	0.057	0.354	1.000	0.23
sales	0.782	0.576	0.228	1.00

**Comments:** As the SRM exam syllabus says, an exam question may provide you with some simple R output and expect you to make use of the given output to perform further analysis. Like in this example, the output should be self-explanatory and no knowledge of R programming is needed.

(a) Determine which of the following statements is/are true.

- I. Newspaper is statistically significant for `sales` when `TV` and `radio` are ignored.
- II. Newspaper is statistically insignificant for `sales` when `TV` and `radio` are ignored.
- III. Newspaper is statistically significant for `sales` in the presence of `TV` and `radio`.
- IV. Newspaper is statistically insignificant for `sales` in the presence of `TV` and `radio`.

- (A) I and II only
- (B) I and III only
- (C) I and IV only
- (D) II and III only
- (E) II and IV only

*Solution.* • In the SLR model of `sales` on `newspaper` (where `TV` and `radio` are ignored), `newspaper` is highly significant.

- In the MLR model of `sales` on `newspaper` (with `TV` and `radio` present), `newspaper` is highly insignificant.

Therefore, only I and IV are true. (**Answer: (C)**) □

(b) [**HARDER!**] Determine which of the following best explains the change in the statistical significance of `newspaper` in the two models.

- (A) `TV` is strongly correlated with `sales`.
- (B) `Radio` is strongly correlated with `sales`.
- (C) `Newspaper` is weakly correlated with `sales`.
- (D) `TV` is weakly correlated with `newspaper`.
- (E) `Radio` is moderately correlated with `newspaper`.

*Solution.* All of the five statements are true, but the one that best explains the change in the statistical significance of `newspaper` in the two models is (E), the moderately high correlation (+0.35) between `radio` and `newspaper`. This means that firms that spend more on radio advertising also tend to spend more on newspaper advertising. Symbolically:

$$\text{radio} \uparrow \quad \begin{array}{c} \text{tendency} \\ \Leftrightarrow \end{array} \quad \text{newspaper} \uparrow$$

On the other hand, if **radio** is predictive of **sales** positively (implied by the MLR model and the high correlation between **radio** and **sales**), then with a higher expenditure on radio advertising, **sales** also tends to be higher:

$$\text{radio} \uparrow \quad \begin{array}{c} \text{tendency} \\ \Leftrightarrow \end{array} \quad \text{sales} \uparrow$$

Combining the two implications, we see that there is a tendency that higher values of **sales** go hand in hand with higher values of **newspaper**:

$$\text{sales} \uparrow \quad \begin{array}{c} \text{tendency} \\ \Leftrightarrow \end{array} \quad \text{newspaper} \uparrow$$

This is suggested by the SLR model, which analyzes **sales** in relation to **newspaper** alone and ignores other information such as **radio**. In this case, we can say that **newspaper** serves as a surrogate for **radio** and gets “credit” for some of the predictive power of **radio**, which is truly related to **sales**. (**Answer: (E)**)  $\square$

- **Prediction.** **Prediction** for future, not-yet-realized values of the response variable parallels what we did in Section 1.4. Given the past observations in the form of  $\{(y_i; x_{i1}, x_{i2}, \dots, x_{ip})\}_{i=1}^n$ , we are interested in predicting a new response value  $y_*$  following the same MLR model as that used in the past data, but realized at a given set of predictor variable values, say  $\mathbf{x}_* = (1, x_{*1}, x_{*2}, \dots, x_{*p})'$ :

$$y_* = \beta_0 + \beta_1 x_{*1} + \dots + \beta_p x_{*p} + \varepsilon_*.$$

The figure below visualizes the setting.

	Explanatory variables				Response variable
	$\underline{x}_1$	$\underline{x}_2$	$\cdots$	$\underline{x}_p$	$\underline{y}$
Observed (past) data	$x_{11}$	$x_{12}$	$\cdots$	$x_{1p}$	$y_1$
	$x_{21}$	$x_{22}$	$\cdots$	$x_{2p}$	$y_2$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	$x_{n1}$	$x_{n2}$	$\cdots$	$x_{np}$	$y_n$
Future observation	$x_{*1}$	$x_{*2}$	$\cdots$	$x_{*p}$	$y_*$
	(known)				(unknown target)

Given the LSEs  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ , a natural **point predictor of  $y_*$**  is

$$\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_{*1} + \dots + \hat{\beta}_p x_{*p} = \mathbf{x}'_* \hat{\boldsymbol{\beta}}.$$

If we examine the distribution of the prediction error given by  $y_* - \hat{y}_* = \varepsilon_* + \mathbf{x}'_*(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$  and use the new t-quantile  $t_{n-p-1, \alpha/2}$  in place of  $t_{n-2, \alpha/2}$ , then we can show that a **100(1 -  $\alpha$ )% prediction interval for  $y_*$**  is

$$\hat{y}_* \pm t_{n-p-1, \alpha/2} \times \text{SE}(y_* - \hat{y}_*) = \hat{y}_* \pm t_{n-p-1, \alpha/2} \sqrt{s^2 [1 + \mathbf{x}'_*(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_*]}. \quad (2.1.7)$$

Here are two remarks about (2.1.7).

1. The standard error of prediction, which involves the matrix product  $\mathbf{x}'_*(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_*$ , looks intimidating! However, even in the MLR framework, the standard error of prediction obeys the same structure as that in SLR, corresponding to the two sources of uncertainty associated with prediction, namely,

$$\widehat{\text{Var}}(y_* - \hat{y}_*) = \underbrace{\widehat{\text{Var}}(y_*)}_{\text{estimated variance of new response}} + \underbrace{\widehat{\text{Var}}(\hat{y}_*)}_{\text{estimated variance of point predictor}}. \quad (2.1.8)$$

While  $\widehat{\text{Var}}(y_*) = s^2$  remains true, the estimated variance of the point predictor can no longer be simplified into an explicit algebraic expression as in (1.4.1) or (1.4.3) for SLR. The use of vectors and matrices is inevitable here. To determine  $\widehat{\text{Var}}(\hat{y}_*)$  explicitly, we apply (2.1.6) with  $\mathbf{A} = \mathbf{x}'_*$  and  $\mathbf{Z} = \hat{\boldsymbol{\beta}}$  to get

$$\text{Var}(\hat{y}_*) = \text{Var}(\mathbf{x}'_* \hat{\boldsymbol{\beta}}) = \mathbf{x}'_* \underbrace{\text{Var}(\hat{\boldsymbol{\beta}})}_{\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}} (\mathbf{x}_*)' = \sigma^2 \mathbf{x}'_*(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_*.$$

Finally, we replace  $\sigma^2$  by the MSE  $s^2$  to get

$$\text{SE}(y_* - \hat{y}_*) = \sqrt{\widehat{\text{Var}}(y_* - \hat{y}_*)} = \sqrt{s^2 [1 + \mathbf{x}'_*(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_*]}.$$

2. Although a simple algebraic expression for the standard error of prediction is not available in the MLR setting, it can be shown that the standard error increases if the predictor values are further away from the means of the predictors. This is consistent with SLR, where the standard error of prediction, given by (1.4.1), increases as  $x_*$  is further away from  $\bar{x}$ .

### ⚠ EXAM NOTE ⚠

The standard error of prediction involves the matrix product  $\mathbf{x}'_*(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_*$ . If (2.1.7) is tested, you will probably be given a simple matrix (e.g., a diagonal matrix) so that the matrix multiplications will be relatively easy, or you are directly given the value of  $\mathbf{x}'_*(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_*$ .



## Practice Problems for Section 2.1

—Highly recommended problems are marked with an asterisk (\*).—

**Problem 2.1.1.**  (SOA Course 120 November 1988 Exam Question 8: Calculation of LSEs given  $(\mathbf{X}'\mathbf{X})^{-1} - \mathbf{I}$ ) You are given the following weekly data on profits, food sales and non-food sales for a supermarket:

Profit (\$000)	Food Sales (\$000)	Non-Food Sales (\$000)
$y$	$x_1$	$x_2$
2	1	1
3	1	1
4	2	2
6	3	2
10	3	4

You are to represent these data by the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where  $\varepsilon$  is a random error term with mean 0 and variance  $\sigma^2$ .

You have determined that

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 1.20 & -0.50 & 0.00 \\ -0.50 & 0.75 & -0.50 \\ 0.00 & -0.50 & 0.50 \end{pmatrix}$$

Calculate  $\hat{\beta}_1$ .

- (A) -2.00                      (B) -0.75                      (C) -0.50  
 (D) 0.50                      (E) 0.75

*Solution.* By (2.1.2), the vector of LSEs is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \begin{pmatrix} 1.20 & -0.50 & 0.00 \\ -0.50 & 0.75 & -0.50 \\ 0.00 & -0.50 & 0.50 \end{pmatrix} \begin{pmatrix} 25 \\ 61 \\ 65 \end{pmatrix} = \begin{pmatrix} * \\ \boxed{0.75} \\ * \end{pmatrix}. \quad (\text{Answer: (E)})$$

□

**Problem 2.1.2.**  (SOA Course 120 May 1988 Question 10: Calculation of LSEs given  $(\mathbf{X}'\mathbf{X})^{-1}$  – II) You are given the following data:

$y$	$x_1$	$x_2$
1	4	1
2	5	1
4	5	2
4	6	3
5	7	3

You are to represent the data by the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

where  $\varepsilon$  is a random error term with mean 0 and variance  $\sigma^2$ .

You have determined:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 87.9 & -18.7 & -32.8 & 6.8 \\ -18.7 & 4.1 & 6.4 & -1.4 \\ -32.8 & 6.4 & 17.1 & -3.1 \\ 6.8 & -1.4 & -3.1 & 0.6 \end{pmatrix}.$$

Calculate  $\hat{\beta}_3$ .

- (A)  $-2.4$                       (B)  $-1.8$                       (C)  $-1.2$   
 (D)  $-0.6$                       (E)  $0.0$

*Solution.* Similar to the preceding problem,

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \\ &= \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ 6.8 & -1.4 & -3.1 & 0.6 \end{pmatrix} \begin{pmatrix} 16 \\ 93 \\ 38 \\ 231 \end{pmatrix} = \begin{bmatrix} * \\ * \\ * \\ \boxed{-0.6} \end{bmatrix}. \quad \text{(Answer: (D))} \end{aligned}$$

□

**Problem 2.1.3.**  (SOA Course 120 November 1989 Question 15: Parameter estimate for SLR vs. MLR) You have collected the following data on the response of a variable,  $y$ , to two other variables:

$y$	$x_1$	$x_2$
6.70	-5	-1
8.08	-3	1
20.09	-1	-1
18.09	1	1
30.00	3	-1
30.85	5	1

You have determined:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.167 & 0.0 & 0.0 \\ 0.0 & 0.016 & -0.016 \\ 0.0 & -0.016 & 0.182 \end{pmatrix}$$

$\hat{\beta}_1$  is the estimate of the coefficient of  $x_1$  in the multiple regression:

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

$\hat{b}_1$  is the estimate of the coefficient of  $x_1$  in the simple regression:

$$y = b_0 + b_1 x_{i1} + \varepsilon_i.$$

Determine  $\hat{\beta}_1 - \hat{b}_1$ .

- (A) -0.4                      (B) -0.2                      (C) 0.0  
 (D) 0.1                        (E) 0.3

*Solution.* • MLR: By (2.1.2),

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} * & * & * \\ 0.0 & 0.016 & -0.016 \\ * & * & * \end{pmatrix} \begin{pmatrix} * \\ 184.51 \\ 0.23 \end{pmatrix} = \begin{pmatrix} * \\ 2.94848 \\ * \end{pmatrix}.$$

- SLR: Using a financial calculator, we also find  $\hat{b}_1 = 2.635857$ .

The required difference is  $\hat{\beta}_1 - \hat{b}_1 = \boxed{0.31}$ . (**Answer: (E)**) □

**Problem 2.1.4.**  (SOA Course 120 May 1990 Question 1: Diagonal  $\mathbf{X}'\mathbf{X} - \mathbf{I}$ ) You are given the following model:

$$\mathbb{E}[y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

You have collected the following data:

$y$	$x_1$	$x_2$
6.3	-3	-1
7.5	-1	-1
8.1	1	-1
10.8	3	-1
4.7	-3	1
6.8	-1	1
8.3	1	1
7.0	3	1

Determine  $\hat{\beta}_2$ .

- (A) -1.5                      (B) -0.7                      (C) 0.6  
 (D) 1.5                        (E) 2.1

*Solution.* Because  $\sum_{i=1}^8 x_1 = \sum_{i=1}^8 x_2 = \sum_{i=1}^8 x_1 x_2 = 0$ , we have

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 \end{pmatrix} = \begin{pmatrix} 8 & 0 & 0 \\ 0 & 40 & 0 \\ 0 & 0 & 8 \end{pmatrix}.$$

The LSE of  $\beta_2$  can be found from

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} 1/8 & 0 & 0 \\ 0 & 1/40 & 0 \\ 0 & 0 & 1/8 \end{pmatrix} \begin{pmatrix} * \\ * \\ -5.9 \end{pmatrix} = \begin{pmatrix} * \\ * \\ \boxed{-0.7375} \end{pmatrix}. \quad \text{(Answer: (B))}$$

□



**Problem 2.1.6.**  (SOA Part 2 November 1981 Exam Question 18: No-intercept MLR model) Consider the regression model:

$$y_i = \beta_1 x_i + \beta_2 w_i + \varepsilon_i, \quad i = 1, \dots, n,$$

and  $(x_1, w_1), \dots, (x_n, w_n)$  are fixed constants. Let the independent random variables  $\varepsilon_1, \dots, \varepsilon_n$  have mean 0 and common variance  $\sigma^2$ . What is the least squares estimate of  $\beta_1$ , if it is given that

$$\sum_{i=1}^n x_i^2 = 1 \quad \text{and} \quad \sum_{i=1}^n w_i^2 = 1?$$

- (A)  $\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$       (B)  $\frac{\sum x_i y_i}{\sum x_i^2}$   
 (C)  $\frac{\sum y_i - \beta_2 \sum w_i}{\sum x_i}$       (D)  $\frac{(\sum y_i w_i)(\sum w_i x_i) - \sum y_i x_i}{(\sum w_i x_i)^2 - 1}$   
 (E)  $\frac{(\sum y_i w_i)(\sum w_i x_i) - \sum y_i w_i}{(\sum w_i x_i)^2 - 1}$

*Solution.* The design matrix is  $\mathbf{X} = \begin{pmatrix} x_1 & w_1 \\ x_2 & w_2 \\ \vdots & \vdots \\ x_n & w_n \end{pmatrix}$ , so

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i w_i \\ \sum_{i=1}^n x_i w_i & \sum_{i=1}^n w_i^2 \end{pmatrix} = \begin{pmatrix} 1 & \sum_{i=1}^n x_i w_i \\ \sum_{i=1}^n x_i w_i & 1 \end{pmatrix}$$

and

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{1 - (\sum_{i=1}^n x_i w_i)^2} \begin{pmatrix} 1 & -\sum_{i=1}^n x_i w_i \\ -\sum_{i=1}^n x_i w_i & 1 \end{pmatrix}.$$

The least squares estimator of  $(\beta_1, \beta_2)$  is

$$\begin{aligned} (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y}) &= \frac{1}{1 - (\sum_{i=1}^n x_i w_i)^2} \begin{pmatrix} 1 & -\sum_{i=1}^n x_i w_i \\ -\sum_{i=1}^n x_i w_i & 1 \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n w_i y_i \end{pmatrix} \\ &= \begin{pmatrix} \frac{\sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i w_i)(\sum_{i=1}^n w_i y_i)}{1 - (\sum_{i=1}^n x_i w_i)^2} \\ \frac{\sum_{i=1}^n w_i y_i - (\sum_{i=1}^n x_i w_i)(\sum_{i=1}^n x_i y_i)}{1 - (\sum_{i=1}^n x_i w_i)^2} \end{pmatrix}. \end{aligned}$$

In particular,

$$\hat{\beta}_1 = \boxed{\frac{\sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i w_i)(\sum_{i=1}^n w_i y_i)}{1 - (\sum_{i=1}^n x_i w_i)^2}}. \quad \text{(Answer: (D))}$$

□

**Problem 2.1.7.**  (SOA Course 120 November 1989 Question 13: Calculation of residual) You are given:

Period	$y$	$x_1$	$x_2$
1	1.3	6	4.5
2	1.5	7	4.6
3	1.8	7	4.5
4	1.6	8	4.7
5	1.7	8	4.6

You are to use the following regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad i = 1, 2, \dots, 5.$$

You have determined:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 1,522.73 & 26.87 & -374.67 \\ 26.87 & 0.93 & -7.33 \\ -374.67 & -7.33 & 93.33 \end{pmatrix}$$

Calculate  $e_2$ .

- (A) -0.2                      (B) -0.1                      (C) 0.0  
 (D) 0.1                      (E) 0.2

*Solution.* The LSEs are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} 1,522.73 & 26.87 & -374.67 \\ 26.87 & 0.93 & -7.33 \\ -374.67 & -7.33 & 93.33 \end{pmatrix} \begin{pmatrix} 7.9 \\ 57.3 \\ 36.19 \end{pmatrix} = \begin{pmatrix} 9.9107 \\ 0.2893 \\ -2.2893 \end{pmatrix}.$$

Thus  $\hat{y}_2 = \hat{\beta}_0 + 7\hat{\beta}_1 + 4.6\hat{\beta}_2 = 1.40502$  and the 2nd residual is

$$e_2 = y_2 - \hat{y}_2 = 1.5 - 1.40502 = \boxed{0.09498}. \quad \text{(Answer: (D))}$$

□

*Remark.* If you take  $e_2 = \hat{y}_2 - y_2$ , you will get Answer (B), which is incorrect!







- (ii) Failure to square  $r = 0.895168$  would result in Answer (D).
- (iii) A longer way to attack this example is to calculate RSS and TSS. Taking the sum of the squared residuals, we have

$$\begin{aligned} \text{RSS} &= (67 - 69.19)^2 + (68 - 67.80)^2 + (69 - 67.82)^2 + (72 - 71.28)^2 + (74 - 73.91)^2 \\ &= 6.755. \end{aligned}$$

Moreover,  $\text{TSS} = 5(2.607681)^2 = 34$ . Thus  $R^2 = 1 - \text{RSS}/\text{TSS} = 1 - 6.755/34 = \boxed{0.8013}$ .  
**(Answer: (B))**

**Problem 2.1.13.**  **(CAS Exam 3L Spring 2012 Question 25: Properties of least squares estimators)** You are using the simple linear model:

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

where

- $y_i, x_i$  are the dependent and independent variables, respectively.
- The  $\varepsilon_i$ 's are independent, identically distributed normal random variables with  $E(\varepsilon_i) = 0$ .

The parameters are estimated by:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

Which of the following characteristics apply to  $\hat{\beta}_0$  and  $\hat{\beta}_1$ ?

- I. Least-squares estimators
  - II. Maximum-likelihood estimators
  - III. Biased estimators
- (A) None                      (B) I and II                      (C) I and III  
 (D) II and III                      (E) I, II and III

*Solution.* Only I and II are true, as the LSEs are unbiased for  $\beta_0$  and  $\beta_1$ . **(Answer: (B))**  $\square$

**Problem 2.1.14.** 🌟 [HARDER!] (CAS VEE Applied Statistics Summer 2005 Question 13: Estimated covariance between LSEs) You fit the model  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$  to twenty observations. You are given:

$$\begin{aligned} \text{Error sum of squares} &= 2000 \\ \sum x_i &= -300 \\ \sum x_i^2 &= 6000 \end{aligned}$$

Determine  $\widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_1)$ .

- (A) 0.7                      (B) 0.8                      (C) 0.9  
(D) 1.0                      (E) 1.1

*Solution.* We didn't have a formula for the (estimated) covariance between  $\hat{\beta}_0$  and  $\hat{\beta}_1$  in Chapter 1, so let's extract it from the matrix  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ . From page 101, we have

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{nS_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix},$$

so the covariance between  $\hat{\beta}_0$  and  $\hat{\beta}_1$  should be the off-diagonal entry of  $(\mathbf{X}'\mathbf{X})^{-1}$  multiplied by  $\sigma^2$ . When  $\sigma^2$  is estimated by  $s^2$ , we get the estimated covariance:

$$\widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{s^2}{S_{xx}}\bar{x}. \quad (2.1.9)$$

Now the MSE is  $s^2 = \text{RSS}/(n-2) = 2,000/(20-2) = 1,000/9$ . With  $\bar{x} = -300/20 = -15$ , the estimated covariance between  $\hat{\beta}_0$  and  $\hat{\beta}_1$  is

$$\widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_1) = -\left[ \frac{1,000/9}{6,000 - 20(-15)^2} \right] (-15) = \boxed{\frac{10}{9} = 1.1111}. \quad (\text{Answer: (E)}) \quad \square$$

**Problem 2.1.15.** 🌟 [HARDER!] (Variance-covariance matrix of fitted response values) Consider the multiple linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

and  $\epsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ . Let  $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ .

Determine the variance-covariance matrix of the fitted response values.

- (A)  $\sigma^2(\mathbf{X}'\mathbf{X})$                       (B)  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$                       (C)  $\sigma^2\mathbf{H}$   
 (D)  $\sigma^2\mathbf{H}^{-1}$                       (E)  $\sigma^2(\mathbf{I} - \mathbf{H})$ , where  $\mathbf{I}$  is the  $n \times n$  identity matrix

*Solution.* The vector of fitted response values is

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y} = \mathbf{H}\mathbf{y}.$$

As  $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ , we can use (2.1.6) to get

$$\text{Var}(\hat{\mathbf{y}}) = \mathbf{H}(\sigma^2\mathbf{I}_n)\mathbf{H}' = \sigma^2\mathbf{H}\mathbf{H}' \stackrel{(\mathbf{H}'=\mathbf{H}=\mathbf{H}^2)}{=} \boxed{\sigma^2\mathbf{H}}. \quad (\text{Answer: (C)}) \quad \square$$

**Problem 2.1.16.**  \* (SOA Course 120 Study Note 120-83-96 Question 4: Standard error of a linear combination of LSEs – I) You fit the multiple regression model  $y_i = \beta_0 + \beta_1x_{i1} + \beta_2x_{i2} + \varepsilon_i$  to a set of data.

You determine:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 6.1333 & -0.0733 & -0.1933 \\ -0.0733 & 0.0087 & -0.0020 \\ -0.1933 & -0.0020 & 0.0087 \end{pmatrix}$$

$$s^2 = 280.1167$$

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

- (A) 1.9                      (B) 2.2                      (C) 2.5  
 (D) 2.8                      (E) 3.1

*Solution.* The estimated variance of  $\hat{\beta}_1 - \hat{\beta}_2$  is

$$\begin{aligned} \widehat{\text{Var}}(\hat{\beta}_1 - \hat{\beta}_2) &= \widehat{\text{Var}}(\hat{\beta}_1) + \widehat{\text{Var}}(\hat{\beta}_2) - 2\widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2) \\ &= 280.1167[0.0087 + 0.0087 - 2(-0.0020)] \\ &= 5.99449738. \end{aligned}$$

The standard error, which is the estimated standard deviation, of  $\hat{\beta}_1 - \hat{\beta}_2$  is

$$\sqrt{\widehat{\text{Var}}(\hat{\beta}_1 - \hat{\beta}_2)} = \sqrt{5.99449738} = \boxed{2.4484}. \quad (\text{Answer: (C)}) \quad \square$$



**Problem 2.1.18.**  (CAS VEE Applied Statistics Exam Summer 2005 Question 2: Standard error of a linear combination of LSEs – III) You are given:

- (i)  $y$  is the annual number of discharges from a hospital.
- (ii)  $x$  is the number of beds in the hospital.
- (iii) Dummy variable  $D$  is 1 if the hospital is private and 0 if the hospital is public.
- (iv) The classical three-variable linear regression model  $y = \beta_0 + \beta_1 x + \beta_2 D + \varepsilon$  is fitted to  $N$  cases using ordinary least squares.
- (v) The matrix of estimated variances and covariances of  $\hat{\beta}_0$ ,  $\hat{\beta}_1$  and  $\hat{\beta}_2$  is:

$$\begin{pmatrix} 1.89952 & -0.00364 & -0.82744 \\ -0.00364 & 0.00001 & -0.00041 \\ -0.82744 & -0.00041 & 2.79655 \end{pmatrix}$$

Determine the standard error of  $\hat{\beta}_0 + 600\hat{\beta}_1$ .

- (A) 1.06
- (B) 1.13
- (C) 1.38
- (D) 1.90
- (E) 2.35

*Solution.* The estimated variance of  $\hat{\beta}_0 + 600\hat{\beta}_1$  is

$$\begin{aligned} \widehat{\text{Var}}(\hat{\beta}_0 + 600\hat{\beta}_1) &= \widehat{\text{Var}}(\hat{\beta}_0) + 600^2 \widehat{\text{Var}}(\hat{\beta}_1) + 2(600)\widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_1) \\ &= 1.89952 + 600^2(0.00001) + 2(600)(-0.00364) \\ &= 1.131520. \end{aligned}$$

The standard error is  $\sqrt{1.131520} = \boxed{1.0637}$ . (Answer: (A)) □

*Remark.* (i) What a wasteful question! The information given in (i) to (iv) is not used at all!

(ii) Note that the matrix given in (v) is  $s^2(\mathbf{X}'\mathbf{X})^{-1}$ , not  $(\mathbf{X}'\mathbf{X})^{-1}$ .

**Problem 2.1.19.**  (CAS Exam ST Fall 2015 Question 21: Selecting significant variables by a t-test – I) You wish to explain  $y$  using the following multiple regression model and 32 observations:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

A linear regression package generates the following table of summary statistics:

	Estimated Coefficient	Standard Error
Intercept	44.200	5.960
$\beta_1$	-0.295	0.118
$\beta_2$	9.110	6.860
$\beta_3$	-8.700	1.200

For the Intercept and each of the betas, you decide to reject the null hypothesis which is that the Estimated Coefficient is zero at  $\alpha = 10\%$  significance.

Which variables have coefficients significantly different from zero?

- (A) Intercept                      (B) Intercept,  $x_1$                       (C) Intercept,  $x_2$   
 (D) Intercept,  $x_1, x_3$                       (E) Intercept,  $x_2, x_3$

**Comments:** Two serious errors are present in this CAS exam question:

- The model equation should read

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \boxed{+\varepsilon}.$$

The (normal) random error term is essential, without which  $y$  is non-random!

- One will never test the null hypothesis that the estimated coefficient  $\hat{\beta}_j$  is zero, which can be true or not, because  $\hat{\beta}_j$  is random depending on the sample of data. We are only interested in the unknown coefficient  $\beta_j$ , which is a model parameter.

*Solution.* The t-statistics for testing  $\beta_j = 0$  for  $j = 0, 1, 2, 3$  are calculated as follows:

$$\begin{aligned} t(\hat{\beta}_0) &= \frac{\hat{\beta}_0}{\text{SE}(\hat{\beta}_0)} = \frac{44.2}{5.960} = 7.4161, \\ t(\hat{\beta}_1) &= \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} = \frac{-0.295}{0.118} = -2.5, \\ t(\hat{\beta}_2) &= \frac{\hat{\beta}_2}{\text{SE}(\hat{\beta}_2)} = \frac{9.110}{6.860} = 1.3280, \\ t(\hat{\beta}_3) &= \frac{\hat{\beta}_3}{\text{SE}(\hat{\beta}_3)} = \frac{-8.700}{1.200} = -7.25. \end{aligned}$$

The null hypothesis is rejected when the observed value of the corresponding t-statistic exceeds  $t_{32-3-1,0.05} = t_{28,0.05} = 1.7011$  in absolute value. All but  $t(\hat{\beta}_2)$  satisfy this. Therefore, all except  $x_2$  have a coefficient significantly different from zero. **(Answer: (D))**  $\square$



**Problem 2.1.21.**  \* (Continuation of Example 2.1.14 (SRM Sample Question 27): What do the t-tests say?) Determine which of the following statements best describes the results of the model, at the 5% significance level.

- (A) Number of weekend days is a significant variable.
- (B) Number of weekend days is a significant variable, in the presence of the intercept, number of weekdays, and average number of members.
- (C) Number of weekend days is an insignificant variable.
- (D) Number of weekend days is an insignificant variable, in the presence of the intercept, number of weekdays, and average number of members.
- (E) None of the above statements is correct.

*Solution.* Because only the p-value of the test corresponding to number of weekend days is greater than 5%, it is insignificant, *in the presence of other variables* included in the MLR model. (Answer: (D)) □

**Problem 2.1.22.**  (SOA Course 120 May 1986 Question 8: Calculation of t-statistic – I) You are given the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where  $\varepsilon$  is a random error term with mean 0 and variance  $\sigma^2$ .

Using this model and a set of observations, you have determined the estimated variance-covariance matrix for the least squares estimators, as follows:

$$\begin{pmatrix} 900 & -144 & 225 \\ -144 & 1600 & 400 \\ 225 & 400 & 625 \end{pmatrix}$$

You have also determined:

$$\hat{\beta}_2 = 40$$

Calculate the  $t$  value used to test  $H_0 : \beta_2 = 0$ .

- (A) 0.6
- (B) 0.8
- (C) 1.0
- (D) 1.3
- (E) 1.6

*Solution.* The t-statistic is

$$t(\hat{\beta}_2) = \frac{40 - 0}{\sqrt{625}} = \boxed{1.6}. \quad (\text{Answer: (E)}) \quad \square$$

**Problem 2.1.23.** \* (SOA Course 120 Study Note 120-82-97 Question 5: Calculation of t-statistic – II) You perform a regression of  $y$  on  $x_1$  and  $x_2$ .

You determine:

$$\hat{y}_i = 20.0 - 1.5x_{i1} - 2.0x_{i2}$$

Source	Sum of Squares	$df$	Mean Square	$F$
Regression	42	2	21	5.25
Error	12	3	4	
Total	54	5		

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 4/3 & -1/4 & -1/3 \\ -1/4 & 1/16 & 0 \\ -1/3 & 0 & 2/3 \end{pmatrix}$$

Calculate the value of the t-statistic for testing the null hypothesis  $H_0: \beta_2 = 1$ .

- (A) -0.9                      (B) -1.2                      (C) -1.8  
 (D) -3.0                      (E) -5.0

*Solution.* The t-statistic for testing  $H_0: \beta_2 = 1$  is

$$t(\hat{\beta}_2) = \frac{\hat{\beta}_2 - 1}{\text{SE}(\hat{\beta}_2)} = \frac{-2.0 - 1}{\sqrt{4(2/3)}} = \boxed{-1.8371}. \quad \text{(Answer: (C))}$$

□

**Problem 2.1.24.**  (SOA Course 120 May 1988 Question 8: C.I. construction) A sample of 25 observations has been represented by a model of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where  $\varepsilon$  is a random error term with mean 0 and variance  $\sigma^2$ .

You are given:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 188.9832 & 0.8578 & -28.0275 \\ 0.8578 & 0.2500 & -0.6000 \\ -28.0275 & -0.6000 & 5.0625 \end{pmatrix}$$

$$s^2 = 0.0361$$

$$\hat{\beta} = \begin{pmatrix} -4.04 \\ 0.14 \\ 0.45 \end{pmatrix}$$

Determine the shortest symmetric 95-percent confidence interval for  $\beta_1$ .

- (A)  $(-0.26, 0.54)$                       (B)  $(-0.12, 0.40)$                       (C)  $(-0.06, 0.34)$   
 (D)  $(0.06, 0.22)$                       (E)  $(0.12, 0.16)$

*Solution.* The estimated variance of  $\hat{\beta}_1$  is

$$\widehat{\text{Var}}(\hat{\beta}_1) = s^2 \times (2,2)\text{th entry of } (\mathbf{X}'\mathbf{X})^{-1} = 0.0361(0.25) = 0.009025.$$

The symmetric 95% confidence interval for  $\beta_1$  is

$$\begin{aligned} \hat{\beta}_1 \pm t_{25-2-1, 0.025} \times \text{SE}(\hat{\beta}_1) &= 0.14 \pm 2.0739\sqrt{0.009025}. \\ &= \boxed{(-0.05703, 0.33703)}. \quad \text{(Answer: (C))} \end{aligned}$$

□

**Problem 2.1.25.** \* (Given a C.I., find the standard error of a linear combination of LSEs) For the multiple linear regression model  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$  with  $i = 1, \dots, 20$ , you are given:

$$(i) (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.20 & -0.15 & -0.15 & -0.17 \\ -0.15 & 0.39 & 0.03 & -0.05 \\ -0.15 & 0.03 & 0.26 & 0.00 \\ -0.17 & -0.05 & 0.00 & 0.65 \end{pmatrix}$$

- (ii) The 98% symmetric confidence interval for  $\beta_2$  is  $(-0.36, 2.55)$ .



*Solution.* From the ANOVA table, we have  $s^2 = 162.86/2 = 81.43$ . With  $\mathbf{x}_* = (1 \ 20 \ 30)'$ , the standard error of the prediction error is

$$\begin{aligned} \sqrt{s^2[1 + \mathbf{x}_*'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*]} &= \sqrt{81.43 \left( 1 + \underbrace{(1 \ 20 \ 30) \begin{pmatrix} 9.300 & -0.080 & -0.300 \\ -0.080 & 0.009 & -0.004 \\ -0.300 & -0.004 & 0.016 \end{pmatrix} \begin{pmatrix} 1 \\ 20 \\ 30 \end{pmatrix}}_{1.3} \right)} \\ &= \boxed{13.69}. \quad (\text{Answer: (B)}) \end{aligned}$$

□

*\*\*Use the following information for the next two problems.\*\**

You are examining the relationship between the number of fatal car accidents on a tollway each month and three other variables: precipitation, traffic volume and the occurrence of a holiday weekend during the month.

You are using the following model:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where

- $y$  = the number of fatal car accidents,
- $x_1$  = precipitation, in inches,
- $x_2$  = traffic volume, in thousands of cars,
- $x_3$  = 1, if a holiday weekend occurs during the month, and  
0, otherwise, and
- $\varepsilon$  is a random error term with mean 0 and variance  $\sigma^2$ .

The following data were collected for a 12-month period:

Month	$y$	$x_1$	$x_2$	$x_3$
1	1	3	1	1
2	3	2	1	1
3	1	2	1	0
4	2	5	2	1
5	4	4	2	1
6	1	1	2	0
7	3	0	2	1
8	2	1	2	1
9	0	1	3	1
10	2	2	3	1
11	1	1	4	0
12	3	4	4	1

You have determined:

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{6,506} \begin{pmatrix} 257 & -82 & -446 \\ -82 & 254 & -364 \\ -446 & -364 & 2,622 \end{pmatrix}$$

$$s^2 = 1.45241$$

**Problem 2.1.27.**  (SOA Part 3 November 1985 Question 7) Determine  $\hat{\beta}_1$ .

(Answer to the nearest 0.01)

- (A) -0.07                      (B) 0.15                      (C) 0.24  
 (D) 0.70                      (E) 1.30

*Solution.* By (2.1.2),

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \overbrace{\begin{pmatrix} \sum x_{i1}y_i \\ \sum x_{i2}y_i \\ \sum x_{i3}y_i \end{pmatrix}}^{\text{calculate from data patiently!}} = \frac{1}{6,506} \begin{pmatrix} 257 & -82 & -446 \\ * & * & * \\ * & * & * \end{pmatrix} \begin{pmatrix} 57 \\ 51 \\ 20 \end{pmatrix} = \begin{pmatrix} 0.2378 \\ * \\ * \end{pmatrix},$$

so  $\hat{\beta}_1 = \boxed{0.2378}$ . (Answer: (C)) □

**Problem 2.1.28.**  (SOA Part 3 November 1985 Question 8) You have predicted the number of fatal car accidents when:

- (i) precipitation is 3 inches,
- (ii) traffic volume is 4000 cars, and
- (iii) a holiday weekend occurs during the month.

Determine the estimated variance of the prediction.

(Answer to nearest 0.01)

- (A) 0.32
- (B) 0.77
- (C) 1.22
- (D) 1.45
- (E) 1.77

*Solution.* With  $\mathbf{x}_* = (3 \ 4 \ 1)'$ , the estimated variance of the prediction error is

$$\begin{aligned}
 & s^2[1 + \mathbf{x}'_*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*] \\
 = & 1.45241 \left[ 1 + \frac{1}{6,506} \times (3 \ 4 \ 1) \underbrace{\begin{pmatrix} 257 & -82 & -446 \\ -82 & 254 & -364 \\ -446 & -364 & 2,622 \end{pmatrix}}_{1,443} \begin{pmatrix} 3 \\ 4 \\ 1 \end{pmatrix} \right] \\
 = & \boxed{1.7745}. \quad (\text{Answer: (E)})
 \end{aligned}$$

□

*Remark.* If you omit  $s^2 = 1.45241$  and just compute  $s^2\mathbf{x}'_*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*$ , you will end up with the incorrect Answer (A).

**Problem 2.1.29.** (SOA Course 120 November 1988 Question 9: Length of P.I.)

You are given the following information:

$y$	$x_1$	$x_2$
1	12	6.8
2	13	7.2
3	13	7.4
4	14	7.1
5	14	7.0
6	15	7.4

$$s^2 = 0.035$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 188.9832 & 0.8578 & -28.0275 \\ 0.8578 & 0.2528 & -0.5963 \\ -28.0275 & -0.5963 & 5.0459 \end{pmatrix}$$

Determine the length of the 95-percent prediction interval for  $y_*$ , where  $(x_{*0}, x_{*1}, x_{*2}) = (1, 13, 6.9)$ .

- (A) 0.7                      (B) 1.0                      (C) 1.2  
 (D) 1.5                      (E) 1.8

*Solution.* We first compute

$$\begin{aligned} s^2 \mathbf{x}'_* (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_* &= 0.035 \begin{pmatrix} 1 & 13 & 6.9 \end{pmatrix} \underbrace{\begin{pmatrix} 188.9832 & 0.8578 & -28.0275 \\ 0.8578 & 0.2528 & -0.5963 \\ -28.0275 & -0.5963 & 5.0459 \end{pmatrix}}_{0.488779} \begin{pmatrix} 1 \\ 13 \\ 6.9 \end{pmatrix} \\ &= 0.017107. \end{aligned}$$

The length of the 95% prediction interval for  $y_*$  is

$$\begin{aligned} \underbrace{2 t_{6-3, 0.025}}_{3.1824} \sqrt{s^2 [1 + \mathbf{x}'_* (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_*]} &= 2(3.1824) \sqrt{0.035 + 0.017107} \\ &= \boxed{1.4529}. \quad (\text{Answer: (D)}) \end{aligned}$$

□

*Remark.* The raw dataset plays no role in this problem.

**Problem 2.1.30.** \* (Construction of a prediction interval given summary output)

For a multiple linear regression model  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$  for  $i = 1, 2, \dots, 6$ , you are given:

(i) The fitted regression function is  $\hat{y} = 34.5 - 0.304x_1 + 0.383x_2$ .

(ii) The (incomplete) ANOVA table:

Source	Sum of Squares	df
Regression	270.09	?
Error	?	?
Total	290.00	?

$$(iii) (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 25.0487 & -0.8457 & 0.2864 \\ -0.8457 & 0.0294 & -0.0104 \\ 0.2864 & -0.0104 & 0.0040 \end{pmatrix}$$

Calculate the upper end-point of the 95% symmetric prediction interval for  $y_*$  when  $x_{*1} = 60$  and  $x_{*2} = 80$ .

- (A) 54                                      (B) 55                                      (C) 56  
 (D) 57                                      (E) 58

*Solution.* The point predictor is

$$\hat{y}_* = 34.5 - 0.304(60) + 0.383(80) = 46.9.$$

With  $\mathbf{x}_* = (1 \ 60 \ 80)'$ , the standard error of prediction is

$$\begin{aligned} \text{SE}(y_* - \hat{y}_*) &= \sqrt{s^2[1 + \mathbf{x}'_* (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_*]} \\ &= \sqrt{6.64 \left( 1 + \underbrace{\begin{bmatrix} 1 & 60 & 80 \end{bmatrix} \begin{bmatrix} 25.0487 & -0.8457 & 0.2864 \\ -0.8457 & 0.0294 & -0.0104 \\ 0.2864 & -0.0104 & 0.0040 \end{bmatrix} \begin{bmatrix} 1 \\ 60 \\ 80 \end{bmatrix}}_{0.9887} \right)} \\ &= 3.633864. \end{aligned}$$

Thus a 95% prediction interval for  $y_*$  is

$$46.9 \pm \underbrace{t_{6-3, 0.025}}_{3.1824} \times \text{SE}(y_* - \hat{y}_*) = [35.34, \boxed{58.46}]. \quad \text{(Answer: (E))}$$

□

## Prelude

Now that you have been well “trained” on this manual, you need to be exposed to “unseen tests” to avoid overfitting and identify areas in which you need more “training.” To this end, here are six comprehensive practice exams designed to assess your understanding of the whole SRM exam syllabus and boost your chance of passing the real exam.

**What are these practice exams like?** These practice exams have the following characteristics:

- Each exam has 35 multiple-choice questions distributed in line with the weights of the five topics in the SRM exam syllabus. In particular, each exam has about 7 questions set on decision trees in view of their increased weight (20-25%) effective from the May 2023 sitting.
- They represent a nice combination of quantitative (roughly 40%) and qualitative (roughly 60%) exam items, as many students who took Exam SRM recently have suggested. The amount of calculations required by the computational questions should be reasonable (not too tedious, not trivial).
- For your convenience, the questions in these practice exams are sorted  $\downarrow$  according to the 10 chapters of this manual. That is, Question 1 is set on Chapter 1 (SLR) and Question 35 on Chapter 10 (cluster analysis). This way, you can easily see which topics are your weak spots and identify additional practice questions for those topics. Questions in the real SRM exam will appear in a random order.
- The six exams have more or less the same level of difficulty, so you need not work them out in order. You can start with Practice Exam 6 if you like.

**How to use these practice exams?** To make the most of these exams, here are our recommendations:

- Attempt them when and only when you have completed the core of this study manual. Working on the practice exams when you are not fully ready defeats their purpose.
- Set aside exactly 3.5 hours and work on each exam in a simulated exam environment detached from distractions. Put away your notes and phone—no Facebook , Instagram , Twitter , or Snapchat  for 3.5 hours. You can only have the SRM tables, scratch papers, and your calculator  with you, as if this was a real exam.
- Budget your time wisely.  Don’t spend a disproportionate amount of time (say, more than 10 minutes) on a single question, no matter how difficult it seems, and don’t be afraid to skip questions. For a 210-minute exam with 35 questions, you should spend about 6 minutes on each question.
- When you are done, check your answers with the detailed illustrative solutions we provide. If you miss a question, it is important to understand the cause. Is it due to a lack of familiarity with the syllabus material, carelessness, or just bad luck? In quite a number of computational questions, the wrong answers are distractors that come with some rationale. (I tried to anticipate what mistakes students can make. The SOA will do the same! ) Even if you get a question right, it is beneficial to look at our solutions, which may be shorter or neater than yours, and may include some problem-solving remarks.

**⚠ IMPORTANT NOTE ⚠**

The experiences of students who took SRM recently and the main author of this manual (who has experienced SRM first-hand) suggest that these practice exams are likely more difficult than the real exam, so if you do well (say, you get **at least 25 out of 35 questions** correct in each exam), you should be on your way to passing SRM with ease. Good luck!

# Practice Exam 1

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

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

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

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

Calculate the least squares estimate of  $\beta_0$ .

(A) 15

(B) 18

(C) 20

(D) 22

(E) 25

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

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

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

(iii) The least squares estimate of  $\beta_1$  is  $-0.75$ .

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

(A) 131

(B) 132

(C) 133

(D) 134

(E) 135

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

- (i) The least squares estimate of  $\beta_1$  is 4.5.
- (ii) The sample standard deviation of the explanatory variable is 5.
- (iii) The residual standard error is 50.

You use a t-test to test  $H_0 : \beta_1 = 0$  against  $H_a : \beta_1 \neq 0$ .

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

- (A) Do not reject  $H_0$  at the 0.100 significance level.
  - (B) Reject  $H_0$  at the 0.100 significance level, but not at the 0.050 significance level.
  - (C) Reject  $H_0$  at the 0.050 significance level, but not at the 0.025 significance level.
  - (D) Reject  $H_0$  at the 0.025 significance level, but not at the 0.010 significance level.
  - (E) Reject  $H_0$  at the 0.010 significance level.
4. You fit the simple linear regression model  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$  to 10 observed values  $(x_i, y_i)$ .

You are given:

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

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

- (A) 1
- (B) 2
- (C) 3
- (D) 4
- (E) 5

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

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

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

(iii) The 98% symmetric confidence interval for  $\beta_2$  is (9.638, 20.362).

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

- (A) (79, 112)
- (B) (75, 116)
- (C) (71, 120)
- (D) (67, 124)
- (E) (63, 128)

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

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

Determine which of the following inequalities is always correct.

- (A)  $R^2 \leq r_{12}^2$
- (B)  $r_{12}^2 \leq R^2$
- (C)  $R^2 \leq r_1^2$
- (D)  $r_1^2 \leq R^2$
- (E)  $R^2 \leq r_2^2$

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

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

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

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

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

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

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

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

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

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

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

- (A) 2.2
- (B) 3.3
- (C) 4.4
- (D) 5.5
- (E) 6.6

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

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

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

(iii)

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

Calculate the weighted least squares estimate of  $\beta_1$ .

- (A) 0.15  
(B) 0.18  
(C) 0.20  
(D) 0.22  
(E) 0.25
9. Using ordinary least squares, Steve has fitted a simple linear regression model to predict an individual's weight from the individual's height. It turns out that some of the individuals in the study are members of the same family and so have been exposed to the same environmental factors.
- Determine which of the following statements about Steve's model is/are true.
- I. The estimated standard errors of the coefficient estimates are lower than they should be.  
II. Confidence and prediction intervals are narrower than they should be.  
III. He may be led to erroneously conclude that height is a statistically significant predictor of weight when it is not.
- (A) I only  
(B) II only  
(C) III only  
(D) I, II, and III  
(E) The correct answer is not given by (A), (B), (C), or (D).

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

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

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

- I. It is for a quantitative response variable.
- II. It refers to the squared discrepancy between the response variable of a previously unseen observation and the prediction of a model fitted to a fixed training set, averaged over a large number of test observations  $(x_0, y_0)$ .
- III.  $\text{Var} \left( \hat{f}(x_0) \right) + \left[ \text{Bias} \left( \hat{f}(x_0) \right) \right]^2$  is known as the irreducible error.
  - (A) I only
  - (B) II only
  - (C) III only
  - (D) I, II, and III
  - (E) The correct answer is not given by (A), (B), (C), or (D).

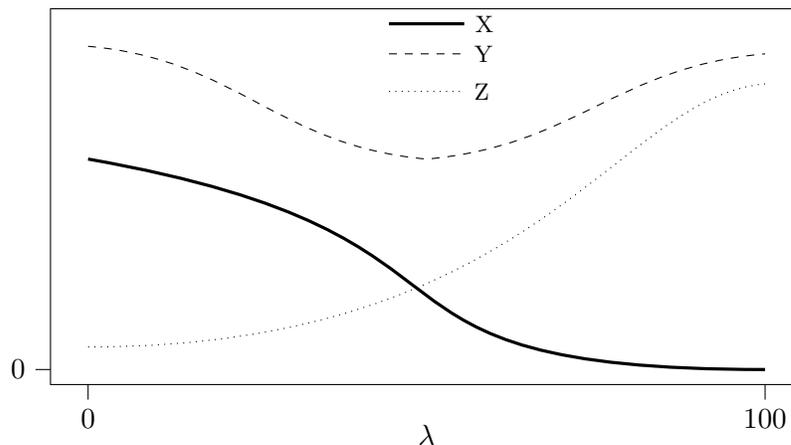
11.  Determine which of the following statements about different resampling methods is/are true.

- I. The validation set approach is a special case of  $k$ -fold cross-validation (CV).
- II. LOOCV has lower bias than  $k$ -fold CV when  $k < n$ .
- III.  $k$ -fold CV is generally less computationally expensive than LOOCV when  $k < n$ .
  - (A) I and II only
  - (B) I and III only
  - (C) II and III only
  - (D) I, II, and III
  - (E) The correct answer is not given by (A), (B), (C), or (D)

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

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

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



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

- |     | <u>X</u>     | <u>Y</u>     | <u>Z</u>     |
|-----|--------------|--------------|--------------|
| (A) | Squared bias | Test MSE     | Training MSE |
| (B) | Squared bias | Variance     | Training MSE |
| (C) | Variance     | Squared bias | Training MSE |
| (D) | Squared bias | Test MSE     | Variance     |
| (E) | Variance     | Test MSE     | Squared bias |

13.  A multiple linear regression model  $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \varepsilon$  is fitted, leading to the following table of parameter estimates:

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

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

- (A) Intercept
  - (B)  $x_1$
  - (C)  $x_2$
  - (D)  $x_3$
  - (E) None should be dropped from the model
14.  You are given the following information about a GLM:
- The model uses four categorical explanatory variables:
    - (a)  $x_1$  is a categorical variables with three levels.
    - (b)  $x_2, x_3$  are categorical variables with two levels.
    - (c)  $x_4$  is a categorical variable with six levels.
  - The model also uses a continuous explanatory variable  $x_5$  modeled with a first order polynomial.
  - There is only one interaction in the model, which is between  $x_1$  and  $x_5$ .

Determine the maximum number of parameters in this model.

- (A) Less than 13
- (B) 13
- (C) 14
- (D) 15
- (E) At least 16

15.  You are given the following GLM output:

Response variable	Pure Premium	
Response distribution	Gamma	
Link	Log	
Scale parameter	1.1	
Parameter	df	$\hat{\beta}$
Intercept	1	4.78
Risk Group	2	
Group 1	0	0.00
Group 2	1	-0.20
Group 3	1	-0.35
Vehicle Symbol	1	
Symbol 1	0	0.00
Symbol 2	1	0.42

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

- (A) 84
- (B) 92
- (C) 7044
- (D) 7749
- (E) 591253

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

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

Calculate  $B$ .

- (A) 211
  - (B) 212
  - (C) 213
  - (D) 214
  - (E) 215
17.  Determine which of the following statements about GLMs with normal responses and identity link function is/are true.
- I. A large deviance indicates a poor fit for a model.
  - II. The deviance reduces to the residual sum of squares.
  - III. The deviance residual is the same as the Pearson residual.
- (A) I only
  - (B) II only
  - (C) III only
  - (D) I, II, and III
  - (E) The correct answer is not given by (A), (B), (C), or (D).

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

- (i) The model uses two explanatory variables:
- (a) Age group, which is treated as a continuous explanatory variable taking values of 1, 2 and 3, modeled with a second order polynomial
  - (b) Sex, which is a categorical explanatory variable with two levels
- (ii) Parameter estimates:

Parameter	$\hat{\beta}$
Intercept	-1.1155
Sex	
Female	0.0000
Male	-0.4192
Age group	1.2167
(Age group) <sup>2</sup>	-0.5412

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

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

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

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

19.  You are given:

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

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

where  $x_i$  refers to the value of the explanatory variable of the  $i$ th observation.

- Analysis of the data produced the following output:

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

Calculate the deviance of the model.

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

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

1    1.5    1.6    1.4    1.5    1.7

Calculate the sample lag-3 autocorrelation.

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

21.  You are given:

(i) The random walk model

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

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

(ii) The following eight observed values of  $y_t$ :

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

(iii)  $y_0 = 0$

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

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

- (A) (20, 40)
- (B) (18, 42)
- (C) (16, 44)
- (D) (14, 46)
- (E) (12, 48)

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

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

Calculate the mean absolute percentage error.

- (A) 30
- (B) 35
- (C) 40
- (D) 45
- (E) 50

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

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

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

(iii)  $y_{2018} = 10$

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

(A) (2.5, 5.3)

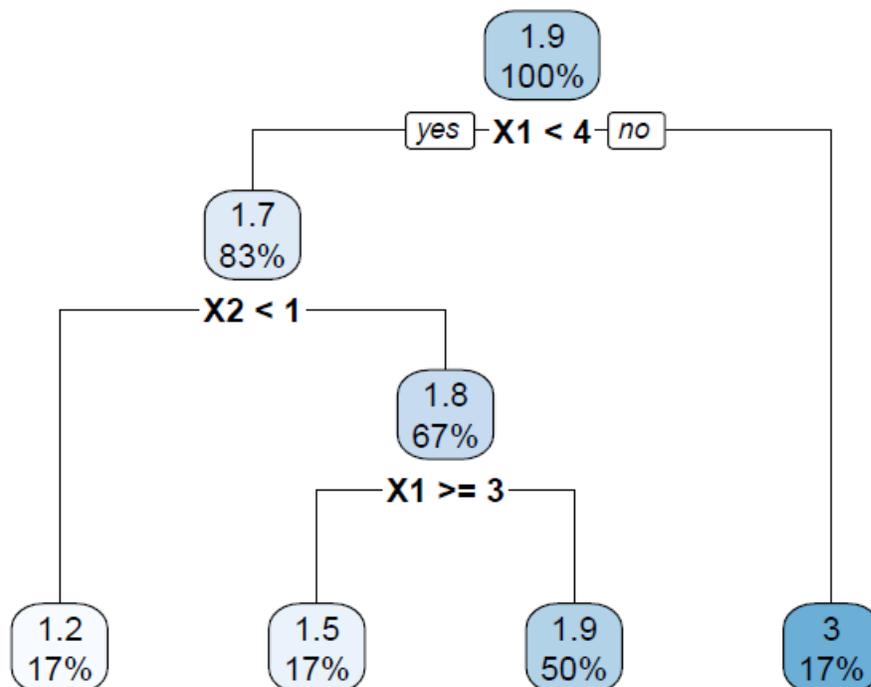
(B) (2.4, 5.4)

(C) (2.3, 5.5)

(D) (2.2, 5.6)

(E) (2.1, 5.7)

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



In each node, the first number is the response mean of the training observations in that node. Determine which of the following statements is/are true.

- I. The predicted value when  $X_1 = 5$  and  $X_2 = 2$  is 3.
  - II. The predicted value when  $X_1 = 3$  and  $X_2 = 0$  is 1.2.
  - III. The predicted value when  $X_1 = 2$  and  $X_2 = 3$  is 1.5
- (A) None
  - (B) I and II only
  - (C) I and III only
  - (D) II and III only
  - (E) The correct answer is not given by (A), (B), (C), or (D).

- 25.** Determine which of the following statements about recursive binary splitting for decision trees is/are true.
- I. It is a greedy algorithm.
  - II. It is a top-down approach.
  - III. In making each split, one of the predictors is randomly chosen as the split variable.
- (A) I only  
(B) II only  
(C) III only  
(D) I, II, and III  
(E) The correct answer is not given by (A), (B), (C), or (D).
- 26.** Determine which of the following statements about cost complexity pruning for decision trees is/are true.
- I. A tree split is made so long as the decrease in node impurity due to that split exceeds some threshold.
  - II. It has the effect of increasing the variance compared to an unpruned tree.
  - III. It is also known as strongest link pruning.
- (A) None  
(B) I and II only  
(C) I and III only  
(D) II and III only  
(E) The correct answer is not given by (A), (B), (C), or (D).
- 27.** Consider a classification tree with a binary response variable. Determine which of the following statements is/are true.
- I. A pure node is characterized by a classification error rate close to zero.
  - II. A pure node is characterized by a Gini index close to zero.
  - III. The classification error is always bounded from above by the Gini index.
- (A) I only  
(B) II only  
(C) III only  
(D) I, II, and III  
(E) The correct answer is not given by (A), (B), (C), or (D).

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

- I. Decision trees can be displayed graphically.
  - II. Decision trees can capture interactions between variables without the use of additional features.
  - III. Decision trees tend to be non-robust in the sense that a small change in the data can cause a large change in the final estimated tree.
- (A) I only  
 (B) II only  
 (C) III only  
 (D) I, II, and III  
 (E) The correct answer is not given by (A), (B), (C), or (D).

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

0.1, 0.15, 0.15, 0.25, 0.55, 0.6, 0.6, 0.65, 0.7, 0.75.

Determine which of the following combinations is correct.

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

- 30.**  Determine which of the following statements about bagging and random forests is/are true.
- I. Bagging can only be used for regression problems while random forests can be used for both regression and classification problems.
  - II. Only bagging (but not random forests) involves randomly selecting predictors when making a split.
  - III. Only the test error of a bagged model (but not that of a random forest) can be estimated by out-of-bag estimation.
- (A) I only
  - (B) II only
  - (C) III only
  - (D) I, II, and III
  - (E) The correct answer is not given by (A), (B), (C), or (D).
- 
- 31.**  Determine which of the following statements about principal components analysis is/are true.
- I. Scaling the variables has no effect on the results of principal components analysis.
  - II. Each principal component loading vector is unique.
  - III. If the number of principal components is one less than the number of observations, then the representation of observed data in terms of principal components is exact.
- (A) I only
  - (B) II only
  - (C) III only
  - (D) I, II, and III
  - (E) The correct answer is not given by (A), (B), (C), or (D).

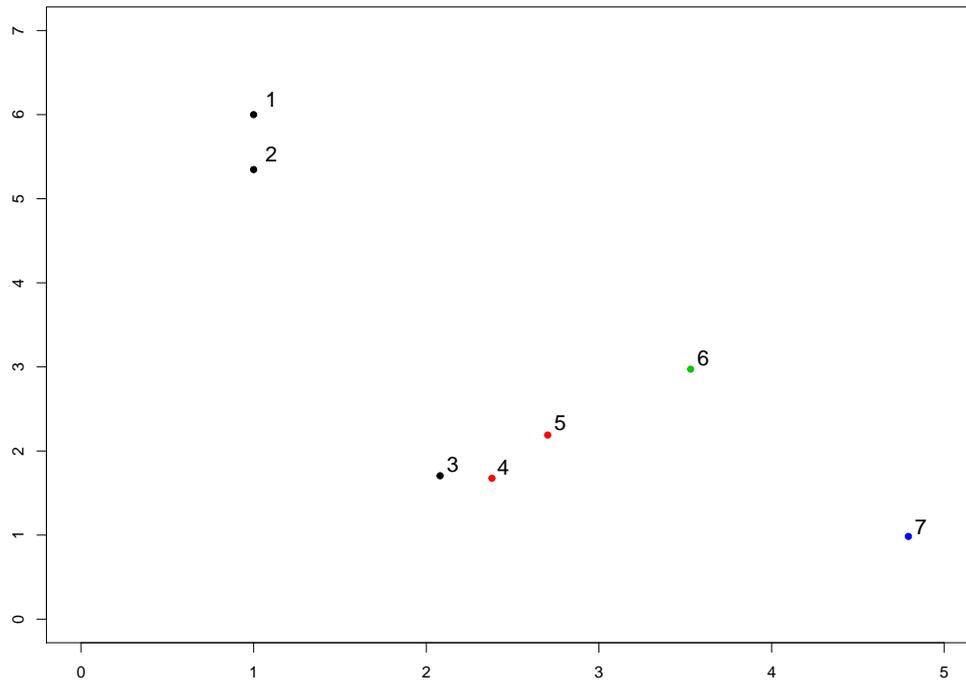
**32.**  Determine for which of the following statistical learning methods the variables are typically standardized.

- I. Shrinkage methods
  - II. Principal components analysis
  - III.  $K$ -means clustering
- (A) None
  - (B) I and II only
  - (C) I and III only
  - (D) II and III only
  - (E) The correct answer is not given by (A), (B), (C), or (D).

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

- I. If the number of clusters is known a priori, then  $K$ -means clustering is always preferred over hierarchical clustering.
  - II. If hierarchical clustering is applied to  $n$  observations and  $n$  clusters are desired, then each cluster contains only one observation.
  - III. In each step of the  $K$ -means clustering algorithm, only one observation can move to another cluster.
- (A) I only
  - (B) II only
  - (C) III only
  - (D) I, II, and III
  - (E) The correct answer is not given by (A), (B), (C), or (D).

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



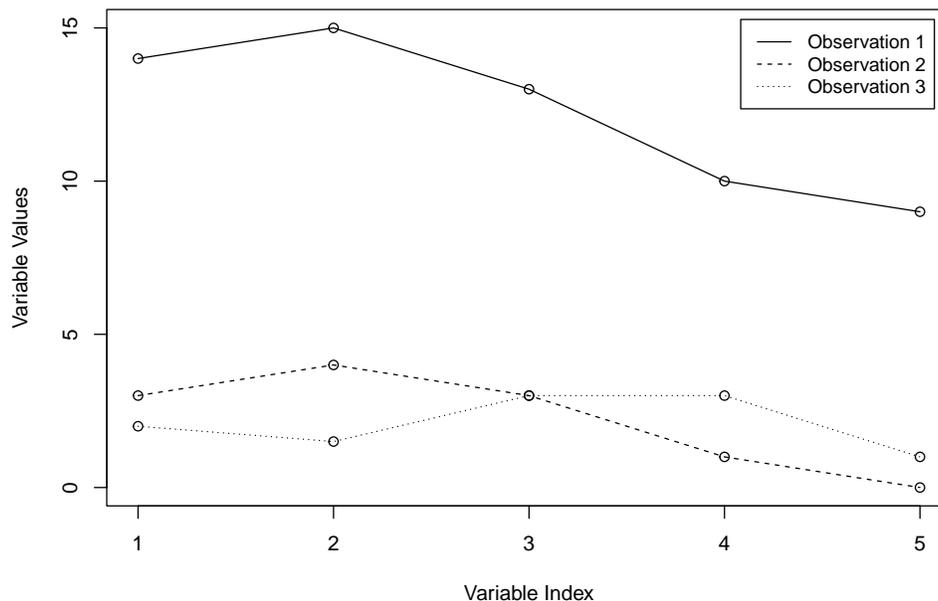
Hierarchical clustering with single linkage is used to determine the clusters. If four clusters are desired, then they are given by

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

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

- (A)  $C_1$  and  $C_2$
- (B)  $C_1$  and  $C_3$
- (C)  $C_1$  and  $C_4$
- (D)  $C_2$  and  $C_3$
- (E)  $C_2$  and  $C_4$

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



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

- I. Observations 1 and 2 are the closest in terms of Euclidean distance.
  - II. Observations 2 and 3 are the closest in terms of Euclidean distance.
  - III. Observations 1 and 2 are the closest in terms of correlation-based distance.
- (A) None  
 (B) I and II only  
 (C) I and III only  
 (D) II and III only  
 (E) The correct answer is not given by (A), (B), (C), or (D).

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

**Solutions to Practice Exam 1****Answer Key**

Question #	Answer
1	A
2	A
3	B
4	C
5	D
6	D
7	D
8	E
9	D
10	A
11	C
12	E
13	B
14	B
15	D
16	C
17	D
18	A
19	D
20	D

Question #	Answer
21	C
22	E
23	E
24	B
25	E
26	A
27	D
28	D
29	A
30	E
31	C
32	E
33	B
34	D
35	D

**1. (LSEs of  $\beta_0$  and  $\beta_1$  in SLR)**

*Solution.* By (1.1.6), the LSE of  $\beta_1$  is

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

Then by (1.1.4), the LSE of  $\beta_0$  is

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

□

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

*Solution 1 (Better).* By (1.1.6),

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

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

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

□

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

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

□

**3. (Result of a t-test for various  $\alpha$ )**

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

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

The t-statistic for  $H_0 : \beta_1 = 0$  is

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

which is between  $t_{18,0.050} = 1.7341$  and  $t_{18,0.025} = 2.1009$ . The critical region of the two-sided test is  $\{|t(\hat{\beta}_1)| > t_{18,\alpha/2}\}$ , where  $\alpha$  is the size of the test. Thus  $H_0 : \beta_1 = 0$  is rejected at  $\alpha = 2(0.050) = 0.10$ , but not at  $\alpha = 2(0.025) = 0.05$ . (**Answer: (B)**) □

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

(ii) Answer (C) is for the one-sided alternative  $H_a : \beta_1 > 0$ .