

An Integrated
Component
of Your Study
Manual Program



a/s/m

Exam SRM Study Manual



4th Edition

Abraham Weishaus, Ph.D., FSA, CFA, MAAA

a/s/m

Actuarial Study Materials

Learning Made Easier

a/s/m

Exam SRM Study Manual

4th Edition

Abraham Weishaus, Ph.D., FSA, CFA, MAAA



TO OUR READERS:

Please check A.S.M.'s web site at www.studymanuals.com for errata and updates. If you have any comments or reports of errata, please e-mail us at mail@studymanuals.com.

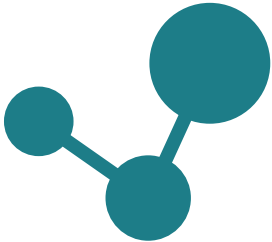
©Copyright 2023, Actuarial Study Materials, a division of ArchiMedia Advantage Inc.

All rights reserved. Reproduction in whole or in part without express written permission from the publisher is strictly prohibited.



Welcome to Actuarial University

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



You can find integrated topics using this network icon.


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

1. Login to: www.actuarialuniversity.com

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

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

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

 Pareto Distribution ×

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

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

and cdf

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

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

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

and

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

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

Related Topics ▾

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

Unlocked Products are the products that you own.

ACTEX Manual for P



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

Probability for Risk Management, 3rd Edition



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

Instructional Videos*

Topic Search

Planner

Formula & Review Sheet

Make your study session more efficient with our Planner!

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

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



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 B 235 C 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.

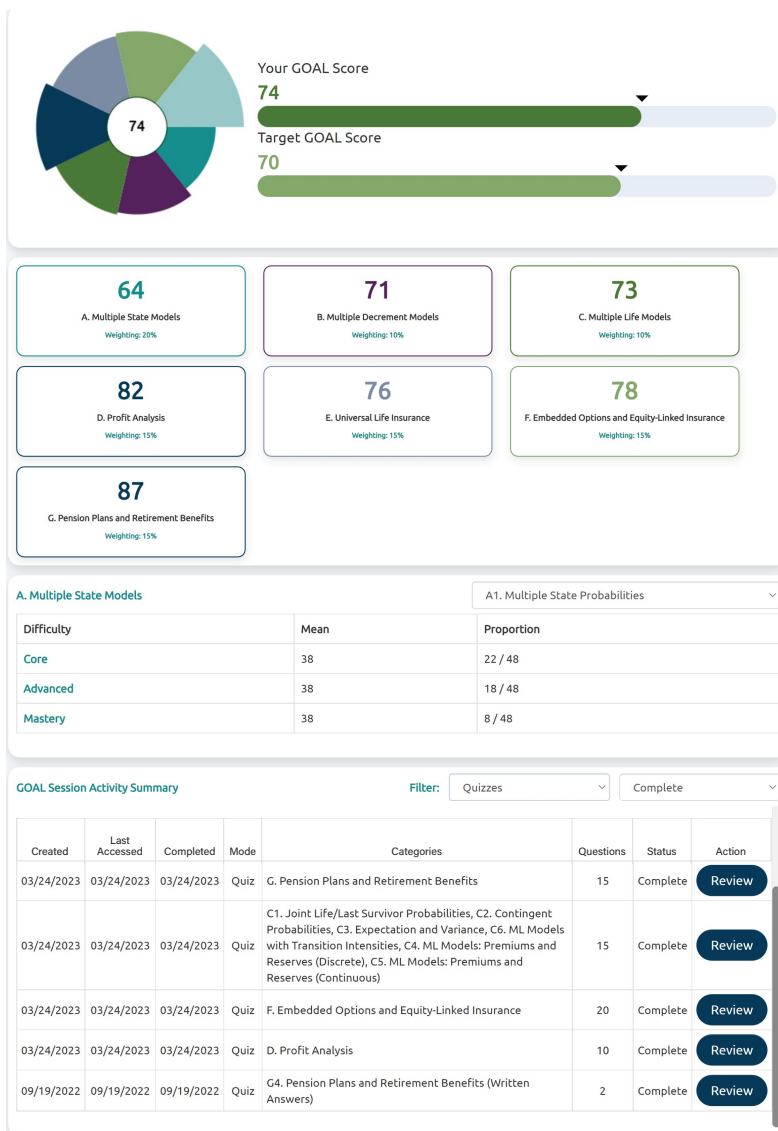


Track your exam readiness with GOAL Score!

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

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

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



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

1	Basics of Statistical Learning	1
1.1	Statistical learning	1
1.2	Types of variables	3
1.3	Graphs	3
	Exercises	7
	Solutions	7
I	Linear Regression	9
2	Linear Regression: Estimating Parameters	11
2.1	Basic linear regression	11
2.2	Multiple linear regression	14
2.3	Alternative model forms	15
	Exercises	16
	Solutions	27
3	Linear Regression: Standard Error, R^2, and t statistic	33
3.1	Residual standard error of the regression	33
3.2	R^2 : the coefficient of determination	35
3.3	t statistic	36
3.4	Added variable plots and partial correlation coefficients	38
	Exercises	40
	Solutions	52
4	Linear Regression: F	59
	Exercises	61
	Solutions	71
5	Linear Regression: Validation	79
5.1	Validating model assumptions	79
5.2	Outliers and influential points	81
5.3	Collinearity of explanatory variables; VIF	83
	Exercises	86
	Solutions	94
6	Resampling Methods	101
6.1	Validation set approach	101
6.2	Cross-validation	102
	Exercises	105
	Solutions	107
7	Linear Regression: Subset Selection	109
7.1	Subset selection	109
7.2	Choosing the best model	111
	Exercises	115
	Solutions	125

8	Linear Regression: Shrinkage and Dimension Reduction	131
8.1	Shrinkage methods	131
8.1.1	Ridge regression	131
8.1.2	The lasso	133
8.2	Dimension reduction methods	135
8.2.1	Principal components regression	136
8.2.2	Partial least squares	136
8.3	The curse of dimensionality	137
	Exercises	139
	Solutions	148
9	Linear Regression: Predictions	155
	Exercises	156
	Solutions	159
10	Interpreting Regression Results	163
10.1	Statistical significance	163
10.2	Uses of regression models	163
10.3	Variable selection	163
10.4	Data collection	164
II	Generalized Linear Model	165
11	Generalized Linear Model: Basics	167
11.1	Linear exponential family	167
11.2	Link function	169
11.3	Estimation	171
11.4	Overdispersion	173
	Exercises	174
	Solutions	188
12	Generalized Linear Model: Categorical Response	193
12.1	Binomial response	193
12.2	Nominal response	197
12.3	Ordinal response	199
	Exercises	201
	Solutions	217
13	Generalized Linear Model: Count Response	225
13.1	Poisson response	225
13.2	Overdispersion and negative binomial models	225
13.3	Other count models	226
13.3.1	Zero-inflated models	226
13.3.2	Hurdle models	227
13.3.3	Heterogeneity models	227
13.3.4	Latent models	228
	Exercises	228
	Solutions	234
14	Generalized Linear Model: Measures of Fit	237
14.1	Pearson chi-square	237
14.2	Likelihood ratio tests	238
14.3	Deviance	238

14.4	Penalized loglikelihood tests	241
14.5	Max-scaled R^2 and pseudo- R^2	242
14.6	Residuals	242
	Exercises	245
	Solutions	254
III	Other Statistical Learning Methods	261
15	K-Nearest Neighbors	263
15.1	The Bayes classifier	263
15.2	KNN classifier	264
15.3	KNN regression	265
	Exercises	266
	Solutions	271
16	Decision Trees	277
16.1	Building decision trees	277
16.2	Bagging, random forests, boosting	281
16.2.1	Bagging	281
16.2.2	Random forests	282
16.2.3	Boosting	282
	Exercises	285
	Solutions	299
17	Principal Components Analysis	305
17.1	Loadings and scores	305
17.2	Biplots	307
17.3	Approximation	310
17.4	Scaling	310
17.5	Proportion of variance explained	312
	Exercises	315
	Solutions	324
18	Cluster Analysis	329
18.1	K-means clustering	329
18.2	Hierarchical clustering	331
18.3	Issues with clustering	336
	Exercises	338
	Solutions	345
IV	Time Series	351
19	Time Series: Basics	353
19.1	Introduction	353
19.2	Mean and variance	354
19.3	White noise	355
19.4	Random walks	355
19.5	Control charts	356
19.6	Evaluating forecasts	356
	Exercises	358
	Solutions	364

20 Time Series: Autoregressive Models	371
Exercises	373
Solutions	377
21 Time Series: Forecasting Models	379
21.1 Moving average smoothing	379
21.2 Exponential smoothing	379
21.3 Seasonal models	380
21.4 Unit root tests	381
21.5 ARCH and GARCH models	382
Exercises	383
Solutions	386
V Practice Exams	391
1 Practice Exam 1	393
2 Practice Exam 2	403
3 Practice Exam 3	415
4 Practice Exam 4	425
5 Practice Exam 5	437
6 Practice Exam 6	447
Appendices	457
A Solutions to the Practice Exams	459
Solutions for Practice Exam 1	459
Solutions for Practice Exam 2	464
Solutions for Practice Exam 3	469
Solutions for Practice Exam 4	474
Solutions for Practice Exam 5	480
Solutions for Practice Exam 6	486
B Cross Reference Tables	493

Preface

Welcome to Statistics for Risk Modeling!

This course gives you an introduction to statistical learning and data science. It is a prerequisite for the Predictive Analytics exam.

You should have some knowledge of calculus, probability, and mathematical statistics, and you should know what matrix multiplication, transposition, and inversion means. However, the technical part of this course is light; you should know what we're talking about when we mention testing a null hypothesis H_0 against an alternative, or Student's t test, and it would be nice if you know what maximum likelihood estimation is, but we don't go very deeply into mathematical statistics.

Download the syllabus for the exam. At this writing (July 2023), the September 2023 syllabus is at

<https://www.soa.org/498cac/globalassets/assets/files/edu/2023/fall/syllabi/2023-09-exam-srm-syllabus.pdf>

For more recent syllabi, go to <https://www.soa.org/education/exam-req/edu-exam-srm-detail> and select the syllabus you want.

The syllabus has two links at the bottom. The second one links to sample questions and solutions. There are 64 sample questions.

The syllabus includes the following topics and weights:

Topic	Weight	Lessons	Sample Questions	Number of Sample Questions
Basics of Statistical Learning	7.5–12.5%	1		0
Linear Models	40–50%	2–15	7, 8, 11, 12, 13, 14, 17, 18, 19, 20, 23, 24, 27, 28, 42, 44, 45, 47, 49, 52, 53, 54, 56, 61, 62	25
Time Series Models	10–15%	19–21	3, 4, 21, 22, 31, 38, 46, 55, 58, 64	10
Decision Trees	20–25%	16	9, 10, 25, 26, 33, 39, 41, 48, 50, 51, 57, 63	12
Unsupervised Learning Techniques	10–15%	17–18	1, 2, 5, 6, 15, 16, 29, 30, 32, 34, 35, 36, 37, 40, 43, 59, 60	17

The sample questions are classified by topic in this table.

Here are some comments on the table:

1. It is interesting that the topic “ K nearest neighbors” is not a separate topic of the syllabus; rather, it is included in linear models. In this manual, there is a separate lesson on K nearest neighbors. None of the sample questions are on this topic. It is possible that they never test on it.
2. Linear models is the largest topic, but includes generalized linear models and various other topics.
3. The distribution of the sample questions is different from the syllabus weights. Part of this may be to provide more questions on topics which have not appeared on exams before SRM, such as cluster analysis.

The contents of the syllabus did not change in the September 2023 syllabus, but the weights changed. Here is a comparison of the old and new weights, and the expected number of questions on the exam:

Topic	Pre-Sept. 2023 Syllabus		Sept. 2023 Syllabus	
	Weight	Questions	Weight	Questions
Basics of Statistical Learning	7.5–12.5%	3.5	5–10%	2.625
Linear Models	40–50%	15.75	40–50%	15.75
Time Series Models	12.5%–17.5%	5.25	10–15%	4.375
Decision Trees	10–15%	4.375	20–25%	7.875
Unsupervised Learning Techniques	12.5–22.5%	6.125	10–15%	4.375

The current syllabus puts a much higher weight on decision trees, and removes about one question apiece from basics and time series and two questions from unsupervised learning techniques. Before Sept. 2023, unsupervised learning techniques were split into Principal Components Analysis (2.5–7.5%) and Cluster Analysis (10–15%). Only one lesson in this manual (and one chapter of the syllabus readings) deals with decision trees; you should spend a lot of time on it.

About 60% of the sample questions are conceptual; no calculations are needed. Some of these questions are taken from obscure passages in the Frees textbook. All of the concepts in these questions are covered in this manual, but in some cases very briefly. However, there is no guarantee that they won't ask a question on something that I didn't include. I believe that knowing the information in this manual will be enough to get a 10, but not necessarily enough to answer every exam question.

You should also download the tables that are linked to the bottom of the syllabus. The tables include the normal distribution, critical values for the t distribution, and critical values for the chi-square distribution. The SOA hasn't provided rounding rules for the normal table, but you won't be using it that heavily anyway. Possibly the SOA will provide a Prometric calculator or a spreadsheet instead of providing tables.

There are two textbooks on the syllabus. There is some overlap between the two textbooks, as they both discuss linear regression and generalized linear models. The styles of the two textbooks are different.

The first textbook is *Regression Modeling with Actuarial and Financial Applications* by Edward Frees, an actuary. This book covers the linear models and time series parts of the syllabus. The author comes across as a scholar who is very familiar with his material and tries to get it across by showing many practical examples of its use. Practical examples means computer outputs. To make the book more readable, technical detail is usually placed in a section at the end of each chapter (and those sections are not on the SRM syllabus). Despite the author's good intentions, I found this book somewhat difficult to read for the following reasons:

1. The book has a fairly large number of errors. The errata list for the book is at

<https://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/RegressionFreesErrata12September.pdf>

The errata list must be taken into account, since many formulas in the book are incorrect.

2. The book lacks a useful index. And the table of contents only lists chapters and sections, not subsections. Isn't it interesting that the index does not have anything starting with F ? (The F -ratio *is* discussed in this textbook, although the author seems to prefer using t -ratios.) If you wanted to know something about Cook's distance, where would you find it? (If you were really clever and knew that Cook's distance had something to do with leverage, you could look under leverage and find Cook's distance there, but hey - an index should be usable by dummies!) If some practice question mentioned "Dickey-Fuller" and you wanted to know what it is, where would you look? (If you knew that the full name of the test is the "Dickey-Fuller unit root test", you would still not find "unit root tests" in the index, but you would find it in the table of contents. Hurrah!) I was frustrated by the difficulty of finding things in the book.
3. It is often very hard to understand what the author is saying without knowing the technical background, and often the technical background is not provided in the last section of the chapter.

I have omitted some of the more obscure topics from this textbook, and I don't think they would appear on an exam.¹

The second textbook is *An Introduction to Statistical Learning*, coauthored by four non-actuarial authors. This textbook covers all parts of the syllabus except time series, but does not discuss logistic models.² This book is available free as a download, and I encourage you to download it! The style of this book is enthusiasm; these authors are excited about this topic and want you to be excited as well! You can read this book in bed, and I challenge you to find an error in it³. *An Introduction to Statistical Learning* avoids technical details as much as possible, and rather than have you do calculations, shows you how to use R to carry out the modeling.

You will find *An Introduction to Statistical Learning* easier to read than this manual. However, this manual will still help you in the following ways:

1. It summarizes the material. You can pick up the material faster by reading this manual, although you will not be motivated as much. You may even want to read *An Introduction to Statistical Learning* once and then this manual many times to review the material.
2. It provides you with exam-like examples and questions. *An Introduction to Statistical Learning* is interested in teaching you practical uses of the material. It only provides a small number of simple small-scale examples, and you may want more. Calculation questions on Exam SRM are simple and small-scale.
3. On rare occasions *An Introduction to Statistical Learning* is not completely clear.

Much of the material on this exam does not lend itself to calculation by hand. Therefore, much of this exam will contain knowledge questions rather than calculation questions. Knowledge questions are three-way or four-way true/false questions. On three-way true/false questions, the SOA enforces symmetry of the answer choices. The only two sets of answer choices that you will encounter on an exam are

(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).

and

(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).

In both cases, there is symmetry among the three statements I, II, and III, and also symmetry in how likely a statement is to be true. Apparently symmetry is not required for four-way true/false questions. For either type, choice E should be the correct choice about 1/5 of the time only.

Exercises and practice exams

All questions on the practice exams in this manual are original. However, a small number of the calculation questions are slight variations of the SRM sample questions. For the knowledge questions, there was a limit to the number of different tricky statements I could come up, so you may find them slightly repetitive. For decision trees, coming up with 7 questions per exam was hard, and you may find these questions a bit repetitive as well.

Many exercises are original, but I also used SRM sample questions and questions from old CAS MAS-I, MAS-II, and S exams. Let's discuss how appropriate CAS questions are.

First of all, the CAS style is in general different from the SOA style. It tends to be less precise. And answer choices are usually ranges rather than specific answers. On three-way true/false questions, they used to not insist on symmetry, but for the last few years they do; all the MAS-I and MAS-II three-way true/false questions follow the symmetry rules.

In terms of syllabus, the CAS uses the same textbook, *An Introduction to Statistical Learning*, for the topics that are covered in that textbook. They use it even for logistic regression. However, the CAS uses textbooks for linear models and time series *different* from the ones the SOA uses. The time series coverage on MAS-I has very little overlap

¹In particular, I do not discuss the multinomial logit model

²Actually it does discuss logistic models, but the chapter that discusses them is not on the syllabus.

³Because of the non-technical style, you will find inexact statements, but I don't consider those errors.

with the time series coverage on SRM. For generalized linear models, different authors use different terminology and different parametrizations. For example, what we call “linear exponential family” is called “exponential family in canonical form” in the CAS textbook, and is parametrized differently as well. When different terminology and parametrizations were used, I translated CAS questions whenever possible, otherwise I did not use the question.

New for this edition

The newer SOA sample questions, up to question 64, were added to the appropriate lessons.

The practice exams were updated to reflect the updated syllabus weights, including the heavier weight on decision trees. Questions that were removed from the practice exams were moved to the appropriate lessons.

The R language

This manual does not cover R, and you won’t need it for SRM. However, you will need it for the PA exam. You should read the labs in *An Introduction to Statistical Learning* to learn how to use R to carry out the statistical learning methods you learn in this course.

Cross-reference tables

Note that Appendix B has cross-reference tables showing you which section of the manual corresponds to each section in the textbooks. As discussed in that appendix, these may be helpful when you are studying for Exam PA.

Errata

Please report all errors to the author. You may send them to the publisher at mail@studymaterials.com or directly to me at errata@aceyourexams.net.

An errata list will be posted at <http://errata.aceyourexams.net>. Check this errata list frequently.

Acknowledgements

I would like to thank the CAS for allowing me to use questions from their old exams, and the SOA for allowing me to use its sample questions.

The creators of \TeX , \LaTeX , and its multitude of packages all deserve thanks for making possible the professional typesetting of this mathematical material.

I’d like to thank Michael Bean for his diligent job proofreading this manual, as well as advice on how to improve the content.

I’d like to thank the following correspondents who submitted errata: Hiu Tung Chan, Cheng Chen, Joel Cheung, Maria Doran, Neil Xavier Elpa, Lingyi Fang, Natalie Jacobsen, Boren Jiang, Dan Kamka, Drew Lehe, Yingxin Liu, Mario Mendiola, Li Kee Ong, Greg Schlottbohm, Aaron Shotkin, Tara Starling, Ryan Talley, Chan Hiu Tung, Isaac Zhang, Wei Zhao, Dihui Zhu.

Lesson 2

Linear Regression: Estimating Parameters

Reading: *Regression Modeling with Actuarial and Financial Applications* 1.3, 2.1–2.2, 3.1–3.2; *An Introduction to Statistical Learning* 3.1–3.2, 3.3.2, 3.3.3

In a **linear regression model**, we have a variable y that we are trying to explain using variables x_1, \dots, x_k .¹ We have n observations of sets of k **explanatory variables** and their **responses**: $\{y_i, x_{i1}, x_{i2}, \dots, x_{ik}\}$ with $i = 1, \dots, n$. We would like to relate y to the set of $x_j, j = 1, \dots, k$ as follows:

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

where ε_i is an **error term**. We estimate the vector $\beta = (\beta_0, \beta_1, \dots, \beta_k)$ by selecting the vector that minimizes $\sum_{i=1}^n \varepsilon_i^2$.

For statistical purposes, ε_i is a **random variable**. We make the following assumptions about these random variables:

1. $E[\varepsilon_i] = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$. In other words, the variance of each error term is the same. This assumption is called **homoscedasticity** (sometimes spelled homoskedasticity).
2. ε_i are independent.
3. ε_i follow a normal distribution.

If these assumptions are valid, then for any set of values of the k variables $\{x_1, x_2, \dots, x_k\}$, the resulting value of y will be normally distributed with mean $\beta_0 + \sum_{i=1}^k \beta_i x_i$ and variance σ^2 . Moreover, the estimate of β is the **maximum likelihood estimate**.

Notice that our linear model has k parameters $\beta_1, \beta_2, \dots, \beta_k$ in addition to the constant β_0 . Thus we are really estimating $k + 1$ parameters. Some authors refer to “ $k + 1$ variable regression”. I’ve never been sure whether this is because $k + 1$ β s are estimated or because the response variable is counted as a variable.

2.1 Basic linear regression

When $k = 1$, the model is called “basic linear regression” or “**simple linear regression**”.² In this case, the formulas for the estimators of β_0 and β_1 are

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \quad (2.1)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (2.2)$$

Often we use Latin letters for the estimators of Greek parameters, so we can write b_i instead of $\hat{\beta}_i$.³

The formula for $\hat{\beta}_1$ can be expressed as the quotient of the covariance of x and y over the variance of x . The **sample covariance** is

$$cov_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

and the **sample variance** is

¹*Regression Modeling with Actuarial and Financial Applications* uses k for the number of variables, but *An Introduction to Statistical Learning* uses p .

²*Regression Modeling with Actuarial and Financial Applications* calls it basic linear regression and *An Introduction to Statistical Learning* calls it simple linear regression. As indicated in the previous paragraph, some authors call it “2 variable regression”, and while this terminology is not used by either textbook, you may find it on old exam questions.

³*Regression Modeling with Actuarial and Financial Applications* uses b_i , while *An Introduction to Statistical Learning* uses $\hat{\beta}_i$.

$$s_x^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

The $n - 1$ s cancel when division is done, so they may be ignored. Then equation (2.1) becomes

$$\hat{\beta}_1 = \frac{cv_{xy}}{s_x^2}$$

You may use the usual shortcuts to calculate **variance** and **covariance**:

$$\text{Cov}(X, Y) = \mathbf{E}[XY] - \mathbf{E}[X] \mathbf{E}[Y]$$

$$\text{Var}(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2$$

In the context of sample data, if we use the biased sample variance and covariance with division by n rather than $n - 1$ (It doesn't really matter whether biased or unbiased is used, since the denominators of the sums, whether they are n or $n - 1$, will cancel when one is divided by the other.), these formulas become

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{\sum x_i \sum y_i}{n} = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{(\sum x_i)^2}{n} = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum_{i=1}^n x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

Let s_x, s_y , be the **sample standard deviations** of x and y , and let r_{xy} be the **sample correlation** of x and y , defined as follows:

$$r_{xy} = \frac{cv_{xy}}{s_x s_y}$$

From formula (2.1), we have $\hat{\beta}_1 = \frac{r_{xy} s_x s_y}{s_x^2}$, or

$$\hat{\beta}_1 = r_{xy} \frac{s_y}{s_x} \quad (2.3)$$

so $\hat{\beta}_1$ is proportional to the correlation of x and y .

EXAMPLE 2A You are given the linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ to fit to the following data:

x	2	2	3	4	5	5	7
y	34	38	38	53	50	60	70

Determine the least squares estimate of β_1 . ■

SOLUTION: First we calculate $\sum x_i^2$ and $\sum x_i y_i$, then we subtract $n\bar{x}^2$ and $n\bar{x}\bar{y}$. We obtain:

$$\sum x_i^2 = 132$$

$$\sum x_i y_i = 1510$$

$$\bar{x} = \frac{28}{7} = 4$$

$$\bar{y} = \frac{343}{7} = 49$$

$$\sum (x_i - \bar{x})^2 = 132 - 7(4^2) = 20$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 1510 - 7(4)(49) = 138$$

$$\hat{\beta}_1 = \frac{138}{20} = \mathbf{6.9}$$

Although not required by the question, we can easily calculate $\hat{\beta}_0$:

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ &= 49 - (6.9)(4) = 21.4\end{aligned}$$

□

You would never go through the calculations of the previous example since your calculator can carry out the regression. On the TI-30XS, use data, ask for 2-Var statistics. In those statistics, item D is β_1 (with the unusual name a) and item E is β_0 (with the unusual name b). You can try this out on this quiz:



Quiz 2-1 For a new product released by your company, revenues for the first 4 months, in millions, are:

Month 1	27
Month 2	34
Month 3	48
Month 4	59

Revenues are assumed to follow a linear regression model of the form

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

where x_i is the month and y_i is revenues.

Estimate β_1 for this model.

More likely, an exam question would give you summary statistics only and you'd use the formulas to get $\hat{\beta}_0$ and $\hat{\beta}_1$.

EXAMPLE 2B For 8 observations of X and Y , you are given:

$$\bar{x} = 6 \qquad \bar{y} = 8 \qquad \sum x_i^2 = 408 \qquad \sum x_i y_i = 462$$

Perform a simple linear regression of Y on X :

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

Determine $\hat{\beta}_0$.

SOLUTION:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} \\ &= \frac{462 - 8(6)(8)}{408 - 8(6^2)} = 0.65\end{aligned}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 8 - 0.65(6) = \mathbf{4.1}$$

The next example illustrates predicting an observation using the regression model.

EXAMPLE 2C Experience for four cars on an automobile liability coverage is given in the following chart:

Miles Driven	7,000	10,000	11,000	12,000
Aggregate Claim Costs	600	2000	1000	1600

A least squares model relates aggregate claims costs to miles driven.

Calculate predicted aggregate claims costs for a car driven 5000 miles.

SOLUTION: We let x_i be miles driven and y_i aggregate claim costs. It is convenient to drop thousands both in miles driven and aggregate claim costs.

$$\begin{aligned}\bar{x} &= \frac{7 + 10 + 11 + 12}{4} = 10 & \bar{y} &= \frac{0.6 + 2 + 1 + 1.6}{4} = 1.3 \\ \sum x_i^2 &= 7^2 + 10^2 + 11^2 + 12^2 = 414 & \sum x_i y_i &= (7)(0.6) + (10)(2) + (11)(1) + (12)(1.6) = 54.4 \\ \text{denominator} &= 414 - (4)(10^2) = 14 & \text{numerator} &= 54.4 - (4)(10)(1.3) = 2.4 \\ \hat{\beta}_1 &= \frac{2.4}{14} = \frac{6}{35} & \hat{\beta}_0 &= 1300 - \left(\frac{6}{35}\right)(10000) = -\frac{2900}{7}\end{aligned}$$

Notice that we multiplied back by 1000 when calculating $\hat{\beta}_0$.

The predicted value is therefore $-\frac{2900}{7} + \frac{6}{35}(5000) = \mathbf{442.8571}$. □

The fitted value of y_i , or $\hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_{ij}$, is denoted by \hat{y}_i . The difference between the actual and fitted values of y_i , or $\hat{\varepsilon}_i = y_i - \hat{y}_i$, is called the *residual*. As a result of the equations that are used to solve for $\hat{\beta}$, *the sum of the residuals $\sum_{i=1}^n \hat{\varepsilon}_i$ on the training set is always 0*. As with $\hat{\beta}_i$, we may use Latin letters instead of hats and denote the residual by e_i .

2.2 Multiple linear regression

We started the lesson with this equation:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \quad (*)$$

Let's now discuss *multiple regression*, the case when $k > 1$. We then have k *explanatory variables* plus an *intercept* and n values for each one. We can arrange these into an $n \times (k + 1)$ matrix:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

Notice how the intercept was turned into a variable of 1s. Set $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$. Then equation (*) can be

written like this:

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$$

\mathbf{X} is called the *design matrix*.⁴ The generalized formulas for linear regression use matrices. We will use lower case boldface letters for column and row vectors and upper case boldface letters for matrices with more than one row and column. We will use a prime on a matrix to indicate its transpose. The *least squares estimate of $\boldsymbol{\beta}$* is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (2.4)$$

and then the fitted value of y is $\hat{y} = \mathbf{X}\hat{\boldsymbol{\beta}}$. I doubt you'd be expected to use formula (2.4) on an exam, unless you were given $(\mathbf{X}'\mathbf{X})^{-1}$, since it involves inverting a large matrix. In fact, I doubt you will be asked any questions requiring matrix multiplication.

The $(\mathbf{X}'\mathbf{X})^{-1}$ matrix is singular (non-invertible) if there is a linear relationship among the column vectors of \mathbf{X} . Therefore, it is important that the column vectors not be collinear. Even if the variables are only "almost" collinear, the regression is unstable. We will discuss tests for *collinearity* in Section 5.3.

⁴The reason for this name is that in some scientific experiments, the points \mathbf{x} are chosen for the experiment. But this will generally not be the case for insurance studies.

As with simple linear regression, the sum of the residuals is 0.

When an explanatory variable is a **categorical variable** with m possible values, you must include $m - 1$ indicator variables in the model. Sometimes indicator variables are called “**dummy variables**”. Each indicator variable corresponds to one possible value of the categorical variable. It is equal to 1 if the variable is equal to that value, 0 otherwise.

For example, if one of the explanatory variables is sex (male or female), you would set up one indicator variable for either male or female. If the indicator variable is for female, it would equal 0 if male and 1 if female. If one of the explanatory variables is age bracket and there are 5 age brackets, you would set up 4 indicator variables for 4 of the 5 age brackets. Notice that if you set up 5 variables, their sum would equal 1. The sum would be identical to x_0 , the first column vector of X , resulting in a linear relationship among columns of the matrix, which would make it singular. Thus one variable must be omitted. The omitted variable is called the **base level** or **reference level**. You should select the value that occurs most commonly as the base level. If you select a value that is almost always 0, then the sum of the other indicator variables will almost always be 1, making the computation of the inverse of $X'X$ less stable.

A special case is a variable with only two categories. The indicator variable is then a binary variable.

2.3 Alternative model forms

Even though regression is a linear model, it is possible to incorporate nonlinear explanatory variables. Powers of variables may be included in the model. For example, you can estimate

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i1}^3 + \varepsilon_i$$

You can include **interaction** between explanatory variables by including a term multiplying them together:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$$

Another possibility is a regression with an exponential:

$$y_i = \beta_0 + \beta_1 e^{x_i} + \varepsilon_i$$

Linear regression assumes homoscedasticity, linearity, and normality. If these assumptions aren't satisfied, sometimes a few adjustments can be made to make the data satisfy these conditions.

Suppose the variance of the observations varies in a way that is known in advance. In other words, we know that $\text{Var}(\varepsilon_i) = \sigma^2/w_i$, with w_i varying by observation, although we don't necessarily know what σ^2 is. Then w_i is the **precision of observation** i , with $w_i = 0$ for an observation with no precision (which we would have to discard) and $w_i \rightarrow \infty$ for an exact observation. We can then multiply all the variables in observation i by $\sqrt{w_i}$. After this multiplication, all observations will have the same variance. Let W be the diagonal matrix with w_i in the i^{th} position on the diagonal, 0 elsewhere. Then equation (2.4) would be modified to

$$\hat{\beta}^* = (X'WX)^{-1}X'W\mathbf{y} \quad (2.5)$$

The estimator $\hat{\beta}^*$ is called the **weighted least squares estimator**.

One may also transform y to levelize the variance or to remove skewness. If variance appears to be proportional to y , logging y may levelize the variance:

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

which is equivalent to

$$y_i = e^{\beta_0 + \beta_1 x_i + \varepsilon_i}$$

In this model, $\ln y_i$ is assumed to have a normal distribution, which means that y_i is lognormal. A **lognormal distribution** is skewed to the right, so logging y may remove skewness.

A general family of power transformations is the **Box-Cox family of transformations**:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases} \quad (2.6)$$

This family includes taking y to any power, positive or negative, and logging. Adding a constant and dividing by a constant does not materially affect the form of a linear regression; it merely changes the intercept and scales the β coefficients. So $(y^\lambda - 1)/\lambda$ could just as well be y^λ . The only reason to subtract 1 and divided by λ is so that as $\lambda \rightarrow 0$, $(y^\lambda - 1)/\lambda \rightarrow \ln y$.

I doubt that the exam will require you to calculate parameters of regression models. Do a couple of the calculation exercises for this lesson just in case, but don't spend too much time on them.

Table 2.1: Summary of Linear Model Formulas


For a simple regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$	
	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (2.2)$
	$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \quad (2.1)$
	$\hat{\beta}_1 = r_{xy} \frac{s_y}{s_x} \quad (2.3)$
For a multiple variable regression model	$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (2.4)$
For any regression	$\sum_{i=1}^n e_i = 0$
For a weighted least squares model	$\hat{\beta}^* = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y} \quad (2.5)$
Box-Cox power transformations	$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases} \quad (2.6)$

Exercises

- 2.1. You are given the linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ to fit to the following data:


x	-2	-1	0	1	2
y	3	5	8	9	10

Determine the least squares estimate of β_0 .

- 2.2.  You are fitting a linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ to 18 observations. You are given the following:


- (i) $\sum_{i=1}^{18} x_i = 216$
- (ii) $\sum_{i=1}^{18} x_i^2 = 3092$
- (iii) $\sum_{i=1}^{18} y_i = 252$
- (iv) $\sum_{i=1}^{18} y_i^2 = 4528$
- (v) $\sum_{i=1}^{18} x_i y_i = 3364$

Determine the least squares estimate of β_1 .

- 2.3.  [SRM Sample Question #17] The regression model is $y = \beta_0 + \beta_1 x + \varepsilon$. There are six observations. The summary statistics are:


$$\sum y_i = 8.5 \quad \sum x_i = 6 \quad \sum x_i^2 = 16 \quad \sum x_i y_i = 15.5 \quad \sum y_i^2 = 17.25$$

Calculate the least squares estimate of β_1 .

- (A) 0.1 (B) 0.3 (C) 0.5 (D) 0.7 (E) 0.9
- 2.4.  [SRM Sample Question #47] You are given the following summary statistics:


$$\begin{aligned} \bar{x} &= 3.500 \\ \bar{y} &= 2.840 \\ \sum (x_i - \bar{x})^2 &= 10.820 \\ \sum (x_i - \bar{x})(y_i - \bar{y}) &= 2.677 \\ \sum (y_i - \bar{y})^2 &= 1.125 \end{aligned}$$

Determine the equation of the regression line, using the least squares method.

- (A) $y = 1.97 + 0.25x$
 (B) $y = 0.78 + 0.59x$
 (C) $y = 0.57 + 0.65x$
 (D) $y = 0.39 + 0.70x$
 (E) The correct answer is not given by (A), (B), (C), or (D).
- 2.5.  [SRM Sample Question #53] Determine which of the following statements is NOT true about the equation

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- (A) β_0 is the expected value of Y .
- (B) β_1 is the average increase in Y associated with a one-unit increase in X .
- (C) The error term, ε is typically assumed to be independent of X and Y .
- (D) The equation defines the population regression line.
- (E) The method of least squares is commonly used to estimate the coefficients β_0 and β_1 .

2.6.  [SRM Sample Question #23] Toby observes the following coffee prices in his company cafeteria:


- 12 ounces for 1.00
- 16 ounces for 1.20
- 20 ounces for 1.40

The cafeteria announces that they will begin to sell any amount of coffee for a price that is the value predicted by a simple linear regression using least squares of the current prices on size.

Toby and his co-worker Karen want to determine how much they would save each day, using the new pricing, if, instead of each buying a 24-ounce coffee, they bought a 48-ounce coffee and shared it.

Calculate the amount they would save.


- (A) It would cost them 0.40 more.
- (B) It would cost the same.
- (C) They would save 0.40.
- (D) They would save 0.80.
- (E) They would save 1.20.

2.7.  [MAS-I-F18:29] An ordinary least squares model with one variable (Advertising) and an intercept was fit to the following observed data in order to estimate Sales:

Observation	Advertising	Sales
1	5.5	100
2	5.8	110
3	6.0	112
4	5.9	115
5	6.2	117


Calculate the residual for the third observation.

- (A) Less than -2
- (B) At least -2 , but less than 0
- (C) At least 0, but less than 2
- (D) At least 2, but less than 4
- (E) At least 4

2.8.  You are fitting the linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ to the following data:

x	2	5	8	11	13	15	16	18
y	-10	-9	-4	0	4	5	6	8

Determine the least squares estimate of β_1 .

2.9.  You are fitting the linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ to the following data:

x	3	5	7	8	9	10
y	2	5	7	8	9	11

Determine the fitted value of y corresponding to $x = 6$.

2.10. You are fitting the linear regression model $\hat{y}_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. You are given:

- (i) $\sum_{i=1}^{28} x_i = 392$
- (ii) $\sum_{i=1}^{28} y_i = 924$
- (iii) $\sum_{i=1}^{28} x_i y_i = 13,272$
- (iv) $\hat{\beta}_0 = -23$

Determine $\sum_{i=1}^{28} x_i^2$.

2.11. [3-F84:5] You are fitting the linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ to 10 points of data. You are given:

$$\begin{aligned}\sum x_i &= 100 \\ \sum y_i &= 200 \\ \sum x_i y_i &= 2000 \\ \sum x_i^2 &= 2000 \\ \sum y_i^2 &= 5000\end{aligned}$$

Calculate the least-squares estimate of β_1 .

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

2.12. The following model relates Y to X :

$$Y_i = \exp(\beta_0 + \beta_1 X_i + \varepsilon_i)$$

For 10 observations, you are given:

- (i) $\sum X_i = 13.5$
- (ii) $\sum \ln Y_i = 9.681$
- (iii) $\sum X_i^2 = 19.57$
- (iv) $\sum (\ln Y_i)^2 = 9.690$
- (v) $\sum X_i \ln Y_i = 13.697$

Calculate the prediction of Y when $X = 2$.

- (A) 3.2 (B) 3.6 (C) 4.0 (D) 4.4 (E) 4.8

- 2.13. [3L-S05:27] Given the following information:

$$\begin{aligned}\sum x_i &= 144 \\ \sum y_i &= 1,742 \\ \sum x_i^2 &= 2,300 \\ \sum y_i^2 &= 312,674 \\ \sum x_i y_i &= 26,696 \\ n &= 12\end{aligned}$$

Determine the least squares equation for the following model:

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

- (A) $\hat{y}_i = -0.73 + 12.16x_i$
 (B) $\hat{y}_i = -8.81 + 12.16x_i$
 (C) $\hat{y}_i = 283.87 + 10.13x_i$
 (D) $\hat{y}_i = 10.13 + 12.16x_i$
 (E) $\hat{y}_i = 23.66 + 10.13x_i$
- 2.14. [120-F90:6] You are estimating the linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. You are given


i	1	2	3	4	5
x_i	6.8	7.0	7.1	7.2	7.4
y_i	0.8	1.2	0.9	0.9	1.5

Determine $\hat{\beta}_1$.

- (A) 0.8 (B) 0.9 (C) 1.0 (D) 1.1 (E) 1.2
- 2.15. [120-S90:11] Which of the following are valid expressions for b , the slope coefficient in the simple linear regression of y on x ?

I. $\frac{(\sum x_i y_i) - \bar{y} \sum x_i}{(\sum x_i^2) - \bar{x} \sum x_i}$
 II. $\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum x_i^2 - \bar{x}^2}$
 III. $\frac{\sum x_i(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

- (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).


- 2.16.  [Old exam] For the linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon$ with 30 observations, you are given:

- (i) $r_{xy} = 0.5$
- (ii) $s_x = 7$
- (iii) $s_y = 5$


where r_{xy} is the sample correlation coefficient.

Calculate the estimated value of β_1 .

- (A) 0.4 (B) 0.5 (C) 0.6 (D) 0.7 (E) 0.8

- 2.17.  [110-S83:14] In a bivariate distribution the regression of the variable y on the variable x is $1500 + b(x - 68)$ for some constant b . If the correlation coefficient is 0.81 and if the standard deviations of y and x are 220 and 2.5 respectively, then what is the expected value of y , to the nearest unit, when x is 70?

- (A) 1357 (B) 1515 (C) 1517 (D) 1643 (E) 1738


- 2.18.  [120-82-97:7] You are given the following information about a simple regression model fit to 10 observations:

$$\begin{aligned}\sum x_i &= 20 \\ \sum y_i &= 100 \\ s_x &= 2 \\ s_y &= 8\end{aligned}$$

You are also given that the correlation coefficient $r_{xy} = -0.98$.

Determine the predicted value of y when $x = 5$.

- (A) -10 (B) -2 (C) 11 (D) 30 (E) 37

- 2.19.  In a simple regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, you are given

$$\begin{aligned}\sum x_i &= 30 & \sum y_i &= 450 \\ \sum x_i^2 &= 270 & \sum x_i y_i &= 8100 \\ n &= 15 & & \\ x_5 &= 3 & y_5 &= 40\end{aligned}$$

Calculate the fifth residual, $\hat{\varepsilon}_5$.

2.20. [120-F89:13] 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:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 1522.73 & 26.87 & -374.67 \\ 26.87 & 0.93 & -7.33 \\ -374.67 & -7.33 & 93.33 \end{pmatrix}$$

Calculate $\hat{\varepsilon}_2$.

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


2.21. You are fitting the following data to a linear regression model of the form $y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$:

y	5	3	10	4	3	5
x_1	0	1	0	1	0	1
x_2	1	0	1	1	0	1
x_3	0	1	1	0	0	0

You are given that

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{30} \begin{pmatrix} 26 & -10 & -18 & -12 \\ -10 & 20 & 0 & 0 \\ -18 & 0 & 24 & 6 \\ -12 & 0 & 6 & 24 \end{pmatrix}$$

Determine the least squares estimate of β_1 .

- 2.22.  [120-82-94:11] An automobile insurance company wants to use gender ($x_1 = 0$ if female, 1 if male) and traffic penalty points (x_2) to predict the number of claims (y). The observed values of these variables for a sample of six motorists are given by:

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


You are to use the following model:

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

You have determined

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

Determine $\hat{\beta}_2$.


- (A) -0.25 (B) 0.25 (C) 1.25 (D) 2.00 (E) 4.25
- 2.23.  You are fitting the following data to the linear regression model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$:

y	1	2	6	5	1	2	3
x_1	0	0	1	-1	0	1	1
x_2	0	-1	0	0	1	-1	0
x_3	1	1	4	0	0	0	1

You are given that

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{30} \begin{pmatrix} 7 & 0 & 1.5 & -2.5 \\ 0 & 1.2 & 3 & -3 \\ 1.5 & 3 & 11.25 & -0.75 \\ -2.5 & -3 & -0.75 & 3.25 \end{pmatrix}.$$

Determine the fitted value of y for $x_1 = x_2 = x_3 = 1$.

- 2.24.  **[Old exam]** 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

x_3 = 1, if a holiday weekend occurs during the month, and 0 otherwise


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:

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{6506} \begin{pmatrix} 257 & -82 & -446 \\ -82 & 254 & -364 \\ -446 & -364 & 2622 \end{pmatrix}$$


Determine $\hat{\beta}_1$.

- (A) -0.07 (B) 0.15 (C) 0.24 (D) 0.70 (E) 1.30
- 2.25.  **[S-F15:35]** You are given a regression model of liability claims with the following potential explanatory variables only:

- Vehicle price, which is a continuous variable modeled with a third order polynomial
- Average driver age, which is a continuous variable modeled with a first order polynomial
- Number of drivers, which is a categorical variable with four levels
- Gender, which is a categorical variable with two levels
- There is only one interaction in the model, which is between gender and average driver age.

Determine the maximum number of parameters in this model.

- (A) Less than 9 (B) 9 (C) 10 (D) 11 (E) At least 12

- 2.26.  [MAS-I-S18:37] You fit a linear model using the following two-level categorical variables:

$$X_1 = \begin{cases} 1 & \text{if Account} \\ 0 & \text{if Monoline} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if Multi-Car} \\ 0 & \text{if Single Car} \end{cases}$$

with the equation

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

This model produced the following parameter estimates:

$$\begin{aligned} \beta_0 &= -0.10 \\ \beta_1 &= -0.25 \\ \beta_2 &= 0.58 \\ \beta_3 &= -0.20 \end{aligned}$$

Another actuary modeled the same underlying data, but coded the variables differently as such:

$$X_1 = \begin{cases} 0 & \text{if Account} \\ 1 & \text{if Monoline} \end{cases}$$

$$X_2 = \begin{cases} 0 & \text{if Multi-Car} \\ 1 & \text{if Single Car} \end{cases}$$

with the equation

$$E[Y] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1 X_2$$

Afterwards you make a comparison of the individual parameter estimates in the two models.

Calculate how many pairs of coefficient estimates $(\hat{\alpha}_i, \hat{\beta}_i)$ switched signs, and how many pairs of estimates stayed identically the same, when results of the two models are compared.

- (A) 1 sign change, 0 identical estimates
- (B) 1 sign change, 1 identical estimate
- (C) 2 sign changes, 0 identical estimates
- (D) 2 sign changes, 1 identical estimate
- (E) The correct answer is not given by (A), (B), (C), or (D).

2.27. [MAS-I-S19:32] You are fitting a linear regression model of the form:

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

and are given the following values used in this model:

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 1 & 9 \\ 1 & 1 & 1 & 15 \\ 1 & 1 & 1 & 8 \\ 0 & 1 & 1 & 7 \\ 0 & 1 & 1 & 6 \\ 0 & 0 & 1 & 6 \end{pmatrix}; \mathbf{y} = \begin{pmatrix} 21 \\ 32 \\ 19 \\ 17 \\ 15 \\ 15 \end{pmatrix}; \mathbf{X}'\mathbf{X} = \begin{pmatrix} 3 & 2 & 3 & 32 \\ 2 & 4 & 4 & 36 \\ 3 & 4 & 6 & 51 \\ 32 & 36 & 51 & 491 \end{pmatrix}; \mathbf{X}'\mathbf{X}^{-1} = \begin{pmatrix} 1.38 & 0.25 & 0.54 & -0.16 \\ 0.25 & 0.84 & -0.20 & -0.06 \\ 0.54 & -0.20 & 1.75 & -0.20 \\ -0.16 & -0.06 & -0.20 & 0.04 \end{pmatrix}$$

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

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

Calculate the modeled estimate of the intercept parameter.

- (A) Less than 0
- (B) At least 0, but less than 1
- (C) At least 1, but less than 2
- (D) At least 2, but less than 3
- (E) At least 3

- 2.28. [MAS-I-S19:29] Tim uses an ordinary least squares regression model to predict salary based on Experience and Gender. Gender is a qualitative variable and is coded as follows:

$$\text{Gender} = \begin{cases} 1 & \text{if Male} \\ 0 & \text{if Female} \end{cases}$$

His analysis results in the following output:

Coefficients	Estimate	Std. Error	t-value	Pr(> t)
Intercept	18169.300	212.2080	85.62027	2.05E-14
Experience	1110.233	59.8224	18.55881	1.75E-08
Gender	169.550	162.9177	10.38285	2.62E-06

Abby uses the same data set but codes gender as follows:

$$\text{Gender} = \begin{cases} 1 & \text{if Female} \\ 0 & \text{if Male} \end{cases}$$

Calculate the value of the Intercept in Abby's model.

- (A) At most 18,169.3
 (B) Greater than 18,169.3, but at most 18,400.0
 (C) Greater than 18,400.0, but at most 18,600.0
 (D) Greater than 18,600.0
 (E) The answer cannot be computed from the information given

Solutions

2.1. $\bar{x} = 0$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, so $\hat{\beta}_0 = \bar{y} = \mathbf{7}$.

2.2.

$$\sum (x_i - \bar{x})^2 = 3092 - \frac{216^2}{18} = 500$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 3364 - \frac{(216)(252)}{18} = 340$$

$$\hat{\beta}_1 = \frac{340}{500} = \mathbf{0.68}$$

- 2.3. The least squares estimate of β_1 is the covariance of x and y divided by the variance of x . In the following calculation, the numerator is n times the covariance and the denominator is n times the variance; the n s cancel. We have $n = 6$ observations.

$$b_1 = \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sum x_i^2 - (\sum x_i)^2 / n}$$

$$= \frac{15.5 - (6)(8.5)/6}{16 - 6^2/6} = \mathbf{0.7} \quad (\text{D})$$

- 2.4. Using equations (2.1) and (2.2),

$$\hat{\beta}_1 = \frac{2.677}{10.820} = 0.2474$$

$$\hat{\beta}_0 = 2.840 - 0.2474(3.500) = 1.9741 \quad (\mathbf{A})$$

- 2.5. β_0 is *not* the expected value of Y . In fact, $\beta_0 = \bar{Y} - \beta_1 \bar{X}$. (A)
- 2.6. The observations already lie in a straight line; each 4 ounce increase raises the price 0.20. The slope is therefore $0.2/4 = 0.05$ and the intercept (using 12 ounces = 1 = $0.05(12) + \beta_0$) is 0.4. By buying 48 ounces, one intercept, or **0.40**, is saved. (C)
- 2.7. An exam question like this asking you to carry out a linear regression is rare. You can carry out a linear regression on your calculator without knowing the formulas. But anyway, here is the calculation, with X being advertising and Y being sales.

$$\begin{aligned} \sum X_i &= 29.4 \\ \sum Y_i &= 554 \\ \sum X_i^2 - \frac{(\sum X_i)^2}{5} &= 0.268 \\ \sum X_i Y_i - \frac{(\sum X_i \sum Y_i)}{5} &= 6.38 \\ \hat{\beta} &= \frac{6.38}{0.268} = 23.806 \\ \hat{a} &= \frac{554}{5} - 23.806 \left(\frac{29.4}{5} \right) = -29.179 \\ \hat{y}_3 &= -29.179 + 23.806(6.0) = 113.657 \end{aligned}$$

The third residual is $112 - 113.657 = \mathbf{-1.657}$. (B)

- 2.8. In the following, on the third line, because $\bar{y} = 0$, $\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum(x_i - \bar{x})y_i$.

$$\begin{aligned} \bar{x} &= 11 \\ \sum x_i^2 &= 1188 & \sum (x_i - \bar{x})^2 &= 1188 - 8(11^2) = 220 \\ \bar{y} &= 0 & \sum (x_i - \bar{x})(y_i - \bar{y}) &= 2(-10) + 5(-9) + \dots + 18(8) = 270 \\ \hat{\beta}_1 &= \frac{270}{220} = \mathbf{1.2273} \end{aligned}$$


2.9.

$$\begin{aligned} \bar{x} = \bar{y} &= 7 \\ \sum (x_i - \bar{x})^2 &= 4^2 + 2^2 + 0^2 + 1^2 + 2^2 + 3^2 = 34 \\ \sum x_i y_i &= (3)(2) + (5)(5) + (7)(7) + (8)(8) + (9)(9) + (10)(11) = 335 \\ \sum (x_i - \bar{x})(y_i - \bar{y}) &= 335 - 6(7)(7) = 41 \\ \hat{\beta}_1 &= \frac{41}{34} \\ \hat{\beta}_0 &= 7 - \frac{41}{34}(7) = -\frac{49}{34} \\ \hat{y}(6) &= -\frac{49}{34} + 6 \left(\frac{41}{34} \right) = \frac{197}{34} = \mathbf{5.7941} \end{aligned}$$

2.10.

$$\begin{aligned} \bar{y} &= \frac{924}{28} = 33 \\ \bar{x} &= \frac{392}{28} = 14 \\ \sum (x_i - \bar{x})(y_i - \bar{y}) &= 13272 - 28(33)(14) = 336 \end{aligned}$$


Practice Exam 1

1.  A life insurance company is underwriting a potential insured as Preferred or Standard, for the purpose of determining the premium. Insureds with lower expected mortality rates are Preferred. The company will use factors such as credit rating, occupation, and blood pressure. The company constructs a decision tree, based on its past experience, to determine whether the potential insured is Preferred or Standard.

Determine, from a statistical learning perspective, which of the following describes this underwriting method.

- I. Classification setting
- II. Parametric
- III. Supervised

- (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).


2.  An insurance company is modeling the probability of a claim using logistic regression. The explanatory variable is vehicle value. Vehicle value is banded, and the value of the variable is 1, 2, 3, 4, 5, or 6, depending on the band. Band 1 is the reference level.

The fitted value of the β corresponding to band 4 is -0.695 .

Let O_1 be the odds of a claim for a policy in band 1, and O_4 the odds of a claim for a policy in band 4.

Determine O_4/O_1 .

- (A) 0.30 (B) 0.35 (C) 0.40 (D) 0.45 (E) 0.50

3.  Auto liability claim size is modeled using a generalized linear model. Based on an analysis of the data, it is believed that the coefficient of variation of claim size is constant.

Which of the following response distributions would be most appropriate to use?

- (A) Poisson (B) Normal (C) Gamma (D) Inverse Gamma (E) Inverse Gaussian

4. You are given the following output from a GLM to estimate loss size:

- (i) Distribution selected is Inverse Gaussian.
- (ii) The link is $g(\mu) = 1/\mu^2$.

Parameter	β
Intercept	0.00279
Vehicle Body	
Coupe	0.002
Sedan	-0.001
SUV	0.003
Vehicle Value (000)	-0.00007
Area	
B	-0.025
C	0.015
D	0.005

Calculate mean loss size for a sedan with value 25,000 from Area A.

- (A) 80 (B) 160 (C) 320 (D) 640 (E) 1280

5. For a generalized linear model,

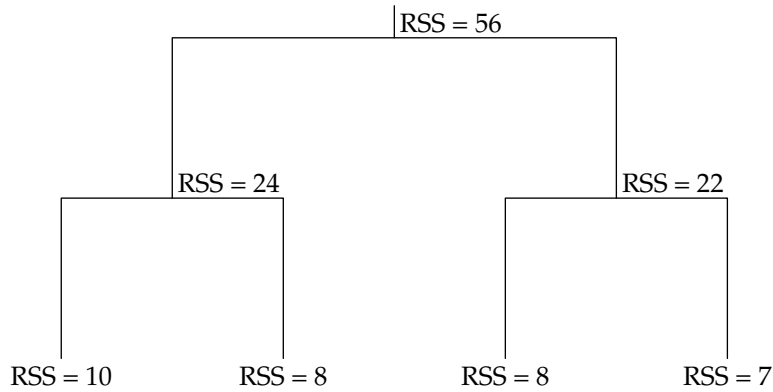
- (i) There are 72 observations.
- (ii) There are 25 parameters.
- (iii) The loglikelihood is -361.24

You are considering adding a cubic polynomial variable.

Determine the lowest loglikelihood for which this additional variable would not be rejected at 1% significance.

- (A) -358 (B) -356 (C) -354 (D) -352 (E) -350


6. In the following decision tree, the RSS of the observations at each node are displayed. For example, at the lowest left node, $\sum(y_i - \bar{y})^2 = 10$ for the observations y_i that are at that node.



The tree is pruned using cost complexity pruning.

Determine the lowest α for which two branches are pruned.

- (A) 6 (B) 7 (C) 10 (D) 12 (E) 13
7. Determine which of the following statements is/are true.
- I. The lasso is a more flexible approach than linear regression.
 - II. Flexible approaches lead to more accurate predictions.
 - III. Generally, more flexible approaches result in less bias.
- (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).


8.  A generalized linear model for automobile insurance with 40 observations has the following explanatory variables:

SEX (male or female)
 AGE (4 levels)
 TYPE OF VEHICLE (sedan, coupe, SUV, van)
 MILES DRIVEN (continuous variable)
 USE (business, pleasure, farm)

Model I includes all of these variables and an intercept. Model II is the same as Model I except that it excludes USE. You have the following statistics from these models:

	Deviance	AIC
Model I	23.12	58.81
Model II		62.61

Using the likelihood ratio test, which of the following statements is correct?


- (A) Reject Model II at 0.5% significance.
 (B) Reject Model II at 1.0% significance but not at 0.5% significance.
 (C) Reject Model II at 2.5% significance but not at 1.0% significance.
 (D) Reject Model II at 5.0% significance but not at 2.5% significance.
 (E) Do not reject Model II at 5.0% significance.
9.  In a classification decision tree, there is one predictor, X . In one region, there are five observations with the following values:

X	3	5	7	9	11
Y	0	0	1	0	1

This region is split using classification error as the criterion.

The classification error of a split is the sum of the probabilities of the subregions times the classification errors of the subregions.

Determine the classification error of the optimal split.


- (A) 0.1 (B) 0.2 (C) 0.3 (D) 0.4 (E) 0.5
10.  A normal linear model with 2 variables and an intercept is based on 45 observations. \hat{y}_j is the fitted value of y_j , and $\hat{y}_{j(i)}$ is the fitted value of y_j if observation i is removed. You are given:


(i) $\sum_{j=1}^{45} (\hat{y}_j - \hat{y}_{j(1)})^2 = 4.1$.

(ii) The leverage of the first observation is 0.15.

Determine $|\hat{\epsilon}_1|$, the absolute value of the first residual of the regression with no observation removed.




- (A) 3.9 (B) 4.4 (C) 4.9 (D) 5.4 (E) 5.9


11.  A least squares model with a large number of predictors is fitted to 90 observations. To reduce the number of predictors, forward stepwise selection is performed.
- For a model with k predictors, $RSS = c_k$.
- The estimated variance of the error of the fit is $\hat{\sigma}^2 = 40$.
- Determine the value of $c_d - c_{d+1}$ for which you would be indifferent between the $d + 1$ -predictor model and the d -predictor model based on Mallows's C_p .
- (A) 40 (B) 50 (C) 60 (D) 70 (E) 80

12.  A classification response variable has three possible values: A, B, and C.
- A split of a node with 100 observations in a classification tree resulted in the following two groups:

Group	Number of A	Number of B	Number of C
I	40	10	10
II	5	25	10

Calculate the cross-entropy for this split.


- (A) 0.72 (B) 0.76 (C) 0.80 (D) 0.84 (E) 0.88
13.  Determine which of the following statements are true regarding cost complexity pruning.
- I. A higher α corresponds to higher MSE for the training data.
 - II. A higher α corresponds to higher bias for the test data.
 - III. A higher α corresponds to a higher $|T|$.
- (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).
14.  Determine which of the following constitutes data snooping.
- (A) Using personal data without authorization of the individuals.
 - (B) Using large amounts of low-quality data.
 - (C) Using an excessive number of variables to fit a model.
 - (D) Fitting an excessive number of models to one set of data.
 - (E) Validating a model with a large number of validation sets.
15.  Determine which of the following statements are true regarding K -nearest neighbors (KNN) regression.
- I. KNN tends to perform better as the number of predictors increases.
 - II. KNN is easier to interpret than linear regression.
 - III. KNN becomes more flexible as $1/K$ increases.
- (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).

16.  A department store is conducting a cluster analysis to help focus its marketing. The store sells many different products, including food, clothing, furniture, and computers. Management would like the clusters to group together customers with similar shopping patterns.

Determine which of the following statements regarding cluster analysis for this department store is/are true.


- I. The clusters will depend on whether the input data is units sold or dollar amounts sold.
- II. Hierarchical clustering would be preferable to K -means clustering.
- III. If a correlation-based dissimilarity measure is used, frequent and infrequent shoppers will be grouped together.

- (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).

17.  Determine which of the following statements regarding principal components analysis is/are true.

- I. Principal components analysis is a method to visualize data.
- II. Principal components are in the direction in which the data is most variable.
- III. Principal components are orthogonal.

- (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.  A random walk is the cumulative sum of a white noise process c_t . You are given that c_t is normally distributed with mean 0 and variance σ^2 .

Which of the following statements are true?

- I. The mean of the random walk does not vary with time.
- II. At time 50, the variance is $50\sigma^2$.
- III. Differences of the random walk form a stationary time series.

- (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 the following regression model, based on 22 observations.


$$\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \beta_4\mathbf{x}_4 + \beta_5\mathbf{x}_5 + \varepsilon$$

The residual sum of squares for this model is 156.

If the variables \mathbf{x}_4 and \mathbf{x}_5 are removed, the error sum of squares is 310.

Calculate the F ratio to determine the significance of the variables \mathbf{x}_4 and \mathbf{x}_5 .

- (A) 3.9 (B) 4.9 (C) 5.9 (D) 6.9 (E) 7.9

20.  You are given the following time series:

$$20, 22, 21, 24, 23$$

The time series is fitted to an AR(1) process with $y_t = 20.325 + 0.1y_{t-1}$.

Calculate the estimated variance of the residuals.






- (A) 1.3 (B) 1.7 (C) 2.1 (D) 2.5 (E) 2.9

21. Determine which of the following algorithms is greedy.
- I. Hierarchical clustering algorithm
 - II. Recursive binary splitting algorithm for decision trees
 - III. Forward subset selection algorithm
- (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).
22. Determine which of the following statements about boosting is/are true.
- I. Selecting B too high can result in overfitting.
 - II. Selecting a low shrinkage parameter tends to lead to selecting a lower B .
 - III. If $d = 1$, the model is an additive model.
- (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).
23. To validate a time series model based on 20 observations, the first 15 observations were used as a model development subset and the remaining 5 observations were used as a validation subset. The actual and fitted values for those 5 observations are

t	y_t	\hat{y}_t
16	7	10
17	9	12
18	12	14
19	18	16
20	22	18

Calculate the MSE.

- (A) 7.4 (B) 8.4 (C) 9.5 (D) 10.5 (E) 11.5
24. In a hurdle model, the probability of overcoming the hurdle is 0.7. If the hurdle is overcome, the count distribution is $kg(j)$, where $g(j)$ is the probability function of a Poisson distribution with parameter $\lambda = 0.6$. Calculate the probability of 1.
- (A) 0.23 (B) 0.31 (C) 0.39 (D) 0.45 (E) 0.51
25. For a generalized linear model, you are given
- (i) The negative loglikelihood of the model is 74.88.
 - (ii) The deviance of the model is 8.70.
 - (iii) The maximized loglikelihood of the minimal model is -90.31 .
- Calculate the pseudo- R^2 statistic.
- (A) 0.64 (B) 0.68 (C) 0.71 (D) 0.74 (E) 0.78

26.  The number of policies sold by an agent in a year, y , is modeled as a function of the number of years of experience, x . The model is a Poisson regression with a log link. The fitted coefficient of x is $\beta_1 = 0.06$.
The expected number of policies sold after 2 years of experience is a and the expected number of policies sold after 5 years of experience is b .
Calculate b/a .
- (A) 1.18 (B) 1.19 (C) 1.20 (D) 1.21 (E) 1.22
27.  Which of the following statements are true?
- I. Partial Least Squares is a supervised method of dimension reduction.
II. Partial Least Squares directions are linear combinations of the original variables.
III. Partial Least Squares can be used for feature selection.
- (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).
28.  Disability income claims are modeled using linear regression. The model has two explanatory variables:
1. *Occupational class*. This may be (1) professional with rare exposure to hazards, (2) professional with some exposure to hazards, (3) light manual labor, (4) heavy manual labor.
 2. *Health*. This may be (1) excellent, (2) good, (3) fair.
- The model includes an intercept and all possible interactions.
Determine the number of interaction parameters β_i in the model.
- (A) 6 (B) 8 (C) 9 (D) 11 (E) 12
29.  Consider the vector $\{5, -3, 8, -2, 4\}$.
Calculate the absolute difference between the ℓ_2 norm and ℓ_1 norm of this vector.
- (A) Less than 12
(B) At least 12, but less than 15
(C) At least 15, but less than 18
(D) At least 18, but less than 21
(E) At least 21
30.  For a simple linear regression of y on x :
- (i) There are 25 observations.
 - (ii) $\bar{x} = 32$
 - (iii) The unbiased sample variance of x is 20.
 - (iv) $x_4 = 22$
- Calculate the leverage of x_4 .
- (A) 0.21 (B) 0.23 (C) 0.25 (D) 0.27 (E) 0.29

31. You are given the time series

182, 138, 150, 192, 177

The series is smoothed using exponential smoothing with $w = 0.8$.

Calculate the sum of squared one-step prediction errors.

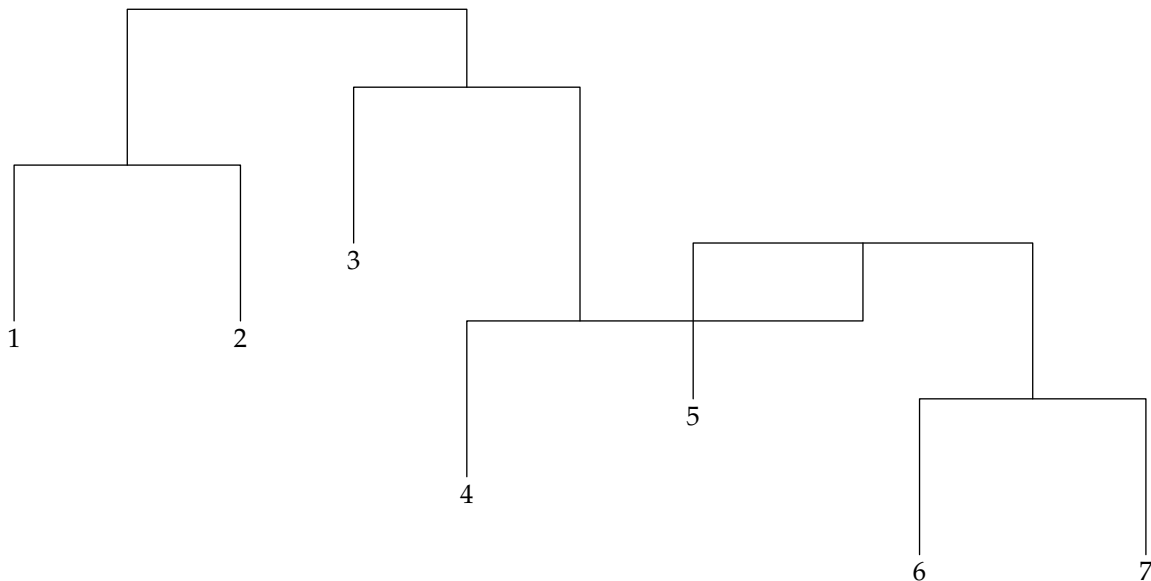
- (A) 2042 (B) 2555 (C) 3038 (D) 3589 (E) 3966

32. Determine which of the following statements about classification trees is/are true.

- I. Classification error is not sensitive enough for growing trees.
- II. Classification error is not sensitive enough for pruning trees.
- III. The predicted values of two terminal nodes coming out of a split are different.

- (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).


33. Hierarchical clustering is performed on 7 observations, resulting in the following dendrogram:



Determine which of the following statements is/are true.

- I. Centroid linkage was used.
- II. Observation 3 is closer to observation 4 than to observation 7.
- III. Observations 3 and 4 are closer to each other than observations 1 and 2.

- (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.  A decision tree model is fitted to five observations (x, y) :

(22,6) (27,10) (13,2) (35,19) (31,12)

Bagging is used with $B = 6$. The resulting predictions for $x = 22$ are

Bootstrap sample	y when $x = 22$
(13,27,13,31,27)	10
(22,13,35,31,35)	7
(27,35,22,31,22)	12
(31,13,27,35,31)	2
(27,13,22,13,27)	5
(27,13,31,22,35)	6

Calculate the OOB prediction for $x = 22$.

- (A) 6 (B) 7 (C) 8 (D) 9 (E) 10
35.  Determine which of the following are results of overfitting models.
- I. The residual standard error may increase.
 - II. The model may be more difficult to interpret.
 - III. The variables may be collinear.
- (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).

Solutions to the above questions begin on page 459.

Appendix A. Solutions to the Practice Exams

Answer Key for Practice Exam 1

1	C	11	E	21	E	31	C
2	E	12	E	22	C	32	A
3	C	13	B	23	B	33	A
4	B	14	D	24	E	34	A
5	B	15	E	25	E	35	D
6	B	16	D	26	C		
7	C	17	D	27	B		
8	C	18	D	28	A		
9	B	19	E	29	A		
10	B	20	D	30	C		

Practice Exam 1

- [Lesson 1]** Classification setting—the company is choosing a class. Supervised—there is something being predicted. But decision trees are not parametric. (C)
- [Lesson 12]** In logistic regression, $g(\mu)$ is the logarithm of the odds, so we must exponentiate β to obtain odds ratio.

$$e^{-0.695} = \mathbf{0.4991} \quad (\text{E})$$

- [Section 11.1]** The square of the coefficient of variation is the variance divided by the square of the mean. If it is constant, then variance is proportional to mean squared. This is true for a gamma distribution. (C)
- [Section 11.1]** Area A is the base level, so nothing is added to $g(\mu)$ for it.

$$g(\mu) = 0.00279 - 0.001 + 25(-0.00007) = 0.00004$$

$$\frac{1}{\mu^2} = 0.00004$$

$$\mu = \sqrt{\frac{1}{0.00004}} = \mathbf{158.11} \quad (\text{B})$$

- [Section 14.2]** A cubic polynomial adds 3 parameters. The 99th percentile of chi-square at 3 degrees of freedom is 11.345. Twice the difference in loglikelihoods must exceed 11.345, so the loglikelihood must increase by 5.67. Then $-361.24 + 5.67 = \mathbf{-355.57}$. (B)
- [Section 16.1]** The cost of pruning the left node is $24 - (10 + 8) = 6$ and the cost of pruning the right node is $22 - (8 + 7) = 7$, so the left node is pruned first when $\alpha = 6$. Next the right node is pruned, and that happens when $\alpha = \mathbf{7}$. (B)

Check out my exclusive video solutions for the FAM, ALTAM, and ASTAM Manuals!

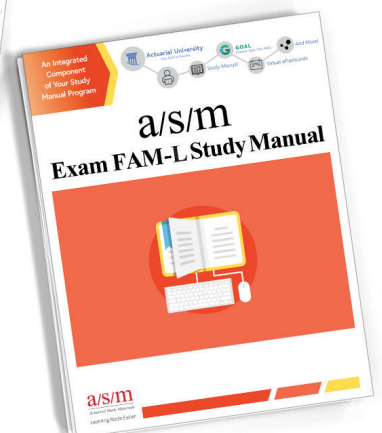
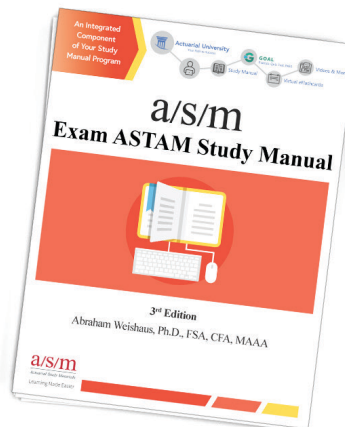
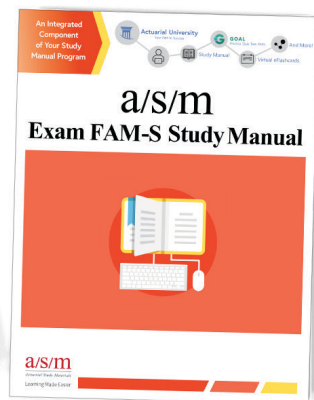
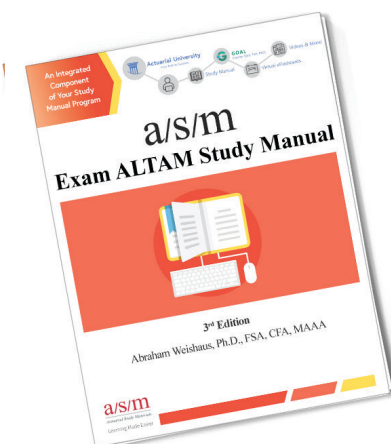
a/s/m

Have you heard of GOAL: Guided Online Actuarial Learning? Use it for practice, quizzes, creating custom exams, and tracking your study progress with GOAL Score.

Connect with me on the ASM discussion group in *Actuarial University*.



Abraham Weishaus
Ph.D., FSA, CFA, MAAA

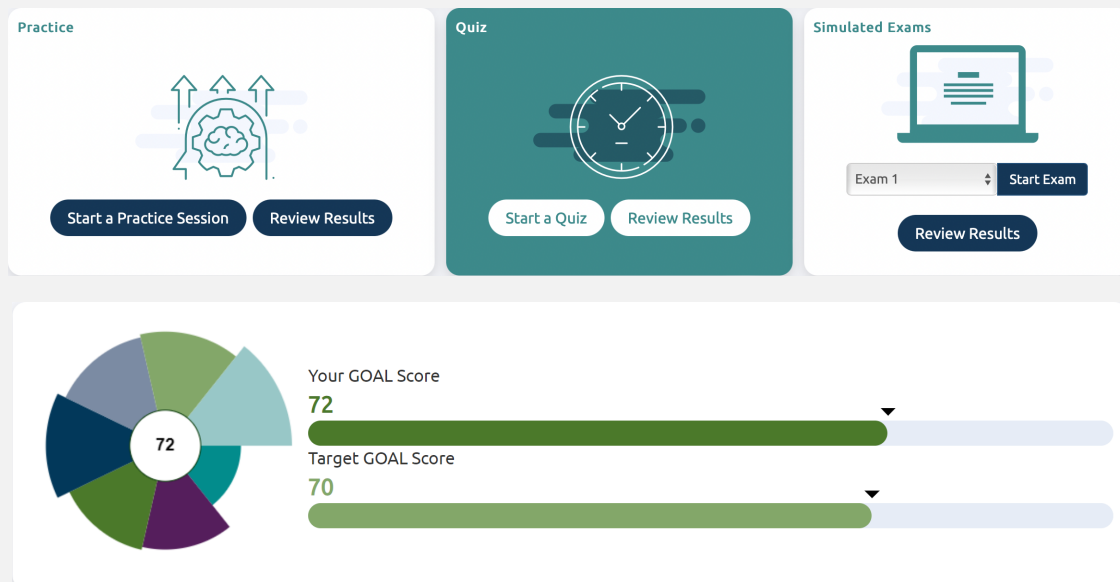


a/s/m Study Manuals

Learning Made Easier.

What's Your **GOAL**?

We'll help you break it down.



Customizable Exam Prep

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

Practice. Predict. Pass.

Using GOAL Score to gauge your readiness

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

