ACTEX

Exam SRM Study Manual

Spring 2019 Edition

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

ACTEX Learning | Learn Today. Lead Tomorrow.

ACTEX

Exam SRM Study Manual

Spring 2019 Edition

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

ACTEX Learning | Learn Today. Lead Tomorrow.

ACTEX Learning
Learn Today. Lead Tomorrow.

Actuarial & Financial Risk Resource Materials
Since 1972

ISBN: 978-1-63588-704-4

Printed in the United States of America.

YOUR OPINION IS IMPORTANT TO US

ACTEX is eager to provide you with helpful study material to assist you in gaining the necessary knowledge to become a successful actuary. In turn we would like your help in evaluating our manuals so we can help you meet that end. We invite you to provide us with a critique of this manual by sending this form to us at your convenience. We appreciate your time and value your input.

Publication:

ACTEX SRM Study Manual, Spring 2019 Edition

I found Actex by: (Check one)

□ A Professor □ School/Internship Program □ Employer □ Friend □ Facebook/Twitter

In preparing for my exam I found this manual: (Check one)

□ Very Good □ Good □ Satisfactory □ Unsatisfactory

I found the following helpful:

____

____

____

____

____

I found the following problems: (Please be specific as to area, i.e., section, specific item, and/or page number.)

____

____

____

____

To improve this manual I would:

____

____

____

____

Name: ____

Address: ____

____

Phone: ______________________ E-mail: ______________________

(Please provide this information in case clarification is needed.)

Send to: Stephen Camilli
ACTEX Learning
P.O. Box 715
New Hartford, CT 06057

Or visit our website at www.ActexMadRiver.com to complete the survey on-line. Click on the "Send Us Feedback" link to access the online version. You can also e-mail your comments to Support@ActexMadRiver.com.

Contents

Preface

Exam SRM (Statistics for Risk Modeling) is a relatively new exam which was offered for the first time in September 2018 by the Society of Actuaries (SOA). In 2019, it will be delivered via computer-based testing (CBT) in January (January 3 to 9), May (May 10 to 16), and September (September 5 to 11). The registration deadlines are November 27, 2018, April 9, 2019, and August 6, 2019, respectively. This new exam is a replacement of the old Validation by Educational Experience (VEE) Applied Statistics requirement and serves as the formal prerequisite for the new Predictive Analytics (PA) exam. The construction of Exam SRM, which revolves around making use of various statistical models to draw inferences and make predictions for the future, is an important step that the SOA takes to incorporate more statistics, most notably predictive modeling, into the modern-day actuarial curriculum. You will considerably sharpen your statistics toolkit as a result of taking (and, with the use of this study manual, passing!) Exam SRM.

It is assumed that you have taken a calculus-based mathematical statistics course (e.g., the one you use to fulfill your VEE Mathematical Statistics requirement) and are no stranger to concepts like (maximum likelihood) estimators, confidence intervals, and hypothesis tests. In Exam SRM, we will make intensive use of these terms to perform point/interval estimation/prediction and hypothesis tests. There will also be instances (mostly in Chapters 2, 3, and 9) in which you will perform matrix multiplication and inversion, which you should have learned from your linear algebra class. Prior exposure to the R programming language, however, is not required.

Syllabus

The syllabus of Exam SRM is very broad (but not necessarily deep) in scope, covering miscellaneous topics in linear regression models, generalized linear models, time series analysis, and data mining techniques, many of which are new topics not tested in any SOA past exams. As a rough estimate, you need at least *three months* of intensive study to master the material in this exam. The six main topics of the syllabus along with their approximate weights and where they are covered in this manual are shown below:

Topic	**Range of Weight**	**Relevant Chapters of This Manual**
1. Basics of Statistical Learning	7.5–12.5%	Chapter 4
2. Linear Models	40–50%	Chapters 1–5
3. Time Series Models	12.5–17.5%	Chapters 6–7
4. Principal Components Analysis	2.5–7.5%	Chapter 9
5. Decision Trees	10–15%	Chapter 8
6. Cluster Analysis	10–15%	Chapter 10

Historically, Topics 2 and 3 on linear models and time series models, which account for more than 50% of the exam, have been on the syllabuses of SOA exams for long (well before the authors of this study manual were born!). They were tested in the 1980s and 1990s in Course 120 (Applied Statistical Methods). From 2000 to 2004, they entered the syllabus of Course 4 (Actuarial Modeling), which was the predecessor of the current Exam C/STAM. From 2005 to June 2018, they were not formally examined but became part of the VEE Applied Statistics requirement. Effective from July 2018, they returned to the exam arena through the newly created Exam SRM, with a significant coverage of non-linear models added. In this study manual, we have extracted virtually all relevant exam questions on linear models and time series models from the above past exams that apply to the current syllabus. Despite the seniority of these past exam questions and that different syllabus texts were used when these exams were offered, they are by no means obsolete and will prove instrumental in illustrating some otherwise obscure concepts in the current syllabus and consolidating your understanding as you progress along this manual.

The SRM syllabus does feature a number of contemporary material. Topics 1, 4, 5, 6, and part of Topic 2 are completely new topics that are introduced to the SOA curriculum for the first time. They pertain to the discipline of statistical learning and predictive analytics, which are very much in vogue nowadays.

Exam Format

Exam SRM is a three and one-half hour computer-based exam consisting of 35 multiple-choice questions. Each question includes five answer choices identified by the letters (A), (B), (C), (D), and (E), only one of which is correct. No credit will be given for omitted answers and no credit will be lost for wrong answers; hence, you should answer all questions, even those for which you have to guess.

According to the SOA, the pass mark for the September 2018 sitting was 70%, which means that candidates needed to answer about **23 to 24 out of 33 to 34 graded questions** correctly to earn a pass (in the CBT environment, one or two questions may be pilot questions that are not graded).

The SOA has released 28 sample questions, which can be accessed from `https://www.soa.org/Files/Edu/2018/exam-srm-sample-questions.pdf`. Although Exam SRM is a new exam, you can expect that many of the exam questions will fall into the following three categories, as the SRM sample questions indicate:

1. *Simple computational questions given a small raw dataset:* In some exam questions (e.g., Sample Questions 1, 3, 4, 9, 11, 15, 23, 28), you will be asked to do some simple calculations using a small dataset, with a size of not more than 10 observations. While many statistical models in the exam syllabus require computers to implement, the fact that the dataset is so small makes it possible to perform at least part of the analysis. Why should the SOA make these unrealistic exam questions? Shouldn't we all use computer to do the work? Although you probably will not have the chance to perform hand calculations in the workplace, these computational 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 and which formulas to use—and are instructive from an educational point of view.

2. *Simple computational/analytical questions given summarized model output:* Constrained by its multiple-choice nature and the absence of computing technology in the CBT environment, the exam will not ask you to use software packages to analyze a large dataset from scratch, nor will it require that you work out bookwork proofs. Rather, you should expect to see

some questions (e.g., Sample Questions 17, 18, 19, 24, 27) in which the model concerned has already been fitted by computers. Given some summarized model output[i] such as tables of parameter estimates and/or plots, you are then 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 the model, all of which require only pen-and-paper calculations.

3. *Conceptual/True-or-false questions:* According to students' comments, the majority of the questions in the most recent SRM exams are conceptual (also known as true-or-false) items, designed to test the uses, motivations, pros and cons, and do's and don'ts of different statistical methods. Sample Questions 2, 5, 6, 7, 8, 10, 12, 13, 14, 16, 20, 21, 22, 25, 26 all belong to this type of questions. You are typically given three statements and asked to pick the correct one(s). The generic structure of these questions is as follows:

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

   I. (blah blah blah...)
   II. (blah blah blah...)
   III. (blah blah blah...)

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

   or

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

   Do not be under the impression that these conceptual questions are easy. The conceptual items being tested can be tricky and at times controversial: Rather than an absolute "yes" or "no," the statement is more a matter of extent. Sadly, if you get any of Statements I, II, or III incorrect, you will likely be led to the incorrect final answer. By the way, Answer (E) occasionally turns out to be the right answer—it is not a filler!

Syllabus Texts

Exam SRM has two required textbooks:

1. *Regression Modeling with Actuarial and Financial Applications*, by Edward W. Frees, 2010 (referred to as Frees in the sequel). The web page of the book is `http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/home.html`.

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

2. *An Introduction to Statistical Learning: With Applications in R*, by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, corrected 8th printing, freely available at `http://www-bcf.usc.edu/~gareth/ISL/` (referred to as "James et al." in the sequel). Although written by four renowned statisticians, this book is designed for non-statisticians and de-emphasizes technical details (formulas and proofs in particular). In fact, one of the selling points of the book is to facilitate the implementation of the statistical learning techniques introduced in the book using R (a black box approach, however!). .

Among the six topics in the exam syllabus, Frees covers Topics 2, 3, and part of Topic 1, while James et al. covers Topics 4, 5, 6, and most of Topic 1. These two texts duplicate to a certain extent when it comes to the chapters on linear regression models. In this study manual, we have streamlined the material in both texts to result in more coherent expositions without unnecessary repetition. As far as possible, we have followed the notation in the two texts. You should note that exam questions can freely use symbols in the texts without defining the symbols for you.

What is Special about This Study Manual?

We fully understand that you have an acutely limited amount of time for study and that the exam syllabus is insanely broad. With this in mind, the overriding objective of this study manual is to help you grasp the material in Exam SRM, which is a relatively new exam, effectively and efficiently, and *pass it with considerable ease.* Here are some of the invaluable features of this manual for achieving this all-important goal:

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

- The learning outcomes of the syllabus are then demystified by precise and concise expositions synthesized from the syllabus readings, helping you acquire a deep and solid understanding of the subject matter.

- Formulas and results of utmost importance are $\boxed{\text{boxed}}$ for easy identification and memorization.

- Mnemonics and shortcuts are emphasized, so are highlights of important exam items and common mistakes committed by students.

- To succeed in any (actuarial) exam, the importance of practicing a wide variety of non-trivial problems to sharpen your understanding and develop proficiency, as always, cannot be overemphasized. This study manual embraces this learning by doing approach and intersperses its expositions with more than **180 in-text examples** and **300 end-of-chapter problems** (the harder ones are labeled as [**HARDER!**]), which are original or taken/adapted from relevant SOA/CAS past exams, all with step-by-step solutions and problem-solving remarks, to consolidate your understanding and give you a sense of what you can expect to see in the real exam. As you read this manual, skills are honed and confidence is built. As a general guide, you should *study all of the in-text examples and work out at least half of the end-of-chapter/section problems.*

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

- Although this study manual is self-contained in the sense that studying this manual is sufficient for the purpose of passing the exam, relevant chapters and sections of the two syllabus texts are referenced at the beginning of each section, for those who would like to learn more.

- Two original full-length practice exams designed to mimic the real exam conclude this study manual giving you a holistic review of the syllabus material.

What is New in this May 2019 Edition?

- While there have been updates throughout the entire manual in terms of content, clarity, and exam focus, the improvements are particularly significant in Sections 1.2, 1.4.1, 4.1.3, 4.2, 4.3.2, 4.3.3, 4.3.4, 5.1.3, and 5.3.2. More remarkably, the three data-mining chapters, Chapters 8, 9, and 10, are thoroughly rewritten.

- Compared to the September 2018 edition, the May 2019 edition features more than 40 new in-text examples and 60 new end-of-chapter problems. They include:

Examples:	1.2.6, 1.3.3, 1.3.6, 1.4.1, 2.1.2, 2.1.6, 3.4.3, 4.1.4, 4.1.5, 4.1.6, 4.2.1, 4.2.2, 4.2.3, 4.2.6, 4.3.2, 4.3.3, 4.3.9, 4.4.3, 4.4.7, 5.1.4, 5.1.18, 5.1.20, 5.2.1, 5.2.11, 5.3.5, 6.1.3, 6.2.2 and most examples in Chapters 8, 9, and 10
Problems:	1.6.8, 1.6.10, 1.6.22, 1.6.23, 1.6.28, 1.6.29, 1.6.31, 1.6.34, 1.6.38, 1.6.50, 2.5.11, 3.5.8, 3.5.17, 4.5.2, 4.5.3, 4.5.7, 4.5.8, 4.5.10, 4.5.11, 4.5.12, 4.5.13, 4.5.14, 4.5.21, 4.5.22, 4.5.23, 4.5.24, 4.5.25, 4.5.29, 4.5.31, 4.5.32, 4.5.36, 4.5.37, 4.5.38, 5.4.4, 5.4.5, 5.4.6, 5.4.14, 5.4.26, 5.4.33 and most problems in Chapters 8, 9, and 10

  Many of these examples and problems are of the true-or-false type, which, according to students' comments, has figured prominently in recent SRM exams.

- Relevant questions from the very recent Fall 2018 MAS-I and MAS-II exams of the CAS have also been included.

- A new practice exam (Practice Exam 2) has been designed to give you an additional opportunity to do an overall review of the exam syllabus. This new exam, along with Practice Exam 1, has a nice combination of computational and conceptual exam items.

- All known typographical errors have been fixed.

Exam Tables

In the real exam, you will be supplied with three statistical tables, namely, the standard normal distribution, t-distribution, and chi-square distribution tables. They are available for download at `https://www.soa.org/Files/Edu/2018/exam-srm-tables.pdf` and will be intensively used during your study (especially in Parts I and II of this study manual) as well as in the exam. You should not hesitate to print out a copy and learn how to locate the relevant entries in these tables as you work out examples and problems in this manual.

Acknowledgments

We would like to thank our colleagues, Professor Elias S. W. Shiu and Dr. Michelle A. Larson, at the University of Iowa for sharing with us many pre-2000 SOA/CAS exam papers. These hard-earned old exam papers, of which the Society of Actuaries and Casualty Actuarial Society own the sole copyright, have proved invaluable in illustrating a number of less commonly tested exam topics. Ambrose Lo is also grateful to students in his VEE Applied Statistics course (STAT:4510) in Fall 2016 and Fall 2017 for class testing earlier versions of this study manual.

Errata

While we go to great lengths to polish and proofread this manual, some mistakes will inevitably go unnoticed. We would like to apologize in advance for any errors, typographical or otherwise, and would greatly appreciate it if you could bring them to our attention via email so that they can be fixed in a future edition of the manual.

- For questions about Chapters 1 to 7, please email Ambrose Lo at ambrose-lo@uiowa.edu.
- For questions about Chapters 8 and 10, please email Daniël Linders at dlinders@illinois.edu (and c.c. ambrose-lo@uiowa.edu).
- For questions about Chapter 9, please email Runhuan Feng at rfeng@illinois.edu (and c.c. ambrose-lo@uiowa.edu).

Compliments and criticisms are also welcome. The authors will try their best to respond to any inquiries as soon as possible and an ongoing errata list will be maintained online at `https://sites.google.com/site/ambroseloyp/publications/SRM`. Students who report errors will be entered into a quarterly drawing for a $100 in-store credit.

Runhuan Feng
Daniël Linders
Ambrose Lo
February 2019

About the Authors

Runhuan Feng, FSA, CERA, is an associate professor and the Director of Actuarial Science Program at the University of Illinois at Urbana-Champaign. He obtained his Ph.D 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 his efforts 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 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 Ph.D 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 courses on Predictive Analytics, Life Contingencies, Pension Financing and Risk Measurement. He is currently teaching at the University of Illinois and is guest lecturer at the University of Leuven and the ISM-Adonaï, Benin.

Ambrose Lo, FSA, CERA, is currently Assistant Professor of Actuarial Science at the Department of Statistics and Actuarial Science at the University of Iowa. He received his Ph.D. in Actuarial Science from the University of Hong Kong in 2014, with dependence structures, risk measures, and optimal reinsurance being his research interests. 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 has taught courses on financial derivatives, mathematical finance, life contingencies, credibility theory, advanced probability theory, and regression and time series analysis. His emphasis in teaching is always placed on the development of a thorough understanding of the subject matter complemented by concrete problem-solving skills. He is also the sole author of the *ACTEX Study Manual for CAS MAS-I* (Spring 2019 Edition) and the textbook *Derivative Pricing: A Problem-Based Primer* (2018) published by Chapman & Hall/CRC Press.

Part I

Regression Models

Chapter 1

Simple Linear Regression

EXAM SRM LEARNING OBJECTIVES

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:

b) Estimate parameters using least squares and maximum likelihood.

d) Select an appropriate model, considering:

- t and F tests

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

OPTIONAL SYLLABUS READING(S)

- Frees, Chapter 2 (except Section 2.6)
- James et al., Section 3.1

Chapter overview: This chapter examines in detail *simple linear regression*[i] (SLR), arguably the simplest statistical model in the entire Exam SRM, where we seek to understand the linear relationship between a pair of variables. In this somewhat simplistic framework, virtually all of the essential ideas of linear regression, such as parameter estimation, hypothesis testing, construction of confidence intervals, and prediction, can be well illustrated. In addition, by restricting ourselves to the two-dimensional setting, relationships between variables can be displayed graphically and valuable intuition about regression techniques gained.

[i] Frees refers to simple linear regression as "basic" linear regression. However, both James et al. and the SOA sample questions use the more common term "simple" linear regression, and we follow this usage. An alternative but somewhat unprofessional name for simple linear regression is *two-variable* regression, which was used in some old SOA problems. This, however, should not be confused with regression with two explanatory variables.

This chapter is organized as follows. Section 1.1 walks you through a simple motivating example that gives you some sense of linear regression that is valuable for and beyond taking Exam SRM. The SLR model is then set up and the basic statistical terminology that will be used throughout this study manual is introduced. In Section 1.2, we discuss how the SLR model can be fitted to a dataset by means of the least squares method. Section 1.3 assesses the goodness of fit of the regression model and the significance of the explanatory variable in "explaining" the response variable. The results can be conveniently tabulated in a so-called ANOVA table and summarized by a simple proportion measure known as the coefficient of determination. Section 1.4 proceeds to draw inference about the underlying regression parameters. Confidence intervals are constructed and hypothesis tests performed. Finally, Section 1.5 concludes this chapter with the practically important task of predicting future responses. The subtle differences between estimation and prediction are also pointed out.

1.1 Overture

OPTIONAL SYLLABUS READING(S)

- Frees, Sections 2.1 and 2.2
- James et al., Subsection 3.1.1

1.1.1 A Motivating Example

The following dataset records the overall examination scores,[ii] correct to the nearest integer, of 20 students who took Course Y (a notoriously difficult actuarial course):

78	89	90	72	89	77	66	85	84	86
77	88	61	87	96	44	84	62	84	80

Figure 1.1.1 gives a *scatter plot* of the scores.

Question: Predict the exam score of the next student who will take Course Y.

"Naive" answer: Use the average of the 20 scores, namely $\bar{y} = 78.80$. Observe that the exam scores scatter around the sample mean but are subject to considerable fluctuations. The use of $\bar{y}$ is justifiable if the exam scores are, for instance, independent and identically distributed (i.i.d.). In the absence of further information, this seems to be the best we can do.

Is the i.i.d. model suitable in this context? It is, *only if* the students are relatively homogeneous in nature. Given the diversity of students in this day and age, the i.i.d. assumption appears untenable.

Can the exam scores of these students in another course be of use? Suppose that the exam scores of Course X of these 20 students are also available in Table 1.2. Because both Course X and Course Y were taught by the same devilish instructor, Ambrose Lo, and Course X serves as a prerequisite for Course Y, it seems plausible that the Course X scores will be useful in predicting

[ii] These are real exam scores at the University of Iowa.

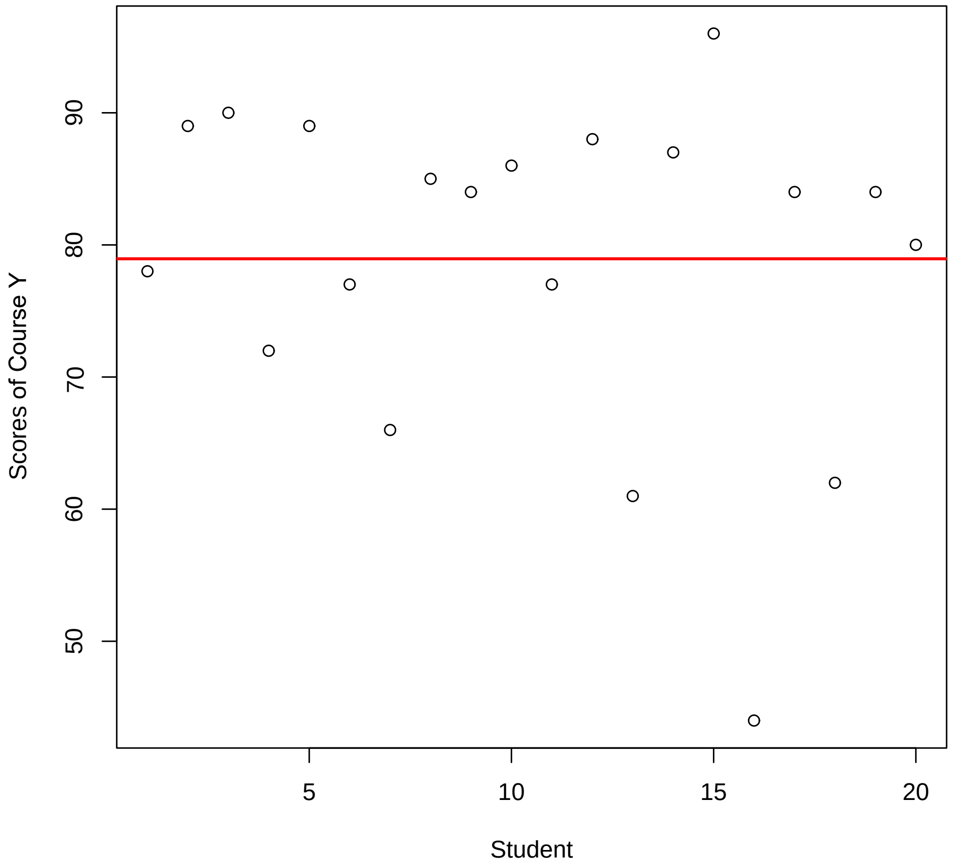

Figure 1.1.1: Scores of 20 students who took Course Y. The red horizontal line represents the sample mean level of $\bar{y} = 78.80$.

Course X Score	Course Y Score	Course X Score	Course Y Score
70	78	79	77
87	89	86	88
94	90	58	61
82	72	92	87
87	89	101	96
75	77	52	44
77	66	81	84
95	85	75	62
86	84	99	84
90	86	58	80

Table 1.2: Exam scores of Course X and Course Y for 20 students.

Course Y scores (or else the prerequisite can be lifted!). Now each observation in the dataset consists of the values of two variables of a student:[iii]

$$(x, y) := (\text{Course X score, Course Y score}).$$

Figure 1.1.2 plots the scores of Course Y (y) against the scores of Course X (x) for the dataset. We can observe a pretty strong linear relationship between x and y (the strength of this linear relationship will be formally quantified using techniques in Section 1.3), with students scoring high in Course X having a tendency to perform well in Course Y too. As far as prediction is concerned, it seems more reliable to assume a linear function relating the scores of students in the two courses, and predict the score of the next student in Course Y based on his/her score in Course X than to use $\bar{y}$.

Figure 1.1.2 also fits a sloped straight line to the scatter plot (by the least squares method, to be discussed in Section 1.2). This straight line summarizes the linear relationship between scores of Course X and Course Y, and can be used for predicting future students' scores in Course Y on the basis of how they performed in Course X. Compared to Figure 1.1.1, the fluctuations of the 20 observations around the sloped straight line appear much smaller. Thus it seems fair to say that a function linear in the scores of Course X (the sloped straight line) can better account for the observed variation in scores of Course Y than a simple constant function (the horizontal line). A crucial question of interest to the instructor of Course Y is: How much is the "sloped straight line" model better than the "i.i.d." model (or "horizontal line" model)? Taking this a step further, can one assert that a student doing well in Course X tends to do well in Course Y? These questions will be addressed in Sections 1.3 and 1.4 in a statistical framework.

Linear regression. The above example highlights the essence of *regression*, which is a statistical technique of employing data on some other variables (e.g. scores of Course X) relevant to the main variable of interest (e.g., scores of Course Y) in order to better explain the observed variation in the latter. It modifies the "i.i.d. assumption" typically used in the VEE Mathematical Statistics course by keeping "independent" but removing "identically distributed"—the 20 students now differ in terms of distribution according to their scores of Course X. In particular, regression involving the use of linear functions to summarize the relationship between variables is called *linear regression*, which is the focus of Chapters 1 to 4 of this manual and a main topic of Exam SRM. In the exam scores example above, we assumed that

$$y = \beta_0 + \beta_1 x + \varepsilon$$

for some unknown parameters β_0 and β_1, and some random deviation ε.

In regression analysis, each observation consists of measurements on a number of variables related to an individual experimental/observational unit sampled from the population. To make our studies in Exam SRM more systematic, there are two common ways to classify variables, by their role in the study, or by their nature:

- *Response vs. explanatory variables:* We designate the variable of primary interest as the *response* variable (or *dependent* variable)—because we are interested in their "response"—and those which might provide supplementary information useful for explaining the behavior of the response variable as the *explanatory* variables. Alternative names commonly used for

[iii] Throughout this study manual, the symbol ":=" means "is defined as."

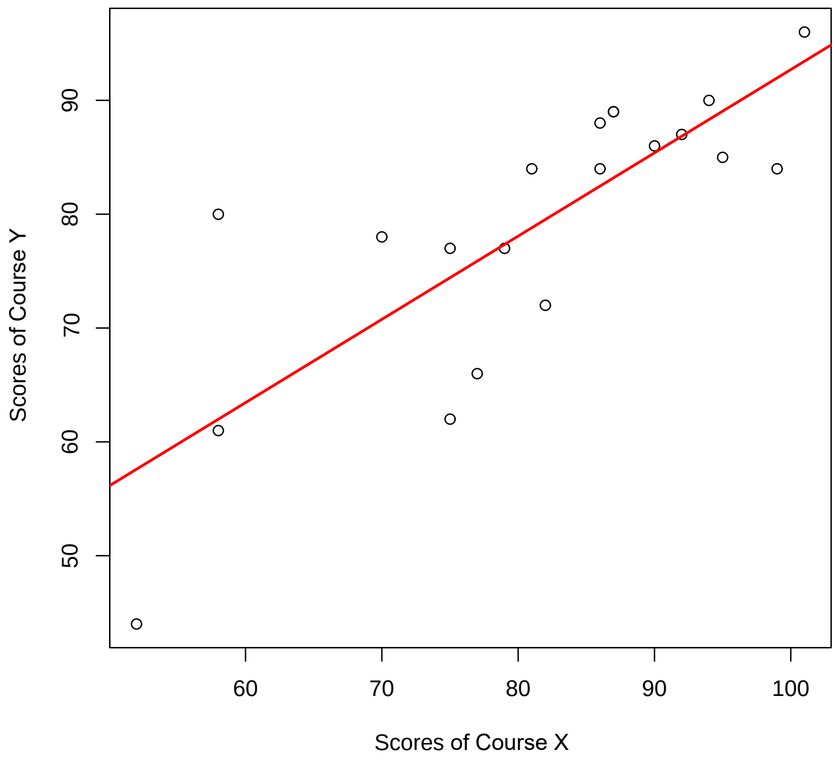

Figure 1.1.2: A plot of scores of Course Y against scores of Course X. The red sloped line is fitted by the method of least squares.

"explanatory" variables are *independent* variables, *predictors*, *regressors,* and *features*, and these terms are used interchangeably in the two SRM texts and in this manual.

Here are some common examples of response and explanatory variables:

Response Variable	**Explanatory Variable**
Opinion	Sex, age, educational level, etc.
House price	Building age, facilities, location
Insurance premium	Sex, age, living style, health conditions
Voltage	Current

Typical questions one wishes to answer by linear modeling include:

1. Does a certain explanatory variable affect the response significantly? If so, is the effect a positive or negative one?
2. Is the regression model adequate for explaining the relation between the response and the explanatory variables?
3. Can we predict a future response based on the values of the explanatory variables?

- *Continuous and categorical variables:* This will be treated in detail in Subsection 2.3.1.

1.1.2 Simple Linear Regression

Model equation. In an SLR model, it is postulated that the response variable y is related to the single explanatory variable x via the (approximately[iv]) linear relationship

$$\boxed{y = \beta_0 + \beta_1 x + \varepsilon,} \tag{1.1.1}$$

where

> β_0 and β_1 are unknown *regression coefficients* (or *regression parameters*), about which inference is to be made later in this chapter, and
>
> ε is the unobservable *random error term* (also called the *noise term*) that accounts for the fluctuation of y about the regression line $\beta_0 + \beta_1 x$.

Among the two SRM syllabus texts, Frees denotes variables in lowercase letters, as in (1.1.1), whereas James et al. uses capital letters, e.g., X and Y. In this manual, we mostly follow Frees since it is the main text that covers regression and time series models.

In (1.1.1), we say that y is *regressed on* x. The straight line $\beta_0 + \beta_1 x$ is called the *regression function*, which is the primary target of interest in regression analysis. In particular, β_0 is called the *intercept*, which captures the value of $\mathbb{E}[y]$ when $x = 0$, and β_1 is the *slope* parameter, which measures the increase in $\mathbb{E}[y]$ per unit increase in x. Because all of the observations from the SLR model share the same parameters β_0 and β_1, the regression function is also known as the *systematic component* of the model. In contrast, ε is referred to as the idiosyncratic part of the model, with different observations having different random errors.

From this SLR model, suppose that we are given n independent (but not identically distributed—why?) copies of y, say $y_1, y_2, \ldots, y_n$, observed at $x = x_1, x_2, \ldots, x_n$, respectively. In other words, we have n pairs of observations, $\{(x_i, y_i)\}_{i=1}^n$, where each y_i is generated according to

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{1.1.2}$$

In spreadsheet form, the data structure can be depicted as:

Observation	x	y
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

It is expected that the response values y_i's fluctuate about their means $\beta_0 + \beta_1 x_i$ by the random errors ε_i. A plot of y_i against x_i is expected to exhibit a linear trend, subject to such random errors (e.g., Figure 1.1.2).

Model assumptions. The SLR model relies on a number of assumptions, including:

A1. The y_i's are realizations of random variables, while the x_i's are nonrandom (i.e., known, measured without error).

[iv] The linear relationship is only approximate due to the presence of the random error ε.

A2. The n random errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent with $\mathbb{E}[\varepsilon_i] = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$ for all $i = 1, 2, \ldots, n$. This, together with Assumption 1, implies that $y_1, y_2, \ldots, y_n$ are also independent with

$$\mathbb{E}[y_i] = \beta_0 + \beta_1 x_i \quad \text{and} \quad \text{Var}(y_i) = \sigma^2.$$

Note that the mean of each y_i is linear in the explanatory variable x ("simple") as well as in the parameters β_0 and β_1 ("linear"), hence the term "simple linear regression."

In the next section, we will answer the question of how the parameters β_0 and β_1 should be "optimally" selected based on the observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

1.2 Model Fitting by the Least Squares Method

OPTIONAL SYLLABUS READING(S)

- Frees, Sections 2.1 and 2.2
- James et al., Subsection 3.1.1

This section is devoted to the following question, which inevitably arises before the SLR model can be put to use:

> How to find the estimates $\hat{\beta}_0, \hat{\beta}_1$[v] for β_0, β_1 such that the *fitted regression line*[vi]
>
> $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$
>
> "best" fits the observations?

There are many criteria defining how the estimates should be optimally chosen, the most common one in regression settings being the method of least squares.

Least squares method. As its name suggests, the least squares method consists in choosing the estimates of β_0 and β_1 in order to make the sum of the vertical "squared" differences between the observed values and the corresponding points on the fitted regression line the "least," i.e., the *least squares estimators* (LSEs) $\hat{\beta}_0$ and $\hat{\beta}_1$ are such that they minimize

$$\text{SS}(\beta_0, \beta_1) := \sum_{i=1}^{n} [\underbrace{y_i}_{\text{obs. value}} - (\underbrace{\beta_0 + \beta_1 x_i}_{\text{candidate fitted value}})]^2 \tag{1.2.1}$$

over all candidate values β_0 and β_1. By calculus, the optimal solutions solve $\frac{\partial}{\partial \beta_0}\text{SS}(\hat{\beta}_0, \hat{\beta}_1) = \frac{\partial}{\partial \beta_1}\text{SS}(\hat{\beta}_0, \hat{\beta}_1) = 0$ and are given by

$$\boxed{\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},} \tag{1.2.2}$$

[v] Frees denotes the estimators of β_0 and β_1 by b_0 and b_1, respectively. However, the symbols $\hat{\beta}_0$ and $\hat{\beta}_1$, adopted by James et al., are more popular in the regression literature. SRM exam questions can use either $\hat{\beta}_0, \hat{\beta}_1$ or b_0, b_1.

[vi] Note that the fitted regression line is not the same as the true regression line $\mathbb{E}[y] = \beta_0 + \beta_1 x$. The former serves to estimate the latter.

where $\bar{x} = \sum_{i=1}^n x_i/n$ and $\bar{y} = \sum_{i=1}^n y_i/n$ are the sample means of x and y, respectively, and

$$\boxed{S_{xy} := \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \quad \text{and} \quad S_{xx} := \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2.}$$

Here, "S" suggests "sum" of squares of the quantities indicated in the subscript *with their sample means corrected*, so S_{xy} returns the corrected sum of squares of x multiplied by the corrected sum of squares of y, and S_{xx} gives the corrected sum of squares of x multiplied by itself.

How to calculate the least squares estimates efficiently? The calculation of the LSEs is the first step of a regression analysis and hence can be an important exam item. There are two ways that the LSEs can be computed by hand in an exam environment:

- *Case 1: Given the raw data*

  You may be given the raw data $\{(x_i, y_i)\}_{i=1}^n$ with a relatively small sample size n (e.g., less than 10). In this case, the two LSEs can be calculated by directly applying (1.2.2). Alternatively and much more efficiently, they can also be computed by entering the data into your financial calculator and reading the output from its statistics mode. In the case of the *BA-II Plus Professional* calculator, for instance, follow these steps: (other financial calculators have similar steps)

  1. Press `[2ND][DATA]` (you may need to first clear the memory of the calculator by pressing `[2ND][DATA][2ND][CE/C]`).
  2. Enter the data values by the following keystroke:

     `(value of` x_1`)[ENTER][↓](value of` y_1`)[ENTER][↓]`

     $\vdots$

     `(value of` x_n`)[ENTER][↓](value of` y_n`)[ENTER][↓]`

     (*Warning:* Make sure that you enter the value of x followed by the value of y! If you mix up the order, the parameter estimates would be different; see Example 1.3.6 on page 26.)
  3. Press `[2ND][STAT]`, followed by `[↓]` until you see "`a`" and "`b`".[vii] These are the values of $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively.

  The knowledge of (1.2.2) is not required in this case!

Example 1.2.1. (SOA Course 120 November 1990 Question 6: Calculation of LSE given raw data) You are estimating a simple regression of the form

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

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

[vii] Some write the model equation of an SLR model as $y = \alpha + \beta x + \varepsilon$.

Determine $\hat{\beta}_1$.

(A) 0.8

(B) 0.9

(C) 1.0

(D) 1.1

(E) 1.2

Solution. Following the steps above (which you should practice!), you will be able to get $\hat{\beta}_1 = \boxed{0.9}$ from your financial calculator. **(Answer: (B))** □

- *Case 2: Given summarized data in the form of various sums*

  Instead of the full dataset, you may be given only summarized information such as the values of

  $$\sum_{i=1}^{n} x_i, \quad \sum_{i=1}^{n} y_i, \quad \sum_{i=1}^{n} x_i^2, \quad \sum_{i=1}^{n} y_i^2, \quad \sum_{i=1}^{n} x_i y_i.$$

  In this case, the use of (1.2.2) is necessary. To calculate the LSEs, it is most convenient to expand the products in the two sums that appear in (1.2.2) and use the alternative form

  $$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}.} \tag{1.2.3}$$

Example 1.2.2. (SOA Exam SRM Sample Question 17: Calculation of LSE given various sums) 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

Solution. As $\bar{x} = 6/6 = 1$ and $\bar{y} = 8.5/6 = 17/12$, the LSE of β_1, by (1.2.3), is

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{15.5 - 6(1)(17/12)}{16 - 6(1)^2} = \boxed{0.7}. \qquad \textbf{(Answer: (D))}$$

□

Example 1.2.3. (SOA Course 120 May 1990 Question 8: When you have a careless assistant!) Your assistant was to estimate the parameters of a simple regression model of the form:

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad x_i = i, \quad i = 1, \ldots, 5$$

Your assistant determined that the parameter estimates were:

$$\begin{aligned} \hat{\beta}_0 &= 7 \\ \hat{\beta}_1 &= 4 \end{aligned}$$

Later, you learned that your assistant inadvertently found the parameter estimates for the transformed variable $z = 2y - 3$.

Determine the parameter estimates of the correct regression.

(A) $\hat{\beta}_0 = 4$, $\hat{\beta}_1 = 2$

(B) $\hat{\beta}_0 = 4$, $\hat{\beta}_1 = 8$

(C) $\hat{\beta}_0 = 5$, $\hat{\beta}_1 = 2$

(D) $\hat{\beta}_0 = 5$, $\hat{\beta}_1 = 8$

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

Solution. Note that $z_i = 2y_i - 3$ for $i = 1, \ldots, 5$, and $\bar{z} = 2\bar{y} - 3$. The LSE of the slope when z is regressed on x is

$$\hat{\beta}_1^{z\sim x} = \frac{S_{xz}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(z_i - \bar{z})}{\sum(x_i - \bar{x})^2} = \frac{2\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = 4,$$

so the LSE of the slope when y is regressed on x is

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

Moreover, $\hat{\beta}_0^{z\sim x} = \bar{z} - \hat{\beta}_1^{z\sim x}\bar{x} = (2\bar{y} - 3) - 2\hat{\beta}_1^{y\sim x}\bar{x} = 2(\bar{y} - \hat{\beta}_1^{y\sim x}\bar{x}) - 3 = 7$, so that

$$\hat{\beta}_0^{y\sim x} = \bar{y} - \hat{\beta}_1^{y\sim x}\bar{x} = \boxed{5}. \qquad \textbf{(Answer: (C))}$$

□

Remark. The fact that $x_i = i$ for $i = 1, \ldots, 5$ is not required for solving this problem.

An alternative formula for $\hat{\beta}_1$ in terms of the sample correlation. There is another way to express $\hat{\beta}_1$ that is less commonly seen in mainstream regression textbooks, but is stated on page 28 of Frees and comes in useful occasionally. It reads

$$\boxed{\hat{\beta}_1 = r \times \frac{s_y}{s_x}, \qquad \left(\text{Warning: Not } r \times \frac{s_x}{s_y}!\right)} \tag{1.2.4}$$

where

- s_x and s_y (with lowercase "s") are the sample standard deviations of x and y given respectively by

$$s_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{S_{xx}}{n-1} \qquad \text{and} \qquad s_y = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{S_{yy}}{n-1}.$$

- r is the sample correlation between x and y given by

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Formula (1.2.4) can be easily shown by writing

$$\hat{\beta}_1 \overset{(1.2.2)}{=} \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \times \sqrt{\frac{S_{yy}}{S_{xx}}} = r \times \sqrt{\frac{S_{yy}/(n-1)}{S_{xx}/(n-1)}} = r \times \frac{s_y}{s_x}$$

and is especially useful when summarized information involving the sample correlation between x and y is given in an exam question.

Example 1.2.4. (SOA Course 120 November 1985 Question 5: Calculation of LSE given r) You are given 30 pairs of observations (x_i, y_i) which are to be represented by the following model:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where ε is a random error term with mean 0 and variance σ^2.

You have determined:

$$\begin{aligned} r &= 0.5 \\ s_x &= 7.0 \\ s_y &= 5.0 \end{aligned}$$

Calculate the least squares estimate of β_1.

(Answer to nearest 0.1)

(A) 0.4

(B) 0.5

(C) 0.6

(D) 0.7

(E) 0.8

Solution. Using (1.2.4), we have

$$\hat{\beta}_1 = r \times \frac{s_y}{s_x} = 0.5 \times \frac{5}{7} = \boxed{0.3571}. \qquad \textbf{(Answer: (A))}$$

□

Remark. Incorrectly computing $\hat{\beta}_1$ as $r \times s_x/s_y = 0.5 \times 7/5 = 0.7$ leads to Answer (D).

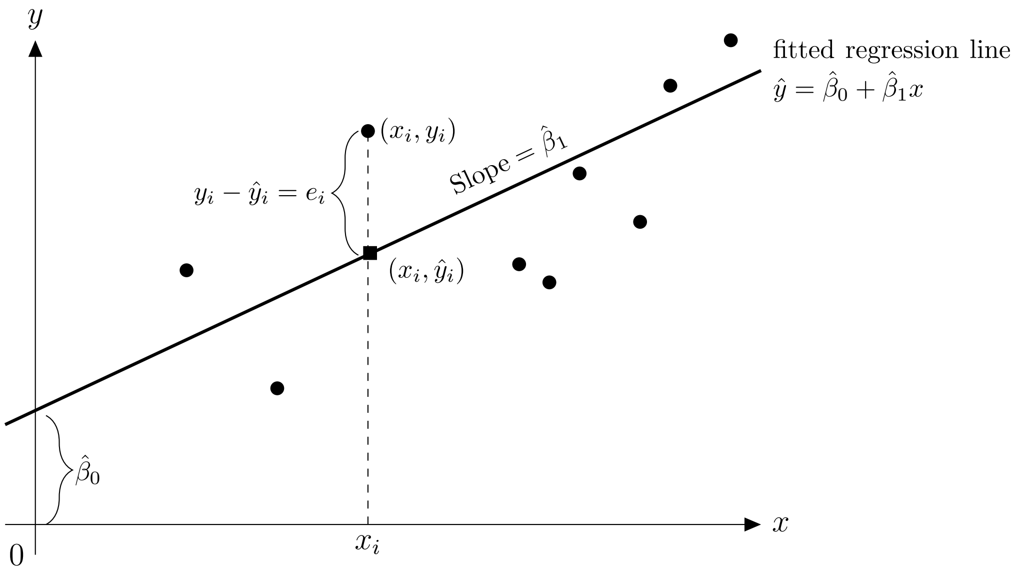

Figure 1.2.1: Graphical illustration of the fitted regression line and the definitions of the fitted value and residual. The black dots denote the observed data and the square denotes the fitted value of y at $x = x_i$.

Fitted values and residuals. Having found the LSEs $\hat{\beta}_0$ and $\hat{\beta}_1$, we can compute, for each observation:

- The *fitted value* (or *predicted value*) $\boxed{\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i}$, for $i = 1, \ldots, n$

  These values are obtained from the model equation (1.1.1) with the unknown parameters β_0 and β_1 replaced by the LSEs $\hat{\beta}_0$ and $\hat{\beta}_1$ and with the random error replaced by its expected value of zero. Ideally, we would like the fitted value of each observation to be as close to the observed value as possible.

- The *residual*[viii] $\boxed{e_i = y_i - \hat{y}_i}$ (note: not $\hat{y}_i - y_i$!), which captures the discrepancy between the observed value and the fitted value

  Note that residuals and the random errors are completely different entities. The former are computable from the data (through $\hat{\beta}_0, \hat{\beta}_1, x_i$ and y_i) and serve to approximate the latter, which are unobservable. Some authors call the residuals the "observed" errors to distinguish them from the unobservable random errors.

Figure 1.2.1 depicts the fitted regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ for a set of sample data and how the fitted value and residual are defined for the ith observation.

Example 1.2.5. (SOA Exam SRM Sample Question 23: Calculation of fitted values given raw data) Toby observes the following coffee prices in his company cafeteria:

- 12 ounces for 1.00

[viii] The symbol e_i is used in both Frees and James et al. In our opinion, the self-explanatory symbol $\hat{\varepsilon}_i$ is more indicative of the role played by the residuals in approximating the unknown random errors. Nevertheless, we shall follow the notation of Frees and James et al.

- 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.

Solution. We are given $(x_1, y_1) = (12, 1)$, $(x_2, y_2) = (16, 1.2)$, and $(x_3, y_3) = (20, 1.4)$. To determine the fitted regression line, we can calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ by (1.2.2). Alternatively and more efficiently, we can observe that the three data points already lie on a straight line, which in turn becomes the fitted regression line—there is no error in this case! The equation of the line is $y = 0.4 + 0.05x$.

To determine the price of a 24-ounce coffee and a 48-ounce coffee, we set $x = 24$ and $x = 48$ to get, respectively, $\hat{y} = 0.4 + 0.05(24) = 1.6$ and $\hat{y} = 0.4 + 0.05(48) = 2.8$. Compared to buying two cups of 24-ounce coffee, which costs $2(1.6) = 3.2$, buying one cup of 48-ounce coffee costs only 2.8. The amount of saving is $3.2 - 2.8 = \boxed{0.4}$. **(Answer: (C))** □

Remark. If you can observe that the amount that Toby and Karen can save equals the intercept of the fitted regression line, which is 0.4 in this case, then the above calculations can be shortened.

Example 1.2.6. (CAS Exam MAS-I Fall 2018 Question 29: Calculation of residual given a small set of raw data) 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 3$^{\text{rd}}$ 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

Solution. Inputting $\{(\text{Advertising}_i, \text{Sales}_i)\}_{i=1}^5$ (note: not $\{(\text{Sales}_i, \text{Advertising}_i)\}_{i=1}^5$!) into our financial calculator, we get $\hat{\beta}_0 = -29.1791$ and $\hat{\beta}_1 = 23.8060$. Hence $\hat{y}_3 = \hat{\beta}_0 + 6\hat{\beta}_1 = 113.6567$ and $e_3 = y_3 - \hat{y}_3 = 112 - 113.6567 = \boxed{-1.6567}$. **(Answer: (B))** □

Remark. If you mistakenly compute e_3 as $\hat{y}_3 - y_3$, you will end up with Answer (C), which is incorrect!

Sum-to-zero constraints on residuals. Whenever the SLR model is fitted by the method of least squares, the residuals can be shown to satisfy the following sum-to-zero constraints (see Exercises 2.14 and 2.15 of Frees):

1. $\sum_{i=1}^n e_i = 0$, provided that the intercept term β_0 is included in the model. This is a desirable property because it implies that the residuals offset one another to produce a zero sum. An implication is that the residuals are negatively correlated.

2. $\sum_{i=1}^n x_i e_i = 0$, which is true no matter whether the intercept is present or not.

These two facts can be easily shown by realizing that $\hat{\beta}_0$ and $\hat{\beta}_1$, as the minimizer of $\text{SS}(\beta_0, \beta_1)$, satisfy

$$\begin{aligned}
\frac{\partial}{\partial \beta_0}\text{SS}(\hat{\beta}_0, \hat{\beta}_1) &= -2\sum_{i=1}^n \overbrace{[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]}^{e_i} = 0, \\
\frac{\partial}{\partial \beta_1}\text{SS}(\hat{\beta}_0, \hat{\beta}_1) &= -2\sum_{i=1}^n x_i \underbrace{[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]}_{e_i} = 0.
\end{aligned}$$

Example 1.2.7. (SOA Course 120 Study Note 120-82-97 Question 1: Given the LSE, deduce the observation) You fit the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ to the following data:

i	1	2	3
x_i	1	3	4
y_i	2	y_2	5

You determine that $\hat{\beta}_0 = 5/7$.
Calculate y_2.

(A) 0

(B) 1

(C) 2

(D) 3

(E) 4

Solution. First,

$$\begin{aligned}\sum_{i=1}^{3} e_i &= \left[2-\left(\frac{5}{7}+\hat{\beta}_1\right)\right]+\left[y_2-\left(\frac{5}{7}+3\hat{\beta}_1\right)\right]+\left[5-\left(\frac{5}{7}+4\hat{\beta}_1\right)\right]=0\\ 7y_2-56\hat{\beta}_1 &= -34. \qquad (1.2.5)\end{aligned}$$

Second,

$$\begin{aligned}\sum_{i=1}^{3} x_i e_i &= \left[2-\left(\frac{5}{7}+\hat{\beta}_1\right)\right]+3\left[y_2-\left(\frac{5}{7}+3\hat{\beta}_1\right)\right]+4\left[5-\left(\frac{5}{7}+4\hat{\beta}_1\right)\right]=0\\ 21y_2-182\hat{\beta}_1 &= -114. \qquad (1.2.6)\end{aligned}$$

Solving (1.2.5) and (1.2.6) gives $y_2 = \boxed{2}$ (and $\hat{\beta}_1 = 6/7$). **(Answer: (C))** □

Remark. As a check, you can input $\{(x_i, y_i)\}_{i=1}^3$ into your financial calculator with $y_2 = 2$ and see whether you can get $\hat{\beta}_0 = 5/7$.

1.3 Assessing the Goodness of Fit of the Model

OPTIONAL SYLLABUS READING(S)

- Frees, Section 2.3
- James et al., Subsections 3.1.3 and 3.2.2 (P. 75-76)

From now onward, we assume that the random errors $\varepsilon_1, \ldots, \varepsilon_n$ are normally distributed, i.e., $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} \text{N}(0, \sigma^2)$ for some unknown variance σ^2. While not necessary for least squares estimation, this normality assumption is crucial to much of the statistical inference (e.g., constructing confidence intervals for and testing hypotheses on regression coefficients of interest) and prediction that follow.

1.3.1 Partitioning the Sum of Squares

Sum of squares partition. After an SLR model (and, more generally, a general linear model) is fitted, the most pressing issue we face is to assure ourselves that the model does help us better understand the behavior of the response variable (than the i.i.d. model) and, more importantly, how much better. To this end, we have to check the quality of the regression fit and quantify the strength of the relationship between the response and explanatory variables.

To begin with, note that for each observed response value y_i, we have two candidate "predictions":

(1) *The sample mean $\bar{y}$ suggested by the i.i.d. model $y = \beta_0 + \varepsilon$*

In the absence of the knowledge of x, the sample mean of the y-values is the best fitted value for each y_i, as we have seen in Subsection 1.1.1 (see Problem 1.6.6 on page 46 for rigorous justification). Doing so makes $y_i - \bar{y}$ the departure between the ith response value y_i and the ith fitted value (under the i.i.d. model).

(2) *The fitted value $\hat{y}_i$ under the SLR model*

With the knowledge of x, each y_i can be predicted by the point on the fitted regression line at $x = x_i$, that is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. The departure between the ith response value y_i and the ith fitted value then becomes $e_i = y_i - \hat{y}_i$, which is the ith residual introduced in Section 1.2.

Intuitively, if the incorporation of x is worthwhile, then the sum of the squares of the departures under the SLR model should be much less than that under the naive i.i.d. model. To quantify the improvement of the SLR model over the i.i.d. model, consider the telescoping decomposition

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}), \qquad i = 1, \ldots, n.$$

The left-hand side can be viewed as the ith residual of the i.i.d. model and the term $y_i - \hat{y}_i$ is the ith residual of the fitted SLR model. Now we square both sides of the preceding equation and sum over all $i = 1, \ldots, n$ to obtain[ix]

$$\begin{aligned}\sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \underbrace{2\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{0 \quad \text{(see footnote)}} \\ &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.\end{aligned}$$

In summary, we get the decomposition formula for various sums of squares:

$$\boxed{\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{\text{RSS}} + \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\text{Reg SS}}.} \tag{1.3.1}$$

Here, the three kinds of sums of squares are defined in Table 1.3.

The two required SRM texts, Frees and James et al., are at odds with each other in terms of how to designate and denote the three sums of squares. The abbreviations in Table 1.3 follow James et al. and the SRM sample questions, while Frees uses the symbols "Total SS", "Error SS", and

[ix] A direct algebraic proof for SLR goes as follows:

$$\begin{aligned}\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y}) \\ &= \sum_{i=1}^n [y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})][\hat{\beta}_1(x_i - \bar{x})] = \hat{\beta}_1 S_{xy} - \hat{\beta}_1^2 S_{xx} = 0.\end{aligned}$$

This proof, however, does not carry over to a general linear model, where we do not have explicit algebraic expressions for the individual LSEs.

Sum of Squares	**Abbreviation**	**Definition**	**What Does It Measure?**
Total SS	TSS	Variation of the response values about the sample mean $\bar{y}$	Amount of variability inherent in the response prior to performing regression
Residual SS or Error SS	RSS	Variation of the response values about the fitted regression line	• Goodness of fit of the SLR model (the lower, the better) • Amount of variability of the response left unexplained even after the introduction of x
Regression SS	Reg SS	Variation explained by the SLR model (or the knowledge of x)	How effective the SLR model is in explaining the variation in y

Table 1.3: The three sums of squares that constitute (1.3.1).

"Regression SS" (sometimes "Regress SS"). You should be cautioned that RSS does *not* refer to the regression sum of squares, but the residual sum of squares.

Back to the three sums of squares, note that as soon as the response values $y_1, \ldots, y_n$ have been obtained, TSS is a characteristic that does not depend on any regression model you are using (it does not involve any fitted values $\hat{y}_i$'s!); only RSS and Reg SS vary with the choice of the model. The significance of (1.3.1) is then two-fold:

1. The residual sum of squares of a regression model (given by RSS) must be less than that of the naive i.i.d. model (given by TSS). In other words, any SLR, no matter how useless the explanatory variable is, must perform better than the naive i.i.d. model with respect to the magnitude of the residual sum of squares.

2. Because TSS is kept fixed and both RSS and Reg SS are non-negative (as they are sum of squares) and sum to TSS, the higher the Reg SS of a regression model, the lower its RSS. A good regression model is then characterized by a large Reg SS, or equivalently, a low RSS.

Formally speaking, *analysis of variance* (ANOVA) is an exercise of partitioning the variation in the sample of y-values (TSS) into the variation explained by the fitted regression model (Reg SS) and the residual variation about the fitted line (RSS). It allows us to decide whether Reg SS is large enough for us to declare that the SLR model is effective.

Coefficient of determination. To examine whether Reg SS is high in proportion to TSS, it is informative to look at the *coefficient of determination* defined as

$$\boxed{R^2 = \frac{\text{Reg SS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}.} \tag{1.3.2}$$

This ratio gives an idea of the extent to which the explanatory variable x accounts for or "determines" the response variable y. Note that R^2 is always valued between 0 and 1 (because both RSS and Reg SS are non-negative and must be bounded by TSS) and seeks to measure the proportion of the variation of the response variable (about its mean) that can be explained by the regression model. The higher the value of R^2, the more effective the fitted regression line is in reducing the variation in y.

Example 1.3.1. (SOA Exam SRM Sample Question 18: Going between TSS, RSS, and R^2) For a simple linear regression model the sum of squares of the residuals is $\sum_{i=1}^{25} e_i^2 = 230$ and the R^2 statistic is 0.64.

Calculate the total sum of squares (TSS) for this model.

(A) 605.94

(B) 638.89

(C) 690.77

(D) 701.59

(E) 750.87

Solution. By (1.3.2), we solve

$$0.64 = R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{230}{\text{TSS}},$$

which gives TSS = $\boxed{638.89}$. **(Answer: (B))** □

Specialized formulas for Reg SS and R^2 under SLR. In the particular context of SLR, the regression sum of squares takes the simple form

$$\begin{aligned}\text{Reg SS} &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 \\ &\overset{(1.2.2)}{=} \sum_{i=1}^{n}[(\bar{y} - \hat{\beta}_1\bar{x}) + \hat{\beta}_1 x_i - \bar{y}]^2 \\ &= \boxed{\hat{\beta}_1^2 S_{xx}},\end{aligned} \tag{1.3.3}$$

and, as a result,

$$\text{RSS} = \text{TSS} - \text{Reg SS} = S_{yy} - \hat{\beta}_1^2 S_{xx}.$$

The ingredients used to compute the least squares estimates can therefore be recycled to determine RSS and Reg SS in a single expression. The formula for Reg SS is presented in Exercise 2.13 (b) of Frees and, as a result, it is possible (though not extremely likely) that exam questions are set on the formula. If you are aiming for Grade 10 in Exam SRM, you should not hesitate to memorize it!

The formula has a surprisingly useful consequence:

> In an SLR model, the coefficient of determination is simply the *square* of the *sample* correlation coefficient between x and y.

This follows from

$$\boxed{R^2 = \frac{\text{Reg SS}}{\text{TSS}} = \frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}} \overset{(1.2.2)}{=} \frac{S_{xy}^2}{S_{xx}S_{yy}} = \left(\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\right)^2 = r^2.} \tag{1.3.4}$$

This fact is mentioned on page 39 and followed up in Exercise 2.13 (c) of Frees.

Note that the specialized formulas for RSS, Reg SS, and R^2 above apply only to SLR.

Example 1.3.2. (SOA Course 4 Fall 2002 Question 5: Calculation of R^2 given summarized data) You fit the following model to eight observations:

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

You are given:

$$\begin{aligned} \hat{\beta}_1 &= 2.065 \\ \sum (x_i - \bar{x})^2 &= 42 \\ \sum (y_i - \bar{y})^2 &= 182 \end{aligned}$$

Determine R^2.

(A) 0.48

(B) 0.62

(C) 0.83

(D) 0.91

(E) 0.98

Solution. In terms of $\hat{\beta}_1$, the coefficient of determination is

$$R^2 = \frac{\text{Reg SS}}{\text{TSS}} = \hat{\beta}_1^2 \frac{S_{xx}}{S_{yy}} = 2.065^2 \left(\frac{42}{182}\right) = \boxed{0.984052}. \quad \textbf{(Answer: (E))}$$

□

Calculating ANOVA quantities using a financial calculator. If you are given a small dataset to work with and are asked to calculate ANOVA quantities like TSS, RSS, Reg SS, and R^2, you may want to input the data into your financial calculator and use its built-in functions to speed up your work. After inputting the data $\{(x_i, y_i)\}_{i=1}^n$ into the calculator, press `[2ND][STAT]` and you can find, among other statistics:

- The values of the sample standard deviations of x and y denoted by "`Sx`" and "`Sy`"
- The empirical (with division by n instead of $n-1$) standard deviations of x and y denoted by "σx" and "σy"

- The sample correlation between x and y denoted by "`r`"

By (1.3.4), we can compute the coefficient of determination almost effortlessly as $R^2 = r^2$ (don't forget to square!) and the total sum of squares as

$$\boxed{\text{TSS} = (n-1)s_y^2 \quad \text{or} \quad n\sigma_y^2.}$$

In view of the definitions of RSS and Reg SS, they can be calculated in terms of R^2 and TSS as

$$\boxed{\text{RSS} = \text{TSS}(1-R^2) \qquad \text{and} \qquad \text{Reg SS} = \text{TSS}(R^2).}$$

Now try the following example and see how much work the built-in functions of your calculator can save for you.

Example 1.3.3. (SOA Course 4 2000 Sample Exam Question 29: Calculation of R^2 given raw data) You wish to determine the nature of the relationship between sales (y) and the number of radio advertisements broadcast (x). Data collected on four consecutive days is shown below.

Day	Sales	Number of Radio Advertisements
1	10	2
2	20	2
3	30	3
4	40	3

Using the method of least squares, you determine the estimated regression line:

$$\hat{y} = -25 + 20x$$

Determine the value of R^2 for this model.

(A) .70

(B) .75

(C) .80

(D) .85

(E) .90

Solution 1 (By definition). The fitted values are

$$\hat{y}_1 = \hat{y}_2 = -25 + 20(2) = 15 \quad \text{and} \quad \hat{y}_3 = \hat{y}_4 = -25 + 20(3) = 35.$$

The residual sum of squares is

$$\text{RSS} = (-5)^2 + 5^2 + (-5)^2 + 5^2 = 100.$$

As $\bar{y} = 100/4 = 25$, the total sum of squares is

$$\text{TSS} = (-15)^2 + (-5)^2 + 5^2 + 15^2 = 500.$$

Hence

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{100}{500} = \boxed{0.8}. \quad \textbf{(Answer: (C))}$$

□

Solution 2 (Square of r). Inputting the four pairs of (x, y) into a financial calculator, we can get the sample correlation coefficient $r = 0.894427$. Then $R^2 = r^2 = \boxed{0.8}$. **(Answer: (C))** □

Remark. Forgetting to square and taking $r = 0.894427$ as the final answer would lead to Option (E).

ANOVA table. It is customary and convenient to tabulate the partitioning of the sum of squares using an *ANOVA table*. For an SLR model, the ANOVA table looks like:

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

Here are the features of an ANOVA table:

- Bottom item = sum of items above in the same column
- Each sum of squares (SS) accounts for a source of variation in y
- Each SS is associated with a degree of freedom (*df*). Here are some “informal” rules for counting *df*:
  - ▷ TSS represents n deviations from the sample mean $\bar{y}$, which estimates the population mean μ, and one *df* is lost from n in the process.
  - ▷ Likewise, RSS represents n deviations from the fitted regression line, which has two estimated parameters $\hat{\beta}_0$ and $\hat{\beta}_1$, and possesses $n-2$ *df*.
  - ▷ Reg SS has the leftover *df*: $(n-1)-(n-2) = 1$, corresponding to the single explanatory variable x.
  - ▷ The sum of the *df* of RSS and Reg SS must equal the *df* of TSS, which is $n-1$.
- Dividing an SS by its *df* results in a *mean square* (MS). In particular, dividing RSS by $n-2$ yields the *mean square error* (MSE)

$$\boxed{s^2 = \frac{\text{RSS}}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2},}$$

which can be shown to be an unbiased estimator of the unknown error variance σ^2. The positive square root, $s = \sqrt{s^2}$, is known as the *residual standard deviation* (see page 34 of Frees) or *residual standard error*, or RSE for short (see page 66 of James et al.).

- The "F-value" column will be explained in the next subsection.

Example 1.3.4. (SOA Course 120 Study Note 120-81-95 Question 2: Calculation of R^2 from RSS and TSS) You use simple linear regression and have observed the following five values of the dependent variable, y:

$$1, \quad 2, \quad 3, \quad 4, \quad 5.$$

You determine that $s^2 = 1$.

Calculate R^2.

(A) 0.1

(B) 0.3

(C) 0.5

(D) 0.6

(E) 0.7

Solution. As $s^2 = \text{RSS}/(5-2)$, we have $\text{RSS} = 3(1) = 3$. With $\bar{y} = 3$,

$$\text{TSS} = \sum_{i=1}^{5}(y_i - \bar{y})^2 = (1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2 = 10,$$

or, using a financial calculator,

$$\text{TSS} = 5(1.414214)^2 = 10.$$

It follows that

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{3}{10} = \boxed{0.7}. \quad \textbf{(Answer: (E))}$$

□

1.3.2 F-test

F-statistic: Definition. The F-test[x] is a formal statistical test to judge whether Reg SS is large enough for us to declare the usefulness of the fitted SLR model, with respect to explaining the

[x] In the required portions of the two SRM texts, the F-test is discussed only in James et al. in the context of multiple linear regression models (see Chapter 2 of this manual). This is a somewhat awkward and unfortunate arrangement.

variation in the response y. The formal hypotheses are

$$\boxed{\underbrace{\mathrm{H}_0 : \beta_1 = 0}_{\text{i.i.d. model}} \quad \text{vs} \quad \underbrace{\mathrm{H}_a : \beta_1 \neq 0}_{\text{SLR model}},}$$

and can be assessed by means of the *F-statistic* defined by

$$\boxed{F = \frac{\text{Reg SS}/1}{\text{RSS}/(n-2)},}$$

which is a by-product of the last column of the ANOVA table.

In this study manual, we content ourselves with the following facts (rigorous proofs require deep statistical theory!):

- Under H_0, the F-statistic has an $\boxed{F_{1,n-2}}$ distribution, i.e., an F-distribution with 1 and $n-2$ degrees of freedom. Its expected value is close to *one*. Here the two degrees of freedom are inherited from those of Reg SS and RSS, or the third column of the ANOVA table.

  (Forgot what an F-distribution is? Refer to your mathematical statistics textbook!)

- If H_0 is not true, then the F-statistic tends to take a value which is much higher than what an $F_{1,n-2}$ random variable typically assumes.

Based on the F-statistic, the decision rule is:

> At a fixed significance level of α, we reject H_0 in favor of H_a and conclude that the SLR model is significantly better than the i.i.d. model (or equivalently, the explanatory variable x is statistically significant) using the following two equivalent ways:
>
> - *Critical value approach:* The observed value of the F-statistic is greater than $F_{1,n-2,\alpha}$, which is the α-upper quantile of the F-distribution with 1 and $n-2$ degrees of freedom, i.e., $\mathbb{P}(\underbrace{F_{1,n-2}}_{\text{r.v.}} > \underbrace{F_{1,n-2,\alpha}}_{\text{quantile}}) = \alpha$.
> - *p-value approach:* The p-value $\mathbb{P}(F_{1,n-2} > f)$, where f is the observed value of F, is less than α.
>
> (Note: Recall from your VEE Mathematical Statistics course that the *p-value* of a hypothesis test is the probability of observing a value of the test statistic as extreme as or more extreme than the observed value, under the null hypothesis. It is a measure, on the scale from 0 to 1, of the strength of the evidence against H_0 in favor of H_a; the smaller the p-value, the stronger the evidence we have. *At a fixed significance level α, we reject H_0 in favor of H_a when the p-value is less than α.* Equivalently, the p-value is the smallest significance level at which the null hypothesis would be rejected.)

To our astonishment, the SRM tables do not include one for the F-distribution. Accordingly, if there are any questions in the SRM exam concerning the F-test, the focus should be on calculating the F-statistic. An exam question will need to provide you with the F-quantiles to proceed further.

***F*-statistic in terms of R^2.** One can, if needed, equivalently describe the F-statistic in terms of the coefficient of determination R^2. To this end, we connect the F-statistic and R^2 by dividing the numerator and denominator of the former by TSS, yielding

$$\boxed{F = (n-2)\left(\frac{\text{Reg SS}/\text{TSS}}{\text{RSS}/\text{TSS}}\right) = (n-2)\left(\frac{R^2}{1-R^2}\right).} \tag{1.3.5}$$

There is no need for memorizing this alternative form of the F-statistic. Just remember the trick:

Divide both the numerator and denominator of the F-statistic by TSS.

An added merit of (1.3.5) is that because R^2 can be easily calculated from raw data as the square of the sample correlation, we can also readily compute the F-statistic from raw data—there is no need to deal with ANOVA quantities like RSS, Reg SS, and TSS at all.

Example 1.3.5. (SOA Course 4 Spring 2000 Question 1: Calculation of the *F*-statistic from R^2) You fit the following model to 20 observations:

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

You determine that $R^2 = 0.64$.

Calculate the value of the F statistic used to test for a linear relationship.

(A) Less than 30

(B) At least 30, but less than 33

(C) At least 33, but less than 36

(D) At least 36, but less than 39

(E) At least 39

Solution. Given the value of R^2, the value of the F statistic, by (1.3.5), is

$$F = (n-2)\left(\frac{R^2}{1-R^2}\right) = (20-2)\left(\frac{0.64}{1-0.64}\right) = \boxed{32}. \quad \textbf{(Answer: (B))}$$

□

Example 1.3.6. [HARDER!] (Regressing y on x vs. regressing x on y) Two actuaries are analyzing the same dataset involving a pair of variables (x, y). You are given:

(i) Actuary P fits a simple linear regression model by regressing y on x.

(ii) Actuary Q fits a simple linear regression model by regressing x on y.

Determine which of the following statements about the two models must be true.

I. The estimated slope coefficient of Actuary P's model is the reciprocal of that of Actuary Q's model.

II. The two models have the same value of the coefficient of determination.

III. The two models have the same value of the F-statistic for testing for the significance of the explanatory variable.

(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).

Solution. I. False. The two estimated β_1's are generally not the inverse of each other. To see this, let's write $\hat{\beta}_1^{y\sim x}$ and $\hat{\beta}_1^{x\sim y}$ for the LSEs of the slope coefficient in Actuary P's model (y is regressed on x) and Actuary Q's model (x is regressed on y), respectively. By (1.2.4), we have

$$\hat{\beta}_1^{y\sim x} \times \hat{\beta}_1^{x\sim y} = \left(r \times \frac{s_y}{s_x}\right)\left(r \times \frac{s_x}{s_y}\right) = r^2 \overset{(1.3.4)}{=} R^2,$$

Thus $\hat{\beta}_1^{y\sim x} \neq 1/\hat{\beta}_1^{x\sim y}$, unless $R^2 = 1$.

II. True. This is because in the SLR setting, R^2 equals the square of the sample correlation coefficient between x and y (recall (1.3.4)), whose value remains unchanged if we interchange the role of x and y.

III. True. This follows from Statement II and (1.3.5). **(Answer: (D))**

□

Remark. Statement I is motivated by Exercise 2.6 of Frees.

1.4 Statistical Inference about Regression Coefficients

OPTIONAL SYLLABUS READING(S)

- Frees, Section 2.4 to Subsection 2.5.2
- James et al., Subsection 3.1.2

In SLR analysis, the regression parameters β_0 and β_1 are of primary interest (σ^2, though unknown, is of secondary importance). The slope parameter β_1 is particularly important because it quantifies the direct influence of the explanatory variable x on the response y. The LSEs $\hat{\beta}_0$ and $\hat{\beta}_1$ provide *point* estimates for the parameters, but would vary from sample to sample and not be informative unless accompanied by a standard error to quantify uncertainty. To assess the accuracy of the LSE and draw further inference about β_0 and β_1, the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are warranted.

1.4.1 Sampling Distributions of LSEs

Linear combination formulas. To explore the distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$, it may be convenient to express them as a linear combination of the response values y_i's. For $\hat{\beta}_1$, the representation is

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{S_{xx}} - \underbrace{\frac{\bar{y}\sum_{i=1}^n (x_i - \bar{x})}{S_{xx}}}_{0} = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{S_{xx}},$$

or

$$\boxed{\hat{\beta}_1 = \sum_{i=1}^n w_i y_i, \qquad \text{where } w_i = \frac{x_i - \bar{x}}{S_{xx}} \text{ for } i = 1, 2, \ldots, n.} \tag{1.4.1}$$

Note that the weights satisfy $\sum_{i=1}^n w_i = 0$, $\sum_{i=1}^n w_i x_i = 1$, and $\sum_{i=1}^n w_i^2 = 1/S_{xx}$.

An analogous weighted sum formula exists for $\hat{\beta}_0$, which is the content of Exercise 2.4 of Frees:

$$\hat{\beta}_0 = \sum_{i=1}^n w_{i,0} y_i, \qquad \text{where } w_{i,0} = \frac{1}{n} - \bar{x}w_i \text{ for } i = 1, 2, \ldots, n.$$

This formula can be derived easily from (1.4.1):

$$\hat{\beta}_0 \overset{(1.2.2)}{=} \bar{y} - \hat{\beta}_1\bar{x} \overset{(1.4.1)}{=} \frac{1}{n}\sum_{i=1}^n y_i - \bar{x}\sum_{i=1}^n w_i y_i = \sum_{i=1}^n \underbrace{\left(\frac{1}{n} - \bar{x}w_i\right)}_{w_{i,0}} y_i.$$

The following example, adapted from an old SOA exam problem, shows how the computations of the weights can be tested. Perhaps to your astonishment, calculating these weights requires more work than simply calculating $\hat{\beta}_0$ and $\hat{\beta}_1$.

Example 1.4.1. (SOA Course 4 Fall 2001 Question 13 (Adapted): LSE as a weighted average of response values) You fit the following simple linear regression model to four observations:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, 3, 4$$

You are given:

i	x_i
1	-3
2	-1
3	1
4	3

The least squares estimator of β_1 can be expressed as $\hat{\beta}_1 = \sum_{i=1}^4 w_i y_i$.
Determine (w_1, w_2, w_3, w_4).

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

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

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

(D) $(-0.30, -0.10, \quad 0.10, \quad 0.30)$

(E) $(-0.10, \quad 0.30, -0.30, \quad 0.10)$

Solution. As $\bar{x} = 0$,

$$w_i = \frac{x_i - \bar{x}}{S_{xx}} = \frac{x_i}{\sum_{i=1}^4 x_i^2} = \frac{x_i}{20},$$

so

$$\boxed{w_1 = -0.15, \quad w_2 = -0.05, \quad w_3 = 0.05, \quad w_4 = 0.15.} \qquad \textbf{(Answer: (A))}$$

□

Expectations and variances of LSEs. Because $y_1, \ldots, y_n$ are independent normal random variables, a direct consequence of the linear combination formulas above is that $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed. Furthermore, taking expectations and variances of the formulas allows us to determine the expected values and variances of $\hat{\beta}_0$ and $\hat{\beta}_1$:

- *Expectations:* $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively, i.e., $\mathbb{E}[\hat{\beta}_j] = \beta_j$ for $j = 0, 1$.

- *Variances:* The variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ are

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{nS_{xx}} \qquad \text{and} \qquad \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}.$$

Note that these variances involve the unknown parameter σ^2, which can be estimated unbiasedly by the MSE s^2, leading to the following *estimated* variances:

$$\widehat{\text{Var}}(\hat{\beta}_0) = s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) = \frac{s^2 \sum_{i=1}^n x_i^2}{nS_{xx}} \quad \text{and} \quad \widehat{\text{Var}}(\hat{\beta}_1) = \frac{s^2}{S_{xx}}. \tag{1.4.2}$$

The *estimated* standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$, denoted by $\text{SE}(\hat{\beta}_0)$ and $\text{SE}(\hat{\beta}_1)$, respectively, are called their *standard errors*:

$$\boxed{\text{SE}(\hat{\beta}_0) = \sqrt{s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} = \sqrt{\frac{s^2 \sum_{i=1}^n x_i^2}{nS_{xx}}} \quad \text{and} \quad \text{SE}(\hat{\beta}_1) = \sqrt{\frac{s^2}{S_{xx}}}.} \tag{1.4.3}$$

These are measures of the reliability, or precision, of the LSEs. Observe from (1.4.3) that the standard errors of both $\hat{\beta}_0$ and $\hat{\beta}_1$ are increasing in s^2 but decreasing in S_{xx}. Therefore, other things being equal, the standard errors will be smaller if the observations exhibit a greater tendency to lie closer to the fitted regression line (so that s^2 is smaller), and if the observed values of the explanatory variable are more spread out (so that S_{xx} is larger).

EXAM NOTE

Even though (1.4.1), (1.4.2), and (1.4.3) can be derived as special cases of general results in linear regression models you will learn in the next chapter, it is suggested that you memorize these formulas as deriving them from first principles takes considerable time.

Example 1.4.2. (SOA Course 4 Spring 2001 Question 40: Standard error of $\hat{\beta}_1$) For a two-variable regression based on seven observations, you are given:

(i) $\sum(x_i - \bar{x})^2 = 2000$

(ii) $\sum e_i^2 = 967$

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

(A) 0.26

(B) 0.28

(C) 0.31

(D) 0.33

(E) 0.35

Solution. From (ii), the MSE is $s^2 = \text{RSS}/(n-2) = 967/(7-2) = 193.4$. By (1.4.3),

$$\text{SE}(\hat{\beta}_1) = \sqrt{\frac{s^2}{S_{xx}}} = \sqrt{\frac{193.4}{2,000}} = \boxed{0.3110}. \quad \textbf{(Answer: (C))}$$

□

1.4.2 Hypothesis Tests and Confidence Intervals

***t*-test.** Armed with the distributional results in the preceding subsection, we are now in a position to formulate hypothesis tests on the regression coefficients in the form of $\text{H}_0 : \beta_j = d$,[xi] where d is a user-specified hypothesized value, for $j = 0$ or 1. The hypothesis test of greatest interest is arguably $\text{H}_0 : \beta_1 = 0$ (i.e., $j = 1$ and $d = 0$), in which case the SLR model no longer includes the explanatory variable x. Such a hypothesis test therefore allows us to examine the importance of x using tools in the hypothesis testing framework.

To gauge the plausibility of the general null hypothesis $\text{H}_0 : \beta_j = d$, we examine the proximity

[xi] Never write $\text{H}_0 : \hat{\beta}_j = d$, that is, never state the hypothesis in terms of estimators, which are random variables. Our interest is in the unknown parameter β_j, not the LSE $\hat{\beta}_j$.

Alternative Hypothesis H_a	Decision Rule	p-value (t is the observed value of $t(\hat{\beta}_j)$)
$\beta_j \neq d$	$\lvert t(\hat{\beta}_j)\rvert > t_{n-2,\alpha/2}$	$\mathbb{P}(\lvert t_{n-2}\rvert > \lvert t\rvert) = 2\mathbb{P}(t_{n-2} > \lvert t\rvert)$
$\beta_j > d$	$t(\hat{\beta}_j) > t_{n-2,\alpha}$	$\mathbb{P}(t_{n-2} > t)$
$\beta_j < d$	$t(\hat{\beta}_j) < -t_{n-2,\alpha}$	$\mathbb{P}(t_{n-2} < t)$

Table 1.4: Decision-making procedures for testing $\text{H}_0 : \beta_j = d$ against various alternative hypotheses by means of a t-test.

of $\hat{\beta}_j$ to d, scaled by the standard error of $\hat{\beta}_j$, via the *t-statistic* (or *t-ratio*) defined by

$$\boxed{t(\hat{\beta}_j) = \frac{\text{LSE} - \text{hypothesized value}}{\text{standard error of LSE}} = \frac{\hat{\beta}_j - d}{\text{SE}(\hat{\beta}_j)}, \quad j = 0, 1,}$$

where the denominator is given in (1.4.3). The reason why $t(\hat{\beta}_j)$ is known as the t-statistic is that under H_0, it can be shown that $t(\hat{\beta}_j)$ follows a t-distribution with $n-2$ degrees of freedom[xii], i.e.,

$$\boxed{t(\hat{\beta}_j) \overset{\text{H}_0}{\sim} t_{n-2}.}$$

This forms the basis for the formulation of decision rules for given significance level α, and the computation of p-values for various alternative hypotheses, as shown in Table 1.4. Here, we denote by $t_{n-2,\alpha}$ the α-upper percentile from the t-distribution with $n-2$ degrees of freedom, that is

$$\mathbb{P}(\underbrace{t_{n-2}}_{\text{random variable}} \geq \underbrace{t_{n-2,\alpha}}_{\text{upper percentile}}) = \alpha.$$

To make sense of the decision rule and the formula for the p-value in Table 1.4, consider, for instance, testing H_0 against the one-sided alternative $\text{H}_a : \beta_j > d$. To see what values of the t-statistic constitute evidence against H_0 in support of H_a, we write

$$t(\hat{\beta}_j) = \frac{\hat{\beta}_j - d}{\text{SE}(\hat{\beta}_j)} = \underbrace{\frac{\hat{\beta}_j - \overbrace{\beta_j}^{\text{true parameter}}}{\text{SE}(\hat{\beta}_j)}}_{\sim t_{n-2}\ (\text{always})} + \underbrace{\frac{\beta_j - \overbrace{d}^{\text{hypothesized value}}}{\text{SE}(\hat{\beta}_j)}}_{>0\ (\text{under H}_a:\beta_j>d)}.$$

This seemingly unnecessary way of writing reveals that if the alternative hypothesis is true, then the t-statistic tends to take an observed value which is systematically larger than what a t_{n-2} distribution typically assumes. Therefore, a large t-statistic value is evidence against H_0 in favor of H_a. Similar considerations can be used to justify the decision rule and the formula for the p-value for $\text{H}_a : \beta_j \neq d$ (extremely big or extremely small values are against H_0 in favor of H_a) and $\text{H}_a : \beta_j < d$ (small values are against H_0 in favor of H_a).

In the SRM exam, you may be asked to calculate the value of the t-statistic and, based on which, decide whether to accept or reject H_0 given a significance level α. For the latter task, you will need the upper quantiles of the t_{n-2}-distribution, which you can obtain from the t-table provided in the SRM exam. Part of the table reads:

[xii] In the language of mathematical statistics, the t-statistic $t(\hat{\beta}_j)$ is a pivotal quantity. It is a function of the unknown parameter β_j but has a distribution which is free of β_j.

df	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$
1	3.0777	6.3138	12.7062	31.8205	63.6567
2	1.8856	2.9200	4.3027	6.9646	9.9248
⋮	⋮	⋮	⋮	⋮	⋮

For example, $t_{2,0.025} = 4.3027$.

Example 1.4.3. (SOA Course 4 Fall 2003 Question 5: Calculation of *t*-statistic) For the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $i = 1, 2, \ldots, 10$, you are given:

(i) $x_i = \begin{cases} 1, & \text{if the } i\text{th individual belongs to a specified group 0} \\ 0, & \text{otherwise} \end{cases}$

(ii) 40 percent of the individuals belong to the specified group.

(iii) The least squares estimate of β_1 is $\hat{\beta}_1 = 4$.

(iv) $\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 92$

Calculate the t-statistic for testing $\mathrm{H}_0 : \beta_1 = 0$.

(A) 0.9

(B) 1.2

(C) 1.5

(D) 1.8

(E) 2.1

Solution. To calculate the t-statistic, we need the estimated variance or standard error of $\hat{\beta}_1$. From (iv), $s^2 = \mathrm{RSS}/(n-2) = 92/(10-2) = 11.5$. With $\bar{x} = 4/10 = 0.4$, the estimated variance of $\hat{\beta}_1$ is

$$\widehat{\mathrm{Var}}(\hat{\beta}_1) = \frac{s^2}{S_{xx}} = \frac{11.5}{4(1-0.4)^2 + 6(-0.4)^2} = \frac{115}{24}.$$

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

$$t(\hat{\beta}_1) = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)} = \frac{4}{\sqrt{115/24}} = \boxed{1.8273}. \quad \textbf{(Answer: (D))}$$

□

Remark. (i) Since $t_{8,0.1} = 1.3968$ and $t_{8,0.05} = 1.8595$, the p-value of the test (against the two-sided alternative $\mathrm{H}_a : \beta_1 \neq 0$) is between $2(0.05) = 0.1$ and $2(0.1) = 0.2$.

(ii) The explanatory variable x here is an example of a binary variable; see Subsection 2.3.1.

Confidence intervals for regression coefficients. The fact that $t(\hat{\beta}_j) \overset{\text{H}_0}{\sim} t_{n-2}$, besides underlying the t-test above, can also be exploited to construct confidence intervals for the two regression coefficients β_0 and β_1. Starting with the probability statement

$$\mathbb{P}\left(\underbrace{-t_{n-2,\alpha/2}}_{\text{(by symmetry)}} < \frac{\hat{\beta}_j - \beta_j}{\text{SE}(\hat{\beta}_j)} < t_{n-2,\alpha/2}\right) = 1-\alpha$$

and making the unknown parameter β_j the subject of the event on the left-hand side, we have

$$\mathbb{P}\left(\hat{\beta}_j - t_{n-2,\alpha/2} \times \text{SE}(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + t_{n-2,\alpha/2} \times \text{SE}(\hat{\beta}_j)\right) = 1-\alpha.$$

This shows that a $100(1-\alpha)\%$ *confidence interval* for β_j[xiii] takes the easy-to-remember form:

$$\boxed{\text{LSE} \pm t\text{-quantile} \times \text{Standard error} = \hat{\beta}_j \pm t_{n-2,\alpha/2} \times \text{SE}(\hat{\beta}_j),}$$

where again the standard error is given in (1.4.3). For the t-quantile, make sure that you use $\alpha/2$ as the probability level due to the equal-tailed nature of the confidence interval, i.e., the probability that β_j exceeds the upper bound of the confidence interval and the probability that β_j is less than the lower bound are both equal to $\alpha/2$.

Example 1.4.4. (SOA Course 4 Fall 2002 Question 38: Confidence interval for $\boldsymbol{\beta_0}$) You fit a two-variable linear regression model to 20 pairs of observations.

You are given:

(i) The sample mean of the independent variable is 100.

(ii) The sum of squared deviations from the mean of the independent variable is 2266.

(iii) The ordinary least-squares estimate of the intercept parameter is 68.73.

(iv) The error sum of squares is 5348.

Determine the lower limit of the symmetric 95% confidence interval for the intercept parameter.

(A) −273

(B) −132

(C) −70

(D) −8

(E) −3

[xiii] Never say a $100(1-\alpha)\%$ confidence interval for $\hat{\beta}_j$!

Solution. We need the standard error of $\hat{\beta}_0$. As $s^2 = \text{RSS}/(n-2) = \underbrace{5,348}_{\text{(iv)}}/(20-2) = 2,674/9$, the estimated variance of $\hat{\beta}_0$, by (1.4.2), is

$$\widehat{\text{Var}}(\hat{\beta}_0) = s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) = \frac{2,674}{9}\left(\frac{1}{20} + \frac{100^2}{2,266}\right) = 1,326.0255.$$

The lower limit of the symmetric 95% confidence interval for β_0 is

$$\underbrace{68.73}_{\text{(iii)}} - \underbrace{t_{18,0.025}}_{2.1009}\sqrt{1,326.0255} = \boxed{-7.77}. \qquad \textbf{(Answer: (D))}$$

□

Example 1.4.5. (SOA Course 4 Fall 2001 Question 5: Confidence interval for β_1 – I) You fit the following model to eight observations:

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

You are given:

$$\begin{aligned} \hat{\beta}_1 &= -35.69 \\ \sum(x_i - \bar{x})^2 &= 1.62 \\ \sum(y_i - \hat{y}_i)^2 &= 2394 \end{aligned}$$

Determine the symmetric 90-percent confidence interval for β_1.

(A) $(-74.1, 2.7)$

(B) $(-66.2, -5.2)$

(C) $(-63.2, -8.2)$

(D) $(-61.5, -9.9)$

(E) $(-61.0, -10.4)$

Solution. The MSE is $s^2 = \text{RSS}/(n-2) = 2,394/(8-2) = 399$. By (1.4.2), the estimated variance of $\hat{\beta}_1$ is

$$\widehat{\text{Var}}(\hat{\beta}_1) = \frac{s^2}{S_{xx}} = \frac{399}{1.62} = 246.2963.$$

The symmetric 90% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{6,0.05}\sqrt{\widehat{\text{Var}}(\hat{\beta}_1)} = -35.69 \pm 1.9432\sqrt{246.2963} = \boxed{(-66.18, -5.20)}. \quad \textbf{(Answer: (B))}$$

□

Example 1.4.6. (CAS Exam ST Fall 2014 Question 20: Confidence interval for $\boldsymbol{\beta_1}$ – II) For the linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, you are given:

- $n = 6$
- $\hat{\beta}_1 = 4$
- $\sum_{i=1}^n (x_i - \bar{x})^2 = 50$
- $\sum_{i=1}^n (y_i - \bar{y})^2 - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = 25$

Calculate the upper bound of the 95% confidence interval for β_1.

(A) Less than 5.1

(B) At least 5.1, but less than 5.3

(C) At least 5.3, but less than 5.5

(D) At least 5.5, but less than 5.7

(E) At least 5.7

Solution. We are given in the fourth point that

$$S_{yy} - \hat{\beta}_1 S_{xy} \overset{(1.2.2)}{=} \text{TSS} - \hat{\beta}_1(\hat{\beta}_1 S_{xx}) = \text{TSS} - \text{Reg SS} = \text{RSS} = 25.$$

Thus the MSE is $s^2 = \text{RSS}/(n-2) = 25/(6-2) = 6.25$ and the upper bound of the 95% confidence interval for β_1 is

$$\hat{\beta}_1 + t_{4,0.025} \times \sqrt{\frac{s^2}{S_{xx}}} = 4 + 2.7764 \times \sqrt{\frac{6.25}{50}} = \boxed{4.9816}. \quad \textbf{(Answer: (A))}$$

□

[HARDER!] Relationship between $\boldsymbol{F}$-test and $\boldsymbol{t}$-test for $\text{H}_0 : \beta_1 = 0$**.** Thus far, we have introduced two ways to test $\text{H}_0 : \beta_1 = 0$:

1. By the F-test in Section 1.3, with test statistic

$$F = \frac{\text{Reg SS}/1}{\text{RSS}/(n-2)}.$$

2. By the t-test in this section, with test statistic

$$t(\hat{\beta}_1) = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{s^2/S_{xx}}}.$$

Do these two hypothesis tests always give the same conclusion?

	F-test	t-test
1.	More convenient for testing whether $\beta_1 = 0$.	Equally convenient for testing whether β_1 equals any hypothesized value, e.g., $\beta_1 = 2.5$.
2.	The alternative hypothesis is usually two-sided, e.g., $\text{H}_a : \beta_1 \neq 0$.	The alternative hypothesis can be two-sided or one-sided, e.g, $\text{H}_a : \beta_1 > 0$.

Table 1.5: Differences between the F-test and t-test for testing $\text{H}_0 : \beta_1 = 0$.

It turns out that there is an intimate relationship between these two statistical tests. Specifically, the t-statistic and the F-statistic enjoy a one-to-one relationship given by

$$\boxed{t(\hat{\beta}_1)^2 = \frac{\hat{\beta}_1^2}{s^2/S_{xx}} = \frac{\hat{\beta}_1^2 S_{xx}}{s^2} \overset{(1.3.3)}{=} \frac{\text{Reg SS}/1}{\text{RSS}/(n-2)} = F.}$$

Together with the distributional equality $t_v^2 = F_{1,v}$ for any $v \geq 0$ (a more subtle fact you may have seen in your VEE Mathematical Statistics class), the t-test and F-test indeed have the same rejection region and are equivalent ways of testing $\text{H}_0 : \beta_1 = 0$ against $\text{H}_a : \beta_1 \neq 0$.

Table 1.5 summarizes the differences between the F-test and t-test for testing $\text{H}_0 : \beta_1 = 0$.

Example 1.4.7. (SOA Course 120 Study Note 120-83-96 Question 2: Given ANOVA output, find the t-statistic) You fit the simple linear regression model to 47 observations and determine $\hat{y} = 1.0 + 1.2x$. The total sum of squares (corrected for mean) is 54, and the regression sum of squares is 7.

Determine the value of the t-statistic for testing $\text{H}_0 : \beta_1 = 0$ against $\text{H}_a : \beta_1 \neq 0$.

(A) 0.4

(B) 1.2

(C) 2.2

(D) 2.6

(E) 6.7

Solution. The value of the F-statistic for testing $\text{H}_0 : \beta_1 = 0$ against $\text{H}_a : \beta_1 \neq 0$ is

$$F = \frac{\text{Reg SS}/1}{\text{RSS}/(n-2)} = \frac{7/1}{(54-7)/(47-2)} = \frac{315}{47}.$$

The value of the t-statistic for the same hypotheses is the *positive* square root of F (we take the positive root because $\hat{\beta}_1 = 1.2 > 0$; note that $\hat{\beta}_1$ and $t(\hat{\beta}_1)$ share the same sign), or

$$t(\hat{\beta}_1) = \sqrt{315/47} = \boxed{2.5889}. \qquad \textbf{(Answer: (D))}$$

□

Remark. Here is a solution without using the fact that $t(\hat{\beta}_1)^2 = F$:

As RSS $= 54 - 7 = 47$, the MSE is $s^2 = 47/45$. For SLR, Reg SS $= \hat{\beta}_1^2 S_{xx}$, so $S_{xx} = 7/1.2^2 = 4.861111$. Then

$$t(\hat{\beta}_1) = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} = \frac{1.2}{\sqrt{(47/45)/4.861111}} = 2.5889.$$

1.5 Prediction

OPTIONAL SYLLABUS READING(S)

- Frees, Subsection 2.5.3
- James et al., Section 3.2, P. 81-82

Prediction vs. estimation. Now that statistical inference has been settled in the preceding section, we consider in this section a problem in a similar vein, that is, to *predict* the response variable y when the explanatory variable is set at some known value, say x_*. Note that a prediction problem is fundamentally different from the previous estimation problem in the sense that we are now interested in a *random individual* response, say y_*, in contrast to an unknown parameter β_j. The variability stemming from the random nature of y_* needs to be specifically taken into account in the prediction procedure, especially when formulating prediction intervals. Because of this extra degree of variability, prediction is generally less precise than estimation with a bigger standard error.

Setting. The following diagram visualizes the prediction problem of interest:

	response		known values of explanatory variables
	$\underline{y}$		$\underline{x}$
observed (past) data	y_1 y_2 $\vdots$ y_n		x_1 x_2 $\vdots$ x_n
Unobserved (future) data	y_* (target)	←	x_*

Two assumptions are typically necessary for the validity of our prediction procedure:

1. The future, yet-to-be-realized response value y_* is subject to the same data-generating mechanism (i.e., the SLR model) that governs the currently available observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Mathematically, we have $y_* = \beta_0 + \beta_1 x_* + \varepsilon_*$, where

   β_0 and β_1 are the (unknown) parameters in the same SLR model,

   x_* is the x-value of interest, and

ε_* is the normal error term that underlies y_*.

The fact that y_* comes from the same SLR model allows us to make use of the information about the model based on the realized observations, particularly the estimates of $\beta_0, \beta_1, \sigma^2$.

2. The future error term ε_* and the past error terms $\varepsilon_1, \ldots \varepsilon_n$ are independent. This is equivalent to the independence between the future response y_* and the past response values $y_1, \ldots, y_n$. This independence assumption is crucial to decomposing the variance of the prediction error into two distinguishing parts, as will be shown below.

Prediction intervals. Given the LSEs $\hat{\beta}_0$ and $\hat{\beta}_1$, a sensible point predictor of $y_* = \beta_0 + \beta_1 x_* + \varepsilon_*$ is obtained by replacing $\beta_0, \beta_1, \varepsilon_*$ by $\hat{\beta}_0, \hat{\beta}_1, 0$, respectively:

$$\boxed{\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*.}$$

As soon as the data values are observed, $\hat{\beta}_0$ and $\hat{\beta}_1$ can be readily computed to yield a single point prediction. To provide a range of reliability, it is often informative to accompany the point prediction with an interval prediction enclosing y_* with a specified probability.

The construction of such an interval prediction is a more complicated task. To this end, we look at the sampling distribution of the *prediction error* $y_* - \hat{y}_*$, which can be decomposed algebraically as

$$\underbrace{y_*}_{\text{future (random)}} - \underbrace{\hat{y}_*}_{\text{past}} = \underbrace{\varepsilon_*}_{\text{deviation inherent in } y_*} + \underbrace{[(\beta_0 + \beta_1 x_*) - (\hat{\beta}_0 + \hat{\beta}_1 x_*)]}_{\text{error in estimating the regression line at } x_*} .$$

Because $\hat{y}_*$ is calculated from the observed past data but y_* relates only to the unobserved future response, $\hat{y}_*$ and y_* are independent and follow their respective normal distributions. It follows that the prediction error is also normally distributed with mean

$$\mathbb{E}[y_* - \hat{y}_*] = (\beta_0 + \beta_1 x_*) - (\beta_0 + \beta_1 x_*) = 0,$$

i.e., our point predictor $\hat{y}_*$ is accurate on average, and with variance

$$\begin{aligned}\mathrm{Var}(y_* - \hat{y}_*) &= \mathrm{Var}(\varepsilon_*) + \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_*) \\ &= \vdots \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}\right],\end{aligned}$$

i.e.,

$$y_* - \hat{y}_* \sim \mathrm{N}\left(0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}\right]\right).$$

Upon estimating σ^2 unbiasedly by s^2, the standard error of prediction[xiv] is

$$\mathrm{SE}(y_* - \hat{y}_*) = \sqrt{s^2 \left[1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}\right]}, \tag{1.5.1}$$

[xiv] This is the formula stated at the bottom of page 40 of Frees without proof. Note that the standard error of prediction does not equal the standard error of $\hat{y}_*$ because our target y_* itself is also random.

and upon studentization, we have

$$\frac{y_* - \hat{y}_*}{\text{SE}(y_* - \hat{y}_*)} \sim t_{n-2}.$$

With this distributional result, a $100(1-\alpha)\%$ *prediction interval*[xv] for y_* is

$$\boxed{\hat{y}_* \pm t_{n-2,\alpha/2} \times \text{SE}(y_* - \hat{y}_*) = (\hat{\beta}_0 + \hat{\beta}_1 x_*) \pm t_{n-2,\alpha/2}\sqrt{s^2\left[1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}\right]}.} \quad (1.5.2)$$

Example 1.5.1. (SOA Course 120 May 1991 Question 7: Estimated variance of prediction error) You are representing 10 pairs of observations (x_i, y_i) by the following model:

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

where ε is a random error term with mean 0 and variance σ^2.

You have determined:

$$\begin{aligned}\sum_{i=1}^{10} x_i &= 50\\ \sum_{i=1}^{10} x_i^2 &= 750\\ s^2 &= 100\end{aligned}$$

Calculate the estimated variance of the predicted value of y when $x = 10$.

(A) 100

(B) 105

(C) 110

(D) 115

(E) 120

Comments: The phrase "estimated variance of the predicted value of y" is misleading. Literally it means $\widehat{\text{Var}}(\hat{y}_*)$. What the question really requests is the estimated variance of the "prediction error."

Solution. With $S_{xx} = \sum x_i^2 - n\bar{x}^2 = 750 - 10(5)^2 = 500$, the estimated variance of the prediction error for $x = 10$ is

$$s^2\left[1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}\right] = 100\left[1 + \frac{1}{10} + \frac{(10-5)^2}{500}\right] = \boxed{115}. \quad \textbf{(Answer: (D))}$$

□

[xv] A $100(1-\alpha)\%$ prediction interval for a random variable Y is defined to be a *random* interval $[A, B]$ such that $\mathbb{P}(A \le Y \le B) = 1 - \alpha$. Parenthetically, a $100(1-\alpha)\%$ confidence interval for $\mathbb{E}[y_*] = \beta_0 + \beta_1 x_*$ is obtained by replacing $\text{SE}(y_* - \hat{y}_*)$ by $\text{SE}(\hat{y}_*)$, treating as if your target y_* has no variability.

Example 1.5.2. (SOA Course 120 Study Note 120-83-96 Question 3: Width of a prediction interval) You fit the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ to 10 observed values (x_i, y_i).

You determine:

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

Determine the width of the shortest symmetric 95% prediction interval for y when $x = 8$.

(A) 0.9

(B) 1.3

(C) 1.5

(D) 1.7

(E) 2.9

Solution. The MSE is

$$s^2 = \frac{\text{RSS}}{n-2} = \frac{2.79}{10-2} = 0.34875.$$

The width of the 95% prediction interval for y when $x = 8$ is

$$\begin{aligned}2t_{8,0.025}\sqrt{s^2\left[1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}\right]} &= 2(2.3060)\sqrt{0.34875\left[1 + \frac{1}{10} + \frac{(8-6)^2}{180}\right]}\\ &= \boxed{2.8853}. \qquad \textbf{(Answer: (E))}\end{aligned}$$

□

Remark. The values of $\sum(y_i - \bar{y})^2$ and $\bar{y}$ are not needed.

Some remarks on the structure of the prediction interval. Although the formula for the prediction interval above looks formidable, you can make sense of its structure by looking at the expression of the estimated variance of the prediction, which is

$$\widehat{\text{Var}}(y_* - \hat{y}_*) = \underbrace{s^2}_{\textcircled{1}} + \underbrace{s^2\left[\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}\right]}_{\textcircled{2}}. \tag{1.5.3}$$

Loosely speaking, there are two sources of uncertainty associated with prediction:

1. *Estimation of the true regression line at x_*:* The LSEs $\hat{\beta}_0$ and $\hat{\beta}_1$ are only estimates of β_0 and β_1, and are subject to sampling fluctuations. The intrinsic variability of $\hat{\beta}_0$ and $\hat{\beta}_1$ is

reflected in ②, which is essentially $\widehat{\text{Var}}(\hat{y}_*) = \widehat{\text{Var}}(\hat{\beta}_0 + \hat{\beta}_1 x_*)$. This part of the variability of the prediction also depends critically on the value of x_*, at which prediction is made. The variance of the prediction error is minimized when x_* equals the sample mean $\bar{x}$ of the explanatory variable and increases quadratically as x_* moves away from $\bar{x}$. In other words, prediction will become less and less accurate in the region far from the center of the observed data.

2. *Variability of the random error ε_*:* Even if we know the true values of β_0 and β_1, the future response value y_* still cannot be predicted perfectly because of the inherent random error ε_* with variance σ^2, which is ① $= \widehat{\text{Var}}(\varepsilon_*)$. The extra s^2 that appears in (1.5.3) is a measure of the contribution of this source of uncertainty, which has nothing to do with the parameter estimation process.

Example 1.5.3. (SOA Course 120 November 1990 Question 4: Decomposing the variance of the prediction error into two parts) You have performed a simple regression of the form $y = \beta_0 + \beta_1 x + \varepsilon$.

You are given:

$$\begin{aligned} n &= 25 \\ x_* &= 3\bar{x} \\ \sum (x_i - \bar{x})^2 &= 25\bar{x}^2 \end{aligned}$$

Determine the fraction of the variance in $y_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$ that is due to the variability of the least squares line about the regression line.

(A) $\dfrac{1}{6}$

(B) $\dfrac{4}{25}$

(C) $\dfrac{21}{46}$

(D) $\dfrac{5}{6}$

(E) The correct answer cannot be determined from the data given.

Solution. The estimated variance due to the variability of the least squares line is

$$s^2 \left[\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}} \right] = s^2 \left[\frac{1}{25} + \frac{(3\bar{x} - \bar{x})^2}{25\bar{x}^2} \right] = \frac{s^2}{5},$$

while the estimated variance due to the variability of y_* is s^2. By division, the required fraction is $(1/5)/(1 + 1/5) = \boxed{1/6}$. **(Answer: (A))** □