ACTEX

SOA Exam SRM Study Manual

With *StudyPlus*+

***StudyPlus*+ gives you digital access* to:**

- Actuarial Exam & Career Strategy Guides
- Technical Skill eLearning Tools
- Samples of Supplemental Textbook
- And more!

**See inside for keycode access and login instructions*

Fall 2018 Edition

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

ACTEX Learning | Learn Today. Lead Tomorrow.

Actuarial & Financial Risk Resource Materials
Since 1972

Printed in the United States of America.

Contents

Preface

Exam SRM (Statistics for Risk Modeling) is a brand new exam which will be offered for the first time in September 2018 by the Society of Actuaries. It will be delivered via computer-based testing (CBT) from September 6 to September 12. The registration deadline is July 31, 2018. This new exam is a replacement of the old Validation by Educational Experiences (VEE) Applied Statistics requirement and serves as the formal prerequisite for the new Predictive Analytics exam to be offered in December 2018. The construction of Exam SRM is an important step that the SOA takes to incorporate more statistics, most notably predictive modeling, into the actuarial curriculum and to enhance the statistical literacy of actuaries in this day and age. You will considerably sharpen your statistics toolkit as a result of taking (and, in all likelihood, passing!) Exam SRM.

It is assumed that you have taken a 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, hypothesis tests, and p-values. In Exam SRM, we will make intensive use of these terms as well perform point/interval estimation, hypothesis tests, and predictions. Although not strictly necessary, prior exposure to the R programming language will be beneficial.

Syllabus

The syllabus of Exam SRM, available from `https://www.soa.org/Files/Edu/2018/2018-exam-srm-syllabi.pdf`, is very broad (but not necessarily deep) in scope, covering miscellaneous topics in linear regression models, generalized linear models, statistical learning, and time series analysis, many of which are new topics not tested in any SOA past exams. As a rough estimate, you need at least *three months* of study to master the material in this exam. The six main topics of the syllabus along with their approximate weights in the exam and the relevant chapters of 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 10
5. Decision Trees	10–15%	Chapters 8 and 9
6. Cluster Analysis	10–15%	Chapter 11

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 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 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.

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 either of the following two categories, as the SRM sample questions indicate:

1. *Simple computational questions:* Many statistical models in the exam syllabus require computers to implement. Constrained by its multiple-choice nature, however, the exam will not ask you to use software packages to analyze a large dataset, nor will it require that you work out bookwork proofs. Rather, you should expect to see many questions 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 or interval predictions, and assessing the goodness of the model, all of which require only pen-and-paper calculations.

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

2. *Conceptual questions:* A small number of questions may be set on conceptual items, meant to test the mechanics, motivations, and pros and cons of different statistical methods. Sample Questions 7, 8, 16, 20, 22, 25, for instance, belong to this type. You may be given several statements and asked to pick the correct ones.

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`.

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)

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 somewhat in 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 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 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.

- To succeed in any (actuarial) exam, the importance of practicing a wide variety of non-trivial problems to sharpen your understanding and to develop proficiency, as always, cannot be overemphasized. This study manual embraces this learning by doing approach and intersperses its expositions with more than **160 in-text examples** and **250 end-of-chapter/section problems** (the harder ones are labeled as [**HARDER!**]), which are original or taken from relevant SOA/CAS past exams, all with step-by-step solutions and problem-solving remarks, to consolidate your understanding and to 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 will not shy away from explaining the meaning of various formulas in the syllabus. The interpretations and insights we present will foster a genuine understanding of the syllabus material and discourage slavish memorization. At times, we will present brief derivations in the hope that they can help you appreciate the 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.

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

- 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.

- A full-length practice exam designed to mimic the real exam concludes this study manual giving you a holistic review of the syllabus material.

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 from `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 by *sending any errors you identify to* `ambrose-lo@uiowa.edu` *and c.c.* `support@actexmadriver.com`. Compliments and criticisms are also welcome. The authors will try their best to respond to any inquiries within 48 hours 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
June 2018

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* (Fall 2018 Edition) and the textbook *Derivative Pricing: A Problem-Based Primer*.

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

[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.

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.

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.

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

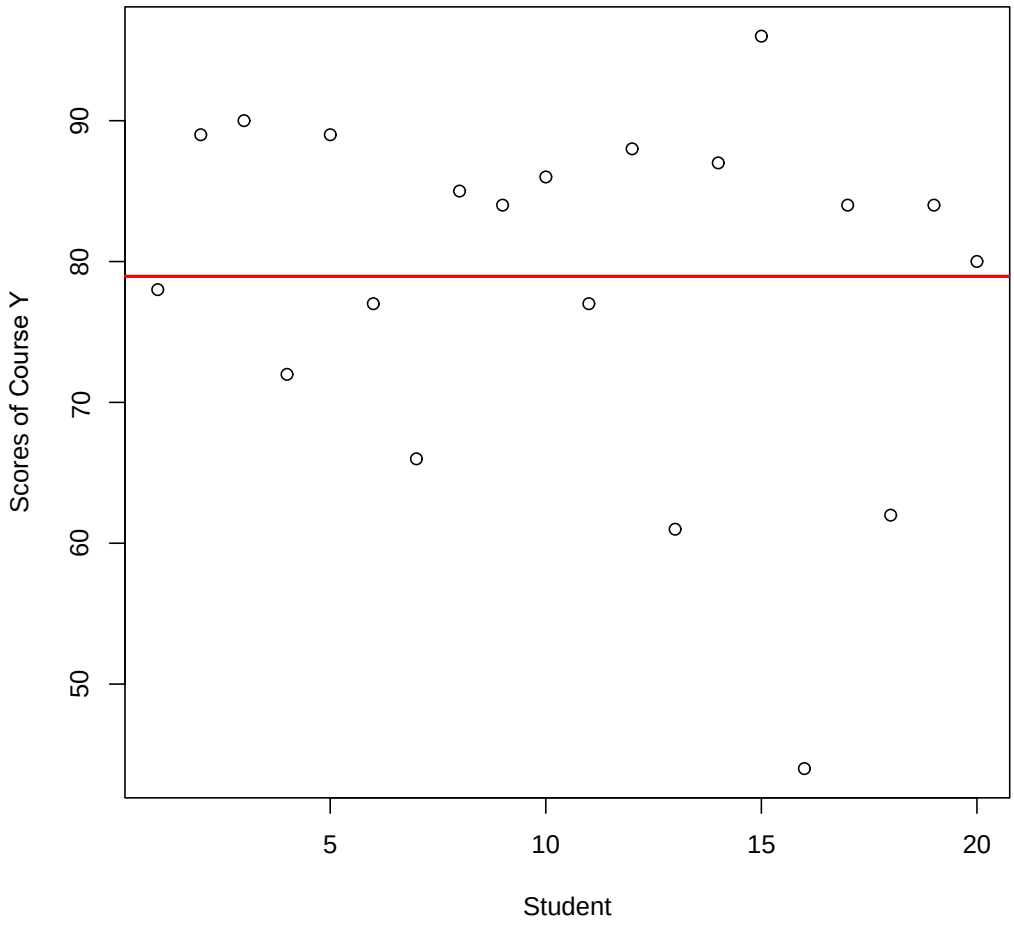


Figure 1.1.1: Scores of 20 students who took Course Y. The red horizontal line represents the sample mean level of $\bar{y} = 78.80$.

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.

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

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

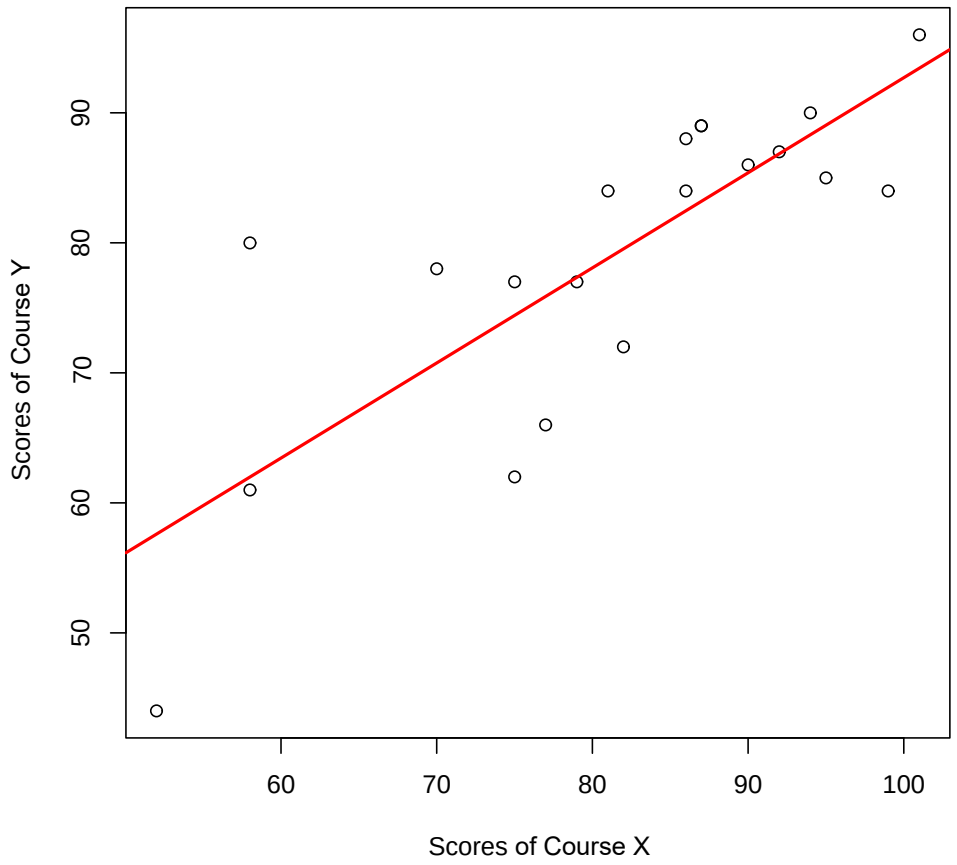


Figure 1.1.2: A plot of scores of Course Y against scores of Course X. The red sloped line is fitted by the method of least squares.

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)? This will be addressed in Sections 1.3 and 1.4.

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 Part I 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 "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

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

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 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).

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 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 $\frac{\partial}{\partial \beta_0}\text{SS}(\hat{\beta}_0, \hat{\beta}_1) = \frac{\partial}{\partial \beta_1}\text{SS}(\hat{\beta}_0, \hat{\beta}_1) = 0$ 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.

[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

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

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 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:

  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][↓]
     ⋮
     (*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. Never mix up the order!)
  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!

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:

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

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

Determine $\hat{\beta}_1$.

(A) 0.8

(B) 0.9

(C) 1.0

(D) 1.1

(E) 1.2

Solution. Following the steps above (which you should try!), 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, where all of the observations are available, you may be given only summarized information such as the values of

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

  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 expression of $\hat{\beta}_1$ in (1.2.2) and get

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \qquad \text{and} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},} \tag{1.2.3}$$

  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.

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 (1.2.2), 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:

$$\hat{\beta}_0 = 7, \qquad \hat{\beta}_1 = 4$$

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 $\bar{z} = 2\bar{y} - 3$. Consider

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

which gives

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

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.

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 are obtained from the model equation (1.1.1) with the 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 1.2.4. (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
- 16 ounces for 1.20

[viii] The symbol e_i is used in both Frees and James et al. In my 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.

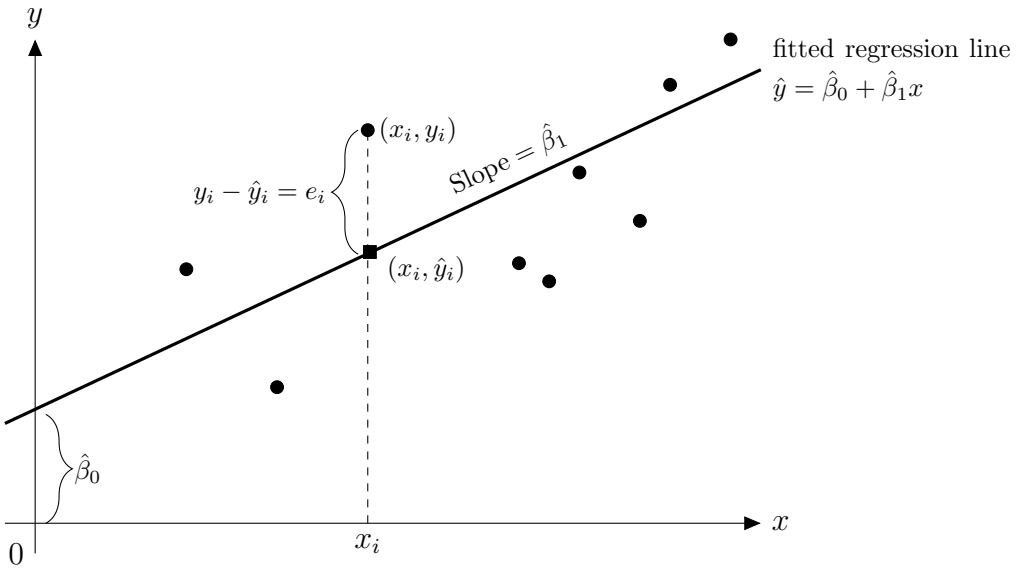


Figure 1.2.1: Graphical illustration of the fitted regression line and the definitions of the fitted value and residual. The black dots denote the observed data and the square denotes the fitted value at $x = x_i$.

- 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 1.2.5. (CAS Exam 3L Spring 2013 Question 25: Calculation of residual given raw data) You are given the following data set with two variables, x and y:

x	y
10	22
13	20
20	6
15	18
5	10

Using the least squares method for a linear regression with y as the dependent variable, calculate the absolute value of the residual where the x and y values are 12 and 18 respectively.

(A) Less than 1

(B) At least 1, but less than 2

(C) At least 2, but less than 3

(D) At least 3, but less than 4

(E) At least 4

Solution. Inputting the data $\{(x_i, y_i)\}_{i=1}^{5}$ into a financial calculator, we find $\hat{\beta}_0 = 18.9838$ and $\hat{\beta}_1 = -0.3003$. The fitted value when $x = 12$ is $\hat{y} = \hat{\beta}_0 + 12\hat{\beta}_1 = 15.3802$. Then the absolute value of the residual when $x = 12$ and $y = 18$ is

$$|y - \hat{y}| = |18 - 15.3802| = \boxed{2.6198}. \qquad \textbf{(Answer: (C))}$$

□

Remark. Strictly speaking, residuals are not defined for an arbitrary (x, y); they are defined for a particular observation.

Sum-to-zero constraints on residuals. Provided that 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 will offset one another to produce a zero sum. More formally speaking, the residuals are negatively correlated.

2. $\sum_{i=1}^{n} x_i e_i = 0$.

These two facts can be easily shown by realizing that $\hat{\beta}_0$ and $\hat{\beta}_1$ 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 1.2.6. (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,

$$\begin{aligned}
\sum_{i=1}^{3} e_i &= \left[2 - \left(\frac{5}{7} + \hat{\beta}_1\right)\right] + \left[y_2 - \left(\frac{5}{7} + 3\hat{\beta}_1\right)\right] + \left[5 - \left(\frac{5}{7} + 4\hat{\beta}_1\right)\right] = 0 \\
7y_2 - 56\hat{\beta}_1 &= -34. \qquad (1.2.4)
\end{aligned}$$

Second,

$$\begin{aligned}\sum_{i=1}^{3} x_i e_i &= \left[2-\left(\frac{5}{7}+\hat{\beta}_1\right)\right]+3\left[y_2-\left(\frac{5}{7}+3\hat{\beta}_1\right)\right]+4\left[5-\left(\frac{5}{7}+4\hat{\beta}_1\right)\right]=0\\ 21y_2-182\hat{\beta}_1 &= -114. \qquad (1.2.5)\end{aligned}$$

Solving (1.2.4) and (1.2.5) 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]

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

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 "Regression SS" (sometimes "Regress SS"). You should be

[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 (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).

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))** □

Example 1.3.2. (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. 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

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

Remark. The fact that $R^2 = 0.8$ means that 80% of the variation of the response about the sample mean can be explained by the fitted SLR model.

□

Specialized formulas for RSS and Reg SS 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 as well. The formula for Reg SS is presented in Exercise 2.13 (b) of Frees and, as a result, exam questions may be set on the formula. If you are aiming for Grade 10 in Exam SRM, you should not hesitate to memorize it!

The formula has two consequences:

1. The regression sum of squares is directly proportional to the square of the least squares estimator of the slope parameter. Other things equal, the higher the value of $\hat{\beta}_1$ in absolute value, the better the goodness of fit is, as expected.

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

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

   This fact, mentioned on page 39 and followed up in Exercise 2.13 (c) of Frees, comes in useful occasionally.

Note that these specialized formula for RSS and Reg SS apply only to SLR.

Example 1.3.3. (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

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

□

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 in 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 given raw data) 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 $\bar{y} = 3$,

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

It follows that

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

□

Remark. In theory, we can calculate RSS ourselves by computing the least squares estimates of β_0, β_1, the fitted values $\hat{y}_i$'s and finally

$$\text{RSS} = \sum_{i=1}^{5}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{5}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

This question shows mercy by making this rather large amount of calculations unnecessary.

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 variation in the response y. The formal hypotheses are

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

[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.

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 (difficult to justify completely rigorously!):

- 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. 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 what you learned in 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. They 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 an F-test 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 F-statistic by TSS, yielding

$$\boxed{F = (n-2)\left(\frac{\text{Reg SS/TSS}}{\text{RSS/TSS}}\right) = (n-2)\left(\frac{R^2}{1-R^2}\right).} \tag{1.3.4}$$

Then the critical region of the F-test of size α, in terms of R^2, is

$$(n-2)\left(\frac{R^2}{1-R^2}\right) > F_{\alpha,1,n-2} \quad \Leftrightarrow \quad R^2 > \underbrace{\frac{F_{\alpha,1,n-2}}{n-2+F_{\alpha,1,n-2}}}_{\substack{\text{critical value of } R^2 \text{ if used as the} \\ \text{test statistic to conduct } F\text{-test}}} .$$

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

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

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

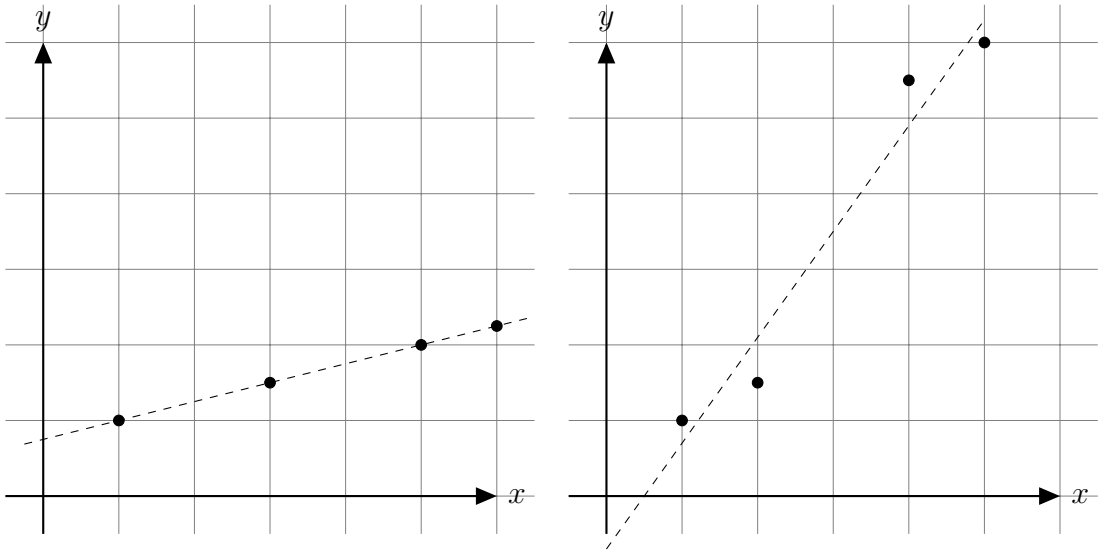
(E) At least 39

Solution. Given the value of R^2, the value of the F statistic, by (1.3.4), 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!] (Subtle differences between $\hat{\beta}_1$, F-statistics, and R^2) A simple linear regression (SLR) model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with independent normal errors is fit to each of following two sets of data.



Dataset 1 Dataset 2

The dotted lines represent the respective fitted regression lines. The two scatter plots are on the same scale.

Without any calculation, compare the two SLR models with respect to the magnitude of:

(a) $\hat{\beta}_1$, the least squares estimate of β_1

(b) The coefficient of determination R^2

(c) The F-statistic for testing $\mathrm{H}_0 : \beta_1 = 0$

Comments: This tricky example demonstrates the somewhat surprising fact that the strength of the linear relationship between x and y and the quality of the regression fit are subtly different concepts.

Solution. (a) Note that $\hat{\beta}_1$ is the slope of the fitted regression line. From the two scatter plots, Dataset 2 has a higher slope and therefore a higher value of $\hat{\beta}_1$ than Dataset 1.

(b) Observe that Dataset 2 has a very good fit to the observed data, but Dataset 1 has even a perfect fit—the fitted regression line passes through all four points. Thus $\mathrm{RSS}_2 > \mathrm{RSS}_1 = 0$ and

$$R_1^2 = 1 - \frac{\mathrm{RSS}_1}{\mathrm{TSS}_1} = 1 < 1 - \frac{\mathrm{RSS}_2}{\mathrm{TSS}_2} = R_2^2.$$

(c) By (1.3.4), $F = (n-2) \times R^2/(1-R^2)$, which increases with R^2. As Dataset 1 has a higher R^2, it also has a higher F-statistic. In fact, the F-statistic of Dataset 1 is infinite.

In conclusion:

		Dataset 1	Dataset 2
(a)	$\hat{\beta}_1$	Smaller	Bigger
(b)	R^2	1	Close to 1
(c)	F	$+\infty$	Very positive

□

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

On pages 35 and 36 of Frees, you can find straightforward algebraic proofs of the distributional properties of the two LSEs. These proofs are somewhat tedious, uninspiring, and most importantly, are confined to the SLR framework. Instead of following the algebraic manipulations there, we find it much more instructive and efficient to employ matrix algebra to derive the desired results, which will be performed . For the time being, we simply state that $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed with means and variances described as follows:

- *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

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

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:

$$\widehat{\text{Var}}(\hat{\beta}_0) = s^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right) = \frac{s^2\sum_{i=1}^n x_i^2}{nS_{xx}} \quad \text{and} \quad \widehat{\text{Var}}(\hat{\beta}_1) = \frac{s^2}{S_{xx}}. \tag{1.4.1}$$

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*:

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

These are measures of the reliability, or precision, of the LSEs. Observe from (1.4.2) 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) and (1.4.2) can be derived as special cases of general results in linear regression models, it is suggested that you memorize (1.4.1) and (1.4.2) as deriving them from first principles takes considerable time.

Example 1.4.1. (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 (1.4.2),

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

□

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 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 (1.4.2). 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

[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 $\hat{\beta}_j$.

[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.

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 $H_0: \beta_j = d$ against various alternative hypotheses by means of a t-test.

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 $H_a: \beta_j \neq d$ (extremely big or extremely small values are against H_0 in favor of H_a) and $H_a: \beta_j < d$ (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:

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 1.4.2. (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

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

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 $t(\hat{\beta}_j) \overset{\mathrm{H}_0}{\sim} t_{n-2}$, 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

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

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

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

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 (1.4.2). 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 1.4.3. (SOA Course 4 Fall 2002 Question 38: Confidence interval for β_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

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 (1.4.1), 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

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

□

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

Example 1.4.4. (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 (1.4.1), 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

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

□

Example 1.4.5. (CAS Exam ST Fall 2014 Question 20: Confidence interval for β_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

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

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 F-test and 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?

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

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

	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., $H_a : \beta_1 \neq 0$.	The alternative hypothesis can be two-sided or one-sided, e.g, $H_a : \beta_1 > 0$.

Table 1.5: Differences between the F-test and t-test for testing $H_0 : \beta_1 = 0$.

where the third equality follows from (1.3.3). Together with the distributional equality $t_v^2 = F_{1,v}$ for any $v \geq 0$, the t-test and F-test indeed have the same rejection region and are equivalent ways of testing $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$.

Table 1.5 summarizes the differences between the F-test and t-test for testing $H_0 : \beta_1 = 0$.

Example 1.4.6. (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 $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$.

(A) 0.4

(B) 1.2

(C) 2.2

(D) 2.6

(E) 6.7

Solution. The value of the F-statistic for testing $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ 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 $\hat{\beta}_1 = 1.2 > 0$), 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	x_1
	y_2	x_2
	$\vdots$	$\vdots$
	y_n	x_n
Unobserved (future) data	y_* (target)	$\leftarrow$ 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 same (unknown) parameters in the 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

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

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

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

and upon studentization, we have

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

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

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

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

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

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 $\widehat{\mathrm{Var}}(\hat{y}_*) = \widehat{\mathrm{Var}}(\hat{\beta}_0 + \hat{\beta}_1 x_*)$. 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.

[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 y_* itself is also random.

[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 $\mathrm{SE}(y_* - \hat{y}_*)$ by $\mathrm{SE}(\hat{y}_*)$, treating as if your target y_* has no variability.

2. *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 $\textcircled{1} = \widehat{\text{Var}}(\varepsilon_*)$. 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 1.5.1. (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

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

□

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) 1.9

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.

1.6 Problems

Model fitting by least squares method

Problem 1.6.1. (SOA Course 120 May 1990 Question 11: Different forms of $\hat{\beta}_1$) Which of the following are valid expressions for $\hat{\beta}_1$, the slope coefficient in the simple linear regression of y on x?

I. $\dfrac{\sum x_i y_i - \bar{y}\sum x_i}{\sum x_i^2 - \bar{x}\sum x_i}$

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

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

(A) I and II only

(B) I and III only

(C) II and III only

(D) I, II and III

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

Solution. The first formula is equivalent to

$$\text{(I)} = \frac{\sum x_i y_i - \bar{y}\sum x_i}{\sum x_i^2 - \bar{x}\sum x_i} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}} = \hat{\beta}_1.$$

The third formula is also correct, because

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

Only the second formula is incorrect, as $\bar{x}^2$ in the denominator should be multiplied by n. **(Answer: (B))** □

Problem 1.6.2. (SOA Course 120 May 1988 Question 5: Calculation of LSE given raw data) You are given the following data about a simple regression of the form $y = \beta_0 + \beta_1 x + \varepsilon$, where ε is a random error term with mean 0 and variance σ^2:

x	y
1.0	10.7
2.0	4.9
3.0	9.3
4.0	14.6
5.0	15.7
6.0	11.8

Determine the least squares estimate for the constant term β_0.

(A) -3

(B) 1

(C) 7

(D) 11

(E) 14

Solution. Direct use of a financial calculator shows that $\hat{\beta}_0 = \boxed{6.8467}$. **(Answer: (C))** □

Problem 1.6.3. (SOA Course 120 Study Note 120-81-95 Question 10: Calculation of LSE given raw data for an SLR model in disguise) You use the method of least squares to fit the model $y_i = \beta_0 + \beta_1(x_i + x_i^2) + \varepsilon_i$ to the following data:

x_i	0	0	1	1	2	2
y_i	-3	-2	0	1	8	10

Determine $\hat{\beta}_0$.

(A) -2.9

(B) -2.4

(C) -1.9

(D) -1.4

(E) -0.9

Solution. The given model can be viewed as an SLR model with explanatory variable $x + x^2$:

$x_i + x_i^2$	0	0	2	2	6	6
y_i	-3	-2	0	1	8	10

Inputting these six pairs of data into a financial calculator yields $\hat{\beta}_0 = \boxed{-20/7}$ (and $\hat{\beta}_1 = 109/56$). **(Answer: (A))** □

Problem 1.6.4. (SOA Part 3 November 1984 Question 5: Calculation of LSE given summarized data – I) You are given 10 pairs of values (x_i, y_i) which will 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:

$$\begin{aligned}\sum_{i=1}^{10} x_i &= 100\\ \sum_{i=1}^{10} y_i &= 200\\ \sum_{i=1}^{10} x_i y_i &= 2000\\ \sum_{i=1}^{10} x_i^2 &= 2000\\ \sum_{i=1}^{10} y_i^2 &= 5000\end{aligned}$$

Calculate the least squares estimate of β_1.

(Answer to nearest 0.1)

(A) 0.0

(B) 0.1

(C) 0.2

(D) 0.3

(E) 0.4

Solution. By (1.2.3),

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{2,000 - 10(10)(20)}{2,000 - 10(10)^2} = \boxed{0}. \quad \textbf{(Answer: (A))}$$

□

Problem 1.6.5. (CAS Exam 3L Spring 2005 Question 27: Calculation of LSE given summarized data – II) Given the following information:

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

Determine the least squares equation for the following model:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon.$$

(A) $\hat{y}_i = -0.73 + 12.16x_i$

(B) $\hat{y}_i = -8.81 + 12.16x_i$

(C) $\hat{y}_i = 283.87 + 10.13x_i$

(D) $\hat{y}_i = 10.13 + 12.16x_i$

(E) $\hat{y}_i = 23.66 + 10.13x_i$

Solution. By (1.2.3),

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{26,696 - 12(144/12)(1,742/12)}{2,300 - 12(144/12)^2} = 10.1259$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1,742}{12} - 10.1259\left(\frac{144}{12}\right) = 23.6559.$$

Therefore, the fitted regression line is

$$\boxed{\hat{y}_i = 23.6559 + 10.1259x_i.} \quad \textbf{(Answer: (E))}$$

□

Problem 1.6.6. (LSE for naive i.i.d. model) Consider the linear model

$$y_i = \beta + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2), \quad i = 1, 2, \ldots, n,$$

where β and σ^2 are unknown parameters.

Determine the least squares estimator of β.

(A) 0

(B) $\bar{y}$

(C) $\bar{y} - \hat{\beta}_1 \bar{x}$

(D) $\sum_{i=1}^n (y_i - \bar{y})^2/(n-2)$

(E) $\sum_{i=1}^n (y_i - \bar{y})^2/(n-1)$

Comments: This problem confirms that the residual sum of squares of the i.i.d. model is the total sum of squares in the ANOVA table.

Solution. Differentiating $\sum_{i=1}^{n}(y_i-\beta)^2$ with respect to β, we consider

$$\frac{\partial}{\partial\beta}\sum_{i=1}^{n}(y_i-\beta)^2=-2\sum_{i=1}^{n}(y_i-\beta)\begin{cases}<0, & \text{if } \beta<\bar{y},\\ =0, & \text{if } \beta=\bar{y},\\ >0, & \text{if } \beta>\bar{y}.\end{cases}$$

Thus $\sum_{i=1}^{n}(y_i-\beta)^2$ is minimized at

$$\hat{\beta}=\frac{\sum_{i=1}^{n}y_i}{n}=\boxed{\bar{y}}. \qquad \textbf{(Answer: (B))}$$

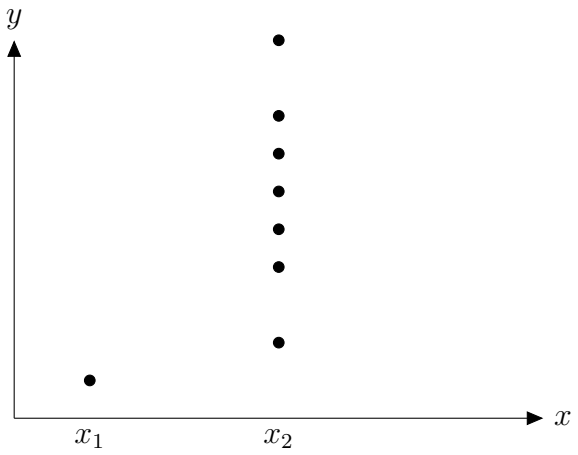
□

Remark. The following spurious solution is unacceptable:

Since $\hat{\beta}=\bar{y}-\hat{\beta}_1 x$ and $\beta_1=0$ in the model, we have $\hat{\beta}=\bar{y}-(0)\bar{x}=\bar{y}$.

What are your $\hat{\beta}_1$ and x here?

Problem 1.6.7. [HARDER!] (SOA Course 120 May 1990 Question 7: Determining the fitted regression line from a scatter plot) Eight pairs of observations (x_i, y_i) are represented by the following scatter diagram:



The least squares line passes through the point $\left(\frac{1}{2}(x_1+x_2), k\right)$.
Determine k.

(A) $\frac{3}{8}y_1+\frac{5}{8}\bar{y}$

(B) $\frac{3}{7}y_1+\frac{4}{7}\bar{y}$

(C) $\frac{1}{2}y_1+\frac{1}{2}\bar{y}$

(D) $\frac{4}{7}y_1+\frac{3}{7}\bar{y}$

(E) $\frac{5}{8}y_1+\frac{3}{8}\bar{y}$

Comments: You can work out this problem by inspection without using any formula. Think about how the straight line should be positioned to minimize the sum of squares of residuals. The result of Problem 1.6.6 may also help.

Solution. Because of the special form of the data (only two distinct x-values), the fitted regression line, which is designed to minimize the sum of squared errors, must pass through (x_1, y_1), so that the residual at x_1 is exactly zero. It remains to determine the value of the fitted regression line at x_2. Because the seven y-values, denoted by $y_2, y_3, \ldots, y_8$, are all situated at x_2, they constitute an i.i.d. sample. By virtue of Problem 1.6.6, the fitted y-value at x_2 should be the sample mean of $y_2, y_3, \ldots, y_8$, or $(y_2+\cdots+y_8)/7$. Symbolically, we have

$$\hat{y}_1 = y_1 \quad \text{and} \quad \hat{y}_2 = \hat{y}_3 = \cdots = \hat{y}_8 = \frac{y_2 + \cdots + y_8}{7}.$$

The value on the fitted regression line at $(x_1 + x_2)/2$ is

$$\begin{aligned} k &= \frac{y_1 + (y_2 + \cdots + y_8)/7}{2} \\ &= \frac{y_1}{2} + \frac{1}{14}\left(\sum_{i=1}^{8} y_i - y_1\right) \\ &= \frac{y_1}{2} + \frac{1}{14}(8\bar{y} - y_1) \\ &= \boxed{\frac{3}{7}y_1 + \frac{4}{7}\bar{y}}. \qquad \textbf{(Answer: (B))} \end{aligned}$$

□

Remark. There is a (lesser-known) fact in the SLR framework: $(\bar{x}, \bar{y})$ must be a point on the fitted regression line. This is because

$$\hat{\beta}_0 + \hat{\beta}_1\bar{x} = (\bar{y} - \hat{\beta}_1\bar{x}) + \hat{\beta}_1\bar{x} = \bar{y}.$$

Because $(x_1 + x_2)/2$ is the sample mean of the two x-values, the corresponding point on the fitted regression line, in view of the fact above, should be $\bar{y}$. But we have found that $k = 3y_1/7 + 4\bar{y}/7$, not $\bar{y}$. Is there any contradiction? Why or why not?

Problem 1.6.8. (SOA Part 3 November 1984 Question 10: LSE for a no-intercept regression model) You are given the following model:

$$y = \beta x^2 + \varepsilon$$

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

Determine the least squares estimator of β.

(A) $\dfrac{\sum x_i y_i}{\sum x_i^2}$

(B) $\dfrac{\sum x_i^2 y_i}{\sum x_i^4}$

(C) $\dfrac{\sum (x_i - \bar{x})^2 y_i^2}{\sum (x_i - \bar{x})^2}$

(D) $\dfrac{\sum (x_i - \bar{x})^2 y_i}{\sum (x_i - \bar{x})^4}$

(E) $\dfrac{\sum (x_i - \bar{x})^2 y_i^2}{\sum (x_i - \bar{x})^4}$

Solution. We minimize $\sum (y_i - \beta x_i)^2$ by differentiating it with respect to β, giving

$$2\sum(-x_i^2)(y_i - \beta x_i^2) = 0 \quad \Rightarrow \quad \hat{\beta} = \boxed{\frac{\sum x_i^2 y_i}{\sum x_i^4}}. \qquad \textbf{(Answer: (B))}$$

□

Remark. Regression models without the intercept, sometimes known as regression through the origin, are presented in Exercise 2.8 of Frees. In general, the LSE of β in the no-intercept SLR model $y = \beta x + \varepsilon$ is

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}.$$

Replacing x_i by x_i^2 yields the answer above.

Problem 1.6.9. (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

Comments: Before computing the answer, we derive an expression for $\hat{\beta}_1$ in terms of r, the sample correlation between x and y. By (1.2.2),

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \times \sqrt{\frac{S_{yy}}{S_{xx}}} = r \times \sqrt{\frac{S_{yy}/(n-1)}{S_{xx}/(n-1)}} = \boxed{r \times \frac{s_y}{s_x}}, \tag{1.6.1}$$

where $s_x = \sqrt{S_{xx}/(n-1)}$ and $s_y = \sqrt{S_{yy}/(n-1)}$ (in lower case letters) are the sample standard deviations of x and y, respectively. This formula for $\hat{\beta}_1$ is not as common as (1.2.2), but is stated on page 28 of Frees.

Solution. Using the formula above, we have

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

□

Problem 1.6.10. (SOA Course 120 Study Note 120-82-97 Question 7: Calculation of fitted value given r) You are given the following information about a simple regression model fit to 10 observations:

$$\begin{aligned} \sum_{i=1}^{10} x_i &= 20 \\ \sum_{i=1}^{10} y_i &= 100 \\ s_x &= 2 \\ s_y &= 8 \end{aligned}$$

You are also given that the sample correlation coefficient $r = -0.98$.

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

(A) -10

(B) -2

(C) 11

(D) 30

(E) 37

Solution. By (1.2.2) and (1.6.1),

$$\hat{\beta}_1 = r \times \frac{s_y}{s_x} = (-0.98) \times \frac{8}{2} = \boxed{-3.92}$$

and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 10 - (-3.92)(2) = 17.84$. Then the predicted value of y when $x = 5$ is

$$\hat{y} = \hat{\beta}_0 + 5\hat{\beta}_1 = 17.84 + 5(-3.92) = \boxed{-1.76}. \quad \textbf{(Answer: (B))}$$

□

Problem 1.6.11. [HARDER!] (SOA Course 120 November 1986 Question 6 Regressing y on x vs regressing x on y) A set of observations (x_i, y_i) has been represented by the following model:

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

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

You are given:

$$\begin{aligned} \bar{x} &= 4.00 \\ \text{Sample standard deviation of } x &= 2.00 \\ \text{Sample standard deviation of } y &= 4.00 \\ \hat{\beta}_0 &= 6.00 \\ \hat{\beta}_1 &= 1.50 \end{aligned}$$

You believe that a better representation of the observed data could be obtained by regressing x on y, using the following model:

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

where ε' is a random error with mean 0 and variance $(\sigma')^2$.

Calculate $\hat{\beta}_1'$.

(A) 0.25

(B) 0.38

(C) 0.67

(D) 0.75

(E) 1.50

(Lesson from this problem: Never mix up x and y when you perform real data analysis!!)

Solution. By (1.2.2), we have

$$\frac{S_{xy}}{S_{xx}} = 1.5 \quad \Rightarrow \quad S_{xy} = 1.5 S_{xx}.$$

From the sample standard deviations of x and y, we further have

$$\begin{cases} S_{xx}/(n-1) = 2^2 \\ S_{yy}/(n-1) = 4^2 \end{cases} \quad \Rightarrow \quad \frac{S_{yy}}{S_{xx}} = 4.$$

It follows that

$$\hat{\beta}_1' = \frac{S_{xy}}{S_{yy}} = \frac{1.5 S_{xx}}{S_{yy}} = \frac{1.5}{4} = \boxed{0.375}. \quad \textbf{(Answer: (B))}$$

□

Remark. Let $\hat{\beta}_1^{x\sim y}$ and $\hat{\beta}_1^{x\sim y}$ be the LSEs of β_1 when x is regressed on y and y is regressed on x, respectively. By (1.6.1),

$$\hat{\beta}_1' \times \hat{\beta}_1 = \left(r \times \frac{s_y}{s_x} \right) \left(r \times \frac{s_x}{s_y} \right) = r^2 = R^2,$$

which is Exercise 2.6 of Frees. Thus in general, $\hat{\beta}_1^{x\sim y} \neq 1/\hat{\beta}_1^{x\sim y}$, unless $R^2 = 1$.

Problem 1.6.12. (SOA Course 120 November 1990 Question 8: Given the LSEs, deduce the observation) Your assistant had estimated the parameters of a two-variable regression of the form

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

as follows:

$$\hat{\beta}_0 = 3.0, \qquad \hat{\beta}_1 = 2.0$$

The data had been partially obliterated:

Obs.	y	x
1	2.7	0.0
2	7.3	2.0
3	11.5	4.0
4	14.8	6.0
5		8.0
6		10.0

Calculate y_5.

(A) 17.7

(B) 18.2

(C) 18.7

(D) 19.2

(E) 19.7

Solution. The two sum-to-zero conditions

$$\sum_{i=1}^{6} e_i = [2.7 + 7.3 + 11.5 + 14.8 + (y_5 + y_6)] - [3(6) + 2(2 + 4 + 6 + 8 + 10)] = 0$$

and

$$\sum_{i=1}^{6} x_i e_i = 2(7.3 - 7) + 4(11.5 - 11) + 6(14.8 - 15) + 8(y_5 - 19) + 10(y_6 - 23) = 0$$

result in two linear equations in y_5 and y_6:

$$\begin{cases} y_5 + y_6 = 41.7 \\ 8y_5 + 10y_6 = 380.6 \end{cases}.$$

This can be solved to yield $y_5 = \boxed{18.2}$ and $y_6 = 23.5$. **(Answer: (B))** □

Assessing the goodness of fit

Problem 1.6.13. (SOA Course 120 November 1988 Question 6: Given R^2) You are given the following information concerning a simple linear regression:

$$\begin{aligned} \sum_{i=1}^{8} (y_i - \bar{y})^2 &= 435.0 \\ R^2 &= 0.65 \end{aligned}$$

Calculate s^2.

(A) 19.0

(B) 21.8

(C) 25.4

(D) 40.4

(E) 47.1

Solution. The residual sum of squares is

$$\text{RSS} = (1 - R^2)\text{TSS} = (1 - 0.65)(435) = 152.25.$$

Then the MSE is

$$s^2 = \frac{\text{RSS}}{n-2} = \frac{152.25}{8-2} = \boxed{25.375}. \qquad \textbf{(Answer: (C))}$$

□

Problem 1.6.14. (SOA Course 120 November 1989 Question 8: Calculation of R^2 given raw data) You are given:

x	y
6.8	0.8
7.0	1.2
7.1	0.9
7.2	0.9
7.4	1.5

Determine R^2, the coefficient of determination for the regression of y on x.

(A) 0.3

(B) 0.4

(C) 0.5

(D) 0.6

(E) 0.7

Solution. The easiest way to calculate R^2 is to input the five pairs of data into a financial calculator and square the sample correlation between x and y. Then we get $R^2 = r^2 = 0.698535^2 = \boxed{0.4880}$. **(Answer: (C))** □

Problem 1.6.15. (CAS Exam S Fall 2017 Question 34: Calculation of R^2 given summarized data – I) For an ordinary linear regression with 5 parameters and 50 observations, you are given:

- The total sum of squares, $S_{yy} = 996$.
- The unbiased estimate for the constant variance, σ^2, is $s^2 = 2.47$.

Calculate the coefficient of determination.

(A) Less than 0.65

(B) At least 0.65, but less than 0.75

(C) At least 0.75, but less than 0.85

(D) At least 0.85, but less than 0.95

(E) At least 0.95

Solution. From the third itemized point, we have

$$s^2 = \frac{\text{RSS}}{n-k-1} \Rightarrow \text{RSS} = (n-k-1)s^2 = (50 - \underbrace{5}_{\text{note!}})(2.47) = 111.15.$$

Then

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{111.15}{996} = \boxed{0.8884}. \quad \textbf{(Answer: (D))}$$

□

Remark. Even if you (incorrectly) take $k = 5$, you will get $R^2 = 0.8909$ and Answer (D).

Problem 1.6.16. (SOA Course 120 May 1989 Question 7: Calculation of R^2 given summarized data – II) You are interested in the relationship between the price movements of XYZ Corporation and the "market" during the fourth quarter of 1987.

You have used the least squares criterion to fit the following line to 14 weekly closing values of XYZ stock (y_i) and the Dow Jones Industrial Average (x_i) during the period of interest:

$$\hat{y}_i = -116.607 + 0.195x_i.$$

You are given:

$$\begin{aligned} \sum_{i=1}^{14}(x_i - \bar{x})^2 &= 17,807.080 \\ \sum_{i=1}^{14}(y_i - \bar{y})^2 &= 949.388 \end{aligned}$$

Determine the percentage of variation in the value of XYZ stock that was "explained" by variations of the Dow.

(A) 50

(B) 60

(C) 70

(D) 80

(E) 90

Solution. The regression sum of squares is

$$\text{Reg SS} = \hat{\beta}_1^2 S_{xx} = 0.195^2 \times 17,807.080 = 677.114217.$$

Then the coefficient of determination is

$$R^2 = \frac{\text{Reg SS}}{\text{TSS}} = \frac{677.114217}{949.388} = \boxed{0.7132}. \quad \textbf{(Answer: (C))}$$

□

Problem 1.6.17. (SOA Course 120 May 1991 Question 6: Calculation of R^2 as a correlation) A bank is examining the relationship between income (x) and savings (y). A survey of six randomly selected depositors yielded the following sample means, sample variances and sample covariance:

$$\begin{aligned} \bar{x} &= 27.5 \\ \bar{y} &= 3.0 \\ s_x^2 &= 87.5 \\ s_y^2 &= 3.5 \\ s_{xy} &= 17.0 \end{aligned}$$

Determine the proportion of the variation in y about $\bar{y}$ that is explained by the linear relationship of y with x.

(A) 0.1

(B) 0.2

(C) 0.3

(D) 0.7

(E) 0.9

Solution. This problem uses an indirect way to ask for R^2, which is

$$R^2 = r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{17^2}{87.5(3.5)} = \boxed{0.9437}. \quad \textbf{(Answer: (E))}$$

□

Problem 1.6.18. (SOA Course 120 May 1990 Question 14: Basic facts about R^2) Which of the following statements are true about simple linear regression?

I. The coefficient of determination is the fraction of the variation in y about $\bar{y}$ that is explained by the linear relationship of y with x.

II. The coefficient of determination is the ratio of the regression sum of squares to the total sum of squares.

III. The residual mean square provides an estimate of the variance of y for a given x based on $(n-1)$ degrees of freedom.

(A) I and II only

(B) I and III only

(C) II and III only

(D) I, II and III

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

Solution. I and II are correct by definition, while III is not. III would be correct if $n-1$ is replaced by $n-2$. **(Answer: (A))** □

Problem 1.6.19. (SOA Course 120 Study Note 120-82-97 Question 3: Calculation of R^2 given residuals) You fit a simple linear regression model to five pairs of observations. The residuals for the first four observations are $0.4, -0.3, 0.0, .-0.7$, and the estimated variance of the dependent variable y is $\widehat{\text{Var}}(y) = 1.5$.

Calculate the coefficient of determination, R^2.

(A) 0.82

(B) 0.84

(C) 0.86

(D) 0.88

(E) 0.90

Solution. Due to the sum-to-zero constraint, the firth residual is $e_5 = -\sum_{i=1}^{4} e_i = 0.6$. The residual sum of squares is

$$\text{RSS} = \sum_{i=1}^{5} e_i^2 = 0.4^2 + (-0.3)^2 + 0^2 + (-0.7)^2 + 0.6^2 = 1.1.$$

Considering the estimated variance of y, we have

$$\frac{\sum_{i=1}^{5}(y_i - \bar{y})^2}{4} = 1.5 \quad \Rightarrow \quad \text{TSS} = \sum_{i=1}^{5}(y_i - \bar{y})^2 = 6.$$

The coefficient of determination R^2 is

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{1.1}{6} = \boxed{0.8167}. \qquad \textbf{(Answer: (A))}$$

□

Problem 1.6.20. (SOA Course 120 November 1985 Question 10: Going from R^2 and TSS to MSE) You are representing ten 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:

$$\begin{aligned} R^2 &= 0.6 \\ \sum_{i=1}^{10} y_i &= 30 \\ \sum_{i=1}^{10} y_i^2 &= 690 \end{aligned}$$

Calculate s^2.

(Answer to nearest integer)

(A) 27

(B) 30

(C) 36

(D) 45

(E) 50

Solution. As $\text{TSS} = \sum_{i=1}^{10} y_i^2 - 10\bar{y}^2 = 690 - 10(3)^2 = 600$,

$$\text{RSS} = (1 - R^2)(\text{TSS}) = (1 - 0.6)(600) = 240.$$

Hence

$$s^2 = \frac{\text{RSS}}{n-2} = \frac{240}{10-2} = \boxed{30}. \quad \textbf{(Answer: (B))}$$

□

Problem 1.6.21. (SOA Course 120 Study Note 120-82-94 Question 8: Calculation of F-statistic) Twenty pairs of observations (x_i, y_i) are represented by the following model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, 20$$

where the ε_i are independent, normally distributed random variables with mean 0 and variance σ^2.

You have determined:

$$\begin{aligned} \sum x_i &= 100 \\ \sum y_i &= 120 \\ \sum x_i^2 &= 2100 \\ \sum y_i^2 &= 7120 \\ \sum x_i y_i &= 2520 \\ \hat{\beta}_1 &= 1.2 \end{aligned}$$

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

(A) 2

(B) 4

(C) 6

(D) 8

(E) 10

Solution. Since $S_{xx} = \sum x_i^2 - 20\bar{x}^2 = 2100 - 20(5)^2 = 1600$, we have Reg SS $= \hat{\beta}_1^2 S_{xx} = (1.2)^2(1600) = 2304$. Moreover, TSS $= \sum y_i^2 - 20\bar{y}^2 = 7120 - 20(6)^2 = 6400$, so RSS $=$ TSS $- \hat{\beta}_1^2 S_{xx} = 6400 - (1.2)^2(1600) = 4096$. The F-statistic for testing $\mathrm{H}_0 : \beta_1 = 0$ is

$$F = \frac{\text{Reg SS}/1}{\text{RSS}/(n-2)} = \frac{2304/1}{4096/(20-2)} = \boxed{10.125}. \quad \textbf{(Answer: (E))}$$

□

Problem 1.6.22. (CAS Exam ST Spring 2016 Question 21: Going from R^2 to F-statistic) You are given the following linear regression model which is fitted to 11 observations:

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

The coefficient of determination is $R^2 = 0.25$.

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

(A) Less than 1.5

(B) At least 1.5, but less than 2.5

(C) At least 2.5, but less than 3.5

(D) At least 3.5, but less than 4.5

(E) At least 4.5

Solution. Given the value of R^2, the value of the F statistic, by (1.3.4), is

$$F = (n-2)\left(\frac{R^2}{1-R^2}\right) = (11-2)\left(\frac{0.25}{1-0.25}\right) = \boxed{3}. \quad \textbf{(Answer: (C))}$$

□

Problem 1.6.23. (SOA Course 120 May 1986 Question 10: Going from F-statistic to R^2) You are given 20 pairs of values (x_i, y_i) which are 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 that the F ratio for testing the hypothesis $\mathrm{H}_0 : \beta_1 = 0$ is equal to 12.

Calculate R^2.

(A) 0.3

(B) 0.4

(C) 0.5

(D) 0.6

(E) 0.7

Solution. Considering

$$F = (n-2)\left(\frac{R^2}{1-R^2}\right) = (20-2)\left(\frac{1}{1-R^2} - 1\right) = 12,$$

we get $R^2 = \boxed{0.4}$. **(Answer: (B))** □

Statistical inference about regression coefficients

Problem 1.6.24. (SOA Course 120 May 1986 Question 5: Simple true-or-false statements about variances) Which of the following statements about simple regression models are true?

I. The variance of $\hat{\beta}_0$ is a function of the number of observations.

II. The variance of $\hat{\beta}_1$ can be reduced by using a wider range of the explanatory variable.

III. The variance associated with the predicted values of individual observations is smallest for values of the explanatory variable closest to the mean.

(A) I and II only

(B) I and III only

(C) II and III only

(D) I, II and III

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

Solution. All three statements are correct. For II, note that $\text{Var}(\hat{\beta}_1) = \sigma^2/S_{xx}$, which decreases with S_{xx}. For III, recall that

$$\text{Var}(\text{prediction error}) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}\right],$$

which is smallest when x_* is closest to $\bar{x}$. **(Answer: (D))** □

Problem 1.6.25. (SOA Course 120 Study Note 120-81-95 Question 1: $\widehat{\text{Var}}(\hat{\beta}_1)$) You fit a least squares regression line to five pairs of observations (x_i, y_i) using the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

You determine:

$$\begin{aligned} \sum x_i &= 10 \\ \sum x_i^2 &= 30 \\ \sum (y_i - \hat{y}_i)^2 &= 15 \end{aligned}$$

Determine the estimated variance of $\hat{\beta}_1$.

(A) 0.3

(B) 0.5

(C) 1.0

(D) 2.0

(E) 3.0

Solution. We need both s^2 and S_{xx}.

- The MSE can be obtained as $s^2 = \text{RSS}/(n-2) = 15/(5-2) = 5$.
- Given $\sum x_i$ and $\sum x_i^2$, we have $S_{xx} = \sum x_i^2 - n\bar{x}^2 = 30 - 5(2)^2 = 10$.

By (1.4.1), the estimated variance of $\hat{\beta}_1$ is

$$\widehat{\text{Var}}(\hat{\beta}_1) = \frac{s^2}{S_{xx}} = \frac{5}{10} = \boxed{0.5}. \qquad \textbf{(Answer: (B))}$$

□

Problem 1.6.26. [HARDER!] (SOA Course 120 November 1990 Question 12: Given $\hat{\beta}_1, r$, find SE$(\hat{\beta}_1)$) You are given a simple regression of the form:

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

Based on 18 observations of the variables x and y, you have determined that the correlation coefficient of the two variables is 0.6 and that the estimated slope coefficient $\hat{\beta}_1$ is 3.0.

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

(A) 0.1

(B) 0.4

(C) 1.0

(D) 1.2

(E) 1.5

Solution. We are given that

$$\hat{\beta}_1 = r \times \frac{s_y}{s_x} = 0.6 \times \frac{s_y}{s_x} = 3,$$

so $s_y/s_x = 5$. Since $S_{yy} = \text{TSS} = \text{RSS} + \text{Reg SS} = \text{RSS} + \hat{\beta}_1^2 S_{xx}$, we have

$$\widehat{\text{Var}}(\hat{\beta}_1) = \frac{s^2}{S_{xx}} = \frac{(S_{yy} - \hat{\beta}_1^2 S_{xx})/16}{S_{xx}} = \frac{S_{yy}/S_{xx} - \hat{\beta}_1^2}{16} = \frac{5^2 - 3^2}{16} = 1.$$

The standard error is the square root of 1, or $\boxed{1}$. **(Answer: (C))** □

Problem 1.6.27. (SOA Course 120 May 1991 Question 8: 95th percentile of $\hat{\beta}_1$) The model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where the ε_i are uncorrelated normal random variables with mean 0 and variance $\sigma^2 = 1$, is to be fitted to observations at $x = -2, -1, 0, 1$ and 2.

Using this model, determine q such that $\mathbb{P}(\hat{\beta}_1 - \beta_1 < q) = 0.95$.

(A) 0.2

(B) 0.3

(C) 0.5

(D) 0.6

(E) 0.7

Solution. Recall that $\hat{\beta}_1$ is normally distributed with mean β_1 and variance $\sigma^2/S_{xx} = 1/\sum_{i=1}^5 (x_i - \bar{x})^2 = 0.1$ (note that $\sigma^2 = 1$ is known and does not need estimation), i.e., $\hat{\beta}_1 \sim \text{N}(\beta_1, 0.1)$. Thus $q = \sqrt{0.1} \times$ 95th percentile of $\text{N}(0,1) = \sqrt{0.1} \times 1.645 = \boxed{0.5202}$. **(Answer: (C))** □

Problem 1.6.28. (SOA Course 120 May 1986 Question 6: Calculation of t-statistic) You are representing ten observations (x_i, y_i) by the following model

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

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

You have determined:

$$\begin{aligned} \hat{\beta}_1 &= 1.3 \\ \sum_{i=1}^{10} x_i &= 50 \\ \sum_{i=1}^{10} x_i^2 &= 650 \\ s^2 &= 169 \end{aligned}$$

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

(A) 2

(B) 4

(C) 7

(D) 10

(E) 13

Solution. We need the standard error of $\hat{\beta}_1$, which is

$$\mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{s^2}{S_{xx}}} = \sqrt{\frac{169}{650 - 10(5)^2}} = 0.65.$$

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

$$t(\hat{\beta}_1) = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)} = \frac{1.3}{0.65} = \boxed{2}. \quad \textbf{(Answer: (A))}$$

□

Problem 1.6.29. (CAS Exam ST Spring 2015 Question 22: Result of a two-sided t-test) You are given the following linear regression model fitted to 12 observations:

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

The results of the regression are as follows:

Parameter	Estimate	Standard Error
β_0	15.52	3.242
β_1	0.40	0.181

Determine the results of the hypothesis test $H_0 : \beta_1 = 0$ against the alternative $H_a : \beta_1 \neq 0$.

(A) Reject at $\alpha = 0.01$

(B) Reject at $\alpha = 0.02$, Do Not Reject at $\alpha = 0.01$

(C) Reject at $\alpha = 0.05$, Do Not Reject at $\alpha = 0.02$

(D) Reject at $\alpha = 0.10$, Do Not Reject at $\alpha = 0.05$

(E) Do Not Reject at $\alpha = 0.10$

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

$$t(\hat{\beta}_1) = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)} = \frac{0.40}{0.181} = 2.2099.$$

Under H_0, $t(\hat{\beta}_1)$ has a $t_{12-2} = t_{10}$ distribution. From the t-table, we find $t_{10,0.05} = 1.8125$ and $t_{10,0.025} = 2.2281$, so the p-value, which is $\mathbb{P}(|t_{10}| > 2.2099) = 2\mathbb{P}(t_{10} > 2.2099)$, lies between $2(0.025) = 0.05$ and $2(0.05) = 0.1$. Equivalently, we reject the null hypothesis in favor of the alternative when $\alpha = 0.1$, but not when $\alpha = 0.05$. **(Answer: (D))** □

Problem 1.6.30. (Result of a one-sided t-test) For a simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$, you are given:

(i) There are 20 observations.

(ii) $\hat{\beta}_1 = 0.7314$

(iii) The standard error of $\hat{\beta}_1$ is 0.1314.

You are interested in testing the null hypothesis $H_0 : \beta_1 = 0.5$ against the alternative hypothesis $H_a : \beta_1 > 0.5$.

Determine the result of the test.

(A) Reject the null hypothesis at the 1% significance level

(B) Reject the null hypothesis at the 2.5% significance level, but not the 1% significance level

(C) Reject the null hypothesis at the 5% significance level, but not the 2.5% significance level

(D) Reject the null hypothesis at the 10% significance level, but not the 5% significance level

(E) Do not reject the null hypothesis at the 10% significance level

Solution. The t-statistic for testing $\mathrm{H}_0 : \beta_1 = 0.5$ is

$$t(\hat{\beta}_1) = \frac{\hat{\beta}_1 - 0.5}{\mathrm{SE}(\hat{\beta}_1)} = \frac{0.7314 - 0.5}{0.1314} = 1.761035.$$

Under H_0, the t-statistic follows a $t_{20-2} \equiv t_{18}$ distribution. For the one-sided alternative $\mathrm{H}_a : \beta_1 > 0.5$, we will reject the null hypothesis in favor of H_a whenever the t-statistic is greater than the α-upper quantile from the t_{18} distribution. From the t-table, we find that $t_{18,0.05} = 1.7341$ and $t_{18,0.025} = 2.1009$. Because $t(\hat{\beta}_1) \in (t_{18,0.05}, t_{18,0.025})$, we conclude that H_0 is rejected when $\alpha = 0.05$ but not when $\alpha = 0.025$. **(Answer: (C))** □

Remark. Equivalently, the p-value of the one-sided test is between 0.025 and 0.05.

Problem 1.6.31. (CAS Exam ST Spring 2014 Question 20: p-value of a two-sided t-test) The model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ was fit using 6 observations. The estimated parameters are as follows:

- $\hat{\beta}_0 = 2.31$
- $\hat{\beta}_1 = 1.15$
- $\mathrm{SE}(\hat{\beta}_0) = 0.057$
- $\mathrm{SE}(\hat{\beta}_1) = 0.043$

The following hypothesis test is performed:

- $\mathrm{H}_0 : \beta_1 = 1$
- $\mathrm{H}_a : \beta_1 \neq 1$

Determine the minimum significance level at which the null hypothesis would be rejected.

(A) Less than 0.01

(B) At least 0.01, but less than 0.02

(C) At least 0.02, but less than 0.05

(D) At least 0.05, but less than 0.10

(E) At least 0.10

Solution. The t-statistic for testing $\mathrm{H}_0 : \beta_1 = 1$ is

$$t(\hat{\beta}_1) = \frac{\hat{\beta}_1 - 1}{\mathrm{SE}(\hat{\beta}_1)} = \frac{1.15 - 1}{0.043} = 3.4884.$$

From the t-table, we have $t_{4,0.025} = 2.7764$ and $t_{4,0.01} = 3.7469$. The minimum significance level at which the null hypothesis would be rejected is simply the p-value of the two-sided test, which is between $2(0.01) = 0.02$ and $2(0.025) = 0.05$. **(Answer: (C))** □

Problem 1.6.32. (SOA Course 4 Fall 2000 Question 5: Confidence interval for β_1 – I) You are investigating the relationship between per capita consumption of natural gas and the price of natural gas. You gathered data from 20 cities and constructed the following model:

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

where y is per capita consumption, x is the price, and ε is a normal random error term.

You have determined:

$$\begin{aligned} \hat{\beta}_0 &= 138.561 \\ \hat{\beta}_1 &= -1.104 \\ \sum x_i^2 &= 90,048 \\ \sum y_i^2 &= 116,058 \\ \sum (x_i - \bar{x})^2 &= 10,668 \\ \sum (y_i - \bar{y})^2 &= 20,838 \\ \sum (y_i - \hat{y}_i)^2 &= 7,832 \end{aligned}$$

Determine the shortest 95% confidence interval for β_1.

(A) $(-2.1, -0.1)$

(B) $(-1.9, -0.3)$

(C) $(-1.7, -0.5)$

(D) $(-1.5, -0.7)$

(E) $(-1.3, -0.9)$

Solution. The MSE is $s^2 = 7,832/(20-2) = 435.1111$. The shortest 95% confidence interval for β_1 is

$$\begin{aligned} \hat{\beta}_1 \pm t_{18,0.025}\sqrt{\frac{s^2}{S_{xx}}} &= -1.104 \pm 2.101\sqrt{\frac{435.1111}{10,668}} \\ &= \boxed{(-1.5283, -0.6797)}. \quad \textbf{(Answer: (D))} \end{aligned}$$

□

Remark. More than enough information is supplied in this question, possibly to confuse you.

Problem 1.6.33. (SOA Course 120 November 1989 Question 4: Confidence interval for β_1 – II) You are studying the average return on sales as a function of the

number of firms in an industry. You have collected data for 1969-1988 and performed a simple linear regression of the form:

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

The data are displayed in the following table:

Observation	Year	Number of Firms (x_i)	Return on Sales (y_i)
1	1969	138	0.167
2	1970	130	.155
3	1971	134	.211
4	1972	152	.166
5	1973	180	.121
6	1974	216	.100
7	1975	256	.147
8	1976	294	.058
9	1977	328	.092
10	1978	354	.146
11	1979	367	.069
12	1980	368	.041
13	1981	356	.116
14	1982	332	.085
15	1983	299	.076
16	1984	260	.087
17	1985	221	.155
18	1986	184	.142
19	1987	155	.199
20	1988	136	.119

You have obtained the following summary statistics from these data:

$$\sum x_i = 4,860 \qquad \sum y_i = 2.452 \qquad \sum x_i y_i = 539.309$$
$$\sum x_i^2 = 1,330,224 \qquad \sum y_i^2 = 0.341804$$

The associated ANOVA table is:

Source	SS	d.f.	MS	F Ratio
Regression	0.0214	1	0.0214	19.45
Residual	0.0198	18	0.0011	
Total (about mean)	0.0412	19		

Determine the upper bound of the shortest 95-percent confidence interval for the regression coefficient β_1.

(A) -0.0010

(B) -0.0008

(C) -0.0006

(D) -0.0004

(E) -0.0002

Solution. First, we find

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{539.309 - 20(243)(0.1226)}{\underbrace{1,330,224 - 20(243)^2}_{149,244}} = -0.000379.$$

With the standard error of $\hat{\beta}_1$ being

$$\text{SE}(\hat{\beta}_1) = \sqrt{\frac{s^2}{S_{xx}}} = \sqrt{\frac{0.0011}{149,244}} = 0.000086,$$

the upper bound of the symmetric 95% confidence interval for β_1 is

$$\begin{aligned} \hat{\beta}_1 + t_{18,0.025} \times \text{SE}(\hat{\beta}_1) &= -0.000379 + 2.101(0.000086) \\ &= \boxed{-0.000198}. \qquad \textbf{(Answer: (E))} \end{aligned}$$

□

Remark. You can also deduce from the F-statistic that the t-statistic for testing $\text{H}_0 : \beta_1 = 0$ (against $\text{H}_a : \beta_1 \neq 0$) is $t(\hat{\beta}_1) = \sqrt{F} = \sqrt{19.45} = 4.410215$. Then from $\hat{\beta}_1$ or s.e.$(\hat{\beta}_1)$, you can get the other.

Problem 1.6.34. (SOA Course 120 Study Note 120-81-95 Question 9: Confidence interval for β_1 – III) The following table shows the average rating (x) given by customers to 8 salespeople, and their average monthly salary (y) in thousands:

Salesperson	1	2	3	4	5	6	7	8
Rating	3.9	1.5	2.4	1.7	3.0	2.5	1.9	4.0
Salary	2.8	1.7	2.0	1.6	2.3	2.5	1.2	1.8

You use a simple linear regression model to test the relationship of salary to a salesperson's rating and have determined:

$$\begin{aligned} \bar{x} &= 2.612 \\ \bar{y} &= 1.987 \\ \sum (x_i - \bar{x})^2 &= 6.36 \\ \sum (y_i - \bar{y})^2 &= 1.91 \\ \sum (x_i - \bar{x})(y_i - \bar{y}) &= 2.08 \end{aligned}$$

Determine the upper bound of the shortest symmetric 90% confidence interval for β_1.

(A) 0.35

(B) 0.39

(C) 0.58

(D) 0.63

(E) 0.68

Solution. The least squares estimator of β_1 is

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{2.08}{6.36} = 0.327044.$$

Because the MSE is

$$s^2 = \frac{\text{RSS}}{n-2} = \frac{1.91 - 0.327044^2(6.36)}{8-2} = 0.204958,$$

the standard error of $\hat{\beta}_1$ is

$$\text{SE}(\hat{\beta}_1) = \sqrt{\frac{s^2}{S_{xx}}} = \sqrt{\frac{0.204958}{6.36}} = 0.179516.$$

Finally, a 90% confidence interval for β_1 is

$$\begin{aligned} \hat{\beta}_1 \pm t_{6,0.05} \times \text{SE}(\hat{\beta}_1) &= 0.327044 \pm 1.943 \times 0.179516 \\ &= [-0.0218, \boxed{0.6758}]. \qquad \textbf{(Answer: (E))} \end{aligned}$$

□

Problem 1.6.35. (SOA Course 120 Study Note 120-81-98 Question 3: Confidence interval for β_0 – I) 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% confidence interval for β_0.

(A) 1.1

(B) 1.2

(C) 1.3

(D) 1.4

(E) 1.5

Solution. Given RSS = 2.79, we have $s^2 = \text{RSS}/(n-2) = 2.79/8 = 0.34875$. By (1.4.2), the standard error of $\hat{\beta}_0$ is

$$\text{SE}(\hat{\beta}_0) = \sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} = \sqrt{0.34875\left(\frac{1}{10} + \frac{6^2}{180}\right)} = 0.323458.$$

The width of the symmetric 95% confidence interval for β_0 is

$$2t_{8,0.025}\text{SE}(\hat{\beta}_0) = 2(2.3060)(0.323458) = \boxed{1.4918}. \qquad \textbf{(Answer: (E))}$$

□

Problem 1.6.36. (CAS Exam ST Spring 2015 Question 20: Confidence interval for β_0 – II) You are given the following loss ratios for two lines of business, A and B, over a four-year period:

A	80	75	75	90
B	70	75	60	95

You are asked to set up a linear regression model for estimating loss ratios of line B using loss ratios of line A.

Based on the above sample, the estimated slope is $\frac{11}{6}$, and the unbiased estimate of the error variance is 72.92.

Calculate the upper bound of the symmetric 95 percent confidence interval for the intercept of the regression line.

(A) Less than -100

(B) At least -100, but less than 0

(C) At least 0, but less than 100

(D) At least 100, but less than 200

(E) At least 200

Solution. Note that loss ratios of line B are the response variables while loss ratios of line A are the explanatory variables (do not mix them up!). Using a financial calculator, we

find $\hat{\beta}_0 = -215/3$. By (1.4.1), the upper bound of the symmetric 95% confidence interval for β_0 is

$$\begin{aligned}\hat{\beta}_0 + t_{2,0.025}\sqrt{s^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)} &= -\frac{215}{3} + 4.3027\sqrt{72.92\left(\frac{1}{4}+\frac{80^2}{25,750-4(80)^2}\right)} \\ &= \boxed{169.03}. \qquad \textbf{(Answer: (D))}\end{aligned}$$

□

Remark. The value of $\hat{\beta}_1$ is not needed for solving this problem.

Problem 1.6.37. (SOA Course 120 May 1989 Question 6: Given the C.I. for β_1, find the C.I. for β_0) Your assistant fitted a set of observations to a two-variable regression model with independent variable $x_i = i$ for $i = 1$ to 5.

When you returned, you found that coffee had obliterated most of the information. The only things that you could make out were that the shortest 95% confidence interval for β_1 was $(-1.5, 2.5)$ and that $\hat{\beta}_0 = 3$.

Determine the upper bound of the shortest 95% confidence interval for β_0.

(A) 5.0

(B) 6.6

(C) 8.8

(D) 9.6

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

Comments: It seems that the examiners of the SOA are very prone to spilling coffee over their work...

Solution. Recalling (1.4.1), observe that

$$\widehat{\text{Var}}(\hat{\beta}_0) = \widehat{\text{Var}}(\hat{\beta}_1) \times \frac{\sum_{i=1}^n x_i^2}{n} = \widehat{\text{Var}}(\hat{\beta}_1) \times \frac{1^2+2^2+3^2+4^2+5^2}{5} = 11 s^2_{\hat{\beta}_1}.$$

From the confidence interval for β_1, we deduce that $t_{3,0.025}\sqrt{\widehat{\text{Var}}(\hat{\beta}_1)} = [2.5-(-1.5)]/2 = 2$. It follows that the shortest 95% confidence interval for β_0 is

$$\hat{\beta}_0 \pm t_{3,0.025}\sqrt{\widehat{\text{Var}}(\hat{\beta}_1)} = 3 \pm \sqrt{11} \times 2 = (-3.3633, \boxed{9.6332}). \quad \textbf{(Answer: (D))}$$

□

Problem 1.6.38. (SOA Part 3 November 1984 Question 6: Calculation of t-statistic from ANOVA output) You are given 10 pairs of values (x_i, y_i) which will 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:

$$\begin{aligned} \sum_{i=1}^{10} (x_i - \bar{x})^2 &= 400 \\ \sum_{i=1}^{10} (y_i - \bar{y})^2 &= 425 \\ \sum_{i=1}^{10} (\hat{y}_i - \bar{y})^2 &= 225 \end{aligned}$$

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

(Answer to the nearest 0.5)

(A) 1.5

(B) 2.0

(C) 2.5

(D) 3.0

(E) 3.5

Solution. We are given that TSS = 425 and Reg SS = 225, so RSS = TSS − Reg SS = 200. It follows that the F-statistic for testing $\mathrm{H}_0 : \beta_1 = 0$ against $\mathrm{H}_a : \beta_1 \neq 0$ is

$$F = \frac{\text{Reg SS}/1}{\text{RSS}/(n-2)} = \frac{225/1}{200/(10-2)} = 9.$$

The t-statistic, taken as the positive square root of F, is $t(\hat{\beta}_1) = \sqrt{9} = \boxed{3}$. **(Answer: (D))** □

Remark. (i) If the question does not specify the alternative hypothesis, assume a two-sided one.

(ii) Because there is no information suggesting that $\hat{\beta}_1$ must be positive, $t(\hat{\beta}_1) = -3$ should also be an answer. By the way, we can deduce the value of $\hat{\beta}_1^2$ via

$$\hat{\beta}_1^2 = \frac{\text{Reg SS}}{S_{xx}} = \frac{225}{400} = \frac{9}{16}.$$

Prediction

Problem 1.6.39. (SOA Course 120 May 1990 Question 5: Standard error of prediction – I) You have performed a simple linear regression of the number of hours required to fill a production order on the size of the order.

You have analyzed a sample of 25 orders with a mean order size of five.

You have determined:

$$\begin{aligned} \text{RSS} &= 92 \\ \widehat{\text{Var}}(\hat{\beta}_1) &= 1.0 \end{aligned}$$

Determine the estimated standard deviation of the number of hours required to fill an order of four units.

(A) 1.0

(B) 1.1

(C) 1.2

(D) 2.2

(E) 2.3

Solution. The MSE is $s^2 = \text{RSS}/(n-2) = 92/(25-2) = 4$. By (1.5.1), the standard error of prediction is

$$\sqrt{s^2\left[1+\frac{1}{n}+\frac{(x_*-\bar{x})^2}{S_{xx}}\right]} = \sqrt{4\left(1+\frac{1}{25}\right)+1.0(4-5)^2} = \boxed{2.2716}. \quad \textbf{(Answer: (E))}$$

□

Problem 1.6.40. (SOA Course 120 November 1985 Question 6: Standard error of prediction – II) You are given 20 pairs or values (x_i, y_i) which will 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:

$$\begin{aligned} \sum_{i=1}^{20} x_i &= 200 \\ \sum_{i=1}^{20} y_i &= 400 \\ \sum_{i=1}^{20} x_i^2 &= 2,400 \\ \sum_{i=1}^{20} y_i^2 &= 8,600 \\ s^2 &= 100 \end{aligned}$$

Determine the estimated standard deviation of the predicted value of y when $x = 26$.
(Answer to nearest integer)

(A) 10

(B) 11

(C) 12

(D) 13

(E) 14

Solution. The standard error of the prediction error is

$$\begin{aligned}\sqrt{s^2\left(1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right)} &= \sqrt{100\left[1+\frac{1}{20}+\frac{(26-10)^2}{2,400-20(10)^2}\right]} \\ &= \boxed{13}. \quad \textbf{(Answer: (D))}\end{aligned}$$

□

Problem 1.6.41. (SOA Course 120 Study Note 120-82-94 Question 6: Width of a prediction interval) Two observations on y were obtained for each of $x = 1, 2, 3, 4, 5$. A simple linear regression model fitted to these data resulted in the estimated regression line $\hat{y} = 4 + x$ and the following ANOVA table:

Source	SS	DF	MS
Regression	20.00	1	20.00
Residual	10.00	8	1.25
Total	30.00	9	

Assuming that y is normally distributed, a 95-percent prediction interval for y when $x = 5$ is in the form $9 \pm K$.
Determine K.

(A) 1.4

(B) 2.0

(C) 2.6

(D) 2.9

(E) 3.3

Solution. Note that there are $n = 5(2) = 10$ observations, and that $\bar{x} = 3$. Then

$$\begin{aligned} K &= t_{8,0.025}\sqrt{s^2\left[1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}\right]} \\ &= 2.3060\sqrt{1.25\left[1+\frac{1}{10}+\frac{(5-3)^2}{20}\right]} \\ &= \boxed{2.94}. \qquad \textbf{(Answer: D)} \end{aligned}$$

□

Problem 1.6.42. (SOA Course 120 Study Note 120-81-95 Question 4: Upper limit of PI) The following table gives the average annual yield (bushels/acre) and rainfall (inches) for the years 1970-1983:

Year	70	71	72	73	74	75	76	77	78	79	80	81	82	83
Yield (y)	17	13	19	10	12	13	6	8	9	17	18	28	26	11
Rain (x)	11	9	17	10	12	14	9	13	11	14	15	30	17	10

A simple linear regression analysis yields:

$$\begin{aligned} \hat{y} &= 1.35 + 0.98x \\ \sum(y_i - \bar{y})^2 &= 546.36 \\ \sum(\hat{y}_i - \bar{y})^2 &= 363.59 \end{aligned}$$

Determine the upper limit of the shortest symmetric 95% prediction interval for the yield in a year with 15 inches of rainfall.

(A) 19

(B) 21

(C) 23

(D) 25

(E) 27

Solution. The point prediction is $\hat{y}_* = 1.35 + 0.98(15) = 16.05$. The error sum of squares is

$$\text{RSS} = \text{TSS} - \text{Reg SS} = 546.36 - 363.59 = 182.77,$$

so the MSE is $s^2 = 182.77/(14-2) = 15.2308$. To find S_{xx}, use the relationship

$$\text{Reg SS} = \hat{\beta}_1^2 S_{xx} \quad \Rightarrow \quad S_{xx} = \frac{363.59}{0.98^2} = 378.5818.$$

Therefore, the standard error of the prediction error is

$$\sqrt{s^2\left(1+\frac{1}{n}+\frac{(x_*-\bar{x})^2}{S_{xx}}\right)}=\sqrt{15.2308\left(1+\frac{1}{14}+\frac{(15-13.7143)^2}{378.5818}\right)}=4.047866.$$

The upper limit of the shortest symmetric 95% prediction interval for y_* at $x_*=15$ is

$$16.05+\underbrace{t_{12,0.025}}_{2.179}(4.047866)=\boxed{24.87}. \quad \textbf{(Answer: (D))}$$

□

Remark. The only use of the raw data set is to find $\bar{x}=13.7143$.

Problem 1.6.43. (CAS Exam S Spring 2017 Question 36: Prediction interval in SLR) You are given the following information for a model fitted using ordinary least squares (OLS):

Response variable	Rating
Response distribution	normal
Link	identity
MSE	6.993

Parameter	df	$\hat{\beta}$	standard error ($\hat{\beta}$)
Intercept	1	14.37632	6.61999
Complaints	1	0.75461	0.09753

Summary Statistic	Rating	Complaints
N	30	30
Min	40	37
Max	85	90
Sum	1939	1998
Median	65.5	65.0
Sample Std. Dev.	12.1725619	13.3147572

Calculate the upper bound of the 95% prediction interval for Rating, for an observation with a Complaints value of 50.

(A) Less than 50

(B) At least 50, but less than 60

(C) At least 60, but less than 70

(D) At least 70, but less than 80

(E) At least 80

Solution. Here, "rating" (y) is the response variable and "complaints" (x) is the single explanatory variable. We are given from the table that $\hat{\beta}_0 = 14.37632$, $\hat{\beta}_1 = 0.75461$, $s^2 = 6.993$, $n = 30$ (sample size), and $\sum_{i=1}^{30} x_i = 1998$. To apply (1.5.2), it remains to find

$$S_{xx} = \sum_{i=1}^{30}(x_i - \bar{x})^2 = 29\left(\sqrt{\frac{1}{29}\sum_{i=1}^{30}(x_i - \bar{x})^2}\right)^2 = 29(13.3147572)^2 = 5141.2000.$$

A 95% prediction interval for Rating, for an observation with a Complaints value of $x_* = 50$, is

$$\begin{aligned}
& (\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{28,0.025}\sqrt{s^2\left[1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}\right]} \\
= \; & (14.37632 + 0.75461 \times 50) \pm 2.0484\sqrt{6.993\left(1 + \frac{1}{30} + \frac{(50 - 1998/30)^2}{5141.2000}\right)} \\
= \; & 52.10682 \pm 5.646286 \\
= \; & (46.46, \boxed{57.75}). \qquad \textbf{(Answer: (B))}
\end{aligned}$$

□

Remark. (i) Do not take $\bar{x} = 1939/30$, which should be for $\bar{y}$!

(ii) For the time being, ignore the term "link." It will be explained in Subsection 5.1.1 in the context of generalized linear models.

Problem 1.6.44. [HARDER!] (Construction of a PI from scratch) The following table shows the inflation rate (y) prevailing in a certain country when the interest rate is set by the central bank at x:

y (%)	2.2	2.8	3.1	3.6	3.2
x (%)	2.0	2.5	3.0	3.5	4.0

A simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ is considered.

Calculate the lower bound of the 90% symmetric prediction interval for the prevailing inflation rate when the central bank sets the interest rate at 3.2%.

(A) 1

(B) 2

(C) 3

(D) 4

(E) 5

Solution. Using a financial calculator, we can find the LSEs: $\hat{\beta}_0 = 1.3$ and $\hat{\beta}_1 = 0.56$. With $\text{TSS} = \sum_{i=1}^{5} y_i^2 - 5\bar{y}^2 = 45.49 - 5(2.98)^2 = 1.088$ and $\text{Reg SS} = \hat{\beta}_1^2 S_{xx} = 0.56^2[47.5 - 5(3)^2] = 0.784$, we have $\text{RSS} = 1.088 - 0.784 = 0.304$. Hence the MSE is

$$s^2 = \frac{\text{RSS}}{n-2} = \frac{0.304}{5-2} = 0.101333.$$

The 90% symmetric prediction interval for y_* observed at $x_* = 3.2\%$ is

$$\begin{aligned}
& \hat{y}_* \pm t_{3,0.05} \times \sqrt{s^2 \left[1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}\right]} \\
= \; & [1.3 + 0.56(3.2)] \pm 2.3534 \times \sqrt{0.101333 \left[1 + \frac{1}{5} + \frac{(3.2-3)^2}{2.5}\right]} \\
= \; & 3.092 \pm 0.825970 \\
= \; & [\boxed{2.27}, 3.92]. \qquad \textbf{(Answer: (B))}
\end{aligned}$$

□