Statistical Modeling: Regression, Survival Analysis, and Time Series Analysis

Lawrence M. Leemis
Department of Mathematics
William & Mary
Williamsburg, Virginia

Library of Congress Cataloging-in-Publication Data

Leemis, Lawrence M.
Statistical Modeling: Regression, Survival Analysis, and Time Series Analysis /
Lawrence M. Leemis
Includes bibliographic references and index.
ISBN 978-0-9829174-3-5
1. Statistics
QA 273.L44 2023

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

ISBN 978-0-9829174-3-5

Statistical Modeling: Regression, Survival Analysis, and Time Series Analysis

Short Contents

Detailed Contents

Preface

This book provides a brief introduction to three statistical modeling techniques: regression, survival analysis, and time series analysis. My motivation for writing this book came from a recent article in *Nature* that indicated that the paper introducing the product–limit estimator by American statisticians Edward Kaplan and Paul Meier in 1958 and the paper introducing the proportional hazards model written by British statistician David Cox in 1972 were the two most cited papers in the statistical literature. Yet most undergraduates majoring in applied mathematics, statistics, data science, systems engineering, and management science do not encounter the statistical models developed in either of these two pivotal papers. This book provides an elementary introduction to these two statistical procedures, and many others.

This book is designed as a one-semester introduction to regression, survival analysis, and time series analysis for advanced undergraduates or first-year graduate students. The pre-requisites for this book are (a) a course in linear algebra, (b) a calculus-based introduction to probability, and (c) a course in mathematical statistics that covers point estimation, interval estimation, and hypothesis testing. The book is not comprehensive and is not a replacement for a full-semester class on each of the topics. It contains only brief introductions to the three topics.

Three chapters are devoted to each of the three topics. The initial two chapters move at about the pace one would expect in a full-semester course. The third chapter on each of the topics is like a "further reading" section which briefly introduces some topics that would be covered in depth in a full-semester course. An instructor might choose to skip or expand on these topics.

The material in the book can be covered at the ambitious pace of one chapter per week. An instructor could also choose to move more slowly if some of this material is part of a course covering another topic.

Most of the data sets that are used for examples in the book are given as clear text on the website `www.math.wm.edu/~leemis/data/topics`.

The text is organized into chapters, sections, and subsections. When there are several topics within a subsection, they are set off by **boldface** headings. Definitions and theorems are boxed; examples are indented; proofs are terminated with a box, like this: □. Proofs are included when they are instructive to the material being presented. Exercises are numbered sequentially at the end of each chapter. Computer code is set in monospace font, and is not punctuated. Indentation is used to indicate nesting in code and pseudocode. An index is included. Italicized page numbers in the index correspond to the primary source of information on a topic.

The term *estimator* is used to describe a point estimator in the abstract or as a random variable; the term *estimate* is used to describe a point estimator that assumes a specific value estimated from a realization of data values. In some instances the case is altered to highlight this distinction. The sample mean $\bar{X}$, for example, is a point estimator for the population mean μ. A numerical value of the sample mean calculated from data values is sometimes denoted by the point estimate $\bar{x}$.

The R language is used throughout the text for graphics, computation, and Monte Carlo simulation. In many of the examples involving computations, the results are computed arithmetically, then confirmed in R, and then computed a third time using an R built-in function (such as `lm` for computing the coefficients in a regression model, `coxph` from the `survival` package for computing the regression coefficients in a Cox proportional hazards model, `survfit` from the `survival` package to calculate the step heights in the Kaplan–Meier product–limit estimator, or `arima` to fit a univariate time series). This three-step process is used to avoid treating R functions as black boxes without considering what goes on underneath the hood. R can be downloaded for free at `r-project.org`.

There are no references cited in the text for readability. The sources of materials in the various chapters are cited in the paragraphs below.

Chapter 1 notes: The quote by George Box is from page 202 of the book chapter: Box, G.E.P. (1979), "Robustness in the Strategy of Scientific Model Building," from Robustness in Statistics, edited by R.L. Launer and G.N. Wilkinson, New York: Academic Press, pages 201–236. The data pairs associated with the boiling points and barometric pressures in Example 1.11 are from Forbes, J. (1857), "Further Experiments and Remarks on the Measurement of Heights and Boiling Point of Water," *Transactions of the Royal Society of Edinburgh*, Volume 21, Issue 2, pages 235–243.

Chapter 2 notes: The four sets of data pairs known as *Anscombe's quartet* are from Anscombe, F.J. (1973), "Graphs in Statistical Analysis," *The American Statistician*, Volume 27, Number 1, pages 17–21. The housing data set in Example 2.9 is from De Cock, D. (2011), "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project," *Journal of Statistics Education*, Volume 19, Number 3, pages 1–15. The Shapiro–Wilk test for normality (and related tests) are overviewed in Razali, N., and Wah, Y.B. (2011), "Power Comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling Tests," *Journal of Statistical Modeling and Analytics*, Volume 2, Number 1, pages 21–33.

Chapter 3 notes: The chemical data from Example 3.1 is from Bennett, N.A., and Franklin, N.L. (1954), *Statistical Analysis in Chemistry and the Chemical Industry*, New York: Wiley. Cook's distances are derived in Cook, R.D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, Volume 19, Number 1, pages 15–18. The U.S. National debt over time is from `https://www.thebalance.com/national-debt-by-year-compared-to-gdp-and-major-events-3306287`. The original paper introducing ridge regression is Hoerl, A.E., and Kennard, R.W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, Volume 12, Number 1, pages 55–67.

Chapter 4 notes: Early references on the Weibull distribution include Fisher, R.A., and Tippett, L.H.C. (1928), "Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample," *Proceedings of the Cambridge Philosophical Society*, Volume 24, Issue 2, pages 180–190, Weibull, W. (1939), "A Statistical Theory of the Strength of Materials," *Ingeniors Vetenskaps Akademien Handlingar*, Number 153, and Weibull, W. (1951), "A Statistical Distribution Function of Wide Applicability," *Journal of Applied Mechanics*, Volume 18, pages 293–297. The moment ratio diagrams given in Section 4.5 are adapted from those given in Vargo, E., Pasupathy, R., and Leemis, L. (2010), "Moment-Ratio Diagrams for Univariate Distributions," *Journal of Quality Technology*, Volume 42, Number 3, pages 1–11. The Cox proportional hazards model was formulated in Cox, D.R. (1972), "Regression Models and Life-Tables" (with discussion), *Journal of the Royal Statistical Society B*, Volume 34, Number 2, pages 187–220.

Chapter 5 notes: The ball bearing failure times from Example 5.5 are from Lieblein, J., and Zelen, M. (1956), "Statistical Investigation of the Fatigue Life of Deep-Groove Ball Bearings," *Journal of Research of the National Bureau of Standards*, Volume 57, Number 5, pages 273–316. The 48.48 data value in the ball bearing data set is given as 48.40 on page 99 of Lawless, J.F. (2003), *Statistical Models and Methods for Lifetime Data*, Second Edition, Hoboken, NJ: John

Wiley & Sons, Inc., and page 4 of Meeker, W.Q., and Escobar, L.A. (2022), *Statistical Methods for Reliability Data*, Second Edition, New York: John Wiley & Sons, Inc., but is listed as 48.48 in Caroni, C. (2002), "The Correct 'Ball Bearings' Data," *Lifetime Data Analysis*, Volume 8, Number 4, pages 395–399. The 6–MP data set is from Gehan, E.A. (1965), "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples," *Biometrika*, Volume 52, Parts 1 and 2, pages 203–223. The automotive a/c switch failure times are from pages 253–254 of Kapur, K.C., and Lamberson, L.R. (1977), *Reliability in Engineering Design*, New York: John Wiley & Sons, Inc. The initial estimator for the Weibull shape parameter κ from a complete data set is given by Menon, M.V. (1963), "Estimation of the Shape and Scale Parameters of the Weibull Distribution," *Technometrics*, Volume 5, Number 2, pages 175–182.

Chapter 6 notes: The Clopper–Pearson confidence interval was introduced by Clopper, C.J., and Pearson, E.S. (1934), "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, Volume 26, Number 4, pages 404–413. The Wilson–Score confidence interval was introduced by Wilson, E.B. (1927), "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, Volume 22, Number 158, pages 209–212. The Jeffreys confidence interval is described by Brown, L.D., Cai, T.T., and DasGupta, A. (2001), "Interval Estimation for a Binomial Proportion," *Statistical Science*, Volume 16, Number 2, pages 101–133. The Agresti–Coull confidence interval was introduced by Agresti, A., and Coull, B.A. (1998), "Approximate is Better than 'Exact' for Interval Estimation of Binomial Proportions," *The American Statistician*, Volume 52, Number 2, pages 119–126. The product–limit estimator was devised by Kaplan, E.L., and Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, Volume 53, Number 282, pages 457–481. The earliest reference to Greenwood's formula comes from Greenwood, M. (1926), "The Natural Duration of Cancer," *Reports on Public Health and Medical Subjects*, Her Majesty's Stationery Office, London, Volume 33, pages 1–26. The proof of Theorem 6.1 is given in Appendix C of Leemis, L.M. (2009), *Reliability: Probability Models and Statistical Methods*, Second Edition, Lightning Source. A lucid presentation of Poisson processes and nonhomogeneous Poisson processes is given by Ross, S.M. (2019), *Introduction to Probability Models*, Twelfth Edition, London: Academic Press.

Chapter 7 notes: The monthly international airline traveler counts from Example 7.2 are Series G from Box, G.E.P., and Jenkins, G.M. (1976), Time Series Analysis: Forecasting and Control, Oakland, CA: Holden–Day. Introductions to the basics of time series analysis are given in Chatfield, C. (2004), *The Analysis of Time Series: An Introduction*, Sixth Edition, Boca Raton, FL: Chapman & Hall/CRC, and Brockwell, P.J., and Davis, R.A. (2016), *Introduction to Time Series and Forecasting*, Third Edition, Springer International Publishing Switzerland.

Chapter 8 notes: The details associated with the Box–Pierce test are given in Box, G.E.P., and Pierce, D.A. (1970), "Distribution of Residual Auto-Correlations in Autoregressive-Integrated Moving Average Time-Series Models," *Journal of the American Statistical Association*, Volume 65, Number 332, pages 1509–1526. The details associated with the Ljung–Box test are given in Ljung, G.M., and Box, G.E.P. (1978), "On a Measure of Lack of Fit in Time Series Models," *Biometrika*, Volume 65, Number 2, pages 297–303. The turning point test was first devised by Bienaymé, I.–J. (1874), "Sur Une Question de Probabilités," Bulletin de la Société Mathématique de France, Volume 2, pages 153–154. The Akaike Information Criterion was formulated in Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, Volume 19, Number 6, pages 716–723. The corrected Akaike Information Criterion was formulated in Hurvich, C.M., and Tsai, C–L. (1989), "Regression and Time Series Model Selection in Small Samples," *Biometrika*, Volume 76, Number 2, pages 297–307.

Chapter 9 notes: The graph displaying the stationarity region in terms of ϕ_1 and ϕ_2 shown in

Figure 9.13 is adapted from a figure in Stralkowski, C.M. (1968), "Lower Order Autoregressive-Moving Average Stochastic Models and Their Use for the Characterization of Abrasive Cutting Tools," PhD Thesis, The University of Wisconsin. Rom Lipscus from the Virginia Institute of Marine Science helped me find the source of the Lake Huron lake level data given in Example 9.14. Lauren M. Fry from the NOAA Great Lakes Environmental Research Laboratory provided the source of the Lake Huron levels on pages 151–154 of

`https://tidesandcurrents.noaa.gov/publications/Monthly_Annual_Averages_IGLD55_1860thru1985.PDF`

The time series of 210 consecutive chemical production yields from Example 9.35 are from pages 120–121 of Box, G.E.P., Hunter, J.S., and Hunter, W.G. (2005), *Statistics for Experimenters: Design, Innovation, Discovery*, Second Edition, Hoboken, NJ: John Wiley & Sons. The data values are IGLD55, which means that they are the sea level (in feet) above the level of the Atlantic Ocean. The shortened version of the lynx pelt sales from the Hudson's Bay Company that was fit to the transformed AR(3) model in Example 9.29 was suggested by Wei, W.W.S. (2006), *Time Series Analysis: Univariate and Multivariate Methods*, Second Edition, Boston: Pearson/Addison–Wesley. An early comprehensive treatment of ARIMA modeling is given in Box, G.E.P., and Jenkins, G.M. (1976), Time Series Analysis: Forecasting and Control, Oakland, CA: Holden–Day.

There are dozens of people to thank for making this book possible. Carrie Cooper, Lisa Nickel, Tami Back, Rosie Liljenquist, and all of the librarians and generous donors at the Swem Library at William & Mary made this book possible through their Library Scholar position. Thanks also goes to the William & Mary statisticians Ed Chadraa, Flip deCamp, Greg Hunt, Ross Iaci, Rui Pereira, Heather Sasinowska, and Guannan Wang, who have helped brainstorm about the topics and their sequencing in the book. I am grateful for Olivia Ding, Kexin Feng, Robert Jackson, Yuxin Qin, Chris Weld, and Hailey Young taking the time to read all or portions of an early draft of the text and providing helpful feedback. Barry Lawson from Bates College helped with the inset and lines in Figure 7.25. Five special people have made extraordinary contributions to this book: Heather Sasinowska edited the regression and time series chapters, Robert Lewis edited most of the entire textbook, Raghu Pasupathy provided keen insight concerning the presentation of the time series material and the moment ratio diagrams, Rosie Liljenquist edited the first two chapters just before having her sixth baby, and my wife Jill helped me push the book over the finish line. Finally, thanks goes to Drea George for the handsome book cover.

Since this is an open educational resource, this book is a work in progress. Please e-mail any typographical errors or suggested alterations that you spot to me. Thank you.

Williamsburg, VA

Larry Leemis
January 2023

Part I

REGRESSION

Chapter 1

Simple Linear Regression

Regression is a statistical technique that involves describing the relationship between one or more *independent variables* and a single *dependent variable*. For simplicity, assume for now that there is just a single independent variable. To establish some notation, let

- X be an independent variable, also called an explanatory variable, predictor variable, or regressor, which is typically assumed to take on fixed values (that is, X is *not* a random variable) which can be observed without error, and
- Y be a dependent variable, also called a response variable, which is typically a continuous random variable.

The relationship between the independent variable X and the dependent variable Y is often established by collecting n data pairs denoted by (X_1, Y_1), (X_2, Y_2), $\ldots$, (X_n, Y_n), plotting these pairs on a pair of axes, and looking for a pattern that can be translated to a mathematical form. This process establishes an empirical mathematical model for the underlying relationship between the independent variable X and the dependent variable Y.

1.1 Deterministic Models

Regression analysis establishes a functional relationship between X and Y. The simplest type of relationship between X and Y is a *deterministic* relationship $Y = f(X)$. In this rare case, the value of Y can be determined without error once the value of X is known, so Y is not a random variable when the relationship between X and Y is deterministic. The *deterministic model* is described by $Y = f(X)$. Deterministic relationships are uncommon in real-world applications because there is typically uncertainty in the dependent variable. If data pairs (X_1, Y_1), (X_2, Y_2), $\ldots$, (X_n, Y_n) are collected and the deterministic relationship $Y = f(X)$ establishes the correct functional relationship between X and Y, then all of the data pairs will fall on the graph of the function $Y = f(X)$.

> **Example 1.1** Bob is a salesman. The independent variable X is the *number* of sales that he makes per week. Bob receives a \$50 commission for each sale, regardless of the amount of each sale. The dependent random variable Y is the total weekly commission that Bob receives. Find the deterministic relationship between X and Y.

In this setting, the independent variable X is a fixed constant which is measured without error, and the deterministic relationship between X and Y is

$$Y = f(X) = 50X.$$

This deterministic relationship expresses Y as a linear function of X. If the next three weeks of Bob's sales activity result in the three data pairs

$$(X_1, Y_1) = (6, 300), \qquad (X_2, Y_2) = (8, 400), \qquad \text{and} \qquad (X_3, Y_3) = (2, 100),$$

then all three of these data pairs will fall on the graph of the deterministic relationship $Y = f(X) = 50X$. The X_i values are distinct for these data pairs, but this need not necessarily be the case. Bob could have had weeks in which he made the same number of sales multiple times. Figure 1.1 shows the deterministic relationship and the three data values that fall on the line. Notice that the graph of $Y = f(X) = 50X$ passes through the origin, $(0, 0)$, because zero weekly sales results in no weekly commissions. In this particular example, a line is plotted even though X can only take on integer values.

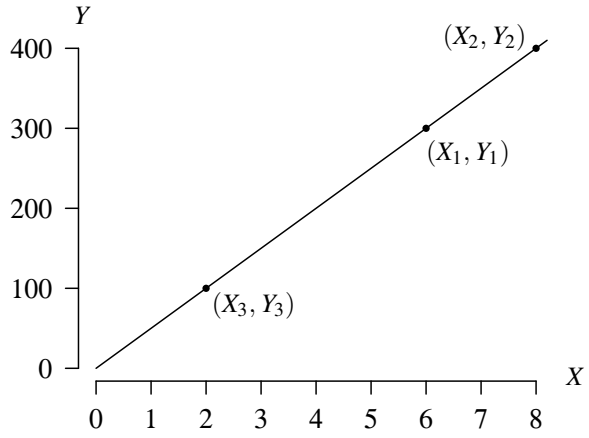

Figure 1.1: A deterministic linear relationship between X and Y.

Determining the relationship between the number of sales per week X and the commissions paid per week Y did not require the collection of any data to determine the function $Y = f(X)$. That linear relationship was implicit in the problem statement. Other cases can arise, such as (a) the relationship is deterministic but requires data to determine its functional form, or (b) the relationship is deterministic, but unlike the relationship in the previous example, it is not linear. The following example illustrates a nonlinear deterministic relationship between the independent variable X and the dependent variable Y.

Example 1.2 Alice purchases a five-year certificate of deposit paying 8% annually with an initial deposit of \$1000. Let the independent variable X be the number of months that the certificate of deposit has been held at a bank. Let the dependent variable Y be the associated balance. Find the deterministic relationship between X and Y assuming that the interest on the certificate of deposit is compounded monthly.

Under these assumptions, the balance on Alice's certificate of deposit at month X is

$$Y = f(X) = 1000\left(1 + \frac{0.08}{12}\right)^X.$$

(This relationship between X and Y makes three somewhat minor simplifying assumptions: (1) $Y = f(X)$ gives the instantaneous value of the CD after X months have passed even though interest is paid monthly, making this a continuous function rather than a step function, (2) all 12 months are assumed to have the same number of days, and (3) all years have the same number of days, which is not the case because of leap years. The violation of these assumptions are minor, and the relationship given here is very close to the balance Y after X months have passed.)

The curve in Figure 1.2 associated with the deterministic relationship is concave upward because of compounding. The three points plotted on the curve are

$$(X_1, Y_1) = (0, 1000), \quad (X_2, Y_2) = (12, 1083.00), \quad \text{and} \quad (X_3, Y_3) = (60, 1489.85).$$

The first data pair corresponds to the initial \$1000 deposit into the certificate of deposit at $X = 0$. The second data pair corresponds to the account balance after one year, or $X = 12$ months. The balance after 12 months is slightly more than the annual simple interest balance $\$1000 \cdot (1 + 0.08) = \$1080$ because of the monthly compounding. The third data pair corresponds to the final balance of \$1489.85 after 60 months. As was the case with the sales commissions in the previous example, the three data pairs were not necessary to establish the deterministic relationship between the independent variable X and the dependent variable Y. Their relationship is implicit in the problem statement. In both examples, the three data pairs fall on the graph of the deterministic relationship.

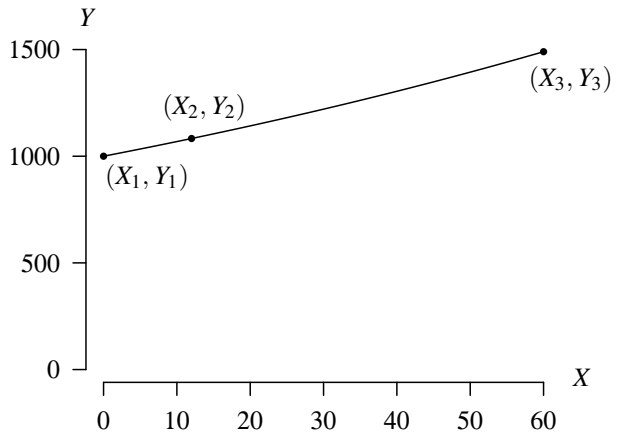

Figure 1.2: A deterministic nonlinear relationship between X and Y.

In most applications, the relationship between the independent variable X and the dependent variable Y is not deterministic because Y is typically a random variable. The next section introduces some of the thinking behind the development of a *statistical model* that describes the relationship between X and Y.

1.2 Statistical Models

The goal in constructing a statistical model is to write a formula that adequately captures the governing probabilistic relationship between an independent variable X and a dependent variable Y. This formula might be used subsequently for prediction or some other form of statistical inference. In this section, we assume that the dependent variable Y is a continuous random variable that can assume a range of values associated with a particular setting of the independent variable X. The relationship

$$Y = f(X)$$

that was used in the previous section is no longer adequate because X is assumed to be observed without error, and this formula results in a value of Y which is deterministic rather than random. One way of overcoming this problem is to replace the left-hand side of this equation by the expected value of Y, which is a constant, resulting in

$$E[Y] = f(X).$$

To be a little more careful about what is meant by this statistical relationship, the left-hand side is actually a conditional expectation, namely

$$E[Y \,|\, X = x] = f(x).$$

In words, given that the independent variable X assumes the value x, the transformation $f(x)$ gives the conditional expected value of the dependent variable Y. Notice that this statistical model does not specify the *distribution* of the random variable Y for a particular value of X; it only tells us the expected value of Y for a particular value of X. This statistical regression model defines a *hypothesized* relationship between the observed value of X on the right-hand side of the model and the conditional expected value of Y on the left-hand side of the model. The hypothesized relationship might be adequate for modeling or it might need some refining. There is typically no model that perfectly captures the relationship between X and Y. This was recognized by George Box, who wrote:

All models are wrong; some models are useful.

In a statistical model that involves parameters, the estimation of the model parameters will be followed by assessments to determine whether the model holds in an empirical sense. If the model needs refining, the new set of parameters are estimated and new assessments are made to see if the refined model is an improvement over the previous model. Regression modeling is an iterative process.

There is a second way to write a statistical model that is equivalent to the statistical model described in the previous paragraph. The model can be written as

$$Y = f(X) + \varepsilon,$$

where the error term ε (also known as the "noise" or "disturbance" term) is a random variable that accounts for the fact that the independent variable cannot predict the dependent variable with certainty. This term makes the relationship between X and Y a random (or *statistical* or *stochastic*) relationship rather than a deterministic relationship. If the probability distribution of the error term is specified, then not only is the expected value of Y conditioned on the value of X determined, but also the entire probability distribution of Y conditioned on the value of X is specified. It is common practice to assume that the expected value of ε is zero. The probability distribution of ε establishes

the nature and magnitude of the scatter of the data values about the regression function. When the population variance of ε is small, the values of Y are tightly clustered about the regression function $f(X)$; when the population variance of ε is large, the values of Y stray further from the regression function $f(X)$.

Regression modeling involves determining the functional form of $f(X)$ from a data set of n data pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. The statistical model for X and Y in a general sense also applies to each of the data points, so

$$Y_i = f(X_i) + \varepsilon_i$$

for $i = 1, 2, \ldots, n$. The sign of ε_i indicates whether the observed data pair (X_i, Y_i) falls above ($\varepsilon_i > 0$) or below ($\varepsilon_i < 0$) the conditional expected value of Y_i, for $i = 1, 2, \ldots, n$.

The function $f(X)$ is called the *regression function*, and was first referred to in print as such by Sir Francis Galton (1822–1911), a British anthropologist and meteorologist, in his 1885 paper titled "Regression Toward Mediocrity in Hereditary Stature" published in the *Journal of the Anthropological Institute*. He established a regression function relating the adult height of an offspring, Y, as a function of an average of the parent's heights, X, which had been adjusted for gender.

The regression function $Y = f(X)$ can be either linear or nonlinear. The next section focuses on the easier case, a linear regression function. In this case, the model is typically referred to as a *simple linear regression* model, which is often abbreviated as an SLR model. The model is *simple* because there is only one independent variable X that is used to predict the dependent variable Y. The model is *linear* because the regression function $f(X) = \beta_0 + \beta_1 X$ is assumed to be linear in the parameters β_0 and β_1. The more complicated cases of multiple linear regression, which involve more than one independent variable, and nonlinear regression, in which $f(X)$ is not a linear function, will be introduced later.

1.3 Simple Linear Regression Model

A simple linear regression model assumes a linear relationship between an independent variable X and a dependent variable Y. In this section, the more general regression model

$$Y = f(X) + \varepsilon$$

is reduced to the simple linear regression model given in the definition below.

Definition 1.1 A *simple linear regression model* is given by

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

where

- X is the independent variable, assumed to be a fixed value observed without error,
- Y is the dependent variable, which is a continuous random variable,
- β_0 is the population intercept of the regression line, which is an unknown constant,
- β_1 is the population slope of the regression line, which is an unknown constant, and
- ε is the error term, a continuous random variable with population mean zero and positive, finite population variance σ^2 that accounts for the randomness in the relationship between X and Y.

Stating the simple linear regression model in this fashion will not seem natural from probability theory. As a non-regression illustration from probability theory, $W \sim N(\mu, \sigma^2)$ indicates that W has a normal distribution with population mean μ and population variance σ^2. Although much less compact, the probability distribution of W can also be written as $W = \mu + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$. This illustration reflects the essence behind writing the simple linear regression model in the form $Y = \beta_0 + \beta_1 X + \varepsilon$ in Definition 1.1.

The formulation of the simple linear model from Definition 1.1 involves a random variable ε on the right-hand side of the model. In some settings, this model might be viewed as a transformation of a random variable, but this is not the correct interpretation of the model in this setting. The simple linear regression model defines a hypothesized relationship between the random variable on the left-hand side of the model and terms on the right-hand side of the model. This probability model is hypothesized to govern the relationship between X and Y. The goal in constructing a simple linear regression model is to determine if it adequately captures the probabilistic relationship between X and Y. Estimation of the model parameters will be followed by assessment to see if the model holds in an empirical sense.

The assumption that the random variable ε has population mean zero and population variance σ^2 in the most basic simple linear regression model in Definition 1.1 allows for mathematically tractable statistical inference. In models that allow for confidence intervals and hypothesis testing concerning the estimated slope and intercept, the error term is assumed to have a specific distribution, which is typically the normal distribution. The error term models all sources of variation, both known and unknown, other than the variation in Y associated with the particular level of X. Notice that σ^2 is constant over all values of X.

The assumption that the independent variable X is not subject to random variability is not always satisfied in practice. The fitting procedure becomes more complicated when X is considered to be a random variable. For this reason, we assume that the observed value of X is either exact or that the variation of X is small enough so that its observed value can be assumed to be exact.

The assumption of a *linear relationship* between X and Y might also be flawed. In some cases it might not be a perfectly linear relationship, but a linear relationship provides a close enough approximation between X and Y to be useful for associated statistical inference. In other cases, a linear relationship might be appropriate for some range of values of X, known as the *scope* of the model, but not others. One important step in establishing a simple linear regression model is to specify the values of X for which the simple linear regression model is valid.

The procedure for establishing a simple linear regression model that relates the dependent variable Y to the independent variable X is given below.

1. **Collect the data pairs**. The data pairs are denoted by (X_1, Y_1), (X_2, Y_2), $\ldots$, (X_n, Y_n). In some settings, it is possible to exert some control over the X_i values. As will be seen later, there are advantages to having the X_i values spread out as much as possible in terms of the precision of the fitted regression model.

2. **Make a scatterplot of data pairs**. A *scatterplot* is just a plot of the points (X_1, Y_1), (X_2, Y_2), $\ldots$, (X_n, Y_n) on a set of axes. The purpose of the scatterplot is to see if the linear relationship between X and Y is appropriate and to visually assess the spread of the data values about the regression function. With modern statistical software, scatterplots are easy to generate.

3. **Inspect the scatterplot**. Although this step is subjective, it is important to visually assess (a) whether the relationship between X and Y appears to be linear or nonlinear, (b) whether the spread of the data pairs about the regression function is small or large, and (c) whether the

variability of the data pairs about the regression function remains constant over the range of X values that have been collected.

4. **State the regression model**. In this chapter, the regression model is assumed to be the simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$. Nonlinear regression models, such as the quadratic model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$, and multiple regression models with more than one independent variable, such as $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, will be considered later.

5. **Fit the regression model to the data pairs**. The method of *least squares*, which will be described in the next section, is commonly used to estimate the parameters in the regression model. The least squares criterion is to choose the regression model that minimizes the sum of the squares of the vertical differences between data points and the fitted regression model.

6. **Assess the adequacy of the fitted regression model**. Visual assessment techniques for assessing the fitted regression model include superimposing the fitted regression model onto the scatterplot of the data pairs and examining a plot of the residuals. A residual is the signed vertical distance between a data pair and its associated value on the regression function. In addition, there are statistical methods that can be applied to the fitted regression model to see if it adequately describes the relationship between X and Y.

7. **Perform statistical inference**. Once the fitted regression model is deemed an acceptable approximation to the relationship between X and Y, it can be used for statistical inference. One simple example of statistical inference that occurs often in practice is the prediction of a future value of Y for a particular level of X.

The seven steps for establishing a regression model are not necessarily performed in the order given here. Many times the fitted regression model is rejected in Step 6, and it is necessary to return to Step 4 in order to formulate an alternative model. Steps 4 through 6 might need to be repeated several times before arriving at an acceptable model for statistical inference.

The simple linear regression model given in Definition 1.1 implies that all of the (X_i, Y_i) pairs also follow the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

for $i = 1, 2, \ldots, n$, where

- (X_i, Y_i) are the data pairs, for $i = 1, 2, \ldots, n$,
- X_i is the value of the independent variable for observation i, which is observed without error, for $i = 1, 2, \ldots, n$,
- Y_i is the value of the dependent variable for observation i, which is a continuous random variable, for $i = 1, 2, \ldots, n$,
- β_0 is the population intercept of the regression line,
- β_1 is the population slope of the regression line, and
- ε_i is the random error term for observation i which satisfies
  - $E[\varepsilon_i] = 0$ for $i = 1, 2, \ldots, n$,
  - $V[\varepsilon_i] = \sigma^2$ for $i = 1, 2, \ldots, n$,

– the random ε_i values are mutually independent random variables, which implies that their variance–covariance matrix is diagonal.

When the simple linear regression model is stated in this fashion, four properties become apparent. First, Y_i is a random variable that can be broken into two components: a deterministic component $\beta_0 + \beta_1 X_i$, and a random component ε_i, for $i = 1, 2, \ldots, n$. Second, Y_i has population mean

$$E[Y_i] = E[\beta_0 + \beta_1 X_i + \varepsilon_i] = \beta_0 + \beta_1 X_i$$

for $i = 1, 2, \ldots, n$ and population variance

$$V[Y_i] = V[\beta_0 + \beta_1 X_i + \varepsilon_i] = V[\varepsilon_i] = \sigma^2$$

for $i = 1, 2, \ldots, n$. Using slightly different notation, it would be reasonable to write the population mean and variance as the conditional expectations

$$E[Y_i \,|\, X_i] = \beta_0 + \beta_1 X_i \qquad \text{and} \qquad V[Y_i \,|\, X_i] = \sigma^2$$

for $i = 1, 2, \ldots, n$. The property that the variance does not change with X_i is known as *homoscedasticity*. Temporarily dropping the subscripts, the line

$$E[Y] = \beta_0 + \beta_1 X,$$

with β_0 and β_1 replaced by the associated estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$, is oftentimes superimposed onto the scatterplot to visualize the fitted regression model. Third, each data pair (X_i, Y_i) has a Y_i value that misses the regression function by the error term ε_i, for $i = 1, 2, \ldots, n$. Fourth, the values of the observed dependent variables $Y_1, Y_2, \ldots, Y_n$ must be mutually independent random variables because the error terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are mutually independent random variables.

1.4 Least Squares Estimators

We now turn to the question of estimating the intercept β_0 and the slope β_1 by the method of least squares. German mathematician Carl Friedrich Gauss (1777–1855) invented the least squares method and French mathematician Adrien–Marie Legendre (1752–1833) first published the method in 1805. The least squares method determines the values of β_0 and β_1 that minimize the sum of the squares of the errors, where the error is the vertical distance between the Y_i value and the fitted regression line. The term *estimator* will be used here to refer to a generic formula for $\hat{\beta}_0$ or $\hat{\beta}_1$; the term *estimate* will be used to refer to a specific numeric value for $\hat{\beta}_0$ or $\hat{\beta}_1$.

One bit of notation that will make the expressions of the point estimators more compact is

$$\begin{aligned} S_{XY} &= \sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) \\ &= \sum_{i=1}^{n} \left(X_i Y_i - X_i \bar{Y} - \bar{X} Y_i + \bar{X}\bar{Y}\right) \\ &= \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y} - n\bar{X}\bar{Y} + n\bar{X}\bar{Y} \\ &= \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}. \end{aligned}$$

Similarly,

$$S_{XX} = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$$

and

$$S_{YY} = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2.$$

This new notation allow us to express nS_{XY}, nS_{XX}, and nS_{YY} as

$$nS_{XY} = n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i,$$

$$nS_{XX} = n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2,$$

and

$$nS_{YY} = n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2.$$

Using this notation, the least squares estimators for the slope and intercept of the model, denoted by $\hat{\beta}_1$ and $\hat{\beta}_0$, are given in the following theorem. Notice that the term *normal equations* in the theorem is not related to the normal distribution.

Theorem 1.1 Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be n data pairs with at least two distinct X_i values. The *least squares estimators* of β_0 and β_1 minimize the sum of the squared deviations between Y_i and the associated fitted value $\hat{\beta}_0 + \hat{\beta}_1 X_i$ in the simple linear regression model. The least squares estimators are the solution to the simultaneous *normal equations*

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$
$$\hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i$$

and are given by

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$$

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$

where $\bar{X}$ and $\bar{Y}$ are the sample means

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \qquad \text{and} \qquad \bar{Y} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}.$$

Proof The deviation of Y_i from the associated value on the population regression line is

$$Y_i - (\beta_0 + \beta_1 X_i),$$

for $i = 1, 2, \ldots, n$. The sum of the squared deviations is

$$S = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

The least squares estimators are those that minimize S respect to β_0 and β_1; that is,

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Using calculus to minimize S with respect to β_0 and β_1 requires taking the partial derivatives of S with respect to β_0 and β_1:

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0$$
$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) = 0.$$

Simplifying and using the hat notation to denote the estimators results in the simultaneous normal equations

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$
$$\hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i.$$

The normal equations are a system of two linear equations in the two unknowns $\hat{\beta}_0$ and $\hat{\beta}_1$. Solving these equations simultaneously yields the point estimator for the slope

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}}.$$

Dividing the first normal equation by the sample size n yields the point estimator for the intercept

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

The next step is to show that the *least* squares estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ *minimize* S. This will be done by showing that the Hessian matrix is positive definite. The Hessian matrix $\mathbf{H}$ is the matrix of second partial derivatives of S with respect to β_0 and β_1:

$$\mathbf{H} = \begin{bmatrix} \dfrac{\partial^2 S}{\partial \beta_0^2} & \dfrac{\partial^2 S}{\partial \beta_0 \partial \beta_1} \\ \dfrac{\partial^2 S}{\partial \beta_1 \partial \beta_0} & \dfrac{\partial^2 S}{\partial \beta_1^2} \end{bmatrix} = \begin{bmatrix} 2n & 2\displaystyle\sum_{i=1}^{n} X_i \\ 2\displaystyle\sum_{i=1}^{n} X_i & 2\displaystyle\sum_{i=1}^{n} X_i^2 \end{bmatrix}.$$

The $\mathbf{H}$ matrix is unchanged when evaluated at the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. To show that this matrix is positive definite at the least squares estimators, it is sufficient to show that the upper-left-hand element and the determinant of $\mathbf{H}$ are both positive. The

upper-left-hand element is positive for all values of the sample size n. The determinant of $\mathbf{H}$ is

$$|\mathbf{H}| = \begin{vmatrix} 2n & 2\sum_{i=1}^{n} X_i \\ 2\sum_{i=1}^{n} X_i & 2\sum_{i=1}^{n} X_i^2 \end{vmatrix} = 4n\sum_{i=1}^{n} X_i^2 - 4\left(\sum_{i=1}^{n} X_i\right)^2.$$

This expression is positive when there are at least two distinct X_i values by the Cauchy–Schwartz inequality. The Cauchy–Schwartz inequality (a special case of the triangle inequality) states that for real numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$,

$$\left(a_1^2 + a_2^2 + \cdots + a_n^2\right) \cdot \left(b_1^2 + b_2^2 + \cdots + b_n^2\right) \geq \left(a_1 b_1 + a_2 b_2 + \cdots + a_n b_n\right)^2,$$

where equality is satisfied if and only if $a_1 = a_2 = \cdots = a_n$ and $b_1 = b_2 = \cdots = b_n$. Letting $a_i = 1$ and $b_i = x_i$ in the Cauchy–Schwartz inequality indicates that the determinant of $\mathbf{H}$ is positive when there are at least two distinct X_i values. Hence, the Hessian matrix $\mathbf{H}$ is positive definite and the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize S. □

The requirement that there are at least two distinct X_i values in Theorem 1.1 is consistent with intuition. Figure 1.3 shows $n = 5$ data pairs in which the independent variable assumes the same value for each pair: $X_1 = X_2 = X_3 = X_4 = X_5 = 3$. It is not possible to estimate the slope of the regression line in this particular setting. This is the geometric reason for the requirement that there are at least two distinct X_i values. In addition, the denominator in $\hat{\beta}_1 = S_{XY}/S_{XX}$ is zero when all X_i values are equal, which gives the associated algebraic reason for the requirement. From this point forward, whenever the simple linear regression model is used, it is assumed that the associated data pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ have at least two distinct X_i values.

Figure 1.4 shows the geometric interpretation associated with the estimated intercept $\hat{\beta}_0$ and estimated slope $\hat{\beta}_1$. The $n = 9$ data pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_9, Y_9)$ are plotted as points, along with the associated estimated regression line $Y = \hat{\beta}_0 + \hat{\beta}_1 X$. The y-intercept of the graph $\hat{\beta}_0$ is the height of the estimated regression line at $X = 0$. The "rise over run" interpretation of the slope is illustrated by the right triangle with legs consisting of dotted lines.

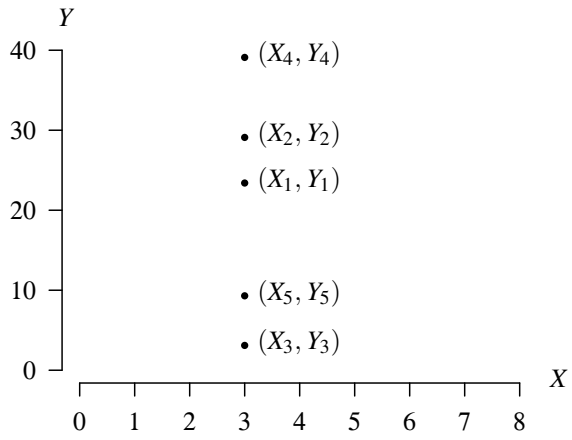

Figure 1.3: Identical independent variable values for all $n = 5$ data pairs.

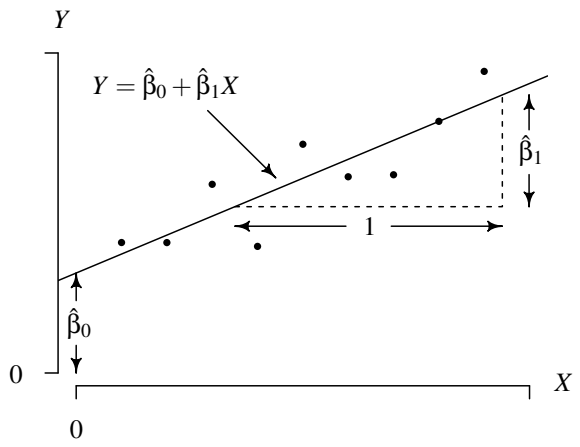

Figure 1.4: Geometry associated with $\hat{\beta}_0$ and $\hat{\beta}_1$.

The next example illustrates the mechanics associated with calculating the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. In order to focus on the calculations performed by hand, a small sample size of $n = 3$ data pairs is used. The numbers have been handpicked in order to make the resulting parameter estimates come out to whole numbers. A sample size of $n = 2$ is too simplistic in that two points determine a line, and the estimated regression line will always pass through those two points.

Example 1.3 Cheryl sells farm equipment and supplies. Let X be the number of sales she completes in a week, which will serve as the independent variable in this example. Each sale that she completes results in an associated random amount of revenue to the company that can be attributed to Cheryl's sales prowess. The dependent random variable Y is the associated total revenue to the company from Cheryl's sales for that week, in thousands of dollars. The data pairs for the past $n = 3$ weeks are

$$(X_1, Y_1) = (6, 2), \qquad (X_2, Y_2) = (8, 9), \qquad \text{and} \qquad (X_3, Y_3) = (2, 2).$$

Find the least squares estimates of the population intercept β_0 and population slope β_1 for the simple linear regression model from these data pairs and plot the fitted regression line and the data pairs on a single plot.

A scatterplot for this data set is generated using the `plot` function in the R commands

```
x = c(6, 8, 2)
y = c(2, 9, 2)
plot(x, y, xlim = c(0, 8), ylim = c(0, 9))
```

and is displayed in Figure 1.5. Your immediate reaction to the scatterplot might be to conclude that this is certainly not a linear relationship between X and Y. But this conclusion might not be warranted because of the small number of data pairs collected. One thing that is unusual about this data set is that Cheryl generated six sales in the first week, resulting in just $2000 in revenue, and then two sales in the third week, also resulting in $2000 in revenue. Clearly the sales transacted during the first week were

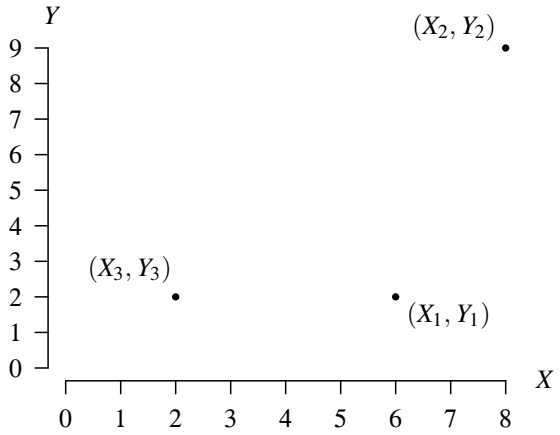

Figure 1.5: A scatterplot of the sales data pairs.

much smaller in size, on average, than those in the third week. Since the purpose of this example is to illustrate the calculations for computing $\hat{\beta}_0$ and $\hat{\beta}_1$, we will proceed as if the linear model were appropriate. Assessing a simple linear regression model with only $n = 3$ data pairs is nearly impossible.

The least squares estimates for β_0 and β_1 will be calculated in three different fashions. First, they will be calculated by hand, with all of the calculations displayed here. Second, they will be calculated in R using an approach that mirrors the hand calculations.

Third, they will be calculated in R using the `lm` (for linear model) function, which automates the process of estimating β_0 and β_1.

Table 1.1 contains the data pairs and calculations necessary to compute the estimated slope and intercept of the regression line. The sample means of the independent and dependent variables are

$$\bar{X} = \frac{16}{3} \qquad \text{and} \qquad \bar{Y} = \frac{13}{3}.$$

Although $\bar{X}$ and $\bar{Y}$ are set in upper case, it is important to remember that the X_i values are observed without error and the Y_i values are the associated random responses. The

Observation number i	Number of sales X_i	Total revenue Y_i	$(X_i - \bar{X})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	6	2	$(6-16/3)^2$	$(6-16/3)(2-13/3)$
2	8	9	$(8-16/3)^2$	$(8-16/3)(9-13/3)$
3	2	2	$(2-16/3)^2$	$(2-16/3)(2-13/3)$
Sum	16	13	168/9	168/9

Table 1.1: Data pairs and calculated values for estimating β_0 and β_1.

sums in the bottom row of Table 1.1 give the sums of squares

$$S_{XX} = \sum_{i=1}^{3} (X_i - \bar{X})^2 = \frac{168}{9} \qquad \text{and} \qquad S_{XY} = \sum_{i=1}^{3} (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{168}{9}.$$

The fact that $S_{XX} = S_{XY}$ is coincidental, and is typically not the case in practice. Using Theorem 1.1, the least squares estimates of β_1 and β_0 are

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{168/9}{168/9} = 1 \qquad \text{and} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{13}{3} - 1 \cdot \frac{16}{3} = -1.$$

A second way to calculate the least squares estimates $\hat{\beta}_1$ and $\hat{\beta}_0$ uses the R code below to implement the formulas given in Theorem 1.1. The code is generic in the sense that once the two vectors `x` and `y` are defined using the first two commands, the last four commands will calculate the point estimates $\hat{\beta}_1$ and $\hat{\beta}_0$ for any number of (X_i, Y_i) pairs.

```
x        = c(6, 8, 2)
y        = c(2, 9, 2)
sxx      = sum((x - mean(x)) ^ 2)
sxy      = sum((x - mean(x)) * (y - mean(y)))
beta1hat = sxy / sxx
beta0hat = mean(y) - beta1hat * mean(x)
```

This code also returns the point estimates

$$\hat{\beta}_1 = 1 \qquad \text{and} \qquad \hat{\beta}_0 = -1.$$

As you might imagine, these calculations are performed so often by statisticians that R has a built-in function to estimate β_1 and β_0.

A third way to calculate the least squares estimates of β_1 and β_0 via use of the R `lm` function.

```
x = c(6, 8, 2)
y = c(2, 9, 2)
lm(y ~ x)$coefficients
```

The `lm` function takes a formula for an argument, in this case `y`$\sim$`x`, and returns a list. One component of the list returned by `lm` is named `coefficients`, and it contains the estimated regression coefficients $\hat{\beta}_1 = 1$ and $\hat{\beta}_0 = -1$.

The fitted regression line is added to the scatterplot in Figure 1.6 using the R code below. The `plot` function plots the data pairs, the `lm` function estimates the intercept and slope of the regression line via least squares, and the `abline` function plots the fitted regression line. The labels on the data pairs can be added with the `text` function. The regression line plotted in Figure 1.6 is the line which minimizes the sum of the squares of the vertical distances between the points associated with the data pairs and the fitted regression line.

```
x = c(6, 8, 2)
y = c(2, 9, 2)
plot(x, y, xlim = c(0, 8), ylim = c(-1, 9))
fit = lm(y ~ x)
abline(fit$coefficients)
```

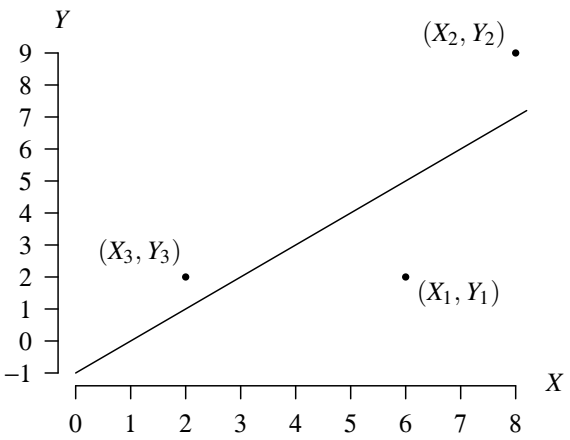

Figure 1.6: A scatterplot of the sales data pairs with the fitted regression line.

The fitted regression line has intercept $\hat{\beta}_0 = -1$ and slope $\hat{\beta}_1 = 1$. The fact that the intercept is $\hat{\beta}_0 = -1$ rather than $\hat{\beta}_0 = 0$ (because $X = 0$ sales in a week should result in $Y = 0$ revenue in that week) is due to random sampling variability. Section 3.1 investigates how to force a regression line through the origin, which would be appropriate in this setting. The interpretation of the estimated slope $\hat{\beta}_1 = 1$ is that the average amount of revenue generated from each sale that Cheryl completes is \$1000.

Figure 1.7 makes two embellishments to Figure 1.6. First, the axes have been adjusted so that the length of one unit on the vertical axis is the same as the length of one unit on the horizontal axis. Second, three shaded squares have been added to the plot. Each square has one vertex at a data pair, and a second vertex at the associated point on the fitted regression line. The numbers in each square give the area of the square. For these data pairs, the total area is the sum of squares

$$\begin{aligned} S &= (Y_1 - \hat{\beta}_0 - \hat{\beta}_1 X_1)^2 + (Y_2 - \hat{\beta}_0 - \hat{\beta}_1 X_2)^2 + (Y_3 - \hat{\beta}_0 - \hat{\beta}_1 X_3)^2 \\ &= (2+1-6)^2 + (9+1-8)^2 + (2+1-2)^2 \\ &= 9+4+1 \\ &= 14. \end{aligned}$$

The fitted least squares line is unique in the following sense. The squares illustrated in Figure 1.7 for any line having an intercept and/or slope that differ from $\hat{\beta}_0 = -1$ and $\hat{\beta}_1 = 1$ will have a total area that exceeds $S = 14$. The fitted least squares line is that line which minimizes S. If a different line were selected and plotted, some of the squares might become smaller, but at least one of the squares would become larger, and the total area of the squares would exceed 14.

Another way to view the minimization of S is to consider contours, or level surfaces, of the sum of squares as a function of the intercept β_0 and the slope β_1. The point

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = (-1, 1)$$

in Figure 1.8 corresponds to the fitted least squares line with a sum of squares $S = 14$ for the three data pairs. The concentric contours corresponding to $S = 15$, $S = 18$, $S = 23$,

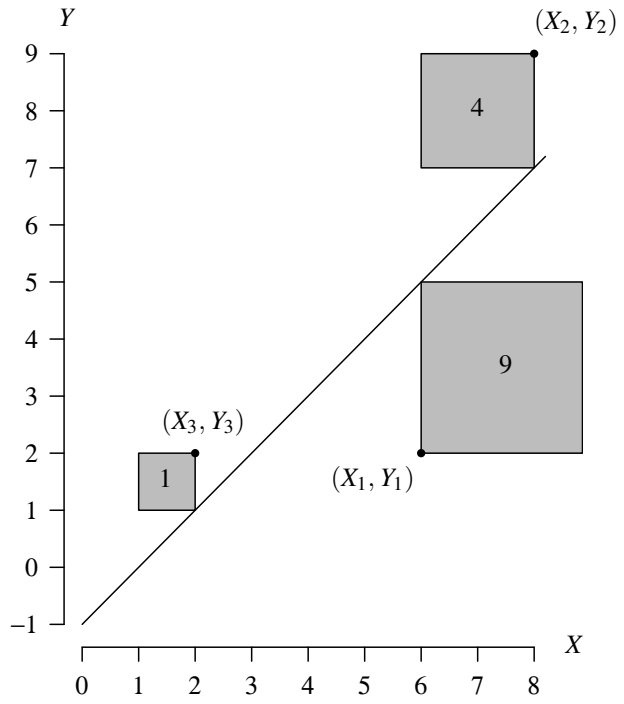

Figure 1.7: A scatterplot of the sales data pairs with the fitted regression line.

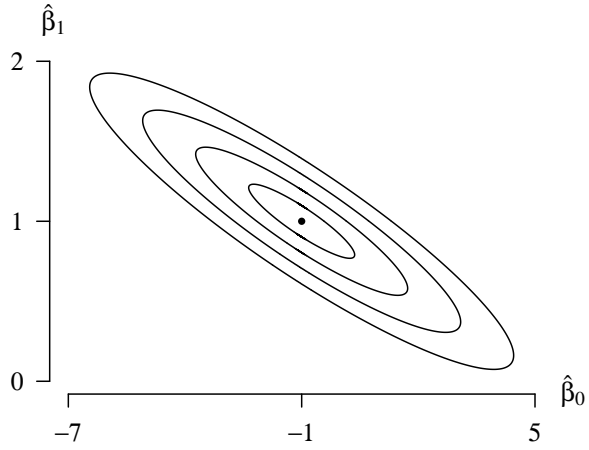

Figure 1.8: Level surfaces of the sum of squares.

and $S = 30$ show how the sum of squares increases as the intercept and slope stray from the least squares estimates.

1.5 Properties of Least Squares Estimators

The least squares estimators of β_0 and β_1 possess several properties which are important for statistical inference. The four properties established in this section are:

- the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1,
- the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ can be written as linear combinations of the dependent variables $Y_1, Y_2, \ldots, Y_n$,
- the variance–covariance matrix of $\hat{\beta}_0$ and $\hat{\beta}_1$ can be written in closed form, and
- the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ have the smallest population variance among all unbiased estimators that can be expressed as linear combinations of the dependent variables.

Proofs of the associated results are included in each of the following subsections.

1.5.1 $\hat{\beta}_0$ and $\hat{\beta}_1$ are Unbiased Estimators of β_0 and β_1

A key property associated with the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ is that their expected values equal the associated population values β_0 and β_1. The next result establishes the unbiasedness of the two point estimators.

Theorem 1.2 The least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ in the simple linear regression model are unbiased estimators of β_0 and β_1, respectively.

Proof To show that $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased estimators of β_1 and β_0, it is sufficient to show that

$$E\left[\hat{\beta}_1\right] = \beta_1 \qquad \text{and} \qquad E\left[\hat{\beta}_0\right] = \beta_0.$$

The denominator of the expression for $\hat{\beta}_1$, which is S_{XX}, is a constant because the values of the independent variables $X_1, X_2, \ldots, X_n$ are assumed to be observed without error in the simple linear regression model. Thus, the expected value of $\hat{\beta}_1$ is

$$\begin{aligned}
E\left[\hat{\beta}_1\right] &= E\left[\frac{S_{XY}}{S_{XX}}\right] \\
&= E\left[\frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^n X_i^2 - n\bar{X}^2}\right] \\
&= \frac{\sum_{i=1}^n X_i E\left[Y_i\right] - n\bar{X}E\left[\bar{Y}\right]}{\sum_{i=1}^n X_i^2 - n\bar{X}^2} \\
&= \frac{\sum_{i=1}^n X_i\left(\beta_0 + \beta_1 X_i\right) - n\bar{X}\left(\beta_0 + \beta_1\bar{X}\right)}{\sum_{i=1}^n X_i^2 - n\bar{X}^2} \\
&= \frac{\beta_0\sum_{i=1}^n X_i + \beta_1\sum_{i=1}^n X_i^2 - \beta_0\sum_{i=1}^n X_i - n\beta_1\bar{X}^2}{\sum_{i=1}^n X_i^2 - n\bar{X}^2} \\
&= \beta_1.
\end{aligned}$$

The expected value of $\hat{\beta}_0$ is

$$E\left[\hat{\beta}_0\right] = E\left[\bar{Y} - \hat{\beta}_1\bar{X}\right] = \beta_0 + \beta_1\bar{X} - \beta_1\bar{X} = \beta_0.$$

Therefore, $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased estimators of β_1 and β_0. □

The fact that the least squares estimators of the slope and intercept of the regression line are unbiased will be supported by a Monte Carlo simulation experiment in the next example. Unlike the typical simple linear regression setting in which data pairs (X_1, Y_1), (X_2, Y_2), ... , (X_n, Y_n) are used to estimate the *unknown* parameters β_0 and β_1, the simulation will generate data pairs and associated regression lines for *known* parameters β_0 and β_1.

Example 1.4 Consider the simple linear regression model

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

where

- the population intercept is $\beta_0 = 1$,
- the population slope is $\beta_1 = 1/2$, and
- the error term ε has a $U(-1, 1)$ distribution.

The population parameters have been chosen arbitrarily. The error term distribution has population mean zero and finite population variance, so it satisfies the conditions of a simple linear regression model from Definition 1.1. The uniform error term distribution is not likely to occur in practice, however, because it cuts off at -1 and 1. Probability distributions with tails, such as the normal distribution, are used more often in practice. Conduct a Monte Carlo simulation with 5000 replications that analyzes the probability distribution of the estimated intercept $\hat{\beta}_0$ and estimated slope $\hat{\beta}_1$ for $n = 10$ data pairs. Assume that the X_i values are equally likely to be one of the integers $0, 1, 2, \ldots, 9$. The independent variable X happens to assume discrete values in this example, but it would pose no difficulty if it took on continuous values.

One problem that might arise in the Monte Carlo experiment is that the X_i values might all be equal. This would violate the assumption in Theorem 1.1 that at least two X_i values must be distinct. Even though this event occurs with the low probability

$$10 \cdot \left(\frac{1}{10}\right)^{10} = 10^{-9},$$

an `if` statement will be included in the Monte Carlo simulation code to catch this problem if it occurs.

The R code below conducts 5000 replications of the Monte Carlo experiment. The commands prior to the `for` loop set the number of replications to 5000, set the sample size to $n = 10$, set the population intercept to $\beta_0 = 1$, set the population slope to $\beta_1 = 1/2$, define the vectors `beta0hat` and `beta1hat` to hold the simulated estimated intercepts and slopes, and establish the random number stream with the `set.seed` function with arbitrary argument. Within the `for` loop, `x` contains the values of the independent variables, `y` contains the values of the associated dependent variables, and `fit` is the list that stores the results of the regression analysis generated by the call to the `lm` function.

```
nrep     = 5000
n        = 10
```

```
beta0    = 1
beta1    = 1 / 2
beta0hat = numeric(nrep)
beta1hat = numeric(nrep)
set.seed(100)
for (i in 1:nrep) {
  x = sample(0:9, n, replace = TRUE)
  if (min(x) == max(x)) stop("All x values are equal")
  y = beta0 + beta1 * x + runif(n, -1, 1)
  fit = lm(y ~ x)
  beta0hat[i] = fit$coefficients[1]
  beta1hat[i] = fit$coefficients[2]
}
```

Figure 1.9 shows the scatterplot and the fitted regression line for the first replication of the simulation. Notice that having tied values for the independent variables poses no difficulty for calculating the estimates of the intercept and slope of the fitted regression line. This first fitted regression line has intercept $\hat{\beta}_0 = 1.398$ which exceeds the population intercept $\beta_0 = 1$; this first fitted regression line has slope $\hat{\beta}_1 = 0.399$ which is less than the population slope $\beta_1 = 0.5$. Each of the 5000 replications will yield unique values of $\hat{\beta}_0$ and $\hat{\beta}_1$. Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1 by Theorem 1.2, the 5000 simulated point estimates will hover around their population counterparts.

Figure 1.10 contains four lines. The thick, solid line is the population regression line with intercept $\beta_0 = 1$ and slope $\beta_1 = 1/2$. The other three dashed lines correspond to the fitted regression lines for the first three replications of the simulation. As expected, the estimated intercepts and slopes differ from the associated population values from one replication to the next.

When the simulation is run for all 5000 replications, there are 5000 $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ pairs generated. The *additional* R commands below plot a histogram of the 5000 $\hat{\beta}_0$ values on

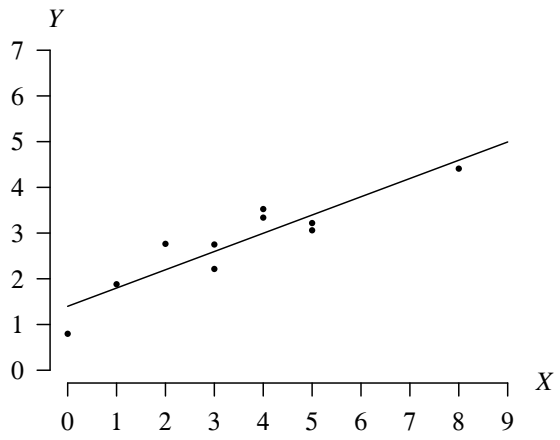

Figure 1.9: Scatterplot of simulated data pairs and fitted regression line (replication 1).

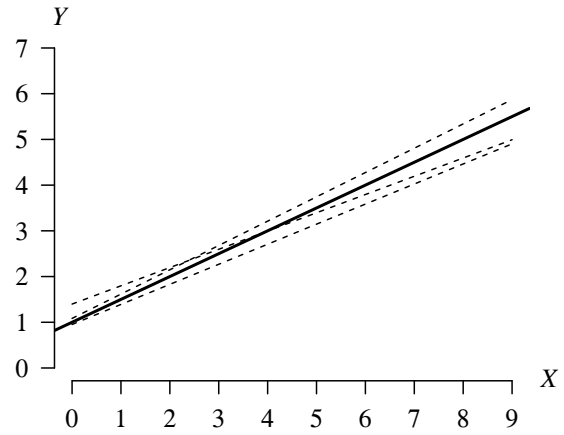

Figure 1.10: Population and fitted regression lines (replications 1–3).

the left and a histogram of the 5000 $\hat{\beta}_1$ values on the right. The `mfrow` (multiple frame by row) argument in `par` function sets up a 1×2 array of plots, and the `hist` function plots the histograms. Figure 1.11 contains the two histograms. The vertical axes have been suppressed because only the center and shape of the histogram is of interest.

```
par(mfrow = c(1, 2))
hist(beta0hat)
hist(beta1hat)
```

As predicted by Theorem 1.2, the histogram of the $\hat{\beta}_0$ values is centered around $\beta_0 = 1$ and the histogram of the $\hat{\beta}_1$ values is centered around $\beta_1 = 1/2$. Both histograms have a bell shape, indicating that the extreme values for the intercepts and slopes are less likely as you move further away from the population values. Although the error terms in the model are mutually independent $U(-1, 1)$ random variables, the summations

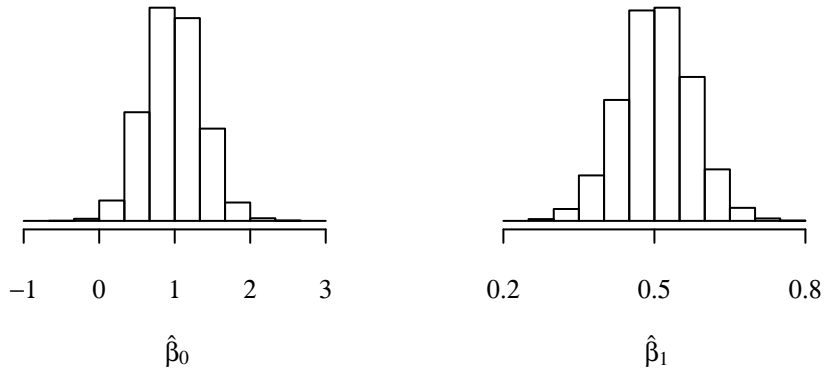

Figure 1.11: Histograms of estimated intercepts (left) and estimated slopes (right).

involved with the computation of $\hat{\beta}_0$ and $\hat{\beta}_1$ allow the central limit theorem to produce a histogram shape that is quite close to that of a normal probability density function.

The two histograms in Figure 1.11 do not indicate whether $\hat{\beta}_0$ and $\hat{\beta}_1$ are independent or dependent random variables. The *additional* R command

```
plot(beta0hat, beta1hat)
```

plots the 5000 $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ pairs, which is displayed in Figure 1.12. The Monte Carlo simulation indicates that the estimated intercepts and slopes are negatively correlated. They tend to be on the opposite sides of their respective means. A larger-than-usual slope is likely to be associated with a smaller-than-usual intercept, and vice versa.

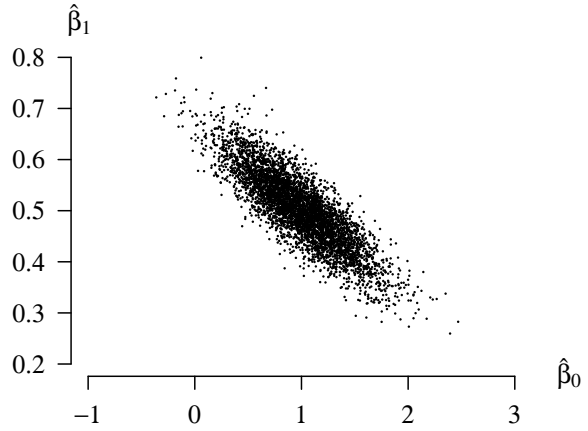

Figure 1.12: Estimated intercepts and slopes for 5000 Monte Carlo simulation replications.

The two key take-aways from this Monte Carlo experiment are:

- $\hat{\beta}_0$ and $\hat{\beta}_1$ being unbiased estimators of β_0 and β_1 via Theorem 1.2 is supported by the histograms in Figure 1.11, and
- $\hat{\beta}_0$ and $\hat{\beta}_1$ appear to be negatively correlated for this particular simple linear regression model by Figure 1.12.

1.5.2 $\hat{\beta}_0$ and $\hat{\beta}_1$ are Linear Combinations of $Y_1, Y_2, \ldots, Y_n$

Theorem 1.2, which states that $E\left[\hat{\beta}_0\right] = \beta_0$ and $E\left[\hat{\beta}_1\right] = \beta_1$, concerns the *accuracy* of the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. These estimators are "on target" in the sense that their expected values equal their associated population values. The histograms in Figure 1.11 show that the estimators for β_0 and β_1 do not systematically deviate above or below their population values.

The *precision* of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ is also of interest. This requires that we also compute their population variances. Before doing so, it is helpful to see that both of these point estimators can be written as linear combinations of the values of the dependent variables $Y_1, Y_2, \ldots, Y_n$.

It is not immediately apparent from the formula for the point estimator for the slope of the regression line $\hat{\beta}_1 = S_{XY}/S_{XX}$, but the estimator can be written as a linear combination of the dependent variables:

$$\begin{aligned}\hat{\beta}_1 &= \frac{S_{XY}}{S_{XX}} \\ &= \frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2} \\ &= \frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)Y_i}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}\end{aligned}$$

because $\bar{Y}\sum_{i=1}^{n}\left(X_i-\bar{X}\right) = \bar{Y}\left(n\bar{X}-n\bar{X}\right) = 0$. This formula indicates that the point estimator for the slope of the regression line is the linear combination

$$\hat{\beta}_1 = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n,$$

where

$$a_i = \frac{X_i-\bar{X}}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}$$

for $i = 1, 2, \ldots, n$.

The coefficients $a_1, a_2, \ldots, a_n$ in the linear combination $\hat{\beta}_1 = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n$ satisfy three properties. First, $\sum_{i=1}^{n} a_i = 0$ because

$$\sum_{i=1}^{n} a_i = \frac{1}{S_{XX}}\sum_{i=1}^{n}\left(X_i-\bar{X}\right) = \frac{n\bar{X}-n\bar{X}}{S_{XX}} = 0.$$

Second, $\sum_{i=1}^{n} a_iX_i = 1$ because

$$\sum_{i=1}^{n} a_iX_i = \frac{1}{S_{XX}}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)X_i = \frac{1}{S_{XX}}\left[\sum_{i=1}^{n}X_i^2 - n\bar{X}^2\right] = \frac{S_{XX}}{S_{XX}} = 1.$$

Third, $\sum_{i=1}^{n} a_i^2 = 1/S_{XX}$ because

$$\sum_{i=1}^{n} a_i^2 = \frac{1}{S_{XX}^2}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 = \frac{S_{XX}}{S_{XX}^2} = \frac{1}{S_{XX}}.$$

These properties can be useful in deriving results associated with the simple linear regression model.

Likewise, the least squares point estimator for the intercept of the regression line is also a linear combination of the Y_i values:

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1\bar{X} \\ &= \frac{1}{n}\sum_{i=1}^{n}Y_i - \bar{X}\sum_{i=1}^{n}\frac{X_i-\bar{X}}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}Y_i \\ &= \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}\cdot\frac{X_i-\bar{X}}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}\right)Y_i.\end{aligned}$$

This formula indicates that the point estimator for the intercept of the regression line can also be written as a linear combination:

$$\hat{\beta}_0 = c_1 Y_1 + c_2 Y_2 + \cdots + c_n Y_n,$$

where

$$c_i = \frac{1}{n} - \bar{X} \cdot \frac{X_i - \bar{X}}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}$$

for $i = 1, 2, \ldots, n$. This derivation constitutes a proof of the following result.

Theorem 1.3 The least squares estimators of the parameters β_0 and β_1 in the simple linear regression model can be written as linear combinations of the dependent variables:

$$\hat{\beta}_0 = c_1 Y_1 + c_2 Y_2 + \cdots + c_n Y_n$$

and

$$\hat{\beta}_1 = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n,$$

where

$$c_i = \frac{1}{n} - \bar{X} \cdot \frac{X_i - \bar{X}}{S_{XX}} \qquad \text{and} \qquad a_i = \frac{X_i - \bar{X}}{S_{XX}}$$

for $i = 1, 2, \ldots, n$, and

$$\sum_{i=1}^{n} a_i = 0, \qquad \sum_{i=1}^{n} a_i X_i = 1, \qquad \text{and} \qquad \sum_{i=1}^{n} a_i^2 = \frac{1}{S_{XX}}.$$

These formulas will be illustrated for the small data set consisting of $n = 3$ data pairs.

Example 1.5 Consider again the $n = 3$ data pairs

$$(X_1, Y_1) = (6, 2), \qquad (X_2, Y_2) = (8, 9), \qquad \text{and} \qquad (X_3, Y_3) = (2, 2)$$

from Example 1.3. Recall that the independent variable X is Cheryl's number of sales per week. Each sale results in a random amount of revenue to the company. The dependent random variable Y is the associated total revenue from the sales that Cheryl completes for a particular week, in thousands of dollars. Find the least squares estimates of the intercept β_0 and slope β_1 for the simple linear regression model using the formulas that express the estimates as linear combinations of Y_1, Y_2, Y_3 from Theorem 1.3.

The sample mean of the independent variables is

$$\bar{X} = \frac{6+8+2}{3} = \frac{16}{3}.$$

The value of S_{XX} is

$$S_{XX} = \sum_{i=1}^{3} \left(X_i - \bar{X}\right)^2 = \left(6 - \frac{16}{3}\right)^2 + \left(8 - \frac{16}{3}\right)^2 + \left(2 - \frac{16}{3}\right)^2 = \frac{4}{9} + \frac{64}{9} + \frac{100}{9} = \frac{56}{3}.$$

The coefficients for the linear combination associated with $\hat{\beta}_1$ are

$$a_i = \frac{X_i - \bar{X}}{S_{XX}}$$

for $i = 1, 2, 3$, or

$$a_1 = \frac{6 - 16/3}{56/3} = \frac{1}{28}, \qquad a_2 = \frac{8 - 16/3}{56/3} = \frac{1}{7}, \qquad a_3 = \frac{2 - 16/3}{56/3} = -\frac{5}{28}.$$

You might want to check that the three properties associated with the coefficients a_1, a_2, and a_3 from Theorem 1.3, namely $a_1 + a_2 + a_3 = 0$, $a_1X_1 + a_2X_2 + a_3X_3 = 1$, and $a_1^2 + a_2^2 + a_3^2 = 1/S_{XX}$, are all satisfied as expected. The least squares estimate of the slope of the regression line is

$$\hat{\beta}_1 = a_1Y_1 + a_2Y_2 + a_3Y_3 = \frac{1}{28} \cdot 2 + \frac{1}{7} \cdot 9 - \frac{5}{28} \cdot 2 = \frac{1}{14} + \frac{9}{7} - \frac{5}{14} = 1.$$

The R code for performing these calculations is given below.

```
x = c(6, 8, 2)
y = c(2, 9, 2)
a = (x - mean(x)) / sum((x - mean(x)) ^ 2)
beta1hat = sum(a * y)
```

The coefficients for the linear combination associated with $\hat{\beta}_0$ are

$$c_i = \frac{1}{n} - \bar{X} \cdot \frac{X_i - \bar{X}}{S_{XX}} = \frac{1}{n} - \bar{X} \cdot a_i$$

for $i = 1, 2, 3$, or

$$c_1 = \frac{1}{3} - \frac{16}{3} \cdot \frac{1}{28} = \frac{1}{7}, \qquad c_2 = \frac{1}{3} - \frac{16}{3} \cdot \frac{1}{7} = -\frac{3}{7}, \qquad c_3 = \frac{1}{3} - \frac{16}{3} \cdot \frac{-5}{28} = \frac{9}{7}.$$

The least squares estimate of the intercept of the regression line is

$$\hat{\beta}_0 = c_1Y_1 + c_2Y_2 + c_3Y_3 = \frac{1}{7} \cdot 2 - \frac{3}{7} \cdot 9 + \frac{9}{7} \cdot 2 = \frac{2}{7} - \frac{27}{7} + \frac{18}{7} = -1.$$

The R code for performing these calculations follows.

```
x = c(6, 8, 2)
y = c(2, 9, 2)
n = length(x)
c = 1 / n - mean(x) * (x - mean(x)) / sum((x - mean(x)) ^ 2)
beta0hat = sum(c * y)
```

In both cases the point estimates match the associated values calculated by the standard formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ from Theorem 1.1 that were used in Example 1.3, as expected.

1.5.3 Variance–Covariance Matrix of $\hat{\beta}_0$ and $\hat{\beta}_1$

Theorem 1.2 states that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1 because $E\left[\hat{\beta}_0\right] = \beta_0$ and $E\left[\hat{\beta}_1\right] = \beta_1$. This result concerns the *accuracy* of the least squares estimators, but does not address the *precision* of the least squares estimators. We now return to the question of assessing the precision

of the point estimators. Being able to express the point estimators of the least squares estimators as linear combinations of the dependent variables as summarized in Theorem 1.3 will be very useful as we proceed. In order to assess the precision of $\hat{\beta}_0$ and $\hat{\beta}_1$, it is necessary to compute $V\left[\hat{\beta}_0\right]$ and $V\left[\hat{\beta}_1\right]$. More generally, we will compute the variance–covariance matrix of $\hat{\beta}_0$ and $\hat{\beta}_1$ in this subsection. Returning to the Monte Carlo simulation in Example 1.4, the magnitudes of the diagonal elements of the variance–covariance matrix reflect the spread of the histograms in Figure 1.11, and the off-diagonal elements of the variance–covariance matrix give the population covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$ which is apparent in the simulation results displayed in Figure 1.12. The general form for the population covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$ will indicate whether the negative sample covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$ that was encountered in the Monte Carlo simulation was due to the particular values of the parameters in the simple linear regression model or whether the negative covariance is generally the case.

We begin with the lower-right-hand element of the variance–covariance matrix of $\hat{\beta}_0$ and $\hat{\beta}_1$. In the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

for $i = 1, 2, \ldots, n$, the error terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are assumed to be mutually independent random variables. This implies that the dependent variables $Y_1, Y_2, \ldots, Y_n$ are also mutually independent random variables. Using the fact that $\hat{\beta}_1$ can be written as a linear combination of the dependent variables from Theorem 1.3, the population variance of $\hat{\beta}_1$ is

$$\begin{aligned} V\left[\hat{\beta}_1\right] &= V\left[a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n\right] \\ &= \sum_{i=1}^{n} V\left[a_i Y_i\right] \\ &= \sum_{i=1}^{n} a_i^2 V\left[Y_i\right] \\ &= \left(\sum_{i=1}^{n} a_i^2\right) \sigma^2 \\ &= \frac{\sigma^2}{S_{XX}} \end{aligned}$$

because $\sum_{i=1}^{n} a_i^2 = 1/S_{XX}$ by Theorem 1.3. Although the experimenter typically has no control over σ^2, the experimenter may have control over selecting the values of $X_1, X_2, \ldots, X_n$ in some applications of simple linear regression. In order to make $V\left[\hat{\beta}_1\right]$ as small as possible, the experimenter should make S_{XX} as large as possible. Spreading the X_i values as much as possible gives the most stability to the estimated slope of the regression line. Simple linear regression modeling can still be performed when the X_i values are tightly clustered together, but the estimated slope will be less stable, and the scope of the model will be limited. As an extreme example of spreading the X_i values, consider clustering all of the X_i values at a left-most and a right-most extreme possible values for the independent variable. The good news is that this will give you the largest possible S_{XX} and the associated smallest possible $V\left[\hat{\beta}_1\right]$. The bad news is that you will not be able to assess linearity in this case because you have observed the dependent variable at only two values of the independent variable. A multitude of functions can model the average of the dependent variables at these two extreme values of the independent variable. So the usual practice is to select the X_i values in an approximately uniform fashion over as wide a range as possible. This gives the experimenter the opportunity to assess linearity and also achieves a large S_{XX}, resulting in an associated small $V\left[\hat{\beta}_1\right]$.

The next step is to calculate the upper-left-hand element of the variance–covariance matrix of $\hat{\beta}_0$ and $\hat{\beta}_1$. Before calculating the population variance of $\hat{\beta}_0$, it is necessary to establish that $\bar{Y}$ and $\hat{\beta}_1$ are uncorrelated. Since $Y_1, Y_2, \ldots, Y_n$ are mutually independent random variables, each with population variance $V[Y_i] = \sigma^2$, the population covariance between $\bar{Y}$ and $\hat{\beta}_1$ is

$$\begin{aligned}
\mathrm{Cov}\left(\bar{Y}, \hat{\beta}_1\right) &= \mathrm{Cov}\left(\frac{Y_1}{n} + \frac{Y_2}{n} + \cdots + \frac{Y_n}{n}, a_1Y_1 + a_2Y_2 + \cdots + a_nY_n\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}\left(\frac{Y_i}{n}, a_jY_j\right) \\
&= \sum_{i=1}^{n} \mathrm{Cov}\left(\frac{Y_i}{n}, a_iY_i\right) \\
&= \sum_{i=1}^{n} \frac{a_i}{n} V[Y_i] \\
&= \frac{\sigma^2}{n}\sum_{i=1}^{n} a_i \\
&= 0
\end{aligned}$$

because $\sum_{i=1}^{n} a_i = 0$ by Theorem 1.3. So $\bar{Y}$ and $\hat{\beta}_1$ are uncorrelated.

Based on the fact that the population covariance between $\bar{Y}$ and $\hat{\beta}_1$ is zero, the population variance of $\hat{\beta}_0$ is

$$\begin{aligned}
V\left[\hat{\beta}_0\right] &= V\left[\bar{Y} - \hat{\beta}_1\bar{X}\right] \\
&= V\left[\bar{Y}\right] + \bar{X}^2 V\left[\hat{\beta}_1\right] \\
&= \frac{\sigma^2}{n} + \frac{\bar{X}^2\sigma^2}{S_{XX}} \\
&= \left[\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right]\sigma^2 \\
&= \left[\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 + n\bar{X}^2}{n\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\right]\sigma^2 \\
&= \frac{\sum_{i=1}^{n} X_i^2}{nS_{XX}}\sigma^2.
\end{aligned}$$

The last step is to calculate the off-diagonal elements of the variance–covariance matrix of $\hat{\beta}_0$ and $\hat{\beta}_1$. Since $\mathrm{Cov}\left(\bar{Y}, \hat{\beta}_1\right) = 0$, the population covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$ is

$$\begin{aligned}
\mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) &= \mathrm{Cov}\left(\bar{Y} - \hat{\beta}_1\bar{X}, \hat{\beta}_1\right) \\
&= \mathrm{Cov}\left(\bar{Y}, \hat{\beta}_1\right) - \mathrm{Cov}\left(\hat{\beta}_1\bar{X}, \hat{\beta}_1\right) \\
&= -\mathrm{Cov}\left(\hat{\beta}_1\bar{X}, \hat{\beta}_1\right) \\
&= -\bar{X}\mathrm{Cov}\left(\hat{\beta}_1, \hat{\beta}_1\right) \\
&= -\bar{X}V\left[\hat{\beta}_1\right] \\
&= -\frac{\bar{X}\sigma^2}{S_{XX}}.
\end{aligned}$$

All of the elements of the variance–covariance matrix have now been established, which constitutes a proof of the following theorem.

Theorem 1.4 The least squares estimators of the parameters β_0 and β_1 in the simple linear regression model have variance–covariance matrix

$$\begin{bmatrix} V[\hat{\beta}_0] & \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_0) & V[\hat{\beta}_1] \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} X_i^2/(nS_{XX}) & -\bar{X}/S_{XX} \\ -\bar{X}/S_{XX} & 1/S_{XX} \end{bmatrix} \sigma^2.$$

There are two important observations that can be made from Theorem 1.4. First, the elements of the variance–covariance matrix of $\hat{\beta}_0$ and $\hat{\beta}_1$ are a function of only the X_i values and the typically unknown population error variance σ^2; the values of $Y_1, Y_2, \ldots, Y_n$ do not play a role. Recall from Definition 1.1 that the independent variable observations $X_1, X_2, \ldots, X_n$ are assumed to be observed without error. Second, since $S_{XX} > 0$ because at least two of the X_i values are distinct, the population covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$ takes the opposite sign of $\bar{X}$. This provides an explanation of why $\hat{\beta}_0$ and $\hat{\beta}_1$ appeared to have negative covariance in the results of the 5000 simulated estimates plotted in Figure 1.12.

Example 1.6 Consider again the simple linear regression model

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

from Example 1.4 in which

- the population intercept is $\beta_0 = 1$,
- the population slope is $\beta_1 = 1/2$, and
- the error term ε has a $U(-1, 1)$ distribution.

The error term distribution has population mean zero, so this model satisfies the conditions of a simple linear regression model. Find the variance–covariance matrix for the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ associated with a single Monte Carlo replication of $n = 10$ data pairs. Assume that the X_i values are equally likely to be one of the integers $0, 1, 2, \ldots, 9$.

The R code that follows conducts a single replication of the Monte Carlo experiment. The results of this single replication were illustrated by the fitted regression line in Figure 1.9. Since the error terms are mutually independent $U(-1, 1)$ random variables and the population variance of a $U(a, b)$ random variable is $(b-a)^2/12$, the population variance of the error terms is $\sigma^2 = (1+1)^2/12 = 1/3$. Although the dependent variables are generated and stored in the vector `y`, they are not used in the calculation of the variance–covariance matrix.

```
n        = 10
beta0    = 1
beta1    = 1 / 2
sigma2   = 1 / 3
set.seed(100)

x = sample(0:9, n, replace = TRUE)
```

```
if (min(x) == max(x)) stop("All x values are equal")
y = beta0 + beta1 * x + runif(n, -1, 1)

sxx       = sum((x - mean(x)) ^ 2)
vcm       = matrix(nrow = 2, ncol = 2)
vcm[1, 1] = sum(x ^ 2) / (n * sxx)
vcm[1, 2] = vcm[2, 1] = - mean(x) / sxx
vcm[2, 2] = 1 / sxx
vcm       = vcm * sigma2
print(vcm)
```

The variance–covariance matrix for this single replication of the Monte Carlo simulation experiment, reported to four digits, is

$$\begin{bmatrix} V\left[\hat{\beta}_0\right] & \text{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) \\ \text{Cov}\left(\hat{\beta}_1, \hat{\beta}_0\right) & V\left[\hat{\beta}_1\right] \end{bmatrix} = \begin{bmatrix} 0.1211 & -0.02509 \\ -0.02509 & 0.007168 \end{bmatrix}.$$

If additional Monte Carlo simulation replications were made, this matrix would vary from one replication to the next because the X_i values vary from one replication to the next. Taking the square roots of the diagonal elements yields

$$\sqrt{V\left[\hat{\beta}_0\right]} = 0.3481 \qquad \text{and} \qquad \sqrt{V\left[\hat{\beta}_1\right]} = 0.0847,$$

which are estimates of the standard deviation of the intercept and slope of the regression line, often referred to as the *standard errors* of the estimated parameters. These two standard deviations are roughly in line with the spread of the histograms generated from the 5000 simulation replications depicted in Figure 1.11. The negative values of the off-diagonal elements of the variance–covariance matrix are consistent with the plot of 5000 simulated $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ values given in Figure 1.12.

So far we have found the expected values and the variance–covariance matrix of the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. But there is a lingering doubt as to whether better point estimators for β_0 and β_1 exist. An example of such a better point estimator would be an unbiased estimator of β_0 with a smaller population variance than the least squares estimator of β_0. This lingering doubt will be addressed in the next subsection.

1.5.4 Gauss–Markov Theorem

Recall from Theorem 1.3 that the least squares estimators for the slope and intercept of the regression line were expressed as linear combinations of the dependent variables:

$$\hat{\beta}_1 = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n$$

and

$$\hat{\beta}_0 = c_1 Y_1 + c_2 Y_2 + \cdots + c_n Y_n.$$

But are these linear combinations the best possible linear combinations for estimating β_1 and β_0? The Gauss–Markov theorem is used to show that these estimators have the minimum variance of all possible unbiased estimators which are linear combinations of the dependent variables. These

estimators are known as *Best Linear Unbiased Estimators*, typically abbreviated with the colorful acronym BLUE. The Venn diagram in Figure 1.13 might be helpful in categorizing the various types of estimators. The set L consists of all point estimators for the regression parameters β_0 and β_1 which can be expressed as linear combinations of the dependent variables $Y_1, Y_2, \ldots, Y_n$. The set U consists of all point estimators for the regression parameters β_0 and β_1 which are unbiased estimators of β_0 and β_1. The shaded intersection of L and U (that is, $L \cap U$) is all estimators which are both linear combinations of $Y_1, Y_2, \ldots, Y_n$ and unbiased. An example of an estimator of β_1 which is neither in L nor in U is Y_1^2. The Gauss–Markov theorem states that the least squares estimators have the smallest possible variance among all estimators in $L \cap U$.

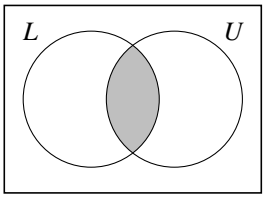

Figure 1.13: Venn diagram of sets L (linear combinations) and U (unbiased estimators).

Theorem 1.5 (Gauss–Markov theorem) The least squares estimators of β_0 and β_1 associated with a simple linear regression model have the smallest population variance among all unbiased estimators that can be expressed as a linear combination of the dependent variables.

Proof (partial proof) This proof will show that $\hat{\beta}_1$ has the smallest population variance among the class of all linear unbiased estimators for β_1. The proof for $\hat{\beta}_0$ is similar but left as an exercise for the reader. Let

$$\hat{\beta}_1 = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n$$

be the unbiased least squares estimator of the population slope β_1 from Theorem 1.3, where $a_i = \left(X_i - \bar{X}\right)/S_{XX}$ for $i = 1, 2, \ldots, n$. Consider another linear combination of the dependent variables which is also an unbiased estimator of β_1 that can be written as

$$\hat{\beta}_1' = k_1 Y_1 + k_2 Y_2 + \cdots + k_n Y_n$$

for some real-valued constants $k_1, k_2, \ldots, k_n$. Since $E[Y_i] = \beta_0 + \beta_1 X_i$, the expected value of $\hat{\beta}_1'$ is

$$\begin{aligned} E\left[\hat{\beta}_1'\right] &= E\left[\sum_{i=1}^{n} k_i Y_i\right] \\ &= \sum_{i=1}^{n} k_i E\left[Y_i\right] \\ &= \sum_{i=1}^{n} k_i \left(\beta_0 + \beta_1 X_i\right) \\ &= \beta_0 \sum_{i=1}^{n} k_i + \beta_1 \sum_{i=1}^{n} k_i X_i. \end{aligned}$$

Since $\hat{\beta}_1'$ is an unbiased estimator of β_1, $E\left[\hat{\beta}_1'\right] = \beta_1$. In order for this to be the case, the following conditions must hold:

$$\sum_{i=1}^{n} k_i = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} k_i X_i = 1.$$

These two conditions will be used in the last step of the derivation that follows. Now let $k_i = a_i + d_i$, for $i = 1, 2, \ldots, n$. We want to find the d_i values that meet the two conditions given above and minimize $V\left[\hat{\beta}_1'\right]$, which is

$$\begin{aligned}
V\left[\hat{\beta}_1'\right] &= V\left[\sum_{i=1}^{n} k_i Y_i\right] \\
&= \sum_{i=1}^{n} k_i^2 V\left[Y_i\right] \\
&= \sum_{i=1}^{n} k_i^2 \sigma^2 \\
&= \sigma^2 \sum_{i=1}^{n} (a_i + d_i)^2 \\
&= \sigma^2 \left[\sum_{i=1}^{n} a_i^2 + \sum_{i=1}^{n} d_i^2 + 2\sum_{i=1}^{n} a_i d_i\right] \\
&= V\left[\hat{\beta}_1\right] + \sigma^2 \sum_{i=1}^{n} d_i^2 + 2\sigma^2 \sum_{i=1}^{n} a_i d_i \\
&= V\left[\hat{\beta}_1\right] + \sigma^2 \sum_{i=1}^{n} d_i^2 + 2\sigma^2 \sum_{i=1}^{n} a_i (k_i - a_i) \\
&= V\left[\hat{\beta}_1\right] + \sigma^2 \sum_{i=1}^{n} d_i^2 + 2\sigma^2 \left(\sum_{i=1}^{n} a_i k_i - \sum_{i=1}^{n} a_i^2\right) \\
&= V\left[\hat{\beta}_1\right] + \sigma^2 \sum_{i=1}^{n} d_i^2 + 2\sigma^2 \left(\sum_{i=1}^{n} k_i \cdot \frac{X_i - \bar{X}}{S_{XX}} - \frac{1}{S_{XX}}\right) \\
&= V\left[\hat{\beta}_1\right] + \sigma^2 \sum_{i=1}^{n} d_i^2 + 2\sigma^2 \left(\frac{\sum_{i=1}^{n} k_i X_i - \bar{X}\sum_{i=1}^{n} k_i - 1}{S_{XX}}\right) \\
&= V\left[\hat{\beta}_1\right] + \sigma^2 \sum_{i=1}^{n} d_i^2.
\end{aligned}$$

In order to minimize $V\left[\hat{\beta}_1'\right]$ the d_i values should be selected to minimize $\sum_{i=1}^{n} d_i^2$. This sum of squares is minimized when $d_1 = d_2 = \cdots = d_n = 0$. Therefore, the least squares estimator $\hat{\beta}_1$, with coefficients $k_i = a_i$ for $i = 1, 2, \ldots, n$, has the smallest variance among all unbiased estimators that can be written as linear combinations of $Y_1, Y_2, \ldots, Y_n$ and is therefore a best linear unbiased estimator. □

The Gauss–Markov theorem indicates that the least squares estimators for β_0 and β_1 have minimal variance among all linear estimators. It does not indicate whether the least squares estimators for β_0 and β_1 have minimal variance among all estimators. The Gauss–Markov theorem extends to

the case of multiple linear regression in which there are several independent variables. The least squares estimators are also the best linear unbiased estimators in this case.

To review the results that have been introduced so far, the simple linear regression model

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

defines a linear statistical relationship between an independent variable X, observed without error, and a random dependent variable Y as given in Definition 1.1. The point estimators for β_1 and β_0 from n data pairs (X_1, Y_1), (X_2, Y_2), $\ldots$, (X_n, Y_n) using the least squares criterion are

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \qquad \text{and} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

as given in Theorem 1.1. The least squares estimators are unbiased estimators of their associated parameters because

$$E\left[\hat{\beta}_1\right] = \beta_1 \qquad \text{and} \qquad E\left[\hat{\beta}_0\right] = \beta_0$$

as given in Theorem 1.2. The least squares estimators of β_0 and β_1 can be expressed as linear combinations of $Y_1, Y_2, \ldots, Y_n$ as

$$\hat{\beta}_0 = c_1 Y_1 + c_2 Y_2 + \cdots + c_n Y_n \qquad \text{and} \qquad \hat{\beta}_1 = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n,$$

with coefficients $c_1, c_2, \ldots, c_n$ and $a_1, a_2, \ldots, a_n$ given in Theorem 1.3. The variance–covariance matrix of $\hat{\beta}_0$ and $\hat{\beta}_1$ is

$$\begin{bmatrix} V\left[\hat{\beta}_0\right] & \mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) \\ \mathrm{Cov}\left(\hat{\beta}_1, \hat{\beta}_0\right) & V\left[\hat{\beta}_1\right] \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} X_i^2 / (nS_{XX}) & -\bar{X}/S_{XX} \\ -\bar{X}/S_{XX} & 1/S_{XX} \end{bmatrix} \sigma^2$$

as given in Theorem 1.4. Finally, the Gauss–Markov theorem given in Theorem 1.5 states that the least squares estimators of β_0 and β_1 have the smallest population variance among all unbiased estimators that can be expressed as a linear combination of $Y_1, Y_2, \ldots, Y_n$.

The next section defines fitted values and residuals. Fitted values are the heights of the regression line associated with the observed values of the independent variable $X_1, X_2, \ldots, X_n$. The residuals are the vertical signed distances between the observed values of the dependent variable $Y_1, Y_2, \ldots, Y_n$ and the associated fitted values that fall on the regression line. Residuals play an analogous role to the error terms in the simple linear regression model.

1.6 Fitted Values and Residuals

The simple linear regression model

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

was introduced in the previous section as a linear statistical model for describing the relationship between an independent variable X and a dependent variable Y. Taking the expected value of both sides of this equation yields

$$E[Y] = \beta_0 + \beta_1 X$$

because $E[\varepsilon] = 0$ and X is a fixed value assumed to be observed without error, which are two key assumptions in Definition 1.1. When the population intercept β_0 and the population slope β_1 are replaced by their associated least squares point estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, the resulting estimated regression line is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

This estimated regression line is typically plotted on a scatterplot that contains the data pairs (X_1, Y_1), (X_2, Y_2), ... , (X_n, Y_n). Seeing the data pairs and the least squares regression line on the same plot often makes the visual assessment of linearity easier. For any value X in which the simple linear regression model is valid, $\hat{Y}$ is the point estimator for the value of the dependent variable based on the data pairs and associated estimated regression line. This equation can be rewritten for the particular values of the independent variable collected as

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

for $i = 1, 2, \ldots, n$. The value $\hat{Y}_i$ is known as the *fitted value* associated with data pair i, for $i = 1, 2, \ldots, n$. When $\hat{Y}_i \neq Y_i$, which is almost always the case in applications, the fitted value does not fall on the estimated regression line; when $\hat{Y}_i = Y_i$, the fitted value falls on the estimated regression line. The next example illustrates the notion of fitted values for the sales data set.

Example 1.7 Consider the sales data set from Example 1.3 with just $n = 3$ data pairs:

$$(X_1, Y_1) = (6, 2), \qquad (X_2, Y_2) = (8, 9), \qquad (X_3, Y_3) = (2, 2).$$

Find the fitted values $\hat{Y}_1$, $\hat{Y}_2$, and $\hat{Y}_3$ associated with the least squares regression line.

From Examples 1.3 and 1.5, the point estimates for the population intercept and population slope are

$$\hat{\beta}_0 = -1 \qquad \text{and} \qquad \hat{\beta}_1 = 1.$$

Hence, the estimated regression line is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$, or

$$\hat{Y} = -1 + X,$$

which is plotted along with the scatterplot of the data pairs in Figure 1.14. So calculating the fitted values is just a matter of using the X_i values as arguments in the estimated regression line:

$$\begin{aligned}
\hat{Y}_1 &= -1 + X_1 = -1 + 6 = 5 \quad &\Rightarrow \quad (X_1, \hat{Y}_1) &= (6, 5) \\
\hat{Y}_2 &= -1 + X_2 = -1 + 8 = 7 \quad &\Rightarrow \quad (X_2, \hat{Y}_2) &= (8, 7) \\
\hat{Y}_3 &= -1 + X_3 = -1 + 2 = 1 \quad &\Rightarrow \quad (X_3, \hat{Y}_3) &= (2, 1).
\end{aligned}$$

The fitted values are also plotted as points that lie on the estimated regression line in Figure 1.14. Recall from the previous section that the fitted least squares line is the line which minimizes the sum of the squares of the lengths of the vertical dashed lines which connect the data pair with its associated fitted value. The fitted values are calculated and stored in a component named `fitted` in the list returned by the R `lm` function. The R code below confirms the fitted values calculated above by hand.

```
x = c(6, 8, 2)
y = c(2, 9, 2)
lm(y ~ x)$fitted
```

The spread of the data pair (X_i, Y_i) from the fitted regression line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ is reflected in the vertical signed distance between the data pair (X_i, Y_i) and the associated fitted value $\left(X_i, \hat{Y}_i\right)$, These signed distances are known as the *residuals*, and are defined by

$$e_i = Y_i - \hat{Y}_i$$

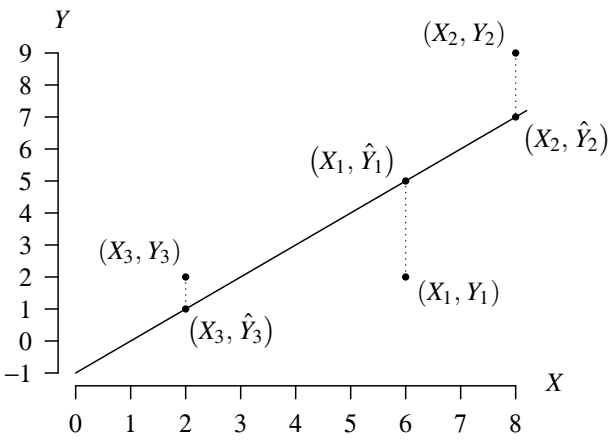

Figure 1.14: A scatterplot of the sales data pairs with the fitted values.

for $i = 1, 2, \ldots, n$. Data pairs that fall above the regression line correspond to positive residuals; data pairs that fall below the regression line correspond to negative residuals. The least squares approach used so far in estimating the intercept and slope of the regression line is a matter of finding the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimize the sum of the squares of the residuals. In other words, minimize

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

The fitted values and residuals are formally defined next.

Definition 1.2 Let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the least squares estimators of the parameters β_0 and β_1 in the simple linear regression model with data pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. The *fitted value* associated with the ith data pair (X_i, Y_i) is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, for $i = 1, 2, \ldots, n$. The *residual* associated with ith data pair (X_i, Y_i) is $e_i = Y_i - \hat{Y}_i$, for $i = 1, 2, \ldots, n$.

Choosing to use the *vertical distance* between the observed value of the dependent variable and the regression line in the definition of the residual was based on the fact that the values of the independent variable $X_1, X_2, \ldots, X_n$ are assumed to be observed without error in Definition 1.1. The mathematics associated with simple linear regression changes substantially if both X and Y are considered to be random variables.

A subtle but important distinction should be drawn between the model error term ε_i for data pair i and the residual e_i for data pair i. The model error terms are defined by

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

for $i = 1, 2, \ldots, n$, and represent the vertical distances between the observed dependent variable Y_i and the *true* (population) regression line $Y = \beta_0 + \beta_1 X$. The simple linear regression model assumes that $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are mutually independent random variables. In nearly all applications, however, β_0 and β_1 are unknown. This means that for a particular data set, these model error terms are also unknown. On the other hand, the residuals are defined by

$$e_i = Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)$$

for $i = 1, 2, \ldots, n$, and represent the error for data pair i when compared to the *estimated* regression line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$, which is calculated from the n data pairs. Thus, $\hat{\varepsilon}_i = e_i$, for $i = 1, 2, \ldots, n$. The $e_1, e_2, \ldots, e_n$ values are *not* mutually independent random variables because they must sum to zero. (This will be proven subsequently in Theorem 1.6.) For a particular data set, these residuals are known. The residuals are calculated for the sales data next.

Example 1.8 Consider again the sales data set from Example 1.3 with $n = 3$ data pairs:

$$(X_1, Y_1) = (6, 2) \qquad (X_2, Y_2) = (8, 9) \qquad (X_3, Y_3) = (2, 2).$$

Calculate the residuals e_1, e_2, and e_3 associated with the least squares regression line and display them on a scatterplot that includes the regression line.

Table 1.2 contains the calculations required to calculate the residuals and their squares. The sum of the squared residuals for these data pairs is

$$S = \sum_{i=1}^{3} e_i^2 = (-3)^2 + 2^2 + 1^2 = 9 + 4 + 1 = 14.$$

This total is consistent with the sum of the areas of the squares from Figure 1.7. The data pairs were handpicked in this example to make the residuals all integers. This will not be the case in nearly all applications of simple linear regression. This value for S which is associated with the estimated regression line is the smallest possible value for the sum of squared residuals. Any other line will be associated with a larger sum of squared residuals.

Figure 1.15 shows the residuals e_1, e_2, and e_3 along with the data pairs and the estimated regression line. Unless all of the data pairs fall in a line (which would correspond to $S = 0$), there will always be one or more data values falling above the line and one or more data values falling below the line.

The values of the residuals are stored in a component named `residuals` in the list returned by the R `lm` function. The R code below calculates and displays the residuals that were calculated by hand and displayed in Table 1.2.

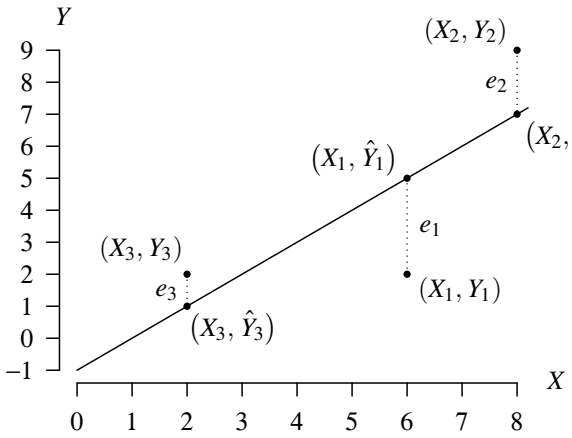

Figure 1.15: A scatterplot of the sales data pairs with the fitted values and residuals.

Observation number i	Number of sales X_i	Total revenue Y_i	Fitted value $\hat{Y}_i$	Residual $e_i = Y_i - \hat{Y}_i$	Squared residual e_i^2
1	6	2	5	-3	9
2	8	9	7	2	4
3	2	2	1	1	1
Sum	16	13	13	0	14

Table 1.2: Data pairs, fitted values, residuals, and squared residuals.

```
x = c(6, 8, 2)
y = c(2, 9, 2)
lm(y ~ x)$residuals
```

A close inspection of the entries in Table 1.2 reveals that there are some curious outcomes that occur, such as

$$\sum_{i=1}^{n} e_i = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i.$$

In other words, (*a*) the sum of the residuals is zero, and (*b*) the sum of the observed values of the dependent variable equals the sum of the fitted values. These were not just a matter of coincidence. The following theorem confirms that these relationships, along with a few other relationships, are true in general.

Theorem 1.6 Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be n data pairs associated with the simple linear regression model

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

Using the notation from Definition 1.2, the fitted values are $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n$ and the residuals are $e_1, e_2, \ldots, e_n$. Then

- $\sum_{i=1}^{n} e_i = 0,$
- $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i,$
- $\sum_{i=1}^{n} X_i e_i = 0,$
- $\sum_{i=1}^{n} \hat{Y}_i e_i = 0,$
- $\left(\bar{X}, \bar{Y}\right)$ is a point that lies on the estimated regression line.

Proof Each of the five results will be proven individually.

- Since $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ from Theorem 1.1, the sum of the residuals is

$$\begin{aligned}
\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right) \\
&= \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) \\
&= \sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} X_i \\
&= \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} Y_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i \\
&= 0.
\end{aligned}$$

- Since $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ from Theorem 1.1, the sum of the fitted values is

$$\begin{aligned}
\sum_{i=1}^{n} \hat{Y}_i &= \sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right) \\
&= n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i \\
&= n\left(\bar{Y} - \hat{\beta}_1 \bar{X}\right) + \hat{\beta}_1 \sum_{i=1}^{n} X_i \\
&= \sum_{i=1}^{n} Y_i.
\end{aligned}$$

  Thus, the sum of the values of the dependent variable always equals the sum of the fitted values.

- The sum of the products of the independent variables and residuals is

$$\begin{aligned}
\sum_{i=1}^{n} X_i e_i &= \sum_{i=1}^{n} X_i \left(Y_i - \hat{Y}_i\right) \\
&= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \hat{Y}_i \\
&= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right) \\
&= \sum_{i=1}^{n} X_i Y_i - \hat{\beta}_0 \sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \\
&= 0.
\end{aligned}$$

  The final step uses the second normal equation from Theorem 1.1.

- Using the first and third result in this theorem, the sum of the products of the fitted values and residuals is

$$\sum_{i=1}^{n} \hat{Y}_i e_i = \sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right) e_i = \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i e_i = \hat{\beta}_0 \cdot 0 + \hat{\beta}_1 \cdot 0 = 0.$$

- The first normal equation from Theorem 1.1 is

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i.$$

Dividing both sides by n,

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{X} = \bar{Y},$$

which indicates that the point $(\bar{X}, \bar{Y})$ lies on the estimated regression line. □

These five results from Theorem 1.6 will be illustrated for the sales data in the example that follows.

Example 1.9 Calculate the quantities given in Theorem 1.6 for the $n = 3$ data pairs from the sales data set from Example 1.3:

$$(X_1, Y_1) = (6, 2) \qquad (X_2, Y_2) = (8, 9) \qquad (X_3, Y_3) = (2, 2).$$

From Examples 1.3 and 1.5, the point estimate for the intercept is $\hat{\beta}_0 = -1$ and the point estimate for the slope is $\hat{\beta}_1 = 1$. Table 1.3 contains the calculations necessary to illustrate the results given in Theorem 1.6. More specifically,

- $\sum_{i=1}^{3} e_i = 0,$
- $\sum_{i=1}^{3} Y_i = \sum_{i=1}^{3} \hat{Y}_i = 13,$
- $\sum_{i=1}^{3} X_i e_i = 0,$
- $\sum_{i=1}^{3} \hat{Y}_i e_i = 0.$

Finally, the point $\left(\bar{X}, \bar{Y}\right) = (16/3, 13/3)$ lies on the estimated regression line $\hat{Y} = -1 + X$, as illustrated in Figure 1.16.

i	X_i	Y_i	$\hat{Y}_i$	e_i	e_i^2	$X_i e_i$	$\hat{Y}_i e_i$
1	6	2	5	−3	9	−18	−15
2	8	9	7	2	4	16	14
3	2	2	1	1	1	2	1
Sum	16	13	13	0	14	0	0

Table 1.3: Calculation of quantities from Theorem 1.6.

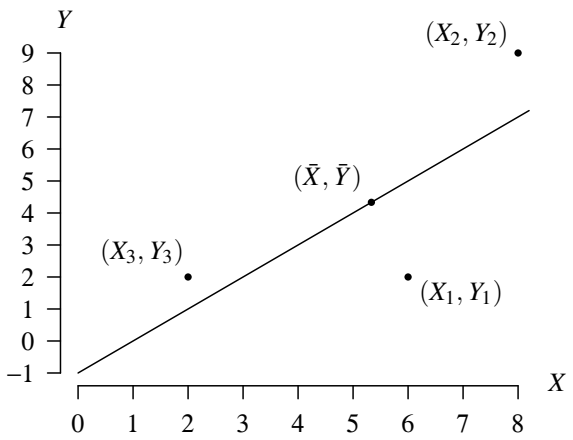

Figure 1.16: The point $(\bar{X}, \bar{Y})$ falls on the estimated regression line.

1.7 Estimating the Variance of the Error Terms

The emphasis so far has been focused on the estimation of the intercept and slope of the regression line. While $\hat{\beta}_0$ and $\hat{\beta}_1$ are the most critical parameters in most applications of a simple linear regression model, there is another parameter, the population variance of the error terms σ^2, which should also be estimated from the data pairs.

To establish a foundation for the estimation of σ^2, assume *for this paragraph only* that there is a univariate, rather than a bivariate, sample of values denoted by $X_1, X_2, \ldots, X_n$. These will not be fixed values observed without error as they were in regression modeling. It is assumed that these values constitute a random sample from a population that has finite population mean μ and finite population variance σ^2. The goal in this paragraph is to estimate σ^2 as a function of the data values. If the population mean μ is known (which is rare in practice), then an unbiased estimator of σ^2 is

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2.$$

If the first $n-1$ deviations between the sample values and the population mean $X_1-\mu$, $X_2-\mu$, $\ldots$, $X_{n-1}-\mu$ were known, the final deviation, $X_n-\mu$, would be free to take on any value. It is in this sense that the sum of squares

$$\sum_{i=1}^{n}(X_i-\mu)^2$$

is said to have n "degrees of freedom." It is common practice in statistics to divide a sum of squares by its degrees of freedom to arrive at a point estimator. In this particular instance, dividing by n makes the point estimator an unbiased estimator of σ^2. The problem that arises more often in practice is to estimate σ^2 when μ is unknown. An unbiased estimator of σ^2 in this case is the sample variance

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2,$$

which is typically denoted by S^2 by statisticians. There are three reasons why the term outside of the summation has $n-1$ in the denominator. The first reason is that this is the appropriate term so

that this estimator is an unbiased estimator of σ^2. This can be stated as $E\left[S^2\right] = \sigma^2$. The second reason is that one can't estimate the dispersion of a distribution from a single data value, so the sample variance is undefined when $n = 1$. The third reason is that the sum of squares has $n-1$ degrees of freedom. One degree of freedom is lost because the sample mean $\bar{X}$ is used to estimate the population mean μ. If the first $n-1$ deviations between the sample values and the sample mean $X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_{n-1} - \bar{X}$ were known, the final deviation, $X_n - \bar{X}$, could be calculated from the other $n-1$ values because

$$\sum_{i=1}^{n} \left(X_i - \bar{X}\right) = \sum_{i=1}^{n} X_i - n\bar{X} = 0.$$

It is in this sense that the sum of squares

$$\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

is said to have $n-1$ degrees of freedom. This ends the discussion of degrees of freedom for a univariate data set.

We now return to the problem of estimating σ^2 in simple linear regression. The independent variables $X_1, X_2, \ldots, X_n$ are once again assumed to be fixed values observed without error as they have been throughout this chapter. Based on the fact that the error terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ in the simple linear regression model are assumed to be mutually independent and identically distributed random variables, each with population mean 0 and finite population variance σ^2, the population variance of the error terms can be estimated with the unbiased estimator

$$\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

if β_0 and β_1 were known. But in practice, the two parameters β_0 and β_1 are estimated from the data pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, so two degrees of freedom are lost and an appropriate point estimator for the population variance σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2.$$

It is important that the population variance of the error terms σ^2 remain constant over the range of X values in which the simple linear regression model is appropriate. One tool for visually assessing this assumption is a scatterplot of the data pairs with the estimated regression line superimposed.

The point estimator for σ^2 when β_0 and β_1 are estimated from the data pairs involves the sum of squares of the residuals, and this is often abbreviated as *SSE*, for *sum of squares for error*:

$$SSE = \sum_{i=1}^{n} e_i^2,$$

which is also known as the *error sum of squares*, *residual sum of squares*, and *sum of squares due to error*. When this quantity is divided by its degrees of freedom, it is known as the *mean square error*, which is abbreviated by *MSE*:

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2.$$

Some good news is provided by the next result, which states that $MSE = \hat{\sigma}^2$ is an unbiased estimator of σ^2.

Theorem 1.7 If $e_1, e_2, \ldots, e_n$ are the residuals in a simple linear regression model, then the point estimator

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

is an unbiased estimator of σ^2.

Proof The simple linear regression model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

for $i = 1, 2, \ldots, n$. Summing both sides of this equation and dividing by n yields

$$\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{\varepsilon}.$$

Taking the difference between the previous two equations results in

$$Y_i - \bar{Y} = \beta_1 (X_i - \bar{X}) + \varepsilon_i - \bar{\varepsilon} \tag{1}$$

for $i = 1, 2, \ldots, n$. The definition of the residual associated with data pair i is

$$e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

for $i = 1, 2, \ldots, n$. Recognizing that the residuals sum to zero via Theorem 1.6, summing both sides of this equation, and dividing by n yields

$$0 = \bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X}.$$

Taking the difference between the previous two equations results in

$$e_i = Y_i - \bar{Y} - \hat{\beta}_1 (X_i - \bar{X}) \tag{2}$$

for $i = 1, 2, \ldots, n$. Substituting the right-hand side of equation (1) for $Y_i - \bar{Y}$ in equation (2) gives

$$\begin{aligned} e_i &= \beta_1 \left(X_i - \bar{X}\right) + \varepsilon_i - \bar{\varepsilon} - \hat{\beta}_1 \left(X_i - \bar{X}\right) \\ &= \left(\beta_1 - \hat{\beta}_1\right)\left(X_i - \bar{X}\right) + (\varepsilon_i - \bar{\varepsilon}) \end{aligned}$$

for $i = 1, 2, \ldots, n$. Squaring both sides of this equation and summing gives

$$\sum_{i=1}^{n} e_i^2 = \left(\hat{\beta}_1 - \beta_1\right)^2 \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 - 2\left(\hat{\beta}_1 - \beta_1\right) \sum_{i=1}^{n} \left(X_i - \bar{X}\right)(\varepsilon_i - \bar{\varepsilon}) + \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2.$$

Taking into account that the X_i values are assumed to be fixed constants in a simple linear regression model, the expected value of both sides of this equation is

$$E\left[\sum_{i=1}^{n} e_i^2\right] = E\left[\left(\hat{\beta}_1 - \beta_1\right)^2\right] \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 - 2E\left[\left(\hat{\beta}_1 - \beta_1\right) \sum_{i=1}^{n} \left(X_i - \bar{X}\right)(\varepsilon_i - \bar{\varepsilon})\right]$$

$$+ E\left[\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2\right]. \tag{3}$$

There are three terms on the right-hand side of equation (3). Each term will be considered separately. The first term contains $E\left[\left(\hat{\beta}_1 - \beta_1\right)^2\right]$, which is an expression for the population variance of $\hat{\beta}_1$ because $\hat{\beta}_1$ is an unbiased estimator for β_1 via Theorem 1.2. This population variance is the lower-right entry of the variance–covariance matrix given in Theorem 1.4. So the first term on the right-hand side of equation (3) reduces to

$$E\left[\left(\hat{\beta}_1 - \beta_1\right)^2\right]\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = V\left[\hat{\beta}_1\right]\cdot S_{XX} = \frac{\sigma^2}{S_{XX}}\cdot S_{XX} = \sigma^2.$$

Before considering the second term on the right-hand side of equation (3), recall from Theorem 1.3 that $\hat{\beta}_1$ can be written as a linear combination of the observations of the dependent variable $Y_1, Y_2, \ldots, Y_n$ as $\hat{\beta}_1 = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n$, where $a_i = (X_i - \bar{X})/S_{XX}$ for $i = 1, 2, \ldots, n$. So an expression for the least squares point estimator of β_1 can be written as

$$\begin{aligned}
\hat{\beta}_1 &= \sum_{i=1}^{n} a_iY_i \\
&= \sum_{i=1}^{n} a_i(\beta_0 + \beta_1X_i + \varepsilon_i) \\
&= \beta_0\sum_{i=1}^{n} a_i + \beta_1\sum_{i=1}^{n} a_iX_i + \sum_{i=1}^{n} a_i\varepsilon_i \\
&= \beta_1 + \sum_{i=1}^{n} a_i\varepsilon_i
\end{aligned}$$

via Theorem 1.3. Temporarily ignoring the -2 coefficient on the second term in equation (3) and using the fact that $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are mutually independent random variables with population mean zero and population variance σ^2, the expected value in the second term on the right-hand side of equation (3) is

$$\begin{aligned}
E\left[\left(\hat{\beta}_1 - \beta_1\right)\sum_{i=1}^{n}\left(X_i - \bar{X}\right)(\varepsilon_i - \bar{\varepsilon})\right] &= E\left[\left(\beta_1 + \sum_{i=1}^{n} a_i\varepsilon_i - \beta_1\right)\sum_{i=1}^{n}\left(X_i - \bar{X}\right)(\varepsilon_i - \bar{\varepsilon})\right] \\
&= E\left[\left(\sum_{i=1}^{n} a_i\varepsilon_i\right)\left(\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\varepsilon_i - \bar{\varepsilon}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\right)\right] \\
&= E\left[\frac{1}{S_{XX}}\left(\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\varepsilon_i\right)\left(\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\varepsilon_i\right)\right] \\
&= \frac{1}{S_{XX}}E\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\varepsilon_i^2 + \sum\sum_{i\neq j}\left(X_i - \bar{X}\right)\left(X_j - \bar{X}\right)\varepsilon_i\varepsilon_j\right] \\
&= \frac{1}{S_{XX}}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2E\left[\varepsilon_i^2\right] \\
&= \sigma^2.
\end{aligned}$$

Finally, consider the third term on the right-hand side of equation (3). Using (i) the shortcut formula for the population variance, (ii) the fact that the expected value operator E is a linear operator, (iii) the fact that the population variance of a sample mean comprised of mutually independent and identically distributed random variables is the ratio of the population variance to the sample size, and (iv) the fact that $E[\varepsilon_i] = 0$ for $i = 1, 2, \ldots, n$ and therefore $E[\bar{\varepsilon}] = 0$, the third term on the right-hand side of equation (3) is

$$\begin{aligned}
E\left[\sum_{i=1}^{n}\left(\varepsilon_i - \bar{\varepsilon}\right)^2\right] &= E\left[\sum_{i=1}^{n}\varepsilon_i^2 - 2\bar{\varepsilon}\sum_{i=1}^{n}\varepsilon_i + n\bar{\varepsilon}^2\right] \\
&= E\left[\sum_{i=1}^{n}\varepsilon_i^2 - 2n\bar{\varepsilon}^2 + n\bar{\varepsilon}^2\right] \\
&= E\left[\sum_{i=1}^{n}\varepsilon_i^2 - n\bar{\varepsilon}^2\right] \\
&= \sum_{i=1}^{n}E\left[\varepsilon_i^2\right] - nE\left[\bar{\varepsilon}^2\right] \\
&= \sum_{i=1}^{n}\left(V\left[\varepsilon_i\right] + E\left[\varepsilon_i\right]^2\right) - n\left(V\left[\bar{\varepsilon}\right] + E\left[\bar{\varepsilon}\right]^2\right) \\
&= \sum_{i=1}^{n}\sigma^2 - n\cdot\frac{\sigma^2}{n} \\
&= n\sigma^2 - \sigma^2 \\
&= (n-1)\sigma^2.
\end{aligned}$$

Combining the three terms together, equation (3) becomes

$$E\left[\sum_{i=1}^{n}e_i^2\right] = \sigma^2 - 2\sigma^2 + (n-1)\sigma^2 = (n-2)\sigma^2.$$

Dividing both sides of this equation by $n-2$ indicates that the *MSE* is an unbiased estimator of σ^2:

$$E\left[\frac{1}{n-2}\sum_{i=1}^{n}e_i^2\right] = \sigma^2. \qquad \square$$

To summarize, there are three parameters in a simple linear regression model: the population intercept β_0, the population slope β_1, and the population variance of the error terms σ^2. These parameters can be estimated from n data pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ by the least squares method. Theorem 1.2 indicates that the least squares point estimator $\hat{\beta}_0$ is an unbiased estimator of β_0 and the least squares point estimator $\hat{\beta}_1$ is an unbiased estimator of β_1. Theorem 1.7 indicates that the *MSE* is an unbiased estimator of σ^2. All three parameter estimators are on target on average. The next three examples illustrate the estimation of σ^2.

Example 1.10 Estimate the variance of the error terms σ^2 for the $n = 3$ data pairs from the sales data set in Example 1.3:

$$(X_1, Y_1) = (6, 2), \qquad (X_2, Y_2) = (8, 9), \qquad (X_3, Y_3) = (2, 2).$$

Using the calculations in Example 1.9, the estimated variance of the error terms is

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{3-2}(9+4+1) = 14.$$

The R code to compute the value of the point estimate for σ^2 is given below.

```
x = c(6, 8, 2)
y = c(2, 9, 2)
n = length(x)
fit = lm(y ~ x)
sum(fit$residuals ^ 2) / (n - 2)
```

The magnitude of the point estimate of σ^2 is a reflection of whether the data points are tightly clustered about the estimated regression line (for small values of $\hat{\sigma}^2$) or whether the data points stray significantly from the estimated regression line (for large values of $\hat{\sigma}^2$). In the previous example involving the sales data pairs, there is significant vertical deviation between the data points and the associated fitted values, as seen in Figure 1.15. The next example illustrates the case in which the data pairs are tightly clustered about the regression line.

Example 1.11 Scottish physicist James Forbes wanted to devise a technique to estimate the altitude above sea level without transporting a fragile mercury barometer to the location of interest. He knew that the altitude could be computed from the barometric pressure, with lower barometric pressures corresponding to higher altitudes. He wanted to see if the boiling point of water could be used as a surrogate to determine the barometric pressure. In the 1840's and 1850's, he gathered $n = 17$ data pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_{17}, Y_{17})$ from various locations at different altitudes in the Alps, where

X_i: the boiling point of water in degrees Fahrenheit at location i, and
Y_i: the adjusted barometric pressure in inches of mercury at location i,

for $i = 1, 2, \ldots, 17$. The data was published in an 1857 article in the *Transactions of the Royal Society of Edinburgh* titled "Further Experiments and Remarks on the Measurement of Heights and Boiling Point of Water." The $n = 17$ data pairs in Forbes' data set are shown in Table 1.4. Make a scatterplot of the data values to determine whether a simple linear regression model is appropriate. If it is an appropriate model, estimate the model parameters β_0, β_1, and σ^2.

A scatterplot of the data is plotted with the R commands given below.

```
x = c(194.5, 194.3, 197.9, 198.4, 199.4, 199.9, 200.9, 201.1, 201.4,
      201.3, 203.6, 204.6, 209.5, 208.6, 210.7, 211.9, 212.2)
y = c(20.79, 20.79, 22.40, 22.67, 23.15, 23.35, 23.89, 23.99, 24.02,
      24.01, 25.14, 26.57, 28.49, 27.76, 29.04, 29.88, 30.06)
plot(x, y)
```

The scatterplot is given in Figure 1.17. On the range of the independent variable X that was collected by Forbes, which is $194.3 \leq X \leq 212.2$, there appears to be a linear relationship between the boiling temperature X and the barometric pressure Y, so it

Boiling point	Barometric pressure
194.5	20.79
194.3	20.79
197.9	22.40
198.4	22.67
199.4	23.15
199.9	23.35
200.9	23.89
201.1	23.99
201.4	24.02
201.3	24.01
203.6	25.14
204.6	26.57
209.5	28.49
208.6	27.76
210.7	29.04
211.9	29.88
212.2	30.06

Table 1.4: Data pairs for Forbes' experiment.

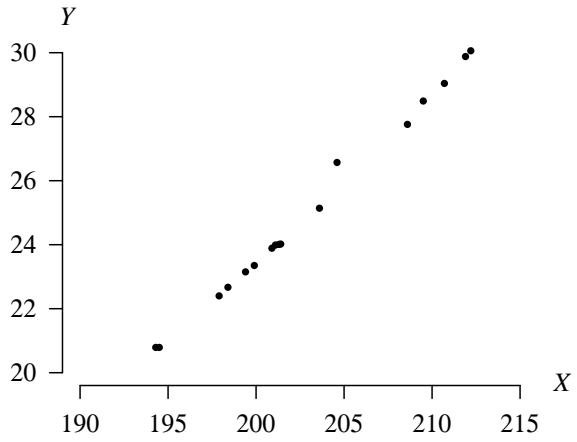

Figure 1.17: Scatterplot of the Forbes data.

is reasonable to proceed with fitting a simple linear regression model. The point that seems to stray slightly from the linear relationship, namely $(X_{12}, Y_{12}) = (204.6, 26.57)$, could be due to (i) random sampling variability, (ii) measurement error associated with the barometric pressure $Y_{12} = 26.57$, or (iii) measurement error associated with the boiling point $X_{12} = 204.6$ even though the simple linear regression model assumes that the boiling points are measured without error.

The R code below plots the fitted regression line on the scatterplot, which is shown in

Figure 1.18. Forbes' data pairs are in a built-in data frame named `forbes` that resides in the `MASS` package. The first column in `forbes` is named `bp` (for boiling point) and the second column is named `pres` (for barometric pressure).

```
library(MASS)
x = forbes$bp
y = forbes$pres
plot(x, y)
fit = lm(y ~ x)
abline(fit$coefficients)
```

Figure 1.18 confirms our conclusion about the linear relationship between X and Y from the scatterplot on the range of X values collected by Forbes. A simple linear regression model seems appropriate in this setting. The *additional* R commands that follow print the estimates for β_0, β_1, and σ^2 for Forbes' $n = 14$ data pairs.

```
n = length(x)
print(fit$coefficients)
print(sum(fit$residuals ^ 2) / (n - 2))
```

These yield the three unbiased point estimates for the simple linear regression model as

$$\hat{\beta}_0 = -81.0637 \qquad \hat{\beta}_1 = 0.5229 \qquad \hat{\sigma}^2 = 0.05421.$$

So the estimated regression line is

$$\hat{Y} = -81.0637 + 0.5229X.$$

Using the usual interpretation of the estimated intercept, when the boiling point of water is 0° Fahrenheit, the barometric pressure is estimated to be -81 inches of mercury. This is obviously an inappropriate conclusion and highlights the fact that this simple linear

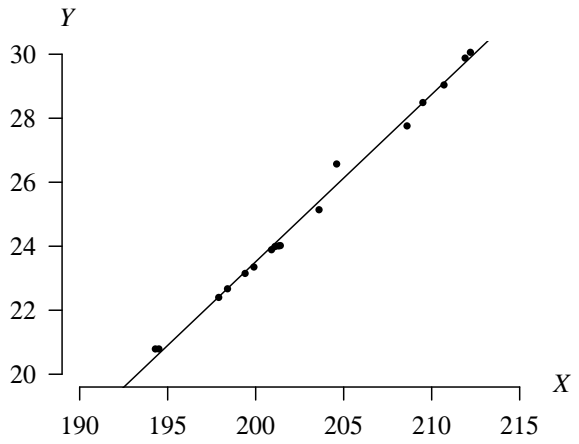

Figure 1.18: Scatterplot of the Forbes data with the estimated regression line.

regression model is only appropriate for a limited range of X values. The interpretation of $\hat{\beta}_1$, however, is meaningful. The barometric pressure increases by an estimated 0.5229 inches of mercury for every degree increase in the boiling point of water over the range of X values collected by Forbes. Finally, the estimated variance of the error terms, $\hat{\sigma}^2 = 0.05421$, is small (particularly relative to the estimated variance of the dependent variable observations, $S_{YY}/(n-1) = 9.12$, calculated with the *additional* R command `var(forbes$pres)`). This small estimated variance indicates that the data values are tightly clustered about the regression line. This is clearly the case in Figure 1.18.

The *additional* R command `plot(fit$residuals)` generates a plot of the residuals. Figure 1.19 shows the $n = 17$ residuals, along with a dashed horizontal line at a residual value of zero to show which observations fall above and below the regression line. (Notice that some of the X_i values are not in increasing order.) Six of the residuals are positive and 11 are negative. The reason that more residuals are negative is that the 12th data pair $(X_{12}, Y_{12}) = (204.6, 26.57)$ exerts a strong upwards "tug" on the fitted regression line, which is reflected in the plot of the residuals in Figure 1.19.

The non-symmetry in the values of $e_1, e_2, \ldots, e_{17}$ will also be reflected in a histogram of the residuals. Although $n = 17$ is a relatively small sample size for drawing a histogram and having a meaningful interpretation, one is displayed in Figure 1.20. This histogram can be generated with the *additional* R command `hist(fit$residuals)`. The histogram reveals a bell-shaped distribution for the residuals, with a single extreme value in the right-hand tail associated with the residual $e_{12} = 0.65$. This is consistent with the plot of the residuals in Figure 1.19.

In conclusion, the regression analysis seems to indicate that Forbes' experiment was a success. The barometric pressure does appear to be a function of the boiling point of water, and furthermore, the relationship between the two variables appears to be reasonably linear on the range of data pairs collected by Forbes. For a particular boiling point X that falls within that range of X values, the barometric pressure can be estimated

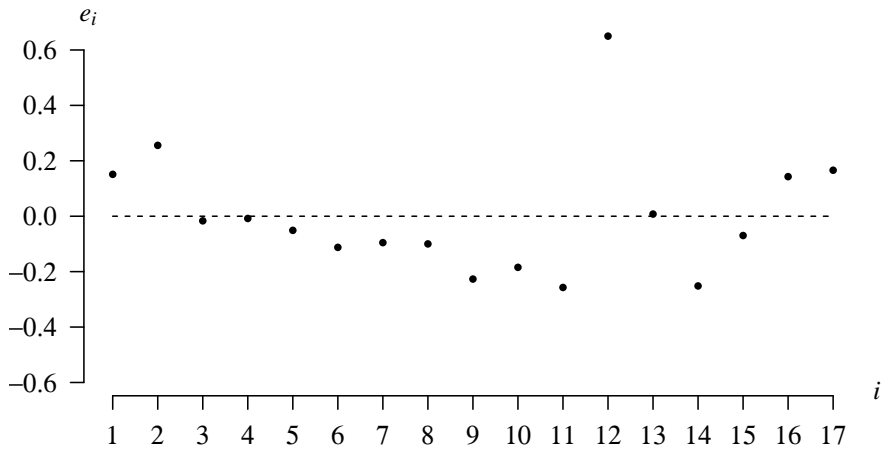

Figure 1.19: Residuals for the Forbes data.

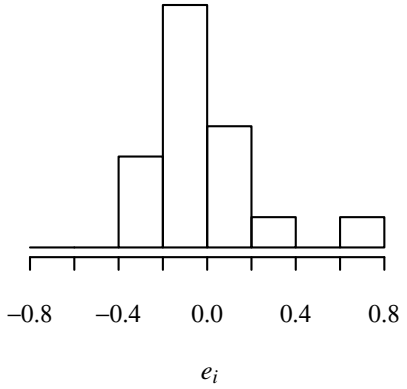

Figure 1.20: Histogram of the residuals for the Forbes data.

by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = -81.0637 + 0.5229X.$$

The altitude can, in turn, be estimated from the estimate of the barometric pressure provided by the regression analysis.

In the two previous examples, point estimates of the population variance of the error terms σ^2 were calculated. In the sales data example, the estimated error term variance $\hat{\sigma}^2 = 14$ indicated that the data pairs strayed a large distance from the estimated regression line, as illustrated in Figure 1.6. In the Forbes data set, the estimated error term variance $\hat{\sigma}^2 = 0.05421$ reflects data pairs that cluster closely to the estimated regression line, as illustrated in Figure 1.18. But these two examples involving individual data sets do not indicate anything about the *distribution* of $\hat{\sigma}^2$. The next example addresses this topic by extending the Monte Carlo simulation experiment from Example 1.4.

Example 1.12 Consider again the simple linear regression model

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

from Example 1.4, where

- the population intercept is $\beta_0 = 1$,
- the population slope is $\beta_1 = 1/2$, and
- the error term ε has a $U(-1, 1)$ distribution.

The focus in this example will be on the estimation of the probability distribution of $\hat{\sigma}^2$. Recall that the error term distribution has population mean zero and finite population variance, so it satisfies the conditions of a simple linear regression model from Definition 1.1. Conduct a Monte Carlo simulation with 5000 replications that estimates the probability distribution of the estimated variance of the error terms $\hat{\sigma}^2$ for $n = 10$ data pairs. Assume that the X_i values are equally likely to be one of the integers $0, 1, 2, \ldots, 9$.

The R code below conducts 5000 replications of the Monte Carlo experiment. The simulated regression model is fit by the `lm` function and the results are stored in the list

named `fit`. The component of the list named `fit$residuals` contains the residuals $e_1, e_2, \ldots, e_{10}$ for a particular simulation replication. The estimator of the variance of the error term in the simple linear regression model is given by the *MSE*:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2,$$

which is an unbiased estimator estimator of σ^2 by Theorem 1.7. The code generates a histogram of the 5000 estimates of the variance of the error terms.

```
nrep    = 5000
n       = 10
beta0   = 1
beta1   = 1 / 2
sig2hat = numeric(nrep)
set.seed(100)
for (i in 1:nrep) {
  x = sample(0:9, n, replace = TRUE)
  if (min(x) == max(x)) stop("All x values are equal")
  y = beta0 + beta1 * x + runif(n, -1, 1)
  fit = lm(y ~ x)
  sig2hat[i] = sum(fit$residuals ^ 2) / (n - 2)
}
hist(sig2hat)
```

The histogram that is produced by this Monte Carlo simulation is given in Figure 1.21. The histogram is centered around the population variance of the error terms

$$\sigma^2 = \frac{(1+1)^2}{12} = \frac{4}{12} = \frac{1}{3}$$

because the population variance of the $U(a, b)$ distribution is

$$\sigma^2 = \frac{(b-a)^2}{12},$$

where $a = -1$ and $b = 1$. So the Monte Carlo simulation supports the fact that $\hat{\sigma}^2$ is an unbiased estimator of σ^2 via Theorem 1.7. Although the distribution of the probability density function is bell-shaped, a careful examination of the histogram indicates that the right-hand tail of the distribution appears to be slightly heavier than the left-hand tail of the distribution. The probability density function of $\hat{\sigma}^2$ is not symmetric. This nonsymmetry is a universal result which extends beyond this particular simple linear regression model. This should not be surprising because the support of $\hat{\sigma}^2$ is the positive real numbers, unlike the support of $\hat{\beta}_0$ and $\hat{\beta}_1$ whose support is the entire real number line.

So the conclusions of the Monte Carlo simulation experiment are that (a) Theorem 1.7 is supported because the histogram in Figure 1.21 is centered around σ^2, and (b) the probability density function of $\hat{\sigma}^2$ is nearly bell-shaped with a slight bit of nonsymmetry.

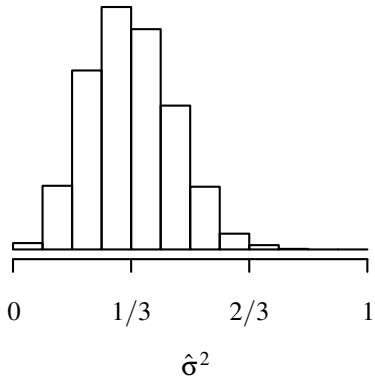

Figure 1.21: Histogram of the error term estimates for the Monte Carlo simulation.

Before leaving the topic of the estimation of σ^2 behind, consider the case of collecting just $n = 2$ data pairs (X_1, Y_1) and (X_2, Y_2), as illustrated in Figure 1.22. One of the assumptions associated with the observations in a simple linear regression model is that there are at least two distinct values of the independent variable observed. So when $n = 2$, it must be the case that $X_1 \neq X_2$. In this case, the least squares regression line will pass through the points (X_1, Y_1) and (X_2, Y_2). This means that the fitted values are identical to the data pairs, and hence, both residuals are zero. So the sum of squares for error is $SSE = e_1^2 + e_2^2 = 0$. But is an SSE of zero an appropriate estimate for the population variance of the spread of the values about the regression line? Can one conclude that this is really a deterministic relationship and any additional data pairs collected will fall on the fitted regression line? Certainly not, because it is not possible to draw that conclusion based on just two data pairs. A third data pair might fall on the regression line or fall significantly off of the regression line, as was the case with the sales data from Example 1.3. The unbiased estimator of σ^2 is undefined because of the $n-2$ in the denominator of the formula for $\hat{\sigma}^2$, as it should be. Two data pairs are adequate

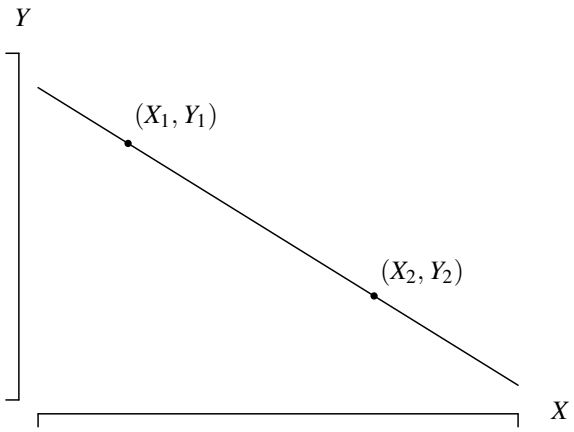

Figure 1.22: Scatterplot and estimated regression line for $n = 2$ data pairs.

for estimating the population slope and population intercept of the regression line, but they are not adequate for estimating σ^2. The mathematics and intuition are consistent in this setting.

1.8 Sums of Squares

Certain sums of squares play a key role in simple linear regression. This section considers three topics related to these sums of squares: (*a*) partitioning the total sum of squares, (*b*) defining and interpreting the coefficient of determination and the coefficient of correlation, and (*c*) displaying the sums of squares in an ANOVA table.

1.8.1 Partitioning the Total Sum of Squares

A topic that is closely related to fitted values and residuals is the partitioning of the total sum of squares. Figure 1.23 provides the geometric framework for the mathematical derivation provided next. There are only three points plotted in Figure 1.23. The first point plotted is (X_i, Y_i), which is a generic data pair. The other $n-1$ data pairs are not plotted in order to keep the figure uncluttered. The estimated regression line associated with the n data pairs, which happens to have a negative slope, is also plotted. The second point plotted is the fitted value $\left(X_i, \hat{Y}_i\right)$ associated with the ith data pair, which is located directly below data pair i and falls on the estimated regression line. The third point plotted is $(\bar{X}, \bar{Y})$, which, by Theorem 1.6, will always fall on the regression line.

Figure 1.23 provides a geometric proof of the relationship

$$Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i$$

for $i = 1, 2, \ldots, n$. The relationship can also be established algebraically by recognizing that the right-hand side of this equation can be determined by just adding and subtracting $\hat{Y}_i$ to the left-hand side of the equation. As will be stated and proved subsequently, squaring both sides of this equation and summing results in

$$\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2.$$

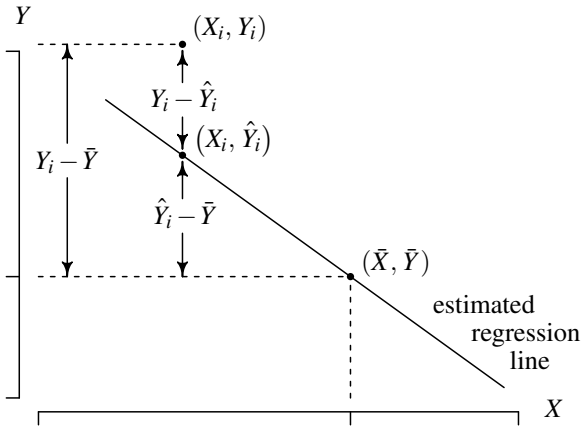

Figure 1.23: Partitioning the total sum of squares.

This equation involves three sums of squares that occur so often in regression analysis that they are given the abbreviations

$$SST = SSR + SSE,$$

where *SST* stands for total sum of squares, *SSR* stands for sum of squares for regression, and *SSE* stands for sum of squares for error. (The sum of squares for error has already been encountered in Theorem 1.7.) This equation expresses the total variation of the observed values of the dependent variable $Y_1, Y_2, \ldots, Y_n$ about their sample mean $\bar{Y}$ in *SST* as the sum of two sums of squares. The first term on the right-hand side, *SSR*, reflects the variation of the fitted values $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n$ about the sample mean $\bar{Y}$. The second term on the right-hand side, *SSE*, reflects the variation of the observed values $Y_1, Y_2, \ldots, Y_n$ about their associated fitted values $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n$. Since all three terms in this equation are sums of squares, all three terms are nonnegative. Notice that $SST/(n-1)$ is the sample variance of $Y_1, Y_2, \ldots, Y_n$.

The equation

$$SST = SSR + SSE$$

partitions *SST* into two pieces: *SSR*, which accounts for the total variability in $Y_1, Y_2, \ldots, Y_n$ that is accounted for by the regression line (that is, the linear relationship between X and Y), and *SSE*, which accounts for the remaining variability that is not associated with the regression line. This is why *SSR* measures the total variability in $Y_1, Y_2, \ldots, Y_n$ "explained" by the relationship between X and Y, whereas *SSE* measures the total variability in $Y_1, Y_2, \ldots, Y_n$ left "unexplained" by the relationship between X and Y. It is reasonable to think of *SSR* as measuring the "signal" associated with the linear relationship and *SSE* as measuring the "noise" associated with the linear relationship. The result is stated formally and proven next.

Theorem 1.8 Let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the least squares estimators of the parameters β_0 and β_1 in the simple linear regression model fitted to the data pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. Let $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ be the fitted value associated with data pair i, for $i = 1, 2, \ldots, n$. Let $\bar{Y}$ be the sample mean of $Y_1, Y_2, \ldots, Y_n$. Then

$$\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2,$$

or, equivalently,

$$SST = SSR + SSE.$$

Proof Beginning with $Y_i - \bar{Y}$, adding and subtracting $\hat{Y}_i$ gives

$$Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i$$

for $i = 1, 2, \ldots, n$. Grouping the two terms on the right-hand side of this equation as $\left(\hat{Y}_i - \bar{Y}\right)$ and $\left(Y_i - \hat{Y}_i\right)$, squaring both sides of the equation, and summing gives

$$\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + 2\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)\left(Y_i - \hat{Y}_i\right) + \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2.$$

The middle summation on the right-hand side of this equation is zero because

$$2\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)\left(Y_i - \hat{Y}_i\right) = 2\sum_{i=1}^{n} \hat{Y}_i\left(Y_i - \hat{Y}_i\right) - 2\bar{Y}\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right) = 2\sum_{i=1}^{n} \hat{Y}_i e_i - 2\bar{Y}\sum_{i=1}^{n} e_i = 0$$

by Theorem 1.6, which proves the result. □

1.8.2 Coefficients of Determination and Correlation

There are two measures that are helpful in assessing the degree of the linear relationship between X and Y in a simple linear regression model. The *coefficient of determination* and the *coefficient of correlation* are defined next. The thinking behind the way that the coefficient of determination $R^2 = SSR/SST$ is defined is as follows. The value of SST reflects the variability in $Y_1, Y_2, \ldots, Y_n$ when the values of the associated independent variables $X_1, X_2, \ldots, X_n$ are ignored. The value of SSE reflects the variability in $Y_1, Y_2, \ldots, Y_n$ when a fitted regression model uses $X_1, X_2, \ldots, X_n$ as predictors. Their difference, $SSR = SST - SSE$, reflects the reduction in variability associated with using the regression model. The ratio SSR/SST captures the fraction of that reduction in variability.

Definition 1.3 Let SST, SSR, and SSE for a simple linear regression model be defined as in Theorem 1.8. The *coefficient of determination* is

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

when $SST \neq 0$. The *coefficient of correlation* (a.k.a. the *sample correlation coefficient*) is

$$r = \pm\sqrt{R^2},$$

where the sign associated with r is positive (negative) when the slope of the estimated regression line is positive (negative).

The coefficient of determination R^2 is the fraction of the variation in $Y_1, Y_2, \ldots, Y_n$ about $\bar{Y}$ that is accounted for by the linear relationship between X and Y. Based on the result from Theorem 1.8, $SST = SSR + SSE$, the coefficient of determination must satisfy $0 \leq R^2 \leq 1$. Likewise, the coefficient of correlation must satisfy $-1 \leq r \leq 1$, which is true for all population and sample correlations.

Values of R^2 that are near 1 indicate that nearly all of the variation in $Y_1, Y_2, \ldots, Y_n$ about $\bar{Y}$ can be explained by the linear relationship between X and Y. This in turn implies that X is a useful predictor for Y. On the other hand, values of R^2 that are near 0 indicate that very little of the variation in $Y_1, Y_2, \ldots, Y_n$ about $\bar{Y}$ can be explained by the linear relationship between X and Y. This in turn implies that X is not a useful predictor for Y. It is in this sense that R^2 is a measure of the strength of the linear relationship between X and Y.

There are some important limitations associated with R^2 and r. First, it is important to remember that the linear relationship between X and Y might only be appropriate on a limited range of X values. Second, even a relatively large value of R^2 might not provide the precision necessary for a particular application. Third, regardless of the value of R^2, the scatterplot of the data pairs must always be inspected to see if a simple linear regression model is warranted. Both high and low values of R^2 can be associated with a strong *nonlinear* relationship between X and Y. Fourth, in the case in which the experimenter can control the values of $X_1, X_2, \ldots, X_n$, the magnitude of R^2 depends on the choices of the independent variables, which clouds its interpretation. Fifth, the usual interpretation of the coefficient of correlation r as an estimator of $\rho = \mathrm{Cov}(X, Y) / (\sigma_X \sigma_Y)$ is only appropriate when X and Y are random variables, which is not the case in simple linear regression because X is assumed to be observed without error.

It is a useful thought experiment to consider the scatterplots associated with the values of SST, SSR, and SSE at their extremes. These three extreme cases will be described in the next three paragraphs.

The first of these extreme cases is illustrated for $n = 7$ in Figure 1.24 in which

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} e_i^2 = 0.$$

The only way to achieve a sum of squares for error of zero is to have the data pairs (X_1, Y_1), (X_2, Y_2), $\ldots$, (X_n, Y_n) all fall on a line, which is the regression line. Using the result from Theorem 1.8 that $SST = SSR + SSE$, in this case $SST = SSR$, which implies that $R^2 = 1$. Therefore, *all* of the variation in $Y_1, Y_2, \ldots, Y_n$ is explained by the linear relationship between X and Y. In addition, $r = -1$ if the slope of the regression line is negative and $r = 1$ if the slope of the regression line is positive.

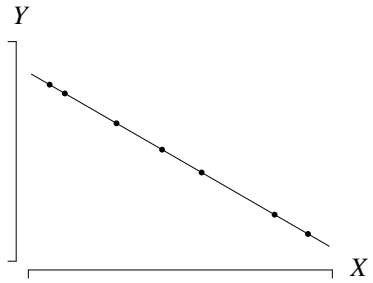

Figure 1.24: Data pairs with $SSE = 0$ and $\hat{\beta}_1 \neq 0$ (which implies that $SST = SSR$ and $R^2 = 1$).

The second of these extreme cases is illustrated for $n = 7$ in Figure 1.25 in which

$$SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 = 0.$$

The only way to achieve a sum of squares for regression of zero is to have an estimated regression line with slope zero. Using the result from Theorem 1.8 that $SST = SSR + SSE$, in this case $SST = SSE$, which implies that $R^2 = 0$. This means that *none* of the variation in $Y_1, Y_2, \ldots, Y_n$ is explained by the linear relationship between X and Y. In addition, $r = 0$.

The third of these extreme cases is illustrated for $n = 7$ in Figure 1.26 in which

$$SST = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = 0.$$

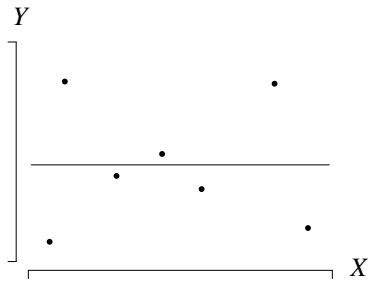

Figure 1.25: Data pairs with $SSR = 0$ (which implies that $SST = SSE$ and $R^2 = 0$).

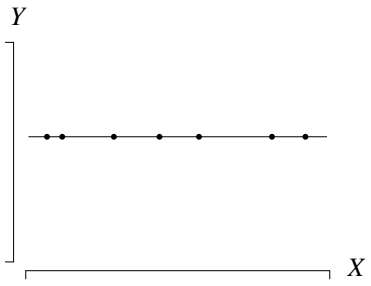

Figure 1.26: Data pairs with $SST = 0$ (which implies that $SSR = SSE = 0$ and R^2 is undefined).

The only way to achieve a total sum of squares of zero is to have an estimated regression line with slope zero and all points lying on the estimated regression line. Using the result from Theorem 1.8 that $SST = SSR + SSE$, in this case $SSR = SSE = 0$, and the coefficient of determination and coefficient of correlation are undefined because the denominator is zero.

Each of the sums of squares has an associated degrees of freedom. The total sum of squares

$$SST = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$$

has $n-1$ degrees of freedom for either of two reasons: (1) one degree of freedom is lost because $\bar{Y}$ is used to estimate the population mean, and (2) the terms in the summation above are subject to the one constraint—they must sum to zero. The sum of squares for regression

$$SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$$

has 1 degree of freedom because each of the $\hat{Y}_i$ values is calculated from the same regression line which has two degrees of freedom, but is subject to the additional constraint $\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right) = 0$ by Theorem 1.6. The sum of squares for error

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

has $n-2$ degrees of freedom for the reasons outlined just before Theorem 1.7.

An alternative definition for computing the coefficient of correlation r can save on computation time, as given in the following theorem.

Theorem 1.9 The coefficient of correlation r is

$$r = \hat{\beta}_1 \sqrt{\frac{S_{XX}}{S_{YY}}}.$$

Proof Recall from Definition 1.3 that the coefficient of correlation is

$$r = \pm \sqrt{\frac{SSR}{SST}},$$

where the sign associated with r is the same as the sign of $\hat{\beta}_1$. Since $\sum_{i=1}^n \hat{Y}_i = \sum_{i=1}^n Y_i$ by Theorem 1.6, this can be rewritten as

$$\begin{aligned}
r &= \pm\sqrt{\frac{SSR}{SST}} \\
&= \pm\sqrt{\frac{\sum_{i=1}^n \left(\hat{Y}_i - \bar{Y}\right)^2}{S_{YY}}} \\
&= \pm\sqrt{\frac{\sum_{i=1}^n \hat{Y}_i^2 - 2\bar{Y}\sum_{i=1}^n \hat{Y}_i + n\bar{Y}^2}{S_{YY}}} \\
&= \pm\sqrt{\frac{\sum_{i=1}^n \hat{Y}_i^2 - 2\bar{Y}\sum_{i=1}^n Y_i + n\bar{Y}^2}{S_{YY}}} \\
&= \pm\sqrt{\frac{\sum_{i=1}^n \hat{Y}_i^2 - n\bar{Y}^2}{S_{YY}}} \\
&= \pm\sqrt{\frac{\sum_{i=1}^n \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)^2 - n\bar{Y}^2}{S_{YY}}} \\
&= \pm\sqrt{\frac{\sum_{i=1}^n \left(\bar{Y} - \hat{\beta}_1\bar{X} + \hat{\beta}_1 X_i\right)^2 - n\bar{Y}^2}{S_{YY}}} \\
&= \pm\sqrt{\frac{n\bar{Y}^2 + 2\bar{Y}\hat{\beta}_1\sum_{i=1}^n (X_i - \bar{X}) + \hat{\beta}_1^2\sum_{i=1}^n (X_i - \bar{X})^2 - n\bar{Y}^2}{S_{YY}}} \\
&= \pm\sqrt{\frac{\hat{\beta}_1^2\sum_{i=1}^n (X_i - \bar{X})^2}{S_{YY}}} \\
&= \hat{\beta}_1\sqrt{\frac{S_{XX}}{S_{YY}}},
\end{aligned}$$

which proves the theorem. □

1.8.3 The ANOVA Table

The three sums of squares for the simple linear regression model and their associated degrees of freedom can be summarized in an analysis of variance (ANOVA) table. The four columns in the generic ANOVA table shown in Table 1.5 are (a) the source of variation, (b) the sum of squares, (c) the degrees of freedom, and (d) the mean square. The sums of squares and the degrees of freedom add to the values in the row labeled "Total". The mean square is the ratio of the sum of

Source	SS	df	MS
Regression	SSR	1	MSR
Error	SSE	$n-2$	MSE
Total	SST	$n-1$	

Table 1.5: Partial ANOVA table for simple linear regression.

squares to the associated degrees of freedom. The regression mean square is $MSR = SSR/1 = SSR$. The mean square error is $MSE = SSE/(n-2)$. The mean square entries do not add. Tradition dictates that the mean square associated with SST is not reported in an ANOVA table, but it does have meaning as the sample variance of $Y_1, Y_2, \ldots, Y_n$. More information on how the ANOVA table can be used for hypothesis testing concerning the population slope β_0 by adding a fifth column to the ANOVA table will be given in the next chapter.

Example 1.13 Consider the Forbes data set from Example 1.11 in which the independent variable X is the boiling point of water in degrees Fahrenheit and the dependent variable Y is the adjusted barometric pressure in inches of mercury. There are $n = 17$ data pairs collected from various locations. Calculate the three sums of squares (SST, SSR, and SSE), show that Theorem 1.8 is satisfied, calculate R^2 and r, and present the results in an ANOVA table.

The scatterplot with the estimated regression line superimposed from Example 1.11 is reproduced in Figure 1.27. The R commands below calculate the three sums of squares.

```
library(MASS)
x   = forbes$bp
y   = forbes$pres
fit = lm(y ~ x)
sst = sum((y - mean(y)) ^ 2)
ssr = sum((fit$fitted - mean(y)) ^ 2)
sse = sum(fit$residuals ^ 2)
print(c(sst, ssr, sse))
```

These commands result in the following values for the three sums of squares:

$$SST = 145.9378 \qquad SSR = 145.1246 \qquad SSE = 0.8131.$$

Ignoring the roundoff error in the fourth digit after the decimal point, these values sat-

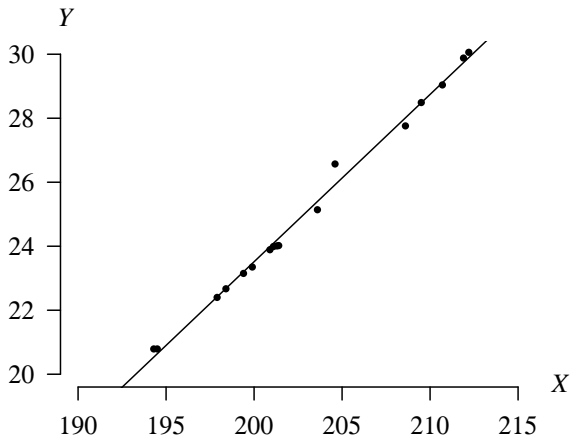

Figure 1.27: Scatterplot of the Forbes data with the estimated regression line.

isfy the result in Theorem 1.8:

$$SST = SSR + SSE.$$

The fact that *SSR* is more than two orders of magnitude greater than *SSE* indicates that there is much more of the total variation in $Y_1, Y_2, \ldots, Y_n$ that is explained by the relationship between X and Y than unexplained. This interpretation is consistent with the scatterplot and estimated regression line given in Figure 1.27.

The value of the coefficient of determination and the coefficient of correlation for this data set can be calculated by the *additional* R commands

```
R2 = ssr / sst
r  = sign(fit$coefficients[2]) * sqrt(R2)
print(c(R2, r))
```

via Definition 1.3 or

```
sxx = sum((x - mean(x)) ^ 2)
syy = sum((y - mean(y)) ^ 2)
R2  = ssr / sst
r   = fit$coefficients[2] * sqrt(sxx / syy)
print(c(R2, r))
```

via Theorem 1.9. Both code segments print the values

$$R^2 = 0.9944 \qquad \text{and} \qquad r = 0.9972.$$

The proper interpretation of R^2 is that 99.44% of the total variation in $Y_1, Y_2, \ldots, Y_n$ can be explained by the linear relationship between X and Y. This high percentage is consistent with the scatterplot and estimated regression line in Figure 1.27, which shows a nearly perfect linear relationship between boiling point of water and the barometric pressure, and data values that lie very close to the estimated regression line. Table 1.6 contains the sums of squares, degrees of freedom, and mean squares for the $n = 17$ data pairs collected by Forbes. This ANOVA table can be generated with the *additional* R command

```
anova(fit)
```

The `anova` function returns a data frame, and values in that data frame can be extracted using the `$` extractor. The degrees of freedom for the sum of squares for error, for example, can be extracted with the R command

Source	SS	df	MS
Regression	145.1246	1	145.1246
Error	0.8131	15	0.0542
Total	145.9378	16	

Table 1.6: Partial ANOVA table for the Forbes data.

```
anova(fit)$Df[2]
```

The definitions and theorems that are associated with fitted values, residuals, estimating the population variance σ^2, partitioning the sums of squares, the coefficient of determination, the coefficient of correlation, and the ANOVA table are briefly reviewed here. The simple linear regression model

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

from Definition 1.1 establishes a linear statistical relationship between an independent variable X and a dependent random variable Y. The error term ε has population mean 0 and finite population variance σ^2. The n data pairs collected are denoted by $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. The fitted values $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n$ are the values on the estimated regression line associated with the independent variables $X_1, X_2, \ldots, X_n$:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

for $i = 1, 2, \ldots, n$, as established in Definition 1.2. The associated residuals are defined by

$$e_i = Y_i - \hat{Y}_i$$

for $i = 1, 2, \ldots, n$, as established in Definition 1.2. An unbiased estimator of the population variance of the error terms is

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2$$

as given in Theorem 1.7. The total sum of squares SST can be partitioned into the regression sum of squares SSR and the sum of squares for error SSE as

$$SST = SSR + SSE$$

or

$$\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

as given in Theorem 1.8. Two quantities that measure the linear association between X and Y are the coefficient of determination

$$R^2 = \frac{SSR}{SST},$$

which satisfies $0 \leq R^2 \leq 1$, and the coefficient of correlation

$$r = \pm\sqrt{R^2},$$

which satisfies $-1 \leq r \leq 1$ as defined in Definition 1.3. The coefficient of determination is the fraction of variation in $Y_1, Y_2, \ldots, Y_n$ that is explained by the linear relationship with X. The sums of squares are often presented in an ANOVA table, which includes columns for the source of variation, the sum of squares, the associated degrees of freedom, and the mean squares. An additional column will be added to the ANOVA table in the next chapter, when statistical inference in simple linear regression is introduced.

The point estimators for β_0, β_1, and σ^2 in the simple linear regression model have now all been established and many of their properties have been surveyed. But without additional assumptions, it is not possible to easily obtain interval estimators or perform hypothesis testing concerning these parameters. The next chapter addresses this issue.

1.9 Exercises

1.1 Establish a linear deterministic relationship between the independent variable X, the temperature in degrees Fahrenheit, and the dependent variable Y, the associated temperature in degrees Celsius.

1.2 Establish a nonlinear deterministic relationship between the independent variable X, the distance between two objects with fixed masses m_1 and m_2, and the dependent variable Y, the gravitational force acting between the two objects, using Newton's Law of Universal Gravitation.

1.3 For the following interpretations of the independent and dependent variables, predict whether the estimated slope $\hat{\beta}_1$ in a simple linear regression model will be positive or negative.

(a) The independent variable X is a car's speed and the dependent variable Y is the car's stopping distance.

(b) The independent variable X is a car's weight and the dependent variable Y is the car's fuel efficiency measured in miles per gallon.

(c) The independent variable X is a husband's height and the dependent variable Y is the wife's height for a married couple.

(d) The independent variable X is the average annual unemployment rate and the dependent variable Y is the annual GDP for a particular country.

1.4 For the simple linear regression model, show that solving the 2×2 set of linear normal equations

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$
$$\hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i$$

for $\hat{\beta}_0$ and $\hat{\beta}_1$ gives the expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ given in Theorem 1.1.

1.5 Consider the simple linear regression model

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

where

- the population intercept is $\beta_0 = 1$,
- the population slope is $\beta_1 = 1/2$, and
- the error term ε has a $U(-1, 1)$ distribution.

Assume that $n = 10$ data pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_{10}, Y_{10})$ are collected. The values of the independent variable X are equally likely to be one of the integers $0, 1, 2, \ldots, 9$, What are the minimum and maximum values that the estimated parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ can assume?

1.6 For the values of the independent variables $X_1, X_2, \ldots, X_n$, show that

$$\sum_{i=1}^{n} \left(X_i - \bar{X}\right) = 0.$$

1.7 Write R commands to plot contours of the sum of squares for the sales data pairs

$$(X_1, Y_1) = (6, 2), \qquad (X_2, Y_2) = (8, 9), \qquad (X_3, Y_3) = (2, 2)$$

in the (β_0, β_1) plane.

1.8 The least squares criterion applied to a simple linear regression model minimizes

$$S = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

If instead the *least absolute deviation* criterion (also known as the minimum absolute deviation or MAD criterion) were applied to a simple linear regression model to minimize

$$S = \sum_{i=1}^{n} |Y_i - \beta_0 - \beta_1 X_i| ,$$

what are the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ for the sales data pairs

$$(X_1, Y_1) = (6, 2) \qquad (X_2, Y_2) = (8, 9) \qquad (X_3, Y_3) = (2, 2)?$$

1.9 Write a Monte Carlo simulation experiment that uses the same parameters as those in Example 1.4 (that is, $\beta_0 = 1$, $\beta_1 = 1/2$, $\varepsilon \sim U(-1, 1)$, $n = 10$) for 5000 replications, but this time selects the independent variable values to be equally likely integers from -5 and 5. Produce analogous figures to those of Figure 1.11 and Figure 1.12. Comment on your figures and how they relate to the variance–covariance matrix from Theorem 1.4.

1.10 For a simple linear regression model with $X_1 = 1, X_2 = 2, \ldots, X_n = n$ and $\sigma^2 = 1$, find the variance–covariance matrix of $\hat{\beta}_0$ and $\hat{\beta}_1$.

1.11 Use Theorems 1.2 and 1.4 to show that the least squares estimator of the intercept of the regression line β_0 in the simple linear regression model is a consistent estimator of β_0.

1.12 Example 1.6 calculates the variance–covariance matrix for a single replication of a Monte Carlo simulation experiment. Conduct this experiment for 5000 replications and report the average of the values in the variance–covariance matrix.

1.13 Let L be the set of all linear estimators of the slope β_1 in a simple linear regression model. Let U be the set of all unbiased estimators of the slope β_1 in a simple linear regression model. Give an example of an estimator of β_1 in $L \cap U'$.

1.14 Show that the fitted simple linear regression model

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

for $i = 1, 2, \ldots, n$ can be written as

$$\hat{Y}_i - \bar{Y} = \hat{\beta}_1 (X_i - \bar{X}) ,$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimators of β_0 and β_1 and $\bar{X}$ and $\bar{Y}$ are the sample means of the observed values of the independent and dependent variables.

1.15 Write a paragraph that argues why a fitted least squares regression line cannot pass through all data pairs except for one of the data pairs.

1.16 One of the most common error distributions used in simple linear regression is the normal distribution with population mean 0 and finite population variance σ^2, which has probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)} \qquad -\infty < x < \infty.$$

An alternative error distribution is the Laplace distribution with probability density function

$$f(x) = \frac{1}{\sqrt{2}\sigma} e^{-\sqrt{2}|x-\mu|/\sigma} \qquad -\infty < x < \infty.$$

Since the error distribution must have expected value zero by assumption, this reduces to

$$f(x) = \frac{1}{\sqrt{2}\sigma} e^{-\sqrt{2}|x|/\sigma} \qquad -\infty < x < \infty.$$

As parameterized here, the Laplace distribution has population variance σ^2. Both of these distributions are symmetric and centered about zero.

(a) Plot the normal and Laplace error probability density functions on $-3 < x < 3$ and comment on any differences between the two error distributions. Use $\sigma = 1$ for the plots.

(b) Plot the normal and Laplace error probability density functions on $4 < x < 5$ and comment on any differences between the tails of the two error distributions.

(c) Fit both of these error distributions (that is, find $\hat{\sigma}^2$ for each distribution) for the `forbes` data set from the `MASS` package in R using the simple linear regression model.

1.17 Let the independent variable X be a car's speed and the dependent variable Y be the car's stopping distance, which are going to be modeled with a simple linear regression model. In which of the following scenarios do you expect to have a larger population variance of the error term?

(a) The data pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_{20}, Y_{20})$ are $n = 20$ new cars that are all of the same make and model.

(b) The data pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_{20}, Y_{20})$ are $n = 20$ new cars from $n = 20$ different car manufacturers.

1.18 Show that the sum of squares for regression in a simple linear regression model can be written as

$$SSR = \hat{\beta}_1 S_{XY}.$$

1.19 Show that the sum of squares for regression in a simple linear regression model can be written as

$$SSR = \hat{\beta}_1^2 S_{XX}.$$

1.20 Consider the data pairs in the `Formaldehyde` data set built into the base R language. Use the `help` function in R to determine the interpretation of the independent and dependent variables. Fit a simple linear regression model to the data pairs and interpret the meaning of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$. Also, calculate SST, SSR, and SSE for this data set.

1.21 Consider the data pairs collected by James Forbes that are given in the data frame `forbes` contained in the `MASS` package in R. The independent variable is the boiling point (in degrees Fahrenheit) and the dependent variable is the barometric pressure (in inches of mercury). For a simple linear regression model, calculate

- the fitted values,
- the residuals,
- the sum of squares for error, and
- the mean square error

without using the `lm` function. Then use the `lm` function to check the correctness of the values that you calculate.

1.22 This exercise investigates the effect of controllable values of $X_1, X_2, \ldots, X_n$ on the coefficient of determination R^2 in simple linear regression. Consider the simple linear regression model

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

where

- the population intercept is $\beta_0 = 1$,
- the population slope is $\beta_1 = 1/2$, and
- the error term ε has a $N(0, 1)$ distribution.

Conduct a Monte Carlo simulation with 40,000 replications that estimates the expected coefficient of determination for $n = 10$ data pairs under the following two ways of setting the values of $X_1, X_2, \ldots, X_{10}$.

(a) Let $X_i = i$ for $1, 2, \ldots, 10$.

(b) Let $X_1 = X_2 = \cdots = X_5 = 5$ and $X_6 = X_7 = \cdots = X_{10} = 6$.

1.23 Let S_X and S_Y be the sample standard deviations of the independent and dependent variables, respectively. Show that the following four definitions of the coefficient of correlation are equivalent.

(a) $r = \dfrac{1}{n-1} \displaystyle\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S_X} \right) \left(\frac{Y_i - \bar{Y}}{S_Y} \right)$

(b) $r = \pm \sqrt{\dfrac{SSR}{SSE}}$

(c) $r = \dfrac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$

(d) $r = \hat{\beta}_1 \sqrt{\dfrac{S_{XX}}{S_{YY}}}$