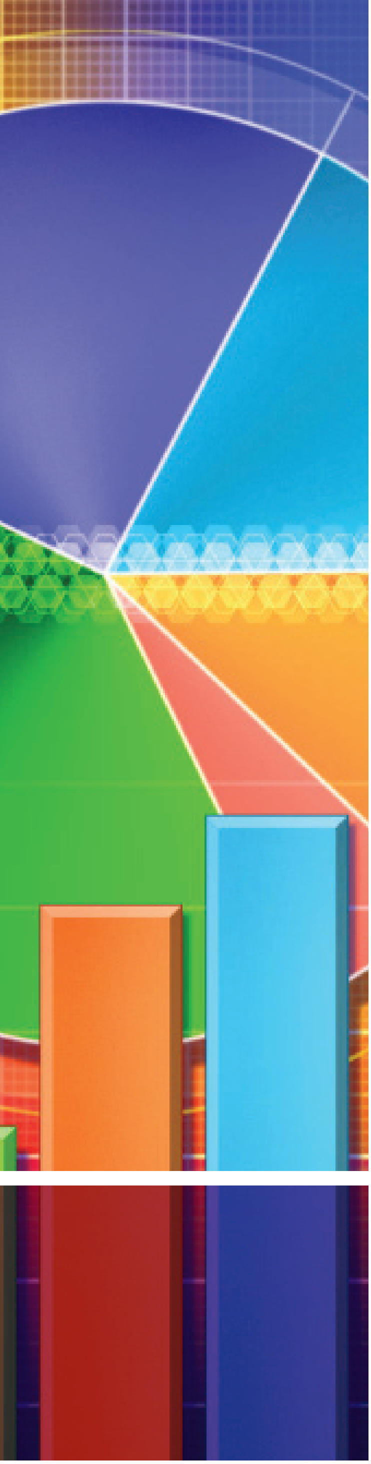


QUANTITATIVE APPROACHES TO DECISION MAKING COLLECTION

Donald N. Stengel, *Editor*

Regression Analysis

Unified Concepts, Practical Applications, and Computer Implementation

Bruce L. Bowerman

Richard T. O'Connell

Emily S. Murphree

BUSINESS EXPERT PRESS

Regression Analysis

Regression Analysis

Unified Concepts, Practical Applications, and Computer Implementation

Bruce L. Bowerman, Richard T. O'Connell, and
Emily S. Murphree

BEP BUSINESS EXPERT PRESS

Regression Analysis: Unified Concepts, Practical Applications, and Computer Implementation

First published in 2015 by
Business Expert Press, LLC
222 East 46th Street, New York, NY 10017
www.businessexpertpress.com

ISBN-13: 978-1-60649-950-4 (paperback)
ISBN-13: 978-1-60649-951-1 (e-book)

Business Expert Press Quantitative Approaches to Decision Making Collection

Collection ISSN: 2163-9515 (print)
Collection ISSN: 2163-9582 (electronic)

Cover and interior design by Exeter Premedia Services Private Ltd., Chennai, India

First edition: 2015

10 9 8 7 6 5 4 3 2 1

Printed in the United States of America.

Abstract

Regression Analysis: Unified Concepts, Practical Applications, and Computer Implementation is a concise and innovative book that gives a complete presentation of applied regression analysis in approximately one-half the space of competing books. With only the modest prerequisite of a basic (non-calculus) statistics course, this text is appropriate for the widest possible audience.

Keywords

logistic regression, model building, model diagnostics, multiple regression, regression model, simple linear regression, statistical inference, time series regression

Contents

Preface ix

Chapter 1 An Introduction to Regression Analysis 1

Chapter 2 Simple and Multiple Regression:
An Integrated Approach...5

Chapter 3 More Advanced Regression Models.............................97

Chapter 4 Model Building and Model Diagnostics.....................159

Appendix A Statistical Tables...253

References...261

Index 263

Preface

Regression Analysis: Unified Concepts, Practical Applications, and Computer Implementation is a concise and innovative book that gives a complete presentation of applied regression analysis in approximately one-half the space of competing books. With only the modest prerequisite of a basic (non-calculus) statistics course, this text is appropriate for the widest possible audience—including college juniors, seniors, and first year graduate students in business, the social sciences, the sciences, and statistics, as well as professionals in business and industry. The reason that this text is appropriate for such a wide audience is that it takes a very unique and integrative approach to teaching regression analysis. Most books, after a short chapter introducing regression, cover simple linear regression and multiple regression in roughly four chapters by beginning with a chapter reviewing basic statistical concepts and then having chapters on simple linear regression, matrix algebra, and multiple regression. In contrast, this book, after a short chapter introducing regression, covers simple linear regression and multiple regression in a single cohesive chapter, Chapter 2, by efficiently integrating the discussion of the two techniques. In addition, the same Chapter 2 teaches both the necessary basic statistical concepts (for example, hypothesis testing) and the necessary matrix algebra concepts as they are needed in teaching regression. We believe that this approach avoids the needless repetition of traditional approaches and does the best job of getting a wide variety of readers (who might be students with different backgrounds in the same class) to the same level of understanding.

Chapter 3 continues the integrative approach of the book by discussing more advanced regression models, including models using squared and interaction terms, models using dummy variables, and logistic regression models. The book concludes with Chapter 4, which organizes the techniques of model building, model diagnosis, and model improvement into a cohesive six step procedure. Whereas many competing texts spread such modeling techniques over a fairly large number of chapters that can seem unrelated to the novice, the six step procedure organizes both standard and more advanced modeling techniques into a unified presentation. In addition, each chapter features motivating examples (many real world, all realistic) and concludes with a section showing how to use SAS followed by a set of exercises. Excel, MINITAB, and SAS outputs are used throughout the text, and the book's website contains more exercises for each chapter. The book's website also houses Appendices B, C, and D. Appendix B gives careful derivations of most of the applied results in the text. These derivations are referenced in the main text as the applied results are discussed. Appendix C includes an applied discussion extending the basic treatment of logistic regression given in the main text. This extended discussion covers binomial logistic regression, generalized (multiple category) logistic regression, and Poisson regression. Appendix D extends the basic treatment of modeling time series data given in the main text. The Box-Jenkins methodology and its use in regression analysis are discussed

Author Bruce Bowerman would like to thank Professor David Nickerson of the University of Central Florida for motivating the writing of this book. All three authors would like to thank editor Scott Isenberg, production manager Destiny Hadley, and permissions editor Marcy Schneidewind, as well as the fine people at Exeter, for their hard work. Most of all we are indebted to our families for their love and encouragement over the years.

Bruce L. Bowerman
Richard T. O'Connell
Emily S. Murphree

CHAPTER 1

An Introduction to Regression Analysis

1.1 Observational Data and Experimental Data

In many statistical studies a variable of interest, called the *response variable* (or *dependent variable*), is identified. Data are then collected that tell us about how one or more *factors* might influence the variable of interest. If we cannot control the factor(s) being studied, we say that the data are *observational*. For example, suppose that a natural gas company serving a city collects data to study the relationship between the city's weekly natural gas consumption (the response variable) and two factors—the average hourly atmospheric temperature and the average hourly wind velocity in the city during the week. Because the natural gas company cannot control the atmospheric temperatures or wind velocities in the city, the data collected are observational.

If we can control the factors being studied, we say that the data are *experimental*. For example, suppose that an oil company wishes to study how three different gasoline types (A, B, and C) affect the mileage obtained by a popular midsized automobile model. Here the response variable is gasoline mileage, and the company will study a single factor—gasoline type. Since the oil company can control which gasoline type is used in the midsized automobile, the data that the oil company will collect are experimental.

1.2 Regression Analysis and Its Objectives

Regression analysis is a statistical technique that can be used to analyze both observational and experimental data, and it tells us how the factors under consideration might affect the response (dependent) variable. In regression analysis the factors that might affect the dependent variable are most often referred to as independent, or predictor, variables. We denote the dependent variable in regression analysis by the symbol y, and we denote the independent variables that might affect the dependent variable by the symbols $x_1, x_2, \ldots, x_k$. The objective of regression analysis is to build a *regression model or prediction equation*—an equation relating y to $x_1, x_2, \ldots, x_k$. We use the model to *describe, predict*, and *control y* on the basis of the independent variables. When we predict y for a particular set of values of $x_1, x_2, \ldots, x_k$, we will wish to place a bound on the *error of prediction.* The goal is to build a regression model that produces an error bound that will be small enough to meet our needs.

A regression model can employ *quantitative independent variables*, or *qualitative independent* variables, or both. *A quantitative independent variable* assumes numerical values corresponding to points on the real line. A *qualitative independent variable* is nonnumerical. The levels of such a variable are defined by describing them. As an example, suppose that we wish to build a regression model relating the dependent variable

y = demand for a consumer product

to the independent variables

x_1 = the price of the product,
x_2 = the average industry price of competitors' similar products,
x_3 = advertising expenditures made to promote the product, and
x_4 = the type of advertising campaign (television, radio, print media, etc.) used to promote the product.

Here x_1, x_2, and x_3 are quantitative independent variables. In contrast, x_4 is a qualitative independent variable, since we would define the levels of x_4 by describing the different advertising campaigns. After constructing an appropriate regression model relating y to x_1, x_2, x_3, and x_4, we would use the model

1. to *describe* the relationships between y and x_1, x_2, x_3, and x_4. For instance, we might wish to describe the effect that increasing advertising expenditure has on the demand for the product. We might also wish to determine whether this effect depends upon the price of the product;

2. to *predict* future demands for the product on the basis of future values of x_1, x_2, x_3, and x_4;
3. to *control* future demands for the product by controlling the price of the product, advertising expenditures, and the types of advertising campaigns used.

Note that we cannot control the price of competitors' products, nor can we control competitors' advertising expenditures or other factors that affect demand. Therefore we cannot perfectly control or predict future demands.

We develop a regression model by using observed values of the dependent and independent variables. If these values are observed over time, the data are called *time series* data. On the other hand, if these values are observed at one point in time, the data are called *cross-sectional* data. For example, suppose we observe values of the demand for a product, the price of the product, and the advertising expenditures made to promote the product. If we observe these values in one sales region over 30 consecutive months, the data are time series data. If we observe these values in thirty different sales regions for a particular month of the year, the data are cross-sectional data.

CHAPTER 2

Simple and Multiple Regression: An Integrated Approach

2.1 The Simple Linear Regression Model, and the Least Squares Point Estimates

2.1.1 The Simple Linear Regression Model

The *simple linear regression model* relates the dependent variable, which is denoted y, to a single independent variable, which is denoted x, and assumes that the relationship between y and x can be approximated by a straight line. We can tentatively decide whether there is an approximate straight-line relationship between y and x by making a *scatter diagram*, or *scatter plot*, of y versus x. First, data concerning the two variables are observed in pairs. To construct the scatter plot, each value of y is plotted against its corresponding value of x. If the y values tend to increase or decrease in a straight-line fashion as the x values increase, and if there is a scattering of the (x, y) points around the straight line, then it is reasonable to describe the relationship between y and x by using the simple linear regression model. We illustrate this in the following example, which shows how regression analysis can help a natural gas company improve its gas ordering process.

Example 2.1

When the natural gas industry was deregulated in 1993, natural gas companies became responsible for acquiring the natural gas needed to heat the homes and businesses in the cities they serve. To do this, natural gas companies purchase natural gas from marketers (usually through long-term contracts) and periodically (daily, weekly, monthly, or the like) place orders for natural gas to be transmitted by pipeline transmission systems to their cities. There are hundreds of pipeline transmission systems in the United States, and many of these systems supply a large number of cities.

To place an order (called a *nomination*) for an amount of natural gas to be transmitted to its city over a period of time (day, week, month), a natural gas company makes its best prediction of the city's natural gas needs for that period. The company then instructs its marketer(s) to deliver this amount of gas to its pipeline transmission system. If most of the natural gas companies being supplied by the transmission system can predict their cities' natural gas needs with reasonable accuracy, then the overnominations of some companies will tend to cancel the undernominations of other companies. As a result, the transmission system will probably have enough natural gas to efficiently meet the needs of the cities it supplies.

In order to encourage natural gas companies to make accurate transmission nominations and to help control costs, pipeline transmission systems charge, in addition to their usual fees, transmission fines. A natural gas company is charged a transmission fine if it substantially undernominates natural gas, which can lead to an excessive number of unplanned transmissions, or if it substantially overnominates natural gas, which can lead to excessive storage of unused gas. Typically, pipeline transmission systems allow a certain percentage nomination error before they impose a fine. For example, some systems do not impose a fine unless the actual amount of natural gas used by a city differs from the nomination by more than 10 percent. Beyond the allowed percentage nomination error, fines are charged on a sliding scale—the larger the nomination error, the larger the transmission fine.

Suppose, we are analysts in a management consulting firm. The natural gas company serving a small city has hired the consulting firm to develop an accurate way to predict the amount of fuel (in millions of cubic feet–MMcf–of natural gas) that will be required to heat the city. Because the pipeline transmission system supplying the city evaluates nomination errors and assesses fines weekly, the natural gas company wants predictions of future weekly fuel consumptions. Moreover, since the pipeline transmission system allows a 10 percent nomination error before assessing a fine, the company would like the actual and predicted weekly fuel consumptions to differ by no more than 10 percent. Our experience suggests that weekly fuel consumption substantially depends on the average hourly temperature (in degrees Fahrenheit) measured in the city during the week. Therefore, we will try to predict the *dependent (response) variable* weekly fuel consumption (y) on the basis of the *independent (predictor) variable* average hourly temperature (x) during the week. To this end, we observe values of y and x for eight weeks. The data are given in the Excel output of Figure 2.1, along with a scatter plot of y versus x. This plot shows (1) a tendency for the fuel consumptions to decrease in a straight line fashion as the temperatures increase and (2) a scattering of points around the straight line.

To begin to find a regression model that represents characteristics (1) and (2) of the data plot, consider a specific average hourly temperature x. For example, consider the average hourly temperature 28°F, which was observed in week one, or consider the average hourly temperature 45.9°F, which was observed in week five (there is nothing special about these two average hourly temperatures, but we will use them throughout this example to help explain the idea of a regression model). For the specific average hourly temperature x that we consider, there are, in theory, many weeks that could have this temperature. However, although these weeks each have the same average hourly temperature, other factors that affect fuel consumption could vary from week to week. For example, these weeks might have different average hourly wind velocities, different thermostat settings, and so forth. Therefore, the weeks could have different fuel consumptions. It follows that there is a population of weekly fuel

	A	B	C	D	E	F	G	H
1	TEMP	FUELCONS						
2	28	12.4						
3	28	11.7						
4	32.5	12.4						
5	39	10.8						
6	45.9	9.4						
7	57.8	9.5						
8	58.1	8						
9	62.5	7.5						
10								
11								
12								
13								
14								

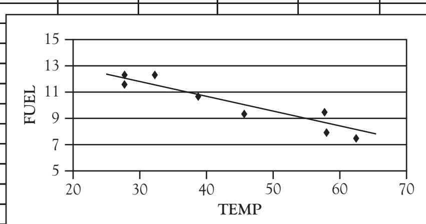


Figure 2.1 The fuel consumption data, and a scatter plot

consumptions that could be observed when the average hourly temperature is x. Furthermore, this population has a mean, which we denote as $\mu_{y|x}$ (pronounced **mu of *y* given *x***).

We can represent the straight-line tendency we observe in Figure 2.1 by assuming that $\mu_{y|x}$ is related to x by the equation

$$\mu_{y|x} = \beta_0 + \beta_1 x$$

This is the equation of a straight line with ***y*-intercept** β_0 (pronounced **beta zero**) and **slope** β_1 (pronounced **beta one**). To better understand the straight line and the meanings of β_0 and β_1, we must first realize that the values of β_0 and β_1 determine the precise value of the mean weekly fuel consumption $\mu_{y|x}$ that corresponds to a given value of the average hourly temperature x. We cannot know the true values of β_0 and β_1, and in the next section we will learn how to estimate these values. However, for illustrative purposes, let us suppose that the true value of β_0 is 15.77 and the true value of β_1 is $-.1281$. It would then follow, for example, that the mean of the population of all weekly fuel consumptions that could be observed when the average hourly temperature is 28°F is

$$\begin{aligned}\mu_{y|28} &= \beta_0 + \beta_1(28) \\ &= 15.77 - .1281(28) \\ &= 12.18 \text{ MMcf of natural gas}\end{aligned}$$

As another example, it would also follow that the mean of the population of all weekly fuel consumptions that could be observed when the average hourly temperature is 45.9°F is

$$\begin{aligned}\mu_{y|45.9} &= \beta_0 + \beta_1(45.9) \\ &= 15.77 - .1281(45.9) \\ &= 9.89 \text{ MMcf of natural gas}\end{aligned}$$

When we say that the equation $\mu_{y|x} = \beta_0 + \beta_1 x$ is the equation of a straight line, we mean that the different mean weekly fuel consumptions that correspond to different average hourly temperatures lie exactly on a straight line. For example, consider the eight mean weekly fuel consumptions that correspond to the eight average hourly temperatures in Figure 2.1. In Figure 2.2 we depict these mean weekly fuel consumptions as triangles that lie exactly on the straight line defined by the equation

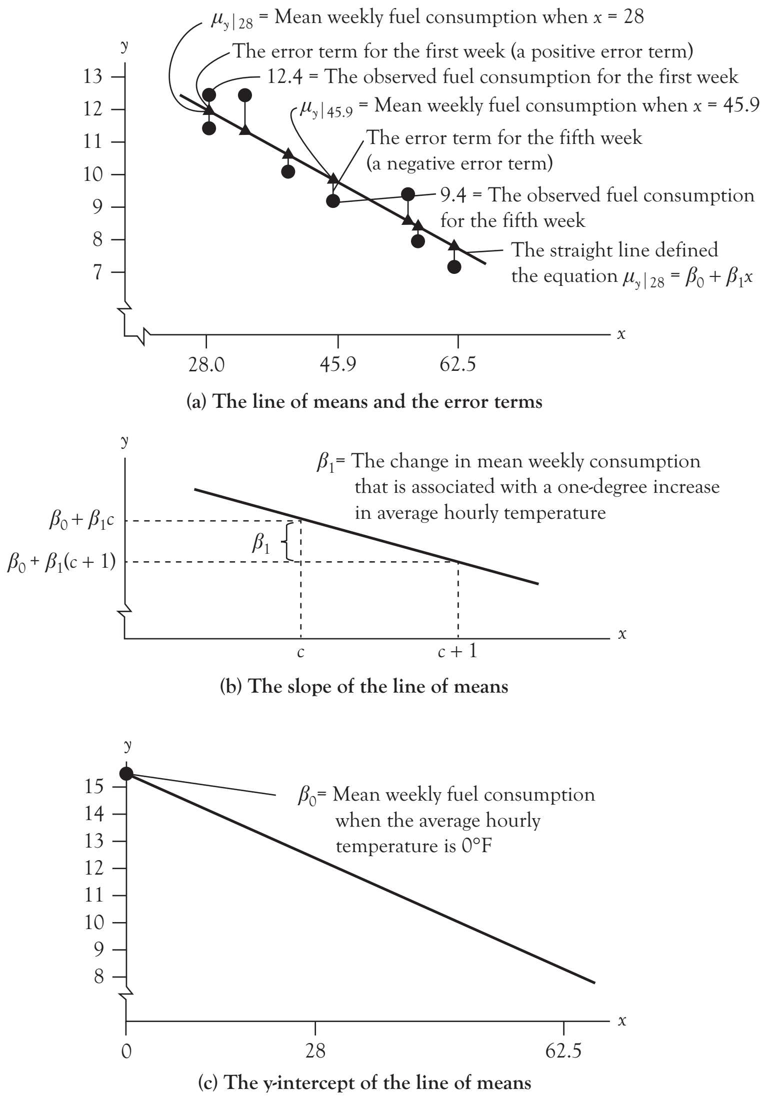


Figure 2.2 The simple linear regression model relating weekly fuel consumption to average hourly temperature

$\mu_{y|x} = \beta_0 + \beta_1 x$. Furthermore, in this figure we draw arrows pointing to the triangles that represent the previously discussed means $\mu_{y|28}$ and $\mu_{y|45.9}$. Sometimes we refer to the straight line defined by the equation $\mu_{y|x} = \beta_0 + \beta_1 x$ as the *line of means*.

In order to interpret the slope β_1 of the line of means, consider two different weeks. Suppose that for the first week the average hourly temperature is c. The mean weekly fuel consumption for all such weeks is

$$\beta_0 + \beta_1 (c)$$

For the second week, suppose that the average hourly temperature is $(c+1)$. The mean weekly fuel consumption for all such weeks is

$$\beta_0 + \beta_1 (c + 1)$$

It is easy to see that the difference between these mean weekly fuel consumptions is β_1. Thus, as illustrated in Figure 2.2(b), the slope β_1 is the change in mean weekly fuel consumption that is associated with a one-degree increase in average hourly temperature. To interpret the meaning of the y-intercept β_0, consider a week having an average hourly temperature of 0°F. The mean weekly fuel consumption for all such weeks is

$$\beta_0 + \beta_1 (0) = \beta_0$$

Therefore, as illustrated in Figure 2.2(c), the y-intercept β_0 is the mean weekly fuel consumption when the average hourly temperature is 0°F. However, because we have not observed any weeks with temperatures near zero, we have no data to tell us what the relationship between mean weekly fuel consumption and average hourly temperature looks like for temperatures near zero. Therefore, the interpretation of β_0 is of dubious practical value. More will be said about this later.

Now recall that the observed weekly fuel consumptions are not exactly on a straight line. Rather, they are scattered around a straight line. To represent this phenomenon, we use the **simple linear regression model**

$$\begin{aligned} y &= \mu_{y|x} + \varepsilon \\ &= \beta_0 + \beta_1 x + \varepsilon \end{aligned}$$

This model says that the weekly fuel consumption y observed when the average hourly temperature is x differs from the mean weekly fuel consumption $\mu_{y|x}$ by an amount equal to ε (*epsilon*). Here ε is called an *error term*. The error term describes the effect on y of all factors other than the average hourly temperature. Such factors would include the average hourly wind velocity and the average hourly thermostat setting in the city. For example, Figure 2.2(a) shows that the error term for the first week is positive. Therefore, the observed fuel consumption $y = 12.4$ in the first week was above the corresponding mean weekly fuel consumption for all weeks when $x = 28$. As another example, Figure 2.2(a) also shows that the error term for the fifth week was negative. Therefore, the observed fuel consumption $y = 9.4$ in the fifth week was below the corresponding mean weekly fuel consumption for all weeks when $x = 45.9$. Of course, since we do not know the true values of β_0 and β_1, the relative positions of the quantities pictured in the figure are only hypothetical.

With the fuel consumption example as background, we are ready to define the *simple linear regression model relating the dependent variable y to the independent variable x*. We suppose that we have gathered n observations—each observation consists of an observed value of x and its corresponding value of y. Then:

The simple linear regression model

The *simple linear* (or *straight-line*) *regression model* is

$$y = \mu_{y|x} + \varepsilon = \beta_0 + \beta_1 x + \varepsilon$$

Here

1. $\mu_{y|x} = \beta_0 + \beta_1 x$ is the *mean value* of the dependent variable y when the value of the independent variable is x.
2. β_0 is the y *-intercept*. β_0 is the mean value of y when x equals 0.

The simple linear regression model (Continued)

3. β_1 is the *slope*. β_1 is the change (amount of increase or decrease) in the mean value of y associated with a one-unit increase in x. If β_1 is positive, the mean value of y increases as x increases. If β_1 is negative, the mean value of y decreases as x increases.
4. ε is an error term that describes the effects on y of all factors other than the value of the independent variable x.

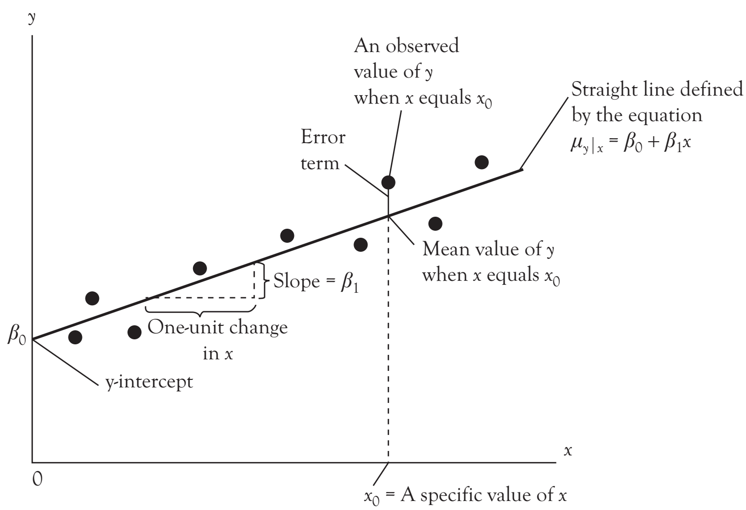


Figure 2.3 The simple linear regression model $(\beta_1 > 0)$

This model is illustrated in Figure 2.3 (note that x_0 in this figure denotes a specific value of the independent variable x). The y-intercept β_0 and the slope β_1 are called *regression parameters.* We will see how to estimate these parameters in the next subsection. Then, we will see how to use these estimates to predict y.

2.1.2 The Least Squares Point Estimates

Suppose that we have gathered n observations (x_1, y_1), (x_2, y_2), . . , (x_n, y_n), where each observation consists of a value of an independent variable x and a corresponding value of a dependent variable y. Also, suppose that a scatter plot of the n observations indicates that the simple linear regression model relates y to x. In order to estimate the y-intercept β_0 and the slope β_1 of the line of means of this model, we could visually draw a line—called an estimated regression line—through the scatter plot. Then, we could read the y-intercept and slope off the *estimated regression line* and use these values as the point estimates of β_0 and β_1. Unfortunately, if different people visually drew lines through the scatter plot, their lines would probably differ from each other. What we need is the *best line* that can be drawn through the scatter plot. Although there are various definitions of what this best line is, one of the most useful best lines is the *least squares line*.

To understand the least squares line, we let $\hat{y} = b_0 + b_1 x$ denote the general equation of an estimated regression line drawn through a scatter plot. Here, since we will use this line to predict y on the basis of x, we call $\hat{y}$ *the predicted value of* y when the value of the independent variable is x. In addition, b_0 is the y-intercept and b_1 is the slope of the estimated regression line. When we determine numerical values for b_0 and b_1, these values will be the point estimates of the y-intercept β_0 and the slope β_1 of the line of means. To explain which estimated regression line is the least squares line, we begin with the fuel consumption situation. Figure 2.4 shows an estimated regression line drawn through a scatter plot of the fuel consumption data. In this figure the dots represent the eight observed fuel consumptions and the squares represent the eight predicted fuel consumptions given by the estimated regression line. Furthermore, the line segments drawn between the dots and squares represent *residuals*, which are

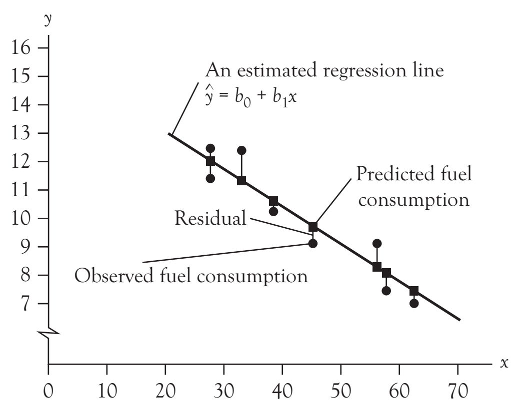


Figure 2.4 An estimated regression line drawn through the fuel consumption scatter plot

the differences between the observed and predicted fuel consumptions. Intuitively, if a particular estimated regression line provides a good "fit" to the fuel consumption data, it will make the predicted fuel consumptions "close" to the observed fuel consumptions, and thus the residuals given by the line will be small. The *least squares line* is the line that minimizes the sum of squared residuals. That is, the least squares line is the line positioned on the scatter plot so as to minimize the sum of the squared vertical distances between the observed and predicted fuel consumptions.

To define the least squares line in a general situation, consider an arbitrary observation (x_i, y_i) in a sample of n observations. For this observation, the predicted value of the dependent variable y given by an estimated regression line is $\hat{y}_i = b_0 + b_1 x_i$. Furthermore, the *prediction error* (also called the *residual*) for this observation is

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$$

Then, the *least squares line* is the line that minimizes the sum of the squared prediction errors (that is, the *sum of squared residuals*)

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

To find the least squares line, we find the values of the y-intercept b_0 and slope b_1 that give values of $\hat{y}_i = b_0 + b_1 x_i$ that minimize SSE. These values of b_0 and b_1 are called the *least squares point estimates* of β_0 and β_1. Using calculus (see Section B.1 in Appendix B), we can show that the least squares point estimates are as follows:

The least squares point estimates

For the *simple linear regression model*:

1. The *least squares point estimate of the slope* β_1 is $b_1 = \dfrac{SS_{xy}}{SS_{xx}}$, where[1]

[1]In order to simplify notation, we will often drop the limits on summations in this and subsequent chapters. That is, instead of using the summation $\sum_{i=1}^{n}$ we will simply write $\sum$.

The least squares point estimates (Continued)

$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

and

$$SS_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

2. The **least squares point estimate of the y-intercept** β_0 is $b_0 = \bar{y} - b_1\bar{x}$, where

$$\bar{y} = \frac{\sum y_i}{n} \text{ and } \bar{x} = \frac{\sum x_i}{n}$$

Here n is the number of observations (an observation is an observed value of x and its corresponding value of y).

Example 2.2

In order to calculate least squares point estimates of the parameters β_1 and β_0 in the fuel consumption model

$$\begin{aligned} y &= \mu_{y|x} + \varepsilon \\ &= \beta_0 + \beta_1 x + \varepsilon \end{aligned}$$

we first consider the summations that are shown in Table 2.1. Using these summations, we calculate SS_{xy} and SS_{xx} as follows:

$$\begin{aligned} \mathrm{SS}_{xy} &= \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} \\ &= 3413.11 - \frac{(351.8)(81.7)}{8} = -179.6475 \\ \mathrm{SS}_{xx} &= \sum x_i^2 - \frac{(\sum x_i)^2}{n} \\ &= 16{,}874.76 - \frac{(351.8)^2}{8} = 1404.355 \end{aligned}$$

Table 2.1 The calculation of the point estimates b_0 and b_1 of the parameters in the fuel consumption model $y = \mu_{y|x} + \varepsilon = \beta_0 + \beta_1 x + \varepsilon$

y_i	x_i	x_i^2	$x_i y_i$
12.4	28.0	$(28.0)^2 = 784$	$(28.0)(12.4) = 347.2$
11.7	28.0	$(28.0)^2 = 784$	$(28.0)(11.7) = 327.6$
12.4	32.5	$(32.5)^2 = 1{,}056.25$	$(32.5)(12.4) = 403$
10.8	39.0	$(39.0)^2 = 1{,}521$	$(39.0)(10.8) = 421.2$
9.4	45.9	$(45.9)^2 = 2{,}106.81$	$(45.9)(9.4) = 431.46$
9.5	57.8	$(57.8)^2 = 3{,}340.84$	$(57.8)(9.5) = 549.1$
8.0	58.1	$(58.1)^2 = 3{,}375.61$	$(58.1)(8.0) = 464.8$
7.5	62.5	$(62.5)^2 = 3{,}906.25$	$(62.5)(7.5) = 468.75$
$\Sigma y_i = 81.7$	$\Sigma x_i = 351.8$	$\Sigma x_i^2 = 16{,}874.76$	$\Sigma x_i y_i = 3{,}413.11$

It follows that the least squares point estimate of the slope β_1 is

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{-179.6475}{1404.355} = -.1279$$

Furthermore, because

$$\bar{y} = \frac{\sum y_i}{8} = \frac{81.7}{8} = 10.2125 \text{ and } \bar{x} = \frac{\sum x_i}{8} = \frac{351.8}{8} = 43.98$$

the least squares point estimate of the y-intercept β_0 is

$$b_0 = \bar{y} - b_1\bar{x} = 10.2125 - (-.1279)(43.98) = 15.84$$

Since $b_1 = -.1279$, we estimate that mean weekly fuel consumption decreases (since b_1 is negative) by .1279 MMcf of natural gas when average hourly temperature increases by one degree. Since $b_0 = 15.84$, we estimate that mean weekly fuel consumption is 15.84 MMcf of natural gas when average hourly temperature is 0°F. However, we have not observed any weeks with temperatures near zero, so making this interpretation of b_0 might be dangerous. We discuss this point more fully in the next section.

Table 2.2 Predictions using the least squares point estimates $b_0 = 15.84$ and $b_1 = -.1279$

Week, i	x_i	y_i	$\hat{y}_i = 15.84 - .1279x_i$	$e_i = y_i - \hat{y}_i$
1	28.0	12.4	12.2560	.1440
2	28.0	11.7	12.2560	−.5560
3	32.5	12.4	11.6804	.7196
4	39.0	10.8	10.8489	−.0489
5	45.9	9.4	9.9663	−.5663
6	57.8	9.5	8.4440	1.0560
7	58.1	8.0	8.4056	−.4056
8	62.5	7.5	7.8428	−.3428

$\text{SSE} = \sum_{i=1}^{8} e_i^2 = 2.568$

The least squares line

$$\hat{y} = b_0 + b_1 x = 15.84 - .1279x$$

is sometimes called the *least squares prediction equation*. In Table 2.2 we summarize using this prediction equation to calculate the predicted fuel consumptions and the residuals for the eight weeks of fuel consumption data. For example, since in week one the average hourly temperature was 28°F, the predicted fuel consumption for week one is

$$\hat{y}_1 = 15.84 - .1279(28) = 12.2560$$

It follows, since the observed fuel consumption in week one was $y_1 = 12.4$, that the residual for week one is

$$e_1 = y_1 - \hat{y}_1 = 12.4 - 12.2560 = .1440$$

If we consider all of the residuals in Table 2.4 and add their squared values, we find that SSE, the sum of squared residuals, is 2.568. If we calculated SSE by using any point estimates of β_0 and β_1 other than the least squares point estimates $b_0 = 15.84$ and $b_1 = -.1279$, we would obtain a larger value of *SSE*. The SSE of 2.568 given by the least squares point estimates will be used throughout this chapter.

We next define the *experimental region* to be the range of the previously observed values of the average hourly temperature x. Referring to Figure 2.1, we see that the experimental region consists of the range of average hourly temperatures from 28°F to 62.5°F. The simple linear regression model relates weekly fuel consumption y to average hourly temperature x for values of x that are in the experimental region. For such values of x, the least squares line is the estimate of the line of means. This implies that the point on the least squares line that corresponds to the average hourly temperature x

$$\begin{aligned}\hat{y} &= b_0 + b_1 x \\ &= 15.84 - .1279x\end{aligned}$$

is the point estimate of $\mu_{y|x} = \beta_0 + \beta_1 x$, the mean of all weekly fuel consumptions that could be observed when the average hourly temperature is x. In addition, we predict the error term ε to be zero. Therefore, $\hat{y}$ is also the *point prediction of an individual value* $y = \beta_0 + \beta_1 x + \varepsilon$, which is the amount of fuel consumed in a single week that has an average hourly temperature of x. Note that the reason we predict the error term ε to be zero is that, because of several *regression assumptions* to be discussed in Section 2.3, ε has a 50 percent chance of being positive and a 50 percent chance of being negative.

Now suppose a weather forecasting service predicts that the average hourly temperature in the next week will be 40°F. Because 40°F is in the experimental region,

$$\begin{aligned}\hat{y} &= 15.84 - .1279(40) \\ &= 10.72 \text{ MMcf of natural gas}\end{aligned}$$

is (1) the point estimate of the mean weekly fuel consumption when the average hourly temperature is 40°F and (2) the point prediction of an individual weekly fuel consumption when the average hourly temperature is 40°F. This says that (1) we estimate that the average of all possible weekly fuel consumptions that could potentially be observed when the average hourly temperature is 40°F equals 10.72 MMcf of natural gas, and (2) we predict that the fuel consumption in a single week when the average hourly temperature is 40°F will be 10.72 MMcf of natural gas.

To conclude this example, note that Figure 2.5 illustrates both the point prediction $\hat{y} = 10.72$ and the potential danger of using the least squares line to predict outside the experimental region. In the figure, we extrapolate the least squares line far beyond the experimental region to obtain a prediction for a temperature of −10°F. As shown in Figure 2.1, for values of x in the experimental region, the observed values of y tend to decrease in a straight-line fashion as the values of x increase. However, for temperatures lower than 28°F the relationship between y and x might become curved. If it does, extrapolating the straight-line prediction equation to obtain a prediction for $x = -10$ might badly underestimate mean weekly fuel consumption (see Figure 2.5).

The previous example illustrates that when we are using a least squares regression line, we should not estimate a mean value or predict an individual value unless the corresponding value of x is in the *experimental region*—the range of the previously observed values of x. Often the value

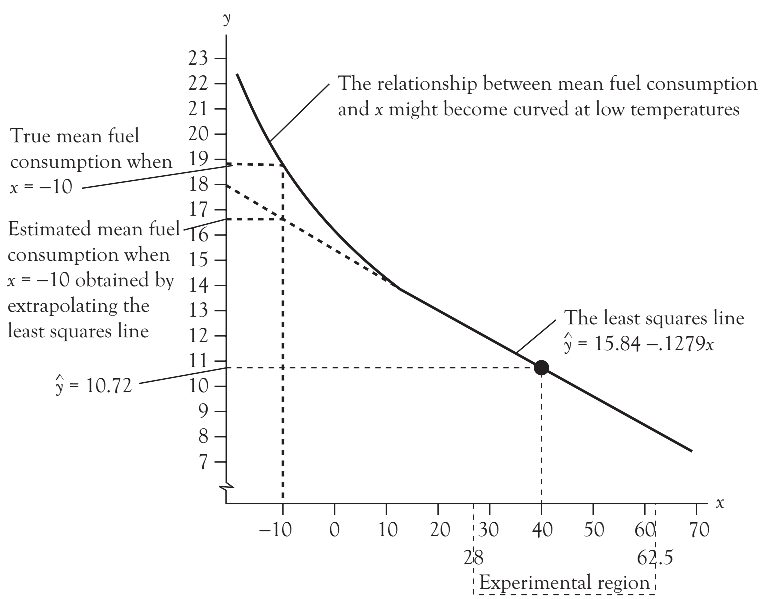


Figure 2.5 The point prediction $\hat{y} = 10.72$ and the danger of extrapolation

$x = 0$ is not in the experimental region. For example, consider the fuel consumption problem. Figure 2.5 illustrates that the average hourly temperature 0°F is not in the experimental region. In such a situation, it would not be appropriate to interpret the y-intercept b_0 as the estimate of the mean value of y when x equals zero. In the case of the fuel consumption problem, it would not be appropriate to use $b_0 = 15.84$ as the point estimate of the mean weekly fuel consumption when average hourly temperature is zero. Therefore, because it is not meaningful to interpret the y-intercept in many regression situations, we often omit such interpretations.

2.2 The (Multiple) Linear Regression Model, and the Least Squares Point Estimates Using Matrix Algebra

2.2.1 The (Multiple) Linear Regression Model

Regression models that employ more than one independent variable are called multiple regression models. We begin our study of these models by considering the following example.

Example 2.3

Part 1: The Data and a Regression Model

Consider the fuel consumption problem in which the natural gas company wishes to predict weekly fuel consumption for its city. In Section 2.1 we used the single predictor variable x, average hourly temperature, to predict y, weekly fuel consumption. We now consider predicting y on the basis of average hourly temperature and a second predictor variable—the chill index. The chill index for a given average hourly temperature expresses the combined effects of all other major weather-related factors that influence fuel consumption, such as wind velocity, cloud cover, and the passage of weather fronts. The chill index is expressed as a whole number between 0 and 30. A weekly chill index near zero indicates that, given the average hourly temperature during the week, all other major weather-related factors will only slightly increase weekly fuel consumption. A weekly chill index near 30 indicates that, given the average hourly temperature during the week, other weather-related factors will greatly increase weekly fuel consumption.

The company has collected data concerning weekly fuel consumption (y), average hourly temperature (x_1), and the chill index (x_2) for the last eight weeks. These data are given in Table 2.3. Figure 2.6 presents a scatter plot of y versus x_1. (Note that the y and x_1 values given in Table 2.3 are the same as the y and x values given in Figure 2.1). This plot shows that y tends to decrease in a straight-line fashion as x_1 increases. This suggests that if we wish to predict y on the basis of x_1 only, the simple linear regression model (having a negative slope)

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

relates y to x_1. Figure 2.6 also presents a scatter plot of y versus x_2. This plot shows that y tends to increase in a straight-line fashion as x_2 increases. This suggests that if we wish to predict y on the basis of x_2 only, the simple linear regression model (having a positive slope)

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

relates y to x_2. Since we wish to predict y on the basis of both x_1 and x_2, it seems reasonable to combine these models to form the model

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

Table 2.3 Fuel consumption data

Week	Average hourly temperature, x_1	Chill index, x_2	Fuel consumption, y (MMcf)
1	28.0	18	12.4
2	28.0	14	11.7
3	32.5	24	12.4
4	39.0	22	10.8
5	45.9	8	9.4
6	57.8	16	9.5
7	58.1	1	8.0
8	62.5	0	7.5

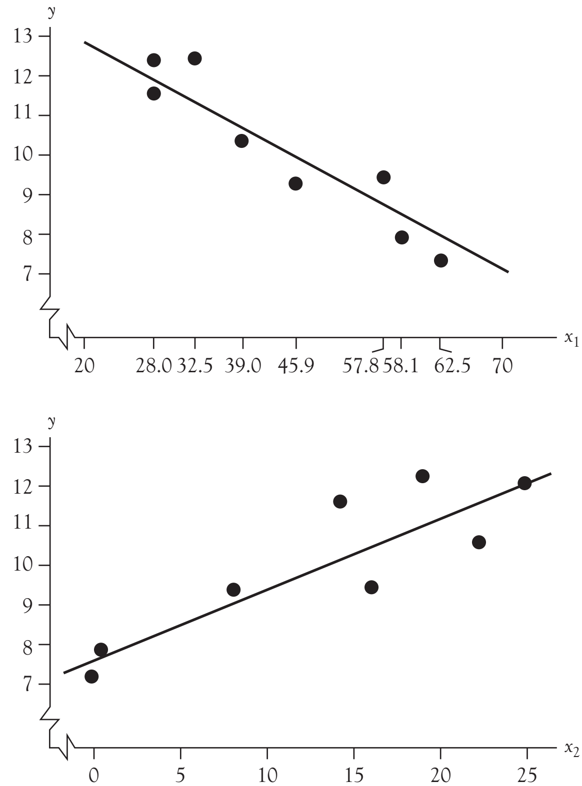


Figure 2.6 Scatter plots of y versus x_1 and y versus x_2

to relate y to x_1 and x_2. Here we have arbitrarily placed the $\beta_1 x_1$ term first and the $\beta_2 x_2$ term second, and we have renumbered β_1 and β_2 to be consistent with the subscripts on x_1 and x_2. This regression model says that

1. $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ is the mean value of y when the average hourly temperature is x_1 and the chill index is x_2. For instance,

$$\beta_0 + \beta_1(45.9) + \beta_2(8)$$

   is the average fuel consumption for all weeks having an average hourly temperature equal to 45.9 and a chill index equal to 8.

2. β_0, β_1, and β_2 are regression parameters relating the mean value of y to x_1 and x_2.
3. ε is an error term that describes the effects on y of all factors other than x_1 and x_2.

Part 2: Interpreting the Regression Parameters β_0, β_1, and β_2

The exact interpretations of the parameters β_0, β_1, and β_2 are quite simple. First, suppose that $x_1 = 0$ and $x_2 = 0$. Then

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$$

So β_0 is the mean weekly fuel consumption for all weeks having an average hourly temperature of 0°F and a chill index of zero. The parameter β_0 is called the *intercept* in the regression model. One might wonder whether β_0 has any practical interpretation, since it is unlikely that a week having an average hourly temperature of 0°F would also have a chill index of zero. Indeed, sometimes the parameter β_0 and other parameters in a regression analysis do not have practical interpretations because the situations related to the interpretations would not be likely to occur in practice. In fact, sometimes each parameter does not, by itself, have much practical importance. Rather, the parameters relate the mean of the dependent variable to the independent variables in an overall sense.

We next interpret the individual meanings of β_1 and β_2. To examine the interpretation of β_1, consider two different weeks. Suppose that for the first week the average hourly temperature is c and the chill index is d. The mean weekly fuel consumption for all such weeks is

$$\beta_0 + \beta_1(c) + \beta_2(d)$$

For the second week, suppose that the average hourly temperature is $c+1$ and the chill index is d. The mean weekly fuel consumption for all such weeks is

$$\beta_0 + \beta_1(c+1) + \beta_2(d)$$

It is easy to see that the difference between these mean fuel consumptions is β_1. Since weeks one and two differ only in that the average hourly temperature during week two is one degree higher than the average hourly temperature during week one, we can interpret the parameter β_1 as the change in mean weekly fuel consumption that is associated with a one-degree increase in average hourly temperature when the chill index does not change.

The interpretation of β_2 can be established similarly. We can interpret β_2 as the change in mean weekly fuel consumption that is associated with a one-unit increase in the chill index when the average hourly temperature does not change.

Part 3: A Geometric Interpretation of the Regression Model

We now interpret our fuel consumption model geometrically. We begin by defining the *experimental region* to be the range of the combinations of the observed values of x_1 and x_2. From the data in Table 2.3, it is reasonable to depict the experimental region as the shaded region in Figure 2.7.

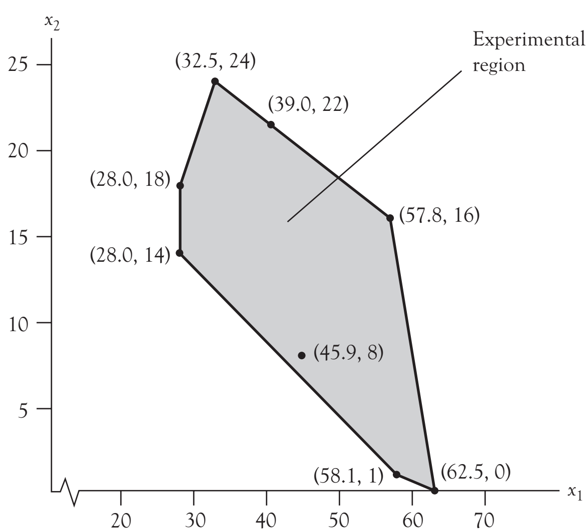


Figure 2.7 The experimental region

Here the combinations of x_1 and x_2 values are the ordered pairs in the figure.

We next write the mean value of y when the average hourly temperature is x_1 and the chill index is x_2 as $\mu_{y|x_1,x_2}$ (pronounced *mu of y given* x_1 *and* x_2) and consider the equation

$$\mu_{y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

which relates mean fuel consumption to x_1 and x_2. Since this is a linear equation in two variables, geometry tells us that this equation is the equation of a plane in three-dimensional space. We sometimes refer to this plane as the *plane of means*, and we illustrate the portion of this plane corresponding to the (x_1, x_2) combinations in the experimental region in Figure 2.8. As illustrated in this figure, the model

$$\begin{aligned} y &= \mu_{y|x_1,x_2} + \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \end{aligned}$$

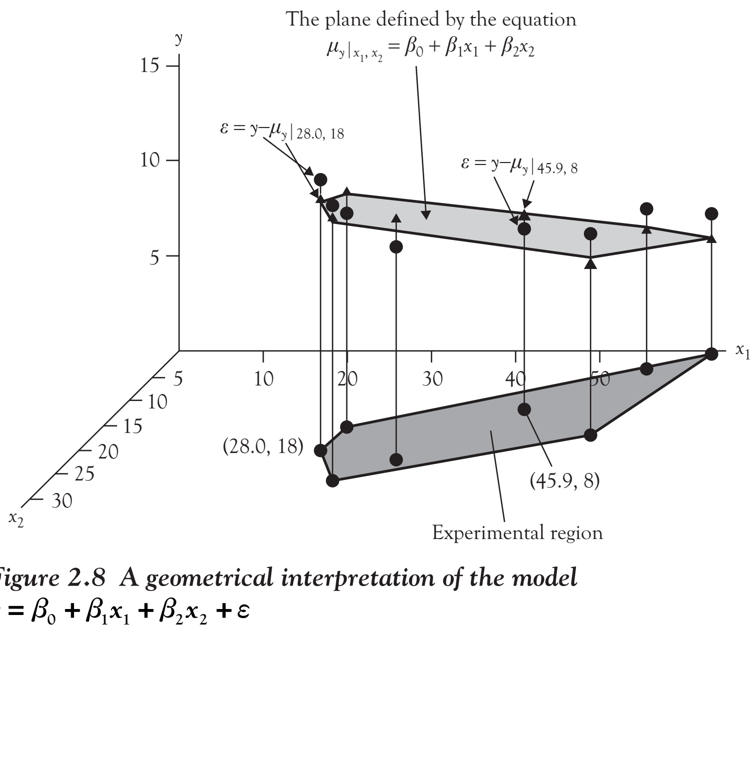


Figure 2.8 A geometrical interpretation of the model $\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon}$

says that the eight error terms cause the eight observed fuel consumptions (the dots in the upper portion of the figure) to deviate from the eight mean fuel consumptions (the triangles in the figure), which exactly lie on the plane of means

$$\mu_{y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

For example, consider the data for week one in Table 2.3 (y = 12.4, x_1 = 28.0, x_2 = 18). Figure 2.8 shows that the error term for this week is positive, causing y to be higher than $\mu_{y|28.0,18}$ (mean fuel consumption when $x_1 = 28$ and $x_2 = 18$). Here factors other than x_1 and x_2 (for instance, thermostat settings that are higher than usual) have resulted in a positive error term. As another example, the error term for week 5 in Table 2.3 ($y = 9.4$, $x_1 = 45.9$, $x_2 = 8$) is negative. This causes y for week five to be lower than $\mu_{y|45.9,8}$ (mean fuel consumption when $x_1 = 45.9$ and $x_2 = 8$). Here factors other than x_1 and x_2 (for instance, lower-than-usual thermostat settings) have resulted in a negative error term.

The fuel consumption model expresses the dependent variable as a function of two independent variables. In general, we can use a multiple regression model to express a dependent variable as a function of any number of independent variables. For example, the Cincinnati Gas and Electric Company predicts daily natural gas consumption as a function of four independent variables—average temperature, average wind velocity, average sunlight, and change in average temperature from the previous day. The general form of a multiple regression model expresses the dependent variable y as a function of k independent variables $x_1, x_2, \ldots, x_k$. We call this general form the (multiple) *linear regression model* and express it as shown in the following box.

The linear regression model

The linear regression model relating y to $x_1, x_2, \ldots, x_k$ is

$$y = \mu_{y|x_1, x_2, \ldots, x_k} + \varepsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

The linear regression model (Continued)

Here

1. $\mu_{y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ is the mean value of the dependent variable y when the values of the independent variables are $x_1, x_2, \ldots, x_k$.
2. $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are (unknown) *regression parameters* relating the mean value of y to $x_1, x_2, \ldots, x_k$.
3. ε is an *error term* that describes the effects on y of all factors other than the values of the independent variables $x_1, x_2, \ldots, x_k$.

2.2.2 The Least Squares Point Estimates Using Matrix Algebra

The regression parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ in the linear regression model are unknown. Therefore, they must be estimated from sample data. We assume that we have obtained n observations, with each observation consisting of an observed value of y and corresponding observed values of $x_1, x_2, \ldots, x_k$. For $i = 1, 2, \ldots, n$, we let y_i denote the ith observed value of y, and we let $x_{i1}, x_{i2}, \ldots, x_{ik}$ denote the ith observed values of $x_1, x_2, \ldots, x_k$. If $b_0, b_1, b_2, \ldots, b_k$ denote point estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ then a point prediction of

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

is

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik}$$

Here, since the regression assumptions to be discussed in Section 2.3 imply that the error term ε_i has a 50 percent chance of being positive and a 50 percent chance of being negative, we predict ε_i to be 0. Intuitively, if any particular values of $b_0, b_1, b_2, \ldots, b_k$ are good point estimates, they will make (for $i = 1, 2, \ldots, n$) $\hat{y}_i$ close to y_i and thus the *residual*

$e_i = y_i - \hat{y}_i$ small. We define the *least squares points estimates* to be the values $b_0, b_1, b_2, \ldots, b_k$ that minimize the *sum of squared residuals*

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Using calculus (see Section B.2), it can be shown that the least squares point estimates can be calculated by using a formula involving *matrix algebra*. We now discuss matrix algebra and explain the formula.

A *matrix* is rectangular array of numbers (called *elements*) that is composed of rows and columns. Matrices are denoted by boldface letters. For example, we will use two matrices to calculate the least squares point estimates of the parameters β_0, β_1 and β_2 in the fuel consumption model

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

These matrices are

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix} = \begin{bmatrix} 12.4 \\ 11.7 \\ 12.4 \\ 10.8 \\ 9.4 \\ 9.5 \\ 8.0 \\ 7.5 \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \\ 1 & x_{51} & x_{52} \\ 1 & x_{61} & x_{62} \\ 1 & x_{71} & x_{72} \\ 1 & x_{81} & x_{82} \end{bmatrix} = \begin{bmatrix} 1 & 28.0 & 18 \\ 1 & 28.0 & 14 \\ 1 & 32.5 & 24 \\ 1 & 39.0 & 22 \\ 1 & 45.9 & 8 \\ 1 & 57.8 & 16 \\ 1 & 58.1 & 1 \\ 1 & 62.5 & 0 \end{bmatrix}$$

Here, the matrix $\mathbf{y}$ consists of a single column containing the eight observed weekly fuel consumptions $y_1 = 12.4$, $y_2 = 11.7$, … , $y_8 = 7.5$ (see Table 2.3). In addition, the matrix $\mathbf{X}$ consists of three columns containing the observed values of the independent variables corresponding to (that is, multiplied by) the three parameters in the model. Therefore, since the number 1 is multiplied by β_0, the column of the $\mathbf{X}$ matrix corresponding to β_0 is a column of 1s. Since the independent variable x_1 is multiplied by β_1, the column of the $\mathbf{X}$ matrix corresponding to β_1 is a column containing the observed average hourly temperatures $x_{11} = 28,\ x_{21} = 28, \ldots, x_{81} = 62.5$. The independent variable x_2 is multiplied by β_2, and thus the column of the **X** matrix corresponding to β_2 is a column containing the observed chill indices $x_{12} = 18,\ x_{22} = 14, \ldots, x_{82} = 0$.

The *dimension* of a matrix is determined by the number of rows and columns in the matrix. Since the matrix **X** has eight rows and three columns, this matrix is said to have dimension 8 by 3 (commonly written 8×3). In general, a matrix with m rows and n columns is said to have dimension $m \times n$. As another example, the matrix **y** has eight rows and one column. In general, a matrix having one column is called a *column vector*. In order to use the matrix **X** and column vector **y** to calculate the least squares point estimates, we first define the *transpose of* **X**.

The *transpose* of a matrix is formed by interchanging the rows and columns of the matrix. For example, the transpose of the matrix **X**, which we denote as **X′** is

$$\mathbf{X}' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 28.0 & 28.0 & 32.5 & 39.0 & 45.9 & 57.8 & 58.1 & 62.5 \\ 18 & 14 & 24 & 22 & 8 & 16 & 1 & 0 \end{bmatrix}$$

We next multiply **X′** by **X** and **X′** by **y**. To see how to do this, we need to discuss how to multiply two matrices together. Consider two matrices **A** and **B** where the number of *columns* in **A** equals the number of *rows* in **B**. Then the *product of the two matrices* **A** *and* **B** is a matrix calculated so that the element in row i and column j of the product is obtained by multiplying the elements in row i of matrix **A** by the corresponding elements in column j of matrix **B** and adding the resulting products.

In general, we can multiply a matrix **A** with m rows and r columns by a matrix **B** with r rows and n columns and obtain a matrix **C** with m rows and n columns. Moreover, c_{ij}, the number in the product in row i and column j, is obtained by multiplying the elements in row i of **A** by the corresponding elements in column j of **B** and adding the resulting products. Note that the number of columns in **A** must equal the number of rows in **B** in order for this multiplication procedure to be defined. The multiplication procedure is illustrated in Figure 2.9.

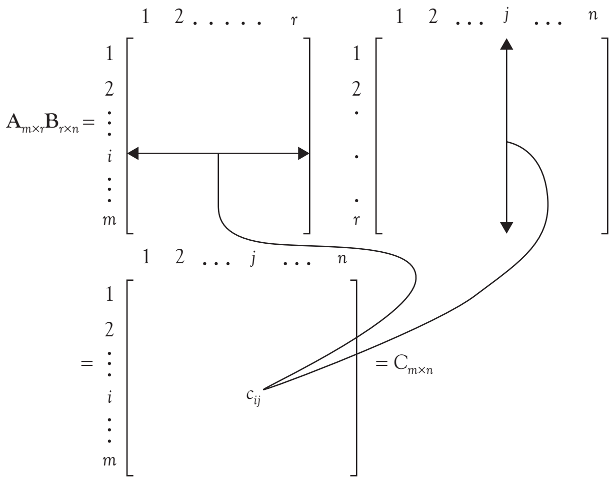


Figure 2.9 An illustration of matrix multiplication

We multiply $\mathbf{X}'$ by $\mathbf{X}$ as follows:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 28.0 & 28.0 & 32.5 & 39.0 & 45.9 & 57.8 & 58.1 & 62.5 \\ 18 & 14 & 24 & 22 & 8 & 16 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 28.0 & 18 \\ 1 & 28.0 & 14 \\ 1 & 32.5 & 24 \\ 1 & 39.0 & 22 \\ 1 & 45.9 & 8 \\ 1 & 57.8 & 16 \\ 1 & 58.1 & 1 \\ 1 & 62.5 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 8.0 & 351.8 & 103.0 \\ 351.8 & 16874.76 & 3884.1 \\ 103.0 & 3884.1 & 1901.0 \end{bmatrix}$$

To understand this matrix multiplication, note that $\mathbf{X}'$ has three rows and eight columns and that $\mathbf{X}$ has eight rows and three columns. Therefore, since the number of columns of $\mathbf{X}'$ equals the number of rows of $\mathbf{X}$, we can multiply the two matrices together. Furthermore, since $\mathbf{X}'$ has three rows and $\mathbf{X}$ has three columns, multiplying $\mathbf{X}'$ by $\mathbf{X}$ will result in a matrix $\mathbf{X}'\mathbf{X}$ that has three rows and three columns. To obtain the element in row 1 and column 1 of $\mathbf{X}'\mathbf{X}$, we multiply the elements in row 1 of $\mathbf{X}'$ by the corresponding elements in column l of **X** and add up the resulting products as follows:

$$(1)(1)+(1)(1)+(1)(1)+(1)(1)+(1)(1)+(1)(1)+(1)(1)+(1)(1)=8$$

To obtain the element in row 1 and column 2 of $\mathbf{X'X}$, we multiply the elements in row 1 of $\mathbf{X'}$ by the corresponding elements in column 2 of **X** and add up the resulting products as follows:

$$(1)(28.0) + (1)(28.0) + (1)(32.5) + (1)(39.0) + (1)(45.9) + (1)(57.8) + (1)(58.1) + (1)(62.5) = 351.8$$

Continuing this process, we obtain all the elements of $\mathbf{X'X}$. As one final example, we obtain the element in row 2 and column 3 of $\mathbf{X'X}$ by multiplying the elements in row 2 of $\mathbf{X'}$ by the corresponding elements in column 3 of **X** and adding up the resulting products as follows:

$$(28.0)(18) + (28.0)(14) + (32.5)(24) + (39.0)(22) + (45.9)(8) + (57.8)(16) + (58.1)(1) + (62.5)(0) = 3{,}884.1$$

We continue using matrix multiplication and multiply $\mathbf{X'}$ by $\mathbf{y}$ as follows:

$$\mathbf{X'y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 28.0 & 28.0 & 32.5 & 39.0 & 45.9 & 57.8 & 58.1 & 62.5 \\ 18 & 14 & 24 & 22 & 8 & 16 & 1 & 0 \end{bmatrix} \begin{bmatrix} 12.4 \\ 11.7 \\ 12.4 \\ 10.8 \\ 9.4 \\ 9.5 \\ 8.0 \\ 7.5 \end{bmatrix}$$

$$= \begin{bmatrix} 81.7 \\ 3413.11 \\ 1157.4 \end{bmatrix}$$

We next consider the matrix

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 5.43405 & -.085930 & -.118856 \\ -.085930 & .00147070 & .00165094 \\ -.118856 & .00165094 & .00359276 \end{bmatrix}$$

This matrix is called the *inverse* of $\mathbf{X}'\mathbf{X}$ because if we multiply $\mathbf{X}'\mathbf{X}$ by this matrix we obtain the *identity matrix*

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

In general, for a matrix $\mathbf{A}$ to have an inverse, it must be *square* (that is, its number of rows must equal its number of columns) and it must have *linearly independent columns* (that is, no one column can be expressed as a linear combination of the other columns). Then, the inverse of $\mathbf{A}$, denoted $\mathbf{A}^{-1}$, is another matrix such that if we multiply $\mathbf{A}$ by this other matrix we obtain the *identity matrix* (that is, a square matrix with 1s running down the main diagonal—from the upper left to the lower right—and 0s elsewhere). To intuitively illustrate the idea of linear independence, consider the following matrix $\mathbf{A}$ and the following vectors $\mathbf{c}$ and $\mathbf{d}$:

$$\mathbf{A} = \begin{bmatrix} 3 & 1 & 2 \\ 1 & .5 & 0 \\ 2 & 0 & 4 \end{bmatrix} \quad \mathbf{c} = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \quad \mathbf{d} = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}$$

The elements in the column vector $\mathbf{c}$ are obtained by multiplying the elements in the second column of the matrix $\mathbf{A}$ by 2, and the elements in the column vector $\mathbf{d}$ are obtained by multiplying the elements in the third column of the matrix $\mathbf{A}$ by .5. Moreover, the elements in the first column of $\mathbf{A}$ are found by adding the corresponding elements of $\mathbf{c}$ and $\mathbf{d}$ together. This implies that all of the columns of $\mathbf{A}$ are not linearly independent and thus $\mathbf{A}$ does not have an inverse. However, in this book we define each matrix $\mathbf{X}$ in regression analysis so that all of its columns are linearly independent. This can be shown to imply that all of the columns of $\mathbf{X'X}$ are linearly independent and thus $\mathbf{X'X}$ has an inverse. We obtain the inverse by using a statistical software package (there is a hand calculation procedure for obtaining inverses, but we will not discuss it).

In order to obtain the least squares point estimates b_0, b_1, and b_2 of the parameters β_0, β_1, and β_2 in the fuel consumption model

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

we multiply $(\mathbf{X'X})^{-1}$ by $\mathbf{X'y}$ as follows:

$$\begin{aligned}
\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} &= \mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y} \\
&= \begin{bmatrix} 5.43405 & -.085930 & -.118856 \\ -.085930 & .00147070 & .00165094 \\ -.118856 & .00165094 & .00359276 \end{bmatrix} \begin{bmatrix} 81.7 \\ 3413.11 \\ 1157.4 \end{bmatrix} \\
&= \begin{bmatrix} 13.1087 \\ -.09001 \\ .08249 \end{bmatrix}
\end{aligned}$$

We will interpret the meanings of these least squares point estimates in the next example. First, however, we give a general matrix algebra formula for calculating the least squares point estimates $b_0, b_1, b_2, \ldots, b_k$ of the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ in the linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

The general matrix algebra formula uses the following matrices:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & . & . & . & k \\ & x_1 & x_2 & . & . & . & x_k \end{array} \\ \begin{bmatrix} 1 & x_{11} & x_{12} & \cdot & \cdot & \cdot & x_{1k} \\ 1 & x_{21} & x_{22} & \cdot & \cdot & \cdot & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdot & \cdot & \cdot & x_{nk} \end{bmatrix} \end{array}$$

Here, $\mathbf{y}$ is a column vector containing the n observed values of the dependent variable, $y_1, y_2, \ldots, y_n$. Moreover, because the linear regression model uses $(k+1)$ parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$, the matrix $\mathbf{X}$ consists of $(k+1)$ columns. The columns in the matrix $\mathbf{X}$ contain the observed values of the independent variables corresponding to (that is, multiplied by) the $(k+1)$ parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$. The columns of this matrix are numbered in the same manner as the parameters are numbered (see the preceding $\mathbf{X}$ matrix). The general matrix algebra formula is then as follows:

The least squares point estimates

The *least squares point estimates* $b_0, b_1, b_2, \ldots, b_k$ are calculated by using the formula

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

We have demonstrated using this formula in calculating the least squares point estimates of the parameters in the fuel consumption model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. It is also important to note that when we use the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ to relate a dependent variable y to a single independent variable x, then the column vector $\mathbf{y}$ and the matrix $\mathbf{X}$ used to calculate the least squares point estimates b_0 *and* b_1 of the parameters β_0 and β_1 are

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Here, $y_1, y_2, \ldots, y_n$ are the n observed values of y, and $x_1, x_2, \ldots, x_k$ are the n observed values of x. By using this $\mathbf{y}$ vector and $\mathbf{X}$ matrix it can be shown that

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} \bar{y} - b_1\bar{x} \\ \dfrac{SS_{xy}}{SS_{xx}} \end{bmatrix}$$

These are the same formulas for b_0 and b_1 that we presented in Section 2.1.

Example 2.4

Figure 2.10 is the Minitab output of a regression analysis of the fuel consumption data in Table 2.3 by using the model

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

This output shows that the least squares point estimates of $\beta_0, \beta_1,$ and β_2 are $b_0 = 13.1087$, $b_1 = -.09001$, and $b_2 = .08249$, as have been calculated previously using matrices.

The point estimate $b_1 = -.09001$ of β_1 says we estimate that mean weekly fuel consumption decreases (since b_1 is negative) by .09001 MMcf of natural gas when average hourly temperature increases by one degree and the chill index does not change. The point estimate $b_2 = .08249$ of β_2 says we estimate that mean weekly fuel consumption increases (since b_2 is positive) by .08249 MMcf of natural gas when there is a one-unit increase in the chill index and average hourly temperature does not change.

The equation

$$\begin{aligned}\hat{y} &= b_0 + b_1 x_1 + b_2 x_2 \\ &= 13.1087 - .09001 x_1 + .08249 x_2\end{aligned}$$

is called the *least squares prediction equation*. It is obtained by replacing $\beta_0, \beta_1,$ and β_2 by their estimates $b_0, b_1,$ and b_2 and by predicting the error term to be zero. This equation is given on the Minitab output (labeled as the "regression equation"—note that $b_0, b_1,$ and b_2 have been rounded to 13.1, –.0900, and .0825). We can use this equation to compute a prediction for any observed value of y. For instance, a point prediction of y_1 = 12.4 (when $x_1 = 28.0$ and $x_2 = 18$) is

$$\hat{y}_1 = 13.1087 - .09001(28.0) + .08249(18) = 12.0733$$

This results in a residual equal to

$$e_1 = y_1 - \hat{y}_1 = 12.4 - 12.0733 = .3267$$

Table 2.4 gives the point prediction obtained using the least squares prediction equation and the residual for each of the eight observed fuel consumption values. In addition, this table tells us that the *SSE* equals .674.

The least squares prediction equation is the equation of a plane that we sometimes call the *least squares plane*. For combinations of values of x_1 and x_2 that are in the experimental region, the *least squares plane* is the estimate of the *plane of means* (see Figure 2.8). This implies that the point on the least squares plane corresponding to the average hourly temperature x_1 and the chill index x_2

$$\begin{aligned}\hat{y} &= b_0 + b_1 x_1 + b_2 x_2 \\ &= 13.1087 - .09001 x_1 + .08249 x_2\end{aligned}$$

is the point estimate of $\mu_{y|x_1,x_2}$, the mean of all the weekly fuel consumptions that could be observed when the average hourly temperature is x_1 and the chill index is x_2. In addition, since we predict the error term to be zero, $\hat{y}$ is also the point prediction of $y = \mu_{y|x_1,x_2} + \varepsilon$, which is the

Table 2.4 Predictions and residuals using the least squares point estimates $b_0 = 13.1$, $b_1 = -.0900$, ***and*** $b_2 = .0825$

Week	x_1	x_2	y	$\hat{y} = 13.1 - .0900x_1 + 0.0825x_2$	$e = y - \hat{y}\hat{y}$
1	28.0	18	12.4	12.0733	.3267
2	28.0	14	11.7	11.7433	–.0433
3	32.5	24	12.4	12.1632	.2368
4	39.0	22	10.8	11.4131	–.6131
5	45.9	8	9.4	9.6371	–.2371
6	57.8	16	9.5	9.2259	.2741
7	58.1	1	8.0	7.9614	.0386
8	62.5	0	7.5	7.4829	.0171
SSE = $(.3267)^2 + (-.0433)^2 + \ldots + (.0171)^2 = .674$					

amount of fuel consumed in a single week when the average hourly temperature is x_1 and the chill index is x_2.

For example, suppose a weather forecasting service predicts that in the next week the average hourly temperature will be 40°F and the chill index will be 10. Since this combination is inside the experimental region (see Figure 2.7), we see that

$$\begin{aligned}\hat{y} &= 13.1087 - .09001(40) + .08249(10) \\ &= 10.333 \ \text{ MMcf of natural gas}\end{aligned}$$

is

1. The point estimate of the mean weekly fuel consumption when the average hourly temperature is 40°F and the chill index is 10.
2. The point prediction of the amount of fuel consumed in a single week when the average hourly temperature is 40°F and the chill index is 10.

Notice that $\hat{y} = 10.333$ is given at the bottom of the Minitab output in Figure 2.10. Also, note that Figure 2.11 is the Minitab output that results from using the data in Figure 2.1 and the simple linear regression model

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

to relate y = weekly fuel consumption to the single independent variable x = average hourly temperature. This output gives the least squares point estimates $b_0 = 15.837$ and $b_1 = -.12792$ that we have calculated in Example 2.2, as well as $\hat{y} = 15.837 - .12792(40) = 10.721$, the point estimate of mean weekly fuel consumption and the point prediction of an individual weekly fuel consumption when average hourly temperature is 40°F. Of course, the values of x = average hourly temperature in Figure 2.1 that are used to help fit the model $y = \beta_0 + \beta_1 x_1 + \varepsilon$ are the same as the values of x = average hourly temperature in Table 2.3 that are used to help fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. Throughout the rest of this chapter we will use the Minitab outputs in Figures 2.10 and 2.11 to help compare these models and assess whether the extra independent variable x_2 = the chill index makes the second model more likely to give a more accurate prediction of future weekly fuel consumptions.

```
The regression equation is
FUELCONS = 13.1 - 0.0900 TEMP + 0.0825 CHILL

Predictor        Coef    SE Coef[d]      T[e]       P[f]
Constant   13.1087[a]      0.8557     15.32      0.000
TEMP      -0.09001[b]     0.01408     -6.39      0.001
CHILL      0.08249[c]     0.02200      3.75      0.013

S = 0.367078[g]    R-Sq = 97.4%[h]    R-Sq(adj) = 96.3%

Analysis of Variance

Source            DF          SS          MS          F          P
Regression         2   24.875[i]      12.438   92.30[l]   0.000[m]
Residual Error     5    0.674[j]       0.135
Total              7   25.549[k]

       Fit[n]  SE Fit[o]      95% CI[p]              95% PI[q]
     10.333      0.170    (9.895, 10.771)       (9.293, 11.374)
```

Figure 2.10 Minitab output of a regression analysis using the fuel consumption model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

[a] b_0 [b] b_1 [c] b_2 [d] s_{b_j} [e] t-statistics [f] p-values for t-statistics [g] s = standard error [h] R^2 [i] Explained variation [j] SSE = unexplained variation [k] Total variation [l] F(model) statistic [m] p-value for F(model) [n] $\hat{y}$ [o] $s_{\hat{y}}$ [p] 95% confidence interval when $x_1 = 40$ and $x_2 = 10$ [q] 95% prediction interval when $x_1 = 40$ and $x_2 = 10$

```
The regression equation is
FUELCONS = 15.8 - 0.128 TEMP

Predictor        Coef     SE Coef          T          P[g]
Constant   15.8379[a]   0.8018[c]   19.75[e]       0.000
TEMP      -0.12792[b]  0.01746[d]   -7.33[f]       0.000

S = 0.654209[h]    R-Sq = 89.9%[i]    R-Sq(adj) = 88.3%

Analysis of Variance

Source            DF          SS          MS          F          P
Regression         1   22.981[j]      22.981   53.69[m]   0.000[n]
Residual Error     6    2.568[k]       0.428
Total              7   25.549[l]

       Fit[o]  SE Fit[p]      95% CI[q]              95% PI[r]
     10.721     0.241    (10.130, 11.312)      (9.015, 12.427)
```

Figure 2.11 Minitab output of a regression analysis using the fuel consumption model $y = \beta_0 + \beta_1 x + \varepsilon$, where x = average hourly temperature

[a] b_0 [b] b_1 [c] s_{b_0} [d] s_{b_1} [e] t for testing $H_0: \beta_0 = 0$ [f] t for testing $H_0: \beta_1 = 0$ [g] p-values for t-statistics [h] s = standard error [i] r^2 [j] Explained variation [k] SSE = Unexplained variation [l] Total variation [m] F(model) statistic [n] p-value for F(model) [o] $\hat{y}$ when $x = 40$ [p] $s_{\hat{y}}$ [q] 95% confidence interval when $x = 40$ [r] 95% prediction interval when $x = 40$

Point estimation and point prediction in multiple regression

Let $b_0, b_1, b_2, \ldots, b_k$ be the least squares point estimates of the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ in the linear regression model, and suppose that $x_{01}, x_{02}, \ldots, x_{0k}$ are specified values of the independent variables $x_1, x_2, \ldots, x_k$. If the combination of specified values is inside the experimental region, then

$$\hat{y} = b_0 + b_1 x_{01} + b_2 x_{02} + \ldots + b_k x_{0k}$$

is the *point estimate* of the *mean value of the dependent variable* when the values of the independent variables are $x_{01}, x_{02}, \ldots, x_{0k}$. In addition, $\hat{y}$ is the *point prediction* of an *individual value of the dependent variable* when the values of the independent variables are $x_{01}, x_{02}, \ldots, x_{0k}$. Here we predict the error term to be zero.

Example 2.5

Suppose the sales manager of a company wishes to evaluate the performance of the company's sales representatives. Each sales representative is solely responsible for one sales territory, and the manager decides that it is reasonable to measure the performance, y, of a sales representative by using the yearly sales of the company's product in the representative's sales territory. The manager feels that sales performance y substantially depends on five independent variables:

x_1 = number of months the representative has been employed by the company (Time)
x_2 = sales of the company's product and competing products in the sales territory (MktPoten)
x_3 = dollar advertising expenditure in the territory (Adver)
x_4 = weighted average of the company's market share in the territory for the previous four years (MktShare)
x_5 = change in the company's market share in the territory over the previous four years (Change)

In Table 2.5(a) we present values of y and x_1 through x_5 for 25 randomly selected sales representatives. To understand the values of y and x_2 in the table, note that sales of the company's product or any competing product are measured in hundreds of units of the product sold. Therefore, for example, the first sales figure of 3669.88 in Table 2.5(a) means that the first randomly selected sales representative sold 366,988 units of the company's product during the year.

Plots of y versus x_1 through x_5 are given in Table 2.5(b). Since each plot has an approximate straight-line appearance, it is reasonable to relate y to x_1 through x_5 by using the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

Here, $\mu_{y|x_1,x_2,\ldots,x_5} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$ is, intuitively, the mean sales in all sales territories where the values of the previously described five independent variables are x_1, x_2, x_3, x_4, and x_5. Furthermore, for example, the parameter β_3 equals the increase in mean sales that is associated with a \$1 increase in advertising expenditure (x_3) when the other four independent variables do not change. The main objective of the regression analysis is to help the sales manager evaluate sales performance by comparing actual performance to predicted performance. The manager has randomly selected the 25 representatives from all the representatives the company considers to be effective and wishes to use a regression model based on effective representatives to evaluate questionable representatives. Questionable representatives whose performance is substantially lower than performance predictions will get special training aimed at improving their sales techniques.

By using the data in Table 2.5(a) we define the column vector $\mathbf{y}$ and matrix $\mathbf{X}$ as follows:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{25} \end{bmatrix} = \begin{bmatrix} 3669.88 \\ 3473.95 \\ \vdots \\ 2799.97 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 & x_2 & x_3 & x_4 & x_5 \\ 1 & 43.10 & 74065.11 & 4582.88 & 2.51 & .34 \\ 1 & 108.13 & 58117.30 & 5539.78 & 5.51 & .15 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 21.14 & 22809.53 & 3552.00 & 9.14 & -.74 \end{bmatrix}$$

Table 2.5 Sales territory performance data, data plots, and regression

(a) The data

Sales	Time	Mkt-Poten	Adver	Mkt-Share	Change
3,669.88	43.10	74.065.11	4,582.88	2.51	0.34
3,473.95	108.13	58,117.30	5,539.78	5.51	0.15
2,295.10	13.82	21,118.49	2,950.38	10.91	–0.72
4,675.56	186.18	68,521.27	2,243.07	8.27	0.17
6,125.96	161.79	57,805.11	7,747.08	9.15	0.50
2,134.94	8.94	37,806.94	402.44	5.51	0.15
5,031.66	365.04	50,935.26	3,140.62	8.54	0.55
3,367.45	220.32	35,602.08	2,086.16	7.07	–0.49
6,519.45	127.64	46,176.77	8,846.25	12.54	1.24
4,876.37	105.69	42,053.24	5,673.11	8.85	0.31
2,468.27	57.72	36,829.71	2,761.76	5.38	0.37
2,533.31	23.58	33,612.67	1,991.85	5.43	–0.65
2,408.11	13.82	21,412.79	1,971.52	8.48	0.64
2,337.38	13.82	20,416.87	1,737.38	7.80	1.01
4,586.95	86.99	36,272.00	10,694.20	10.34	0.11
2,729.24	165.85	23,093.26	8,618.61	5.15	0.04
3,289.40	116.26	26,878.59	7,747.89	6.64	0.68
2,800.78	42.28	39,571.96	4,565.81	5.45	0.66
3,264.20	52.84	51,866.15	6,022.70	6.31	–0.10
3,453.62	165.04	58,749.82	3,721.10	6.35	–0.03
1,741.45	10.57	23,990.82	860.97	7.37	–1.63
2,035.75	13.82	25,694.86	3,571.51	8.39	–0.43
1,578.00	8.13	23,736.35	2,845.50	5.15	0.04
4,167.44	58.54	34,314.29	5,060.11	12.88	0.22
2,799.97	21.14	22,809.53	3,552.00	9.14	–0.74

Source: This dataset is from a research study published by Cravens, Woodruff, and Stamper (1972). We have updated the situation in our case study to be more modern.

(b) Data plots

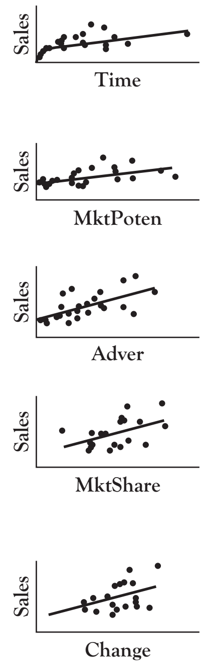


Table 2.5 (*Continued*)

(c) SAS output of a regression analysis using the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	37862659[a]	7572532	40.91[d]	<.0001[e]
Error	19	3516890[b]	185099		
Corrected Total	24	41379549[c]			

Root MSE	430.23189[k]	R-Square	0.9150[f]
Dependent Mean	3374.56760	Adj R-Sq	0.8926
Coeff Var	12.74924		

Parameter Estimates

Variable	Label	DF	Parameter[g] Estimate	Standard[h] Error	t Value[i]	Pr > |t|[j]
Intercept	Intercept	1	-1113.78788	419.88690	-2.65	0.0157
Time	Time	1	3.61210	1.18170	3.06	0.0065
MktPoten	MktPoten	1	0.04209	0.00673	6.25	<.0001
Adver	Adver	1	0.12886	0.03704	3.48	0.0025
MktShare	MktShare	1	256.95554	39.13607	6.57	<.0001
Change	Change	1	324.53345	157.28308	2.06	0.0530

Obs	Dep Var Sales	Predicted[l] Value	Std Error[o] Mean Predict	95% CL Mean[m]		95% CL Predict[n]	
26	.	4182	141.8220	3885	4479	3234	5130

[a]Explained variation [b]SSE = unexplained variation [c]Total variation [d]F(model) [e]p-value for F(model) [f]R^2 [g]b_j [h]s_{b_j} [i]t-statistic [j]p-value for t-statistic [k]s = standard error [l]$\hat{y}$ [m]95 percent confidence interval for mean [n]95 percent prediction interval [o]$s_{\hat{y}}$

If the appropriate matrix calculations are then done, the equation $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ then tells us that the least squares point estimates of the parameters $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$, and β_5 in the sales territory performance regression model are $b_0 = -1113.7879$, $b_1 = 3.6121$, $b_2 = .0421$, $b_3 = .1289$, $b_4 = 256.9555$, and $b_5 = 324.5335$. These point estimates are shown in Table 2.5(c), which is the SAS output of a regression analysis using the sales territory performance regression model. On this output x_1, x_2, x_3, x_4, and x_5 are denoted as Time, MktPoten, Adver, MktShare, and Change, respectively. Recalling that the sales values in Table 2.5(a) are measured in hundreds of units of the product sold, the point estimate $b_3 = .1289$ says we estimate that mean sales increase by .1289 hundreds of units—that is, by 12.89 units—for each dollar increase in advertising expenditure when the other four independent variables do not change. If the company sells each unit for \$1.10, this implies that we estimate that mean sales revenue increases by $(\$1.10)(12.89) = \$14.18$ for each dollar increase in advertising expenditure when the other four independent variables do not change. The other β values in the model can be interpreted similarly.

Consider a questionable sales representative for whom Time = 85.42, MktPoten = 35,182.73, Adver = 7281.65, MktShare = 9.64, and Change = .28. The point prediction of the sales corresponding to this combination of values of the independent variables is

$$\begin{aligned}\hat{y} &= -1113.7879 + 3.6121(85.42) + .0421(35{,}182.73)\\ &\quad + .1289(7281.65) + 256.9555(9.64) + 324.5335(.28)\\ &= 4182 \text{ (that is, 418,200 units)}\end{aligned}$$

which is given on the SAS output. The actual sales for the questionable sales representative were 3088. This sales figure is 1094 less than the point prediction $\hat{y} = 4182$. However, we will have to wait until we study *prediction intervals* to determine whether there is strong evidence that the actual sales figure is unusually low. In the exercises, the reader will further analyze the sales territory performance data by using techniques (including prediction intervals) that will be discussed in the rest of this chapter.

2.3 Model Assumptions, Sampling, and the Standard Error

2.3.1 Model Assumptions

In order to perform hypothesis tests and set up various types of intervals when using the linear regression model

$$\begin{aligned} y &= \mu_{y|x_1,x_2,\ldots,x_k} + \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \end{aligned}$$

we need to make certain assumptions about the error term ε. At any given combination of values of $x_1, x_2, \ldots, x_k$, there is a population of error term values that could potentially occur. These error term values describe the different potential effects on y of all factors other than the given combination of values of $x_1, x_2, \ldots, x_k$. Therefore, these error term values explain the variation in the y values that could be observed at the given combination of values of $x_1, x_2, \ldots, x_k$. Our statement of the linear regression model assumes that $\mu_{y|x_1,x_2,\ldots,x_k}$, the mean of the population of all y values that could be observed when the independent variables are $x_1, x_2, \ldots, x_k$, is $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$. This model also implies that $\varepsilon = y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$, so this is equivalent to assuming that the mean of the population of potential error term values that could occur at a given combination of values of $x_1, x_2, \ldots, x_k$, is zero. In total, we make four assumptions—called the *regression assumptions*—about the linear regression model. Stated in terms of potential error term values, these assumptions are as follows.

Assumptions for the linear regression model

1. At any given combination of values of $x_1, x_2, \ldots, x_k$, the population of potential error term values has a mean equal to 0.
2. *Constant variance assumption*: At any given combination of values of $x_1, x_2, \ldots, x_k$, the population of potential error term values has a variance that does not depend on the combination of values of $x_1, x_2, \ldots, x_k$. That is, the different populations of potential error term values corresponding to different combinations of values of $x_1, x_2, \ldots, x_k$ have equal variances. We denote the constant variance as σ^2.

3. *Normality assumption*: At any given combination of values of $x_1, x_2, \ldots, x_k$, the population of potential error term values has a *normal distribution*.
4. *Independence assumption*: Any one value of the error term ε is *statistically independent* of any other value of ε. That is, the value of the error term ε corresponding to an observed value of y is statistically independent of the error term corresponding to any other observed value of y.

Taken together, the first three regression assumptions say that at any given combination of values of $x_1, x_2, \ldots, x_k$, the population of potential error term values is normally distributed with mean zero and a variance σ^2 that does not depend on the combination of values of $x_1, x_2, \ldots, x_k$. The model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

implies that at any given combination of values of $x_1, x_2, \ldots, x_k$, the variation in the y values is caused by and thus is the same as the variation in the ε values. Therefore, the first three regression assumptions imply that at any given combination of values of $x_1, x_2, \ldots, x_k$, the population of y values that could be observed is normally distributed with mean $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ and a variance σ^2 that does not depend on the combination of values of $x_1, x_2, \ldots, x_k$. These three assumptions are illustrated in Figure 2.12 in the context of the simple linear regression

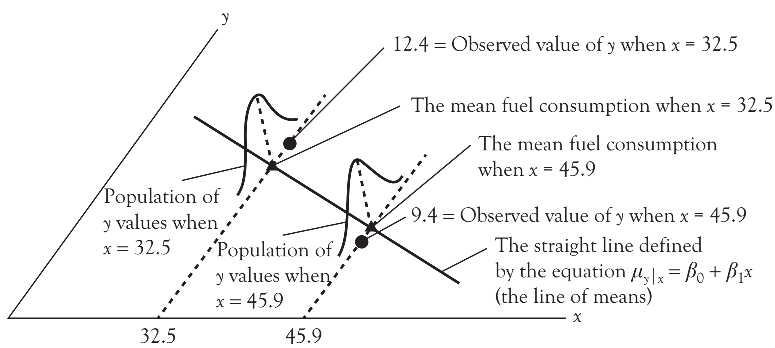


Figure 2.12 An illustration of the regression assumptions

model $y = \mu_{y|x} + \varepsilon = \beta_0 + \beta_1 x + \varepsilon$ relating y = weekly fuel consumption to x = average hourly temperature. Specifically, this figure depicts the populations of weekly fuel consumptions corresponding to two values of average hourly temperature—32.5 and 45.9. Note that these populations are shown to be normally distributed with different means (each of which is on the line of means) and with the same variance (or spread) σ^2. To illustrate the first three regression assumptions using the two independent variable fuel consumption model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, consider for example, the following two populations: The population of all weekly fuel consumptions that could be observed when the average hourly temperature is 32.5°F and the chill index is 24, and the population of all weekly fuel consumptions that could be observed when the average hourly temperature is 45.9°F and the chill index is 8. Then, the first three regression assumptions say that, although these two populations have different means of, respectively, $\beta_0 + \beta_1(32.5) + \beta_2(24)$ and $\beta_0 + \beta_1(45.9) + \beta_2(8)$, both populations are normally distributed with the same variance σ^2.

The independence assumption is most likely to be violated when time series data are utilized in a regression study. Intuitively, this assumption says that there is no pattern of positive error terms being followed (in time) by other positive error terms, and there is no pattern of positive error terms being followed by negative error terms. That is, there is no pattern of higher-than-average y values being followed by other higher-than-average y values, and there is no pattern of higher-than-average y values being followed by lower-than-average y values.

It is important to point out that the regression assumptions very seldom, if ever, hold exactly in any practical regression problem. However, it has been found that regression results are not extremely sensitive to mild departures from these assumptions. In practice, only pronounced departures from these assumptions require attention. In Chapter 4 we show how to check the regression assumptions. Until then, we will suppose that the assumptions are valid in our examples.

In Sections 2.1 and 2.2 we stated that when we predict an individual value of the dependent variable, we predict the error term to be zero. To see why we do this, note that the regression assumptions state that at any given value of the independent variable, the population of all error term values that can potentially occur is normally distributed with a mean equal to zero. Since we also assume that successive error terms (observed over time) are statistically independent, each error term has a 50 percent chance of being positive and a 50 percent chance of being negative. Therefore**,** it is reasonable to predict any particular error term value to be zero.

2.3.2 *Sampling and the Unbiased Least Squares Point Estimates*

The least squares point estimates $b_0, b_1, b_2, \ldots, b_k$ of the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ of the *linear regression model* are calculated by using the matrix algebra equation $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and thus depend upon the n observed values $y_1, y_2, \ldots, y_n$ of the dependent variable y. Considered before y_i was actually observed, y_i could have been any value in the normally distributed population of all possible values of the dependent variable that could be observed when the values of the independent variables are $x_{i1}, x_{i2}, \ldots, x_{ik}$. This is true for each of $y_1, y_2, \ldots, y_n$, and thus there are an infinite number of different possible samples (or sets) of n values $y_1, y_2, \ldots, y_n$ of the dependent variable that could have been observed. Because each of these samples would yield its own unique values of $b_0, b_1, b_2, \ldots, b_k$, there is an infinite population of potential values of each of these least squares point estimates.

For example, consider the fuel consumption regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. Corresponding to each of the eight observed combinations of the average hourly temperature and the chill index, there is a normally distributed population of possible weekly fuel consumptions that could be observed. For example, (1) there is a normally distributed population of possible weekly fuel consumptions that could be observed when the average hourly temperature is 28.0 and the chill index is 18 (as occurred in week 1); (2) there is a normally distributed population of possible weekly fuel consumptions that could be observed when the average hourly temperature is 28.0 and the chill index is 14 (as occurred in week 2); . . . ; (8) there is a normally distributed population of possible weekly fuel consumptions that could be observed when the average hourly temperature is 62.5 and the chill index is 0 (as occurred in week 8). Sample 1 in Table 2.6 is the sample of eight weekly fuel consumptions that we have actually observed from the eight normally distributed populations of possible weekly fuel consumptions. In Section 1.2 we have used sample 1 to calculate the least squares point estimates $b_0 = 13.1087$, $b_1 = -.09001$, and $b_2 = .08249$, which are shown following sample 1 in Table 2.6. Samples 2 and 3 in Table 2.6 are two other samples of eight weekly fuel consumptions that we could have

Table 2.6 Three samples of weekly fuel consumptions and their least squares point estimates

Week	Average hourly temperature, x_1	The chill index, x_2	Sample 1	Sample 2	Sample 3
1	28.0	18	$y_1 = 12.4$	$y_1 = 12.0$	$y_1 = 10.7$
2	28.0	14	$y_2 = 11.7$	$y_2 = 11.8$	$y_2 = 10.2$
3	32.5	24	$y_3 = 12.4$	$y_3 = 12.3$	$y_3 = 10.5$
4	39.0	22	$y_4 = 10.8$	$y_4 = 11.5$	$y_4 = 9.8$
5	45.9	8	$y_5 = 9.4$	$y_5 = 9.1$	$y_5 = 9.5$
6	57.8	16	$y_6 = 9.5$	$y_6 = 9.2$	$y_6 = 8.9$
7	58.1	1	$y_7 = 8.0$	$y_7 = 8.5$	$y_7 = 8.5$
8	62.5	0	$y_8 = 7.5$	$y_8 = 7.2$	$y_8 = 8.0$
			$\mathbf{b}_0 = 13.1087$	$\mathbf{b}_0 = 12.949$	$\mathbf{b}_0 = 11.593$
			$\mathbf{b}_1 = -.09001$	$\mathbf{b}_1 = -.0882$	$\mathbf{b}_1 = -.0548$
			$\mathbf{b}_2 = .08249$	$\mathbf{b}_2 = .0876$	$\mathbf{b}_2 = .0256$

observed from the eight normally distributed populations of possible weekly fuel consumptions. Below each sample are given the least squares point estimates b_0, b_1, and b_2 that would be calculated by using the sample. Because there are an infinite number of possible samples of eight weekly fuel consumptions that could be observed from the eight populations of possible weekly fuel consumptions, there is an infinite population of potential values of each of the least squares point estimates b_0, b_1, and b_2.

In general, let β_j denote any particular one of the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ of the linear regression model, and let b_j denote the least squares point estimate of β_j. For example, if $j = 1$, we are considering β_1 and b_1. If $j = 2$, we are considering β_2 and b_2. It is, of course, highly unlikely that the least squares point estimate b_j of β_j that we calculate using the sample of n observed values $y_1, y_2, \ldots, y_n$ of the dependent variable equals the true value of β_j. However, it can be shown (see Section B.3) that μ_{b_j}, the mean of the population of all possible values of b_j that could be calculated from all possible samples of n values of the dependent variable, is equal to β_j. Because $\mu_{b_j} = \beta_j$, we say that b_j is an *unbiased point estimate* of β_j.

2.3.3 *The Mean Square Error and the Standard Error*

We next wish to find point estimates of σ^2 and σ, the constant variance and standard deviation of each of the different populations of possible values of the dependent variable. We have seen that, for $i = 1, 2, \ldots, n$, σ^2 measures the variation—around the mean $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$ of all the possible values of the dependent variable that could be observed when the values of the independent variables are $x_{i1}, x_{i2}, \ldots, x_{ik}$. Because the point estimate of the mean $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$ is $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik}$, it seems natural to use the sum of squared residuals $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ to help construct a point estimate of σ^2. It can be shown that if we divide *SSE* by $n - (k+1)$, which is called the *number of degrees of freedom* associated with *SSE*, then we obtain an unbiased point estimate of σ^2 (see Section B.3). That is, let $s^2 = SSE / [n - (k + 1)]$, which we call the *mean square error*, be the point estimate of σ^2. Then, it can be shown that μ_{s^2}, the mean of all possible values of s^2 that could be calculated from all possible samples, is equal to σ^2. Moreover, let $s = \sqrt{s^2}$, which we call the *standard error*, be the point estimate of $\sigma = \sqrt{\sigma^2}$. Unfortunately, s is not an unbiased point estimate of σ. However, we use s as the point estimate of σ because it is intuitive to do so and because there is no easy way to calculate an unbiased point estimate of σ. We summarize the point estimates of σ^2 and σ as follows:

The mean square error and the standard error

Suppose that the linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

utilizes k independent variables and thus has $(k + 1)$ parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$. Then, if the regression assumptions are satisfied, and if *SSE* denotes the sum of squared residuals for the model:

1. A point estimate of σ^2 is the *mean square error*

$$s^2 = \frac{SSE}{n - (k+1)}$$

2. A point estimate of σ is the *standard error*

$$s = \sqrt{\frac{SSE}{n - (k+1)}}$$

The mean square error and the standard error (Continued)

Furthermore, the sum of squared residuals

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

can be calculated by the alternative formula

$$SSE = \sum_{i=1}^{n} y_i^2 - \mathbf{b}'\mathbf{X}'\mathbf{y}$$

Here, $\mathbf{b}' = [b_0, b_1, b_2, \ldots, b_k]$ is a row vector (the transpose of **b**) containing the least squares point estimates, and $\mathbf{X}'\mathbf{y}$ is the column vector used in calculating the least squares point estimates.

We will see in Section 2.7 that if a particular regression model gives a small standard error s, then the model will give short *prediction intervals* and thus accurate predictions of individual y values. For example, Table 2.4 shows that *SSE* for the fuel consumption model

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

is .674. To calculate *SSE* by the alternative formula, recall that

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 81.7 \\ 3413.11 \\ 1157.40 \end{bmatrix}$$

It follows that

$$\begin{aligned} \mathbf{b}'\mathbf{X}'\mathbf{y} &= \begin{bmatrix} 13.1087 & -.09001 & .08249 \end{bmatrix} \begin{bmatrix} 81.7 \\ 3413.11 \\ 1157.40 \end{bmatrix} \\ &= 13.1087(81.7) + (-.09001)3413.11) + (.08249)(1157.40) \\ &= 859.236 \end{aligned}$$

Furthermore, the eight observed fuel consumptions (see Table 2.1) can be used to calculate

$$\sum_{i=1}^{8} y_1^{\,2} = y_1^2 + y_2^2 + \ldots + y_8^2$$
$$= (12.4)^2 + (11.7)^2 + \ldots + (7.5)^2 = 859.91$$

Therefore SSE can be calculated in the following alternative fashion:

$$\begin{aligned} SSE &= \sum_{i=1}^{8} y_i^2 - \mathbf{b}'\mathbf{X}'\mathbf{y} \\ &= 859.91 - 859.236 \\ &= .674 \end{aligned}$$

Since the aforementioned fuel consumption model utilizes $k = 2$ independent variables and thus has $k + 1 = 3$ parameters $\left(\beta_0, \beta_1, \text{ and } \beta_2\right)$, a point estimate of σ^2 for this model is the mean square error

$$s^2 = \frac{SSE}{n-(k+1)} = \frac{.674}{8-3} = \frac{.674}{5} = .1348$$

and a point estimate of σ is the standard error $s = \sqrt{.1348} = .3671$. Note that $SSE = .674$, $s^2 = .1348 \approx .135$, and $s = .3671$ are given on the Minitab output in Figure 2.10.

Also, note that Table 2.4 tells us that $SSE = 2.57$ for the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ relating y = weekly fuel consumption to x = average hourly temperature. Since the simple linear regression model utilizes $k = 1$ independent variable and thus has $k + 1 = 2$ parameters (β_0 and β_1), a point estimate of σ^2 for this model is

$$s^2 = \frac{SSE}{n-(k+1)} = \frac{2.57}{8-2} = \frac{2.57}{6} = .428$$

and a point estimate of σ is $s = \sqrt{.428} = .6542$. Here, $SSE = 2.57$, $s^2 = .428$, and $s = .6542$ are given on the Minitab output in Figure 2.11.

Moreover, notice that $s = .3671$ for the model using both the average hourly temperature and the chill index is less than $s = .6542$ for the model using only the average hourly temperature. Therefore, we have evidence that the two independent variable model will give more accurate predictions of future weekly fuel consumptions

2.4 Coefficients of Determination and Correlation

We indicated in the previous section that if a regression model gives a small s, then the model will accurately predict individual y values. For this reason, s is one measure of the usefulness, or utility, of a regression model. In this section we discuss several other ways to assess the utility of a regression model.

2.4.1 Measures of Variation, R^2, and R

The *coefficient of determination* is a measure of the usefulness of the linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$. To define this quantity, we need to develop several *measures of variation*. Therefore, suppose that we have observed n combinations of values of the dependent variable y and the independent variables $x_1, x_2, \ldots, x_k$. If $b_0, b_1, b_2, \ldots, b_k$ denote the least squares point estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$, then $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik}$ is the point prediction of y_i, the ith observed value of the dependent variable. Moreover, let $\bar{y}$ denote the mean of the n observed values of the dependent variable. Then, it follows that $(y_i - \bar{y})$, the *total deviation* of y_i from $\bar{y}$, can be partitioned into a deviation, $(\hat{y}_i - \bar{y})$, that is *explained* by the linear regression model, plus a deviation $(y_i - \hat{y}_i)$ that is left *unexplained* by the linear regression model. That is,

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

To understand this partitioning consider Figure 2.13, which shows the partitioning using the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$. For this model, the least squares line fitted to the observed data gives the point prediction $\hat{y}_i = b_0 + b_1 x_i$ of y_i. Moreover, Figure 2.13 shows that

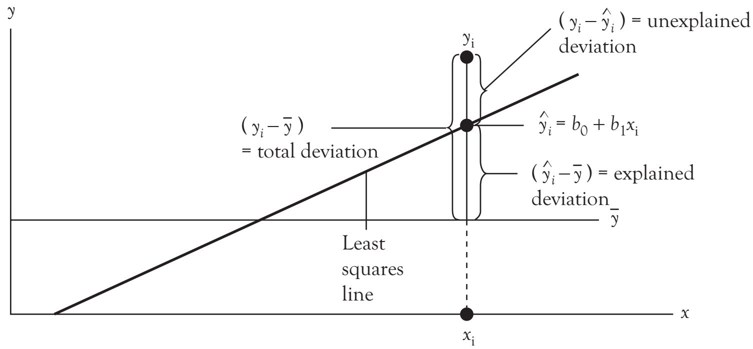


Figure 2.13 The total, explained, and unexplained deviations

the *total deviation* $(y_i - \bar{y})$, which is the total vertical distance from $\bar{y}$ to y_i, equals the *explained deviation* $(\hat{y}_i - \bar{y})$, which is the vertical distance from $\bar{y}$ to the point $\hat{y}_i$ on the least squares line, plus the *unexplained deviation* $(y_i - \hat{y}_i)$, which is the vertical distance from $\hat{y}_i$ to y_i—a vertical distance left unexplained by the least squares line. In addition, it can be shown (see Section B.4) that for the linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The sum of the squared total deviations, $\sum (y_i - \bar{y})^2$, is called the *total variation* and measures the variation of the y_i values around their mean $\bar{y}$. The sum of the squared explained deviations, $\sum (\hat{y}_i - \bar{y})^2$, is called the *explained variation* and measures the amount of the total variation that is explained by the linear regression model. The sum of the squared unexplained deviations, $\sum (y_i - \hat{y}_i)^2$, is called the *unexplained variation* (this is another name for *SSE*) and measures the amount of the total variation that is left unexplained by the linear regression model. We now define the *coefficient of determination*, denoted by R^2, to be the ratio of the explained variation to the total variation. That is R^2 = (explained variation)/(total variation), and we say that R^2 is the proportion of the total variation in the *n* observed values of *y* that is explained by the linear regression model. Neither the explained variation nor the total variation can be negative (both quantities are sums of squares). Therefore, R^2 is greater than or equal to 0. Because the explained variation must be less than or equal to the total variation, R^2 cannot be greater than one. The nearer R^2 for a particular regression model is to one, the larger is the proportion of the total variation that is explained by the model, and the greater is the potential utility of the model in predicting y. If a model's value of R^2 is not reasonably close to one, the model will probably not provide accurate predictions of y. In such a case we need to find a better model.

The coefficient of determination, R^2

For the linear regression model:

1. *Total variation* $= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$
2. *Explained variation* $= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \mathbf{b}'\mathbf{X}'\mathbf{y} - n\bar{y}^2$
3. *Unexplained variation* $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} y_i^2 - \mathbf{b}'\mathbf{X}'\mathbf{y}$
4. *Total variation = Explained variation + Unexplained variation*
5. *The coefficient of determination is*

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

6. R^2 is the proportion of the total variation in the n observed values of the dependent variable that is explained by the overall regression model.

At the end of this section we will discuss some special facts about the coefficient of determination, R^2, when using the simple linear regression model. When using a multiple linear regression model (a model with more than one independent variable), we sometimes refer to R^2 as the *multiple coefficient of determination*, and we define the *multiple correlation coefficient* to be $R = \sqrt{R^2}$. For example, consider the fuel consumption model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.

Using the fuel consumption data, we previously made the following calculations:

$$\sum_{i=1}^{8} y_i^2 = 859.91 \quad \mathbf{b'X'y} = 859.236 \quad \bar{y} = \frac{\sum_{i=1}^{8} y_i}{8} = 10.2125$$

$$\text{Unexplained variation} = SSE = \sum_{i=1}^{8} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{8} y_i^2 - \mathbf{b'X'y}$$
$$= 859.91 - 859.236 = .674$$

We can calculate the total variation to be

$$\text{Total variation} = \sum_{i=1}^{8} (y_i - \bar{y})^2 = \sum_{i=1}^{8} y_i^2 - 8\bar{y}^2$$
$$= 859.91 - 8(10.2125)^2$$
$$= 25.549$$

Moreover, we can calculate the explained variation by either of the following two methods:

$$\text{Explained variation} = \text{Total variation} - \text{Unexplained variation}$$
$$= 25.549 - .674 = 24.875$$

or

$$\text{Explained variation} = \sum_{i=1}^{8} (\hat{y}_i - \bar{y})^2$$
$$= \mathbf{b'X'y} - 8\bar{y}^2$$
$$= 859.236 - 8(10.2125)^2 = 24.875$$

The Minitab output in Figure 2.10 tells us that the total, explained, and unexplained variations for this model are, respectively, 25.549, 24.875, and .674. This output also tells us that the multiple coefficient of determination is

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{24.875}{25.549} = .974$$

The multiple correlation coefficient is $R = \sqrt{.974} = .9869$. The value of $R^2 = .974$ says that the fuel consumption model with two independent variables explains 97.4 percent of the total variation in the eight observed fuel consumptions.

2.4.2 *Adjusted R^2*

Even if the independent variables in a regression model are unrelated to the dependent variable, they will make R^2 somewhat greater than zero. To avoid overestimating the importance of the independent variables, many analysts recommend calculating an *adjusted* coefficient of determination. To understand this idea, suppose that the values of the k independent variables are completely random (that is, randomly chosen from a population of numbers). It can be shown that these independent variables will still explain enough of the total variation in the observed values of the dependent variable to make R^2 equal to, on the average, $k/(n-1)$. Therefore, our first step in adjusting R^2 is to subtract this random explanation and form the quantity $R^2 - k/(n-1)$.

If the values of the independent variables are completely random, then this adjusted version of R^2 is (on average) equal to zero. However, if the values of the independent variables are not completely random, then this quantity reduces R^2 too much. To see why, note that if R^2 is equal to 1, then $R^2 - k/(n-1)$ is not equal to 1 but is equal to $1 - k/(n-1) = (n-k-1)/(n-1)$, which is less than 1, since $n-k-1 < n-1$. To define an adjusted R^2 that is equal to 1 if R^2 is equal to 1, we multiply $R^2 - k/(n-1)$ by $(n-1)/(n-k-1)$. This gives the following *adjusted coefficient of determination (adjusted R^2)*.

Adjusted R^2

The adjusted coefficient of determination (adjusted R^2) is

$$\bar{R}^2 = \left(R^2 - \frac{k}{n-1}\right)\left(\frac{n-1}{n-k-1}\right)$$

When using a multiple linear regression model, we sometimes refer to the adjusted coefficient of determination as the *adjusted multiple coefficient of determination*. For example, consider the fuel consumption model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. Because we have seen that the multiple coefficient of determination for this model is $R^2 = .974$, it follows that the adjusted multiple coefficient of determination for this model is

$$
\begin{aligned}
\bar{R}^2 &= \left(R^2 - \frac{k}{n-1}\right)\left(\frac{n-1}{n-k-1}\right) \\
&= \left(.974 - \frac{2}{8-1}\right)\left(\frac{8-1}{8-2-1}\right) \\
&= .963
\end{aligned}
$$

Note that $\bar{R}^2 = .963$ is given on the Minitab output in Figure 2.10.

If R^2 is less than $k/(n-1)$ (which can happen), then $\bar{R}^2$ will be negative. In this case, statistical software systems set $\bar{R}^2$ equal to zero. Historically, $\bar{R}^2$ and R^2 have been popular measures of model utility—possibly because they are unitless and between 0 and 1. In general, we desire R^2 and $\bar{R}^2$ to be near one. However, sometimes even if a regression model has an R^2 and an $\bar{R}^2$ that are near one, the standard error s is still too large for the model to predict accurately. The best that can be said for an R^2 and an $\bar{R}^2$ near one is that they give us hope that the model will predict accurately. Of course, the only way to know is to see if s is small enough. In other words, since we usually are judging a model's ability to predict, s is a better measure of model utility than are R^2 and $\bar{R}^2$. We will say more later about using s, R^2, and $\bar{R}^2$ to help choose a regression model.

2.4.3 Simple Coefficients of Determination and Correlation, r^2 and r

When we are using the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$, we sometimes refer to R^2 and $\bar{R}^2$ as, respectively, the *simple coefficient of determination* and the *adjusted simple coefficient of determination*. Moreover, we sometimes denote these quantities as r^2 and $\bar{r}^2$. For example, the Minitab output in Figure 2.11 tells us that for the simple linear regression model relating y = weekly fuel consumption to x = average hourly temperature, the explained variation is 22.981 and the total variation is 25.549. It follows that the simple coefficient of determination is $r^2 = 22.981/25.549 = .899$ and the adjusted simple coefficient of determination is

$$\begin{aligned}\bar{r}^2 &= \left(r^2 - \frac{k}{n-1}\right)\left(\frac{n-1}{n-k-1}\right)\\ &= \left(.899 - \frac{1}{8-1}\right)\left(\frac{8-1}{8-1-1}\right)\\ &= .883\end{aligned}$$

These quantities are shown on the Minitab output in Figure 2.11. They are not as large as the R^2 of .974 and the $\bar{R}^2$ of .963 given by the regression model that uses both the average hourly temperature and the chill index as predictor variables. We next define the *simple correlation coefficient* as follows.

The simple correlation coefficient

The simple correlation coefficient between y and x, denoted by r, *is*

$$r = +\sqrt{r^2} \text{ if } b_1 \text{ is positive and } r = -\sqrt{r^2} \text{ if } b_1 \text{ is negative}$$

where b_1 is the slope of the least squares line relating y to x. This correlation coefficient *measures the strength of the linear relationship between* y *and* x.

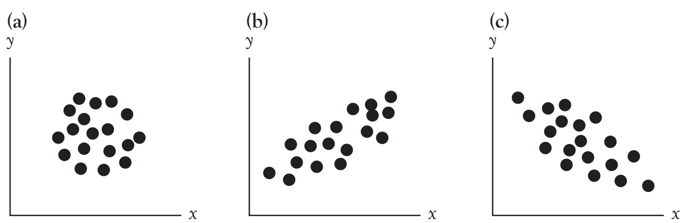


***Figure 2.14** Some types of linear correlation (a) little correlation (b) positive correlation (c) negative correlation*

Because r^2 is always between 0 and 1, the simple correlation coefficient r is between -1 and 1. A value of r near 0 implies little linear relationship or (correlation) between y and x as illustrated in Figure 2.14(a). A value of r close to 1 says that y and x have a strong tendency to move together in a straight-line fashion with a positive slope and, therefore, that y and x are highly related and *positively correlated*. Positive correlation is illustrated in Figure 2.14(b). *A* value of r close to -1 says that y and x have a strong tendency to move together in a straight-line fashion with a negative slope and, therefore, that y and x are highly related and *negatively correlated*. Negative correlation is illustrated in Figure 2.14(c). For the simple linear regression model relating y = weekly fuel consumption to x = average hourly temperature, we have found that $b_1 = -.1279$ and $r^2 = .899$. Therefore,

$$r = -\sqrt{r^2} = -\sqrt{.899} = -.948$$

This simple correlation coefficient says that x and y have a strong tendency to move together in a linear fashion with a negative slope. We have seen this tendency in Figure 2.1, which indicates that y and x are negatively correlated.

If we have computed the least squares slope b_1 and r^2, the method given in the previous box provides the easiest way to calculate r. The simple correlation coefficient can also be calculated using the formula

$$r = \frac{\text{SS}_{xy}}{\sqrt{\text{SS}_{xx}\text{SS}_{yy}}}$$

Here SS_{xy} and SS_{xx} have been defined in Section 2.1, and SS_{yy} denotes the total variation, which has been defined in this section. Futhermore, this formula for r automatically gives r the correct (+ or $-$) sign. For instance, in the fuel consumption problem, $\text{SS}_{xy} = -179.6475$, $\text{SS}_{xx} = 1404.355$, and $\text{SS}_{yy} = 25.549$ (see Table 2.1 and Figure 2.11). Therefore,

$$r = \frac{\text{SS}_{xy}}{\sqrt{\text{SS}_{xx}\text{SS}_{yy}}} = \frac{-179.6475}{\sqrt{(1404.355)(25.549)}} = -.948$$

It is important to point out that *high correlation does not imply that a cause-and-effect relationship exists.* When r indicates that y and x are highly correlated, this says that y and x have a strong tendency to move together in a straight-line fashion. The correlation does not mean that changes in x cause changes in y. Instead, some other variable (or variables) could be causing the apparent relationship between y and x. For example, suppose that college students' grade point averages and college entrance exam scores are highly positively correlated. This does not mean that earning a high score on a college entrance exam causes students to receive a high grade point average. Rather, other factors such as intellectual ability, study habits, and attitude probably determine both a student's score on a college entrance exam and a student's college grade point average. In general, while the simple correlation coefficient can show that variables tend to move together in a straight-line fashion, scientific theory must be used to establish cause-and-effect relationships.

2.5 The Overall *F*-Test

In previous sections, we have shown that s, R^2, and $\bar{R}^2$ help us to assess the utility of a regression model. In this and the next section we will discuss several *hypothesis tests* that help us to evaluate the importance of the independent variables in a regression model. To begin, note that the linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

assumes that $\mu_{y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$. If each of $\beta_1, \beta_2, \ldots$, and β_k equals zero, then $\mu_{y|x_1,x_2,\ldots,x_k} = \beta_0$. In this case, the mean value of y does not depend upon x_1 or x_2 or…or x_k, and we would say that there is no *overall regression relationship* between the dependent variable y and the independent variables $x_1, x_2, \ldots, x_k$. On the other hand, if at least one of β_1 or β_2 or....or β_k does not equal zero, then the mean value of y depends upon at least one of x_1 or x_2 or...or x_k, and we would say that there is an overall regression relationship between y and $x_1, x_2, \ldots, x_k$. To test for an overall regression relationship between y and $x_1, x_2, \ldots, x_k$, we test the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

which says that no overall regression relationship exists, versus the alternative hypothesis

$$H_a\text{: At least one of } \beta_1, \beta_2, \ldots, \beta_k \text{ does not equal 0}$$

which says that an overall regression relationship does exist. To test H_0 versus H_a, we use the *test statistic*

$$F(\text{model}) = \frac{(\text{Explained variation})/k}{(\text{Unexplained variation})/[n-(k+1)]}$$

A large value of F(model) would be caused by an explained variation that is large compared to the unexplained variation. This would occur if the mean value of the dependent variable y depends upon at least one of the independent variables $x_1, x_2, \ldots, x_k$, which would imply that $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ is false and H_a: At least one of $\beta_1, \beta_2, \ldots, \beta_k$ does not equal 0 is true. To decide exactly how large F(model) has to be to reject H_0, we consider the *probability of a Type I error* for the hypothesis test. A Type I error is committed if we reject $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ when H_0 is true. This means that we would conclude that an overall regression relationship exists when it does not. To perform the hypothesis test, we set the probability of a Type I error (also called the *level of significance*) for the hypothesis test equal to a specified value α. The smaller the value α at which we can reject H_0, the smaller is the probability that we have concluded that an overall regression relationship exists when it does not. Therefore, the stronger is the evidence that we have made the correct decision in concluding that an overall regression relationship exists.

In practice we usually choose α to be between .10 and .01, with .05 being the most common value of α. Note that we rarely set α lower than .01 because doing so would mean that the probability of a Type II error (failing to conclude that an overall regression relationship exists when it does exist) would be unacceptably large.

2.5.1 Using a Rejection Point

In order to set the level of significance for testing $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ equal to a specified value α, we use the fact that if H_0 is true, then the population of all possible values of F(model) is described by a probability distribution called the *F-distribution*. (This fact is proven in Appendix B.5) The curve of the F-distribution is skewed with a tail to the right (see Figure 2.15), and the exact shape of this curve is determined by two parameters—the *numerator degrees of freedom* and the *denominator degrees of freedom* of the F-distribution. The F-distribution describing the population of all possible values of F(model) has k numerator degrees of freedom and $n-(k+1)$ denominator degrees of freedom . This leads to the following procedure for testing $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ at level of significance α:

- Place the level of significance α in the right-hand tail of the curve of the F-distribution having k numerator and $n-(k+1)$ denominator degrees of freedom, and use the F table (see Table A1 in Appendix A) to find the *rejection point* $F_{[\alpha]}$. Here, $F_{[\alpha]}$ is the point on the horizontal axis under the curve of this F-distribution so that the tail area to the right of this point is α. (see Figure 2.15[a]).
- Reject H_0 if and only if the test statistic F(model) is greater than $F_{[\alpha]}$.

For example, consider the fuel consumption model

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

The Minitab output in Figure 2.10 tells us that the explained and unexplained variations for this model are, respectively, 24.875 and .674. It follows, since there are $k = 2$ independent variables, that

$$F(\text{model}) = \frac{(\text{Explained variation})/k}{(\text{Unexplained variation})/[n-(k+1)]}$$

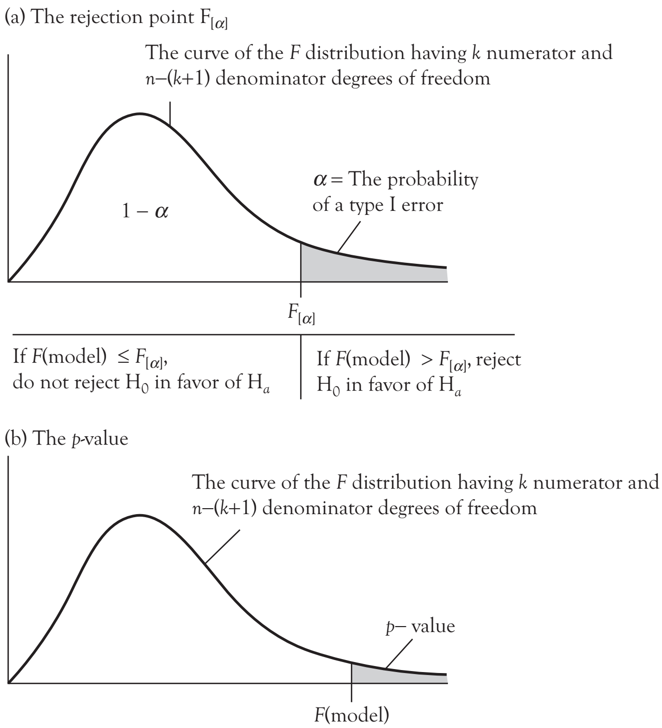


Figure 2.15 An F-test for the linear regression model

$$= \frac{24.875/2}{.674/[8-(2+1)]} = \frac{12.438}{.135}$$
$$= 92.30$$

Note that this F(model) statistic is given on the Minitab output . If we wish to test $H_0 : \beta_1 = \beta_2 = 0$ versus H_a : At least one of β_1 or β_2 does not equal 0 at level of significance $\alpha = .05$, we use the rejection point $F_{[\alpha]} = F_{[.05]}$ based on $k = 2$ numerator and $n - (k+1) = 8 - (2+1) = 5$ denominator degrees of freedom. Using Table A1 in Appendix A, we find that $F_{[.05]} = 5.79$. Since $F(\text{model}) = 92.30 > F_{[.05]} = 5.79$, we can reject $H_0 : \beta_1 = \beta_2 = 0$ at level of significance .05.

In general, if we can reject $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k$ at a small level of significance α, we conclude at the small level of significance α that the *overall regression relationship* (or *regression model*) *is significant*. This is the same as concluding at the small level of significance α that *at least one of the independent variables* $x_1, x_2, \ldots, x_k$ in the regression model is *significantly related to the dependent variable* . Statistical practice has shown that

1. If we can reject H_0 at the *.05 level of significance*, then we have *strong evidence that the regression model is significant*;
2. If we can reject H_0 at the *.01 level significance*, then we have *very strong evidence that the regression model is significant*;
3. If we can reject H_0 at the .001 level of significance, then we have *extremely strong evidence that the regression model is significant.*

If we wish to use rejection points to test $H_0 : \beta_1 = \beta_2 = 0$ for the fuel consumption model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ at the .01 and .001 levels of significance, we would need to compare F(model) = 92.30 with $F_{[.01]}$ and $F_{[.001]}$ based on two numerator and five denominator degrees of freedom. While tables of values of $F_{[.01]}$ and $F_{[.001]}$ are readily available in books of statistical tables, and values of both $F_{[.01]}$ and $F_{[.001]}$ can be found using statistical software packages (including Excel), the *p-value approach* is an easier and more informative way to test a hypothesis.

2.5.2 Using a *p*-Value

The p-value for testing $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ *is defined to be the area under the curve of the F-distribution having k numerator and* $n-(k+1)$ *denominator degrees of freedom to the right of F(model).* This p-value is illustrated in Figure 2.15(b). When testing $H_0 : \beta_1 = \beta_2 = 0$ in the fuel consumption model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, the p-value is the area under the curve of the F-distribution having $k = 2$ numerator and $n - (k + 1) = 8 - (2 + 1) = 5$ denominator degrees of freedom to the right of F(model) = 92.30. The Minitab output in Figure 2.10 says that this p-value is .000. When Minitab says that a p-value is .000, it means that the p-value is less than .001. If we use Excel, we can find that the p-value in this situation is .0000215. Interpreted as a probability, the p-value of .0000215 says that if the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ is true, then only about 2 in 100,000 of all F(model) statistics that could be observed are at least as large as 92.30 Thus the p-value of .0000215 leads us to reach one of two possible conclusions. The first conclusion is that $H_0: \beta_1 = \beta_2 = 0$ is true and we have observed an F(model) statistic that is so rare that only .0000215 of all possible F(model) statistics are at least as large as this observed F(model) statistic. The second conclusion is that $H_0: \beta_1 = \beta_2 = 0$ is false. A reasonable person would probably make the second conclusion. In general, how small does the p-value have to be before we reject H_0? It depends upon the level of significance α that we set for the hypothesis test. Moreover, once we have computed the p-value, we immediately know for any particular level of significance α whether we can reject H_0. It turns out that we can reject H_0 if the p-value is less than α. *To understand this, suppose that the p-value, which is the area to the right of F(model), is less than* α*, which is the area to right of* $F_{[\alpha]}$. *Comparing Figures 2.15 (a) and (b), we see that this implies that F(model) is greater than* $F_{[\alpha]}$. *But F(model) being greater than* $F_{[\alpha]}$ *is the previously discussed rejection point condition, and thus we can reject* H_0 *at level of significance* α. When testing $H_0: \beta_1 = \beta_2 = 0$ in the fuel consumption model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, the p-value of .0000215 is less than the α values .05, .01, and .001. Therefore, we can reject H_0 at levels of significance .05, .01, and .001. It follows that we have extremely strong evidence that the fuel consumption model is significant. That is, we have extremely strong evidence that at least one of the independent variables x_1 and x_2 in the model is significantly related to y.

We summarize the hypothesis test for the significance of the linear regression model as follows.

An F-test for the linear regression model

Suppose that the regression assumptions hold and that the linear regression model has $(k + 1)$ parameters, and consider testing

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

versus

$$\mathrm{H}_a: \text{At least one of } \beta_1, \beta_2, \ldots, \beta_k \text{ does not equal } 0$$

An F-test for the linear regression model (Continued)

Define the *overall F-statistic* to be

$$F(\text{model}) = \frac{(\text{Explained variation})/k}{(\text{Unexplained variation})/[n-(k+1)]}$$

Also, define the *p-value* related to *F*(model) to be the area under the curve of the F-distribution having k numerator and $n-(k+1)$ denominator degrees of freedom to the right of *F*(model). Then, we can reject H_0 in favor of H_a at level of significance α if either of the following equivalent conditions holds:

1. $F(\text{model}) > F_{[\alpha]}$
2. p-value $< \alpha$

Here the *rejection point* $F_{[\alpha]}$ is the point on the horizontal axis under the curve of the F distribution having k numerator and $n-(k+1)$ denominator degrees of freedom so that the tail area to the right of this point is α.

In general, the *overall F-test* just summarized is usually regarded as a *preliminary test of significance*. To understand this, suppose that the overall *F*-test allows us at a small value of α (say, .05) to reject H_0 and thus conclude that at least one of the independent variables in the regression model under consideration is significantly related to the dependent variable. Statisticians then regard this result as a *license to use individual t tests* to decide *which independent variables* in the regression model are significantly related to the dependent variable. Such individual *t* tests are discussed next.

2.6 Individual *t* Tests

Consider the linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

In order to gain information about which independent variables significantly affect y, we can test the significance of a single independent variable. We arbitrarily refer to this variable as x_j and assume that it is multiplied by the parameter β_j. For example, if $j = 1$, we are testing the significance of x_1, which is multiplied by β_1; if $j = 2$, we are testing the significance of x_2, which is multiplied by β_2. To test the significance of x_j, we test the null hypothesis $H_0 : \beta_j = 0$. We usually test H_0 versus the *two-sided* alternative hypothesis $H_a : \beta_j \neq 0$, which says that a nonzero change in the mean value of the dependent variable is associated with an increase in the value of the independent variable x_j. In some situations we would know whether this change in the mean value of the dependent variable would be an increase or a decrease, and in such situations it would be appropriate to use a *one-sided* alternative hypothesis. For example, in the fuel consumption model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, we can say that if β_1 is not zero, then it must be negative. A negative β_1 would say that mean fuel consumption decreases as average hourly temperature x_1 increases. Because of this, it would be appropriate to test $H_0 : \beta_1 = 0$ versus the *less than alternative* $H_a : \beta_1 < 0$. Similarly, we can say that if β_2 is not zero, then it must be positive. A positive β_2 would say that mean fuel consumption increases as the chill index x_2 increases. Because of this, it would be appropriate to test $H_0 : \beta_2 = 0$ versus the *greater than alternative* $H_a : \beta_2 > 0$. Although it can be shown that using the appropriate one-sided alternative is slightly more effective than using a two-sided alternative, in some regression models it is difficult to know whether the appropriate one-sided alternative should be a greater than alternative or a less than alternative. Moreover, even if we do know the appropriate one-sided alternative, there is little practical difference between using the appropriate one-sided alternative and using a two-sided alternative. For these reasons, statistical software packages (such as Minitab, SAS, and Excel) present results for testing the two-sided alternative, and, thus, we will emphasize testing the two-sided alternative. It follows that it is reasonable to conclude that the independent variable x_j is significantly related to the dependent variable y in the regression model under consideration if we can reject $H_0 : \beta_j = 0$ in favor of $H_a : \beta_j \neq 0$ at a small level of significance α.

Here the phrase *in the regression model under consideration* is very important. This is because it can be shown that whether x_j is significantly related to y in a particular regression model can depend on what other independent variables are included in the model. This issue is discussed in detail in Chapter 4.

It can be proved (see Section B.6) that if the regression assumptions hold, the population of all possible values of the least squares point estimate b_j is normally distributed with mean β_j and standard deviation

$$\sigma_{b_j} = \sigma\sqrt{c_{jj}}$$

Here, σ is the constant standard deviation of the different error term populations (or different populations of possible values of the dependent variable), and c_{jj} is the jth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ (we illustrate how to find c_{jj} in the next example). We denote the point estimate of σ_{b_j} by s_{b_j} and refer to s_{b_j} as the *standard error of the estimate b_j*. Since we estimate σ by s, it follows that

$$s_{b_j} = s\sqrt{c_{jj}}$$

In order to test $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$, we divide b_j by s_{b_j} and form the *test statistic*

$$t = \frac{b_j}{s_{b_j}} = \frac{b_j - 0}{s_{b_j}}$$

This test statistic measures the distance between b_j and zero (the value that makes the null hypothesis $H_0 : \beta_j = 0$ true). If the absolute value of t is large, this implies that the distance between b_j and zero is large and provides evidence that we should reject $H_0 : \beta_j = 0$. Before discussing how large in absolute value t must be in order to reject $H_0 : \beta_j = 0$ at level of significance α, we first show how to calculate this test statistic.

Example 2.6

Consider the fuel consumption model

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

We have previously found that

$$
\begin{array}{rc}
 & \text{column} \\
 & \begin{array}{cccc} \textit{row} & 0 & 1 & 2 \end{array} \\
(\mathbf{X}'\mathbf{X})^{-1} = & \begin{array}{c} 0 \\ 1 \\ 2 \end{array}\begin{bmatrix} 5.43405 & -.085930 & -.118856 \\ -.085930 & .00147070 & .00165094 \\ -.118856 & .00165094 & .00359276 \end{bmatrix} \\
= & \begin{bmatrix} c_{00} & & \\ & c_{11} & \\ & & c_{22} \end{bmatrix}
\end{array}
$$

Here, we have numbered the rows and columns of $(\mathbf{X}'\mathbf{X})^{-1}$ as 0, 1, and 2 because the β's in the fuel consumption model are denoted as β_0, β_1, and β_2. Thus, the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to

$$1.\ \beta_0 \textit{ is } c_{00} = 5.43405 \approx 5.434$$
$$2.\ \beta_1 \textit{ is } c_{11} = .00147070 \approx .00147$$
$$3.\ \beta_2 \textit{ is } c_{22} = .00359276 \approx .0036$$

Since we have seen in Section 2.3 that s = .3671, it follows that we calculate $s_{b_0}, s_{b_1}, s_{b_2}$, and the associated t-statistics for testing $H_0: \beta_0 = 0$, $H_0: \beta_1 = 0$, and $H_0: \beta_2 = 0$ as shown in Table 2.7. The s_{b_j} values and t statistics shown in Table 2.7 are also given in the Minitab output in Figure 2.10.

2.6.1 Using a Rejection Point

It can be shown that, if the regression assumptions hold, then the population of all possible values of $\left(b_j - \beta_j\right)/ s_{b_j}$ is described by a probability distribution called the *t-distribution*. The curve of the t distribution is symmetrical and bell-shaped and centered at zero (see Figure 2.16), and the spread of this curve is determined by a parameter called the *number of degrees of freedom* of the t-distribution. The t-distribution describing the population of all possible values of $\left(b_j - \beta_j\right)/ s_{b_j}$ has

Table 2.7 Calculations of the standard errors of the b_j values and the t-Statistics for testing $H_0 : \beta_0 = 0$, $H_0 : \beta_1 = 0$, and $H_0 : \beta_2 = 0$ in the fuel consumption model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

Independent variable	b_j	$s_{b_j} = s\sqrt{c_{jj}}$	$t = \frac{b_j}{s_{b_j}}$	**p - value**
Intercept	$b_0 = 13.1087$	$s_{b_0} = s\sqrt{c_{00}}$ $= .3671\sqrt{5.434}$ $= .8557$	$t = \frac{13.1087}{.8557} = 15.32$	.000
x_1	$b_1 = -.09001$	$s_{b_1} = s\sqrt{c_{11}}$ $= .3671\sqrt{.00147}$ $= .01408$	$t = \frac{-.09001}{.01408} = -6.39$	.001
x_2	$b_2 = .08249$	$s_{b_2} = s\sqrt{c_{22}}$ $= .3671\sqrt{.0036}$ $= .0220$	$t = \frac{.08249}{.0220} = 3.75$	.013

$n-(k+1)$ degrees of freedom. It follows that, if the null hypothesis $H_0 : \beta_j = 0$ is true, then the population of all possible values of the test statistic $t = (b_j - 0)/s_{b_j} = b_j / s_{b_j}$ is described by a t-distribution having $n-(k+1)$ degrees of freedom. This leads to the following procedure for testing $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$ at level of significance α:

- Divide the level of significance α in half, and place the area $\alpha/2$ in the right-hand tail of the curve of the t-distribution having $n-(k+1)$ degrees of freedom. Then, use the t table (see Table A2 in Appendix A) to find the *rejection point* $t_{[\alpha/2]}$. Here, $t_{[\alpha/2]}$ is the point on the horizontal axis under the curve of the t distribution having $n-(k+1)$ degrees of freedom so that the tail area to the right of this point is $\alpha/2$ (see Figure 2.16[a]).
- Reject H_0 if and only if $|t|$, the absolute value of the test statistic $t = b_j / s_{b_j}$ is greater than $t_{[\alpha/2]}$-that is, if $t = b_j / s_{b_j}$ is either greater than $t_{[\alpha/2]}$ or less than $-t_{[\alpha/2]}$.

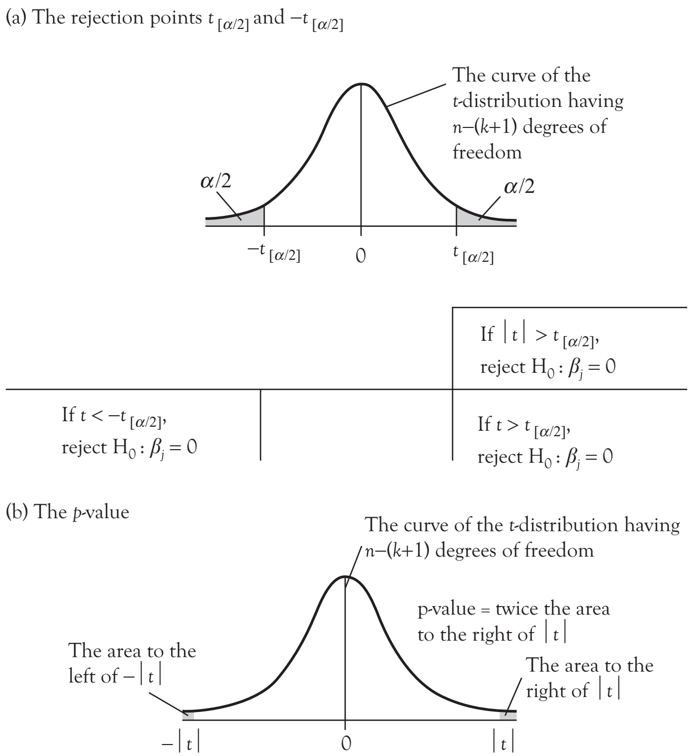


Figure 2.16 A t-test of $H_0: \beta_j = 0$ versus $H_a: \beta_j \neq 0$

For example, consider the fuel consumption model

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

We can test each of the null hypotheses $H_0: \beta_0 = 0$, $H_0: \beta_1 = 0$, and $H_0: \beta_2 = 0$, at level of significance $\alpha = .05$ by using the rejection point $t_{[\alpha/2]} = t_{[.05/2]}$ based on $n - (k + 1) = 8 - (2 + 1) = 5$ degrees of freedom. Utilizing Table A2 in Appendix A, we find that $t_{[.025]} = 2.57$. Table 2.7 tells us that the test statistics for testing $H_0: \beta_0 = 0$, $H_0: \beta_1 = 0$, and $H_0: \beta_2 = 0$, are, respectively, $t = 15.32$, $t = -6.39$, and $t = 3.75$. Because the absolute value of each of these test statistics is greater than $t_{[.025]} = 2.571$, we can reject each of $H_0: \beta_0 = 0$, $H_0: \beta_1 = 0$, and $H_0: \beta_2 = 0$, at the .05 level of *significance*.

In general, consider the parameter β_j that is multiplied by the independent variable x_j in the linear regression model. The smaller the level of significance α at which we can reject $H_0 : \beta_j = 0$, the smaller is the probability that we have mistakenly concluded that the independent variable x_j is significantly related to the dependent variable y in the regression model under consideration. Thus, the stronger is the evidence that x_j is significantly related to y in the regression model. Statistical practice has shown that

1. If we can reject $H_0 : \beta_j = 0$ at the .05 level of significance, we have strong evidence that the independent variable x_j is significantly related to y in the regression model;
2. If we can reject $H_0 : \beta_j = 0$ at the .01 level of significance, we have very strong evidence that x_j is significantly related to y in the regression model;
3. If we can reject $H_0 : \beta_j = 0$ at the .001 level of significance, we have extremely strong evidence that x_j is significantly related to y in the regression model.

We can test $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$ at different levels of significance α (for example, at α values of .05, .01, and .001) by looking up the appropriate different rejection points $t_{[\alpha/2]}$ (for example, $t_{[.025]}$, $t_{[0.005]}$, and $t_{[.0005]}$) in a t-table. However, it is easier and more informative to use a p-value.

2.6.2 *Using a p-Value*

We define the p-value for testing $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$ to be twice the area under the curve of the t-distribution having $n-(k+1)$ degrees of freedom to the right of $|t|$, the absolute value of $t = b_j / s_{b_j}$. This p-value is illustrated in Figure 2.16(b). For example, Table 2.7 tells us that the value of the test statistic for testing $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ in the fuel consumption model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ is $t = -6.39$. Using Excel, we can find that the area under the curve of the t distribution having $n - (k + 1) = 5$ degrees of freedom to the right of

$|t| = |-6.39| = 6.39$ is .0007. Therefore, the p-value, which is twice this area, is $2(.0007) = .0014$. (Note from Figure 2.10 that Minitab rounds this p-value to .001.) The symmetry of the curve of the t-distribution implies that the p-value, which is twice the area to the right of $|t| = 6.39$, equals the area to the right of $|t| = 6.39$ plus the area to the left of $-|t| = -6.39$ (see Figure 2.16[b]). It follows that the p-value of .0014 says that, if we are to believe that $H_0 : \beta_1 = 0$ is true, we must believe that we have observed a test statistic value ($t = -6.39$) that is so rare that only 14 in 10,000 of all possible test statistic values are at least as far from zero (positively or negatively) as this observed test statistic value. It is very difficult to believe that we have observed such a rare test statistic value. Moreover, in general, once we have computed the p-value, we immediately know for any particular level of significance α whether we can reject $H_0 : \beta_j = 0$. It turns out *we can reject* H_0 *if the* p*-value is less than* α. *To understand this, note that if the* p*-value, which is twice the area to right of* $|t|$, *is less than* α, *then the area to the right of* $|t|$ *is less than* $\alpha/2$. *But this implies (examining Figures 2.16[a] and [b]) that* $|t|$ *is greater than* $t_{[\alpha/2]}$. *Therefore, we can reject* $H_0 : \beta_j = 0$ *in favor of* $H_a : \beta_j \neq 0$ *at level of significance* α. When testing $H_0 : \beta_1 = 0$ in the fuel consumption model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, the p-value of .0014 is less than .01 but not less than .001. Therefore, we can reject $H_0 : \beta_1 = 0$ at the .01 level of significance but not at the .001 level of significance. It follows that we have very strong evidence, but not extremely strong evidence, that x_1(the average hourly temperature) is significantly related to y in the fuel consumption regression model. Similarly, the p-value for testing $H_0 : \beta_2 = 0$ can be calculated to be .013 (see the Minitab output in Figure 2.10). Because the p-value of .013 is less than .05 but not less than .01, we can reject $H_0 : \beta_2 = 0$ at the .05 level of significance but not at the .01 level of significance. It follows that we have strong evidence, but not very strong evidence, that x_2 (the chill index) is significantly related to y in the fuel consumption regression model. Lastly, the p-value for testing $H_0 : \beta_0 = 0$ can be calculated to be less than .001, which implies that we can reject $H_0 : \beta_0 = 0$ at the .001 level of significance. Therefore, we have extremely strong evidence that the intercept β_0 is significant in the fuel consumption regression model.

We summarize the hypothesis test of $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$ in the linear regression model as follows.

Testing the significance of the independent variable x_j

Define the test statistic

$$t = \frac{b_j}{s_{b_j}}$$

where $s_{b_j} = s\sqrt{c_{jj}}$, and suppose that the regression assumptions hold. Also, define the p-value related to t to be twice the area under the curve of the t-distribution having $n-(k+1)$ degrees of freedom to the right of $|t|$, the absolute value of t. Then we can reject $H_0 : \beta_j = 0$ in favor of $H_a : \beta_j \neq 0$ at level of significance α if either of the following equivalent conditions hold:

1. $|t| > t_{[\alpha/2]}$ – that is, if $t > t_{[\alpha/2]}$ or $t < -t_{[\alpha/2]}$
2. p-value $< \alpha$

Here the *rejection point* $t_{[\alpha/2]}$ is the point on the horizontal axis under the curve of the t-distribution having $n-(k+1)$ degrees of freedom so that the tail area to the right of this point is $\alpha/2$.

Not every independent variable that we initially include in a regression model will make the model better in terms of helping us to accurately describe, predict, and control the dependent variable. One of the main uses of the individual t tests of this section is to help decide which independent variables should be retained in a regression model. Statistical practice indicates that if we can reject $H_0 : \beta_j = 0$ at the .05 level of significance and thus conclude that there is strong evidence that the independent variable x_j in a regression model is significantly related to the dependent variable y, then retaining x_j in the model is likely to make the model better. Throughout this book we will discuss various ways to help us determine the "best" regression model.

We have seen in Section 2.5 that the intercept β_0 is the mean value of the dependent variable when all of the independent variables $x_1, x_2, \ldots, x_k$ equal zero. In some situations it might seem logical that β_0 would equal zero. For example, if we were using the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ to relate x, the number of items processed at a naval installation, to y, the number of labor hours required to process the items, then it might seem logical that β_0, the mean number of hours required to process zero items, is zero. Therefore, if we fail to reject $H_0 : \beta_0 = 0$ and cannot conclude that the intercept is significant at the .05 level of significance, it might be reasonable to set β_0 equal to zero and remove it from the regression model. This would give us the model $y = \beta_1 x + \varepsilon$, and we would say that we are performing a *regression analysis through the origin***.** We will give some specialized formulas for doing this in Section 2.9. In general, to perform a regression analysis through the origin in (multiple) linear regression (that is, to set the intercept β_0 equal to zero), we would fit the model by leaving the column of 1's out of the **X** matrix. However, in general, logic seeming to indicate that β_0 equals zero can be faulty. For example, the intercept β_0 in the model $y = \beta_0 + \beta_1 x + \varepsilon$ relating the number of items processed to processing time might represent a mean basic "set up" time to process any number of items. This would imply that β_0 might not be zero. In fact, many statisticians (including the authors) believe that leaving the intercept in a regression model will give the model more "modeling flexibility" and is appropriate, no matter what the t test of $H_0 : \beta_0 = 0$ says about the significance of the intercept.

We next consider how to calculate a confidence interval for a regression parameter.

A confidence interval for the regression parameter β_j

If the regression assumptions hold, a $100(1-\alpha)$ *percent confidence interval for the regression parameter* β_j is

$$\left[b_j \pm t_{[\alpha/2]} s_{b_j} \right]$$

Example 2.9 Consider the fuel consumption model

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

The Minitab output in Figure 2.10 tells us that $b_1 = -.09001$ and $s_{b_1} = .01408$.

If we wish to calculate a 95 percent confidence interval for β_1, then $100(1 - \alpha)\% = 95\%$, which implies $1 - \alpha = .95$ and $\alpha = .05$. Therefore, we use the t point $t_{[\alpha/2]} = t_{[.05/2]} = t_{[.025]} = 2.571$ that is based on $n - (k + 1) = 8 - (2 + 1) = 5$ degrees of freedom. It follows that a 95 percent confidence interval for β_1 is

$$\begin{aligned}\left[b_1 \pm t_{[.025]} s_{b_1}\right] &= [-.09001 \pm 2.571(.01408)] \\ &= [-.1262, -.0538]\end{aligned}$$

This interval says we are 95 percent confident that if average hourly temperature increases by one degree and the chill index does not change, then mean weekly fuel consumption will decrease by at least .0538 MMcf of natural gas and by at most .1262 MMcf of natural gas. Furthermore, since this 95 percent confidence interval does not contain 0, we can reject $H_0 : \beta_1 = 0$ in favor of $\mathrm{H}_a : \beta_1 \neq 0$ at the .05 level of significance.

To conclude this subsection, note that because we calculate the least squares point estimates by using the matrix algebra equation $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, the least squares point estimate b_j of β_j is a linear function of $y_1, y_2, \ldots, y_n$. For this reason, we call the least squares point estimate b_j a *linear point estimate* (which, since $\mu_{b_j} = \beta_j$, is also an *unbiased point estimate*) of β_j. An important theorem called the *Gauss-Markov Theorem* says that if regression assumptions 1, 2, and 4 hold, then the variance (or spread around β_j) of all possible values (from all possible samples) of the least squares point estimate b_j is smaller than the variance of all possible values of any other unbiased, linear point estimate of β_j. This theorem is important because it says that the actual value of the least squares point estimate b_j that we obtain from the actual sample we observe is likely to be nearer the true β_j than would be the actual value of any other unbiased, linear point estimate of β_j (we prove the Gauss-Markov Theorem in Sections B.6 and B.9).

2.6.3 Tests For β_0 and β_1 in the Simple Linear Regression Model

For the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$, the t statistics used to test $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$ are, respectively,

$$t = \frac{b_0}{s_{b_0}} \text{ and } t = \frac{b_1}{s_{b_1}}$$

where

$$s_{b_0} = s\sqrt{c_{00}} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}} \text{ and } s_{b_1} = s\sqrt{c_{11}} = \frac{s}{\sqrt{SS_{xx}}}$$

Because the simple linear regression model uses $k = 1$ independent variable, we can reject $H_0 : \beta_1 = 0$ in favor of $H_a : \beta_1 \neq 0$ at level of significance α if $|t| = |b_1 / s_{b_1}|$ is greater than $t_{[\alpha/2]}$, which is based on $n - (k+1) = n - (1+1) = n - 2$ degrees of freedom. A second way to test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ is to reject H_0 at level of significance α if the F(model) statistic for the simple linear regression model

$$F(\text{model}) = \frac{(\text{Explained variation})/k}{(\text{Unexplained variation})/[n-(k+1)]}$$
$$= \frac{(\text{Explained variation})}{(\text{Unexplained variation})/n-2}$$

is greater than $F_{[\alpha]}$, which is based on $k = 1$ numerator and $n - (k+1) = n - 2$ denominator degrees of freedom. Moreover, these two ways to test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ are equivalent. Specifically, it can be shown that $(t)^2 = F(\text{model})$ and that $(t_{[\alpha/2]})^2$, which is based on $n - 2$ degrees of freedom, equals $F_{[\alpha]}$ based on 1 numerator and $n - 2$ denominator degrees of freedom. It follows that the rejection point condition $|t| > t_{[\alpha/2]}$ for the t test will hold if and only if the rejection point condition F(model) > $F_{[\alpha]}$ for the F test holds. Furthermore, the p-values related to t and F(model) can be shown to be equal.

For example, for the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ relating y = weekly fuel consumption to x = average hourly temperature, we have found in Example 2.2 that $b_1 = -.1279$ and $SS_{xx} = 1404.35$. Also, the Minitab output in Figure 2.11 tells us that the explained variation equals 22.981, the unexplained variation (SSE) equals 2.568, and s equals .6542. It follows that $s_{b_1} = s / \sqrt{SS_{xx}} = .6542/\sqrt{1404.35} = .01746$, and thus the t statistic for testing $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ is $t = b_1 / s_{b_1} = -.1279 / .01746 = -7.3277$. Using Excel, we find that the area under the curve of the t distribution having $n - (k + 1) = 8 - 2 = 6$ degrees of freedom to the right $|t| = 7.3277$ is .00015, and therefore the p-value for the t test is 2(.00015) = .0003. It also follows that the (unexplained variation) / $(n - 2)$ equals $2.568 / (8 - 2)$, or .428. Consequently, since the explained variation equals 22.981, the F(model) statistic for testing $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ is $22.981 / .428 = 53.6949$. Using Excel, we find that the area under the curve of the F distribution having $k = 1$ numerator and $n - (k + 1) = 8 - 2 = 6$ denominator degrees of freedom to the right of F(model) = 53.6949 is .0003. This is the p-value for the F test and is the same as the p-value for the t-test. In addition, $(t)^2 = (-7.3277)^2 = 53.6949 = F(\text{model})$.

The Minitab output in Figure 2.11 gives $t = b_1 / s_{b_1}$, F(model), and the corresponding p-value, which Minitab says is .000 (meaning less than .001). It follows that we can reject $H_0 : \beta_1 = 0$ in favor of $H_a : \beta_1 \neq 0$ at the .001 level of significance. Therefore, we have extremely strong evidence that x (average hourly temperature) is significantly related to y in the simple linear regression model.

2.6.4 A Test for the Population Correlation Coefficient

It can be shown that the t statistic $t = b_1 / s_{b_1}$ for testing $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ in the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ equals

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

where r is the previously defined simple correlation coefficient between the n observed x and y values. The latter t statistic is the statistic that has historically been used to test the null hypothesis $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$, where ρ is the *population correlation coefficient*. Here ρ can intuitively be regarded as equaling what r would equal if we calculated r using the population of all possible observed combinations of values of x and y. More precisely, let x and y be random variables (for example, average hourly temperature and weekly fuel consumption). Also, let μ_x and σ_x denote the mean and the standard deviation of all possible values of x, and let μ_y and σ_y denote the mean and the standard derivation of all possible values of y. We then define the population correlation coefficient ρ to be $\operatorname{cov}(x, y) / (\sigma_x \sigma_y)$, where $\operatorname{cov}(x, y)$ is the covariance between x and y. That is, $\operatorname{cov}(x, y)$ is the mean of all possible values of $(x - \mu_x)(y - \mu_y)$ that correspond to all possible observed combinations of x and y. In order for the test of $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$ to be valid, the population of all possible observed combinations of values of x and y must be described by a *bivariate normal probability distribution*. The formula for this probability distribution is

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_x}{\sigma_x} \right)^2 - 2\rho \left(\frac{x-\mu_x}{\sigma_x} \right) \left(\frac{y-\mu_y}{\sigma_y} \right) + \left(\frac{y-\mu_y}{\sigma_y} \right)^2 \right] \right\}$$

Assuming that the population of all possible observed combinations of values of the average hourly temperature, x, and the weekly fuel consumption, y are described by a bivariate normal probability distribution, and recalling that r for the $n = 8$ observed combinations of x and y is $-.948$, we calculate.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-.948\sqrt{8-2}}{\sqrt{1-(-.948)^2}} = -7.3277$$

This t statistic for testing $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$ equals the t statistic $t = b_1 / s_{b_1}$ for testing $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ that is given on the

Minitab output in Figure 2.11. Moreover, the p-value for both tests is the same, and the Minitab output tells us that this p-value is less than .001. It follows that we can reject $H_0 : \rho = 0$ in favor of $H_a : \rho \neq 0$ at the .001 level of significance. Therefore, we have extremely strong evidence of a nonzero population correlation coefficient between the average hourly temperature and weekly fuel consumption. In Chapter 4 we will use tests of population correlation coefficients between the dependent variable and the potential independent variables and between just the potential independent variables themselves to help us "build" an appropriate regression model.

To conclude this section, note that it can be shown that for large samples $(n \geq 25)$, an approximate $100(1-\alpha)$ percent confidence interval for $(1/2)\ln[(1+\rho)/(1-\rho)]$ is

$$\left[\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \pm z_{[\alpha/2]}\sqrt{\frac{1}{n-3}}\right]$$

Moreover, if this interval is calculated to be $[\mathrm{a},\ \mathrm{b}]$, it further follows that a $100(1-\alpha)$ percent confidence interval for ρ is

$$\left[\frac{e^{2a}-1}{e^{2a}+1}, \frac{e^{2b}-1}{e^{2b}+1}\right]$$

Note that, in calculating the first interval, $z_{[\alpha/2]}$ is the point on the horizontal axis under the curve of the standard normal distribution so that the tail area to the right of this point is $\alpha/2$. Table A3 in Appendix A is a table of areas under the standard normal curve. For example, suppose that the sample correlation coefficient between the productivities and aptitude test scores of $n = 250$ word processing specialists is .84. To find a 95 percent confidence interval for $(1/2)\ln[(1+\rho)/(1-\rho)]$, we use $z_{[.025]}$. Because the standard normal curve tail area to the right of $z_{[.025]}$ is .025, the standard normal curve area between 0 and $z_{[.025]}$ is $.5 - .025 = .475$. Looking up .475 in the body of Table A3, we find that $z_{[.025]} = 1.96$. Therefore, the desired confidence interval is

$$\left[\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \pm z_{[.025]}\sqrt{\frac{1}{n-3}}\right] = \left[\frac{1}{2}\ln\left(\frac{1+.84}{1-.84}\right) \pm 1.96\sqrt{\frac{1}{250-3}}\right]$$

$$= [1.0965, 1.3459]$$

It follows that a 95 percent confidence interval for ρ is

$$\left[\frac{e^{2(1.0965)}-1}{e^{2(1.0965)}+1}, \frac{e^{2(1.3459)}-1}{e^{2(1.3459)}+1}\right] = [.80, .87]$$

2.7 Confidence Intervals and Prediction Intervals

We have seen that

$$\hat{y} = b_0 + b_1 x_{01} + b_2 x_{02} + \ldots + b_k x_{0k}$$

is

1. The point estimate of

$$\mu_{y|x_{01},x_{02},\ldots,x_{0k}} = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \ldots + \beta_k x_{0k}$$

   the mean value of the dependent variable y when the values of the independent variables are $x_{01}, x_{02}, \ldots, x_{0k}$.
2. The point prediction of

$$\begin{aligned} y &= \mu_{y|x_{01},x_{02},\ldots,x_{0k}} + \varepsilon \\ &= \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \ldots + \beta_k x_{0k} + \varepsilon \end{aligned}$$

   an individual value of the dependent variable y when the values of the independent variables are $x_{01}, x_{02}, \ldots, x_{0k}$.

Because different samples give different values of the least squares point estimates $b_0, b_1, b_2, \ldots, b_k$, different samples give different values of the point estimate and point prediction $\hat{y}$. Unless we are extremely lucky, the value of $\hat{y}$ that we calculate using the sample we observe will not exactly equal the mean value of y or an individual value of y. Therefore, it is important to calculate a *confidence interval for the mean value of* y and a *prediction interval for an individual value of* y. Both of these intervals are based on a quantity called the *distance value*. We first define this quantity, show how to calculate it, and explain its intuitive meaning. Then, we find the confidence interval and prediction interval based on the distance value.

The Distance Value

The **distance value** is

$$\text{Distance value} = \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$$

where $\mathbf{x}_0' = [1 \;\; x_{01} \;\; x_{02} \;\; \ldots \;\; x_{0k}]$ is a row vector containing the numbers multiplied by $b_0, b_1, b_2, \ldots, b_k$ in the equation for $\hat{y} = b_0 + b_1 x_{01} + b_2 x_{02} + \ldots + b_k x_{0k}$.

Example 2.7

In the fuel consumption problem, recall that a weather forecasting service has told us that the average hourly temperature in the future week will be $x_{01} = 40.0$ and the chill index in the future week will be $x_{02} = 10$. We saw in Example 2.4 that

$$\begin{aligned}\hat{y} &= b_0 + b_1 x_{01} + b_2 x_{02} \\ &= 13.1087 - .09001(40.0) + .08249(10) \\ &= 10.333 \text{ MMcf of natural gas}\end{aligned}$$

is the point estimate of the mean fuel consumption when x_1 equals 40 and x_2 equals 10, and is the point prediction of the individual fuel consumption in a single week when x_1 equals 40 and x_2 equals 10. To calculate the

$$\text{Distance value} = \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$$

note that $\mathbf{x}_0'$ is a row vector containing the numbers multiplied by the least squares point estimates b_0, b_1, and b_2 in the point estimate (and prediction) $\hat{y}$. Since 1 is multiplied by $b_0, x_0 = 40.0$ is multiplied by b_1, and $x_{02} = 10$ is multiplied by b_2, it follows that

$$\mathbf{x}_0' = [1 \;\; x_{01} \;\; x_{02}] = [1 \; 40 \; 10]$$

and

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \end{bmatrix} = \begin{bmatrix} 1 \\ 40 \\ 10 \end{bmatrix}$$

Hence, since we have previously calculated $(\mathbf{X}'\mathbf{X})^{-1}$ (see Example 2.3), it follows that

$$\text{Distance value} = \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$$

$$= \begin{bmatrix} 1 & 40 & 10 \end{bmatrix} \begin{bmatrix} 5.43405 & -.085930 & -.118856 \\ -.085930 & .00147070 & .00165094 \\ -.118856 & .00165094 & .00359276 \end{bmatrix} \begin{bmatrix} 1 \\ 40 \\ 10 \end{bmatrix}$$

$$= \begin{bmatrix} .80828 & -.0105926 & -.0168908 \end{bmatrix} \begin{bmatrix} 1 \\ 40 \\ 10 \end{bmatrix} = .2157$$

To intuitively understand the distance value, first note that the averages of the observed average hourly temperatures and the observed chill indices in Table 2.3 are $\bar{x}_1 = 43.98$ and $\bar{x}_2 = 12.88$. The point $(\bar{x}_1, \bar{x}_2) = (43.98, 12.88)$ is shown in Figure 2.17 and is regarded as the center of the experimental region shown in that figure. Figure 2.17 also shows the point $(x_{01}, x_{02}) = (40, 10)$ representing the average hourly temperature and the chill index for which we wish to estimate the mean weekly fuel consumption and predict an individual weekly fuel consumption. The length of the line segment drawn between the

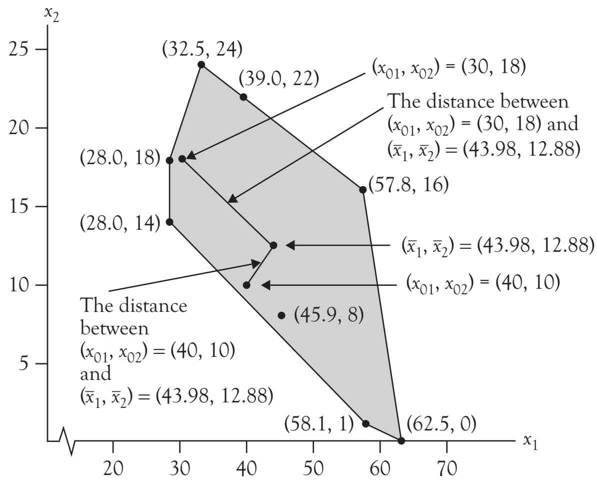


Figure 2.17 Distances in the experimental region

point $(x_{01}, x_{02}) = (40, 10)$ and the point $(\bar{x}_1, \bar{x}_2) = (43.98, 12.88)$ is the **distance** in two-dimensional space between these points. It can be shown that the distance value $\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = .2157$ is reflective of this distance. That is, in general, the greater the distance is between a point (x_{01}, x_{02}) and the center $(\bar{x}_1, \bar{x}_2) = (43.98, 12.88)$ of the experimental region, the greater is the distance value. For example, Figure 2.17 shows that the distance between the point $(x_{01}, x_{02}) = (30, 18)$ and $(\bar{x}_1, \bar{x}_2) = (43.98, 12.88)$ is greater than the distance between the point $(x_{01}, x_{02}) = (40, 10)$ and $(\bar{x}_1, \bar{x}_2) = (43.98, 12.88)$. Consequently, the distance value corresponding to the point $(x_{01}, x_{02}) = (30, 18)$, which is calculated using $\mathbf{x}_0' = [1 \;\; x_{01} \;\; x_{02}] = [1 \; 30 \; 18]$ and equals $\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = .2701$, is greater than the distance value corresponding to the point $(x_{01}, x_{02}) = (40, 10)$, which is calculated using $\mathbf{x}_0' = [1 \;\; x_{01} \;\; x_{02}] = [1 \; 40 \; 10]$ and equals .2157.

In general, let $x_{01}, x_{02}, \ldots, x_{0k}$ be the values of the independent variables $x_1, x_2, \ldots, x_k$ for which we wish to estimate the mean value of the dependent variable and predict an individual value of the dependent variable. Also, define the center of the experimental region to be the point $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k)$, where $\bar{x}_1$ is the average of the previously observed x_1 values, $\bar{x}_2$ is the average of the previously observed x_2 values, and so forth. Then, it can be shown that the greater the distance is (in k-dimensional space) between the point $x_{01}, x_{02},..., x_{0k}$ and $(\bar{x}_1, \bar{x}_2,..., \bar{x}_k)$, the greater is the distance value $\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$, where $\mathbf{x}_0' = [1 \;\; x_{01} \;\; x_{02} \;\; ... \;\; x_{0k}]$.

It can also be shown (see Section B.7) that, if the regression assumptions hold, then the population of all possible values of the point estimate $\hat{y} = b_0 + b_1x_{01} + b_2x_{02} + ... + b_kx_{0k}$ is normally distributed with mean $\mu_{y|x_{01},x_{02},....,x_{0k}}$ and standard deviation $\sigma_{\hat{y}} = \sigma\sqrt{\text{Distance value}}$. Since the standard error s is the point estimate of σ, the point estimate of $\sigma_{\hat{y}}$ is $s_{\hat{y}} = s\sqrt{\text{Distance value}}$, which is called the *standard error of the estimate* $\hat{y}$. Using this standard error, we can form a confidence interval. Note that the $t_{[\alpha/2]}$ point used in the confidence interval (and in the prediction interval to follow) are based on $n-(k + 1)$ degrees of freedom.

A Confidence Interval For a Mean Value of *y*

If the regression assumptions hold, a **$100(1-\alpha)$ percent confidence interval for the mean value of *y*** when the values of the independent variables are $x_{01}, x_{02},..., x_{0k}$ is

$$\left[\hat{y} \pm t_{[\alpha/2]} s\sqrt{\text{Distance value}}\right]$$

We develop a prediction interval for an individual value of y when the values of the independent variables are $x_{01}, x_{02},..., x_{0k}$ by considering the *prediction error* $y - \hat{y}$. After observing a particular sample from the infinite population of all possible samples and calculating a point prediction $\hat{y}$ based on this sample, we could observe any one of an infinite number of different individual values of $y = \mu_{y|x_{01},x_{02},...,x_{0k}} + \varepsilon$ (because of different possible error terms). Therefore, there are an infinite number of different prediction errors that could be observed. If the regression assumptions hold, it can be shown (see Section B.7) that the population of all possible prediction errors is normally distributed with mean 0 and standard deviation $\sigma_{(y-\hat{y})} = \sigma\sqrt{1+\text{Distance value}}$. The point estimate of $\sigma_{(y-\hat{y})}$ is $s_{(y-\hat{y})} = s\sqrt{1+\text{Distance value}}$, which is called the *standard error of the prediction error*. Using this quantity we obtain a *prediction interval* as follows.

A Prediction interval for an individual value of y

If the regression assumptions hold, a **$100(1-\alpha)$ percent prediction interval for an individual value of y** when the values of the independent variables are $x_{01}, x_{02}, \ldots, x_{0k}$ is

$$\left[\hat{y} \pm t_{[\alpha/2]} s\sqrt{1 + \text{Distance value}}\right]$$

Comparing the formula $[\hat{y} \pm t_{[\alpha/2]} s\sqrt{\text{Distance value}}]$ for a confidence interval for the mean value $\mu_{y|x_{01}, x_{02}, \ldots, x_{0k}}$ with the formula $[\hat{y} \pm t_{[\alpha/2]} s\sqrt{1 + \text{Distance value}}]$ for a prediction interval for an individual value $y = \mu_{y|x_{01}, x_{02}, \ldots, x_{0k}} + \varepsilon$, we note that the formula for the prediction interval has an "extra 1" under the radical. This makes the prediction interval longer than the confidence interval. Intuitively, the reason for the extra 1 under the radical is that, although we predict the error term to be zero when computing the point prediction $\hat{y}$ of an individual value $y = \mu_{y|x_{01}, x_{02}, \ldots, x_{0k}} + \varepsilon$, the error term will probably not be zero. The extra 1 under the radical accounts for the added uncertainly that the error term causes, and thus the prediction interval is longer. Also, note the larger the distance value is, the longer are the confidence interval and the prediction interval. Said another way, when $(x_{01}, x_{02}, \ldots, x_{0k})$ is farther from the center of the observed data, $\hat{y} = b_0 + b_1 x_{01} + b_2 x_{02} + \ldots + b_k x_{0k}$ is likely to be less accurate as a point estimate and point prediction.

Before considering an example, consider the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$. For this model $\hat{y} = b_0 + b_1 x_0$ is the point estimate of the mean value of y when x is x_0 and is the point prediction of an individual value of y when x is x_0. Therefore, since 1 is multiplied by b_0 and x_0 is multiplied by b_1 in the expression $\hat{y} = b_0 + b_1 x_0$, it follows that $\mathbf{x}_0' = [1 \;\; x_{01}]$. If we use $\mathbf{x}_0'$ to calculate the distance value, it can be shown that

$$\text{Distance value} = \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}$$

Example 2.8

In Example 2.7 we have seen that

$$\begin{aligned}\hat{y} &= 13.1087 - .09001x_{01} + .08249x_{02}\\ &= 13.1087 - .09001(40) + .08249(10)\\ &= 10.333 \text{ MMcf of natural gas}\end{aligned}$$

is the point estimate of mean weekly fuel consumption when x_1 equals 40 and x_2 equals 10, and is the point prediction of the individual fuel consumption in a single week (next week) when x_1 equals 40 and x_2 equals 10. We have also seen that the distance value equals .2157. Therefore, since we recall from Section 2.3 that the standard error, s, is .3671, it follows that a 95 percent confidence interval for the mean fuel consumption is

$$\begin{aligned}\left[\hat{y} \pm t_{[.025]} s\sqrt{\text{Distance value}}\right] &= \left[10.333 \pm 2.571(.3671)\sqrt{.2157}\right]\\ &= [10.333 \pm .438]\\ &= [9.895, 10.771]\end{aligned}$$

Here, $t_{[.025]} = 2.571$ is based on $n - (k + 1) = 8 - 3 = 5$ degrees of freedom. This interval says we are 95 percent confident that mean weekly fuel consumption for all weeks having an average hourly temperature of 40°F and a chill index of 10 is between 9.895 MMcf of natural gas and 10.771 MMcf of natural gas. Furthermore, a 95 percent prediction interval for the individual fuel consumption is

$$\begin{aligned}\left[\hat{y} \pm t_{[.025]} s\sqrt{1 + \text{Distance value}}\right] &= \left[10.333 \pm 2.571(.3671)\sqrt{1.2157}\right]\\ &= [10.333 \pm 1.04]\\ &= [9.293, 11.374]\end{aligned}$$

This interval says that we are 95 percent confident that the amount of fuel consumed in a single week (next week) when the average hourly temperature is 40°F and the chill index is 10 will be between 9.293 MMcf of natural gas and 11.374 MMcf of natural gas.

The point prediction $\hat{y} = 10.333$ of next week's fuel consumption would be the natural gas company's transmission nomination (order of natural gas from the pipeline transmission service) for next week, This point prediction is the midpoint of the 95 percent prediction interval, [9.293, 11.374], for next week's fuel consumption. As previously calculated, the half-length of this interval is 1.04, and the 95 percent prediction interval can be expressed as [10.333 ± 1.04]. Therefore, since 1.04 is (1.04/10.333)100% = 10.07% of the transmission nomination of 10.333, the model makes us 95 percent confident that the actual amount of natural gas that will be used by the city next week will differ from the natural gas company's transmission nomination by no more than 10.07 percent. That is, we are 95 percent confident that the natural gas company's percentage nomination error will be less than or equal to 10.07 percent. It follows that this error will probably be within the 10 percent allowance granted by the pipeline transmission system, and it is unlikely that the natural gas company will be required to pay a transmission fine.

The bottom of the Minitab output in Figure 2.10 gives the point estimate and prediction $\hat{y} = 10.333$, along with the just calculated confidence and prediction intervals. Moreover, although the Minitab output does not directly give the distance value, it does give $s_{\hat{y}} = s\sqrt{\text{Distance value}}$ under the heading "SE Fit." Specifically, since the Minitab output tells us that $s_{\hat{y}}$ equals .170 and also tells us that s equals .3671, the Minitab output tells us that the distance value equals $(s_{\hat{y}} / s)^2 = (.170 / .3671)^2 = .2144515$. The reason that this value differs slightly from the value calculated using matrices is that the values of $s_{\hat{y}}$ and s on the Minitab output are rounded.

In order to use the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ to predict next week's fuel consumption on the basis of just the average hourly temperature of 40°F, recall from Example 2.2 that $b_0 = 15.84$, $b_1 = -.1279$, $\bar{x} = 43.98$, and $SS_{xx} = 1404.355$. Also recall from Section 2.3 that $s = .6542$. The simple linear regression model's point prediction of next week's fuel consumption is $\hat{y} = 15.84 - .1279(40) = 10.72$ MMcf of natural gas. Furthermore, we compute the distance value to be $(1/n) + (x_0 - \bar{x})^2 / SS_{xx} = (1/8) + (40 - 43.98)^2 / 1404.355 = .1362$. Since $t_{[.025]}$ based on $n - (k+1) = 8 - (1+1) = 6$ degrees of freedom is 2.447, a 95 percent prediction interval for next week's fuel consumption is

$$\left[\hat{y} \pm t_{[.025]} s\sqrt{1 + \text{Distance value}}\right] = \left[10.72 \pm 2.447(.6542)\sqrt{1.1362}\right]$$
$$= \left[10.72 \pm 1.71\right]$$
$$= \left[9.01, 12.43\right]$$

Now, consider using the point prediction $\hat{y} = 10.72$ given by the simple linear regression model as the natural gas company's transmission nomination for next week. Also, note that the half-length of the 95 percent prediction interval given by this model is 1.71, which is (1.71/10.72)100% = 15.91% of the transmission nomination. In this case we would be 95 percent confident that the actual amount of natural gas that will be used by the city next week will differ from the natural gas company's transmission nomination by no more than 15.91 percent. That is, we would be 95 percent confident that the natural gas company's percentage nomination error will be less than or equal to 15.91 percent. It follows that we would not be confident that the company's percentage nomination error will be within the 10 percent allowance granted by the pipeline transmission system. Consequently, the natural gas company needs to base its natural gas nomination on the point prediction $\hat{y}$ = 10.333 MMcf of natural gas given by the two independent variable fuel consumption model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.

To conclude this example, consider Figure 2.18. This figure illustrates-in the context of the fuel consumption model $y = \beta_0 + \beta_1 x + \varepsilon$ that uses only the average hourly temperature x - the effect of the distance value on the lengths of confidence intervals and prediction intervals. Specifically, this figure shows that as an individual value x_0 of x moves away from the center of the experimental region ($\bar{x}$ = 43.98), the distance value gets larger, and thus both the confidence interval for the mean value of y and the prediction interval for an individual value of y get longer.

2.8 Inverse Prediction In Simple Linear Regression

Ott and Longnecker (2010) present an example where an engineer wishes to calibrate a flow meter used on a liquid-soap production line. To perform the calibration, the engineer fixes the flow rate x on the production line at 10 different values—1, 2, 3, 4, 5, 6, 7, 8, 9, and 10—and observes the corresponding readings (y)—1.4, 2.3, 3.1, 4.2, 5.1, 5.8, 6.8, 7.6, 8.7, and 9.5—given

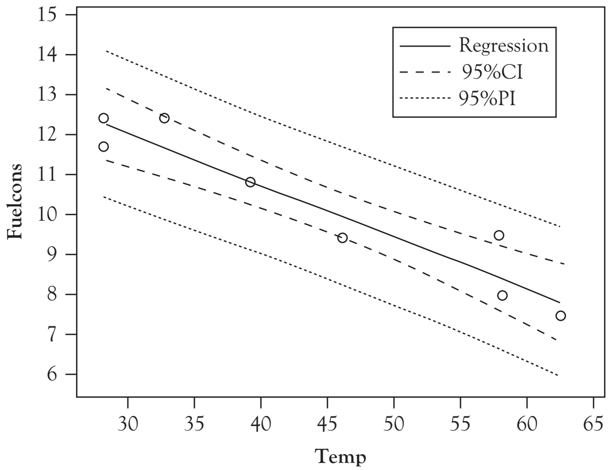


Figure 2.18 Confidence and prediction intervals for the fuel consumption model $y = \beta_0 + \beta_1 x + \varepsilon$

by the flow meter. If we consider fitting the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ to these data, we find that $\bar{x} = 5.5$, $\bar{y} = 5.45$, $SS_{xy} = 74.35$, and $SS_{xx} = 82.5$. This implies that $b_1 = SS_{xy} / SS_{xx} = 74.35 / 82.5 = .9012$ and $b_0 = \bar{y} - b_1 \bar{x} = 5.45 - .9012(5.5) = .4934$. Moreover, we find that $SSE = .0608$, $s^2 = SSE / (n-2) = .0608 / (10-2) = .0076$, $s = \sqrt{.0076} = .0872$, $s_{b_1} = s / \sqrt{SS_{xx}} = .0872 / \sqrt{82.5} = .0096$, and the t statistic for testing $H_0 : \beta_1 = 0$ is $t = b_1 / s_{b_1} = .9012 / .0096 = 93.87$. The *inverse prediction* problem asks us to predict the x value that corresponds to a particular y value. That is, sometime in the future the liquid soap production line will be in operation, we will make a meter reading y of the flow rate and we would like to know the actual flow rate x. The point prediction of and a 100 $(1 - \alpha)$ percent prediction interval for x are as follows.

Inverse Prediction

If the regression assumptions are satisfied for the simple linear regression model, then

1. A point prediction of the x value that corresponds to a particular y value is $\hat{x} = (y - b_0) / b_1$.

Inverse Prediction (Continued)

2. A $100(1-\alpha)$ percent prediction interval for the x value that corresponds to a particular y value is $[\hat{x}_L, \hat{x}_U]$, where

$$\hat{x}_L = \bar{x} + \frac{1}{1-c^2}\left[(\hat{x}-\bar{x}) - d\right]$$

$$\hat{x}_U = \bar{x} + \frac{1}{1-c^2}\left[(\hat{x}-\bar{x}) + d\right]$$

$$d = \frac{t_{[\alpha/2]}s}{b_1}\sqrt{\frac{(n+1)}{n}(1-c^2) + \frac{(\hat{x}-\bar{x})}{SS_{xx}}} \text{ and } c^2 = \frac{(t_{[\alpha/2]})^2 s^2}{b_1^2\, SS_{xx}}$$

Here $t_{[\alpha/2]}$ is based on $n-2$ degrees of freedom.

In order to discuss the prediction interval, note that

$$c = t_{[\alpha/2]}\; s / b_1\sqrt{SS_{xx}} \text{ can be shown to equal } t_{[\alpha/2]} / t$$

where $t = b_1 / s_{b_1}$. To use the prediction interval, we require that $|t| > t_{[\alpha/2]}$, which implies that $|c| < 1$, $c^2 < 1$, and $(1 - c^2)$ in the prediction interval formula is greater than zero and less than one. For example, suppose that we wish to have a point prediction of and a $100(1-\alpha)\% = 95\%$ prediction interval for the actual flow rate x that corresponds to a meter reading of $y = 4$. The point prediction of x is $\hat{x} = (y - b_0)/b_1 = (4 - .4934)/.9012 = 3.8910$. Moreover, $t_{[\alpha/2]} = t_{[.025]}$ (based on $n - 2 = 10 - 2 = 8$ degrees of freedom) is 2.306. Because t has been previously calculated to be 93.87 and because $|t| = 93.87 > 2.306 = t_{[.025]}$ we can calculate a 95 percent prediction interval for x as follows:

$$c^2 = \frac{(t_{[\alpha/2]})^2 s^2}{b_1^2\ \mathrm{SS}_{xx}} = \frac{(2.306)^2(.0076)}{(.9012)^2(82.5)} = .0006$$

$$1 - c^2 = .9994 \quad \bar{x} = 5.5 \quad s = .0872$$

$$\hat{x}_U = 5.5 + \frac{1}{.9994}\left[(3.8910 - 5.5) + \frac{2.306(.0872)}{.9012}\sqrt{\frac{11}{10}(.9994) + \frac{(3.8910-5.5)^2}{82.5}}\right]$$

$$= 5.5 + \frac{1}{.9994}(-1.6090 + .2373) = 4.1274$$

$$\hat{x}_L = 5.5 + \frac{1}{.9994}(-1.6090 - .2373) = 3.6526$$

Therefore, we are 95 percent confident that the actual flow rate when the meter reading is $y = 4$ is between 3.6526 and 4.1274.

2.9 Regression Through the Origin in Simple Linear Regression

It can be shown that the least squares point estimate of β_1 in the model $y = \beta_1 x + \varepsilon$ is $b_1 = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2$. We reject $H_0 : \beta_1 = 0$ in favor of $H_a : \beta_1 \neq 0$ at level significance α if $t = b_1 / s_{b_1}$ is greater in absolute value than $t_{[\alpha/2]}$, which is based on $(n - 1)$ degrees of freedom. Here $s_{b_1} = s / \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}$, where $s = \sqrt{SSE/(n-1)}$ and $SSE = \sum_{i=1}^{n}(y_i - b_1 x_i)^2$. If x_0 is an individual value of x, then a $100(1-\alpha)$ percent confidence interval for the mean value of y is $[\hat{y} \pm t_{[\alpha/2]} s (x_0^2 / \sum_{i=1}^{n} x_i^2)^{1/2}]$, and a $100(1-\alpha)$ percent prediction interval for an individual value of y is $[\hat{y} \pm t_{[\alpha/2]} s (1 + x_0^2 / \sum_{i=1}^{n} x_i^2)^{1/2}]$. Here, $\hat{y} = b_1 x_0$.

2.10 Using SAS

In Figure 2.19 we present the SAS program needed to carry out a multiple regression analysis of the sales territory performance data in Table 2.5(a). This program gives the SAS output in Table 2.5(c).

2.11 Exercises

Exercise 2.1

Ott (1984) presents twelve observations concerning y = weight loss of a compound (in pounds), x_1 = the amount of time the compound was exposed to the air (in hours), and x_2 = the relative humidity of the

```
DATA TERR;
INPUT Sales Time MktPoten Adver Mktshare Change;
DATALINES;
3669.88   43.10   74065.11   4582.88   2.51   .34
3473.95  108.13   58117.30   5539.78   5.51   .15
                     ⋮
2799.97   21.14   22809.53   3552.00   9.14  −.74
.         85.42   35182.73   7281.65   9.64   .28
PROC PRINT;
PROC REG DATA = TERR;
MODEL Sales = Time MktPoten Adver MktShare Change/P CLM CLI;
```

→ Sales territory performance data (See Table 2.5)

(Note: If we do not wish to have an intercept β_0 in the model, we would add in the command "NOINT" after the slash in the MODEL statement).

Figure 2.19 Sales territory performance data SAS program

environment during exposure. The twelve observations of y are 4.3, 5.5, 6.8, 8.0, 4.0, 5.2, 6.6, 7.5, 2.0, 4.0, 5.7, and 6.5. The corresponding observations of x_1 are 4, 5, 6, 7, 4, 5, 6, 7, 4, 5, 6, and 7. The corresponding observations of x_2 are .20, .20, .20, .20, .30, .30, .30, .30, .40, .40, .40, and .40. If we use the regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ to relate y to x_1, and x_2, then we define the following $\mathbf{y}$ vector and $\mathbf{X}$ matrix and make the following calculations:

$$\mathbf{y}=\begin{bmatrix}4.3\\5.5\\\vdots\\6.5\end{bmatrix} \quad \mathbf{X}=\begin{bmatrix}1 & 4 & .20\\1 & 5 & .20\\\vdots & \vdots & \vdots\\1 & 7 & .40\end{bmatrix} \quad \mathbf{X'X}=\begin{bmatrix}12 & 66 & 3.6\\66 & 378 & 19.8\\3.6 & 19.8 & 1.16\end{bmatrix}$$

$$(\mathbf{X'X})^{-1}=\begin{bmatrix}3.2250 & -0.3667 & -3.7500\\-0.3667 & 0.0667 & 0.0000\\-3.7500 & 0.0000 & 12.5000\end{bmatrix} \quad \mathbf{X'y}=\begin{bmatrix}66.1\\383.3\\19.19\end{bmatrix}$$

Using the data given and these matrices, show that (within rounding):

(a) $b_0 = 66667, b_1 = 1.31667,$ and $b_2 = -8.0$; also, interpret the meaning of these least squares point estimates.

(b) $SSE = \sum_{i=1}^{12} y_i^2 - \mathbf{b'X'y} = 1.3450$; $s^2 = .14944$; $s = .38658$.

(c) $\bar{y} = 5.50833$; Explained variation = $\mathbf{b'X'y} - n\bar{y}^2 = 31.12417$

(d) Total variation = $\sum_{i=1}^{12} y_i^2 - n\bar{y}^2 = 32.46917$.

(e) $R^2 = .958576$; also calculate $\bar{R}^2$.

(f) F(model) = 104.13; also test $H_0 : \beta_1 = \beta_2 = 0$ by setting α equal to .05 and using a rejection point; what does the test tell you?

(g) $s_{b_0} = s\sqrt{c_{00}} = .69423$, $t = b_0 / s_{b_0} = .96$; $s_{b_1} = s\sqrt{c_{11}} = .09981$, $t = b_1 / s_{b_1} = 13.19$; $s_{b_2} = s\sqrt{c_{22}} = 1.36677$, $t = b_2 / s_{b_2} = -5.85$; also test each of $H_0 : \beta_0 = 0$ versus $H_a : \beta_0 \neq 0$, $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$, and $H_0 : \beta_2 = 0$ versus $H_a : \beta_2 \neq 0$ by setting α equal to .05 and using a rejection point. What does each test tell you?

(h) Calculate 95 percent confidence intervals for β_0, β_1, and β_2. Interpret what these intervals say.

(i) Suppose that we are considering exposing the compound to the air for 6.5 hours at 35 percent relative humidity. Since we will expose many amounts of the same weight of the compound to the air, the mean weight loss per amount is of interest (because this mean multiplied by the number of amounts exposed approximates the total weight loss). Verify that $\hat{y}$ = 6.425 is a point estimate of and [6.05269, 6.79731] is a 95 percent confidence interval for the mean weight loss when x_1 = 6.5 and x_2 = .35. Are we 95 percent confident that the mean weight loss when x_1 = 6.5 and x_2 = .35 is less than 7 pounds. Explain. Find a point prediction of and a 95 percent prediction interval for the weight loss of an individual amount of the compound when x_1 = 6.5 and x_2 = .35.

Exercise 2.2

Recall that Figure 2.11 is the SAS output of a regression analysis of the sales territory performance data in Table 2.5 by using the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

(a) Show how F(model) = 40.91 has been calculated by using other quantities on the output. The SAS output tells us that the p-value related to F(model) is less than .0001. What does this say?

(b) The SAS output tells us that the p-values for testing the significance of the independent variables Time, MktPoten, Adver, MktShare, and Change are, respectively, .0065, < .0001, .0025, < .0001, and .0530. Interpret what these p-values say. Note: Although the p-value of .0530 for testing the significance of Change is larger than .05, we will see in Chapter 4 that retaining Change (x_2) in the model makes the model better.

(c) Consider a questionable sales representative for whom Time = 85.42, MktPoten = 35,182.73, Adver = 7281.65, MktShare = 9.64, and Change = .28. In Example 2.5 we have seen that the point prediction of the sales corresponding to this combination of values of the independent variables is $\hat{y} = 4182$ (that is, 418,200 units). In addition to giving $\hat{y} = 4182$, the SAS output tells us that $s_{\hat{y}} = s\sqrt{\text{Distance value}}$ (shown under the heading "Std Error Predict") is 141.8220. Since the SAS output also tells us that s for the sales territory performance model equals 430.23188, the distance value equals $(s_{\hat{y}}/s)^2$ = $(141.8220/430.23188)^2$ = .109. Specify what row vector $\mathbf{x}_0'$ SAS used to calculate the distance value by the matrix algebra expression $\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$. Then, use $\hat{y}$, the distance value, s, and $t_{[.025]}$ based on $n - (k+1) = 25 - (5+1) = 19$ degrees of freedom to verify that (within rounding) the 95 percent prediction interval for the sales corresponding to the questionable sales representative's values of the independent variables is [3234, 5130]. This interval is given on the SAS output. Recalling that the actual sales for the questionable representative were 3082, why does the prediction interval provide strong evidence that these actual sales were unusually low?

Exercise 2.3

Consider the model $y = \beta_1 x + \varepsilon$ describing regression through the origin in simple linear regression. For this model, the $\mathbf{y}$ column vector is a column vector containing the n observed values $y_1, y_2, \ldots, y_n$ of the dependent variable, and the matrix $\mathbf{X}$ is a column vector containing the n observed values $x_1, x_2, \ldots, x_n$ of the independent variable. Show that $\mathbf{X'X}$ equals $\sum_{i=1}^{n} x_i^2$, which implies that $(\mathbf{X'X})^{-1} = 1/\sum_{i=1}^{n} x_i^2$. Then show that the matrix algebra formula $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ gives the least squares point estimate $b_1 = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2$ of β_1.

CHAPTER 3

More Advanced Regression Models

3.1 Using Squared and Interaction Terms

One useful form of the linear regression model is what we call the *quadratic regression model.* Assuming that we have obtained n observations—each consisting of an observed value of y and a corresponding value of x—the model is as follows.

The quadratic regression model

The *quadratic regression model* relating y to x is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

where

1. $\beta_0 + \beta_1 x + \beta_2 x^2$ is $\mu_{y|x}$, the mean value of the dependent variable y when the value of the independent variable is x.
2. β_0, β_1, and β_2 are (unknown) *regression parameters* relating the mean value of y to x.
3. ε is an error term that describes the effects on y of all factors other than x and x^2.

The quadratic equation $\mu_{y|x} = \beta_0 + \beta_1 x + \beta_2 x^2$ that relates $\mu_{y|x}$ to x is the equation of a ***parabola***. Two parabolas are shown in Figure 3.1(a) and (b) and help to explain the meanings of the parameters β_0, β_1, and β_2. Here β_0 is the y-intercept of the parabola (the value of $\mu_{y|x}$ when $x = 0$).

Furthermore, β_1 is the *shift parameter* of the parabola: the value of β_1 shifts the parabola to the left or right. Specifically, increasing the value of β_1 shifts the parabola to the left. Lastly, β_2 is the *rate of curvature* of the parabola. If β_2 is greater than 0, the parabola opens upward (see Figure 3.1[a]). If β_2 is less than 0, the parabola opens downward (see Figure 3.1[b]). If a scatter plot of y versus x shows points scattered around a parabola, or a part of a parabola (some typical parts are shown in Figure 3.1[c], [d], [e], and [f]), then the quadratic regression model might appropriately relate y to x.

It is important to note that although the quadratic model employs the squared term x^2 and therefore assumes a curved relationship between the mean value of y and x, this model is a *linear regression model*. This is because the expression $\beta_0 + \beta_1 x + \beta_2 x^2$ expresses the mean value of y as a *linear function of the parameters* β_0, β_1, and β_2. In general, as long as the mean value of y is a *linear function of the regression parameters,* we are using a linear regression model.

Example 3.1

An oil company wishes to improve the gasoline mileage obtained by cars that use its premium unleaded gasoline. Company chemists suggest that an additive, ST-3000, be blended with the gasoline. In order to study the

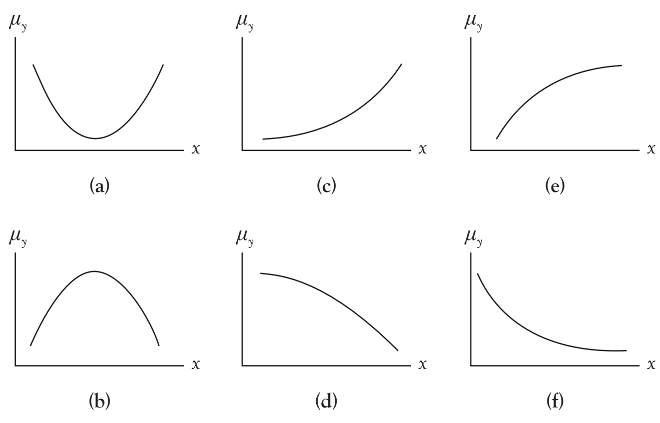


***Figure 3.1** The mean value of y changing in a quadratic fashion as x increases*

effects of this additive, mileage tests are carried out in a laboratory using test equipment that simulates driving under prescribed conditions. The amount of additive ST-3000 blended with the gasoline is varied, and the gasoline mileage for each test run is recorded. Table 3.1 gives the results of the test runs. Here the dependent variable y is gasoline mileage (in miles per gallon, mpg) and the independent variable x is the amount of additive ST-3000 used (measured as the number of units of additive added to each gallon of gasoline). One of the study's goals is to determine the number of units of additive that should be blended with the gasoline to maximize gasoline mileage. The company would also like to predict the maximum mileage that can be achieved using additive ST-3000.

Figure 3.2 gives a scatter plot of y versus x. Since the scatter plot has the appearance of a quadratic curve (that is, part of a parabola), it seems reasonable to relate y to x by using the quadratic model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Table 3.1 Gasoline mileage data

Additive units (x)	Gasoline mileage (y)
0	25.8
0	26.1
0	25.4
1	29.6
1	29.2
1	29.8
2	32.0
2	31.4
2	31.7
3	31.7
3	31.5
3	31.2
4	29.4
4	29.0
4	29.5

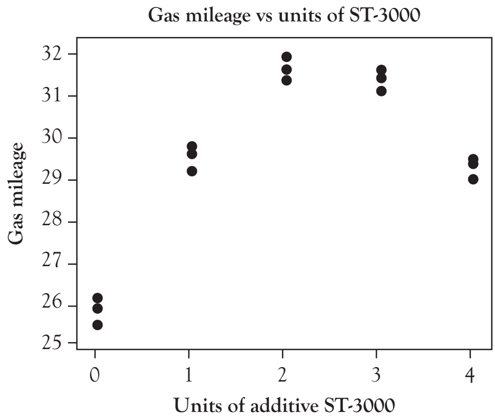


Figure 3.2 Scatter plot of gasoline mileage(y) versus number of units (x) of additive ST-3000

Figure 3.3 gives the MINITAB output of a regression analysis of the data using this quadratic model. Here the squared term x^2 is denoted as UnitsSq on the output. The MINITAB output tells us that the least squares point estimates of the model parameters are $b_0 = 25.7152$, $b_1 = 4.9762$, and $b_2 = -1.01905$. These estimates give us the least squares prediction equation

$$\hat{y} = 25.7152 + 4.9762x - 1.01905x^2$$

This is the equation of the best quadratic curve that can be fitted to the data plotted in Figure 3.2. The MINITAB output also tells us that the p-values related to x and x^2 are less than .001. This implies that we have very strong evidence that each of these model components is significant. The fact that x^2 seems significant confirms the graphical evidence that there is a quadratic relationship between y and x. Once we have such confirmation, we usually retain the linear term x in the model no matter what the size of its p-value. The reason is that geometrical considerations indicate that it is best to use both x and x^2 to model a quadratic relationship.

The oil company wishes to find the value of x that results in the highest predicted mileage. Using calculus, it can be shown that the value $x = 2.44$ maximizes predicted gas mileage. Therefore, the oil company can maximize

```
The regression equation is
Milleage = 25.7 + 4.98 Units - 1.02 UnitsSq

Predictor       Coef  SE Coef       T      P
Constant     25.7152   0.1554  165.43  0.000
Units         4.9762   0.1841   27.02  0.000
UnitsSq     -1.01905  0.04414  -23.09  0.000

S = 0.286079   R-Sq = 98.6%   R-Sq(adj) = 98.3%

Analysis of Variance
Source          DF      SS      MS       F      P
Regression       2  67.915  33.958  414.92  0.000
Residual Error  12   0.982   0.082
Total           14  68.897
    Fit  SE Fit        95% CI             95% PI
31.7901  0.1111 (31.5481, 32.0322) (31.1215, 32.4588)
```

Figure 3.3 MINITAB output for the gasoline mileage quadratic regression model

predicted mileage by blending 2.44 units of additive ST-3000 with each gallon of gasoline. This will result in a predicted gas mileage equal to

$$\begin{aligned}\hat{y} &= 25.7152 + 4.9762(2.44) - 1.01905(2.44)^2 \\ &= 31.7901 \text{ miles per gallon}\end{aligned}$$

This predicted mileage is the point estimate of the mean mileage that would be obtained by all gallons of the gasoline (when blended as just described) and is the point prediction of the mileage that would be obtained by an individual gallon of the gasoline. Note that $\hat{y} = 31.7901$ is given at the bottom of the MINITAB output in Figure 3.3. In addition, the MINITAB output tells us that a 95% confidence interval for the mean mileage that would be obtained by all gallons of the gasoline is [31.5481, 32.0322]. If the test equipment simulates driving conditions in a particular automobile, this confidence interval implies that an owner of the automobile can be 95% confident that he or she will average between 31.5481 mpg and 32.0322 mpg when using a very large number of gallons of the gasoline. The MINITAB output also tells us that a 95% prediction interval for the mileage that would be obtained by an individual gallon of the gasoline is [31.1215, 32.4588].

Multiple regression models often contain *interaction variables*. We form an interaction variable by multiplying two independent variables together. For instance, if a regression model includes the independent variables x_1 and x_2, then we can form the interaction variable x_1x_2. It is appropriate to employ an interaction variable if the relationship between the dependent variable y and one of the independent variables depends upon the value of the other independent variable. In the following example we consider a multiple regression model that uses a linear variable, a squared variable, and an interaction variable.

Example 3.2

Enterprise Industries produces *Fresh*, a brand of liquid laundry detergent. In order to more effectively manage its inventory and make revenue projections, the company would like to better predict demand for Fresh. To develop a prediction model, the company has gathered data concerning demand for Fresh over the last 30 sales periods (each sales period is defined to be a four-week period). The demand data are presented in Table 3.2. Here, for each sales period,

y = the demand for the large size bottle of Fresh (in hundreds of thousands of bottles) in the sales period

x_1 = the price (in dollars) of Fresh as offered by Enterprise Industries in the sales period

x_2 = the average industry price (in dollars) of competitors' similar detergents in the sales period

x_3 = Enterprise Industries' advertising expenditure (in hundreds of thousands of dollars) to promote Fresh in the sales period

$x_4 = x_2 - x_1$ = the "price difference" in the sales period

To begin our analysis, suppose that Enterprise Industries believes on theoretical grounds that the single independent variable x_4 adequately describes the effects of x_1 and x_2 on y. That is, perhaps demand for Fresh depends more on how the price for Fresh compares to competitors' prices than it does on the absolute levels of the prices for Fresh and other competing detergents. This makes sense since most consumers must buy a certain amount of detergent no matter what the price might be.

Figures 3.4 and 3.5 present scatter plots of y versus x_4 and y versus x_3. Because the plot in Figure 3.4 shows a linear relationship between y

Table 3.2 Historical data, including price differences, concerning demand for Fresh detergent

Sales period	Price for Fresh x_1($)	Average industry price, x_2 ($)	Price difference, $x_4 = x_2 - x_1$ ($)	Advertising expenditure for Fresh, x_3 (×$100,000)	Demand for Fresh, y (×100,000 bottles)
1	3.85	3.80	−.05	5.50	7.38
2	3.75	4.00	.25	6.75	8.51
3	3.70	4.30	.60	7.25	9.52
4	3.70	3.70	0	5.50	7.50
5	3.60	3.85	.25	7.00	9.33
6	3.60	3.80	.20	6.50	8.28
7	3.60	3.75	.15	6.75	8.75
8	3.80	3.85	.05	5.25	7.87
9	3.80	3.65	−.15	5.25	7.10
10	3.85	4.00	.15	6.00	8.00
11	3.90	4.10	.20	6.50	7.89
12	3.90	4.00	.10	6.25	8.15
13	3.70	4.10	.40	7.00	9.10
14	3.75	4.20	.45	6.90	8.86
15	3.75	4.10	.35	6.80	8.90
16	3.80	4.10	.30	6.80	8.87
17	3.70	4.20	.50	7.10	9.26
18	3.80	4.30	.50	7.00	9.00
19	3.70	4.10	.40	6.80	8.75
20	3.80	3.75	−.05	6.50	7.95
21	3.80	3.75	−.05	6.25	7.65
22	3.75	3.65	−.10	6.00	7.27
23	3.70	3.90	.20	6.50	8.00
24	3.55	3.65	.10	7.00	8.50
25	3.60	4.10	.50	6.80	8.75
26	3.65	4.25	.60	6.80	9.21
27	3.70	3.65	−.05	6.50	8.27
28	3.75	3.75	0	5.75	7.67
29	3.80	3.85	.05	5.80	7.93
30	3.70	4.25	.55	6.80	9.26

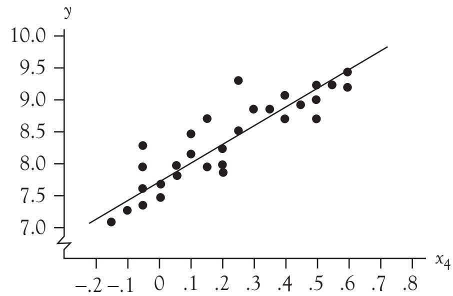


Figure 3.4 Plot of y versus x_4

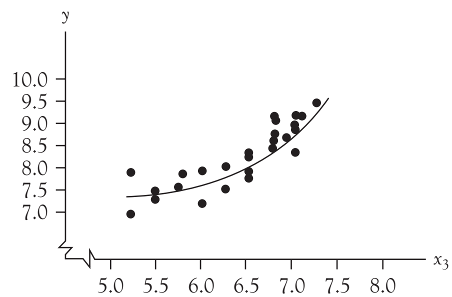


Figure 3.5 Plot of y versus x_3

and x_4, we should use x_4 to predict y. Because the plot in Figure 3.5 shows a quadratic relationship between y and x_3, we should use x_3 and x_3^2 to predict y. Moreover, if x_4 and x_3 interact, then we should use the interaction term x_4x_3 to predict y. This gives the model

$$y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \varepsilon$$

By using the data in Table 3.2, we define the column vector

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{30} \end{bmatrix} = \begin{bmatrix} 7.38 \\ 8.51 \\ 9.52 \\ \vdots \\ 9.26 \end{bmatrix}$$

and the matrix

$$\begin{array}{cc} & \begin{array}{ccccc} 1 & x_4 & x_3 & x_3^2 & x_4x_3 \end{array} \\ \mathbf{X} = & \begin{bmatrix} 1 & -.05 & 5.50 & (5.50)^2 & (-.05)(5.50) \\ 1 & .25 & 6.75 & (6.75)^2 & (.25)(6.75) \\ 1 & .60 & 7.25 & (7.25)^2 & (.60)(7.25) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & .55 & 6.80 & (6.80)^2 & (.55)(6.80) \end{bmatrix} \end{array}$$

$$\begin{array}{cc} & \begin{array}{ccccc} 1 & x_4 & x_3 & x_3^2 & x_4x_3 \end{array} \\ = & \begin{bmatrix} 1 & -.05 & 5.50 & 30.25 & -.275 \\ 1 & .25 & 6.75 & 45.5625 & 1.6875 \\ 1 & .60 & 7.25 & 52.5625 & 4.35 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & .55 & 6.80 & 46.24 & 3.74 \end{bmatrix} \end{array}$$

Thus we can calculate the least squares point estimates of β_0, β_1, β_2, and β_3 to be

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$= \begin{bmatrix} 1315.261 & 543.4463 & -433.586 & 35.50156 & -83.4036 \\ 543.4463 & 464.2447 & -179.952 & 14.80313 & -69.5252 \\ -433.586 & -179.952 & 143.1914 & 11.7449 & 27.67939 \\ 35.50156 & 14.80313 & -11.7449 & 0.965045 & -2.28257 \\ -83.4036 & -69.5252 & 27.67939 & -2.28257 & 10.45448 \end{bmatrix} \begin{bmatrix} 251.48 \\ 57.646 \\ 1632.781 \\ 10677.4 \\ 397.7442 \end{bmatrix}$$

$$= \begin{bmatrix} 29.11329 \\ 11.13423 \\ -7.60801 \\ 0.6712472 \\ -1.47772 \end{bmatrix}$$

Figure 3.6 presents the SAS output obtained by using the interaction model to perform a regression analysis of the Fresh demand data. This output shows that each of the *p*-values for testing the significance of the intercept and the independent variables is less than .05. Therefore, we have strong

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	12.39419	3.09855	72.78	<.0001
Error	25	1.06440	0.04258		
Corrected Total	29	13.45859			

Root MSE	0.20634	R-Square	0.9209
Dependent Mean	8.38267	Adj R-Sq	0.9083
Coeff Var	2.46150		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	29.11329	7.48321	3.89	0.0007
x4	PriceDif	1	11.13423	4.44585	2.50	0.0192
x3	AdvExp	1	-7.60801	2.46911	-3.08	0.0050
x3SQ	x3 ** 2	1	0.67125	0.20270	3.31	0.0028
x4x3	x4 * x3	1	-1.47772	0.66716	-2.21	0.0361

Obs	Dep Var Demand	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict	
31	.	8.3272	0.0563	8.2112	8.4433	7.8867	8.7678

Figure 3.6 *SAS output of a regression analysis of the Fresh demand data using the interaction model* $y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \varepsilon$

evidence that the intercept and each of x_4, x_3, x_3^2, and x_4x_3 are significant. In particular, since the p-value related to x_4x_3 is .0361, we have strong evidence that the interaction variable x_4x_3 is important. This confirms that the interaction between x_4 and x_3 that we suspected really does exist.

Suppose that Enterprise Industries wishes to predict demand for Fresh in a future sales period when the price difference will be \$.20 ($x_4 = .20$) and when the advertising expenditure for Fresh will be \$650,000 ($x_3 = 6.50$). Using the least squares point estimates in Figure 3.6, the needed point prediction is

$$\begin{aligned}\hat{y} &= 29.11329 + 11.13423(.20) - 7.60801(6.50) + .67125(6.50)^2 \\ &\quad - 1.47772(.20)(6.50) \\ &= 8.3272 \text{ (832,720 bottles)}\end{aligned}$$

This point prediction is given on the SAS output of Figure 3.6, which also tells us that the 95% confidence interval for mean demand when x_4 equals .20 and x_3 equals 6.50 is [8.2112, 8.4433] and that the 95% prediction interval for an individual demand when x_4 equals .20 and x_3 equals 6.50 is [7.8867, 8.7678]. Here, since

$$\mathbf{x}_0' = [1 \;\; .20 \;\; 6.50 \;\; (6.50)^2 \;\; (.20)(6.50)] = [1 \; .20 \; 6.50 \; 42.25 \; 1.3]$$

the distance value can be computed to be $\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = .07366$. Since $s = .20634$ and $n-(k+1) = 30-5 = 25$, the 95% prediction interval for the demand is

$$\begin{aligned}\left[\hat{y} \pm t_{[.025]}s\sqrt{1+\text{Distance value}}\right] &= \left[8.3272 \pm 2.060(.20634)\sqrt{1+.07366}\right] \\ &= [7.8867, 8.7678]\end{aligned}$$

This interval says that we are 95 percent confident that the actual demand in the future sales period will be between 788,670 bottles and 876,780 bottles. The upper limit of this interval can be used for inventory control. It says that if Enterprise Industries plans to have 876,780 bottles on hand to meet demand in the future sales period, then the company can be very confident that it will have enough bottles. The lower limit of the interval can be used to better understand Enterprise Industries' cash flow situation. It says the company can be very confident that it will sell at least 788,670 bottles in the future sales period.

To investigate the nature of the interaction between x_3 and x_4, consider the prediction equation

$$\hat{y} = 29.11329 + 11.13423x_4 - 7.60801x_3 + .67125x_3^2 - 1.47772x_4 x_3$$

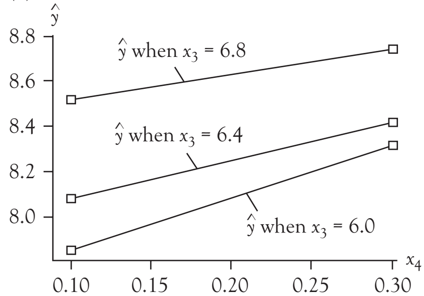
obtained from the least squares point estimates in Figure 3.6. Also, consider the six combinations of price difference x_4 and advertising expenditure x_3 obtained by combining the x_4 values .10 and .30 with the x_3 values 6.0, 6.4, and 6.8. When we use the prediction equation to predict the demands for Fresh corresponding to these six combinations, we obtain the predicted demands ($\hat{y}$) shown in Figure 3.7(a) (Note that we consider *two* x_4 values because there is a *linear* relationship between y and x_4, and we consider *three* x_3 values because there is a *quadratic* relationship between y and x_3). Now

1. If we fix x_3 at 6.0 in Figure 3.7(a) and plot the corresponding $\hat{y}$ values 7.86 and 8.31 versus the x_4 values .10 and .30, we obtain the two squares connected by the lowest line in Figure 3.7(b). Similarly, if we fix x_3 at 6.4 and plot the corresponding $\hat{y}$ values 8.08 and 8.42 versus the x_4 values .10 and .30, we obtain the two squares connected by the middle line in Figure 3.7(b). Also, if we fix x_3 at 6.8 and plot the corresponding $\hat{y}$ values 8.52 and 8.74 versus the x_4 values .10 and .30, we obtain the two squares connected by the highest line in Figure 3.7(b). Examining the three lines relating $\hat{y}$ to x_4, we see that the slopes of these lines decrease as x_3 increases from 6.0 to 6.4 to 6.8. This says that as the price difference x_4 increases from .10 to .30 (that is, as Fresh becomes less expensive compared to its competitors), the *rate of increase* of predicted demand $\hat{y}$ is slower when advertising expenditure x_3 is higher than when advertising expenditure x_3 is lower. Moreover, this might be logical because it says that when a higher advertising expenditure makes more customers aware of Fresh's cleaning abilities and thus causes customer demand for Fresh to be higher, there is less opportunity for an increased price difference to increase demand for Fresh.

(a)

x_3 \ x_4	.10	.30
6.0	7.86	8.31
6.4	8.08	8.42
6.8	8.52	8.74

(b)



(c)

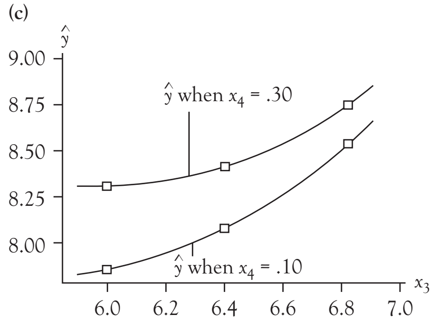


Figure 3.7 Interaction between x_4 and x_3 *(a) predicted demands ($\hat{y}$ values) (b) plots of $\hat{y}$ versus x_4 for different x_3 values (c) plots of $\hat{y}$ versus x_3 for different x_4 values*

2. If we fix x_4 at .10 in Figure 3.7(a) and plot the corresponding $\hat{y}$ values 7.86, 8.08, and 8.52 versus the x_3 values 6.0, 6.4, and 6.8, we obtain the three squares connected by the lower quadratic curve in Figure 3.7(c). Similarly, if we fix x_4 at .30 and plot the corresponding $\hat{y}$ values 8.31, 8.42, and 8.74 versus the x_3 values 6.0, 6.4, and 6.8, we obtain the three squares connected by the higher quadratic curve in Figure 3.7(c). The nonparallel quadratic curves in Figure 3.7(c) say that as advertising expenditure x_3 increases from 6.0 to 6.4 to 6.8, the rate of increase of predicted demand $\hat{y}$ is slower when the price difference x_4 is larger (that is, $x_4 = .30$) than when the price difference x_4 is smaller (that is, $x_4 = .10$). Moreover, this might be logical because it says that when a larger price difference causes customer demand for Fresh to be higher, there is less opportunity for an increased advertising expenditure to increase demand for Fresh.

To summarize the nature of the interaction between x_4 and x_3, we might say that a higher value of each of these independent variables somewhat weakens the impact of the other independent variable on predicted demand. In Exercise 3.1 we will consider a situation where a higher value of each of two independent variables somewhat strengthens the impact of the other independent variable on the predicted value of the dependent variable. Moreover, if the *p*-value related to x_4x_3 in the Fresh detergent situation had been large and thus we had removed x_4x_3 from the model (that is, *no interaction*), then the plotted lines in Figure 3.7(b) would have been *parallel* and the plotted quadratic curves in Figure 3.7(c) would have been *parallel*. This would mean that predicted demand always responds in the same way to a change in one independent variable, regardless of the other independent variable's value.

As another example, if we perform a regression analysis of the fuel consumption data by using the model

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

we find that the *p*-value for testing $H_0 : \beta_3 = 0$ is .787. Therefore, we conclude that the interaction term x_1x_2 is not needed and that there is little or no interaction between the average hourly temperature and the chill index.

A final comment is in order. If a *p*-value indicates that an interaction term (say, x_1x_2) is important, then it is usual practice to retain the corresponding linear terms (x_1 and x_2) in the model no matter what the size of their *p*-values. The reason is that doing so can be shown to give a model that will better describe the interaction between x_1 and x_2.

3.2 Using Dummy Variables to Model Qualitative Independent Variables

The levels (or values) of a quantitative independent variable are numerical, whereas the levels of a *qualitative* independent variable are defined by describing them. For instance, the type of sales technique used by a door-to-door salesperson is a qualitative independent variable. Here we might define three different levels—high pressure, medium pressure, and low pressure.

We can model the effects of the different levels of a qualitative independent variable by using what we call *dummy variables* (also called *indicator variables*). Such variables are usually defined so that they take on two values—either 0 or 1. To see how we use dummy variables, we begin with an example.

Example 3.3

Part 1: The Data and Data Plots

Suppose that Electronics World, a chain of stores that sells audio and video equipment, has gathered the data in Table 3.3. These data concern store sales volume in July of last year (y, measured in thousands of dollars), the number of households in the store's area (x, measured in thousands), and the location of the store (on a suburban street or in a suburban shopping mall—a qualitative independent variable). Figure 3.8 gives a data plot of y versus x. Stores having a street location are plotted as solid dots, while stores having a mall location are plotted as asterisks. Notice that the line relating y to x for mall locations has a higher y-intercept than does the line relating y to x for street locations.

Table 3.3 The electronics world sales volume data

Store	Number of households, x ($\times 1000$)	Location	Sales volume, y ($\times 1000$)
1	161	Street	157.27
2	99	Street	93.28
3	135	Street	136.81
4	120	Street	123.79
5	164	Street	153.51
6	221	Mall	241.74
7	179	Mall	201.54
8	204	Mall	206.71
9	214	Mall	229.78
10	101	Mall	135.22

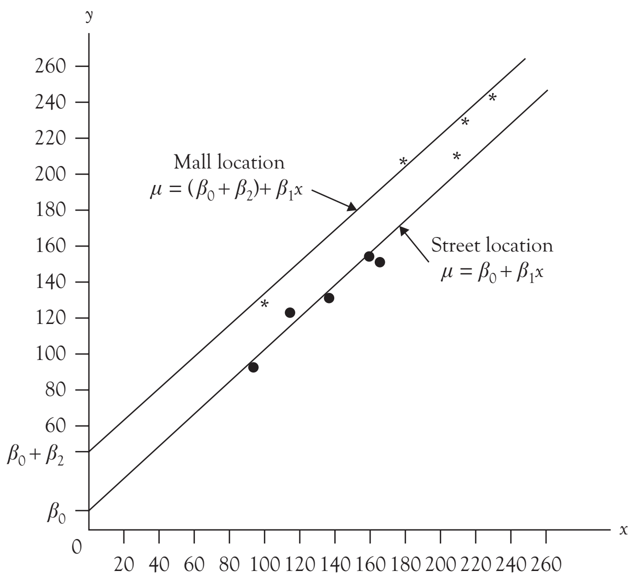


Figure 3.8 Plot of the sales volume data and a geometrical interpretation of the model $y = \beta_0 + \beta_1 x + \beta_2 D_M + \varepsilon$

Part 2: A Dummy Variable Model

In order to model the effects of the street and shopping mall locations, we define a dummy variable denoted D_M as follows:

$$D_M = \begin{cases} 1 & \text{if a store is in a mall location} \\ 0 & \text{otherwise} \end{cases}$$

Using this dummy variable, we consider the regression model

$$y = \beta_0 + \beta_1 x + \beta_2 D_M + \varepsilon$$

This model and the definition of D_M imply that

1. For a street location, mean sales volume equals

$$\begin{aligned} \beta_0 + \beta_1 x + \beta_2 D_M &= \beta_0 + \beta_1 x + \beta_2 (0) \\ &= \beta_0 + \beta_1 x \end{aligned}$$

2. For a mall location, mean sales volume equals

$$\begin{aligned}\beta_0 + \beta_1 x + \beta_2 D_M &= \beta_0 + \beta_1 x + \beta_2(1) \\ &= (\beta_0 + \beta_2) + \beta_1 x\end{aligned}$$

Thus the dummy variable allows us to model the situation illustrated in Figure 3.8. Here, the lines relating mean sales volume to x for street and mall locations have different y intercepts—β_0 and $(\beta_0 + \beta_2)$—and the same slope β_1. It follows that this dummy variable model assumes no interaction between x and store location—note the *parallel* data patterns for the street and mall locations in Figure 3.8. Also, note that β_2 is the difference between the mean monthly sales volume for stores in mall locations and the mean monthly sales volume for stores in street locations, when all these stores have the same number of households in their areas. If we use a computer software package, we find that the least squares point estimate of β_2 is $b_2 = 29.216$ and that the associated p-value is .0012. The point estimate says that for any given number of households in a store's area, we estimate that the mean monthly sales volume in a mall location is \$29,216 greater than the mean monthly sales volume in a street location.

Part 3: A Dummy Variable Model for Comparing Three Locations

In addition to the data concerning street and mall locations in Table 3.3, Electronics World has also collected data concerning downtown locations. The complete data set is given in Table 3.4 and plotted in Figure 3.9. Here, stores having a downtown location are plotted as open circles. A model describing these data is

$$y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \varepsilon$$

Here, the dummy variable D_M is as previously defined, and the dummy variable D_D is defined as follows:

$$D_D = \begin{cases} 1 & \text{if a store is in a downtown location} \\ 0 & \text{otherwise} \end{cases}$$

Table 3.4 The complete electronics world sales volume data

Store	Number of households, x (×1000)	Location	Sales volume, y (×1000)
1	161	Street	157.27
2	99	Street	93.28
3	135	Street	136.81
4	120	Street	123.79
5	164	Street	153.51
6	221	Mall	241.74
7	179	Mall	201.54
8	204	Mall	206.71
9	214	Mall	229.78
10	101	Mall	135.22
11	231	Downtown	224.71
12	206	Downtown	195.29
13	248	Downtown	242.16
14	107	Downtown	115.21
15	205	Downtown	197.82

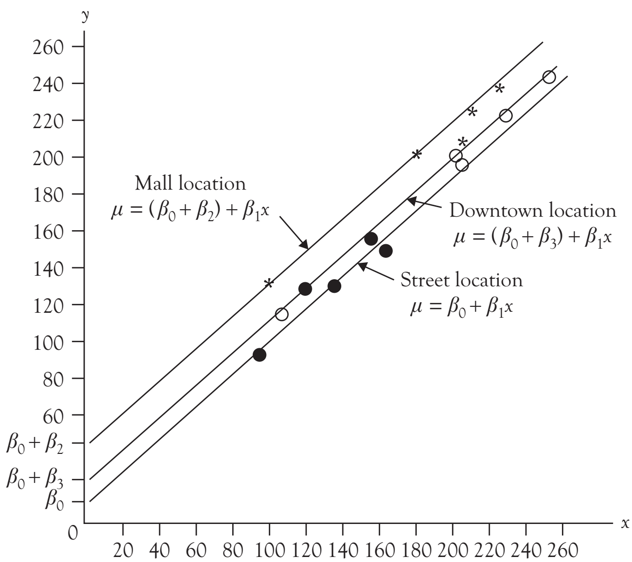


Figure 3.9 Plot of the complete Electronics World sales volume data and a geometrical interpretation of the model $y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \varepsilon$

It follows that

1. for a street location, mean sales volume equals

$$\begin{aligned}\beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D &= \beta_0 + \beta_1 x + \beta_2(0) + \beta_3(0)\\ &= \beta_0 + \beta_1 x\end{aligned}$$

2. for a mall location, mean sales volume equals

$$\begin{aligned}\beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D &= \beta_0 + \beta_1 x + \beta_2(1) + \beta_3(0)\\ &= (\beta_0 + \beta_2) + \beta_1 x\end{aligned}$$

3. for a downtown location, mean sales volume equals

$$\begin{aligned}\beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D &= \beta_0 + \beta_1 x + \beta_2(0) + \beta_3(1)\\ &= (\beta_0 + \beta_3) + \beta_1 x\end{aligned}$$

Thus the dummy variables allow us to model the situation illustrated in Figure 3.9. Here the lines relating mean sales volume to x for street, mall, and downtown locations have different y-intercepts—β_0, $(\beta_0 + \beta_2)$, and $(\beta_0 + \beta_3)$—and the same slope β_1. It follows that this dummy variable model assumes no interaction between x and store location.

In order to find the least squares point estimates of β_0, β_1, β_2, and β_3 in the dummy variable model, we use the data in Table 3.4 to define the column vector **y** and matrix **X** that are shown in Figure 3.10. It then follows that the least squares point estimates of β_0, β_1, β_2, and β_3 are

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} 14.978 \\ .8686 \\ 28.374 \\ 6.864 \end{bmatrix}$$

Part 4: Comparing the Locations

To compare the effects of the street, shopping mall, and downtown locations, consider comparing three means, which we denote as $\mu_{h,S}$, $\mu_{h,M}$, and $\mu_{h,D}$.

$$\mathbf{y} = \begin{bmatrix} 157.27 \\ 93.28 \\ 136.81 \\ 123.79 \\ 153.51 \\ 241.74 \\ 201.54 \\ 206.71 \\ 229.78 \\ 135.22 \\ 224.71 \\ 195.29 \\ 242.16 \\ 115.21 \\ 197.82 \end{bmatrix} \qquad \mathbf{X} = \begin{array}{c} \begin{array}{cccc} 1 & x & D_M & D_D \end{array} \\ \begin{bmatrix} 1 & 161 & 0 & 0 \\ 1 & 99 & 0 & 0 \\ 1 & 135 & 0 & 0 \\ 1 & 120 & 0 & 0 \\ 1 & 164 & 0 & 0 \\ 1 & 221 & 1 & 0 \\ 1 & 179 & 1 & 0 \\ 1 & 204 & 1 & 0 \\ 1 & 214 & 1 & 0 \\ 1 & 101 & 1 & 0 \\ 1 & 231 & 0 & 1 \\ 1 & 206 & 0 & 1 \\ 1 & 248 & 0 & 1 \\ 1 & 107 & 0 & 1 \\ 1 & 205 & 0 & 1 \end{bmatrix} \end{array}$$

Figure 3.10 The column vector* y *and matrix* X *using the data in Table 3.4 and the model $y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \varepsilon$

These means represent the mean sales volumes at stores having h households in the area and located on streets, in shopping malls, and downtown, respectively. If we set $x = h$, it follows that

$$\begin{aligned} \mu_{h,S} &= \beta_0 + \beta_1 h + \beta_2(0) + \beta_3(0) \\ &= \beta_0 + \beta_1 h \\ \mu_{h,M} &= \beta_0 + \beta_1 h + \beta_2(1) + \beta_3(0) \\ &= \beta_0 + \beta_1 h + \beta_2 \end{aligned}$$

and

$$\begin{aligned} \mu_{h,D} &= \beta_0 + \beta_1 h + \beta_2(0) + \beta_3(1) \\ &= \beta_0 + \beta_1 h + \beta_3 \end{aligned}$$

In order to compare street and mall locations, we look at

$$\mu_{h,M} - \mu_{h,S} = (\beta_0 + \beta_1 h + \beta_2) - (\beta_0 + \beta_1 h) = \beta_2$$

which is the difference between the mean sales volume for stores in mall locations having h households in the area and the mean sales volume for stores in street locations having h households in the area. Figure 3.11 gives the MINITAB output of a regression analysis of the data in Table 3.4 by using the dummy variable model. The output tells us that the least squares point estimate of β_2 is $b_2 =$ 28.374. This says that for any given number

```
The regression equation is
y = 15.0 + 0.869 x + 28.4 DM +  6.86 DD

Predictor       Coef    SE Coef       T       P
Constant      14.978      6.188    2.42   0.034
x            0.86859    0.04049   21.45   0.000
DM            28.374      4.461    6.36   0.000
DD             6.864      4.770    1.44   0.178

S = 6.34941   R-Sq = 98.7%   R-Sq(adj) = 98.3%

Analysis of Variance

Source            DF      SS      MS        F       P
Regression         3   33269   11090   275.07   0.000
Residual Error    11     443      40
Total             14   33712

    Fit  SE Fit         95% CI              95% PI
 217.07    2.91   (210.65, 223.48)   (201.69, 232.45)
```

Figure 3.11 MINITAB output of a regression analysis of the sales volume data using the model $\boldsymbol{y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \varepsilon}$

of households in a store's area, we estimate that the mean monthly sales volume in a mall location is \$28,374 greater than the mean monthly sales volume in a street location. Furthermore, since the output tells us that $s_{b_2} = 4.461$, and since $t_{[.025]}$ based on $n-(k+1)=15-(3+1)=11$ degrees of freedom is 2.201, a 95 percent confidence interval for β_2 is

$$\begin{aligned}[b_2 \pm t_{[.025]} s_{b_2}] &= [28.374 \pm 2.201(4.461)] \\ &= [18.554,\ 38.193]\end{aligned}$$

This interval says we are 95 percent confident that for any given number of households in a store's area, the mean monthly sales volume in a mall location is between \$18,554 and \$38,193 greater than the mean monthly sales volume in a street location. The MINITAB output also shows that the t-statistic for testing $H_0 : \beta_2 = 0$ versus $H_a :\ \beta_2 \neq 0$ equals 6.36 and that the related p-value is less than .001. Therefore, we have very strong evidence that there is a difference between the mean monthly sales volumes in mall and street locations.

In order to compare downtown and street locations, we look at

$$\mu_{h,D} - \mu_{h,S} = (\beta_0 + \beta_1 h + \beta_3) - (\beta_0 + \beta_1 h) = \beta_3$$

Since the MINITAB output in Figure 3.11 tells us that $b_3 = 6.864$, we estimate that for any given number of households in a store's area, the mean monthly sales volume in a downtown location is \$6,864 greater than the mean monthly sales volume in a street location. Furthermore, since the output tells us that $s_{b_3} = 4.770$, a 95 percent confidence interval for β_3 is

$$\begin{aligned}[b_3 \pm t_{[.025]} s_{b_3}] &= [6.864 \pm 2.201(4.770)] \\ &= [-3.636, 17.363]\end{aligned}$$

This says we are 95 percent confident that for any given number of households in a store's area, the mean monthly sales volume in a downtown location is between \$3,636 less than and \$17,363 greater than the mean monthly sales volume in a street location. The MINITAB output also shows that the t-statistic and *p-value* for testing $H_0 : \beta_3 = 0$ versus $H_a : \beta_3 \neq 0$ are $t = 1.44$ and p-value = .178. Therefore, we do not have strong evidence that there is a difference between the mean monthly sales volumes in downtown and street locations.

In order to compare mall and downtown locations, we look at

$$\mu_{h,M} - \mu_{h,D} = (\beta_0 + \beta_1 h + \beta_2) - (\beta_0 + \beta_1 h + \beta_3) = \beta_2 - \beta_3$$

The least squares point estimate of this difference is

$$b_2 - b_3 = 28.374 - 6.864 = 21.51$$

This says that for any given number of households in a store's area we estimate that the mean monthly sales volume in a mall location is \$21,510 greater than the mean monthly sales volume in a downtown location. There are two approaches for calculating a confidence interval for $\mu_{h,M} - \mu_{h,D}$ and for testing the null hypothesis $H_0 : \mu_{h,M} - \mu_{h,D} = 0$. Because $\mu_{h,M} - \mu_{h,D}$ equals the *linear combination* $\beta_2 - \beta_3$ of the β_j's in the model $y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \varepsilon$, one approach shows how to make statistical inferences about a linear combination of β_j's. This approach is discussed in Section 3.5. The other approach, discussed near the end of this section, involves specifying an alternative dummy variable regression model which is such that $\mu_{h,M} - \mu_{h,D}$ is equal to a single β_j in that model. Using either approach, we will find that there is very strong evidence that the mean monthly sales volume in a mall location is greater than the mean monthly sales volume in a downtown location. In summary, the mall location seems to give a greater mean monthly sales volume than either the street or downtown location.

Part 5: Predicting a Future Sales Volume

Suppose that Electronics World wishes to predict the sales volume in a future month for an individual store that has 200,000 households in its area and is located in a shopping mall. The needed point prediction is (since $D_M = 1$ and $D_D = 0$ when a store is in a shopping mall)

$$\begin{aligned}\hat{y} &= b_0 + b_1(200) + b_2(1) + b_3(0)\\ &= 14.978 + .8686(200) + 28.374(1)\\ &= 217.07\end{aligned}$$

which is given at the bottom of the MINITAB output in Figure 3.11. Furthermore, since $\mathbf{x}_0' = [1 \quad 200 \quad 1 \quad 0]$, the distance value can be computed to be $\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = .21063$. Since $s = 6.34941$, a 95 percent prediction interval for the sales volume is

$$\begin{aligned}\left[\hat{y} \pm t_{[.025]} s\sqrt{1 + \text{Distance value}}\right] &= [217.07 \pm 2.201(6.34941)\sqrt{1 + .21063}]\\ &= [201.69,\ 232.45]\end{aligned}$$

This prediction interval, which is also given on the MINITAB output, says we are 95 percent confident that the sales volume in a future sales period for an individual mall store that has 200,000 households in its area will be between \$201,690 and \$232,450.

Part 6: An Interaction Model

In modeling the sales volume data we might consider using the model

$$y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \beta_4 x D_M + \beta_5 x D_D + \varepsilon$$

This model implies that

1. for a street location, mean sales volume equals (since $D_M = 0$ and $D_D = 0$)

$$\beta_0 + \beta_1 x + \beta_2(0) + \beta_3(0) + \beta_4 x(0) + \beta_5 x(0)$$
$$= \beta_0 + \beta_1 x$$

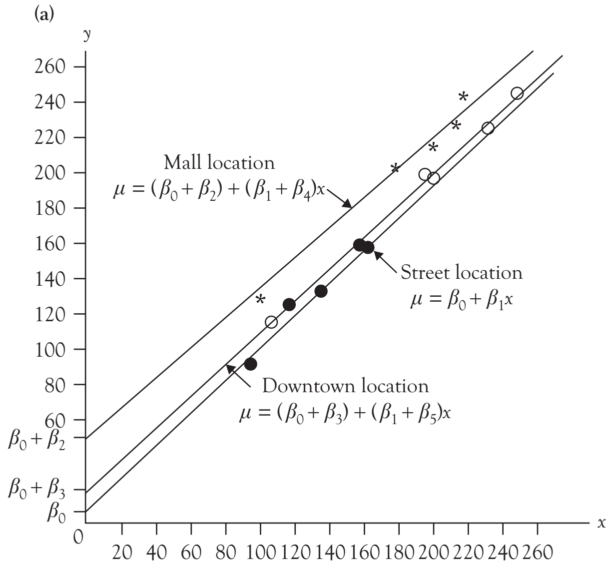
2. for a mall location, mean sales volume equals (since $D_M = 1$ and $D_D = 0$)

$$\beta_0 + \beta_1 x + \beta_2(1) + \beta_3(0) + \beta_4 x(1) + \beta_5 x(0)$$
$$= (\beta_0 + \beta_2) + (\beta_1 + \beta_4)x$$

3. for a downtown location, mean sales volume equals (since $D_M = 0$ and $D_D = 1$)

$$\beta_0 + \beta_1 x + \beta_2(0) + \beta_3(1) + \beta_4 x(0) + \beta_5 x(1)$$
$$= (\beta_0 + \beta_3) + (\beta_1 + \beta_5)x$$

As illustrated in Figure 3.12(a), if we use this model, then the straight lines relating mean sales volume to x for the street, mall, and downtown locations have different y-intercepts and different slopes. The different slopes imply that this model assumes **interaction** between x and store location. Specifically, note that the differently sloped lines in Figure 3.12(a) move closer together as x increases. This implies that the differences between the mean sales volumes in the street, mall, and downtown locations get smaller as the number of households in a store's area increases. Of course, the opposite type of interaction, in which differently sloped lines move farther apart as x increases, is also possible. This type of interaction would imply that the differences between the mean sales volumes in the street, mall, and downtown locations get larger as the number of households in a store's area increases. Figure 3.12(b) gives a partial SAS output of a regression analysis of the sales volume data using the *interaction model*, which is also called the *unequal slopes model*. Note that D_M, D_D, xD_M, and xD_D are labeled as DM, DD, xDM, and xDD, respectively, on the output. The



(b)

Root MSE	6.79953	R-Square	0.9877
Dependent Mean	176.98933	Adj R-Sq	0.9808
Coeff Var	3.841777		

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	7.90042	17.03513	0.46	0.6538
x	0.92070	0.12343	7.46	<.0001
DM	42.72974	21.50420	1.99	0.0782
DD	10.25503	21.28319	0.48	0.6414
XDM	-0.09172	0.14163	-0.65	0.5334
XDD	-0.03363	0.13819	-0.24	0.8132

Figure 3.12 Regression analysis of the sales volume data using the model $y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \beta_4 x D_M + \beta_5 x D_D + \varepsilon$ (a) Geometrical interpretation of the model (b) Partial SAS output

SAS output tells us that the *p-value*s related to the significance of xD_M and xD_D are large—.5334 and .8132, respectively. Therefore, these interaction terms do not seem to be important. In addition, the SAS output tells as that the standard error s for the interaction model is $s = 6.79953$, which is larger than the s of 6.34941 for the *no-interaction model* $y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \varepsilon$ (see Figure 3.11). It follows that the no-interaction model, which is sometimes called the *parallel slopes model*, seems to be the better model describing the sales volume data. Recall that this no-interaction model implies that $\mu_{h,M} - \mu_{h,S} = \beta_2$, $\mu_{h,D} - \mu_{h,S} = \beta_3$, and $\mu_{h,M} - \mu_{h,D} = \beta_2 - \beta_3$. That is, the no-interaction model implies that the differences between the mean sales volumes in the street, mall, and downtown locations do not depend upon the value h of x, the number of households in the area. Therefore, the previous and future statistical inferences for these differences made by using the no-interaction model are valid.

In general, if we wish to model the effect of a qualitative independent variable having a levels, we use $a-1$ dummy variables. Consider the kth such dummy variable D_k (k = one of the values 1, 2, ..., $a-1$). The parameter β_k multiplying D_k represents the mean difference between the level of y when the qualitative variable assumes level k and when it assumes the level a (where the level a is the level which we do not use a dummy variable to represent). For example, if we wish to use a confidence interval and a hypothesis test to compare the mall and downtown locations in the Electronics World example, we can use the model $y = \beta_0 + \beta_1 x + \beta_2 D_S + \beta_3 D_M + \varepsilon$. Here the dummy variable D_M is as previously defined, and D_S is a dummy variable that equals 1 if a store is in a street location and 0 otherwise. Because this model does not use a dummy variable to represent the downtown location, the parameter β_2 expresses the effect on mean sales of a street location compared to a downtown location, and the parameter β_3 expresses the effect on mean sales of a mall location compared to a downtown location. That is $\beta_2 = \mu_{h,S} - \mu_{h,D}$ and $\beta_3 = \mu_{h,M} - \mu_{h,D}$. The Excel output tells us that the least squares point estimate of β_3 is 21.51 and that the standard error of this estimate is 4.0651. It follows that a 95 percent confidence interval for $\mu_{h,M} - \mu_{h,D}$ is

$$[21.51 \pm 2.201(4.0651)] = [\ 12.563, 30.457]$$

This says we are 95 percent confident that for any given number of households in a store's area, the mean monthly sales volume in a mall location is between \$12,563 and \$30,457 greater than the mean monthly sales volume in a downtown location. The Excel output also shows that the t-statistic and p-value for testing the significance of $\mu_{h,M} - \mu_{h,D}$ are, respectively, 5.29 and 0.000256. Therefore, we have very strong evidence

	Coefficients	Standard Error	t Stat	P-value
Intercept	21.84147001	8.55847513	2.552028216	0.026897774
x	0.868588415	0.040489928	21.45196249	2.51663E-10
DS	-6.863776795	4.770476502	-1.438803187	0.178046589
DM	21.50997928	4.065091975	5.291388094	0.00025577

Figure 3.13 Partial Excel output for the model $y = \beta_0 + \beta_1 x + \beta_2 D_S + \beta_3 D_M + \varepsilon$

that there is a difference between the mean monthly sales volumes in mall and downtown locations.

3.3 The Partial F-Test

We now present a *partial F-test* that allows us to test the significance of a set of independent variables in a regression model. That is, we can use this *F*-test to test the significance of a *portion* of a regression model. For example, recall that in the previous section we decided that the no-interaction (or paralled slopes) model

$$y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \varepsilon$$

describes the sales volume data better than does the interaction (or unequal slopes) model

$$y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \beta_4 x D_M + \beta_5 x D_D + \varepsilon$$

The reasons for this decision were that the no-interaction model has the smaller standard error *s* and the *p*-values related to the significance of xD_M and xD_D in the interaction model are large—.5334 and .8132—indicating that these interaction terms are not important. Another way to decide which of these models is best is to test the significance of the *interaction portion* of the interaction model. We do this by testing the null hypothesis

$$H_0 : \beta_4 = \beta_5 = 0$$

which says that neither of the interaction terms significantly affects sales volume, versus the alternative hypothesis

$$H_a: \text{At least one of } \beta_4 \text{ and } \beta_5 \text{ does not equal } 0$$

which says that at least one of the interaction terms significantly affects sales volume.

In general, consider the regression model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_g x_g + \beta_{g+1} x_{g+1} + \ldots + \beta_k x_k + \varepsilon$$

Suppose we wish to test the null hypothesis

$$H_0: \beta_{g+1} = \beta_{g+2} = \ldots = \beta_k = 0$$

which says that *none of the independent variables* $x_{g+1}, x_{g+2}, \ldots, x_k$ *affects* y, versus the alternative hypothesis

$$H_a: \text{At least one of } \beta_{g+1}, \beta_{g+2}, \ldots, \beta_k \text{ does not equal } 0$$

which says that *at least one of the independent variables* $x_{g+1}, x_{g+2}, \ldots, x_k$ *affects* y. If we can reject H_0 in favor of H_a by specifying a *small* probability of a Type I error, then it is reasonable to conclude that at least one of $x_{g+1}, x_{g+2}, \ldots, x_k$ *significantly* affects y. In this case we should use t-statistics and other techniques to determine which of $x_{g+1}, x_{g+2}, \ldots, x_k$ significantly affects y. To test H_0 versus H_a, consider the following two models:
Complete model: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_g x_g + \beta_{g+1} x_{g+1} + \ldots + \beta_k x_k + \varepsilon$
Reduced model: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_g x_g + \varepsilon$
Here the complete model is assumed to have k independent variables, the reduced model is the complete model under the assumption that H_0 is true, and $(k - g)$ denotes the number of regression parameters we have set equal to 0 in the statement of H_0.

To carry out this test, we calculate SSE_C, *the unexplained variation for the complete model, and* SSE_R, *the unexplained variation for the reduced model*. The appropriate test statistic is based on the difference

$$SSE_R - SSE_C$$

which is called *the drop in the unexplained variation attributable to the independent variables* $x_{g+1}, x_{g+2}, \ldots, x_k$. In the following box we give the formula for the test statistic and show how to carry out the test. (The valdity of the test is proven in section B.8.)

The partial F-test: An F-test for a portion of a regression model

Suppose that the regression assumptions hold and consider testing

$$H_0 : \beta_{g+1} = \beta_{g+2} = \ldots = \beta_k = 0$$

versus

$$H_a : \text{ At least one of } \beta_{g+1}, \beta_{g+2}, \ldots, \beta_k \text{ does not equal } 0$$

We define the *partial F-statistic* to be

$$F = \frac{(SSE_R - SSE_C) / (k - g)}{SSE_C \; / [n - (k + 1)]}$$

Also define the *p-value* related to F to be the area under the curve of the F distribution [having $k - g$ and $n - (k + 1)$ degrees of freedom] to the right of F. Then, we can reject H_0 in favor of H_a at level of significance α if either of the following equivalent conditions holds:

1. $F > F_{[\alpha]}$
2. p-value < α

Here the rejection point $F_{[\alpha]}$ is based on $k - g$ numerator and $n - (k + 1)$ denominator degrees of freedom.

It can be shown that the "extra" independent variables $x_{g+1}, x_{g+2}, \ldots, x_k$ will always explain some of the variation in the observed y values and, therefore, will always make SSE_C somewhat smaller than SSE_R. Condition 1 says that we should reject H_0 if

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C / [n-(k+1)]}$$

is large. This is reasonable because a large value of F would result from a large value of $SSE_R - SSE_C$, which would be obtained if at least one of the independent variables $x_{g+1}, x_{g+2}, \ldots, x_k$ makes SSE_C substantially smaller than SSE_R. This would suggest that H_0 is false and that H_a is true.

Before looking at an example, we should point out that testing the significance of a single independent variable by using a partial F-test is equivalent to carrying out this test by using the previously discussed t-test. It can be shown that when we test $H_0: \beta_j = 0$ versus $H_a: \beta_j \neq 0$ using a partial F-test

$$F = t^2 \text{ and } F_{[\alpha]} = (t_{[\alpha/2]})^2$$

Here $F_{[\alpha]}$ is based on 1 numerator and $n-(k+1)$ denominator degrees of freedom and $t_{[\alpha/2]}$ is based on $n-(k+1)$ degrees of freedom. Hence, the rejection conditions

$$|t| > t_{[\alpha/2]} \text{ and } F > F_{[\alpha]}$$

are equivalent. It can also be shown that in this case the *p-value* related to t equals the *p-value* related to F.

Example 3.4

In order to test $H_0: \beta_4 = \beta_5 = 0$ in the Electronics World interaction model, we regard this model as the complete model:

Complete Model: $y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \beta_4 x D_M + \beta_5 x D_D + \varepsilon$

Although the partial SAS output in Figure 3.12 (b) does not show the unexplained variation for this complete model, SAS can be used to show that this unexplained variation is 416.1027. That is, $SSE_C = 416.1027$. If the null hypothesis $H_0: \beta_4 = \beta_5 = 0$ is true, the complete model becomes the following reduced model:

Reduced Model: $y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \varepsilon$

which is the no-interaction (parallel slopes) model and has an unexplained variation of 443.4650. That is, $SSE_R = 443.4650$. There are $n = 15$ observations in the Electronics World data set (see Table 3.4), and the complete model uses $k = 5$ independent variables. In addition, because two parameters $(\beta_4$ and $\beta_5)$ are set equal to 0 in the statement of $H_0: \beta_4 = \beta_5 = 0$, we have that $k - g = 2$. Therefore:

$$
\begin{aligned}
F &= \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C / [n-(k+1)]} \\
&\frac{(443.4650 - 416.1027)/2}{416.1027/(15-6)} \\
&= .2959
\end{aligned}
$$

If we wish to set α equal to .05, we compare $F = .2959$ with $F_{[.05]} = 4.26$, which is based on $k - g = 2$ numerator and $n - (k + 1) = 15 - 6 = 9$ denominator degrees of freedom. Since $F = .2959$ is less than $F_{[.05]} = 4.26$, we cannot reject $H_0: \beta_4 = \beta_5 = 0$ at the .05 level of significance, and thus we do not have strong evidence that at least one of the interaction terms significantly affects sales volume. This is further evidence that the no-interaction model is the better model. Also, recalling that the no-interaction model is sometimes called the parallel slopes model, the partial F-test just performed is sometimes called a *test for parallel slopes*.

In Example 3.3 we used the no-interaction model

$$y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \varepsilon$$

to make pairwise comparisons of the street, mall, and downtown store locations by carrying out a t-test for each of the parameters β_2, β_3, and $\beta_2 - \beta_3$. There is a theoretical problem with this because, although we can set the probability of a Type I error equal to .05 for each individual test, it is possible to show that the probability of falsely rejecting H_0 in *at least* one of these tests is greater than .05. Because of this problem, many statisticians feel that before making pairwise comparisons we should test for differences between the effects of the locations by testing the single hypothesis

$$H_0: \mu_{h,S} = \mu_{h,M} = \mu_{h,D}$$

which says that the street, mall, and downtown locations have the same effects on mean sales volume (no differences between locations).

To carry out this test we consider the following:

$$\textit{Complete model}: \ y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \varepsilon$$

In Example 3.3 we saw that for this model

$$\beta_2 = \mu_{h,M} - \mu_{h,S} \text{ and } \beta_3 = \mu_{h,D} - \mu_{h,S}$$

It follows that the null hypothesis $H_0 : \mu_{h,S} = \mu_{h,M} = \mu_{h,D}$ is equivalent to $H_0 : \beta_2 = \beta_3 = 0$ and that the alternative hypothesis

$$H_a : \text{At least two of } \mu_{h,S}, \mu_{h,M}, \text{ and } \mu_{h,D} \text{ differ}$$

which says that at least two locations have different effects on mean sales volume, is equivalent to

$$H_a : \text{At least one of } \beta_2 \text{ and } \beta_3 \text{ does not equal } 0$$

Because of these equivalencies, we can test H_0 versus H_a by using a partial F-test. For the just given complete model (which has $k = 3$ independent variables), we obtain an unexplained variation equal to $SSE_C = 443.4650$. The reduced model is the complete model when H_0 is true. Therefore, we obtain

$$\textit{Reduced model}: \ y = \beta_0 + \beta_1 x + \varepsilon$$

For this model the unexplained variation is $SSE_R = 2467.8067$. Noting that two parameters $(\beta_2 \text{ and } \beta_3)$ are set equal to 0 in the statement of $H_0 : \beta_2 = \beta_3 = 0$, we have $k - g = 2$. Therefore, the needed partial F-statistic is

$$\begin{aligned} F &= \frac{(SSE_R - SSE_C) / (k - g)}{SSE_C / [n - (k + 1)]} \\ &= \frac{(2467.8067 - 443.4650) / 2}{443.4650 / [15 - 4]} \\ &= 25.1066 \end{aligned}$$

If we wish to set α equal to .05, we compare $F = 25.1066$ with $F_{[.05]} = 3.98$, which is based on $k - g = 2$ numerator and $n - (k+1) = 15 - 4 = 11$ denominator degrees of freedom. Since $F = 25.1066$ is greater than $F_{[.05]} = 3.98$, we can reject H_0 at the .05 level of significance, and we have very strong statistical evidence that at least two locations have different effects on mean sales volume. Having reached this conclusion, it makes sense to compare the effects of specific pairs of locations. We have already done this in Example 3.3. It should also be noted that even if H_0 were not rejected, some practitioners feel that pairwise comparisons should still be made. This is because there is always a possibility that we have erroneously decided to not reject H_0.

We next consider two statistics that provide descriptive information that supplements the information provided by a partial F-test.

Partial Coefficients of Determination and Correlation

1. The *partial coefficient of determination* is

$$R^2(x_{g+1},\ldots,x_k \mid x_1,\ldots,x_g) = \frac{SSE_R - SSE_C}{SSE_R}$$

= the proportion of the unexplained variation in the reduced model that is explained by the extra independent variables in the complete model

2. The *partial coefficient of correlation* is

$$R(x_{g+1},\ldots,x_k \mid x_1,\ldots,x_g) = \sqrt{R^2(x_{g+1},\ldots,x_k \mid x_1,\ldots,x_g)}$$

For example, consider the Electronics World situation. If we consider the complete model to be the model $y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \varepsilon$ and the reduced model to be the model $y = \beta_0 + \beta_1 x + \varepsilon$, then we have seen that $SSE_C = 443.4640$ and $SSE_R = 2467.8067$. It follows that

$$
\begin{aligned}
R^2(D_M, D_D \mid x) &= \frac{SSE_R - SSE_C}{SSE_R} \\
&= \frac{2467.8067 - 443.4650}{2467.8067} \\
&= .8206
\end{aligned}
$$

That is, D_M and D_D in the complete model explain 82.06 percent of the unexplained variation in the reduced model. Also, $R(D_M, D_D \mid x) = \sqrt{.8206} = .9059$

3.4 Statistical Inference for a Linear combination of Regression parameters

Consider the Electronics World dummy variable model

$$y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \varepsilon$$

In Example 3.3 we have seen that $\beta_2 - \beta_3$ is the difference between the mean monthly sales volumes in mall and downtown locations. In order to make statistical inferences about $\beta_2 - \beta_3$, we express this difference as a linear combination of the parameters β_0, β_1, β_2, and β_3 in the dummy variable model. Specifically, letting l denote the linear combination, we write

$$l = \beta_2 - \beta_3 = (0)\beta_0 + (0)\beta_1 + (1)\beta_2 + (-1)\beta_3$$

In general, let

$$l = \lambda_0\beta_0 + \lambda_1\beta_1 + \lambda_2\beta_2 + \ldots + \lambda_k\beta_k$$

be a linear combination of regression parameters. A point estimate of l is

$$\hat{l} = \lambda_0 b_0 + \lambda_1 b_1 + \lambda_2 b_2 + \ldots + \lambda_k b_k$$

If the regression assumptions are satisfied, it can be shown (see Section B.9) that the population of all possible values of $\hat{l}$ is normally distributed with mean l and standard deviation

$$\sigma_{\hat{l}} = \sigma\sqrt{\lambda'(\mathbf{X'X})^{-1}\lambda}$$

Here $\lambda' = \left[\lambda_0 \lambda_1 \lambda_2 \ldots \lambda_k\right]$ is a row vector containing the numbers multiplied by the β's in the equation for l. Since we estimate σ by s, it follows that

$$s_{\hat{l}} = s\sqrt{\lambda'\left(\mathbf{X'X}\right)^{-1}\lambda}$$

We use $s_{\hat{l}}$ to calculate the t-statistic for testing $H_0 : l = 0$ and to calculate confidence intervals for l.

The t-statistic for testing $H_0 : l = 0$ versus $H_a : l \neq 0$ is

$$t = \frac{\hat{l}}{s_{\hat{l}}} = \frac{\hat{l}}{s\sqrt{\lambda'\left(\mathbf{X'X}\right)^{-1}\lambda}}$$

A $100(1-\alpha)\%$ confidence interval for l is

$$\left[\hat{l} \pm t_{[\alpha/2]} s_{\hat{l}}\right] = \left[\hat{l} \pm t_{[\alpha/2]} s\sqrt{\lambda'\left(\mathbf{X'X}\right)^{-1}\lambda}\right]$$

Example 3.5

Consider the Electronics World dummy variable model

$$y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \varepsilon$$

Since we have seen in Example 3.3 that the least squares point estimates of β_2 and β_3 are b_2 = 28.374 and b_3 = 6.864, the point estimate of $l = \beta_2 - \beta_3$ is

$$\hat{l} = b_2 - b_3 = 28.374 - 6.864 = 21.51$$

Noting that

$$l = \beta_2 - \beta_3 = (0)\beta_0 + (0)\beta_1 + (1)\beta_2 + (-1)\beta_3$$

it follows that

$$\lambda' = \begin{bmatrix} 0 & 0 & 1 & -1 \end{bmatrix}$$

and

$$\lambda = \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix}$$

Using λ' and λ, $\lambda'(\mathbf{X'X})^{-1}\lambda$ can be computed to be .409898. Therefore, since $s = 6.34941$ and $n-(k+1) = 15-4 = 11$, a 95 percent confidence interval for $l = \beta_2 - \beta_3$ is

$$\begin{aligned}\left[\hat{l} \pm t_{[.025]} s\sqrt{\lambda'(\mathbf{X'X})^{-1}\lambda}\right] &= \left[21.51 \pm 2.201(6.34941)\sqrt{.409898}\right] \\ &= \left[21.51 \pm 2.201(4.0651)\right] \\ &= [12.5627, 30.4573]\end{aligned}$$

This says that we are 95 percent confident that for any given number of households in a store's area the mean monthly sales volume in a mall location is between \$12,563 and \$30,457 greater than the mean monthly sales volume in a downtown location.

We next point out that almost all of the SAS regression outputs we have looked at to this point were obtained by using a SAS procedure called PROC REG. This procedure will not carry out statistical inference for linear combinations of regression parameters (such as $\beta_2 - \beta_3$). However, another SAS procedure called PROC GLM (GLM stands for "General Linear Model") will do this. Figure 3.14 gives a partial

Parameter	Estimate	T for HO: Parameter=0	Pr > |T|	Std Error of Estimate
MUMALL - MUSTR	28.37375607	6.36	0.0001	4.46130660
MUDOWNTN - MUSTR	6.86377679	1.44	0.1780	4.77047650
MUMALL - MUDOWNTN	21.50997928	5.29	0.0003	4.06509197

Figure 3.14 Partial SAS PROC GLM output for the model $y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \varepsilon$

PROC GLM output of a regression analysis of the sales volume data using the previously given dummy variable model. On the output, the parameters β_2, β_3, and $\beta_2 - \beta_3$ are labeled as MUMALL—MUSTR, MUDOWNTN—MUSTR, and MUMALL—MUDOWNTN. Notice that the point estimates, standard errors, t statistics, and p-values we have used to analyze β_2 and β_3 are given on the output corresponding to MUMALL—MUSTR and MUDOWNTN—MUSTR. The point estimate, standard error of the estimate, t statistic, and p-value for analyzing $\beta_2 - \beta_3$ are given on the output corresponding to MUMALL—MUDOWNTN. Here, as calculated previously, the point estimate of $\beta_2 - \beta_3$ is $b_2 - b_3 = 21.51$ and the standard error of this estimate is 4.0651. This allows us to calculate the 95 percent confidence interval for $\beta_2 - \beta_3$ as $[21.51 \pm 2.201(4.0651)] = [12.5627, 30.4573]$. The SAS output also tells us that the t statistic and p-value for testing the significance of the linear combination $\beta_2 - \beta_3$ are, respectively, $t = 21.51 / 4.0651 = 5.29$ and p-value $= .0003$. Therefore, we have very strong evidence that there is a difference between the mean monthly sales volumes in mall and downtown locations. In summary, the mall location seems superior to both street and downtown locations. Of course, this conclusion (and other interpretations in this situation) assumes that the regression relationships between y and x and the store locations apply to future months, and other stores. Thus we assume that there are no trends, seasonal, or other time-related influences affecting store sales volume.

3.5 Simultaneous Confidence Intervals

Each of the confidence and prediction intervals we have studied uses the t point $t_{[\alpha/2]}$ and is based on *individual* $100(1-\alpha)$ percent confidence.

The *Bonferroni procedure* tells us that if we wish to calculate g confidence and/or prediction intervals such that we are $100(1-\alpha)$ percent confident that all g intervals simultaneously meet their objectives (that is, contain the parameters that they are supposed to contain—in the case of confidence intervals—or are such that the future y value of interest falls in the interval—in the case of prediction intervals), we should calculate each interval based on individual $100(1-\alpha/g)$ percent confidence. (This result is proven in Section B.10.)

For example, using the Electronics World model $y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \varepsilon$, which has $k = 3$ independent variables and is fit to the $n = 15$ store location observations, we have previously calculated confidence intervals for $\mu_{h,M} - \mu_{h,S} = \beta_2$, $\mu_{h,D} - \mu_{h,S} = \beta_3$, and $\mu_{h,M} - \mu_{h,D} = = \beta_2 - \beta_3$ based on individual 95 percent confidence and using $t_{[\alpha/2]} = t_{[.025]} = 2.201$ [based on $n-(k+1) = 15-(3+1) = 11$ degrees of freedom]. If we wish to be 95 percent confident that all $g = 3$ confidence intervals simultaneously contain the parameters they are attempting to estimate, we should base each interval on individual $100(1-\alpha/g)\% = 100(1-.05/3)\% = 100(.983333)\% = 98.3333\%$ confidence, and thus use $t_{[\alpha/2g]} = t_{[.05/6]} = t_{[.0083333]}$. We would have to find $t_{[.0083333]}$ using a computer. Using the Excel look up menu, we find that $t_{[.0083333]} = 2.82004$. Since this t point is larger than $t_{[.025]} = 2.201$, the Bonferroni simultaneous 95 percent confidence intervals are wider than the individual 95 percent confidence intervals. Figure 3.14 tells us that the point estimates of $\mu_{h,M} - \mu_{h,S} = \beta_2, \mu_{h,D} - \mu_{h,S} = \beta_3$, and $\mu_{h,M} - \mu_{h,D} = \beta_2 - \beta_3$ are respectively, $b_2 = 28.374$, $b_3 = 6.864$, and $\hat{l} = b_2 - b_3 = 21.51$. This figure also tells us that the standard errors of these point estimates are $s_{b_2} = 4.461$, $s_{b_3} = 4.770$, and $s_{\hat{l}} = 4.065$. If follows that Bonferroni simultaneous 95 percent confidence intervals for $\mu_{h,M} - \mu_{h,S} = \beta_2$, $\mu_{h,D} - \mu_{h,S} = \beta_3$, and $\mu_{h,M} - \mu_{h,D} = \beta_2 - \beta_3$ are:

$$[28.374 \pm 2.82004(4.461)] = [15.794, 40.954]$$
$$[6.864 \pm 2.82004(4.770)] = [-6.588, 20.316]$$

and

$$[21.51 \pm 2.82004(4.065)] = [10.046, 32.974]$$

These simultaneous 95 percent confidence intervals are wider than the previously calculated individual 95 percent confidence intervals, which were, respectively $[18.554, 38.193]$, $[-3.636, 17.363]$ and $[12.563, 30.457]$. However, the first and third simultaneous 95 percent confidence intervals still consist of all positive numbers and those make us *simultaneously* 95 percent confident that $\mu_{h,M}$ is greater than $\mu_{h,S}$ and is greater than $\mu_{h,D}$. More specifically, the lower ends of the first and third simultaneous 95 percent confidence intervals make us *simultaneously* 95 percent confident that for any given number of households in a store's area the mean monthly sales volume in a mall location is at least \$15,794 more than the mean monthly sales volume in a street location *and* is at least \$10,046 more than the monthly sales volume in a downtown location.

3.6 Logistic Regression

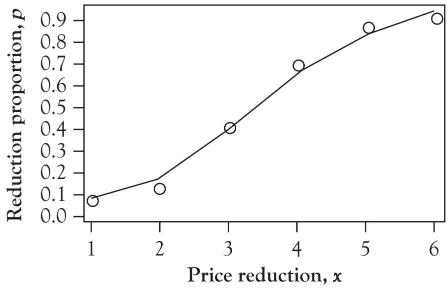
Suppose that in a study of the effectiveness of offering a price reduction on a given product, 300 households having similar incomes were selected. A coupon offering a price reduction, x, on the product, as well as advertising material for the product, was sent to each household. The coupons offered different price reductions (10, 20, 30, 40, 50, and 60 dollars), and 50 homes were assigned at random to each price reduction. Table 3.5 summarizes the number, y, and proportion, $\hat{p}$, of households redeeming coupons for each price reduction, x (expressed in units of \$10). In the middle of the left side of Table 3.5, we plot the $\hat{p}$ values versus the x values and draw a hypothetical curve through the plotted points. A theoretical curve having the shape of the curve in Table 3.5 is the *logistic curve*

$$p(x) = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}}$$

where $p(x)$ denotes the probability that a household receiving a coupon having a price reduction of x will redeem the coupon. The MINITAB output at the bottom of Table 3.5 tells us that the point estimates of β_0

Table 3.5 The price reduction data and logistic regression

x	1	2	3	4	5	6
y	4	7	20	35	44	46
$\hat{p}$	.08	.14	.40	.70	.88	.92



Price reduction, x	Probability Estimate
1	0.066943
2	0.178920
3	0.398256
4	0.667791
5	0.859260
6	0.948831

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P
Constant	-3.7456	0.434355	-8.62	0.000
x	1.1109	0.119364	9.31	0.000

and β_1 are $b_0 = -3.7456$ and $b_1 = 1.1109$. (Estimation in logistic regression is usually done by *maximum likehood estimation*. This technique and extensions of logistic regression are discussed in Appendix C.) Using these estimates, it follows that, for example,

$$\hat{p}(5) = \frac{e^{(-3.7456+1.1109(5))}}{1+e^{(-3.7456+1.1109(5))}} = \frac{6.1037}{1+6.1037} = .8593$$

That is, $\hat{p}(5) = .8593$ is the point estimate of the probability that a household receiving a coupon having a price reduction of \$50 will redeem the coupon. The middle of the right side of Table 3.5 gives the values of $\hat{p}(x)$ for $x =$ 1, 2, 3, 4, 5, and 6.

The *general logistic regression model* relates the probability that an event (such as redeeming a coupon) will occur to k independent variables $x_1, x_2, \ldots, x_k$. This general model is

$$p(x_1, x_2, \ldots, x_k) = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k)}}$$

where $p(x_1, x_2, \ldots, x_k)$ is the probability that the event will occur when the values of the independent variables are $x_1, x_2, \ldots, x_k$. In order to estimate $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$, we obtain n observations, with each observation consisting of observed values of $x_1, x_2, \ldots, x_k$, and of a dependent variable y. Here, y is a *dummy variable* that equals 1 if the event has occurred and 0 otherwise.

For example, suppose that the personnel director of a firm has developed two tests to help determine whether potential employees would perform successfully in a particular position. To help estimate the usefulness of the tests, the director gives both tests to 43 employees that currently hold the position. If an employee is performing successfully, we set the dependent variable *Group* equal to l; if the employee is performing unsuccessfully, we set *Group* equal to 0. Let x_1 and x_2 denote the scores of an employee on tests l and 2, and let $p(x_1, x_2)$ denote the probability that the employee having the scores x_1 and x_2 will perform successfully in the position. We can estimate the relationship between $p(x_1, x_2)$ and x_1 and x_2 by using the logistic regression model

$$p(x_1, x_2) = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$

Of the 43 employees tested by the personnel director, 23 are performing successfully and 20 are performing unsuccessfully in the particular position. Each of the 23 successfully performing employees is assigned a *Group* value of 1, and the combinations of scores on tests 1 and 2 for the 23 successfully performing employees are (96, 85), (96, 88), (91, 81), (95, 78), (92, 85), (93, 87), (98, 84), (92, 82), (97, 89), (95, 96), (99, 93), (89, 90), (94, 90), (92, 94), (94, 84), (90, 92), (91, 70), (90, 81), (86, 81), (90, 76), (91, 79), (88, 83), and (87, 82). Each of the 20 unsuccessfully performing employees is assigned a *Group* value of 0, and the combinations of scores on tests 1 and 2 for the 20 unsuccessfully performing employees are (93, 74), (90, 84), (91, 81), (91, 78), (88, 78), (86, 86), (79, 81), (83, 84),

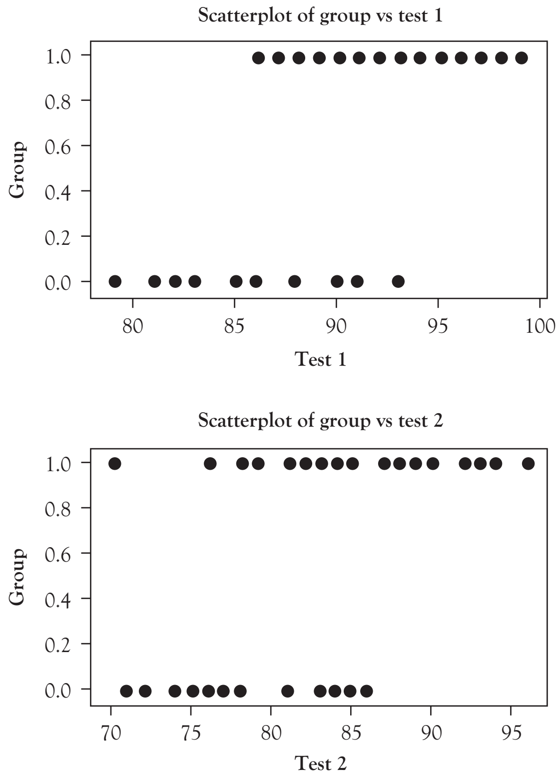


Figure 3.15 Scatterplots of group versus x_1 and group versus x_2

(79, 77), (88, 75), (81, 85), (85, 83), (82, 72), (82, 81), (81, 77), (86, 76), (81, 84), (85, 78), (83, 77), and (81, 71). The source of the data for this example is Dielman (1996), and Figure 3.15 shows scatterplots of Group versus x_1 (the score on test 1) and Group versus x_2(the score on test 2).

The MINITAB output in Figure 3.16 tells us that the point estimates of β_0, β_1, and β_2 are $b_0 = -56.17$, $b_1 = .4833$, and $b_2 = .1652$. Consider, therefore, a potential employee who scores a 93 on test 1 and an 84 on test 2. It follows that a point estimate of the probability that the potential employee will perform successfully in that position is

$$\hat{p}(93,84) = \frac{e^{(-56.17+.4833(93)+.1652(84))}}{1+e^{(-56.17+.4833(93)+.1652(84))}} = \frac{14.206506}{15.206506} = .9342$$

```
                                                      Odds     95% CI
Predictor       Coef    SE Coef       Z       P     Ratio   LOwer  Upper
Constant    -56.1704    17.4516   -3.22   0.001
Test 1      0.483314   0.157779    3.06   0.002      1.62    1.19   2.21
Test 2      0.165218   0.102070    1.62   0.106      1.18    0.97   1.44

Log-Likelihood = -13.959
Test that all slopes are zero: G = 31.483, DF = 2, p-value = 0.000
```

Figure 3.16 MINITAB output of logistic regression of the performance data

To further analyze the logistic regression output, we consider several hypothesis tests that are based on the *chi-square distribution*.[1] We first consider testing $H_0 : \beta_1 = \beta_2 = 0$ versus H_a : At least one of β_1 or β_2 does not equal 0. The *p*-value for this test is the area under the chi-square curve having $k = 2$ degrees of freedom to the right of the test statistic value $G = 31.483$. Although the calculation of G is too complicated to demonstrate in this book, the MINITAB output gives the value of G and the related *p*-value, which is less than .001. This *p-value* implies that we have extremely strong evidence that at least one of β_1 or β_2 does not equal zero. The *p*-value for testing $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ is the area under the chi-square curve having one degree of freedom to the right of the square of $z = (b_1 / s_{b_1}) = (.4833 / .1578) = 3.06$. The MINITAB output tells us that this *p*-value is .002, which implies that we have very strong evidence that the score on test 1 is related to the probability of a potential employee's success. The *p*-value for testing $H_0 : \beta_2 = 0$ versus $H_a : \beta_2 \neq 0$ is the area under the chi-square curve having one degree of freedom to the right of the square of $z = (b_2 / s_{b_2}) = (.1652 / .1021) = 1.62$. The MINITAB output tells us that this *p*-value is .106, which implies that we do not have strong evidence that the score on test 2 is related to the probability of a potential employee's success. In the exercises we will consider a logistic regression model that uses only the score on test 1 to estimate the probability of a potential employee's success.

The *odds* of success for a potential employee is defined to be the probability of success divided by the probability of failure for the employee. That is,

[1] Like the curve of the *F*-distribution, the curve of the chi-square distribution is skewed with a tail to the right. The exact shape of a chi-square distribution curve is determined by the (single) number of degrees of freedom associated with the chi-square distribution under consideration.

$$\text{odds} = \frac{p(x_1, x_2)}{1 - p(x_1, x_2)}$$

For the potential employee who scores a 93 on test 1 and an 84 on test 2, we estimate that the odds of success are $.9342 / (1 - .9342) = 14.2$. That is, we estimate that the odds of success for the potential employee are about 14 to 1. It can be shown that $e^{b_1} = e^{.4833} = 1.62$ is a point estimate of the *odds ratio for* x_1, which is the proportional change in the odds (for any potential employee) that is associated with an increase of one in x_1 when x_2 stays constant. This point estimate of the odds ratio for x_1 is shown on the MINITAB output and says that, for every one point increase in the score on test 1 when the score on test 2 stays constant, we estimate that a potential employee's odds of success increases by 62 percent. Furthermore, the 95 percent confidence interval for the odds ratio for x_1, [1.19, 2.21], does not contain 1. Therefore, as with the (equivalent) chi-square test of $H_0 : \beta_1 = 0$, we conclude that there is strong evidence that the score on test 1 is related to the probability of success for a potential employee. Similarly, it can be shown that $e^{b_2} = e^{.1652} = 1.18$ is a point estimate of the *odds ratio for* x_2, which is the proportional change in the odds (for any potential employee) that is associated with an increase of one in x_2 when x_1 stays constant. This point estimate of the odds ratio for x_2 is shown on the MINITAB output and says that, for every one point increase in the score on test 2 when the score on test l stays constant, we estimate that a potential employee's odds of success increases by 18 percent. However, the 95 percent confidence interval for the odds ratio for x_2—[.97, 1.44]—contains l. Therefore, as with the equivalent chi-square test of $H_0 : \beta_2 = 0$, we cannot conclude that there is strong evidence that the score on test 2 is related to the probability of success for a potential employee.

To better understand the odds ratio, consider the general logistic regression model

$$p(x_1, x_2, \ldots, x_k) = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}$$

where $p(x_1, x_2, \ldots, x_k)$ is the probability that the event under consideration will occur when the values of the independent variables are $x_1, x_2, \ldots, x_k$. The *odds* of the event occurring, which we will denote as $\text{odds}(x_1, x_2, \ldots, x_k)$, is defined to be $p(x_1, x_2, \ldots, x_k) / (1 - p(x_1, x_2, ..., x_k))$, which is the probability that the event will occur divided by the probability that the event will not occur. Now, $1 - p(x_1, x_2, \ldots, x_k)$ equals

$$
\begin{aligned}
&1 - \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}} \\
&= \frac{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)} - e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}} \\
&= \frac{1}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}
\end{aligned}
$$

Therefore, $\text{odds}(x_1, x_2, \ldots, x_k)$ equals

$$
\begin{aligned}
&\frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)} \Big/ \left[1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}\right]}{1 \Big/ \left[1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}\right]} \\
&= e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}
\end{aligned}
$$

If the jth independent variable x_j increases by 1 and the other independent variables remain constant, the odds ratio for x_j is $\text{odds}\left(x_1, x_2, \ldots, x_j + 1, \ldots, x_k\right) \Big/ \text{odds}\left(x_1, x_2, \ldots, x_j, \ldots, x_k\right)$, which equals

$$
\begin{aligned}
&\frac{e^{\left[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_j \left(x_j + 1\right) + \ldots + \beta_k x_k\right]}}{e^{\left[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_j x_j + \ldots + \beta_k x_k\right]}} \\
&= \frac{e^{\left[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{j-1} x_{j-1} + \beta_{j+1} x_{j+1} + \ldots + \beta_k x_k\right]} e^{\beta_j \left(x_j + 1\right)}}{e^{\left[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{j-1} x_{j-1} + \beta_{j+1} x_{j+1} + \ldots + \beta_k x_k\right]} e^{\beta_j x_j}} \\
&= \left[e^{\beta_j (x_j + 1)}\right] \left[e^{-\beta_j x_j}\right] \\
&= e^{\beta_j \left(x_j + 1\right) - \beta_j x_j} \\
&= e^{\beta_j}
\end{aligned}
$$

This says that e^{b_j} is the point estimate of the *odds ratio for* x_j, which is the proportional change in the odds that is associated with a one unit increase in x_j when the other independent variables stay constant. Also, note that the natural logarithm of the odds is $(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$, which is called the *logit*. If $b_0, b_1, b_2, \ldots, b_k$ are the point estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$, the point estimate of the logit, denoted by $\widehat{lg}$, is $(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k)$. It follows that the point estimate of the probability that the event will occur is

$$\hat{p}(x_1, x_2, \ldots, x_k) = \frac{e^{\widehat{lg}}}{1 + e^{\widehat{lg}}} = \frac{e^{(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k)}}{1 + e^{(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k)}}$$

To conclude this section, note that logistic regression can be used to find a confidence interval for $p(x_1, x_2, \ldots, x_k)$, the probability than an event will occur. For example, in the employee performance example, consider an employee who scores a 93 on test 1 and an 84 on test 2. The SAS output of a logistic regression of the performance data is given in Figure 3.17. The "Wald Chi-Square" for a variable on this output equals the [(Parameter Estimate)/(Standard Error)]^2. The output tells us that a point estimate of and a 95 percent confidence interval for the probability that the employee will perform successfully in the particular position are, respectively, .93472 and [.69951, .98877]. That is, our best single estimate of the probability that the employee will perform successfully is .93472. Moreover, we are 95 percent confident that the probability that the employee will perform successfully is between .69951 and .98877.

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Odds Ratio
INTERCPT	-56.2601	17.4495	10.3952	0.0013	.
TEST1	0.4842	0.1576	9.4438	0.0021	1.62
TEST2	0.1653	0.1023	2.6136	0.1060	1.18

OBS	Group	TEST 1	TEST 2	PREDICT	CLLOWER	CLUPPER
44	.	93	84	0.93472	0.69951	0.98877
45	.	85	82	0.17609	0.04489	0.49286

***Figure 3.17** SAS output of a logistic regression of the performance data*

3.7 Using SAS

In Exercises 3.3 through 3.9 we analyze the Fresh detergent demand data in Table 3.2 and Table 3.7 (on page 148) by using two models:

$$\text{Model 1: } y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 \mathrm{D_B} + \beta_6 \mathrm{D_C} + \varepsilon$$

$$\text{Model 2: } y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 \mathrm{D_B} + \beta_6 \mathrm{D_C} + + \beta_7 x_3 \mathrm{D_B} + \beta_8 x_3 \mathrm{D_C} + \varepsilon$$

Here, three advertising campaigns—A, B, and C—were used in the 30 sales periods. For example, Table 3.7 tells us that advertising campaign B was used in sales periods 1, 2, and 3; advertising campaign A was used in sales period 4; advertising campaign C was used in sales period 5; and advertising campaign C was used in sales period 30. Advertising campaign C will also be used in a future sales period. In the above model, $D_B = 1$ if advertising campaign B is used in a sales period and 0 otherwise; $D_C = 1$ if advertising campaign C is used in a sales period and 0 otherwise. Figure 3.18 presents the SAS program that gives the outputs used in Exercises 3.3 through 3.9.

```
DATA DETR;
INPUT Y X4 X3 DB DC;
X3SQ = X3*X3;
X43 = X4*X3;
X3DB = X3*DB;
X3DC = X3*DC;

DATALINES;
7.38  -0.05   5.50   1  0
8.51   0.25   6.75   1  0
9.52   0.60   7.25   1  0
7.50   0.00   5.50   0  0
9.33   0.25   7.00   0  1
              .
              .
              .
9.26   0.55   6.80   0  1
 .     0.20   6.50   0  1   }→ Future sales period

PROC REG;
MODEL Y =  X4  X3  X3SQ  X43  DB   DC/P  CLM   CLI;
T1: TEST DB=0, DC=0;}Performs partial F test of H0 : β5 = β6 = 0
```

Figure 3.18 SAS programs for fitting models 1 and 2 (Continued)

```
PROC GLM;
MODEL Y =  X4  X3  X3SQ  X43  DB   DC/P CLI;
ESTIMATE 'MUDAB-MUDAA' DB 1; }Estimates β5
ESTIMATE 'MUDAC-MUDAA' DC 1; }Estimates β6
ESTIMATE 'MUDAC-MUDAB' DB -1 DC 1; }Estimates β6–β5
PROC REG;
MODEL Y =  X4  X3  X3SQ  X43  DB  DC  X3DB  X3DC/P CLM CLI;
T2: TEST DB=0, DC=0, X3DB=0, X3DC=0;}→
                                   Tests H0 : β5 = β6 = β7 = β8 = 0
T3: TEST X3DB=0, X3DC=0; }→Tests H0 : β7 = β8 = 0
PROC GLM;
MODEL Y =  X4  X3  X3SQ  X43 DB DC X3DB X3DC/P  CLI;
ESTIMATE 'DIFF1' DC 1 X3DC 6.2; }→ Estimates β6 + β8 (6.2)
ESTIMATE 'DIFF2' DC 1 X3DC 6.6; }→ Estimates β6 + β8 (6.6)
ESTIMATE 'DIFF3' DC 1 DB -1 X3DC 6.2 X3DB -6.2;}→
                                   Estimates β6 – β5 + β8 (6.2) – β7 (6.2)
ESTIMATE 'DIFF4' DC 1 DB -1 X3DC 6.6 X3DB -6.6;}→
                                   Estimates β6 – β5 + β8 (6.6) – β7 (6.6)
```

Figure 3.18 SAS programs for fitting models 1 and 2

```
data;
input Group Test1 Test2;
datalines;
1 96 85
1 96 85
    .
    .
1 87 82
2 93 74
2 90 84
    .
    .
2 81 71
. 93 84
. 85 82
proc logistic;
model Group = Test1 Test2;
output out=results P=PREDICT L=CLLOWER U=CLUPPER;
proc print;
```

Note: The 0's (unsuccessful employees) must be a "higher number" than the 1's (successful employees) when using SAS. So we used 2's to represent the unsuccessful employees.

Figure 3.19 SAS program for performing logistic regression using the performance data

3.8 Exercises

Exercise 3.1

In the article "Integrating Judgment With a Regression Appraisal", published in *The Real Estate Appraiser and Analyst* (1986), R. L. Andrews and J. T. Ferguson present ten observations concerning y = sales price of a house (in thousands of dollars), x_1 = home size (in hundreds of square feet), and x_2 = rating (an overall "niceness rating" for the house expressed on a scale from 1 [worst] to 10 [best], and provided by the real estate agency). The sales prices of the ten observed houses are 180, 98.1, 173.1, 136.5, 141, 165.9, 193.5, 127.8, 163.5, and 172.5. The corresponding square footages are 23, 11, 20, 17, 15, 21, 24, 13, 19, and 25, and the corresponding niceness ratings are 5, 2, 9, 3, 8, 4, 7, 6, 7, and 2. If we fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_1 x_2 + \varepsilon$ to the observed data, we find that the least squares point estimates of the model parameters and their associated p-values (given in parentheses) are $b_0 = 27.438(<.001)$, $b_1 = 5.0813(<.001)$, $b_2 = 7.2899(<.001)$, $b_3 = -.5311(.001)$, and $b_4 = .11473(.014)$.

(a) A point prediction of and a 95 percent prediction interval for the sales price of a house having 2000 square feet $(x_1 = 20)$ and a niceness rating of $8(x_2 = 8)$ are 171.751 ($171,751) and [168.836, 174.665]. Using the above model, show how the point prediction is calculated.

(b) Table 3.6 gives predictions of sales prices of houses for six combinations of x_1 and x_2, and Figure 3.20 gives plots of the predictions needed to interpret the interaction between x_1 and x_2. Carefully interpret this interaction.

Table 3.6 Predicted real estate sales prices

x_2 \ x_1	13	22
2	108.933	156.730
5	124.124	175.019
8	129.756	183.748

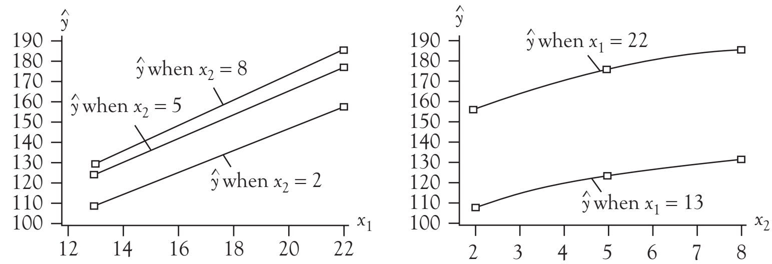
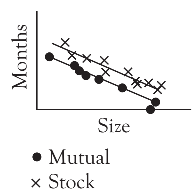


Figure 3.20 Predicted sales price interaction plots

Exercise 3.2

Kutner, Nachtsheim, and Li (2005) present twenty observations which they use to relate the speed, y, with which a particular insurance innovation is adopted to the size of the insurance firm, x, and the type of firm. The dependent variable y is measured by the number of months elapsed between the time the first firm adopted the innovation and the time the firm being considered adopted the innovation. The size of the firm, x, is measured by the total assets of the firm (in millions of dollars) and the type of firm—a qualitative independent variable—is either a mutual company or a stock company. The data consist of ten mutual companies, which have y values of 17, 26, 21, 30, 22, 0, 12, 19, 4, and 16 and corresponding x values of 151, 92, 175, 31, 104, 277, 210, 120, 290, and 238. The data also consists of ten stock companies, which have y values of 28, 15, 11, 38, 31, 21, 20, 13, 30, and 14 and corresponding x values of 164, 272, 295, 68, 85, 224, 166, 305, 124, and 246.

(a) Discuss why the data plot on the side of this exercise part indicates that the model $y = \beta_0 + \beta_1 x + \beta_2 D_S + \varepsilon$ might appropriately describe the obtained data. Here, D_S equals 1 if the firm is a stock firm and 0 if the firm is a mutual firm



(b) The model of part (a) implies that the mean adoption time of an insurance innovation by mutual companies having an asset size x equals $\beta_0 + \beta_1 x + \beta_2(0) = \beta_0 + \beta_1 x$ and that the mean adoption time by stock companies having an asset size x equals $\beta_0 + \beta_1 x + \beta_2(1) = \beta_0 + \beta_1 x + \beta_2$. What does β_2 represent?

(c) If we fit the model of part (a) to the data, we find that the least squares point estimates of β_0, β_1, and β_2 and their associated p-values (given in parentheses) are $b_0 = 33.8741(< .001)$, $b_1 = -.1017(< .001)$, and $b_2 = 8.0555\ (< .001)$. Interpret the meaning of $b_2 = 8.0555$.

(d) If we add the interaction term xD_S, to the model of part *a,* we find that the p-value related to this term is .9821. What does this imply?

Exercise 3.3

Recall from Example 3.2 that Enterprise Industries has observed the historical data in Table 3.2 concerning y(demand for Fresh liquid laundry detergent), x_4(the price difference), and x_3 (Enterprise Industries' advertising expenditure for Fresh). To ultimately increase the demand for Fresh, Enterprise Industries' marketing department is comparing the effectiveness of three different advertising campaigns. These campaigns are denoted as campaigns *A*, *B*, and *C*. Campaign *A* consists entirely of television commercials, campaign *B* consists of a balanced mixture of television and radio commercials, and campaign *C* consists of a balanced mixture of television, radio, newspaper, and magazine ads. To conduct the study, Enterprise Industries has randomly selected one advertising campaign to be used in each of the 30 sales periods in Table 3.2. Although logic would indicate that each of campaigns *A*, *B*, and *C* should be used in 10 of the 30 sales periods, Enterprise Industries has made previous commitments to the advertising media involved in the study. As a result, campaigns *A*, *B*, and *C* were randomly assigned to, respectively, 9, 11, and 10 sales periods. Furthermore, advertising was done in only the first three weeks of each sales period, so that the carryover effect of the campaign used in a sales period to the next sales period would be minimized, Table 3.7 lists the campaigns used in the sales periods.

To compare the effectiveness of advertising campaigns *A*, *B,* and *C*, we define two dummy variables. Specifically, we define the dummy variable

Table 3.7 Advertising campaigns used by enterprise industries

Sales period	Advertising campaign	Sales period	Advertising campaign
1	B	16	B
2	B	17	B
3	B	18	A
4	A	19	B
5	C	20	B
6	A	21	C
7	C	22	A
8	C	23	A
9	B	24	A
10	C	25	A
11	A	26	B
12	C	27	C
13	C	28	B
14	A	29	C
15	B	30	C

D_B to equal 1 if campaign *B* is used in a sales period and 0 otherwise. Furthermore, we define the dummy variable D_C to equal 1 if campaign *C* is used in a sales period and 0 otherwise. Figure 3.21 presents the SAS PROG REG output of a regression analysis of the Fresh demand data by using the model

$$y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 D_B + \beta_6 D_C + \varepsilon$$

To compare the advertising campaigns, consider comparing three means, denoted $\mu_{[d,a,A]}, \mu_{[d,a,B]}$, and $\mu_{[d,a,C]}$. These means represent the mean demands for Fresh when the price difference is d, the advertising expenditure is a, and we use advertising campaigns *A, B*, and *C*, respectively. If we set $x_4 = d$ and $x_3 = a$ in the expression

$$\beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 D_B + \beta_6 D_C$$

it follows that

```
                          Analysis of Variance
                                   Sum of           Mean
Source                    DF      Squares         Square    F Value    Pr > F

Model                      6     13.06502        2.17750     127.25    <.0001
Error                     23      0.39357        0.01711
Corrected Total           29     13.45859

           Root MSE              0.13081    R-Square     0.9708
           Dependent Mean        8.38267    Adj R-Sq     0.9631
           Coef Var              1.56050

                         Parameter Estimates
                                Parameter      Standard
Variable    Label         DF     Estimate         Error    t value    Pr > |t|

Intercept   Intercept     1      25.61270       4.79378       5.34      <.0001
X4          X4            1       9.05868       3.03170       2.99      0.0066
X3          X3            1      -6.53767       1.58137      -4.13      0.0004
X3SQ        X3 ** 2       1       0.58444       0.12987       4.50      0.0002
X4X3        X4 *  X3      1      -1.15648       0.45574      -2.54      0.0184
DB          DB            1       0.21369       0.06215       3.44      0.0022
DC          DC            1       0.38178       0.06125       6.23      <.0001

       Dep Var Predicted      Std Error
 Obs         Y     Value   Mean Predict   95% CL Mean         95% CL Predict
  31         .    8.5007         0.0469  8.4037     8.5977   8.2132    8.7881
```

Figure 3.21 SAS PROC REG output of a regression analysis of the fresh demand data using the model $y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 D_B + \beta_6 D_C + \varepsilon$

Parameter Estimates

Variable	Parameter Estimate	Standard Error	t value	Pr > |t|
Intercept	25.82638	4.79456	5.39	<.0001
X3	-6.53767	1.58137	-4.13	0.0004
X4	9.05868	3.03170	2.99	0.0066
X3SQ	0.58444	0.12987	4.50	0.0002
X4X3	-1.15648	0.45574	-2.54	0.0184
DA	-0.21369	0.06215	-3.44	0.0022
DC	0.16809	0.06371	2.64	0.0147

Figure 3.22 SAS PROC REG output for the fresh demand model $y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 D_B + \beta_6 D_C + \varepsilon$

$$\begin{aligned}\mu_{[d,a,A]} &= \beta_0 + \beta_1 d + \beta_2 a + \beta_3 a^2 + \beta_4 da + \beta_5(0) + \beta_6(0)\\ &= \beta_0 + \beta_1 d + \beta_2 a + \beta_3 a^2 + \beta_4 da\\ \mu_{[d,a,B]} &= \beta_0 + \beta_1 d + \beta_2 a + \beta_3 a^2 + \beta_4 da + \beta_5(1) + \beta_6(0)\\ &= \beta_0 + \beta_1 d + \beta_2 a + \beta_3 a^2 + \beta_4 da + \beta_5\end{aligned}$$

and

$$\begin{aligned}\mu_{[d,a,C]} &= \beta_0 + \beta_1 d + \beta_2 a + \beta_3 a^2 + \beta_4 da + \beta_5(0) + \beta_6(1)\\ &= \beta_0 + \beta_1 d + \beta_2 a + \beta_3 a^2 + \beta_4 da + \beta_6\end{aligned}$$

These equations imply that: $\mu_{[d,a,B]} - \mu_{[d,a,A]} = \beta_5$

$$\mu_{[d,a,C]} - \mu_{[d,a,A]} = \beta_6 \text{ and } \mu_{[d,a,C]} - \mu_{[d,a,B]} = \beta_6 - \beta_5$$

(a) Use the least squares point estimates of the model parameters to find a point estimate of each of the three differences in means. Also, find a 95 percent confidence interval for and test the significance of each of the first two differences in means.

(b) The prediction results at the bottom of the SAS output correspond to a future period when the price difference will be $x_4 = .20$, the advertising expenditure $x_3 = 6.50$, and campaign C will be used. Show how $\hat{y} = 8.5007$ is calculated. Identify and interpret a 95 percent confidence interval for the mean demand and a 95 percent prediction interval for an individual demand when $x_4 = .20$, $x_3 = 6.50$, and campaign C is used.

(c) Consider the alternative model

$$y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 D_A + \beta_6 D_C + \varepsilon$$

Here D_A equals 1 if advertising campaign A is used and 0 otherwise. The SAS PROC REG output of the least squares point estimates of the parameters of this model is given in Figure 3.22. Since β_6 compares the effect of advertising campaign C with respect to the effect of advertising campaign B, β_6 equals $\mu_{[d,a,C]} - \mu_{[d,a,B]}$. Find a 95 percent confidence interval for and test the significance of $\mu_{[d,a,C]} - \mu_{[d,a,B]}$.

(d) Figure 3.23 presents the SAS output using the model

$$y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 D_B + \beta_6 D_C + \beta_7 x_3 D_B + \beta_8 x_3 D_C + \varepsilon$$

When there are many independent variables in a model, we might not be able to trust the p-values to tell us what is important. This is because of a condition called *multicollinearity*, which is discussed in Section 4.1. Note, however, that the p-value for $x_3 D_C$ is the smallest of the p-values for the independent variables D_B, D_C, $x_3 D_B$, and $x_3 D_C$. This might be regarded as "some evidence" that "some interaction" exists between advertising expenditure and advertising campaign. To further investigate this interaction, note that the model utilizing $x_3 D_B$ and $x_3 D_C$ implies that

$$\mu_{[d,a,A]} = \beta_0 + \beta_1 d + \beta_2 a + \beta_3 a^2 + \beta_4 da + \beta_5(0) + \beta_6(0) + \beta_7 a(0) + \beta_8 a(0)$$
$$\mu_{[d,a,B]} = \beta_0 + \beta_1 d + \beta_2 a + \beta_3 a^2 + \beta_4 da + \beta_5(1) + \beta_6(0) + \beta_7 a(1) + \beta_8 a(0)$$
$$\mu_{[d,a,C]} = \beta_0 + \beta_1 d + \beta_2 a + \beta_3 a^2 + \beta_4 da + \beta_5(0) + \beta_6(1) + \beta_7 a(0) + \beta_8 a(1)$$

(1) Using these equations verify that $\mu_{[d,a,C]} - \mu_{[d,a,A]}$ equals $\beta_6 + \beta_8 a$. (2) Using the least squares point estimates in Figure 3.23, show that a point estimate of $\mu_{[d,a,C]} - \mu_{[d,a,A]}$ equals .3266 when $a = 6.2$ and equals .4080 when $a = 6.6$. (3) Verify that $\mu_{[d,a,C]} - \mu_{[d,a,B]}$ equals

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	28.68734	5.12847	5.59	<.0001
X3	X3	1	-7.41146	1.66169	-4.46	0.0002
X4	X4	1	10.82532	3.29880	3.28	0.0036
X3SQ	X3 ** 2	1	0.64584	0.13460	4.80	<.0001
X4X3	X3 * X4	1	-1.41562	0.49287	-2.87	0.0091
DB	DB	1	-0.48068	0.73089	-0.66	0.5179
DC	DC	1	-0.93507	0.83572	-1.12	0.2758
X3DB	X3 * DB	1	0.10722	0.11169	0.96	0.3480
X3DC	X3 * DC	1	0.20349	0.12882	1.58	0.1291

Obs	Dep Var y	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict	
31	.	8.5118	0.0479	8.4123	8.6114	8.2249	8.7988

Figure 3.23 Partial SAS PROC REG output for the fresh demand model

$$y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 D_B + \beta_6 D_C + \beta_7 x_3 D_B + \beta_8 x_3 D_C + \varepsilon$$

$\beta_6 - \beta_5 + \beta_8 a - \beta_7 a$. (4) Using the least squares point estimates, show that a point estimate of $\mu_{[d,a,C]} - \mu_{[d,a,B]}$ equals .14266 when $a = 6.2$ and equals .18118 when $a = 6.6$ (5) Discuss why these results imply that the larger the advertising expenditure a is, then the larger is the improvement in mean sales that is obtained by using advertising campaign C rather than advertising campaign A or B.

(e) Figures 3.21 and 3.23 give 95 percent prediction intervals of demand for Fresh in a future sales period when the price difference will be $x_4 =$ .20, the advertising expenditure will be $x_3 = 6.50$, and campaign C will be used. Which model—the one in Figure 3.21 that assumes that no interaction exists between advertising expenditure and advertising campaign, or the one in Figure 3.23 that assumes that such interaction does exist—gives the shortest 95 percent prediction interval?

(f) Using all the information in this exercise, discuss why it might be reasonable to conclude that a small amount of interaction exists between advertising expenditure and advertising campaign.

In Exercises 3.4 through 3.6 you will perform partial F tests by using the following three Fresh detergent models:

$$\text{Model 1}: y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \varepsilon$$
$$\text{Model 2}: y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 D_B + \beta_6 D_C + \varepsilon$$
$$\text{Model 3}: y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 D_B + \beta_6 D_C + \beta_7 x_3 D_B + \beta_8 x_3 D_C + \varepsilon$$

The values of SSE for models 1, 2, and 3 are, respectively, 1.0644, .3936, and .3518.

***Exercise* 3.4** In Model 2, test $H_0: \beta_5 = \beta_6 = 0$ by setting α equal to .05. Reason that testing $H_0: \beta_5 = \beta_6 = 0$ is equivalent to testing $H_0: \mu_{[d,a,A]} = \mu_{[d,a,B]} = \mu_{[d,a,C]}$. Interpret what this says.

***Exercise* 3.5** In Model 3, test $H_0: \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$ by setting α equal to .05. Interpret your results.

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
MUDAB - MUDAA	0.21368626	3.44	0.0022	0.06215362
MUDAC - MUDAA	0.38177617	6.23	0.0001	0.06125253
MUDAC - MUDAB	0.16808991	2.64	0.0147	0.06370664

Figure 3.24 Partial SAS PROC GLM output for the fresh demand model $y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 D_B + \beta_6 D_C + \varepsilon$

***Exercise* 3.6** In Model 3, test $H_0: \beta_7 = \beta_8 = 0$ by setting α equal to .05. Interpret your results.

***Exercise* 3.7** Figure 3.24 presents a partial SAS PROC GLM output obtained by using the model

$$y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 D_B + \beta_6 D_C + \varepsilon$$

to analyze the Fresh demand data. On the output, MUDAB–MUDAA $= \mu_{[d,a,B]} - \mu_{[d,a,A]} = \beta_5$, MUDAC–MUDAA $= \mu_{[d,a,C]} - \mu_{[d,a,A]} = \beta_6$, and MUDAC–MUDAB $= \mu_{[d,a,C]} - \mu_{[d,a,B]} = \beta_6 - \beta_5$. The point estimate of $\ell = \beta_6 - \beta_5$ is $\hat{\ell} = b_6 - b_5 = .38177617 - .21368626 = .16808991$, which is given on the SAS output, and the standard error of this point estimate is $s_{\hat{\ell}} = s\sqrt{\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda}} = .06370664$, which is also given on the SAS output. Specify what the row vector $\boldsymbol{\lambda}'$ equals and calculate a 95% confidence interval for $\mu_{[d,a,C]} - \mu_{[d,a,B]} = \beta_6 - \beta_5$. Is this interval the same interval (within rounding) that you obtained using the alternative dummy variable model in part (c) of Exercise 3.3?

***Exercise* 3.8** Use the information in Figure 3.24 to calculate Bonferroni simultaneous 95 percent confidence intervals for $\mu_{[d,a,B]} - \mu_{[d,a,A]} = \beta_5$, $\mu_{[d,a,C]} - \mu_{[d,a,A]} = \beta_6$, and $\mu_{[d,a,C]} - \mu_{[d,a,B]} = \beta_{6,} - \beta_5$. Interpret these intervals.

***Exercise* 3.9** Recall from Exercise 3.3 that we have used the Fresh detergent demand model

$$y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 D_B + \beta_6 D_C + \beta_7 x_3 D_B + \beta_8 x_3 D_C + \varepsilon$$

to relate y to x_4, x_3, and the advertising strategy used to promote Fresh. Here D_B equals 1 if advertising strategy B is used and 0 otherwise; D_C equals 1 if advertising strategy C is used and 0 otherwise. Table 3.7 gives the advertising strategies used in the 30 sales periods. Noting that the advertising strategies employed in periods 1, 2, 3, 4, and 30 were B, B, B, A, and C, we use a column vector $\mathbf{y}$ containing the 30 demands in Table 3.2 and the matrix $\mathbf{X}$ given in Figure 3.25 to calculate the least squares point estimates. Figure 3.25 also presents a partial PROC GLM output of a regression analysis using these matrices.

(a) Using $(\mathbf{X'X})^{-1}$ and $\mathbf{X'y}$, show how the least squares point estimates have been calculated.

(b) Consider a single sales period when the price difference is \$.20, advertising expenditure is \$650,000, and advertising strategy C is used. The SAS output tells us that a point prediction of demand for Fresh in this sales period is (see Observation 31)

$$\begin{aligned}\hat{y} &= b_0 + b_1(.20) + b_2(6.50) + b_3(6.50)^2 + b_4(.20)(6.50) \\ &\quad + b_5(0) + b_6(1) + b_7(6.50)(0) + b_8(6.50)(1) \\ &= 8.5118\end{aligned}$$

The SAS output also tells us that a 95 percent prediction interval for demand for Fresh in this sales period is [8.2249, 8.7988]. What is the row vector $\mathbf{x}_0'$ that is used to calculate this prediction interval by the formula $[\hat{y} \pm t_{[\alpha/2]} s\sqrt{1 + \mathbf{x}_0'(\mathbf{X'X})^{-1}\mathbf{x}_0}]$?

(c) D1FF1, DIFF2, DIFF3, and DIFF4 on the SAS output are

$$\begin{aligned}\text{DIFF1} &= \mu_{[d,a,C]} - \mu_{[d,a,A]} = \beta_6 + \beta_8(6.2) \\ \text{DIFF2} &= \mu_{[d,a,C]} - \mu_{[d,a,A]} = \beta_6 + \beta_8(6.6) \\ \text{DIFF3} &= \mu_{[d,a,C]} - \mu_{[d,a,B]} = \beta_6 - \beta_5 + \beta_8(6.2) - \beta_7(6.2) \\ \text{DIFF4} &= \mu_{[d,a,C]} - \mu_{[d,a,B]} = \beta_6 - \beta_5 + \beta_8(6.6) - \beta_7(6.6)\end{aligned}$$

$$\mathbf{X} = \begin{array}{c} \begin{array}{ccccccccc} 1 & x_4 & x_3 & x_3^2 & x_4x_3 & D_B & D_C & x_3D_B & x_3D_C \end{array} \\ \begin{bmatrix} 1 & -.05 & 5.50 & (5.50)^2 & (-.05)(5.50) & 1 & 0 & (5.50)(1) & (5.50)(0) \\ 1 & .25 & 6.75 & (6.75)^2 & (.25)(6.75) & 1 & 0 & (6.75)(1) & (6.75)(0) \\ 1 & .60 & 7.25 & (7.25)^2 & (.60)(7.25) & 1 & 0 & (7.25)(1) & (7.25)(0) \\ 1 & 0 & 5.50 & (5.50)^2 & (0)(5.50) & 0 & 0 & (5.50)(0) & (5.50)(0) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & .55 & 6.80 & (6.80)^2 & (.55)(6.80) & 0 & 1 & (6.80)(0) & (6.80)(1) \end{bmatrix} \end{array} \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 251.48 \\ 57.646 \\ 1632.781 \\ 10677.40275 \\ 397.74425 \\ 93.12 \\ 84.31 \\ 608.815 \\ 538.857 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} 28.687341618 \\ 10.825323968 \\ -7.411462373 \\ 0.6458377529 \\ -1.415623462 \\ -0.480676068 \\ -0.935072983 \\ 0.1072216076 \\ 0.2034866904 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 1570.2166784 & 715.91258349 & -507.5709271 & 40.848930939 & -108.8802964 & -46.06948829 & -99.05615848 & 7.1382273495 & 14.985137276 \\ 715.91258349 & 649.67279264 & -233.3895714 & 18.922149559 & -96.91280936 & 0.8689183154 & -57.76335611 & 0.3095802122 & 8.7293799345 \\ -507.5709271 & -233.3895714 & 164.84861139 & -13.32510731 & 35.568798578 & 11.401693141 & 28.239059064 & -1.781312674 & -4.266435431 \\ 40.848930939 & 18.922149559 & -13.32510731 & 1.0815492382 & -2.890482718 & -0.669108229 & -1.994271188 & 0.1055660411 & 0.3003816013 \\ -108.8802964 & -96.91280936 & 35.568798578 & -2.890482718 & 14.502515125 & 0.0050715341 & 8.4814570462 & -0.067312284 & -1.279797485 \\ -46.06948829 & 0.8689183154 & 11.401693141 & -0.669108229 & 0.0050715341 & 31.892710178 & 21.543187306 & -4.856450129 & -3.282203044 \\ -99.05615848 & -57.76335611 & 28.239059064 & -1.994271188 & 8.4814570462 & 21.543187306 & 41.6971799 & -3.311305375 & -6.410044917 \\ 7.1382273495 & 0.3095802122 & -1.781312674 & 0.1055660411 & -0.067313284 & -4.856450129 & -3.311305375 & 0.7448064027 & 0.5069898915 \\ 14.985137276 & 8.7293799345 & -4.266453431 & 0.3003816013 & -1.279797485 & -3.282203044 & -6.410044917 & 0.5069898915 & 0.9906624167 \end{bmatrix}$$

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
DIFF1	0.32654450	4.66	0.0001	0.07013744
DIFF2	0.40793917	6.46	0.0001	0.06311786
DIFF3	0.14244660	2.04	0.0547	0.06999803
DIFF4	0.18095263	2.81	0.0106	0.06447170

Observation	Observed Value	Predicted Value	Residual	Lower 95% CL for Individual	Upper 95% CL for Individual
31	.	8.51182605	.	8.22486409	8.79878801

Figure 3.25 The matrix X and a partial SAS PROC GLM output using the model
$y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 D_B + \beta_6 D_C + \beta_7 x_3 D_B + \beta_8 x_3 D_C + \varepsilon$

Each of these differences is a *linear combination of regression parameters* (that is, the β_j's). The point estimate of $l = \text{DIFF4} = \beta_6 - \beta_5 + \beta_8(6.6) - \beta_7(6.6)$ is

$$\begin{aligned}\hat{l} &= b_6 - b_5 + b_8(6.6) - b_7(6.6)\\ &= -.93507 - (-.48068) + .20349(6.6) - .10722(6.6)\\ &= .18095\end{aligned}$$

which is given on the SAS output. Moreover, note that

$$\begin{aligned}l = \beta_6 - \beta_5 + \beta_8(6.6) - \beta_7(6.6) = (0)\beta_0 + (0)\beta_1 + (0)\beta_2 + (0)\beta_3 + 0(\beta_4)\\ +(-1)\beta_5 + (1)\beta_6 + (-6.6)\beta_7 + (6.6)\beta_8 = \boldsymbol{\lambda}'\boldsymbol{\beta},\end{aligned}$$

where $\boldsymbol{\lambda}' = [0\ 0\ 0\ 0\ 0\ -1\ 1\ -6.6\ 6.6]$. It follows that the standard error of the estimate $\hat{l}$, denoted $s_{\hat{l}}$, is calculated by the equation $s_{\hat{l}} = s\sqrt{\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda}}$. Here $s = .1294$ is the standard error for the model (that is, $s = \sqrt{SSE/(n-(k+1))}$), and $\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda}$ for DIFF4 can be calculated to be .2482388. Therefore, $s_{\hat{l}}$ for DIFF4 is $s\sqrt{\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda}} = .1294\sqrt{.2482388}$, or .06447170 (see Figure 3.25). Find $\boldsymbol{\lambda}'$ for DIFF1, DIFF2, and DIFF3. Then, using the fact that $\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\lambda}$ for DIFF1, DIFF2, and DIFF3 can be calculated to be .2937858, .2379223, and .2926191, calculate $s_{\hat{l}}$ for DIFF1, DIFF2, and DIFF3. Also, calculate 95 percent confidence intervals for DIFF1, DIFF2, DIFF3, and DIFF4. Interpret what these intervals say.

***Exercise* 3.10** If we use the logistic regression model $p(x_1) = e^{(\beta_0+\beta_1 x_1)} / [1 + e^{(\beta_0+\beta_1 x_1)}]$ to analyze the performance data in Section 3.6, we obtain maximum likelihood estimates of β_0 and β_1 equal to –43.3684 and $b_1 = .4897$. We also find that a point estimate of and a 95 percent confidence interval for the probability of successful performance for (1) a potential employee who scores a 93 on test 1 are .89804 and [.67987, .97336]; (2) a potential employee who scores 85 on test 1 are .14905 and [.03915, .42951]. Show how the point estimates have been calculated, and compare the lengths of the confidence intervals with the lengths of the corresponding confidence intervals in Figure 3.17. Also, calculate and interpret a point estimate of the odds ratio for x_1.

***Exercise* 3.11** Mendenhall and Sinicich (2011) present data that can be used to investigate allegations of gender discrimination in the hiring practices of a particular firm. Of the twenty-eight candidates who applied for employment at the firm, nine were hired. The combinations of education x_1, (in years), experience x_2, (in years), and gender x_3 (a dummy variable that equals 1 if the potential employee was a male and 0 if the potential employee was a female) for the nine hired candidates were (6, 6, 1), (6, 3, 1), (8, 3, 0), (8, 10, 0), (4, 5, 1), (6, 1, 1), (8, 5, 1), (4, 10, 1), and (6, 12, 0). For the nineteen candidates that were not hired, the combinations of values of x_1, x_2, and x_3 were (6, 2, 0), (4, 0, 1), (4, 1, 0), (4, 2, 1), (4, 4, 0), (6, 1, 0), (4, 2, 1), (8, 5, 0), (4, 2, 0), (6, 7, 0), (6, 4, 0), (8, 0, 1), (4, 7, 0), (4, 1, 1), (4, 5, 0), (6, 0, 1), (4, 9, 0)(8, 1, 0), and (6, 1, 0). If $p(x_1, x_2, x_3)$ denotes the probability of a potential employee being hired, and if we use the logistic regression model $p(x_1, x_2, x_3) = e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)} / [1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}]$ to analyze these data, we find that the point estimates of the model parameters and their associated p-values (given in parentheses) are $b_0 = -14.2483(.0191)$, $b_1 = 1.1549(.0552)$, $b_2, = .9098(.0341)$, and $b_3 = 5.6037(.0313)$.

(a) Consider a potential employee having 4 years of education and 5 years of experience. Find a point estimate of the probability that the potential employee will be hired if the potential employee is a male, and find a point estimate of the probability that the potential employee will be hired if the potential employee is a female.

(b) Using $b_3 = 5.6037$, find a point estimate of the odds ratio for x_3. Interpret this odds ratio. Using the p-value describing the importance of x_3, can we conclude that there is strong evidence that gender is related to the probability that a potential employee will be hired?

CHAPTER 4

Model Building and Model Diagnostics

4.1 Step 1: Preliminary Analysis and Assessing Multicollinearity

Recall the sales territory performance data in Table 2.5. These data consist of values of the dependent variable y (Sales) and of the independent variables x_1 (Time), x_2 (MktPoten), x_3 (Adver), x_4 (MktShare), and x_5 (Change). The complete sales territory performance data analyzed by Cravens, Woodruff, and Stomper (1972) consists of the data presented in Table 2.5 and data concerning three additional independent variables. These three additional variables are x_6 = number of accounts handled by the representative (Accts); x_7 = average workload per account, measured by using a weighting based on the sizes of the orders by the accounts and other workload-related criteria (Wkload); and x_8 = an aggregate rating on eight dimensions of the representative's performance, made by a sales manager and expressed on a 1 to 7 scale (Rating).

Table 4.1 gives the observed values of x_6, x_7, and x_8, and Figure 4.1 presents the MINITAB output of a *correlation matrix* for the sales territory performance data. Examining the first column of the matrix, we see that the simple correlation coefficient between Sales and Wkload is –.117 and that the p-value for testing the significance of the relationship between Sales and Wkload is .577. This indicates that there is little or no relationship between Sales and Wkload. However, the simple correlation coefficients between Sales and the other seven independent variables range from .402 to .754, with associated p-values ranging from .046 to .000. This indicates the existence of potentially useful relationships between Sales and these seven independent variables.

Although simple correlation coefficients (and scatter plots) give us a preliminary understanding of the data, they cannot be relied upon

Table 4.1 Values of Accts, Wkload, and Rating

Accounts, x_6	Workload, x_7	Rating, x_8
74.86	15.05	4.9
107.32	19.97	5.1
96.75	17.34	2.9
195.12	13.40	3.4
180.44	17.64	4.6
104.88	16.22	4.5
256.10	18.80	4.6
126.83	19.86	2.3
203.25	17.42	4.9
119.51	21.41	2.8
116.26	16.32	3.1
142.28	14.51	4.2
89.43	19.35	4.3
84.55	20.02	4.2
119.51	15.26	5.5
80.49	15.87	3.6
136.58	7.81	3.4
78.86	16.00	4.2
136.58	17.44	3.6
138.21	17.98	3.1
75.61	20.99	1.6
102.44	21.66	3.4
76.42	21.46	2.7
136.58	24.78	2.8
88.62	24.96	3.9

alone to tell us which independent variables are significantly related to the dependent variable. One reason for this is a condition called multicollinearity. *Multicollinearity* is said to exist among the independent variables in a regression situation if these independent variables are related to or dependent upon each other. One way to investigate multicollinearity is to examine the correlation matrix. To understand this, note that all of the simple correlation coefficients not located in the first column of this matrix measure the *simple correlations between the independent variables*. For example, the simple correlation coefficient between Accts and

	Sales	Time	Mkt Poten	Adver	Mkt Share	Change	Accts	WkLoad
Time	0.623							
	0.001							
MktPoten	0.598	0.454						
	0.002	0.023						
Adver	0.596	0.249	0.174					
	0.002	0.230	0.405					
MktShare	0.484	0.106	-0.211	0.264				
	0.014	0.613	0.312	0.201				
Change	0.489	0.251	0.268	0.377	0.085			
	0.013	0.225	0.195	0.064	0.685			
Accts	0.754	0.758	0.479	0.200	0.403	0.327		
	0.000	0.000	0.016	0.338	0.046	0.110		
WkLoad	-0.117	-0.179	-0.259	-0.272	0.349	-0.288	-0.199	
	0.577	0.391	0.212	0.188	0.087	0.163	0.341	
Rating	0.402	0.101	0.359	0.411	-0.024	0.549	0.229	-0.277
	0.046	0.631	0.078	0.041	0.911	0.004	0.272	0.180

Cell contents: Pearson correlation
P-Value

Figure 4.1 MINITAB output of the correlation matrix

Time is .758, which says that the Accts values increase as the Time values increase. Such a relationship makes sense because it is logical that the longer a sales representative has been with the company the more accounts he or she handles. Statisticians often regard multicollinearity in a dataset to be severe if at least one simple correlation coefficient between the independent variables is at least .9. Since the largest such simple correlation coefficient in Figure 4.1 is .758, this is not true for the sales territory performance data. Note, however, that even moderate multicollinearity can be a potential problem. This will be demonstrated later using the sales territory performance data.

Another way to measure multicollinearity is to use *variance inflation factors*. Consider a regression model relating a dependent variable y to a set of independent variables $x_1,\ldots,x_{j-1},x_j,x_{j+1},\ldots,x_k$. The *variance inflation factor* for the independent variable x_j in this set is denoted VIF_j and is defined by the equation

$$VIF_j = \frac{1}{1-R_j^2}$$

where R_j^2 is the multiple coefficient of determination for the regression model that relates x_j to all the other independent variables $x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_k$ in the set. For example, Figure 4.2 gives the SAS

Predictor	Coef	SE Coef	T	P	VIF
Constant	-1507.8	778.6	-1.94	0.071	
Time	2.010	1.931	1.04	0.313	3.343
MktPoten	0.037205	0.008202	4.54	0.000	1.978
Adver	0.15099	0.04711	3.21	0.006	1.910
MktShare	199.02	67.03	2.97	0.009	3.236
Change	290.9	186.8	1.56	0.139	1.602
Accts	5.551	4.776	1.16	0.262	5.639
WkLoad	19.79	33.68	0.59	0.565	1.818
Rating	8.2	128.5	0.06	0.950	1.809

Figure 4.2 The t statistics, p-values, and variance inflation factors for the eight independent variables model

output of the t-statistics, p-values, and variance inflation factors for the sales territory performance model that relates y to all eight independent variables. The largest variance inflation factor is $VIF_6 = 5.639$. To calculate VIF_6, SAS first calculates the multiple coefficient of determination for the regression model that relates x_6 to x_1, x_2, x_3, x_4, x_5, x_7, and x_8 to be $R_6^2 = .822673$. It then follows that

$$VIF_6 = \frac{1}{1 - R_6^2} = \frac{1}{1 - .822673} = 5.639$$

VIF_j is called the variance inflation factor because it can be shown that $\sigma_{b_j}^2$, the variance of the population of all possible values of the least squares point estimate b_j is related to VIF_j by the equation $\sigma_{b_j}^2 = \sigma^2 (VIF_j) / SS_{x_j x_j}$ where $SS_{x_j x_j} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$. If R_j^2 =0, that is, if x_j is not related to the other independent variables $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$ through a multiple regression model that relates x_j to $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$, then the variance inflation factor $VIF_j = 1/(1 - R_j^2)$ equals 1. In this case $\sigma_{b_j}^2 = \sigma^2 / SS_{x_j x_j}$. If $R_j^2 > 0, x_j$ is related to the other independent variables. This implies that $1 - R_j^2$ is less than 1, and $VIF = 1/(1 - R_j^2)$ is greater than 1. Therefore, $\sigma_{b_j}^2 = \sigma^2 (VIF_j) / SS_{x_j x_j}$ is inflated beyond the value of $\sigma_{b_j}^2$ when $R_j^2 = 0$. Usually, the multicollinearity between independent variables is considered (1) severe if the largest variance inflation factor is greater than 10 and (2) moderately strong if the largest variance inflation factor is greater than five. Moreover, if the mean of the variance inflation factors is substantially greater than one (sometimes a difficult criterion to assess), multicollinearity might be problematic. In the sales territory performance model, the largest variance inflation factor, $VIF_6 = 5.639$, is greater than five. Therefore, we might classify the multicollinearity as being moderately strong.

If there is strong multicollinearity, then two slightly different samples of values of the dependent variable can yield two substantially different values of b_j. To intuitively understand why strong multicollinearity can significantly affect the least squares point estimates, consider the so-called *picket fence* display in Figure 4.3. This figure depicts two independent variables (x_1 and x_2) exhibiting strong multicollinearity (note that as x_1 increases, x_2 increases). The heights of the pickets on the fence represent the y observations. If we assume that the model

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

adequately describes this data, then calculating the least squares point estimates is like fitting a plane to the points on the top of the picket fence. Clearly, this plane would be quite unstable. That is, a slightly different height of one of the pickets (a slightly different y value) could cause the slant of the fitted plane (and the least squares point estimates that determine this slant) to radically change. It follows that when strong multicollinearity exists, sampling variation can result in least squares point estimates that differ substantially from the true values of the regression parameters. In fact, some of the least squares point estimates may have a sign (positive or negative) that differs from the sign of the true value of the parameter (we will see an example of this in the exercises). Therefore, when strong multicollinearity exists, it is dangerous to individually interpret the least squares point estimates.

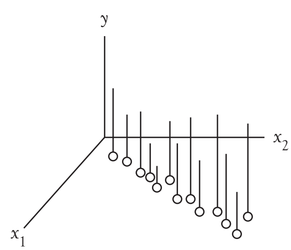


Figure 4.3 The picket fence display

The most important problem caused by multicollinearity is that even when the multicollinearity is not severe, it can hinder our ability to use the t-statistics and related p-values to assess the importance of the independent variables. Recall that we can reject $H_0 : \beta_j = 0$ in favor of $H_a : \beta_j \neq 0$ at level of significance α if and only if the absolute value of the corresponding t-statistic is greater than $t_{[\alpha/2]}$, or equivalently, if and only if the related p-value is less than α. Thus the larger (in absolute value) the t-statistic is and the smaller the p-value is, the stronger is the evidence that we should reject $H_0 : \beta_j = 0$ and the stronger is the evidence that the independent variable x_j is significant. When multicollinearity exists, the sizes of the t-statistic and of the related p-value *measure the additional importance of the independent variable* x_j *over the combined importance of the other independent variables in the regression model.* Since two or more correlated independent variables contribute redundant information, multicollinearity often causes the t-statistics obtained by relating a dependent variable to a set of correlated independent variables to be smaller (in absolute value) than the t-statistics that would be obtained if separate regression analyses were run, where each separate regression analysis relates the dependent variable to a smaller set (for example, only one) of the correlated independent variables. Thus, multicollinearity can cause some of the correlated independent variables to appear less important—in terms of having small absolute t-statistics and large p-values—than they really are. Another way to understand this is to note that since multicollinearity inflates σ_{b_j} it inflates the point estimate s_{b_j} of σ_{b_j}. Since $t = b_j / s_{b_j}$, an inflated value of s_{b_j} can (depending on the size of b_j) cause t to be small (and the related p-value to be large). This would suggest that x_j is not significant even though x_j may really be important.

For example, Figure 4.2 tells us that when we perform a regression analysis of the sales territory performance data using a model that relates y to all eight independent variables, the p-values related to Time, MktPoten, Adver, MktShare, Change, Accts, Wkload, and Rating are, respectively, .3134, .0003, .0055, .0090, .1390, .2621, .5649, and .9500. By contrast, recall from Table 2.5c that when we perform a regression analysis of the sales territory performance data using a model that relates y to the first five independent variables, the p-values related to Time, MktPoten, Adver, MktShare, and Change are, respectively, .0065, .0001,

.0025, .0001, and .0530. Note that Time (p-value = .0065) seems highly significant and Change (p-value = .0530) seems *somewhat significant* in the five-independent-variable model. However, when we consider the model that uses all eight independent variables, Time (p-value = .3134) seems *insignificant* and Change (p-value = .1390) seems *somewhat insignificant*. The reason that Time and Change seem more significant in the model with five independent variables is that since this model uses fewer variables, Time and Change contribute less overlapping information and thus have additional importance in this model.

4.2 Step 2: Comparing Regression Models: Model Comparison Statistics

We have seen that when multicollinearity exists in a model, the p-value associated with an independent variable in the model measures the additional importance of the variable over the combined importance of the variables in the model. Therefore, it can be difficult to use the p-values to determine which variables to retain in and which variables to remove from a model. The implication is that we need to evaluate more than the *additional importance* of each independent variable in a regression model. We also need to evaluate how well the independent variables *work together* to accurately describe, predict, and control the dependent variable. One way to do this is to determine if the *overall* model gives a high R^2 and $\bar{R}^2$, a small s, and short prediction intervals.

It can be proved that *adding any independent variable to a regression model, even an unimportant independent variable, will decrease the unexplained variation and will increase the explained variation.* Therefore, since the total variation $\sum (y_i - \bar{y})^2$ depends only on the observed y values and thus remains unchanged when we add an independent variable to a regression model, it follows that *adding any independent variable to a regression model will increase the coefficient of determination* $R^2 = (\textit{Explained variation}) / (\textit{Total variation})$. This implies that R^2 cannot tell us (by decreasing) that adding an independent variable is undesirable. That is, although we wish to obtain a model with a large R^2, there are better criteria than R^2 that can be used to compare regression models.

One better criterion is the standard error

$$s = \sqrt{\frac{SSE}{n-(k+1)}}$$

When we add an independent variable to a regression model, the number of model parameters $(k+1)$ increases by one, and thus the number of degrees of freedom $n-(k+1)$ decreases by one. If the decrease in $n-(k+1)$, which is used in the denominator to calculate s, is proportionally more than the decrease in SSE (the unexplained variation) that is caused by adding the independent variable to the model, then s will increase. *If s increases, this tells us that we should not add the independent variable to the model.* To see one reason why, consider the formula for the prediction interval for y

$$[\hat{y} \pm t_{[\alpha/2]} s \sqrt{1 + \text{Distance value}}]$$

Since adding an independent variable to a model decreases the number of degrees of freedom, adding the variable will increase the $t_{[\alpha/2]}$ point used to calculate the prediction interval. To understand this, look at any column of the *t*-table in Table A2 and scan from the bottom of the column to the top—you can see that the *t*-points increase as the degrees of freedom decrease. It can also be shown that adding any independent variable to a regression model will not decrease (and usually increases) the distance value. Therefore, since adding an independent variable increases $t_{[\alpha/2]}$ and does not decrease the distance value, *if s increases, the length of the prediction interval for y will increase.* This means the model will predict less accurately and thus we should not add the independent variable.

On the other hand, if adding an independent variable to a regression model decreases s, the length of a prediction interval for y will decrease if and only if the decrease in s is enough to offset the increase in $t_{[\alpha/2]}$ and the (possible) increase in the distance value. Therefore, *an independent variable should not be included in a final regression model unless it reduces s enough to reduce the length of the desired prediction interval for y*. However, we must balance the length of the prediction interval, or in general, the goodness of any criterion, against the difficulty and expense of using the model. For instance, predicting y requires knowing the corresponding values of the independent variables. So we must decide whether including an independent variable reduces s and prediction interval lengths enough to offset the potential errors caused by possible inaccurate determination of values of the independent variables, or the possible expense of determining these values. If adding an independent variable provides prediction intervals that are only slightly shorter while making the model more difficult and more expensive to use, we might decide that including the variable is not desirable.

Since a key factor is the length of the prediction intervals provided by the model, one might wonder why we do not simply make direct comparisons of prediction interval lengths (without looking at s). It is useful to compare interval lengths, but these lengths depend on the distance value, which depends on how far the values of the independent variables we wish to predict for are from the center of the experimental region. We often wish to compute prediction intervals for several different combinations of values of the independent variables (and thus for several different values of the distance value). Thus we would compute prediction intervals having slightly different lengths. However, the standard error s is a constant factor with respect to the length of prediction intervals (as long as we are considering the same regression model). Thus it is common practice to compare regression models on the basis of s(and s^2). Finally, note that it can be shown that *the standard error s decreases if and only if $\bar{R}^2$ (adjusted R^2) increases.* It follows that if we are comparing regression models*, the model that gives the smallest s gives the largest $\bar{R}^2$.*

Example 4.1

Figure 4.4 gives MINITAB output resulting from calculating R^2, $\bar{R}^2$, and s for all possible regression models based on all possible combinations of the eight independent variables in the sales territory performance situation (the values of C_p on the output will be explained after we complete this example). The MINITAB output gives the two best models of each size in terms of s and $\bar{R}^2$—the two best one-variable models, the two best two-variable models, and so on. Examining Figure 4.4, we see that the three models having the smallest values of s and the largest values of $\bar{R}^2$ are

Vars	R-Sq	R-Sq(adj)	Mallows C-P	S	Time	MktPoten	Adver	MktShare	Change	Accts	WkLoad	Rating
1	56.8	55.0	67.6	881.09						X		
1	38.8	36.1	104.6	1049.3	X							
2	77.5	75.5	27.2	650.39			X			X		
2	74.6	72.3	33.1	691.11		X		X				
3	84.9	82.7	14.0	545.51		X	X	X				
3	82.8	80.3	18.4	582.64		X	X			X		
4	90.0	88.1	5.4	453.84		X	X	X		X		
4	89.6	87.5	6.4	463.95	X	X	X	X				
5	91.5	89.3	4.4	430.23	X	X	X	X	X			
5	91.2	88.9	5.0	436.75		X	X	X	X	X		
6	92.0	89.4	5.4	428.00	X	X	X	X	X	X		
6	91.6	88.9	6.1	438.20		X	X	X	X	X	X	
7	92.2	89.0	7.0	435.67	X	X	X	X	X	X	X	
7	92.0	88.8	7.3	440.30	X	X	X	X	X	X		X
8	92.2	88.3	9.0	449.03	X	X	X	X	X	X	X	X

Figure 4.4 MINITAB output of the two best sales territory performance regression models of each size

1. the six-variable model that contains

   Time, MktPoten, Adver, MktShare, Change, Accts

   and has $s = 428.00$ and $\bar{R}^2 = 89.4$; we refer to this model as Model 1;
2. the five-variable model that contains

   Time, MktPoten, Adver, MktShare, Change

   and has $s = 430.23$ and $\bar{R}^2 = 89.3$; we refer to this model as Model 2;
3. the seven-variable model that contains

   Time, MktPoten, Adver, MktShare, Change, Accts, Wkload

   and has $s = 435.67$ and $\bar{R}^2 = 89.0$; we refer to this model as Model 3.

To see that s can increase when we add an independent variable to a regression model, note that s increases from 428.00 to 435.67 when we add Wkload to Model 1 to form Model 3. In this case, although it can be verified that adding Wkload decreases the unexplained variation from 3,297,279.3342 to 3,226,756.2751, this decrease has not been enough to offset the change in the denominator of

$$s^2 = \frac{SSE}{n-(k+1)}$$

which decreases from 25 – 7 = 18 to 25 – 8 = 17. To see that prediction interval lengths might increase even though s decreases, consider adding Accts to Model 2 to form Model 1. This decreases s from 430.23 to 428.00. However, consider a questionable sales representative for whom Time = 85.42, MktPoten = 35,182.73, Adver = 7281.65, MktShare = 9.64, Change = .28, and Accts = 120.61. The 95 percent prediction interval given by Model 2 for sales corresponding to this combination of values of the independent variables is [3234, 5130] (see Table 2.5c) and has length $5130 - 3234 = 1896$. The 95 percent prediction interval given by Model 1 for such values can be found to be [3194, 5093] and has length 5093 − 3194 = 1899. In other words, the slight decrease in s accomplished by adding Accts to Model 2 to form Model 1 is not enough to offset the increases in $t_{[\alpha/2]}$ and the distance value (which can be shown to increase from .109 to .115), and thus the length of the prediction interval given by Model 1 increases. In addition, the extra independent variable Accts in Model 1 can be verified to have a *p-value* of .2881. Therefore, we conclude that Model 2 is better than Model 1 and is, in fact, the "best" sales territory performance model (using only linear terms).

Another quantity that can be used for comparing regression models is called the *C-statistic* (also often called the *C_k-statistic*). This criterion evaluates the *total mean squared error* of the n fitted $\hat{y}_i$ values for each possible regression model. In general, we know that if a particular regression model using k independent variables satisfies the regression assumptions, then $\mu_{\hat{y}_i}$, the mean of all possible $\hat{y}_i$ values equals

$$\mu_{y_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

the mean y_i value for the k independent variable model. If the k independent variable model has been misspecified and the *true model* describing y_i uses perhaps more independent variables that imply that the true mean y_i value is μ_{y_i}(True), we would want to consider the expected value of

$$(\hat{y}_i - \mu_{y_i}(\text{True}))^2 = [(\hat{y}_i - \mu_{\hat{y}_i}) + (\mu_{\hat{y}_i} - \mu_{y_i}(\text{True}))]^2$$

This expected value, which is called the mean squared error of the fitted value $\hat{y}_i$ can be shown to equal

$$[\mu_{\hat{y}_i} - \mu_{y_i}(\text{True})]^2 + \sigma^2_{\hat{y}_i}$$

where $[\mu_{\hat{y}_i} - \mu_{y_i}(\text{True})]^2$ represents the squared *bias* of the *k* independent variable model and $\sigma^2_{\hat{y}_i}$ is the variance of $\hat{y}_i$ for the *k* independent variable model. The *total mean squared error* for all *n* fitted $\hat{y}_i$ values is the sum of the *n* individual mean squared errors

$$\sum_{i=1}^{n} [\mu_{\hat{y}_i} - \mu_{y_i}(\text{True})]^2 + \sum_{i=1}^{n} \sigma^2_{\hat{y}_i}$$

The theoretical criterion behind the C statistic is

$$\Gamma = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n} [\mu_{\hat{y}_i} - \mu_{y_i}(\text{True})]^2 + \sum_{i=1}^{n} \sigma^2_{\hat{y}_i}\right]$$

where σ^2 is the true error variance. To estimate Γ, we first note that, if $\mathbf{x}_\mathbf{i}' = [1 \ \ x_{i1} \ \ x_{i2} \ldots x_{ik}]$, then

$$\sum_{i=1}^{n} \sigma^2_{\hat{y}_i} = \sum_{i=1}^{n} \sigma^2 [\mathbf{x}_\mathbf{i}' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_\mathbf{i}] = \sigma^2 \sum_{i=1}^{n} \mathbf{x}_\mathbf{i}' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_\mathbf{i} = (k+1)\sigma^2$$

Here, it can be proven that $\sum_{i=1}^{n} \mathbf{x}_\mathbf{i}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_\mathbf{i} = (k+1)$ for a model that uses *k* independent variables. It can also be proven that if *SSE* denotes the unexplained variation for the model using *k* independent variables, then

$$\mu_{SSE} = \sum_{i=1}^{n} [\mu_{\hat{y}_i} - \mu_{y_i}(\text{True})]^2 + [n - (k+1)]\sigma^2$$

This implies that

$$\sum_{i=1}^{n}[\mu_{\hat{y}_i} - \mu_{y_i}(\text{True})]^2 = \mu_{SSE} - [n-(k+1)]\sigma^2$$

and thus we have that

$$\begin{aligned}\Gamma &= \frac{1}{\sigma^2}\left[\mu_{SSE} - [n-(k+1)]\sigma^2 + (k+1)\sigma^2\right] \\ &= \frac{\mu_{SSE}}{\sigma^2} - [n-2(k+1)]\end{aligned}$$

If we estimate μ_{SSE} by SSE, *the unexplained variation for the model using k independent variables,* and if we estimate σ^2 by s_p^2, *the mean square error for the model using all p potential independent variables,* then the estimate of Γ for the model using k independent variables is called the C statistic and is defined by the equation:

$$C = \frac{\text{SSE}}{s_p^2} - [n-2(k+1)]$$

For example, consider the sales territory performance case. It can be verified that the mean square error for the model using all $p = 8$ independent variables is 201,621.21 and that the SSE for the model using the first $k = 5$ independent variables (Model 2 in the previous example) is 3,516,812.7933. It follows that the C-statistic for this latter model is

$$C = \frac{3{,}516{,}812{,}7933}{201{,}621.21} - [25 - 2(5+1)] = 4.4$$

Since the C-statistic for a given model is a function of the model's SSE, and since we want SSE to be small, *we want C to be small.* Although adding an unimportant independent variable to a regression model will decrease SSE, adding such a variable can increase C. This can happen when the decrease in SSE caused by the addition of the extra independent variable is not enough to offset the decrease in $n-2(k+1)$ caused by the addition of the extra independent variable (which increases k by 1). It should be noted that although adding an unimportant independent variable to a regression model can increase both s^2 and C, there is no exact relationship between s^2 and C.

Although we want C to be small, note that if a particular model using k independent variable has no bias, then $\Gamma = k+1$ and the expected value of C is close to $k+1$. Therefore, *we also wish to find a model for which the C-statistic roughly equals $k+1$*, the number of parameters in the model. *If a model has a C-statistic substantially greater than $k+1$, this model has substantial bias and is undesirable.* Thus, although we want to find a model for which C is as small as possible, if C for such a model is substantially greater than $k+1$, we may prefer to choose a different model for which C is slightly larger and more nearly equal to the number of parameters in that (different) model. *If a particular model has a small value of C and C for this model is less than $k+1$, then the model should be considered desirable.* Finally, it should be noted that for the model that includes all p potential independent variables (and thus utilizes $p+1$ parameters), it can be shown that $C = p+1$.[1]

If we examine Figure 4.4, we see that Model 2 of the previous example has the smallest *C*-statistic. The *C*-statistic for this model equals 4.4. Since $C = 4.4$ is less than $k+1=6$, the model is not biased. Therefore, this model should be considered best with respect to the *C*-statistic.

Thus far, we have considered how to find the best model using linear independent variables. In later discussions we illustrate, using the sales territory performance case, a procedure for deciding which squared and interaction terms to include in a regression model. We have found that this procedure often identifies important squared and interaction terms that are not identified by simply using scatter and residual plots.

4.2.2 Stepwise Regression and Backward Elimination

In some situations it is useful to employ an *iterative model selection procedure*, where at each step a single independent variable is added to or,

[1] That fact that $C = p+1$ for the model using all p potential independent variables is not a recommendation for choosing this model as the best model but a consequence of estimating σ^2 by s_p^2, which means that we are assuming that this model has no bias.

deleted from a regression model, and a new regression model is evaluated. We begin by discussing one such procedure—*stepwise regression.*

Stepwise regression begins by considering all of the one-independent-variable models and choosing the model for which the *p*-value related to the independent variable in the model is the smallest. If this *p*-value is less than α_{entry}, an α value for entering a variable, the independent variable is the first variable entered into the stepwise regression model and stepwise regression continues. Stepwise regression then considers the remaining independent variables not in the stepwise model and chooses the independent variable which, when paired with the first independent variable entered, has the smallest *p*-value. If this *p*-value is less than α_{entry}, the new variable is entered into the stepwise model. Moreover, the stepwise procedure checks to see if the *p*-value related to the first variable entered into the stepwise model is less than α_{stay}, an α value for allowing a variable to stay in the stepwise model. This is done because multicollinearity could have changed the *p*-value of the first variable entered into the stepwise model. The stepwise procedure continues this process and concludes when no new independent variable can be entered into the stepwise model. It is common practice to set both α_{entry} and α_{stay} equal to .05 or .10.

For example, again consider the sales representative performance data. We let x_1, x_2, x_3, x_4, x_5, x_6, x_7, and x_8 be the eight potential independent variables employed in the stepwise procedure. Figure 4.5a gives the MINITAB output of the stepwise regression employing these independent variables where both α_{entry} and α_{stay} have been set equal to .10. The stepwise procedure (1) adds Accts (x_6) on the first step; (2) adds Adver (x_3) and retains Accts on the second step; (3) adds MktPoten (x_2) and retains Accts and Adver on the third step; and (4) adds MktShare (x_4) and retains Accts, Adver, and MktPoten on the fourth step. The procedure terminates after step 4 when no more independent variables can be added. Therefore, the stepwise procedure arrives at the model that utilizes x_2, x_3, x_4, and x_6. Note that this model is not the model using x_1, x_2, x_3, x_4, and x_5 that was obtained by evaluating all possible regression models and that has the smallest C statistic of 4.4. In general, stepwise regression can miss finding the best regression model but is useful in *data mining,* where a massive number of independent variables exist and all possible regression models cannot be evaluated.

In contrast to stepwise regression, *backward elimination* is an iterative model selection procedure that begins by considering the model that contains all of the potential independent variables and then attempts to remove independent variables one at a time from this model. On each step an independent variable is removed from the model if it has the largest *p*-value of any independent variable remaining in the model and if its *p*-value is greater than α_{stay}, an α value for allowing a variable to stay in the model. Backward elimination terminates when all the *p*-values for the independent variables remaining in the model are less than α_{stay}. For example, Figure 4.5b gives the MINITAB output of a backward elimination of the sales territory performance data. Here the backward elimination uses $\alpha_{stay} = .05$, begins with the model using all eight independent variables, and removes (in order) Rating (x_8), then Wkload (x_7), then Accts (x_6), and finally Change (x_5). The procedure terminates when no independent variable remaining can be removed—that is, when no independent variable has a related *p*-value greater than $\alpha_{stay} = .05$—and arrives at a model that uses Time (x_1), MktPoten (x_2), Adver (x_3), and MktShare (x_4). Similar to stepwise regression, backward elimination has not arrived at the model using x_1, x_2, x_3, x_4, and x_5 that was obtained by evaluating all possible regression models and that has the smallest *C* statistic of 4.4. However, note that the model found in step 4 by backward elimination is the model using x_1, x_2, x_3, x_4, and x_5 and is the final model that would have been obtained by backward elimination if α_{stay} had been set at .10.

The sales territory performance example brings home two important points. First, the models obtained by backward elimination and stepwise regression depend on the choices of α_{entry} and α_{stay} (whichever is appropriate). Second, it is best not to think of these methods as "automatic model-building procedures." Rather, they should be regarded as processes that allow us to find and evaluate a variety of model choices.

4.2.3 Model Building with Squared and Interaction Terms

We have concluded that perhaps the best sales representative performance model using only linear independent variables is the model using Time, MktPoten, Adver, MktShare, and Change. We have also seen that using squared variables (which model quadratic curvature) and interaction

(a) Stepwise regression ($\alpha_{entry} = \alpha_{stay}$=.10)

```
Step                1       2        3         4
Constant       709.32   50.30  -327.23  -1441.94

Accts            21.7    19.0     15.6       9.2
T-Value          5.50    6.41     5.19      3.22
P-Value         0.000   0.000    0.000     0.004

Adver                   0.227    0.216     0.175
T-Value                  4.50     4.77      4.74
P-Value                 0.000    0.000     0.000

MktPoten                        0.0219    0.0382
T-Value                           2.53      4.79
P-Value                          0.019     0.000

MktShare                                     190
T-Value                                     3.82
P-Value                                    0.001

S                 881     650      583       454
R-Sq            56.85   77.51    82.77     90.04
R-Sq(adj)       54.97   75.47    80.31     88.05
Mallows C-P      67.6    27.2     18.4       5.4
```

(b) Backward elimination (α_{stay}=.05)

```
Step               1        2        3        4        5
Constant       -1508    -1486    -1165    -1114    -1312

Time             2.0      2.0      2.3      3.6      3.8
T-Value         1.04     1.10     1.34     3.06     3.01
P-Value        0.313    0.287    0.198    0.006    0.007

MktPoten      0.0372   0.0373   0.0383   0.0421   0.0444
T-Value         4.54     4.75     5.07     6.25     6.20
P-Value        0.000    0.000    0.000    0.000    0.000

Adver          0.151    0.152    0.141    0.129    0.152
T-Value         3.21     3.51     3.66     3.48     4.01
P-Value        0.006    0.003    0.002    0.003    0.001

MktShare         199      198      222      257      259
T-Value         2.97     3.09     4.38     6.57     6.15
P-Value        0.009    0.007    0.000    0.000    0.000

Change           291      296      285      325
T-Value         1.56     1.80     1.78     2.06
P-Value        0.139    0.090    0.093    0.053

Accts            5.6      5.6      4.4
T-Value         1.16     1.23     1.09
P-Value        0.262    0.234    0.288

WkLoad            20       20
T-Value         0.59     0.61
P-Value        0.565    0.550

Rating             8
T-Value         0.06
P-Value        0.950

S                449      436      428      430      464
R-Sq           92.20    92.20    92.03    91.50    89.60
R-Sq(adj)      88.31    88.99    89.38    89.26    87.52
Mallows C-P      9.0      7.0      5.4      4.4      6.4
```

Figure 4.5 MINITAB iterative procedures for the sales territory performance problem variables can improve a regression model. In Figure 4.6a we present the five squared variables and the ten (pairwise) interaction variables that can be formed using Time, MktPoten, Adver, MktShare, and Change. Consider having MINITAB evaluate all possible models involving these squared and interaction variables, where the five linear variables are included in each possible model. If we have MINITAB do this and find the best model of each size in terms of s, we obtain the output in Figure 4.6b. (Note that we do not include values of the C statistic on the output because it can be shown that this statistic can give misleading results when using squared and interaction variables). Examining the output, we see that the model that uses 12 squared and interaction variables (or a total of 17 variables, including the 5 linear variables) has the smallest s (174.6) of any model. If we desire a somewhat simpler model, note that s does not increase substantially until we move from a model having seven squared and interaction variables to a model having six such variables. Moreover, we might subjectively conclude that the s of 210.70 for the model using seven squared and interaction variables is not that much larger than the s of 174.6 for the model using 12 squared and interaction variables. In addition, if we fit the model having seven squared and interaction variables to the sales territory performance data, it can be verified that the p-value for each and every independent variable in this model is less than .05. Therefore, we might subjectively conclude that this model represents a good balance between having a small s, having small p-values, and being simple (having fewer independent variables). Finally, note that the s of 210.70 for this model is considerably smaller than the s of 430.23 for the model using only linear independent variables (see Table 2.5c). This smaller s yields shorter 95 percent prediction intervals, and thus more precise predictions for evaluating the performance of questionable sales representatives. For example, consider the questionable sales representative discussed in Example 2.5. The 95 percent prediction interval for the sales of this representative given by the model using only linear variables is [3234, 5130] (see Obs 26 in Table 2.5c), whereas the 95 percent prediction interval for the sales of this representative given by the seven squared and interaction variable model in Figure 4.6b is much shorter—[3979.4, 5007.8] (see Obs 26 in Figure 4.6c).

(a) The five squared variables and the ten (pairwise) interaction variables

```
SQT  = TIME*TIME                TC   = TIME*CHANGE
SQMP = MKTPOTEN*MKTPOTEN        MPA  = MKTPOTEN*ADVER
SQA  = ADVER*ADVER              MPMS = MKTPOTEN*MKTSHARE
SQMS = MKTSHARE*MKTSHARE        MPC  = MKTPOTEN*CHANGE
SQC  = CHANGE*CHANGE            AMS  = ADVER*MKTSHARE
TMP  = TIME*MKTPOTEN            AC   = ADVER*CHANGE
TA   = TIME*ADVER               MSC  = MKTSHARE*CHANGE
TMS  = TIME*MKTSHARE
```

(b) MINITAB comparisons (note: all models include the 5 linear variables)

Total Vars	Squared and interaction Vars	R-Sq	R-Sq(adj)	S	SQT	SQMP	SQA	SQMS	SQC	TMP	TA	TMS	TC	MPA	MPMS	MPC	AMS	AC	MSC
6	1	94.2	92.2	365.87								X							
7	2	95.8	94.1	318.19	X							X							
8	3	96.5	94.7	301.61	X							X		X					
9	4	97.0	95.3	285.53	X						X	X			X				
10	5	97.5	95.7	272.05	X						X	X			X		X		
11	6	98.1	96.5	244.00	X		X				X	X			X			X	
12	7	98.7	97.4	210.70	X	X					X	X			X		X	X	
13	8	99.0	97.8	193.95	X	X			X		X	X			X		X	X	
14	9	99.2	98.0	185.44	X	X		X			X	X			X		X	X	X
15	10	99.3	98.2	175.70	X	X		X			X	X			X	X	X	X	X
16	11	99.4	98.2	177.09	X	X		X		X	X	X			X	X	X	X	X
17	12	99.5	98.2	174.60	X	X		X	X	X	X	X			X	X	X	X	X
18	13	99.5	98.1	183.22	X	X	X		X	X	X	X	X		X	X	X	X	X
19	14	99.6	97.9	189.77	X	X		X	X	X	X	X	X	X	X	X	X	X	X
20	15	99.6	97.4	210.78	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

(c) Predicted sales performance using the seven squared and interaction variable model

Obs	Dep Var SALES	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict	Upper95% Predict
26	.	4493.6	106.306	4262.0	4725.2	3979.4	5007.8

***Figure 4.6** Sales territory performance model building using squared and interaction variables*

4.3 Step 3: Diagnosing and Remedying Violations of Regression Assumptions 1, 2, and 3

4.3.1 *Residual Analysis*

As discussed in Section 2.3, four regression assumptions must at least approximately hold if statistical inferences made using the linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

are to be valid. The first three regression assumptions say that, at any given combination of values of the independent variables $x_1, x_2, \ldots, x_k$, the population of error terms that could potentially occur

1. has mean zero;
2. has a constant variance σ^2 (a variance that does not depend upon $x_1, x_2, \ldots, x_k$);
3. is normally distributed.

The fourth regression assumption says that any one value of the error term is statistically independent of any other value of the error term. To assess whether the regression assumptions hold in a particular situation, note that the regression model implies that the error term ε is given by the equation $\varepsilon = y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$. The point estimate of this error term is the residual

$$e = y - \hat{y} = y - (b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k)$$

where $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$ is the predicted value of the dependent variable y. Therefore, since the n residuals are the point estimates of the n error terms in the regression analysis, we can use the residuals to check the validity of the regression assumptions about the error terms. One useful way to analyze residuals is to plot them versus various criteria. The resulting plots are called *residual plots.* To construct a residual plot, we compute the residual for each observed y value. The calculated residuals are then plotted versus some criterion. To validate the regression assumptions, we make residual plots against (1) values of each of the independent variables $x_1, x_2, \ldots, x_k$; (2) values of $\hat{y}$, the predicted value of the dependent variable; and (3) the time order in which the data have been observed (if the regression data are time series data).

Example 4.2

Quality Home Improvement Center (QHIC) operates five stores in a large metropolitan area. The marketing department at QHIC wishes to study the relationship between x, home value (in thousands of dollars), and y, yearly expenditure on home upkeep (in dollars). A random sample of 40 homeowners is taken and survey participants are asked to estimate their expenditures during the previous year on the types of home upkeep products and services offered by QHIC. Public records of the county auditor are used to obtain the previous year's assessed values of the homeowner's homes. Figure 4.7 gives the resulting values of x (see value) and y (see upkeep) and a scatter plot of these values. The least squares point estimates of the y-intercept β_0 and the slope β_1 of the simple linear regression model describing the QHIC data are $b_0 = -348.3921$ and $b_1 = 7.2583$. Moreover, Figure 4.7 presents the predicted home upkeep expenditures and residuals that are given by the regression model. Here each residual is computed as

$$e = y - \hat{y} = y - (b_0 + b_1 x) = y - (-348.3921 + 7.2583x)$$

Home	Value	Unkeep	Predicted	Residual
1	237.00	1,412.080	1,371.816	40.264
2	153.08	797.200	762.703	34.497
3	184.86	872.480	993.371	–120.891
4	222.06	1,003.420	1,263.378	–259.958
5	160.68	852.900	817.866	35.034
6	99.68	288.480	375.112	–86.632
7	229.04	1,288.460	1,314.041	–25.581
8	101.78	423.080	390.354	32.726
9	257.86	1,351.740	1.523.224	–171.484
10	96.28	378.040	350.434	27.606
11	171.00	918.080	892.771	25.309
12	231.02	1,627.240	1,328.412	298.828

Figure 4.7 The QHIC data and residuals, and a scatter plot (Continued)

13	228.32	1,204.760	1308.815	–104.055
14	205.90	857.040	1,146.084	–289.044
15	185.72	775.000	999.613	–224.613
16	168.78	869.260	876.658	–7.398
17	247.06	1,396.000	1,444,835	–48.835
18	155.54	711.500	780.558	–69.056
19	224.20	1,475.180	1,278.911	196.269
20	202.04	1,413.320	1.118.068	295.252
21	153.04	849.140	762.413	86.727
22	232.18	1,313.840	1.336.832	–22.992
23	125.44	602.060	562.085	39.975
24	169.82	642.140	884.206	–242.066
25	177.28	1.038.800	938.353	100.447
26	162.82	697.000	833.398	–136.398
27	120.44	324.340	525.793	–201.453
28	191.10	965.100	1,038.662	–73.562
29	158.78	920.140	804.075	116.065
30	178.50	950.900	947.208	3.692
31	272.20	1,670.320	1,627.307	43.013
32	48.90	125.400	6.537	118.863
33	104.56	479.780	410.532	69.248
34	286.18	2,010.640	1,728.778	281.862
35	83.72	368.360	259.270	109.090
36	86.20	425.600	277.270	148.330
37	133.58	626.900	621.167	5.733
38	212.86	1,316.940	1,196.602	120.338
39	122.02	390.160	537.261	–147.101
40	198.02	1,090.840	1,088.889	1.951

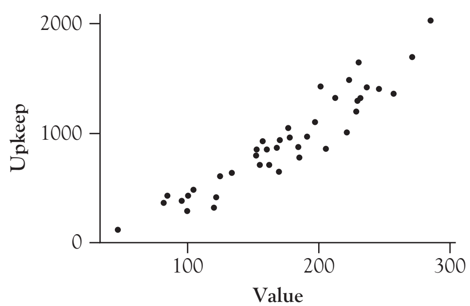


Figure 4.7 The QHIC data and residuals, and a scatter plot

For instance, for the first home, when $y = 1{,}412.08$ and $x = 237.00$, the residual is

$$\begin{aligned} e &= 1{,}412.08 - (-348.3921 + 7.2583(237)) \\ &= 1{,}412.08 - 1{,}371.816 = 40.264 \end{aligned}$$

The MINITAB output in Figure 4.8a and 4.8b gives plots of the residuals for the QHIC simple linear regression model against values of x (value) and $\hat{y}$ (predicted upkeep). To understand how these plots are constructed, recall that for the first home $y = 1{,}412.08$, $x = 237.00$, $\hat{y} = 1{,}371.816$, and the residual is 40.264. It follows that the point plotted in Figure 4.8a corresponding to the first home has a horizontal axis coordinate of the x value 237.00 and a vertical axis coordinate of the residual 40.264. It also follows that the point plotted in Figure 4.8b corresponding to the first home has a horizontal axis coordinate of the $\hat{y}$ value 1,371.816, and a vertical axis coordinate of the residual 40.264. Finally, note that the QHIC data are cross-sectional data, not time series data. Therefore, we cannot make a residual plot versus time.

4.3.2 The Constant Variance Assumption

To check the validity of the constant variance assumption, we examine plots of the residuals against values of x, $\hat{y}$ and time (if the regression data are time series data). When we look at these plots, the pattern of the residuals' fluctuation around 0 tells us about the validity of the constant variance assumption. A residual plot that *fans out* (as in Figure 4.9a) suggests that the error terms are becoming more spread out as the horizontal plot value increases and that the constant variance assumption is violated. Here we would say that an *increasing error variance* exists. A residual plot that *funnels in* (as in Figure 4.9b) suggests that the spread of the error terms is decreasing as the horizontal plot value increases and that again the constant variance assumption is violated. In this case we would say that a *decreasing error variance* exists. A residual plot with a horizontal band appearance (as in Figure 4.9c) suggests that the spread of the error terms around 0 is not changing much as the horizontal plot value increases. Such a plot tells us that the constant variance assumption (approximately) holds.

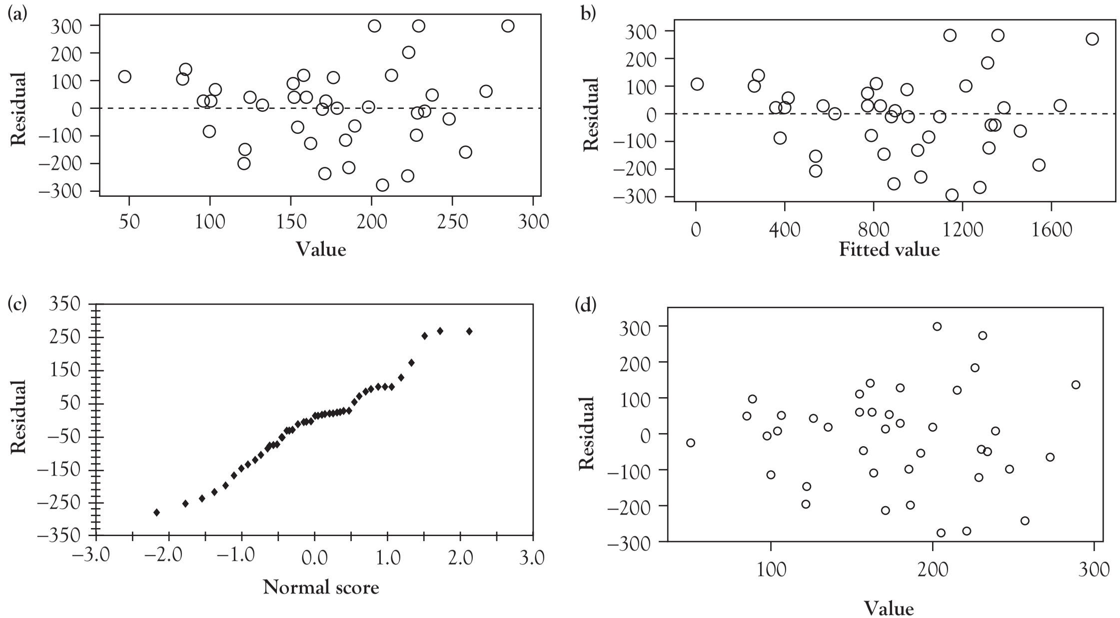


Figure 4.8 Residual analysis for QHIC data models (a) Simple linear regression model residual plot versus x(value) (b) Simple linear regression model residual plot versus $\hat{y}$ (predicted upkeep) (c) Simple linear regression model normal plot (d) Quadratic regression model residual plot versus x(value)

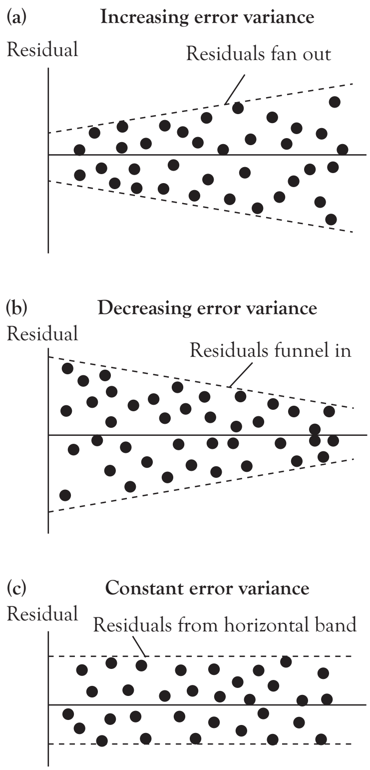


Figure 4.9 Residual plots and the constant variance assumption

As an example, consider the QHIC case and the residual plot in Figure 4.8a. This plot appears to fan out as x increases, indicating that the spread of the error terms is increasing as x increases. That is, an increasing error variance exists. This is equivalent to saying that the variance of the population of potential yearly upkeep expenditures for houses worth x (thousand dollars) appears to increase as x increases. The reason is that the model $y = \beta_0 + \beta_1 x + \varepsilon$ says that the variation of y is the same as the variation of ε. For example, the variance of the population of potential yearly upkeep expenditures for houses worth \$200,000 would be larger than the variance of the population of potential yearly upkeep expenditures for houses worth \$100,000. Increasing variance makes some intuitive sense because people with more expensive homes generally have more discretionary income. These people can choose to spend either a substantial amount or a much smaller amount on home upkeep, thus causing a relatively large variation in upkeep expenditures.

Another residual plot showing the increasing error variance in the QHIC case is Figure 4.8b. This plot tells us that the residuals appear to fan out as $\hat{y}$ (predicted y) increases, which is logical because $\hat{y}$ is an increasing function of x. Also, note that the original scatter plot of y versus x in Figure 4.7 shows the increasing error variance—the y values appear to fan out as x increases. In fact, one might ask why we need to consider residual plots when we can simply look at scatter plots. One answer is that, in general, because of possible differences in scaling between residual plots and scatter plots, one of these types of plots might be more informative in a particular situation. Therefore, we should always consider both types of plots.

When the constant variance assumption is violated, we cannot use the regression formulas presented in this book to make statistical inferences. Later in this section we will learn how to remedy violations of the constant variance assumption.

4.3.3 The Assumption of Correct Functional Form

Consider the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$. If for any value of x in this model the population of potential error terms has a mean of 0 (regression assumption 1), then the population of potential y values has a mean of $\mu_{y|x} = \beta_0 + \beta_1 x$. But this is the same as saying that for different values of x the corresponding values of $\mu_{y|x}$ lie on a straight line (rather than, for example, a curve). Thus for the simple linear regression model we call regression assumption 1 the assumption of *correct functional form*. If we mistakenly use a simple linear regression model when the true relationship between y and x is curved, the residual plot will have a curved appearance. For example, the scatter plot of upkeep expenditure, y, versus home value, x, in Figure 4.7 has either a straight-line or slightly curved appearance. We used a simple linear regression model to describe the relationship between y and x, but note that there is a *dip* or slightly curved appearance, in the upper left portion of the residual plots against x and $\hat{y}$ in Figure 4.8a and 4.8b. Therefore, both the scatter plot and residual plots indicate that there might be a slightly curved relationship between y and x. One remedy for the simple linear regression model's violation of the correct functional form assumption is to fit the quadratic regression model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ to the QHIC data. When we do this and plot the model's residuals versus x (value), we obtain the residual plot in Figure 4.8d. The fact that this residual plot does not have any curved appearance implies that the quadratic regression model has remedied the violation of the correct functional form assumption. However, note that the residuals fan out as x increases, indicating that the constant variance assumption is still being violated.

If we generalize the above ideas to the multiple linear regression model, we can say that if a residual plot against a particular independent variable x_j or against the predicted value of the dependent variable $\hat{y}$ has a curved appearance, then this indicates a violation of regression assumption 1 and says that the multiple linear regression model does not have the correct functional form. Specifically, the multiple linear regression model may need additional squared or interaction variables, or both. To give an illustration of using residual plots in multiple linear regression, consider the sales territory performance data in Table 2.5a and recall that Table 2.5c gives the SAS output of a regression analysis of these data using the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

The least squares point estimates on the output give the prediction equation

$$\hat{y} = -1113.7879 + 3.6121x_1 + .0421x_2 + .1289x_3 + 256.9555x_4 + 324.5335x_5$$

Using this prediction equation, we can calculate predicted sales values and residuals for the 25 sales representatives. For example, observation 10 in this data set corresponds to a sales representative for whom $x_1 = 105.69$, $x_2 = 42{,}053.24$, $x_3 = 5673.11$, $x_4 = 8.85$, and $x_5 = .31$. If we insert these values into the prediction equation, we obtain a predicted sales value of $\hat{y}_{10} = 4143.597$. Since the actual sales for the sales representative are $y_{10} = 4876.370$, the residual e_{10} equals the difference between

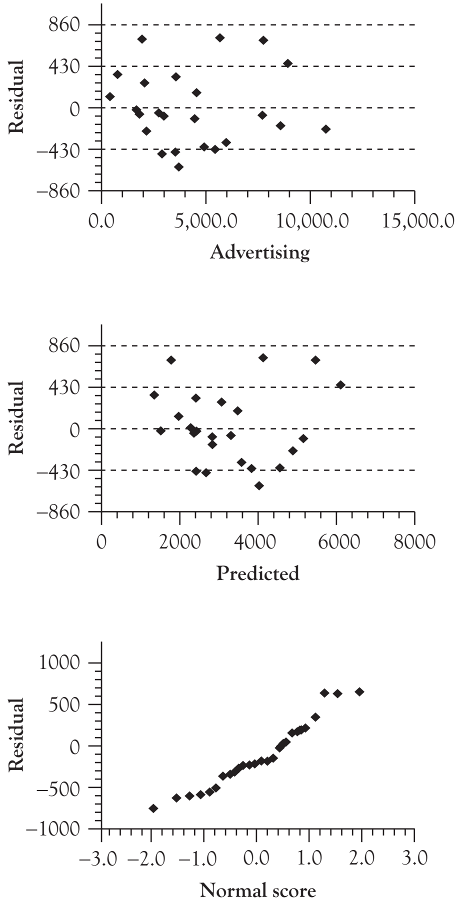


Figure 4.10 Sales territory performance residual analysis

y_{10} = 4876.370 and $\hat{y}_{10}$ = 4143.597, which is 732.773. A plot of all 25 residuals versus each of the independent variables x_1, x_2, x_3, x_4, and x_5 can be verified to have a horizontal band appearance (the plot of the residuals versus x_3, advertising, is shown in Figure 4.10), as does the plot of these residuals versus predicted sales (again, see Figure 4.10). Therefore, the constant variance and correct functional form assumptions do not appear to be violated. Recall from Section 4.2, however, that adding seven squared and interaction variables (see Figure 4.6) to the above model (that uses only the five linear terms) gives a model with a much smaller s that yields more accurate predictions. This illustrates that we need to use all of the model building and model diagnostic procedures in this book to find an appropriate final regression model.

4.3.4 The Normality Assumption

If the normality assumption holds, a histogram or stem-and-leaf display of the residuals should look reasonably bell-shaped and reasonably symmetric about 0, and a normal plot of the residuals should have a straight line appearance. To construct a normal plot, we first arrange the residuals in order from smallest to largest. Letting the ordered residuals be denoted as $e_{(1)}, e_{(2)}, \ldots, e_{(n)}$, we denote the ith residual in the ordered listing as $e_{(i)}$. We plot $e_{(i)}$ on the vertical axis against a normal point $z_{(i)}$ on the horizontal axis. Here $z_{(i)}$ is defined to be the point on the horizontal axis under the standard normal curve so that the area under this curve to the left of $z_{(i)}$ is $(3i-1)/(3n+1)$. For example, recall in the QHIC case that there are $n=$ 40 residuals in Figure 4.7. It follows that, when $i=1$, $(3i-1)/(3n+1) = [3(1)-1]/[3(40)+1] = .0165$. Using Table A3 to look-up the normal point $z_{(i)}$, which has a standard normal curve area to its left of .0165 and thus an area of $.5-.0165=.4835$ between itself and 0, we find that $z_{(1)}=-2.13$. Because the smallest residual in Figure 4.7 is -289.044, the first point plotted is $e_{(1)}=-289.044$ on the vertical axis versus $z_{(1)}=-2.13$ on the horizontal axis. Plotting the other ordered residuals $e_{(2)}, e_{(3)}, \ldots, e_{(40)}$ against their corresponding normal points in the same way, we obtain the normal plot in Figure 4.8c. In a similar fashion, if we use the residuals for the sales territory performance model $y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\varepsilon$, we obtain the normal plot in Figure 4.10. Both normal plots essentially have a straight line appearance. Therefore, there appears to be no violation of the normality assumption in either case.

It is important to realize that violations of the constant variance and correct functional form assumptions can often cause a histogram and/or stem-and-leaf display of the residuals to look nonnormal and can cause the normal plot to have a strongly curved appearance. Because of this, it is usually a good idea to use residual plots to check for nonconstant variance and incorrect functional form before making any final conclusions about the normality assumption.

4.3.5 *Handling Unequal Variances, and Weighted Least Squares*

Consider the linear regression model

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

If the variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ of the error terms are unequal and known, then the variances can be equalized by using the transformed model

$$\frac{y_i}{\sigma_i} = \beta_0\left(\frac{1}{\sigma_i}\right) + \beta_1\left(\frac{x_{i1}}{\sigma_i}\right) + \beta_2\left(\frac{x_{i2}}{\sigma_i}\right) + \ldots + \beta_k\left(\frac{x_{ik}}{\sigma_i}\right) + \eta_i$$

where $\eta_i = \varepsilon_i / \sigma_i$. This transformed model has the same parameters as the original model and also satisfies the constant variance assumption. This is because the properties of the variance tell us that the variance of the error term η_i for the transformed model is $\sigma^2_{\eta_i} = \sigma^2_{(\varepsilon_i/\sigma_i)} = (1/\sigma_i)^2\, \sigma^2_{\varepsilon_i} = \sigma_i^2 / \sigma_i^2 = 1$. The least squares point estimates $b_0, b_1, b_2, \ldots, b_k$ of the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ of the transformed model are calculated by using the equation $\mathbf{b} = \left(\mathbf{X}'_* \mathbf{X}_*\right)^{-1} \mathbf{X}'_* \mathbf{y}_*$, where

$$\mathbf{y}_* = \begin{bmatrix} \frac{y_1}{\sigma_1} \\ \frac{y_2}{\sigma_2} \\ \vdots \\ \frac{y_n}{\sigma_n} \end{bmatrix} \quad \text{and} \quad \mathbf{X}_* = \begin{bmatrix} \frac{1}{\sigma_1} & \frac{x_{11}}{\sigma_1} & \frac{x_{12}}{\sigma_1} & \cdots & \frac{x_{1k}}{\sigma_1} \\ \frac{1}{\sigma_2} & \frac{x_{21}}{\sigma_2} & \frac{x_{22}}{\sigma_2} & \cdots & \frac{x_{2k}}{\sigma_2} \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{1}{\sigma_n} & \frac{x_{n1}}{\sigma_n} & \frac{x_{n2}}{\sigma_n} & \cdots & \frac{x_{nk}}{\sigma_n} \end{bmatrix}$$

Letting $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik}$, the least squares point estimates $b_0, b_1, b_2, \ldots, b_k$ of the parameters of the transformed model minimize the following sum of squared residuals

$$\begin{aligned}SSE_* &= \sum_{i=1}^{n}(y_i / \sigma_i - \hat{y}_i / \sigma_i)^2 \\ &= \sum_{i=1}^{n}(1/\sigma_i)^2[y_i - \hat{y}_i]^2 \\ &= \sum_{i=1}^{n}(1/\sigma_i)^2[y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik})]^2\end{aligned}$$

Now, if we consider the original, untransformed model

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

the estimates $b_0(w), b_1(w), b_2(w), \ldots, b_k(w)$ of the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ that minimize

$$SSE_{\mathrm{W}} = \sum_{i=1}^{n} w_i [y_i - \{b_0(w) + b_1(w)x_{i1} + b_2(w)x_{i2} + \ldots + b_k(w)x_{ik}\}]^2$$

are called the *weighted least squares point estimates* of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$. Comparing the expression for SSE_* with the expression for SSE_{W}, we see that the (ordinary) least squares point estimates $b_0, b_1, b_2, \ldots, b_k$ of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ using the transformed model equal the weighted least squares point estimates $b_0(w), b_1(w), b_2(w), \ldots, b_k(w)$ of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ using the original model, if we let the weight w_{i} *equal* $(1/\sigma_i)^2$ *for* $i = 1, 2, \ldots, n$. This is important because it gives us two equivalent ways to remedy violations of the constant variance assumption and make appropriate statistical inferences:

1. Use the transformed model to calculate the ordinary least squares point estimates and make statistical inferences based on these point estimates.
2. Use the original, untransformed model to calculate the weighted least squares point estimates, where $w_{\mathrm{i}} = (1/\sigma_i)^2$, and make statistical inferences based on these point estimates.

With respect to (2), statisticians have shown that the formula for the weighted least squares point estimates is

$$\begin{bmatrix} b_0(w) \\ b_1(w) \\ b_2(w) \\ \vdots \\ b_k(w) \end{bmatrix} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$$

Here, **y** and **X** are defined in Section 2.2 for the original, untransformed model, and

$$\mathbf{W} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}$$

In addition, formulas exist for the hypothesis test statistics, confidence intervals, and prediction intervals based on the weighted least squares point estimates. We will not present these formulas here, but sophisticated statistical software systems such as SAS carry out weighted least squares regression analysis. If one is using a statistical software system that does not do this analysis, the transformed model can be used.

We will demonstrate using both the transformed model approach and the weighted least squares approach, but first note that we almost never know the true values of the error term variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$. However, we can sometimes use the following three step procedure to estimate these variances and remedy a violation of the constant variance assumption:

Step 1: Fit the original, untransformed regression model using ordinary least squares and assuming equal variances.

Step 2: Plot the residuals from the fitted regression model against each independent variable. If the residual plot against increasing values of the independent variable x_j fans out, plot the absolute values of the residuals versus the x_{ij} values. If the plot shows a straight line relationship, fit the simple linear regression model $|e_i| = \beta_0' + \beta_1' x_{ij} + \varepsilon_i'$ to the absolute values of the residuals and predict the absolute value of the ith residual to be

$$pabe_i = b_0' + b_1' x_{ij}$$

Step 3: Use $pabe_i$ as the point estimate of σ_{i} and use ordinary least squares to fit the transformed model

$$\frac{y_i}{pabe_i} = \beta_0\left(\frac{1}{pabe_i}\right) + \beta_1\left(\frac{x_{i1}}{pabe_i}\right) + \beta_2\left(\frac{x_{i2}}{pabe_i}\right) + \ldots + \beta_k\left(\frac{x_{ik}}{pabe_i}\right) + \eta_i$$

or, equivalently, use weighted least squares to fit the original, untransformed model, where $w_i = \left(1 / pabe_i\right)^2$.

Note that if in step 2 the plot of the absolute values of the residuals versus the x_{ij} values did not have a straight line appearance, but a plot of the squared residuals versus the x_{ij} values did have a straight line appearance, we would fit the simple linear regression model $e_i^2 = \beta_0' + \beta_1' x_{ij} + \varepsilon_i$ to the squared residuals and predict the squared value of the *i*th residual to be $psqe_i = b_0' + b_1' x_{ij}$. In this case we estimate σ_i^2 by $psqe_i$ and σ_i by $\sqrt{psqe_i}$, which implies that we should specify a transformed regression model by dividing all terms in the original regression model by $\sqrt{psqe_i}$. Alternatively, we can fit the original regression model using weighted least squares where $w_i = 1 / psqe_i$.

For example, recall that Figure 4.8d shows that when we fit the quadratic regression model $y = \beta_0 + \beta_1 x \ + \beta_2 x^2 + \varepsilon$ to the QHIC data, the model's residuals fan out as x increases. A plot of the absolute values of the model's residuals versus the x values can be verified to have a straight line appearance. Figure 4.11 shows that when we use the simple linear regression model to relate the model's absolute residuals to x, we obtain the equation $pabe_i = 22.23055 + \ .49067 x_i$ for predicting the absolute values of the model's residuals. For example, because the value x of the first home in Figure 4.7 is 237, the prediction of the absolute value of the quadratic model's residual for home 1 is $pabe_1 = 22.23055 + .40967(237) = 138.519$. This and the other predicted absolute residuals are shown in Figure 4.11. Figures 4.12 and 4.13 are the partial SAS outputs that are obtained if we use ordinary least squares to fit the transformed model

$$\frac{y_i}{pabe_i} = \beta_0\left(\frac{1}{pabe_i}\right) + \beta_1\left(\frac{x_i}{pabe_i}\right) + \beta_2\left(\frac{x_i^2}{pabe_i}\right) + \eta_i$$

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	22.23055	41.72626	0.53	0.5973
Value	1	0.49067	0.22774	2.15	0.0376

Obs	$pabe_i$	Obs	$pabe_i$	Obs	$pabe_i$	Obs	$pabe_i$
1	138.519	11	106.135	21	97.323	31	155.791
2	97.342	12	135.585	22	136.154	32	46.224
3	112.936	13	134.260	23	83.780	33	73.535
4	131.189	14	123.260	24	105.556	34	162.651
5	101.071	15	113.358	25	109.217	35	63.309
6	71.141	16	105.046	26	102.122	36	64.526
7	134.614	17	143.456	27	81.327	37	87.774
8	72.171	18	98.549	28	115.998	38	126.675
9	148.755	19	132.239	29	100.139	39	82.102
10	69.472	20	121.366	30	109.815	40	119.393
						41	130.178

***Figure 4.11** Partial SAS output of a simple linear regression analysis using the model $|e_i| = \beta_0' + \beta_1' x_{ij} + \varepsilon_i'$, and the predictions $pabe_i = 22.23055 + .49067x_i$ of the absolute values of the residuals*

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
inv_pabe	1	-41.63220	107.18869	-0.39	0.6999
Value_star	1	3.23363	1.55100	2.08	0.0440
Val_Sq_star	1	0.01178	0.00510	2.31	0.0267

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict	
41	.	9.5252	0.2570	9.0045	10.0459	6.9211	12.1293

***Figure 4.12** Partial SAS output when using ordinary least squares to fit the transformed model $y_i / pabe_i = \beta_0 (1/pabe_i) + \beta_1 (x_i / pabe_i) + \beta_2 (x_i^2 / pabe_i) + \eta_i$*

and weighted least squares, where $w_i = (1/pabe_i)^2$, to fit the original model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$ to the QHIC data. A plot of the residuals versus the x_i values for the transformed model has a horizontal band

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-41.63220	107.18869	-0.39	0.6999
Value	1	3.23363	1.55100	2.08	0.0440
Val_Sq	1	0.01178	0.00510	2.31	0.0267

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict	
41	.	1240	33.4562	1172	1308	900.9750	1579

Figure 4.13 Partial SAS output when using weighted least squares to fit the original model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$, ***where*** $w_i = (1/pabe_i)^2$

appearance, showing that the constant variance assumption approximately holds for the transformed model.

Suppose that QHIC has decided to send an advertising brochure to a home if the point prediction of y_0, the yearly upkeep expenditure for the home, is at least \$500. QHIC will also send a special, more elaborate advertising brochure to a home if its value makes QHIC 95 percent confident that μ_0, the mean yearly upkeep expenditure for all homes having this value, is at least \$1,000. Consider a home with a value of \$220,000. That is, the x value for this home is $x_0 = 220$. The predicted absolute residual for a home for which $x_0 = 220$ is $pabe_0 = 22.2305 + .49067(220) = 130.178$, as shown in Figure 4.11. Therefore, the point prediction of $y_0/130.178$ and point estimate of $\mu_0/130.178$ obtained from the transformed model is

$$\begin{aligned}\frac{\hat{y}_0}{130.178} &= b_0\left(\frac{1}{130.178}\right) + b_1\left(\frac{x_0}{130.178}\right) + b_2\left(\frac{x_0^2}{130.178}\right)\\ &= -41.63220\left(\frac{1}{130.178}\right) + 3.23363\left(\frac{220}{130.178}\right)\\ &\quad +.01178\left[\frac{(220)^2}{130.178}\right]\\ &= 9.5252\end{aligned}$$

Figure 4.12 shows that $\hat{y}_0/130.178 = 9.5252$. It follows that $\hat{y}_0 = 9.5252(130.178) = 1240$, which is shown in Figure 4.13 and can be obtained directly from the weighted least squares prediction equation as follows:

$$\hat{y}_0 = -41.63220 + 3.23363(220) + .01178(220)^2$$
$$= 1240$$

Because the point prediction $\hat{y}_0 = \$1240$ of the home's yearly upkeep expenditure is at least \$500, QHIC will send the home an advertising brochure. Figure 4.12 also shows that a 95 percent confidence interval for $\mu_0 / 130.178$ is $[9.0045,\ 10.0459]$. It follows that a 95 percent confidence internal for μ_0 is $[9.0045(130.178),\ 10.0459(130.178)] = [\$1172,\ \$1308]$, which is shown on the weighted least squares output in Figure 4.13. Because this interval says that QHIC is 95 percent confident that μ_0 is at least \$1172, QHIC is more than 95 percent confident that μ_0 is at least \$1000. Therefore, a home with a value of \$220,000 will also be sent the special, more elaborate advertising brochure.

4.3.6 *Fractional Power Transformations of the Dependent Variable*

To conclude this section, note that if a data or residual plot indicates that the error variance of a regression model increases as an independent variable or the predicted value of the dependent variable increases, then another way that is sometimes successful in remedying the situation involves transforming the dependent variable by taking each y value to a fractional power. As an example, we might use a transformation in which we take the square root (or one-half power) of each y value. Letting y^* denote the value obtained when the transformation is applied to y, we would write the *square root transformation* as $y^* = y^{.5}$. Another commonly used transformation is the *quartic root transformation*. Here we take the y value to the one-fourth power. That is, $y^* = y^{.25}$.

If we consider a transformation that takes each y value to a fractional power (such as .5, .25, or the like), as the power approaches 0, the transformed value y^* approaches the natural logarithm of y (commonly written $ln\ y$). In fact, we sometimes use the *logarithmic transformation* $y^* = ln\ y$, which takes the natural logarithm of each y value.

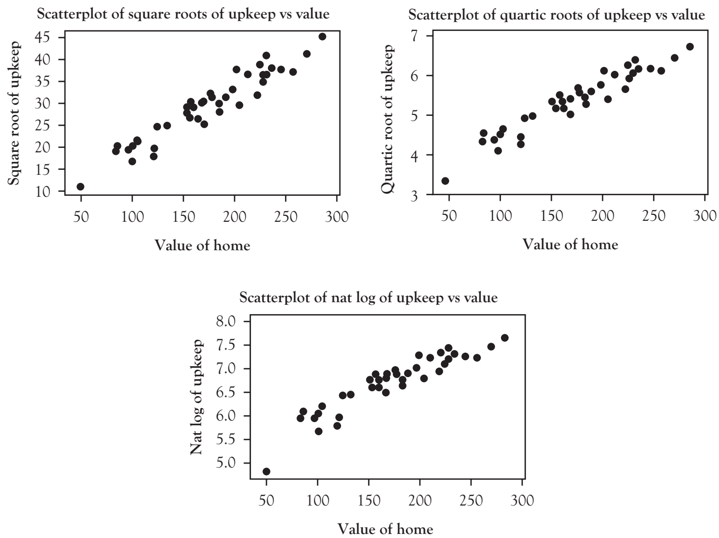


Figure 4.14 Fractional power transformations of the upkeep expenditures

For example, consider the QHIC upkeep expenditures in Figure 4.7. In Figure 4.14 we show the plots that result when we take the square root, quartic root, and natural logarithmic transformations of the upkeep expenditures and plot the transformed values versus the home values. To interpret these plots, note that when we take a fractional power (including the natural logarithm) of the dependent variable, the transformation not only tends to equalize the error variance but also tends to straighten out certain types of nonlinear data plots. Specifically, if a data plot indicates that the dependent variable is increasing at an increasing rate as the independent variable increases (this is true of the QHIC data plot in Figure 4.7), then a fractional power transformation tends to straighten out the data plot. A factional power transformation can also help to remedy a violation of the normality assumption. Because we cannot know which fractional power to use before we actually take the transformation, we recommend taking all of the square root, quartic root, and natural logarithm transformations and seeing which one best equalizes the error variance and (possibly) straightens out a nonlinear data plot. This is what we have done in Figure 4.14, and examining this figure, it seems that the square root transformation best equalizes the error variance and straightens out the curved data plot in Figure 4.7. Note that the natural logarithm transformation seems to overtransform the data—the error variance tends to decrease as the home value increases and the data plot seems to bend down. The plot of the quartic roots indicates that the quartic root transformation also seems to overtransform the data (but not by as much as the logarithmic transformation). In general, as the fractional power gets smaller, the transformation gets stronger. Different fractional powers are best in different situations.

Because the plot in Figure 4.14 of the square roots of the upkeep expenditures versus the home values has a straight-line appearance, we consider the model $y^* = \beta_0 + \beta_1 x + \varepsilon$, where $y^* = y^{.5}$. If we fit this model to the QHIC data, we find that the least squares point estimates of β_0 and β_1 are $b_0 = 7.201$ and $b_1 = .127047$. Moreover, a plot of the transformed model's residuals versus x has a horizontal band appearance. Consider a home worth \$220,000. Using the least squares point estimates, a point prediction of y^* for such a home is $\hat{y}^* = 7.201 + .127047(220) = 35.151$. This point prediction is given on the MINITAB output in Figure 4.15, as is the 95 percent prediction interval for y^*, which is [30.348, 39.954]. It follows that a point prediction of the upkeep expenditure for a home worth \$220,000 is $(35.151)^2 = \$1{,}235.59$ and that a 95 percent prediction interval for this upkeep expenditure is $[(30.348)^2, (39.954)^2] = [\$921.00, \$1596.32]$. Recall that QHIC will send an advertising brochure to any home that has a predicted upkeep expenditure of at least \$500. It follows that a home worth \$220,000 will be sent an advertising brochure. This is because the predicted yearly upkeep expenditure for such a home is (as just calculated) \$1,235.59. Also, recall that QHIC will send a special, more elaborate advertising brochure to a home if its value makes QHIC 95 percent confident that μ_0, the mean yearly upkeep expenditure for all homes having this value, is at least \$1000. We were able to find a 95 percent confidence interval for μ_0 using the transformed quadratic regression model of the previous subsection. However, although Figure 4.15 gives a 95 percent confidence interval for the mean of the square roots of the upkeep expenditures, the mean of these square roots is not equal to $\sqrt{\mu_0}$, and thus we cannot square both ends of the confidence interval in Figure 4.15 to find a 95 percent confidence interval for μ_0. This is a

```
Predicted Values for New Observations
New Obs     Fit   SE Fit       95% CI            95% PI
      1  35.151   0.474   (34.191, 36.111)  (30.348, 39.954)
```

Figure 4.15 MINITAB output of prediction using the model $y^* = \beta_0 + \beta_1 x + \varepsilon$ where $y^* = y^{.5}$

disadvantage of using a fractional power transformation. However, if we are mainly interested in predicting an individual value of the dependent variable (as will be true in the time series prediction examples of the next subsection), then the fractional power transformation technique can be very successful.

4.3.7 A Lack of Fit Test, and an Introduction to Nonlinear Regression

When a beam of light is passed through a chemical solution, a certain fraction of the light will be either absorbed or reflected and the remainder of the light will be transmitted. Graybill and Iyer (1994) give $n = 12$ observations resulting from an experiment where the concentration, x, of a chemical is fixed at 12 values and corresponding optical readings of the amount, y, of transmitted light are made. The 12 fixed chemical concentration x values are 0, 0, 1, 1, 2, 2, 3, 3, 4, 4, 5, 5 6, and 6, and the corresponding optical reading y values are 2.86, 2.64, 1.57, 1.24, .45, 1.02, .65, .18, .15, .01, .04, and .36. The upper plot of y versus x in Figure 4.16 implies that $\mu_{y|x}$, the mean amount of transmitted light corresponding to chemical concentration x, steadily decreases at a slower and slower rate as x increases and ultimately approaches a constant value. Hence, it does not seem appropriate to describe the data by using the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$. However, noting that the data consists of a set of repeated y values for each x value, we can use the data and this model to demonstrate what is called a *lack of fit test*.

In general, the *lack of fit test* tests the hypothesis H_0 that the functional form of a particular regression model is correct versus the alternative hypothesis H_a that the functional form of the model is not correct. To carry out the test we start by calculating SS_{PE}, the *sum of squares due to pure error*. To find SS_{PE}, we find the *deviation* (Dev) of each y value from the appropriate *set mean* of y values, square each deviation, and sum the squared deviations. The appropriate *set mean* of y values for a particular y value is the mean of all of the y values that correspond to the same x value as does the particular y value. For the light data, the optical readings corresponding to the x values 0 and 0 are 2.86 and 2.64, which have a set mean of $(2.86+2.64)/2=2.75$ and associated deviations of $2.86-2.75=.11$ and $2.64-2.75=-.11$. The optical readings corresponding to the x values 1 and 1 are 1.57 and 1.24, which have a set mean of 1.405 and associated deviations of $1.57-1.405=.165$ and $1.24-1.405=-.165$. The optical readings corresponding to the x values 2 and 2 are .45 and 1.02, which have a set mean of .735 and associated deviations of −.285 and .285. The optical readings corresponding to the x values 3 and 3 are .65 and .18, which have a set mean of .415 and associated deviations of .235 and −.235. The optical readings corresponding to the x values 4 and 4 are .15 and .01, which have a set mean of .08 and associated deviations of .07 and −.07. The optical readings corresponding to the x values 5 and 5 are .04 and .36, which have a set mean of .20 and associated deviations of −.16 and .16. The sum of squares due to pure error for the light data, SS_{PE}, is the sum of the squares of the 12 deviations that we have calculated and equals .4126. Also, if we fit the simple linear regression model to the data, we find that SSE, the sum of squared residuals, is 2.3050. In general to perform a lack of fit test, we let the symbol m denote the number of *distinct* x values for which there is at least one y value ($m=6$ for the light data), and we let n denote the total number of observations ($n=12$ for the light data). We then calculate the following *lack of fit statistic*, the value of which we show for the light data:

$$\begin{aligned} F(LF) &= \frac{SS_{LF}/(m-2)}{SS_{PE}/(n-m)} = \frac{(SSE-SS_{PE})/(m-2)}{SS_{PE}/(n-m)} \\ &= \frac{(2.3050-.4126)/(6-2)}{.4126/(12-6)} = \frac{1.8924/4}{.4126/6} \\ &= 6.88 \end{aligned}$$

Because $F(LF)=6.88$ is greater than $F_{[.05]}=4.53$, based on $m-2=6-2=4$ numerator and $n-m=12-6=6$ denominator degrees of freedom, we reject the null hypothesis H_0 that the functional form of the simple linear regression model is correct. Note that to test the null hypothesis that the functional form of a *multiple* regression model is correct, we use $[m-(k+1)]$ as the numerator degrees of freedom in $F(LF)$. Here, k is the number of independent variables in the multiple regression model, and m is the number of distinct combinations of the k independent variables for which there is at least one y value. Moreover, in computing SS_{PE}, the set mean of y values for a particular y value is the mean of all of the y values that correspond to the same combination of values of the k independent variables as does the particular y value.

One approach to remedying the lack of fit of the simple linear regression model to the light data is to transform the dependent variable by taking the natural logarithm of each y value. The lower plot in Figure 4.16 shows that the natural logarithms decrease in a straight line fashion but with increasing variation as x increases. If the variation of the original, decreasing y values had been decreasing as x increases, the natural logarithm transformation would have possibly equalized the variation. But, since the variation of the original, decreasing y values is reasonably constant as x increases (see the upper plot in Figure 4.16), the natural logarithm transformation has caused the variation of the decreasing natural logarithms to increase as x increases. Therefore, it is not appropriate to fit the simple linear regression model $ln\, y = \beta_0' + \beta_1' x + \varepsilon'$ to the natural logarithms, because this model assumes that the variation of the error terms and thus of the natural logarithms is constant as x increases.

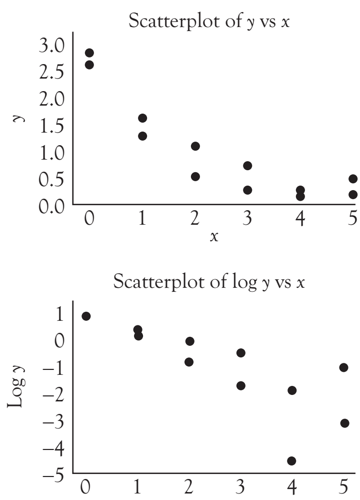


Figure 4.16 Plots of the light data

Note that we use the special symbols β_0', β_1', and ε' to represent the y-intercept, slope, and the error term in the simple linear regression model $\ln y = \beta_0' + \beta_1' x + \varepsilon'$ because, although this model is not appropriate, it can lead us to find an appropriate model. The reason is that the model $\ln y = \beta_0' + \beta_1' x + \varepsilon'$ is equivalent to the model

$$
\begin{aligned}
y &= e^{(\beta_0' + \beta_1' x + \varepsilon')} = (e^{\beta_0'})(e^{\beta_1' x})(e^{\varepsilon'}) \\
&= \beta_2 e^{-\beta_3 x}\,\eta
\end{aligned}
$$

where $\beta_2 = e^{\beta_0'}, -\beta_3 = \beta_1'$, and $\eta = e^{\varepsilon'}$. Just as the expression $\beta_0' + \beta_1' x$ models the straight line decreasing pattern in the natural logarithms of the y's, the expression $\beta_2 e^{-\beta_3 x}$ measures the curvilinear (or exponential) decreasing pattern in the y's themselves (see the upper plot in Figure 4.16). However, the error term $\eta = e^{\varepsilon'}$ is *multiplied* by the expression $\beta_2 e^{-\beta_3 x}$ in the model $y = \beta_2 e^{-\beta_3 x}\eta$. Therefore, this model incorrectly assumes that as x increases and thus $\beta_2 e^{-\beta_3 x}$ decreases, the variation in the y's themselves decreases. To model the fact that as x increases and thus $\beta_2 e^{-\beta_3 x}$ decreases, the variation of the y's stays constant (as we can see is true from the upper plot in Figure 4.16), we can change the multiplicative error $\eta = e^{\varepsilon'}$ to an additive error term ε. In addition, although the upper plot in Figure 4.16 implies that the mean amount of transmitted light $\mu_{y|x}$ might be approaching zero as x increases, we will add an additional parameter β_1 into the final model to allow the possibility that $\mu_{y|x}$ might be approaching a nonzero value β_1 as x increases. This gives us the *final model*

$$
y = \beta_1 + \beta_2 e^{-\beta_3 x} + \varepsilon
$$

The final model is not linear in the parameters β_1, β_2, *and* β_3, and neither is the previously discussed similar model $y = \beta_2 e^{-\beta_3 x}\eta$. However, by taking natural logarithms, the model $y = \beta_2 e^{-\beta_3 x}\eta$ can be *linearized* to the previously discussed logarithmic model as follows:

$$
\begin{aligned}
\ln y &= \ln(\beta_2 e^{-\beta_3 x}\eta) = \ln \beta_2 + \ln e^{-\beta_3 x} + \ln(e^{\varepsilon'}) \\
&= \ln \beta_2 - \beta_3 x + \varepsilon' = \beta_0' + \beta_1' x + \varepsilon'
\end{aligned}
$$

where $\beta_0' = \ln \beta_2$ and $\beta_1' = -\beta_3$. If we fit this simple linear regression model to the natural logarithms of the transmitted light values, we find that the least squares point estimates of β_0' and β_1' are $b_0' = 1.02$ and $b_1' = -.7740$. Considering the models $\ln y = \beta_0' + \beta_1' x + \varepsilon$ and $y = \beta_2 e^{-\beta_3 x}\eta$, since $\beta_0' = \ln \beta_2$, it follows that $\beta_2 = e^{\beta_0'}$, and thus a point estimate of β_2 is $b_2 = e^{b_0'} = e^{1.02} = 2.77$. Moreover, since $\beta_1' = -\beta_3$, it follows that $\beta_3 = -\beta_1'$, and thus a point estimate of β_3 is $b_3 = -b_1' = -(-.7740) = .7740$. Although the nonlinear model $y = \beta_1 + \beta_2 e^{-\beta_3 x} + \varepsilon$ cannot be *linearized* (by using, for example, a natural logarithm transformation), recall that it is reasonable to conclude that β_1 might be near zero. Therefore, we can use 0 as a *preliminary estimate* of β_1 and the estimates $b_2 = 2.77$ and $b_3 = .7740$ for the model $y = \beta_2 e^{-\beta_3 x}\eta$ as *preliminary estimates* of β_2 and β_3 in the model $y = \beta_1 + \beta_2 e^{-\beta_3 x} + \varepsilon$. These preliminary (or initial) estimates are needed because we cannot use the usual matrix algebra formula $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ to calculate the least squares point estimates of the parameters of a nonlinear regression model. Rather, statistical software systems start with user-specified preliminary estimates of the parameters of the nonlinear model and do an iterative search in an attempt to find the least squares point estimates. Figure 4.17a shows the results of the iterative search when we begin with the preliminary estimates 0, 2.77, and .7740 for

(a) The interative search

Iter	beta1	beta2	beta3	Sum of Squares
0	0	2.7700	0.7740	0.5741
1	0.0352	2.7155	0.6797	0.4611
2	0.0288	2.7232	0.6828	0.4604
3	0.0288	2.7233	0.6828	0.4604

(b) The final estimates and statistital inference

Parameter	Estimate	Approx Std Error	Approx 95% Conf Limits	
beta1	0.0288	0.1715	-0.3593	0.4168
beta2	2.7233	0.2105	2.2470	3.1996
beta2	0.6828	0.1417	0.3623	1.0032

Figure 4.17 Partial MINITAB output of nonlinear estimation for the light data.

β_1, β_2, and β_3. Figure 4.17b shows that the final estimates obtained are $b_1 = .0288$, $b_2 = 2.7233$, and $b_3 = .6828$. Because the approximate 95 percent confidence intervals for β_2 and β_3 do not contain zero, we have strong evidence that β_2 and β_3 are significant in the model. Because the 95 percent confidence interval for β_1 does contain zero, we do not have strong evidence that β_1 is significant in the model, However, we will arbitrarily leave β_1 in the model and form the prediction equation $\hat{y} = .0288 + 2.7233e^{-.6828x}$. A practical use of this equation would be to pass a beam of light through a solution of the chemical that has an unknown chemical concentration x, make an optical reading (call it y^*) of the amount of transmitted light, set y^* equal to $.0288 + 2.7233e^{-.6828x}$, and solve for the chemical concentration x.

4.4 Step 4: Diagnosing and Remedying Violations of the Independence Assumption

4.4.1 *Trend, Seasonal Patterns, and Autocorrelation*

Regression Assumption 4, the independence assumption, is most likely to be violated when the regression data are *time series data*–that is, data that have been collected in a time sequence. Time series data can exhibit *trend* and/or *seasonal patterns. Trend* refers to the upward or downward movement that characterizes a time series over time. Thus trend reflects the longrun growth or decline in the time series. Trend movements can represent a variety of factors. For example, long-run movements in the sales of a particular industry might be determined by changes in consumer tastes, increases in total population, and increases in per capita income. *Seasonal variations* are periodic patterns in a time series that complete themselves within a calendar year or less and then are repeated on a regular basis. Often seasonal variations occur yearly. For example, soft drink sales and hotel room occupancies are annually higher in the summer months, while department store sales are annually higher during the winter holiday season. Seasonal variations can also last less than one year. For example, daily restaurant patronage might exhibit within-week seasonal variation, with daily patronage higher on Fridays and Saturdays.

As an example, Figure 4.18 presents a time series of hotel room occupancies observed by Traveler's Rest, Inc., a corporation that operates four hotels in a midwestern city. The analysts in the operating division of the corporation were asked to develop a model that could be used to obtain short-term forecasts (up to one year) of the number of occupied rooms in the hotels. These forecasts were needed by various personnel to assist in hiring additional help during the summer months, ordering materials that have long delivery lead times, budgeting of local advertising expenditures, and so on. The available historical data consisted of the number of occupied rooms during each day for the previous 14 years. Because it was desired to obtain monthly forecasts, these data were reduced to monthly averages by dividing each monthly total by the number of days in the month. The monthly room averages for the previous 14 years are the time series values given in Figure 4.18. A time series plot of these values in Figure 4.18 shows that the monthly room averages follow a strong trend and have a seasonal pattern with one major and several minor peaks during the year. Note that the major peak each year occurs during the high summer travel months of June, July, and August. Moreover, there seems to be some possible curvature in the trend, with the hotel room averages possibly increasing at an increasing rate over time. Also, the seasonal variation appears to fan out over time. To attempt to straighten out the trend and remedy the violation of the constant variance assumption, we will try a square root, a quartic root, and a natural logarithm transformation. The uppermost plot in Figure 4.19 shows that the square roots ($y_t^* = y_t^{.5}$) of the room averages still fan out over time indicating that the square root transformation is not strong enough. The middle plot in Figure 4.19 shows that the quartic roots ($y_t^* = y_t^{.25}$) of the room averages exhibit an approximately straight line trend with approximately constant variation, indicating that the quartic root transformation is appropriate. The lowest plot in Figure 4.19 shows that the natural logarithms ($y_t^* = \ln y_t$) of the room averages might be increasing at a slightly decreasing rate and might be exhibiting slightly decreasing variation over time, as is evidenced by seasonal swings that slightly funnel in over time. Therefore, we might conclude that the natural logarithm transformation is too strong and over-transforms the data. In summary, the quartic root transformation seems best. Letting y_t denote the hotel room

Year	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1	501	488	504	578	545	632	728	725	585	542	480	530
2	518	489	528	599	572	659	739	758	602	587	497	558
3	555	523	532	623	598	683	774	780	609	604	531	592
4	578	543	565	648	615	697	785	830	645	643	551	606
5	585	553	576	665	656	720	826	838	652	661	584	644
6	623	553	599	657	680	759	878	881	705	684	577	656
7	645	593	617	686	679	773	906	934	713	710	600	676
8	645	602	601	709	706	817	930	983	745	735	620	698
9	665	626	649	740	729	824	937	994	781	759	643	728
10	691	649	656	735	748	837	995	1040	809	793	692	763
11	723	655	658	761	768	885	1067	1038	812	790	692	782
12	758	709	715	788	794	893	1046	1075	812	822	714	802
13	748	731	748	827	788	937	1076	1125	840	864	717	813
14	811	732	745	844	833	935	1110	1124	868	860	762	877

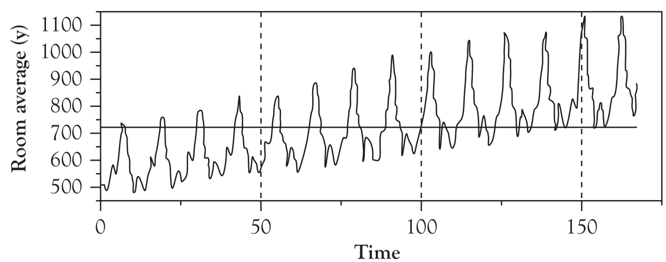


Figure 4.18 Hotel room averages and a time series plot of the hotel room averages.

average observed in time period t, a regression model describing the quartic root of y_t is

$$y_t^{.25} = \beta_0 + \beta_1 t + \beta_{M1}M_1 + \beta_{M2}M_2 + \ldots + \beta_{M11}M_{11} + \varepsilon_t$$

The expression $(\beta_0 + \beta_1 t)$ models the linear trend evident in the middle plot of Figure 4.19. Furthermore, $M_1, M_2, \ldots, M_{11}$ are *seasonal dummy variables* defined for months January (month 1) through November (month 11). For example, M_1 equals 1 if a monthly room average was observed in January, and 0 otherwise; M_2 equals 1 if a monthly room average was observed in February, and 0 otherwise. Note that we have

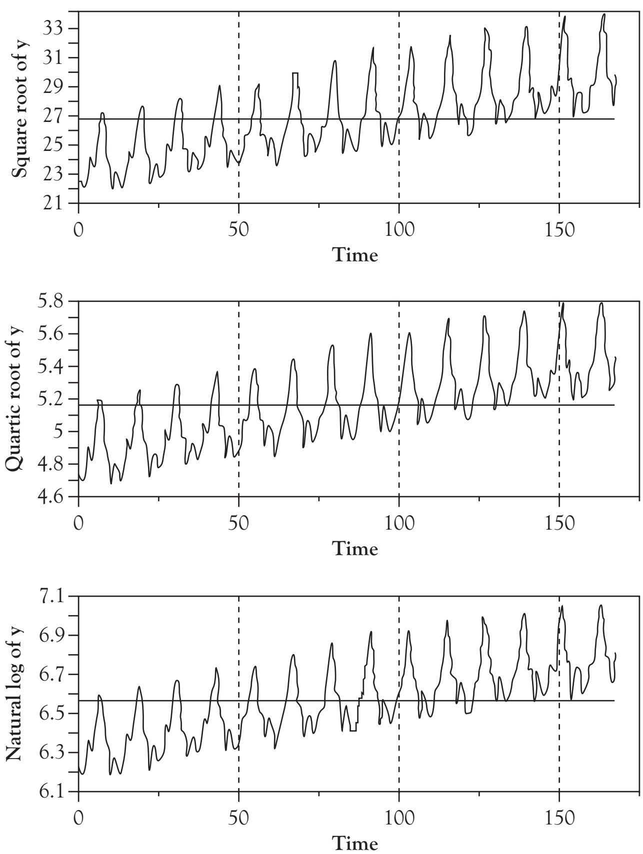


Figure 4.19 Time series plots of the square roots, quartic roots, and natural logarithms of the hotel room averages

not defined a dummy variable for December (month 12). It follows that the regression parameters β_{M1}, $\beta_{M2},\ldots,$ β_{M11} compare January through November with December. Intuitively, for example, β_{M1}, is the difference, excluding trend, between the level of the time series ($y_t^{.25}$) in January and the level of the time series in December. A positive β_{M1} would imply that, excluding trend, the value of the time series in January can be expected to be greater than the value in December. A negative β_{M1} would imply that, excluding trend, the value of the time series in January can be expected to be smaller than the value in December. In general, a trend component such as $\beta_1 t$ and seasonal dummy variables such as $M_1, M_2, \ldots, M_{11}$ are called *time series variables*, whereas an independent variable (such as Traveler's Rest monthly advertising expenditure) that might have a cause and effect relationship with the dependent variable (monthly hotel room average) is called a *causal variable.* We should use whatever time series variables and causal variables that we think might significantly affect the dependent variable when analyzing time series data. As another example, if we plot the demands for Fresh detergent in Table 3.2 versus time (or the sales period number), there is a clear lack of any trend or seasonal patterns. Therefore, it does not seem necessary to add any time series variables into the previously discussed Fresh demand model $y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \varepsilon$. Further verifying this conclusion is Figure 4.20, which shows that a plot of the model's residuals versus time has no trend or seasonal patterns.

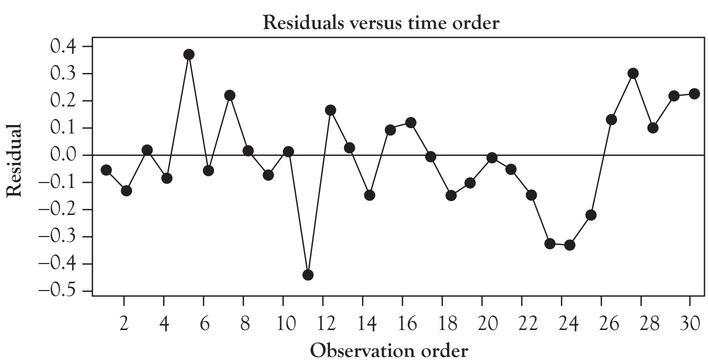


Figure 4.20 Residual plot versus time for the fresh detergent model $\boldsymbol{y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \varepsilon}$

Even when we *think* we have done our best to include the important time series and causal variables in a regression model describing a dependent variable that has been observed over time, the time-ordered error terms in the regression model can still be *autocorrelated.* Intuitively, we say that error terms occurring over time have *positive autocorrelation* when positive error terms tend to be followed over time by positive error terms and when negative error terms tend to be followed over time by negative error terms. Positive autocorrelation in the error terms is depicted in Figure 4.21, which illustrates that *positive autocorrelation can produce a cyclical error term pattern over time.* Because the residuals are point estimates of the error terms, if a plot of the residuals versus the data's time sequence has a cyclical appearance, we have evidence that the error terms are positively autocorrelated and thus that the independence assumption is violated. Another type of autocorrelation that sometimes exists is *negative autocorrelation*, where positive error terms tend to be followed over time by negative error terms and negative error terms tend to be followed over time by positive error terms. *Negative autocorrelation can produce an alternating error term pattern over time* (see Figure 4.22) and is suggested by an alternating pattern in a plot of the time ordered-residuals. Both positive and negative autocorrelation can be caused by leaving important independent variables out of a regression model. For example,

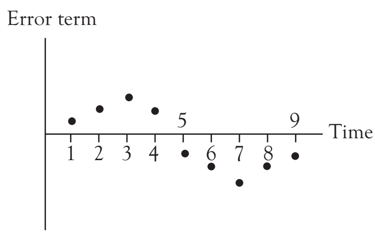


Figure 4.21 Positive autocorrelation

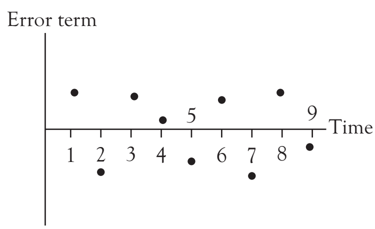


Figure 4.22 Negative autocorrelation

the Fresh demand model $y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \varepsilon$ does not include an independent variable that measures the advertising expenditure for a possible main competitor's laundry detergent. Suppose that such a competitor advertises in a cyclical fashion, with, say, large advertising expenditures for five sales periods, followed by small advertising expenditures for five sales periods, followed by a repeating pattern of this advertising behavior. This cyclical pattern might cause smaller than predicted Fresh demands for five sales periods (see the five negative residuals in periods 21 through 25 in Figure 4.20) followed by larger than predicted Fresh demands for the next five sales periods (see the five positive residuals in periods 26 through 30 in Figure 4.20) followed by a repeating pattern of this Fresh demand behavior. The residual plot in Figure 4.20 has an approximately random, horizontal band appearance until period 21, when a possible cyclical pattern (as just described) begins. It follows that it is questionable as to whether the error terms for the Fresh demand model satisfy the independence assumption or exhibit some possible positive autocorrelation. To remedy the possible positive autocorrelation might seem difficult, because the competing laundry detergent's maker would not wish to tell us what its advertising expenditures have been in the past and what (for the purposes of our predicting future demands for Fresh) its advertising expenditures will be in the future. Moreover, in some situations we cannot identify what independent variable is causing positive or negative autocorrelation. However, we will see at the end of this section that we can account for such autocorrelation by specifying a model that simply describes the relationship between the error terms without discovering the reason for the relationship. Finally, it can be verified that a plot of the residuals from the hotel room average regression model versus time does not have any apparent cyclical or alternating patterns. However, in the next subsection we will see that there is in fact both positive and negative autocorrelation of a rather complex kind in the model's error terms.

4.4.2 The Durbin-Watson Test and Modeling Autocorrelated Errors

One type of positive or negative autocorrelation is called *first-order autocorrelation*. It says that ε_t, the error term in time period t, is related to ε_{t-1}, the error term in time period $t-1$. To check for first-order autocorrelation, we can use the *Durbin–Watson statistic*. To calculate this statistic, we use the time ordered residuals $e_1, e_2, \ldots, e_n$. For example, the residuals $e_1, e_2, \ldots, e_{29}$, and e_{30} from fitting the Fresh demand model $y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \varepsilon$ to the fresh demand data in Table 3.2 are $e_1 = -.044139$, $e_2 = -.122850$, ..., $e_{29} = .234223$, and $e_{30} = .245527$. The definition of the Durbin Watson statistic and its value using the Fresh demand model residuals (where $n = 30$) is as follows:

$$
\begin{aligned}
d &= \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \\
&= \frac{[-.122850-(-.044139)]^2 + \ldots + [.245527 - .234223]^2}{(-.044139)^2 + (-.122850)^2 + \ldots + (.245527)^2} \\
&= 1.512
\end{aligned}
$$

Intuitively, small values of d lead us to conclude that there is positive autocorrelation. This is because, if d is small, the differences $(e_t - e_{t-1})$ are small. This indicates that the adjacent residuals e_t and e_{t-1} are of the same magnitude, which in turn says that the adjacent error terms ε_t and ε_{t-1} are positively correlated. Consider testing the null hypothesis H_0 *that the error terms are not autocorrelated* versus the alternative hypothesis H_a *that the error terms are positively autocorrelated.* Durbin and Watson have shown that there are points (denoted $d_{L,\alpha}$ and $d_{U,\alpha}$) such that, if α is the probability of a Type I error, then

1. If $d < d_{L,\alpha}$, we reject H_0.
2. If $d > d_{U,\alpha}$, we do not reject H_0.
3. If $d_{L,\alpha} \le d \le d_{U,\alpha}$, the test is inconclusive.

Table A4 give values of $d_{L,\alpha}$ and $d_{U,\alpha}$ for $\alpha = .05$ and different values of k, the number of independent variables used by the regression model, and n, the number of observations. (Tables of $d_{L,\alpha}$ and $d_{U,\alpha}$ for different values of α can be found in more detailed books of statistical tables). Since there are $n = 30$ Fresh demands in Table 3.2 and $k = 4$ independent variables in the

Fresh demand model, Table A4 tells us that $d_{L,.05} = 1.14$ and $d_{U,.05} = 1.74$. Since $d = 1.512$ for the Fresh demand model is between these points, the test for positive autocorrelation is inconclusive (as is the residual plot in Figure 4.20).

It can be shown that the Durbin–Watson statistic d is always between 0 and 4. Large values of d (and hence small values of $4 - d$) lead us to conclude that there is negative autocorrelation because if d is large, this indicates that the differences $(e_t - e_{t-1})$ are large. This says that the adjacent error terms ε_t and ε_{t-1} are negatively autocorrelated. Consider testing the null hypothesis H_0 *that the error terms are not autocorrelated* versus the alternative hypothesis H_a *that the error terms are negatively autocorrelated.* Durbin and Watson have shown that based on setting the probability of a Type I error equal to α, the points $d_{L,\alpha}$ and $d_{U,\alpha}$ are such that

1. If $(4-d) < d_{L,\alpha}$, we reject H_0.
2. If $(4-d) > d_{U,\alpha}$, we do not reject H_0.
3. If $d_{L,\alpha} \leq (4-d) \leq d_{U,\alpha}$ the test is inconclusive.

For example, for the fresh demand model we see that $(4-d) = (4-1.512) = 2.488$ is greater than $d_{U,.05} = 1.74$. Therefore, on the basis of setting α equal to .05, we do not reject the null hypothesis of no autocorrelation. That is, there is no evidence of negative (first-order) autocorrelation.

We can also use the Durbin–Watson statistic to test for positive or negative autocorrelation. Specifically, consider testing the null hypothesis H_0 *that the error terms are not autocorrelated* versus the alternative hypothesis H_a *that the error terms are positively or negatively autocorrelated.* Durbin and Watson have shown that, based on setting the probability of a Type I error equal to α**,** we perform both the above described test for positive autocorrelation and the above described test for negative autocorrelation by using the critical values $d_{L,\alpha/2}$ and $d_{U,\alpha/2}$ for each test. If *either test* says to reject H_0, then we reject H_0. If *both tests* say to not reject H_0, then we do not reject H_0. Finally, if *either test* is inconclusive, then the overall test is inconclusive.

As another example of testing for positive autocorrelation, consider the $n = 168$ hotel room averages in Figure 4.18 and note that when we fit the *quartic root room average model*

$$y_t^{.25} = \beta_0 + \beta_1 t + \beta_{M1} M_1 + \beta_{M2} M_2 + \ldots + \beta_{M11} M_{11} + \varepsilon_t$$

to these data, we find that the Durbin-Watson statistic is $d = 1.26$. Because the above model uses $k = 12$ independent variables and there are $n = 168$ observations, the points $d_{L,.05}$ and $d_{U,.05}$ are not in Table A4. However $d = 1.26$ is fairly small and thus indicative of possible positive autocorrelation in the error terms. One approach to dealing with autocorrelation in the error terms is to predict a future error term ε_t by using an *autoregressive model* that relates ε_t to past error terms $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$. One way to find such a model is to use SAS PROC AUTOREG. This procedure begins by fitting the quartic root room average model to the $n = 168$ hotel room averages and then performs a *backward elimination* on the residuals of this model to choose an appropriate autoregressive model describing the residuals. This model is an estimate of the model describing the error terms. The user must supply what is called a *maximum lag* q and level of significance (denoted α_{stay}) in order to use the backward elimination procedure. The procedure begins by assuming that ε_t is described by the autoregressive model

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \ldots + \phi_q \varepsilon_{t-q} + a_t$$

where the a_t' s, which are called *random shocks*, are assumed to be numerical values that have been randomly and independently selected from a normally distributed population of numerical values having mean 0 and a variance that does not depend on t. Estimates of the autoregressive model parameters are obtained by using all terms in the autoregressive model. Then the error term with the smallest (in absolute value) t statistic is selected. If the t statistic indicates that this term is significant at the α_{stay} level (that is, the related p-value is less than α_{stay}), then the procedure terminates by choosing the error structure including all q terms. If this term is not significant at the α_{stay} level, it is removed from the model, and estimates of the model parameters are obtained by using an autoregressive model containing all the remaining terms. The procedure continues by removing terms one at a time from the model describing the error structure. At each step a term is removed if it has the smallest (in absolute value) t statistic of the terms remaining in the model and if it is not significant at the α_{stay} level. The procedure terminates when none of the terms remaining can be removed. The experience of the authors indicates that choosing α_{stay} equal to .15 is effective and when monthly data is being analyzed, choosing $q = 18$ is also effective. When we make these choices to analyze the room average data, Figure 4.23 tells us that SAS PROC AUTOREG chooses the autoregressive model

$$\varepsilon_t = \phi_1\varepsilon_{t-1} + \phi_2\varepsilon_{t-2} + \phi_3\varepsilon_{t-3} + \phi_{12}\varepsilon_{t-12} + \phi_{18}\varepsilon_{t-18}$$

When we use SAS PROC ARIMA to fit the quartic root room average model combined with this autoregressive error term model, we obtain the SAS output of *estimation, diagnostic checking, and forecasting* that is given in Figure 4.24. Without going into the theory of diagnostic checking, it can be shown that because each of the *chi-square p*-values in Figure 4.24b is greater than .05, the combined model has *adequately* accounted for the autocorrelation in the data (see Bowerman et al. 2005). Using the least squares point estimates in Figure 4.24a, we compute a point prediction of $y_{169}^{.25}$, the quartic root of the hotel room average in period 169 (January of next year) to be

$$\begin{aligned}
&b_0 + b_1 t + b_{M1}M_1 + b_{M2}M_2 + \ldots + b_{M11}M_{11} + \hat{\varepsilon}_t \\
&= b_0 + b_1(169) + b_{M1}(1) + b_{M2}(0) + \ldots + b_{M11}(0) + \hat{\varepsilon}_{169} \\
&= b_0 + b_1(169) + b_{M1} + \hat{\phi}_1\hat{\varepsilon}_{168} + \hat{\phi}_2\hat{\varepsilon}_{167} + \hat{\phi}_3\hat{\varepsilon}_{166} + \hat{\phi}_{12}\hat{\varepsilon}_{157} + \hat{\phi}_{18}\hat{\varepsilon}_{151} \\
&= 4.80114 + .0035312(169) + (-.04589) + .30861\,\hat{\varepsilon}_{168} + .12487\,\hat{\varepsilon}_{167} \\
&\quad + (-.26534)\hat{\varepsilon}_{166} + .26437\,\hat{\varepsilon}_{157} + (-.15846)\hat{\varepsilon}_{151} \\
&= 5.3788
\end{aligned}$$

Estimates of the Authoregressive Parameters

Lag	Coefficient	Std Error	t Ratio
1	-0.29654610	0.07645457	-3.878723
2	-0.12575788	0.07918744	-1.588104
3	0.25507527	0.07532157	3.386484
12	-0.22113831	0.07208243	-3.067853
18	0.13817435	0.07054036	1.958799

Figure 4.23 SAS PROC AUTOREG output of using backward elimination to find an autoregressive error term model for the error terms of the quartic root room average model ($\alpha_{\text{stay}} = .15$ and $q = 18$)

(a) Estimation

Parameter	Estimate	T Ratio	Lag	Variable
MU	4.80114	407.21	0	QRY
AR1,1	0.30861	4.05	1	QRY
AR1,2	0.12487	1.57	2	QRY
AR1,3	-0.26534	-3.53	3	QRY
AR1,4	0.26437	3.54	12	QRY
AR1,5	-0.15846	-2.12	18	QRY
NUM1	0.0035312	67.41	0	TIME
NUM2	-0.04589	-4.20	0	M1
NUM3	-0.13005	-9.96	0	M2
NUM4	-0.09662	-6.04	0	M3
NUM5	0.06160	3.56	0	M4
NUM6	0.03528	1.94	0	M5
NUM7	0.19768	10.53	0	M6
NUM8	0.38731	21.32	0	M7
NUM9	0.41322	23.88	0	M8
NUM10	0.07050	4.37	0	M9
NUM11	0.04656	3.60	0	M10
NUM12	-0.14464	-13.45	0	M11

Constant Estimate = 3.48539274

Variance Estimate = 0.00057384
Std Error Estimate = 0.02395493

(b) Diagnostic checking

To Lag	Chi Square	DF	Prob
6	2.55	1	0.110
12	6.69	7	0.462
18	9.44	13	0.739
24	13.32	19	0.822
30	18.14	25	0.836

(c) Predictions of $y_{169}^{.25}$ through $y_{180}^{.25}$

Obs	Forecast	Std Error	Lower 95%	Upper 95%
169	5.3788	0.0240	5.3318	5.4257
170	5.2699	0.0251	5.2207	5.3190
171	5.2921	0.0256	5.2419	5.3423
172	5.4431	0.0259	5.3923	5.4938
173	5.4334	0.0260	5.3824	5.4844
174	5.6009	0.0262	5.5497	5.6522
175	5.8059	0.0262	5.7547	5.8572
176	5.8414	0.0262	5.7901	5.8926
177	5.5012	0.0262	5.4499	5.5525
178	5.4796	0.0262	5.4283	5.5308
179	5.2959	0.0262	5.2446	5.3472
180	5.4573	0.0262	5.4060	5.5086

(d) Predictions of y_{169} through y_{180}

Obs	Y	L95CI	FY	U95CI
169	.	808.17	837.02	866.63
170	.	742.88	771.25	800.42
171	.	755.02	784.37	814.56
172	.	845.46	877.75	910.96
173	.	839.25	871.51	904.69
174	.	948.57	984.10	1020.62
175	.	1096.68	1136.28	1176.94
176	.	1123.94	1164.27	1205.68
177	.	882.17	915.85	950.48
178	.	868.25	901.53	935.76
179	.	756.60	786.63	817.54
180	.	854.12	886.99	920.81

Figure 4.24 Partial SAS PROC ARIMA output of a regression analysis using the quartic root room average model combined with an autoregressive error term model

Here, the predictions $\hat{\varepsilon}_{168}$, $\hat{\varepsilon}_{167}$, $\hat{\varepsilon}_{166}$, $\hat{\varepsilon}_{157}$ and $\hat{\varepsilon}_{151}$ of the error terms ε_{168}, ε_{167}, ε_{166}, ε_{157}, and ε_{151} are the residuals $e_{168}, e_{167}, e_{166}, e_{157}$, and e_{151} obtained by using the quartic root room average model to predict the quartic roots of the room averages in periods 168, 167, 166, 157, and 151. For example, because the quartic root of $y_{167} = 762$ (see Figure 4.18) is 5.253984, and because period 167 is a November with $b_{M11} = -.14464$, we have $\hat{\varepsilon}_{167} = e_{167}$ $= 5.253984 - [4.80114 + .0035312(167) + (-.14464)] = .0077736$. The point prediction 5.3788 of $y_{169}^{.25}$ is given in Figure 4.24c and implies that the point prediction of y_{169} is $(5.3788)^4$ = 837.02 [see Figure 4.24d]. Figure 4.24c also tells us that a 95 percent prediction interval for $y_{169}^{.25}$ is [5.3318, 5.4257], which implies that a 95 percent prediction interval for y_{169} is $[(5.3318)^4, (5.4257)^4]$ = [808.17, 866.63] (see Figure 4.24d). This interval says that Traveler's Rest can be 95 percent confident that the monthly hotel room average in period 169 (January of next year) will be no less than 808.17 rooms per day and no more than 866.63 rooms per day. Lastly, note that Figures 4.24c and 4.24d also give point predictions of and 95 percent prediction intervals for $y_{170}^{.25}, \ldots, y_{180}^{.25}$ and $y_{170}, \ldots, y_{180}$ (the hotel room averages in February through December of next year).

In order to see how least squares point estimates like those in Figure 4.24(a) are calculated, consider, in general, a regression model that describes a time series of y_t values by using k time series and/or causal independent variables. We will call this model the *original regression model*, and to simplify discussions to follow, we will express it by showing only an arbitrary one of its k independent variables. Therefore, we will express this model as $y_t = \beta_0 + \cdots + \beta_j x_{tj} + \cdots + \varepsilon_t$. If the error terms in the model are not statistically independent but are described by the error term model $\varepsilon_t = \varphi_1 \varepsilon_{t-1} + \varphi_2 \varepsilon_{t-2} + \cdots + \varphi_q \varepsilon_{t-q} + a_t$, regression assumption 4 is violated. To remedy this regression assumption violation, we can use the original regression model to write out expressions for $y_t, \varphi_1 y_{t-1}, \varphi_2 y_{t-2}, \ldots, \varphi_q y_{t-q}$ and then consider the transformed regression model

$$
\begin{aligned}
y_t - \varphi_1 y_{t-1} - \varphi_2 y_{t-2} - \cdots - \varphi_q y_{t-q} \\
&= \beta_0 - \beta_0 \varphi_1 - \beta_0 \varphi_2 - \cdots - \beta_0 \varphi_q + \cdots + \\
&\quad \beta_j x_{tj} - \beta_j \varphi_1 x_{t-1,j} - \beta_j \varphi_2 x_{t-2,j} - \cdots - \beta_j \varphi_q x_{t-q,j} \\
&\quad + \cdots + \varepsilon_t - \varphi_1 \varepsilon_{t-1} - \varphi_2 \varepsilon_{t-2} - \cdots - \varphi_q \varepsilon_{t-q}
\end{aligned}
$$

This transformed model can be written concisely as $y_t^* = \beta_0^* + \cdots + \beta_j x_{tj}^* + \cdots + \varepsilon_t^*$, where, for $t = q+1, q+2, \ldots, n$: $y_t^* = y_t - \varphi_1 y_{t-1} - \varphi_2 y_{t-2} - \cdots - \varphi_q y_{t-q}$, $\beta_0^* = \beta_0(1 - \varphi_1 - \varphi_2 - \cdots - \varphi_q)$, $x_{tj}^* = x_{tj} - \varphi_1 x_{t-1,j} - \varphi_2 x_{t-2,j} - \cdots - \varphi_q x_{t-q,j}$ and $\varepsilon_t^* = \varepsilon_t - \varphi_1 \varepsilon_{t-1} - \varphi_2 \varepsilon_{t-2} - \cdots - \varphi_q \varepsilon_{t-q}$. The transformed model has independent error terms. This is because, since the error term model says that $\varepsilon_t = \varphi_1 \varepsilon_{t-1} + \varphi_2 \varepsilon_{t-2} + \cdots + \varphi_q \varepsilon_{t-q} + a_t$, it follows that $\varepsilon_t^* = \varepsilon_t - \varphi_1 \varepsilon_{t-1} - \varphi_2 \varepsilon_{t-2} - \cdots - \varphi_q \varepsilon_{t-q} = a_t$, and the a_t's are the previously discussed random shocks that are assumed to be statistically independent. Unfortunately, we do not know the true values of $\varphi_1, \varphi_2, \ldots, \varphi_q$, and so we need to estimate these φ parameters. The *Cochran-Orcutt procedure* is a three step iterative procedure that estimates both the φ and the β parameters in the original regression model. This procedure (1) uses the original regression model to calculate least squares point estimates $b_0, b_1, \ldots, b_j, \ldots, b_k$ based on the original observed y_t values and calculates residuals using the fitted model; (2) uses the residuals $e_{q+1}, e_{q+2}, \ldots, e_n$ to find the least squares point estimates $\hat{\varphi}_1, \hat{\varphi}_2, \ldots, \hat{\varphi}_q$ of the parameters $\varphi_1, \varphi_2, \ldots, \varphi_q$ in the model $e_t = \varphi_1 e_{t-1} + \varphi_2 e_{t-2} + \cdots + \varphi_q e_{t-q}$; and (3) uses the transformed model $y_t^* = \beta_0^* + \cdots + \beta_j x_{tj}^* + \cdots + \varepsilon_t^*$, where, for $t = q+1, q+2, \ldots, n$: $y_t^* = y_t - \hat{\varphi}_1 y_{t-1} - \hat{\varphi}_2 y_{t-2} - \cdots - \hat{\varphi}_q y_{t-q}$ and $x_{tj}^* = x_{tj} - \hat{\varphi}_1 x_{t-1,j} - \hat{\varphi}_2 x_{t-2,j} - \cdots - \hat{\varphi}_q x_{t-q,j}$ to calculate new least squares point estimates $b_0^*, b_1, \ldots, b_j, \ldots, b_k$. Note that because $\beta_0^* = \beta_0(1 - \varphi_1 - \varphi_2 - \cdots - \varphi_q)$, the new least squares estimate of β_0 is $b_0 = b_0^* / (1 - \hat{\varphi}_1 - \hat{\varphi}_2 - \cdots - \hat{\varphi}_q)$. If the new least squares point estimates are "close" to the original least squares point estimates, the procedure stops and uses $\hat{\varphi}_1, \hat{\varphi}_2, \cdots \hat{\varphi}_q$ and the new least squares point estimates $b_0, b_1, \ldots, b_j, \ldots, b_k$ as the final least squares point estimates. Otherwise, the new least squares point estimates are inserted into the original regression model, new residuals are computed, and steps (2) and (3) are repeated. This iterative procedure continues until the least squares point estimates change little between iterations. Usually, a very small number of iterations is required, but if the procedure does not converge quickly, another procedure should be tried. Also note that the procedure losses information from the first q observations. If n is large, the loss of information is not severe, and there are methods to recoup the lost information. Finally note that although the Cochran-Orcutt procedure is iterative, it can be carried out using ordinary least squares. In contrast, the

Hildreth-Lu procedure does a numerical search to find the combination of estimates of $\varphi_1, \varphi_2, \ldots, \varphi_q, \beta_1, \ldots, \beta_j, \ldots, \beta_k$ that minimizes the sum of squared differences between the y_t^*'s and the predictions of the y_t^*'s given by the transformed regression model. The procedure is not iterative but requires advanced computing techniques. The Cochran-Orcutt procedure, the Hildreth procedure, and other procedures are used by various statistical software systems. For example, SAS PROC ARIMA gives the user a choice between using the *maximum likelihood method,* the *conditional least squares method,* and the *unconditional least squares method* of estimating the β and ϕ parameters. The estimates in Figure 4.24a were obtained by using the conditional least squares method. Appendix D extends the discussion of modeling time series data given here and considers the Box-Jenkins methodology.

4.5 Step 5: Diagnosing and Using Information About Outlying and Influential Observations

An observation that is well separated from the rest of the data is called an outlier, and an observation may be an outlier with respect to its y value or its x values, or both. We illustrate these ideas by considering Figure 4.25, which is a hypothetical plot of the values of a dependent variable y against an independent variable x. Observation 1 in this figure is outlying with respect to its y value, but not with respect to its x value. Observation 2 is outlying with respect to its x value, but because its y value is consistent with the regression relationship displayed by the nonoutlying observations, it is not outlying with respect to its y value. Observation 3 is an outlier with respect to its x value and its y value.

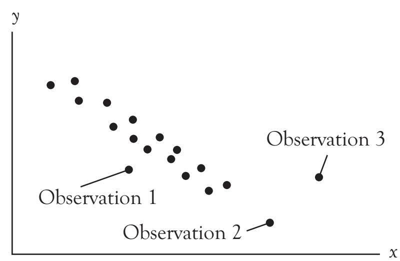


Figure 4.25 Outlying observations

It is important to identify outliers because (as we will see) outliers can have adverse effects on a regression analysis and thus are candidates for removal from a data set. Moreover, in addition to using data plots, we can use more sophisticated procedures to detect outliers. For example, suppose that the U.S. Nary wishes to develop a regression model based on efficiently run Navy hospitals to evaluate the labor needs of questionably run Navy hospitals. Table 4.2 gives labor needs data for 17 Navy hospitals. Specifically, this table gives values of the dependent variable Hours (y, monthly labor hours required) and of the independent variables X-ray (x_1, monthly X-ray exposures), BedDays (x_2, monthly occupied bed days—a hospital has one occupied bed day if one bed is occupied for an entire day), Length (x_3, average length of patients' stay, in days), Load (x_4, average daily patient load), and Pop(x_5, eligible population in the area, in thousands). In the exercises the reader will show that the model describing these data that gives the smallest s and smallest C statistic is the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$. When we fit this model, which we will sometimes call the *original model*, to the data in Table 4.2, we obtain the SAS output of outlying and influential diagnostics in Figure 4.26a and the residual plot in Figure 4.26b. We will now interpret those diagnostics, and in a technical note at the end of this section we will learn how to calculate them.

4.5.1 Leverage Values

The leverage value for an observation is the distance value, discussed in Section 2.7, and is used to calculate a prediction interval for the y value of the observation. This value is a measure of the distance between the observation's x values and the center of the experimental region. The leverage value is labeled as *Hat Diag H* on the SAS output in Figure 4.26a. If the leverage value for an observation is large, the observation is outlying with respect to its x values and thus would have substantial leverage in determining the least squares prediction equation. To intuitively understand this, note that each of observations 2 and 3 in Figure 4.25 is an outlier with respect to its x value and thus would have substantial leverage in determining the position of the least squares line. Moreover, because observations 2 and 3 have inconsistent y values, they would pull

Table 4.2 Hospital labor needs data

Hospital	Hours y	Xray x_1	BedDays x_2	Length x_3	Load x_4	Pop x_5
1	566.52	2463	472.92	4.45	15.57	18.0
2	696.82	2048	1339.75	6.92	44.02	9.5
3	1033.15	3940	620.25	4.28	20.42	12.8
4	1603.62	6505	568.33	3.90	18.74	36.7
5	1611.37	5723	1497.60	5.50	49.20	35.7
6	1613.27	11520	1365.83	4.60	44.92	24.0
7	1854.17	5779	1687.00	5.62	55.48	43.3
8	2160.55	5969	1639.92	5.15	59.28	46.7
9	2305.58	8461	2872.33	6.18	94.39	78.7
10	3503.93	20106	3655.08	6.15	128.02	180.5
11	3571.89	13313	2912.00	5.88	96.00	60.9
12	3741.40	10771	3921.00	4.88	131.42	103.7
13	4026,52	15543	3865.67	5.50	127.21	126.8
14	10343.81	36194	7684.10	7.00	252.90	157.7
15	11732.17	34703	12446.33	10.78	409.20	169.4
16	15414.94	39204	14098.40	7.05	463.70	331.4
17	18854.45	86533	15524.00	6.35	510.22	371.6

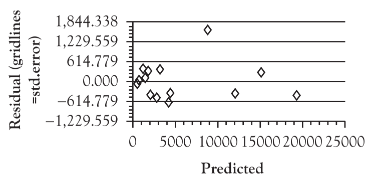
Source: Procedures and Analysis for Staffing Standards Development: Regression Analysis Handbook (San Diego, CA: Navy Manpower and Material Analysis Center. 1979).

the least squares line in opposite directions. A leverage value is considered to be large if it is greater than twice the average of all of the leverage values, which can be shown to be equal to $2(k+1)/\text{n}$. For example, because there are $n = 17$ observations in Table 4.2 and because the model relating y to $x_1, x_2,$ and x_3 utilizes $k = 3$ independent variables, twice the average leverage value is $2(k+1)/n = 2(3+1)/17 = .4706$. Looking at Figure 4.26a, we see that the leverage values for hospitals 15, 16, and 17 are, respectively, .682, .785, and .863. Because these leverage values are greater than .4706, we conclude that hospitals 15, 16, and 17 are outliers with respect to their x values. Intuitively, this is because Table 4.2 indicates that x_2 (monthly occupied bed days) is substantially larger for hospitals 15, 16, and 17 than for hospitals 1 through 14. Also note that both x_1 (monthly X-ray exposures) and x_2 (monthly occupied bed days) are substantially larger for hospital 14 than for hospitals 1 through 13. To

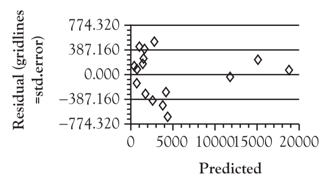
(a) Diagnostics for original model and studentized deleted residuals for options 1 and 2

Obs	Residual	Std Err Residual	Student Residual	Hat Diag H	Rstudent	Option1 Rstudent	Option2 Rstudent	Cook's D	Dffits
1	-121.9	576.469	-0.211	0.1207	-0.2035	-0.3330	-1.4388	0.002	-0.0754
2	-25.0283	540.821	-0.046	0.2261	-0.0445	0.4036	0.2327	0.000	-0.0240
3	67.7570	573.539	0.118	0.1297	0.1136	0.1607	-0.7498	0.001	0.0438
4	431.2	563.870	0.765	0.1588	0.7517	1.2336	0.2025	0.028	0.3266
5	84.5898	588.099	0.144	0.0849	0.1383	0.4249	0.2128	0.000	0.0421
6	-380.6	579.326	-0.657	0.1120	-0.6419	-0.7953	-1.4903	0.014	-0.2280
7	177.6	588.367	0.302	0.0841	0.2911	0.6766	0.6172	0.002	0.0882
8	369.1	588.712	0.627	0.0830	0.6118	1.1171	1.0099	0.009	0.1841
9	-493.2	588.201	-0.838	0.0846	-0.8283	-1.0783	-0.4091	0.016	-0.2518
10	-687.4	576.628	-1.192	0.1203	-1.2136	-1.3591	-0.4002	0.049	-0.4487
11	380.9	590.529	0.645	0.0773	0.6299	1.4612	2.5712	0.009	0.1824
12	-623.1	557.704	-1.117	0.1771	-1.1290	-2.2241	-0.6245	0.067	-0.5237
13	-337.7	594.623	-0.568	0.0645	-0.5526	-0.6851	0.4643	0.006	-0.1451
14	1630.5	567.981	2.871	0.1465	4.5584		1.4058	0.353	1.8882
15	-348.7	346.813	-1.005	0.6818	-1.0059	-0.1375	-2.0492	0.541	-1.4723
16	281.9	284.743	0.990	0.7855	0.9892	1.2537	1.1081	0.897	1.8930
17	-406.0	227.346	-1.786	0.8632	-1.9751	0.5966	-0.6386	5.033	-4.9623

(b) Plot of residuals for original model



(c) Plot of residuals for Option 1



(d) Plot of residuals for Option 2

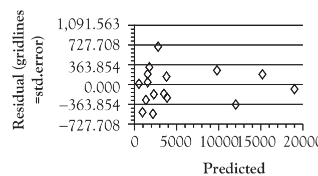


Figure 4.26 Partial SAS output of outlying and influential observation diagnostics

summarize, we might classify hospitals 1 through 13 as small to medium sized hospitals and hospitals 14, 15, 16, and 17 as larger hospitals.

4.5.2 Studentized Residuals and Studentized Deleted Residuals

To identify outliers with respect to their y values, we can use residuals. Any residual that is substantially different from the others is suspect. For example, note from Figure 4.26a that the residual for hospital 14, $e_{14} = 1630.503$, seems much larger than the other residuals. Assuming that the labor hours of 10,343.81 for hospital 14 has not been misrecorded, the residual of 1630.503 says that the labor hours are 1630.503 hours more than predicted by the regression model. If we divide an observation's residual by the residual's standard error, we obtain a studentized residual. For example, Figure 4.26a tells us that the studentized residual (see "Student Residual") for hospital 14 is 2.871. If the studentized residual for an observation is greater than 2 in absolute value, we have some evidence that the observation is an outlier with respect to its y value. However, a better way to identify an outlier with respect to its y value is to use a studentized deleted residual. To introduce this statistic, consider again Figure 4.25 and suppose that we use observation 3 to determine the least squares line. Doing this might draw the least squares line toward observation 3, causing the point prediction $\hat{y}_3$ given by the line to be near y_3 and thus the usual residual $y_3 - \hat{y}_3$ to be small. This would falsely imply that observation 3 is not an outlier with respect to its y value. Moreover, this sort of situation shows the need for computing a deleted residual. For a particular observation, observation i, the deleted residual is found by subtracting from y_i the point prediction $\hat{y}_{(i)}$ computed using least squares point estimates based on all n observations except for observation i. Standard statistical software packages calculate the deleted residual for each observation and divide this residual by its standard error to form the studentized deleted residual. The experience of the authors leads us to suggest that one should conclude that an observation is an outlier with respect to its y value if (and only if) the studentized deleted residual is greater in absolute value than $t_{[.005]}$, which is based on $n - k - 2$ degrees of freedom. For the hospital labor needs model, $n - k - 2 = 17 - 3 - 2 = 12$, and therefore $t_{[.005]} = 3.055$. The studentized deleted residual for hospital 14, which equals 4.5584 (see "Rstudent" in Figure 4.26a), is greater in absolute value than $t_{[.005]} = 3.055$. Therefore, we conclude that hospital 14 is an outlier with respect to its y value.

4.5.3 An Example of Dealing with Outliers

One option for dealing with the fact that hospital 14 is an outlier with respect to its y value is to assume that hospital 14 has been run inefficiently. Because we need to develop a regression model using efficiently run hospitals, based on this assumption we would remove hospital 14 from the data set. If we perform a regression analysis using a model relating y to x_1, x_2, and x_3 with hospital 14 removed from the data set (we call this Option 1), we obtain a standard error of $s = 387.16$. This s is considerably smaller than the large standard error of 614.779 caused by hospital 14's large residual when we use all 17 hospitals to relate y to x_1, x_2, and x_3.

A second option is motivated by the fact that large organizations sometimes exhibit inherent inefficiencies. To assess whether there might be general large hospital inefficiency, we define a dummy variable D_L that equals 1 for the larger hospitals 14 to 17 and 0 for the smaller hospitals 1 to 13. If we fit the resulting regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 D_L + \varepsilon$ to all 17 hospitals (we call this Option 2), we obtain a b_4 of 2871.78 and a p-value for testing $H_0 : \beta_4 = 0$ of .0003. This indicates the existence of a large hospital inefficiency that is estimated to be an extra 2871.78 hours per month. In addition, the dummy variable model's s is 363.854, which is slightly smaller than the s of 387.16 obtained using Option1. In the exercises the reader will use the studentized deleted residual for hospital 14 when using Option 2 (see Figure 4.26a) to show that hospital 14 is not an outlier with respect to its y value. This means that if we remove hospital 14 from the data set and predict y_{14} by using a newly fitted dummy variable model having a large hospital inefficiency estimate based on the remaining large hospitals 15, 16, and 17, the prediction obtained indicates that hospital 14's labor hours are not unusually large. This justifies leaving hospital 14 in the data set when using the dummy variable model. In summary, both Options 1 and 2 seem reasonable. The reader will further compare these options in the exercises.

4.5.4 Cook's D, Dfbetas, and Dffits

If a particular observation, observation i, is an outlier with respect to its y or x values, it might significantly influence the least squares point estimates of the model parameters. To detect such influence, we compute Cook's distance measure (or Cook's D) for observation i, which we denote as D_i. To understand D_i, let $F_{.50}$ denote the 50th percentile of the F distribution based on $(k+1)$ numerator and $n-(k+1)$ denominator degrees of freedom. It can be shown that if D_i is greater than $F_{.50}$, then removing observation i from the data set would significantly change (as a group) the least squares point estimates of the model parameters. In this case we say that observation i is *influential*. For example, suppose that we relate y to $x_1, x_2,$ and x_3 using all $n=17$ observations in Table 4.2 Noting that $k+1 = 4$ and $n-(k+1)=13$, we find (using Excel) that $F_{.50}=.8845$. Figure 4.26a tells us that $D_{16}=.897$ and $D_{17} = 5.033$. Since both $D_{16}=.897$ and $D_{17} = 5.033$ are greater than $F_{.50} = .8845$, it follows that removing either hospital 16 or 17 from the data set would significantly change (as a group) the least squares estimates of the model parameters.

To assess whether a particular least squares point estimate b_j would significantly change, we consider the difference between the least squares point estimate b_j of β_j, computed using all n observations, and the least squares point estimate $b_j^{(i)}$ of β_j, computed using all n observations except for observation i. SAS calculates this difference for each observation and divides the difference by its standard error to form the *difference in estimate of* β_j *statistic*. If the absolute value of this statistic is greater than 2 (a sometimes-used critical value for this statistic), then removing observation i from the data set would substantially change the least squares point estimate of β_j. Figure 4.27 shows the SAS output of the difference in estimate of β_j statistics (Dfbetas) for hospitals 16 and 17. Examining this output we see that for hospital 17 "INTERCEP Dfbetas" (=.0294), "X2 Dfbetas" (=1.2688), and "**X**3 Dfbetas" (=.3155) are all less than 2 in absolute value. This says that individual least squares point estimates of $\beta_0, \beta_2,$ and β_3 probably would not change substantially if hospital 17 were removed from the data set. Similarly, all of the of Dfbetas statistics for hospital 16 and (it can be verified) for the other hospitals (1 to 15) not shown in Figure 4.27 are less than 2 in absolute value. This says that the individual least squares

Obs	INTERCEP Dfbetas	X1 Dfbetas	X2 Dfbetas	X3 Dfbetas
16	0.9880	-1.4289	1.7339	-1.1029
17	0.0294	-3.0114	1.2688	0.3155

Figure 4.27 SAS output for Dfbetas for hospitals 16 and 17

point estimates of $\beta_0, \beta_1, \beta_2,$ and β_3 would not change substantially if any one of hospitals 1 to 16 were removed from the dataset. However, for observation 17 "X1 Dfbetas" (= – 3.0114) is greater than 2 in absolute value and is negative. This implies that removing hospital 17 from the dataset would *significantly decrease* the least squares point estimate of the effect, β_1, of monthly X-ray exposures on monthly labor hours. One possible consequence might then be that our model would *significantly underpredict* the monthly labor hours for a hospital which (like hospital 17—see Table 4.2) has a particularly large number of monthly X-ray exposures.

To assess whether a particular point prediction, $\hat{y}$, would significantly change, consider the difference between the point prediction $\hat{y}_i$ of y_i, computed using least squares point estimates based on all n observations, and the point prediction $\hat{y}_{(i)}$ of y_i, computed using least squares point estimates based on all n observations except for observation i. SAS calculates this difference for each observation and divides the difference by its standard-error to form the *difference in fits statistic*. If the absolute value of this statistic is greater than 2 (a sometimes used critical value for this statistic), then removing observation i from the dataset would substantially change the point prediction of y_i. For example, Figure 4.26a tells us that the difference in fits statistic (Dffits) for hospital 17 equals –4.9623, which is greater than 2 in absolute value and is negative. This implies that removing hospital 17 from the dataset would significantly reduce the point prediction of y_{17}—that is, of the labor hours for a hospital that has the same independent variable values (including the large number of X-ray exposures) as hospital 17. Moreover, although it can be verified that using the previously discussed Option 1 or Option 2 to deal with hospital 14's large residual substantially reduces Cook's D, Dfbetas for x_1, and Dffits for hospital 17, these or similar statistics remain or become somewhat significant for the large hospitals 15, 16, and 17. The practical implication is that if we wish to predict monthly labor hours for questionably run large hospitals, it is very important to keep all of the efficiently run large hospitals 15, 16, and 17 in the data set. (Furthermore, it would be desirable to add information for additional efficiently run large hospitals to the data set.)

4.5.5 Technical Note

Suppose we perform a regression analysis of n observations by using a regression model that utilizes k independent variables. Let SSE and s denote the unexplained variation and the standard error for the regression model and consider the hat matrix:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

which has n rows and n columns. For $i = 1, 2, \ldots, n$ we define the *leverage value* h_i *of the x values* $x_{i1}, x_{i2}, \ldots, x_{ik}$ *to* be the ith diagonal element of $\mathbf{H}$. It can be shown that

$$h_i = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \quad \textit{where} \quad \mathbf{x}_i' = [1 \;\; x_{i1} \;\; x_{i2} \;\ldots\; x_{ik}]$$

is a row vector containing the values of the independent variables in the ith observation. Also, let $e_i = y_i - \hat{y}_i$ denote the usual residual for observation i. In Section B.11 we show that the standard deviation of e_i is $\sigma_{e_i} = \sigma\sqrt{1-h_i}$, and thus the standard error of e_i (that is, the point estimate of σ_{e_i}) is $s_{e_i} = s\sqrt{1-h_i}$. This implies that the *studentized residual* for observation i equals $e_i / \left(s\sqrt{1-h_i}\right)$. Furthermore, let $d_i = y_i - \hat{y}_{(i)}$ denote the *deleted residual* for observation i, where

$$\hat{y}_{(i)} = b_0^{(i)} + b_1^{(i)}x_{i1} + b_2^{(i)}x_{i2} + \ldots + b_k^{(i)}x_{ik}$$

is the point prediction of y, calculated by using least squares point estimates $b_0^{(i)}, b_1^{(i)}, b_2^{(i)}, \ldots, b_k^{(i)}$ which are calculated by using all n observations except for the ith observation. Also, let s_{d_i} denote the standard error of d_i. Then, it can be shown that the deleted residual d_i and the *studentized deleted residual* d_i / s_{d_i} can be calculated by using the equations

$$d_i = \frac{e_i}{1-h_i} \quad and \quad \frac{d_i}{s_{d_i}} = e_i\left[\frac{n-k-2}{SSE(1-h_i)-e_i^2}\right]^{1/2}$$

Next, if D_i denotes the value of the Cook's D statistic for observation i, then D_i is defined by the equation $D_i = (\mathbf{b}-\mathbf{b}^{(i)})'\mathbf{X}'\mathbf{X}(\mathbf{b}-\mathbf{b}^{(i)})/(k+1)s^2$, where

$$\mathbf{b}-\mathbf{b}^{(i)} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} - \begin{bmatrix} b_0^{(i)} \\ b_1^{(i)} \\ b_2^{(i)} \\ \vdots \\ b_k^{(i)} \end{bmatrix} = \begin{bmatrix} b_0 - b_0^{(i)} \\ b_1 - b_1^{(i)} \\ b_2 - b_2^{(i)} \\ \vdots \\ b_k - b_k^{(i)} \end{bmatrix}$$

and it can be shown that

$$D_i = \frac{e_i^2}{(k+1)s^2}\left[\frac{h_i}{(1-h_i)^2}\right]$$

Moreover, let $g_j^{(i)} = b_j - b_j^{(i)}$. If $s_{g_j^{(i)}}$ denotes the standard error of this difference, then the *difference in estimate of the β_j statistic* is defined to be $g_j^{(i)} / s_{g_j^{(i)}}$. It can be shown that

$$\frac{g_j^{(i)}}{s_{g_j^{(i)}}} = \left[\frac{d_i}{s_{d_i}}\right]\left[\frac{r_{j,i}}{\sqrt{(\mathbf{r}_j'\mathbf{r}_j)(1-h_i)}}\right]$$

Here, $r_{j,i}$, is the element in row j and column i of $\mathbf{R} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and $\mathbf{r}_j'$ is row j of $\mathbf{R}$.

Also, let $f_i = \hat{y}_i - \hat{y}_{(i)}$. If s_{f_i} denotes the standard error of this difference, then the *difference in fits statistic* is defined to be f_i / s_{f_i}. It can be shown that

$$\frac{f_i}{s_{f_i}} = \left[\frac{d_i}{s_{d_i}}\right]\left[\frac{h_i}{1-h_i}\right]^{1/2}$$

4.6 Step 6: Validating the Model

When we have used model comparison techniques and model diagnostics to select one or more potential final regression models, it is important to *validate* the models by using them to analyze a data set that differs from the data set used to build the models. For example, Kutner, Neter, Wasserman, Nachtsheim, and Li (2005) consider 108 observations described by the dependent variable y = survival time (in days) after undergoing a particular liver operation and the independent variables x_1 = blood clotting score, x_2 = prognostic index, x_3 = enzyme function test score, x_4 = liver function test score, x_5 = age (in years), $x_6 = 1$ for a female patient and 0 for a male patient, $x_7 = 1$ for a patient who is a moderate drinker and 0 otherwise, and $x_8 = 1$ for a patient who is a heavy drinker and 0 otherwise. A regression analysis relating y to x_1, x_2, x_3, and x_4 based on 54 observations (the training data) had a residual plot that was curved and fanned out, suggesting the need for a natural logarithm transformation. Using all possible regressions on the 54 observations, the models with the smallest PRESS statistic (the sum of squared deleted residuals), smallest C statistic, and largest $\bar{R}^2$ were the following models 1, 2, and 3 (see Table 4.3):

$$\text{Model 1: } \ln y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_8 x_8 + \varepsilon$$
$$\text{Model 2: } \ln y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_6 x_6 + \beta_8 x_8 + \varepsilon$$
$$\text{Model 3: } \ln y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_8 x_8 + \varepsilon$$

Note that although we did not discuss the PRESS statistic in Section 4.2, it is another useful model building statistic.

Each model was fit to the remaining 54 observations (the validation data) and also used to compute

$$\text{MSPR} = \frac{\sum_{i=1}^{n^*} (y_i' - \hat{y}_i)^2}{n^*}$$

when n^* is the number of observations in the validation data set, y_i' is the value of the dependent variable for the i^{th} observation in the validation data set, and $\hat{y}_i$ is the prediction of y_i' using the training data set model.

Table 4.3 Comparisons of Models 1, 2, and 3

	Model 1 Training	Model 1 Validation	Model 2 Training	Model 2 Validation	Model 3 Training	Model 3 Validation
PRESS	2.7378	4.5219	2.7827	4.6536	2.7723	4.8981
C	5.7508	6.2094	5.5406	7.3331	5.7874	8.7166
s^2	0.0445	0.0775	0.0434	0.0777	0.0427	0.0783
$\bar{R}^2$	0.8160	0.6824	0.8205	0.6815	0.8234	0.6787
MSPR	0.0773	–	0.0764	–	0.0794	–

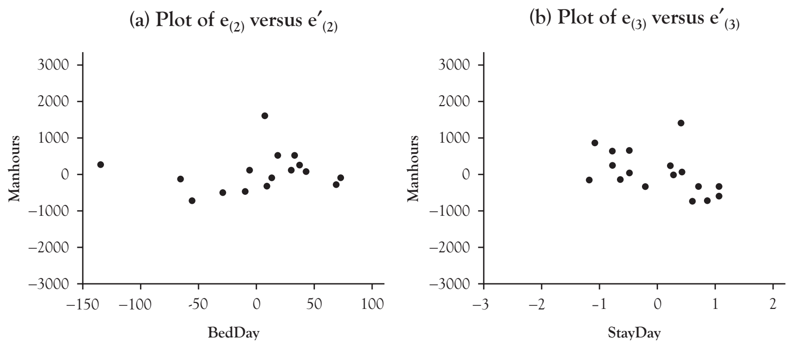


Figure 4.28 Partial leverage residual plots

The values of MSPR for the three above models, as well as the values of PRESS, C, s^2, and $\bar{R}^2$ when the three models are fit to the validation data set, are shown in Table 4.3. Model 3 was eliminated because the sign of the age coefficient changed from a negative $b_5 = -.0035$ to a positive $b_5 = .0025$ as we went from the training data set to the validation data set. Model 1 was chosen as the final model because it had (1) the smallest PRESS for the training data; (2) the smallest PRESS, C, and s^2 for the validation data; (3) the second smallest MSPR; (4) all p-values less than .01 (it was the only model with all p-values less than .10); and (5) the fewest independent variables. The final prediction equation was

$$\widehat{\ln y} = 3.852 + .073x_1 + .0142x_2 + .0155x_3 + .353x_8$$

and thus $\hat{y} = e^{\widehat{\ln y}}$

4.7 Partial Leverage Residual Plots

Suppose that we are attempting to relate the dependent variable y to the independent variables $x_1,\ldots,x_{j-1},x_j,x_{j+1},\ldots,x_k$. Let $b_0, b_1,\ldots,b_{j-1},b_{j+1},\ldots,b_k$ be the least squares point estimates of the parameters in the model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{j-1}x_{j-1} + \beta_{j+1}x_{j+1} + \ldots + \beta_k x_k + \varepsilon$$

and let $b_0', b_1',\ldots,b_{j-1}',b_{j+1}',\ldots,b_k'$ be the least squares point estimates of the parameters in the model

$$x_j = \beta_0' + \beta_1' x_1 + \ldots + \beta_{j-1}' x_{j-1} + \beta_{j+1}' x_{j+1} + \ldots + \beta_k' x_k + \varepsilon$$

Then a *partial leverage residual plot* of

$$e_{(j)} = y - (b_0 + b_1 x_1 + \ldots + \mathrm{b}_{j-1} x_{\mathrm{j}-1} + b_{j+1} x_{j+1} + \ldots + b_k x_k)$$

versus

$$e_{(j)}' = x_j - (b_0' + b_1' x_1 + \ldots + b_{j-1}' x_{j-1} + b_{j+1}' x_{j+1} + \ldots + b_k' x_k)$$

represents a plot of y versus x_j, with the effects of the other independent variables $x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_k$ removed. When strong multicollinearity exists between x_j and the other independent variables, a plot of y versus x_j can reveal an (apparent) significant relationship between y and x_j, while the partial leverage residual plot of $e_{(j)}$ versus $e'_{(j)}$ reveals very little or no relationship between $e_{(j)}$ and $e'_{(j)}$. This is a graphical illustration of the multicollinearity and says that there is very little or no relationship between y and x_j when the effects of the other independent variables are removed. In other words, x_j has little or no importance in describing y over and above the combined importance of the other independent variables. Finally, note that the least squares point estimate of the slope parameter β_j in the simple linear model $e_{(j)} = \beta_0 + \beta_j e_{(j)}' + \varepsilon_{(j)}$ equals the least squares point estimate of the parameter β_j in the model $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j + \ldots + \beta_k x_k + \varepsilon$.

To illustrate partial leverage residual plots, recall that Table 4.2 gives data concerning the need for labor in 17 U.S. Navy hospitals. It can be verified that data plots of y (labor hours) versus x_1 (X-ray exposures), x_2 (BedDays), x_4 (average daily patient load), and x_5 (eligible population) show upward linear relationships. However, in the exercises of this chapter the reader will show that there is extreme multicollinearity between x_2 (BedDays), x_4 (average daily patient load), and x_5 (eligible population). Therefore, the partial leverage residual plots of y versus x_2, x_4, and x_5 do not show much of a relationship. For example, Figure 4.28a is a partial leverage residual plot that shows little relationship between $e_{(2)} = y - (b_0 + b_1x_1 + b_3x_3 + b_4x_4 + b_5x_5)$ and $e'_{(2)} = x_2 - (b'_0 + b'_1x_1 + b'_3x_3 + b'_4x_4 + b'_5x_5)$. In the exercises of this chapter the reader will also show that there is strong (although not extreme) multicollinearity between x_1 (X-ray exposures) and the variables x_2, x_4, and x_5. Correspondingly it can be verified that the partial leverage residual plot of y versus x_1 shows somewhat less of an upward linear relationship than does the usual data plot. Finally, the reader will show in the exercises of this chapter that there is not strong multicollinearity between x_3 (average length of patients' stay) and the other independent variables $\left(x_1, x_2, x_4, \text{ and } x_5\right)$. It can be verified that a data plot shows an upward linear relationship between y and x_3. On the other hand, Figure 4.28b is a partial leverage residual plot that shows a downward linear relationship between $e_{(3)} = y - (b_0 + b_1x_1 + b_2x_2 + b_4x_4 + b_5x_5)$ and $e'_{(3)} = x_3 - (b'_0 + b'_1x_1 + b'_2x_2 + b'_4x_4 + b'_5x_5)$. Moreover, this is consistent with the fact that the point estimate of β_3 in the model $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5 + \varepsilon$ is negative $\left(b_3 = -394.31\right)$. In other words, for two hospitals with the same values of $x_1, x_2, x_4, \textit{and } x_5$ the hospital with a longer average length of patients' stay can be expected to use fewer labor hours, possibly because there is less turnover of patients and thus less initial labor.

4.8 Ridge Regression, the Standardized Regression Model, and a Robust Regression Technique

When strong multicollinearity is present, we can sometimes use *ridge regression* to calculate point estimates that are closer to the true values of the model parameters than are the usual least squares point estimates. We first show how to calculate *ridge point estimates*. Then we discuss the advantage and disadvantages of these estimates.

To calculate the ridge estimates of the parameters in the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

we first consider the *standardized regression model*

$$y_i' = \beta_1' x_{i1}' + \ldots + \beta_k' x_{ik}' + \varepsilon_i'$$

where

$$y_i' = \frac{1}{\sqrt{n-1}}\left(\frac{y_i - \bar{y}}{s_y}\right) \quad \text{and} \quad x_{ij}' = \frac{1}{\sqrt{n-1}}\left(\frac{x_{ij} - \bar{x}_j}{s_{x_j}}\right)$$

Here, $\bar{y}$ and s_y are the mean and the standard derivation of the n observed values of the dependent variable y, and, for $j = 1, 2, \ldots, k$, $\bar{x}_j$ and s_{x_j} are the mean and the standard deviation of the n observed values of the jth independent variable x_j. If we form the matrices

$$\dot{\mathbf{y}} = \begin{bmatrix} y_1' \\ y_2' \\ \cdot \\ \cdot \\ \cdot \\ y_n' \end{bmatrix} \qquad \dot{\mathbf{X}} = \begin{bmatrix} x_{11}' & \cdot & \cdot & \cdot & x_{1k}' \\ x_{21}' & \cdot & \cdot & \cdot & x_{2k}' \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ x_{n1}' & \cdot & \cdot & \cdot & x_{nk}' \end{bmatrix}$$

it can be shown that

$$\dot{\mathbf{X}}'\dot{\mathbf{X}} = \begin{bmatrix} 1 & r_{x_1,x_2} & \cdot & \cdot & \cdot & r_{x_1,x_k} \\ r_{x_2,x_1} & 1 & \cdot & \cdot & \cdot & r_{x_2,x_k} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ r_{x_k,x_1} & r_{x_k,x_2} & \cdot & \cdot & \cdot & 1 \end{bmatrix} \quad \dot{\mathbf{X}}'\dot{\mathbf{y}} = \begin{bmatrix} r_{y,x_1} \\ r_{y,x_2} \\ \cdot \\ \cdot \\ \cdot \\ r_{y,x_k} \end{bmatrix}$$

Because r_{x_j,x'_j} is the simple correlation coefficient between the independent variables x_j and x'_j and r_{y,x_j} is the simple correlation coefficient between the dependent variable y and the independent variable x_j, we say that the above defined quantities y'_i and x'_{ij} are *correlation transformations* of the ith value of the dependent variable y and the ith value of the independent variable x_j.

Ridge Estimation

The *ridge point estimates* of the parameters $\beta_1', \ldots, \beta_k'$ of the standardized regression model are

$$\begin{bmatrix} b'_{1,R} \\ \cdot \\ \cdot \\ \cdot \\ b'_{k,R} \end{bmatrix} = (\dot{\mathbf{X}}' \dot{\mathbf{X}} + c\mathbf{I})^{-1} \dot{\mathbf{X}}' \dot{\mathbf{y}}$$

Here, we use a *biasing constant* $c \geq 0$. Then the ridge point estimates of the parameters $\beta_0, \beta_1, \ldots, \beta_k$ in the original regression model are

$$b_{j,R} = \left(\frac{s_y}{s_{x_j}} \right) b'_{j,R} \qquad j = 1, \ldots, k$$

$$b_{0,R} = \bar{y} - b_{1,R}\bar{x}_1 - b_{2,R}\bar{x}_2 - \ldots - b_{k,R}\bar{x}_k$$

To understand the biasing constant c, first note that if $c = 0$, then the ridge point estimates are the least squares point estimates. Recall that the least squares estimation procedure is unbiased. That is, $\mu_{b_j} = \beta_j$. If $c > 0$, the ridge estimation procedure is not unbiased. That is, $\mu_{b_{j,R}} \neq \beta_j$ if $c > 0$. We define the bias of the ridge estimation procedure to be $\left\{ \mu_{b_{j.R}} - \beta_j \right\}$. To compare a biased estimation procedure with an unbiased estimation procedure, we employ *mean squared errors*. The mean squared error of an estimation procedure is defined to be the average of the squared deviations of the different possible point estimates from the unknown parameter.

This can be proven to be equal to the sum of the *squared bias* of the procedure and the *variance* of the procedure. Here, the variance is the average of the squared deviations of the different possible point estimates from the mean of all possible point estimates. If the procedure is unbiased, the mean of all possible point estimates is the parameter we are estimating. In other words, when the bias is zero, the mean squared error and the variance of the procedure are the same, and thus the mean squared error of the (unbiased) least squares estimation procedure for estimating β_j is the variance $\sigma^2_{b_j}$. The mean squared error of the ridge estimation procedure is

$$[\mu_{b_{j,\mathrm{R}}} - \beta_j]^2 + \sigma^2_{b_{j,\mathrm{R}}}$$

It can be proved that as the biasing constant c increases from zero, the bias of the ridge estimation procedure increases, and the variance of this procedure decreases. It can further be proved that there is some $c > 0$ that makes $\sigma^2_{b_{j,R}}$ so much smaller than $\sigma^2_{b_j}$ that the mean squared error of the ridge estimation procedure is smaller than the mean squared error of the least squares estimation procedure. This is one advantage of ridge estimation. It implies that the ridge point estimates are less affected by multicollinearity than the least squares point estimates. Therefore, for example, they are less affected by small changes in the data. One problem is that the optimum value of *c* differs for different applications and is unknown.

Before discussing how to choose *c*, we note that, in addition to using the standardized regression model to calculate ridge point estimates, some statistical software systems automatically use this model to calculate the usual least squares point estimates. The reason is that when strong multicollinearity exists, the columns of the matrix $\mathbf{X}$ obtained from the usual (multiple) linear regression model are close to being linearly dependent and thus there can be serious rounding errors in calculating $(\mathbf{X'X})^{-1}$. Such errors can also occur when the elements of $\mathbf{X'X}$ have substantially different magnitudes. This occurs when the magnitudes of the independent variables differ substantially. Use of the standardized regression model means that $\dot{\mathbf{X}}'\dot{\mathbf{X}}$ consists of simple correlation coefficients, all elements of which are between –1 and 1. Therefore these elements have the same magnitudes. This can help to eliminate serious rounding errors in calculating $(\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}$ and thus in calculating the least squares point estimates $b_1', \ldots, b_k'$. Of course, the standardized regression model is used to calculate the ridge point estimates for similar reasons.

One way to choose c is to calculate ridge point estimates for different values of c. We usually choose values between 0 and 1. Experience indicates that the ridge point estimates may fluctuate wildly as c is increased slightly from zero. The estimates may even change sign. Eventually, the values of the ridge point estimates begin to change slowly. It is reasonable to choose c to be the smallest value where all of the ridge point estimates begin to change slowly. Here, making a *ridge trace* can be useful. This is a simultaneous plot of the values of all of the ridge point estimates against values of c. Another way to choose c is to note that variance inflation factors related to the ridge point estimates of the parameters in the standardized regression model are the diagonal elements of the matrix

$$(\dot{\mathbf{X}}'\dot{\mathbf{X}} + c\mathbf{I})^{-1}\dot{\mathbf{X}}'\dot{\mathbf{X}}(\dot{\mathbf{X}}'\dot{\mathbf{X}} + c\mathbf{I})^{-1}$$

As c increases from zero, the variance inflation factors initially decrease quickly and then begin to change slowly. Therefore, we might choose c to be a value where the variance inflation factors are sufficiently small. A related way to choose c is to consider the *trace* (the sum of the diagonal elements) of the matrix

$$\mathbf{H}_c = \dot{\mathbf{X}}(\dot{\mathbf{X}}'\dot{\mathbf{X}} + c\mathbf{I})^{-1}\dot{\mathbf{X}}'$$

It can be shown that as c increases from zero, this trace, denoted $tr(H_c)$, initially decreases quickly and then begins to decrease slowly. We might choose c to be the smallest value where $tr(H_c)$ begins to decrease slowly. This is because at this value the multicollinearity in the data begins to have a sufficiently small impact on the ridge point estimates.

One disadvantage of ridge regression is that the choice of c is somewhat subjective. Furthermore, the different ways to choose c often contradict each other. We have discussed only three such methods. Myers (1986) gives an excellent discussion of other methods for choosing c. Another major problem with ridge regression is that the exact probability distribution of all possible values of a ridge point estimate is unknown.

This means that we cannot (easily) perform statistical interference. Ridge regression is very controversial. Our view is that before using ridge regression one should use the various model-building techniques of this book to eliminate severe multicollinearity by identifying redundant independent variables.

As an example of ridge regression, consider the hospital labor needs data in Table 4.2. Table 4.4 shows the ridge point estimates of the parameters in the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$. Here, we have ranged c from 0.00 to 0.20 and also include the values of $tr(H_c)$. Noting the changes in sign in the ridge point estimates, it is certainly not

Table 4.4 The ridge point estimates for the hospital labor needs model

c	$b_{0,R}$	$b_{1,R}$	$b_{2,R}$	$b_{3,R}$	$b_{4,R}$	$b_{5,R}$	$tr(H_c)$
0.00	1962.95	0.0559	1.5896	–394.31	–15.8517	–4.2187	5.0000
0.01	1515.07	0.0600	0.5104	–312.71	14.5765	–2.1732	3.6955
0.02	1122.83	0.0621	0.4664	–236.20	13.5101	0.2488	3.4650
0.03	839.55	0.0634	0.4358	–180.25	12.7104	1.9882	3.2867
0.04	624.89	0.0643	0.4130	–137.25	12.0993	3.2949	3.1427
0.05	456.27	0.0648	0.3951	–102.94	11.6180	4.3098	3.0227
0.06	320.08	0.0652	0.3808	–74.75	11.2286	5.1188	2.9206
0.07	207.65	0.0653	0.3690	–51.05	10.9066	5.7768	2.8320
0.08	113.17	0.0654	0.3591	–30.75	10.6353	6.3209	2.7541
0.09	32.61	0.0654	0.3507	–13.07	10.4031	6.7768	2.6848
0.10	–36.91	0.0654	0.3434	2.50	10.2016	7.1632	2.6225
0.11	–97.52	0.0653	0.3370	16.39	10.0247	7.4937	2.5661
0.12	–150.81	0.0652	0.3313	28.88	9.8679	7.7787	2.5145
0.13	–198.00	0.0651	0.3262	40.21	9.7276	8.0261	2.4671
0.14	–240.04	0.0649	0.3216	50.56	9.6010	8.2422	2.4233
0.15	–277.70	0.0648	0.3175	60.07	9.4860	8.4319	2.3827
0.16	–311.58	0.0646	0.3137	68.85	9.3808	8.5990	2.3447
0.17	–342.18	0.0644	0.3103	77.00	9.2841	8.7469	2.3092
0.18	–369.91	0.0642	0.3071	84.59	9.1948	8.8782	2.2758
0.19	–395.10	0.0640	0.3041	91.69	9.1118	8.9950	2.2443
0.20	–418.03	0.0638	0.3013	98.35	9.0343	9.0992	2.2146

easy to determine the value of c at which they begin to change slowly. We might arbitrarily choose $c = .16$. In contrast, the values of $tr(H_c)$ seem to begin to change slowly at $c = .01$. If we do a finer search by ranging c in increments of .0001 from .0000 to .0010, the values of $tr(H_c)$ begin to change slowly at $c = .0004$. The corresponding ridge point estimates can be calculated to be

$$\begin{array}{lll} b_{0,R} = 2053.33 & b_{1,R} = 12.5411 & b_{2,R} = .0565 \\ b_{3,R} = .6849 & b_{4,R} = -5.4249 & b_{5,R} = -416.09 \end{array}$$

Experience indicates that various criteria for choosing c tend to differ when the data set has one or more observations that are considerably different from the others. Recall from Section 4.5 that we have concluded that hospitals 14, 15, 16, and 17 are considerably larger than hospitals 1 through 13. At any rate, before using the results of ridge regression we should attempt to identify redundant independent variables. The reader will show in the exercises of this chapter that there is extreme multicollinearity between x_2 (BedDays), x_4 (average daily patient load), and x_5 (eligible population) and also that perhaps the best model describing the hospital labor needs data in Table 4.2 is the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$. This model uses only one of x_2, x_4, and x_5 and thus eliminates much multicollinearity. However, the reader will find in the exercises of this chapter that strong multicollinearity still exists in this best model, and thus we could again use ridge regression.

To conclude this section, recall from Section 4.5 that an outlying observation can significantly influence the values of the least squares point estimates. As an alternative to the least squares procedure, which chooses the point estimates that minimize the sum of the squared residuals (differences between the observed and predicted values of the dependent variable), we could dampen the effect of an influential outlier by calculating point estimates that minimize the sum of the absolute values of the residuals.

The reader is referred to Kennedy and Gentle (1980) for a discussion of the computational aspects of such a minimization. Also, note that minimizing the sum of absolute residuals is only one of a variety of *robust regression* procedures. These procedures are intended to yield point estimates that are less sensitive than the least squares point estimates to both outlying observations and failures of the model assumptions. For example, if the populations sampled are not normal but are *heavy tailed,* then we are more likely to obtain a y_i value that is far from the mean y_i value. This value will act much like an outlier, and its effect can be dampened by minimizing the sum of absolute residuals. An excellent discussion of robust regression procedures is given by Myers (1986).

4.9 Regression Trees

Regression trees are a very powerful but conceptually simple method of relating a dependent variable to one or more independent variables without stating a (parameter based) equation relating the dependent variable to the one more independent variables (this is called *nonparametric regression*). Regression trees partition the $(x_1, x_2, \ldots, x_k)$ space into rectangular regions, where each rectangular region has similar y values. Then the mean of the observed y values in each region serves as the prediction of any y value in that region. To illustrate regression trees, we consider an example presented by Kutner, Nachtsheim, Neter, and Li (2005). In this example, we attempt to predict GPA at the end of the freshman year (y) on the basis of ACT entrance test score (x_1) and high school rank (x_2). The data consisted if 705 cases-352 were used for the training data set and 353 for the validation data set. The high school rank was the percentile at which the student graduated in his or her high school graduating class.

In the first step, illustrated in Figure 4.29a, we calculate $\bar{y}$, the average of the 352 GPA's in the training data set. Then we use $\bar{y}$ to calculate

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n} \qquad MSPR = \frac{\sum_{i=1}^{n^*} (y_i' - \bar{y})^2}{n^*}$$

where y_i is the ith GPA among the $n = 352$ GPA's in the training data set and y_i' is the ith GPA among the $n^* = 353$ GPA's in the validation data set. In the second step, we find the dividing point in the $(x_1, x_2) =$ (ACT, H.S. Rank) space that gives the greatest reduction is MSE. As illustrated in Figure 4.29b the dividing point is a high school rank of 81.5, and the new MSE and MSPR are

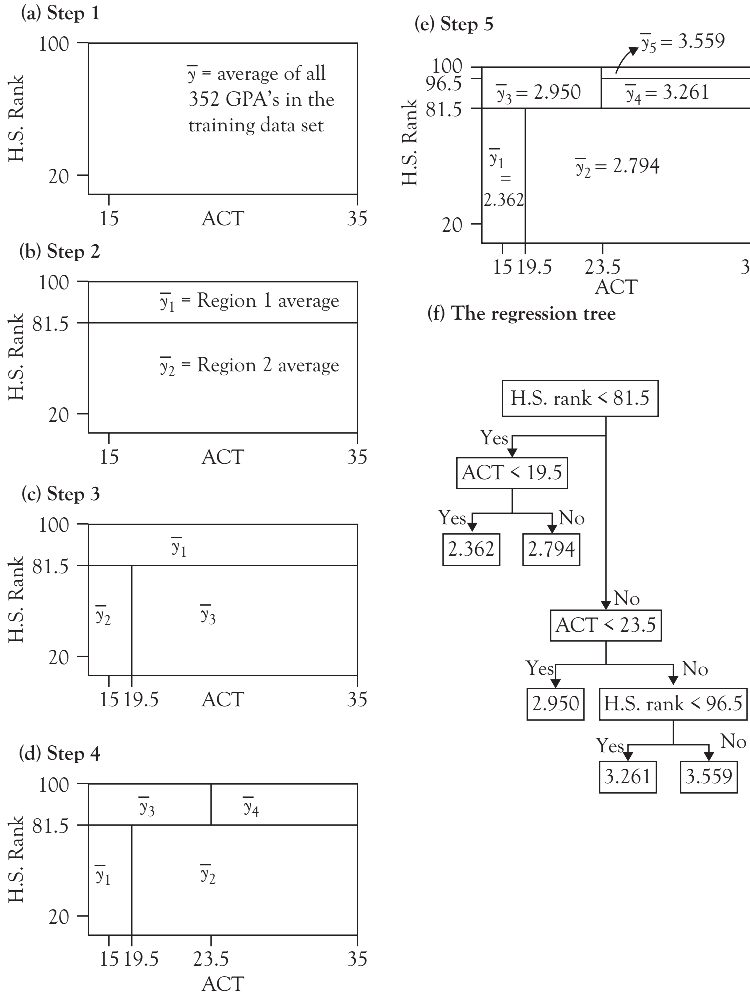


Figure 4.29 Regression tree analysis of the GPA data

$$MSE = \frac{\sum_{i=1}^{n_1}(y_i - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_i - \bar{y}_2)^2}{n}$$

and

$$MSPR = \frac{\sum_{i=1}^{n_1^*}(y_i' - \bar{y}_1)^2 + \sum_{i=1}^{n_2^*}(y_i' - \bar{y}_2)^2}{n}$$

Here, $\bar{y}_1$ is the average of the n_1 GPA's in Region 1 of the training data set and $\bar{y}_2$ is the average of the n_2 GPA's in Region 2 of the training data set. Also, using the high school rank dividing point of 81.5 to divide the validation data set into Region 1 and Region 2, n_1^* denotes the number of GPA's in Region 1 of the validation data set and n_2^* denotes the number of GPA's in Region 2 of the validation data set. As illustrated in Figure 4.29, we continue to find dividing points, where the next dividing point found gives the biggest reduction in MSE. In step 3 the dividing point is an ACT score of 19.5, in step 4 the dividing point is an ACT score of 23.5, and in step 5 the dividing point is a high school rank of 96.5. We could continue to find dividing points indefinitely, until the entire $(x_1, x_2) =$ (ACT, H.S. Rank) space in the training data set is divided into the original 352 GPA's and at each step MSE would decrease. However, there is a step in the dividing process where MSPR will increase, and in this example this occurs when we find the next dividing point after step 5. In general, we stop the dividing process when MSPR increases and use the sample means obtained at the previous step (step 5 in this situation) as the point predictions of the y values in the regions that have been obtained. To make it easy to find the point prediction of a y value in a particular region, statistical software packages present a regression tree such as the one shown in Figure 4.29f.

Using the sample mean predictions given in the regression tree in Figure 4.29f, R^2 for the training data set is .256 and for the validation data set is .157. We conclude that GPA is related to H.S. Rank and ACT, but that the fraction of the variation in GPA explained by the regression tree is not high. If we use parametric regression, our model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon$. This model has an MSE of .333 and an MSPR of .296 as compared to an MSE of .322 and an MSPR of .318 for the regression tree model. Therefore, the regression tree model does about as well as the parametric regression model.

In general, regression trees are useful in exploratory studies when there is an extremely large number of independent variables—as in data mining.

4.10 Using SAS

Figure 4.30 gives the SAS program for making model comparisons using the sales territory performance data in Tables 2.5a and 4.1. Figure 4.31

```
DATA TERR;
INPUT SALES TIME MKTPOTEN ADVER MKTSHARE CHANGE
      ACCTS WKLOAD RATING;
TMP = TIME*MKTPOTEN;
TA = TIME*ADVER;
TMS = TIME*MKTSHARE;
TC = TIME*CHANGE;
MPA = MKTPOTEN*ADVER;
MPMS = MKTPOTEN*MKTSHARE;
MPC= MKTPOTEN*CHANGE;
AMS= ADVER*MKTSHARE;
AC= ADVER*CHANGE;
MSC= MKTSHARE*CHANGE;
SQT= TIME*TIME;
SQMP= MKTPOTEN*MKTPOTEN;
SQA= ADVER*ADVER;
SQMS= MKTSHARE*MKTSHARE;
SQC= CHANGE*CHANGE;
DATALINES;
 3669.88   43.10 74065.11  4582.88  2.51   0.34   24.86  15.05  4.9
 3473.95  108.13 58117.30  5539.78  5.51   0.15  107.32  19.97  5.1
                              .
                              .
 2799.97   21.14 22809.53  3552.00  9.14  -0.74   88.62  24.96  3.9
    .       85.42  35182.73  7281.65  9.64    .28  120.61  15.72  4.5
PROC PLOT;
PLOT SALES*(TIME MKTPOTN ADVER MKTSHARE CHANGE ACCTS WKLOAD RATING);
PROC CORR;
PROC REG;
MODEL SALES = TIME MKTPOTN ADVER MKTSHARE CHANGE ACCTS WKLOAD
RATING/VIF;
PROC REG DATA = TERR;
MODEL SALES=TIME MKTPOTN ADVER MKTSHARE CHANGE ACCTS WKLOAD
RATING/SELECTION=STEPWISE SLENTRY=.10 SLSTAY=.10;
```

(Note: To perform backward elimination with α_{stay} =.10, we would write "SELECTION = BACKWARD SLSTAY = .10")

```
MODEL SALES=TIME MKTPOTN ADVER MKTSHARE CHANGE ACCTS WKLOAD
RATING/SELECTION=RSQUARE RMSE ADJRSQ MSE RMSE CP;
```

(Note: This statement gives all of the one variable models ranked in terms of R^2, then all of the two variable models ranked in terms of R^2, etc. There would be 256 models given. If we added in the statement "BEST = 2" at the end, we would get the two best models of each size ranked in terms of R^2, If after the equal sign following "SELECTION," we started with "ADJRSQ," we would get all 256 models ranked, irrespective of size, in term of $\bar{R}^2$, s^2, and s. If we added in, for example, "BEST = 8," we would get the best 8 models ranked, irrespective of size in terms of $\bar{R}^2$, s^2, and s)

```
MODEL SALES=TIME MKTPOTN ADVER MKTSHARE CHANGE MPMS TMP TA TMS
TC MPA MPC AMS AC MSC SQT AQMP SQA SQMS SQC / SELECTION = RSQUARE RMSE CP
ADJRSQ INCLUDE=5 BEST=1;
```

(Note: This statement gives the single model of each size having the highest R^2, where all five linear independent variables are included in every model).

```
MODEL SALES=TIME MKTPOTN ADVER MKTSHARE CHANGE SQT SQMP
MPMS TA TMS AMS AC / P CLM CLI;
```

Figure 4.30 SAS program for model building using the sales territory performance data

gives the SAS program needed to perform residual analysis and to fit the transformed regression model and a weighted least squares regression model when analyzing the QHIC data in Table 4.7. Figure 4.32 gives the SAS program needed to analyze the hotel room average occupancy data in Figure 4.18. Figure 4.33 gives the SAS program for model building and

```
data qhic;
input value upkeep;
val_sq = value**2;
datalines;

237.00      1412.08
153.08       797.20
         .
         .
122.02       390.16
198.02      1090.84
220               .

Proc reg;
  model upkeep = value val_sq:
   plot r.*value;
  output out = new1 r=resid p = yhat;
```

(Note: This statement places the residuals and the $\hat{y}$ values in a new data set called "new1". The command "r=resid" says that we are giving the name "resid" to the residuals (r). The command "p = yhat" says that we are giving the name "yhat" to the predicted values (p).

```
data new2;
set new1;
 abs_res = abs(resid);
proc plot;
plot abs_res*value;
proc reg;
 model abs_res = value;
 Output out = new3 p = shat;
  proc print;
 var shat;
data new4;
  set new3;
 y_star = upkeep/shat;
 inv_pabe = 1/shat;
 value_star = value/shat;
 val_sq_star = val_sq/shat;
 wt = shat**(-2);
proc reg;
model y_star = inv_pabe value_star val_sq_star / noint clm cli;
 plot r.*p.;
proc reg;
model upkeep = value val_sq / clm cli;
weight wt;
plot r.*p.;
```

Figure 4.31 SAS program for analyzing the QHIC Data

residual analysis and for detecting outlying and influential observations using the hospital labor needs data in Table 4.2 and values of the dummy variable D_L which equals 1 for large hospitals 14, 15, 16, and 17 and equals 0 otherwise. Figure 4.34 gives the SAS program for fitting the nonlinear regression model $y = \beta_1 + \beta_2 e^{-\beta_3 x} + \varepsilon$ to the light data in Figure 4.16.

```
DATA OCCUP;
INPUT Y M;
IF M = 1 THEN M1 = 1;
ELSE M1 = 0;
IF M = 2 THEN M2 = 1;
ELSE M2 = 0;
        •
        •
IF M = 11 THEN M11 = 1;
ELSE M11 = 0;
TIME = _N_;
LNY = LOG(Y);
SRY = Y**.5;
QRY = Y**.25;
PROC PLOT;
PLOT y*TIME;
PLOT LNY*TIME;
PLOT SRY*TIME;
PLOT QRY*TIME;
DATLINES;
501 1
488 2
    •
    •
877 12
 .  1
 .  2
    •
    •
 .  12
```

Defines the dummy variables $M_1, M_2, \ldots, M_{11}$

Hotel room average occupancy data

Predicts next year's monthly room averages

```
PROC REG DATA = OCCUP;
MODEL QRY = TIME M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11/ P DW CLM CLI;
```

(These statements fit the quartic root room average model assuming independent errors and calculate the Durbin-Watson statistic.)

```
PROC AUTOEG DATA=OCCUP;
 MODEL QRY = TIME M1 M2 M3 M4 M5 M6 M7 M8 M9
             M10 M11/ NLAG = 18 BACKSTEP SLSTAY=.15;
```

(These statements perform backward elimination on the quartic root room average model residuals with q=18 and $\alpha_{\text{stay}} = .15$. Recall that the error term model chosen is $\varepsilon_t = \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3} + \phi_{12} \varepsilon_{t-12} + \phi_{18} \varepsilon_{t-18}$. The following commands fit the quartic root room average model combined with this error term model.)

```
PROC ARIMA DATA = OCCUP;
IDENIFY VAR = QRY CROSSCOR = (TIME M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11)
NOPRINT;
ESTIMATE INPUT = (TIME M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11)
P = (1,2,3,12,18) PRINTALL PLOT;
FORECAST LEAD = 12 OUT = FCAST3;
DATA FORE3;
SET FCAST3;
Y = QRY**4;
FY = FORECAST**4;
L95CI = L95**4;
U95CI = U95**4;
PROC PRINT DATA = FORE3;
VAR Y L95CI FY U95CI;
```

Figure 4.32 SAS program to analyze the hotel room average occupancy data

```
DATA HOSP;
INPUT Y X1 X2 X3 X4 X5 D;
DATALINES;
566.52 2463 472.92 4.45 15.57 18.0 0
696.82 2048 1339.75 6.92 44.02 9.5 0
                   .
                   .
4026.52 15543 3865.67 5.50 127.21 126.8 0
10343.81 36194 7684.10 7.00 252.90 157.7 1
11732.17 34703 12446.33 10.78 409.20 169.4 1
15414.94 39204 14098.40 7.05 463.70 331.4 1
18854.45 86533 15524.00 6.35 510.22 371.6 1
. 56194 14077.88 6.89 456.13 351.2 1
PROC PRINT;
PROC CORR;
PROC PLOT;
PLOT Y * (X1 X2 X3 X4 X5 D);
PROC REG;
MODEL Y = X1 X2 X3 X4 X5 D / VIF;
PROC REG;
MODEL Y = X1 X2 X3 X4 X5 D / SELECTION = RSQUARE ADJRSQ
MSE RMSE CP;
MODEL Y = X1 X2 X3 X4 X5 D / SELECTION = STEPWISE
SLENTRY = .10 SLSTAY = .10;
PROC REG;
MODEL Y = X1 X2 X3  D / P R INFLUENCE CLM CLI VIF;     } Detects outlying and influential observations
OUTPUT OUT = ONE PREDICTED = YHAT RESIDUAL = RESID;    } Constructs residual and normal plots
PRC PLOT DATA = ONE;
PLOT RESID * (X1 X2 X3 D YHAT);
PROC UNIVARIATE PLOT DATA = ONE;
VAR RESID;
RUN;
```

Figure 4.33 SAS program for model building and residual analysis and for detecting outlying and influential observations using the hospital labor needs data

4.11 Exercises

Exercise 4.1

Suppose that the United States Navy wishes to develop a regression model based on efficiently run Navy hospitals to evaluate the labor needs of questionably run Navy hospitals. Table 4.2, which has been given in Section 4.5, gives labor needs data for 17 Navy hospitals. Specifically, this table gives values of the dependent variable Hours (y, monthly labor hours required) and of the independent variables X-ray (x_1, monthly X-ray exposures), BedDays (x_2, monthly occupied bed days—a hospital has one occupied bed day if one bed is occupied for an entire day), Length (x_3, average length of patients'

```
DATA TRANSMIS;
INPUT CHEMCON LIGHT;
DATALINES;
0.0   2.86
0.0   2.64
1.0   1.57        light data in Section 4.3
  .
  .
  .
5.0   0.36
PROC NLIN;        NLIN is SAS's nonlinear regression procedure
PARAMETERS BETA1 = 0 BETA2 = 2.77 BETA3 = .774;
MODEL LIGHT = BETA1 + BETA2*EXP(-BETA3*CHEMCON);
```

***Figure 4.34** SAS program for fitting the nonlinear regression model $y = \beta_1 + \beta_2 e^{-\beta_3 x} + \varepsilon$ to the light data*

stay, in days), Load (x_4, average daily patient load), and Pop (x_5, eligible population in the area, in thousands). Figure 4.35 gives MINITAB and SAS outputs of multicollinearity analysis and model building for these data.

(a) Discuss why Figure 4.35a and 4.35b indicate that BedDays, Load, and Pop are most strongly involved in multicollinearity. Note that the negative coefficient (that is, least squares point estimate) of $b_3 = -394.3$ for Length might be intutively reasonable because it might say that, when all other independent variables remain constant, an increase in average length of patients' stay implies less patient turnover and thus fewer start-up hours needed for the initial care of new patients. However, the negative coefficients for Load and Pop do not seem to be intuitively reasonable—another indication of extremely muticollnearity. The extremely strong multicollinearity between BedDays, Load, and Pop implies that we may not need all three in a regression model.

(b) Which model has the highest adjusted R^2, smallest C statistic, and smallest s?

(c) (1) Which model is chosen by stepwise regression in Figure 4.35? (2) If we start with all five potential independent variables and use backward elimination with an α_{stay} of .10, the procedure removes (in order) Load and Pop and then stops. Which model is chosen by backward elimination? (3) Discuss why the model that uses Xray,

(a) MINITAB output of a correlation matrix

	Xray	BedDays	Length	Load	Pop
BedDays	0.907				
	0.000				
Length	0.447	0.671			
	0.072	0.003			
Load	0.907	1.000	0.671		
	0.000	0.000	0.003		
Pop	0.910	0.933	0.463	0.936	
	0.000	0.000	0.061	0.000	
Hours	0.945	0.986	0.579	0.986	0.940
	0.000	0.000	0.015	0.000	0.000

(b) MINITAB output of the variance inflation factors

Predictor	Coef	SE Coef	T	P	VIF
Constant	1963	1071	1.83	0.094	
Xray	0.05593	0.02126	2.63	0.023	7.9
BedDays	1.590	3.092	0.51	0.617	8933.1
Length	-394.3	209.6	-1.88	0.087	4.3
Load	-15.85	97.65	-0.16	0.874	9597.6
Pop	-4.219	7.177	-0.59	0.569	23.3

(C) The SAS output of the best five models

Adjusted R-Square Selection Method

Number in Model	Adjusted R-Square	R-Square	c(p)	Root MSE	Variables in Model
3	0.9878	0.9901	2.9177	614.77942	Xray BedDays Length
4	0.9877	0.9908	4.0263	615.48868	Xray BedDays Length Pop
4	0.9875	0.9906	4.2643	622.09422	Xray Length Load Pop
4	0.9874	0.9905	4.3456	624.33413	Xray BedDays Length Load
3	0.9870	0.9894	3.7142	634.99196	Xray Length Load

Figure 4.35 MINITAB and SAS output of muticollinearity and model building for the hospital labor needs data in Table 4.2

BedDays, and Length seems to be the overall best model. (4) Which of BedDays, Load, and Pop does this best model use?

(d) Consider a questionable hospital for which Xray = 56,194, BedDays = 14,077.88, Length = 6.89, Load = 456.13, and Pop = 351.2. The least squares point estimates and associated p-values (given in parentheses) of the parameters in the best model, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, are $b_0 = 1523.3892(.0749)$, $b_1 = .05299(.0205)$, $b_2 = .97898(< .0001)$ and $b_3 = -320.9508(.0563)$. Using this model, a point prediction of and a 95 percent prediction interval for the labor hours, y_0, of an efficiently run hospital having the same

```
Step                 1        2         3
Constant        -28.13   -68.31   1523.39

BedDays          1.117    0.823     0.978
T-value          22.90     9.92      9.31
p-value          0.000    0.000     0.000

Xray                      0.075     0.053
T-value                    3.91      2.64
p-value                   0.002     0.021

Length                               -321
T-value                             -2.10
p-value                             0.056

S                  958      685       615
R-Sq             97.22    98.67     99.01
R-Sq (adj)       97.03    98.48     98.78
Mallows C-P       20.4      4.9       2.9
```

Figure 4.36 MINITAB output of a stepwise regression of the hospital labor needs data $\left(\alpha_{\text{entry}} = \alpha_{\text{stay}} = .10\right)$

values of the independent variables as the questionable hospital are 16,065 and [14,511, 17,618]. Show how the point prediction has been calculated. If y_0 turned out to be 17,821.65, what would you conclude? If y_0 turned out to be 17,207.31 what would you conclude?

(e) The variance inflation factors for the independent variables $x_1, x_2,$ and x_3 in the best model can be calculated to be 7.737, 11.269, and 2.493. Compare the multicollinearity situation in the best model with the multicollinearity situation in the model using all five independent variables.

Exercise 4.2

Table 4.5 shows data concerning the time, y, required to perform service (in minutes) and the number of laptop computers serviced, x, for 15 service calls. Figure 4.37 shows that the y values tend to increase in a straight line fashion and with increasing variation as the x values increase. If we fit the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ to the data, the model's residuals fan out as x increases (we do not show the residual

Table 4.5 The laptop service time data

Service Time, y	Laptops Serviced, x
92	3
63	2
126	6
247	8
49	2
90	4
119	5
114	6
67	2
115	4
188	6
298	11
77	3
151	10
27	1

plot), indicating a violation of the constant variance assumption. A plot of the absolute values of the model's residuals versus x can be verified to have a straight line appearance, and we obtain the prediction equation $pabe_i = -8.06688 + 6.49919x_i$, which gives the predicted absolute residuals shown in Figure 4.38. Figures 4.39 and 4.40 are partial SAS outputs that are obtained when we use both least squares to fit the transformed regression model $y_i / pabe_i = \beta_0(1/pabe_i) + \beta_1(x_i / pabe_i) + \eta_i$ and weighted least squares to fit the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ to the laptop service time

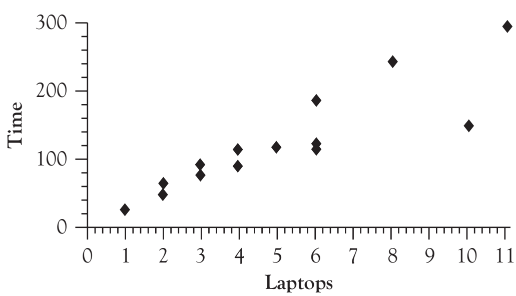


Figure 4.37 Plot of the laptop service time data

Obs	$Pabe_i$	Obs	$Pabe_i$
1	11.4307	9	4.9315
2	4.9315	10	17.9299
3	30.9283	11	30.9283
4	43.9267	12	63.4243
5	4.9315	13	11.4307
6	17.9299	14	56.9251
7	24.4291	15	-1.5677
8	30.9283	16	37.4275

Figure 4.38 SAS output of the $pabe_i$'s

Variable	DF	Parameter Estimate	Standard Error	t value	Pr>|t|
inv_pabe	1	1.66902	3.52841	0.47	0.6440
laptops_star	1	26.57951	2.23770	11.88	<.0001

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict	
16	.	5.0157	0.3401	4.2809	5.7506	2.0768	7.9546

Figure 4.39 Partial SAS output when using least squares to fit the tranformed model $y_i\ /\ pabe_i\ = \beta_0\left(1/\ pabe_i\right)\ + \beta_1\left(x_i\ /\ pabe_i\right)\ + n_i$

Variable	DF	Parameter Estimate	Standard Error	t value	Pr>|t|
Intercept	1	1.66902	3.52841	0.47	0.6440
laptops	1	26.57951	2.23770	11.88	<.0001

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict	
16	.	187.7256	12.7308	160.2224	215.2288	77.7288	297.7224

Figure 4.40 *Partial SAS output when using weighted least squares to fit the orginal model $y_i = \beta_0\ + \beta_1 x_i + \varepsilon_i$*

data. Observation 16 on the SAS output represents a future service call on which seven laptop computers will be serviced. The predicted absolute residual for such a service call is $pabe_0 = -8.06688 + 6.49919(7) = 37.4275$, as shown in Figure 4.38.

(a) Show how the predicted service time $\hat{y}_0 / 37.4275 = 5.0157$ in Figure 4.39 and the predicted service time $\hat{y}_0 = 187.7256$ in Figure 4.40 have been calculated by SAS.

(b) Letting μ_0 represent the mean service time for all service calls on which seven laptops will be serviced, Figure 4.39 says that a 95 percent confidence interval for $\mu_0 / 37.4275$ is [4.2809, 5.7506], and Figure 4.40 says that a 95 percent confidence interval for μ_0 is [160.2224, 215.2288]. If the number of minutes we will allow for the future service call is the upper limit of the 95 percent confidence interval for μ_0, how many minites will we allow?

Exercise 4.3

Western Steakhouses, a fast-food chain, opened 15 years ago. Each year since then the number of steakhouses in operation, y, was recorded. An analyst for the firm wishes to use these data to predict the number of steakhouses that will be in operation next year. The data are given in Table 4.6, and a plot of the data is given in Figure 4.41. Examining the data plot, we see that the number of steakhouses in operation has increased over time at an increasing rate and with increasing variation. A plot of the natural logarithms of the steakhouse values versus time (see Figure 4.42) has a

Table 4.6 The steakhouse data

Year, t	Steakhouses, y
1	11
2	14
3	16
4	22
5	28
6	36
7	46
8	67
9	82
10	99
11	119
12	156
13	257
14	284
15	403

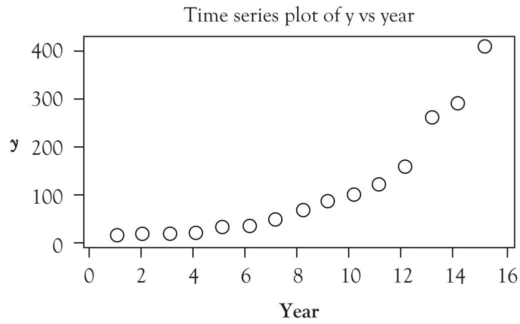


Figure 4.41 Number of steakhouses in operation versus year

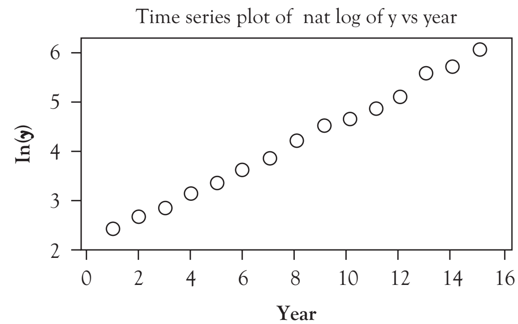


Figure 4.42 Logged steakhouses versus year

straight-line appearance with constant variation. Therefore, we consider the model $\ln y_t = \beta_0 + \beta_1 t + \varepsilon_t$. 1f we use MINITAB, we find that the least squares point estimates of β_0 and β_1 are $b_0 = 2.07012$ and $b_1 = .256880$. We also find that a point prediction of and a 95 percent prediction interval for the natural logarithm of the number of steakhouses in operation next year (year 16) are 6.1802 and [5.9945, 6.3659].

(a) Use the least squares point estimates to calculate the point prediction.
(b) By exponentiating the point prediction and prediction interval–that is by calculating $e^{6.1802}$ and $[e^{5.9945}, e^{6.3659}]$—find a point prediction of and a 95 percent prediction interval for the number of steakhouses in operation next year.
(c) The model $\ln y_t = \beta_0 + \beta_1 t + \varepsilon_t$ is called a growth curve model because it implies that $y_t = e^{(\beta_0 + \beta_1 t + \varepsilon_t)} = (e^{\beta_0})(e^{\beta_1 t})(e^{\varepsilon_t}) = \alpha_0 \alpha_1^t \eta_t$

where $\alpha_0 = e^{\beta_0}$, $\alpha_1 = e^{\beta_1}$ and $\eta_t = e^{\varepsilon_t}$. Here $\alpha_1 = e^{\beta_1}$ is called the *growth rate* of the y values. Noting that the least squares point estimate of β_1 is $b_1 = .256880$, estimate the growth rate α_1.

(d) We see that $y_t = \alpha_0 \alpha_1^t \eta_t = (\alpha_0 \alpha_1^{t-1}) \alpha_1 \eta_t \approx (y_{t-1}) \alpha_1 \eta_t$. This says that y_t is expected to be approximately α_1 times y_{t-1}. Noting this, interpret the growth rate of part (c).

Exercise 4.4

In Section 4.4 we used $\hat{\varepsilon}_{166}$ to help compute a point prediction of $y_{169}^{.25}$, the quartic root of the hotel room average in period 169. Calculate $\hat{\varepsilon}_{166}$.

Exercise 4.5

In Exercise 4.1 you concluded that the best model describing the hospital labor needs data in Table 4.2 is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$. In Section 4.5 we concluded using the studentized deleted residual that hospital 14 is an outlier with respect to its y value. Option 1 for dealing with this outlier is to remove hospital 14 from the data and fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ to the remaining 16 observations. Option 2 is to fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 D_L + \varepsilon$ to all 17 observations. Here, $D_L = 1$ for the larger hospitals 14 to 17 and 0 otherwise.

(a) (1) Use the studentized deleted residuals in Figure 4.26a (see Option 1 Rstudent and Option 2 Rstudent) to see if there are any outliers with respect to their y values when using Options 1 and 2. (2) Is hospital 14 an outlier with respect to its y value when using Option 2? (3) Consider a questionable large hospital ($D_L = 1$) for which Xray = 56.194, BedDays = 14,077.88, and Length = 6.89. Also, consider the labor needs in an efficiently run large hospital described by this combination of values of the independent variables. The 95 percent prediction intervals for these labor needs given by the models of Options 1 and 2 are, respectively, [14,906, 16,886] and [15,175, 17,030]. By comparing these prediction intervals, by analyzing the residual plots for Options 1 and 2 given in Figure 4.26c and 4.26d, and by using your conclusions regarding the studentized deleted residuals, recommend which option should be used. (4) What would you conclude if the questionable large hospital used 17,821.65 monthly labor hours? If it used 17,207.31 monthly labor hours?

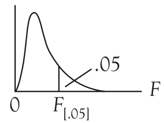
(b) When we remove hospital 14 from the data set and compare all possible regression models, we find that, although the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$ has a slightly smaller s than the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, this latter model has a smaller value of C and gives a slightly shorter 95 percent prediction interval for the monthly labor needs of the questionable hospital. This justifies using the latter model when using Option 1. If we add the dummy variable D_L to the data set and compare all possible regression models using all 17 observations, we find that the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 D_L + \varepsilon$, which is used in Option 2, is the "best model". Justify this conclusion and perform all relevant diagnostic checks by using a statistical software system. Note: The SAS program for doing this is given in Figure 4.33.

APPENDIX A

Statistical Tables

Table A1: An *F* table: Values of $F_{[.05]}$
Table A2: A *t*-table: Values of $t_{[\gamma]}$
Table A3: A table of areas under the standard normal curve
Table A4: Critical values for the Durbin—Watson *d* statistic (α = .05)

Table A1. An F table: Values of $F_{[.05]}$



Denominator Degrees of Freedom (df_2)

df_2 \ df_1	Numerator degrees of freedom (df_1)																	
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.5	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.32	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07

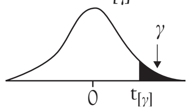
(*Continued*)

Table A1. An F table: Values of $F_{[.05]}$ (Continued)

Denominator Degrees of Freedom (df_2) \ df_1	Numerator degrees of freedom (df_1)																		
df_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Source: Reproduced by permission from Merrington and Thompson (1943) © by the Biometrika Trustees.

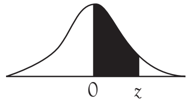
Table A2. A t-table: Values of $t_{[\gamma]}$



df	$t_{[.10]}$	$t_{[.05]}$	$t_{[.025]}$	$t_{[.01]}$	$t_{[.005]}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
inf.	1.282	1.645	1.960	2.326	2.576

Source: Reproduced by permission from Merrington (1941) © by the Biometrika Trustees.

Table A3. Standard normal distribution areas



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.49865	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
4.0	.4999683									

Source: Neter, Wasserman, and Whitmore (1972).

Table A4. Critical values for the Durbin–Watson d statistic (α =.05)

	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
n	$d_{L,.05}$	$d_{U,.05}$	$d_{L,.05}$	$d_{U,.05}$	$d_{L,.05}$	$d_{U,.05}$	$d_{L,.05}$	$d_{U,.05}$	$d_{L,.05}$	$d_{U,.05}$
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	Σ1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77

(Continued)

	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
n	$d_{L,.05}$	$d_{U,.05}$	$d_{L,.05}$	$d_{U,.05}$	$d_{L,.05}$	$d_{U,.05}$	$d_{L,.05}$	$d_{U,.05}$	$d_{L,.05}$	$d_{U,.05}$
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

Source: Reproduced by permission from Durbin and Waston (1951) © by the Biometrika Trustees.

References

Andrews, R.L., and S.T. Ferguson. 1986. "Integrating Judgment With a Regression Appraisal." *The Real Estate Appraiser and Analyst* 52, no. 2, pp. 71–74.

Bowerman, B.L., R.T. O'Connell, and A.B. Koehler. 2005. *Forecasting, Time Series, and Regression.* 4th ed. Belmont, CA: Brooks Cole.

Cravens, D.W., R.B. Woodruff, and J.C. Stomper. January, 1972. "An Analytical Approach for Evaluation of Sales Territory Performance." *Journal of Marketing* 36, no. 1, pp. 31–37.

Dielman, T. 1996. *Applied Regression Analysis for Business and Economics*. Belmont, CA: Duxbury Press.

Durbin,J., and G.S. Waston. 1951. "Testing for Serial Correlation in Least Squares Regression, II." *Biometrika* 30, pp. 159–178.

Freund, R.J., and R.C. Littell. 1991. *SAS System for Regression.* 2nd ed. Cary, NC: SAS Institute Inc.

Kennedy, W.J., and J.E. Gentle. 1980. *Statistical Computing*. New York, NY: Dekker.

Kutner, M.H., C.S. Nachtsheim, J. Neter, and W. Li. 2005. *Applied Linear Statistical Models*. 5th ed. Burr Ridge, IL: McGraw. Hill, Irwin.

Mendenhall, W., and T. Sincich. 2011. *A Second Course in Statistics: Regression Analysis.* 7th ed. Upper Saddle River, NJ: Prentice Hall.

Merrington, M. 1941. "Table of Percentage Points of the *t*-Distribution." *Biometrika* 32, p. 300.

Merrington, M., and Thompson, C.M. April, 1943. "Tables of Percentage Points of the Inverted Beta *(F)*-Distribution." *Biometrika* 33, no. 1, pp. 73–88.

Myers, R. 1986. *Classical and Modern Regression with Applications*. Boston, MA: Duxbury Press.

Neter, J., W. Wasserman, and G.A. Whitmore. 1972. *Fundamental Statistics for Business and Economics.* 4th ed. Boston, MA: Allyn & Bacon, Inc.

Ott, R.L. 1984. *An Introduction to Statistical Methods and Data Analysis.* 2nd ed. Boston, MA: Duxbury Press.

Ott, R.L., and M.L. Longnecker. 2010. *An Introduction to Statistical Methods and Data Analysis*. 6th ed. Belmont, CA: Brooks/Cole.

Index

Adjusted coefficient of determination, 56–57
Autocorrelated errors, 208–216
Autoregressive model, 211

Backward elimination, 172–174, 211
Biasing constant, 231
Bonferroni procedure, 134
Box-Jenkins methodology, 216

Causal variable, 206
Chi-square *p*-values, 212
Cochran-Orcutt procedure, 215
Coefficients of determination, 52–60
Conditional least squares method, 216
Confidence intervals, 81–89, 90
Constant variance assumption, 44, 181–184
Correct functional form, assumption of, 184–187
Correlation, 57–60
Correlation matrix, 159
Cross-sectional data, 2
C-statistic, 169
Curvature, rate of, 98

Dependent (response) variable, 7
  fractional power transformations of, 194–197
Distance value, 82
Dummy variables, 110–123, 137
Durbin–Watson d statistic, 258–259
Durbin–Watson statistic, 209, 210
Durbin–Watson test, 208–216

Error term, 11
Experimental region, 18, 24
Explained deviation, 53

First-order autocorrelation, 208
F table, 254–255

Gauss–Markov theorem, 76
General logistic regression model, 136

Handling unequal variances, 188–194
Hildreth-Lu procedure, 216

Independence assumption, diagnosing and remedying violations of, 45
  autocorrelation, 202–208
  Durbin–Watson test, 208–216
  modeling autocorrelated errors, 208–216
  seasonal patterns, 202–208
  trend, 202–208
Independent (predictor) variable, 7, 74
Indicator variables, 111
Individual *t* tests, 66–69
  population correlation coefficient, 78–81
  simple linear regression model, 77–78
  using p-*value,* 72–76
  using rejection point, 69–72
Individual value
  point prediction of, 18
  prediction interval for, 86
Interaction model, 120
Interaction terms, 97–110, 174–177
Interaction variables, 101
Intercept, 23
Inverse prediction in simple linear regression, 89–91

Lack of fit test, 197–202
Least squares line, 13, 14
Least squares plane, 36

Least squares point estimates, 12–20
using matrix algebra, 27–43
Least squares prediction equation, 17
Leverage values, 217–220
Linear combination of regression parameters, 130–133
Linear regression model, 26–27, 98
assumptions for, 44–45
Line of means, 10
Logistic regression, 135–142

Matrix algebra, least squares point estimates using, 27–43
Maximum likehood estimation, 136
Maximum likelihood method, 216
Mean square error, 48–52, 232
Mean value, confidence interval for, 85
Measures of variation, 52–56
Model assumptions, 44–47
Model building, 159–251
with squared and interaction terms, 174–177
Model comparison statistics, 165–177
Model diagnostics, 159–251
Multicollinearity, 159–165
Multiple linear regression model, 20–26

Negative autocorrelation, 207
No-interaction model, 121
Nonlinear regression, 197–202
Nonparametric regression, 236
Normality assumption, 45, 187–188

Outlying and influential observations
Cook's D, Dfbetas, and Dffits, 222–224
dealing with outliers, 221
leverage values, 217–220
studentized residuals and studentized deleted residuals, 220–221
Overall *F*-test, 60–66
Overall regression relationship, 63

Parabola, equation of, 97
Parallel slopes, 127
Parallel slopes model, 121
Partial coefficient of correlation, 129
Partial coefficient of determination, 129
Partial F-test, 123–131
Partial leverage residual plots, 228–229
Plane of means, 25
Point estimation, 39
Point prediction, 39
Population correlation coefficient, 78–81
Positive autocorrelation, 207
Prediction error, 14
Prediction intervals, 81–89, 90
p-value, 64–66, 72–76

Quadratic regression model, 97
Qualitative independent variable, 110
Quartic root transformation, 194

Regression analysis
cross-sectional data, 2
experimental data, 1
objectives of, 1–3
observational data, 1
qualitative independent variable, 2
quantitative independent variable, 2
Regression assumptions, 18, 45
Regression assumptions, diagnosing and remedying violations of
constant variance assumption, 181–184
correct functional form, assumption of, 184–187
dependent variable, fractional power transformations of, 194–197
handling unequal variances, 188–194
nonlinear regression, 197–202
normality assumption, 187–188
residual analysis, 178–181
weighted least squares, 188–194
Regression model, geometric interpretation of, 24–26
Regression parameters, 12, 23, 75, 97
statistical inference for linear combination of, 130–133

Regression through the origin, 92
Regression trees, 236–239
Rejection point, 69–72
Residual analysis, 178–181
Residual error, 14
Residual plots, 178, 206
Response/dependent variable, 1
Ridge regression, 229–236
Robust regression procedures, 235
Robust regression technique, 229–236

Sampling, 47–48
Scatter diagram/scatter plot, 5, 7
Seasonal dummy variables, 205
Seasonal variations, 202
Shift parameter, 98
Simple coefficient of determination, 57–60
Simple linear regression, 9, 11–12
  inverse prediction in, 89–91
Simple linear regression model, 5–12
Simultaneous confidence intervals, 133–135
Squared and interaction terms, 97–110, 174–177
Square root transformation, 194
Standard error, 48–52
Standardized regression model, 229–236
Standard normal distribution areas, 257
Stepwise regression, 172–174

t-distribution, 69
Time series data, 2
Time series variables, 206
Total deviation, 53
Total mean squared error, 169, 170
t statistics, 162
t-table, 256

Unbiased least squares point estimates, 47–48
Unconditional least squares method, 216
Unequal slopes model, 120
Unexplained deviation, 53

Validating model, 226–227
Variance inflation factors, 161, 162

Wald Chi-Square, 142
Weighted least squares, 188–194

OTHER TITLES IN QUANTITATIVE APPROACHES TO DECISION MAKING COLLECTION

Donald Stengel, California State University, Fresno, Editor

- *Working With Sample Data: Exploration and Inference* by Priscilla Chaffe-Stengel and Donald N. Stengel
- *Business Applications of Multiple Regression* by Ronny Richardson
- *Operations Methods: Waiting Line Applications* by Ken Shaw
- *Regression Analysis: Understanding and Building Business and Economic Models Using Excel* by J. Holton Wilson, Barry P. Keating and Mary Beal-Hodges
- *Forecasting Across the Organization* by Ozgun Caliskan Demirag, Diane Parente and Carol L. Putman
- *Service Mining: Framework and Application* by Wei-Lun Chang

FORTHCOMING IN THIS COLLECTION

- *Effective Applications of Statistical Process Control* by Ken Shaw
- *Leveraging Business Analysis for Project Success* by Vicki James
- *Project Risk: Concepts, Process, and Tools* by Tom R. Wielicki and Donald N. Stengel
- *Effective Applications of Supply Chain Logistics* by Ken Shaw

THE BUSINESS EXPERT PRESS DIGITAL LIBRARIES

EBOOKS FOR BUSINESS STUDENTS

Curriculum-oriented, born-digital books for advanced business students, written by academic thought leaders who translate real-world business experience into course readings and reference materials for students expecting to tackle management and leadership challenges during their professional careers.

POLICIES BUILT BY LIBRARIANS

- *Unlimited simultaneous usage*
- *Unrestricted downloading and printing*
- *Perpetual access for a one-time fee*
- *No platform or maintenance fees*
- *Free MARC records*
- *No license to execute*

The Digital Libraries are a comprehensive, cost-effective way to deliver practical treatments of important business issues to every student and faculty member.

Regression Analysis

Unified Concepts, Practical Applications, and Computer Implementation

Bruce L. Bowerman • Richard T. O'Connell • Emily S. Murphree

This book is a concise and innovative book that gives a complete presentation of applied regression analysis in approximately one-half the space of competing books. With only the modest prerequisite of a basic (non-calculus) statistics course, this text is appropriate for the widest possible audience.

After a short chapter, Chapter 1, introducing regression, this book covers simple linear regression and multiple regressions in a single cohesive chapter, Chapter 2, by efficiently integrating the discussion of these two techniques. Chapter 2 also makes learning easier for students of all backgrounds by teaching the necessary statistical background topics (for example, hypothesis testing) and the necessary matrix algebra concepts as they are needed in teaching regression. Chapter 3 continues the integrative approach of the text by giving a unified presentation of more advanced regression models, including models using squared and interaction terms, models using dummy variables, and logistic regression models.

The book concludes with Chapter 4, which organizes the techniques of model building, model diagnosis, and model improvement into an easy to understand six step procedure.

Bruce L. Bowerman is professor emeritus of decision sciences at Miami University in Oxford, Ohio. He received his PhD degree in statistics from Iowa State University in 1974 and has over forty years of experience teaching basic statistics, regression analysis, time series forecasting, and other courses. He has been the recipient of an Outstanding Teaching award from his students at Miami and an Effective Educator award from the Richard T. Farmer School of Business Administration at Miami.

Richard T. O'Connell is professor emeritus of decision sciences at Miami University, Oxford, Ohio. He has more than 35 years of experience teaching basic statistics, regression analysis, time series forecasting, quality control, and other courses. Professor O'Connell has been the recipient of an Effective Educator award from the Richard T. Farmer School of Business Administration at Miami.

Emily S. Murphree is professor emeritus of statistics at Miami University, Oxford, Ohio. She received her PhD in statistics from the University of North Carolina with a research concentration in applied probability. Professor Murphree received Miami's College of Arts and Sciences Distinguished Education Award and has received various civic awards.

QUANTITATIVE APPROACHES TO DECISION MAKING COLLECTION

Donald N. Stengel, *Editor*

ISBN: 978-1-60649-950-4
90000
9 781606 499504

BEP BUSINESS EXPERT PRESS