13.1 Using Several Variables to Predict a Response
13.2 Extending the Correlation and $R^2$ for Multiple Regression
13.3 Using Multiple Regression to Make Inferences
13.4 Checking a Regression Model Using Residual Plots
13.5 Regression and Categorical Predictors
13.6 Modeling a Categorical Response
Example 1

Predicting the Selling Price of a House

Picture the Scenario

You are saving to buy a home, and you wonder how much it will cost. The House Selling Prices OR data file on the text CD has observations on 200 recent home sales in Corvallis, Oregon. Table 13.1 shows data for two homes.

Table 13.1 Selling Prices and Related Factors for a Sample of Home Sales

<table>
<thead>
<tr>
<th>House</th>
<th>Selling Price</th>
<th>House Size (sq. ft)</th>
<th>Number of Bedrooms</th>
<th>Number of Bathrooms</th>
<th>Lot Size (sq. ft)</th>
<th>Year Built</th>
<th>Garage (Y/N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$232,500</td>
<td>1679</td>
<td>3</td>
<td>1.5</td>
<td>10,019</td>
<td>1976</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>$158,000</td>
<td>1292</td>
<td>1</td>
<td>1</td>
<td>217,800</td>
<td>1958</td>
<td>N</td>
</tr>
</tbody>
</table>

Variables listed are the selling price (in dollars), house size (in square feet), number of bedrooms, number of bathrooms, the lot size (in square feet), and whether or not the house has a garage. Table 13.2 reports the mean and standard deviation of these variables for all 200 home sales.

Table 13.2 Descriptive Statistics for Sales of 200 Homes

<table>
<thead>
<tr>
<th></th>
<th>Selling Price Mean</th>
<th>House Size (sq. ft) Mean</th>
<th>Number of Bedrooms Mean</th>
<th>Number of Bathrooms Mean</th>
<th>Lot Size (sq. ft) Mean</th>
<th>Age Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviation</td>
<td>$115,808</td>
<td>2551</td>
<td>3.08</td>
<td>2.03</td>
<td>23,217</td>
<td>34.75</td>
</tr>
</tbody>
</table>

Questions to Explore

In your community, if you know the values of such variables,

- How can you predict a home’s selling price?
- How can you describe the association between selling price and the other variables?
- How can you make inferences about the population based on the sample data?

Thinking Ahead

You can find a regression equation to predict selling price by treating one of the other variables as the explanatory variable. However, since there are several explanatory variables, you might make better predictions by using all of them at once. That’s the idea behind multiple regression. It uses more than one explanatory variable to predict a response variable. We’ll learn how to apply multiple regression to a variety of analyses in Examples 2, 3, 8, 9, and 10 of this chapter.

Recall

Review the concept of lurking variables in Section 3.4.

Besides helping you to better predict a response variable, multiple regression can help you analyze the association between two variables while controlling (keeping fixed) values of other variables. Such adjustment is important because the effect of an explanatory variable can change considerably after you account for potential lurking variables. A multiple regression analysis provides information...
not available from simple regression analyses involving only two variables at a time.

Section 13.1 presents the basics of multiple regression. Section 13.2 extends the concepts of correlation and $r^2$ for describing strength of association to multiple predictors (explanatory variables). Section 13.3 presents inference methods, and Section 13.4 shows how to check the model. Section 13.5 extends the model further to incorporate categorical explanatory variables. The final section presents a regression model for a categorical response variable.

### 13.1 Using Several Variables to Predict a Response

Chapter 12 modeled the mean $\mu_y$ of a quantitative response variable $y$ and a quantitative explanatory variable $x$ by the straight-line equation $\mu_y = \alpha + \beta x$. We refer to this model with a single predictor as bivariate (two-variable) regression, since it contains only two variables ($x$ and $y$).

Now, suppose there are two predictors, denoted by $x_1$ and $x_2$. The bivariate regression equation generalizes to the multiple regression equation,

$$\mu_y = \alpha + \beta_1 x_1 + \beta_2 x_2.$$ 

In this equation, $\alpha$, $\beta_1$, and $\beta_2$ are parameters. When we substitute values for $x_1$ and $x_2$, the equation specifies the population mean of $y$ for all subjects with those values of $x_1$ and $x_2$. When there are additional predictors, each has a $\beta x$ term.

### Multiple Regression Model

The multiple regression model relates the mean $\mu_y$ of a quantitative response variable $y$ to a set of explanatory variables $x_1, x_2, \ldots$. For three explanatory variables, for example, the multiple regression equation is

$$\mu_y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

and the sample prediction equation is

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + b_3 x_3.$$ 

With sample data, software estimates the multiple regression equation. It uses the method of least squares to find the best prediction equation (the one with the smallest possible sum of squared residuals).

### Example 2

Predicting Selling Price Using House Size and Number of Bedrooms

**Picture the Scenario**

For the house selling price data described in Example 1, MINITAB reports the results in Table 13.3 for a multiple regression analysis with selling price as the response variable and with house size and number of bedrooms as explanatory variables.
Chapter 13  Multiple Regression

Plotting Relationships

Always look at the data before doing a multiple regression analysis. Most software has the option of constructing scatterplots on a single graph for each pair of variables. This type of plot is called a scatterplot matrix.

Figure 13.1 shows a MINITAB scatterplot matrix for selling price, house size, and number of bedrooms. It shows each pair of variables twice. For a given pair, in one plot a variable is on the y-axis and in another it is on the x-axis. For instance, selling price of house is on the y-axis for the plots in the first row, whereas it is on the x-axis for the plots in the first column. Since selling price is

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>60102</td>
<td>18623</td>
<td>3.23</td>
<td>0.001</td>
</tr>
<tr>
<td>House_size</td>
<td>62.983</td>
<td>4.753</td>
<td>13.25</td>
<td>0.000</td>
</tr>
<tr>
<td>Bedrooms</td>
<td>15170</td>
<td>5330</td>
<td>2.85</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Questions to Explore

a. State the prediction equation.
b. The first home listed in Table 13.1 has house size = 1679 square feet, three bedrooms, and selling price $232,500. Find its predicted selling price and the residual (prediction error). Interpret the residual.

Think It Through

a. The response variable is \( y \) = selling price. Let \( x_1 \) = house size and \( x_2 \) = the number of bedrooms. From Table 13.3, the prediction equation is

\[
\hat{y} = 60,102 + 63.0x_1 + 15,170x_2.
\]

b. For \( x_1 = 1679 \) and \( x_2 = 3 \), the predicted selling price is

\[
\hat{y} = 60,102 + 63.0(1679) + 15,170(3) = 211,389, \text{ that is, } $211,389.
\]

The residual is the prediction error,

\[
y - \hat{y} = 232,500 - 211,389 = 21,111.
\]

This result tells us that the actual selling price was $21,111 higher than predicted.

Insight

The coefficients of house size and number of bedrooms are positive. As these variables increase, the predicted selling price increases, as we would expect.

Try Exercise 13.1

In Words

As in bivariate regression, we use the Greek letter \( \beta \) (beta) for parameters describing effects of explanatory variables, with \( \beta_1 \) and \( \beta_2 \) read as “beta one” and “beta two.”

In Practice  The Number of Explanatory Variables You Can Use Depends on the Amount of Data

In practice, you should not use many explanatory variables in a multiple regression model unless you have lots of data. A rough guideline is that the sample size \( n \) should be at least 10 times the number of explanatory variables. For example, to use two explanatory variables, you should have at least \( n = 20 \) observations.

Plotting Relationships

Always look at the data before doing a multiple regression analysis. Most software has the option of constructing scatterplots on a single graph for each pair of variables. This type of plot is called a scatterplot matrix.
the response variable for this example, the plots in the first row (where selling price is on the $y$-axis) are the ones of primary interest. These graphs show strong positive linear relationships between selling price and both house size and number of bedrooms. Because a scatterplot matrix shows each pair of variables twice, you only need to look at the plots in the upper-right triangle.

Each scatterplot portrays only two variables. It’s a two-dimensional picture. A multiple regression equation, which has several variables, is more difficult to portray graphically. Note also that the scatterplot involving the variable number of bedrooms as an explanatory variable has an appearance of columns of points. This is due to the highly discrete nature of this quantitative variable.

**Interpretation of Multiple Regression Coefficients**

The simplest way to interpret a multiple regression equation is to look at it in two dimensions as a function of a single explanatory variable. We can do this by fixing values for the other explanatory variable(s). For instance, let’s fix $x_1 =$ house size at the value 2000 square feet. Then the prediction equation simplifies to one with $x_2 =$ number of bedrooms alone as the predictor,

$$
\hat{y} = 60,102 + 63.0(2000) + 15,170x_2 = 186,102 + 15,170x_2.
$$

For 2000-square-foot houses, the predicted selling price relates to number of bedrooms by $\hat{y} = 186,102 + 15,170x_2$. Since the slope coefficient of $x_2$ is 15,170, the predicted selling price increases by $15,170$ for every bedroom added.

Likewise, we could fix the number of bedrooms, and then describe how the predicted selling price depends on the house size. Let’s consider houses with number of bedrooms $x_2 =$ three bedrooms. The prediction equation becomes

$$
\hat{y} = 60,102 + 63.0x_1 + 15,170(3) = 105,612 + 63.0x_1.
$$

For houses with this number of bedrooms, the predicted selling price increases by $6300$ for every 100 square foot in size increase.

Can we say an increase of one bedroom has a larger impact on the selling price ($15,170$) than an increase of a square foot in house size ($63$ per square foot)?
No, we cannot compare these slopes for these explanatory variables because their units of measurement are not the same. Slopes can’t be compared when the units differ. We could compare house size and lot size directly if they had the same units of square feet.

**Recall**

Section 10.5 showed that in multivariate analyses, we can control a variable statistically by keeping its value constant while we study the association between other variables.

---

**Summarizing the Effect While Controlling for a Variable**

The multiple regression model states that each explanatory variable has a straight line relationship with the mean of $y$, given fixed values of the other explanatory variables. Specifically, the model assumes that the slope for a particular explanatory variable is identical for all fixed values of the other explanatory variables. For instance, the coefficient of $x_1$ in the prediction equation $\hat{y} = 60,102 + 63.0x_1 + 15,170x_2$ is 63.0 regardless of whether we plug in $x_2 = 1$ or $x_2 = 2$ or $x_2 = 3$ for the number of bedrooms. When you fix $x_2$ at these three levels, you can check that:

<table>
<thead>
<tr>
<th>$x_2$</th>
<th>$\hat{y}$ = 60,102 + 63.0$x_1$ + 15,170$x_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\hat{y}$ = 75,272 + 63.0$x_1$</td>
</tr>
<tr>
<td>2</td>
<td>$\hat{y}$ = 90,442 + 63.0$x_1$</td>
</tr>
<tr>
<td>3</td>
<td>$\hat{y}$ = 105,612 + 63.0$x_1$</td>
</tr>
</tbody>
</table>

The slope effect of house size is 63.0 for each equation. Setting $x_2$ at a variety of values yields a collection of parallel lines, each having slope 63.0. See Figure 13.2.

![Figure 13.2 The Relationship Between $\hat{y}$ and $x_1$ for the Multiple Regression Equation $\hat{y} = 60,102 + 63.0x_1 + 15,170x_2$. This shows how the equation simplifies when number of bedrooms $x_2 = 1$, or $x_2 = 2$, or $x_2 = 3$. Question: The lines move upward (to higher $\hat{y}$-values) as $x_2$ increases. How would you interpret this fact?](image-url)

When we fix the value of $x_2$ we are holding it constant: We are controlling for $x_2$. That’s the basis of the major difference between the interpretation of slopes in multiple regression and in bivariate regression:

- In multiple regression, a slope describes the effect of an explanatory variable while controlling effects of the other explanatory variables in the model.
- Bivariate regression has only a single explanatory variable. So a slope in bivariate regression describes the effect of that variable while ignoring all other possible explanatory variables.

For example, the bivariate regression between $y$ = selling price and $x_1$ = house size is $\hat{y} = 97,997 + 66.4x_1$. In this equation, $x_2$ = number of bedrooms and other possible predictors are ignored, not controlled. The equation describes the relationship for all the house sales in the data set. On the other hand, the equation $\hat{y} = 90,442 + 63.0x_1$ we obtained above by substituting $x_2 = 2$ into the multiple
Section 13.1 Using Several Variables to Predict a Response

regression equation applies only for houses that have that number of bedrooms. In this case, number of bedrooms is controlled. This is why the slopes are different, 66.4 for bivariate regression and 63.0 for multiple regression.

One of the main uses of multiple regression is to identify potential lurking variables and control for them by including them as explanatory variables in the model. Doing so can have a major impact on a variable’s effect. When we control a variable, we keep that variable from influencing the associations among the other variables in the study. As we’ve seen before, the direction of the effect can change after we control for a variable. Exercise 13.5 illustrates this for multiple regression modeling.

13.1 Practicing the Basics

13.1 Predicting weight For a study of University of Georgia female athletes, the prediction equation relating \( y = \text{total body weight (in pounds)} \) to \( x_1 = \text{height (in inches)} \) and \( x_2 = \text{percent body fat} \) is \( \hat{y} = -121 + 3.50x_1 + 1.35x_2 \).

a. Find the predicted total body weight for a female athlete at the mean values of 66 and 18 for \( x_1 \) and \( x_2 \).

b. An athlete with \( x_1 = 66 \) and \( x_2 = 18 \) has actual weight \( y = 115 \) pounds. Find the residual and interpret it.

13.2 Does study help GPA? For the Georgia Student Survey file on the text CD, the prediction equation relating \( y = \text{college GPA} \) to \( x_1 = \text{high school GPA} \) and \( x_2 = \text{study time (hours per day)} \), is

\[ \hat{y} = 1.13 + 0.643x_1 + 0.0078x_2. \]

a. Find the predicted college GPA of a student who has a high school GPA of 3.5 and who studies three hours a day.

b. For students with fixed study time, what is the change in predicted college GPA when high school GPA increases from 3.0 to 4.0?

13.3 Predicting college GPA For all students at Walden University, the prediction equation for \( y = \text{college GPA} \) (range 0–4.0) and \( x_1 = \text{high school GPA} \) (range 0–4.0) and \( x_2 = \text{college board score (range 200–800)} \) is

\[ \hat{y} = 0.20 + 0.50x_1 + 0.002x_2. \]

a. Find the predicted college GPA for students having (i) high school GPA = 4.0 and college board score = 800 and (ii) \( x_1 = 2.0 \) and \( x_2 = 200 \).

b. For those students with \( x_2 = 500 \), show that \( \hat{y} = 1.20 + 0.50x_1 \).

c. For those students with \( x_2 = 600 \), show that \( \hat{y} = 1.40 + 0.50x_1 \). Thus, compared to part b, the slope for \( x_1 \) is still 0.50, and increasing \( x_2 \) by 100 (from 500 to 600) shifts the intercept upward by 100 \times (slope for \( x_2 \)) = 100(0.002) = 0.20 units.

13.4 Interpreting slopes on GPA Refer to the previous exercise.

a. Explain why setting \( x_2 \) at a variety of values yields a collection of parallel lines relating \( \hat{y} \) to \( x_1 \). What is the value of the slope for those parallel lines?

b. Since the slope 0.50 for \( x_1 \) is larger than the slope 0.002 for \( x_2 \), does this imply that \( x_1 \) has a larger effect than \( x_2 \) on \( y \) in this sample? Explain.

13.5 Does more education cause more crime? The FL Crime data file on the text CD has data for the 67 counties in Florida on

\[ y = \text{crime rate: Annual number of crimes in county per 1000 population} \]

\[ x_1 = \text{education: Percentage of adults in county with at least a high school education} \]

\[ x_2 = \text{urbanization: Percentage in county living in an urban environment.} \]

The figure shows a scatterplot matrix. The correlations are 0.47 between crime rate and education, 0.68 between crime rate and urbanization, and 0.79 between education and urbanization. MINITAB multiple regression results are also displayed.

Multiple regression for \( y = \text{crime rate} \), \( x_1 = \text{education} \), and \( x_2 = \text{urbanization} \).

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>59.12</td>
<td>28.37</td>
<td>2.08</td>
<td>0.041</td>
</tr>
<tr>
<td>education</td>
<td>-0.5834</td>
<td>0.4725</td>
<td>-1.23</td>
<td>0.221</td>
</tr>
<tr>
<td>urbanization</td>
<td>0.6825</td>
<td>0.1232</td>
<td>5.54</td>
<td>0.000</td>
</tr>
</tbody>
</table>

a. Find the predicted crime rate for a county that has 0% in an urban environment and (i) 70% high school graduation rate and (ii) 80% high school graduation rate.
b. Use results from part a to explain how education affects the crime rate, controlling for urbanization, interpreting the slope coefficient –0.58 of education.

c. Using the prediction equation, show that the equation relating crime rate and education when urbanization is fixed at (i) 0, (ii) 50, and (iii) 100, is as follows:

\[
\begin{array}{c|c|c}
\text{Urbanization} & \hat{y} = 59.1 - 0.58x_1 + 0.68x_2 & \hat{y} = 93.2 - 0.58x_1 \\
0 & \hat{y} = 93.2 - 0.58x_1 & \hat{y} = 127.4 - 0.58x_1 \\
50 & \hat{y} = 93.2 - 0.58x_1 & \quad \\
100 & \quad & \quad \\
\end{array}
\]

Sketch a plot with these lines and use it to interpret the effect of education on crime rate, controlling for urbanization.

d. The scatterplot matrix shows that education has a positive association with crime rate, but the multiple regression equation shows that the association is negative when we keep \(x_2\) = urbanization fixed. The reversal in the association is an example of Simpson’s paradox (See Example 16 in Sec. 3.4 and Example 17 in Sec. 10.5). Consider the hypothetical figure that follows. Sketch lines that represent (i) the bivariate relationship, ignoring the information on urbanization and (ii) the relationship for counties having urbanization = 50. Use this figure and the correlations provided to explain how Simpson’s paradox can happen.

![Hypothetical scatterplot for crime rate and education, labeling by urbanization.](image)

### 13.6 Crime rate and income

Refer to the previous exercise. MINITAB reports the results below for the multiple regression of \(y = \) crime rate on \(x_1 = \) median income (in thousands of dollars) and \(x_2 = \) urbanization.

**Results of regression analysis**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>39.97</td>
<td>16.35</td>
<td>2.44</td>
<td>0.017</td>
</tr>
<tr>
<td>Income</td>
<td>-0.7906</td>
<td>0.8049</td>
<td>-0.98</td>
<td>0.330</td>
</tr>
<tr>
<td>Urbanization</td>
<td>0.6418</td>
<td>0.1110</td>
<td>5.78</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Correlations: crime, income, urbanization**

\[
\begin{array}{c|c|c}
\text{Correlation} & \text{crime} & \text{urbanization} \\
\text{crime} & 0.677 & \text{urbanization} \\
\text{income} & 0.434 & 0.731 \\
\end{array}
\]

a. Report the prediction equations relating crime rate to income at urbanization levels of (i) 0 and (ii) 100.

b. For the bivariate model relating \(y = \) crime rate to \(x = \) income, MINITAB reports

\[
\text{crime} = -11.6 + 2.61 \text{ income}
\]

Interpret the effect of income, according to the sign of its slope. How does this effect differ from the effect of income in the multiple regression equation?

c. The correlation matrix for these three variables is shown in the table. Use these correlations to explain why the income effect seems so different in the models in part a and part b.

d. Do these variables satisfy Simpson’s paradox?

### 13.7 The economics of golf

The earnings of a PGA Tour golfer are determined by performance in tournaments. A study analyzed tour data to determine the financial return for certain skills of professional golfers. The sample consisted of 393 golfers competing in one or both of the 2002 and 2008 seasons. The most significant factors that contribute to earnings were the percent of attempts a player was able to hit the green in regulation (GIR), the number of times that a golfer made par or better after hitting a bunker divided by the number of bunkers hit (SS), and the number of PGA events entered (Events). The resulting coefficients from multiple regression to predict 2008 earnings are:

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>$26,417,000</td>
</tr>
<tr>
<td>GIR</td>
<td>$168,300</td>
</tr>
<tr>
<td>SS</td>
<td>$33,859</td>
</tr>
<tr>
<td>AvePutt</td>
<td>$19,784,000</td>
</tr>
<tr>
<td>Events</td>
<td>$44,725</td>
</tr>
</tbody>
</table>


a. State the regression formula for a PGA Tour golfer’s earnings for 2008.

b. Explain how to interpret the coefficients for AvePutt and Events.

c. Find the predicted total score for a golfer who had a GIR score of 60, SS score of 50, AvePutt is 1.5, for 20 Events.

### 13.8 Comparable number of bedrooms and house size effects

In Example 2, the prediction equation between \(y = \) selling price and \(x_1 = \) house size and \(x_2 = \) number of bedrooms was \(\hat{y} = 60,102 + 63.0x_1 + 15,170x_2\).

a. For fixed number of bedrooms, how much is the house selling price predicted to increase for each square foot increase in house size? Why?

b. For a fixed house size of 2000 square feet, how does the predicted selling price change for two, three, and four bedrooms?

### 13.9 Controlling can have no effect

Suppose that the correlation between \(x_1\) and \(x_2\) equals 0. Then, for multiple regression with those predictors, it can be shown that the slope for \(x_1\) is the same as in bivariate regression when \(x_1\) is the only predictor. Explain why you would expect this to be true. (**Hint:** If you don’t control \(x_2\), would you expect it to have an impact on how \(x_1\) affects \(y\), if \(x_1\) and \(x_2\) have correlation of 0?)
13.2 Extending the Correlation and \( R^2 \) for Multiple Regression

The correlation \( r \) and its square \( r^2 \) describe strength of association in a straight-line regression analysis. These measures can also describe association between \( y \) and a set of explanatory variables that predict \( y \) in a multiple regression model.

**Multiple Correlation**

To summarize how well a multiple regression model predicts \( y \), we analyze how well the observed \( y \) values correlate with the predicted \( \hat{y} \) values. As a set, the explanatory variables are strongly associated with \( y \) if the correlation between the \( y \) and \( \hat{y} \) values is strong. Treating the \( \hat{y} \) as a variable gives us a way of summarizing several explanatory variables by *one* variable, for which we can use its ordinary correlation with the \( y \) values. The correlation between the observed \( y \) values and the predicted \( \hat{y} \) values from the multiple regression model is called the *multiple correlation*.

**Multiple Correlation, \( R \)**

For a multiple regression model, the multiple correlation is the correlation between the observed \( y \) values and the predicted \( \hat{y} \) values. It is denoted by \( R \).

For each subject, the regression equation provides a predicted value \( \hat{y} \). So, each subject has a \( y \) value and a \( \hat{y} \) value. For the two houses listed in Table 13.1, Table 13.4 shows the actual selling price \( y \) and the predicted selling price \( \hat{y} \) from the equation \( \hat{y} = 60,102 + 63.0x_1 + 15,170x_2 \) with house size and number of bedrooms as predictors. The correlation computed between all 200 pairs of \( y \) and \( \hat{y} \) values is the multiple correlation, \( R \). Software tells us that this equals 0.72. The scatterplot in the margin displays these pairs of \( y \) and \( \hat{y} \) values.

**Table 13.4 Selling Prices and Their Predicted Values**

These values refer to the two home sales listed in Table 13.1. The predictors are \( x_1 = \) house size and \( x_2 = \) number of bedrooms.

<table>
<thead>
<tr>
<th>Home</th>
<th>Selling Price</th>
<th>Predicted Selling Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>232,500</td>
<td>( \hat{y} = 60,102 + 63.0(1679) + 15,170(3) = 211,389 )</td>
</tr>
<tr>
<td>2</td>
<td>158,000</td>
<td>( \hat{y} = 60,102 + 63.0(1292) + 15,170(1) = 156,668 )</td>
</tr>
</tbody>
</table>

The larger the multiple correlation, the better are the predictions of \( y \) by the set of explanatory variables. For the housing data, \( R = 0.72 \) indicates a moderately strong association.

The predicted values \( \hat{y} \) cannot correlate negatively with \( y \). Otherwise, the predictions would be worse than merely using \( y \) to predict \( y \). Therefore, \( R \) falls...
between 0 and 1. In this way, the multiple correlation $R$ differs from the bivariate correlation $r$ between $y$ and a single variable $x$, which falls between $-1$ and $+1$.

### Example 3

Predicting House Selling Prices

**Picture the Scenario**

For the 200 observations on $y =$ selling price in thousands of dollars, using, $x_1 =$ house size in thousands of square feet and $x_2 =$ number of bedrooms, Table 13.5 shows the ANOVA (analysis of variance) table that MINITAB reports for the multiple regression model.

### Table 13.5 ANOVA Table and $R^2$ for Predicting House Selling Price (in thousands of dollars) Using House Size (in thousands of square feet) and Number of Bedrooms

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>1399524</td>
</tr>
<tr>
<td>Residual Error</td>
<td>197</td>
<td>1269345</td>
</tr>
<tr>
<td>Total</td>
<td>199</td>
<td>2668870</td>
</tr>
</tbody>
</table>

### Questions to Explore

a. Show how to use the sums of squares in the ANOVA table to find $R^2$ for this multiple regression model. Interpret.

b. Find and interpret the multiple correlation.

**Think It Through**

a. From the sum of squares (SS) column, the total sum of squares is $\Sigma(y - \bar{y})^2 = 2,668,870$. The residual sum of squares from using the multiple regression equation to predict $y$ is $\Sigma(y - \hat{y})^2 = 1,269,345$. The value of $R^2$ is...
Section 13.2 Extending the Correlation and $R^2$ for Multiple Regression

Properties of $R^2$

The example showed that $R^2$ for the multiple regression model was larger than $r^2$ for a bivariate model using only one of the explanatory variables. In fact, a key property of $R^2$ is that it cannot decrease when predictors are added to a model.

The difference $\Sigma(y - \bar{y})^2 - \Sigma(y - \hat{y})^2$ that forms the numerator of $R^2$ also appears in the ANOVA table. It is called the regression sum of squares. A simpler formula for it is

$$\text{Regression SS} = \Sigma(y - \bar{y})^2.$$

If each $\hat{y} = \bar{y}$, then the regression equation predicts no better than $\bar{y}$. Then regression SS $= 0$, and $R^2 = 0$.

The numerical values of the sums of squares reported in Table 13.5 result when $y$ is measured in thousands of dollars (for instance, the first house $y$ is 232.5 rather than 232,500, and $\hat{y}$ is 211.4 rather than 211,389). We truncated the numbers for convenience, because the sums of squares would otherwise be enormous with each one having six more zeros on the end! It makes no difference to the result. Another property of $R$ and $R^2$ is that, like the correlation, their values don't depend on the units of measurement.

In summary, the properties of $R^2$ are similar to those of $r^2$ for bivariate models. Here's a list of the main properties:

<table>
<thead>
<tr>
<th>SUMMARY: Properties of $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$ falls between 0 and 1. The larger the value, the better the explanatory variables collectively predict $y$.</td>
</tr>
</tbody>
</table>
Table 13.6 shows how \( R^2 \) increases as we add explanatory variables to a multiple regression model to predict \( y = \) house selling price. The single predictor in the data set that is most strongly associated with \( y \) is the house size (\( r^2 = 0.505 \)). See the margin table. When we add number of bedrooms as a second predictor, \( R^2 \) goes up from 0.505 to 0.524. As other predictors are added, \( R^2 \) continues to go up, but not by much. Predictive power is not much worse with only house size as a predictor than with all six predictors in the regression model.

### Table 13.6 \( R^2 \) Value for Multiple Regression Models for \( y = \) House Selling Price

<table>
<thead>
<tr>
<th>Explanatory Variables in Model</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>House size</td>
<td>0.505</td>
</tr>
<tr>
<td>House size, Number of bedrooms</td>
<td>0.524</td>
</tr>
<tr>
<td>House size, Number of bedrooms, Lot size</td>
<td>0.524</td>
</tr>
<tr>
<td>House size, Number of bedrooms, Lot size, Number of bathrooms</td>
<td>0.604</td>
</tr>
<tr>
<td>House size, Number of bedrooms, Lot size, Number of bathrooms, Garage</td>
<td>0.608</td>
</tr>
<tr>
<td>House size, Number of bedrooms, Lot size, Number of bathrooms, Garage, Age</td>
<td>0.612</td>
</tr>
</tbody>
</table>

Although \( R^2 \) goes up by only small amounts when we add other variables after house size is already in the model, this does not mean that the other predictors are only weakly correlated with selling price. Because the predictors are themselves highly correlated, once one or two of them are in the model, the remaining ones don’t help much in adding to the predictive power. For instance, lot size is highly positively correlated with number of bedrooms and with size of house. So, once number of bedrooms and size of house are included as predictors in the model, there’s not much benefit to including lot size as an additional predictor.

### In Practice \( R^2 \) Often Does Not Increase Much After a Few Predictors Are in the Model

When there are many explanatory variables but the correlations among them are strong, once you have included a few of them in the model, \( R^2 \) usually doesn’t increase much more when you add additional ones. This does not mean that the additional variables are uncorrelated with the response variable but merely that they don’t add much new power for predicting \( y \), given the values of the predictors already in the model.

As in the bivariate case, a disadvantage of \( R^2 \) (compared to the multiple correlation \( R \)) is that its units are the square of the units of measurement. The \( R^2 \) of 0.524 in Example 3 implies that the estimated variance of \( y \) at fixed values of the predictors is 52.4% less than the overall variance of \( y \).
### 13.2 Practicing the Basics

#### 13.11 Predicting sports attendance

Keeneland Racetrack in Lexington, Kentucky, has been a social gathering place since 1935. Every spring and fall thousands of people come to the racetrack to socialize, gamble, and enjoy the horse races that have become so popular in Kentucky. A study investigated the different factors that affect the attendance at Keeneland. To many people, the social aspect of Keeneland is equally as important as watching the races, if not more. The study investigates which factors significantly contribute to attendance by Keeneland visitors. (Source: Data from Gatton School of Business and Economics publication, Gatton College, 2009.)

**ANOVA table for y = attendance at Keeneland**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>6.00</td>
<td>2360879925</td>
<td>393479988</td>
<td>21.71</td>
<td>0.00</td>
</tr>
<tr>
<td>Residual</td>
<td>110.00</td>
<td>1993805006</td>
<td>18125500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>116.00</td>
<td>4354684931</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Show how $R^2$ is calculated from the SS values, and report its value.
b. Interpret the $R^2$ value. Does the multiple regression equation help us predict the attendance much better than we could without knowing that equation?c. Find the multiple correlation. Interpret.

#### 13.12 Predicting weight

Let’s use multiple regression to predict total body weight (in pounds) using data from a study of University of Georgia female athletes. Possible predictors are $\text{HGT} = \text{height (in inches)}, \% \text{BF} = \text{percent body fat}, \text{and age}$. The display shows correlations among these explanatory variables.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>5.473</td>
<td>0.616</td>
<td>8.64</td>
<td>0.00</td>
</tr>
<tr>
<td>$\text{GDP}$</td>
<td>0.0002779</td>
<td>0.00002779</td>
<td>7.87</td>
<td>0.000</td>
</tr>
<tr>
<td>$\text{CD}$</td>
<td>-1.3849</td>
<td>0.7270</td>
<td>-1.90</td>
<td>0.089</td>
</tr>
<tr>
<td>$S = 9.23490$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2 = 0.55$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Report $R^2$ and show how it is determined by SS values in the ANOVA table.
b. Interpret its value as a proportional reduction in prediction error.

c. Find the multiple correlation. Interpret.

#### 13.13 When does controlling have little effect?

Refer to the previous exercise. Height has a similar estimated slope for each of the three models. Why do you think that controlling for $\% \text{body fat}$ and then age does not change the effect of height much? (Hint: How strongly is height correlated with the other two variables?)

#### 13.14 Internet use

For countries listed in the Twelve Countries data file on the text CD, y = Internet use (percent) is predicted by $x_1 = \text{per capita GDP}$ (gross domestic product, in thousands of dollars) with $r^2 = 0.88$. Adding $x_2 = \text{carbon dioxide emissions per capita}$ to the model yields the results in the following display.

**Regression of Internet use on GDP and carbon dioxide emissions per capita**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>2.843</td>
<td>0.5473</td>
<td>5.21</td>
<td>0.000</td>
</tr>
<tr>
<td>$\text{GDP}$</td>
<td>0.0002779</td>
<td>0.00002779</td>
<td>7.87</td>
<td>0.000</td>
</tr>
<tr>
<td>$\text{CD}$</td>
<td>-1.3849</td>
<td>0.7270</td>
<td>-1.90</td>
<td>0.089</td>
</tr>
<tr>
<td>$S = 9.23490$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2 = 0.55$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Interpret the value of the multiple correlation.

#### 13.15 More Internet use

In the previous exercise, $r^2 = 0.88$ when $x_1$ is the predictor and $R^2 = 0.914$ when both $x_1$ and $x_2$ are predictors. Why do you think that the predictions of $y$ don’t improve much when $x_2$ is added to the model? (The association of $x_2$ with $y$ is $r = 0.5692$.)

#### 13.16 Softball data

For the Softball data set on the text CD, for each game the variables are a team’s number of runs scored (RUNS), number of hits (HT), number of errors (ERR), and the difference (DIFF) between the number of runs scored by that team and by the other team, which is the response variable. MINITAB reports

<table>
<thead>
<tr>
<th>Difference</th>
<th>-4.03 + 0.0260 * Hits + 1.04 Run - 1.22 Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a.$</td>
<td>If you know the team’s number of runs and number of errors in a game, explain why it does not help much to know how many hits the team has.</td>
</tr>
<tr>
<td>$b.$</td>
<td>Explain why the result in part a is also suggested by knowing that $R^2 = 0.4794$ for this model, whereas $R^2 = 0.7593$ when only runs and errors are the explanatory variables in the model.</td>
</tr>
</tbody>
</table>

#### 13.17 Slopes, correlations, and units

In Example 2 on $y = \text{house selling price}, x_1 = \text{house size}, \text{and} x_2 = \text{number of bedrooms},\ y = 60.102 + 63.0x_1 + 15.170x_2, \text{and} R = 0.72$.

a. Interpret the value of the multiple correlation.
b. Suppose house selling prices are changed from dollars to thousands of dollars. Explain why if each house price in Table 13.1 is divided by 1000, then the prediction equation changes to \( y = 60.102 + 0.063x_1 + 15.170x_2 \).

c. In part b, does the multiple correlation change to 0.00072? Justify your answer.

### 13.3 Using Multiple Regression to Make Inferences

We’ve seen that multiple regression uses more than one explanatory variable to predict a response variable \( y \). With it, we can study an explanatory variable’s effect on \( y \) while controlling other variables that could affect the results. Now, let’s turn our attention to using multiple regression to make inferences about the population.

Inferences require the same assumptions as in bivariate regression:

- The regression equation truly holds for the population means.
- The data were gathered using randomization.
- The response variable \( y \) has a normal distribution at each combination of values of the explanatory variables, with the same standard deviation.

The first assumption implies that there is a straight-line relationship between the mean of \( y \) and each explanatory variable, with the same slope at each value of the other predictors. We will see how to check this assumption and the third assumption in Section 13.4.

### Estimating Variability Around the Regression Equation

A check for normality is needed to validate the third assumption; this is discussed in the next section. In addition to normality, a constant standard deviation for each combination of explanatory variables is assumed. This standard deviation is similar to the standard error for the explanatory variable in bivariate data. (Recall Section 12.4.) As in bivariate regression, a standard deviation parameter \( \sigma \) describes variability of the observations around the regression equation. Its sample estimate is

\[
s = \sqrt{\frac{\text{Residual SS}}{df}} = \sqrt{\frac{\sum(y - \hat{y})^2}{n - (\text{number of parameters in regression equation})}}.
\]

This residual standard deviation describes the typical size of the residuals. Its degrees of freedom are \( df = n - \text{number of parameters in the regression equation} \). Software reports \( s \) and its square, the mean square error.

#### Example 4

**Female Athletes’ Weight**

**Picture the Scenario**

The College Athletes data set on the text CD comes from a study of 64 University of Georgia female athletes who participated in Division I sports. The study measured several physical characteristics, including total body...
weight in pounds (TBW), height in inches (HGT), the percent of body fat (%BF), and age. Table 13.7 shows the ANOVA table for the regression of weight on height, % body fat, and age.

**Table 13.7 ANOVA Table for Multiple Regression Analysis of Athlete Weights**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>3</td>
<td>12407.9</td>
<td>4136.0</td>
<td>40.48</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>60</td>
<td>6131.0</td>
<td>102.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>63</td>
<td>18539.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S = 10.1086</td>
<td></td>
<td>R-Sq = 66.9%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Question to Explore**

For female athletes at particular values of height, percent of body fat, and age, estimate the standard deviation of their weights.

**Think It Through**

The SS column tells us that the residual sum of squares is 6131.0. There were \( n = 64 \) observations and 4 parameters in the regression model, so the DF column reports \( df = n - 4 = 60 \) opposite the residual SS. The mean square error is

\[
s^2 = \frac{\text{residual SS}}{df} = \frac{6131.0}{60} = 102.2.
\]

It appears in the mean square (MS) column, in the row labeled “Residual Error.” The residual standard deviation is \( s = \sqrt{102.2} = 10.1 \), identified as S. For athletes with certain fixed values of height, percent body fat, and age, the weights vary with a standard deviation of about 10 pounds.

**Insight**

If the conditional distributions of weight are approximately bell shaped, about 95% of the weight values fall within about \( 2s = 20 \) pounds of the true regression equation. More precisely, software can report *prediction intervals* within which a response outcome has a certain chance of falling. For instance, at \( x_1 = 66, x_2 = 18, \) and \( x_3 = 20 \), which are close to the mean predictor values, software reports \( \hat{y} = 133.9 \) and a 95% prediction interval of 133.9 ± 20.4.

**Try Exercise 13.21**

In the sample, weight has standard deviation = 17 pounds, describing variability around mean weight = 133 pounds (see Table 13.9). How is it that the residual standard deviation could be only 10 pounds? The residual standard deviation is smaller because it refers to variability at *fixed* values of the predictors. Weight varies less at given values of the predictors than it does overall. If we could predict weight perfectly knowing height, percent body fat, and age, the residual standard deviation would be 0.

The ANOVA table also reports another mean square, called a *mean square for regression*, or *regression mean square* for short. We’ll next see how to use the mean squares to conduct a significance test about all the slope parameters together.
The Collective Effect of Explanatory Variables

Do the explanatory variables collectively have a statistically significant effect on the response variable \( y \)? With three predictors in a model, we can check this by testing

\[
H_0: \beta_1 = \beta_2 = \beta_3 = 0.
\]

This hypothesis states that the mean of \( y \) does not depend on any of the predictors in the model. That is, \( y \) is statistically independent of all the explanatory variables. The alternative hypothesis is

\[
H_a: \text{At least one } \beta \text{ parameter is not equal to } 0.
\]

This states that at least one explanatory variable is associated with \( y \).

The null hypothesis that all the slope parameters equal 0 is equivalent to the hypothesis that the population values of the multiple correlation and \( R^2 \) equal 0. The equivalence occurs because the population values of \( R \) and \( R^2 \) equal 0 only in those situations in which all the \( \beta \) parameters equal 0.

The test statistic for \( H_0 \) is denoted by \( F \). It equals the ratio of the mean squares from the ANOVA table,

\[
F = \frac{\text{Mean square for regression}}{\text{Mean square error}}.
\]

We won’t need the formulas for the mean squares here, but the value of \( F \) is proportional to \( R^2/(1 - R^2) \). As \( R^2 \) increases, the \( F \) test statistic increases.

When \( H_0 \) is true, the expected value of the \( F \) test statistic is approximately 1. When \( H_0 \) is false, \( F \) tends to be larger than 1. The larger the \( F \) test statistic, the stronger the evidence against \( H_0 \). The \( P \)-value is the right-tail probability from the sampling distribution of the \( F \) test statistic.

From the ANOVA table in Table 13.7 (shown again in the margin) for the regression model predicting weight, the mean square for regression = 4136.0 and the mean square error = 102.2. Then the test statistic

\[
F = \frac{4136.0}{102.2} = 40.5.
\]

Before conclusions can be made about this test, we must first understand the \( F \) distribution better.

The \( F \) Distribution and Its Properties

The sampling distribution of the \( F \) test statistic is called the \( F \) distribution. The symbol for the \( F \) statistic and its distribution honors the eminent British statistician R. A. Fisher, who discovered the \( F \) distribution in 1922. Like the chi-squared distribution, the \( F \) distribution can assume only nonnegative values and is skewed to the right. See Figure 13.3. The mean is approximately 1.
Section 13.3 Using Multiple Regression to Make Inferences

The precise shape of the \( F \) distribution is determined by two degrees of freedom, denoted by \( df_1 \) and \( df_2 \). These are the \( df \) values in the ANOVA table for the two mean squares whose ratio equals \( F \). The first one is
\[
df_1 = \text{number of explanatory variables in the model.}
\]
The second,
\[
df_2 = n - \text{number of parameters in regression equation},
\]
is the \( df \) for the \( t \) tests about the individual regression parameters. Sometimes \( df_1 \) is called the **numerator df**, because it is listed in the ANOVA table next to the mean square for regression, which goes in the numerator of the \( F \) test statistic. Likewise, \( df_2 \) is called the **denominator df**, because it is listed in the ANOVA table next to the mean square error, which goes in the denominator of the \( F \) test statistic.

Table D at the end of the text lists the \( F \) values that have P-value = 0.05, for various \( df_1 \) and \( df_2 \) values. Table 13.8 shows a small excerpt. For instance, when \( df_1 = 3 \) and \( df_2 = 40 \), \( F = 2.84 \) has P-value = 0.05. When \( F > 2.84 \), the test statistic is farther out in the tail and the P-value < 0.05. (See margin figure.) Regression software reports the actual P-value. The \( F \) value for our example is extremely large (\( F = 40.5 \)). See Table 13.7. With this large value of \( F \) the ANOVA table reports P-value = 0.000.

**Table 13.8 An Excerpt of Table D Displaying \( F \) Values**

These are the values that have right-tail probability equal to 0.05.

<table>
<thead>
<tr>
<th>( df_2 )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>4.17</td>
<td>3.32</td>
<td>2.92</td>
<td>2.69</td>
<td>2.69</td>
</tr>
<tr>
<td>40</td>
<td>4.10</td>
<td>3.23</td>
<td>2.84</td>
<td>2.61</td>
<td>2.45</td>
</tr>
<tr>
<td>60</td>
<td>4.00</td>
<td>3.15</td>
<td>2.76</td>
<td>2.52</td>
<td>2.37</td>
</tr>
</tbody>
</table>

**SUMMARY: \( F \) Test that All the Multiple Regression \( \beta \) Parameters = 0**

1. **Assumptions:**
   - Multiple regression equation holds
   - Data gathered using randomization
   - Normal distribution for \( y \) with same standard deviation at each combination of predictors.

2. **Hypotheses:**
   \[
   H_0: \beta_1 = \beta_2 = \ldots = 0 \quad (\text{all the beta parameters in the model } = 0)
   \]
   \[
   H_a: \text{At least one } \beta \text{ parameter differs from } 0.
   \]

3. **Test statistic:** \( F = \frac{\text{mean square for regression}}{\text{mean square error}} \)

4. **P-value:** Right-tail probability above observed \( F \) test statistic value from \( F \) distribution with
   \[
   df_1 = \text{number of explanatory variables},
   \]
   \[
   df_2 = n - \text{ (number of parameters in regression equation)}.
   \]

5. **Conclusion:** The smaller the P-value, the stronger the evidence that at least one explanatory variable has an effect on \( y \). If a decision is needed, reject \( H_0 \) if P-value \( \leq \) significance level, such as 0.05. Interpret in context.
Chapter 13  Multiple Regression

Inferences About Individual Regression Parameters

For the bivariate model, \( \mu_y = \alpha + \beta x \), there's a \( t \) test for the null hypothesis
\( H_0: \beta = 0 \) that \( x \) and \( y \) are statistically independent. Likewise, a \( t \) test applies to any slope parameter in multiple regression. Let's consider a particular parameter, say \( \beta_1 \).

If \( \beta_1 = 0 \), the mean of \( y \) is identical for all values of \( x_1 \), at fixed values for the other explanatory variables. So, \( H_0: \beta_1 = 0 \) states that \( y \) and \( x_1 \) are statistically independent, controlling for the other variables. This means that once the other

---

**Example 5**

**Athletes’ Weight**

**Picture the Scenario**

For the 64 female college athletes, the ANOVA table for the multiple regression predicting \( y = \) weight using \( x_1 = \) height, \( x_2 = \) percent body fat, and \( x_3 = \) age shows:

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>3</td>
<td>12407.9</td>
<td>4136.0</td>
<td>40.48</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>60</td>
<td>6131.0</td>
<td>102.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Questions to Explore**

a. State and interpret the null hypothesis tested in this table.

b. From the \( F \) table, which \( F \) value would have a P-value of 0.05 for these data?

c. Report the observed test statistic and P-value. Interpret the P-value, and make a decision for a 0.05 significance level.

**Think It Through**

a. Since there are three explanatory variables, the null hypothesis is \( H_0: \beta_1 = \beta_2 = \beta_3 = 0 \). It states that weight is independent of height, percent body fat, and age.

b. In the DF column, the ANOVA table shows \( df_1 = 3 \) and \( df_2 = 60 \). The \( F \) table or software indicates that the \( F \) value with right-tail probability of 0.05 is 2.76. (See also Table 13.8.)

c. From the ANOVA table, the observed \( F \) test statistic value is 40.5. Since this is well above 2.76, the P-value is less than 0.05. The ANOVA table reports P-value = 0.000. If \( H_0 \) were true, it would be extremely unusual to get such a large \( F \) test statistic. We can reject \( H_0 \) at the 0.05 significance level. In summary, we conclude that at least one predictor has an effect on weight.

**Insight**

The \( F \) test tells us that at least one explanatory variable has an effect. The following section discusses how to follow up from the \( F \) test to investigate which explanatory variables have a statistically significant effect on predicting \( y \).

**Try Exercise 13.25**

**Recall**

Section 12.3 showed that the \( t \) test statistic for \( H_0: \beta = 0 \) in the bivariate model is

\[
t = \frac{\text{sample slope} - 0}{\text{std. error of sample slope}} = \frac{b - 0}{se},
\]

with \( df = n - 2 \).
Section 13.3 Using Multiple Regression to Make Inferences

explanatory variables are in the model, it doesn’t help to have $x_1$ in the model. The alternative hypothesis usually is two sided, $H_0: \beta_1 \neq 0$, but one-sided alternative hypotheses are also possible.

The test statistic for $H_0: \beta_1 = 0$ is

$$t = (b_1 - 0)/se,$$

where $se$ is the standard error of the slope estimate $b_1$. Software provides the $se$ value, the $t$ test statistic, and the P-value. If $H_0$ is true and the inference assumptions hold, the $t$ test statistic has the $t$ distribution. The degrees of freedom are

$$df = n - k,$$

where $n$ is the sample size and $k$ is the number of parameters in the regression equation.

The degrees of freedom are also equal to those used to calculate residual standard deviation.

The bivariate model $\mu_y = \alpha + \beta x$ has two parameters ($\alpha$ and $\beta$), so $df = n - 2$, as Section 12.3 used. The model with two predictors, $\mu_y = \alpha + \beta_1 x_1 + \beta_2 x_2$, has three parameters, so $df = n - 3$. Likewise, with three predictors, $df = n - 4$, and so on.

### SUMMARY: Significance Test About a Multiple Regression Parameter (such as $\beta_1$)

1. **Assumptions:**
   - Each explanatory variable has a straight-line relation with $\mu_y$, with the same slope for all combinations of values of other predictors in model
   - Data gathered with randomization (such as a random sample or a randomized experiment)
   - Normal distribution for $y$ with same standard deviation at each combination of values of other predictors in model

2. **Hypotheses:**
   - $H_0: \beta_1 = 0$
   - $H_a: \beta_1 \neq 0$

   When $H_0$ is true, $y$ is independent of $x_1$, controlling for the other predictors.

3. **Test statistic:** $t = (b_1 - 0)/se$. Software supplies the slope estimate $b_1$, its $se$, and the $t$ value.

4. **P-value:** Two-tail probability from $t$ distribution of values larger than observed $t$ test statistic (in absolute value). The $t$ distribution has

   $$df = n - k,$$

   where $k$ is the number of parameters in regression equation (such as $df = n - 3$ when $\mu_y = \alpha + \beta_1 x_1 + \beta_2 x_2$, which has three parameters).

5. **Conclusion:** Interpret P-value in context; compare to significance level if decision needed.

### Example 6

**What Helps Predict a Female Athlete’s Weight?**

**Picture the Scenario**

The College Athletes data set (Examples 4 and 5) measured several physical characteristics, including total body weight in pounds (TBW), height in inches (HGT), the percent of body fat (%BF), and age. Table 13.9 shows summary statistics for these variables.
Table 13.9 Summary Statistics for Study of Female College Athletes
Variables are TBW = total body weight, HGT = height, %BF = percent body fat, and AGE.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>StDev</th>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBW</td>
<td>133.0</td>
<td>17.2</td>
<td>96.0</td>
<td>119.2</td>
<td>131.5</td>
<td>143.8</td>
<td>185.0</td>
</tr>
<tr>
<td>HGT</td>
<td>65.5</td>
<td>3.5</td>
<td>56.0</td>
<td>63.0</td>
<td>65.0</td>
<td>68.2</td>
<td>75.0</td>
</tr>
<tr>
<td>%BF</td>
<td>18.4</td>
<td>4.1</td>
<td>11.2</td>
<td>15.2</td>
<td>18.5</td>
<td>21.5</td>
<td>27.6</td>
</tr>
<tr>
<td>AGE</td>
<td>20.0</td>
<td>1.98</td>
<td>17.0</td>
<td>18.0</td>
<td>20.0</td>
<td>22.0</td>
<td>23.0</td>
</tr>
</tbody>
</table>

Table 13.10 shows results of fitting a multiple regression model for predicting weight using the other variables. The predictive power is good, with $R^2 = 0.669$.

Table 13.10 Multiple Regression Analysis for Predicting Weight
Predictors are HGT = height, %BF = body fat, and age of subject.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-97.69</td>
<td>28.79</td>
<td>-3.39</td>
<td>0.001</td>
</tr>
<tr>
<td>HGT</td>
<td>3.4285</td>
<td>0.3679</td>
<td>9.32</td>
<td>0.000</td>
</tr>
<tr>
<td>%BF</td>
<td>1.3643</td>
<td>0.3126</td>
<td>4.36</td>
<td>0.000</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.9601</td>
<td>0.6483</td>
<td>-1.48</td>
<td>0.144</td>
</tr>
</tbody>
</table>

R-Sq = 66.9%

Questions to Explore

a. Interpret the effect of age on weight in the multiple regression equation.

b. In the population, does age help you to predict weight if you already know height and percent body fat? Show all steps of a significance test, and interpret.

Think It Through

a. Let $\hat{y}$ = predicted weight, $x_1 = \text{height}$, $x_2 = \%\text{body fat}$, and $x_3 = \text{age}$. Then

$$\hat{y} = -97.7 + 3.43x_1 + 1.36x_2 - 0.96x_3.$$ 

The slope coefficient of age is -0.96. The sample effect of age on weight is negative, which may seem surprising, but in practical terms it is small: For athletes having fixed values of $x_1$ and $x_2$, the predicted weight decreases by only 0.96 pounds for a one-year increase in age, and the ages vary only from 17 to 23.

b. If $\beta_3 = 0$, then $x_3 = \text{age}$ has no effect on weight in the population, controlling for height and body fat. The hypothesis that age does not help us better predict weight, if we already know height and body fat, is $H_0: \beta_3 = 0$. Here are the steps:

1. Assumptions: The 64 female athletes were a convenience sample, not a random sample. Although the goal was to make inferences about all female college athletes, inferences are tentative. We’ll discuss the other assumptions and learn how to check them in Section 13.4.

2. Hypotheses: The null hypothesis is $H_0: \beta_3 = 0$. Since there’s no prior prediction about whether the effect of age is positive or negative (for fixed values of $x_1$ and $x_2$), we use the two-sided $H_a: \beta_3 \neq 0$. 


Section 13.3 Using Multiple Regression to Make Inferences

As usual, a test merely tells us whether the null hypothesis is plausible. In Example 6 we saw that $\beta_3$ may equal 0, but what are its other plausible values? A confidence interval answers this question.

- **Test statistic:** Table 13.10 reports a slope estimate of $-0.960$ for age and a standard error of $se = 0.648$. It also reports the $t$ test statistic of 
  $$t = (b_3 - 0)/se = -0.960/0.648 = -1.48.$$ 
  Since the sample size equals $n = 64$ and the regression equation has four parameters, the degrees of freedom are $df = n - 4 = 60$.

- **P-value:** Table 13.10 reports $P$-value $= 0.14$. This is the two-tailed probability of a $t$ statistic below $-1.48$ or above $1.48$, if $H_0$ were true.

- **Conclusion:** The P-value of 0.14 does not give much evidence against the null hypothesis that $\beta_3 = 0$. At common significance levels, such as 0.05, we cannot reject $H_0$. Age does not significantly predict weight if we already know height and percentage of body fat. These conclusions are tentative because the sample of 64 female athletes was selected using a convenience sample rather than a random sample.

**Insight**

By contrast, Table 13.10 shows that $t = 9.3$ for testing the effect of height ($H_0: \beta_1 = 0$) and $t = 4.4$ for testing the effect of %BF ($H_0: \beta_2 = 0$). Both P-values are 0.000. It does help to have each of these variables in the model, given the other two.

**Try Exercise 13.19**

As usual, a test merely tells us whether the null hypothesis is plausible. In Example 6 we saw that $\beta_3$ may equal 0, but what are its other plausible values? A confidence interval answers this question.

**Confidence Interval for a Multiple Regression $\beta$ Parameter**

A 95% confidence interval for a $\beta$ slope parameter in multiple regression equals

$$\text{Estimated slope} \pm t_{0.025}(se).$$

The $t$-score has $df = n - \text{number of parameters in regression equation}$, as in the $t$ test. The assumptions are also the same as for the $t$ test.

**Example 7**

What’s Plausible for the Effect of Age on Weight?

**Picture the Scenario**

For the college athletes data, consider the multiple regression analysis of $y = \text{weight}$ and predictors $x_1 = \text{height}$, $x_2 = \%\text{body fat}$, and $x_3 = \text{age}$.

**Question to Explore**

Find and interpret a 95% confidence interval for $\beta_3$, the effect of age while controlling for height and percent of body fat.

**Think It Through**

From the previous example, $df = 60$. For $df = 60, t_{0.025} = 2.00$. From Table 13.10 (shown partly in the margin), the estimate of $\beta_3$ is $-0.96$, with $se = 0.648$. The 95% confidence interval equals

$$b_3 \pm t_{0.025}(se) = -0.96 \pm 2.00(0.648),$$

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-97.7</td>
<td>28.8</td>
</tr>
<tr>
<td>HGT</td>
<td>3.43</td>
<td>0.368</td>
</tr>
<tr>
<td>%BF</td>
<td>1.36</td>
<td>0.313</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.96</td>
<td>0.648</td>
</tr>
</tbody>
</table>
Unless the sample size is small and the correlation is weak between $y$ and each explanatory variable, the $F$ test usually has a small P-value. If the explanatory variables are chosen sensibly, at least one should have some predictive power. When the P-value is small, we can conclude merely that at least one explanatory variable affects $y$. The more narrowly focused $t$ inferences about individual slopes judge which effects are nonzero and estimate their sizes.

<table>
<thead>
<tr>
<th>Recall</th>
</tr>
</thead>
</table>
The multiple regression equation using all three predictors was
$$y = -97.7 + 3.43x_1 + 1.36x_2 - 0.96x_3.$$ The effects of height and percent body fat are similar in the equation without $x_3 = \text{age}$. |

| Caution |
Procedures for regression model selection must be used with care. Consultation with a trained statistician is recommended. |

| Model Building |
Unless the sample size is small and the correlation is weak between $y$ and each explanatory variable, the $F$ test usually has a small P-value. If the explanatory variables are chosen sensibly, at least one should have some predictive power. When the P-value is small, we can conclude merely that at least one explanatory variable affects $y$. The more narrowly focused $t$ inferences about individual slopes judge which effects are nonzero and estimate their sizes.

| Insight |
The confidence interval contains 0. Age may have no effect on weight, once we control for height and percent body fat. This is in agreement with not rejecting $H_0: \beta_3 = 0$ in favor of $H_a: \beta_3 \neq 0$ at the $\alpha = 0.05$ level in the significance test.

| Try Exercise 13.20 |

| In Practice |
The Overall $F$ Test is Done Before the Individual $t$ Inferences
As illustrated in the previous examples, the $F$ test is typically performed first before looking at the individual $t$ inferences. The $F$ test result tells us if there is sufficient evidence to make it worthwhile to consider the individual effects. When there are many explanatory variables, doing the $F$ test first provides protection from doing lots of $t$ tests and having one of them be significant merely by random variation when, in fact, there truly are no effects in the population.

After finding the highly significant $F$ test result of Example 5, we would study the individual effects, as we did in Examples 6 and 7. When we look at an individual $t$ inference, suppose we find that the plausible values for a parameter are all relatively near 0. This was the case for the effect of age on weight (controlling for the other variables) as shown by the confidence interval in Example 7. Then, to simplify the model, you can remove that predictor, refitting the model with the other predictors. When we do this for $y = \text{weight}$, $x_1 = \text{height}$, and $x_2 = \text{percent body fat}$, we get
$$\hat{y} = -121.0 + 3.50x_1 + 1.35x_2.$$ For the simpler model, $R^2 = 0.66$. This is nearly as large as the value of $R^2 = 0.67$ with age also in the model. The predictions of weight are essentially as good without age in the model.

When you have several potential explanatory variables for a multiple regression model, how do you decide which ones to include? As just mentioned, the lack of statistical and practical significance is one criterion for deleting a term from the model. Many other possible criteria can also be considered. Most regression software has automatic procedures that successively add or delete predictor variables from models according to criteria such as statistical significance. Three of the most common procedures are backward elimination, forward selection, and stepwise regression. It is possible that all three procedures result in identical models; however, this is not always the case. In a situation where you must choose between models it is best to look at $R^2$ values and sensibility of model. These procedures must be used with great caution. There is no guarantee that the final model chosen is sensible. Model selection requires quite a bit of statistical sophistication, and for applications with many potential explanatory variables we recommend that you seek guidance from a statistician.
13.3 Practicing the Basics

13.19 Predicting GPA  For the 59 observations in the Georgia Student Survey data file on the text CD, the result of regressing college GPA on high school GPA and study time follows.

College GPA, high school GPA, and study time

The regression equation is

\[ CGPA = 1.13 + 0.643HSGPA + 0.0078\text{Studytime} \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1.1262</td>
<td>0.5690</td>
<td>1.98</td>
<td>0.053</td>
</tr>
<tr>
<td>HSGPA</td>
<td>0.6434</td>
<td>0.1458</td>
<td>4.41</td>
<td>0.000</td>
</tr>
<tr>
<td>Studytime</td>
<td>0.00776</td>
<td>0.01614</td>
<td>0.48</td>
<td>0.633</td>
</tr>
</tbody>
</table>

\( S = 0.3188 \quad \text{R-Sq} = 25.8\% \)

a. Explain in nontechnical terms what it means if the population slope coefficient for high school GPA equals 0.

b. Show all steps for testing the hypothesis that this slope equals 0.

13.20 Study time help GPA?  Refer to the previous exercise.

a. Report and interpret the P-value for testing the hypothesis that the population slope coefficient for study time equals 0.

b. Find a 95% confidence interval for the true slope for study time. Explain how the result is in accord with the result of the test in part a.

c. Does the result in part a imply that in the corresponding population, study time has no association with college GPA? Explain. (Hint: What is the impact of also having HSGPA in the model?)

13.21 Variability in college GPA  Refer to the previous two exercises.

a. Report the residual standard deviation. What does this describe?

b. Interpret the residual standard deviation by predicting where approximately 95% of the Georgia college GPAs fall when high school GPA = 3.80 and study time = 5.0 hours per day, which are the sample means.

13.22 Does leg press help predict body strength?  Chapter 12 analyzed strength data for 57 female high school athletes. Upper body strength was summarized by the maximum number of pounds the athlete could bench press (denoted BP below, 1RM Bench in file). This was predicted well by the number of times she could do a 60-pound bench press (denoted BP_60 in output, BRTF(60) in file). Can we predict BP even better if we also know how many times an athlete can perform a 200-pound leg press? The table shows results after adding this second predictor (denoted LP_200 in output, LP RTF(200) in file) to the model.

Multiple regression analysis of strength data

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>60.596</td>
<td>2.872</td>
<td>21.10</td>
<td>0.000</td>
</tr>
<tr>
<td>BP_60</td>
<td>1.3318</td>
<td>0.1876</td>
<td>7.10</td>
<td>0.000</td>
</tr>
<tr>
<td>LP_200</td>
<td>0.2110</td>
<td>0.1519</td>
<td>1.39</td>
<td>0.171</td>
</tr>
</tbody>
</table>

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>6473.3</td>
<td>3236.6</td>
<td>51.39</td>
<td>0.000</td>
</tr>
<tr>
<td>Res. Error</td>
<td>54</td>
<td>3401.3</td>
<td>63.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>56</td>
<td>9874.6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( S = 7.93640 \quad \text{R-Sq} = 65.6\% \)

a. Does LP_200 have a significant effect on BP, if BP_60 is also in the model? Show all steps of a significance test to answer this.

b. Show that the 95% confidence interval for the slope for LP_200 equals 0.21 ± 0.30, roughly (−0.1, 0.5). Based on this interval, does LP_200 seem to have a strong impact, or a weak impact, on predicting BP, if BP_60 is also in the model?

c. Given that LP_200 is in the model, provide evidence from a significance test that shows why it does help to add BP_60 to the model.

13.23 Leg press uncorrelated with strength?  The P-value of 0.17 in part a of the previous exercise suggests that LP_200 plausibly had no effect on BP, once BP_60 is in the model. Yet when LP_200 is the sole predictor of BP, the correlation is 0.58 and the significance test for its effect has a P-value of 0.000, suggesting very strong evidence of an effect. Explain why this is not a contradiction.

13.24 Interpret strength variability  Refer to the previous two exercises. The sample standard deviation of BP was 13.3. The residual standard deviation of BP when BP_60 and LP_200 are predictors in a multiple regression model is 7.9.

a. Explain the difference between the interpretations of these two standard deviations.

b. If the conditional distributions of BP are approximately bell shaped, explain why most maximum bench press values fall within about 16 pounds of the regression equation, when the predictors BP_60 and LP_200 are near their sample mean values.

c. At x_1 = 11 and x_2 = 22, which are close to the sample mean values, software reports \( \hat{y} = 80 \) and a 95% prediction interval of 80 ± 16, or (64, 96). Is this interval an inference about where the population BP values fall or where the population mean of the BP values fall (for subjects having x_1 = 11 and x_2 = 22)? Explain.

d. Refer to part c. Would it be unusual for a female athlete with these predictor values to be able to bench press more than 100 pounds? Why?

13.25 Any predictive power?  Refer to the previous three exercises.

a. State and interpret the null hypothesis tested with the F statistic in the ANOVA table given in Exercise 13.22.

b. From the F table (Table D), which F statistic value would have a P-value of 0.05 for these data?

c. Report the observed F test statistic and its P-value. Interpret the P-value, and make a decision for a 0.05 significance level. Explain in nontechnical terms what the result of the test means.
13.26 Predicting pizza revenue  Aunt Erma’s Pizza restaurant keeps monthly records of total revenue, amount spent on TV advertising, and amount spent on newspaper advertising.

a. Specify notation and formulate a multiple regression equation for predicting the monthly revenue. Explain how to interpret the parameters in the equation.

b. State the null hypothesis that you would test if you want to analyze whether TV advertising is helpful, for a given amount of newspaper advertising.

c. State the null hypothesis that you would test if you want to analyze whether at least one of the sources of advertising has some effect on monthly revenue.

13.27 Regression for mental health  A study in Alachua County, Florida, investigated an index of mental health impairment, which had \( \bar{y} = 27.3 \) and \( s = 5.5 \). Two explanatory variables were \( x_1 \) = life events score (mean = 44.4, \( s = 22.6 \)) and \( x_2 \) = SES (socioeconomic status, mean = 56.6, \( s = 25.3 \)). Life events is a composite measure of the number and severity of major life events, such as death in the family, that the subject experienced within the past three years. SES is a composite index based on occupation, income, and education. The table shows data for 6 subjects. The complete data is available as a text file on the text CD. Some regression results are also shown.

a. Find the 95% confidence interval for \( \beta_2 \).

b. Explain why the interval in part a means that an increase of 100 units in life events corresponds to anything from a 4- to 17-unit increase in mean mental impairment, controlling for SES. (This lack of precision reflects the small sample size.)

13.28 Mental health again  Refer to the previous exercise.

a. Report the test statistic and P-value for testing \( H_0: \beta_1 = \beta_2 = 0 \).

b. State the alternative hypothesis that is supported by the result in part a.

c. Does the result in part a imply that necessarily both life events and SES are needed in the model? Explain.

13.29 More predictors for house price  The MINITAB results are shown for predicting selling price using \( x_1 \) = size of home, \( x_2 \) = number of bedrooms, and \( x_3 \) = age.

Regression of selling price on house size, number of bedrooms, and age

The regression equation is

\[
HP = 80.5 + 0.0626 \times House\ Size + 13.5 \times Bedrooms - 0.418 \times Age
\]

a. State the null hypothesis for an \( F \) test, in the context of these variables.

b. The \( F \) statistic equals 74.23, with \( P \)-value = 0.000. Interpret.

c. Explain in nontechnical terms what you learn from the results of the \( t \) tests reported in the table for the three explanatory variables.

13.30 House prices  Use software to do further analyses with the multiple regression model of \( y = \) selling price of home in thousands, \( x_1 = \) size of home, and \( x_2 = \) number of bedrooms, considered in Section 13.1. The data file House Selling Prices OR is on the text CD.

a. Report the \( F \) statistic and state the hypotheses to which it refers. Report its \( P \)-value, and interpret. Why is it not surprising to get a small \( P \)-value for this test?

b. Report and interpret the \( t \) statistic and \( P \)-value for testing \( H_0: \beta_2 = 0 \) against \( H_a: \beta_2 > 0 \).

c. Construct a 95% confidence interval for \( \beta_3 \), and interpret. This inference is more informative than the test in part b. Explain why.
the other explanatory variables), (2) the data were gathered randomly, and (3) \( y \) has a normal distribution with the same standard deviation at each combination of predictors. Now, let’s see how to check the assumptions about the regression equation and the distribution of \( y \).

For bivariate regression, the scatterplot provides a simple visual check of whether the straight-line model is appropriate. For multiple regression, a plot of all the variables at once would require many dimensions. So instead, we study how the predicted values \( \hat{y} \) compare to the observed values \( y \). This is done using the residuals, \( y - \hat{y} \).

### Checking Shape and Detecting Unusual Observations

Consider first the assumption that the conditional distribution of \( y \) is normal, at any fixed values of the explanatory variables. If this is true, the residuals should have approximately a bell-shaped histogram. Nearly all the standardized residuals should fall between about \(-3\) and \(+3\). A standardized residual below \(-3\) or above \(+3\) indicates a potential regression outlier.

When some observations have large standardized residuals, you should think about whether or not there is an explanation. Often this merely reflects skew in the conditional distribution of \( y \), with a long tail in one direction. Other times the observations differ from the others on a variable that was not included in the model. Once that variable is added, those observations cease to be so unusual. For instance, suppose the data file from the weight study of Examples 4–7 also contained observations for a few males. Then we might have observed a few \( y \) values considerably above the others and with very large positive standardized residuals. These residuals would probably diminish considerably once the model also included gender as a predictor.

**Recall**

As discussed in Section 12.4, software can plot a histogram of the residuals or the standardized residuals, which are the residuals divided by their standard errors.

**Example 8**

**House Selling Price**

**Picture the Scenario**

For the House Selling Price OR data set (Examples 1–3), Figure 13.4 is a MINITAB histogram of the standardized residuals for the multiple regression model predicting selling price by the house size and the number of bedrooms.

![Figure 13.4 Histogram of Standardized Residuals for Multiple Regression Model Predicting Selling Price.](image)
Question to Explore
What does Figure 13.4 tell you? Interpret.

Think It Through
The residuals are roughly bell shaped about 0. They fall mostly between about −3 and +3. The shape suggests that the conditional distribution of the response variable may have a bit of skew, but no severe nonnormality is indicated.

Insight
When \( n \) is small, don’t make the mistake of reading too much into such plots. We’re mainly looking for dramatic departures from the assumptions and highly unusual observations that stand apart from the others.

Try Exercise 13.31

Two-sided inferences about slope parameters are robust. The normality assumption is not as important as the assumption that the regression equation approximates the true relationship between the predictors and the mean of \( y \). We consider that assumption next.

Plotting Residuals Against Each Explanatory Variable

Plots of the residuals against each explanatory variable help us check for potential problems with the regression model. We discuss this in terms of \( x_1 \), but you should view a plot for each predictor. Ideally, the residuals should fluctuate randomly about the horizontal line at 0, as in Figure 13.5a. There should be no obvious change in trend or change in variation as the values of \( x_1 \) increase.

By contrast, a scattering of the residuals as in Figure 13.5b suggests that \( y \) is actually nonlinearily related to \( x_1 \). It suggests that \( y \) tends to be below \( \hat{y} \) for very small and very large \( x_1 \) values (giving negative residuals) and above \( \hat{y} \) for medium-sized \( x_1 \) values (giving positive residuals).

A common occurrence is that residuals have increased variation as \( x_1 \) increases, as seen in Figure 13.5c. A pattern like this indicates that the residual standard deviation of \( y \) is not constant—the data display more variability at larger values of \( x_1 \) (see the margin figure). Two-sided inferences about slope parameters still perform well. However, ordinary prediction intervals are invalid. The width of such intervals is similar at relatively high and relatively low \( x_1 \) values, but in reality, observations on \( y \) are displaying more variability at high \( x_1 \) values than at low \( x_1 \) values. Figure 13.5a suggests that \( y \) is linearly related to \( x_1 \). It also suggests there is a constant variation as \( x_1 \) increases. Therefore, Assumption 1 and part of Assumption 3 are met.
Section 13.4 Checking a Regression Model Using Residual Plots

**In Practice** Use Caution in Interpreting Residual Patterns

Residual patterns are often not as neat as the ones in Figure 13.5. Be careful not to let a single outlier or ordinary sampling variability overly influence your reading of a pattern from the plot.

### Example 9

**Plotting residuals**

**House Selling Price**

**Picture the Scenario**

For the House Selling Price OR data set, Figure 13.6 is a residual plot for the multiple regression model relating selling price to house size and to number of bedrooms. It plots the standardized residuals against house size.

![Figure 13.6 Standardized Residuals of Selling Price Plotted Against House Size, for Model With House Size and Number of Bedrooms as Predictors. Questions How does this plot suggest that selling price has more variability at higher house size values, for given number of bedrooms? You don’t see number of bedrooms on the plot, so how do its values affect the analysis?](image)

**Question to Explore**

Does this plot suggest any irregularities with the model?

**Think It Through**

As house size increases, the variability of the standardized residuals seems to increase. This suggests more variability in selling prices when house size is larger, for a given number of bedrooms. We must be cautious, though, because the few points with large negative residuals for the largest houses and the one point with a large positive residual may catch our eyes more than the others. Generally, it’s not a good idea to allow a few points to overly influence your judgment about the shape of a residual pattern. However, there is definite evidence that the variability increases as house size increases. A larger data set would provide more evidence about this.

**Insight**

Nonconstant variability does not invalidate the use of multiple regression. It would, however, make us cautious about using prediction intervals. We
would expect predictions about selling price to have smaller prediction errors when house size is small than when house size is large.

Try Exercise 13.32

We have seen several examples illustrating the components of multiple regression analysis. A summary of the entire process of multiple regression follows.

SUMMARY: The Process of Multiple Regression
Steps should include:
1. Identify response and potential explanatory variables
2. Create a multiple regression model; perform appropriate tests (F and t) to see if and which explanatory variables have a statistically significant effect in predicting y
3. Plot y versus ŷ for resulting models and find R and R² values
4. Check assumptions (residual plot, randomization, residuals histogram)
5. Choose appropriate model
6. Create confidence intervals for slope
7. Make predictions at specified levels of explanatory variables
8. Create prediction intervals

13.4 Practicing the Basics

13.31 Body weight residuals
Examples 4–7 used multiple regression to predict total body weight of college athletes. The figure shows the standardized residuals for another multiple regression model for predicting weight.

a. About which distribution do these give you information—the overall distribution of weight or the conditional distribution of weight at fixed values of the predictors?
b. What does the histogram suggest about the likely shape of this distribution? Why?

13.32 Strength residuals
In Chapter 12, we analyzed strength data for a sample of female high school athletes. The following figure is a residual plot for the multiple regression model relating the maximum number of pounds the athlete could bench press (BP) to the number of 60-pound bench presses (BP_60) and the number of 200-pound leg presses (LP_200). It plots the standardized residuals against the values of LP_200.

a. You don’t see BP_60 on the plot, so how do its values affect the analysis?
b. Explain how the plot might suggest less variability at the lower values of LP_200.
c. Suppose you remove the three points with standardized residuals around –2. Then is the evidence about variability in part b so clear? What does this suggest about cautions in looking at residual plots?
13.33 More residuals for strength  Refer to the previous exercise. The following figure is a residual plot for the model relating maximum bench press to LP_200 and BP_60. It plots the standardized residuals against the values of BP_60. Does this plot suggest any irregularities with the model? Explain.

![Residual plot for Exercise 13.37](image)

13.34 Nonlinear effects of age  Suppose you fit a straight-line regression model to \( y = \text{amount of time sleeping per day} \) and \( x = \text{age of subject} \). Values of \( y \) in the sample tend to be quite large for young children and for elderly people, and they tend to be lower for other people. Sketch what you would expect to observe for (a) the scatterplot of \( x \) and \( y \) and (b) a plot of the residuals against the values of age.

13.35 Driving accidents  Suppose you fit a straight-line regression model to \( x = \text{age of subjects} \) and \( y = \text{driving accident rate} \). Sketch what you would expect to observe for (a) the scatterplot of \( x \) and \( y \) and (b) a plot of the residuals against the values of age.

13.36 Why inspect residuals?  When we use multiple regression, what's the purpose of doing a residual analysis? Why can't we just construct a single plot of the data for all the variables at once in order to tell whether the model is reasonable?

13.37 College athletes  The College Athletes data set on the text CD comes from a study of University of Georgia female athletes. The response variable BP = maximum bench press (1RM in data set) has explanatory variables LBM = lean body mass (which is weight times 1 minus the proportion of body fat) and REP_BP = number of repetitions before fatigue with a 70-pound bench press (REPS70 in data set). Let's look at all the steps of a regression analysis for these data.

a. The first figure shows a scatterplot matrix. Which two plots in the figure describe the associations with BP as a response variable? Describe those associations.

b. Results of a multiple regression analysis are shown in the next column. Write down the prediction equation, and interpret the coefficient of REP_BP.

c. Report \( R^2 \), and interpret its value in the context of these variables.

d. Based on the value of \( R^2 \), report and interpret the multiple correlation.

e. Interpret results of the \( F \) test that BP is independent of these two predictors. Show how to obtain the \( F \) statistic from the mean squares in the ANOVA table.

f. Once REP_BP is in the model, does it help to have LBM as a second predictor? Answer by showing all steps of a significance test for a regression parameter.

13.38 Examine the histogram shown of the residuals for the multiple regression model. What does this describe, and what does it suggest?

h. Examine the plot shown of the residuals plotted against values of REP_BP. What does this describe, and what does it suggest?

i. From the plot in part h, can you identify a subject whose BP value was considerably lower than expected based on the predictor values? Identify by indicating the approximate values of REP_BP and the standardized residual for that subject.

### Regression of maximum bench press on LBM and REP_BP

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>55.012</td>
<td>7.740</td>
<td>7.11</td>
<td>0.000</td>
</tr>
<tr>
<td>LBM</td>
<td>0.167</td>
<td>0.075</td>
<td>2.22</td>
<td>0.030</td>
</tr>
<tr>
<td>REP_BP</td>
<td>1.657</td>
<td>0.109</td>
<td>15.21</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\( S = 7.11957 \)  \( R^2 = 83.2\% \)

### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>15283.0</td>
<td>7641.5</td>
<td>150.75</td>
<td>0.000</td>
</tr>
<tr>
<td>Res.</td>
<td>61</td>
<td>3092.0</td>
<td>50.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>63</td>
<td>18375.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
13.5 Regression and Categorical Predictors

So far, we’ve studied regression models for quantitative variables. Next, we’ll learn how to include a categorical explanatory variable. The final section shows how to perform regression for a categorical response variable.

Indicator Variables

Regression models specify categories of a categorical explanatory variable using artificial variables, called indicator variables. The indicator variable for a particular category is binary. It equals 1 if the observation falls into that category and it equals 0 otherwise.

In the house selling prices (Oregon) data set, the condition of the house is a categorical variable. It was measured with categories (good, not good). The indicator variable $x$ for condition is

\[ x = 1 \text{ if house in good condition} \]
\[ x = 0 \text{ otherwise}. \]

This indicator variable indicates whether or not a home is in good condition. Let’s see how an indicator variable works in a regression equation. To start, suppose condition is the only predictor of selling price. The regression model is then $\mu_y = \alpha + \beta x$, with $x$ as just defined. Substituting the possible values 1 and 0 for $x$,

\[ \mu_y = \alpha + \beta(1) = \alpha + \beta, \text{ if house is in good condition (so } x = 1) \]
\[ \mu_y = \alpha + \beta(0) = \alpha, \text{ if house is not in good condition (so } x = 0). \]

The difference between the mean selling price for houses in good condition and other conditions is

\[ (\mu_y \text{ for good condition}) - (\mu_y \text{ for other}) = (\alpha + \beta) - \alpha = \beta. \]

The coefficient $\beta$ of the indicator variable $x$ is the difference between the mean selling prices for homes in good condition and for homes not in good condition.
Example 10

Including Condition in Regression for House Selling Price

Picture the Scenario
Let’s now fit a regression model for \( y = \) selling price of home using \( x_1 = \) house size and \( x_2 = \) condition of the house. Table 13.11 shows MINITAB output.

Table 13.11 Regression Analysis of \( y = \) Selling Price Using \( x_1 = \) House Size and \( x_2 = \) Indicator Variable for Condition (Good, Not Good)

The regression equation is

\[
\text{House Price} = 96271 + 66.5 \times x_1 + 12927 \times x_2
\]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>96271</td>
<td>13465</td>
<td>7.15</td>
<td>0.000</td>
</tr>
<tr>
<td>House Size</td>
<td>66.463</td>
<td>4.682</td>
<td>14.20</td>
<td>0.000</td>
</tr>
<tr>
<td>Condition</td>
<td>12927</td>
<td>17197</td>
<td>0.75</td>
<td>0.453</td>
</tr>
</tbody>
</table>

\( S = 81787.4 \) R-Sq = 50.6%

Questions to Explore

a. Find and plot the lines showing how predicted selling price varies as a function of house size, for homes in good condition or not in good condition.

b. Interpret the coefficient of the indicator variable for condition.

Think It Through

a. Table 13.11 reports the prediction equation,

\[
\hat{y} = 96,271 + 66.5x_1 + 12,927x_2
\]

For homes not in good condition, \( x_2 = 0 \). The prediction equation for \( y = \) selling price using \( x_1 = \) house size then simplifies to

\[
\hat{y} = 96,271 + 66.5x_1 + 12,927(0) = 96,271 + 66.5x_1
\]

For homes in good condition, \( x_2 = 1 \). The prediction equation then simplifies to

\[
\hat{y} = 96,271 + 66.5x_1 + 12,927(1) = 109,198 + 66.5x_1
\]

Both lines have the same slope, 66.5. For homes in good condition or not in good condition, the predicted selling price increases by \$66.5 for each square-foot increase in house size. Figure 13.7 plots the two prediction equations. The quantitative explanatory variable, house size, is on the \( x \)-axis. The figure portrays a separate line for each category of condition (good, other).

b. At a fixed value of \( x_1 = \) house size, the difference between the predicted selling prices for homes in good (1) versus not good (0) condition is

\[
(109,198 + 66.5x_1) - (96,271 + 66.5x_1) = 12,927
\]

This is precisely the coefficient of the indicator variable, \( x_2 \). For any fixed value of house size, we predict that the selling price is \$12,927 higher for homes that are good versus not in good condition.
A categorical variable having more than two categories uses an additional indicator variable for each category. For instance, if we want to use all three categories good, average, fair in the model, we could use two indicator variables:

\[ x_1 = 1 \text{ for houses in good condition, and } x_1 = 0 \text{ otherwise}, \]

\[ x_2 = 1 \text{ for houses in average condition, and } x_2 = 0 \text{ otherwise}, \]

If \( x_1 = x_2 = 0 \), the house is in fair condition. We don’t need a separate indicator variable for fair, as it would be redundant. We can tell whether a house is in the fair condition merely from seeing the values of \( x_1 \) and \( x_2 \). Generally, a categorical explanatory variable in a regression model uses one fewer indicator variable than the number of categories. For instance, with the two categories (good condition, other), we needed only a single indicator variable.

Why can’t we specify the three categories merely by setting up a variable \( x \) that equals 1 for homes in good condition, 0 for homes in average condition, and −1 for fair condition? Because this would treat condition as quantitative rather than categorical. It would treat condition as if different categories corresponded to different amounts of the variable. But the variable measures which condition, not how much condition. Treating it as quantitative is inappropriate.

**Determining if Interaction Exists**

In Example 10, the regression equation simplified to two straight lines:

\[ \hat{y} = 109.198 + 66.5x_1 \text{ for homes in good condition}, \]

\[ \hat{y} = 96.271 + 66.5x_1 \text{ for homes in other conditions}. \]

Both equations have the same slope. The model forces the effect of \( x_1 = \text{ house size} \) on selling price to be the same for both conditions.
Likewise, in a multiple regression model, the slope of the relationship between the population mean of $y$ and each explanatory variable is identical for all values of the other explanatory variables. Such models are sometimes too simple. The effect of an explanatory variable may change considerably as the value of another explanatory variable in the model changes. The multiple regression model we've studied assumes this does not happen. When it does happen, there is interaction.

**Interaction**

For two explanatory variables, interaction exists between them if their effects on the response variable when the slope of the relationship between $\mu_y$ and one of them changes as the value of the other changes.

Suppose the actual population relationship between $x_1 =$ house size and the mean selling price is

$$\mu_y = 100,000 + 50x_1$$

for homes in good condition, and

$$\mu_y = 80,000 + 35x_1$$

for homes in other conditions.

The slope for the effect of $x_1$ differs (50 versus 35) for the two conditions. There is then interaction between house size and condition in their effects on selling price. See Figure 13.8.

How can you allow for interaction when you do a regression analysis? With two explanatory variables, one quantitative and one categorical (as in Example 10), you can fit a separate line with a different slope between the two quantitative variables for each category of the categorical variable.

**Example 11**

**Comparing Winning High Jumps for Men and Women**

**Picture the Scenario**

Men have competed in the high jump in the Olympics since 1896 and women since 1928. Figure 13.9 shows how the winning high jump in the Olympics has changed over time for men and women. The High Jump data file on
the text CD contains the winning heights for each year. A multiple regression analysis of $y =$ winning height (in meters) as a function of $x_1 =$ number of years since 1928 (when women first participated in the high jump) and $x_2 =$ gender ($1 =$ male, $0 =$ female) gives $\hat{y} = 1.63 + 0.0057x_1 + 0.35x_2$.

**Questions to Explore**

a. Interpret the coefficient of year and the coefficient of gender in the equation.

b. To allow interaction, we can fit equations separately to the data for males and the data for females. We then get $\hat{y} = 1.98 + 0.0055x_1$ for males and $\hat{y} = 1.60 + 0.0065x_1$ for females. Describe the interaction by comparing slopes.

c. Describe the interaction allowed in part b by comparing predicted winning high jumps for males and females in 1928 and in 2008.

**Think It Through**

a. For the prediction equation $\hat{y} = 1.63 + 0.0057x_1 + 0.35x_2$, the coefficient of year is 0.0057. For each gender, the predicted winning high jump increases by 0.0057 meters per year. This seems small, but over a hundred years it projects to an increase of $100(0.0057) = 0.57$ meters, about 27 inches. The model does not allow interaction, as it assumes that the slope of 0.0057 is the same for each gender. The coefficient of gender is 0.35. In a given year, the predicted winning high jump for men is 0.35 meters higher than for women. Because this model does not allow interaction, the predicted difference between men and women is the same each year.

b. The slope of 0.0065 for females is higher than the slope of 0.0055 for males. So the predicted winning high jump increases a bit more for females than for males over this time period.

c. In 1928, $x_1 = 0$, and the predicted winning high jump was $\hat{y} = 1.98 + 0.0055(0) = 1.98$ for males and $\hat{y} = 1.60 + 0.0065(0) = 1.60$ for females, a difference of 0.38 meters. In 2008, $x_1 = 2008 - 1928 = 80$ and the predicted winning high jump was $1.98 + \ldots$
0.0055(80) = 2.42 for males and 1.60 + 0.0065(80) = 2.12 for females, a difference of 0.30 meters. The predicted difference between the winning high jumps of males and females decreased a bit between 1928 and 2008.

**Insight**
When we allow interaction, the estimated slope is a bit higher for females than for males. This is what caused the difference in predicted winning high jumps to be less in 2008 than in 1928. However, the slopes were not dramatically different, as Figure 13.9 shows that the points go up at similar rates for the two genders. The sample degree of interaction was not strong.

---

### 13.5 Practicing the Basics

**13.40 Winning high jump** Refer to Example 11 on winning Olympic high jumps. The prediction equation relating \( y = \) winning height (in meters) as a function of 
\[ x_1 = \text{number of years since 1928} \quad \text{and} \quad x_2 = \text{gender} \] 
(1 = male, 0 = female) is \( \hat{y} = 1.63 + 0.0057x_1 + 0.348x_2 \).

**a.** Using this equation, find the prediction equations relating winning height to year, separately for males and for females.

**b.** Find the predicted winning height in 2012 for (i) females, (ii) males, and show how the difference between them relates to a parameter estimate for the model.

**13.41 Mountain bike prices** The Mountain Bike data file on the text CD shows selling prices for mountains bikes. When \( y = \) mountain bike price ($) is regressed on \( x_1 = \) weight of bike (lbs) and
\[ x_2 = \text{the type of suspension} \] 
(0 = full, 1 = front end), \( \hat{y} = 2741.62 - 53.752x_1 - 643.595x_2 \).

**a.** Interpret the estimated effect of the weight of the bike.

**b.** Interpret the estimated effect of the type of suspension on the mountain bike.

**13.42 Predict using house size and condition** For the House Selling Prices OR data set, when we regress \( y = \) selling price (in thousands) on \( x_1 = \) house size and \( x_2 = \) condition
(1 = Good, 0 = Not Good), we get the results shown.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>96.27</td>
<td>13.46</td>
<td>7.15</td>
<td>0.000</td>
</tr>
<tr>
<td>House Size</td>
<td>0.066463</td>
<td>0.004682</td>
<td>14.20</td>
<td>0.000</td>
</tr>
<tr>
<td>Condition</td>
<td>12.93</td>
<td>17.20</td>
<td>0.75</td>
<td>0.453</td>
</tr>
<tr>
<td>S = 81.7874</td>
<td>R-Sq = 50.6%</td>
<td>R-Sq(adj) = 50.1%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**a.** Report the regression equation. Find and interpret the separate lines relating predicted selling price to house size for good condition homes and for homes in not good condition.

**b.** Sketch how selling price varies as a function of house size, for homes in good condition and for homes in not good condition.

**c.** Estimate the difference between the mean selling price of homes in good and in not good condition, controlling for house size.

**13.43 Quality and productivity** The table shows data from 27 automotive plants on \( y = \) number of assembly defects per 100 cars and \( x = \) time (in hours) to assemble each vehicle. The data are in the Quality and Productivity file on the text CD.
Number of defects in assembling 100 cars and time to assemble each vehicle

<table>
<thead>
<tr>
<th>Plant</th>
<th>Defects</th>
<th>Time</th>
<th>Plant</th>
<th>Defects</th>
<th>Time</th>
<th>Plant</th>
<th>Defects</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>39</td>
<td>27</td>
<td>10</td>
<td>89</td>
<td>17</td>
<td>19</td>
<td>69</td>
<td>54</td>
</tr>
<tr>
<td>2</td>
<td>38</td>
<td>23</td>
<td>11</td>
<td>48</td>
<td>20</td>
<td>20</td>
<td>79</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>42</td>
<td>15</td>
<td>12</td>
<td>38</td>
<td>26</td>
<td>21</td>
<td>29</td>
<td>31</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>17</td>
<td>13</td>
<td>68</td>
<td>21</td>
<td>22</td>
<td>84</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>55</td>
<td>12</td>
<td>14</td>
<td>67</td>
<td>26</td>
<td>23</td>
<td>87</td>
<td>44</td>
</tr>
<tr>
<td>6</td>
<td>56</td>
<td>16</td>
<td>15</td>
<td>69</td>
<td>30</td>
<td>24</td>
<td>98</td>
<td>23</td>
</tr>
<tr>
<td>7</td>
<td>57</td>
<td>18</td>
<td>16</td>
<td>69</td>
<td>32</td>
<td>25</td>
<td>100</td>
<td>25</td>
</tr>
<tr>
<td>8</td>
<td>56</td>
<td>26</td>
<td>17</td>
<td>70</td>
<td>31</td>
<td>26</td>
<td>140</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>61</td>
<td>20</td>
<td>18</td>
<td>68</td>
<td>37</td>
<td>27</td>
<td>170</td>
<td>28</td>
</tr>
</tbody>
</table>


a. The prediction equation is \( \hat{y} = 61.3 + 0.35x \). Find the predicted number of defects for a car having assembly time (i) 12 hours (the minimum) and (ii) 54 hours (the maximum).

b. The first 11 plants were Japanese facilities and the rest were not. Let \( x_1 = \) time to assemble vehicle and \( x_2 = \) whether facility is Japanese (1 = yes, 0 = no). The fit of the multiple regression model is

\[
\hat{y} = 105.0 - 0.78x_1 - 36.0x_2
\]

Interpret the coefficients that estimate the effect of the \( x_1 \) and the effect of \( x_2 \).

c. Explain why part a and part b indicate that Simpson’s paradox has occurred.

d. Explain how Simpson’s paradox occurred. To do this, construct a scatterplot between \( y \) and \( x_1 \) in which points are identified by whether the facility is Japanese. Note that the Japanese facilities tended to have low values for both \( x_1 \) and \( x_2 \).

### 13.44 Predicting hamburger sales

A chain restaurant that specializes in selling hamburgers wants to analyze how \( y \) = sales for a customer (the total amount spent by a customer on food and drinks, in dollars) depends on the location of the restaurant, which is classified as inner city, suburbia, or at an interstate exit.

a. Construct indicator variables \( x_1 \) for inner city and \( x_2 \) for suburbia so you can include location in a regression equation for predicting the sales.

b. For part a, suppose \( \hat{y} = 5.8 - 0.7x_1 + 1.2x_2 \). Find the difference between the estimated mean sales in suburbia and at interstate exits.

### 13.45 Houses, size, and garage

Use the House Selling Prices OR data file on the text CD to regress selling price in thousands on house size and whether the house has a garage.

a. Report the prediction equation. Find and interpret the equations predicting selling price using house size, for homes with and without a garage.

b. How do you interpret the coefficient of the indicator variable for whether the home has a garage?

### 13.46 House size and garage interact?

Refer to Example 11 and Exercise 13.40, with \( \hat{y} = \) predicted winning high jump and \( x_1 = \) number of years since 1928. When equations are fitted separately for males and for females, we get \( \hat{y} = 1.98 + 0.0055x_1 \) for males and \( \hat{y} = 1.60 + 0.0065x_1 \) for females.

a. In allowing the lines to have different slopes, the overall model allows for ________ between gender and year in their effects on the winning high jump. (Fill in the correct word.)

b. Show that both equations yield the same predicted winning high jump when \( x_1 = 380 \) (that is, 380 years after 1928, or in 2308).

c. Is it sensible to use this model to predict that in the year 2036 men and women will have about the same winning high jump? Why or why not?

### 13.47 Equal high jump for men and women

Refer to Example 11 and Exercise 13.40. With \( \hat{y} = \) predicted winning high jump and \( x_1 = \) number of years since 1928, or in 2308).

a. Explain why there is actually a substantial degree of interaction.

b. Sketch a hypothetical scatter diagram, showing points identified by garage or no garage, suggesting that there is a substantial degree of interaction.

### 13.48 Comparing sales

You own a gift shop that has a campus location and a shopping mall location. You want to compare the regressions of \( y = \) daily total sales on \( x = \) number of people who enter the shop, for total sales listed by day at the campus location and at the mall location. Explain how you can do this using regression modeling

a. With a single model, having an indicator variable for location, that assumes the slopes are the same for each location.

b. With separate models for each location, permitting the slopes to be different.

### 13.6 Modeling a Categorical Response

The regression models studied so far are designed for a quantitative response variable \( y \). When \( y \) is categorical, a different regression model applies, called **logistic regression**. Logistic regression can model

- A voter’s choice in an election (Democrat or Republican), with explanatory variables of annual income, political ideology, religious affiliation, and race.
Whether a credit card holder pays his or her bill on time (yes or no), with explanatory variables of family income and the number of months in the past year that the customer paid the bill on time.

We’ll study logistic regression in this section for the special case of binary $y$.

**The Logistic Regression Model**

Denote the possible outcomes for $y$ by 0 and 1. We’ll use the generic terms failure and success for these outcomes. The population mean of the 0 and 1 scores equals the population proportion of 1 outcomes (successes) for the response variable. That is, $\mu_y = p$, where $p$ denotes the population proportion of successes. This proportion also represents the probability that a randomly selected subject has a success outcome. The model describes how $p$ depends on the values of the explanatory variables.

For a single explanatory variable, $x$, the straight-line regression model is

$$p = \alpha + \beta x.$$  

As Figure 13.10 shows, this model implies that $p$ falls below 0 or above 1 for sufficiently small or large $x$-values. However, a proportion must fall between 0 and 1. Although the straight-line model may be valid over a restricted range of $x$-values, it is usually inadequate when there are multiple explanatory variables.

Figure 13.10 also shows a more realistic model. It has a curved, S-shape instead of a straight-line trend. The regression equation that best models this S-shaped curve is known as the **logistic regression equation**.

A regression equation for an S-shaped curve for the probability of success $p$ is

$$p = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}.$$  

This equation for $p$ is called the **logistic regression equation**. Logistic regression is used when the response variable has only two possible outcomes (it’s binary.)

Here, $e$ raised to a power represents the exponential function evaluated at that number. Most calculators have an $e^x$ key that provides values of $e$ raised to a power. The model has two parameters, $\alpha$ and $\beta$. Since the numerator of the formula for $p$ is smaller than the denominator, the model forces $p$ to fall between 0 and 1. With a regression model having this S-shape, the probability $p$ falls between 0 and 1 for all possible $x$ values.

Like the slope of a straight line, the parameter $\beta$ in this model refers to whether the mean of $y$ increases or decreases as $x$ increases. When $\beta > 0$, the probability $p$ increases as $x$ increases. When $\beta < 0$, the probability $p$ decreases as $x$ increases. See the margin figure. If $\beta = 0$, $p$ does not change as $x$ changes, so the curve flattens to a horizontal straight line. The steepness of the curve...
increases as the absolute value of \( \beta \) increases. However, unlike in the straight-line model, \( \beta \) is not a slope—the change in the mean per one-unit change in \( x \). For this S-shaped curve, the rate at which the curve climbs or descends changes according to the value of \( x \).

Software can estimate the parameters \( \alpha \) and \( \beta \) in the logistic regression model. Calculators or software can find estimated probabilities \( \hat{p} \) based on the model fit.

**Example 12**

**Travel Credit Cards**

**Picture the Scenario**

An Italian study with 100 randomly selected Italian adults considered factors associated with whether a person has at least one travel credit card. Table 13.12 shows results for the first 15 people on this response variable and on the person’s annual income, in thousands of euros. The complete data set is in the Credit Card and Income data file on the text CD. Let \( x = \) annual income and let \( y = \) whether the person has a travel credit card (1 = yes, 0 = no).

**Table 13.12 Annual Income (in thousands of euros) and Whether Person Has a Travel Credit Card**

The response \( y \) equals 1 if a person has a travel credit card and equals 0 otherwise. The complete data set is on the text CD.

<table>
<thead>
<tr>
<th>Income</th>
<th>( y )</th>
<th>Income</th>
<th>( y )</th>
<th>Income</th>
<th>( y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>0</td>
<td>14</td>
<td>1</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>14</td>
<td>0</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>15</td>
<td>0</td>
</tr>
</tbody>
</table>

*Source: Data from R. Piccarreta, Bocconi University, Milan (personal communication).*

**Questions to Explore**

Table 13.13 shows what software provides for conducting a logistic regression analysis.

**Table 13.13 Results of Logistic Regression for Italian Credit Card Data**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>( Z )</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-3.5180</td>
<td>0.71034</td>
<td>-4.95</td>
<td>0.000</td>
</tr>
<tr>
<td>income</td>
<td>0.1054</td>
<td>0.02616</td>
<td>4.03</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**a.** State the prediction equation for the probability of owning a travel credit card, and explain whether annual income has a positive or a negative effect.

**b.** Find the estimated probability of having a travel credit card at the lowest and highest annual income levels in the sample, which were \( x = 12 \) and \( x = 65 \).

---

\(^2\)The data were originally recorded in Italian lira but have been changed to euros and adjusted for inflation.
Ways to Interpret the Logistic Regression Model

The sign of β in the logistic regression model tells us whether the probability of the outcome increases or decreases as x increases. How else can we interpret the model parameters and their estimates? Here, we’ll describe three ways.

1. To describe the effect of x, you can compare estimates of p at two different values of x. One possibility is to compare p found at the minimum and maximum values of x, as we did in Example 12. There we saw that the estimated probability changed from 0.09 to 0.97, a considerable change. An alternative is to instead use values of x to evaluate this probability that are not as affected by outliers, such as the first and third quartiles.

2. It’s good to know the value of x at which p = 0.50, that is, that value of x for which each outcome is equally likely. This depends on the logistic regression parameters α and β. It can be shown that x = -α/β when p = 0.50. (Exercise 13.92).

3. The simplest way to use the logistic regression parameter β to interpret the steepness of the curve uses a straight-line approximation. Because the logistic regression formula is a curve rather than a straight line, β itself is no longer the ordinary slope. At the x-value where p = 0.50, the line drawn tangent to the logistic regression curve has slope β/4. See Figure 13.11. The value β/4 represents the approximate change in the probability p for a one-unit increase in x, when x is close to the value at which p = 0.50. Tangent lines at other points have weaker slopes (Exercise 13.91), close to 0 when p is near 0 or 1.

Think It Through

a. Substituting the α and β estimates from Table 13.13 into the logistic regression model formula, \( p = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}} \), we get the equation for the estimated probability \( \hat{p} \) of having a travel credit card,

\[
\hat{p} = \frac{e^{-3.52 + 0.105x}}{1 + e^{-3.52 + 0.105x}}.
\]

Because the estimate 0.105 of β (the coefficient of x) is positive, this sample suggests that annual income has a positive effect: The estimated probability of having a travel credit card is higher at higher levels of annual income.

b. For subjects with income x = 12 thousand euros, the estimated probability of having a travel credit card equals

\[
\hat{p} = \frac{e^{-3.52 + 0.105(12)}}{1 + e^{-3.52 + 0.105(12)}} = \frac{e^{-2.26}}{1 + e^{-2.26}} = \frac{0.104}{1.104} = 0.09.
\]

For x = 65, the highest income level in this sample, you can check that the estimated probability equals 0.97.

Insight

There is a strong effect. The estimated probability of having a travel credit card changes from 0.09 to 0.97 (nearly 1.0) as annual income changes over its range.

Using software, we could also fit the straight-line regression model. Its prediction equation is \( \hat{p} = -0.159 + 0.0188x \). However, its \( \hat{p} \) predictions are quite different at the low end and at the high end of the annual income scale. At x = 65, for instance, it provides the prediction \( \hat{p} = 1.06 \). This is a poor prediction because we know that a proportion must fall between 0 and 1.
Chapter 13  Multiple Regression

Figure 13.11 When the probability $p = 0.50$, for a one-unit change in $x$, $p$ changes by about $\beta/4$. Here, $\beta$ is the coefficient of $x$ from the logistic regression model.

Example 13

Effect of Income on Credit Card Use

Picture the Scenario

For the Italian travel credit card study, Example 12 found that the equation

$$\hat{p} = \frac{e^{-3.52 + 0.105x}}{1 + e^{-3.52 + 0.105x}}$$

estimates the probability $p$ of having such a credit card as a function of annual income $x$.

Questions to Explore

Interpret this equation by (a) finding the income value at which the estimated probability of having a credit card equals 0.50 and (b) finding the approximate rate of change in the probability at that income value.

Think It Through

a. The estimate of $\alpha$ is $-3.52$ and the estimate of $\beta$ is $0.105$. Substituting the estimates into the expression $\frac{-\alpha}{\beta}$ for the value of $x$ at $p = 0.50$, we get $x = \frac{3.52}{0.105} = 33.5$. The estimated probability of having a travel credit card equals 0.50 when annual income equals €33,500.

b. A line drawn tangent to the logistic regression curve at the point where $p = 0.50$ has slope equal to $\beta/4$. The estimate of this slope is $0.105/4 = 0.026$. For each increase of €1000 in annual income near the income value of €33,500, the estimated probability of having a travel credit card increases by approximately 0.026.

Insight

Figure 13.12 shows the estimated logistic regression curve, highlighting what we’ve learned in Examples 12 and 13: The estimated probability increases from 0.09 to 0.97 between the minimum and maximum income values, and it equals 0.50 at an annual income of €33,500.
Inference for Logistic Regression

Software also reports a z test statistic for the hypothesis $H_0: \beta = 0$. When $\beta = 0$, the probability of possessing a travel credit card is the same at all income levels. The test statistic equals the ratio of the estimate $\hat{b}$ of $\beta$ divided by its standard error. Sections 8.2 and 9.2 showed that inference about proportions uses z test statistics rather than t test statistics. From Table 13.13 (shown again in the margin), the z test statistic equals

$$z = \frac{(\hat{b} - 0)/se = (0.105 - 0)/0.0262 = 4.0.}$$

This has a P-value of 0.000 for $H_a: \beta \neq 0$. Since the sample slope is positive, for the population of adult Italians, there is strong evidence of a positive association between annual income and having a travel credit card.

The result of this test is no surprise. We would expect people with higher annual incomes to be more likely to have travel credit cards. Some software can construct a confidence interval for the probability $p$ at various $x$ levels. A 95% confidence interval for the probability of having a credit card equals (0.04, 0.19) at the lowest sample income level of €12,000, and (0.78, 0.996) at the highest sample income level of €65,000.

Multiple Logistic Regression

Just as ordinary regression extends to handle several explanatory variables, so does logistic regression. Also, logistic regression can include categorical explanatory variables using indicator variables.

### Table 13.13

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE</th>
<th>Z</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-3.518</td>
<td>0.710</td>
<td>-4.95</td>
<td>0.000</td>
</tr>
<tr>
<td>income</td>
<td>0.105</td>
<td>0.0262</td>
<td>4.03</td>
<td>0.000</td>
</tr>
</tbody>
</table>

### Example 14

Estimating Proportion of Students Who’ve Used Marijuana

**Picture the Scenario**

Table 13.14 is a three-variable contingency table from a Wright State University survey asking senior high-school students near Dayton, Ohio, whether they had ever used alcohol, cigarettes, or marijuana. We’ll treat marijuana use as the response variable and cigarette use and alcohol use as explanatory variables.

### Table 13.14 Alcohol, Cigarette, and Marijuana Use for High School Seniors

<table>
<thead>
<tr>
<th>Alcohol Use</th>
<th>Cigarette Use</th>
<th>Marijuana Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>911</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>44</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>3</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>2</td>
</tr>
</tbody>
</table>

*Source:* Data from Professor Harry Khamis, Wright State University (personal communication).

**Questions to Explore**

Let $y$ indicate marijuana use, coded 1 = yes, 0 = no. Let $x_1$ be an indicator variable for alcohol use (1 = yes, 0 = no), and let $x_2$ be an indicator variable for cigarette use (1 = yes, 0 = no). Table 13.15 shows MINITAB output for a logistic regression model.
Chapter 13  Multiple Regression

**Table 13.15 Minitab Output for Estimating the Probability of Marijuana Use**

**Based on Alcohol Use and Cigarette Use**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>Z</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-5.30904</td>
<td>0.475190</td>
<td>-11.17</td>
<td>0.000</td>
</tr>
<tr>
<td>alcohol</td>
<td>2.98601</td>
<td>0.464671</td>
<td>6.43</td>
<td>0.000</td>
</tr>
<tr>
<td>cigarettes</td>
<td>2.84789</td>
<td>0.163839</td>
<td>17.38</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**a.** Report the prediction equation, and interpret.

**b.** Find the estimated probability \( \hat{p} \) of having used marijuana (i) for students who have not used alcohol or cigarettes and (ii) for students who have used both alcohol and cigarettes.

**Think It Through**

**a.** From Table 13.15, the logistic regression prediction equation is

\[
\hat{p} = \frac{e^{-5.31 + 2.99x_1 + 2.85x_2}}{1 + e^{-5.31 + 2.99x_1 + 2.85x_2}}.
\]

The coefficient of alcohol use \( x_1 \) is positive (2.99). The indicator variable \( x_1 \) equals 1 for those who’ve used alcohol. Thus, the alcohol users have a higher estimated probability of using marijuana, controlling for whether they used cigarettes. Likewise, the coefficient of cigarette use \( x_2 \) is positive (2.85), so the cigarette users have a higher estimated probability of using marijuana, controlling for whether they used alcohol. Table 13.15 tells us that for each predictor, the test statistic is large. In other words, the estimated effect is a large number of standard errors from 0. The P-values are both 0.000, so there is strong evidence that the corresponding population effects are positive also.

**b.** For those who have not used alcohol or cigarettes, \( x_1 = x_2 = 0 \). For them, the estimated probability of marijuana use is

\[
\hat{p} = \frac{e^{-5.31 + 2.99(0) + 2.85(0)}}{1 + e^{-5.31 + 2.99(0) + 2.85(0)}} = \frac{e^{-5.31}}{1 + e^{-5.31}} = 0.0049 = 0.005.
\]

For those who have used alcohol and cigarettes, \( x_1 = x_2 = 1 \). For them, the estimated probability of marijuana use is

\[
\hat{p} = \frac{e^{-5.31 + 2.99(1) + 2.85(1)}}{1 + e^{-5.31 + 2.99(1) + 2.85(1)}} = 0.629.
\]

In summary, the probability that students have tried marijuana seems highly related on whether they’ve used alcohol and cigarettes.

**Insight**

Likewise, you can find the estimated probability of using marijuana for those who have used alcohol but not cigarettes (let \( x_1 = 1 \) and \( x_2 = 0 \)) and for those who have not used alcohol but have used cigarettes. Table 13.16 summarizes results. We see that marijuana use is unlikely unless a student has used both alcohol and cigarettes.
Checking the Logistic Regression Model

How can you check whether a logistic regression model fits the data well? When the explanatory variables are categorical, you can find the sample proportions for the outcome of interest. Table 13.16 also shows these for the marijuana use example. For instance, for those who had used alcohol and cigarettes, from Table 13.14 (shown again in the margin) the sample proportion who had used marijuana is \( \frac{911}{911 + 538} = 0.629 \). Table 13.16 shows that the sample proportions are close to the estimated proportions generated by the model. The model seems to fit well.

Table 13.16 may suggest the question, why bother to fit the model? Why not merely inspect a table of sample proportions? One important reason is that the model enables us to easily conduct inferences about the effects of explanatory variables, while controlling for other variables. For instance, if we want to test the effect of alcohol use on marijuana use, controlling for cigarette use, we can test \( H_0: \beta_1 = 0 \) in the model with alcohol use and cigarette use as predictors.

### Table 13.16 Estimated Probability of Marijuana Use, by Alcohol Use and Cigarette Use, Based on Logistic Regression Model

<table>
<thead>
<tr>
<th>Alcohol Use</th>
<th>Cigarette Use</th>
<th>Marijuana</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>911</td>
<td>538</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>44</td>
<td>456</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>3</td>
<td>43</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>2</td>
<td>279</td>
</tr>
</tbody>
</table>

### Try Exercises 13.56 and 13.57

13.6 Practicing the Basics

#### 13.49 Income and credit cards

Example 12 used logistic regression to estimate the probability of having a travel credit card when \( x = \) annual income (in thousands of euros). At the mean income of €25,000, show that the estimated probability of having a travel credit card equals 0.29.

#### 13.50 Hall of Fame induction

Baseball’s highest honor is election to the Hall of Fame. The history of the election process, however, has been filled with controversy and accusations of favoritism. Most recently, there is also the discussion about players who used performance enhancement drugs. The Hall of Fame has failed to define what the criteria for entry should be. Several statistical models have attempted to describe the probability of a player being elected into the Hall of Fame. How does hitting 400 or 500 home runs affect a player’s chances of being enshrined? What about having a .300 average or 1500 RBIs? One factor, the number of home runs, is examined by using logistic regression as the probability of being elected:

\[ P(\text{HOF}) = \frac{e^{6.7 + 0.0175HR}}{1 + e^{6.7 + 0.0175HR}} \]

13.51 Horseshoe crabs

A study of horseshoe crabs by zoologist Dr. Jane Brockmann at the University of Florida used logistic regression to predict the probability that a female crab had a male partner nesting nearby. One explanatory variable was \( x = \) weight of the female crab (in kilograms). The results were

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-3.695</td>
</tr>
<tr>
<td>Weight</td>
<td>1.815</td>
</tr>
</tbody>
</table>

The quartiles for weight were \( Q_1 = 2.00, Q_2 = 2.35, \) and \( Q_3 = 2.85. \)

- **a.** Find the estimated probability of a male partner at \( Q_1 \) and at \( Q_3. \)
Chapter 13  Multiple Regression

b. Interpret the effect of weight by estimating how much the probability increases over the middle half of the sampled weights, between Q1 and Q3.

13.52 More crabs  Refer to the previous exercise. For what weight values do you estimate that a female crab has probability (a) 0.50, (b) greater than 0.50, and (c) less than 0.50, of having a male partner nesting nearby?

13.53 Voting and income  A logistic regression model describes how the probability of voting for the Republican candidate in a presidential election depends on x, the voter’s total family income (in thousands of dollars) in the previous year. The prediction equation for a particular sample is

\[ \hat{p} = \frac{e^{N.00+0.02x}}{1 + e^{N.00+0.02x}} \]

Find the estimated probability of voting for the Republican candidate when (a) income = $10,000, (b) income = $100,000. Describe how the probability seems to depend on income.

13.54 Equally popular candidates  Refer to the previous exercise.

a. At which income level is the estimated probability of voting for the Republican candidate equal to 0.50?

b. Over what region of income values is the estimated probability of voting for the Republican candidate (i) greater than 0.50 and (ii) less than 0.50?

c. At the income level for which \( \hat{p} = 0.50 \), give a linear approximation for the change in the probability for each $1000 increase in income.

13.55 Many predictors of voting  Refer to the previous two exercises. When the explanatory variables are \( x_1 = \) family income, \( x_2 = \) number of years of education, and \( x_3 = \) gender (1 = male, 0 = female), suppose a logistic regression reports

\begin{align*}
\text{Predictor} & & \text{Coef} & & \text{SE Coef} \\
\text{Constant} & & -1.40 & & 0.12 \\
\text{income} & & 0.02 & & 0.01 \\
\text{education} & & 0.08 & & 0.05 \\
\text{gender} & & 0.20 & & 0.06 \\
\end{align*}

For this sample, \( x_1 \) ranges from 6 to 157 with a standard deviation of 25, and \( x_2 \) ranges from 7 to 20 with a standard deviation of 3.

a. Interpret the effects using the sign of the coefficient for each predictor.

b. Illustrate the gender effect by finding and comparing the estimated probability of voting Republican for (i) a man with 16 years of education and income $40,000 and (ii) a woman with 16 years of education and income $40,000.

13.56 Graduation, gender, and race  The U.S. Census Bureau lists college graduation numbers by race and gender. The table shows the data for graduating 25-year-olds.

<table>
<thead>
<tr>
<th>College graduation</th>
<th>Sample Size</th>
<th>Graduates</th>
</tr>
</thead>
<tbody>
<tr>
<td>White females</td>
<td>31,249</td>
<td>10,781</td>
</tr>
<tr>
<td>White males</td>
<td>39,583</td>
<td>10,727</td>
</tr>
<tr>
<td>Black females</td>
<td>13,194</td>
<td>2,309</td>
</tr>
<tr>
<td>Black males</td>
<td>17,707</td>
<td>2,054</td>
</tr>
</tbody>
</table>


a. Identify the response variable.

b. Express the data in the form of a three-variable contingency table that cross-classifies whether graduated (yes, no), race, and gender.

c. When we use indicator variables for race (1 = white, 0 = black) and for gender (1 = female, 0 = male), the coefficients of those predictors in the logistic regression model are 0.975 for race and 0.375 for gender. Based on these estimates, which race and gender combination has the highest estimated probability of graduation? Why?

13.57 Death penalty and race  The three-dimensional contingency table shown is from a study of the effects of racial characteristics on whether or not individuals convicted of homicide receive the death penalty. The subjects classified were defendants in indictments involving cases with multiple murders in Florida between 1976 and 1987.

<table>
<thead>
<tr>
<th>Death penalty verdict by defendant's race and victims' race</th>
<th>White</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defendant's Race</td>
<td>Victims' Race</td>
<td>Death Penalty</td>
</tr>
<tr>
<td>-----------</td>
<td>--------------</td>
<td>----------------</td>
</tr>
<tr>
<td>White</td>
<td>White</td>
<td>Yes  53</td>
</tr>
<tr>
<td>Black</td>
<td>White</td>
<td>Yes  11</td>
</tr>
<tr>
<td>White</td>
<td>Black</td>
<td>Yes  4</td>
</tr>
</tbody>
</table>


a. Based on the percentages shown, controlling for victims’ race, for which defendant’s race was the death penalty more likely?

b. Let \( y = \) death penalty verdict (1 = yes, 0 = no), let \( d \) be an indicator variable for defendant’s race (1 = white, 0 = black), and let \( v \) be an indicator variable for victims’ race (1 = white, 0 = black). The logistic regression prediction equation is

\[ \hat{p} = \frac{e^{-3.596 - 0.868d + 2.404v}}{1 + e^{-3.596 - 0.868d + 2.404v}} \]

According to this equation, for which of the four groups is the death penalty most likely? Explain your answer.

13.58 Death penalty probabilities  Refer to the previous exercise.

a. Based on the prediction equation, when the defendant is black and the victims were white, show that the estimated death penalty probability is 0.233.

b. The model-estimated probabilities are 0.011 when the defendant is white and victims were black, 0.113 when the defendant and the victims were white, and 0.027 when the defendant and the victims were black. Construct a table cross-classifying defendant’s race by victims’ race and show the estimated probability of the death penalty for each cell. Use this to interpret the effect of defendant’s race.

c. Collapse the contingency table over victims’ race, and show that (ignoring victims’ race) white defendants were more likely than black defendants to get the death penalty. Comparing this with what happens when you control for victims’ race, explain how Simpson’s paradox occurs.
Chapter Review

ANSWERS TO THE CHAPTER FIGURE QUESTIONS

Figure 13.1 Selling price is the response variable, and the graphs that show it as the response variable are in the first row.

Figure 13.2 The coefficient (15,170) of \( x_1 \) = number of bedrooms is positive, so increasing it has the effect of increasing the predicted selling price (for fixed house size).

Figure 13.3 We use only the right tail because larger \( F \) values provide greater evidence against \( H_0 \).

Figure 13.4 A histogram that shows a few observations that have extremely large or extremely small standardized residuals, well removed from the others.

Figure 13.5 The pattern suggests that \( y \) tends to be below \( \hat{y} \) for very small and large \( x_1 \)-values and above \( \hat{y} \) for medium-sized \( x_1 \)-values. The effect of \( x_1 \) appears to be better modeled by a mathematical function that has a parabolic shape.

Figure 13.6 As house size increases, the variability of the standardized residuals appears to increase, suggesting more variability in selling prices when house size is larger, for a given number of bedrooms. The number of bedrooms affect the analysis because they have an effect on the predicted values for the model, on which the residuals are based.

Figure 13.7 Each line has the same slope.

Figure 13.10 A straight line implies that \( p \) falls below 0 or above 1 for sufficiently small or large \( x \)-values. This is a problem, because a proportion must fall between 0 and 1.

CHAPTER SUMMARY

This chapter generalized regression to include more than one explanatory variable in the model. The multiple regression model relates the mean \( \mu_y \) of a response variable \( y \) to several explanatory variables, for instance,

\[
\mu_y = \alpha + \beta_1 x_1 + \beta_2 x_2
\]

for two predictors. Advantages over bivariate regression include better predictions of \( y \) and being able to study the effect of an explanatory variable on \( y \) while controlling (keeping fixed) values of other explanatory variables in the model.

- The regression parameters (the betas) are slopes that represent effects of the explanatory variables while controlling for the other variables in the model. For instance, \( \beta_1 \) represents the change in the mean of \( y \) for a one-unit increase in \( x_1 \), at fixed values of the other explanatory variables.

- The multiple correlation \( R \) and its square \( (R^2) \) describe the predictability of the response variable \( y \) by the set of explanatory variables. The multiple correlation \( R \) equals the correlation between the observed and predicted \( y \) values. Its square, \( R^2 \), is the proportional reduction in error from predicting \( y \) using the prediction equation instead of using \( \bar{y} \) (and ignoring \( x \)). Both \( R \) and \( R^2 \) fall between 0 and 1, with larger values representing stronger association.

- An \( F \) statistic tests \( H_0: \beta_1 = \beta_2 = \ldots = 0 \), which states that \( y \) is independent of all the explanatory variables in the model. The \( F \) test statistic equals a ratio of mean squares. A small \( P \)-value suggests that at least one explanatory variable affects the response.

- Individual \( t \) tests and confidence intervals for each \( \beta \) parameter analyze separate population effects of each explanatory variable, controlling for the other variables in the model.

- Categorical explanatory variables can be included in a regression model using indicator variables. With two categories, the indicator variable equals 1 when the observation is in the first category and 0 when it is in the second.

- For binary response variables, the logistic regression model describes how the probability of a particular category depends on the values of explanatory variables. An S-shaped curve describes how the probability changes as the value of a quantitative explanatory variable increases.

Table 13.17 summarizes the basic properties and inference methods for multiple regression and those that Chapter 12 introduced for bivariate regression, with quantitative variables.

<table>
<thead>
<tr>
<th>Table 13.17 Summary of Bivariate and Multiple Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bivariate Regression</strong> Model formula ( \mu_y = \alpha + \beta x )</td>
</tr>
<tr>
<td><strong>Simultaneous Effect</strong></td>
</tr>
<tr>
<td>Properties of measures</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>( r^2 ) is proportional reduction in error (PRE)</td>
</tr>
<tr>
<td>Hypotheses of no effect</td>
</tr>
<tr>
<td>Test statistic</td>
</tr>
</tbody>
</table>
Chapter 13  Multiple Regression

SUMMARY OF NOTATION

\( x_1, x_2, \ldots = \text{explanatory variables in multiple regression model} \)
\( R = \text{multiple correlation} = \text{correlation between observed} \ y \ \text{and predicted} \ y \ \text{values} \)
\( R^2 = \text{proportional reduction in prediction error, for multiple regression} \)
\( F = \text{test statistic for testing that all} \ \beta \ \text{parameters} = 0 \ \text{in regression model} \)

\( e^{\alpha + \beta x} = \text{exponential function used in numerator and denominator of logistic regression equation, which models the probability} \ p \ \text{of a binary outcome by} \)
\[ p = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}. \]

CHAPTER PROBLEMS

Practicing the Basics

13.59 House prices  This chapter has considered many aspects of regression analysis. Let’s consider several of them at once by using software with the House Selling Prices OR data file on the text CD to conduct a multiple regression analysis of \( y = \text{selling price of home,} x_1 = \text{size of home,} \ x_2 = \text{number of bedrooms,} \ x_3 = \text{number of bathrooms.} \)

a. Construct a scatterplot matrix. Identify the plots that pertain to selling price as a response variable. Interpret, and explain how the highly discrete nature of \( x_2 \) and \( x_3 \) affects the plots.

b. Fit the model. Write down the prediction equation, and interpret the coefficient of size of home by its effect when \( x_2 \) and \( x_3 \) are fixed.

c. Show how to calculate \( R^2 \) from SS values in the ANOVA table. Interpret its value in the context of these variables.

d. Find and interpret the multiple correlation.

e. Show all steps of the \( F \) test that selling price is independent of these predictors. Explain how to obtain the \( F \) statistic from the mean squares in the ANOVA table.

f. Report the \( t \) statistic for testing \( H_0: \beta_2 = 0 \). Report the \( P \)-value for \( H_0: \beta_2 < 0 \), and interpret. Why do you think this effect is not significant? Does this imply that the number of bedrooms is not associated with selling price?

g. Construct and examine the histogram of the residuals for the multiple regression model. What does this describe, and what does it suggest?

h. Construct and examine the plot of the residuals plotted against size of home. What does this describe, and what does it suggest?

13.60 Predicting body strength  In Chapter 12, we analyzed strength data for a sample of female high school athletes. When we predict the maximum number of pounds the athlete can bench press using the number of times she can do a 60-pound bench press (BP_60), we get \( r^2 = 0.643 \). When we add the number of times an athlete can perform a 200-pound leg press (LP_200) to the model, we get \( \hat{y} = 60.6 + 1.33(\text{BP}_60) + 0.21(\text{LP}_200) \) and \( R^2 = 0.656 \).

a. Find the predicted value and residual for an athlete who has \( \text{BP} = 85, \text{BP}_60 = 10, \) and \( \text{LP}_200 = 20. \)

b. Find the prediction equation for athletes who have \( \text{LP}_200 = 20, \) and explain how to interpret the slope for \( \text{BP}_60. \)

c. Note that \( R^2 = 0.656 \) for the multiple regression model is not much larger than \( r^2 = 0.643 \) for the bivariate model with \( \text{LP}_200 \) as the only explanatory variable. What does this suggest?

13.61 Softball data  Refer to the Softball data set on the text CD. Regress the difference (DIFF) between the number of runs scored by that team and by the other team on the number of hits (HIT) and the number of errors (ERR).

a. Report the prediction equation, and interpret the slopes.

b. From part a, approximately how many hits does the team need so that the predicted value of DIFF is positive (corresponding to a predicted win), if they can play error-free ball (ERR = 0)?

13.62 Violent crime  A MINITAB printout is provided from fitting the multiple regression model to U.S. crime data for the 50 states (excluding Washington, D.C.), on \( y = \text{violent crime rate,} x_1 = \text{poverty rate,} \) and \( x_2 = \text{percent living in urban areas.} \)

a. Predict the violent crime rate for Massachusetts, which has violent crime rate = 476, poverty rate = 10.2%, and urbanization = 92.1%. Find the residual, and interpret.

b. Interpret the effects of the predictors by showing the prediction equation relating \( y \) and \( x_1 \) for states with (i) \( x_2 = 0 \) and (ii) \( x_2 = 100. \) Interpret.

Regression of violent crime rate on poverty rate and urbanization

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-270.7</td>
<td>121.1</td>
<td>-2.23</td>
<td>0.030</td>
</tr>
<tr>
<td>poverty</td>
<td>28.334</td>
<td>7.249</td>
<td>3.91</td>
<td>0.000</td>
</tr>
<tr>
<td>urbanization</td>
<td>5.416</td>
<td>1.035</td>
<td>5.23</td>
<td>0.000</td>
</tr>
</tbody>
</table>

13.63 Effect of poverty on crime  Refer to the previous exercise. Now we add \( x_3 = \text{percentage of single-parent families} \) to the model. The SPSS table on the next page shows results. Without \( x_2 \) in the model, poverty has slope 28.33, and when \( x_2 \) is added, poverty has slope 14.95. Explain the differences in the interpretations of these two slopes.
Violent crime predicted by poverty rate, urbanization, and single-parent family

<table>
<thead>
<tr>
<th>Model</th>
<th>Coefficients*</th>
<th>Unstandardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Constant)</td>
<td>-631.700</td>
<td>149.604</td>
<td>-4.222</td>
<td>.000</td>
</tr>
<tr>
<td>poverty</td>
<td>14.953</td>
<td>7.540</td>
<td>1.983</td>
<td>.053</td>
</tr>
<tr>
<td>urbanization</td>
<td>4.406</td>
<td>.973</td>
<td>4.528</td>
<td>.000</td>
</tr>
<tr>
<td>singleparent</td>
<td>25.362</td>
<td>7.220</td>
<td>3.513</td>
<td>.001</td>
</tr>
</tbody>
</table>

* Dependent Variable: violent crime rate.

13.64 Modeling fertility For the World Data for Fertility and Literacy data file on the text CD, a MINITAB printout follows that shows fitting a multiple regression model for y = fertility, x₁ = adult literacy rate (both sexes), x₂ = combined educational enrollment (both sexes). Report the value of each of the following:

a. r between y and x₁
b. R²
c. Total sum of squares
d. Residual sum of squares
e. Standard deviation of y
f. Residual standard deviation of y
g. Test statistic value for H₀: β₁ = 0
h. P-value for H₀: β₁ = β₂ = 0

Analysis of fertility, literacy, and combined educational enrollment:

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>SE Mean</th>
<th>StDev</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adolescent fertility</td>
<td>142</td>
<td>40.42</td>
<td>3.73</td>
<td>44.50</td>
<td>49.55</td>
</tr>
<tr>
<td>Adult literacy rate</td>
<td>142</td>
<td>80.80</td>
<td>1.64</td>
<td>19.56</td>
<td>88.40</td>
</tr>
<tr>
<td>Combined educ enrol</td>
<td>142</td>
<td>68.59</td>
<td>1.32</td>
<td>15.79</td>
<td>71.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>187.95</td>
<td>12.92</td>
<td>14.55</td>
<td>0.000</td>
</tr>
<tr>
<td>Adult literacy rate</td>
<td>-1.2590</td>
<td>0.2416</td>
<td>-5.21</td>
<td>0.000</td>
</tr>
<tr>
<td>Combined educ enrol</td>
<td>-0.3372</td>
<td>0.2994</td>
<td>-1.26</td>
<td>0.211</td>
</tr>
</tbody>
</table>

S = 33.4554 R-Sq = 44.3%

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>123582</td>
<td>61791</td>
<td>55.21</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>139</td>
<td>155577</td>
<td>1119</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>141</td>
<td>279160</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Correlations:

<table>
<thead>
<tr>
<th>Fertility</th>
<th>Adult literacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult literacy</td>
<td>-0.661</td>
</tr>
<tr>
<td>Combined educ enroll</td>
<td>-0.578</td>
</tr>
</tbody>
</table>

13.65 Significant fertility prediction? Refer to the previous exercise.

a. Show how to construct the F statistic for testing H₀: β₁ = β₂ = 0 from the reported mean squares, report its P-value, and interpret.

b. If these are the only nations of interest to us for this study, rather than a random sample of such nations, is this significance test relevant? Explain.

13.66 Motivation to study medicine What motivates someone to pursue the study of medicine? A student’s interest and motivation to study medicine can depend on the strength of motivation and career-related values and approaches to learning. Validated and reliable questionnaires were used to obtain data from 116 first-year medical students. This study found no differences in strength of motivation based on sex, nationality, or age. It did find that the motivation to enter medical school was based on interpersonal factors such as wanting to help people, being respected and successful, and fulfilling a sense of achievement. Students motivated to perform better were driven either by positive self-esteem or by perceiving medicine as a means of enhancing their social status. Students who set their own goals and work toward these goals are more likely to succeed. A regression analysis for predicting motivation score is shown below.

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Std Error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>44.677</td>
<td>6.749</td>
</tr>
<tr>
<td>Strategic approach to learning</td>
<td>0.252</td>
<td>0.065</td>
</tr>
<tr>
<td>Avoidance of role strain</td>
<td>-0.340</td>
<td>0.135</td>
</tr>
<tr>
<td>Prestige</td>
<td>0.451</td>
<td>0.117</td>
</tr>
<tr>
<td>Lack of purpose</td>
<td>-0.755</td>
<td>0.230</td>
</tr>
<tr>
<td>Income</td>
<td>-0.243</td>
<td>0.114</td>
</tr>
</tbody>
</table>

a. Find the equation relating predicted motivation score to the explanatory factors listed.
b. Interpret the signs of the coefficients in the equation for each of the explanatory variables as the explanatory variable relates to motivation.

13.67 Education and gender in modeling income Consider the relationship between y = annual income (in thousands of dollars) and x₁ = number of years of education, by x₂ = gender. Many studies in the United States have found that the slope for a regression equation relating y to x₁ is larger for men than for women. Suppose that in the population, the regression equations are μ₁ = -10 + 4x₁ for men and μ₂ = -5 + 2x₁ for women. Explain why these equations imply that there is interaction between education and gender in their effects on income.

13.68 Horseshoe crabs and width A study of horseshoe crabs found a logistic regression equation for predicting the probability that a female crab had a male partner nesting nearby using x = width of the carapace shell of the female crab (in centimeters). The results were

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-12.351</td>
<td></td>
</tr>
<tr>
<td>Width</td>
<td>0.497</td>
<td></td>
</tr>
</tbody>
</table>

a. For width, Q₁ = 24.9 and Q₃ = 27.7. Find the estimated probability of a male partner at Q₁ and at Q₃. Interpret the effect of width by estimating the increase in the probability over the middle half of the sampled widths.
b. At which carapace shell width level is the estimated probability of a male partner (i) equal to 0.50, (ii) greater than 0.50, and (iii) less than 0.50?
13.69 AIDS and AZT  In a study (reported in New York Times, February 15, 1991) on the effects of AZT in slowing the development of AIDS symptoms, 338 veterans whose immune systems were beginning to falter after infection with the AIDS virus were randomly assigned either to receive AZT immediately or to wait until their T cells showed severe immune weakness. The study classified the veterans’ race, whether they received AZT immediately, and whether they developed AIDS symptoms during the three-year study. Let \( x_1 \) denote whether used AZT (1 = yes, 0 = no) and let \( x_2 \) denote race (1 = white, 0 = black). A logistic regression analysis for the probability of developing AIDS symptoms gave the prediction equation
\[
\hat{p} = \frac{e^{-1.074 - 0.720x_1 + 0.056x_2}}{1 + e^{-1.074 - 0.720x_1 + 0.056x_2}}.
\]

a. Interpret the sign of the effect of AZT.

b. Show how to interpret the AZT effect for a particular race by comparing the estimated probability of AIDS symptoms for black veterans with and without immediate AZT use.

c. The \( se \) value was 0.279 for the AZT use effect. Does AZT use have a significant effect, at the 0.05 significance level? Show all steps of a test to justify your answer.

13.70 Factors affecting first home purchase  The table summarizes results of a logistic regression model for predictions about first home purchase by young married households. The response variable is whether the subject owns a home (1 = yes, 0 = no). The explanatory variables are husband’s income, wife’s income (each in ten-thousands of dollars), the number of years the respondent has been married, the number of children aged 0–17 in the household, and an indicator variable that equals 1 if the subject’s parents owned a home in the last year the subject lived in the parental home.

a. Explain why, other things being fixed, the probability of home ownership increases with husband’s earnings, wife’s earnings, the number of children, and parents’ home ownership.

b. From the table, explain why the number of years married seems to show little evidence of an effect, given the other variables in the model.

### Results of logistic regression for probability of home ownership

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Husband earnings</td>
<td>0.569</td>
<td>0.088</td>
</tr>
<tr>
<td>Wife earnings</td>
<td>0.306</td>
<td>0.140</td>
</tr>
<tr>
<td>No. years married</td>
<td>-0.039</td>
<td>0.042</td>
</tr>
<tr>
<td>No. children</td>
<td>0.220</td>
<td>0.101</td>
</tr>
<tr>
<td>Parents’ home ownership</td>
<td>0.387</td>
<td>0.176</td>
</tr>
</tbody>
</table>


13.71 Student data  Refer to the FL Student Survey data file on the text CD. Using software, conduct a regression analysis using \( y = \) college GPA and predictors high school GPA and sports (number of weekly hours of physical exercise).

Prepare a report, summarizing your graphical analyses, bivariate models and interpretations, multiple regression models and interpretations, inferences, checks of effects of outliers, and overall summary of the relationships.

13.72 Why regression?  In 100–200 words, explain to someone who has never studied statistics the purpose of multiple regression and when you would use it to analyze a data set or investigate an issue. Give an example of at least one application of multiple regression. Describe how multiple regression can be useful in analyzing complex relationships.

13.73 Modeling salaries  The table shows results of fitting a regression model to data on Oklahoma State University salaries (in dollars) of 675 full-time college professors of different disciplines with at least two years of instructional employment. All of the predictors are categorical (binary), except for years as professor, merit ranking, and market influence. The market factor represents the ratio of the average salary at comparable institutions for the corresponding academic field and rank to the actual salary at OSU. Prepare a summary of the results in a couple of paragraphs, interpreting the effects of the predictors. The levels of ranking for professors are assistant, associate, and full professor from low to high. An instructor ranking is nontenure track. Gender and race predictors were not significant in this study.

### Modeling professor salaries

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>17.870</td>
<td>0.272</td>
</tr>
<tr>
<td>Nonenure track</td>
<td>-0.010</td>
<td>0.011</td>
</tr>
<tr>
<td>Instructor</td>
<td>-0.284</td>
<td>0.018</td>
</tr>
<tr>
<td>Associate professor</td>
<td>0.170</td>
<td>0.013</td>
</tr>
<tr>
<td>Full professor</td>
<td>0.407</td>
<td>0.018</td>
</tr>
<tr>
<td>Years as professor</td>
<td>0.004</td>
<td>0.001</td>
</tr>
<tr>
<td>Average merit rating</td>
<td>0.044</td>
<td>0.005</td>
</tr>
<tr>
<td>Business</td>
<td>0.395</td>
<td>0.015</td>
</tr>
<tr>
<td>Education</td>
<td>0.053</td>
<td>0.015</td>
</tr>
<tr>
<td>Engineering</td>
<td>0.241</td>
<td>0.014</td>
</tr>
<tr>
<td>Fine arts</td>
<td>0.000</td>
<td>0.018</td>
</tr>
<tr>
<td>Social science</td>
<td>0.077</td>
<td>0.013</td>
</tr>
<tr>
<td>Market influence</td>
<td>-7.046</td>
<td>0.268</td>
</tr>
</tbody>
</table>

Note: Dependent variable is the logarithm of annual salary.

Model summary: Adjusted \( R^2 = 0.94; F\)-ratio = 411.76; \( N = 675; P\)-value = 0.001.


13.74 Multiple choice: Interpret parameter  If \( \hat{y} = 2 + 3x_1 + 5x_2 - 8x_3 \), then controlling for \( x_2 \) and \( x_3 \), the change in the estimated mean of \( y \) when \( x_1 \) is increased from 10 to 20

a. equals 30.

b. equals 0.3.

c. Cannot be given—depends on specific values of \( x_2 \) and \( x_3 \).

d. Must be the same as when we ignore \( x_2 \) and \( x_3 \).

13.75 Multiple choice: Interpret indicator  In the model
\[
\mu = \alpha + \beta_1x_1 + \beta_2x_2,
\]
suppose that \( x_2 \) is an indicator variable for gender, equaling 1 for females and 0 for males.

a. We set \( x_2 = 0 \) if we want a predicted mean without knowing gender.
b. The slope effect of $x_1$ is $\beta_1$ for males and $\beta_2$ for females.
c. $\beta_2$ is the difference between the population mean of $y$ for females and for males.
d. $\beta_2$ is the difference between the population mean of $y$ for females and males, for all those subjects having $x_1$ fixed, such as $x_1 = 10$.

13.76 Multiple choice: Regression effects
Multiple regression is used to model $y = \text{annual income}$ using $x_1 = \text{number of years of education}$ and $x_2 = \text{number of years employed in current job}$.

a. It is possible that the coefficient of $x_1$ is positive in a bivariate regression but negative in multiple regression.
b. It is possible that the correlation between $y$ and $x_1$ is 0.30 and the multiple correlation between $y$ and $x_1$ and $x_2$ is 0.26.
c. If the $F$ statistic for $H_0: \beta_1 = \beta_2 = 0$ has a P-value = 0.001, then we can conclude that both predictors have an effect on annual income.
d. If $\beta_2 = 0$, then annual income is independent of $x_2$ in bivariate regression.

13.77 True or false: $R$ and $R^2$
For each of the following statements, indicate whether it is true or false. If false, explain why it is false.

a. The multiple correlation is always the same as the ordinary correlation computed between the values of the response variable and the values $\hat{y}$ predicted by the regression model.
b. The multiple correlation is like the ordinary correlation in that it falls between $-1$ and 1.
c. $R^2$ describes how well you can predict $y$ using $x_1$ when you control for the other variables in the multiple regression model.
d. It’s impossible for $R^2$ to go down when you add explanatory variables to a regression model.

13.78 True or false: Regression
For each of the following statements, indicate whether it is true or false. If false, explain why it is false. In regression analysis:

a. The estimated coefficient of $x_1$ can be positive in the bivariate model but negative in a multiple regression model.
b. When a model is refitted after $y = \text{income}$ is changed from dollars to euros, $R^2$, the correlation between $y$ and $x_1$, the $F$ statistics and the $t$ statistics will not change.
c. If $r^2 = 0.6$ between $y$ and $x_1$ and if $r^2 = 0.6$ between $y$ and $x_2$, then for the multiple regression model with both predictors $R^2 = 1.2$.
d. The multiple correlation between $y$ and $\hat{y}$ can equal $-0.40$.

13.79 True or false: Slopes
For data on $y = \text{college GPA}$, $x_1 = \text{high school GPA}$, and $x_2 = \text{average of mathematics and verbal entrance exam score}$, we get $\hat{y} = 2.70 + 0.45x_1$ for bivariate regression and $\hat{y} = 0.3 + 0.40x_1 + 0.003x_2$ for multiple regression. For each of the following statements, indicate whether it is true or false. Give a reason for your answer.

Chapter Problems 677

a. The correlation between $y$ and $x_1$ is positive.
b. A one-unit increase in $x_1$ corresponds to a change of 0.45 in the predicted value of $y$, controlling for $x_2$.
c. Controlling for $x_1$, a 100-unit increase in $x_2$ corresponds to a predicted increase of 0.30 in college GPA.

13.80 Scores for religion
You want to include religious affiliation as a predictor in a regression model, using the categories Protestant, Catholic, Jewish, Other. You set up a variable $x_1$ that equals 1 for Protestants, 2 for Catholics, 3 for Jewish, and 4 for Other, using the model $\mu_i = \alpha + \beta x_1$. Explain why this is inappropriate.

13.81 Lurking variable
Give an example of three variables for which you expect $\beta \neq 0$ in the model $\mu_i = \alpha + \beta x_1$ but $\beta_i = 0$ in the model $\mu_i = \alpha + \beta_1 x_1 + \beta_2 x_2$. (Hint: The bivariate effect of $x_1$ could be completely due to a lurking variable, $x_2$.)

13.82 Properties of $R^2$
Using its definition in terms of SS values, explain why $R^2 = 1$ only when all the residuals are 0, and $R^2 = 0$ when each $\hat{y} = \bar{y}$. Explain what this means in practical terms.

13.83 Why an $F$ test?
When a model has a very large number of predictors, even when none of them truly have an effect in the population, one or two may look significant in $t$ tests merely by random variation. Explain why performing the $F$ test first can safeguard against getting such false information from $t$ tests.

13.84 Multicollinearity
For the high school female athletes data file, regress the maximum bench press on weight and percent body fat.

a. Show that the $F$ test is statistically significant at the 0.05 significance level.
b. Show that the P-values are both larger than 0.35 for testing the individual effects with $t$ tests. (It seems like a contradiction when the $F$ test tells us that at least one predictor has an effect but the $t$ tests indicate that neither predictor has a significant effect. This can happen when the predictor variables are highly correlated, so a predictor has little impact when the other predictors are in the model. Such a condition is referred to as multicollinearity. In this example, the correlation is 0.871 between weight and percent body fat.)

13.85 Logistic versus linear
For binary response variables, one reason that logistic regression is usually preferred over straight-line regression is that a fixed change in $x$ often has a smaller impact on a probability $p$ when $p$ is near 0 or near 1 than when $p$ is near the middle of its range. Let $y$ refer to the decision to rent or to buy a home, with $p = \text{the probability of buying}$, and let $x = \text{weekly family income}$. In which case do you think an increase of $\$100$ in $x$ has greater effect: when $x = 50,000$ (for which $p$ is near 1), when $x = 0$ (for which $p$ is near 0), or when $x = 500$? Explain how your answer relates to the choice of a linear versus logistic regression model.
678  Chapter 13  Multiple Regression

13.86  Adjusted \( R^2 \)  When we use \( R^2 \) for a random sample to estimate a population \( R^2 \), it’s a bit biased. It tends to be a bit too large, especially when \( n \) is small. Some software also reports

\[
\text{Adjusted } R^2 = R^2 - \frac{p}{n - (p + 1)}(1 - R^2),
\]

where \( p \) = number of predictor variables in the model. This is slightly smaller than \( R^2 \) and is less biased. Suppose \( R^2 = 0.500 \) for a model with \( p = 2 \) predictors. Calculate adjusted \( R^2 \) for the following sample sizes: 10, 100, 1000. Show that the difference between adjusted \( R^2 \) and \( R^2 \) diminishes as \( n \) increases.

13.87  \( R \) can’t go down  The least squares prediction equation provides predicted values \( \hat{y} \) with the strongest possible correlation with \( y \), out of all possible prediction equations of that form. Based on this property, explain why the multiple correlation \( R \) cannot decrease when you add a variable to a multiple regression model. (Hint: The prediction equation for the simpler model is a special case of a prediction equation for the full model that has coefficient 0 for the added variable.)

13.88  Indicator for comparing two groups  Chapter 10 presented methods for comparing means for two groups. Explain how it’s possible to perform a significance test of equality of two population means as a special case of a regression analysis. (Hint: The regression model then has a single explanatory variable—an indicator variable for the two groups being compared. What does \( \mu_1 = \mu_2 \) correspond to in terms of a value of a parameter in this model?)

13.89  Simpson’s paradox  Let \( y = \) death rate and \( x = \) average age of residents, measured for each county in Louisiana and in Florida. Draw a hypothetical scatterplot, identifying points for each state, such that the mean death rate is higher in Florida than in Louisiana when \( x \) is ignored, but lower when it is controlled. (Hint: When you fit a line for each state, the line should be higher for Louisiana, but the \( y \)-values for Florida should have an overall higher mean.)

13.90  Parabolic regression  A regression formula that gives a parabolic shape instead of a straight line for the relationship between two variables is

\[
\mu_i = \alpha + \beta_1 x_i + \beta_2 x_i^2.
\]

a. Explain why this is a multiple regression model, with \( x \) playing the role of \( x_1 \) and \( x^2 \) (the square of \( x \)) playing the role of \( x_2 \).

b. For \( x \) between 0 and 5, sketch the prediction equation (i) \( \hat{y} = 10 + 2x + 0.5x^2 \) and (ii) \( \hat{y} = 10 + 2x - 0.5x^2 \). This shows how the parabola is bowl-shaped or mound-shaped, depending on whether the coefficient \( x^2 \) is positive or negative.

13.91  Logistic slope  At the \( x \) value where the probability of success is some value \( p \), the line drawn tangent to the logistic regression curve has slope \( \beta p(1 - p) \).

a. Explain why the slope is \( \beta/4 \) when \( p = 0.5 \).

b. Show that the slope is weaker at other \( p \) values by evaluating this at \( p = 0.1, 0.3, 0.7, \) and 0.9. What does the slope approach as \( p \) gets closer and closer to 0 or 1? Sketch a curve to illustrate.

13.92  When is \( p = 0.50 ? \)  When \( \alpha + \beta x = 0 \), so that

\[
x = -\alpha/\beta, \text{ show that the logistic regression equation } p = e^{\alpha + \beta x}/(1 + e^{\alpha + \beta x}) \text{ gives } p = 0.50.
\]

Student Activities

13.93  Class data  Refer to the data file your class created in Activity 3 in Chapter 1. For variables chosen by your instructor, fit a multiple regression model and conduct descriptive and inferential statistical analyses. Interpret and summarize your findings, and prepare to discuss these in class.