In the section 14.2 in the textbook, we discussed estimating the coefficients $\alpha, \beta_1, \ldots, \beta_k$ in the model $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$ (using the principle of least squares) and then showed how the usefulness of the model could be confirmed by application of the F test for model utility. If $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ cannot be rejected at a reasonably small level of significance, it must be concluded that the model does not specify a useful relationship between $y$ and any of the predictor variables $x_1, x_2, \ldots, x_k$. The investigator must then search further for a model that does describe a useful relationship, perhaps by introducing different predictors or making variable transformations. Only if $H_0$ can be rejected is it appropriate to proceed further with the chosen model and make inferences based on the estimated coefficients $\alpha, b_1, \ldots, b_k$ and the estimated regression function $\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$. In this section, we will consider two different types of inferential problems. One type involves drawing a conclusion about an individual regression coefficient $\beta_i$—either computing a confidence interval for $\beta_i$ or testing a hypothesis concerning $\beta_i$. The second type of problem involves first fixing values of $x_1, x_2, \ldots, x_k$ and then computing either a point estimate or a confidence interval for the corresponding mean $y$ value or predicting a future $y$ value with a point prediction or a prediction interval.

## Inferences About a Regression Coefficient

A confidence interval for the slope coefficient $\beta$ and a hypothesis test about $\beta$ in a simple linear regression were based on facts about the sampling distribution of the statistic $b$ used to obtain a point estimate of $\beta$. Similarly, in multiple regression, procedures for making inferences about $\beta_i$ are derived from properties of the sampling distribution of $b_i$. Formulas for $b_i$ and its standard deviation $\sigma_{b_i}$ are quite complicated and cannot be stated concisely except by using advanced mathematical notation.

### The Sampling Distribution of $b_i$

Let $b_i$ denote the statistic for estimating (via the principle of least squares) the coefficient $\beta_i$ in the model $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$. Assumptions about this model given in Section 14.1 of the textbook imply the following:

1. $b_i$ has a normal distribution.
2. $\mu_{b_i} = \beta_i$, so $b_i$ is an unbiased statistic for estimating $\beta_i$.
3. The standard deviation of $b_i$, $\sigma_{b_i}$, involves $\sigma^2$ and a complicated function of all the values of $x_1, x_2, \ldots, x_k$ in the sample. The estimated standard deviation $s_{b_i}$ results from replacing $\sigma^2$ with $s^2$ in the formula for $\sigma_{b_i}$.

A consequence of the properties of the sampling distribution of $b_i$ is that the standardized variable

$$t = \frac{b_i - \beta_i}{s_{b_i}}$$

has a $t$ distribution with $df = n - (k + 1)$. This leads to a confidence interval for $\beta_i$. 

---

In the section 14.2 in the textbook, we discussed estimating the coefficients $\alpha, \beta_1, \ldots, \beta_k$ in the model $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$ (using the principle of least squares) and then showed how the usefulness of the model could be confirmed by application of the F test for model utility. If $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ cannot be rejected at a reasonably small level of significance, it must be concluded that the model does not specify a useful relationship between $y$ and any of the predictor variables $x_1, x_2, \ldots, x_k$. The investigator must then search further for a model that does describe a useful relationship, perhaps by introducing different predictors or making variable transformations. Only if $H_0$ can be rejected is it appropriate to proceed further with the chosen model and make inferences based on the estimated coefficients $\alpha, b_1, \ldots, b_k$ and the estimated regression function $\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$. In this section, we will consider two different types of inferential problems. One type involves drawing a conclusion about an individual regression coefficient $\beta_i$—either computing a confidence interval for $\beta_i$ or testing a hypothesis concerning $\beta_i$. The second type of problem involves first fixing values of $x_1, x_2, \ldots, x_k$ and then computing either a point estimate or a confidence interval for the corresponding mean $y$ value or predicting a future $y$ value with a point prediction or a prediction interval.

## Inferences About a Regression Coefficient

A confidence interval for the slope coefficient $\beta$ and a hypothesis test about $\beta$ in a simple linear regression were based on facts about the sampling distribution of the statistic $b$ used to obtain a point estimate of $\beta$. Similarly, in multiple regression, procedures for making inferences about $\beta_i$ are derived from properties of the sampling distribution of $b_i$. Formulas for $b_i$ and its standard deviation $\sigma_{b_i}$ are quite complicated and cannot be stated concisely except by using advanced mathematical notation.

### The Sampling Distribution of $b_i$

Let $b_i$ denote the statistic for estimating (via the principle of least squares) the coefficient $\beta_i$ in the model $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$. Assumptions about this model given in Section 14.1 of the textbook imply the following:

1. $b_i$ has a normal distribution.
2. $\mu_{b_i} = \beta_i$, so $b_i$ is an unbiased statistic for estimating $\beta_i$.
3. The standard deviation of $b_i$, $\sigma_{b_i}$, involves $\sigma^2$ and a complicated function of all the values of $x_1, x_2, \ldots, x_k$ in the sample. The estimated standard deviation $s_{b_i}$ results from replacing $\sigma^2$ with $s^2$ in the formula for $\sigma_{b_i}$.

A consequence of the properties of the sampling distribution of $b_i$ is that the standardized variable

$$t = \frac{b_i - \beta_i}{s_{b_i}}$$

has a $t$ distribution with $df = n - (k + 1)$. This leads to a confidence interval for $\beta_i$.
A Confidence Interval for $\beta_i$

A confidence interval for $\beta_i$ is

$$b_i \pm (t \text{ critical value}) s_{bi}$$

The $t$ critical value is based on $df = n - (k + 1)$.

Any good statistical computer package will provide both the estimated coefficients $a$, $b_1$, …, $b_k$ and their estimated standard deviations $s_a$, $s_{b_1}$, …, $s_{b_k}$.

Example 14.12 Soil and Sediment Characteristics

Soil and sediment adsorption, the extent to which chemicals collect in a condensed form on the surface, is an important characteristic because it influences the effectiveness of pesticides and various agricultural chemicals. The article “Adsorption of Phosphates, Arsenate, Methane Arsenate, and Cacodylate by Lake and Stream Sediments: Comparisons with Soils” (Journal of Environmental Quality [1984]: 499–504) presented the following data consisting of $n = 13$ ($x_1$, $x_2$, $y$) triples and proposed the model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e$$

for relating

$y = \text{phosphate adsorption index}$

$x_1 = \text{amount of extractable iron}$

$x_2 = \text{amount of extractable aluminum}$

<table>
<thead>
<tr>
<th>Observation</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>61</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>175</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>111</td>
<td>24</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>124</td>
<td>23</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>130</td>
<td>64</td>
<td>26</td>
</tr>
<tr>
<td>6</td>
<td>173</td>
<td>38</td>
<td>26</td>
</tr>
<tr>
<td>7</td>
<td>169</td>
<td>33</td>
<td>21</td>
</tr>
<tr>
<td>8</td>
<td>169</td>
<td>61</td>
<td>30</td>
</tr>
<tr>
<td>9</td>
<td>160</td>
<td>39</td>
<td>28</td>
</tr>
<tr>
<td>10</td>
<td>244</td>
<td>71</td>
<td>36</td>
</tr>
<tr>
<td>11</td>
<td>257</td>
<td>112</td>
<td>65</td>
</tr>
<tr>
<td>12</td>
<td>333</td>
<td>88</td>
<td>62</td>
</tr>
<tr>
<td>13</td>
<td>199</td>
<td>54</td>
<td>40</td>
</tr>
</tbody>
</table>
The regression equation is
\[ HPO = -7.35 + 0.113 \text{FE} + 0.349 \text{AL} \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-7.35</td>
<td>3.485</td>
<td>-2.11</td>
<td>0.061</td>
</tr>
<tr>
<td>FE</td>
<td>0.11273</td>
<td>0.02969</td>
<td>3.80</td>
<td>0.004</td>
</tr>
<tr>
<td>AL</td>
<td>0.34900</td>
<td>0.07131</td>
<td>4.89</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ s = 4.379 \quad \text{R-sq} = 94.8\% \quad \text{R-sq(adj)} = 93.8\% \]

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>3529.9</td>
<td>1765.0</td>
<td>92.03</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>10</td>
<td>191.8</td>
<td>19.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>3721.7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The F statistic for the model utility test is 92.03 with an associated \( P \)-value of 0.000, indicating that the model is useful. The values of \( s \) and \( R^2 \) are \( s = 4.379 \) and \( R^2 = .948 \). The values of \( R^2 \) and \( s \) suggest that the chosen model has been very successful in relating \( y \) to the predictors.

Suppose that the investigator had required a 95% confidence interval for \( \beta_1 \), the average change in phosphate adsorption when extractable iron increases by 1 unit and extractable aluminum is held fixed. The necessary quantities are

\[ b_1 = .11273 \quad s_{b_1} = .02969 \quad \text{(from the Coef and Stdev columns in the output)} \]

\[ df = n - (k + 1) = 13 - (2 + 1) = 10 \]

\[ t \text{ critical value} = 2.23 \]

The resulting confidence interval is

\[ b_1 \pm (t \text{ critical value})s_{b_1} = .11273 \pm (2.23)(.02969) = .11273 \pm .06621 = (.047, .179) \]

The standardized variable \( t = \frac{b_i - \beta_i}{s_{b_i}} \) is also the basis for a test statistic for testing hypotheses about \( \beta_i \).

### Testing Hypotheses About \( \beta_i \)

**Null hypothesis:** \( H_0: \beta_i = \text{hypothesized value} \)

**Test statistic:** \[ t = \frac{b_i - \text{hypothesized value}}{s_{b_i}} \]

The test is based on \( df = n - (k + 1) \).

(continued)
Example 14.13 The Price of Fish

Economists are always interested in studying the factors affecting the price paid for a product. The article “Testing for Imperfect Competition at the Fulton Fish Market” (Rand Journal of Economics [1995]: 75–92) considered a regression of $y = \text{price paid for whiting}$ on eight different predictors. (The Fulton Fish Market, located in New York City, is the largest such market in the United States; the many dealers make it highly competitive, and whiting is one of the most frequently purchased types of fish.) There were $n = 54$ observations, and the coefficient of multiple determination was $R^2 = .992$, indicating a highly useful model (model utility $F = 697.5$, $P$-value = .000). One of the predictors used, $x_3$, was an indicator variable for whether a purchaser was Asian. The values $b_3 = .0463$ and $s_{b_3} = .0191$ were given. Provided that all seven other predictors remain in the model, does $x_3$ appear to provide useful information about $y$?

1. $b_3$ is the average difference in price paid by Asian and non-Asian purchasers when all other predictors are held fixed.
2. $H_0: \beta_3 = 0$
3. $H_a: \beta_3 \neq 0$
4. A significance level of $\alpha = .05$ was suggested in the article.
5. Test statistic: $t = \frac{b_3}{s_{b_3}}$
6. Assumptions: Without raw data, we cannot assess the reasonableness of the assumptions. If we had the data, a normal probability plot or boxplot of the standardized residuals would be a good place to start. For purposes of this example, we will assume that it is reasonable to proceed with the test.
7. The test statistic value is
\[ t = \frac{b_3}{s_{b_3}} = \frac{-0.0463}{0.0191} = -2.42 \]

8. The test is based on \( n - (k + 1) = 54 - (8 + 1) = 45 \) df. An examination of the 40 and 60 df columns of Appendix Table 4 shows that the \( P \)-value is roughly 2(.010) = .020.

9. Because \( P \)-value = .020 ≤ .05 = \( \alpha \), \( H_0 \) should be rejected. The predictor that indicates whether a purchaser is Asian does appear to provide useful information about price, over and above the information contained in the other predictors. The author of the article indicated that this result is inconsistent with the model of perfect competition.

**Example 14.14 Soil Characteristics Continued**

Our analysis of the phosphate adsorption data introduced in Example 14.12 has so far focused on the model \( y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e \) in which \( x_1 \) (extractable iron) and \( x_2 \) (extractable aluminum) affect the response separately. Suppose that the researcher wishes to investigate the possibility of interaction between \( x_1 \) and \( x_2 \) through fitting the model with predictors \( x_1, x_2, \) and \( x_3 = x_1 x_2 \). We list a few of the sample values of \( y, x_1, \) and \( x_2 \) along with the corresponding values of \( x_3 \):

<table>
<thead>
<tr>
<th>Observation</th>
<th>( y )</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>61</td>
<td>13</td>
<td>793</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>175</td>
<td>21</td>
<td>3675</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>111</td>
<td>24</td>
<td>2664</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>40</td>
<td>54</td>
<td>10,746</td>
</tr>
</tbody>
</table>

In practice, a statistical software package would automatically compute the \( x_3 \) values upon request once the \( x_1 \) and \( x_2 \) values had been entered, so hand computations would not be necessary. Figure 14.12 displays partial MINITAB output resulting from a request to fit this model. Let’s use the output to see whether inclusion of the interaction predictor is justified.

**Figure 14.12** MINITAB output for model with interaction fit to the phosphate adsorption data of Example 14.12 (\( x_1 = \text{FE}, x_2 = \text{AL}, x_3 = x_1 x_2 = \text{FEAL} \)).

The regression equation is
\[ \text{HPO} = -2.37 + 0.0828 \text{FE} + 0.246 \text{AL} + 0.000528 \text{FEAL} \]

<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>-2.368</td>
<td>7.179</td>
<td>-0.33</td>
<td>0.749</td>
</tr>
<tr>
<td>FE</td>
<td>0.08279</td>
<td>0.04818</td>
<td>1.72</td>
<td>0.120</td>
</tr>
<tr>
<td>AL</td>
<td>0.2460</td>
<td>0.1481</td>
<td>1.66</td>
<td>0.131</td>
</tr>
<tr>
<td>FEAL</td>
<td>0.0005278</td>
<td>0.0006610</td>
<td>0.80</td>
<td>0.445</td>
</tr>
<tr>
<td>( S = 4.461 )</td>
<td>( R-Sq = 95.2% )</td>
<td>( R-Sq(adj) = 93.6% )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. The model is \( y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e \) where \( x_3 = x_1 x_2 \)

2. \( H_0: \beta_3 = 0 \)

3. \( H_a: \beta_3 \neq 0 \)

4. Test statistic:
   \[
   t = \frac{b_3}{s_{b_3}}
   \]

5. Significance level: \( \alpha = .05 \)

6. Assumptions: The residuals and standardized residuals (from MINITAB) for this model are as follows:

<table>
<thead>
<tr>
<th>Obs</th>
<th>FE</th>
<th>HPO</th>
<th>Residual</th>
<th>StResid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>61</td>
<td>4</td>
<td>-2.29844</td>
<td>-0.94279</td>
</tr>
<tr>
<td>2</td>
<td>175</td>
<td>18</td>
<td>-1.22530</td>
<td>-0.32746</td>
</tr>
<tr>
<td>3</td>
<td>111</td>
<td>14</td>
<td>-0.13111</td>
<td>-0.03250</td>
</tr>
<tr>
<td>4</td>
<td>124</td>
<td>18</td>
<td>2.93937</td>
<td>0.71808</td>
</tr>
<tr>
<td>5</td>
<td>130</td>
<td>26</td>
<td>-2.52836</td>
<td>-0.76874</td>
</tr>
<tr>
<td>6</td>
<td>173</td>
<td>26</td>
<td>1.22858</td>
<td>0.29624</td>
</tr>
<tr>
<td>7</td>
<td>169</td>
<td>21</td>
<td>-1.68416</td>
<td>-0.40933</td>
</tr>
<tr>
<td>8</td>
<td>169</td>
<td>30</td>
<td>-2.06905</td>
<td>-0.51864</td>
</tr>
<tr>
<td>9</td>
<td>160</td>
<td>28</td>
<td>4.23516</td>
<td>1.00631</td>
</tr>
<tr>
<td>10</td>
<td>244</td>
<td>36</td>
<td>-8.44065</td>
<td>-2.07895</td>
</tr>
<tr>
<td>11</td>
<td>257</td>
<td>65</td>
<td>3.34931</td>
<td>1.19361</td>
</tr>
<tr>
<td>12</td>
<td>333</td>
<td>62</td>
<td>-0.31378</td>
<td>-0.12606</td>
</tr>
<tr>
<td>13</td>
<td>199</td>
<td>40</td>
<td>6.93843</td>
<td>1.68295</td>
</tr>
</tbody>
</table>

A normal probability plot of the standardized residuals is shown here.

The plot is quite straight, indicating that normality of the random error distribution is plausible.

7. By reading directly from the \( t \) ratio column in Figure 14.12, we find \( t = .80 \).

8. Figure 14.12 shows that the \( P \)-value is .445.

9. Because \( .445 > .05 \), \( H_0 \) cannot be rejected. The interaction predictor does not appear to be useful and can be removed from the model.
An interesting aspect of the computer output for the interaction model in Example 14.14 is that the $t$ ratios for $\beta_1$, $\beta_2$, and $\beta_3$ (1.72, 1.66, and 0.80, respectively) are all relatively small and the corresponding $P$-values are large, yet $R^2 = .952$ is quite large. The high $R^2$ value suggests a useful model (this can be confirmed by the model utility test), yet the size of each $t$ ratio and $P$-value might tempt us to conclude that all three $\beta_i$'s are 0. This sounds like a contradiction, but it involves a misinterpretation of the $t$ ratios. For example, the $t$ ratio for $\beta_2$, $t = \frac{b_2}{s_{b_2}}$, tests the null hypothesis $H_0$: $\beta_2 = 0$ when $x_1$ and $x_2$ are included in the model. The smallness of a given $t$ ratio suggests that the associated predictor can be dropped from the model as long as the other predictors are retained. The fact that all $t$ ratios are small in this example does not, therefore, allow us to simultaneously delete all predictors. The data do suggest deleting $x_3$ because the model that includes $x_1$ and $x_3$ has already been found to give a very good fit to the data and is simple to interpret. In Section 14.4, we comment further on why it might happen that all $t$ ratios are small even when the model seems quite useful.

The model utility test amounts to testing a simultaneous claim about the values of all the $\beta_i$'s—that they are all 0. There is also an $F$ test for testing a hypothesis involving a specified subset consisting of at least two $\beta_i$'s. For example, we might fit the model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + e$$

and then wish to test $H_0$: $\beta_3 = \beta_4 = \beta_5 = 0$ (which says that the second-order predictors contribute nothing to the model). Please see the book by Neter, Wasserman, and Kutner listed in the references at the back of the book for further details.

---

### Inferences Based on an Estimated Model

The estimated regression line $\hat{y} = a + bx$ in simple linear regression was used both to estimate the mean $y$ value when $x$ had a specified value and to predict the associated $y$ value for a single observation made at a particular $x$ value. The estimated regression function for the model $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$ can be used in the same two ways. When fixed values of the predictor variables $x_1, x_2, \ldots, x_k$, are substituted into the estimated regression function

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

the result can be used either as a point estimate of the corresponding mean $y$ value or as a point prediction of the $y$ value that will result from a single observation when the $x_i$'s have the specified values.

---

**Example 14.15** Predicting Transit Times

Precise information concerning bus transit times is important when making transportation planning decisions. The article “Factors Affecting Running Time on Transit Routes” (Transportation Research [1983]: 107–113) reported on an empirical study based on data gathered in Cincinnati, Ohio. The variables of interest were

- $y =$ running time per mile during the morning peak period (in seconds)
- $x_1 =$ number of passenger boardings per mile
- $x_2 =$ number of passenger alightings per mile
- $x_3 =$ number of signaled intersections per mile
- $x_4 =$ proportion of the route on which parking is allowed
The values of $x_1$ and $x_2$ were not necessarily equal, because observations were made over partial segments of routes (so not all passengers entered or exited on the segments). The estimated regression function was

$$\hat{y} = 169.50 + 5.07x_1 + 4.53x_2 + 6.61x_3 + 67.70x_4$$

Consider the predictor variable values $x_1 = 4.5$, $x_2 = 5.5$, $x_3 = 5$, and $x_4 = .1$. Then a point estimate for true average running time per mile when the $x$’s have these values is

$$\hat{y} = 169.50 + 5.07(4.5) + 4.53(5.5) + 6.61(5) + 67.70(.1) = 257.05$$

This value, 257.05 seconds, is also the predicted running time per mile for a single trip when $x_1$, $x_2$, $x_3$, and $x_4$ have the given values.

Remember that before the sample observations $y_1, y_2, \ldots, y_n$ are obtained, $a$ and all $b_i$’s are statistics (because they all involve the $y_i$’s). This implies that for fixed values of $x_1, x_2, \ldots, x_k$ the estimated regression function $a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$ is a statistic (its value varies from sample to sample). To obtain a confidence interval for the mean $y$ value for specified $x_1, x_2, \ldots, x_k$ values, we need some facts about the sampling distribution of this statistic.

### Properties of the Sampling Distribution of $\hat{y} = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$

Assumptions about the model $y = \alpha + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k + \epsilon$ given in Section 14.1 of the textbook imply that for fixed values of the predictors $x_1, x_2, \ldots, x_k$, the statistic $\hat{y} = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$ satisfies the following properties:

1. It has a normal distribution.
2. Its mean value is $\alpha + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k$. That is, the statistic is unbiased for estimating the mean $y$ value when $x_1, x_2, \ldots, x_k$ are fixed.
3. The standard deviation of $\hat{y}$ involves $\sigma^2$ and a complicated function of all the sample predictor variable values. The estimated standard deviation of $\hat{y}$, denoted by $s_\hat{y}$, results from replacing $\sigma^2$ with $s_y^2$ in this function.

The formula for $s_\hat{y}$ has been programmed into the most widely used statistical computer packages, and its value for specified $x_1, x_2, \ldots, x_k$ is available upon request. Manipulation of the $t$ variable as before gives a confidence interval formula. A prediction interval formula is based on a similar standardized variable.

For fixed values of $x_1, x_2, \ldots, x_k$, a **confidence interval for the mean $y$ value**—that is, for $\alpha + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k$—is

$$\hat{y} \pm (t \text{ critical value})s_\hat{y}$$

A **prediction interval for a single $y$ value** is

$$\hat{y} \pm (t \text{ critical value}) \sqrt{s_y^2 + s_\hat{y}^2}$$

The $t$ critical value in both intervals is based on $n - (k + 1)$ df.
Example 14.16 Soil Characteristics—Last Time!

Figure 14.11 shows MINITAB output from fitting the model \( y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e \) to the phosphate data introduced previously. A request for estimation and prediction information when \( x_1 = 150 \) and \( x_2 = 40 \) resulted in the following additional output:

<table>
<thead>
<tr>
<th>Fit</th>
<th>StDev Fit</th>
<th>95.0% CI</th>
<th>95.0% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>23.52</td>
<td>1.31</td>
<td>(20.60, 26.44)</td>
<td>(13.33, 33.70)</td>
</tr>
</tbody>
</table>

The \( t \) critical value for a 95% confidence level when \( df = 10 \) is 2.23, and the corresponding interval is

\[
23.52 \pm (2.23)(1.31) = 23.52 \pm 2.92 = (20.60, 26.44)
\]

Using \( s^2 = 4.379 \) along with the given values of \( \hat{y}, x_i \), and \( t \) critical value in the prediction interval formula results in \((13.33, 33.70)\). Both intervals are centered at \( \hat{y} \) but the prediction interval is much wider than the confidence interval.

The danger of extrapolation in simple linear regression is that if the \( x \) value of interest is much outside the interval of \( x \) values in the sample, the proposed model might no longer be valid, and even if it were, \( \alpha + \beta x \) could still be quite poorly estimated (\( s_{\alpha+\beta x} \) could be large). There is a similar danger in multiple regression, but it is not always obvious when the \( x \) values of interest involve an extrapolation from sample data. As an example, suppose that a single \( y \) observation is made for each of the following 13 \((x_i, x_2)\) pairs:

<table>
<thead>
<tr>
<th>Observation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
<td>10</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>10</td>
<td>5</td>
<td>0</td>
<td>-5</td>
<td>-5</td>
<td>-10</td>
<td>-5</td>
<td>0</td>
<td>5</td>
<td>-5</td>
<td>-5</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The \( x_1 \) values range between -10 and 10, as do the \( x_2 \) values. After fitting the model \( y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e \) we might then want a confidence interval for the mean \( y \) value when \( x_1 = 10 \) and \( x_2 = 10 \). However, whereas each of these values separately is within the range of sample \( x \) values, Figure 14.13 shows that the point \((10, 10)\) is actually far from \((x_i, x_2)\) pairs in the sample. Thus, drawing a conclusion about \( y \) when \( x_1 = 10 \) and \( x_2 = 10 \) involves a substantial extrapolation. In particular, the estimated standard deviation of \( a + b_1 x_1 + b_2 x_2 \) would probably be quite large even if the model was valid near this point.

When more than two predictor variables are included in the model, we cannot tell from a plot like that of Figure 14.13 whether the \( x_1, x_2, \ldots, x_k \) values of interest involve an extrapolation. It is then best to compare \( x_i \)'s based on \( a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \) for the values of interest with the \( x_i \)'s corresponding to \( x_1, x_2, \ldots, x_k \) values in the sample (MINITAB gives these on request). Extrapolation is indicated by a value of the former standard deviation (at the \( x \) values of interest) that is much larger than the standard deviations for sampled values.
14.34 Explain why it is preferable to perform a model utility test before using an estimated regression model to make predictions or to estimate the mean $y$ value for specified values of the independent variables.

14.35 The article “Zoning and Industrial Land Values: The Case of Philadelphia” (AREUEA Journal [1991]: 154–159) considered a regression model to relate the value of a vacant lot in Philadelphia to a number of different predictor variables.

a. One of the predictors was $w_3 = \text{distance from the city’s major east-west thoroughfare, for which } b_3 = -.489$ and $s_{b_3} = .1044$. The model contained seven predictors, and the coefficients were estimated from a sample of 100 vacant lots. Calculate and interpret a confidence interval for $\beta_3$, using a confidence level of 95%.

b. Another predictor was $x_3$, an indicator variable for whether the lot was zoned for residential use, for which $b_1 = -1.83$ and $s_{b_1} = .3055$. Carry out a test of $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$, and interpret the conclusion in the context of the problem.

14.36 Twenty-six observations from the article “Multiple Regression Analysis for Forecasting Critical Fish Influxes at Power Station Intakes” (see Exercise 14.15) were used to fit a multiple regression model relating $y = \text{number of fish at intake to the independent variables } x_1 = \text{water temperature (°C)}, x_2 = \text{number of pumps running}, x_3 = \text{sea state (taking values 0, 1, 2, or 3), and } x_4 = \text{speed (knots)}. Partial MINITAB output follows.

The regression equation is

$$\hat{y} = 92.0 - 2.18x_1 - 19.2x_2 - 9.38x_3 + 2.32x_4$$

**Predictor**  | **Coeff** | **Stdev** | **t-ratio**
--- | --- | --- | ---
Constant | 91.98 | 42.07 | 2.19
X1 | -2.179 | 1.087 | -2.00
X2 | -19.189 | 9.215 | -2.08
X3 | -9.378 | 4.356 | -2.15
X4 | 2.3205 | 0.7686 | 3.02

$s = 10.53$  \(R^2 = 39.0\%\)  \(R^2(\text{adj}) = 27.3\%\)

a. Construct a 95% confidence interval for $\beta_3$, the coefficient of $x_3 = \text{sea state.}$ Interpret the resulting interval.

b. Construct a 90% confidence interval for the mean change in $y$ associated with a $1^\circ$ increase in water temperature when number of pumps, sea state, and speed remain fixed.

14.37 A study reported in the article “Leakage of Intracellular Substances from Alfalfa Roots at Various Subfreezing Temperatures” (Crop Science [1991]: 1575–1578) considered a quadratic regression model to relate $y = \text{MDH activity}$ (a measure of the extent to which cellular membranes suffer extensive damage from freezing) to $x = \text{electrical conductivity}$ (which describes the damage in the early stages of freezing). The estimated regression function was

$$\hat{y} = -.1838 + .0272x + .0446x^2$$

with $R^2 = .860$.

a. Supposing that $n = 50$, does the quadratic model appear to be useful? Test the appropriate hypotheses.

b. Using $s_b = .0103$, carry out a test at significance level .01 to decide whether the quadratic predictor $x^2$ is important.

c. If the standard deviation of the statistic $\hat{y} = a + b_1(40) + b_2(1600)$ is $s_{\hat{y}} = .120$, calculate a confidence interval with confidence level 90% for true average MDH activity when conductivity is 40.

14.38 The first article introduced in Exercise 14.25 of Section 14.2 in the textbook gave data on the dimensions of 27 representative food products. Use the multiple regression model fit in Exercise 14.25.

a. Could any of these variables be eliminated from a regression with the purpose of predicting volume?

b. Predict the volume of a package with a minimum width of 2.5 cm, a maximum width of 3.0 cm, and an elongation of 1.55.

c. Calculate a 95% prediction interval for a package with a minimum width of 2.5 cm, a maximum width of 3.0 cm, and an elongation of 1.55.

14.39 Data from a random sample of 107 students taking a managerial accounting course were used to obtain the estimated regression equation $2.178 + .469x_1 + 3.369x_2 + 3.054x_3$ where $y = \text{student’s exam score}, x_1 = \text{student’s expected score on the exam}, x_2 = \text{time spent studying (hr/week), and } x_3 = \text{student’s grade point average or GPA ("Effort, Expectation and Academic Performance in Managerial Cost Accounting," Journal of Accounting Education [1989]: 57–68).}$ The value of $R^2$ was .686, and the estimated standard deviations of the statistics $b_1$, $b_2$, and $b_3$ were .090, .456, and 1.457, respectively.

a. How would you interpret the estimated coefficient .469?

b. Does there appear to be a useful linear relationship between exam score and at least one of the three predictors?
c. Calculate a confidence interval for the mean change in exam score associated with a 1-hour increase in study time when expected score and GPA remain fixed.

d. Obtain a point prediction of the exam score for a student who expects a 75, has studied 8 hours per week, and has a GPA of 2.8.

e. If the standard deviation of the statistic on which the prediction of Part (d) is based is 1.2 and SSTo = 10,200, calculate a 95% prediction interval for the score of the student described in Part (d).

14.40 Benevolence payments are monies collected by a church to fund activities and ministries beyond those provided by the church to its own members. The article “Optimal Church Size: The Bigger the Better” (Journal for the Scientific Study of Religion [1993]: 231–241) considered a regression of $y$ = benevolence payments on $x_1 =$ number of church members, $x_2 = x_1^2$ and $x_3 =$ an indicator variable for urban versus nonurban churches.

a. The sample size was $n = 300$, and adjusted $R^2 = .774$. Does at least one of the three predictors provide useful information about $y$?

b. The article reported that $b_3 = 101.1$ and $s_{b_3} = 625.8$. Should the indicator variable be retained in the model? Test the relevant hypotheses.

14.41 The estimated regression equation

$$
\hat{y} = 28 - .05x_1 - .003x_2 + .00002x_3
$$

where $y =$ number of indictments disposed of in a given month; $x_1 =$ number of cases on judge’s docket; $x_2 =$ number of cases pending in the criminal trial court; $x_3 =$ an indicator for floodplain stations.

appeared in the article “The Caseload Controversy and the Study of Criminal Courts” (Journal of Criminal Law and Criminology [1979]: 89–101). This equation was based on $n = 367$ observations. The $b_i$’s and their associated standard deviations are given in the accompanying table.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$b_i$</th>
<th>Estimated Standard Deviation of $b_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-.05</td>
<td>.03</td>
</tr>
<tr>
<td>2</td>
<td>-.003</td>
<td>.0024</td>
</tr>
<tr>
<td>3</td>
<td>.00002</td>
<td>.000009</td>
</tr>
</tbody>
</table>

14.42 The accompanying data were obtained from a study of a certain method for preparing pure alcohol from refinery streams (“Direct Hydration of Olefins” Industrial and Engineering Chemistry [1961]: 209–211). The independent variable $x$ is volume, and the dependent variable $y$ is amount of isobutylene converted.

$$
\begin{array}{cccccccc}
\text{x} & 1 & 1 & 2 & 4 & 4 & 4 & 6 \\
\text{y} & 23.0 & 24.5 & 28.0 & 30.9 & 32.0 & 33.6 & 20.0 \\
\end{array}
$$

Minitab output—the result of fitting a quadratic regression model where $x_1 = x$ and $x_2 = x^2$—is given. Would a linear regression have sufficed? That is, is the quadratic predictor important? Use a level .05 test.

The regression equation is

$$
\hat{y} = 13.6 + 11.4x_1 - 1.72x_2
$$

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>13.63</td>
<td>1.89</td>
<td>7.19</td>
</tr>
<tr>
<td>X1</td>
<td>11.40</td>
<td>1.36</td>
<td>8.41</td>
</tr>
<tr>
<td>X2</td>
<td>-1.72</td>
<td>0.20</td>
<td>-8.42</td>
</tr>
<tr>
<td>s</td>
<td>1.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-sq</td>
<td>94.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-sq(adj)</td>
<td>92.1%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

14.43 The article “Bank Full Discharge of Rivers” (Water Resources Journal [1978]: 1141–1154) reported data on $y =$ discharge amount (m$^3$/s), $x_1 =$ flow area (m$^2$) and $x_2 =$ slope of the water surface (m/m) obtained at $n = 10$ floodplain stations. A multiple regression model using $x_1$, $x_2$, and $x_3 =$ $x_1x_2$ was fit to this data, and partial Minitab output appears here.

The regression equation is

$$
\hat{y} = -3.14 + 1.70x_1 + 96.1x_2 + 8.38x_3
$$

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-3.14</td>
<td>14.54</td>
<td>-0.22</td>
</tr>
<tr>
<td>X1</td>
<td>1.69</td>
<td>1.43</td>
<td>1.19</td>
</tr>
<tr>
<td>X2</td>
<td>96.1</td>
<td>702.7</td>
<td>0.14</td>
</tr>
<tr>
<td>X3</td>
<td>8.38</td>
<td>199.0</td>
<td>0.04</td>
</tr>
<tr>
<td>s</td>
<td>17.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-sq</td>
<td>73.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-sq(adj)</td>
<td>59.9%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Perform the model utility test using a significance level of .05.

b. Is the interaction term important? Test using a .05 significance level.
14.44 In the article “An Ultracentrifuge Flour Absorption Method” (Cereal Chemistry [1978]: 96–101), the authors discussed the relationship between water absorption for wheat flour and various characteristics of the flour. The model $y = \alpha + \beta_1x_1 + \beta_2x_2 + e$ was used to relate $y =$ absorption (%) to $x_1 =$ flour protein (%) and $x_2 =$ starch damage (Farrand units). MINITAB output based on $\alpha = .05$ for all tests requested.

Use a significance level of .05 for all tests requested.

The regression equation is

$$Y = 19.4 + 1.44X_1 + .336X_2$$

The effect of manganese (Mn) on wheat growth is examined in the article “Manganese Deficiency and Toxicity Effects on Growth, Development and Nutrient Composition in Wheat” (Agronomy Journal [1984]: 213–217). A quadratic regression model was used to relate $y = $ plant height (cm) to $x = \log ($ added Mn$)$, with $\mu$M as the units for added Mn. The accompanying data were read from a scatterplot in the article. Also given is MINITAB output, where $x_1 = x$ and $x_2 = x^2$. Use a .05 significance to describe the relationship between weight and the predictors length and age.

14.45 Exercise 14.23 gave data on fish weight, length, age, and year caught. A multiple regression model was fit to describe the relationship between weight and the predictors length and age.

**c.** Does it bother you that the model utility test indicates a useful model, but all values in the t-ratio column of the output are small? Explain.

**d.** Calculate and interpret a 95% confidence interval for $\beta_2$.

**e.** Based on the results of Part (c), would you conclude that both independent variables are important? Explain.

**f.** Create a dummy (indicator) variable for year caught. Fit a multiple regression model that includes year, length, and age as predictors of weight. Is there evidence that year is a useful predictor given that length and age are included in the model? Test the relevant hypotheses using $\alpha = .05$.

14.46 The article “Predicting Marathon Time from Anaerobic Threshold Measurements” (The Physician and Sports Medicine [1984]: 95–98) gave data on $y =$ maximum heart rate (beats/min), $x_1 =$ age, and $x_2 =$ weight (kg) for $n = 18$ marathon runners. The estimated regression equation for the model $y = \alpha + \beta_1x_1 + \beta_2x_2 + e$ was $\hat{y} = 179 - .8x_1 + .5x_2$ and SSRegr $= 649.75$, SSResid $= 538.03$.

**a.** Is the model useful for predicting maximum heart rate? Use a significance level of .10.

**b.** Using $s_\beta = .280$, calculate and interpret a 95% confidence interval for $\beta_1$.

**c.** Predict the maximum heart rate of a particular runner who is 43 years old and weighs 65 kg, using a 99% interval. The estimated standard deviation of the statistic $a + b_1(30) + b_2(65)$ is 3.52.

**d.** Use a 90% interval to estimate the average maximum heart rate for all marathon runners who are 30 years old and weigh 77.2 kg. The estimated standard deviation of $a + b_1(30) + b_2(77.2)$ is 2.97.

**e.** Would a 90% prediction interval for a single 30-year-old runner weighing 77.2 kg be wider or narrower than the interval computed in Part (d)? Explain. (You need not compute the interval.)
level for any hypothesis tests needed to answer the questions that follow.

\[ x \begin{array}{cccccccc} -1 & -0.4 & 0 & 2 & 2.8 & 3.2 & 3.4 & 4 \\ y \end{array} \begin{array}{cccccccc} 32 & 37 & 44 & 45 & 46 & 42 & 42 & 40 & 37 & 30 \end{array} \]

The regression equation is

\[ Y = 41.7 + 6.58X_1 - 2.36X_2 \]

Predictor Coef Stdev t-ratio
Constant 41.7422 0.8522 48.98
X1 6.581 1.002 6.57
X2 -2.3621 0.3073 -7.69
s = 1.963 R-sq = 89.8% R-sq(adj) = 86.9%

Analysis of Variance

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>237.520</td>
<td>118.760</td>
</tr>
<tr>
<td>Error</td>
<td>7</td>
<td>26.980</td>
<td>3.854</td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>264.500</td>
<td></td>
</tr>
</tbody>
</table>

**a.** Is the quadratic model useful for describing the relationship between \( y \) and \( x \)?

**b.** Are both the linear and quadratic predictors important? Could either one be eliminated from the model? Explain.

**c.** Give a 95% confidence interval for the mean \( y \) value when \( x = 2 \). The estimated standard deviation of \( a + b_1(2) + b_2(4) \) is 1.037. Interpret the resulting interval.

**d.** Estimate the mean height for wheat treated with 10 \( \mu M \) of Mn using a 90% interval. (Note: The estimated standard deviation of \( a + b_1(1) + b_2(1) \) is 1.031 and \( \log(10) = 1.0\).

14.48 **This exercise requires the use of a computer package.** A study on the effect of applying fertilizer in bands is described in the article “Fertilizer Placement Effects on Growth, Yield, and Chemical Composition of Burley Tobacco” (Agronomy Journal [1984]: 183–188). The accompanying data were taken from a scatterplot appearing in the article, with \( y = \) plant Mn (\( \mu g/g \) dry weight) and \( x = \) distance from the fertilizer band (cm). The authors suggest a quadratic regression model.

\[ x \begin{array}{cccccccc} 0 & 10 & 20 & 30 & 40 \\ y \end{array} \begin{array}{cccccccc} 110 & 90 & 76 & 72 & 70 \end{array} \]

**a.** Use a suitable computer package to find the estimated quadratic regression equation.

**b.** Perform the model utility test.

**c.** Interpret the values of \( R^2 \) and \( s_\varepsilon \).

**d.** Are both the linear and quadratic predictors important? Carry out the necessary hypothesis tests and interpret the results.

**e.** Find a 90% confidence interval for the mean plant Mn for plants that are 30 cm from the fertilizer band.

**Bold** exercises answered in back  ●  Data set available online but not required  ▼  Video solution available

### 14.4 Other Issues in Multiple Regression

Primary objectives in multiple regression include estimating a mean \( y \) value, predicting an individual \( y \) value, and gaining insight into how changes in predictor variable values affect \( y \). Often an investigator has data on a number of predictor variables that might be incorporated into a model to be used for such purposes. Some of these predictor variables may actually be unrelated or only weakly related to \( y \), or they may contain information that duplicates information provided by other predictors. If all these predictors are included in the model, many model coefficients will have to be estimated. This reduces the number of degrees of freedom associated with SS Resid, leading to a deterioration in the degree of precision associated with other inferences (e.g., wide confidence and prediction intervals). A model with many predictors can also be cumbersome to use and difficult to interpret.

In this section, we first introduce some guidelines and procedures for selecting a set of useful predictors. In choosing a model, the analyst should examine the data carefully for evidence of unusual observations or potentially troublesome patterns. It is important to identify unusually deviant or influential observations and to look for possible inconsistencies with model assumptions. Our discussion of multiple regression closes with a brief mention of some diagnostic methods designed for these purposes.
Suppose that an investigator has data on $p$ predictors $x_1, x_2, \ldots, x_p$, which are candidates for use in building a model. Some of these predictors might be specified functions of others, for example, $x_3 = x_1x_2$, $x_4 = x_1^2$, and so on. The objective is then to select a set of these predictors that in some sense specifies a best model (of the general additive form considered in Sections 14.2 in the textbook and Section 14.3 online). Fitting a model that uses a specified $k$ predictor requires that $k + 1$ model coefficients ($\alpha$ and the $k$ corresponding $\beta$'s) be estimated. In general, the number of observations, $n$, should be at least twice the number of predictors in the largest model under consideration to ensure reasonably accurate coefficient estimates and a sufficient number of degrees of freedom associated with $\text{SSResid}$.

If $p$ is not too large, a good statistical computer package can quickly fit a model based on each different subset of the $p$ predictors. Consider the case $p = 4$. There are two possibilities for each predictor—it could be included or not included in a model—so the number of possible models in this case is $2(2)(2)(2) = 16$ (including the model with all four predictors and the model with only the constant term and none of the four predictors). These 16 possibilities are displayed in the following table:

<table>
<thead>
<tr>
<th>Predictors Included</th>
<th>Number of Predictors in Model</th>
<th>Predictors Included</th>
<th>Number of Predictors in Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0</td>
<td>$x_2, x_3$</td>
<td>2</td>
</tr>
<tr>
<td>$x_1$</td>
<td>1</td>
<td>$x_2, x_4$</td>
<td>2</td>
</tr>
<tr>
<td>$x_2$</td>
<td>1</td>
<td>$x_3, x_4$</td>
<td>2</td>
</tr>
<tr>
<td>$x_3$</td>
<td>1</td>
<td>$x_1, x_2, x_3$</td>
<td>3</td>
</tr>
<tr>
<td>$x_4$</td>
<td>1</td>
<td>$x_1, x_2, x_4$</td>
<td>3</td>
</tr>
<tr>
<td>$x_1, x_2$</td>
<td>2</td>
<td>$x_1, x_3, x_4$</td>
<td>3</td>
</tr>
<tr>
<td>$x_1, x_3$</td>
<td>2</td>
<td>$x_2, x_3, x_4$</td>
<td>3</td>
</tr>
<tr>
<td>$x_1, x_4$</td>
<td>2</td>
<td>$x_1, x_2, x_3, x_4$</td>
<td>4</td>
</tr>
</tbody>
</table>

More generally, when there are $p$ candidate predictors, the number of possible models is $2^p$. The number of possible models is therefore substantial if $p$ is even moderately large—for example, 1024 possible models when $p = 10$ and 32,768 possibilities when $p = 15$.

Model selection methods can be divided into two types. First, there are methods based on fitting every possible model, computing one or more summary quantities from each fit, and comparing these quantities to identify the most satisfactory models. Several of the most powerful statistical computer packages have an all subsets option, which gives limited output from fitting each possible model. Methods of the second type are appropriate when $p$ is so large that it is not feasible to examine all subsets. These methods are often referred to as automatic selection or stepwise procedures. The general idea is either to begin with the $p$ predictor model and delete predictors one by one until all remaining predictors are judged important or to begin with no predictors and add predictors until no predictor not in the model seems important. With present-day computing power, the value of $p$ for which examination of all subsets is feasible is surprisingly large, so automatic selection procedures are not as important as they once were.
Suppose, then, that \( p \) is small enough that all subsets can be fit. What characteristics of the estimated models should be examined in the search for a best model? An obvious and appealing candidate is the coefficient of multiple determination, \( R^2 \), which measures the proportion of observed \( y \) variation explained by the model. Certainly a model with a large \( R^2 \) value is preferable to another model that contains the same number of predictors but has a much smaller \( R^2 \) value. Thus, if the model with predictors \( x_1 \) and \( x_2 \) has \( R^2 = .765 \) and the model with predictors \( x_1 \) and \( x_3 \) has \( R^2 = .626 \), the second model would almost surely be eliminated from further consideration.

However, using \( R^2 \) to choose between models containing different numbers of predictors is not so straightforward, because adding a predictor to a model can never decrease the value of \( R^2 \). Let

\[
R^2_{(1)} = \text{largest } R^2 \text{ for any one-predictor model} \\
R^2_{(2)} = \text{largest } R^2 \text{ for any two-predictor model} \\
\vdots
\]

Then \( R^2_{(1)} \leq R^2_{(2)} \leq \cdots \leq R^2_{(p)} \). When statisticians base model selection on \( R^2 \), the objective is not simply to find the model with the largest \( R^2 \) value; the model with \( p \) predictors does that. Instead, we should look for a model that contains relatively few predictors but has a large \( R^2 \) value and is such that no other model containing more predictors gives much of an improvement in \( R^2 \). Suppose, for example, that \( p = 5 \) and that

\[
R^2_{(1)} = .427 \quad R^2_{(2)} = .733 \quad R^2_{(3)} = .885 \quad R^2_{(4)} = .898 \quad R^2_{(5)} = .901
\]

Then the best three-predictor model appears to be a good choice, because it substantially improves on the best one- and two-predictor models, whereas little is gained by using the best four-predictor model or all five predictors.

A small increase in \( R^2 \) resulting from the addition of a predictor to a model can be offset by the increased complexity of the new model and the reduction in the number of degrees of freedom associated with \( \text{SSResid} \). This has led statisticians to consider an adjusted \( R^2 \), which can either decrease or increase when a predictor is added to a model. It follows that the adjusted \( R^2 \) for the best \( k \)-predictor model (i.e., the model with coefficient of multiple determination \( R^2_{(k)} \)) may be larger than the adjusted \( R^2 \) for the best model based on \( k + 1 \) predictors. The adjusted \( R^2 \) formalizes the notion of diminishing returns as more predictors are added: Small increases in \( R^2 \) are outweighed by corresponding decreases in the number of degrees of freedom associated with \( \text{SSResid} \). A reasonable strategy in model selection is to identify the model with the largest value of adjusted \( R^2 \) (the corresponding number of predictors \( k \) is often much smaller than \( p \)) and then consider only that model and any others whose adjusted \( R^2 \) values are nearly as large.

**Example 14.17 Modeling the Price of Industrial Properties**

- The paper “Using Multiple Regression Analysis in Real Estate Appraisal” (*The Appraisal Journal* [2001]: 424–430) reported on a study that aimed to relate the price of a property to various other characteristics of the property. The variables were

\[
\begin{align*}
y & = \text{price per square foot} \\
x_1 & = \text{size of building (square feet)} \\
x_2 & = \text{age of building (years)}
\end{align*}
\]

Step-by-step technology instructions available online  
Data set available online
$x_3 =$ quality of location (measured on a scale of 1 [very poor location] to 4 [very good location])

$x_4 =$ land to building ratio

The study reported the following data for a random sample of nine large industrial properties:

<table>
<thead>
<tr>
<th>$y =$ Price</th>
<th>$x_1 =$ Size</th>
<th>$x_2 =$ Age</th>
<th>$x_3 =$ Location</th>
<th>$x_4 =$ Land/Building Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.89</td>
<td>2,166,600</td>
<td>30</td>
<td>4.0</td>
<td>2.01</td>
</tr>
<tr>
<td>3.49</td>
<td>751,658</td>
<td>30</td>
<td>2.0</td>
<td>3.54</td>
</tr>
<tr>
<td>4.33</td>
<td>2,422,650</td>
<td>28</td>
<td>3.0</td>
<td>3.63</td>
</tr>
<tr>
<td>8.24</td>
<td>224,573</td>
<td>25</td>
<td>1.5</td>
<td>4.65</td>
</tr>
<tr>
<td>5.10</td>
<td>3,917,800</td>
<td>26</td>
<td>4.0</td>
<td>1.71</td>
</tr>
<tr>
<td>2.79</td>
<td>2,866,526</td>
<td>35</td>
<td>4.0</td>
<td>2.27</td>
</tr>
<tr>
<td>5.89</td>
<td>1,698,161</td>
<td>28</td>
<td>3.0</td>
<td>3.12</td>
</tr>
<tr>
<td>6.38</td>
<td>1,046,260</td>
<td>33</td>
<td>4.0</td>
<td>4.77</td>
</tr>
<tr>
<td>5.25</td>
<td>1,108,828</td>
<td>28</td>
<td>4.0</td>
<td>7.56</td>
</tr>
</tbody>
</table>

Consider $x_1$, $x_2$, $x_3$, and $x_4$ as the set of potential predictors ($p = 4$, with no derived predictors, such as squares or interaction terms, as candidates for inclusion). Then there are $2^4 = 16$ possible models, among which 4 consist of a single predictor, 6 involve two predictors, 4 others use three predictors, and 1 includes all four predictor variables. We used a statistical computer package to fit each possible model and extracted both $R^2$ and adjusted $R^2$ from the output. These values are as follows:

**Models with One Predictor**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>32.4</td>
<td>24.7</td>
<td>13.2</td>
<td>11.9</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>22.8</td>
<td>14.0</td>
<td>0.8</td>
<td>0.0</td>
</tr>
</tbody>
</table>

**Models with Two Predictors**

<table>
<thead>
<tr>
<th>Predictors</th>
<th>$x_1$, $x_2$</th>
<th>$x_2$, $x_4$</th>
<th>$x_2$, $x_3$</th>
<th>$x_1$, $x_3$</th>
<th>$x_1$, $x_4$</th>
<th>$x_3$, $x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>52.3</td>
<td>40.2</td>
<td>33.3</td>
<td>25.0</td>
<td>24.9</td>
<td>22.8</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>36.3</td>
<td>20.2</td>
<td>11.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

**Models with Three Predictors**

<table>
<thead>
<tr>
<th>Predictors</th>
<th>$x_1$, $x_2$, $x_3$</th>
<th>$x_1$, $x_2$, $x_4$</th>
<th>$x_2$, $x_3$, $x_4$</th>
<th>$x_1$, $x_3$, $x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>58.6</td>
<td>52.3</td>
<td>40.9</td>
<td>25.7</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>33.7</td>
<td>23.7</td>
<td>5.4</td>
<td>0.0</td>
</tr>
</tbody>
</table>

**Models with All Four Predictors**

<table>
<thead>
<tr>
<th>Predictors</th>
<th>$x_1$, $x_2$, $x_3$, $x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>65.3</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>30.6</td>
</tr>
</tbody>
</table>
It is clear that the best two-predictor model offers considerable improvement with respect to both $R^2$ (52.3%) and adjusted $R^2$ (36.3%) over any model with just a single predictor. The best two-predictor model also has the largest value of adjusted $R^2$—even larger than adjusted $R^2$ for the three-predictor models and the model that uses all four predictors.

This suggests selecting the model that uses $x_1$ and $x_2$ (building size and age) as predictors of price. The MINITAB output resulting from fitting this model is given in Figure 14.14. Note that, although this model may be the best choice among those considered here, the $R^2$ value is not particularly large and the value of $s_e = 1.283$ (in dollars per square foot) is quite large given the range of $y$ values in the data set.

![Figure 14.14](image-url) MINITAB output for the data of Example 14.17.

Various other criteria have been proposed and used for model selection after fitting all subsets. The references by Neter, Wasserman, and Kutner or Kerlinger and Pedhazur can be consulted for more details.

When using particular criteria as a basis for model selection, many of the $2^p$ possible subset models are not serious candidates because of poor criteria values. For example, if $p = 15$, there are 3005 different models consisting of 6 predictor variables, many of which typically have small $R^2$ and adjusted $R^2$ values. An investigator usually wishes to consider only a few of the best models of each different size (a model whose criteria value is close to the best one may be easier to interpret than the best model or may include a predictor that the investigator thinks should be in the selected model). In recent years, statisticians have developed computer programs to select the best models of a given size without actually fitting all possible models. One version of such a program has been implemented in MINITAB and can be used as long as $p \leq 20$. (There are roughly 1 million possible models when $p = 20$, so fitting them all would be out of the question.) The user specifies a number between 1 and 10 as the number of models of each given size for which the output will be provided. In addition to $R^2$ and adjusted $R^2$, values of another criterion, called Mallow’s $C_p$, are included. A good model according to this criterion is one that has small $C_p$ (for accurate predictions) and for which $C_p \approx k + 1$ (for unbiasedness in estimating model coefficients). After the choice of models is narrowed, the analyst can request more detail on each finalist.
Example 14.18 Durable Press Rating of Cotton Fabric

In the article “Applying Stepwise Multiple Regression Analysis to the Reaction of Formaldehyde with Cotton Cellulose” (Textile Research Journal [1984]: 157–165), the investigators looked at the dependent variable

\[ y = \text{durable press rating} \]

which is a quantitative measure of wrinkle resistance. The four independent variables used in the model building process were

\[ x_1 = \text{formaldehyde concentration} \]
\[ x_2 = \text{catalyst ratio} \]
\[ x_3 = \text{curing temperature} \]
\[ x_4 = \text{curing time} \]

In addition to these variables, the investigators considered as potential predictors \( x_1^2, x_2^2, x_3^2, x_4^2 \) and all six interactions \( x_1x_2, \ldots, x_3x_4 \), a total of \( p = 14 \) candidates.

<table>
<thead>
<tr>
<th>Observation</th>
<th>( y )</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
<th>Observation</th>
<th>( y )</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.4</td>
<td>8</td>
<td>4</td>
<td>100</td>
<td>1</td>
<td>16</td>
<td>4.6</td>
<td>4</td>
<td>10</td>
<td>160</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>2.2</td>
<td>2</td>
<td>4</td>
<td>180</td>
<td>7</td>
<td>17</td>
<td>4.3</td>
<td>4</td>
<td>10</td>
<td>100</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>4.6</td>
<td>7</td>
<td>4</td>
<td>180</td>
<td>1</td>
<td>18</td>
<td>4.9</td>
<td>10</td>
<td>10</td>
<td>120</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>4.9</td>
<td>10</td>
<td>7</td>
<td>120</td>
<td>5</td>
<td>19</td>
<td>1.7</td>
<td>5</td>
<td>4</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>4.6</td>
<td>7</td>
<td>4</td>
<td>180</td>
<td>5</td>
<td>20</td>
<td>4.6</td>
<td>8</td>
<td>13</td>
<td>140</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>4.7</td>
<td>7</td>
<td>7</td>
<td>180</td>
<td>1</td>
<td>21</td>
<td>2.6</td>
<td>10</td>
<td>1</td>
<td>180</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>4.6</td>
<td>7</td>
<td>13</td>
<td>140</td>
<td>1</td>
<td>22</td>
<td>3.1</td>
<td>2</td>
<td>13</td>
<td>140</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>4.5</td>
<td>5</td>
<td>4</td>
<td>160</td>
<td>7</td>
<td>23</td>
<td>4.7</td>
<td>6</td>
<td>13</td>
<td>180</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>4.8</td>
<td>4</td>
<td>7</td>
<td>140</td>
<td>3</td>
<td>24</td>
<td>2.5</td>
<td>7</td>
<td>1</td>
<td>120</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>1.4</td>
<td>5</td>
<td>1</td>
<td>100</td>
<td>7</td>
<td>25</td>
<td>4.5</td>
<td>5</td>
<td>13</td>
<td>140</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>4.7</td>
<td>8</td>
<td>10</td>
<td>140</td>
<td>3</td>
<td>26</td>
<td>2.1</td>
<td>8</td>
<td>1</td>
<td>160</td>
<td>7</td>
</tr>
<tr>
<td>12</td>
<td>1.6</td>
<td>2</td>
<td>4</td>
<td>100</td>
<td>3</td>
<td>27</td>
<td>1.8</td>
<td>4</td>
<td>1</td>
<td>180</td>
<td>7</td>
</tr>
<tr>
<td>13</td>
<td>4.5</td>
<td>4</td>
<td>10</td>
<td>180</td>
<td>3</td>
<td>28</td>
<td>1.5</td>
<td>6</td>
<td>1</td>
<td>160</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>4.7</td>
<td>6</td>
<td>7</td>
<td>120</td>
<td>7</td>
<td>29</td>
<td>1.3</td>
<td>4</td>
<td>1</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>4.8</td>
<td>10</td>
<td>13</td>
<td>180</td>
<td>3</td>
<td>30</td>
<td>4.6</td>
<td>7</td>
<td>10</td>
<td>100</td>
<td>7</td>
</tr>
</tbody>
</table>

Figure 14.15 displays output for the best three subset models of each size from \( k = 1 \) to \( k = 9 \) predictor variables and also for the model with all 14 predictors.

The choice of a best model here, as often happens, is not clear-cut. We certainly don’t see the benefit of including more than \( k = 8 \) predictor variables (after that, the adjusted \( R^2 \) begins to decrease), nor would we suggest a model with fewer than five predictors (the adjusted \( R^2 \) is still increasing and \( C_p \) is large). Based on this output, the best six-predictor model is a reasonable choice. The corresponding estimated regression function is

\[
y = -1.218 + 0.9599x_2 - 0.0373x_1^2 - 0.0373x_2^2 + 0.0037x_1x_3 + 0.019x_1x_4 - 0.0013x_2x_3
\]

Another good candidate is the best seven-predictor model. Although this model includes one more predictor than the six-predictor model just suggested, only one of
the seven predictors is an interaction term \((x_1x_3)\), so model interpretation is somewhat easier. (Notice, though, that none of the best three models with seven predictors results simply from adding a single predictor to the best six-predictor model.) Because every good model includes \(x_1, x_2, x_3, \text{ and } x_4\) in some predictor, it appears that formaldehyde concentration, catalyst ratio, curing time, and curing temperature are all important determinants of durable press rating.

The most easily understood and implemented automatic selection procedure is referred to as **backward elimination**. It involves starting with the model that contains all \(p\) potential predictors and then deleting them one by one until all remaining predictors seem important. The first step is to specify the value of a positive constant \(t_{out}\), which is used to decide whether deletion should be continued. After fitting the \(p\) predictor model, the \(t\) ratios \(\frac{b_1}{s_{b_1}}, \frac{b_2}{s_{b_2}}, \ldots, \frac{b_p}{s_{b_p}}\) are examined. The predictor variable whose
Chapter 14  Multiple Regression Analysis

The \( t \) ratio is closest to zero, whether positive or negative, is the obvious candidate for deletion. The corresponding predictor is eliminated from the model if this \( t \) ratio satisfies the inequalities \(-t_{out} \leq t \leq t_{out}\). Suppose that this is the case. The model with the remaining \( p-1 \) predictors is then fit, and again the predictor with the \( t \) ratio closest to zero is eliminated, provided that it satisfies \(-t_{out} \leq t \leq t_{out}\). The procedure continues until, at some stage, no \( t \) ratio satisfies \(-t_{out} \leq t \leq t_{out}\) (all are either greater than \( t_{out} \) or less than \(-t_{out}\)). The chosen model is then the last one fit (though some analysts recommend examining other models of the same size). It is customary to use \( t_{out} = 2 \), since for many different values of \( df \), this corresponds to a two-tailed test with approximate significance level .05.*

Example 14.19  More on Price of Industrial Properties

Figure 14.16 shows MINITAB output resulting from the application of the backward elimination procedure with \( t_{out} = 2 \) to the industrial property price data of Example 14.17. The \( t \) ratio closest to 0 when the model with all four predictors was fit was \(-0.88\), so the corresponding predictor \( x_4 \) was deleted from the model. When the model with the three remaining predictors was fit, the \( t \) ratio closest to 0 was \(0.87\), which satisfies \(-2 \leq 0.87 \leq 2\). As a consequence, the corresponding predictor, \( x_3 \), was eliminated, leaving \( x_1 \) and \( x_2 \). The next predictor to be dropped was \( x_1 \), because its \( t \) ratio was \(-1.58\), and \(-2 \leq -1.58 \leq 2\). When the model with just \( x_2 \) = age was fit, it was also eliminated! This illustrates that automatic selection procedures don’t always work well.

### Figure 14.16  MINITAB output for Example 14.19.

<table>
<thead>
<tr>
<th>Step</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>18.428</td>
<td>14.757</td>
<td>14.022</td>
<td>13.533</td>
<td>5.151</td>
</tr>
<tr>
<td>Size</td>
<td>-0.00000</td>
<td>-0.00000</td>
<td>-0.00000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-Value</td>
<td>-1.68</td>
<td>-1.75</td>
<td>-1.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-Value</td>
<td>0.169</td>
<td>0.141</td>
<td>0.166</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.43</td>
<td>-0.34</td>
<td>-0.27</td>
<td>-0.29</td>
<td></td>
</tr>
<tr>
<td>T-Value</td>
<td>-2.14</td>
<td>-2.01</td>
<td>-1.86</td>
<td>-1.83</td>
<td></td>
</tr>
<tr>
<td>P-Value</td>
<td>0.100</td>
<td>0.100</td>
<td>0.112</td>
<td>0.109</td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>1.19</td>
<td>0.61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-Value</td>
<td>1.22</td>
<td>0.87</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-Value</td>
<td>0.288</td>
<td>0.422</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Land/Bui</td>
<td>-0.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-Value</td>
<td>-0.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-Value</td>
<td>0.429</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>1.34</td>
<td>1.31</td>
<td>1.28</td>
<td>1.41</td>
<td>1.61</td>
</tr>
<tr>
<td>R-Sq</td>
<td>65.29</td>
<td>58.58</td>
<td>52.26</td>
<td>32.43</td>
<td>-0.00</td>
</tr>
<tr>
<td>R-Sq(adj)</td>
<td>30.58</td>
<td>33.72</td>
<td>36.34</td>
<td>22.78</td>
<td>0.00</td>
</tr>
</tbody>
</table>

*Some computer packages base the procedure on the squares of the \( t \) ratios, which are \( F \) ratios, and continue to delete as long as \( F \) ratio \( \leq F_{out} \) for at least one predictor. The predictor with the smallest \( F \) ratio is eliminated. \( F_{out} = 4 \) corresponds to \( t_{out} = 2 \).
Unfortunately, the backward elimination method does not always terminate with a model that is the best of its size, and this is also true of other automatic selection procedures. For example, the authors of the article mentioned in Example 14.18 used an automatic procedure to obtain a six-predictor model with $R^2 = .77$, whereas all of the 10 best six predictor models have $R^2$ values of at least .87. Because of this, we recommend using a statistical software package that will identify best subsets of different sizes whenever possible.

- **Checks on Model Adequacy**

  In Chapter 13 we discussed some informal techniques for checking the adequacy of the simple linear regression model. Most of these were based on plots involving the standardized residuals. The formula for standardizing residuals in multiple regression is quite complicated, but it has been programmed into many statistical computer packages. Once the standardized residuals resulting from the fit of a particular model have been computed, plots similar to those discussed previously are useful in diagnosing model defects. A normal probability plot of the standardized residuals that departs too much from a straight line casts doubt on the assumption that the random deviation $e$ has a normal distribution. Plots of the standardized residuals against each predictor variable in the model—that is, a plot of $(x_1, \text{standardized residual})$ pairs, another of $(x_2, \text{standardized residual})$ pairs, and so on—are analogous to the standardized residual versus $x$ plot discussed and illustrated in Chapter 13. The appearance of any discernible pattern in these plots (e.g., curvature or increasing spread from left to right) points to the need for model modification. If observations have been made over time, a periodic pattern when the standardized residuals are plotted in time order suggests that successive observations were not independent. Models that incorporate dependence of successive observations are substantially more complicated than those we have presented here. They are especially important in econometrics, which involves using statistical methods to model economic data. Please consult the reference by Neter, Wasserman, and Kutner for more information.

  One other aspect of model adequacy that has received much attention from statisticians in recent years is the identification of any observations in the data set that may have been highly influential in estimating model coefficients. Recall that in simple linear regression, an observation with potentially high influence is one whose $x$ value places it far to the right or left of the other points in the scatterplot or standardized residual plot. If a multiple regression model involves only two predictor variables $x_1$ and $x_2$, an observation with potentially large influence can be revealed by examining a plot of $(x_1, x_2)$ pairs. Any point in this plot that is far away from the others corresponds to an observation that, if deleted from the sample, may cause coefficient estimates and other quantities to change considerably. Detecting influential observations when the model contains three predictors or more is more difficult. Recent research has yielded several helpful diagnostic quantities. One of these has been implemented in MINITAB, and a large value of this quantity automatically results in the corresponding observation being identified as one that might have great influence. Deleting the observation and refitting the model reveals the extent of actual influence (which depends on how consistent the corresponding $y$ observation is with the rest of the data).

- **Multicollinearity**

  When the simple linear regression model is fit using sample data in which the $x$ values are all close to one another, small changes in observed $y$ values can cause the values of the estimated coefficients $a$ and $b$ to change considerably. This may well result
in standard deviations \( \sigma_y \) and \( \sigma_x \) that are quite large, so that estimates of \( \beta \) and \( \alpha \) from any given sample are likely to differ greatly from the true values. There is an analogous condition that leads to this same type of behavior in multiple regression: a configuration of predictor variable values that is likely to result in poorly estimated model coefficients.

When the model to be fit includes \( k \) predictors \( x_1, x_2, \ldots, x_k \), there is said to be **multicollinearity** if there is a strong linear relationship between values of the predictors. Severe multicollinearity leads to instability of estimated coefficients and to various other problems. Such a relationship is difficult to visualize when \( k \gg 10 \), so statisticians have developed various quantitative indicators to measure the extent of multicollinearity in a data set. The most straightforward approach involves computing \( R^2 \) values for regressions in which the dependent variable is taken to be one of the \( k \)’s and the predictors are the remaining \(( k - 1)\)’s. For example, when \( k = 3 \), there are three relevant regressions:

1. Dependent variable = \( x_1 \), predictor variables = \( x_2 \) and \( x_3 \)
2. Dependent variable = \( x_2 \), predictor variables = \( x_1 \) and \( x_3 \)
3. Dependent variable = \( x_3 \), predictor variables = \( x_1 \) and \( x_2 \)

Each regression yields an \( R^2 \) value. In general, there are \( k \) such regressions and therefore \( k \) resulting \( R^2 \) values. If one or more of these \( R^2 \) values is large (close to 1), multicollinearity is present. MINITAB prints a message saying that the predictors are highly correlated when at least one of these \( R^2 \) values exceeds .99 and refuses to include a predictor in the model if the corresponding \( R^2 \) value is larger than .9999. Other analysts are more conservative and would judge multicollinearity to be a potential problem if any of these \( R^2 \) values exceeded .9.

When the values of predictor variables are under the control of the investigator, which often happens in scientific experimentation, a careful choice of values precludes multicollinearity. Multicollinearity does frequently occur in social science and business applications of regression analysis, where data result simply from observation rather than from intervention by an experimenter. Statisticians have proposed various remedies for the problems associated with multicollinearity in such situations, but a discussion would take us beyond the scope of this book. (After all, we want to leave something for your next statistics course!)

### Exercises 14.49–14.60

**14.49** The article “The Caseload Controversy and the Study of Criminal Courts” (*Journal of Criminal Law and Criminology* [1979]: 89–101) used multiple regression to analyze a data set consisting of observations on the variables

- \( y \) = length of sentence in trial case (months)
- \( x_1 \) = seriousness of first offense
- \( x_2 \) = dummy variable indicating type of trial (bench or jury)
- \( x_3 \) = number of legal motions
- \( x_4 \) = measure of delay for confined defendants (0 for those not confined)
- \( x_5 \) = measure of judge’s caseload

The estimated regression equation proposed by the authors is

\[
\hat{y} = 12.6 + .59x_1 - 70.8x_2 - 33.6x_3 - 15.5x_4 \\
+ .0007x_5 + 3x_7 - 41.5x_8 
\]

where \( x_6 = x_1^2, x_7 = x_1x_2, \) and \( x_8 = x_3x_5. \) How do you think the authors might have arrived at this particular model?
14.50 The article “Histologic Estimation of Age at Death Using the Anterior Cortex of the Femur” (American Journal of Physical Anthropology [1991]: 171–179) developed multiple regression models to relate $y = \text{age at death}$ to a number of different predictor variables. Stepwise regression (an automatic selection procedure favored over the backward elimination method by many statisticians) was used to identify models shown in the accompanying table, for the case of females.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_6$</td>
<td>.60</td>
<td>.60</td>
</tr>
<tr>
<td>$x_3, x_6$</td>
<td>.66</td>
<td>.66</td>
</tr>
<tr>
<td>$x_3, x_5$</td>
<td>.68</td>
<td>.68</td>
</tr>
<tr>
<td>$x_2, x_3, x_5, x_6$</td>
<td>.69</td>
<td>.68</td>
</tr>
<tr>
<td>$x_2, x_3, x_5, x_6$</td>
<td>.70</td>
<td>.69</td>
</tr>
<tr>
<td>$x_2, x_3, x_5, x_6$</td>
<td>.71</td>
<td>.70</td>
</tr>
<tr>
<td>$x_1, x_2, x_3, x_5, x_6$</td>
<td>.71</td>
<td>.70</td>
</tr>
<tr>
<td>$x_1, x_2, x_3, x_5, x_6$</td>
<td>.71</td>
<td>.71</td>
</tr>
<tr>
<td>$x_1, x_2, x_3, x_5, x_6$</td>
<td>.72</td>
<td>.71</td>
</tr>
<tr>
<td>$x_1, x_2, x_3, x_5, x_6$</td>
<td>.72</td>
<td>.70</td>
</tr>
</tbody>
</table>

Based on the given information, which of these models would you recommend using, and why?

14.51 The article “The Analysis and Selection of Variables in Linear Regression” (Biometrics [1976]: 1–49) reports on an analysis of data taken from issues of Motor Trend magazine. The dependent variable $y$ was gas mileage, there were $n = 32$ observations, and the independent variables were $x_1 = \text{engine type (1 = straight, 0 = V)}$, $x_2 = \text{number of cylinders}$, $x_3 = \text{transmission type (1 = manual, 0 = automatic)}$, $x_4 = \text{number of transmission speeds}$, $x_5 = \text{engine size}$, $x_6 = \text{horsepower}$, $x_7 = \text{number of carburetor barrels}$, $x_8 = \text{final drive ratio}$, $x_9 = \text{weight}$, and $x_{10} = \text{quarter-mile time}$. The $R^2$ and adjusted $R^2$ values are given in the accompanying table for the best model using $k$ predictors for $k = 1, \ldots, 10$. Which model would you select? Explain your choice and the criteria used to reach your decision.

<table>
<thead>
<tr>
<th>$k$</th>
<th>Variables Included</th>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$x_9$</td>
<td>.756</td>
<td>.748</td>
</tr>
<tr>
<td>2</td>
<td>$x_2, x_9$</td>
<td>.833</td>
<td>.821</td>
</tr>
<tr>
<td>3</td>
<td>$x_3, x_9, x_{10}$</td>
<td>.852</td>
<td>.836</td>
</tr>
<tr>
<td>4</td>
<td>$x_3, x_6, x_{10}$</td>
<td>.860</td>
<td>.839</td>
</tr>
<tr>
<td>5</td>
<td>$x_3, x_5, x_6, x_{10}$</td>
<td>.866</td>
<td>.840</td>
</tr>
</tbody>
</table>

14.52 The article “Estimation of the Economic Threshold of Infestation for Cotton Aphid” (Mesopotamia Journal of Agriculture [1982]: 71–75) gave $n = 34$ observations on $y = \text{infestation rate (number of aphids per 100 leaves)}$, $x_1 = \text{mean temperature (°C)}$, $x_2 = \text{mean relative humidity}$

Partial SAS computer output resulting from fitting the model $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e$ to the data given in the article is shown.

- $R^2$ = 0.5008
- Adjusted $R^2$ = 0.4533

<table>
<thead>
<tr>
<th>PARAMETER</th>
<th>STANDARD</th>
<th>T FOR HO:</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEPT</td>
<td>15.667014</td>
<td>84.157420</td>
</tr>
<tr>
<td>TEMP</td>
<td>-0.360928</td>
<td>2.330835</td>
</tr>
<tr>
<td>RH</td>
<td>1.959715</td>
<td>0.576946</td>
</tr>
</tbody>
</table>

If the method of backward elimination is to be employed, which variable would be the first candidate for elimination? Using the criterion of eliminating a variable if its $t$ ratio satisfies $-2 \leq t \leq 2$, can it be eliminated from the model?

14.53 The following statement is from the article “Blood Cadmium Levels in Nonexposed Male Subjects Living in the Rome Area: Relationship to Selected Cardiovascular Risk Factors” (Microchemical Journal [1998]: 173–179):

A multiple regression analyzed ln(blood cadmium level) as a dependent variable, with a model including age, alcohol consumption, daily cigarette consumption, driving habits, body mass index, skinfold thickness, high density lipoprotein cholesterol, non-high density lipoprotein cholesterol, and ln(triglyceride level) as possible predictors. This model explained 30.92% of the total variance in ln(blood cadmium level).

- A total of $n = 1856$ men participated in this study.
- Perform a model utility test to determine whether the
Chapter 14  Multiple Regression Analysis

predictor variables, taken as a group are useful in predicting ln(blood cadmium level).

b. The article went on to state:

Smoking habits, the most important predictor of ln(blood cadmium level), explained 30.26% of the total variation, and alcohol consumption was the only other variable influencing ln(blood cadmium level).

These statements were based on the result of a forward stepwise regression. In forward stepwise regression, variables are added to the model one at a time. At each step, the variable that results in the largest improvement is added to the model. This continues until there is no variable whose addition will result in a significant improvement. Describe what you think happened at each step of the stepwise procedure.

14.54 For the multiple regression model in Exercise 14.4, the estimated standard deviations for each of the coefficients (excluding the constant) are (in order) .01, .01, .03, .06, .01, .01, .03, .04, and .05.

a. If a backward elimination variable selection process were to be used, which of the nine predictors would be the first candidate for elimination? Would it be eliminated?

b. Based on the information given, can you tell what the second candidate for elimination would be? Explain.

14.55 The following table showing the results of two multiple regressions appears in “Does the Composition of the Compensation Committee Influence CEO Compensation Practices?” (Financial Management [1999]: 41–53). The dependent variable is log(compensation).

<table>
<thead>
<tr>
<th>Dependent Variables</th>
<th>1991</th>
<th>1992</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interception</td>
<td>2.508</td>
<td>2.098</td>
</tr>
<tr>
<td>Stock ownership</td>
<td>−0.005</td>
<td>−0.005</td>
</tr>
<tr>
<td>Log of sales</td>
<td>0.372</td>
<td>0.005</td>
</tr>
<tr>
<td>CEO tenure</td>
<td>0.020</td>
<td>0.006</td>
</tr>
<tr>
<td>Stock returns</td>
<td>0.001</td>
<td>0.006</td>
</tr>
<tr>
<td>Return on equity</td>
<td>0.061</td>
<td>0.001</td>
</tr>
<tr>
<td>Insider</td>
<td>0.580</td>
<td>2.163</td>
</tr>
</tbody>
</table>

d. Would the same variable identified in Part (c) be the first to be considered in a backward elimination process for the 1992 regression?

e. Assuming the number of observations is 160, determine the P-value for a test of the hypothesis that $b_1$ in the 1991 model is negative.

14.56 Suppose you were considering a multiple regression analysis with $y =$ house price, $x_1 =$ number of bedrooms, $x_2 =$ number of bathrooms, and $x_3 =$ total number of rooms. Do you think that multicollinearity might be a problem? Explain.

14.57 This exercise requires use of a computer package. Using a statistical computer package, compare the best one-, two-, three-, and four-predictor models for the data given in Exercise 14.58. Does this variable-selection procedure lead you to the same choice of model as the backward elimination used in Exercise 14.58?

14.58 The accompanying data are from the article “Breeding Success of the Common Puffin on Different Habitats at Great Island, Newfoundland” (Ecology Monographs [1972]: 246–252). The variables considered are $y =$ nesting frequency (burrows per 9 m$^2$), $x_1 =$ grass cover (%), $x_2 =$ mean soil depth (cm), $x_3 =$ angle of slope (degrees), and $x_4 =$ distance from cliff edge (m).
MINITAB output resulting from application of the backward elimination procedure is also given. Explain what action was taken at each step and why.

<table>
<thead>
<tr>
<th>STEP</th>
<th>CONSTANT</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>T-RATIO</th>
<th>R-SQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.29</td>
<td>-0.019</td>
<td>-0.043</td>
<td>0.224</td>
<td>-0.182</td>
<td>-1.42</td>
<td>86.20</td>
</tr>
<tr>
<td>2</td>
<td>11.45</td>
<td>2.91</td>
<td>2.94</td>
<td>2.91</td>
<td>2.12</td>
<td>2.55</td>
<td>85.99</td>
</tr>
<tr>
<td>3</td>
<td>13.96</td>
<td>-0.43</td>
<td>-0.043</td>
<td>0.225</td>
<td>-0.203</td>
<td>-0.293</td>
<td>85.16</td>
</tr>
</tbody>
</table>

This exercise requires use of a computer package. The accompanying $n = 25$ observations on $y =$ catch at intake (number of fish), $x_1 =$ water temperature ($^\circ$C), $x_2 =$ minimum tide height (m), $x_3 =$ number of pumps running, $x_4 =$ speed (knots), and $x_5 =$ wind-range of direction (degrees) constitute a subset of the data that appeared in the article “Multiple Regression Analysis for Forecasting Critical Fish Influxes at Power Station Intakes” (Journal of Applied Ecology [1983]: 33–42). Use the variable selection procedures discussed in this section to formulate a model.
14.5 Interpreting and Communicating the Results of Statistical Analyses

Multiple regression is a powerful tool for analyzing multivariate data sets. Unfortunately, the results of such an analysis are often reported in a compact way, and it is often difficult to extract the relevant information from the summary.

What to Look For in Published Data

Here are some things to consider when evaluating a study that has used data to fit a multiple regression model:

- What variables are being used as predictors? Are any of them categorical? If so, have they been incorporated in an appropriate way (using indicator variables)?
- Has a model utility test been performed? Does the conclusion of this test indicate that the reported model is useful?
- Are the values of $R^2$ and adjusted $R^2$ reported? Is there a large difference between these two values? If so, this may indicate that the number of predictors used in the model is excessive, given the number of observations in the data set.
- Has a variable selection procedure been used? If so, what procedure was used? Which variables were selected for inclusion in the final model?

The following example illustrates how the results of a multiple regression analysis might be described. Factors influencing teacher referrals to special education programs were examined in the paper “Stress, Biases, or Professionalism: What Drives Teachers’ Referral Judgments of Students with Challenging Behaviors” (Journal of Emotional and Behavioral Disorders [2002]: 204–212). Teachers were asked to review a student record and to rate on a scale of 1 to 10 the likelihood that they would refer the student for a psycho-educational assessment (1 = very likely; 10 = very unlikely). The following variables were considered in a multiple regression analysis relating teacher likelihood of referral to various student characteristics:

$$y = \text{teacher likelihood of referral}$$
$$x_1 = \text{a measure of off-task behavior}$$
$$x_2 = \text{a measure of problem behavior}$$
$$x_3 = \text{a measure of academic competence}$$

The paper reported that “the final model included three predictor variables (i.e., off-task behavior, problem behavior, and academic competence) that jointly accounted for 51% of the variance in likelihood of student referral.” Other variables, including demographic characteristics (such as gender, race, and age) of both students and teachers, were also considered as potential predictors of $y = \text{likelihood of referral}$, but they were eliminated from consideration using a variable selection procedure described in the paper.

14.60 Suppose that $R^2 = .723$ for the model containing predictors $x_1, x_4, x_5, \text{ and } x_8$ and $R^2 = .689$ for the model with predictors $x_1, x_3, x_5 \text{ and } x_6$.

a. What can you say about $R^2$ for the model containing predictors $x_1, x_4, x_5, x_8, \text{ and } x_9$? Explain.

b. What can you say about $R^2$ for the model containing predictors $x_1$ and $x_4$? Explain.
### Additional Key Concepts and Formulas

<table>
<thead>
<tr>
<th>Term or Formula</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_i \pm (t \text{ critical value})s_{b_i}$</td>
<td>A confidence interval for the regression coefficient $\beta_i$ (based on $n - (k + 1) \text{ df}$).</td>
</tr>
<tr>
<td>$t = \frac{b_i - \text{ hypothesized value}}{s_{b_i}}$</td>
<td>The test statistic for testing hypotheses about $\beta_i$ (based on $n - (k + 1) \text{ df}$). The most important case is $H_0: \beta_i = 0$, according to which the predictor is not useful as long as all other predictors remain in the model.</td>
</tr>
<tr>
<td>$\hat{y} \pm (t \text{ critical value})s_{\hat{y}}$</td>
<td>A confidence interval for a mean $y$ value and a prediction interval for a single $y$ value when $x_1, x_2, \ldots, x_k$ have specified values, where $\hat{y} = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$ and $s_{\hat{y}}$ denotes the estimated standard deviation of $\hat{y}$.</td>
</tr>
</tbody>
</table>

### Model selection

An investigator may have data on many predictors that could be included in a model. There are two different approaches to choosing a model: (1) use an all-subsets procedure to identify the best models of each different size (one-predictor models, two-predictor models, etc.) and then compare these according to criteria such as $R^2$ and adjusted $R^2$; and (2) use an automatic selection procedure, which either successively eliminates predictors until a stopping point is reached (backward elimination) or starts with no predictors and adds them successively until the inclusion of additional predictors cannot be justified.

### Chapter Review Exercises 14.61–14.71

**ThomsonNOW**: Know exactly what to study! Take a pre-test and receive your Personalized Learning Plan.

**14.61** What factors affect the quality of education in our schools? The article “How Much Is Enough? Applying Regression to a School Finance Case” (Statistics and the Law, [New York: Wiley, 1986]: 257–287) used data from $n = 88$ school districts to carry out a regression analysis based on the following variables:

- $y$ = average language score for fourth-grade students

**Peer group variables:**

- $x_3 =$ % Title I enrollment
- $x_4 =$ Hispanic enrollment (= 1 if $> 5\%$ and 0 otherwise)
- $x_5 =$ logarithm of fourth-grade enrollment
- $x_6 =$ prior test score (from 4 years previous to year under study)

**Background variables:**

- $x_1 =$ occupational index (% managerial and professional workers in the community)
- $x_2 =$ median income

**School variables:**

- $x_7 =$ administrator-teacher ratio
- $x_8 =$ pupil-teacher ratio
- $x_9 =$ certified staff-pupil ratio

**Bold exercises answered in back**

- Data set available online but not required
- Video solution available
Chapter 14  ■ Multiple Regression Analysis

\[ x_{10}, x_{11} = \text{indicator variables for average teaching experience} \]  
\[ x_{10} = 1 \text{ if less than 3 years and 0 otherwise, } x_{11} = 1 \text{ if more than 6 years and 0 otherwise} \]

The accompanying table gives estimated coefficients, values of \( t \) ratios (\( b_1/s_{b_1}, b_2/s_{b_2}, \) and so on), and other relevant information.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimated Coefficient</th>
<th>( t ) Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>70.67</td>
<td>6.08</td>
</tr>
<tr>
<td>( x_1 )</td>
<td>0.458</td>
<td>3.08</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>-0.000141</td>
<td>-0.59</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>-58.47</td>
<td>-4.02</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>-3.668</td>
<td>-1.73</td>
</tr>
<tr>
<td>( x_5 )</td>
<td>-2.942</td>
<td>-2.34</td>
</tr>
<tr>
<td>( x_6 )</td>
<td>0.200</td>
<td>3.60</td>
</tr>
<tr>
<td>( x_7 )</td>
<td>17.93</td>
<td>0.42</td>
</tr>
<tr>
<td>( x_8 )</td>
<td>0.689</td>
<td>0.42</td>
</tr>
<tr>
<td>( x_9 )</td>
<td>-0.403</td>
<td>-0.27</td>
</tr>
<tr>
<td>( x_{10} )</td>
<td>-0.614</td>
<td>-1.30</td>
</tr>
<tr>
<td>( x_{11} )</td>
<td>-0.382</td>
<td>-0.58</td>
</tr>
<tr>
<td>( s_6 = 5.57 )</td>
<td>( R^2 = .64 )</td>
<td></td>
</tr>
</tbody>
</table>

a. Is there a useful linear relationship between \( y \) and at least one of the predictors? Test the relevant hypotheses using a .01 significance level.

b. What is the value of adjusted \( R^2 \)?

c. Calculate and interpret a 95% confidence interval for \( \beta_1 \).

d. Does it appear that one or more predictors could be eliminated from the model without discarding useful information? If so, which one would you eliminate first? Explain your reasoning.

e. Suppose you wanted to decide whether any of the peer group variables provided useful information about \( y \) (provided that all other predictors remained in the model). What hypotheses would you test (a single \( H_0 \) and \( H_a ) \)? Could one of the test procedures presented in this chapter be used? Explain.

14.63  ■ The accompanying data on \( y = \) glucose concentration (g/L) and \( x = \) fermentation time (days) for a particular blend of malt liquor were read from a scatterplot in the article “Improving Fermentation Productivity with Reverse Osmosis” (Food Technology [1984]: 92–96):

\[
\begin{align*}
1 & \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \\
74 & \quad 54 \quad 52 \quad 51 \quad 52 \quad 53 \quad 58 \quad 71
\end{align*}
\]

a. Construct a scatterplot for these data. Based on the scatterplot, what type of model would you suggest?

b. MINITAB output resulting from fitting a multiple regression model with \( x_1 = x \) and \( x_2 = x^2 \) is shown here. Does this quadratic model specify a useful relationship between \( y \) and \( x \)?

The regression equation is

\[ Y = 84.5 - 15.9X_1 + 1.77X_2 \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>( t )-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>84.482</td>
<td>4.904</td>
<td>17.23</td>
</tr>
<tr>
<td>( X_1 )</td>
<td>-15.875</td>
<td>2.500</td>
<td>-6.35</td>
</tr>
<tr>
<td>( X_2 )</td>
<td>1.7679</td>
<td>0.2712</td>
<td>6.52</td>
</tr>
<tr>
<td>( s = 3.515 )</td>
<td>R-sq = 89.5%</td>
<td>R-sq(adj) = 85.3%</td>
<td></td>
</tr>
</tbody>
</table>

c. Could the quadratic term have been eliminated? That is, would a simple linear model have sufficed? Test using a .05 significance level.

14.64  Much interest in management circles has focused on how employee compensation is related to various company characteristics. The article “Determinants of R and D Compensation Strategies” (Personnel Psychology [1984]: 635–650) proposed a qualitative scale for \( y = \) base salary for employees of high-tech companies. The following estimated multiple regression equation was then presented:

\[ \hat{y} = 2.60 + .125x_1 + .893x_2 + .057x_3 - .014x_4 \]

where \( x_1 = \) sales volume (in millions of dollars), \( x_2 = \) stage in product life cycles (1 = growth, 0 = mature), \( x_3 = \) profitability (%), and \( x_4 = \) attrition rate (%).
a. There were \( n = 33 \) firms in the sample and \( R^2 = .69 \). Is the fitted model useful?
b. Predict base compensation for a growth stage firm with sales volume \( \$50 \) million, profitability 8\%, and attrition rate 12\%.
c. The estimated standard deviations for the coefficient estimates were .064, .141, .014, and .005 for \( b_1, b_2, b_3, \) and \( b_4 \) respectively. Should any of the predictors be deleted from the model? Explain.
d. \( b_2 \) is the difference between average base compensation for growth stage and mature stage firms when all other predictors are held fixed. Use the information in Part (c) to calculate a 95\% confidence interval for \( b_2 \).

14.65 If \( n = 21 \) and \( k = 10 \), for what values of \( R^2 \) would adjusted \( R^2 \) be negative?

14.66 A study of total body electrical conductivity was described in the article “Measurement of Total Body Electrical Conductivity: A New Method for Estimation of Body Composition” (American Journal of Clinical Nutrition [1983]: 735–739). Nineteen observations were given for the variables \( y = \) total body electrical conductivity, \( x_1 = \) age (years), \( x_2 = \) sex (0 = male, 1 = female), \( x_3 = \) body mass (kg/m\(^2\)), \( x_4 = \) body fat (kg), and \( x_5 = \) lean body mass (kg).

a. The backward elimination method of variable selection was employed, and Minitab output is given here. Explain what occurred at each step in the process.

**STEPWISE REGRESSION OF Y ON 5 PREDICTORS, WITH N = 19**

**STEP 1 2 3**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stddev</th>
<th>t-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>0.31</td>
<td>0.36</td>
<td>0.38</td>
</tr>
<tr>
<td>T-RATIO</td>
<td>1.90</td>
<td>2.36</td>
<td>2.96</td>
</tr>
<tr>
<td>X2</td>
<td>-7.9</td>
<td>-7.5</td>
<td>-7.0</td>
</tr>
<tr>
<td>T-RATIO</td>
<td>-1.93</td>
<td>-1.89</td>
<td>-2.04</td>
</tr>
<tr>
<td>X3</td>
<td>-0.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-RATIO</td>
<td>-0.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X4</td>
<td>0.22</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>T-RATIO</td>
<td>0.78</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>X5</td>
<td>0.365</td>
<td>0.339</td>
<td>0.378</td>
</tr>
<tr>
<td>T-RATIO</td>
<td>2.16</td>
<td>2.09</td>
<td>4.18</td>
</tr>
<tr>
<td>S</td>
<td>5.16</td>
<td>5.07</td>
<td>4.91</td>
</tr>
<tr>
<td>R-SQ</td>
<td>79.53</td>
<td>78.70</td>
<td>78.57</td>
</tr>
</tbody>
</table>

b. Minitab output for the multiple regression model relating \( y \) to \( x_1, x_2, \) and \( x_5 \) is shown here. Interpret the values of \( R^2 \) and \( s_e \).

The regression equation is
\[
y = -15.2 + 0.377X1 - 6.99X2 + 0.378X5
\]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stddev</th>
<th>t-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>-15.175</td>
<td>9.620</td>
<td>-1.58</td>
</tr>
<tr>
<td>X1</td>
<td>0.377</td>
<td>0.1273</td>
<td>2.96</td>
</tr>
<tr>
<td>X2</td>
<td>-6.988</td>
<td>3.425</td>
<td>-2.04</td>
</tr>
<tr>
<td>X5</td>
<td>0.37779</td>
<td>0.09047</td>
<td>4.18</td>
</tr>
<tr>
<td>s</td>
<td>4.914</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-sq</td>
<td>78.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-sq(adj)</td>
<td>74.3%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

c. Interpret the value of \( b_2 \) in the estimated regression equation.

d. If the estimated standard deviation of the statistic \( a + b_1(31) + b_2(1) + b_5(52.7) \) is 1.42, give a 95\% confidence interval for the true mean total body electrical conductivity of all 31-year-old females whose lean body mass is 52.7 kg.

14.67 The article “Creep and Fatigue Characteristics of Ferrocement Slabs” (Journal of Ferrocement [1984]: 309–322) reported data on \( y = \) tensile strength (MPa), \( x_1 = \) slab thickness (cm), \( x_2 = \) load (kg), \( x_3 = \) age at loading (days), and \( x_4 = \) time under test (days) resulting from stress tests of \( n = 9 \) reinforced concrete slabs. The backward elimination method of variable selection was applied. Partial Minitab output follows. Explain what action was taken at each step in the process. Minitab output for the selected model is also given at the top of the page 14-30. Use the estimated regression equation to predict tensile strength for a slab that is 25 cm thick, 150 days old, and is subjected to a load of 200 kg for 50 days.

**STEPWISE REGRESSION OF Y ON 5 PREDICTORS, WITH N = 9**

**STEP 1 2 3**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stddev</th>
<th>t-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>8.496</td>
<td>12.670</td>
<td>12.989</td>
</tr>
<tr>
<td>X1</td>
<td>-0.29</td>
<td>-0.42</td>
<td>-0.49</td>
</tr>
<tr>
<td>T-RATIO</td>
<td>-1.33</td>
<td>-2.89</td>
<td>-3.14</td>
</tr>
<tr>
<td>X2</td>
<td>0.0104</td>
<td>0.0110</td>
<td>0.0116</td>
</tr>
<tr>
<td>T-RATIO</td>
<td>6.30</td>
<td>7.40</td>
<td>7.33</td>
</tr>
<tr>
<td>X3</td>
<td>0.0059</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-RATIO</td>
<td>0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X4</td>
<td>-0.023</td>
<td>-0.023</td>
<td></td>
</tr>
<tr>
<td>T-RATIO</td>
<td>-1.48</td>
<td>-1.53</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>5.33</td>
<td>5.16</td>
<td>5.70</td>
</tr>
<tr>
<td>R-SQ</td>
<td>95.81</td>
<td>95.10</td>
<td>92.82</td>
</tr>
</tbody>
</table>
14.68 A study of pregnant grey seals involved \( n = 25 \) observations on the variables \( y = \) fetus progesterone level (mg), \( x_1 = \) fetus sex (0 = male, 1 = female), \( x_2 = \) fetus length (cm), and \( x_3 = \) fetus weight (g). Minitab output for the model using all three independent variables is given (“Gonadotrophin and Progesterone Concentration in Placenta of Grey Seals,” *Journal of Reproduction and Fertility* [1984]: 521–528).

The regression equation is

\[
Y = -1.98 - 1.87X_1 + 0.234X_2 + 0.0001X_3
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

Would you recommend keeping both \( x_1 \) and \( x_2 \) in the model? Explain.

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = -2.09 - 1.87X_1 + 0.240X_2
\]

The regression equation is

\[
Y = -2.09 - 1.87X_1 + 0.231X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]

The regression equation is

\[
Y = 13.0 - 0.487X_1 + 0.116X_2
\]
### Chapter Review Exercises 14-31

#### a. Fit a multiple regression model using both independent variables.

#### b. Use the $F$ test to determine whether the model provides useful information for predicting profit margin.

#### c. Interpret the values of $R^2$ and $s_e$.

#### d. Would a regression model using a single independent variable ($x_1$ alone or $x_2$ alone) have sufficed? Explain.

#### e. Plot the $(x_1, x_2)$ pairs. Does the plot indicate any sample observation that may have been highly influential in estimating the model coefficients? Explain. Do you see any evidence of multicollinearity? Explain.

---

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$y$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.92</td>
<td>7298</td>
<td>.75</td>
<td>3.78</td>
<td>6672</td>
<td>.84</td>
</tr>
<tr>
<td>3.61</td>
<td>6855</td>
<td>.71</td>
<td>3.82</td>
<td>6890</td>
<td>.79</td>
</tr>
<tr>
<td>3.32</td>
<td>6636</td>
<td>.66</td>
<td>3.97</td>
<td>7115</td>
<td>.70</td>
</tr>
<tr>
<td>3.07</td>
<td>6506</td>
<td>.61</td>
<td>4.07</td>
<td>7327</td>
<td>.68</td>
</tr>
<tr>
<td>3.06</td>
<td>6450</td>
<td>.70</td>
<td>4.25</td>
<td>7546</td>
<td>.72</td>
</tr>
<tr>
<td>3.11</td>
<td>6402</td>
<td>.72</td>
<td>4.41</td>
<td>7931</td>
<td>.55</td>
</tr>
<tr>
<td>3.21</td>
<td>6368</td>
<td>.77</td>
<td>4.49</td>
<td>8097</td>
<td>.63</td>
</tr>
<tr>
<td>3.26</td>
<td>6340</td>
<td>.74</td>
<td>4.70</td>
<td>8468</td>
<td>.56</td>
</tr>
<tr>
<td>3.42</td>
<td>6349</td>
<td>.90</td>
<td>4.58</td>
<td>8717</td>
<td>.41</td>
</tr>
<tr>
<td>3.42</td>
<td>6352</td>
<td>.82</td>
<td>4.69</td>
<td>8991</td>
<td>.51</td>
</tr>
<tr>
<td>3.45</td>
<td>6361</td>
<td>.75</td>
<td>4.71</td>
<td>9179</td>
<td>.47</td>
</tr>
<tr>
<td>3.58</td>
<td>6369</td>
<td>.77</td>
<td>4.78</td>
<td>9318</td>
<td>.32</td>
</tr>
<tr>
<td>3.66</td>
<td>6546</td>
<td>.78</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Bold** exercises answered in back | ● Data set available online but not required | ▼ Video solution available
Answers to Selected Odd-Numbered Exercises

Chapter 14

14.35 a. \((-0.697, -0.281)\). With 95% confidence, the change in the mean value of a vacant lot associated with a one unit increase in distance from the city’s major east-west thoroughfare is a decrease of as little as 0.281 or as much as 0.697 assuming all other predictors remain constant.
b. \(r = -0.599, P\text{-value} = 0.550\), fail to reject \(H_0\)
14.37 a. \(F = 144.36, P\text{-value} < 0.001\), reject \(H_0\) and conclude that the model is useful.
b. \(t = 4.33, P\text{-value} = 0,\) reject \(H_0\), and conclude that the quadratic term is important.
c. (72.06, 72.46)

14.39 a. The value 0.469 is an estimate of the expected change (increase) in the mean score of students associated with a one unit increase in the student’s expected score assuming time spent studying and student’s grade point average are fixed.
b. \(F = 75.01, P\text{-value} < 0.001\), reject \(H_0\) and conclude that the model is useful.
c. (2.46, 4.27)
d. 72.856 e. (61.56, 84.15)
14.41 \(t = 2.22, P\text{-value} = 0.028\), reject \(H_0\) and conclude that the interaction term is important.

14.43 a. \(F = 5.47, 0.01 < P\text{-value} < 0.05\), reject \(H_0\) and conclude that the model is useful.
b. \(t = 0.40, P\text{-value} = 0.924\), fail to reject \(H_0\), and conclude that the interaction term is not needed in the model.
c. No. The model utility test is testing all variables as a group. The \(t\) test is testing the contribution of an individual predictor when used in the presence of the remaining predictors.

14.45 a. \(t = 3.71, P\text{-value} = .002\), reject \(H_0\), length cannot be eliminated; \(t = -0.12, P\text{-value} = 0.904\), fail to reject \(H_0\), age could be eliminated.
b. \(t = -4.49, P\text{-value} = .001\), reject \(H_0\), year is a useful predictor.

14.47 a. \(F = 30.81, P\text{-value} < 0.001\), reject \(H_0\), and conclude that the model is useful.
b. For \(\beta_1: t = 6.57, P\text{-value} = 0,\) reject \(H_0\). For \(\beta_2: t = -7.69, P\text{-value} = 0,\) reject \(H_0\).
c. (43.0, 47.92) d. (44.0, 47.92)
14.49 One possible way would have been to start with the set of predictor variables consisting of all five variables, along with all quadratic terms and all interaction terms. Then, use a selection procedure like backward elimination to arrive at the given estimated regression equation.

14.51 The model using the three variables \(x_3, x_9, x_{10}\) appears to be a good choice.

14.53 a. \(F = 91.8, P\text{-value} < 0.001\), reject \(H_0\), and conclude that the model is useful.
b. The predictor added at the first step was smoking habits. At step 2, alcohol consumption was added. No other variables were added to the model.

14.55 a. Yes b. 0.0567 c. Return on equity, yes d. No. CEO tenure would be considered first.

14.57 The best model, using the procedure of minimizing \(C_p\) would use variables \(x_1, x_4\). These are not the same predictor variables as selected in problem 14.58.

14.61 a. \(F = 12.28, P\text{-value} < 0.001\), reject \(H_0\), and conclude that the model is useful.
b. Adjusted \(R^2 = 0.5879\ c. (0.1606, 0.7554)\)
d. \(x_6\) would be considered first, and it would be eliminated.
e. \(H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0\). None of the procedures presented in this chapter could be used.

14.63 a. A quadratic model b. \(F = 21.25, 0.001 < P\text{-value} < 0.01\), reject \(H_0\), and conclude that the model is useful.
c. \(t = 6.52, P\text{-value} = 0,\) reject \(H_0\), the quadratic term should not be eliminated.

14.65 Adjusted \(R^2\) will be negative for values of \(R^2\) less than 0.5.

14.67 First, the model using all four variables was fit. The variable age at loading \((x_3)\) was deleted because it had the \(t\)-ratio closest to zero and it was between \(-2\) and \(2\). Then, the model using the three variables \(x_1, x_2,\) and \(x_4\) was fit. The variable time \((x_4)\) was deleted because its \(t\)-ratio was closest to zero and was between \(-2\) and \(2\). Finally, the model using the two variables \(x_1\) and \(x_2\) was fit. Neither of these variables could be eliminated since their \(t\)-ratios were greater than \(2\) in absolute magnitude.

The final model, then, includes slab thickness \((x_1)\) and load \((x_2)\). The predicted tensile strength for a slab that is 25 cm thick, 150 days old, and is subjected to a load of 200 kg for 50 days is \(\hat{y} = 13 - 0.487(25) + 0.019(200) = 3.145\).

14.69 b. The claim is reasonable because 14 is close to where the curve has its maximum value.

14.71 a. \(\hat{y} = 1.56 + 0.0237x_1 - 0.000249x_2\ b. F = 70.67, P\text{-value} = 0.000,\) reject \(H_0\), and conclude that the model is useful.
c. 86.5% of the total variation in the observed values for profit margin has been explained by the fitted regression equation. A typical deviation of an observed value from the predicted value is 0.0533. d. No. Both variables have associated \(t\)-ratios that exceed 2 in absolute magnitude and so neither can be eliminated from the model. e. There do not appear to be any influential observations. However, there is substantial evidence of multicollinearity. The plot shows a pronounced linear relationship between \(x_1\) and \(x_2\). This is evidence of multicollinearity between \(x_1\) and \(x_2\).