Click on the desired section below to position to that page. You may also scroll to browse the document.

Supplemental Topic 3: Multiple Regression

- S3.1 The Multiple Linear Regression Model
- S3.2 Inference for Multiple Regression Models
- S3.3 Checking Conditions for Multiple Linear Regression
- Exercises
Supplemental Topic 3: Multiple Regression

A regression equation is used to describe the relationship between one or more explanatory variables ($x$ variables) and a quantitative response variable ($y$). In Chapters 5 and 14 we learned about simple regression in which a straight-line equation is used to describe the relationship between a single quantitative explanatory variable ($x$) and a quantitative response variable ($y$). Remember that the $y$ variable is sometimes called the dependent variable, and because regression models may be used to make predictions, an $x$ variable may also be called a predictor variable.

In multiple regression, we use the values of more than one explanatory variable to predict or describe the values of a response variable. Put another way, the equation for the mean of the response variable ($y$) is a function of two or more explanatory variables. As an example, suppose a researcher is studying the connection between behavioral variables and grade point average for college students. Predictor variables of interest to the researcher are:

- $x_1 = \text{Study hours} = \text{Hours spent studying per week}$
- $x_2 = \text{Classes missed} = \text{Number of classes student misses in typical week}$
- $x_3 = \text{Work hours} = \text{Hours per week that student works in a part or full time job}$

The general form of a linear multiple regression model relating grade point average (GPA) to these three predictor variables is:

$$\text{GPA} = \beta_0 + \beta_1 \text{Study hours} + \beta_2 \text{Classes missed} + \beta_3 \text{Work hours}$$

Numerical estimates of the parameters $\beta_0, \beta_1, \beta_2,$ and $\beta_3$ would be determined using data from a sample of students for whom information on grade point average and the three predictor variables is available.
THE MULTIPLE LINEAR REGRESSION MODEL

S3.1

Suppose \( y \) is the response variable, and \( x_1, x_2, \ldots, x_p \) is a set of \( p - 1 \) explanatory variables. The multiple linear regression equation for the relationship between the mean value of \( y \) and the variables \( x_1, x_2, \ldots, x_p \) is

\[
E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1}
\]

- \( E(Y) \) represents the mean value of \( y \) for individuals in the population who all have the same particular values of \( x_1, x_2, \ldots, x_{p-1} \).
- \( \beta_0, \beta_1, \ldots, \beta_{p-1} \) are referred to as the regression coefficients, or, alternatively, may either be called the \( \beta \) coefficients or the \( \beta \) parameters.

As in simple regression, a useful format for expressing the components of the overall population regression model is

\[
y = \text{Mean} + \text{Deviation}
\]

This conceptual equation states that for any individual, the value of the response variable (\( y \)) can be constructed by combining two components:

1. The mean, which is given by \( E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} \), is the mean \( y \) for individuals having the same specific set of \( x \) values as this individual.
2. An individual’s deviation = \( y - \text{mean} \), which is the difference between an individual’s actual value of \( y \) and the mean \( y \) for individuals having the same specific set of \( x \)-values as this individual.

When writing the regression model for an individual within the population, it’s necessary to use additional notation in order to distinguish one individual from another. Let \( x_{i1} \) represent the value of \( x_1 \) for the \( i \)th individual and let \( y_i \) represent the value of the \( i \)th observation for variable \( y \). The multiple linear regression model is

\[
y_i = \text{Mean} + \text{Deviation} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i
\]

The deviations \( \epsilon_i \) are assumed to follow a normal distribution with mean 0 and standard deviation \( \sigma \). In multiple regression, we make the same assumptions about the deviations that we did in simple regression, and in fact you should notice that the multiple regression model with only one \( x \) variable gives the simple regression model.

EXAMPLE S3.1

Predicting Average August Temperature

The temperature dataset on the CD for this text gives geographic latitude, mean January temperature, mean April temperature, and mean August temperature for 20 cities in the United States. The temperatures are recorded as degrees Fahrenheit. A multiple linear regression equation for the relationship between August temperature and the explanatory variables latitude, January temperature, and April temperature is

\[
E(\text{August temp}) = \beta_0 + \beta_1 \text{Latitude} + \beta_2 \text{January temp} + \beta_3 \text{April temp}
\]

Written more formally using statistical notation, the multiple regression model in this situation is

\[
y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i
\]
where the $\epsilon_i$ are assumed to follow a normal distribution with mean 0 and standard deviation $\sigma$.

We will see shortly that we may not need to use all three explanatory variables in the equation because they are predictable from each other. Data for the 20 cities in the temperature dataset can be used to estimate the $\beta$ coefficients. An important first step when analyzing the data is to examine scatterplots for all pairs of variables in order to determine if any influential outliers are present and whether any of the relationships between the response variable and an explanatory variable might be nonlinear.

---

**EXAMPLE S3.2**

**Blood Pressure of Peruvian Indians**

How would somebody’s blood pressure be affected, in the long run, if they migrate from living in a small village at a high altitude to living in an urban area at a low altitude? Anthropologists examined this question by studying $n = 39$ Peruvian Indian men who had made such a migration. The men in the sample ranged in age from 21 to 54 years old, and the number of years since migration to the urban area ranged from 2 to 40 years. In addition to information on age and years since migrating, the investigators collected data on systolic and diastolic blood pressure, age, body weight, and various skinfold measurements related to amount of body fat. The data are given by Ryan and Joiner (2001) and also are in the Minitab worksheet PERU.MTW (not on the text CD). The researchers also created and analyzed an additional variable not in the dataset, which was the percent of a man’s lifetime that he had lived in the urban area.

In this example, we will use a multiple linear regression model in which the response variable is $y =$ systolic blood pressure and the explanatory variables are

- $x_1 =$ Weight (kg)
- $x_2 =$ Age
- $x_3 =$ Years = Number of years since migrating
- $x_4 =$ Pct_Life = (Years/Age) × 100 = Percent of lifetime living in urban area

Although the principal focus of the study was on how blood pressure was related to the number of years since living at altitude, the explanatory variables weight and age are included because it is well known that blood pressure is related to these variables. Expressed somewhat loosely (in terms of notation), the model is

$$\text{Blood pressure} = \beta_0 + \beta_1 \text{Weight} + \beta_2 \text{Age} + \beta_3 \text{Years} + \beta_4 \text{Pct}_\text{Life}$$

Written using statistical notation, the multiple regression model in this situation is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$$

where $\epsilon_i$ are assumed to follow a normal distribution with mean 0 and standard deviation $\sigma$.

---

**The Regression Equation for the Sample**

The first step in a multiple regression problem is to estimate the $\beta$ coefficients using sample data. Notation for the sample coefficients is $b_0, b_1, \ldots, b_{p-1}$. With this notation, the regression equation for the sample could be written as

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_{p-1} x_{p-1}$$
For the $i$th observation in the sample,

- The **predicted value** (estimated mean) is $\tilde{y}_i = b_0 + b_1x_{i1} + b_2x_{i2} + \cdots + b_{p-1}x_{i,p-1}$
- The **residual** (deviation or error) is $e_i = y_i - \tilde{y}_i$

where $y_i$ = the value of the response variable for the $i$th observation in the sample and $x_{i1}, x_{i2}, \ldots, x_{i,p-1}$ are the corresponding values of variables $x_1, x_2, \ldots, x_{p-1}$.

Numerical values for the sample regression coefficients $b_0, b_1, \ldots, b_{p-1}$ usually are found by applying the least-squares criterion described in Chapter 5. The values of $b_0, b_1, \ldots, b_{p-1}$ determined by this criterion give a smaller value of the sum of squared errors for the sample than would any other possibilities for the values $b_0, b_1, \ldots, b_{p-1}$. Put another way, the value of $\text{SSE} = \sum(y_i - \tilde{y}_i)^2$ is minimized. In practice, statistical software is always used to estimate the coefficients of a regression equation, so we don’t give formulas for calculating the sample regression coefficients “by hand.”

**EXAMPLE S3.1—CONTINUED**

Sample Regression Equation for Predicting August Temperature

For the **temperature** dataset, separate plots of $y = \text{August temperature}$ versus each of $x_1 = \text{geographic latitude}$, $x_2 = \text{January temperature}$, and $x_3 = \text{April temperature}$ show linear associations with no obviously influential observations. Thus, we can use the multiple linear regression model. With statistical software, it can be determined that the sample regression equation is

$$\tilde{y} = -29 + 0.70x_1 - 0.44x_2 + 1.66x_3$$

Two oddities are evident when we consider the values of the sample regression coefficients, and they illustrate how difficult it is to interpret the values of the regression coefficients in a multiple regression. Notice that the coefficient multiplying $x_1 = \text{latitude}$ is positive, indicating that August temperature goes up as latitude increases, which does not seem reasonable. The value of the coefficient multiplying $x_2 = \text{January temperature}$ is negative, indicating that the warmer it is in January the colder it is in August. Again, this is not reasonable. Regardless of month, average temperature tends to decrease as latitude increases in the Northern hemisphere. For instance, in this sample, the correlation between latitude and August temperature is $r = -0.781$, indicating a strong negative correlation. Thus we might expect that the regression coefficient multiplying latitude would be negative, but instead it’s positive ($+0.70$). Similarly, we would expect a positive value for the regression coefficient multiplying $x_2 = \text{January temperature}$ because, in general, relatively warm locations in January tend also to be relatively warm places in August. For our sample, the correlation between January and August temperatures is $+0.622$, so the sign of the corresponding regression coefficient ($-0.44$) is in contrast to the direction of the association.

The difficulty is caused by the degree of correlation among the explanatory variables, which is quite high in this example. Specifically, the correlation between $x_1 = \text{latitude}$ and $x_2 = \text{January temperature}$ is $r = -0.86$, the correlation between latitude and $x_3 = \text{April temperature}$ is $r = -0.96$, and the correlation between $x_2 = \text{January temperature}$ and $x_3 = \text{April temperature}$ is $r = -0.91$.

With such strong associations among the explanatory variables, it becomes nearly impossible to separate their individual effects on the response variable. Each coefficient actually represents the *additional* contribution made by that
variable, given that all of the others are already being used as predictors. The moral of the story is that caution must be employed when interpreting the individual regression coefficients in a multiple regression analysis.

The sample regression can be used to determine a predicted value of August temperature, even when it’s difficult to interpret the sample regression coefficients. As an example, the city of Phoenix has latitude = 33, January temperature = 54, and April temperature = 70. The predicted value of August temperature for Phoenix is found as

$$\hat{y} = -29 + 0.70(33) - 0.44(54) + 1.66(70) = 86.5 \text{ degrees}$$

Phoenix is hotter in August than we would predict based on information about its geographic latitude and January and April temperatures. The actual average August temperature in Phoenix is 92 degrees Fahrenheit. The residual for Phoenix is $e = y - \hat{y} = \text{actual} - \text{predicted} = 92 - 86.5 = 5.1 \text{ degrees}$. ◆

Using statistical software (Minitab, for instance), it can be found that the sample regression equation for the Peruvian Indian blood example is

$$\hat{y} = 116.84 + 0.832 \text{ Weight} - 0.951 \text{ Age} + 2.34 \text{ Years} - 1.08 \text{ Pct_Life}$$

where $\hat{y} =$ predicted systolic blood pressure based on the four explanatory variables. The value of $\hat{y}$ is the estimated mean systolic blood pressure for a population of individuals with the same particular set of values for the explanatory variables, and it can also be used as the predicted value of systolic blood pressure for an individual with a specified set of $x$ values.

In multiple regression, it may be difficult to interpret the values of the sample regression coefficients; that most definitely is the case in this example. In fact, the values of the individual coefficients should not be interpreted here. Notice, for instance, that the coefficient multiplying Age in the equation is equal to $-0.951$. Does blood pressure actually decrease as age increases? A naïve interpretation would be that average systolic blood pressure decreases 0.951 points per each one-year increase in age, given that all other variables in the equation are held constant (not allowed to change). But, it’s not possible to change the value of Age while holding Years and Pct_Life constant. The variable Pct_Life is a function of the variable Age; it must necessarily change if Age is increased while Years is held constant.

The following table gives the values of $\hat{y}$, found using the sample regression equation, for four different sets of specific values for explanatory variables:

<table>
<thead>
<tr>
<th>Set</th>
<th>$x_1$ Weight (kg)</th>
<th>$x_2$ Age</th>
<th>$x_3$ Years</th>
<th>$x_4$ Pct_Life</th>
<th>$\hat{y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>65</td>
<td>30</td>
<td>15</td>
<td>50</td>
<td>123.5</td>
</tr>
<tr>
<td>2</td>
<td>65</td>
<td>50</td>
<td>15</td>
<td>30</td>
<td>126.1</td>
</tr>
<tr>
<td>3</td>
<td>65</td>
<td>50</td>
<td>25</td>
<td>50</td>
<td>127.9</td>
</tr>
<tr>
<td>4</td>
<td>65</td>
<td>50</td>
<td>35</td>
<td>70</td>
<td>129.6</td>
</tr>
</tbody>
</table>

Notice that the weight is 65 kg for all four sets, and that age equals 30 in the first set and equals 50 in the other three sets. For sets 2, 3, and 4, in which age is 50, the predicted values of systolic blood pressure are greater than the predicted value for set 1 where age is 30. This is consistent with the usual medical finding that blood pressure, on average, increases with age. By comparing the predicted
values for sets 2, 3, and 4 we see that for men of the same age and weight, increasing the number of years since migration apparently increases average systolic blood pressure.

Minitab Tip

**tech note** Estimating the Sample Regression Equation

- Use Stat>Regression>Regression.
- In the box labeled “Response” specify the worksheet column with the data for the response variable.
- In the box labeled “Predictors” specify the worksheet columns containing the data for the explanatory variables.

**Estimating the Standard Deviation $\sigma$**

Remember from Section 14.2 that the standard deviation in the regression model measures, roughly, the average deviation of $y$-values from the mean (the regression equation). Expressed another way, the **standard deviation for regression measures the general size of the residuals**. In Section 14.2 we gave a formula for calculating an estimate of the standard deviation for a simple regression model and discussed how to interpret its value. The interpretation is essentially the same in multiple regression as it is in simple regression, and the calculation for multiple regression involves only a minor (but important) modification to the procedure for simple regression.

**Estimating the Standard Deviation for a Multiple Regression Model**

Let $p$ = the number of $\beta$ coefficients in the multiple regression model and $n$ = sample size. The formula for estimating the standard deviation for a multiple regression model is

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n - p}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n - p}}$$

The statistic $s = \sqrt{\text{MSE}}$ is an estimate of the population standard deviation $\sigma$.

**The Proportion of Variation Explained by the Explanatory Variables ($R^2$)**

The statistic denoted by $R^2$ measures what researchers often term the “proportion of variation in the response explained by the predictor variables (the $x$ variables).” This statistic may be called the **multiple squared correlation** or, synonymously, the **multiple coefficient of determination**. It is often expressed as a percent, and, if so, its value is between 0 and 100%. A value of 100% would indicate that the $x$ variables completely explain the observed variation in responses for $y$. A value of 0% would indicate that the explanatory variables explain absolutely nothing about the variation in the responses for $y$. The procedure described in Section 5.3 for simple regression also applies for multiple regression.
Values for $s$ and $R^2$ are given in the multiple regression output of nearly all statistical software packages. For the prediction of August temperature, the value of $R^2 = 0.871$, or 87.1%, and the value of the standard deviation is $s = 2.81$. The value of $R^2$ is high, indicating a generally strong linear relationship between August temperature and the three predictor variables. The standard deviation gives, roughly, the average absolute size of the deviations (residuals) between actual and predicted values in the dataset. On average, actual August temperatures might be off by about $2.81$ degrees Fahrenheit from predicted August temperatures.

In the Peruvian Indian blood pressure example, the value of $R^2 = 0.597$, or 59.7%, and the value of the standard deviation is $s = 8.795$. This is a moderately strong value of $R^2$, indicating that the four explanatory variables explain some, but not all, of the variation in individual systolic blood pressures. The interpretation of the standard deviation is that, roughly, the average absolute size of the deviations (residuals) between actual and predicted systolic blood pressure is about 8.795.

Suppose there is a random sample of $n$ observations of the explanatory variables $x_1, x_2, \ldots, x_p$ and the response variable $y$ from a large population. Let $x_{i,k}$ represent the value of the $i$th observation for variable $x_k$, and $y_i$ represent the value of the $i$th observation for variable $y$. The multiple linear regression model follows:

**Population Version**

Mean: $E(Y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_p x_{p-1}$

Individual: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{i,p-1} + \epsilon_i = E(Y) + \epsilon_i$
S3.2 INFERENCE ABOUT MULTIPLE REGRESSION MODELS

In this section, we’ll discuss four commonly used inference procedures in multiple regression.

1. A hypothesis test of the null hypothesis that an individual population regression coefficient \( b_k \) is equal to 0.
2. A hypothesis test of the null hypothesis that \( b_1 \neq 0, b_2 \neq 0, \ldots, b_p \neq 0 \), which is equivalent to stating that none of the \( x \) variables are related to \( y \).
3. A prediction interval for the value of \( y \) for an individual with a specified set of \( x \)-values.
4. A confidence interval for the mean \( y \) for a population of individuals, all with the same particular set of values for the \( x \)-variables.

### Inference for a Coefficient \( \beta_k \)

The multiple regression output of nearly every statistical software program includes results for testing the null hypothesis \( H_0: \beta_k = 0 \) versus the alternative hypothesis \( H_a: \beta_k \neq 0 \) for each separate \( \beta_k \) in the model. Notice that if \( \beta_k = 0 \) for a \( \beta \) coefficient that multiplies an explanatory variable, then the variable \( x_k \) is not influencing the calculation of predicted \( y \)-values because \( \beta_k x_k = 0 \), no matter what value \( x_k \) might have. Thus, these tests are used to make decisions about which explanatory variables may or may not be helpful predictors of the response variable. Great care must be taken however, when interpreting these tests as they are affected by the magnitude of the correlations between the \( x \) variables. We will discuss this issue in an example later in this section.

#### Testing Null and Alternative Hypotheses for a Regression Coefficient

In the multiple regression model, a test of \( H_0: \beta_k = 0 \) versus \( H_a: \beta_k \neq 0 \) is done using the test statistic

\[
 t = \frac{b_k - 0}{\text{s.e.}(b_k)}
\]

where \( b_k \) = value of the sample estimate of \( \beta_k \) and \( \text{s.e.}(b_k) \) is the standard error of \( b_k \).
If a confidence interval for any of the $b_k$ is desired, it is computed as $b_k \pm t^*\text{s.e.}(b_k)$. The multiplier $t^*$ is found using a $t$-distribution with $n - p$ degrees of freedom, and is such that the probability between $-t^*$ and $+t^*$ equals the confidence level for the interval. Table A.2 can be used to find the value of $t^*$.

**Inference for the Overall Model**

Suppose that each of the parameters $b_1, b_2, \ldots, b_p$ in the multiple regression model were equal to 0. If that is so, then the regression equation would reduce to $E(Y) = \beta_0$. This implies that the response variable $y$ is in fact unrelated to any of the explanatory variables being studied. Although this occurs rarely in practice, it must still be considered as a possibility. Thus, all statistical software provides results for a test of the null hypothesis $H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$.

The interpretation of the null hypothesis is that if it were true, then none of the explanatory variables are useful predictors of the values of $y$. The interpretation of the alternative hypothesis is that at least one of the explanatory variables is related to the response variable.

The test statistic in this situation is called an $F$-statistic. In computer output for multiple regression, it is usually given in a table called an analysis of variance table. The general format of the analysis of variance table for multiple regression follows:

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of Freedom</th>
<th>Sum of Squares</th>
<th>Mean Square $= \text{SS/df}$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression model</td>
<td>$p - 1$</td>
<td>$\text{SSR} = \text{SSTO} - \text{SSE} = \Sigma(\hat{y}_i - \bar{y})^2$</td>
<td>$\text{MSR} = \frac{\text{SSR}}{p - 1}$</td>
<td>$\text{MSR} / \text{MSE}$</td>
</tr>
<tr>
<td>Residual error</td>
<td>$n - p$</td>
<td>$\text{SSE} = \Sigma(y_i - \hat{y}_i)^2$</td>
<td>$\text{MSE} = \frac{\text{SSE}}{n - p}$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$n - 1$</td>
<td>$\text{SSTO} = \Sigma(y_i - \bar{y})^2$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The $p$-value is found using an $F$-distribution. An $F$-distribution is a skewed distribution, with a minimum value of 0, and it has two parameters called degrees of freedom. The first of the two parameters is called the numerator degrees of freedom, while the second is called the denominator degrees of freedom. In multiple regression, the numerator df = $p - 1$ = number of $x$ variables in the equation, and the denominator df = $n - p$ (sample size – number of groups). Thus, the $p$-value could be found as the probability that an $F$-distribution with degrees of freedom $p - 1$ and $n - p$ would have a value larger than the observed $F$-statistic. Table A.4 provides specific numerical values for $F$-distributions, corresponding to the upper 5% and 1% tails of the distributions.
Figure S3.1 displays Minitab output for the multiple linear regression model in which $y$ is August temperature and the three $x$ variables are geographic latitude, January temperature, and April temperature. At the far right of the analysis of variance table given toward the bottom of the output, a $p$-value is given for a test of $H_0: \beta_1 = \beta_2 = \beta_3 = 0$. Its value is 0.000 (bold, underlined), so the null hypothesis is rejected, which means that at least one of the explanatory variables in the model is linearly related to August temperature (no surprise).

Toward the top of the output there is a display of five columns with information about the individual regression coefficients. Remember that in the multiple regression model, a test of $H_0: \beta_k = 0$ versus $H_a: \beta_k \neq 0$ is done using the test statistic

$$t = \frac{b_k - 0}{s.e.(b_k)} = \frac{b_k}{s.e.(b_k)}$$

The $t$-statistic for each of the coefficients can be seen in the column headed “T,” and is formed by taking the ratio of the values in the columns headed “Coef” and “SE Coef.” For instance, for testing $H_0: \beta_1 = 0$, the test statistic is $t = 0.7009/0.4053 = 1.73$.

The $p$-values corresponding to these tests are given in the last column (under “P”) and are used for testing the null hypothesis $\beta_k = 0$, separately for each coefficient in the model. We focus on the $p$-values in the rows labeled with the names of the explanatory variables. In order starting with the latitude row, these are $p$-values for testing $\beta_1 = 0$, $\beta_2 = 0$, and $\beta_3 = 0$. The tests for each of $\beta_2 = 0$ and $\beta_3 = 0$ (last two rows) are statistically significant because they have the $p$-values 0.001 and 0.000, respectively. On the other hand, the $p$-value (in the “Latitude” row) for testing $\beta_1 = 0$ is 0.103, which exceeds the usual 0.05 level of significance, so the test is not statistically significant. In fact, there is a strong connection between latitude and August temperature, but January and April temperatures, which themselves are related to latitude, provide sufficient information to pre-

The regression equation is

$$\text{AugTemp} = -29.0 + 0.701 \text{ latitude} - 0.440 \text{ JanTemp} + 1.66 \text{ AprTemp}$$

<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>-28.99</td>
<td>29.42</td>
<td>-0.99</td>
<td>0.339</td>
</tr>
<tr>
<td>Latitude</td>
<td>0.7009</td>
<td>0.4053</td>
<td>1.73</td>
<td>0.103</td>
</tr>
<tr>
<td>JanTemp</td>
<td>-0.4397</td>
<td>0.1036</td>
<td>-4.24</td>
<td>0.001</td>
</tr>
<tr>
<td>AprTemp</td>
<td>1.6649</td>
<td>0.2964</td>
<td>5.62</td>
<td>0.000</td>
</tr>
</tbody>
</table>

$S = 2.814$  \hspace{1cm} R-Sq = 87.1\%  \hspace{1cm} R-Sq(adj) = 84.7\%$

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>3</td>
<td>858.07</td>
<td>286.02</td>
<td>36.11</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>16</td>
<td>126.73</td>
<td>7.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>984.80</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
dict August temperature. The consequence is that we could consider removing the latitude variable from the model because it is not making a “significant” additional contribution to the prediction of $y$. If we do so, however, we must again use statistical software to estimate the regression coefficients for the model with only two predictor variables.

The test for $b_0 = 0$ (information given in the row labeled “Constant”) is not useful in this situation. The coefficient $b_0$ corresponds to the August temperature when the value of each explanatory variable is 0. That’s not a realistic combination of $x$-values. Latitude = 0 occurs at the equator; the January and April temperatures there certainly are not 0 degrees Fahrenheit. ◆

### EXAMPLE S3.2—CONTINUED

**Inference for the Peruvian Indian Blood Pressure Model**

Multiple regression output for the Peruvian Indian example is given in Figure S3.2. The interpretation of the results for significance tests represented in the output is easy. All of the $p$-values are small and are well below any reasonable standard for level of significance. Thus, we can conclude that all four explanatory variables are making “significant” contributions to the explanation of variation in systolic blood pressure. ◆

### Predicting an Individual Response and Estimating $E(Y)$

In the context of the simple regression model, Sections 14.4 and 14.5 described methods for predicting the value of the response variable for an individual and for estimating the mean value of $y$ for a population of individuals all with the same particular value of $x$. The discussion in those sections applies to the multiple regression model as well; the only technical change is that in multiple regression the degrees of freedom $= n - p$ for finding a $t^*$ multiplier, rather than $n - 2$ as in simple regression. There were two key definitions in Sections 14.4 and 14.5, and they can be modified for multiple regression as follows:

The regression equation is

$$\text{Systol} = 117 + 0.832 \text{ Weight} - 0.951 \text{ Age} + 2.34 \text{ Years} - 1.08 \text{ Pct\_life}$$

<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>116.84</td>
<td>21.98</td>
<td>5.32</td>
<td>0.000</td>
</tr>
<tr>
<td>Weight</td>
<td>0.8324</td>
<td>0.2754</td>
<td>3.02</td>
<td>0.005</td>
</tr>
<tr>
<td>Age</td>
<td>-0.9507</td>
<td>0.3164</td>
<td>-3.00</td>
<td>0.005</td>
</tr>
<tr>
<td>Years</td>
<td>2.3393</td>
<td>0.7714</td>
<td>3.03</td>
<td>0.005</td>
</tr>
<tr>
<td>Pct_life</td>
<td>-1.0807</td>
<td>0.2833</td>
<td>-3.81</td>
<td>0.001</td>
</tr>
</tbody>
</table>

$S = 8.795 \quad R^2 = 59.7\% \quad R^2(\text{adj}) = 55.0\%$

**FIGURE S3.2** Minitab output for the Peruvian Indian blood pressure model
Figure S3.3 (on the next page) gives 95% confidence intervals for mean systolic blood pressure as well as 95% prediction intervals for individual blood pressure for the four sets of specific values of the \( x \) variables described earlier in the discussion of this example. The specific values are listed in the bottom portion of the output, and the confidence intervals and prediction intervals are given in the top portion. For example, the first set of \( x \)-values is Weight = 65, Age = 30, Years = 15, and Pct_Life = 50.0. At the top of the output, we see that a 95% confidence interval is 117.55 to 129.4 for the mean systolic blood pressure of a population of individuals with these values. A 95% prediction interval is 104.65 to 142.30 for the systolic blood pressure of an individual with these values for the \( x \) variables. This prediction interval is wide, reflecting the fact that there is substantial variation among blood pressures of individuals with similar characteristics. The column labeled “Fit” gives the predicted values for the cases, and “SE Fit” gives the standard error of the predicted value for each case.

**Minitab Tip**

**Creating Prediction Intervals for \( Y \) and Confidence Intervals for \( E(Y) \)**

- Use Stat>Regression>Regression, then use the options button of the dialog box.
- In the box labeled “Prediction intervals for new observations,” either specify the numerical values for each explanatory variable (in order of appearance in model) or specify a set of columns that give multiple values for the explanatory variables.
- The confidence level can be specified in the box labeled “Confidence Level.”

**S3.3 CHECKING CONDITIONS FOR MULTIPLE LINEAR REGRESSION**

In Section 14.6 we listed these five conditions that should be present in order to use a regression model for inference:

1. The form of the regression equation must be correct.
2. There should not be any extreme outliers that influence the results unduly.
3. The standard deviation of the $y$ variable (equivalently, the deviations) is the same regardless of the values of the $x$ variables.

4. The distribution of $y$ (equivalently, the deviations) is a normal distribution. This condition can be relaxed if the sample size is large.

5. Observations in the sample should be independent of one another. Essentially, this means the sample should be a random sample.

Chapter 14 covered simple regression, but these five conditions should be present also for multiple linear regression. Graphical methods for checking conditions 1 through 4 were covered in Section 14.6, and those techniques can be used to verify the conditions for multiple regression.

An additional plot not mentioned in Section 14.6 is informative about conditions 1, 2, and 3 in a multiple regression analysis. In this plot, the values of $e_i$ (the residuals) are plotted on the vertical axis against the values of $\hat{y}_i$ (the predicted values of $y$) on the horizontal axis. Often this plot is referred to as “residuals versus fits” because the term fit is a synonym for “predicted value.” Ideally, this plot should appear to be a band of points randomly spread around a horizontal line at residual = 0. A curved pattern in this plot indicates that the form of the regression equation may be wrong (condition 1 is violated). Outliers (extreme residuals) will show up as points that are noticeably distant in a vertical direction from the other points in the plot, in which case, condition 2 may be violated. Condition 3 may be violated if the vertical range of the residuals is distinctly different in one region of predicted values than in another.

Figure S3.3 (on the next page) is a plot of residuals versus fits for a multiple regression equation in which January and April temperatures are used to predict mean August temperature, and latitude has not been used as an $x$ variable. (The sample regression equation is $\bar{y} = 20.8 - 0.415$ January temp + 1.24 April temp.) The plot is roughly a horizontal band, as it should be if conditions 1, 2, and 3 are met. There are two potential outliers, however. One is at 85 along the horizontal axis, with a value for the residual that is somewhat larger than +5. This observation is Phoenix, which is hotter in August than would be predicted by its January and April temperatures. The other possible outlier is at about 70 along the
horizontal axis, with a residual of about $-6$. This observation is San Francisco, which is cooler in August than would be predicted by its January and April temperatures.

**Minitab Tip**

**Graphing Residuals Versus Fits**

- Use Stat>Regression>Regression, then use the graphs button of the dialog box.
- Click on the box labeled “Residuals versus Fits.”

**EXERCISES**

S3.1 Refer to Example S3.2—CONTINUED, Sample Regression Equation for Peruvian Indian Blood Pressure Study in Section S3.1 (p. S3-5). Show how the four values of $\hat{y}$ given in that example were calculated, and verify that the values are correct.

S3.2 Suppose that a college admissions committee plans to use data for total SAT score, high school grade point average, and high school class rank to predict college freshman-year grade point averages for applicants. Write the population multiple linear regression model for this situation, and state any necessary assumptions. Specify what is the response variable and what are the explanatory variables.

S3.3 Describe an example of a multiple linear regression model for a topic or subject matter area that interests you. Specify the response variable, the explanatory variables, and write the multiple regression model for your example.

S3.4 The figure for this exercise is a plot of residuals versus fits for the sample regression model used in Example S3.2 (the Peruvian Indian example). Are the necessary conditions for using a multiple regression model for statistical inference met in this situation? Explain.

S3.5 Consider the Peruvian Indian blood pressure example. The sample regression equation was $\hat{y} = 116.84 + 0.832$ Weight $- 0.951$ Age $+ 2.34$ Years $- 1.08$ Pct_Life. Suppose the statement is made that if the men had not migrated to an urban area their blood pressures would have decreased about 0.951 points per year of aging because Years and Pct_Life would both be 0, making the regres-
The regression equation is

\[ \text{Height} = 20.2 + 0.193 \text{ momheigh} + 0.539 \text{ dadheigh} \]

<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>20.248</td>
<td>6.680</td>
<td>3.03</td>
<td>0.003</td>
</tr>
<tr>
<td>momheigh</td>
<td>0.19294</td>
<td>0.08777</td>
<td>2.20</td>
<td>0.031</td>
</tr>
<tr>
<td>dadheigh</td>
<td>0.53897</td>
<td>0.07449</td>
<td>7.24</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ S = 2.591 \quad \text{R-Sq} = 48.1\% \quad \text{R-Sq(adj)} = 46.7\% \]

**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>454.71</td>
<td>227.35</td>
<td>33.86</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>73</td>
<td>490.14</td>
<td>6.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>75</td>
<td>944.85</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**New Obs**

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Fit</th>
<th>SE Fit</th>
<th>95.0% CI</th>
<th>95.0% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>70.670</td>
<td>0.349</td>
<td>(69.976, 71.365)</td>
<td>(65.460, 75.881)</td>
</tr>
</tbody>
</table>

**Values of Predictors for New Observations**

<table>
<thead>
<tr>
<th>New Obs</th>
<th>momheigh</th>
<th>dadheigh</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>63.0</td>
<td>71.0</td>
</tr>
</tbody>
</table>
S3-16 SUPPLEMENTAL TOPIC 3

b. In this situation, explain what conclusion can be made about the hypotheses written in part (a). What does this conclusion indicate about the relationship between student height and parents’ heights?

S3.12 Locate the values of s and R{squared} in the output.

a. What is the value of s? Write a sentence that interprets this value in the context of this situation.

b. What is the value of R{squared}? Write a sentence that interprets this value in the context of this situation.

The output gives a confidence interval and a prediction interval for the case where mother’s height is 63 inches and father’s height = 71 inches.

S3.13 Locate the “Analysis of Variance” in the output below.

- a. State the 95% confidence interval that’s given, and write a sentence that explains what is estimated by this interval.
- b. State the 95% prediction interval that’s given, and write a sentence that explains what is predicted by this interval.
- c. The value of the “Fit” for this combination of parents’ heights is reported to equal 70.67. Explain what this value estimates, and show how it was calculated.

S3.14 The figure for this exercise (next column) is a plot of residuals versus fits for the regression model in which student height is the response variable and mother’s and father’s heights are two explanatory variables. Are the necessary conditions for using the regression model for statistical inference met in this situation? Explain.

Exercises S3.15 to S3.17. The physical dataset on the CD for the text includes measurements of forearm and foot lengths (cm) for n = 55 college students. The output for this exercise is for a multiple regression model with y = left forearm length, x₁ = left foot length, and x₂ = right foot length.

S3.15 Locate the “Analysis of Variance” in the output below.

- a. Write the null and alternative hypotheses that are tested using the p-value given in the analysis of variance.
- b. In this situation, explain what conclusion can be made about the hypotheses written in part (a). What does this conclusion indicate about the relationship between left forearm length and the lengths of the right and left feet?

S3.16 Locate the results for the (separate) tests of H₀: β₁ = 0 and H₀: β₂ = 0.

- a. Write a conclusion for each test. Justify your answers using information in the output.
- b. Explain why the conclusions for part (a) of this exercise do not contradict the conclusion for part (b) of the previous exercise.

S3.17 Locate the values of s and R{squared} in the output.

- a. What is the value of s? Write a sentence that interprets this value in the context of this situation.
- b. What is the value of R{squared}? Write a sentence that interprets this value in the context of this situation.

S3.18 In the Peruvian Indian dataset (not on the CD), the variable Calf gives measurements of calf skinfold measurements for the 39 men in the sample. The output for this exercise gives the results for t-tests of H₀: β₅ = 0 when Calf is included in the regression model as an explanatory variable along with the four explanatory variables used in example S3.2.

The regression equation is

\[ \text{LeftArm} = 11.7 + 0.352 \text{ LeftFoot} + 0.185 \text{ RtFoot} \]

<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>11.71</td>
<td>2.518</td>
<td>4.65</td>
<td>0.000</td>
</tr>
<tr>
<td>LeftFoot</td>
<td>0.3519</td>
<td>0.2961</td>
<td>1.19</td>
<td>0.240</td>
</tr>
<tr>
<td>RtFoot</td>
<td>0.1850</td>
<td>0.2816</td>
<td>0.66</td>
<td>0.514</td>
</tr>
</tbody>
</table>

S = 1.796 R-Sq = 36.9% R-Sq(adj) = 34.4%

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>98.019</td>
<td>49.010</td>
<td>15.19</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>52</td>
<td>167.785</td>
<td>3.227</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>54</td>
<td>265.804</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>119.58</td>
<td>21.57</td>
<td>5.54</td>
<td>0.000</td>
</tr>
<tr>
<td>Weight</td>
<td>0.9636</td>
<td>0.2817</td>
<td>3.42</td>
<td>0.002</td>
</tr>
<tr>
<td>Age</td>
<td>-1.0903</td>
<td>0.3217</td>
<td>-3.39</td>
<td>0.002</td>
</tr>
<tr>
<td>Years</td>
<td>2.6295</td>
<td>0.7762</td>
<td>3.39</td>
<td>0.002</td>
</tr>
<tr>
<td>Pct_life</td>
<td>-1.2080</td>
<td>0.2884</td>
<td>-4.19</td>
<td>0.000</td>
</tr>
<tr>
<td>Calf</td>
<td>-0.6413</td>
<td>0.4022</td>
<td>-1.59</td>
<td>0.120</td>
</tr>
</tbody>
</table>

- a. Give numerical values for each of the sample regression coefficients b₀, b₁, b₂, b₅, and b₆.
- b. Use information in the output to explain whether or not it is useful to include the variable Calf as an explanatory variable in the model.
- c. What is the value of the t-statistic for testing H₀: β₅ = 0? Show how this value is calculated from other values reported in the output.
S3.19 Refer to the previous exercise. The sample regression coefficient for the variable \textit{Calf} is a negative value (-0.6413). Explain why this should not be interpreted to mean that as calf skinfold increases, average systolic blood pressure decreases.

S3.20 The data source for this exercise is the CD dataset \textit{UC-Davis1}, which gives data collected from college students. The output below is for a multiple linear regression analysis of the connection between \( y \) = grade point average and three predictor variables: \( x_1 \) = weekly time spent using the computer, \( x_2 \) = weekly time spent watching television, and \( x_3 \) = weekly time spent exercising.

a. Locate the value of \( R^2 \) reported in the output. What is the value? Write a sentence that interprets the value in the context of this situation.

b. Using the results of the significance tests reported in the output, discuss whether you think there is any relationship between grade point average and the three \( x \) variables used in the multiple regression model.

S3.21 Suppose that a multiple linear regression model includes two explanatory variables.

a. Write the population regression model using appropriate statistical notation.

b. Explain the difference between what is represented by the symbol \( b_2 \) and what is represented by the symbol \( \beta_2 \).

S3.22 In a regression analysis, what are the two different interpretations for a predicted value \( \hat{y} \)?

S3.23 Use the \textit{bears-female} dataset on the CD and statistical software for this exercise. The variable \textit{Weight} is the weight of a bear (pounds), \textit{Neck} is the bear’s neck girth (inches), and \textit{Chest} is the bear’s chest girth (inches). In this exercise, \textit{Weight} will be the response variable (\( y \)).

a. Draw scatterplots of \textit{Weight} versus \textit{Neck} and \textit{Weight} versus \textit{Chest}, using \textit{Weight} as the \( y \)-variable in each instance. Write a short description of each plot. Do the relationships appear to be linear? Are there any outliers?

b. Omit the sixth observation in the dataset. It’s an outlier with values that don’t seem legitimate. Then, find the estimate of the multiple linear regression equation using \textit{Weight} as the response variable, and \textit{Neck} and \textit{Chest} as the explanatory variables. What is the sample regression equation?

c. Based on the relevant \( t \)-tests, explain whether both explanatory variables are “significant” predictors of \textit{Weight}.

d. What is the value of \( R^2 \)? Write a sentence that interprets the value in the context of this situation.

\begin{tabular}{|l|c|c|c|c|}
\hline
Predictor   & Coef & SE Coef & \( T \) & \( P \) \\
\hline
Constant    & 3.0424 & 0.1013 & 30.04 & 0.000 \\
Computer    & -0.005594 & 0.004042 & -1.38 & 0.168 \\
TV          & -0.002773 & 0.004507 & -0.62 & 0.539 \\
Exercise    & -0.00201 & 0.01073 & -0.19 & 0.851 \\
\hline
\end{tabular}

\( S = 0.5924 \) \hspace{1cm} \text{R-Sq = 1.5%} \hspace{1cm} \text{R-Sq(adj) = 0.0%}

Analysis of Variance

\begin{tabular}{|l|c|c|c|c|c|}
\hline
Source & DF & SS & MS & \( F \) & \( P \) \\
\hline
Regression & 3 & 0.8120 & 0.2707 & 0.77 & 0.512 \\
Residual Error & 157 & 55.0912 & 0.3509 & \\
Total & 160 & 55.9032 & & \\
\hline
\end{tabular}

REFERENCE