As we learned in Chapter 9 ("Bivariate Tables"), the differential access to the Internet is real and persistent. Celeste Campos-Castillo’s (2015) research confirmed the impact of gender and race on the digital divide. Pew researchers Andrew Perrin and Maeve Duggan (2015) documented other sources of the divide. For example, Americans with college degrees continue to have higher rates of Internet use than Americans with less than a college degree. Though less-educated adults have increased their Internet use since 2000, the percentage who use the Internet is still lower than the percentage of college graduates.¹

In this chapter, we apply regression and correlation techniques to examine the relationship between interval-ratio variables. **Correlation** is a measure of association used to determine the existence and strength of the relationship between variables and is similar to the proportional reduction of error (PRE) measures reviewed in Chapter 10 ("The Chi-Square Test and Measures of Association"). **Regression** is a linear prediction model, using one or more independent variables to predict the values of a dependent variable. We will present two basic models: (1) **Bivariate regression** examines how changes in one independent variable affects the value of a dependent variable, while (2) **multiple regression** estimates how several independent variables affect one dependent variable.

We begin with calculating the bivariate regression model for educational attainment and Internet hours per week. We will use *years of educational attainment* as our independent variable (*X*) to predict *Internet hours per week* (our dependent variable or *Y*). Fictional data are presented for a sample of 10 individuals in Table 12.1.

**THE SCATTER DIAGRAM**

One quick visual method used to display the relationship between two interval-ratio variables is the **scatter diagram** (or **scatterplot**). Often used as a first exploratory step in regression analysis, a scatter diagram can suggest whether two variables are associated.
Table 12.1  Educational Attainment and Internet Hours per Week, $N = 10$

<table>
<thead>
<tr>
<th>Educational Attainment ($X$)</th>
<th>Internet Hours per Week ($Y$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
</tr>
<tr>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>21</td>
<td>8</td>
</tr>
<tr>
<td>Mean = 14.80</td>
<td>Mean = 4.50</td>
</tr>
<tr>
<td>Variance = 23.07</td>
<td>Variance = 9.17</td>
</tr>
<tr>
<td>Range = 23 − 9 = 14</td>
<td>Range = 9 − 0 = 9</td>
</tr>
</tbody>
</table>

The scatter diagram showing the relationship between educational attainment and Internet hours per week is shown in Figure 12.1. In a scatter diagram, the scales for the two variables form the vertical and horizontal axes of a graph. Usually, the independent variable, $X$, is arrayed along the horizontal axis and the dependent variable, $Y$, along the vertical axis.

In Figure 12.1, each dot represents a person; its location lies at the exact intersection of that person’s years of education and Internet hours per week. Note that individuals with lower educational attainment have fewer hours of Internet use, while individuals with higher educational attainment spend more time on Internet per week. Educational attainment and Internet hours are positively associated.

Scatter diagrams may also reveal a negative association between two variables or no relationship at all. We will review a negative relationship between two variables later in this chapter. Nonlinear relationships are explained in A Closer Look 12.1.

**LINEAR RELATIONSHIPS AND PREDICTION RULES**

Though we can use a scatterplot as a first step to explore a relationship between two interval-ratio variables, we need a more systematic way to express the relationship between two interval-ratio variables. One way is to express them as a linear relationship. A linear relationship allows us to approximate the observations displayed in a scatter diagram with a straight line. In a perfectly linear relationship, all the observations (the dots) fall along a straight line (a perfect relationship is sometimes called a deterministic relationship), and
the line itself provides a predicted value of $Y$ (the vertical axis) for any value of $X$ (the horizontal axis). For example, in Figure 12.3, we have superimposed a straight line on the scatterplot originally displayed in Figure 12.1. Using this line, we can obtain a predicted value of Internet hours per week for any individual by starting with a value from the education axis and then moving up to the Internet hours per week axis (indicated by the dotted lines). For example, the predicted value of Internet hours per week for someone with 12 years of education is approximately 3 hours.

**Finding the Best-Fitting Line**

As indicated in Figure 12.3, the actual relationship between years of education and Internet hours is not perfectly linear; that is, although some of individual points lie very close to the line, none fall exactly on the line. Most relationships we study in the social sciences are not deterministic, and we are not able to come up with a linear equation that allows us to predict $Y$ from $X$ with perfect accuracy. We are much more likely to find relationships approximating linearity, but in which numerous cases don’t follow this trend perfectly.

The relationship between educational attainment and Internet hours, as depicted in Figure 12.3, can also be described with the following algebraic equation, an equation for a straight line:

$$Y = a + b(X)$$  \[(12.1)\]
The regression examples we present in this chapter reflect two assumptions.

The first assumption is that the dependent and independent variables are interval-ratio measurements. In fact, regression models often include ordinal measures such as social class, income, and attitudinal scales. (Later in the chapter we feature a regression model based on ordinal attitudinal scales.) Dummy variable techniques (creating a dichotomous variable, coded one or zero) permit the use of nominal variables, such as sex, race, religion, or political party affiliation. For example, in measuring gender, males could be coded as 0 and females coded as 1. Dummy variable techniques will not be elaborated here.

Our second assumption is that the variables have a linear or straight-line relationship. For the most part, social science relationships can be approximated using a linear equation. It is important to note, however, that sometimes a relationship cannot be approximated by a straight line and is better described by some other, nonlinear function. For example, Figure 12.2 shows a nonlinear relationship between age and hours of reading (hypothetical data). Hours of reading increase with age until the twenties, remain stable until the forties, and then tend to decrease with age.

There are regression models for many nonlinear relationships, for nominal or dichotomous dependent variables, or even when there are multiple dependent variables. These advanced regression techniques will not be covered in this text.

**Figure 12.2 A Nonlinear Relationship Between Age and Hours of Reading per Week**

![Graph showing a nonlinear relationship between age and hours of reading per week.](image)

where

- \( Y \) = the predicted score on the dependent variable
- \( X \) = the score on the independent variable
- \( a \) = the \( Y \)-intercept, or the point where the line crosses the \( Y \)-axis; therefore, \( a \) is the value of \( Y \) when \( X \) is 0
- \( b \) = the slope of the regression line, or the change in \( Y \) with a unit change in \( X \).
LEARNING CHECK

For each of these four lines, as X goes up by 1 unit, what does Y do? Be sure you can answer this question using both the equation and the line.
DEFINING ERROR The best-fitting line is the one that generates the least amount of error, also referred to as the residual. Look again at Figure 12.3. For each education level, the line (or the equation that this line represents) predicts a value of Internet hours. For example, with 21 years of education, the predicted value for $Y$ is 8.34 hours. But we know from Table 12.1 that the actual value for 21 years of education is 8.0 hours. Thus, we have two values for $Y$: (1) a predicted $\hat{Y}$, which we symbolize as $\hat{Y}$, and which is generated by the prediction equation, also called the linear regression equation $Y = a + b(X)$, and (2) the observed $Y$, symbolized simply as $Y$. Thus, for someone with 21 years of education, $\hat{Y} = 8.34$, whereas $Y = 8.0$.

We can think of the residual as the difference between the observed $Y$ and the predicted $\hat{Y}$. If we symbolize the residual as $e$, then

$$e = Y - \hat{Y}$$

The residual is $8.34 - 8.0 = 0.34$ hours.

THE RESIDUAL SUM OF SQUARES ($\Sigma e^2$) Our goal is to identify a line or a prediction equation that minimizes the error for each individual observation. However, any line we choose will minimize the residual for some observations but may maximize it for others. We want to find a prediction equation that minimizes the residuals over all observations.

There are many mathematical ways of defining the residuals. For example, we may take the algebraic sum of residuals $\Sigma (Y - \hat{Y})$, the sum of the absolute residuals $\Sigma |Y - \hat{Y}|$, or the sum of the squared residuals $\Sigma (Y - \hat{Y})^2$. For mathematical reasons, statisticians prefer to work with the third method—squaring and summing the residuals over all observations. The result is the residual sum of squares, or $\Sigma e^2$. Symbolically, $\Sigma e^2$ is expressed as

$$\Sigma e^2 = \Sigma (Y - \hat{Y})^2$$

THE LEAST SQUARES LINE The best-fitting regression line is that line where the sum of the squared residuals, or $\Sigma e^2$, is at a minimum. Such a line is called the least squares line (or best-fitting line), and the technique that produces this line is called the least squares method. The technique involves choosing $a$ and $b$ for the equation such that $\Sigma e^2$ will have the smallest possible value. In the next section, we use the data from the 10 individuals to find the least squares equation.

Computing $a$ and $b$

Through the use of calculus, it can be shown that to figure out the values of $a$ and $b$ in a way that minimizes $\Sigma e^2$, we need to apply the following formulas:

$$b = \frac{\sum XY}{\sum X^2}$$  \hspace{1cm} (12.2)

$$a = \bar{Y} - b(\bar{X})$$  \hspace{1cm} (12.3)

where

- $s_{XY} = \text{covariance of } X \text{ and } Y$
- $s_X^2 = \text{variance of } X$
- $\bar{Y} = \text{mean of } Y$
- $\bar{X} = \text{mean of } X$
\[ a = \text{the } Y\text{-intercept} \]
\[ b = \text{the slope of the line} \]

These formulas assume that \( X \) is the independent variable and \( Y \) is the dependent variable.

Before we compute \( a \) and \( b \), let’s examine these formulas. The denominator for \( b \) is the variance of the variable \( X \). It is defined as follows:

\[
\text{Variance } (X) = s_X^2 = \frac{\sum (X - \bar{X})^2}{N-1}
\]

This formula should be familiar to you from Chapter 4 (“Measures of Variability”). The numerator \((s_{XY})\), however, is a new term. It is the covariance of \( X \) and \( Y \) and is defined as

\[
\text{Covariance } (X, Y) = s_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N-1}
\]

The covariance is a measure of how \( X \) and \( Y \) vary together. Essentially, the covariance tells us to what extent higher values of one variable are associated with higher values of the second variable (in which case we have a positive covariation) or with lower values of the second variable (which is a negative covariation). Based on the formula, we subtract the mean of \( X \) from each \( X \) score and the mean of \( Y \) from each \( Y \) score, and then take the product of the two deviations. The results are then summed for all the cases and divided by \( N - 1 \).

In Table 12.2, we show the computations necessary to calculate the values of \( a \) and \( b \) for our 10 individuals. The means for educational attainment and Internet hours per week are obtained by summing Column 1 and Column 2, respectively, and dividing each sum by \( N \). To calculate the covariance, we first subtract from each \( X \) score (Column 3) and from each \( Y \) score (Column 5) to obtain the mean deviations. We then multiply these deviations for every observation. The products of the mean deviations are shown in Column 7.

The covariance is a measure of the linear relationship between two variables, and its value reflects both the strength and the direction of the relationship. The covariance will be close to zero when \( X \) and \( Y \) are unrelated; it will be larger than zero when the relationship is positive and smaller than zero when the relationship is negative.

Now, let’s substitute the values for the covariance and the variance from Table 12.2 to calculate \( b \):

\[
b = \frac{s_{XY}}{s_X^2} = \frac{14.22}{23.07} = 0.62
\]

Once \( b \) has been calculated, we can solve for \( a \), the intercept:

\[
a = \bar{Y} - b(\bar{X}) = 4.5 - 0.62(14.8) = -4.68
\]

The prediction equation is therefore

\[
\hat{Y} = -4.68 + 0.62(X)
\]

This equation can be used to obtain a predicted value for Internet hours per week given an individual’s years of education. For example, for a person with 15 years of education, the predicted Internet hours is

\[
\hat{Y} = -4.68 + 0.62(15) = 4.62
\]
### Table 12.2 Worksheet for Calculating \(a\) and \(b\) for the Regression Equation

<table>
<thead>
<tr>
<th>Educational Attainment</th>
<th>Internet Hours per Week</th>
<th>((X - X)^2)</th>
<th>((Y - Y)^2)</th>
<th>((Y - Y)^2)</th>
<th>((X - X)(Y - Y))</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
<td>-4.8</td>
<td>-3.5</td>
<td>12.25</td>
<td>16.80</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>-5.8</td>
<td>-4.5</td>
<td>20.25</td>
<td>26.10</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>-2.8</td>
<td>-1.5</td>
<td>2.25</td>
<td>4.00</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>-1.8</td>
<td>-0.5</td>
<td>.25</td>
<td>0.90</td>
</tr>
<tr>
<td>19</td>
<td>7</td>
<td>4.2</td>
<td>2.5</td>
<td>6.25</td>
<td>10.50</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>-3.8</td>
<td>-2.5</td>
<td>6.25</td>
<td>9.50</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>1.2</td>
<td>1.5</td>
<td>2.25</td>
<td>1.80</td>
</tr>
<tr>
<td>23</td>
<td>9</td>
<td>8.2</td>
<td>4.5</td>
<td>20.25</td>
<td>36.90</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td>-0.8</td>
<td>0.5</td>
<td>0.25</td>
<td>-0.40</td>
</tr>
<tr>
<td>21</td>
<td>8</td>
<td>6.2</td>
<td>3.5</td>
<td>12.25</td>
<td>21.70</td>
</tr>
</tbody>
</table>

\[ \sum X = 148 \quad \sum Y = 45 \]

\[ X \bar{X} = \frac{\sum X}{N} = \frac{148}{10} = 14.8 \]

\[ Y \bar{Y} = \frac{\sum Y}{N} = \frac{45}{10} = 4.5 \]

\[ s_x = \sqrt{23.07} = 4.80 \]

\[ s_y = \sqrt{9.17} = 3.03 \]

\[ s_{yx} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N - 1} = \frac{128}{9} = 14.22 \]

**Note:**

a. Answers may differ due to rounding; however, the exact value of these column totals, properly calculated, will always be equal to zero.
A CLOSER LOOK 12.2
UNDERSTANDING THE COVARIANCE

Let’s say we have a set of eight data points for which the mean of
\( X \) is 6 and the mean of \( Y \) is 3.

So the covariance in this case will be positive, giving us a positive \( b \) and a positive \( r \).

Now let’s say we have a set of eight points that look like this:

So the covariance in this case will be negative, giving us a negative \( b \).
We can plot the straight-line graph corresponding to the regression equation. To plot a straight line, we need only two points, where each point corresponds to an X, Y value predicted.

**Interpreting a and b**

The $b$ coefficient is equal to 0.62. This tells us that with each additional year of educational attainment, Internet hours per week is predicted to increase by 0.62 hours.

Note that because the relationships between variables in the social sciences are inexact, we don’t expect our regression equation to make perfect predictions for every individual case. However, even though the pattern suggested by the regression equation may not hold for every individual, it gives us a tool by which to make the best possible guess about how Internet usage is associated, on average, with educational attainment. We can say that the slope of 0.62 is the estimate of this relationship.

The $Y$ intercept $a$ is the predicted value of $Y$, when $X = 0$. Thus, it is the point at which the regression line and the $Y$-axis intersect. The $Y$ intercept can have positive or negative values. In this instance, it is unusual to consider someone with 0 years of education. As a general rule, be cautious when making predictions for $Y$ based on values of $X$ that are outside the range of the data, such as the -4.68 intercept calculated for our model. The intercept may not have a clear substantive interpretation.

We can plot the regression equation with two points: (1) the mean of $X$ and the mean of $Y$ and (2) 0 and the value of $a$. We’ve displayed this regression line in Figure 12.4.

![Figure 12.4 The Best-Fitting Line for Educational Attainment and Internet Hours per Week](image-url)
LEARNING CHECK

Use the prediction equation to calculate the predicted values of Y if X equals 9, 11, or 14. Verify that the regression line in Figure 12.3 passes through these points.

A NEGATIVE RELATIONSHIP: AGE AND INTERNET HOURS PER WEEK

Pew researchers Perrin and Duggan (2015) also documented how older adults have lagged behind younger adults in their Internet adoption. The majority of seniors, about 58%, currently use the Internet. In this section, we’ll examine the relationship between respondent age and Internet hours per week, defining Internet hours as the dependent variable (Y) and age as the independent variable (X). The fictional data are presented in Table 12.3 and the corresponding scatter diagram in Figure 12.5.

The scatter diagram reveals that age and Internet hours per week are linearly related. It also illustrates that these variables are negatively associated; that is, as age increases the number of hours of Internet access decreases. (Compare Figure 12.5 with Figure 12.4. Notice how the regression lines are in opposite direction—one positive, the other negative.)

For a more systematic analysis of the association, we will estimate the least squares regression equation for these data. Table 12.3 shows the calculations necessary to find $a$ and $b$ for our data on age and hours spent weekly on the Internet.

Now, let’s substitute the values for the covariance and the variance from Table 12.3 to calculate $b$:

$$b = \frac{s_{XY}}{s_X^2} = \frac{-34.44}{167.73} = -.205 = -.21$$

Interpreting the slope, we can say with each one year increase in age, Internet hours per week will decline by .21. This indicates a negative relationship between age and Internet hours. Once $b$ has been calculated, we can solve for $a$, the intercept:

$$a = \bar{Y} - b(\bar{X}) = 4.5 - (-.21)(37.8) = 12.44$$

The prediction equation is therefore

$$\hat{Y} = 12.44 - .20(X)$$

This equation can be used to obtain a predicted value for Internet hours per week given respondent’s age. In Figure 12.5, the regression line is plotted over our original scatter diagram.

METHODS FOR ASSESSING THE ACCURACY OF PREDICTIONS

So far, we calculated two regression equations that help us predict Internet usage per week based on educational attainment or age. In both cases, our predictions are far from perfect. If we examine Figures 12.4 and 12.5, we can see that we fail to make accurate predictions in every case. Though some of the individual points lie fairly close to the regression line, not all
### Table 12.3  Age and Internet Hours per Week, $N = 10$; Worksheet for Calculating $a$ and $b$ for the Regression Equation

<table>
<thead>
<tr>
<th>Age</th>
<th>Internet Hours per Week</th>
<th>$(X - \bar{X})$</th>
<th>$(X - \bar{X})^2$</th>
<th>$(Y - \bar{Y})$</th>
<th>$(Y - \bar{Y})^2$</th>
<th>$(X - \bar{X})(Y - \bar{Y})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>1</td>
<td>17.2</td>
<td>295.84</td>
<td>-3.5</td>
<td>12.25</td>
<td>-60.2</td>
</tr>
<tr>
<td>60</td>
<td>0</td>
<td>22.2</td>
<td>492.84</td>
<td>-4.5</td>
<td>20.25</td>
<td>-89.9</td>
</tr>
<tr>
<td>45</td>
<td>3</td>
<td>7.2</td>
<td>51.84</td>
<td>-1.5</td>
<td>2.25</td>
<td>10.8</td>
</tr>
<tr>
<td>35</td>
<td>4</td>
<td>-2.8</td>
<td>7.84</td>
<td>-0.5</td>
<td>.25</td>
<td>1.4</td>
</tr>
<tr>
<td>23</td>
<td>7</td>
<td>-14.8</td>
<td>219.04</td>
<td>2.5</td>
<td>6.25</td>
<td>-37</td>
</tr>
<tr>
<td>40</td>
<td>2</td>
<td>2.20</td>
<td>4.84</td>
<td>-2.5</td>
<td>6.25</td>
<td>-5.5</td>
</tr>
<tr>
<td>22</td>
<td>6</td>
<td>-15.8</td>
<td>249.64</td>
<td>1.5</td>
<td>2.25</td>
<td>-23.7</td>
</tr>
<tr>
<td>27</td>
<td>9</td>
<td>-10.8</td>
<td>116.64</td>
<td>4.5</td>
<td>20.25</td>
<td>-48.6</td>
</tr>
<tr>
<td>41</td>
<td>5</td>
<td>3.2</td>
<td>10.24</td>
<td>0.5</td>
<td>.25</td>
<td>1.6</td>
</tr>
<tr>
<td>30</td>
<td>8</td>
<td>-7.8</td>
<td>60.64</td>
<td>3.5</td>
<td>12.25</td>
<td>-27.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>$\sum X = 378$</strong></td>
<td></td>
<td><strong>$\sum Y = 45$</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$$\bar{X} = \frac{\sum X}{N} = \frac{378}{10} = 37.8$$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{45}{10} = 4.5$$

$$s_x = \sqrt{167.73} = 12.96$$

$$s_y = \sqrt{9.17} = 3.03$$

$$s_{xy} = \sqrt{\frac{\sum (X - \bar{X})(Y - \bar{Y})}{N-1}} = \frac{-310}{9} = -34.44$$

**Note:**

a. Answers may differ due to rounding; however, the exact value of these column totals, properly calculated, will always be equal to zero.
lie directly on the line—an indication that some prediction error made. We have a model that helps us make predictions, but how can we assess the accuracy of these predictions?

We saw earlier that one way to judge our accuracy is to review the scatterplot. The closer the observations are to the regression line, the better the fit between the predictions and the actual observations. Still we need a more systematic method for making such a judgment. We need a measure that tells us how accurate a prediction the regression model provides. The coefficient of determination, or \( r^2 \), is such a measure. The coefficient of determination measures the improvement in the prediction error based on our use of the linear prediction equation. The coefficient of determination is a \( PRE \) measure of association. Recall from Chapter 10 that \( PRE \) measures adhere to the following formula:

\[
PRE = 1 - \frac{E_1 - E_2}{E_1}
\]

where

\[
E_1 = \text{prediction errors made when the independent variable is ignored}
\]

\[
E_2 = \text{prediction errors made when the prediction is based on the independent variable}
\]

Applying this to the regression model, we have two prediction rules and two measures of error. The first prediction rule is in the absence of information on \( X \), predict \( \overline{Y} \). The error of
prediction is defined as \( Y - \bar{Y} \). The second rule of prediction uses \( X \) and the regression equation to predict \( \hat{Y} \). The error of prediction is defined as \( Y - \hat{Y} \).

To calculate these two measures of error for all the cases in our sample, we square the deviations and sum them. Thus, for the deviations from the mean of \( Y \) we have

\[
\Sigma (Y - \bar{Y})^2
\]

The sum of the squared deviations from the mean is called the total sum of squares, or \( SST \):

\[
SST = \Sigma (Y - \bar{Y})^2
\]

To measure deviation from the regression line, or \( \hat{Y} \), we have

\[
\Sigma (Y - \hat{Y})^2
\]

The sum of squared deviations from the regression line is denoted as the residual sum of squares, or \( SSE \):

\[
SSE = \Sigma (Y - \hat{Y})^2
\]

(We discussed this error term, the residual sum of squares, earlier in the chapter.)

The predictive value of the linear regression equations can be assessed by the extent to which the residual sum of squares, or \( SSE \), is smaller than the total sum of squares, \( SST \). By subtracting \( SSE \) from \( SST \) we obtain the regression sum of squares, or \( SSR \), which reflects improvement in the prediction error resulting from our use of the linear prediction equation. \( SSR \) is defined as

\[
SSR = SST - SSE
\]

Let’s calculate \( r^2 \) for our regression model.

**Calculating Prediction Errors**

Figure 12.6 displays the regression line we calculated for educational attainment (\( X \)) and the Internet hours per week (\( Y \)) for 10 individuals, highlighting the prediction of \( Y \) for the person with 16 years of education, Subject A. Suppose we didn’t know the actual \( Y \), the number of Internet hours per week. Suppose further that we did not have knowledge of \( X \), Subject A’s years of education. Because the mean minimizes the sum of the squared errors for a set of scores, our best guess for \( Y \) would be \( \bar{Y} \), or 4.5 hours. The horizontal line in Figure 12.6 represents this mean. Now, let’s compare actual \( Y \), 6 hours with this prediction:

\[
Y - \bar{Y} = 6 - 4.5 = 1.5
\]

With an error of 1.5, our prediction of the average score for Subject A is not accurate.

Let’s see if our predictive power can be improved by using our knowledge of \( X \)—the years of education—and its linear relationship with \( Y \)—Internet hours per week. If we insert Subject A’s 16 years of education into our prediction equation, as follows:
We can now recalculate our new error of prediction by comparing the predicted $\hat{Y}$ with the actual $Y$:

$$Y - \hat{Y} = 6 - 5.24 = .76$$

Although this prediction is by no means perfect, it is a slight improvement of .73 (1.5 - 0.76 = 0.74) over our earlier prediction. This improvement is illustrated in Figure 12.6.

Note that this improvement is the same as $\hat{Y} - \bar{Y} = 5.24 - 4.5 = .74$. This quantity represents the improvement in the prediction error resulting from our use of the linear prediction equation.

Let's calculate these terms for our data on educational attainment ($X$) and Internet use ($Y$). We already have from Table 12.3 the total sum of squares:

$$SST = \sum (Y - \bar{Y})^2 = 82.50$$

To calculate the errors sum of squares, we will calculate the predicted $\hat{Y}$ for each individual, subtract it from the observed $Y$, square the differences, and sum these for all 10 individuals. These calculations are presented in Table 12.4.
The residual sum of squares is thus

\[ SSE = \sum (Y - \hat{Y})^2 = 3.59 \]

\( SSR \) is then given as

\[ SSR = SST - SSE = 82.50 - 3.59 = 78.91 \]

We have all the elements we need to construct a \( PRE \) measure. Because \( SST \) measures the prediction errors when the independent variable is ignored, we can define

\[ E_1 = SST \]

Similarly, because \( SSE \) measures the prediction errors resulting from using the independent variable, we can define

\[ E_2 = SSE \]

We are now ready to define the coefficient of determination \( r^2 \). It measures the \( PRE \) associated with using the linear regression equation as a rule for predicting \( Y \):

\[
PRE = \frac{E_1 - E_2}{E_1} = \frac{\sum (Y - \bar{Y})^2 - \sum (Y - \hat{Y})^2}{\sum (Y - \bar{Y})^2} \tag{12.5}
\]

<p>| Table 12.4 Worksheet for Calculating Errors Sum of Squares (SSE) |
|-------------------|------------------|-----------------|-----------------|----------------|-----------------|
| (1) | (2) | (3) | (4) | (5) |</p>
<table>
<thead>
<tr>
<th>Educational Attainment</th>
<th>Internet Hours per Week</th>
<th>Predicted Y</th>
<th>Y - ( \hat{Y} )</th>
<th>( (Y - \hat{Y})^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Y</td>
<td>( \hat{Y} )</td>
<td>Y - ( \hat{Y} )</td>
<td>( (Y - \hat{Y})^2 )</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1.52</td>
<td>-0.52</td>
<td>0.27</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0.90</td>
<td>-0.90</td>
<td>0.81</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>2.76</td>
<td>0.24</td>
<td>0.06</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>3.38</td>
<td>0.62</td>
<td>0.38</td>
</tr>
<tr>
<td>19</td>
<td>7</td>
<td>7.10</td>
<td>-0.10</td>
<td>0.01</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>2.14</td>
<td>-0.14</td>
<td>0.02</td>
</tr>
<tr>
<td>16</td>
<td>5</td>
<td>5.24</td>
<td>-0.76</td>
<td>0.58</td>
</tr>
<tr>
<td>23</td>
<td>9</td>
<td>9.58</td>
<td>-0.58</td>
<td>0.34</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td>4.00</td>
<td>1</td>
<td>1.0</td>
</tr>
<tr>
<td>21</td>
<td>8</td>
<td>8.34</td>
<td>-0.34</td>
<td>0.12</td>
</tr>
<tr>
<td>( \sum X = 148 )</td>
<td>( \sum Y = 45 )</td>
<td></td>
<td>( \sum (Y - \hat{Y})^2 = 3.59 )</td>
<td></td>
</tr>
</tbody>
</table>
For our example,

\[ r^2 = \frac{82.50 - 3.59}{82.50} = \frac{78.91}{82.50} = 0.96 \]

The coefficient of determination \((r^2)\) reflects the proportion of the total variation in the dependent variable, \(Y\), explained by the independent variable, \(X\). An \(r^2\) of 0.96 means that by using educational attainment and the linear prediction rule to predict Internet hours per week—we have reduced the error of prediction by 96%. We can also say that the independent variable (educational attainment) explains about 96% of the variation in the dependent variable (Internet hours per week) as illustrated in Figure 12.7.

The coefficient of determination ranges from 0.0 to 1.0. An \(r^2\) of 1.0 means that by using the linear regression model, we have reduced uncertainty by 100%. It also means that the independent variable accounts for 100% of the variation in the dependent variable. With an \(r^2\) of 1.0, all the observations fall along the regression line, and the prediction error is equal to 0.0. An \(r^2\) of 0.0 means that using the regression equation to predict \(Y\) does not improve the prediction of \(Y\). Figure 12.8 shows \(r^2\) values near 0.0 and near 1.0. In Figure 12.8a, where \(r^2\) is approximately 1.0, the regression model provides a good fit. In contrast, a very poor fit is evident in Figure 12.8b, where \(r^2\) is near zero. An \(r^2\) near zero indicates either poor fit or a well-fitting line with a \(b\) of zero.

**CALCULATING \(r^2\)** Another method for calculating \(r^2\) uses the following equation:

\[
 r^2 = \frac{\text{Covariance}(X,Y)^2}{\text{Variance}(X)\text{Variance}(Y)} = \frac{\hat{r}^2_{XY}}{\hat{r}^2_{XY} + \hat{r}^2_{Y}}
\]

(12.6)

This formula tells us to divide the square of the covariance of \(X\) and \(Y\) by the product of the variance of \(X\) and the variance of \(Y\).
To calculate $r^2$ for our example, we can go back to Table 12.2, where the covariance and the variances for the two variables have already been calculated:

$$s_{XY} = 14.22$$

$$s_X^2 = 23.07$$

$$s_Y^2 = 9.17$$

Therefore,

$$r^2 = \frac{14.22^2}{23.07 \times 9.17} = \frac{202.21}{211.55} = .96$$

Since we are working with actual values for educational attainment, its metric, or measurement, values are different from the metric values for the dependent variable, Internet hours per week. While this hasn’t been an issue until now, we must account for this measurement difference if we elect to use the variances and covariance to calculate $r^2$ (Formula 12.6). The remedy is actually quite simple. All we have to do is multiply our obtained $r^2$, 0.96, by 100 to obtain 96. Why multiply the obtained $r^2$ by 100?

We can multiply $r^2$ by 100 to obtain the percentage of variation in the dependent variable explained by the independent variable. An $r^2$ of 0.96 means that by using educational attainment and the linear prediction rule to predict $Y$, Internet hours per week—we have reduced uncertainty of prediction by 96%. We can also say that the independent variable explains 96% of the variation in the dependent variable, as illustrated in Figure 12.7.

**LEARNING CHECK**

*Calculate $r$ and $r^2$ for the age and Internet hours regression model. Interpret your results.*
TESTING THE SIGNIFICANCE OF $r^2$ USING ANOVA

Like other descriptive statistics, $r^2$ is an estimate based on sample data. Once $r^2$ is obtained, we should assess the probability that the linear relationship between median household income and the percentage of state residents with a bachelor's degree, as expressed in $r^2$, is really zero in the population (given the observed sample coefficient). In other words, we must test $r^2$ for statistical significance. ANOVA (analysis of variance), presented earlier in Chapter 11 (“Analysis of Variance”), can easily be applied to determine the statistical significance of the regression model as expressed in $r^2$. In fact, when you look closely, ANOVA and regression analysis can look very much the same. In both methods, we attempt to account for variation in the dependent variable in terms of the independent variable, except that in ANOVA the independent variable is a categorical variable (nominal or ordinal, e.g., gender or social class) and with regression, it is an interval-ratio variable (e.g., income measured in dollars).

With ANOVA, we decomposed the total variation in the dependent variable into portions explained (SSB) and unexplained (SSW) by the independent variable. Next, we calculated the mean squares between ($SSB/df_b$) and mean squares within ($SSW/df_w$). The statistical test, $F$, is the ratio of the mean squares between to the mean squares within as shown in Formula 12.7.

$$F = \frac{\text{Mean squares between}}{\text{Mean squares within}} = \frac{SSB / df_b}{SSW / df_w} \quad (12.7)$$

With regression analysis, we decompose the total variation in the dependent variable into portions explained, SSR, and unexplained, SSE. Similar to ANOVA, the mean squares regression and the mean squares residual are calculated by dividing each sum of squares by its corresponding degrees of freedom ($df$). The degrees of freedom associated with SSR ($df_r$) are equal to $K$, which refers to the number of independent variables in the regression equation.

$$\text{Mean squares regression} = \frac{SSR}{df_r} = \frac{SSR}{K} \quad (12.8)$$

For SSE, degrees of freedom ($df_e$) is equal to $[N - (K + 1)]$, with $N$ equal to the sample size.

$$\text{Mean squares residual} = \frac{SSE}{df_e} = \frac{SSE}{[N - (K + 1)]} \quad (12.9)$$

In Table 12.5 for example, we present the ANOVA summary table for educational attainment and Internet hours.

In the table, under the heading Source of Variation are displayed the regression, residual, and total sums of squares. The column marked $df$ shows the degrees of freedom associated with both the regression and residual sum of squares. In the bivariate case, SSR has 1 degree of freedom associated with it. The degrees of freedom associated with SSE is $[N - (K + 1)]$, where $K$ refers to the number of independent variables in the regression equation. In the bivariate case, with one independent variable—median household income—SSE has $N - 2$ degrees of freedom associated with it $[N - (1 + 1)]$. Finally, the mean squares regression (MSR) and the mean squares residual (MSE) are calculated by dividing each sum of squares by its corresponding degrees of freedom. For our example,
The $F$ statistic together with the mean squares regression and the mean squares residual compose the obtained $F$ ratio or $F$ statistic. The $F$ statistic is the ratio of the mean squares regression to the mean squares residual:

$$F = \frac{\text{Mean squares regression}}{\text{Mean squares residual}} = \frac{SSR / df_r}{SSE / df_e}$$

The $F$ ratio, thus, represents the size of the mean squares regression relative to the size of the mean squares residual. The larger the mean squares regression relative to the mean squares residual, the larger the $F$ ratio and the more likely that $r^2$ is significantly larger than zero in the population. We are testing the null hypothesis that $r^2$ is zero in the population.

The $F$ ratio of our example is

$$F = \frac{\text{Mean squares regression}}{\text{Mean squares residual}} = \frac{78.91}{0.45} = 175.36$$

### Making a Decision

To determine the probability of obtaining an $F$ statistic of 175.36, we rely on Appendix E, Distribution of $F$. Appendix E lists the corresponding values of the $F$ distribution for various degrees of freedom and two levels of significance, .05 and .01. We will set alpha at .05, and thus, we will refer to the table marked “$p < .05$.” Note that Appendix E includes two $df$s. For the numerator, $df_r$ refers to the $df$, associated with the mean squares regression; for the denominator, $df_e$ refers to the $df$, associated with the mean squares residual. For our example, we compare our obtained $F$ (175.36) to the $F$ critical. When the $df$s are 1 (numerator) and 8 (denominator), and $\alpha < .05$, the $F$ critical is 5.32. Since our obtained $F$ is larger than the $F$ critical (175.36 > 5.32), we can reject the null hypothesis that $r^2$ is zero in the population. We conclude that the linear relationship between educational attainment and Internet hours per week as expressed in $r^2$ is probably greater than zero in the population (given our observed sample coefficient).

### Table 12.5 ANOVA Summary Table for Educational Attainment and Internet Hours per Week

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>$df$</th>
<th>Mean Squares</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>78.93</td>
<td>1</td>
<td>78.91</td>
<td>175.36</td>
</tr>
<tr>
<td>Residual</td>
<td>3.57</td>
<td>8</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>82.5</td>
<td>9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
LEARNING CHECK

Test the null hypothesis that there is a linear relationship between Internet hours and age. The mean squares regression is 63.66 with 1 degree of freedom. The mean squares residual is 2.355 with 8 degrees of freedom. Calculate the F statistic and assess its significance.

Pearson's Correlation Coefficient ($r$)

The square root of $r^2$, or $r$—known as Pearson's correlation coefficient—is most often used as a measure of association between two interval-ratio variables:

$$r = \sqrt{r^2}$$

Pearson's $r$ is usually computed directly by using the following definitional formula:

$$r^2 = \frac{\text{Covariance}(X,Y)^2}{\text{Standard deviation (X)} \times \text{Standard deviation (Y)}}$$

Thus, $r$ is defined as the ratio of the covariance of $X$ and $Y$ to the product of the standard deviations of $X$ and $Y$.

**CHARACTERISTICS OF PEARSON’S $r$**

Pearson's $r$ is a measure of relationship or association for interval-ratio variables. Like gamma (introduced in Chapter 10), it ranges from 0.0 to ±1.0, with 0.0 indicating no association between the two variables. An $r$ of +1.0 means that the two variables have a perfect positive association; –1.0 indicates that it is a perfect negative association. The absolute value of $r$ indicates the strength of the linear association between two variables. (Refer to A Closer Look 10.2 for an interpretational guide.) Thus, a correlation of –0.75 demonstrates a stronger association than a correlation of 0.50. Figure 12.9 illustrates a strong positive relationship, a strong negative relationship, a moderate positive relationship, and a weak negative relationship.

Unlike the $b$ coefficient, $r$ is a symmetrical measure. That is, the correlation between $X$ and $Y$ is identical to the correlation between $Y$ and $X$.

To calculate $r$ for our example of the relationship between educational attainment and Internet hours per week, let’s return to Table 12.2, where the covariance and the standard deviations for $X$ and $Y$ have already been calculated:

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{14.22}{4.80(3.03)} = 0.98$$

A correlation coefficient of 0.98 indicates that there is a strong positive linear relationship between median household income and the percentage of state residents with a bachelor’s degree.

Note that we could have taken the square root of $r^2$ to calculate $r$, because $r = \sqrt{r^2}$ or $\sqrt{0.96} = 0.98$. Similarly, if we first calculate $r$, we can obtain $r^2$ simply by squaring $r$ (be careful not to lose the sign of $r^2$).
Thus far, we have used examples that involve only two interval-ratio variables: (1) a dependent variable and (2) an independent variable. Multiple regression is an extension of bivariate regression, allowing us to examine the effect of two or more independent variables on the dependent variable.\(^6\)

The general form of the multiple regression equation involving two independent variables is

$$
\hat{Y} = a + b_1 X_1 + b_2 X_2
$$

(12.11)

where

- \(\hat{Y}\) = the predicted score on the dependent variable
- \(X_1\) = the score on independent variable \(X_1\)
- \(X_2\) = the score on independent variable \(X_2\)
- \(a\) = the \(Y\)-intercept, or the value of \(Y\) when both \(X_1\) and \(X_2\) are equal to zero
- \(b_1\) = the partial slope of \(Y\) and \(X_1\), the change in \(Y\) with a unit change in \(X_1\), when the other independent variable \(X_2\) is controlled
- \(b_2\) = the partial slope of \(Y\) and \(X_2\), the change in \(Y\) with a unit change in \(X_2\), when the other independent variable \(X_1\) is controlled

**Figure 12.9** Scatter Diagrams Illustrating Weak, Moderate, and Strong Relationships as Indicated by the Absolute Value of \(r\)

- \(r = 0.82\), strong positive relationship
- \(r = -0.82\), strong negative relationship
- \(r = 0.52\), moderate positive relationship
- \(r = -0.22\), weak negative relationship
It is important to note that the existence of a correlation only denotes that the two variables are associated (they occur together or covary) and not that they are causally related. The well-known phrase “correlation is not causation” points to the fallacy of inferring that one variable causes the other based on the correlation between the variables. Such relationship is sometimes said to be spurious because both variables are influenced by a causally prior control variable, and there is no causal link between them. We can also say that a relationship between the independent and dependent variables is confounded by a third variable.

There are numerous examples in the research literature of spurious or confounded relationships. For instance, in a 2004 article, Michael Benson and his colleagues discuss the issue of domestic violence as a correlate of race. Studies and reports have consistently found that rates of domestic abuse are higher in communities with a higher percentage of African American residents. Would this correlation indicate that race and domestic violence are causally related? To suggest that African Americans are more prone to engage in domestic violence would be erroneous if not outright racist. Benson and colleagues argue that the correlation between race and domestic violence is confounded by the level of economic distress in the community. Economically distressed communities are typically occupied by a higher percentage of African Americans. Also, rates of domestic violence tend to be higher in such communities. We can say that the relationship between race and domestic violence is confounded by a third variable—level of economic distress in the community.

Similarly, to test for the confounding effect of community economic distress on the relationship between race and domestic violence, Benson and Fox calculated rates of domestic violence for African Americans and whites in communities with high and low levels of economic distress. They found that the relationship between race and domestic violence is not significant when the level of economic distress is constant. That is, the difference in the base rate of domestic violence for African Americans and whites is reduced by almost 50% in communities with high distress levels. In communities with low distress level (and high income), the rate of domestic violence of African Americans is virtually identical to that of whites. The results showed that the correlation between race and domestic violence is accounted for in part by the level of economic distress of the community.

Uncovering spurious or confounded relations between an independent and a dependent variable can also be accomplished by using multiple regression. Multiple regression, an extension of bivariate regression, helps us examine the effect of an independent variable on a dependent variable while holding constant one or more additional variables.

\[ b \] = the partial slope of \( Y \) and \( X_i \), the change in \( Y \) with a unit change in \( X_i \), when the other independent variable \( X_j \) is controlled

Notice how the slopes are referred to as partial slopes. Partial slopes reflect the amount of change in \( Y \) for a unit change in a specific independent variable while controlling or holding constant the value of the other independent variables.

To illustrate, let’s combine our investigation of Internet hours per week, educational attainment and age. We hypothesize that individuals with higher levels of education will have higher levels of Internet use per week and that older individuals have lower hours of Internet use. We will estimate the multiple regression model data using SPSS. SPSS output are presented in Figure 12.10.
The partial slopes are reported in the Coefficients table, under the column labeled B. The intercept is also in Column B, on the (Constant) row. Putting it all together, the multiple regression equation that incorporates both educational attainment and respondent age as predictors of Internet hours per week is

\[
\hat{Y} = -0.605 + 0.491(X_1) + -0.057(X_2)
\]

where

\[
\hat{Y} = \text{number of Internet hours per week} \\
X_1 = \text{educational attainment} \\
X_2 = \text{age}
\]

This equation tells us that Internet hours increases by 0.49 per each year of education \((X_1)\), holding age \((X_2)\) constant. On the other hand, Internet hours decreases by 0.06 with each year increase in age \((X_2)\) when we hold educational attainment \((X_1)\) constant. Controlling for the effect of one variable, while examining the effect of the other, allows us to separate out the effects of each predictor independently of the other. For example, given two individuals with the same years of education, the person who might be a year older than the other is expected to use Internet 0.06 hours less. Or given two individuals of the same age, the person who has one more year of education will have 0.49 hours more of Internet use than the other.

Finally, the value of \(a (-0.60)\) reflects Internet hours per week when both education and age are equal to zero. Though this \(Y\)-intercept doesn’t lend itself to a meaningful interpretation, the value of \(a\) is a baseline that must be added to the equation for Internet hours to be properly estimated.
When a regression model includes more than one independent variable, it is likely that the units of measurement will vary. A multiple regression model could include income (dollars), highest degree (years), and number of children (individuals), making it difficult to compare their effects on the dependent variable. The **standardized slope coefficient or beta** (represented by the Greek letter, \( \beta \)) converts the values of each score into a Z score, standardizing the units of measurement so we can interpret their relative effects. Beta, also referred to as beta weights, range from 0 to ±1.0. The largest \( \beta \) value (whether negative or positive) identifies the independent variable with the strongest effect. Beta is reported in the SPSS Coefficient table, under the column labeled “Standardized Coefficient/Beta.”

A standardized multiple regression equation can be written as

\[
\hat{Y} = a + \beta_1X_1 + \beta_2X_2
\]

Based on this data example, the equation is

\[
\hat{Y} = -0.605 + 0.779X_1 + -0.245X_2
\]

We can conclude that education has the strongest effect on Internet hours—education—as indicated by the \( \beta \) value of 0.779 (compared with the other beta of −0.245 for age).

Like bivariate regression, multiple regression analysis yields a **multiple coefficient of determination**, symbolized as \( R^2 \) (corresponding to \( r^2 \) in the bivariate case). \( R^2 \) measures the PRE that results from using the linear regression model. It reflects the proportion of the total variation in the dependent variable that is explained jointly by two or more independent variables. We obtained an \( R^2 \) of 0.977 (in the Model Summary table, in the column labeled R Square). This means that by using educational attainment and age, we reduced the error of predicting Internet hours by 97.7 or 98%. We can also say that the independent variables, educational attainment and age, explain 98% of the variation in Internet hours per week.

Including respondent age in our regression model did not improve the prediction of Internet hours per week. As we saw earlier, educational attainment accounted for 96% of the variation in Internet hours per week. The addition of age to the prediction equation resulted in a 2% increase in the percentage of explained variation.

As in the bivariate case, the square root of \( R^2 \), or \( R \), is **Pearson’s multiple correlation coefficient**. It measures the linear relationship between the dependent variable and the combined effect of two or more independent variables. For our model \( R = 0.988 \) or 0.99. This indicates that there is a strong relationship between the dependent variable and both independent variables.

**LEARNING CHECK**

Use the prediction equation describing the relationship between Internet hours per week and both educational attainment and age to calculate Internet hours per week for someone with 20 years of education who is 35 years old.

SPSS can also produce a correlation matrix, a table that presents the Pearson’s correlation coefficient for all pairs of variables in the multiple regression model. A correlation matrix
provides a baseline summary of the relationships between variables, identifying relationships or hypotheses that are usually the main research objective. Extensive correlation matrices are often presented in social science literature, but in this example, we have three pairs: (1) Internet hours with educational attainment, (2) Internet hours with age, and (3) educational attainment with age. Refer to Figure 12.11.

The matrix reports variable names in columns and rows. Note the diagonal from upper left corner to the lower right corner reporting a correlation value of 1 (there are three 1s). This is the correlation of each variable with itself. This diagonal splits the matrix in half, creating mirrored correlations. We’re interested in the intersection of the row and column variables, the cells that report their correlation coefficient for each pair. For example, the correlation coefficient for Internet and age, \(-0.878\), is reported twice at the upper right-hand corner and at the lower left-hand corner. The other two correlations are also reported twice.

We calculated the correlation coefficients for Internet hours with educational attainment and Internet hours with age earlier in this chapter. The negative correlation between Internet hours and age is confirmed in Figure 12.10. We conclude that there is a strong negative relationship \((-0.813\) between these two variables. We also know that there is a strong positive correlation of 0.978 between Internet hours and educational attainment. The matrix also reports the significance of each correlation.

### ANOVA FOR MULTIPLE LINEAR REGRESSION

The ANOVA summary table for multiple regression is nearly identical to the one for bivariate linear regression, except that the degrees of freedom are adjusted to reflect the number of independent variables in the model.

We conducted an ANOVA test to assess the probability that the linear relationship between Internet hours per week, educational attainment, and age as expressed by \(R^2\), is really zero. The results of this test are reported in Figure 12.10. The obtained \(F\) statistic of 147.87 is
shown in this table. With 2 and 7 degrees of freedom, we would need an $F$ of 9.55 to reject the null hypothesis that $R^2 = 0$ at the .01 level. Since our obtained $F$ exceeds that value (147.87 > 9.55), we can reject the null hypothesis with $p < .01$.

**READING THE RESEARCH LITERATURE**

**Academic Intentions and Support**

Katherine Purswell, Ani Yazedjian, and Michelle Toews (2008)\(^7\) utilized regression analysis to examine academic intentions (intention to perform specific behaviors related to learning engagement and positive academic behaviors), parental support, and peer support as predictors of self-reported academic behaviors (e.g., speaking in class, completed assignments on time during their freshman year) of first- and continuing-generation college students. The researchers apply social capital theory, arguing that relationships with others (parents and peers) would predict positive academic behaviors.

They estimated three separate multiple regression models for first-generation students (Group 1), students with at least one parent with college experience but with no degree (Group 2), and students with at least one parent with a bachelor’s degree or higher (Group 3). The regression models are presented in Table 12.6. All of the variables included in the analysis are ordinal measures, with responses coded on a *strongly disagree* to *strongly agree* scale.

Each model is presented with partial and standardized slopes. No intercepts are reported. The multiple correlation coefficient and $F$ statistic are also reported for each model. The asterisk indicates significance at the .05 level.

The researchers summarize the results of each model.

| Table 12.6 Regression Analyses Predicting Behavior by Intention, Parental Support, and Peer Support |
|-----------------------------------------------|---------------------|---------------|------------------|---------------|
| First Generation Students N = 44                      | Group 2 N = 82       | Group 3 N = 203 |
| Intention                                             | $b$ | $\beta$ | $b$ | $\beta$ | $b$ | $\beta$ |
| Parental Support                                      | .00 | .00   | .07 | .04   | .06* | .13 |
| Peer Support                                          | −.02 | −.02 | .26* | .26 | −.20* | −.16 |
| $R^2$                                                 | .24 | .18   | .23 |
| $F$                                                   | 3.82* | 5.77* | 18.74* |


*p < .05.*
The regression model was significant for all three groups ($p < .05$). For FGCS (first generation college students), the model predicted $24\%$ of the variance in behavior. However, intention was the only significant predictor for this group. For the second group, the model predicted $18\%$ of the variance, with peer support significantly predicting academic behavior. Finally, the model predicted $23\%$ of the variance in behavior for those in the third group with all three independent variables—intention, parental support, and peer support—predicting academic behavior.8

DATA AT WORK

**Shinichi Mizokami: Professor**

Dr. Mizokami is a professor of psychology and pedagogy at Kyoto University, Japan. Pedagogy is a discipline that examines educational theories and teaching methods. His current research involves two areas of study: (1) student learning and development and (2) identity formation in adolescence and young adulthood.

In 2013, his research team launched a 10-year transition survey with $45,000$ second-year high school students. He uses multiple regression techniques to examine students’ transition from school to work. “My team administers the surveys with the questions regarding what attitudes and distinctions competent students have or what activities they are engaged in. We analyze the data controlling the variables of gender, social class, major, kinds of university (doctoral, master’s, or baccalaureate university), and find the results. In the multiple regression analysis, we carefully look at the bivariate tables and correlations between the used variables, and go back and forth between those descriptive statistics and the results of the multiple regression analysis.”

He would be pleased to learn that you are enrolled in an undergraduate statistics course. According to Mizokami, “Many people will not have enough time to learn statistics after they start to work, so it may be worthwhile to study it in undergraduate education. Learning statistics can expand the possibilities of your job and provide many future advantages. . . . This can happen not only in academic fields but also in business. Good luck!”

MAIN POINTS

- A scatter diagram (also called scatterplot) is a quick visual method used to display relationships between two interval-ratio variables.
- Equations for all straight lines have the same general form:
  $$\hat{Y} = a + b(X)$$
- The best-fitting regression line is that line where the residual sum of squares, or $\Sigma e^2$, is at a minimum. Such a line is called the least squares line, and the technique that produces this line is called the least squares method.
- The coefficient of determination ($r^2$) and Pearson’s correlation coefficient ($r$) measure how well the regression model fits the data. Pearson’s $r$ indicates the strength of the association between the two variables. The coefficient of determination is a $PRE$ measure, identifying the reduction of error based on the regression model.
The general form of the multiple regression equation involving two independent variables is \( \hat{Y} = a + b_1X_1 + b_2X_2 \). The multiple coefficient of determination \( (R^2) \) measures the proportional reduction of error based on the multiple regression model.

The standardized multiple regression equation is \( \hat{Y} = \alpha + \beta_1X_1 + \beta_2X_2 \). The beta coefficients allow us to assess the relative strength of all the independent variables.

KEY TERMS

- bivariate regression
- coefficient of determination \( (r^2) \)
- correlation
- deterministic (perfect) linear relationship
- least squares line (best-fitting line)
- least squares method
- linear relationship
- mean squares regression
- mean squares residual
- multiple coefficient of determination \( (R^2) \)
- multiple regression
- partial slopes \( (b^*) \)
- Pearson's correlation coefficient \( (r) \)
- Pearson's multiple correlation coefficient \( (R) \)
- regression
- regression sum of squares \( (SSR) \)
- residual sum of squares \( (SSE) \)
- scatter diagram
- slope \( (b) \)
- standardized slope coefficient or beta
- \( Y \)-intercept \( (a) \)

Sharpen your skills with SAGE edge at http://edge.sagepub.com/frankfort8e. SAGE edge for students provides a personalized approach to help you accomplish your coursework goals in an easy-to-use learning environment.

SPSS DEMONSTRATIONS [GSS14SSDS-A]

Demonstration 1: Producing Scatterplots (Scatter Diagrams)

Do people with more education work more hours per week? Some may argue that those with lower levels of education are forced to work low-paying jobs, thereby requiring them to work more hours per week to make ends meet. Others may rebut this argument by saying those with higher levels of education are in greater positions of authority, which requires more time to ensure operations run smoothly. This question can be explored with SPSS using the techniques discussed in this chapter for interval-ratio data because \textit{hours worked last week} (HRS1) and \textit{number of years of education} (EDUC) are both coded at an interval-ratio level in the GSS14SSDS-A file.

We begin by looking at a scatterplot of these two variables. The Scatter procedure can be found under the \textit{Graphs} menu choice. In the opening dialog box, click Legacy Dialogs then Scatter/Dot (which means we want to produce a standard scatterplot with two variables), select the icon for Simple Scatter, and then click Define.

The Scatterplot dialog box requires that we specify a variable for both the \( X \)- and \( Y \)-axes. We place EDUC (number of years of education) on the \( X \)-axis because we consider it the independent variable and HRS1 (number of hours worked last week) on the \( Y \)-axis because it is the dependent variable. Then, click OK.
You can edit it to change its appearance by double-clicking on the chart in the viewer. The action of double-clicking displays the chart in a chart window. You can edit the chart from the menus, from the toolbar, or by double-clicking on the object you want to edit.

It is difficult to tell whether a relationship exists just by looking at points in the scatterplot, so we will ask SPSS to include the regression line. To add a regression line to the plot, we start by double-clicking on the scatterplot to open the Chart Editor. Click **Elements** from the main menu, then **Fit Line at Total**. In the section of the dialog box headed “Fit Method,” select **Linear**. Click **Apply** and then **Close**. Finally, in the Chart Editor, click **File** and then **Close**. The result of these actions is shown in Figure 12.12.

Since the regression line clearly rises as number of years of education increases, we observe the positive relationship between education and number of hours worked last week. The predicted value for those with 20 years of education is about 44 hours, compared with 39.76 hours for those with 10 years of education. However, because there is a lot of scatter around the line (the points are not close to the regression line), the predictive power of the model is weak.

**Demonstration 2: Producing Correlation Coefficients**

To further quantify the effect of education on hours worked, we request a correlation coefficient. This statistic is available in the Bivariate procedure, which is located by clicking on **Analyze**, **Correlate**, then **Bivariate** (Figure 12.13). Place the variables you are interested in correlating, EDUC and HRS1, in the Variable(s) box, then click **OK**.
SPSS produces a matrix of correlations, shown in Figure 12.14. We are interested in the correlation in the bottom left-hand cell, .084. The correlation is significant at the .05 level (two-tailed). We see that this correlation is closer to 0 than to 1, which tells us that education is not a very good predictor of hours worked, even if it is true that those with more education work more hours per week. The number under the correlation coefficient, 895, is the number of valid cases (N)—those respondents who gave a valid response to both questions. The number is reduced because not everyone in the sample is working.
Demonstration 3: Producing a Regression Equation

Next, we will use SPSS to calculate the best-fitting regression line and the coefficient of determination. This procedure is located by clicking on Analyze, Regression, then Linear. The Linear Regression dialog box (Figure 12.15) provides boxes in which to enter the dependent variable, HRS1, and the independent variable, EDUC (regression allows more than one). After you place the variables in their appropriate places, click OK to generate the output. The Linear Regression dialog box offers many other choices, but the default output from the procedure contains all that we need.

SPSS produces a great deal of output, which is typical for many of the more advanced statistical procedures in the program. The output is presented in Figure 12.16. Under the Model Summary, the coefficient of determination is labeled “R square.” Its value is .007, which is very weak. Educational attainment explains little of the variation in hours worked, less than 1%.

The regression equation coefficients are presented in the Coefficients table. The regression equation coefficients are listed in the column headed “B.” The coefficient for EDUC, or $b$, is about .417; the intercept term, or $a$, identified in the “(Constant)” row, is 35.589. Thus, we would predict that every additional year of education increases the number of hours worked each week by about 25 minutes. Or we could predict that those with a high school level of education work, on average, 35.589 + (.417)(12) hours, or 40.59 hours.

The ANOVA table provides the results of the analysis of variance test. The table includes regression and residual sum of squares, as well as mean squares. To test the null hypothesis that $r^2$ is zero, you will only need the statistic shown in the last column labeled “Sig.” This is the $p$ value associated with the $F$ ratio listed in the column head “$F$.” The $F$ statistic is 6.368, and its associated $p$ value is .012. This means that there is a little probability (.012) that $r^2$ is really zero in the population, given the observed $r^2$ of .007. The model, though not reducing much of the variance in predicting work hours, is significant. We are therefore able to reject the null hypothesis at the .05 level.
Demonstration 4: Producing a Multiple Regression Equation

What other variables, in addition to education, affect the number of hours worked per week? One possible answer to this question is that age (AGE) has something to do with the number of hours worked per week. To answer this question, we will use SPSS to calculate a multiple regression equation and a multiple coefficient of determination. This procedure is similar to the one used to generate the bivariate regression equation. Click Analyze, Regression, then Linear. We place EDUC (number of years of education) and AGE (age in years) in the box for the independent variables and HRS1 (the number of hours worked last week) in the box for the dependent variable, and click OK. The output is presented in Figure 12.17.

Under the Model Summary, the multiple correlation coefficient labeled “R” is .109. This tells us that education and age are weakly associated with hours worked last week. The coefficient of determination is labeled “R square.” Its value is .012. An $R^2$ of .012 means that educational attainment and age jointly explain just 1% of the variation in hours worked last week. In addition, SPSS provides an “adjusted R square,” which is .01. The “adjusted R square” adjusts the $R^2$ coefficient for the number of predictors in the equation. Generally, the adjusted $R^2$ will be lower, relative to $R^2$, the larger the number of predictors.

The regression equation coefficients are listed in the Coefficients table. The regression equation coefficients are listed in the column headed “B.” The coefficient for EDUC is about .433, and for AGE it is −.077. The intercept term, or $a$, identified in the “(Constant)” row, is 38.815. Thus, we would predict that, holding age constant, every additional year of education increases the number of hours worked the previous week by about 26 minutes (0.43 × 60).
1. Explore the relationship between the number of siblings a respondent has (SIBS) and his or her number of children (CHILDS).
   a. Construct a scatterplot of these two variables in SPSS, and place the best-fit linear regression line on the scatterplot. Describe the relationship between the number of siblings a respondent has (IV) and the number of his or her children (DV).
   b. Calculate the regression equation predicting CHILDS with SIBS. What are the intercept and the slope? What are the coefficient of determination and the correlation coefficient?
   c. What is the predicted number of children for someone with three siblings?
   d. What is the predicted number of children for someone without any siblings?

2. Use the same variables as in Exercise 1, but do the analysis separately for men and women. Begin by locating the variable SEX. Click Data, Split File, and then select Organize Output by Groups. Insert SEX into the box and click OK. Now, SPSS will split your results by sex.
   a. Calculate the regression equation for men and women. (Note: You will need to scroll down through your output to find the results for men and women.) How similar are they?
b. What is the predicted number of children for a man with six siblings? For a woman with the same number of siblings? Which group has the higher predicted number of children?

3. Use the same variables as in Exercise 1, but do the analysis separately for white and black respondents. Click Data, Split File, and then select Organize Output by Groups. Insert RACECEN1 into the box and click OK. SPSS will split your results by RACECEN1 (focusing your analysis only on the categories for whites and blacks).
   a. Is there any difference between the regression equations for whites and blacks?
   b. What is the predicted number for whites and blacks with the same number of siblings: one sibling, four siblings, and seven siblings?

4. Use the same variables as in Exercise 1, but do the analysis separately for married and divorced respondents. Begin by locating the variable MARITAL. Click Data, Split File, and then select Organize Output by Groups. Insert MARITAL into the box and click OK. SPSS will split your results by marital status.
   a. Is there any difference between the regression equations for married and divorced respondents?
   b. What is the predicted number of children for married and divorced respondents with the following number of siblings: one sibling, four siblings, and seven siblings?
   c. What differences, if any, do you find? Is the number of siblings a better predictor of number of children for married respondents or for women?

5. Investigate the relationship between the respondent’s education (EDUC) and the education received by his or her father and mother (PAEDUC and MAEDUC, respectively).
   a. Calculate the correlation coefficient, the coefficient of determination, and the regression equation predicting the respondent’s education with father’s education only. Interpret your results.
   b. Determine the multiple correlation coefficient, the multiple coefficient of determination, and the regression equation predicting the respondent’s education with father’s and mother’s education. Interpret your results.
   c. Did taking into account the respondent’s mother’s education improve our prediction? Discuss this on the basis of the results from 5b.
   d. Using the regression equation from 5a, calculate the predicted number of years of education for a person with a father with 12 years of education. Then, repeat this procedure, adding in a mother’s 12 years of education and using the regression equation from 5b.
   e. Review the ANOVA results. Can you reject the null hypothesis that $R^2 = 0$?

CHAPTER EXERCISES

1. Concerns over climate change, pollution, and a growing population has led to the formation of social action groups focused on environmental policies nationally and around the globe. A large number of these groups are funded through donor support. Based on the following eight countries, examine the data to determine the extent of the relationship between simply being concerned about the environment and actually giving money to environmental groups.
<table>
<thead>
<tr>
<th>Country</th>
<th>Percentage Concerned</th>
<th>Percentage Donating Money</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>33.8</td>
<td>22.8</td>
</tr>
<tr>
<td>Austria</td>
<td>35.5</td>
<td>27.8</td>
</tr>
<tr>
<td>The Netherlands</td>
<td>30.1</td>
<td>44.8</td>
</tr>
<tr>
<td>Slovenia</td>
<td>50.3</td>
<td>10.7</td>
</tr>
<tr>
<td>Russia</td>
<td>29.0</td>
<td>1.6</td>
</tr>
<tr>
<td>Philippines</td>
<td>50.1</td>
<td>6.8</td>
</tr>
<tr>
<td>Spain</td>
<td>35.9</td>
<td>7.4</td>
</tr>
<tr>
<td>Denmark</td>
<td>27.2</td>
<td>22.3</td>
</tr>
</tbody>
</table>

**Source:** International Social Survey Programme, 2000.

a. Construct a scatterplot of the two variables, placing percentage concerned about the environment on the horizontal or X-axis and the percentage donating money to environmental groups on the vertical or Y-axis.

b. Does the relationship between the two variables seem linear? Describe the relationship.

c. Find the value of the Pearson correlation coefficient that measures the association between the two variables and offer an interpretation.

2. In this exercise, we will investigate the relationships between adolescent fertility rate and female labor force participation in South America. Data are presented for 2014.

<table>
<thead>
<tr>
<th>Country</th>
<th>Adolescent Fertility Rate</th>
<th>Female Labor Force Participation Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>40.4</td>
<td>63.9</td>
</tr>
<tr>
<td>Bolivia</td>
<td>44.6</td>
<td>71.1</td>
</tr>
<tr>
<td>Brazil</td>
<td>43.8</td>
<td>67.3</td>
</tr>
<tr>
<td>Chile</td>
<td>40.8</td>
<td>48.1</td>
</tr>
<tr>
<td>Colombia</td>
<td>42.6</td>
<td>51.7</td>
</tr>
<tr>
<td>Ecuador</td>
<td>40.4</td>
<td>76.2</td>
</tr>
<tr>
<td>Paraguay</td>
<td>39.2</td>
<td>58.0</td>
</tr>
<tr>
<td>Peru</td>
<td>45.3</td>
<td>49.7</td>
</tr>
<tr>
<td>Uruguay</td>
<td>44.5</td>
<td>56.5</td>
</tr>
<tr>
<td>Venezuela</td>
<td>39.9</td>
<td>79.7</td>
</tr>
</tbody>
</table>

3. Construct a scatterplot for adolescent fertility rate and labor force participation rate. Do you think the scatterplot can be characterized by a linear relationship?

b. Calculate the coefficient of determination and correlation coefficient.

c. Describe the relationship between the variables based on your calculations.

3. Let’s examine the relationship between a country’s gross national product (GNP) and the percentage of respondents willing to pay higher prices for goods to protect the environment. The following table displays information for five countries selected at random.

a. Calculate the correlation coefficient between a country’s GNP and the percentage of its residents willing to pay higher prices to protect the environment. What is its value?

b. Provide an interpretation for the coefficient.

<table>
<thead>
<tr>
<th>Country</th>
<th>GNP per Capita</th>
<th>Percentage Willing to Pay</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>29.24</td>
<td>44.9</td>
</tr>
<tr>
<td>Ireland</td>
<td>18.71</td>
<td>53.3</td>
</tr>
<tr>
<td>The Netherlands</td>
<td>24.78</td>
<td>61.2</td>
</tr>
<tr>
<td>Norway</td>
<td>34.31</td>
<td>40.7</td>
</tr>
<tr>
<td>Sweden</td>
<td>25.58</td>
<td>32.6</td>
</tr>
</tbody>
</table>


4. In 2010, a U.S. Census Bureau report revealed that approximately 14.3% of all Americans were living below the poverty line in 2009. This figure is higher than in 2000, when the poverty rate was 12.2%. Individuals and families living below the poverty line face many obstacles, the least of which is access to health care. In many cases, those living below the poverty line are without any form of health insurance. Using data from the U.S. Census Bureau, analyze the relationship between living below the poverty line and access to health care for a random sample of 12 states. (The health insurance data are pre-Affordable Care Act implementation.)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>17.9</td>
<td>16.9</td>
</tr>
<tr>
<td>California</td>
<td>14.2</td>
<td>20</td>
</tr>
<tr>
<td>Idaho</td>
<td>14.3</td>
<td>15.2</td>
</tr>
<tr>
<td>Louisiana</td>
<td>17.3</td>
<td>16</td>
</tr>
<tr>
<td>New Jersey</td>
<td>9.4</td>
<td>15.8</td>
</tr>
<tr>
<td>New York</td>
<td>14.2</td>
<td>14.8</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>12.5</td>
<td>11.4</td>
</tr>
<tr>
<td>Rhode Island</td>
<td>11.5</td>
<td>12.3</td>
</tr>
</tbody>
</table>

(Continued)
(Continued)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>South Carolina</td>
<td>17.1</td>
<td>17.0</td>
</tr>
<tr>
<td>Texas</td>
<td>17.2</td>
<td>26.1</td>
</tr>
<tr>
<td>Washington</td>
<td>12.3</td>
<td>12.9</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>12.4</td>
<td>9.5</td>
</tr>
</tbody>
</table>


a. Construct a scatterplot, predicting the percentage without health insurance with the percentage living below the poverty level. Does it appear that a straight-line relationship will fit the data?

b. Calculate the regression equation with percentage of the population without health insurance as the dependent variable, and draw the regression line on the scatterplot. What is its slope? What is the intercept? Has your opinion changed about whether a straight line seems to fit the data? Are there any states that fall far from the regression line? Which one(s)?

5. We test the hypothesis that as an individual’s years of education increases, the individual will have fewer children. Based on a subsample from the GSS 2014, we present a scatterplot and regression output for the variables EDUC and CHILDS. Interpret the results.

Scatterplot of Number of Children by Education

\[ y = 3.54 - 0.12X \]

R² Linear=0.048
6. We present SPSS output examining the relationship between education (measured in years) and television viewing per day (measured in hours) based on a GSS 2014 subsample. We hypothesize that as educational attainment increases, hours of television viewing will decrease, indicating a negative relationship between the two variables. Discuss the significance of the overall model based on $F$ and its $p$ values. Is the relationship between education and television viewing significant?
7. Based on the following SPSS output describe the regression model for educational attainment and amount of money given to charity based on GSS 2014.

a. Assess the significance of the overall model based on its $F$ and $p$ values. What is the relationship between the two variables?

b. Calculate the predicted charitable amount for a respondent with 14 years of education and for a respondent with 20 years of education.

8. Research on social mobility, status, and educational attainment has provided convincing evidence on the relationship between parents’ and children’s socioeconomic achievement. The GSS 2014 measures the educational level of respondents and their mothers. Use the scatterplot and regression output to describe the relationship between mothers’ education and respondent’s education.
Scatterplot of Respondent Level of Education by Mother's Level of Education

\[ Y = 9.78 + 0.36^*X \]

R\(^2\) linear = 0.212

Linear Regression Output Specifying the Relationship Between Respondent's Education by Mother's Education

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.461</td>
<td>.212</td>
<td>.212</td>
<td>2.758</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), maeduc HIGHEST YEAR SCHOOL COMPLETED, MOTHER

ANOVA*

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>2188.518</td>
<td>1</td>
<td>2188.518</td>
<td>287.745</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>8122.994</td>
<td>1068</td>
<td>7.606</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>10311.522</td>
<td>1069</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Dependent Variable: educ HIGHEST YEAR OF SCHOOL COMPLETED
b. Predictors: (Constant), maeduc HIGHEST YEAR SCHOOL COMPLETED, MOTHER

table

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>maeduc HIGHEST YEAR SCHOOL COMPLETED, MOTHER</td>
<td>9.784</td>
<td>.259</td>
<td>.37731</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>.360</td>
<td>.021</td>
<td>.461</td>
<td>16.963</td>
</tr>
</tbody>
</table>

a. Dependent Variable: educ HIGHEST YEAR OF SCHOOL COMPLETED

Copyright ©2017 by SAGE Publications, Inc.

This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.
9. We further explore the relationship between respondent’s education and mother’s education, computing regression models separately for males and females.
   a. Calculate the regression equation for each.
   b. What is the predicted value of respondent’s education when mother’s education is 20 years?
   c. For which gender group is the relationship between respondent’s education and mother’s education strongest? Explain.

---

**Linear Regression Output Specifying the Relationship Between Respondent Level of Education by Mother’s Level of Education: Males Only**

*Model Summary*

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.450</td>
<td>.202</td>
<td>.201</td>
<td>2.834</td>
</tr>
</tbody>
</table>

- a. sex RESPONDENTS SEX = 1 MALE
- b. Predictors: (Constant), maeduc HIGHEST YEAR SCHOOL COMPLETED, MOTHER

*ANOVA*

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>976.149</td>
<td>1</td>
<td>976.149</td>
<td>121.55</td>
<td>.000</td>
</tr>
<tr>
<td>Residual</td>
<td>3846.500</td>
<td>479</td>
<td>8.030</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4822.649</td>
<td>480</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- a. sex RESPONDENTS SEX = 1 MALE
- b. Dependent Variable: educ HIGHEST YEAR OF SCHOOL COMPLETED
- c. Predictors: (Constant), maeduc HIGHEST YEAR SCHOOL COMPLETED, MOTHER

*Coefficients*

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>9.768</td>
</tr>
<tr>
<td>maeduc HIGHEST YEAR SCHOOL COMPLETED, MOTHER</td>
<td>.355</td>
<td>.032</td>
</tr>
</tbody>
</table>

- a. sex RESPONDENTS SEX = 1 MALE
- b. Dependent Variable: educ HIGHEST YEAR OF SCHOOL COMPLETED
In Exercise 6, we examined the relationship between years of education and hours of television watched per day. We saw that as education increases, hours of television viewing decreases. The number of children a family has could also affect how much television is viewed per day. Having children may lead to more shared and supervised viewing and thus increases the number of viewing hours. The following SPSS output displays the relationship between television viewing (measured in hours per day) and both education (measured in years) and number of children. We hypothesize that whereas more education may lead to less viewing, the number of children has the opposite effect: Having more children will result in more hours of viewing per day.
a. What is the \( b \) coefficient for education? For number of children? Interpret each coefficient. Is the relationship between each independent variable and hours of viewing as hypothesized?

b. Using the multiple regression equation with both education and number of children as independent variables, calculate the number of hours of television viewing for a person with 16 years of education and two children. Using the equation from Exercise 6, how do the results compare between a person with 16 years of education (number of children not included in the equation) and a person with 16 years of education with two children?

c. Compare the \( r^2 \) value from Exercise 6 with the \( R^2 \) value from this regression. Does using education and number of children jointly reduce the amount of error involved in predicting hours of television viewed per day?

---

11. We return to our chapter analysis of Internet hours per week (WWHR), educational attainment (EDUC), and respondent age (AGE), presenting the multiple regression model and correlation matrix based on GSS 2014 data.

a. What is the \( b \) coefficient for education? For age? Interpret each coefficient. Is the relationship between education and Internet hours as hypothesized in our chapter example? For age and Internet hours?

b. Using the multiple regression equation with both education and age as independent variables, calculate the number of Internet hours per week for a person with 16 years of education and 55 years of age.

---
c. Using the standardized multiple regression equation, identify which independent variable has the strongest effect on Internet hours per week.

d. Interpret the multiple coefficient of determination.

e. Interpret each correlation coefficient (based on the correlation matrix).
We revisit Katherine Purswell, Ani Yazedjian, and Michelle Toews (2008)’ research regarding the relationship between academic intentions (intention to perform specific behaviors related to learning engagement and positive academic behaviors), parental support, and peer support and self-reported academic behaviors (e.g., speaking in class, completed assignments on time during their freshman year) of first- and continuing-generation college students.

They estimated three separate models for first-generation students (Group 1), students with at least one parent with college experience but with no degree (Group 2), and students with at least one parent with a bachelor’s degree or higher (Group 3). The correlation matrix is presented below.

All of the variables included in the analysis are ordinal measures, with responses coded on a strongly disagree to strongly agree scale.

<table>
<thead>
<tr>
<th>Intercorrelations Between Variables Based on Parental Education Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td><strong>First Generation Students (n = 44)</strong></td>
</tr>
<tr>
<td>1. Intention</td>
</tr>
<tr>
<td>2. Parental Support</td>
</tr>
<tr>
<td>3. Peer Support</td>
</tr>
<tr>
<td>4. Behavior</td>
</tr>
<tr>
<td><strong>Group 2 (n = 82)</strong></td>
</tr>
<tr>
<td>1. Intention</td>
</tr>
<tr>
<td>2. Parental Support</td>
</tr>
<tr>
<td>3. Peer Support</td>
</tr>
<tr>
<td>4. Behavior</td>
</tr>
<tr>
<td><strong>Group 3 (n = 203)</strong></td>
</tr>
<tr>
<td>1. Intention</td>
</tr>
<tr>
<td>2. Parental Support</td>
</tr>
<tr>
<td>3. Peer Support</td>
</tr>
<tr>
<td>4. Behavior</td>
</tr>
</tbody>
</table>


*p < .05, **p < .01.

a. Which group has the most significant correlations? Which group has the least?

b. Interpret the correlation for intention and behavior for the three groups. For which group is the relationship the strongest?

c. The correlation for peer support and intention is highest for which group? Explain.
We expand on the model presented in Exercise 10, adding work hours (HRS1) and respondent age (AGE) as independent variables. Data are based on the GSS 2014.

a. Write the multiple regression equation for the model. Interpret the slope for each independent variable.

b. Based on their beta scores, rank the independent variables according to the strength of their effect on TVHOURS (from the highest to the lowest).

c. Interpret the multiple coefficient of determination.