MORE ABOUT REGRESSION

We learned in Chapter 5 that often a straight line describes the pattern of a relationship between two quantitative variables. For instance, in Example 5.1 we explored the relationship between the hand-spans (cm) and heights (inches) of 167 college students, and found that the pattern of the relationship in this sample could be described by the equation

\[ \text{Average hand-span} = -3 + 0.35 \text{Height} \]

An equation like the one relating hand-span to height is called a regression equation, and the term simple regression is sometimes used to describe the analysis of a straight-line relationship (linear relationship) between a response variable (y-variable) and an explanatory variable (x-variable).

In Chapter 5, we only used regression methods to describe a sample and did not make statistical inferences about the larger population. Now, we consider how to make inferences about a relationship in the population represented by the sample. Some questions involving the population that we might ask when analyzing a relationship are:

1. Does the observed relationship also occur in the population? For example, is the observed relationship between hand-span and height strong enough to conclude that the relationship also holds in the population?
2. For a linear relationship, what is the slope of the regression line in the population? For example, in the larger population, what is the slope of the regression line that connects hand-spans to heights?
3. What is the mean value of the response variable (y) for individuals with a specific value of the explanatory variable (x)? For example, what is the mean hand-span in a population of people 65 inches tall?
4. What interval of values predicts the value of the response variable (y) for an individual with a specific value of the explanatory variable (x)? For example, what interval predicts the hand-span of an individual 65 inches tall?

14.1 Sample and Population Regression Models

A regression model describes the relationship between a quantitative response variable (the y-variable) and one or more explanatory variables (x-variables). The y-variable is sometimes called the dependent variable, and because regression models may be used to make predictions, the x-variables may be called the predictor variables. The labels response variable and explanatory variable may be used for the variables on the y-axis and x-axis, respectively, even if there is not an obvious way to assign these labels in the usual sense.

Any regression model has two important components. The most obvious component is the equation that describes how the mean value of the y-variable is connected to specific values of the x-variable. The equation stated before for the connection between hand-span and height, \( \text{Average hand-span} = -3 + 0.35 \text{Height} \), is an example. In this Chapter, we focus on linear relationships so a straight-line equation will be used, but it is important to note that some relationships are curvilinear.

The second component of a regression model describes how individuals vary from the regression line. Figure 14.1, which is identical to Figure 5.6, displays the raw data for the sample of n=167 hand-spans and heights along with the regression line that estimates how the mean hand-span is connected to specific heights. Notice that most individuals vary from the line. When
we examine sample data, we will find it useful to estimate the general size of the deviations from the line. When we consider a model for the relationship within the population represented by a sample, we will state assumptions about the distribution of deviations from the line.

If the sample represents a larger population, we need to distinguish between the regression line for the sample and the regression line for the population. The observed data can be used to determine the regression line for the sample, but the regression line for the population can only be imagined. Because we do not observe the whole population, we will not know numerical values for the intercept and slope of the regression line in the population. As in nearly every statistical problem, the statistics from a sample are used to estimate the unknown population parameters, which in this case are the slope and intercept of the regression line.

Figure 14.1  Regression Line Linking Hand-Span and Height for a Sample of College Students

The Regression Line for the Sample

In Chapter 5, we introduced this notation for the regression line that describes sample data:

\[ \hat{y} = b_0 + b_1 x. \]

In any given situation, the sample is used to determine values for \( b_0 \) and \( b_1 \).

- \( \hat{y} \) is spoken as “y-hat” and it is also referred to either as predicted \( y \) or estimated \( y \).
- \( b_0 \) is the intercept of the straight line. The intercept is the value of \( \hat{y} \) when \( x = 0 \).
- \( b_1 \) is the slope of the straight line. The slope tells us how much of an increase (or decrease) there is for \( \hat{y} \) when the \( x \)-variable increases by one unit. The sign of the slope tells us whether \( \hat{y} \) increases or decreases when \( x \) increases. If the slope is 0, there is no linear relationship between \( x \) and \( y \) because \( \hat{y} \) is the same for all values of \( x \).

The equation describing the relationship between hand-span and height for the sample of college students can be written as

\[ \hat{y} = -3 + 0.35 x. \]

In this equation:

- \( \hat{y} \) estimates the average hand-span for any specific height \( x \). If height=70 inches, for instance, \( \hat{y} = -3 + 0.35(70) = 21.5 \) cm.
Chapter 14

- The intercept is \( b_0 = -3 \). While necessary for the line, this value does not have a useful statistical interpretation in this example. It estimates the average hand-span for individuals who have height = 0 inches, an impossible height far from the range of the observed heights. It also is an impossible hand span.
- The slope is \( b_1 = 0.35 \). This value tells us that the average increase in hand-span is 0.35 centimeters for every one-inch increase in height.

**Reminder: The Least-Squares Criterion**
In Chapter 5, we described the least-squares criterion. This mathematical criterion is used to determine numerical values of the intercept and slope of a sample regression line. The least-squares line is the line, among all possible lines, that has the smallest sum of squared differences between the sample values of \( y \) and the corresponding values of \( \hat{y} \).

**Deviations from the Regression Line in the Sample**
The terms “random error,” “residual variation,” and “residual error” all are used as synonyms for the term “deviation.” Most commonly, the word residual is used to describe the deviation of an observed \( y \)-value from the sample regression line. A residual is easy to compute. It simply is the difference between the observed \( y \)-value for an individual and the value of \( \hat{y} \) determined from the \( x \)-value for that individual.

**Example 1. Residuals in the Hand-Span and Height Regression**
Consider a person 70 inches tall whose hand-span is 23 centimeters. The sample regression line is \( \hat{y} = -3 + 0.35x \), so \( \hat{y} = -3 + 0.35(70) = 21.5 \) cm for this person. The residual = observed \( y \)-predicted \( y \) = \( y - \hat{y} = 23 - 21.5 = 1.5 \) cm. Figure 14.2 illustrates this residual.

For an observation \( y_i \) in the sample, the residual is

\[
e_i = y_i - \hat{y}_i.
\]

\( y_i \) = the value of the response variable for the observation.
\( \hat{y}_i = b_0 + b_1 x_i \) where \( x_i \) is the value of the explanatory variable for the observation.

**Technical Note**: The sum of the residuals is 0 for any least-squares regression line. The "least squares" formulas for determining the equation always result in \( \sum y_i = \sum \hat{y}_i \), so \( \sum e_i = 0 \).
Figure 14.2 Residual for a person 70 inches tall with a hand span = 23 centimeters. The residual is the difference between observed $y=23$ and $\hat{y}=21.5$, the predicted value for a person 70 inches tall.

The Regression Line for the Population

The regression equation for a simple linear relationship in a population can be written as:

$$E(Y) = \beta_0 + \beta_1 x$$

- $E(Y)$ represents the mean or expected value of $y$ for individuals in the population who all have the same particular value of $x$. Note that $\hat{y}$ is an estimate of $E(Y)$.
- $\beta_0$ is the intercept of the straight line in the population.
- $\beta_1$ is the slope of the line in the population. Note that if the slope $\beta_1 = 0$, there is no linear relationship in the population.

Unless we measure the entire population, we cannot know the numerical values of $\beta_0$ and $\beta_1$. These are population parameters that we estimate using the corresponding sample statistics. In the hand-span and height example, $b_1 = 0.35$ is a sample statistic that estimates the population parameter $\beta_1$, and $b_0 = -3$ is a sample statistic that estimates the population parameter $\beta_0$.

Deviations from the Regression Line in the Population

To make statistical inferences about the population, two assumptions about how the $y$-values vary from the population regression line are necessary. First, we assume that the general size of the deviation of $y$-values from the line is the same for all values of the explanatory variable ($x$), an assumption called the constant variance assumption. This assumption may or may not be correct in any particular situation, and a scatter plot should be examined to see if it is reasonable or not. In Figure 14.1, the constant variance assumption looks reasonable because the magnitude of the deviation from the line appears to be about the same across the range of observed heights.

The second assumption about the population is that for any specific value of $x$, the distribution of $y$-values is a normal distribution. Equivalently, this assumption is that deviations from the population regression line have a normal curve distribution. Figure 14.3 illustrates this assumption along with the other elements of the population regression model for a linear
relationship. The line $E(Y) = \beta_0 + \beta_1 x$ describes the mean of $y$, and the normal curves describe deviations from the mean.

**Figure 14.3 Regression Model for Population**

Summary of the Simple Regression Model

A useful format for expressing the components of the population regression model is

$$Y = \text{MEAN} + \text{DEVIAITION}.$$  

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

- The **MEAN**, which in the population is the line $E(Y) = \beta_0 + \beta_1 x$ if the relationship is linear. There are other possible relationships, such as curvilinear, a special case of which is a quadratic relationship, $E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2$. Relationships that are not linear will not be discussed in this book.
- The individual's **DEVIAITION** = $Y - \text{MEAN}$, which is what is left unexplained after accounting for the mean $y$-value at that individual's $x$-value.

This format also applies to the sample, although technically we should use the term "estimated mean" when referring to the sample regression line.

**Example 1 Continued. MEAN and DEVIAITION for Height and Hand-Span Regression.**

Recall that the sample regression line for hand spans and heights is $\hat{y} = -3 + 0.35x$. Although it is not likely to be true, let's assume for convenience that this equation also holds in the population.

If your height is $x=70$ inches and your hand span is $y=23$ cm., then:

- **MEAN** = $-3 + .35(70) = 21.5$,
- **DEVIAITION** = $Y - \text{MEAN} = 23 - 21.5 = 1.5$, and
- $y = 23 = \text{MEAN} + \text{DEVIAITION} = 21.5 + 1.5$

In other words, your handspan is 1.5 cm above the mean for people with your height.
In the theoretical development of procedures for making statistical inferences for a regression model, the collection of all deviations in the population is assumed to have a normal distribution with mean 0 and standard deviation $\sigma$ (so, the variance is $\sigma^2$). The value of the standard deviation $\sigma$ is an unknown population parameter that is estimated using the sample. This standard deviation can be interpreted in the usual way that we interpret a standard deviation. It is, roughly, the average distance between individual values of $y$ and the mean of $y$ as described by the regression line. In other words, it is roughly the size of the average deviation across all individuals in the range of $x$-values.

Keeping the regression notation straight for populations and samples can be confusing. Although we have not yet introduced all relevant notation, a summary at this stage will help you keep it straight.

<table>
<thead>
<tr>
<th>Simple Linear Regression Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>For $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, a sample of $n$ observations of the explanatory variable $x$ and the response variable $y$ from a large population, the simple linear regression model describing the relationship between $y$ and $x$ is:</td>
</tr>
<tr>
<td><strong>Population version</strong></td>
</tr>
<tr>
<td><strong>Mean:</strong> $E(Y) = \beta_0 + \beta_1 x$</td>
</tr>
<tr>
<td><strong>Individual:</strong> $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = E(Y) + \varepsilon_i$</td>
</tr>
<tr>
<td>The deviations $\varepsilon_i$ are assumed to follow a normal distribution with mean 0 and standard deviation $\sigma$.</td>
</tr>
<tr>
<td><strong>Sample version</strong></td>
</tr>
<tr>
<td><strong>Mean:</strong> $\hat{y} = b_0 + b_1 x$</td>
</tr>
<tr>
<td><strong>Individual:</strong> $y_i = b_0 + b_1 x_i + e_i = \hat{y} + e_i$</td>
</tr>
<tr>
<td>where $e_i$ is the residual for individual $i$. The sample statistics $b_0$ and $b_1$ estimate the population parameters $\beta_0, \beta_1$. The mean of the residuals is 0, and the residuals can be used to estimate the population standard deviation $\sigma$.</td>
</tr>
</tbody>
</table>

### 14.2 Estimating the Standard Deviation From the Mean

Recall that the standard deviation in the regression model measures, roughly, the average deviation of $y$-values from the mean (the regression line). Expressed another way, the **standard deviation for regression measures the general size of the residuals**. This is an important and useful statistic for describing individual variation in a regression problem, and it also provides information about how accurately the regression equation might predict $y$-values for individuals. A relatively small standard deviation from the regression line indicates that individual data points generally fall close to the line, so predictions based on the line will be close to the actual values.

The calculation of the estimate of standard deviation is based on the sum of the squared residuals for the sample. This quantity is called the **sum of squared errors** and is denoted by $\text{SSE}$. Synonyms for “sum of squared errors” are **residual sum of squares** or **sum of squared residuals**. To find the $\text{SSE}$, residuals are calculated for all observations, then the residuals are squared and summed. The standard deviation for the sample is

$$s = \sqrt{\frac{\text{Sum of Squared Residuals}}{n-2}} = \sqrt{\frac{\text{SSE}}{n-2}} ,$$

and this sample statistic estimates the population standard deviation $\sigma$. 

### Estimating the Standard Deviation for a Simple Regression Model

\[ SSE = \sum (y_i - \hat{y}_i)^2 = \sum e_i^2 \]

\[ s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} \]

The statistic \( s \) is an estimate of the population standard deviation \( \sigma \).

Remember that in the regression context, \( \sigma \) is the standard deviation of the \( y \)-values at each \( x \), not the standard deviation of the whole population of \( y \)-values.

### Example 2. Relationship Between Height and Weight for College Men

Figure 14.4 displays regression results from the Minitab program and a scatter plot for the relationship between \( y = \text{weight (pounds)} \) and \( x = \text{height (inches)} \) in a sample of \( n=43 \) men in a Penn State statistics class. The regression line for the sample is \( \hat{y} = -318 + 7x \), and this line is drawn onto the plot. We see from the plot that there is considerable variation from the line at any given height. The standard deviation, shown in the row of computer output immediately above the plot, is "\( s=24.00. \)" This value roughly measures, for any given height, the general size of the deviations of individual weights from the mean weight for the height.

The standard deviation from the regression line can be interpreted in conjunction with the Empirical Rule for bell-shaped data stated in Section 2.7. Recall, for instance, that about 95\% of individuals will fall within two standard deviations of the mean. As an example, consider men who are 72 inches tall. For men with this height, the estimated average weight determined from the regression equation is \( -318 + 7.00(72) = 186 \) pounds. The estimated standard deviation from the regression line is \( s=24 \) pounds, so we can estimate that about 95\% of men 72 inches tall have weights within \( 2 \times 24 = 48 \) pounds of 186 pounds, which is 186 ± 48, or 138 to 234 pounds. Think about whether this makes sense for all the men you know who are 72 inches (6 feet) tall.
The regression equation is
\[
\text{Weight} = -318 + 7.00 \times \text{Height}
\]

<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>-317.9</td>
<td>110.9</td>
<td>-2.87</td>
<td>0.007</td>
</tr>
<tr>
<td>Height</td>
<td>6.996</td>
<td>1.581</td>
<td>4.42</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ S = 24.00 \quad R^2 = 32.3\% \quad R^2(\text{adj}) = 30.7\% \]

The Proportion of Variation Explained by \( x \)

In Chapter 5, we learned that a statistic denoted as \( r^2 \) is used to measure how well the explanatory variable actually does explain the variation in the response variable. This statistic is also denoted as \( R^2 \) (rather than \( r^2 \)), and the value is commonly expressed as a percent.

Researchers typically use the phrase “proportion of variation explained by \( x \)” in conjunction with the value of \( r^2 \). For example, if \( r^2 = 0.60 \) (or 60%), the researcher may write that the explanatory variable explains 60% of the variation in the response variable.

The formula for \( r^2 \) presented in Chapter 5 was
\[
\frac{SSTO - SSE}{SSTO}
\]

The quantity \( SSTO \) is the sum of squared differences between observed \( y \) values and the sample mean \( \bar{y} \). It measures the size of the deviations of the \( y \)-values from the overall mean of \( y \), whereas \( SSE \) measures the deviations of the \( y \)-values from the predicted values \( \hat{y} \).
Example 2 Continued. R² Heights and Weights of College Men
In Figure 14.4, we can find the information the "R-sq = 32.3%" for the relationship between weight and height. A researcher might write “the variable height explains 32.3% of the variation in the weights of college men.” This isn’t a particularly impressive statistic. As we noted before, there is substantial deviation of individual weights from the regression line so a prediction of a college man's weight based on height may not be particularly accurate.

Example 3. Driver Age and Highway Sign Reading Distance
In Example 5.2, we examined data for the relationship between y=maximum distance (feet) at which a driver can read a highway sign and x = the age of the driver. There were n=30 observations in the data set. Figure 14.5 displays Minitab regression output for these data. The equation describing the linear relationship in the sample is

$$\text{Average distance} = 577 - 3.01 \times \text{Age}$$

From the output, we learn that the standard deviation from the regression line is $s=49.76$ and $R$-sq=64.2%. Roughly, the average deviation from the regression line is about 50 feet, and the proportion of variation in sign reading distances explained by age is 0.642, or 64.2%.

Figure 14.5 Minitab Output: Sign Reading Distance and Driver Age

<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>576.68</td>
<td>23.47</td>
<td>24.57</td>
<td>0.000</td>
</tr>
<tr>
<td>Age</td>
<td>-3.0068</td>
<td>0.4243</td>
<td>-7.09</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 49.76 \quad R-Sq = 64.2\% \quad R-Sq(adj) = 62.9\%

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>1</td>
<td>124333</td>
<td>124333</td>
<td>50.21</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>28</td>
<td>69334</td>
<td>2476</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
<td>193667</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unusual Observations

<table>
<thead>
<tr>
<th>Obs</th>
<th>Age</th>
<th>Distance</th>
<th>Fit</th>
<th>SE Fit</th>
<th>Residual</th>
<th>St Resid</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>75.0</td>
<td>460.00</td>
<td>351.17</td>
<td>13.65</td>
<td>108.83</td>
<td>2.27R</td>
</tr>
</tbody>
</table>

R denotes an observation with a large standardized residual

The "Analysis of Variance" table provides the pieces needed to compute $r^2$ and $s$:

$$SSE=69334$$

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{69334}{28}} = 49.76.$$  

$$SSTO=193667$$

$$SSTO-SSE = 193667-69334=124333$$

$$r^2 = \frac{124333}{193667} = .642 \text{ or } 64.2\%$$
14.3 Inference about the Linear Regression Relationship

When researchers do a regression analysis, they occasionally know based on past research or common sense that the variables are indeed related. In some instances, however, it may be necessary to do a hypothesis test in order to make the generalization that two variables are related in the population represented by the sample. The statistical significance of a linear relationship can be evaluated by testing whether or not the slope is 0. Recall that if the slope is 0 in a simple regression model, the two variables are not related because changes in the x-variable will not lead to changes in the y-variable.

The usual null hypothesis and alternative hypotheses about \( \beta_1 \), the slope of the population line \( E(Y) = \beta_0 + \beta_1 x \), are:

- \( H_0: \beta_1 = 0 \) (the population slope is 0, so y and x are not linearly related)
- \( H_a: \beta_1 \neq 0 \) (the population slope is not 0, so y and x are linearly related)

The alternative hypothesis may be one-sided or two-sided, although most statistical software uses the two sided alternative.

The test statistic used to do the hypothesis test is a \( t \) statistic with the same general format that we saw in Chapter 13. That format, and its application to this situation, is

\[
t = \frac{\text{sample statistic} - \text{null value}}{\text{standard error}} = \frac{b_1 - 0}{s.e.(b_1)}
\]

This is a standardized statistic for the difference between the sample slope and 0, the null value. Notice that a large value of the sample slope (either positive or negative) relative to its standard error will give a large value of \( t \). If the mathematical assumptions about the population model described in Section 14.1 are correct, the statistic has a \( t \) distribution with \( n-2 \) degrees of freedom. The p-value for the test is determined using that distribution.

“By hand” calculations of the sample slope and its standard error are cumbersome. Fortunately, the regression analysis of most statistical software includes a \( t \)-statistic and a p-value for this significance test.

**Technical Note:** In case you ever need to compute the values by hand, here are the formulas for the sample slope and its standard error:

\[
b_1 = r \frac{s_y}{s_x}
\]

\[
s.e.(b_1) = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}, \text{ where } s = \sqrt{\frac{\text{SSE}}{n-2}}
\]

In the formula for the sample slope, \( s_x \) and \( s_y \) are the sample standard deviations of the x and y values respectively, and \( r \) is the correlation between x and y.

**Example 3 Continued: Driver Age and Highway Sign Reading Distance**

Figure 14.5 presents the Minitab output for the regression of sign reading distance and driver age. The sample estimate of the slope is \( b_1 = -3.01 \). This sample slope is different than 0, but is it enough different to enable us to generalize that a linear relationship exists in the population represented by this sample?

The part of the Minitab output that can be used to test the statistical significance of the relationship is shown in bold in Figure 14.5, and the relevant p-value is underlined (by the authors of this text, not by Minitab). This line of the output provides information about the sample slope, the standard error of the sample slope, the \( t \) statistic for testing statistical significance and the p-value for the test of:
H₀: β₁ = 0 (the population slope is 0, so y and x are not linearly related.)
H₁: β₁ ≠ 0 (the population slope is not 0, so y and x are linearly related.)
The test statistic is:
\[ t = \frac{b₁ - 0}{s.e.(b₁)} = -3.0068 - 0 \]
\[ 0.4243 \]
\[ = -7.09 \]
The p-value is, to 3 decimal places, 0.000. This means the probability is virtually 0 that the observed slope could be as far from 0 or farther than it is if there is no linear relationship in the population. So, as we might expect for these variables, we can conclude that the relationship between the two variables in the sample represents a real relationship in the population.

Confidence Interval for the Population Slope

The significance test of whether or not the population slope is 0 only tells us if we can declare the relationship to be statistically significant. If we decide that the true slope is not 0, we might ask, “What is the value of the slope?” We can answer this question with a confidence interval for β₁, the population slope.

The format for this confidence interval is the same as the general format used in Chapters 10 and 12, which is

\[ \text{sample estimate} \pm \text{multiplier} \times \text{standard error} \]

The estimate of the population slope β₁ is b₁, the slope of the least-squares regression line for the sample. As shown already, the standard error formula is complicated and we’ll usually rely on statistical software to determine this value. The “multiplier” will be labeled t* and is determined using a t-distribution with df = n-2. Table 12.1 can be used to find the multiplier for the desired confidence level.

Example 3 Continued. 95% Confidence Interval for Slope Between Age and Sign Reading Distance

In Figure 14.4, we see that the estimated slope is b₁ = -3.01 and s.e.(b₁) = 0.4243. There are n=30 observations so df = 28 for finding t*. For a 95% confidence level, t* = 2.05 (see Table 12.1). The 95% confidence interval for the population slope is

\[-3.01 \pm 2.05 \times 0.4243\]
\[= -3.01 \pm 0.87\]
\[= -3.88 \text{ to } -2.14\]

With 95% confidence, we can estimate that in the population of drivers represented by this sample, the mean sign reading distance decreases somewhere between 3.88 and 2.14 feet for each one-year increase in age.
Testing Hypotheses about the Correlation Coefficient

In Chapter 5, we learned that the correlation coefficient is 0 when the regression line is horizontal. In other words, if the slope of the regression line is 0, the correlation is 0. This means that the results of a hypothesis test for the population slope can also be interpreted as applying to equivalent hypotheses about the correlation between x and y in the population.

As we did for the regression model, we use different notation to distinguish between a correlation computed for a sample and a correlation within a population. It is commonplace to use the symbol \( \rho \) (pronounced “rho”) to represent the correlation between two variables within a population. Using this notation, null and alternative hypotheses of interest are:

- \( H_0: \rho = 0 \) (x and y are not correlated)
- \( H_a: \rho \neq 0 \) (x and y are correlated)

The results of the hypothesis test described before for the population slope \( \beta_1 \) can be used for these hypotheses as well. If we reject \( H_0: \beta_1 = 0 \), we also reject \( H_0: \rho = 0 \). If we decide in favor of \( H_a: \beta_1 \neq 0 \), we also decide in favor of \( H_a: \rho \neq 0 \).

Many statistical software programs, including Minitab, will give a p-value for testing whether the population correlation is 0 or not. This p-value will be the same as the p-value given for testing whether the population slope is 0 or not.

In the following Minitab output for the relationship between pulse rate and weight in a sample of 35 college women, notice that 0.292 is given as the p-value for testing that the slope is 0 (look under P in the regression results) and for testing that the correlation is 0. Because this is not a small p-value, we can reject the null hypotheses for the slope and the correlation.

```
Regression Analysis: Pulse versus Weight

The regression equation is
Pulse = 57.2 + 0.159 Weight

Predictor        Coef     SE Coef          T        P
Constant        57.17       18.51       3.09    0.004
Weight         0.1591      0.1487       1.07    0.292

Correlations: Pulse, Weight
Pearson correlation of Pulse and Weight = 0.183
P-Value = 0.292
```

The Effect of Sample Size on Significance

The size of a sample always affects whether a specific observed result achieves statistical significance. For example, \( r = 0.183 \) is not a statistically significant correlation for a sample size of \( n=35 \), as in the pulse and weight example, but it would be statistically significant if \( n=1,000 \). With very large sample sizes, weak relationships with low correlation values can be statistically significant. The “moral of the story” here is that with a large sample size, it may not be saying much to say that two variables are significantly related. This only means that we think the correlation is not 0. To assess the practical significance of the result, we should carefully examine the observed strength of the relationship.
14.4 Predicting the Value of Y for an Individual

An important use of a regression equation is to estimate or predict the unknown value of a response variable for an individual with a known specific value of the explanatory variable. Using the data described in Example 3, for instance, we can predict the maximum distance at which an individual can read a highway sign by substituting his or her age for x in the sample regression equation. Consider a person 21 years old. The predicted distance is approximately
\[ \hat{y} = 577 - 3 \times 21 = 514 \text{ feet}. \]

There will be variation among 21 year-olds with regard to the sign reading distance, so the predicted distance of 514 feet is not likely to be the exact distance for the next 21 year old who views the sign. Rather than predicting that the distance will be exactly 514 feet, we should instead predict that the distance will be within a particular interval of values. A 95% prediction interval for the value of the response variable (y) accounts for the variation among individuals with a particular value of x. This interval can be interpreted in two equivalent ways.

- The 95% prediction interval estimates the central 95% of the values of y for members of the population with a specified value of x.
- The probability is 0.95 that a randomly selected individual from the population with a specified value of x falls into the corresponding 95% prediction interval.

Notice that a prediction interval differs conceptually from a confidence interval. A confidence interval estimates an unknown population parameter, which is a numerical characteristic or summary of the population. An example in this Chapter is a confidence interval for the slope of the population line. A prediction interval, however, does not estimate a parameter; instead it estimates the potential data value for an individual. Equivalently, it describes an interval into which a specified percentage of the population may fall.

As with most regression calculations, the “by hand” formulas for prediction intervals are formidable. Statistical software can be used to create the interval. Figure 14.6 shows Minitab output that includes the 95% prediction intervals for three different ages (21 years old, 30 years old, and 45 years old). The intervals are toward the bottom right side of the display in a column labeled "95% PI" and are highlighted with bold type. (Note: The term “Fit” is a synonym for \( \hat{y} \), the estimate of the average response at the specific x value.) Here is what we can conclude:

- The probability is 0.95 that a randomly selected 21 year-old will read the sign at somewhere between roughly 407 and 620 feet.
- The probability is 0.95 that a randomly selected 30 year-old will read the sign at somewhere between roughly 381 and 592 feet.
- The probability is 0.95 that a randomly selected 45 year-old will read the sign at somewhere between roughly 338 and 545 feet.

We can also interpret each interval as an estimate of the sign reading distances for the central 95% of a population of drivers with a specified age. For instance, about 95% of all drivers 21 years old will be able to read the sign at a distance somewhere between 407 and 620 feet.
Chapter 14

Figure 14.6  Minitab output showing prediction interval of distance

The regression equation is
Distance = 577 - 3.01 Age

<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>576.68</td>
<td>23.47</td>
<td>24.57</td>
<td>0.000</td>
</tr>
<tr>
<td>Age</td>
<td>-3.0068</td>
<td>0.4243</td>
<td>-7.09</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 49.76   R-Sq = 64.2%   R-Sq(adj) = 62.9%

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>1</td>
<td>124333</td>
<td>124333</td>
<td>50.21</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>28</td>
<td>69334</td>
<td>2476</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
<td>193667</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unusual Observations

<table>
<thead>
<tr>
<th>Obs</th>
<th>Age</th>
<th>Distance</th>
<th>Fit</th>
<th>SE Fit</th>
<th>Residual</th>
<th>St Resid</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>75.0</td>
<td>460.00</td>
<td>351.17</td>
<td>13.65</td>
<td>108.83</td>
<td>2.27R</td>
</tr>
</tbody>
</table>

R denotes an observation with a large standardized residual

Predicted Values for New Observations

<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>513.54</td>
<td>15.64</td>
<td>(481.50, 545.57)</td>
<td>(406.69, 620.39)</td>
</tr>
<tr>
<td>2</td>
<td>486.48</td>
<td>12.73</td>
<td>(460.41, 512.54)</td>
<td>(381.26, 591.69)</td>
</tr>
<tr>
<td>3</td>
<td>441.37</td>
<td>9.44</td>
<td>(422.05, 460.70)</td>
<td>(337.63, 545.12)</td>
</tr>
</tbody>
</table>

Values of Predictors for New Observations

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21.0</td>
</tr>
<tr>
<td>2</td>
<td>30.0</td>
</tr>
<tr>
<td>3</td>
<td>45.0</td>
</tr>
</tbody>
</table>

We're not limited to using only 95% prediction intervals. With Minitab, we can describe any central percentage of the population that we wish. For example, here are 50% prediction intervals for the sign reading distance at the three specific ages we considered above.

<table>
<thead>
<tr>
<th>Age</th>
<th>Fit</th>
<th>50.0% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>513.54</td>
<td>(477.89, 549.18)</td>
</tr>
<tr>
<td>30</td>
<td>486.48</td>
<td>(451.38, 521.58)</td>
</tr>
<tr>
<td>45</td>
<td>441.37</td>
<td>(406.76, 475.98)</td>
</tr>
</tbody>
</table>

For each specific age, the 50% prediction interval estimates the central 50% of the maximum sign reading distances in a population of drivers with that age. For example, we can estimate that 50% of drivers 21 years old would have a maximum sign reading distance somewhere between about 478 feet and 549 feet. The distances for the other 50% of 21 year-old drivers would be predicted to be outside this range with 25% beyond 549 feet and 25% below 478 feet.

Interpretation of a Prediction Interval

A prediction interval estimates the value of y for an individual with a particular value of x, or equivalently, the range of values of the response variable for a specified central percentage of a population with a particular value of x.
Technical Note:
The formula for the prediction interval for \( y \) at a specific \( x \) is:

\[
\hat{y} \pm t^* \sqrt{s^2 + \text{s.e.}(\text{fit})^2}
\]

where \( s.e.(\text{fit}) = s \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \).

The multiplier \( t^* \) is found using a t-distribution with \( n-2 \) degrees of freedom, and is such that the probability between \(-t^* \) and \(+t^* \) equals the desired level for the interval.

Note:
- The \( s.e.(\text{fit}) \), and thus the width of the interval, depends upon how far the specified \( x \)-value is from \( \bar{x} \). The further the specific \( x \) is from the mean, the wider the interval.
- When \( n \) is large, \( s.e.(\text{fit}) \) will be small, and the prediction interval will be approximately \( \hat{y} \pm t^* s \).

14.5 Estimating the Mean \( Y \) at a Specified \( X \)

In the previous section, we focused on the estimation of the values of the response variable for individuals. A researcher may instead want to estimate the mean value of the response variable for individuals with a particular value of the explanatory variable. We might ask, “What is the mean weight for college men who are 6 feet tall?” This question only asks about the mean weight in a group with a common height, and it is not concerned with the deviations of individuals from that mean.

In technical terms, we wish to estimate the population mean \( E(Y) \) for a specific value of \( x \) that is of interest to us. To make this estimate, we use a confidence interval. This format for this confidence interval is again:

\[
\text{sample estimate} \pm \text{multiplier} \times \text{standard error}
\]

The sample estimate of \( E(Y) \) is the value of \( \hat{y} \) determined by substituting the \( x \)-value of interest into \( \hat{y} = b_0 + b_1 x \), the least-squares regression line for the sample. The standard error of \( \hat{y} \) is the \( s.e.(\text{fit}) \) shown in the Technical Note in the previous section, and its value is usually provided by statistical software. The multiplier is found using a t-distribution with \( \text{df} = n-2 \), and Appendix A-3 can be used to determine its value.

Example 2 Revisited. Estimating Mean Weight of College Men at Various Heights

Based on the sample of \( n=43 \) college men in Example 2, let’s estimate the mean weight in the population of college men for each of three different heights: 68 inches, 70 inches, and 72 inches. Figure 14.7 shows Minitab output that includes the three different confidence intervals for these three different heights. These intervals are toward the bottom of the display in a column labeled “95% CI”. The first entry in that column is the estimate of the population mean weight for men who are 68 inches tall. With 95% confidence, we can estimate that mean weight of college men 68 inches tall is somewhere between 147.78 and 167.81 pounds. The second row under “95% CI” contains the information that the 95% confidence interval for the mean weight of college men 70 inches tall is 164.39 to 179.19 pounds. The 95% confidence interval for the mean weight for men 72 inches tall is 176.25 to 195.31 pounds.

Again, it is important to realize that the confidence intervals for \( E(Y) \) do not describe the variation among individuals. They only are estimates of the mean weights for specific heights.

The prediction intervals for individual responses describe the variation among individuals. You may have noticed that 95% prediction intervals, labeled “95% PI”, are next to the confidence
intervals in the output. Among men 70 inches tall, for instance, we would estimate that 95% of the individual weights would be in the interval from about 122 to about 221 pounds.

Figure 14.7  Minitab Output with Confidence Intervals For Mean Weight

The regression equation is
Weight = - 318 + 7.00 Height

<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>-317.9</td>
<td>110.9</td>
<td>-2.87</td>
<td>0.007</td>
</tr>
<tr>
<td>Height</td>
<td>6.996</td>
<td>1.581</td>
<td>4.42</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 24.00  R-Sq = 32.3%  R-Sq(adj) = 30.7%

--- Some Output Omitted ----

Predicted Values for New Observations

<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>157.80</td>
<td>4.96</td>
<td>(147.78, 167.81)</td>
<td>(108.31, 207.29)</td>
</tr>
<tr>
<td>2</td>
<td>171.79</td>
<td>3.66</td>
<td>(164.39, 179.19)</td>
<td>(122.76, 220.82)</td>
</tr>
<tr>
<td>3</td>
<td>185.78</td>
<td>4.72</td>
<td>(176.25, 195.31)</td>
<td>(136.38, 235.18)</td>
</tr>
</tbody>
</table>

Values of Predictors for New Observations

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68.0</td>
</tr>
<tr>
<td>2</td>
<td>70.0</td>
</tr>
<tr>
<td>3</td>
<td>72.0</td>
</tr>
</tbody>
</table>

14.6 Checking Conditions for Using Regression Models for Inference

There are a few conditions that should be at least approximately true when we use a regression model to make an inference about a population. Of the five conditions that follow, the first two are particularly crucial.

Conditions for Linear Regression

1. The form of the equation that links the mean value of y to x must be correct. For instance, we won’t make proper inferences if we use a straight line to describe a curved relationship.
2. There should not be any extreme outliers that influence the results unduly.
3. The standard deviation of the values of y from the mean y is the same regardless of the value of the x variable. In other words, y values are similarly spread out at all values of x.
4. For individuals in the population with the same particular value of x, the distribution of the values of y is a normal distribution. Equivalently, the distribution of deviations from the mean value of y is a normal distribution. This condition can be relaxed if the sample size is large.
5. Observations in the sample are independent of each other.
Checking the Conditions with Plots
A scatter plot of the raw data and plots of the residuals provide information about the validity of the assumptions. Remember that a residual is the difference between an observed value and the predicted value for that observation, and that some assumptions made for a linear regression model have to do with how y-values deviate from the regression line. If the properties of the residuals for the sample appear to be consistent with the mathematical assumptions made about deviations within the population, we can use the model to make statistical inferences.

Conditions 1, 2 and 3 can be checked using two useful plots:

- A scatter plot of y versus x for the sample (y vs x)
- A scatter plot of the residuals versus x for the sample (resids vs x)

If Condition 1 holds for a linear relationship, then:
- The plot of y vs x should show points randomly scattered around an imaginary straight line.
- The plot of resids vs x should show points randomly scattered around a horizontal line at resid = 0.

If Condition 2 holds, extreme outliers should not be evident in either plot. If condition 3 holds, neither plot should show increasing or decreasing spread in the points as x increases.

Example 2 Continued. Checking the Conditions for the Weight and Height Problem

Figure 14.4 displayed a scatter plot of the weights and heights of n=43 college men. In that plot, it appears that a straight-line is a suitable model for how mean weight is linked to height. In Figure 14.8 there is a plot of the residuals (e_i) versus the corresponding values of height for these 43 men. This plot is further evidence that the right model has been used. If the right model has been used, the way in which individuals deviate from the line (residuals) will not be affected by the value of the explanatory variable. The somewhat random looking blob of points in Figure 14.8 is the way a plot of residuals versus x should look if the right equation for the mean has been used. If the right model has been used, both plots (Figures 14.4 and 14.8) also show that there are no extreme outliers and that the heights have approximately the same variance across the range of heights in the sample. Therefore, Conditions 2 and 3 appear to be met.

Figure 14.8  Plot of Residuals versus X for Example 2. The Absence of a Pattern Indicates the Right Model Has Been Used
Chapter 14

*Condition 4*, which is that deviations from the regression line are normally distributed, is difficult to verify but it is also the least important of the conditions because the inference procedures for regression are *robust*. This means that if there are no major outliers or extreme skewness, the inference procedures work well even if the distribution of y-values is not a normal distribution. In Chapters 12 and 13, we saw that confidence intervals and hypothesis tests for a mean or a difference between two means also were robust.

To examine the distribution of the deviations from the line, a histogram of the residuals is useful although for small samples a histogram may not be informative. A more advanced plot called a normal probability plot can also be used to check whether the residuals are normally distributed, but we do not provide the details in this text. Figure 14.9 displays a histogram of the residuals for Example 2. It appears that the residuals are approximately normally distributed, so Condition 4 is met.

**Figure 14.9 Histogram of Residuals for Example 2**

*Condition 5* follows from the data collection process. It is met as long as the units are measured independently. It would not be met if the same individuals were measured across the range of x-values, such as if x=average speed and y=gas mileage were to be measured for multiple tanks of gas on the same cars. More complicated models are needed for dependent observations, and those models will not be discussed in this book.

**Corrections When Conditions Are Not Met**

There are some steps that can be taken if Conditions 1, 2 or 3 are not met. If Condition 1 is not met, more complicated models can be used. For instance, Figure 14.10 shows a typical plot of residuals that occurs when a straight-line model is used to describe data that are curvilinear. It may help to think of the residuals as prediction errors that would occur if we use the regression line to predict the value of y for the individuals in the sample. In the plot shown in Figure 14.10, the “prediction errors” are all negative in the central region of X and nearly all positive for outer values of X. This occurs because the wrong model is being used to make the predictions. A curvilinear model, such as the quadratic model discussed earlier, may be more appropriate.

**Figure 14.10 A Residual Plot Indicating the Wrong Model Has Been Used**
Condition 2, that there are no influential outliers, can be checked graphically with the scatter plot of \( y \) versus \( x \) and the plot of residuals versus \( x \). The appropriate correction if there are outliers depends on the reason for the outliers. The same considerations and corrective action discussed in Chapter 2 would be taken, depending on the cause of the outlier. For instance, Figure 14.11 shows a scatter plot and a residual plot for the data of Exercise 38 in Chapter 5. A potential outlier is seen in both plots.

In this example, the \( x \)-variable is weight and the \( y \)-variable is time to chug a beverage. The outlier probably represents a legitimate data value. The relationship appears to be linear for weights ranging up to about 210 pounds, but then it appears to change. It could either become quadratic, or it could level off. We do not have enough data to determine what happens for higher weights. The solution in this case would be to remove the outlier, and use the linear regression relationship only for body weights under about 210 pounds. Determining the relationship for higher body weights would require a larger sample of individuals in that range.
If either Condition 1 or Condition 3 is not met, a transformation may be required. This is equivalent to using a different model. Fortunately, often the same transformation will correct problems with Conditions 1, 3, and 4. For instance, when the response variable is monetary, such as salaries, it is often more appropriate to use the relationship
\[ \ln(y) = b_0 + b_1 x + e \]
In other words, to assume that there is a linear relationship between the natural log of y and the x-values. This is called a log transformation on the y's. We will not pursue transformations further in this book.