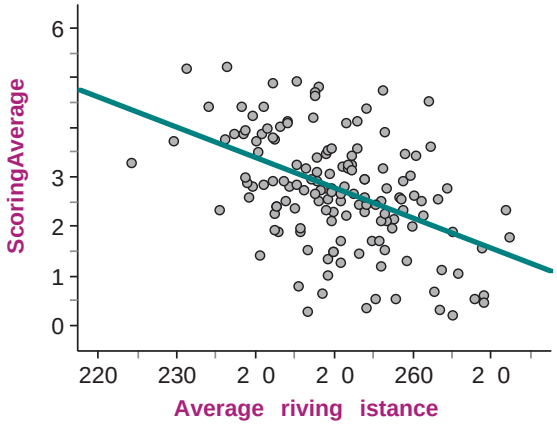
12

# Multiple Regression

## Hit It Long or Hit It Straight? Why Not Both?

In Chapter 10, we asked which variable, average driving distance or driving accuracy, was a better predictor of scoring average on the women's professional golf (LPGA) tour. After looking at scatterplots and calculating correlations, we decided that the relationship between average driving distance and scoring average was stronger than the relationship between driving accuracy and scoring average. In other words, average driving distance is better than driving accuracy as a predictor of scoring average.

Here is a scatterplot showing the relationship between **average driving distance** and **scoring average** for 146 LPGA golfers in 2009. Remember, lower scores are better in golf!




### SPORTS TERMS

In golf, **average driving distance** is the average number of yards a golfer hits her tee shot on holes where she is trying to hit the ball as far as she can. In 2009, Vicky Hurst led the LPGA with a 272.5 driving average.

**Scoring average** is the average number of strokes it takes a golfer to complete an 18-hole course. Lorena Ochoa had the best (lowest) scoring average in 2009 with 70.16 strokes.

The least-squares regression line for these data is $\hat{y} = 87.95 - 0.0609x$. The **slope** of the line indicates that for each additional yard an LPGA golfer

**REVIEW**

The **slope** of a least-squares regression line describes the predicted change in the response variable for each one-unit increase in the explanatory variable. The **standard deviation of the residuals** ($s$) is an estimate of the typical distance between the actual values of the response variable and their corresponding predicted values.

drives the ball, her predicted scoring average will go *down* by about 0.06 strokes. The **standard deviation of the residuals** is $s = 1.01$, which means that when we use driving distance to predict a golfer's score, we will typically be off by about 1.01 strokes.

Of course, there is more to golf than just driving distance! Other variables, such as driving accuracy and putting average, also contribute to a golfer's scoring average. If we can incorporate these additional variables into our predictions of scoring average, we should be able to reduce the standard deviation of the residuals. Fortunately, we can use a technique called *multiple regression* to predict the value of a numerical response variable using more than one explanatory variable.

## ADDING A SECOND EXPLANATORY VARIABLE: DRIVING ACCURACY

Suppose we want to use *both* average driving distance ($x_1$) and **driving accuracy** ($x_2$) to predict scoring average ($y$). Unfortunately, it is very difficult to make a scatterplot showing the relationship among all three variables. However, using the principle of least squares, we can calculate a model in the form $\hat{y} = a + b_1x_1 + b_2x_2$ that makes the sum of the squared residuals as small as possible. For these data, the best model is

$$\hat{y} = 100.62 - 0.0844x_1 - 0.097412x_2$$

**SPORTS TERM**

In golf, **driving accuracy** is the percentage of times a golfer's tee shot lands in the fairway (measured from 0 to 100). For example, Mi Hyun Kim led the LPGA in 2009 with a driving accuracy of 83.4, so 83.4% of her tee shots landed in the fairway.

To see how to use this model, let's consider the 2009 performances of Lorena Ochoa. In 2009, Ochoa had an average driving distance of 265.2 yards and a driving accuracy of 71.8%. This produces a predicted scoring average of

$$\hat{y} = 100.62 - 0.0844(265.2) - 0.097412(71.8) = 71.24 \text{ strokes}$$

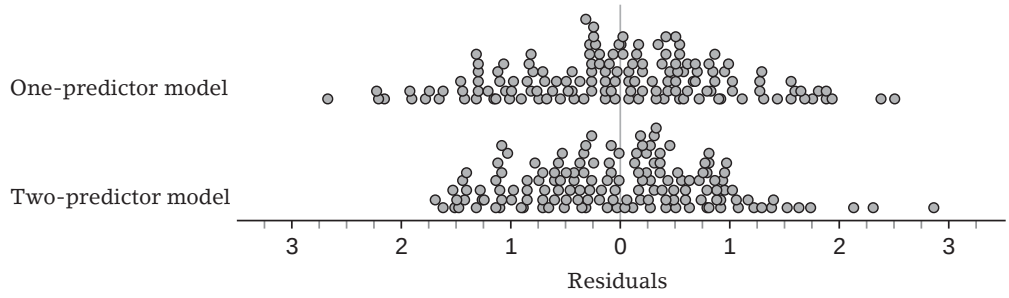
Her actual scoring average in the 2009 season was 70.16, giving a residual of

$$y - \hat{y} = 70.16 - 71.24 = -1.08 \text{ strokes}$$

This means that Lorena Ochoa averaged 1.08 fewer strokes than expected, based on her average driving distance and driving accuracy. Remember, fewer strokes are better in golf!

Overall, the standard deviation of the residuals is $s = 0.88$ strokes, meaning that when using our new model, our typical prediction error is about 0.88 strokes. Because the previous model using only average driving distance had a standard deviation of $s = 1.01$ strokes, this is definitely an improvement.

Here are dotplots showing the residuals when using the one-predictor model and the residuals when using the two-predictor model. As you can see, the residuals are typically a little smaller when using the two-predictor model. That is, when we use two explanatory variables, the residuals tend to be smaller than when we use one explanatory variable.




If using two predictors is better than one, perhaps using three predictors will be better than two. Let's find out how adding a third variable affects the model.

## ADDING A THIRD EXPLANATORY VARIABLE: PUTTING AVERAGE

Adding a third variable, **putting average** ($x_3$), into the model gives the following equation:

$$\hat{y} = 77.35 - 0.0901x_1 - 0.098499x_2 + 0.8246x_3$$

In 2009, Lorena Ochoa's putting average was 29.48 putts per round. This leads to a revised prediction for her scoring average of

$$\begin{aligned}\hat{y} &= 77.35 - 0.0901(265.2) - 0.098499(71.8) + 0.8246(29.48)\\ &= 70.69 \text{ strokes}\end{aligned}$$

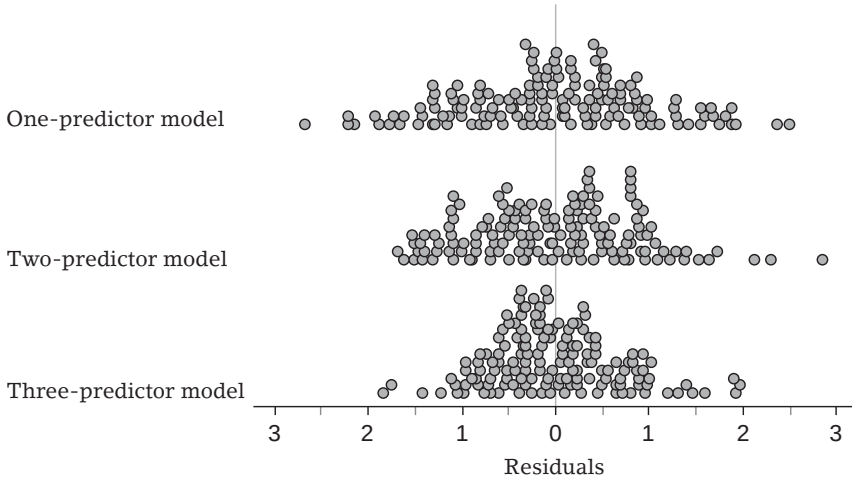
**SPORTS TERM**

**Putting average** is the average number of putts needed to complete an 18-hole course. Inbee Park had the best (lowest) putting average in 2009 with 28.36.

Ochoa's residual is now $y - \hat{y} = 70.16 - 70.69 = -0.53$ strokes. So, even after factoring in average driving distance, driving accuracy, and putting average, we are still predicting a higher scoring average than she actually had, by more than half a stroke per round. This could be due to other variables that have yet to be included in the model, or it could simply be due to ***RANDOM CHANCE***.

Overall, this three-predictor model has a standard deviation of the residuals of $s = 0.70$ strokes, so our predictions will now typically be off by only 0.70 strokes. This definitely represents an improvement over the one-predictor model ($s = 1.01$) and the two-predictor model ($s = 0.88$).

Here are dotplots showing the residuals when using the one-, two-, and three-predictor models. As you can see, on average, the residuals decrease in size as the number of predictors increases.




Now that we know how to use multiple regression models to make predictions, let's learn how to interpret the coefficients of each variable.

## INTERPRETING MULTIPLE REGRESSION COEFFICIENTS

When using a single explanatory variable to predict a response variable, we calculated linear models in the form $\hat{y} = a + bx$. The coefficient of the $x$ variable ($b$) was called the *slope* of the least-squares regression line. For example, when we used average driving distance to predict scoring average, the equation of the least-squares regression line was $\hat{y} = 87.95 - 0.0609x$. The slope of $-0.0609$ means that for each additional yard that a golfer can drive the ball, her *predicted* scoring average would go down by 0.0609 strokes.

In multiple regression, however, interpreting the coefficients is a little more complicated. For example, we can no longer call the coefficients "slopes" because we are no longer working with a two-dimensional scatterplot. However, the basic information conveyed by the coefficients is very similar to the information conveyed by the slope of a least-squares regression line.

Here is our model again, using $x_1$ = average driving distance, $x_2$ = driving accuracy, $x_3$ = putting average, and $\hat{y}$ = predicted scoring average:

$$\hat{y} = 77.35 - 0.0901x_1 - 0.098499x_2 + 0.8246x_3$$

Suppose that a golfer hits 70% of her fairways and has a putting average of 30 putts per round. As long as these values don't change, the model is now

$$\hat{y} = 77.35 - 0.0901x_1 - 0.098499(70) + 0.8246(30)$$
$$\hat{y} = 95.19 - 0.0901x_1$$

This means that for each additional yard that a golfer can drive the ball, her predicted scoring average would go down by 0.0901 strokes,

*assuming that the values of the other variables do not change.* If the other variables don't remain the same, then the predicted scoring average would change by something other than 0.0901.

Likewise, the coefficient of driving accuracy, −0.098499, indicates that for each increase of 1% in driving accuracy, a golfer's predicted scoring average will decrease by 0.098499 strokes, assuming that the values of the other variables do not change.

Finally, the coefficient of putting average, 0.8246, tells us that for each additional putt a golfer averages, the predicted scoring average will go up by 0.8246, assuming that the values of the other variables remain the same.

You might be wondering why the coefficient of putting average isn't exactly 1, because adding an additional putt should add exactly 1 stroke to a golfer's score. This is probably due to the fact that better putters tend to be worse in other areas and vice versa, so players who average 1 putt more than others do slightly better in other aspects of the game. This also means that our assumption about other variables remaining the same when we interpret coefficients is probably not realistic.

### THINK ABOUT IT

On the LPGA Web site, driving accuracy is recorded as a proportion rather than a percent. For example, Lorena Ochoa's driving accuracy was listed as 0.718 rather than 71.8%. To make it easier to interpret the coefficient of driving accuracy, these proportions were converted to percents by multiplying each by 100. Otherwise, saying that driving accuracy "increases by 1" would mean that driving accuracy goes up by 100%.

In general, whenever the values of an explanatory variable are recorded as a proportion from 0 to 1, it is a good idea to convert them to percents from 0 to 100 before using the variable in a multiple regression model.

Here is another example that uses a multiple regression model, this time in football.

## How Bad Are Turnovers for Offenses in the NFL?

**EXAMPLE** Using team data from the 2009 NFL regular season, the following multiple regression model was calculated using $y$ = points scored, $x_1$ = yards gained, and $x_2$ = number of turnovers given up:

$$\hat{y} = -132.83 + 0.0991x_1 - 2.0297x_2$$

**PROBLEM:** In 2009, the Miami Dolphins scored 360 points, gained 5401 yards, and had 29 turnovers. Calculate and interpret their residual.

**SOLUTION:** Using the multiple regression equation, the Dolphins' predicted points scored is

$$\hat{y} = -132.83 + 0.0991(5401) - 2.0297(29) = 343.5 \text{ points}$$

and so their residual is

$$y - \hat{y} = 360 - 343.5 = 16.5 \text{ points}$$

Thus, the Dolphins scored 16.5 more points than expected, based on the number of yards they gained and the number of turnovers they had.

**PROBLEM:** Interpret the coefficients of $x_1$ and $x_2$ in the context of this problem.

**SOLUTION:** For each additional yard gained, the predicted number of points scored will increase by 0.0991, assuming the number of turnovers doesn't change.

For each additional turnover, the predicted number of points scored will go down by 2.0297, assuming the number of yards gained doesn't change. So each turnover costs the offense about 2 potential points.

**PROBLEM:** Interpret the value $s = 39.3$ for this model.

**SOLUTION:** When predicting points scored using yards gained and turnovers given up, our predictions will typically be off by about 39.3 points.

This example illustrates one of the most common uses of multiple regression—trying to isolate the effects of a particular variable after taking other variables into account. Of course, even when using multiple regression, the coefficients are just *estimates*. Because ***RANDOM CHANCE*** is always part of athletic ***PERFORMANCE***, we will never be able to know the exact effect of a variable. Also, the value of a coefficient can change quite a bit, depending on which other explanatory variables are included in the model, especially if there is a relationship between the explanatory variables.

It is also possible to use categorical variables as explanatory variables, as you will see in the next example.

## Using Indicator Variables

What factors affect attendance in the NBA? It seems reasonable that winning teams will attract more fans. Also, teams that play in bigger cities could reasonably assume they will draw bigger crowds. Finally, it seems possible that teams could enjoy a bump in attendance from making it to the NBA Finals sometime in the last few seasons.

The following table lists the 30 NBA teams and data from the 2008–2009 regular season. In this case, the response variable is their average attendance ($y$). The first explanatory variable is their

number of wins ($x_1$). The second explanatory variable, market share ($x_2$), measures the size of the team's television market, measured as a percentage of all households in the United States. For example, Atlanta has 2.07% of all households in the United States, while Charlotte only has 0.981% of households.[1]

Finally, the third explanatory variable, Finals ($x_3$), is called an **indicator variable** because it simply indicates whether or not a team has been in the NBA Finals in the last five years. In this case, a value of 1 indicates that the team has been to the Finals in the last five years, and a value of 0 indicates that they haven't been to the Finals in the last five years.

**KEY TERM**

An **indicator variable** is a categorical variable with two possible outcomes. These outcomes are coded numerically so they can be included in regression calculations. Typically, a success is reported as a "1" and a failure is recorded as a "0."

| TEAM | WINS ($x_1$) | MARKET SHARE ($x_2$) | FINALS ($x_3$) | AVERAGE ATTENDANCE ($y$) |
|---|---|---|---|---|
| Atlanta | 47 | 2.070 | 0 | 16,748 |
| Boston | 62 | 2.105 | 1 | 18,624 |
| Charlotte | 35 | 0.981 | 0 | 14,526 |
| Chicago | 41 | 3.052 | 0 | 21,197 |
| Cleveland | 66 | 1.332 | 1 | 20,010 |
| Dallas | 50 | 2.175 | 1 | 20,042 |
| Denver | 54 | 1.332 | 0 | 17,223 |
| Detroit | 39 | 1.684 | 1 | 21,877 |
| Golden State | 29 | 2.164 | 0 | 18,942 |
| Houston | 53 | 1.840 | 0 | 17,482 |
| Indiana | 36 | 0.974 | 0 | 14,182 |
| LA Clippers | 19 | 4.940 | 0 | 16,170 |
| LA Lakers | 65 | 4.940 | 1 | 18,997 |
| Memphis | 24 | 0.589 | 0 | 12,745 |
| Miami | 43 | 1.352 | 1 | 18,229 |
| Milwaukee | 34 | 0.791 | 0 | 15,389 |
| Minnesota | 24 | 1.512 | 0 | 14,505 |
| New Jersey | 34 | 6.495 | 0 | 15,147 |
| New Orleans | 49 | 0.527 | 0 | 16,968 |
| New York | 32 | 6.495 | 0 | 19,287 |
| Oklahoma City | 23 | 0.600 | 0 | 18,693 |
| Orlando | 59 | 1.281 | 1 | 17,043 |
| Philadelphia | 41 | 2.578 | 0 | 15,802 |
| Phoenix | 46 | 1.622 | 0 | 18,422 |
| Portland | 54 | 1.027 | 0 | 20,524 |
| Sacramento | 17 | 1.223 | 0 | 12,571 |
| San Antonio | 54 | 0.715 | 1 | 18,269 |
| Toronto | 33 | 4.900 | 0 | 18,773 |
| Utah | 48 | 0.803 | 0 | 19,903 |
| Washington | 19 | 2.028 | 0 | 16,612 |

The least-squares model for these data is

$$\hat{y} = 13{,}820 + 69.06x_1 + 269x_2 + 1011x_3$$
$$s = 2108$$

The coefficient of $x_1$ tells us that for each additional win, the predicted average attendance will increase by about 69 fans, assuming that the other variables in the model remain the same. Likewise, the coefficient of $x_2$ tells us that for each increase of 1% in market share, the predicted average attendance will increase by 269 fans, assuming that the values of the other variables remain the same.

Because the value of $x_3$ can only be 0 or 1, the coefficient of the indicator variable Finals ($x_3$) indicates that teams which have made it to the Finals ($x_3 = 1$) experience an increase of 1011 in their predicted average attendance compared to teams that have not made it to the Finals ($x_3 = 0$), assuming that the values of the other variables remain the same. That is, if you have two teams that have the same number of wins and the same market share, a team that has recently made the Finals will average about 1011 more fans than a team that has not recently made the Finals.

### THINK ABOUT IT

We can also use the indicator variable to create two different equations, one for teams that made the Finals in the last five years ($x_3 = 1$) and one for teams that didn't make the Finals ($x_3 = 0$). These equations are the same except for the constant term, and the difference in the constant terms is exactly the coefficient of the indicator variable.

Teams that made the Finals: $\hat{y} = 13{,}820 + 69.06x_1 + 269x_2 + 1011(1)$
$= 14{,}831 + 69.06x_1 + 269x_2$

Teams that didn't make the Finals: $\hat{y} = 13{,}820 + 69.06x_1 + 269x_2 + 1011(0)$
$= 13{,}820 + 69.06x_1 + 269x_2$

We have seen that there is a home field advantage in sports such as football and baseball, but is there a home field advantage in the Olympics? Read the next example to find out.

## Is There a Home Field Advantage in the Winter Olympics?

**EXAMPLE** To investigate whether there is a home field advantage in the Winter Olympics, the total number of medals for each of the 95 countries participating in the 2010 Vancouver Winter Games was recorded. To model each country's *PERFORMANCE* in the 2010 Olympics, we will use two explanatory variables. The first variable, number of medals won in the 2006 Winter Olympics in Torino, Italy, will be

included to estimate a country's overall *ABILITY* to win medals in the Winter Olympics. The second variable will be an indicator variable to identify countries that are close to home (1 for Canada and the United States, 0 for all other countries). Italy and neighboring countries (France, Switzerland, Austria, and Slovenia) were removed because they were "at home" in the 2006 Olympics and their home field advantage in Torino might affect the model.

Using $y$ = number of medals won in 2010, $x_1$ = number of medals won in 2006, and $x_2$ = whether or not a country was at home, the least-squares model is

$$\hat{y} = 0.1283 + 0.9569x_1 + 7.9270x_2$$
$$s = 1.53$$

**PROBLEM:** Interpret the coefficient of $x_1$.

**SOLUTION:** For each additional medal a country won in 2006, the country's predicted number of medals won in 2010 will go up by 0.9569, assuming that its home field status remains the same.

**PROBLEM:** Interpret the coefficient of $x_2$. Does this suggest that there was a home field advantage in the 2010 Winter Olympics?

**SOLUTION:** If two countries won the same number of medals in the 2006 Olympics and one of these countries is close to home in the 2010 Olympics, we predict that the country that is close to home will win 7.9270 medals more than the country that is not close to home. So it appears that there was a home field advantage in the 2010 Winter Olympics.

**PROBLEM:** Interpret the constant term of the model, 0.1283.

**SOLUTION:** If a country won 0 medals in 2006 and wasn't a "home" country in 2010, then I predict that it will win 0.1283 medals.

**PROBLEM:** In Chapter 11, we learned about regression to the mean—the tendency for great *PERFORMANCES* to be followed by not-as-great *PERFORMANCES* and poor *PERFORMANCES* to be followed by not-as-poor *PERFORMANCES*. Explain how the coefficient of $x_1$ and the constant term illustrate regression to the mean.

**SOLUTION:** Because the coefficient of $x_1$ is less than 1 and the constant term is very close to 0, I predict that countries with great *PERFORMANCES* in 2006 will win fewer medals in 2010 (except for the home teams). Likewise, countries with poor *PERFORMANCES* in 2006 (0 medals) are predicted to do a little better ($\hat{y} = 0.1283$) in 2010.

**PROBLEM:** Calculate and interpret the residual for the United States, which won 25 medals in 2006 and 37 medals in 2010.

**SOLUTION:** $\hat{y} = 0.1283 + 0.9569(25) + 7.9270(1) = 31.9778$ medals

$$\text{Residual} = y - \hat{y} = 37 - 31.9778 = 5.0222 \text{ medals}$$

Even after accounting for its Winter Olympic *ABILITY* and its home field advantage, the United States won about 5 more medals than expected.

**PROBLEM:** In the 2010 Winter Olympics, Iceland did not win a single medal. If the next winter Olympics were held in Iceland, will the home field advantage guarantee them at least 7 medals? Explain.

**SOLUTION:** Not necessarily. Even though countries that are close to home tend to win more medals, this is only an association, not a cause-and-effect relationship. Even with multiple regression, we cannot establish cause-and-effect relationships without a designed experiment.

As you might guess, calculating the equation of a multiple regression model is more complicated than calculating the equation of a least-squares regression line, which uses just a single explanatory variable. Although there are programs that can be used on a TI-84 to calculate multiple regression models,* it is definitely better to use a computer because it is much easier to import large data sets to a computer for analysis.†

*Search for "multiple regression" at www.ticalc.org.
†See Appendix A on the book's Web site for tips on importing data.

# Building Multiple Regression Models

Let's use the NBA data from the previous section to calculate and evaluate a multiple regression model.

1. Go to highschool.bfwpub.com/sris2e and launch the *Multiple Regression* applet.
2. Enter "Wins," "Market," and "Finals" as the explanatory variables and "Attendance" as the response variable. You may need to press the + button to enter additional explanatory variables. Also, make sure that the boxes are checked for each explanatory variable to include them in the model.
3. Enter the 30 data values for each variable, separated by spaces or commas. Make sure to keep the data in order!

**Multiple Regression**

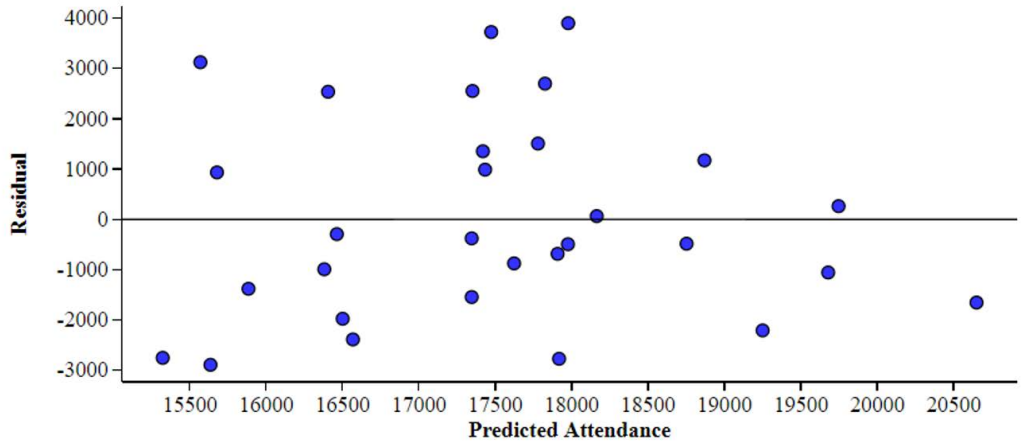
| Variable | Name | Observations (separated by commas or spaces) *Keep individuals in the same order.* | Included in model | |
|---|---|---|---|---|
| Explanatory 1 | Wins | 47 62 35 41 66 50 54 39 29 53 36 19 65 24 43 34 24 34 49 | ☑ | - |
| Explanatory 2 | Market | 2.070 2.105 0.981 3.052 1.332 2.175 1.332 1.684 2.164 1.8 | ☑ | - |
| Explanatory 3 | Finals | 0 1 0 0 1 1 0 1 0 0 0 0 1 0 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 | ☑ | - |
| | | | | + |
| Response | Attendance | 16748 18624 14526 21197 20010 20042 17223 21877 189 | | |

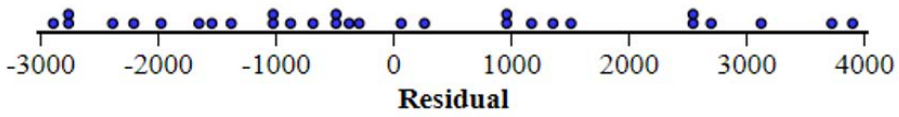
4. Press the "Begin analysis" button to generate the model, summary statistics ($r^2$ and $s$), a residual plot, and a dotplot of the residuals.

**Linear Regression Model**

Attendance = 13819.704069 + 69.06253(Wins) + 269.451556(Market) + 1010.605794(Finals)
$r^2 = 0.306$
$s = 2108.498$

**Residual Plot**






5. To make a prediction, fill in the values of each explanatory variable in the boxes under the graph and press the "Compute predicted value" button.

**Make a Prediction**

Wins = 50
Market = 2.0
Finals = 1
Compute predicted value  Predicted value of Attendance = 18822.339

Now that we know how to calculate least-squares models using multiple regression, it is important that we think about which variables to include in the model. As we will find out, adding more variables isn't always a good idea.

## VARIABLE SELECTION

Each time we add a variable to a multiple regression model, we reduce the sum of squared residuals. Variables that provide additional useful information, such as putting average in the LPGA example, can reduce the sum of squared residuals—and the standard deviation of the residuals—by quite a bit. However, if the variable doesn't provide much additional useful information, the sum of squared residuals won't go down by much. So it is important to carefully consider which variables to include in the model.

Adding variables that don't provide additional useful information will increase the complexity of the model and may not reduce the standard deviation of the residuals at all. This is because increasing the number of explanatory variables ($k$) reduces the denominator in the formula for the standard deviation of the residuals. If the variables aren't useful, then the reduction in the sum of squared residuals ($rss$) in the numerator won't be enough to compensate for the reduction in the denominator:

$$s = \sqrt{\frac{rss}{n - (k + 1)}}$$

For example, let's add the length of a city's name to our model of NBA attendance. Clearly, adding a nonsense variable like name length shouldn't provide any new information about attendance in the NBA! However, adding this variable actually reduces the sum of squared residuals to 114,693,390, due to a chance association between name length and attendance. Although the sum of squared residuals decreases a little, the standard deviation of the residuals $s$ actually *increases* to 2142 fans because we also had to reduce the denominator by 1:

$$s = \sqrt{\frac{114{,}693{,}390}{30 - (4 + 1)}} = 2142 \text{ fans}$$

### STATS 101

Another way to evaluate a multiple regression model is to use the coefficient of determination $r^2$. The value of $r^2$ measures what percentage of the variability in the response variable is accounted for by the explanatory variables included in the model. Because it is calculated using the sum of squared residuals, whenever the sum of squared residuals goes down, the value of $r^2$ will increase.

Because $r^2$ will increase even if a nonsense variable is added to the model, statisticians will often use a modified version of $r^2$, called adjusted-$r^2$:

$$\text{adj-}r^2 = 1 - \frac{s^2}{s_y^2}$$

where $s$ is the standard deviation of the residuals and $s_y$ is the standard deviation of the response variable $y$. Using the standard deviation of the residuals means that the value of adjusted-$r^2$ will actually decrease when a nonsense variable is added to a model. Consequently, adjusted-$r^2$ is a better way to evaluate the contributions of additional variables.

In general, it isn't worth adding variables that don't have a big enough impact on the sum of squared residuals to compensate for the reduction in the denominator of the formula for the standard deviation of the residuals.

Variable selection is an art and highly dependent on the context being studied. Even starting with the same data, different statisticians will probably make different decisions about which variables to include. Here are some strategies that you should consider when deciding which variables to include.

1. Use common sense. It seems reasonable to think a team with more wins will attract bigger crowds, but it doesn't make sense to think that the length of a city's name will have any impact on attendance. Other reasonable variables to consider adding to our NBA model include average ticket price and whether or not the team has a superstar player.

2. Try to limit the number of variables in the model as much as possible. Simpler models with fewer variables are generally easier to use and easier to interpret. And, to get reliable estimates for the coefficients, the number of observations ($n$) should be at least 10 times the number of explanatory variables ($k$). For example, to create a multiple regression model with three explanatory variables, you should have at least 30 observations.

3. Include only the variables that make the standard deviation of the residuals go down. The easiest way to identify these variables is to calculate the model with the variable included and without the variable included to see how much the value of $s$ changes.

4. Don't pick explanatory variables that measure essentially the same thing as the response variable. For example, if you are trying to predict the number of runs a baseball team will score, don't use RBIs (runs batted in) as an explanatory variable. Likewise, don't use the number of touchdowns to predict the number of points for football teams.

5. Don't pick explanatory variables that measure the same thing as each other. For example, if you are predicting the number of wins for a basketball teams, don't use the number of points and the number of field goals, as these two variables mostly tell you the same thing. Likewise, in baseball, don't use both batting average and number of hits to predict runs scored, because batting average is calculated from the number of hits. When two explanatory variables have a strong association, it can have weird effects on the coefficients. Statisticians call this *multicollinearity*.

6. Pick explanatory variables that have linear relationships with the response variable. This can be verified by making separate scatterplots for each explanatory variable. If the relationship is nonlinear, then a more complicated model is needed.

Ideally, each explanatory variable that you include in the multiple regression model will have a strong, linear relationship with the response variable but have no relationship with any other explanatory variable. When an explanatory variable meets both of these requirements, it will almost certainly reduce the standard deviation of the residuals quite a bit and therefore produce better predictions of the response variable!

### STATS 101

In the previous chapter, we conducted tests to determine whether the value of the slope was statistically significant. If the *p*-value was small, then we had convincing evidence that there was an association between the explanatory and response variables.

When multiple regression models are calculated using statistical software, a *p*-value is calculated for each explanatory variable. If the *p*-value is small, then the reduction in the sum of squared residuals when the variable is added to the model is more than we would expect to happen by *RANDOM CHANCE*. In other words, when the *p*-value is small, there is convincing evidence that the additional explanatory variable provides additional useful information about the response variable, even after accounting for all of the other explanatory variables.

Here is one final example that illustrates how to build a multiple regression model.

### SPORTS TERM

**Earned run average**, or **ERA**, measures how many earned runs a pitcher gives up every 9 innings, on average. Unearned runs, which are runs that score because of a fielding error, are not included in a pitcher's ERA:

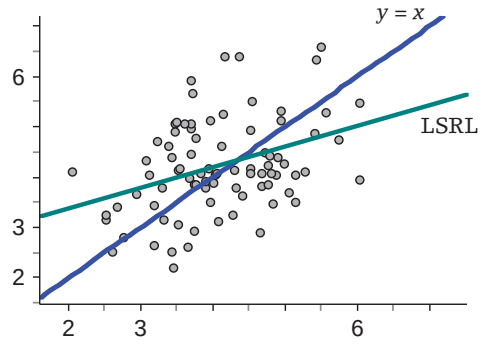
$$\text{ERA} = \frac{\text{number of earned runs}}{\text{number of innings pitched}} \times 9$$

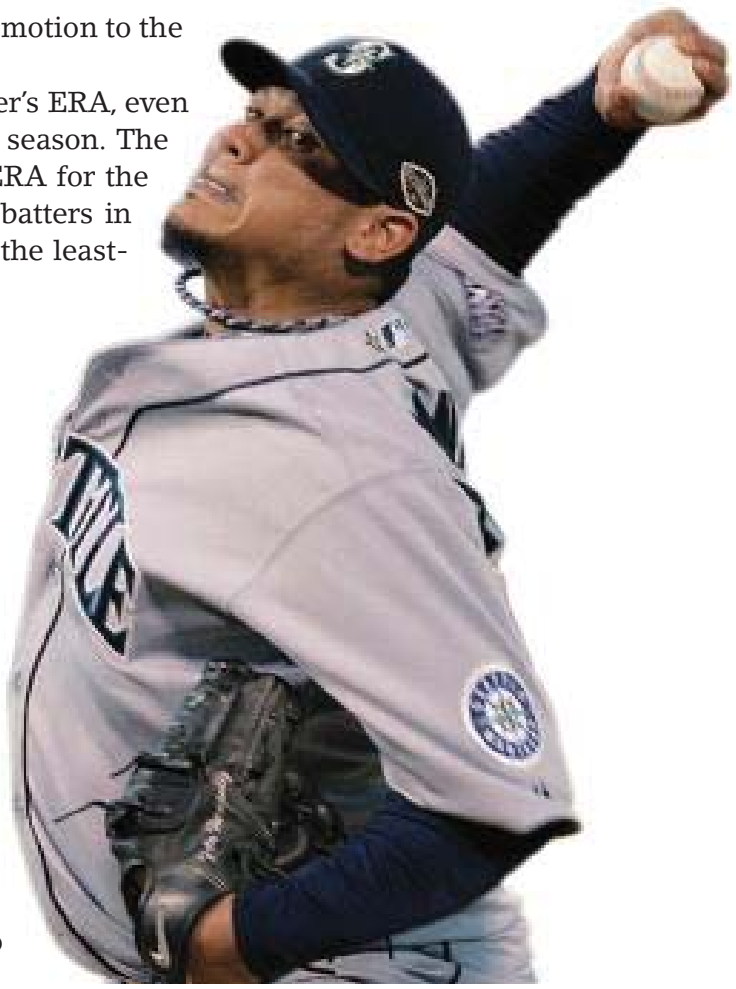
## Putting It All Together: Building a Model to Predict ERA

In baseball, a pitcher's **earned run average (ERA)** measures how many earned runs he gives up every nine innings, on average. A good ERA is 3.00 or below, meaning the pitcher gives up three or fewer earned runs per nine innings, on average. On the other hand,

ERAs above 6.00 will usually earn the pitcher a demotion to the minor leagues.

Unfortunately, it is very hard to predict a pitcher's ERA, even using the same pitcher's ERA from the previous season. The scatterplot below shows the 2008 ERA and 2009 ERA for the 83 Major League pitchers who faced at least 500 batters in both seasons. Also included on the scatterplot are the least-squares regression line and the line $y = x$.



Because the least-squares regression line is much flatter than the line $y = x$, the 2009 ERA values regressed to the mean quite a bit. In other words, pitchers who had low ERAs in 2008 tended to have low ERAs in 2009, but not quite as low. Likewise, pitchers who had high ERAs in 2008 tended to have high ERAs in 2009, but not quite as high.

One of the reasons it is so difficult to predict a pitcher's ERA is that ERA is affected by other factors besides the pitcher's ***ABILITY***. One obvious nonpitching factor that affects ERA is defense. Better defenders will help the pitcher prevent earned runs and vice versa. So any change in the ***ABILITY*** of the defense behind a pitcher will affect his ERA.

One of the biggest factors that affects a pitcher's ERA is ***RANDOM CHANCE***. Groundbreaking research published in the early 2000s showed that pitchers have very little influence on the outcome of a batted ball, other than home runs.* That is, a pitcher's **batting average on balls in play (BABIP)** shows very little correlation from year to year. For example, Kevin Millwood of the Texas Rangers gave up hits on 35.5% of the balls in play in 2008, but gave up hits on only 27.3% of balls in play in 2009. That is, when batters hit fair balls against Millwood, a much higher percentage of these batters reached base in 2008 than in 2009.

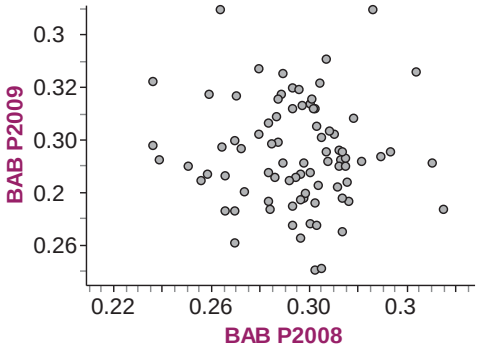
**SPORTS TERM**

**Batting average on balls in play (BABIP)** measures the proportion of times that a pitcher allows a batter to reach base when the batter hits the ball into the field of play. This excludes home runs, strikeouts, and walks. A typical BABIP is about 0.300:

$$\text{BABIP} = \frac{\text{hits} - \text{home runs}}{\text{batters faced} - \text{home runs} - \text{strikeouts} - \text{walks}}$$

*Conduct an Internet search for "defense-independent pitching statistics" or DIPS.

Here is a scatterplot of the 2008 BABIP and 2009 BABIP, using the same 83 pitchers as before:



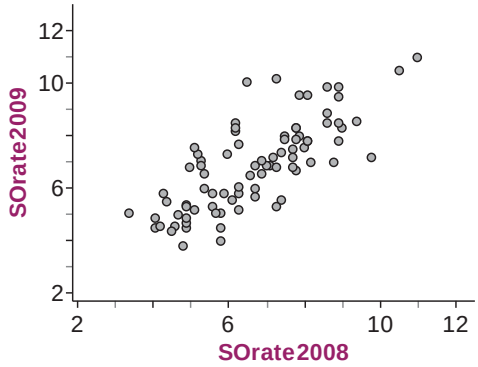

As you can see, there is virtually no correlation ($r = 0.05$) between BABIP in 2008 and BABIP in 2009. Thus, the difference in a pitcher's BABIP from one year to the next is almost entirely due to ***RANDOM CHANCE***! If a pitcher is unlucky and his BABIP is high, then he allows more hits and consequently more runs, hurting his ERA. On the other hand, if a pitcher is lucky, then his BABIP is low and so is his ERA.

**SPORTS TERM**

A pitcher's **strikeout rate** is the number of strikeouts he gets per 9 innings, on average. Likewise, his **walk rate** is the average number of walks he gives up per 9 innings and his **home run rate** is the average number of home runs he gives up per 9 innings.

Because ERA is greatly affected by ***RANDOM CHANCE***, a better way to predict a pitcher's future ***PERFORMANCE*** is to focus on variables over which the pitcher has control. These include the pitcher's **strikeout (SO) rate**, **walk rate**, and **home run rate**.

For example, here is a scatterplot showing the strikeout rates for each of the 83 pitchers in 2008 and 2009. As you can see, there is a much stronger correlation ($r = 0.76$) between strikeout rates in consecutive years than there is in either ERA in consecutive years or BABIP in consecutive years. This means that strikeout rate is a much more predictable variable from year to year.




The least-squares model for predicting a pitcher's 2009 ERA ($y$) using his 2008 strikeout rate ($x_1$), 2008 walk rate ($x_2$), and 2008 home run rate ($x_3$) is as follows:

$$\hat{y} = 4.1280 - 0.2022x_1 + 0.1984x_2 + 0.8486x_3$$
$$s = 0.83$$

Unfortunately, even using all three variables, our typical error will still be 0.83 earned runs per 9 innings. However, this is still better than if we used a pitcher's 2008 ERA to predict his 2009 ERA ($s = 0.87$). Even using three explanatory variables that are largely controlled by the

pitcher, it is difficult to predict ERA due to the *RANDOM CHANCE* involved in the game.

Of course, we can always try to decrease the standard deviation of the residuals $s$ by including additional variables to the model. For example, when adding the number of hits allowed in 2008 as a fourth explanatory variable, $s$ went down to 0.82. But when adding the pitcher's 2008 ERA as a fourth explanatory variable, the value of $s$ actually went up to 0.84, showing that the previous year's ERA doesn't provide any additional useful information about next year's ERA that wasn't already provided by strikeout rate, walk rate, and home run rate.

The table at right shows several different variables and the value of $s$ when each of these is used as a fourth explanatory variable.

Among these variables, only hits allowed provided any additional unique information about a pitcher's future ERA. So, other than possibly adding hits allowed to the model, there is no need to make an unnecessarily complicated model with lots of additional variables.

| VARIABLE (2008 VALUE) | $s$ |
|---|---|
| Hits allowed | 0.82 |
| Innings pitched | 0.83 |
| Wild pitches | 0.83 |
| Win percentage | 0.83 |
| Age | 0.83 |
| Lefty (1 = lefty, 0 = righty) | 0.84 |
| ERA | 0.84 |

## THINK ABOUT IT

Among the original three variables, which is the most important component for predicting a pitcher's ERA? It is difficult to tell by looking at the coefficients because they are measured on different scales. To solve this problem, we can use standardized scores ($z$-scores), which measure how many standard deviations a *PERFORMANCE* is above or below average:

$$z = \frac{\textit{PERFORMANCE} - \text{mean}}{\text{standard deviation}}$$

After standardizing each pitcher's strikeout rate ($z_{x1}$), walk rate ($z_{x2}$), and home run rate ($z_{x3}$), the multiple regression model is

$$\hat{y} = 4.1827 - 0.3277(z_{x1}) + 0.1661(z_{x2}) + 0.2379(z_{x3})$$
$$s = 0.83$$

Now that the explanatory variables are on the same scale, it makes sense to compare their coefficients. If a pitcher is 1 standard deviation better than average in strikeout rate ($z_{x1} = 1$), his predicted ERA will go down by 0.3277. However, if he is 1 standard deviation better than average ($z_{x2} = -1$) in walk rate, his predicted ERA will decrease by only 0.1661; if he is 1 standard deviation better than average ($z_{x3} = -1$) in home run rate, his predicted ERA will decrease by only 0.2379. Remember that it is better to be below average in walk rate and home run rate!

After considering the coefficients of the standardized variables, we can conclude that strikeout rate is about twice as important as walk rate and about 1.5 times more important than home run rate, as long as all three variables remain in the model. If one of these variables is removed or if other variables are added, the relative importance of each variable will change somewhat.*

* For an interesting discussion of this issue in football, see the following article: www.advancednflstats.com/2007/07/what-makes-teams-win-part-1.html.

## Connections: Looking Forward . . . Looking Back

Building on many of the concepts introduced in Chapters 10 and 11, in this chapter we used models with more than one explanatory variable to predict the values of a response variable. We continued to evaluate the quality of our predictions by calculating residuals and the standard deviation of the residuals *s*. In most cases, using more than one explanatory variable decreases the standard deviation of the residuals, meaning that our predictions will typically be closer to the actual values when using a multiple regression model.

We also interpreted the coefficients of a multiple regression model in much the same way that we interpreted the slope of a least-squares regression line in Chapter 11—the predicted change in the response variable for each one-unit increase in the explanatory variable. With multiple regression, however, we need to add the disclaimer that the values of the other variables must remain the same.

A new idea introduced in this chapter is the use of a categorical indicator variable to help predict the value of a numerical response variable. In the next chapter, we will reverse these roles and use a numerical variable to predict the value of a categorical indicator variable. We will also learn what to do when the association between two numerical variables is nonlinear.

# Chapter Summary

- A **multiple regression** model uses more than one explanatory variable to predict the value of a numerical response variable.
- The **coefficient** of a numerical explanatory variable describes how much the predicted value of the response variable will change for each one-unit increase in the explanatory variable, assuming that the values of the other explanatory variables stay the same.
- An **indicator variable** is a categorical explanatory variable with two possible outcomes. These outcomes are coded numerically so they can be included in regression calculations. Typically, a success is reported as a "1" and a failure is recorded as a "0."
- The coefficient of an indicator variable measures how much the predicted value of the response variable will change when the value of the indicator variable changes from 0 to 1, assuming that the values of the other explanatory variables stay the same.
- The **standard deviation of the residuals** ***s*** measures the typical distance from the predicted values to the actual values of the response variable:

$$s = \sqrt{\frac{rss}{n - (k + 1)}}$$

  where $rss$ = the sum of squared residuals, $n$ = the number of observations, and $k$ = the number of explanatory variables.
- To select which variables to include in a multiple regression model, use variables that decrease the standard deviation of the residuals $s$ when they are added to the model.

# For Practice

*Use the following information for Exercises 1–2.*

In Chapter 11, we used a "Pythagorean" model to predict the winning percentage of teams in the NFL using their points scored and their points allowed. We can also use multiple regression to predict the number of wins for a team ($y$) based on their points scored ($x_1$) and their points allowed ($x_2$). Using data from the 2010 NFL regular season, the following model was obtained:

$$\hat{y} = 9.2177 + 0.0264x_1 - 0.0298x_2$$
$$s = 1.30$$

**1.** In the 2010 regular season, the New York Jets scored 367 points and allowed 304 points.

**(a)** Predict the number of wins for the 2010 Jets.

**(b)** The Jets actually won 11 games in the 2010 regular season. Calculate and interpret their residual.

**(c)** Interpret the value $s = 1.30$.

**(d)** Explain how to determine which model, the multiple regression model or the Pythagorean model, gives better predictions.

**2.** In the 2010 regular season, the San Francisco 49ers scored 305 points and allowed 346 points.

**(a)** Predict the number of wins for the 2010 49ers.

**(b)** The 49ers actually won 6 games in the 2010 regular season. Calculate and interpret their residual.

**(c)** What is the benefit of using multiple regression with two explanatory variables rather than simply calculating a least-squares regression line using one explanatory variable?

*Use the following information for Exercises 3–6.*

In his column *Prospectus Hit and Run* on August 4, 2009,[2] author Jay Jaffe presents a multiple regression model to predict a baseball team's winning percentage ($y$) based on their winning percentage one year ago ($x_1$), winning percentage two years ago ($x_2$), and winning percentage three years ago ($x_3$). The model is

$$\hat{y} = 15.57 + 0.4517x_1 + 0.1401x_2 + 0.0968x_3$$

**3.** The St. Louis Cardinals posted a winning percentage of 53.1% in 2008, 48.1% in 2007, and 51.6% in 2006.

**(a)** Calculate the predicted winning percentage for the 2009 St. Louis Cardinals.

**(b)** Calculate and interpret the residual for the 2009 Cardinals who actually had a winning percentage of 56.2% in 2009.

**4.** The Kansas City Royals posted a winning percentage of 46.3% in 2008, 42.6% in 2007, and 38.3% in 2006.

**(a)** Calculate the predicted winning percentage for the 2009 Kansas City Royals.

**(b)** Calculate and interpret the residual for the 2009 Royals who actually had a winning percentage of 40.1% in 2009.

**5.** Suppose that a team won exactly 50% of their games in 2006, 2007, and 2008.

**(a)** Without doing any calculations, predict the team's winning percentage for 2009. Explain your reasoning.

**(b)** Calculate their predicted winning percentage to verify your answer to part (a).

**(c)** Based on the model, which would lead to a higher predicted winning percentage in 2009—winning 60% in 2008 and 50% in the other two years or winning 60% in 2006 and 50% in the other two years? Explain why this makes sense.

**6.** We know that consecutive *PERFORMANCES* for athletes often regress to the mean. What about the *PERFORMANCES* of teams?

**(a)** Predict the winning percentage of a team whose winning percentage was 60% in each of the last three years.

**(b)** Predict the winning percentage of a team whose winning percentage was 40% in each of the last three years.

**(c)** Explain how the results of parts (a) and (b) illustrate the concept of regression to the mean.

**7.** What attracts fans to hockey games? A winning team? Rough play on the ice? A multiple regression model was calculated using $y$ = average attendance, $x_1$ = number of wins, and $x_2$ = number of power play opportunities. Power plays occur when a player is sent to the penalty box for rough play, resulting in a numerical advantage for the opposing team. Using data from the 2010–2011 regular season, here is the least-squares model:

$$\hat{y} = 10{,}146 + 161.6x_1 + 0.603x_2$$
$$s = 2352$$

**(a)** Calculate and interpret the residual for the Toronto Maple Leafs, who had 37 wins, 601 power play opportunities, and an average attendance of 19,534.

**(b)** Interpret the value $s = 2352$.

**(c)** Interpret the coefficients of $x_1$ and $x_2$.

**(d)** Which seems to be a bigger attraction, winning games or rough play?

**8.** What attracts fans to baseball games? A winning team? Lots of home runs? A multiple regression model was calculated using $y$ = average attendance, $x_1$ = number of wins, and $x_2$ = number of home runs hit in home games by both teams. Using data from the 2010 regular season, here is the least-squares model:

$$\hat{y} = 4012 + 290.1x_1 + 17.04x_2$$
$$s = 8429$$

**(a)** Calculate and interpret the residual for the Tampa Bay Rays, who had 96 wins, 162 home runs hit in their home stadium, and an average attendance of 22,758.

**(b)** Interpret the value $s = 8429$.

**(c)** Interpret the coefficients of $x_1$ and $x_2$.

**(d)** Which seems to be a bigger attraction, winning games or home runs?

**9.** Can we predict the maximum bench press for athletes based on just their age and body mass index (BMI)? Using a sample of 57 female high school athletes from Georgia, the following model was calculated using $y$ = maximum bench press (in pounds), $x_1$ = age, and $x_2$ = BMI:

$$\hat{y} = -11.13 + 4.262x_1 + 1.149x_2$$
$$s = 11.53$$

**(a)** After controlling for BMI, can older athletes bench press more weight? Explain.

**(b)** After controlling for age, can athletes with larger BMI values bench press more weight? Explain.

**(c)** Explain why it isn't a good idea to interpret the value −11.13 in this context.

**(d)** Calculate and interpret the residual for the athlete who was 14 years old, had a BMI of 21.1, and had a maximum bench press of 80 pounds.

**10.** Can we predict the maximum bench press for athletes based on just their height and weight? Using a sample of 64 female Division I athletes from the University of Georgia, the following model was calculated using $y$ = maximum bench press (in pounds), $x_1$ = height (in inches), and $x_2$ = weight (in pounds):

$$\hat{y} = 188.12 - 3.0544x_1 + 0.8289x_2$$
$$s = 14.43$$

**(a)** If you compared two athletes who were the same height, would a heavier or lighter athlete be predicted to bench press more weight? Explain.

**(b)** If you compared two athletes who were the same weight, would a taller or shorter athlete be predicted to bench press more weight? Explain.

**(c)** Explain why it isn't a good idea to interpret the value 188.12 in this context.

**(d)** Calculate and interpret the residual for the athlete who was 68.5 inches tall, weighed 138 pounds, and had a maximum bench press of 100 pounds.

*Use the following information for Exercises 11–12.*

How much better is a home run than a double? How much better is a single than a walk? How much do stolen bases contribute to runs scored? These questions can be addressed by creating a multiple regression model to predict a team's runs ($y$) based on its number of singles ($x_1$), doubles ($x_2$), triples ($x_3$), home runs ($x_4$), walks ($x_5$), stolen bases ($x_6$), and outs made ($x_7$). Using data from the 2006–2010 Major League Baseball seasons,[3] the following model was obtained:

$$\hat{y} = -113.42 + 0.5218x_1 + 0.8743x_2 + 1.3239x_3 + 1.4472x_4 + 0.2886x_5 + 0.1346x_6 - 0.0843x_7$$
$$s = 21.6$$

**11.** In 2009, the Los Angeles Angels had 1105 singles, 293 doubles, 33 triples, 173 home runs, 547 walks, 148 stolen bases, and made 4070 outs.

**(a)** How many runs does the model predict that they will score?

**(b)** In 2009, the Angels actually scored 883 runs. Calculate and interpret their residual.

**(c)** Interpret the coefficient of singles and the coefficient of walks. Which event is more valuable? Explain why this might make sense.

**(d)** Interpret the value $s = 21.6$.

**12.** In 2009, the Los Angeles Dodgers had 1049 singles, 278 doubles, 39 triples, 145 home runs, 607 walks, 116 stolen bases, and made 4125 outs.

**(a)** How many runs does the model predict that they will score?

**(b)** In 2009, the Dodgers actually scored 780 runs. Calculate and interpret their residual.

**(c)** Interpret the coefficient of doubles and the coefficient of stolen bases. Which event is more valuable, getting a double or getting a single and stealing second base? Explain why this might make sense.

**(d)** When a batter is hit by a pitch, he is awarded first base, much like receiving a walk. When the variable $x_8$ = number of times hit by pitch was added to the model, the standard deviation of the residuals changed to $s = 21.1$. Does including this variable improve the model? Explain.

**13.** Do NHL hockey teams that play in Canada get an attendance boost because of the Canadian obsession with hockey? In Exercise 7, we used the number of wins and number of power play opportunities to predict the attendance of NHL teams in the 2010–2011 regular season. Adding an indicator variable ($x_3$) for teams in Canada (1 = Canadian, 0 = American), the new multiple regression model is

$$\hat{y} = 13{,}787 + 192.8x_1 - 8.922x_2 + 3083.7x_3$$
$$s = 2041$$

**(a)** Calculate and interpret the residual for the Toronto Maple Leafs, who had 37 wins, 601 power play opportunities, and an average attendance of 19,534.

**(b)** Interpret the coefficient of $x_3$.

**(c)** A different way to estimate the "Canada effect" on attendance would be to compare the average attendance for Canadian teams and the average attendance for American teams. Explain why using multiple regression is preferable to simply comparing the averages to estimate the "Canada effect."

**(d)** Did adding the indicator variable for Canadian teams improve the model? Explain how you know.

**(e)** If an NHL owner wanted to improve the attendance at his or her team's games, will moving to Canada ensure that attendance will go up? Explain.

**14.** Do baseball fans prefer going to games in air-conditioned, domed stadiums? In Exercise 8, we used the number of wins and number of home runs hit in home games to predict the attendance of MLB teams in the 2010 regular season. Adding an indicator variable ($x_3$) for teams that play in domed stadiums (1 = dome or retractable roof, 0 = open air), the new multiple regression model is

$$\hat{y} = 2496 + 270.0x_1 + 43.92x_2 - 5874x_3$$
$$s = 8319$$

**(a)** Calculate and interpret the residual for the Tampa Bay Rays, who had 96 wins, 162 home runs hit in their home stadium, play in a dome, and had an average attendance of 22,758.

**(b)** Interpret the coefficient of $x_3$.

**(c)** A different way to estimate the "dome effect" on attendance would be to compare the average attendance for teams that play in domes and the average attendance for teams that do not play in domes. Explain why using multiple regression is preferable to simply comparing the averages to estimate the "dome effect."

**(d)** Did adding the indicator variable for teams that play in domes improve the model? Explain how you know.

**(e)** If an MLB owner wanted to improve the attendance at his or her team's games, will moving to an open-air stadium ensure that attendance will go up? Explain.

**15.** Using data from 176 male golfers on the 2009 PGA tour, a least-squares model was calculated to predict scoring average ($y$) using $x_1$ = driving average (in yards), $x_2$ = driving accuracy (measured as a percent from 0 to 100), $x_3$ = putts per green-in-regulation (the average number of putts *per hole* when the green is reached in regulation), $x_4$ = sand save percent (the percentage of times a player hits his ball into a sand trap and successfully makes it into the hole in two or fewer additional strokes, measured as a percent from 0 to 100), and $x_5$ = over 40 years old (1 if over 40, 0 if 40 or younger). Remember that lower scores are better in golf! The model is

$$\hat{y} = 69.6 - 0.041x_1 - 0.072x_2 + 10.603x_3 - 0.0265x_4 - 0.026x_5$$
$$s = 0.34$$

**(a)** Interpret the value of $s = 0.34$.

**(b)** Predict the 2009 scoring average for Tiger Woods, who had a driving average of 298.4, a driving accuracy of 64.3%, a putts per green-in-regulation average of 1.743, a sand save percentage of 61.9%, and who is younger than 40 years old in 2009.

**(c)** Tiger Woods actually had a scoring average of 68.84 in 2009. Calculate and interpret his residual.

**(d)** Interpret the coefficients of driving average and putts per green-in-regulation.

**(e)** After factoring in other variables, does the model suggest that it helps to be older? Explain. Why might this be?

**16.** Using data from the 20 teams in the four-man bobsled competition at the 2010 Winter Olympics, a least-squares model was calculated to predict $y$ = race time (in seconds) using $x_1$ = average start time and $x_2$ = experience (1 if the team's country won a medal in the 2006 Winter Olympics and 0 otherwise).[4] Remember that smaller times are better! The model is

$$\hat{y} = 127.1855 + 16.4981x_1 - 0.4860x_2$$
$$s = 1.01$$

**(a)** Interpret the value of $s = 1.01$.

**(b)** Predict the race time for the gold-medal–winning United States team, which had an average start time of 4.7525 seconds and did not win a medal in 2006.

**(c)** The team from the United States finished with a time of 204.46 seconds. Calculate and interpret their residual.

**(d)** Interpret the coefficients of average start time and experience.

**(e)** We added a third variable, average weight of the four men on the team. When this variable was added, the coefficient was 0.01. Does this suggest that it helps or hurts to be heavier? Explain.

*Use the following data from the 2010 NBA playoffs for Exercises 17–20.*

| TEAM | REGULAR SEASON WINNING PERCENTAGE | TOP 4 MVP | MADE PLAYOFFS IN 2009 | PLAYOFF WINNING PERCENTAGE |
|---|---|---|---|---|
| Boston Celtics | 61.0 | 0 | 1 | 62.5 |
| Cleveland Cavaliers | 74.4 | 1 | 1 | 54.5 |
| Milwaukee Bucks | 56.1 | 0 | 0 | 42.9 |
| Chicago Bulls | 50.0 | 0 | 1 | 20.0 |
| Orlando Magic | 72.0 | 1 | 1 | 71.4 |
| Atlanta Hawks | 64.6 | 0 | 1 | 36.4 |
| Miami Heat | 57.3 | 0 | 1 | 20.0 |
| Charlotte Bobcats | 53.7 | 0 | 0 | 0.0 |
| Denver Nuggets | 64.6 | 0 | 1 | 33.3 |
| Utah Jazz | 64.6 | 0 | 1 | 40.0 |
| Portland Trail Blazers | 61.0 | 0 | 1 | 33.3 |
| Oklahoma City Thunder | 61.0 | 1 | 0 | 33.3 |
| Los Angeles Lakers | 69.5 | 1 | 1 | 69.6 |
| Phoenix Suns | 65.9 | 0 | 0 | 62.5 |
| Dallas Mavericks | 67.1 | 0 | 1 | 33.3 |
| San Antonio Spurs | 61.0 | 0 | 1 | 40.0 |

**17.** Does having a superstar player help teams perform better in the NBA playoffs? Create a multiple regression model to predict a team's winning percentage in the 2010 playoffs ($y$) based on the overall team quality measured by their regular season winning percentage ($x_1$) and whether or not they had a player who finished in the top four of the MVP voting ($x_2$). In 2010, the top four in the MVP voting were LeBron James (Cleveland), Kevin Durant (Oklahoma City), Kobe Bryant (Los Angeles), and Dwight Howard (Orlando).

**(a)** State the equation of the least-squares model.

**(b)** Calculate and interpret the standard deviation of the residuals $s$.

**(c)** Calculate and interpret the residual for the NBA champion Los Angeles Lakers.

**(d)** Does the model suggest that having a superstar helps in the playoffs? Explain.

**18.** Does having previous playoff experience help teams perform better in the NBA playoffs? Create a multiple regression model to predict a team's winning percentage in the 2010 playoffs ($y$) based on the overall team quality measured by their regular season winning percentage ($x_1$) and whether or not the team made the playoffs in 2009 ($x_2$).

**(a)** State the equation of the least-squares model.

**(b)** Calculate and interpret the standard deviation of the residuals $s$.

**(c)** Calculate and interpret the residual for the Oklahoma City Thunder.

**(d)** Does the model suggest that having previous playoff experience helps in the playoffs? Explain.

**19.** In Exercise 17, you calculated a multiple regression model to predict playoff winning percentage from regular season winning percentage and whether or not a team had a superstar player. Should the superstar variable stay in the model? Calculate the model, including the standard deviation of the residuals, with and without the superstar variable. Then explain whether the superstar variable provides additional useful information.

**20.** In Exercise 18, you calculated a multiple regression model to predict playoff winning percentage from regular season winning percentage and whether or not a team had playoff experience. Should the playoff experience variable stay in the model? Calculate the model, including the standard deviation of the residuals, with and without the playoff experience variable. Then explain whether the playoff experience variable provides additional useful information.

## CHAPTER REVIEW EXERCISES

**21.** Earlier in the chapter, we learned that an offensive turnover (losing the ball to the other team) in the NFL costs the offense about 2 potential points. To measure the offensive *benefit* of getting a defensive turnover (when the other team loses the ball), a multiple regression model was calculated using the number of points scored by an NFL team in 2009 ($y$), the number of yards gained ($x_1$), and the number of defensive turnovers ($x_2$). The least-squares model is

$$\hat{y} = -291.93 + 0.1009x_1 + 3.4687x_2$$
$$s = 35.0$$

**(a)** Calculate and interpret the residual for the 2009 Pittsburgh Steelers, who scored 368 points, gained 5941 yards, and gained 22 turnovers.

**(b)** Interpret the coefficients of $x_1$ and $x_2$.

**(c)** Using the model for offensive turnovers earlier in the chapter and the model for defensive turnovers in this exercise, what is the overall effect of a turnover in the NFL, considering both the loss to the team that turned the ball over and the gain to the team that caused the turnover?

**22.** In Chapter 10, we discovered that there was a positive relationship between a baseball team's winning percentage in spring training (exhibition) games and the team's regular season winning percentage. Of course, it isn't surprising that teams with good *ABILITY* will play well in both spring training and the regular season. So does a team's spring training performance tell us anything about teams we didn't already know? To find out, we can use a team's winning percentage from 2008 ($x_1$) and their spring training winning percentage in 2009 ($x_2$) to predict their winning percentage in the 2009 regular season ($y$). The multiple regression model is

$$\hat{y} = 20.60 + 0.3854x_1 + 0.2029x_2$$
$$s = 5.99\%$$

**(a)** Interpret the value $s = 5.99\%$.

**(b)** Which variable, 2008 winning percentage or 2009 spring training winning percentage, seems to be more important? Explain.

**(c)** If we were to drop a team's spring training winning percentage from the model, so that we were using only a team's 2008 winning percentage to predict their 2009 winning percentage, $s = 6.28\%$. Does this suggest that a team's spring training winning percentage provides additional useful information about their *PERFORMANCE* in the regular season?

**23.** In 2010, South Africa hosted the World Cup soccer tournament. The following table displays data for the 32 countries in the tournament. The variables are the average number of goals scored per match, shooting percentage, average number of passes per match, whether or not the team played in the 2006 World Cup, and whether or not the team is from Africa.

| COUNTRY | GOALS PER MATCH | SHOOTING PERCENTAGE | PASSES PER MATCH | EXPERIENCE IN 2006 | FROM AFRICA |
|---|---|---|---|---|---|
| Algeria | 0 | 0 | 353 | 0 | 1 |
| Argentina | 2 | 23 | 459 | 1 | 0 |
| Australia | 1 | 21 | 348 | 1 | 0 |
| Brazil | 1.8 | 29 | 451 | 1 | 0 |
| Cote d'Ivoire | 1.33 | 20 | 375 | 1 | 1 |
| Cameroon | 0.67 | 13 | 400 | 0 | 1 |
| Chile | 0.75 | 16 | 360 | 0 | 0 |
| Denmark | 1 | 19 | 345 | 0 | 0 |
| England | 0.75 | 10 | 395 | 1 | 0 |
| France | 0.33 | 9 | 335 | 1 | 0 |
| Germany | 2.29 | 38 | 409 | 1 | 0 |
| Ghana | 1 | 16 | 329 | 1 | 1 |
| Greece | 0.67 | 13 | 281 | 0 | 0 |
| Honduras | 0 | 0 | 252 | 0 | 0 |
| Italy | 1.33 | 22 | 407 | 1 | 0 |

| COUNTRY | GOALS PER MATCH | SHOOTING PERCENTAGE | PASSES PER MATCH | EXPERIENCE IN 2006 | FROM AFRICA |
|---|---|---|---|---|---|
| Japan | 1 | 15 | 223 | 1 | 0 |
| Korea DPR | 0.33 | 9 | 268 | 0 | 0 |
| Korea Republic | 1.5 | 27 | 332 | 1 | 0 |
| Mexico | 1 | 22 | 400 | 1 | 0 |
| Netherlands | 1.71 | 26 | 381 | 1 | 0 |
| New Zealand | 0.67 | 67 | 221 | 0 | 0 |
| Nigeria | 1 | 33 | 266 | 0 | 1 |
| Paraguay | 0.6 | 12 | 311 | 1 | 0 |
| Portugal | 1.75 | 33 | 365 | 1 | 0 |
| Serbia | 0.67 | 18 | 358 | 0 | 0 |
| Slovakia | 1.25 | 45 | 309 | 0 | 0 |
| Slovenia | 1 | 21 | 305 | 0 | 0 |
| South Africa | 1 | 17 | 353 | 0 | 1 |
| Spain | 1.14 | 17 | 543 | 1 | 0 |
| Switzerland | 0.33 | 11 | 277 | 1 | 0 |
| Uruguay | 1.57 | 24 | 270 | 0 | 0 |
| USA | 1.25 | 19 | 294 | 1 | 0 |

**(a)** Create a multiple regression model to predict $y$ = number of goals per match using $x_1$ = shooting percentage, $x_2$ = passes per match, and $x_3$ = experience in the 2006 World Cup. State the equation of the multiple regression model.

**(b)** Calculate and interpret the standard deviation of the residuals $s$ for the model you created in part (a).

**(c)** According to the model, did it help to have previous World Cup experience? Explain.

**(d)** Another way to measure the effect of previous experience would be to compare the average number of goals per match for teams that played in the previous World Cup and teams that did not play in the previous World Cup. Explain why it is preferable to use multiple regression to estimate the value of experience rather than simply comparing the averages.

**(e)** Is there a home-continent advantage in the World Cup? Recalculate the model, including $x_4$ = From Africa. State the equation of the multiple regression model and use it to determine whether there seems to be a home-continent advantage.

**(f)** Calculate and interpret the standard deviation of the residuals $s$ for the model you created in part (e). Was it helpful to include the variable From Africa in the model? Explain.

## OTHER APPLICATIONS

**24.** What factors affect the number of calories in food products at McDonald's?[5] Using 112 food items, a multiple regression model was created to predict $y$ = calories using $x_1$ = grams of fat, $x_2$ = grams of carbs, and $x_3$ = grams of protein. The multiple regression model is

$$\hat{y} = 0.22 + 9.00x_1 + 4.00x_2 + 3.95x_3$$
$$s = 4.15$$

**(a)** Calculate and interpret the residual of the Big Mac®, which has 540 calories, 29 grams of fat, 45 grams of carbs, and 25 grams of protein.

**(b)** Interpret the value $s = 4.15$.

**(c)** Interpret the coefficients of $x_1$, $x_2$, and $x_3$.

**(d)** If McDonald's wanted to reduce the calories in one of its products, would it be best to reduce the fat, reduce the carbs, or reduce the protein? Explain.

**25.** Are refrigerators with top freezers more energy-efficient than refrigerators with bottom freezers? Using data on 64 refrigerators from the May 2010 edition of *Consumer Reports*, the following model was calculated using $y$ = annual energy cost (in dollars), $x_1$ = claimed capacity (in cubic feet), and $x_2$ = top (1 = top, 0 = bottom):

$$\hat{y} = 73.15 + 0.0581x_1 - 18.40x_2$$
$$s = 17.67$$

After controlling for capacity, does the model indicate that refrigerators with top freezers are more energy-efficient? Explain.

**26.** How can you determine how much your used Honda Civic is worth? The following table gives information on 27 Honda Civics listed at www.CarMax.com. The variables are the asking price, age, number of miles, whether or not the car is an EX model, and whether or not the car is red.

| PRICE | AGE | MILES | EX | RED |
|---|---|---|---|---|
| 18,599 | 3 | 34,000 | 1 | 0 |
| 16,599 | 4 | 43,000 | 1 | 1 |
| 15,998 | 4 | 60,000 | 1 | 0 |
| 18,998 | 5 | 45,000 | 1 | 0 |
| 16,450 | 3 | 43,000 | 1 | 0 |
| 16,849 | 4 | 57,000 | 1 | 1 |
| 16,849 | 3 | 29,000 | 0 | 0 |
| 17,450 | 2 | 19,000 | 0 | 0 |
| 16,998 | 4 | 17,000 | 0 | 0 |

| PRICE | AGE | MILES | EX | RED |
|---|---|---|---|---|
| 17,998 | 2 | 23,000 | 0 | 0 |
| 10,998 | 8 | 99,000 | 1 | 0 |
| 18,998 | 4 | 59,000 | 1 | 0 |
| 15,599 | 4 | 56,000 | 1 | 1 |
| 15,599 | 3 | 50,000 | 0 | 1 |
| 17,599 | 3 | 31,000 | 0 | 0 |
| 17,599 | 3 | 42,000 | 1 | 0 |
| 15,998 | 3 | 28,000 | 0 | 0 |
| 16,599 | 3 | 34,000 | 0 | 0 |
| 19,998 | 2 | 55,000 | 1 | 0 |
| 18,599 | 3 | 40,000 | 1 | 0 |
| 18,998 | 2 | 35,000 | 1 | 0 |
| 19,998 | 1 | 6,000 | 0 | 0 |
| 17,998 | 2 | 24,000 | 0 | 0 |
| 17,998 | 3 | 29,000 | 1 | 0 |
| 15,998 | 5 | 45,000 | 1 | 0 |
| 16,998 | 3 | 30,000 | 0 | 0 |
| 21,599 | 1 | 9,000 | 1 | 0 |

**(a)** Create a multiple regression model to predict $y$ = asking price using $x_1$ = age, $x_2$ = miles, $x_3$ = EX, and $x_4$ = red. State the equation of the multiple regression model.

**(b)** Calculate and interpret the standard deviation of the residuals $s$ for the model you created in part (a).

**(c)** Interpret the coefficients of age and miles.

**(d)** According to the model, how much extra should you expect to pay for an EX model? Explain.

**(e)** Another way to estimate the extra cost of an EX model would be to compare the average cost of the EX models and the average cost of the LX models. Explain why it is preferable to use a multiple regression model to estimate the extra cost of an EX model rather than simply comparing the averages.

**(f)** According to the model, should you expect to pay more for a red Civic? Explain.

**(g)** Recalculate the model excluding $x_4$ = red and recalculate the standard deviation of the residuals. Does the variable red provide additional useful information about the price of a used Honda Civic? Explain.

## FOR INVESTIGATION

Use multiple regression to create a model for predicting a response variable that uses at least three explanatory variables. In your written report you should:

**(a)** Write a brief introduction identifying the context of your research and the reasons why you initially chose the variables that you used in the model.

**(b)** Include a table with the raw data (cite the source of the data).

**(c)** State the equation of the multiple regression model.

**(d)** Demonstrate how to use the model by making a prediction and calculating and interpreting the residual.

**(e)** Calculate and interpret the standard deviation of the residuals $s$.

**(f)** Interpret each of the coefficients.

**(g)** Discuss the value of each of the explanatory variables. In other words, if you removed a variable, what effect would this have on the standard deviation of the residuals $s$?