9 Two-Variable Regression Models

OVERVIEW

Regression models are the workhorses of data analysts in a wide range of fields in the social sciences. We begin this chapter with a discussion of fitting a line to a scatter plot of data, and then we discuss the additional inferences that can be made when we move from a correlation coefficient to a two-variable regression model. We include discussions of measures of goodness-of-fit and on the nature of hypothesis testing and statistical significance in regression models. Throughout this chapter, we present important concepts in text, mathematical formulae, and graphical illustrations. This chapter concludes with a discussion of the assumptions of the regression model and minimal mathematical requirements for estimation.

9.1 TWO-VARIABLE REGRESSION

In Chapter 8 we introduced three different bivariate hypothesis tests. In this chapter we add a fourth, two-variable regression. This is an important first step toward the multiple regression model – which is the topic of Chapter 10 – in which we are able to "control for" another variable (Z) as we measure the relationship between our independent variable of interest (X) and our dependent variable (Y). It is crucial to develop an in-depth understanding of two-variable regression before moving to multiple regression. In the sections that follow, we begin with an overview of the two-variable regression model, in which a line is fit to a scatter plot of data. We then discuss the uncertainty associated with the line and how we use various measures of this uncertainty to make inferences about the underlying population. This chapter concludes with a discussion of the assumptions of the regression model and the minimal mathematical requirements for model estimation.

9.2 FITTING A LINE: POPULATION ⇔ SAMPLE

The basic idea of two-variable regression is that we are fitting the "best" line through a scatter plot of data. This line, which is defined by its slope and *y*-intercept, serves as a **statistical model** of reality. In this sense, two-variable regression is very different from the three hypothesis-testing techniques that we introduced in Chapter 8; although those techniques allow hypothesis testing, they do not produce a statistical model. You may remember from a math course the formula for a line expressed as

$$Y = mX + b,$$

where b is the y-intercept and m is the slope – often explained as the "rise-over-run" component of the line formula. For a one-unit increase (run) in X, m is the corresponding amount of rise in Y (or fall in Y, if m is negative). Together these two elements (m and b) are described as the line's **parameters**.[1] You may remember exercises from junior high or high school math classes in which you were given the values of m and b and then asked to draw the resulting line on graph paper. Once we know these two parameters for a line, we can draw that line across any range of X values.[2]

In a two-variable regression model, we represent the y-intercept parameter by the Greek letter alpha (α) and the slope parameter by the Greek letter beta (β).[3] As foreshadowed by all of our other discussions of variables, Y is the dependent variable and X is the independent variable. Our theory about the underlying population in which we are interested is expressed in the **population regression model**:

$$Y_i = \alpha + \beta X_i + u_i.$$

Note that in this model there is one additional component, u_i, which does not correspond with what we are used to seeing in line formulae from math classes. This term is the **stochastic** or "random" component of our dependent variable. We have this term because we do not expect all of our data points to line up perfectly on a straight line. This corresponds directly with our discussion in earlier chapters about the probabilistic (as opposed to deterministic) nature of causal theories about political phenomena. We are, after all, trying to explain processes that involve human behavior. Because human beings are complex, there is bound to be a fair amount

[1] The term "parameter" is a synonym for "boundary" with a more mathematical connotation. In the description of a line, the parameters (m and b in this case) are fixed whereas the variables (X and Y in this case) vary.

[2] If this is not familiar to you, or if you merely want to refresh your memory, you may want to complete Exercise 1 at the end of this chapter before you continue reading.

[3] Different textbooks on regression use slightly different notation for these parameters, so it is important not to assume that all textbooks use the same notation when comparing across them.

of random noise in our measures of their behavior. Thus we think about the values of our dependent variable Y_i as having a systematic component, $\alpha + \beta X_i$, and a stochastic component, u_i.

As we have discussed, we rarely work with population data. Instead, we use sample data to make inferences about the underlying population of interest. In two-variable regression, we use information from the **sample regression model** to make inferences about the unseen population regression model. To distinguish between these two, we place hats ($\hat{}$) over terms in the sample regression model that are estimates of terms from the unseen population regression model. Because they have hats, we can describe $\hat{\alpha}$ and $\hat{\beta}$ as being **parameter estimates**. These terms are our best guesses of the unseen population parameters α and β. Thus the sample regression model is written as

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i.$$

Note that, in the sample regression model, α, β, and u_i get hats, but Y_i and X_i do not. This is because Y_i and X_i are values for cases in the population that ended up in the sample. As such, Y_i and X_i are values that are *measured* rather than *estimated*. We use them to estimate α, β, and the u_i values. The values that define the line are the estimated systematic components of Y. For each X_i value, we use $\hat{\alpha}$ and $\hat{\beta}$ to calculate the predicted value of Y_i, which we call $\hat{Y}_i$, where

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i.$$

This can also be written in terms of expectations,

$$E(Y|X_i) = \hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i,$$

which means that the expected value of Y given X_i (or $\hat{Y}_i$) is equal to our formula for the two-variable regression line. So we can now talk about each Y_i as having an estimated systematic component, $\hat{Y}_i$, and an estimated stochastic component, $\hat{u}_i$. We can thus write our model as

$$Y_i = \hat{Y}_i + \hat{u}_i,$$

and we can rewrite this in terms of $\hat{u}_i$ to get a better understanding of the estimated stochastic component:

$$\hat{u}_i = Y_i - \hat{Y}_i.$$

From this formula, we can see that the estimated stochastic component ($\hat{u}_i$) is equal to the difference between the actual value of the dependent variable (Y_i) and the predicted value of the dependent variable from our two-variable regression model. Another name for the estimated stochastic component is the **residual**. “Residual” is another word for “leftover,” and this is appropriate, because $\hat{u}_i$ is the leftover part of Y_i after we have drawn the line defined by $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$. Another way to refer to $\hat{u}_i$, which follows from the formula $\hat{u}_i = Y_i - \hat{Y}_i$, is to call it the **sample error term**. Because $\hat{u}_i$ is an estimate of u_i, a corresponding way of referring to u_i is to call it the **population error term**.

9.3 WHICH LINE FITS BEST? ESTIMATING THE REGRESSION LINE

Consider the scatter plot of data in Figure 9.1. Our task is to draw a straight line that describes the relationship between our independent variable X and our dependent variable Y. By "straight line," we mean a line with a single slope that does not change as we move from left to right in our figure. So, for instance, consider the line that we've drawn through this plot of data in Figure 9.2. It certainly meets the criteria of having a single slope that doesn't change. In fact, we can see from the figure that the formula for this line is $Y_i = 51 - 0.6X_i$. But, if we look around Figure 9.2, we can see that there are a lot of points that this line misses by a long distance. In fact, we can see a pattern: the points that are furthest from the line in Figure 9.2 are all in the lower-left and upper-right quadrants. This is because, as we know from our work with these same data in Chapter 8, the relationship between growth and presidential vote is positive.

So, how do we draw a better line? We clearly want to draw a line that comes as close as possible to the cases in our scatter plot of data. And because the data have a general pattern from lower-left to upper-right, we

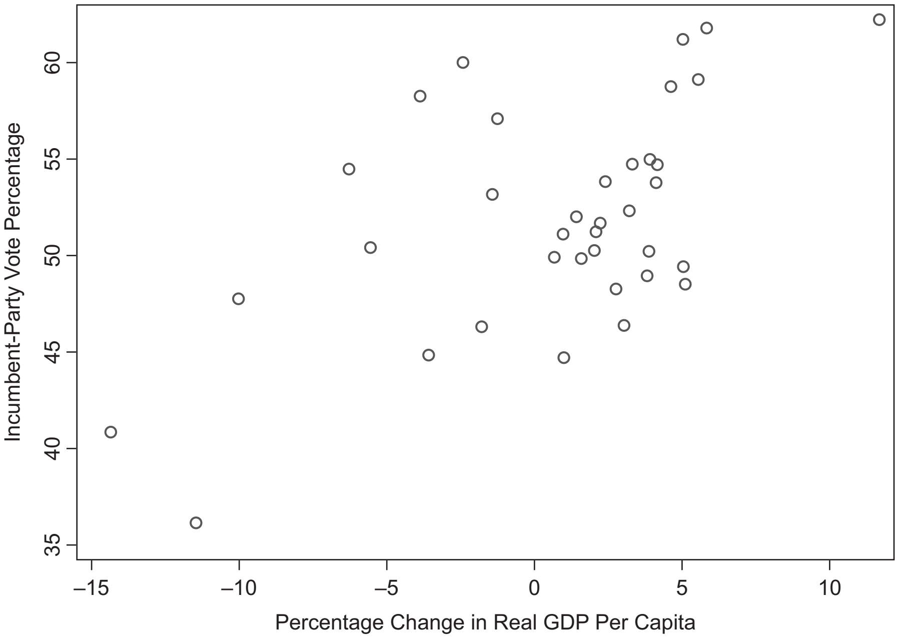


Figure 9.1 Scatter plot of change in GDP and incumbent-party vote share

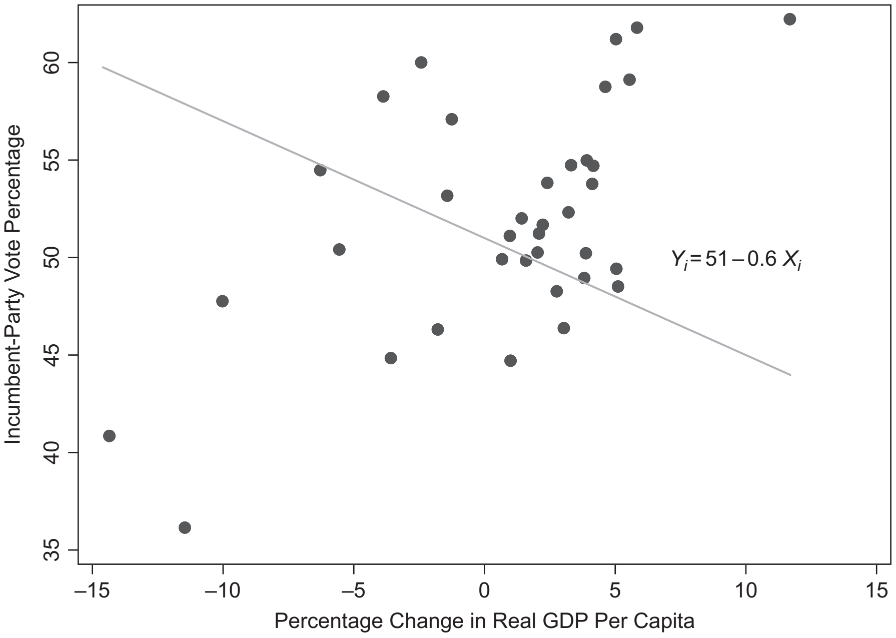


Figure 9.2 Scatter plot of change in GDP and incumbent-party vote share with a negatively sloped line

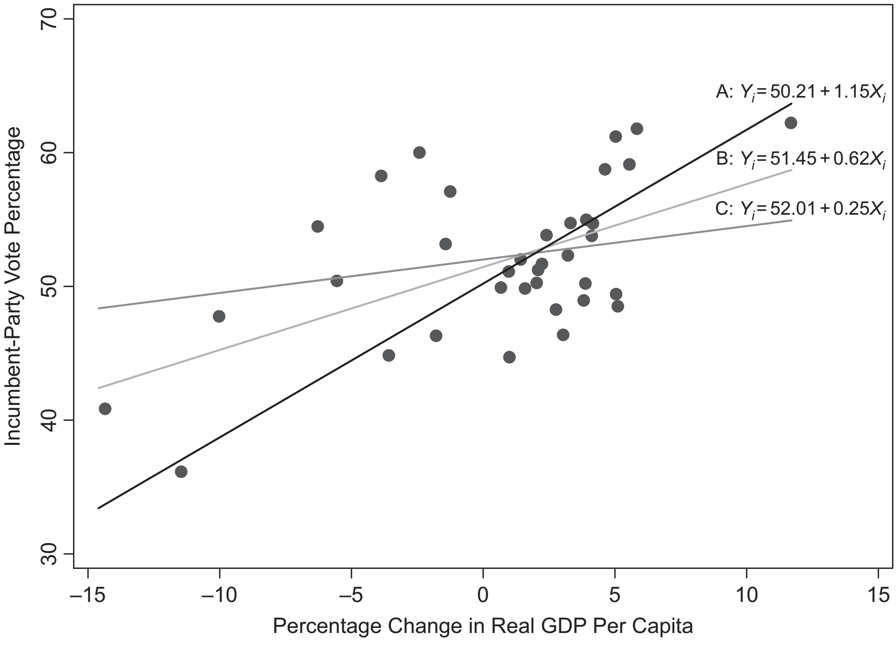


Figure 9.3 Three possible regression lines

know that our slope will be positive. In Figure 9.3, we have drawn three lines with positive slopes – labeled A, B, and C – through the scatter plot of growth and vote and written the corresponding parametric formula above each line on the right-hand side of the figure. So, how do we decide which

Table 9.1 Measures of total residuals for three different lines

Line	Parametric formula	$\sum_{i=1}^{n} \lvert\hat{u}_i\rvert$	$\sum_{i=1}^{n} \hat{u}_i^2$
A	$Y_i = 50.21 + 1.15X_i$	150.18	1085.58
B	$Y_i = 51.45 + 0.62X_i$	139.17	792.60
C	$Y_i = 52.01 + 0.25X_i$	148.22	931.68

line "best" fits the data that we see in our scatter plot of X_i and Y_i values? Because we are interested in explaining our dependent variable, we want our residual values, $\hat{u}_i$, which are vertical distances between each Y_i and the corresponding $\hat{Y}_i$, to be as small as possible. But, because these vertical distances come in both positive and negative values, we cannot just add them up for each line and have a good summary of the "fit" between each line and our data.[4]

So we need a method of assessing the fit of each line in which the positive and negative residuals do not cancel each other out. One possibility is to add together the absolute value of the residuals for each line:

$$\sum_{i=1}^{n} |\hat{u}_i|.$$

Another possibility is to add together the squared value of each of the residuals for each line:

$$\sum_{i=1}^{n} \hat{u}_i^2.$$

With either choice, we want to choose the line that has the smallest total value. Table 9.1 presents these calculations for the three lines in Figure 9.3.

From both calculations, we can see that line B does a better job of fitting the data than lines A and C. Although the absolute-value calculation is just as valid as the squared residual calculation, statisticians have tended to prefer the latter (both methods identify the same line as being "best"). Thus we draw a line that minimizes the sum of the *squared* residuals $\sum_{i=1}^{n} \hat{u}_i^2$. This technique for estimating the parameters of a regression model is known as **ordinary least-squares** (OLS) regression. For a two-variable OLS regression, the formulae for the parameter estimates of the line that meet this criterion are[5]

[4] Initially, we might think that we would want to minimize the sum of our residuals. But the line that minimizes the sum of the residuals is actually a flat line parallel to the x-axis. Such a line does not help us to explain the relationship between X and Y.

[5] The formulae for OLS parameter estimates come from setting the sum of squared residuals equal to zero and using differential calculus to solve for the values of $\hat{\beta}$ and $\hat{\alpha}$.

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$
$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}.$$

If we examine the formula for $\hat{\beta}$, we can see that the numerator is the same as the numerator for calculating the covariance between X and Y. Thus the logic of how each case contributes to this formula, as displayed in Figure 9.3, is the same. The denominator in the formula for $\hat{\beta}$ is the sum of squared deviations of the X_i values from the mean value of X ($\bar{X}$). Thus, for a given covariance between X and Y, the more (less) spread out X is, the less (more) steep the estimated slope of the regression line.

One of the mathematical properties of OLS regression is that the line produced by the parameter estimates goes through the sample mean values of X and Y. This makes the estimation of $\hat{\alpha}$ fairly simple. If we start out at the point defined by the mean value of X and the mean value of Y and then use the estimated slope ($\hat{\beta}$) to draw a line, the value of X where Y equals zero is $\hat{\alpha}$. Figure 9.4 shows the OLS regression line through the scatter plot of data. We can see from this figure that the OLS regression line passes through the point where the line depicting the mean value of X meets the line depicting the mean value of Y.

Using the data presented in Table 8.12 in the preceding formulae, we have calculated $\hat{\alpha} = 51.45$ and $\hat{\beta} = 0.62$, making our sample regression

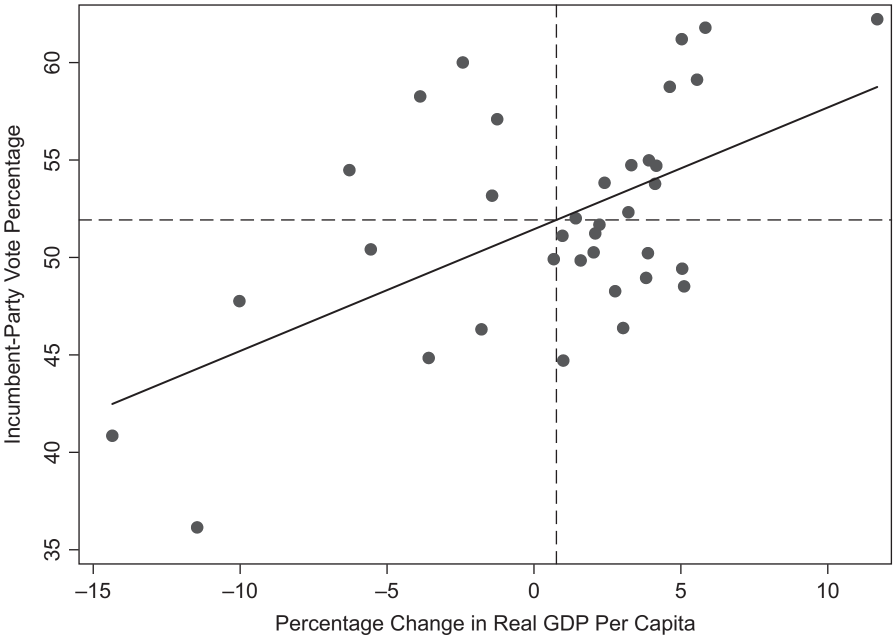


Figure 9.4 OLS regression line through scatter plot with mean-delimited quadrants

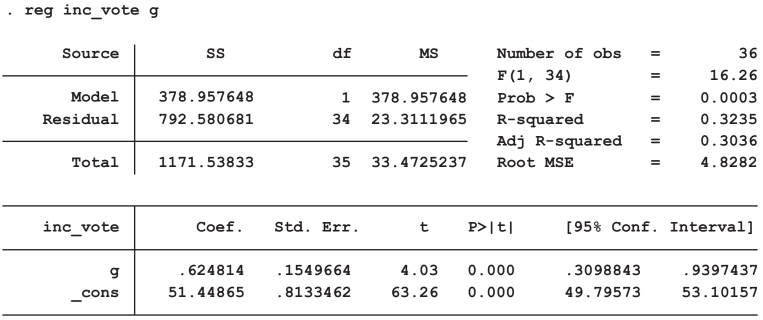
line formula $Y = 51.45 + 0.62X$. If we think about what this tells us about politics, we first need to remember that Y is the incumbent party's share of the major party vote, and X is the real per capita growth in GDP. So, if our measure of growth equals zero, we would expect the incumbent party to obtain 51.45 percent of the vote. If growth is not equal to zero, we multiply the value of growth by 0.62 and add (or subtract, if growth is negative) the result to 51.45 to obtain our best guess of the value of the vote. Moving to the right or the left along our sample regression line in Figure 9.4 means that we are increasing or decreasing the value of growth. For each right–left movement, we see a corresponding rise or decline in the value of the expected level of incumbent vote. If we go back to the logic of rise-over-run, our estimated slope parameter answers the question of how much change in Y we expect to see from a one-unit increase in X. In other words, a one-unit increase in our independent variable, growth, is expected to lead to a 0.62 increase in our dependent variable, incumbent vote.[6]

We can tell from Figure 9.4 that there are points that lie above and below our regression line. We therefore know that our model does not perfectly fit the real world. In the next section we discuss a series of inferences that we can make about the uncertainty associated with our sample regression model.

9.4 MEASURING OUR UNCERTAINTY ABOUT THE OLS REGRESSION LINE

As we have seen in Chapters 7 and 8, inferences about the underlying population of interest from sample data are made with varying degrees of uncertainty. In Chapter 8 we discussed the role of p-values in expressing this uncertainty. With an OLS regression model, we have several different ways in which to quantify our uncertainty. We discuss these measures in terms of the overall fit between X and Y first and then discuss the uncertainty about individual parameters. Our uncertainty about individual parameters is used in the testing of our hypotheses. Throughout this discussion, we refer to our example of fitting a regression line to our data on US presidential elections in order to test the theory of economic voting. Numerical results from Stata for this model are displayed in Figure 9.5. These numerical results can be partitioned into three separate areas. The

[6] Be sure not to invert the independent and dependent variables in describing results. It is *not* correct to interpret these results to say "for every 0.62-point change in growth rate in the US economy, we should expect to see, on average, an extra 1 percent in vote percentage for the incumbent party in presidential elections." Be sure that you can see the difference between those descriptions.



```
. reg inc_vote g

      Source |       SS           df       MS      Number of obs   =        36
-------------+----------------------------------   F(1, 34)        =     16.26
       Model |  378.957648         1  378.957648   Prob > F        =    0.0003
    Residual |  792.580681        34  23.3111965   R-squared       =    0.3235
-------------+----------------------------------   Adj R-squared   =    0.3036
       Total |  1171.53833        35  33.4725237   Root MSE        =    4.8282

------------------------------------------------------------------------------
    inc_vote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           g |    .624814   .1549664     4.03   0.000     .3098843    .9397437
       _cons |   51.44865   .8133462    63.26   0.000     49.79573    53.10157
------------------------------------------------------------------------------
```

Figure 9.5 Stata results for two-variable regression model between "vote" (inc_vote) and "growth" (g): inc_vote $= \alpha + \beta \times$ g

table in the upper-left corner of Figure 9.5 gives us measures of the variation in our model. The set of statistics listed in the upper-right corner of Figure 9.5 gives us a set of summary statistics about the entire model. Across the bottom of Figure 9.5 we get a table of statistics on the model's parameter estimates. The name of the dependent variable, "inc_vote," is displayed at the top of this table. Underneath we see the name of our independent variable, "g," which is short for "growth," and "_cons," which is short for "constant" (another name for the y-intercept term), which we also know as $\hat{\alpha}$. Moving to the right in the table at the bottom of Figure 9.5, we see that the next column heading here is "Coef.," which is short for "coefficient," which is another name for parameter estimate. In this column we see the values of $\hat{\beta}$ and $\hat{\alpha}$, which are 0.62 and 51.45 when we round these results to the second decimal place.[7]

9.4.1 Goodness-of-Fit: Root Mean-Squared Error

Measures of the overall fit between a regression model and the dependent variable are called goodness-of-fit measures. One of the most intuitive of these measures (despite its name) is **root mean-squared error** (root MSE). This statistic is sometimes referred to as the standard error of the regression model. It provides a measure of the average accuracy of the model in the metric of the dependent variable. This statistic ("Root MSE" in Figure 9.5) is calculated as

$$\text{root MSE} = \sqrt{\frac{\sum_{i=1}^{n} \hat{u}_i^2}{n}}.$$

[7] The choice of how many decimal places to report should be decided based on the value of the dependent variable. In this case, because our dependent variable is a vote percentage, we have chosen the second decimal place. Political scientists usually do not report election results beyond the first two decimal places.

The squaring and then taking the square root of the quantities in this formula are done to adjust for the fact that some of our residuals will be positive (points for which Y_i is above the regression line) and some will be negative (points for which Y_i is below the regression line). Once we realize this, we can see that this statistic is basically the average distance between the data points and the regression line.

From the numeric results depicted in Figure 9.5, we can see that the root MSE for our two-variable model of incumbent-party vote is 4.83. This value is found on the sixth line of the column of results on the right-hand side of Figure 9.5. It indicates that, on average, our model is off by 4.83 points in predicting the percentage of the incumbent party's share of the major party vote. It is worth emphasizing that the root MSE is always expressed in terms of the metric in which the dependent variable is measured. The only reason why this particular value corresponds to a percentage is because the metric of the dependent variable is vote percentage.

YOUR TURN: Evaluating a root MSE

In your opinion, is that root MSE "good"? Why or why not?

9.4.2 Goodness-of-Fit: *R*-Squared Statistic

Another commonly used indicator of the model's goodness-of-fit is the ***R*-squared statistic** (typically written as R^2). The R^2 statistic ranges between zero and one, indicating the proportion of the variation in the dependent variable that is accounted for by the model. The basic idea of the R^2 statistic is shown in Figure 9.6, which is a Venn diagram depiction

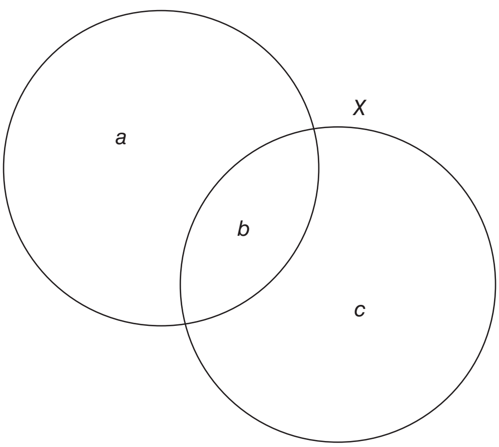


Figure 9.6 Venn diagram of variance and covariance for *X* and *Y*

of variation in X and Y as well as covariation between X and Y. The idea behind this diagram is that we are depicting variation in each variable with a circle. The larger the circle for a particular variable, the larger the variation for that variable. In this figure, the variation in Y consists of two areas, a and b, and variation in X consists of areas b and c. Area a represents variation in Y that is not related to variation in X, and area b represents covariation between X and Y. In a two-variable regression model, area a is the residual or stochastic variation in Y. The R^2 statistic is equal to area b over the total variation in Y, which is equal to the sum of areas a and b. Thus smaller values of area a and larger values of area b lead to a larger R^2 statistic. The formula for total variation in Y (areas a and b in Figure 9.6), also known as the total sum of squares (TSS), is

$$\text{TSS} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

The formula for the residual variation in Y, area a that is not accounted for by X, called the residual sum of squares (RSS), is

$$\text{RSS} = \sum_{i=1}^{n} \hat{u}_i^2.$$

Once we have these two quantities, we can calculate the R^2 statistic as

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

The formula for the other part of TSS that is not the RSS, called the model sum of squares (MSS), is

$$\text{MSS} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2.$$

This can also be used to calculate R^2 as

$$R^2 = \frac{\text{MSS}}{\text{TSS}}.$$

From the numeric results depicted in Figure 9.5, we can see that the R^2 statistic for our two-variable model of incumbent-party vote is 0.324. This number appears on the fourth line of the column of results on the right-hand side of Figure 9.5. It indicates that our model accounts for about 32 percent of the variation in the dependent variable. We can also see in Figure 9.5 the values for the MSS, RSS, and TSS under the column labeled “SS” in the table in the upper-left-hand corner.

YOUR TURN: Evaluating an *R*-squared statistic

In your opinion, is that *R*-squared "good"? Why or why not?

9.4.3 Is That a "Good" Goodness-of-Fit?

A logical question to ask when we see a measure of a model's goodness-of-fit is "What is a good or bad value for the root MSE and/or R^2?" This is not an easy question to answer. In part, the answer depends on what you are trying to do with the model. If you are trying to predict election outcomes, saying that you can predict the outcome with a typical error of 4.83 may not seem very good. After all, most presidential elections are fairly close and, in the scheme of things, 4.83 percent is a lot of votes. In fact, we can see that in 21 of the 36 elections that we are looking at, the winning margin was less than 4.83 percent, making over one-half of our sample of elections too close to call with this model. On the other hand, looking at this another way, we can say that we are able to come this close and, in terms of R^2, explain just over 32 percent of the variation in incumbent vote from 1876 to 2016 with just one measure of the economy. When we start to think of all of the different campaign strategies, personalities, scandals, wars, and everything else that is not in this simple model, this level of accuracy is rather impressive. In fact, we would suggest that this tells us something pretty remarkable about politics in the United States – the economy is massively important.

9.4.4 Uncertainty about Individual Components of the Sample Regression Model

Before we go through this section, we want to warn you that there are a lot of formulae in it. To use a familiar metaphor, as you go through the formulae in this section it is important to focus on the contours of the forest and not to get caught up in the details of the many trees that we will see along the way. Instead of memorizing each formula, concentrate on what makes the overall values generated by these equations larger or smaller.

A crucial part of the uncertainty in OLS regression models is the degree of uncertainty about individual estimates of population parameter values from the sample regression model. We can use the same logic that we discussed in Chapter 7 for making inferences from sample values about population values for each of the individual parameters in a sample regression model.

One estimate that factors into the calculations of our uncertainty about each of the population parameters is the estimated variance of the population stochastic component, u_i. This unseen variance, σ^2, is estimated from the residuals ($\hat{u}_i$) after the parameters for the sample regression model have been estimated by the following formula:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2}.$$

Looking at this formula, we can see two components that play a role in determining the magnitude of this estimate. The first component comes from the individual residual values ($\hat{u}_i$). Remember that these values (calculated as $\hat{u}_i = Y_i - \hat{Y}_i$) are the vertical distance between each observed Y_i value and the regression line. The larger these values are, the further the individual cases are from the regression line. The second component of this formula comes from n, the sample size. By now, you should be familiar with the idea that the larger the sample size, the smaller the variance of the estimate. This is the case with our formula for $\hat{\sigma}^2$.

Once we have estimated $\hat{\sigma}^2$, the variance and standard errors for the slope parameter estimate ($\hat{\beta}$) are then estimated from the following formulae:

$$\operatorname{var}(\hat{\beta}) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

$$\operatorname{se}(\hat{\beta}) = \sqrt{\operatorname{var}(\hat{\beta})} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}.$$

Both of these formulae can be broken into two components that determine their magnitude. In the numerator, we find $\hat{\sigma}$ values. So the larger these values are, the larger will be the variance and standard error of the slope parameter estimate. This makes sense, because the farther the points representing our data are from the regression line, the less confidence we will have in the value of the slope. If we look at the denominator in this equation, we see the term $\sum_{i=1}^{n}(X_i - \bar{X})^2$, which is a measure of the variation of the X_i values around their mean ($\bar{X}$). The greater this variation, the smaller will be the variance and standard error of the slope parameter estimate. This is an important property; in real-world terms it means that the more variation we have in X, the more precisely we will be able to estimate the relationship between X and Y.

The variance and standard errors for the intercept parameter estimate ($\hat{\alpha}$) are then estimated from the following formulae:

$$\operatorname{var}(\hat{\alpha}) = \frac{\hat{\sigma}^2 \sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n}(X_i - \bar{X})^2},$$

$$\operatorname{se}(\hat{\alpha}) = \sqrt{\operatorname{var}(\hat{\alpha})} = \sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n}(X_i - \bar{X})^2}}.$$

The logic for taking apart the components of these formulae is slightly more complicated because we can see that the sum of the squared X_i values appears in the numerator. We can see, however, that the denominator contains the measure of the variation of the X_i values around their mean ($\bar{X}$) multiplied by n, the number of cases. Thus the same basic logic holds for these terms: The larger the $\hat{u}_i$ values are, the larger will be the variance and standard error of the intercept parameter estimate; and the larger the variation of the X_i values around their mean, the smaller will be the variance and standard error of the intercept parameter estimate.

Less obvious – but nevertheless true – from the preceding formulae is the fact that larger sample sizes will also produce smaller standard errors.[8] So, just as we learned about the effects of sample size when calculating the standard error of the mean in Chapter 7, there is an identical effect here. Larger sample sizes will, other things being equal, produce smaller standard errors of our estimated regression coefficients.

9.4.5 Confidence Intervals about Parameter Estimates

In Chapter 7 we discussed how we use the normal distribution (supported by the central limit theorem) to estimate confidence intervals for the unseen population mean from sample data. We go through the same logical steps to estimate confidence intervals for the unseen parameters from the population regression model by using the results from the sample regression model. The formulae for estimating confidence intervals are

$$\hat{\beta} \pm [t \times \text{se}(\hat{\beta})],$$
$$\hat{\alpha} \pm [t \times \text{se}(\hat{\alpha})],$$

where the value for t is determined from the t-table in Appendix B. So, for instance, if we want to calculate a 95 percent confidence interval, this means that we are looking down the column for 0.025.[9] Once we have determined the appropriate column, we select our row based on the number of degrees of freedom. The number of degrees of freedom for this t-test is equal to the number of observations (n) minus the number of parameters estimated (k). In the case of the regression model presented in Figure 9.5, $n = 36$ and $k = 2$, so our degrees of freedom equal 34. Looking down the column for 0.025 and across the row for 30 in the t-table, we can see that $t = 2.042$. However, because we have 34 degrees

[8] It is true because the numerator of the expression contains $\hat{\sigma}$, which, as seen previously, has the sample size n in its denominator.

[9] To understand this, think back to Chapter 7, where we introduced confidence intervals. A 95 percent confidence interval would mean that would leave a total of 5 percent in the tails. Because there are two tails, we are going to use the 0.025 column.

of freedom, the t-values that leave 0.025 in each tail is 2.032.[10] Thus our 95 percent confidence intervals are

$$\hat{\beta} \pm [t \times \text{se}(\hat{\beta})] = 0.624814 \pm (2.032 \times 0.1549664) = 0.31 \text{ to } 0.94,$$

$$\hat{\alpha} \pm [t \times \text{se}(\hat{\alpha})] = 51.44865 \pm (2.032 \times 0.8133462) = 49.80 \text{ to } 53.10.$$

These values are displayed in the lower right-hand corner of the table at the bottom of Figure 9.5.

The traditional approach to hypothesis testing with OLS regression is that we specify a null hypothesis and an **alternative hypothesis** and then compare the two. Although we can test hypotheses about either the slope or the intercept parameter, we are usually more concerned with tests about the slope parameter. In particular, we are usually concerned with testing the hypothesis that the population slope parameter is equal to zero. The logic of this hypothesis test corresponds closely with the logic of the bivariate hypothesis tests introduced in Chapter 8. We observe a sample slope parameter, which is an estimate of the population slope. Then, from the value of this parameter estimate, the confidence interval around it, and the size of our sample, we evaluate how likely it is that we observe this sample slope if the true but unobserved population slope is equal to zero. If the answer is "very likely," then we conclude that the population slope is equal to zero.

To understand why we so often focus on a slope value of zero, think about what this corresponds to in the formula for a line. Remember that the slope is the change in Y from a one-unit increase in X. If that change is equal to zero, resulting in a flat line, then there is no covariation between X and Y, and we have failed to clear our third causal hurdle.

These types of tests are either one- or two-tailed. Most statistical computer programs report the results from two-tailed hypothesis tests that the parameter in question is not equal to zero. Despite this, many political science theories are more appropriately translated into one-tailed hypothesis tests, which are sometimes referred to as "directional" hypothesis tests. We review both types of hypothesis tests with the example regression from Figure 9.5.

9.4.6 Two-Tailed Hypothesis Tests

The most common form of statistical hypothesis tests about the parameters from an OLS regression model is a two-tailed hypothesis test that the slope parameter is equal to zero. It is expressed as

[10] The exact value of t is calculated automatically by statistical packages. For an online tool that gives exact values of t, go to https://www.danielsoper.com/statcalc/calculator.aspx?id=10.

$$H_0:\quad \beta = 0,$$
$$H_1:\quad \beta \neq 0,$$

where H_0 is the null hypothesis and H_1 is the alternative hypothesis. Note that these two rival hypotheses are expressed in terms of the slope parameter from the population regression model. To test which of these two hypotheses is supported, we calculate a ***t*-ratio** in which β is set equal to the value specified in the null hypothesis (in this case zero because H_0: $\beta = 0$), which we represent as β^*:

$$t_{n-k} = \frac{\hat{\beta} - \beta^*}{\text{se}(\hat{\beta})}.$$

For the slope parameter in the two-variable regression model presented in Figure 9.5, we can calculate this as

$$t_{34} = \frac{\hat{\beta} - \beta^*}{\text{se}(\hat{\beta})} = \frac{0.624814 - 0}{0.1549664} = 4.03.$$

From what we have seen in previous chapters, we can tell that this *t*-ratio is quite large. Remember that a typical standard for statistical significance in the social sciences is when the *p*-value is less than 0.05. If we look across the row for degrees of freedom equal to 30 in Appendix B, we can see that, to have a *p*-value of less than 0.05, we would need a *t*-ratio of 2.042 or larger (2.032 if we use the exact degrees of freedom). We clearly have exceeded this standard.[11] In fact, if we look at the far-right-hand column in Appendix B for 30 degrees of freedom, we can see that this *t*-ratio exceeds the value for *t* needed for *p* to be less than 0.002 (we get this by looking down the column labeled "0.001" and seeing a required *t*-value of at least 3.385 for 30 degrees of freedom). This means that it is extremely unlikely that H_0 is the case, which in turn greatly increases our confidence in H_1. If we look at the table at the bottom of Figure 9.5, we can see that the *t*-ratio and resulting *p*-value for this hypothesis test are presented in the fourth and fifth columns of the growth g row. It is worth noting that, although the reported *p*-value is 0.000, this does not mean that the probability of the null hypothesis being the case is actually equal to zero. Instead, this means that it is a very small number that gets rounded to zero when we report it to three decimal places.

The exact same logic is used to test hypotheses about the *y*-intercept parameter. The formula for this *t*-ratio is

[11] Because this is a two-tailed hypothesis test, for the standard of $p < 0.05$ we need to look down the column labeled "0.025." This is the case because we are going to leave 0.025 in each tail.

$$t_{n-k} = \frac{\hat{\alpha} - \alpha^*}{\text{se}(\hat{\alpha})}.$$

In Figure 9.5 we see the calculation for the following null hypothesis and alternative:

$$H_0: \quad \alpha = 0,$$
$$H_1: \quad \alpha \neq 0.$$

The resulting t-ratio is a whopping 63.26! This makes sense when we think about this quantity in real-world terms. Remember that the y-intercept is the expected value of the dependent variable Y when the independent variable X is equal to zero. In our model, this means we want to know the expected value of incumbent-party vote when growth equals zero. Even when the economy is shrinking, there are always going to be some diehard partisans who will vote for the incumbent party. Thus it makes sense that the null hypothesis H_0: $\alpha = 0$ would be pretty easy to reject.

Perhaps a more interesting null hypothesis is that the incumbents would still obtain 50 percent of the vote if growth were equal to zero. In this case,

$$H_0: \quad \alpha = 50,$$
$$H_1: \quad \alpha \neq 50.$$

The corresponding t-ratio is calculated as

$$t_{34} = \frac{\hat{\alpha} - \alpha^*}{\text{se}(\hat{\alpha})} = \frac{51.44865 - 50}{0.8133462} = 1.78.$$

Looking at the row for degrees of freedom equal to 30 in the t-table, we can see that this t-ratio is smaller than 2.042, which is the value for $p < 0.05$ (from the column labeled "0.025") but is larger than the 1.697 value for $p < 0.10$ (from the column labeled "0.05"). With a more detailed t-table or a computer, we could calculate the exact p-value for this hypothesis test, which is 0.08. Thus from these results, we are in a bit of a gray area. We can be pretty confident that the intercept is not equal to 50, but we can only reject the null hypothesis (H_0: $\alpha = 50$) at the 0.10 level instead of the widely accepted standard for statistical significance of 0.05. Let's think for a second, however, about our interest in the value of 50 for the intercept. While the hypothesis test for the alternative hypothesis that we just tested (H_1: $\alpha \neq 50$) is of interest to us, might we be more interested in whether or not incumbents would "win" the popular vote if the growth equaled zero? Before we approach this question, we will explain the relationship between confidence intervals and two-tailed hypothesis tests.

9.4.7 The Relationship between Confidence Intervals and Two-Tailed Hypothesis Tests

In the previous three sections, we introduced confidence intervals and hypothesis tests as two of the ways for making inferences about the parameters of the population regression model from our sample regression model. These two methods for making inferences are mathematically related to each other. We can tell this because they each rely on the t-table. The relationship between the two is such that, if the 95 percent confidence interval does not include a particular value, then the null hypothesis that the population parameter equals that value (a two-tailed hypothesis test) will have a p-value smaller than 0.05. We can see this for each of the three hypothesis tests that we discussed in the section on two-tailed hypothesis tests:

- Because the 95 percent confidence interval for our slope parameter does not include 0, the p-value for the hypothesis test that $\beta = 0$ is less than 0.05.
- Because the 95 percent confidence interval for our intercept parameter does not include 0, the p-value for the hypothesis test that $\alpha = 0$ is less than 0.05.
- Because the 95 percent confidence interval for our intercept parameter does include 50, the p-value for the hypothesis test that $\alpha = 50$ is greater than 0.05.

9.4.8 One-Tailed Hypothesis Tests

As we pointed out in previous sections, the most common form of statistical hypothesis tests about the parameters from an OLS regression model is a two-tailed hypothesis test that the slope parameter is equal to zero. That this is the most common test is something of a fluke. By default, most statistical computer programs report the results of this hypothesis test. In reality, though, *most political science hypotheses are that a parameter is either positive or negative and not just that the parameter is different from zero*. This is what we call a **directional hypothesis**. Consider, for instance, the theory of economic voting and how we would translate it into a hypothesis about the slope parameter in our current example. Our theory is that the *better* the economy is performing, the *higher* will be the vote percentage for the incumbent-party candidate. In other words, we expect to see a positive relationship between economic growth and the incumbent-party vote percentage, meaning that we expect β to be greater than zero.

When our theory leads to such a directional hypothesis, it is expressed as

$$\begin{aligned} H_0&: \ \beta \leq 0, \\ H_1&: \ \beta > 0, \end{aligned}$$

where H_0 is the null hypothesis and H_1 is the alternative hypothesis. As was the case with the two-tailed test, these two rival hypotheses are expressed in terms of the slope parameter from the population regression model. To test which of these two hypotheses is supported, we calculate a t-ratio where β is set equal to the value specified in the null hypothesis[12] (in this case zero because H_0: $\beta \leq 0$), which we represent as β^*:

$$t_{n-k} = \frac{\hat{\beta} - \beta^*}{\text{se}(\hat{\beta})}.$$

For the slope parameter in the two-variable regression model presented in Figure 9.5, we can calculate this as

$$t_{34} = \frac{\hat{\beta} - \beta^*}{\text{se}(\hat{\beta})} = \frac{0.624814 - 0}{0.1549664} = 4.03.$$

Do these calculations look familiar to you? They should, because this t-ratio is calculated exactly the same way that the t-ratio for the two-sided hypothesis about this parameter was calculated. The difference comes in how we use the t-table in Appendix B to arrive at the appropriate p-value for this hypothesis test. Because this is a one-tailed hypothesis test, we use the column labeled "0.05" instead of the column labeled "0.025" to assess whether we have achieved a t-ratio such that $p < 0.05$. In other words, we would need a t-ratio of only 1.697 for 30 degrees of freedom (1.691 for 34 degrees of freedom) to achieve this level of significance for a one-tailed hypothesis test. For a two-tailed hypothesis test, we needed a t-ratio of 2.047 (2.032).

Now, returning to our hypothesis test about the intercept and the value of 50, if we change from a two-tailed to a one-tailed hypothesis test,

$$\begin{aligned} H_0&: \ \alpha \leq 50, \\ H_1&: \ \alpha > 50, \end{aligned}$$

we still get

$$t_{34} = \frac{\hat{\alpha} - \alpha^*}{\text{se}(\hat{\alpha})} = \frac{51.44865 - 50}{0.8133462} = 1.78.$$

[12] We choose 0 when the null hypothesis is H_0: $\beta \leq 0$ because this is the critical value for the null hypothesis. Under this null hypothesis, zero is the threshold, and evidence that β is equal to any value less than or equal to zero is supportive of this null hypothesis.

But, with 34 degrees of freedom, this one-tailed hypothesis test yields a p-value of 0.04. In other words, this is a case where the formulation of our hypothesis test as one-tailed versus two-tailed makes a pretty major difference, especially since many scholars judge 0.05 to be the standard for statistical significance.

We can see from these examples and from the t-table that, when we have a directional hypothesis, we can more easily reject a null hypothesis. One of the quirks of political science research is that, even when they have directional hypotheses, researchers often report the results of two-tailed hypothesis tests. We'll discuss the issue of how to present regression results in greater detail in Chapter 10.

9.5 ASSUMPTIONS, MORE ASSUMPTIONS, AND MINIMAL MATHEMATICAL REQUIREMENTS

If assumptions were water, you'd need an umbrella right now. Any time that you estimate a regression model, you are implicitly making a large set of assumptions about the unseen population model. In this section, we break these assumptions into assumptions about the population stochastic component and assumptions about our model specification. In addition, there are some minimal mathematical requirements that must be met before a regression model can be estimated. In this final section we list these assumptions and requirements and briefly discuss them as they apply to our working example of a two-variable regression model of the impact of economic growth on incumbent-party vote.

9.5.1 Assumptions about the Population Stochastic Component

The most important assumptions about the population stochastic component u_i are about its distribution. These can be summarized as

$$u_i \sim N(0, \sigma^2),$$

which means that we assume that u_i is distributed normally ($\sim N$) with the mean equal to zero and the variance equal to σ^2.[13] This compact mathematical statement contains three of the five assumptions that we make about the population stochastic component any time we estimate a regression model. We now go over each one separately.

[13] Strictly speaking we do not need to make all of these assumptions to estimate the parameters of an OLS model. But we do need to make all of these assumptions to interpret the results from an OLS model in the standard fashion.

u_i Is Normally Distributed

The assumption that u_i is normally distributed allows us to use the t-table to make probabilistic inferences about the population regression model from the sample regression model. The main justification for this assumption is the central limit theorem that we discussed in Chapter 7.

$E(u_i) = 0$: No Bias

The assumption that u_i has a mean or expected value equal to zero is also known as the assumption of zero bias. Consider what it would mean if there was a case for which $E(u_i) \neq 0$. In other words, this would be a case for which we would *expect* our regression model to be off. If we have cases like this, we would essentially be ignoring some theoretical insight that we have about the underlying causes of Y. Remember, this term is supposed to be random. If $E(u_i) \neq 0$, then there must be some nonrandom component to this term. It is important to note here that we do not expect *all* of our u_i values to equal zero because we know that some of our cases will fall above and below the regression line. But this assumption means that our best guess or expected value for each individual u_i value is zero.

If we think about the example in this chapter, this assumption means that we do not have any particular cases for which we expect our model, with economic growth as the independent variable, to overpredict or underpredict the value of the incumbent-party vote percentage. If, on the other hand, we had some expectation along these lines, we would not be able to make this assumption. Say, for instance, that we expected that during times of war the incumbent party would fare better than we would expect them to fare based on the economy. Under these circumstances, we would not be able to make this assumption. The solution to this problem would be to include another independent variable in our model that measured whether or not the nation was at war at the time of each election. Once we control for all such potential sources of bias, we can feel comfortable making this assumption. The inclusion of additional independent variables is the main subject covered in Chapter 10.

u_i Has Variance σ^2: Homoscedasticity

The assumption that u_i has variance equal to σ^2 seems pretty straightforward. But, because this notation for variance does not contain an i subscript, it means that the variance for every case in the underlying population is assumed to be the same. The word for describing this situation is "homoscedasticity," which means "uniform error variance." If this assumption does not hold, we have a situation in which the variance of u_i is σ_i^2, known as "heteroscedasticity," which means "unequal error variance.” When we have heteroscedasticity, our regression model fits some of the cases in the population better than others. This can potentially cause us problems when we are estimating confidence intervals and testing hypotheses.

In our example for this chapter, this assumption would be violated if, for some reason, some elections were harder than others for our model to predict. In this case, our model would be heteroscedastic. It could, for instance, be the case that elections that were held after political debates became televised are harder to predict with our model in which the only independent variable is economic performance. Under these circumstances, the assumption of homoscedasticity would not be reasonable.

No Autocorrelation

We also assume that there is no autocorrelation. Autocorrelation occurs when the stochastic terms for any two or more cases are systematically related to each other. This clearly cuts against the grain of the idea that these terms are stochastic or random. Formally, we express this assumption as

$$\mathrm{cov}_{u_i,u_j} = 0 \quad \forall\, i \neq j;$$

in words, this means that the covariance between the population error terms u_i and u_j is equal to zero for all i not equal to j (for any two unique cases).

The most common form of autocorrelation occurs in regression models of time-series data. As we discussed in Chapter 4, time-series data involve measurement of the relevant variables across time for a single spatial unit. In our example for this chapter, we are using measures of economic growth and incumbent-party vote percentage measured every four years for the United States. If, for some reason, the error terms for adjacent pairs of elections were correlated, we would have autocorrelation.

X Values Are Measured without Error

At first, the assumption that X values are measured without error may seem to be out of place in a listing of assumptions about the population stochastic component. But this assumption is made to greatly simplify inferences that we make about our population regression model from our sample regression model. By assuming that X is measured without error, we are assuming that any variability from our regression line is due to the stochastic component u_i and not to measurement problems in X. To put it another way, if X also had a stochastic component, we would need to model X before we could model Y, and that would substantially complicate matters.

With just about any regression model that we estimate with real-world data, we will likely be pretty uncomfortable with this assumption. In the example for this chapter, we are assuming that we have exactly correct measures of the percentage change in real GDP per capita from 1876 to 2016. If we think a little more about this measure, we can think of all kinds of potential errors in measurement. What about illegal economic activities that are hard for the government to measure? Because this is per capita, how confident are we that the denominator in this calculation, population, is measured exactly correctly?

Despite the obvious problems with this assumption, we make it every time that we estimate an OLS model. Unless we move to considerably more complicated statistical techniques, this is an assumption that we have to live with and keep in the back of our minds as we evaluate our overall confidence in what our models tell us.

Recall from Chapter 5, when we discussed measuring our concepts of interest, that we argued that measurement is important because if we mismeasure our variables we may make incorrect causal inferences about the real world. This assumption should make the important lessons of that chapter crystal clear.

9.5.2 Assumptions about Our Model Specification

The assumptions about our model specification can be summarized as a single assumption that we have *the* correct model specification. We break this into two separate assumptions to highlight the range of ways in which this assumption might be violated.

No Causal Variables Left Out; No Noncausal Variables Included

This assumption means that if we specify our two-variable regression model of the relationship between X and Y there cannot be some other variable Z that also causes Y.[14] It also means that X must cause Y. In other words, this is just another way of saying that the sample regression model that we have specified *is* the true underlying population regression model.

As we have gone through the example in this chapter, we have already begun to come up with additional variables that we theorize to be causally related to our dependent variable. To comfortably make this assumption,

[14] One exception to this is the very special case in which there is a Z variable that is causally related to Y but Z is uncorrelated with X and u_i. In this case, we would still be able to get a reasonable estimate of the relationship between X and Y despite leaving Z out of our model. More on this in Chapter 10.

we will need to include all such variables in our model. Adding additional independent variables to our model is the subject of Chapter 10.

Parametric Linearity

The assumption of parametric linearity is a fancy way of saying that our population parameter β for the relationship between X and Y does not vary. In other words, the relationship between X and Y is the same across all values of X.

In the context of our current example, this means that we are assuming that the impact of a one-unit increase in change in real GDP per capita is always the same. So moving from a value of -10 to -9 has the same effect as moving from a value of 1 to 2. In Chapter 11 we discuss some techniques for relaxing this assumption.

9.5.3 Minimal Mathematical Requirements

For a two-variable regression model, we have two minimal requirements that must be met by our sample data before we can estimate our parameters. We will add to these requirements when we expand to multiple regression models.

X Must Vary

Think about what the scatter plot of our sample data would look like if X did not vary. Basically, we would have a stack of Y values at the same point on the x-axis. The only reasonable line that we could draw through this set of points would be a straight line parallel to the y-axis. Remember that our goal is to explain our dependent variable Y. Under these circumstances we would have failed miserably because any Y value would be just as good as any other given our single value of X. Thus we need some variation in X in order to estimate an OLS regression model.

$n > k$

To estimate a regression model, the number of cases (n) must exceed the number of parameters to be estimated (k). Thus, as a minimum, when we estimate a two-variable regression model with two parameters (α and β) we must have *at least* three cases.

9.5.4 How Can We Make All of These Assumptions?

The mathematical requirements to estimate a regression model aren't too severe, but a sensible question to ask at this point is, "How can we reasonably make all of the assumptions just listed every time that we run a regression model?" To answer this question, we refer back to the discussion in Chapter 1 of the analogy between models and maps. We *know* that all of our assumptions cannot possibly be met. We also know that we are trying to simplify complex realities. The only way that we can do this is to make a large set of unrealistic assumptions about the world. It is crucial, though, that we never lose sight of the fact that we are making these assumptions. In the next chapter we relax one of these most unrealistic assumptions made in the two-variable regression model by controlling for a second variable, *Z*.

CONCEPTS INTRODUCED IN THIS CHAPTER

- alternative hypothesis – the theory-based expectation that is the opposite of the null hypothesis
- directional hypothesis – an alternative hypothesis in which the expected relationship is either positive or negative
- ordinary least-squares – often abbreviated to "OLS," the most popular method for computing sample regression models
- parameter – a synonym for "boundary" with a more mathematical connotation; in the context of statistics, the value of an unknown population characteristic
- parameter estimate – a sample-based calculation of a population characteristic
- population error term – in the population regression model, the difference between the model-based predicted value of the dependent variable and the true value of the dependent variable
- population regression model – a theoretical formulation of the proposed linear relationship between at least one independent variable and a dependent variable
- residual – same as population error term
- root mean-squared error – sometimes shortened to "root MSE," a calculation of goodness-of-fit made by squaring each sample error term, summing them up, dividing by the number of cases, and then taking the square root; also known as the "model standard error"
- *R*-squared statistic – a goodness-of-fit measure that varies between 0 and 1 representing the proportion of variation in the dependent variable that is accounted for by the model
- sample error term – in the sample regression model, the sample-based estimate of the residual
- sample regression model – a sample-based estimate of the population regression model

- statistical model – a numerical representation of a relationship between at least one independent variable and a dependent variable
- stochastic – random
- t-ratio – the ratio of an estimated parameter to its estimated standard error

EXERCISES

1. Draw an X–Y axis through the middle of a 10×10 grid. The point where the X and Y lines intersect is known as the "origin" and is defined as the point at which both X and Y are equal to zero. Draw each of the following lines across the values of X from -5 to 5 and write the corresponding regression equation:

   (a) y-intercept = 2, slope = 2;
   (b) y-intercept = –2, slope = 2;
   (c) y-intercept = 0, slope = –1;
   (d) y-intercept = 2, slope = –2.

2. Solve each of the following mathematical expressions to yield a single component of the two-variable sample regression model:

   (a) $\hat{\alpha} + \hat{\beta}X_i + \hat{u}_i$
   (b) $Y_i - E(Y|X_i)$
   (c) $\hat{\beta}X_i + \hat{u}_i - Y_i$

3. Using the data set "state_data.dta" (which is available on the textbook's web site at www.cambridge.org/fpsr), we estimated a two-variable regression model using data from each US state and the District of Columbia with per capita income ("*pcinc*" in our data set) as our dependent variable and the percentage of state residents with a college degree ("*pctba*" in our data set) as the independent variable. The estimated equation was:

$$pcinc_i = 11519.78 + 1028.96 pctba_i.$$

   Interpret the parameter estimate for the effect of a state's level of education on average income levels.

4. In the data set described in Exercise 3, the value of *pctba* for Illinois equals 29.9. What is the model's predicted per capita income for Illinois?

5. The estimated standard error for the slope parameter in the model described in Exercise 3 was 95.7. Construct a 95 percent confidence interval for this parameter estimate. Show your work. What does this tell you about the estimated relationship?

6. Test the hypothesis that the parameter for *pctba* is not equal to zero. Show your work. What does this tell you about the estimated relationship?

7. Test the hypothesis that the parameter for *pctba* is greater than zero. Show your work. What does this tell you about the estimated relationship?

8. The R-squared statistic for the model described in Exercise 3 is 0.70 and the root MSE = 3773.8. What do these numbers tell us about our model?

9. Estimate and interpret the results from a two-variable regression model different from the model in Exercise 3 using the data set "state_data.dta."

10. Think through the assumptions that you made when you carried out Exercise 9. Which do you feel least and most comfortable making? Explain your answers.

11. In Exercise 10 for Chapter 8, you calculated a correlation coefficient for the relationship between two continuous variables. Now, estimate a two-variable regression model for these same two variables. Produce a table of the results and write about what this table tells you about politics in the United Kingdom in 2005.

10 Multiple Regression: the Basics

OVERVIEW

Despite what we have learned in the preceding chapters on hypothesis testing and statistical significance, we have not yet crossed all four of our hurdles for establishing causal relationships. Recall that all of the techniques we have learned in Chapters 8 and 9 are simply bivariate, X- and Y-type analyses. But, to fully assess whether X *causes* Y, we need to control for other possible causes of Y, which we have not yet done. In this chapter, we show how multiple regression – which is an extension of the two-variable model we covered in Chapter 9 – does exactly that. We explicitly connect the formulae that we include to the key issues of research design that tie the entire book together. We also discuss some of the problems in multiple regression models when key causes of the dependent variable are omitted, which ties this chapter to the fundamental principles presented in Chapters 3 and 4. Lastly, we will incorporate an example from the political science literature that uses multiple regression to evaluate causal relationships.

10.1 MODELING MULTIVARIATE REALITY

From the very outset of this book, we have emphasized that almost all interesting phenomena in social reality have more than one cause. And yet most of our theories are simply bivariate in nature.

We have shown you (in Chapter 4) that there are distinct methods for dealing with the nature of reality in our designs for social research. If we are fortunate enough to be able to conduct an experiment, then the process of randomly assigning our participants to treatment groups will automatically "control for" those other possible causes that are not a part of our theory.

But in observational research – which represents the vast majority of political science research – there is no automatic control for the other possible causes of our dependent variable; we have to control for them statistically. The main way that social scientists accomplish this is through multiple regression. The math in this model is an extension of the math involved in the two-variable regression model you just learned in Chapter 9.

In this book, we have made quite a big deal out of the need to "control for" alternative explanations. And before we introduce the idea of *statistical* control for Z – which we'll do starting in the next section – we want to distinguish between the *statistical control* that you're about to learn about and the *experimental control* that arises from controlling and randomly assigning values of X in an experiment.[1] Both terms, of course, feature the word "control," and therefore you might be tempted to equate "statistical control" with "experimental control." Experimental control is the far stronger version of control; in fact, as we have emphasized, it is the gold standard for scientific investigations of causality. Statistical control is an imperfect kind of control, and considerably less strong than its experimental counterpart. We'll draw your attention to this again below where it is appropriate, but we want to be sure you're on the lookout for those signs.

10.2 THE POPULATION REGRESSION FUNCTION

We can generalize the population regression model that we learned in Chapter 9,

$$\text{bivariate population regression model:} \quad Y_i = \alpha + \beta X_i + u_i,$$

to include more than one systematic cause of Y, which we have been calling Z throughout this book:

$$\text{multiple population regression model:} \quad Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + u_i.$$

The interpretation of the slope coefficients in the three-variable model is similar to interpreting bivariate coefficients, with one very important difference. In both, the coefficient in front of the variable X (β in the two-variable model, β_1 in the multiple regression model) represents the "rise-over-run" effect of X on Y. In the multiple regression case, though, β_1 actually represents the effect of X on Y *while holding constant the effects of* Z. If this distinction sounds important, it is. We show how these differences arise in the next section.

[1] You will recall, from Chapter 4, that the two components of the definition of an experiment are that the researcher be in control of the values of X that the participants are exposed to, and that those values are assigned to the participants randomly.

10.3 FROM TWO-VARIABLE TO MULTIPLE REGRESSION

Recall from Chapter 9 that the formula for a two-variable regression line (in a sample) is

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i.$$

And recall that, to understand the nature of the effect that X has on Y, the estimated coefficient $\hat{\beta}$ tells us, on average, how many units of change in Y we should expect given a one-unit increase in X. The formula for $\hat{\beta}$ in the two-variable model, as we learned in Chapter 9, is

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Given that our goal is to control for the effects of some third variable, Z, how exactly is that accomplished in regression equations? If a scatter plot in two dimensions (X and Y) suggests the formula for a *line*, then adding a third dimension suggests the formula for a *plane*. And the formula for that plane is

$$Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i.$$

That might seem deceptively simple. A formula representing a plane simply adds the additional $\beta_2 Z_i$ term to the formula for a line.[2]

Pay attention to how the notation has changed. In the two-variable formula for a line, there were no numeric subscripts for the β coefficient – because, well, there was only one of them. But now we have two independent variables, X and Z, that help to explain the variation in Y, and therefore we have two different β coefficients, and so we subscript them β_1 and β_2 to be clear that the values of these two effects are different from one another.[3]

The key message from this chapter is that, in the preceding formula, the coefficient β_1 represents more than the effect of X on Y; in the multiple regression formula, it represents *the effect of X on Y while controlling for the effect of Z*. Simultaneously, the coefficient β_2 represents *the effect of Z on Y while controlling for the effect of X*. And in observational research,

[2] All of the subsequent math about adding one more independent variable, Z, generalizes quite easily to adding still more independent variables. We use the two-independent-variable case for ease of illustration.

[3] In many other textbooks on regression analysis, just as we distinguish between β_1 and β_2, the authors choose to label their independent variables X_1, X_2, and so forth. We have consistently used the notation of X, Y, and Z to emphasize the concept of controlling for other variables while examining the relationship between an independent and a dependent variable of theoretical interest. Therefore we will stick with this notation throughout this chapter.

this is the key to crossing our fourth causal hurdle that we introduced all the way back in Chapter 3.

How is it the case that the coefficient for β_1 actually controls for Z? After all, β_1 is not connected to Z in the formula; it is, quite obviously, connected to X. The first thing to realize here is that the preceding multiple regression formula for β_1 is different from the two-variable formula for β from Chapter 9. (We'll get to the formula shortly.) The key consequence of this is that the value of β derived from the two-variable formula, representing the effect of X on Y, will almost always be different – perhaps only trivially different, or perhaps wildly different – from the value of β_1 derived from the multiple regression formula, representing the effect of X on Y while controlling for the effects of Z.

But how does β_1 control for the effects of Z? Let's assume that X and Z are correlated. They need not be related in a *causal* sense, and they need not be related *strongly*. They simply have to be correlated with one another – that is, for this example, their covariance is not exactly equal to zero. Now, assuming that they are related somehow, we can write their relationship just like that of a two-variable regression model:

$$X_i = \hat{\alpha}' + \hat{\beta}' Z_i + \hat{e}_i.$$

Note some notational differences here. Instead of the parameters $\hat{\alpha}$ and $\hat{\beta}$, we are calling the estimated parameters $\hat{\alpha}'$ and $\hat{\beta}'$ just so you are aware that their values will be different from the $\hat{\alpha}$ and $\hat{\beta}$ estimates in previous equations. And note also that the residuals, which we labeled $\hat{u}_i$ in previous equations, are now labeled $\hat{e}_i$ here.

If we use Z to predict X, then the predicted value of X (or $\hat{X}$) based on Z is simply

$$\hat{X}_i = \hat{\alpha}' + \hat{\beta}' Z_i,$$

which is just the preceding equation, but without the error term, because it is expected (on average) to be zero. Now, we can just substitute the left-hand side of the preceding equation into the previous equation, and get

$$X_i = \hat{X}_i + \hat{e}_i$$

or, equivalently,

$$\hat{e}_i = X_i - \hat{X}_i.$$

These $\hat{e}_i$, then, are the exact equivalents of the residuals from the two-variable regression of Y on X that you learned from Chapter 9. So their interpretation is identical, too. That being the case, the $\hat{e}_i$ are the portion of the variation in X that Z cannot explain. (The portion of X that Z *can* explain is the predicted portion – the $\hat{X}_i$.)

So what have we done here? We have just documented the relationship between Z and X and partitioned the variation in X into two parts – the portion that Z *can* explain (the $\hat{X}_i$) and the portion that Z *cannot* explain (the $\hat{e}_i$). Hold this thought.

We can do the exact same thing for the relationship between Z and Y that we just did for the relationship between Z and X. The process will look quite similar, with a bit of different notation to distinguish the processes. So we can model Y as a function of Z in the following way:

$$Y_i = \hat{\alpha}^* + \hat{\beta}^* Z_i + \hat{v}_i.$$

Here, the estimated slope is $\hat{\beta}^*$ and the error term is represented by $\hat{v}_i$.

Just as we did with Z and X, if we use Z to predict Y, then the predicted value of Y (or $\hat{Y}$) (which we will label $\hat{Y}^*$) based on Z is simply

$$\hat{Y}_i^* = \hat{\alpha}^* + \hat{\beta}^* Z_i,$$

which, as before, is identical to the preceding equation, but without the error term, because the residuals are expected (on average) to be zero. And again, as before, we can substitute the left-hand side of the preceding equation into the previous equation, and get

$$Y_i = \hat{Y}_i^* + \hat{v}_i$$

or, equivalently,

$$\hat{v}_i = Y_i - \hat{Y}_i^*.$$

These $\hat{v}_i$, then, are interpreted in an identical way to that of the preceding $\hat{e}_i$. They represent the portion of the variation in Y that Z cannot explain. (The portion of Y that Z *can* explain is the predicted portion – the $\hat{Y}_i^*$.)

Now what has this accomplished? We have just documented the relationship between Z and Y and partitioned the variation in Y into two parts – the portion that Z *can* explain and the portion that Z *cannot* explain.

So we have now let Z try to explain X and found the residuals (the $\hat{e}_i$ values); similarly, we have also now let Z try to explain Y, and found those residuals as well (the $\hat{v}_i$ values). Now back to our three-variable regression model that we have seen before, with Y as the dependent variable, and X and Z as the independent variables:

$$Y_i = \hat{\alpha} + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i + \hat{u}_i.$$

The formula for $\hat{\beta}_1$, representing the effect of X on Y while controlling for Z, is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{e}_i \hat{v}_i}{\sum_{i=1}^{n} \hat{e}_i^2}.$$

Now, we know what $\hat{e}_i$ and $\hat{v}_i$ are from the previous equations. So, substituting, we get

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \hat{X}_i)(Y_i - \hat{Y}_i^*)}{\sum_{i=1}^{n}(X_i - \hat{X}_i)^2}.$$

Pay careful attention to this formula. The "hatted" components in these expressions are from the two-variable regressions involving Z that we previously learned about. The key components of the formula for the effect of X on Y, while controlling for Z, are the $\hat{e}_i$ and $\hat{v}_i$, which, as we just learned, are the portions of X and Y (respectively) that Z cannot account for. And that is how, in the multiple regression model, the parameter β_1, which represents the effects of X on Y, *controls for* the effects of Z. How? Because the only components of X and Y that it uses are components that Z cannot account for – that is, the $\hat{e}_i$ and $\hat{v}_i$.

Comparing this formula for estimating β_1 with the two-variable formula for estimating β is very revealing. Instead of using the factors $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ in the numerator, which were the components of the *two-variable* regression of Y on X from Chapter 9, in the multiple regression formula that controls for Z, the factors in the numerator are $(X_i - \hat{X}_i)$ and $(Y_i - \hat{Y}_i^{\,*})$, where, again, the hatted portions represent X as predicted by Z and Y as predicted by Z.

Note something else in the comparison of the two-variable formula for estimating β and the multiple regression formula for estimating β_1. The result of $\hat{\beta}$ in the two-variable regression of Y and X and $\hat{\beta}_1$ in the three-variable regression of Y on X while controlling for Z will be different almost all the time. In fact, it is quite rare – though mathematically possible in theory – that those two values will be identical.[4]

And the formula for estimating β_2, likewise, represents the effects of Z on Y while controlling for the effects of X. These processes, in fact, happen simultaneously.

It's been a good number of chapters – six of them, to be precise – between the first moment when we discussed the importance of controlling for Z and the point, just above, when we showed you precisely how to do it. The fourth causal hurdle has never been too far from front-and-center since Chapter 3, and now you know the process of crossing it for observational data.

Don't get too optimistic too quickly, though. First, as we noted, the three-variable setup we just mentioned can easily be generalized to more than three variables. But the formula for estimating β_1 controls only for

[4] Later in this chapter, you will see that there are two situations in which the two-variable and multiple regression parameter estimates of β will be the same.

the effects of the Z variable that are included in the regression equation. It does not control for other variables that are not measured and not included in the model. And what happens when we fail to include a relevant cause of Y in our regression model? Bad things. (Those bad things will come into focus a bit later in this chapter.)

Second, as we foreshadowed at the beginning of this chapter, the type of control that we have just introduced for observational studies, what we call "statistical control," is not as strong as the experimental control that we described in Chapter 4. We hope that you can see that the type of control that is present in multiple regression is more akin to an accounting device based on the amounts of shared variance between X, Y, and Z. "Controlling for" Z in the regression sense involves identifying the variation that is shared between Z and the other two variables, and discounting it, and then looking at the relationship that remains between X and Y after the shared variation with Z is removed. This does represent a type of control, to be sure, but it is not as powerful as randomly assigning values of X to participants in an experiment. As we described back in Chapter 4, the reason that experimental control is so powerful is that we know exactly the process that generates values of X. (That process is simple randomness, and nothing more.) With statistical control in observational studies, by contrast, we do not know anything specific about the data-generating process of X. In such studies without experimental control, participants might choose or somehow acquire their own values of X, or there might be a complex causal process that sorts cases into different values of X. And it is possible that that very causal process is somehow polluted by some Z that we have failed to control for, or by Y (and, if this is true, it has even more severely negative consequences). All of this, though, is a normal part of the scientific process. It is always possible that there is another, uncontrolled-for Z out there. But, as a part of this process, it is best to put the results out in the open and see how well they stand the test of time.

10.4 INTERPRETING MULTIPLE REGRESSION

For an illustration of how to interpret multiple regression coefficients, let's return to our example from Chapter 9, in which we showed you the results of a regression of US presidential election results on the previous year's growth rate in the US economy (see Figure 9.5). The model we estimated, you will recall, was $\text{inc\_vote} = \alpha + \beta \times \text{g}$, where inc_vote is "vote" and g is "growth," and the estimated coefficients there were $\hat{\alpha} = 51.45$ and $\hat{\beta} = 0.62$. For the purposes of this example, we need to drop the observation from the presidential election of 1876. Doing this

Table 10.1 Three regression models of US presidential elections

	A	B	C
Growth	0.65*	—	0.58*
	(0.15)	—	(0.15)
Good News	—	0.87*	0.63*
	—	(0.32)	(0.28)
Intercept	51.61*	47.63*	48.47*
	(0.81)	(1.87)	(1.58)
R-squared	0.35	0.18	0.44
Number of cases	35	35	35

Notes: The dependent variable is the percentage of the two major parties' vote for the incumbent party's candidate.
Standard errors are in parentheses.
$*p < 0.05$, two-tailed t-test.

changes our estimates slightly so that $\hat{\alpha} = 51.61$ and $\hat{\beta} = 0.65$.[5] Those results are in column A of Table 10.1.

In column A, you see the parameter estimate (0.65) for the annual growth rate in the US economy (in the row labeled "Growth"), and the standard error of that estimated slope, 0.15. In the row labeled "Intercept," you see the estimated y-intercept for that regression, 51.61, and its associated standard error, 0.81. Both parameter estimates are statistically significant, as indicated by the asterisk and the note at the bottom of the table.

Recall that the interpretation of the slope coefficient in a two-variable regression indicates that, for every one-unit increase in the independent variable, we expect to see β units of change in the dependent variable. In the current context, $\hat{\beta} = 0.65$ means that, for every extra one percentage point in growth rate in the US economy, we expect to see, on average, an extra 0.65 percent in the vote percentage for the incumbent party in presidential elections.

But recall our admonition, throughout this book, about being too quick to interpret any bivariate analysis as evidence of a causal relationship. We have not shown, in column A of Table 10.1, that higher growth rates in the economy *cause* incumbent-party vote totals to be higher. To be sure, the evidence in column A is consistent with a causal connection,

[5] We had to drop 1876 because Fair's data do not include a measure for the new variable that we are adding in this example, "Good News," for that year. When making comparisons across different models of the same data, it is extremely important to have exactly the same cases.

but it does not *prove* it. Why not? Because we have not controlled for other possible causes of election outcomes. Surely there are other causes, in addition to how the economy has (or has not) grown in the last year, of how well the incumbent party will fare in a presidential election. Indeed, we can even imagine other *economic* causes that might bolster our statistical explanation of presidential elections.[6]

Consider the fact that the growth variable accounts for economic growth over the past year. But perhaps the public rewards or punishes the incumbent party for *sustained* economic growth over the long run. In particular, it does not necessarily make sense for the public to reelect a party that has presided over three years of subpar growth in the economy but a fourth year with solid growth. And yet, with our single measure of growth, we are assuming – rather unrealistically – that the public would pay attention to the growth rate only in the past year. Surely the public does pay attention to recent growth, but the public might also pay heed to growth over the long run.

In column B of Table 10.1, we estimate another two-variable regression model, this time using the number of consecutive quarters of strong economic growth leading up to the presidential election – the variable is labeled "Good News" – as our independent variable.[7] (Incumbent-party vote share remains our dependent variable.) In the row labeled "Good News," we see that the parameter estimate is 0.87, which means that, on average, for every additional consecutive quarter of good economic news, we expect to see a 0.87 percent increase in incumbent-party vote share. The coefficient is statistically significant at the usual standard of 0.05.

Our separate two-variable regressions each show a relationship between the independent variable in the particular model and incumbent-party vote shares. But none of the parameter estimates in columns A or B was estimated while controlling for the other independent variable. We rectify that situation in column C, in which we estimate the effects of both the "Growth" and "Good News" variables on vote shares simultaneously.

Compare column C with columns A and B. In the row labeled "Good News," we see that the estimated parameter of $\hat{\beta} = 0.63$ indicates that, for every extra quarter of a year with strong growth rates, the incumbent party should expect to see an additional 0.63 percent of the national vote share, *while controlling for the effects of Growth*. Note the additional clause in the interpretation as well as the emphasis that we place on it. Multiple

[6] And, of course, we can imagine variables relating to success or failure in foreign policy, for example, as other, noneconomic causes of election outcomes.

[7] Fair's operationalization of this variable is "the number of quarters in the first 15 quarters of the administration in which the growth rate of real per capita GDP is greater than 3.2 percent."

regression coefficients always represent the effects of a one-point increase in that particular independent variable on the dependent variable, *while controlling for the effects of all other independent variables in the model*. The higher the number of quarters of continuous strong growth in the economy, the higher the incumbent-party vote share should be in the next election, controlling for the previous year's growth rate.

But, critical to this chapter's focus on multiple regression, notice in column C how including the "Good News" variable changes the estimated effect of the "Growth" variable from an estimated 0.65 in column A to 0.58 in column C. The effect in column C is different because it *controls for the effects of Good News*. That is, when the effects of long-running economic expansions are controlled for, the effects of short-term growth falls a bit. The effect is still quite strong and is still statistically significant, but it is more modest once the effects of long-term growth are taken into account.[8] Note also that the R^2 statistic rises from 0.35 in column A to 0.44 in column C, which means that adding the "Good News" variable increased the proportion of the variance of our dependent variable that we have explained by 9 percent.[9]

In this particular example, the whole emphasis on controlling for other causes might seem like much ado about nothing. After all, comparing the three columns in Table 10.1 did not change our interpretation of whether short-term growth rates affect incumbent-party fortunes at the polls. But we didn't know this until we tested for the effects of long-term growth. And later in this chapter, we will see an example in which controlling for new causes of the dependent variable substantially changes our interpretations about causal relationships. We should be clear about one other

[8] And we can likewise compare the bivariate effect of Good News on vote shares in column B with the multivariate results in column C, noting that the effect of Good News, in the multivariate context, appears to have fallen by approximately one-fourth.

[9] It is important to be cautious when reporting contributions to R^2 statistics by individual independent variables, and this table provides a good example of why this is the case. If we were to estimate Model A first and C second, we might be tempted to conclude that Growth explains 35 percent of Vote and Good News explains 9 percent. But if we estimated Model B first and then C, we might be tempted to conclude that Growth explains 26 percent of Vote and Good News explains 18 percent. Actually, both of these sets of conclusions are faulty. The R^2 is always a measure of the overall fit of the model to the dependent variable. So, all that we can say about the R^2 for Model C is that Growth, Good News, and the intercept term together explain 44 percent of the variation in Vote. So, although we can talk about how the addition or subtraction of a particular variable to a model increases or decreases the model's R^2, we should not be tempted to attribute particular values of R^2 to specific independent variables. If we examine Figure 10.1 (in Section 10.7), we can get some intuition on why this is the case. The R^2 statistic for the model represented in this figure is $(f + d + b)/(a + f + d + b)$. It is the presence of area d that confounds our ability to make definitive statements about the contribution of individual variables to R^2.

thing regarding Table 10.1: Despite controlling for another variable, we still have a long way to go before we can say that we've controlled for all other possible causes of the dependent variable. As a result, we should be cautious about interpreting those results as proof of causality. However, as we continue to add possibly confounding independent variables to our regression model, we inch closer and closer to saying that we've controlled for every other possible cause that comes to mind. Recall that, all the way back in Chapter 1, we noted that one of the "rules of the road" of the scientific enterprise is to always be willing to consider new evidence. New evidence – in the form of controlling for other independent variables – can change our inferences about whether any particular independent variable is causally related to the dependent variable.

10.5 WHICH EFFECT IS "BIGGEST"?

In the preceding analysis, we might be tempted to look at the coefficients in column C of Table 10.1 for Growth (0.58) and for Good News (0.63) and conclude that the effect for Good News is larger than the effect for Growth. As tempting as such a conclusion might be, it must be avoided, for one critical reason: The two independent variables are measured in different metrics, which makes that comparison misleading. Short-run growth rates are measured in a different metric – ranging from negative numbers for times during which the economy shrunk, all the way through stronger periods during which growth exceeded 5 percent per year – than are the number of quarters of consecutive strong growth – which ranges from 0 in the data set through 10. That makes comparing the coefficients misleading.

Because the coefficients in Table 10.1 each exist in the native metric of each variable, they are referred to as **unstandardized coefficients**. Although they are normally not easy to compare to one another, there is a rather simple method to remove the metric of each variable to make them comparable with one another. As you might imagine, such coefficients, because they are on a standardized metric, are referred to as **standardized coefficients**. We compute them, quite simply, by taking the unstandardized coefficients and taking out the metrics – in the forms of the standard deviations – of both the independent and dependent variables:

$$\hat{\beta}_{\text{Std}} = \hat{\beta}\frac{s_X}{s_Y},$$

where $\hat{\beta}_{\text{Std}}$ is the standardized regression coefficient, $\hat{\beta}$ is the unstandardized coefficient (as in Table 10.1), and s_X and s_Y are the standard deviations of X and Y, respectively. The interpretation of the standardized coefficients changes, not surprisingly. Whereas the unstandardized coefficients represent the expected change in Y given a one-unit increase in X, the standardized coefficients represent the expected *standard deviation* change in Y given a *one-standard-deviation* increase in X. Now, because all parameter estimates are in the same units – that is, in expected standard deviation changes of the dependent variable – they become more readily comparable.

Implementing this formula for the unstandardized coefficients in column C of Table 10.1 produces the following results. First, for Growth, where standard deviations are calculated using the last equation in Subsection 6.4.2, we have

$$\hat{\beta}_{\text{Std}} = 0.58\left(\frac{5.50}{6.02}\right) = 0.53.$$

Next, for Good News,

$$\hat{\beta}_{\text{Std}} = 0.63\left(\frac{2.95}{6.02}\right) = 0.31.$$

These coefficients would be interpreted as follows: For a one-standard-deviation increase in Growth, on average, we expect a 0.53-standard-deviation increase in the incumbent-party vote share, controlling for the effect of Good News. And for a one-standard-deviation increase in Good News, we expect to see, on average, a 0.31-standard-deviation increase in the incumbent-party vote shares, controlling for the effect of Growth. Note how, when looking at the unstandardized coefficients, we might have mistakenly thought that the effect of Good News was larger than the effect of Growth. But the standardized coefficients (correctly) tell the opposite story: The estimated effect of Growth is 170 percent of the size of the effect of Good News.[10]

YOUR TURN: Interpreting standardized coefficients

What would be the substantive interpretation for the effect of Good News if $\hat{\beta}_{\text{Std}} = -0.31$?

[10] Some objections have been raised about the use of standardized coefficients (King, 1986). From a technical perspective, because standard deviations can differ across samples, this makes the results of standardized coefficients particularly sample specific. Additionally, and from a broader perspective, one-unit or one-standard-deviation shifts in different independent variables have different substantive meanings regardless of the metrics in which the variables are measured. We might therefore logically conclude that there isn't much use in trying to figure out which effect is biggest.

10.6 STATISTICAL AND SUBSTANTIVE SIGNIFICANCE

Related to the admonition about which effect is "biggest," consider the following, seemingly simpler, question: Are the effects found in column C of Table 10.1 "big"? A tempting answer to that question is "Well of course they're big. Both coefficients are statistically significant. Therefore, they're big."

That logic, although perhaps appealing, is faulty. Recall the discussion from Chapter 7 on the effects of sample size on the magnitude of the standard error of the mean. And we noted in Chapter 9 that the same effects of sample size are present on the magnitude of the standard error of our regression coefficients. What this means is that, even if the strength of the relationship (as measured by our coefficient estimates) remains constant, by merely increasing our sample size we can affect the statistical significance of those coefficients. Why? Because statistical significance is determined by a *t*-test in which the standard error is in the denominator of that quotient. What you can remember is that larger sample sizes will shrink standard errors and therefore make finding statistically significant relationships more likely.[11] It is also apparent from Appendix B that, when the number of degrees of freedom is greater, it is easier to achieve statistical significance.

We hope that you can see that arbitrarily increasing the size of a sample, and therefore finding statistically significant relationships, does not in any way make an effect "bigger" or even "big." Recall, such changes to the standard errors have no bearing on the rise-over-run nature of the slope coefficients themselves.

How, then, should you judge whether an effect of one variable on another is "big?" One way is to use the method just described – using standardized coefficients. By placing the variances of X and Y on the same metric, it is possible to come to a judgment about how big an effect is. This is particularly helpful when the independent variables X and Z, or the dependent variable Y, or both, are measured in metrics that are unfamiliar or artificial.

When the metrics of the variables in a regression analysis are intuitive and well known, however, rendering a judgment about whether an effect is large or small becomes something of a matter of interpretation. For example, in Chapter 11, we will see an example relating the effects of changes in the unemployment rate (X) on a president's approval rating (Y). It is very simple to interpret that a slope coefficient of, say, -1.51, means

[11] To be certain, it's not always possible to increase sample sizes, and, even when possible, it is nearly always costly to do so. The research situations in which increasing sample size is most likely, albeit still expensive, is in mass-based survey research.

that, for every additional point of unemployment, we expect approval to go down by 1.51 points, controlling for other factors in the model. Is that effect large, small, or moderate? There is something of a judgment call to be made here, but at least, in this case, the metrics of both *X* and *Y* are quite familiar; most people with even an elementary familiarity with politics will need no explanation as to what unemployment rates mean or what approval polls mean. Independent of the statistical significance of that estimate – which, you should note, we have not mentioned here – discussions of this sort represent attempts to judge the **substantive significance** of a coefficient estimate. Substantive significance is more difficult to judge than statistical significance because there are no numeric formulae for making such judgments. Instead, substantive significance is a judgment call about whether or not statistically significant relationships are "large" or "small" in terms of their real-world impact.

From time to time we will see a "large" parameter estimate that is not statistically significant. Although it is tempting to describe such a result as substantively significant, it is not. We can understand this by thinking about what it means for a particular result to be statistically significant. As we discussed in Chapter 9, in most cases we are testing the null hypothesis that the population parameter is equal to zero. In such cases, even when we have a large parameter estimate, if it is statistically insignificant this means that it is not statistically distinguishable from zero. Therefore a parameter estimate can be substantively significant only when it is also statistically significant.

10.7 WHAT HAPPENS WHEN WE FAIL TO CONTROL FOR *Z*?

Controlling for the effects of other possible causes of our dependent variable *Y*, we have maintained, is critical to making the correct causal inferences. Some of you might be wondering something like the following: "How does omitting *Z* from a regression model affect my inference of whether *X* causes *Y*? *Z* isn't *X*, and *Z* isn't *Y*, so why should omitting *Z* matter?"

Consider the following three-variable regression model involving our now-familiar trio of *X*, *Y*, and *Z*:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + u_i.$$

And assume, for the moment, that this is the *correct* model of reality. That is, the only systematic causes of *Y* are *X* and *Z*; and, to some degree, *Y* is also influenced by some random error component, *u*.

Now let's assume that, instead of estimating this correct model, we fail to estimate the effects of *Z*. That is, we estimate

$$Y_i = \alpha + \beta_1^* X_i + u_i^*.$$

As we previously hinted, the value of β_1 in the correct, three-variable equation and the value of β_1^* will not be identical under most circumstances. (We'll see the exceptions in a moment.) And that, right there, should be enough to raise red flags. For, if we know that the three-variable model is the *correct* model – and what that means, of course, is that the estimated value of β_1 that we obtain from the data will be equal to the true population value – and if we know that β_1 will not be equal to β_1^*, then there is a problem with the estimated value of β_1^*. That problem is a statistical problem called **bias**, which means that the expected value of the parameter estimate that we obtain from a sample will not be equal to the true population parameter. The specific type of bias that results from the failure to include a variable that belongs in our regression model is called **omitted-variables bias**.

Let's get specific about the nature of omitted-variables bias. If, instead of estimating the true three-variable model, we estimate the incorrect two-variable model, the formula for the slope β_1^* will be

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Notice that this is simply the two-variable formula for the effect of X on Y. (Of course, the model we just estimated is a two-variable model, in spite of the fact that we know that Z, as well as X, affects Y.) But because we know that Z *should* be in the model, and we know from Chapter 9 that regression lines travel through the mean values of each variable, we can figure out that the following is true:

$$(Y_i - \bar{Y}) = \beta_1(X_i - \bar{X}) + \beta_2(Z_i - \bar{Z}) + (u_i - \bar{u}).$$

We can do this because we know that the plane will travel through each variable's mean.

Now notice that the left-hand side of the preceding equation, the $(Y_i - \bar{Y})$, is identical to one portion of the numerator of the slope for $\hat{\beta}_1^*$. Therefore we can substitute the right-hand side of the preceding equation – yes, that entire mess – into the numerator of the formula for $\hat{\beta}_1^*$.

The resulting math isn't anything that is beyond your skills in algebra, but it is cumbersome, so we won't derive it here. After a few lines of multiplying and reducing, though, the formula for $\hat{\beta}_1^*$ will reduce to

$$E(\hat{\beta}_1^*) = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

This might seem like a mouthful – a fact that's rather hard to deny – but there is a very important message in it. What the equation says is that the estimated effect of X on Y, $\hat{\beta}_1^*$, in which we do not include the effects of Z on Y (but should have), will be equal to the true β_1 – that is, the effect with Z taken into account – plus a bundle of other stuff. That other stuff, strictly speaking, is bias. And because this bias came about as a result of omitting a variable (Z) that should have been in the model, this type of bias is known as omitted-variables bias.

Obviously, we'd like the expected value of our $\hat{\beta}_1^*$ (estimated without Z) to equal the true β_1 (as if we had estimated the equation with Z). And if the product on the right-hand side of the "+" sign in the preceding equation equals zero, it will. When will that happen?[12] In two circumstances, neither of which is particularly likely. First, $\hat{\beta}_1^* = \beta_1$ if $\beta_2 = 0$. Second, $\hat{\beta}_1^* = \beta_1$ if the large quotient at the end of the equation, the

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

is equal to zero. What is that quotient? It should look familiar; in fact, it is the bivariate slope parameter of a regression of Z on X.

In the first of these two special circumstances, the bias term will equal zero if and only if the effect of Z on Y – that is, the parameter β_2 – is zero. Okay, so it's safe to omit an independent variable from a regression equation if it has no effect on the dependent variable. (If that seems obvious to you, good.) The second circumstance is a bit more interesting: It's safe to omit an independent variable Z from an equation if it is entirely unrelated to the other independent variable X. Of course, if we omit Z in such circumstances, we'll still be deprived of understanding how Z affects Y; but at least, so long as Z and X are absolutely unrelated, omitting Z will not adversely affect our estimate of the effect of X on Y.[13]

We emphasize that this second condition is unlikely to occur in practice. Therefore, if Z affects Y, and Z and X are related, then if we omit Z from our model, our bias term will not equal zero. In the end, omitting Z will cause us to misestimate the effect of X on Y.

This result has many practical implications. Foremost among them is the fact that, even if you aren't interested theoretically in the connection between Z and Y, you need to control for it, statistically, in order to get an unbiased estimate of the impact of X, which is the focus of the theoretical investigation.

[12] To be very clear, for a mathematical product to equal zero, either one or both of the components must be zero.

[13] Omitting Z from our regression model also drives down the R^2 statistic.

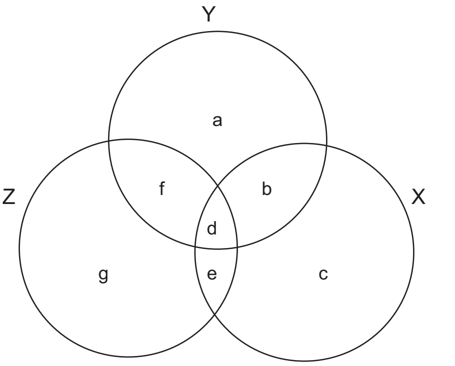


Figure 10.1 Venn diagram in which X, Y, and Z are correlated

That might seem unfair, but it's true. If we estimate a regression model that omits an independent variable (Z) that belongs in the model, then the effects of that Z will somehow work their way into the parameter estimates for the independent variable that we do estimate (X) and pollute our estimate of the effect of X on Y.

The preceding equation also suggests when the magnitude of the bias is likely to be large and when it is likely to be small. If either or both of the components of the bias term

$$\beta_2 \quad \text{and} \quad \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

are *close to* zero, then the bias is likely to be small (because the bias term is the product of both components); but if both are likely to be large, then the bias is likely to be quite large.

Moreover, the equation also suggests the likely *direction* of the bias. All we have said thus far is that the coefficient $\hat{\beta}_1^*$ will be biased – that is, it will not equal its true value. But will it be too large or too small? If we have good guesses about the values of β_2 and the correlation between X and Z – that is, whether or not they are positive or negative – then we can suspect the direction of the bias. For example, suppose that β_1, β_2, and the correlation between X and Z are all positive. That means that our estimated coefficient $\hat{\beta}_1^*$ will be larger than it is supposed to be, because a positive number plus the product of two positive numbers will be a still-larger positive number. And so on.[14]

To better understand the importance of controlling for other possible causes of the dependent variable and the importance of the relationship (or the lack of one) between X and Z, consider the following graphical illustrations. In Figure 10.1, we represent the total variation of Y, X, and Z each with a circle.[15] The covariation between any of these two variables – or among all three – is represented by the places where the circles overlap.

[14] With more than two independent variables, it becomes more complex to figure out the direction of the bias.

[15] Recall from Chapter 9 how we introduced Venn diagrams to represent variation (the circles) and covariation (the overlapping portion of the circles).

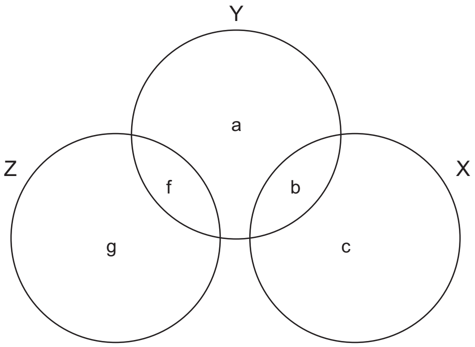


Figure 10.2 Venn diagram in which X and Z are correlated with Y, but not with each other

Thus, in the figure, the total variation in Y is represented as the sum of the area $a+b+d+f$. The covariation between Y and X is represented by the area $b+d$.

Note in Figure 10.1, though, that the variable Z is related to both Y and X (because the circle for Z overlaps with both Y and X). In particular, the relationship between Y and Z is accounted for by the area $f+d$, and the relationship between Z and X is accounted for by the area $d+e$. As we have already seen, d is also a portion of the relationship between Y and X. If, hypothetically, we erased the circle for Z from the figure, we would (incorrectly) attribute all of the area $b+d$ to X, when in fact the d portion of the variation in Y is shared by *both* X and Z. This is why, when Z is related to both X and Y, if we fail to control for Z, we will end up with a biased estimate of X's effect on Y.

Consider the alternative scenario, in which both X and Z affect Y, but X and Z are completely unrelated to one another. That scenario is portrayed graphically in Figure 10.2. There, the circles for both X and Z overlap with the circle for Y, but they do not overlap at all with one another. In that case – which, we have noted, is unlikely in applied research – we can safely omit consideration of Z when considering the effects of X on Y. In that figure, the relationship between X and Y, the area b, is unaffected by the presence (or absence) of Z in the model.[16]

10.7.1 An Additional Minimal Mathematical Requirement in Multiple Regression

We outlined a set of assumptions and minimal mathematical requirements for the two-variable regression model in Chapter 9. In multiple regression, all of these assumptions are made and all of the same minimal mathematical requirements remain in place. In addition to those, however, we need to add one more minimal mathematical requirement to be able to estimate our multiple regression models: It must be the case that *there is no exact linear relationship* between any two or more of our independent variables (which we have called X and Z). This is also called the assumption of

[16] For identical reasons, we could safely estimate the effect of Z on Y, the area f, without considering the effect of X.

no **perfect multicollinearity** (by which we mean that X and Z cannot be *perfectly* collinear, with a correlation coefficient of $r = 1.0$).

What does it mean to say that X and Z cannot exist in an exact linear relationship? Refer back to Figure 10.1. If X and Z had an *exact* linear relationship, instead of having some degree of overlap – that is, some imperfect degree of correlation – the circles would be exactly on top of one another. In such cases, it is literally impossible to estimate the regression model, as separating out the effects of X on Y from the effects of Z on Y is impossible.

This is not to say that we must assume that X and Z are entirely uncorrelated with one another (as in Figure 10.2). In fact, in almost all applications, X and Z will have some degree of correlation between them. Things become complicated only as that correlation approaches 1.0; and when it hits 1.0, the regression model will fail to be estimable with both X and Z as independent variables. In Chapter 11 we will discuss these issues further.

10.8 AN EXAMPLE FROM THE LITERATURE: COMPETING THEORIES OF HOW POLITICS AFFECTS INTERNATIONAL TRADE

What are the forces that affect international trade? Economists have long noted that there are economic forces that shape the extent to which two nations trade with one another.[17] The size of each nation's economy, the physical distance between them, and the overall level of development have all been investigated as economic causes of trade.[18] But in addition to economic forces, does politics help to shape international trade?

Morrow, Siverson, and Tabares (1998) investigate three competing (and perhaps complementary) political explanations for the extent to which two nations engage in international trade. The first theory is that states with friendly relations are more likely to trade with one another than are states engaged in conflict. Conflict, in this sense, need not be militarized disputes (though it may be).[19] Conflict, they argue, can dampen trade in several ways. First, interstate conflict can sometimes produce embargoes

[17] Theories of trade and, indeed, many theories about other aspects of international trade are usually developed with pairs of nations in mind. Thus all of the relevant variables, like trade, are measured in terms of pairs of nations, which are often referred to as "dyads" by international relations scholars. The resulting **dyadic data** sets are often quite large because they encompass each relevant pair of nations.

[18] Such models are charmingly referred to as "gravity models," because, according to these theories, the forces driving trade resemble the forces that determine gravitational attraction between two physical objects.

[19] See Pollins (1989) for an extended discussion of this theory.

(or prohibitions on trade). Second, conflict can reduce trade by raising the risks for firms that wish to engage in cross-border trading.

The second theory is that trade will be higher when both nations are democracies and lower when one (or both) is an autocracy.[20] Because democracies have more open political and judicial systems, trade should be higher between democracies because firms in one country will have greater assurance that any trade disputes will be resolved openly and fairly in courts to which they have access. In contrast, firms in a democratic state may be more reluctant to trade with nondemocratic countries, because it is less certain how any disagreements will be resolved. In addition, firms may be wary of trading with nondemocracies for fear of having their assets seized by the foreign government. In short, trading with an autocratic government should raise the perceived risks of international trade.

The third theory is that states that are in an alliance with one another are more likely to trade with one another than are states that are not in such an alliance.[21] For states that are not in an alliance, one nation may be reluctant to trade with another nation if the first thinks that the gains from trade may be used to arm itself for future conflict. In contrast, states in an alliance stand to gain from each other's increased wealth as a result of trade.

To test these theories, Morrow, Siverson, and Tabares (1998) look at trade among all of the major powers in the international system – the United States, Britain, France, Germany, Russia, and Italy – during most of the twentieth century. They consider each pair of states – called *dyads* – separately and examine exports to each country on an annual basis.[22] Their dependent variable is the amount of exports in every dyadic relationship in each year.

Table 10.2 shows excerpts from the analysis of Morrow, Siverson, and Tabares.[23] In column A, they show that, as the first theory predicts, increases in interstate peace are associated with higher amounts of trade between countries, controlling for economic factors. In addition, the larger the economy in general, the more trade there is. (This finding is consistent across all estimation equations.) The results in column B indicate that pairs of democracies trade at higher rates than do pairs involving at least one nondemocracy. Finally, the results in column C show that trade is higher

[20] See Dixon and Moon (1993) for an elaboration of this theory.

[21] See Gowa (1989) and Gowa and Mansfield (1993) for an extended discussion, including distinctions between bipolar and multipolar organizations of the international system.

[22] This research design is often referred to as a time-series cross-section design, because it contains both variation between units and variation across time. In this sense, it is a hybrid of the two types of quasi-experiments discussed in Chapter 3.

[23] Interpreting the precise magnitudes of the parameter estimates is a bit tricky in this case, because the independent variables were all transformed by use of natural logarithms.

Table 10.2 Excerpts from Morrow, Siverson, and Tabares's table on the political causes of international trade

	A	B	C	D
Peaceful relations	1.12*	—	—	1.45*
	(0.22)	—	—	(0.37)
Democratic partners	—	1.18*	—	1.22*
	—	(0.12)	—	(0.13)
Alliance partners	—	—	0.29*	−0.50*
	—	—	(0.03)	(0.16)
GNP of exporter	0.67*	0.57*	0.68*	0.56*
	(0.07)	(0.07)	(0.07)	(0.08)
R^2	0.77	0.78	0.77	0.78
N	2631	2631	2631	2631

Notes: Other variables were estimated as a part of the regression model but were excluded from this table for ease of presentation.
Standard errors are in parentheses.
$^*p < 0.05$.

between alliance partners than between states that are not in an alliance with one another. All of these effects are statistically significant.

So far, each of the theories received at least some support. But, as you can tell from looking at the table, the results in columns A through C do not control for the other explanations. That is, we have yet to see results of a full multiple regression model, in which the theories can compete for explanatory power. That situation is rectified in column D, in which all three political variables are entered in the same regression model. There, we see that the effects of reduced hostility between states is actually enhanced in the multiple regression context – compare the coefficient of 1.12 with the multiple regression 1.45. Similarly, the effects of democratic trading partners remains almost unchanged in the fully multivariate framework. However, the effect of alliances changes. Before controlling for conflict and democracy, the effect of alliances was (as expected) positive and statistically significant. However, in column D, in which we control for conflict and democracy, the effect flips signs and is now *negative* (and statistically significant), which means that, when we control for these factors, states in an alliance are less (not more) likely to trade with one another.

The article by Morrow, Siverson, and Tabares (1998) represents a case in which synthesizing several competing explanations for the same phenomenon – international trade – produces surprising findings. By using a data set that allowed them to test all three theories simultaneously, Morrow, Siverson, and Tabares were able to sort out which theories received support and which did not.

10.9 MAKING EFFECTIVE USE OF TABLES AND FIGURES

At this point in your class, it's likely that you've spent time in a computer lab learning how to conduct your own analyses. We understand – because we experienced it ourselves when we were your age – that it can feel like a pretty big leap to go from understanding how a statistical formula works algebraically from a book or a class presentation, to understanding how to critique how these methods are applied in work like that by Morrow, Siverson, and Tabares (1998) that we just described in the previous section, to understanding how things work when you're looking at statistical software output on your own computer.

We realize, too, that many of you have interests in conducting your own analyses to investigate problems that you find to be interesting. Good! Perhaps you have an independent study or an honors thesis to work on, or some other project that you want to include as a writing sample for applications to graduate school. And you want to learn to communicate your ideas and findings clearly for your intended audience. That's what this section is about.

We strongly recommend that you spend a lot of time constructing the tables and figures that you include in your projects. When readers first encounter your written work, many of them will take a quick look at the title and introduction and then go directly to your tables and figures. This is certainly a reasonable thing to do when someone is trying to evaluate whether or not they should invest further time reviewing your work. Thus, although they may appear at the back of your project, tables and figures often determine the first impression that potential readers have of your project. As such, we recommend that you construct your tables and figures so they stand on their own and draw readers in. With these two considerations in mind, we have a set of recommendations for what you should and should not do as you put your tables and figures together. We also recommend that you tell readers in the text of your project what they should see in your tables and figures. Some of this can be learned by reading other scholars' work on similar subjects: Take time, when you read, to think about what works and what doesn't work in terms of other scholars' use of tables and figures.

10.9.1 Constructing Regression Tables

As we have made clear, multiple regression analyses are the main tool that researchers in political science use to test their causal claims in observational research. Consumers of political science research are well-trained to read regression tables and make assessments based on what they see in them. In addition to making assessments about the specific results presented in a table, readers will also use what they see – and don't see – in regression tables to make assessments about the technical competence of the person who has constructed the table. Since this will have a major impact on the overall assessment of your project, you will want to be careful and thorough in your construction of regression tables.

The construction of regression tables involves moving back and forth between results in a statistics program and the table-making facilities in whatever word-processing program you are using. The easiest and *worst* way to do this is to simply copy and paste your statistical output into your word-processing program. This is a bad way to proceed for at least six reasons. First of all, it just doesn't look good, and (if you do this) makes you look transparently lazy. Second, statistical programs tend to give you an overabundance of information when you estimate a regression model. This information is often way more than what you will need to report in your regression table. Third, the default reporting of results that the statistical program reports may be different from what is appropriate for your purposes. For instance, as we discussed in Chapter 9, almost all statistical programs report the results from two-tailed hypothesis tests when most of our hypotheses in political science are directional (and thus should be assessed with one-tailed tests). Fourth, statistical programs report the names of your variables as they appear in your data sets. While the abbreviations that you have chosen for your variables probably make sense to you, they will almost surely be confusing to your readers. Fifth, computer programs usually report statistics with a number of digits past the decimal point that go way beyond what you need to report. We recommend rounding to two decimal places. And sixth, computer programs report model results with variables in a particular order, but that order may not be the best for emphasizing the important aspects of your results.

Having established what you *should not* do in constructing your tables, let's now talk about what you *should* do. Remember that your goals are to make your table of results stand on its own and draw potential readers in. As such, you want your tables to transmit to other researchers what you have done. Your regression table should include:

- a title that communicates the purpose of the model and/or the most important implications,
- names for the independent variables that are as clear as possible,
- a listing of your independent variables in an order that suits your purposes (usually with your main theoretical variable(s) at the top and control variables listed below),

- the estimated effect of each independent variable (usually the estimated parameter),
- some indication of the uncertainty/statistical significance of each estimated effect (standard errors or t-statistics in parentheses underneath a parameter estimate),
- some indication of which results have been found to be statistically significant according to a particular standard (e.g., putting stars next to results for which $p < 0.05$),
- some indication of what is the dependent variable,
- some overall diagnostics to communicate the model's fit and the number of cases on which the model was estimated,
- a set of notes to help readers decode anything they need to decode (e.g., that "**" means "$p < 0.01$)," and
- any other information that needs to be communicated in order to convey the importance of the findings.

As an example of a table of regression results, consider Table 10.3.[24] If we go through the list of what a table should contain, we can evaluate how well this table does with each item. The title is fairly informative about what is going on in the model depicted in the table, but certainly conveys the most important implications. The names of the independent variables could certainly be more clear. For instance, we don't know exactly what "Growth" or "Unemployment" represent, though we could probably make a good guess. We also don't know from the table alone what "Government Change" is, and it would be hard to make a good guess. The table clearly contains parameter estimates and an indication (in the form of standard errors) of the uncertainty about them. In addition, we can tell from the note beneath the table that the stars in the table convey different levels of statistical significance. The notes beneath the table also make it fairly clear what the dependent variable is, though we would have to figure out on our own that these data are from monthly surveys. So, overall, while this table is fairly clear, it could certainly be improved upon.

As we have seen in this chapter, it is often the case that we will want to report the results from several regression models in the same table. When we do this, it is important to make sure that we are setting up our comparisons across models in a fashion that conveys exactly what we want. There are two types of comparisons that we typically make when we are presenting multiple regression models in the same table: comparisons of different model specifications with the same sample of data or comparisons of the same model specification across different samples of data. In tables

[24] Tables 10.3 and 10.5 are based on tables contained in Palmer, Whitten, and Williams (2013).

Table 10.3 Economic models of monthly UK government support, 2004–2011 objective economic measures only

Independent variable	Parameter estimate (standard error)
Growth	0.25** (0.11)
Unemployment	0.07 (0.20)
Δ Inflation	−2.72*** (0.75)
Government Change	12.46*** (2.27)
Support_{t-1}	0.78*** (0.06)
Constant	6.37*** (2.13)
R^2	0.81
N	89

Notes: The dependent variable is the percentage of each sample that reported that they would vote for the government if an election was held at the time of the survey.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$ (two-tailed tests, despite directional hypotheses).

that show the results from multiple models, it is important to only make one of these two types of changes at a time.

Consider, for instance, Tables 10.1 and 10.2. In these tables we presented *different* model specifications across the *same* sample. What we can see very well as we move across the columns in these tables is the changes in the estimated effects of our variables as we change our model. But, it is important to note that, if the sample in Table 10.1 or 10.2 was not *exactly* the same across the columns, we would not know why the estimated effects were changing. In such a case, changes could be due to a change in the sample or a change in the model.

As an example of the second type of comparison, where we look at the same model specification but across different samples, consider Tables 10.4 and 10.5. Both of these tables are examples of the type of research strategy discussed in Chapter 2 where we are interested in differences across subpopulations of cases in terms of the relationship between X and Y. The key variable of interest in the first table is how coolly or warmly (on a 0-to-100 scale) survey respondents report feeling about a particular

Table 10.4 Alternative presentation of the effects of gender and feelings toward the women's movement on Hillary Clinton Thermometer scores

	Sample		
Independent variable	**All**	**Male**	**Female**
Women's Movement Thermometer	0.70***	0.75***	0.62***
	(0.03)	(0.05)	(0.04)
Intercept	8.52	1.56	16.77***
	(2.10)	(3.03)	(2.89)
n	1466	656	810
R^2	0.25	0.27	0.21

Notes: The dependent variable in both models is the respondent's thermometer score for Hillary Clinton.
Standard errors in parentheses.
Two-sided t-tests: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

person or group – in this case feelings about Hillary Clinton. Table 10.4 shows such a comparison looking at the relationship between Women's Movement Thermometer scores across men and women.[25] We can see from this table that, although the sample changes across the columns, the model specification is the same. And we can tell from this comparison that there are differences across the columns in terms of the estimated relationships. The key variable in Table 10.5 is the percentage of a sample in the UK that reported that, were an election held that day, they would vote for the party that currently controls the government. The table shows that when we estimate the model for three different subpopulations defined by their income levels, we also see substantial differences in the ways in which the economic variables, the main Xs in this model, impact support for the government.

10.9.2 Writing about Regression Tables

Although our goal in constructing tables is to make them stand well on their own, when writing about regression tables, it is important to do

[25] As we will show in Chapter 11, we can also get leverage on this type of difference in the relationship between X and Y across subpopulations through the use of an interactive model specification. But here we show this difference in the relationship between X and Y by presenting the bivariate regression model with thermometer scores for Hillary Clinton as the dependent variable and Women's Movement Thermometer scores as the independent variable on the entire sample, and then subsamples of cases defined by the gender of the respondent.

Table 10.5 Economic models of monthly UK government support across groups of voters, 2004–2011 objective economic measures only

	Sample			
Independent variable	All	Upper income	Middle income	Low income
Growth	0.25**	0.61***	0.35**	0.33*
	(0.11)	(0.21)	(0.15)	(0.20)
Unemployment	0.07	1.18**	−0.24	−1.76***
	(0.20)	(0.47)	(0.31)	(0.51)
Δ Inflation	−2.72***	−3.40**	−4.21***	−3.38**
	(0.75)	(1.46)	(1.12)	(1.59)
Government Change	12.46***	19.60***	6.28*	−5.11
	(2.27)	(4.56)	(3.42)	(4.84)
Support_{t-1}	0.78***	0.58***	0.56***	0.28***
	(0.06)	(0.09)	(0.08)	(0.10)
Constant	6.37***	5.30**	15.95***	34.61***
	(2.13)	(2.65)	(3.66)	(5.74)
R^2	0.81	0.66	0.58	0.48
N	89	89	89	89

Notes: The dependent variable is the percentage of each sample that reported that they would vote for the government if an election was held at the time of the survey.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$ (two-tailed tests, despite directional hypotheses).

a little bit of handholding. In other words, *tell* your readers what they should take away from each table. Consider the way in which we just ended the above section. Although this table is competently constructed, we don't know for sure which parts of the table are going to catch the eye of our readers. All that we have told readers is that there are substantial differences across groups. Instead of leaving this up to chance, we should tell them what they should see from this table – for instance, that the largest effect of growth appears to happen among the high income group. We should also point out that the effect of unemployment is in the opposite direction of our theoretical expectations for the highest income group, statistically insignificant for the middle income group, and statistically significant in the expected direction for the lowest income group. We should point out that the effects of inflation are roughly the same across the three groups, all statistically significant in the expected (negative) direction, while for only the high income group is there a statistically significant and positive effect for the switch in government from the Labour Party to the Conservative/Liberal Democratic coalition represented by the variable named "Government Change." Finally, we should point out that these effects that we just discussed are only the short-term effects and that all of these variables have long-term effects as well, because these models include a lagged dependent variable, labeled "Support_{t-1}," in the table.[26]

The bottom line with writing about regression tables is that you want to tell your readers what they should see. This will help you to maximize the impact of what you have found and to keep your audience focused on what you are trying to communicate.

10.10 IMPLICATIONS AND CONCLUSIONS

What are the implications of this chapter? The key take-home point – that failing to control for all relevant independent variables will often lead to mistaken causal inferences for the variables that do make it into our models – applies in several contexts. If you are reading a research article in one of your other classes, and it shows a regression analysis between two variables, but fails to control for the effects of some other possible cause of the dependent variable, then you have some reason to be skeptical about the reported findings. In particular, if you can think of another independent variable that is likely to be related to *both* the independent variable and the dependent variable, then the relationship that the article does show that fails to control for that variable is likely to be plagued with bias. And if that's the case, then there is substantial reason to doubt the findings. The findings *might* be right, but you can't know that from the evidence presented in the article; in particular, you'd need to control for the omitted variable to know for sure.

But this critical issue isn't just encountered in research articles. When you read a news article from your favorite media web site that reports a relationship between some presumed cause and some presumed effect – news articles don't usually talk about "independent variables" or "dependent variables" – but fails to account for some other cause that you can imagine might be related to both the independent and dependent variables, then you have reason to doubt the conclusions.

It might be tempting to react to omitted-variables bias by saying, "Omitted-variables bias is such a potentially serious problem that I don't want to use regression analysis." That would be a mistake. In fact, the logic of omitted-variables bias applies to any type of research, no matter what type of statistical technique is used – in fact, no matter whether the research is qualitative or quantitative.

[26] We will learn more about the way to discuss time-series models in Chapter 12.

Sometimes, as we have seen, controlling for other causes of the dependent variable changes the discovered effects only at the margins. That happens on occasion in applied research. At other times, however, failure to control for a relevant cause of the dependent variable can have serious consequences for our causal inferences about the real world.

In Chapters 11 and 12, we present you with some crucial extensions of the multiple regression model that you are likely to encounter when consuming or conducting research.

CONCEPTS INTRODUCED IN THIS CHAPTER

- bias – a statistical problem that occurs when the expected value of the parameter estimate that we obtain from a sample will not be equal to the true population parameter
- dyadic data – data that reflect the characteristics of pairs of spatial units and/or the relationships between them
- omitted-variables bias – the specific type of bias that results from the failure to include a variable that belongs in our regression model
- perfect multicollinearity – when there is an exact linear relationship between any two or more of a regression model's independent variables
- standardized coefficients – regression coefficients such that the rise-over-run interpretation is expressed in standard-deviation units of each variable
- substantive significance – a judgment call about whether or not statistically significant relationships are "large" or "small" in terms of their real-world impact
- unstandardized coefficients – regression coefficients such that the rise-over-run interpretation is expressed in the original metric of each variable

EXERCISES

1. Identify an article from a prominent web site that reports a causal relationship between two variables. Can you think of another variable that is related to both the independent variable and the dependent variable? Print and turn in a copy of the article with your answers.
2. In Exercise 1, estimate the direction of the bias resulting from omitting the third variable.
3. Fill in the values in the third column of Table 10.6.
4. In your own research you have found evidence from a bivariate regression model that supports your theory that your independent variable X_i is positively related to your dependent variable Y_i (the slope parameter for X_i was statistically significant and positive when you estimated a bivariate regression

Table 10.6 Bias in $\hat{\beta}_1$ when the true population model is $Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + u_i$ but we leave out Z

β_2	$\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$	Resulting bias in $\hat{\beta}_1$
0	+	?
0	−	?
+	0	?
−	0	?
+	+	?
−	−	?
+	−	?
−	+	?

Table 10.7 Three regression models of teacher salaries in the US states and the District of Columbia

	A	B	C
Percentage of state residents with a college degree	704.02*	—	24.56
	(140.22)	—	(231.72)
Per capita income	—	0.68*	0.66*
	—	(0.11)	(0.19)
Intercept	28768.01*	21168.11*	21161.07*
	(3913.27)	(4102.40)	(4144.96)
R^2	0.34	0.47	0.47
N	51	51	51

Notes: The dependent variable is the average salary of public elementary and secondary school teachers.
Standard errors are in parentheses.
$^*p < 0.05$ (two-tailed t-test).

model). You go to a research presentation in which other researchers present a theory that their independent variable Z_i is negatively related to their dependent variable Y_i. They report the results from a bivariate regression model in which the slope parameter for Z_i was statistically significant and negative. Your Y_i and their Y_i are the same variable. What would be your reaction to these findings under each of the following circumstances?

(a) You are confident that the correlation between Z_i and X_i is equal to zero.

(b) You think that the correlation between Z_i and X_i is positive.

(c) You think that the correlation between Z_i and X_i is negative.

5. Using the results depicted in Table 10.7, interpret the results of the bivariate models displayed in columns A and B.
6. Using the results depicted in Table 10.7, interpret the results of the multiple regression model displayed in column C. Compare the results in column C with those in both columns A and B.
7. Draw a Venn diagram that depicts what is going on between the three variables based on the results in Table 10.7.

11 Multiple Regression Model Specification

OVERVIEW

In this chapter we provide introductory *discussions of* and *advice for* commonly encountered research scenarios involving multiple regression models. Issues covered include dummy independent variables, interactive specifications, influential cases, and multicollinearity.

All models are wrong, but some are useful.
—George E.P. Box

11.1 EXTENSIONS OF ORDINARY LEAST-SQUARES

In the previous two chapters we discussed in detail various aspects of the estimation, interpretation, and presentation of OLS regression models. In this chapter we go through a series of research scenarios commonly encountered by political science researchers as they attempt to test their hypotheses within the OLS framework. The purpose of this chapter is twofold – first, to help you to identify when you encounter these issues and, second, to help you to figure out what to do to continue on your way.

We begin with a discussion of "dummy" independent variables and how to properly use them to make inferences. We then discuss how to test interactive hypotheses with dummy variables. We next turn our attention to two frequently encountered problems in OLS – outliers and multicollinearity. With both of these topics, at least half of the battle is identifying that you have the problem.

11.2 BEING SMART WITH DUMMY INDEPENDENT VARIABLES IN OLS

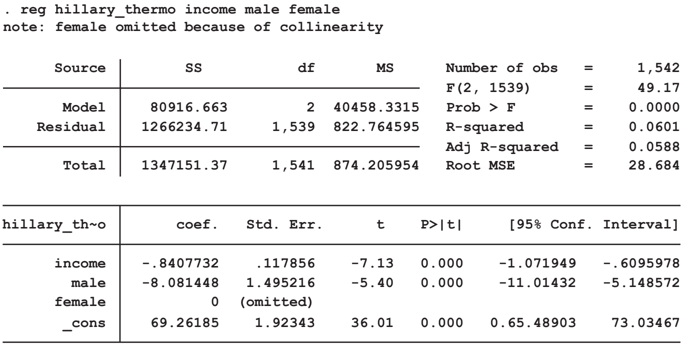
In Chapter 5 we discussed how an important part of knowing your data involves knowing the metric in which each of your variables is measured.

Throughout the examples that we have examined thus far, almost all of the variables, both the independent and dependent variables, have been continuous. This is not by accident. We chose examples with continuous variables because they are, in many cases, easier to interpret than models in which the variables are noncontinuous. In this section, though, we consider a series of scenarios involving independent variables that are *not* continuous. We begin with a relatively simple case in which we have a categorical independent variable that takes on one of two possible values for all cases. Categorical variables like this are commonly referred to as **dummy variables**. Although any two values will do, the most common form of dummy variable is one that takes on values of one or zero. These variables are also sometimes referred to as "indicator variables" when a value of one indicates the presence of a particular characteristic and a value of zero indicates the absence of that characteristic. After considering dummy variables that reflect two possible values, we then consider more complicated examples in which we have an independent variable that is categorical with more than two values. We conclude this section with an examination of how to handle models in which we have multiple dummy variables representing multiple and overlapping classifications of cases.

11.2.1 Using Dummy Variables to Test Hypotheses about a Categorical Independent Variable with Only Two Values

During the 1996 US presidential election between incumbent Democrat Bill Clinton and Republican challenger Robert Dole, Clinton's wife Hillary was a prominent and polarizing figure. Throughout the next couple of examples, we will use her "thermometer ratings" by individual respondents to the National Election Study (NES) survey as our dependent variable. As we discussed briefly in Chapter 10, a thermometer rating is a survey respondent's answer to a question about how they *feel* (as opposed to how they *think*) toward particular individuals or groups on a scale that typically runs from 0 to 100. Scores of 50 indicate that the individual feels neither warm nor cold about the individual or group in question. Scores from 50 to 100 represent increasingly warm (or favorable) feelings, and scores from 50 to 0 represent increasingly cold (or unfavorable) feelings.

During the 1996 campaign, Ms. Clinton was identified as being a left-wing feminist. Given this, we theorize that there may have been a causal relationship between a respondent's family income and their thermometer rating of Ms. Clinton – with wealthier individuals, holding all else constant, liking her less – as well as a relationship between a respondent's gender and their thermometer rating of Ms. Clinton – with women, holding



```
. reg hillary_thermo income male female
note: female omitted because of collinearity

      Source |       SS           df       MS      Number of obs   =     1,542
-------------+----------------------------------   F(2, 1539)      =     49.17
       Model |   80916.663         2  40458.3315   Prob > F        =    0.0000
    Residual |  1266234.71     1,539  822.764595   R-squared       =    0.0601
-------------+----------------------------------   Adj R-squared   =    0.0588
       Total |  1347151.37     1,541  874.205954   Root MSE        =    28.684

------------------------------------------------------------------------------
hillary_th~o |      coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |  -.8407732     .117856    -7.13   0.000    -1.071949   -.6095978
        male |  -8.081448    1.495216    -5.40   0.000    -11.01432   -5.148572
      female |          0  (omitted)
       _cons |   69.26185     1.92343    36.01   0.000   0.65.48903    73.03467
------------------------------------------------------------------------------
```

Figure 11.1 Stata output when we include both gender dummy variables in our model

all else constant, liking her more. For the sake of this example, we are going to assume that both our dependent variable and our income independent variable are continuous.[1] Each respondent's gender was coded as equaling either 1 for "male" or 2 for "female." Although we could leave this gender variable as it is and run our analyses, we chose to use this variable to create two new dummy variables, "male" equaling 1 for "yes" and 0 for "no," and "female" equaling 1 for "yes" and 0 for "no."

Our first inclination is to estimate an OLS model in which the specification is the following:

$$\text{Hillary Thermometer}_i = \alpha + \beta_1 \text{Income}_i + \beta_2 \text{Male}_i + \beta_3 \text{Female}_i + u_i.$$

But if we try to estimate this model, our statistical computer program will revolt and give us an error message.[2] Figure 11.1 shows a screen shot of what this output looks like in Stata. We can see that Stata has reported the results from the following model instead of what we asked for:

$$\text{Hillary Thermometer}_i = \alpha + \beta_1 \text{Income}_i + \beta_3 \text{Female}_i + u_i.$$

Instead of the estimates for β_2 on the second row of parameter estimates, we get a note that this variable was "dropped." This is the case because we have failed to meet the additional minimal mathematical criteria that we introduced when we moved from two-variable OLS to multiple OLS in Chapter 10 – "no perfect multicollinearity." The reason that we have failed to meet this is that, for two of the independent variables in our model, Male_i and Female_i, it is the case that

$$\text{Male}_i + \text{Female}_i = 1 \quad \forall\, i.$$

[1] In this survey, a respondent's family income was measured on a scale ranging from 1 to 24 according to which category of income ranges they chose as best describing their family's income during 1995.

[2] Most programs will throw one of the two variables out of the model and report the results from the resulting model along with an error message.

Table 11.1 Two models of the effects of gender and income on Hillary Clinton Thermometer scores

Independent variable	Model 1	Model 2
Male	—	−8.08***
		(1.50)
Female	8.08***	—
	(1.50)	
Income	−0.84***	−0.84***
	(0.12)	(0.12)
Intercept	61.18***	69.26***
	(2.22)	(1.92)
R^2	0.06	0.06
n	1542	1542

Notes: The dependent variable in both models is the respondent's thermometer score for Hillary Clinton.
Standard errors in parentheses.
Two-sided t-tests: ***$p < 0.01$; **$p < 0.05$; *$p < 0.10$.

In other words, our variables "Male" and "Female" are perfectly correlated: If we know a respondent's value on the "Male" variable, then we know their value on the "Female" variable with perfect certainty.

When this happens with dummy variables, we call this situation the **dummy-variable trap**. To avoid the dummy-variable trap, we have to omit one of our dummy variables. But we want to be able to compare the effects of being male with the effects of being female to test our hypothesis. How can we do this if we have to omit one of our two variables that measures gender? Before we answer this question, let's look at the results in Table 11.1 from the two different models in which we omit one of these two variables. We can learn a lot by looking at what is and what is not the same across these two models. In both models, the parameter estimate and standard error for income are identical. The R^2 statistic is also identical. The parameter estimate and the standard error for the intercept are different across the two models. The parameter estimate for male is −8.08, whereas that for female is 8.08, although the standard error for each of these parameter estimates is 0.12. If you're starting to think that all of these similarities cannot have happened by coincidence, you are correct. In fact, these two models are, mathematically speaking, the same model. All of the $\hat{Y}$ values and residuals for the individual cases are *exactly* the same. With income held constant, the estimated difference between being male and being female is 8.08. The sign on this parameter estimate switches from positive to negative when we go from Model 1 to Model 2 because we are phrasing the question differently across the two models:

- For Model 1: “What is the estimated difference for a female compared with a male?”
- For Model 2: “What is the estimated difference for a male compared with a female?”

So why are the intercepts different? Think back to our discussions in Chapters 9 and 10 about the interpretation of the intercept – it is the estimated value of the dependent variable when the independent variables are all equal to zero. In Model 1 this means the estimated value of the dependent variable for a low-income man. In Model 2 this means the estimated value of the dependent variable for a low-income woman. And the difference between these two values – you guessed it – is $61.18 - 69.26 = -8.08$!

What does the regression line from Model 1 or Model 2 look like? The answer is that it depends on the gender of the individual for which we are plotting the line, but that it does not depend on which of these two models we use. For men, where $\text{Female}_i = 0$ and $\text{Male}_i = 1$, the predicted values are calculated as follows:

Model 1 for Men:

$$\hat{Y}_i = 61.18 + (8.08 \times \text{Female}_i) - (0.84 \times \text{Income}_i)$$

$$\hat{Y}_i = 61.18 + (8.08 \times 0) - (0.84 \times \text{Income}_i)$$

$$\hat{Y}_i = 61.18 - (0.84 \times \text{Income}_i);$$

Model 2 for Men:

$$\hat{Y}_i = 69.26 - (8.08 \times \text{Male}_i) - (0.84 \times \text{Income}_i)$$

$$\hat{Y}_i = 69.26 - (8.08 \times 1) - (0.84 \times \text{Income}_i)$$

$$\hat{Y}_i = 61.18 - (0.84 \times \text{Income}_i).$$

So we can see that, for men, regardless of whether we use the results from Model 1 or Model 2, the formula for predicted values is the same. For women, where $\text{Female}_i = 1$ and $\text{Male}_i = 0$, the predicted values are calculated as follows:

Model 1 for Women:

$$\hat{Y}_i = 61.18 + (8.08 \times \text{Female}_i) - (0.84 \times \text{Income}_i)$$

$$\hat{Y}_i = 61.18 + (8.08 \times 1) - (0.84 \times \text{Income}_i)$$

$$\hat{Y}_i = 69.26 - (0.84 \times \text{Income}_i);$$

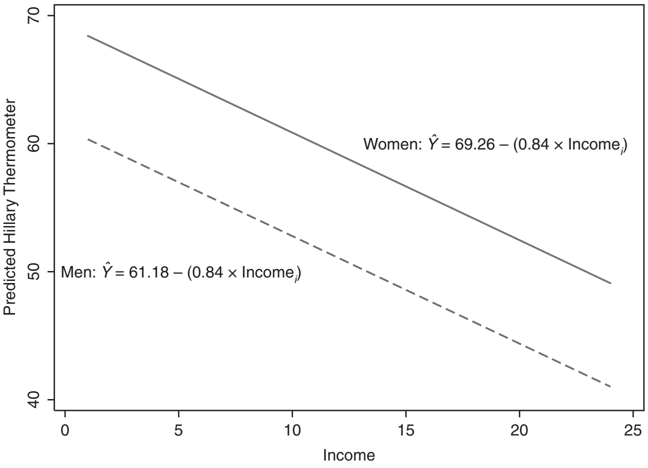


Figure 11.2 Regression lines from the model with a dummy variable for gender

Model 2 for Women:

$$\hat{Y}_i = 69.26 - (8.08 \times \text{Male}_i) - (0.84 \times \text{Income}_i)$$
$$\hat{Y}_i = 69.26 - (8.08 \times 0) - (0.84 \times \text{Income}_i)$$
$$\hat{Y}_i = 69.26 - (0.84 \times \text{Income}_i).$$

Again, the formula from Model 1 is the same as the formula from Model 2 for women. To illustrate these two sets of predictions, we have plotted them in Figure 11.2. Given that the two predictive formulae have the same slope, it is not surprising to see that the two lines in this figure are parallel to each other with the intercept difference determining the space between them.

11.2.2 Using Dummy Variables to Test Hypotheses about a Categorical Independent Variable with More Than Two Values

As you might imagine, when we have a categorical variable with more than two categories and we want to include it in an OLS model, things get more complicated. We'll keep with our running example of modeling Hillary Clinton Thermometer scores as a function of individuals' characteristics and opinions. In this section we work with a respondent's religious affiliation as an independent variable. The frequency of different responses to this item in the 1996 NES is displayed in Table 11.2.

Could we use the Religious Identification variable as it is in our regression models? That would be a bad idea. Remember, this is a categorical variable, in which the values of the variable are not ordered from lowest to highest. Indeed, there is no such thing as "lowest" or "highest" on

Table 11.2 Religious identification in the 1996 NES

Assigned numeric value	Category	Frequency	Percent
0	Protestant	683	39.85
1	Catholic	346	20.19
2	Jewish	22	1.28
3	Other	153	8.93
4	None	510	29.75
	Totals	1714	100

this variable. So running a regression model with these data as they are would be meaningless. But beware: *Your statistics package does not know that this is a categorical variable*. It will be more than happy to estimate the regression and report parameter estimates to you, even though these estimates will be nonsensical.

In the previous section, in which we had a categorical variable (Gender) with only two possible values, we saw that, when we switched which value was represented by "1" and "0," the estimated parameter switched signs. This was the case because we were asking a different question. With a categorical independent variable that has more than two values, we have more than two possible questions that we can ask. Because using the variable as is is not an option, the best strategy for modeling the effects of such an independent variable is to include in our regression a dummy variable for each value of that independent variable *except one*.[3] The value of the independent variable for which we do not include a dummy variable is known as the **reference category**. This is the case because the parameter estimates for all of the dummy variables representing the other values of the independent variable are estimated with reference to that value of the independent variable. So let's say that we choose to estimate the following model:

$$\begin{aligned}\text{Hillary Thermometer}_i = \alpha + \beta_1\text{Income}_i + \beta_2\text{Protestant}_i + \beta_3\text{Catholic}_i \\ + \beta_4\,\text{Jewish}_i + \beta_5\text{Other}_i + u_i.\end{aligned}$$

For this model we would be using "None" as our reference category for religious identification. This would mean that $\hat{\beta}_2$ would be the estimated effect of being Protestant *relative to* being nonreligious, and we

[3] If our theory was that only one category, such as Catholics, was different from all of the others, then we would collapse the remaining categories of the variable in question together and we would have a two-category independent variable. We should do this only if we have a theoretical justification for doing so.

Table 11.3 The same model of religion and income on Hillary Clinton Thermometer scores with different reference categories

Independent variable	Model 1	Model 2	Model 3	Model 4	Model 5
Income	−0.97***	−0.97***	−0.97***	−0.97***	−0.97***
	(0.12)	(0.12)	(0.12)	(0.12)	(0.12)
Protestant	−4.24*	−6.66*	−24.82***	−6.30**	—
	(1.77)	(2.68)	(6.70)	(2.02)	—
Catholic	2.07	−0.35	−18.51**	—	6.30**
	(2.12)	(2.93)	(6.80)	—	(2.02)
Jewish	20.58**	18.16**	—	18.51**	24.82***
	(6.73)	(7.02)	—	(6.80)	(6.70)
Other	2.42	—	−18.16**	0.35	6.66*
	(2.75)	—	(7.02)	(2.93)	(2.68)
None	—	−2.42	−20.58**	−2.07	4.24*
	—	(2.75)	(6.73)	(2.12)	(1.77)
Intercept	68.40***	70.83***	88.98***	70.47***	64.17***
	(2.19)	(2.88)	(6.83)	(2.53)	(2.10)
R^2	0.06	0.06	0.06	0.06	0.06
n	1542	1542	1542	1542	1542

Notes: The dependent variable in both models is the respondent's thermometer score for Hillary Clinton.

Standard errors in parentheses.

Two-sided t-tests: ***$p < 0.01$; **$p < 0.05$; *$p < 0.10$.

could use this value along with its standard error to test the hypothesis that this effect was statistically significant, controlling for the effects of income. The remaining parameter estimates ($\hat{\beta}_3$, $\hat{\beta}_4$, and $\hat{\beta}_5$) would all also be interpreted as the estimated effect of being in each of the remaining categories relative to "None." The value that we choose to use as our reference category does not matter, as long as we interpret our results appropriately. But we can use the choice of the reference category to focus on the relationships in which we are particularly interested. For each possible pair of categories of the independent variable, we can conduct a separate hypothesis test. The easiest way to get all of the p-values in which we are interested is to estimate the model multiple times with different reference categories. Table 11.3 displays a model of Hillary Clinton Thermometer scores with the five different choices of reference categories. It is worth emphasizing that this is *not* a table with five different models, but that this *is* a table with the same model displayed five different ways. From this table we can see that, when we control for the effects of income, some of the categories

Table 11.4 Model of bargaining duration

Independent variable	Parameter estimate
Ideological Range of the Government	2.57* (1.95)
Number of Parties in the Government	−15.44*** (2.30)
Post-Election	5.87** (2.99)
Continuation Rule	−6.34** (3.34)
Intercept	19.63*** (3.82)
R^2	0.62
n	203

Notes: The dependent variable is the number of days before each government was formed.
Standard errors in parentheses.
One-sided t-tests: $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

of religious affiliation are statistically different from each other in their evaluations of Hillary Clinton whereas others are not. This raises an interesting question: Can we say that the effect of religious affiliation, controlling for income, is statistically significant? The answer is that it depends on which categories of religious affiliation we want to compare.

11.2.3 Using Dummy Variables to Test Hypotheses about Multiple Independent Variables

It is often the case that we will want to use multiple dummy independent variables in the same model. Consider the model presented in Table 11.4, which was estimated from data from a paper by Lanny Martin and Georg Vanberg (2003) on the length of time that it takes for coalition governments to form in Western Europe.[4] The dependent variable is the number of days that a government took to form. The model has two continuous independent variables (“Ideological Range of the Government” and

[4] The model that we present in Table 11.4 has been changed from what Martin and Vanberg present in their paper. This model contains fewer variables than the main model of interest in that paper. This model was also estimated using OLS regression whereas the models presented by the original authors were estimated as proportional hazard models. And, we have not reported the results for a technical variable (labeled “Number of Government Parties ∗ ln(T)” by the authors) from the original specification. All of these modifications were made to make this example more tractable.

Table 11.5 Two overlapping dummy variables in models by Martin and Vanberg

		Continuation rule?	
		No (0)	Yes (1)
Post-	No (0)	61	25
Election?	Yes (1)	76	41

Note: Numbers in cells represent the number of cases.

"Number of Parties in the Government") measuring characteristics of the government that eventually formed and two dummy independent variables reflecting the circumstances under which bargaining took place. The variable "Post-Election" identifies governments that were formed in the immediate aftermath of an election while "Continuation Rule" identifies bargaining that took place in settings where the political parties from the outgoing government had the first opportunity to form a new government. As Table 11.5 indicates, all four possible combinations of these two dummy variables occurred in the sample of cases on which the model presented in Table 11.4 was estimated.

So, how do we interpret these results? It's actually not as hard as it might first appear. Remember from Chapter 10 that when we moved from a bivariate regression model to a multiple regression model, we had to interpret each parameter estimate as the estimated effect of a one-point increase in that particular independent variable on the dependent variable, *while controlling for the effects of all other independent variables in the model*. This has not changed. Instead, what is a little different from the examples that we have considered before is that we have two dummy independent variables that can vary independently of each other. So, when we interpret the estimated effect of each continuous independent variable, we interpret the parameter estimate as the estimated effect of a one-point increase in that particular independent variable on the dependent variable, while controlling for the effects of all other independent variables in the model, including the two dummy variables. And, when we interpret the estimated effect of each dummy independent variable, we interpret the parameter estimate as the estimated effect of that variable having a value of one versus zero on the dependent variable, while controlling for the effects of all other independent variables in the model, including the other dummy variable. For instance, the estimated effect of a one-unit increase in the Ideological Range of the Government, holding everything else constant, is a 2.57 day increase in the amount of bargaining time. And, the estimated effect of bargaining in the aftermath of an election (versus at a different time), holding everything else constant, is a 5.87 day increase in the amount of bargaining time.

11.3 TESTING INTERACTIVE HYPOTHESES WITH DUMMY VARIABLES

All of the OLS models that we have examined so far have been what we could call "additive models." To calculate the $\hat{Y}$ value for a particular case from an additive model, we simply multiply each independent variable value for that case by the appropriate parameter estimate and *add* these values together. In this section we explore some **interactive models**. Interactive models contain at least one independent variable that we create by multiplying together two or more independent variables. When we specify interactive models, we are testing theories about how the effects of one independent variable on our dependent variable may be contingent on the value of another independent variable. We will continue with our running example of modeling a respondent's thermometer score for Hillary Clinton. We begin with an additive model with the following specification:

$$\begin{aligned}\text{Hillary Thermometer}_i = \alpha + \beta_1 &\text{Women's Movement Thermometer}_i \\ &+ \beta_2 \text{Female}_i + u_i.\end{aligned}$$

In this model we are testing theories that a respondent's feelings toward Hillary Clinton are a function of their feelings toward the women's movement and their own gender. This specification seems pretty reasonable, but we also want to test an additional theory that the effect of feelings toward the women's movement have a stronger effect on feelings toward Hillary Clinton among women than they do among men. Notice the difference in phrasing there. In essence, we want to test the hypothesis that the slope of the line representing the relationship between Women's Movement Thermometer and Hillary Clinton Thermometer is *steeper* for women than it is for men. To test this hypothesis, we need to create a new variable that is the product of the two independent variables in our model and include this new variable in our model:

$$\begin{aligned}&\text{Hillary Thermometer}_i \\ &\quad = \alpha + \beta_1 \text{Women's Movement Thermometer}_i \\ &\qquad + \beta_2 \text{Female}_i + \beta_3 (\text{Women's Movement Thermometer}_i \times \text{Female}_i) + u_i.\end{aligned}$$

By specifying our model as such, we have essentially created two different models for women and men. So we can rewrite our model as follows:

For Men (Female = 0):

$$\text{Hillary Thermometer}_i = \alpha + \beta_1 \text{Women's Movement Thermometer}_i + u_i;$$

For Women (Female = 1):

$$\begin{aligned}\text{Hillary Thermometer}_i = \alpha &+ \beta_1 \text{Women's Movement Thermometer}_i \\ &+ (\beta_2 + \beta_3)(\text{Women's Movement Thermometer}_i) \\ &+ u_i.\end{aligned}$$

And we can rewrite the formula for women as:

For Women (Female = 1):

$$\begin{aligned}\text{Hillary Thermometer}_i = (\alpha + \beta_2) &+ (\beta_1 + \beta_3) \\ &(\text{Women's Movement Thermometer}_i) + u_i.\end{aligned}$$

What this all boils down to is that we are allowing our regression line to be different for men and women. For men, the intercept is α and the slope is β_1. For women, the intercept is $\alpha + \beta_2$ and the slope is $\beta_1 + \beta_3$. However, if $\beta_2 = 0$ and $\beta_3 = 0$, then the regression lines for men and women will be the same. Table 11.6 shows the results for our additive and interactive models of the effects of gender and feelings toward the women's movement on Hillary Clinton Thermometer scores. We can see from the interactive model that we can reject the null hypothesis that $\beta_2 = 0$ and the null hypothesis that $\beta_3 = 0$, so our regression lines for men and women are different. We can also see that the intercept for the line for women $(\alpha + \beta_2)$ is higher than the intercept for men (α). But, contrary to our expectations, the estimated effect of the Women's Movement Thermometer

Table 11.6 The effects of gender and feelings toward the women's movement on Hillary Clinton Thermometer scores

Independent variable	Additive model	Interactive model
Women's Movement Thermometer	0.68***	0.75***
	(0.03)	(0.05)
Female	7.13***	15.21***
	(1.37)	(4.19)
Women's Movement Thermometer × Female	—	−0.13**
		(0.06)
Intercept	5.98**	1.56
	(2.13)	(3.04)
R^2	0.27	0.27
n	1466	1466

Notes: The dependent variable in both models is the respondent's thermometer score for Hillary Clinton.

Standard errors in parentheses.

Two-sided t-tests: $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

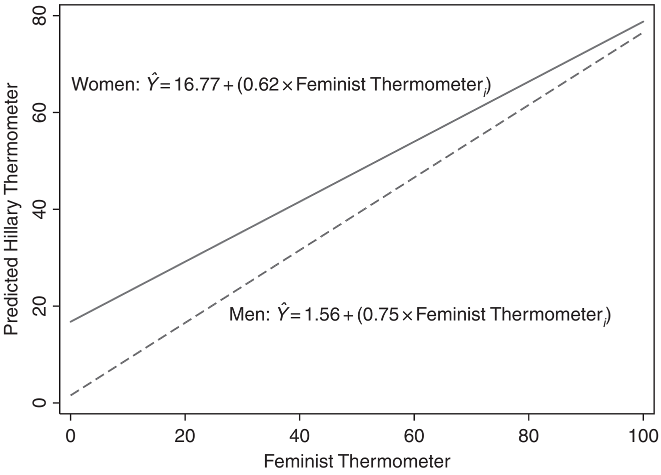


Figure 11.3 Regression lines from the interactive model

for men is greater than the effect of the Women's Movement Thermometer for women.

The best way to see the combined effect of all of the results from the interactive model in Table 11.6 is to look at them graphically in a figure such as Figure 11.3. From this figure we can see the regression lines for men and for women across the range of the independent variable. It is clear from this figure that, although women are generally more favorably inclined toward Hillary Clinton, this gender gap narrows when we compare those individuals who feel more positively toward the feminist movement.

11.4 OUTLIERS AND INFLUENTIAL CASES IN OLS

In Chapter 6 we advocated using descriptive statistics to identify outlier values for each continuous variable. In the context of a single variable, an outlier is an extreme value relative to the other values for that variable. But in the context of an OLS model, when we say that a single case is an outlier, we could mean several different things. For this reason, we prefer to use the term "influential" instead of "outlier" in the context of a regression model.

As we discussed in Chapter 6, we should always strive to know our data well. This means looking at individual variables one at a time before we estimate a regression with them and identifying univariate outliers. But just because a case is an outlier in the univariate sense does not necessarily imply that it will be an **influential case** in a regression. Nonetheless, we should look for outliers in the single-variable sense before we estimate our models and make sure that they are actual values and not values created by some type of data management mistake.

In the regression setting, individual cases can be influential in several different ways:

1. They can have unusual independent variable values. This is known as a case having large **leverage**. This can be the result of a single case having an unusual value for a single variable. A single case can also have large leverage because it has an unusual *combination* of values across two or more variables. There are a variety of different measures of leverage, but they all make calculations across the values of independent variables in order to identify individual cases that are particularly different.
2. They can have large residual values (usually we look at squared residuals to identify outliers of this variety).
3. They can have both large leverage and large residual values.

The relationship among these different concepts of influence for a single case in OLS is often summarized as

$$\text{influence}_i = \text{leverage}_i \times \text{residual}_i.$$

As this formula indicates, the influence of a particular case is determined by the combination of its leverage and residual values. There are a variety of different ways to measure these different factors. We explore a couple of them in the following sections with a controversial real-world example.

11.4.1 Identifying Influential Cases

One of the most famous cases of outliers and influential cases in political data comes from the 2000 US presidential election in Florida. In an attempt to measure the extent to which ballot irregularities may have influenced election results, a variety of models were estimated in which the raw vote numbers for candidates across different counties were the dependent variables of interest. These models were fairly unusual because the parameter estimates and other quantities that are most often the focus of our model interpretations were of little interest. Instead, these were models for which the most interesting quantities were the diagnostics of influential cases. As an example of such a model, we will work with the following:

$$\text{Buchanan}_i = \alpha + \beta\text{Gore}_i + u_i.$$

In this model the cases are individual counties in Florida, the dependent variable (Buchanan_i) is the number of votes in each Florida county for the independent candidate Patrick Buchanan, and the independent variable is the number of votes in each Florida county for the Democratic Party's nominee Al Gore (Gore_i). Such models are unusual in the sense that there is no claim of an underlying causal relationship between the

Table 11.7 Votes for Gore and Buchanan in Florida counties in the 2000 US presidential election

Independent variable	Parameter estimate
Votes for Gore	0.004***
	(0.0005)
Intercept	80.63*
	(46.4)
R^2	0.48
n	67

Notes: The dependent variable is the number of votes for Patrick Buchanan.
Standard errors in parentheses.
Two-sided t-tests: $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

independent and dependent variables. Instead, the theory behind this type of model is that there should be a strong systematic relationship between the number of votes cast for Gore and those cast for Buchanan across the Florida counties.[5] There was a suspicion that the ballot structure used in some counties – especially the infamous "butterfly ballot" – was such that it confused some voters who intended to vote for Gore into voting for Buchanan. If this was the case, we should see these counties appearing as highly influential after we estimate our model.

We can see from Table 11.7 that there was indeed a statistically significant positive relationship between Gore and Buchanan votes, and that this simple model accounts for 48 percent of the variation in Buchanan votes across the Florida counties. But, as we said before, the more interesting inferences from this particular OLS model are about the influence of particular cases. Figure 11.4 presents a Stata lvr2plot (short for "leverage-versus-residual-squared plot") that displays Stata's measure of leverage on the vertical dimension and a normalized measure of the squared residuals on the horizontal dimension. The logic of this figure is that, as we move to the right of the vertical line through this figure, we are seeing cases with unusually large residual values, and that, as we move above the horizontal line through this figure, we are seeing cases with unusually large leverage values. Cases with both unusually large residual and leverage values are highly influential. From Figure 11.4 it is apparent that Pinellas,

[5] Most of the models of this sort make adjustments to the variables (for example, logging the values of both the independent and dependent variables) to account for possibilities of nonlinear relationships. In the present example we avoided doing this for the sake of simplicity.

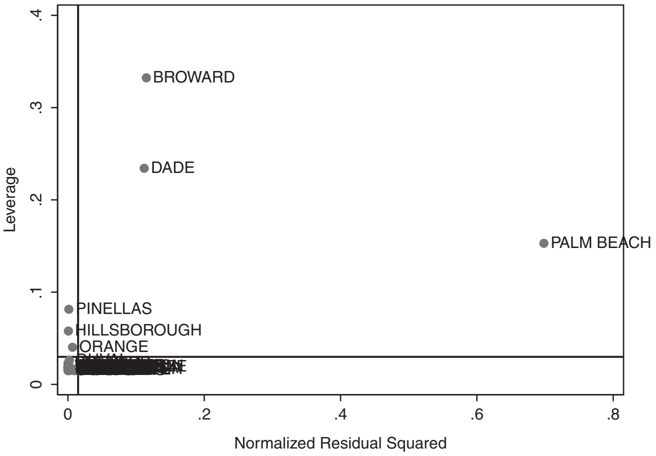


Figure 11.4 Stata lvr2plot for the model presented in Table 11.7

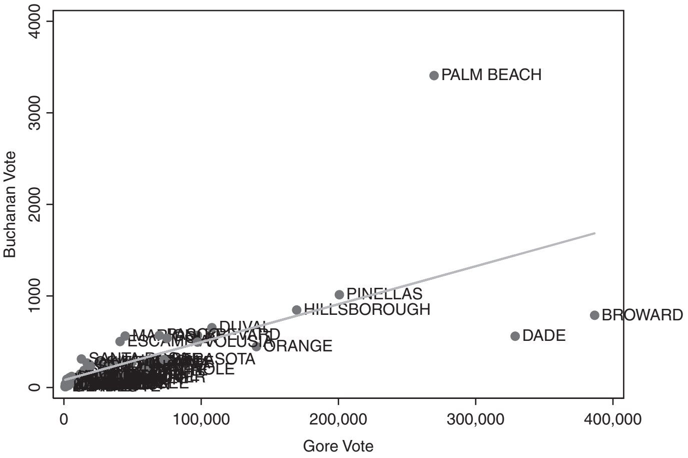


Figure 11.5 OLS line with scatter plot for Florida 2000

Hillsborough, and Orange counties had large leverage values but not particularly large squared residual values, whereas Dade, Broward, and Palm Beach counties were highly influential with both large leverage values and large squared residual values.

We can get a better idea of the correspondence between Figure 11.4 and Table 11.7 from Figure 11.5, in which we plot the OLS regression line through a scatter plot of the data. From this figure it is clear that Palm Beach was well above the regression line whereas Broward and Dade counties were well below the regression line. By any measure, these three cases were quite influential in our model.

A more specific method for detecting the influence of an individual case involves estimating our model with and without particular cases to

Table 11.8 The five largest (absolute-value) DFBETA scores for β from the model presented in Table 11.7

County	DFBETA
Palm Beach	6.993
Broward	−2.514
Dade	−1.772
Orange	−0.109
Pinellas	0.085

see how much this changes specific parameter estimates. The resulting calculation is known as the **DFBETA score** (Belsley, Kuh, and Welsch, 1980). DFBETA scores are calculated as the difference in the parameter estimate without each case divided by the standard error of the original parameter estimate. Table 11.8 displays the five largest absolute values of DFBETA for the slope parameter (β) from the model presented in Table 11.7. Not surprisingly, we see that omitting Palm Beach, Broward, or Dade has the largest impact on our estimate of the slope parameter.

11.4.2 Dealing with Influential Cases

Now that we have discussed the identification of particularly influential cases on our models, we turn to the subject of what to do once we have identified such cases. The first thing to do when we identify a case with substantial influence is to double-check the values of all variables for such a case. We want to be certain that we have not "created" an influential case through some error in our data management procedures. Once we have corrected for any errors of data management and determined that we still have some particularly influential case(s), it is important that we report our findings about such cases along with our other findings. There are a variety of strategies for doing so. Table 11.9 shows five different models that reflect various approaches to reporting results with highly influential cases. In Model 1 we have the original results as reported in Table 11.7. In Model 2 we have added a dummy variable that identifies and isolates the effect of Palm Beach County. This approach is sometimes referred to as **dummying out** influential cases. We can see why this is called "dummying out" from the results in Model 3, which is the original model with the observation for Palm Beach County dropped from the analysis. The parameter estimates and standard errors for the intercept and slope parameters are identical from Models 2 and 3. The only differences are the model R^2 statistic, the number of cases, and the additional parameter estimate reported in Model 2 for the Palm Beach County dummy variable.[6]

[6] This parameter estimate was viewed by some as an estimate of how many votes the ballot irregularities cost Al Gore in Palm Beach County. But if we look at Model 4, where we include dummy variables for Broward and Dade counties, we can see the basis for an argument that in these two counties there is evidence of bias in the opposite direction.

Table 11.9 Votes for Gore and Buchanan in Florida counties in the 2000 US presidential election

Independent variable	Model 1	Model 2	Model 3	Model 4	Model 5
Gore	0.004*** (0.0005)	0.003*** (0.0002)	0.003*** (0.0002)	0.005*** (0.0003)	0.005*** (0.0003)
Palm Beach dummy	—	2606.3*** (150.4)	—	2095.5*** (110.6)	—
Broward dummy	—	—	—	−1066.0*** (131.5)	—
Dade dummy	—	—	—	−1025.6*** (120.6)	—
Intercept	80.6* (46.4)	110.8*** (19.7)	110.8*** (19.7)	59.0*** (13.8)	59.0*** (13.8)
R^2	0.48	0.91	0.63	0.96	0.82
n	67	67	66	67	64

Notes: The dependent variable is the number of votes for Patrick Buchanan.
Standard errors in parentheses.
Two-sided t-tests: $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

In Model 4 and Model 5, we see the results from dummying out the three most influential cases and then from dropping them out of the analysis.

Across all five of the models shown in Table 11.9, the slope parameter estimate remains positive and statistically significant. In most models, this would be the quantity in which we are most interested (testing hypotheses about the relationship between X and Y). Thus the relative robustness of this parameter across model specifications would be comforting. Regardless of the effects of highly influential cases, it is important first to know that they exist and, second, to report accurately what their influence is and what we have done about them.

11.5 MULTICOLLINEARITY

When we specify and estimate a multiple OLS model, what is the interpretation of each individual parameter estimate? It is our best guess of the causal impact of a one-unit increase in the relevant independent variable on the dependent variable, controlling for all of the other variables in the model. Another way of saying this is that we are looking at the impact of a one-unit increase in one independent variable on the dependent variable when we “hold all other variables constant.” We know from Chapter 10 that a minimal mathematical property for estimating a multiple OLS model is that there is no perfect multicollinearity. Perfect multicollinearity, you will recall, occurs when one independent variable is an exact linear function of one or more other independent variables in a model.

In practice, perfect multicollinearity is usually the result of a small number of cases relative to the number of parameters we are estimating, limited independent variable values, or model misspecification. As we have noted, if there exists perfect multicollinearity, OLS parameters cannot be estimated. A much more common and vexing issue is **high multicollinearity**. As a result, when people refer to multicollinearity, they almost always mean "high multicollinearity." From here on, when we refer to "multicollinearity," we will mean "high, but less-than-perfect, multicollinearity." This means that two or more of the independent variables in the model are extremely highly correlated with one another.

11.5.1 How Does Multicollinearity Happen?

Multicollinearity is induced by a small number of degrees of freedom and/or high correlation between independent variables. Figure 11.6 provides a Venn diagram illustration that is useful for thinking about the effects of multicollinearity in the context of an OLS regression model. As you can see from this figure, X and Z are fairly highly correlated. Our regression model is

$$Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + u_i.$$

Looking at Figure 11.6, we can see that the R^2 from our regression model will be fairly high,

$$R^2 = \frac{f+d+b}{a+f+d+b}.$$

But we can also see from this figure that the areas for the estimation of our two slope parameters – area f for β_1 and area b for β_2 – are pretty small. Because of this, our standard errors for our slope parameters will tend to be fairly large, which makes discovering statistically significant relationships more difficult, and we will have difficulty making precise inferences about the impacts of both X and Z on Y. It is possible that because of this problem we would conclude neither X nor Z has much of an impact on Y. But clearly this is not the case. As we can see from the diagram, both X and Z *are* related to Y. The problem is that much of the covariation between X and Y and between Z and Y is also covariation

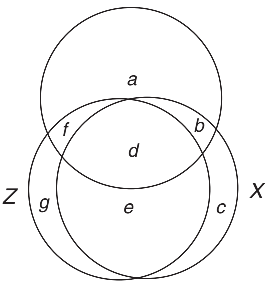


Figure 11.6 Venn diagram with multicollinearity

between X and Z. In other words, it is the size of area d that is causing us problems. We have precious little area in which to examine the effect of X on Y while holding Z constant, and likewise, there is precious little area in which to examine the effect of Z on Y while controlling for X.

It is worth emphasizing at this point that multicollinearity is not a statistical problem (examples of statistical problems include autocorrelation, bias, and heteroscedasticity). Rather, multicollinearity is a data problem. It is possible to have multicollinearity even when all of the assumptions of OLS from Chapter 9 are valid and all of the minimal mathematical requirements for OLS from Chapters 9 and 10 have been met. So, you might ask, what's the big deal about multicollinearity? To underscore the notion of multicollinearity as a data problem instead of a statistical problem, Christopher Achen (1982) has suggested that the word "multicollinearity" should be used interchangeably with **micronumerosity**. Imagine what would happen if we could double or triple the size of the diagram in Figure 11.6 without changing the relative sizes of any of the areas. As we expanded all of the areas, areas f and b would eventually become large enough for us to precisely estimate the relationships of interest.

11.5.2 Detecting Multicollinearity

It is very important to know when you have multicollinearity. In particular, it is important to distinguish situations in which estimates are statistically insignificant because the relationships just aren't there from situations in which estimates are statistically insignificant because of multicollinearity. The diagram in Figure 11.6 shows us one way in which we might be able to detect multicollinearity: If we have a high R^2 statistic, but none (or very few) of our parameter estimates is statistically significant, we should be suspicious of multicollinearity. We should also be suspicious of multicollinearity if we see that, when we add and remove independent variables from our model, the parameter estimates for other independent variables (and especially their standard errors) change substantially. If we estimated the model represented in Figure 11.6 with just one of the two independent variables, we would get a statistically significant relationship. But, as we know from the discussions in Chapter 10, this would be problematic. Presumably we have a theory about the relationship between each of these independent variables (X and Z) and our dependent variable (Y). So, although the estimates from a model with just X or just Z as the independent variable would help us to detect multicollinearity, they would suffer from bias. And, as we argued in Chapter 10, omitted-variables bias is a severe problem.

A more formal way to diagnose multicollinearity is to calculate the **variance inflation factor** (VIF) for each of our independent variables. This calculation is based on an **auxiliary regression model** in which one independent variable, which we will call X_j, is the dependent variable and all of the other independent variables are independent variables.[7] The R^2 statistic from this auxiliary model, R_j^2, is then used to calculate the VIF for variable j as follows:

$$\text{VIF}_j = \frac{1}{(1 - R_j^2)}.$$

Many statistical programs report the VIF and its inverse (1/VIF) by default. The inverse of the VIF is sometimes referred to as the tolerance index measure. The higher the VIF_j value, or the lower the tolerance index, the higher will be the estimated variance of X_j in our theoretically specified model. Another useful statistic to examine is the square root of the VIF. Why? Because the VIF is measured in terms of variance, but most of our hypothesis-testing inferences are made with standard errors. Thus the square root of the VIF provides a useful indicator of the impact the multicollinearity is going to have on hypothesis-testing inferences.

11.5.3 Multicollinearity: a Simulated Example

Thus far we have made a few scattered references to simulation. In this section we make use of simulation to better understand multicollinearity. Almost every statistical computer program has a set of tools for simulating data. When we use these tools, we have an advantage that we do not ever have with real-world data: we can *know* the underlying "population" characteristics (because we create them). When we know the population parameters for a regression model and draw sample data from this population, we gain insights into the ways in which statistical models work.

So, to simulate multicollinearity, we are going to create a population with the following characteristics:

1. Two variables X_{1i} and X_{2i} such that the correlation $r_{X_{1i},X_{2i}} = 0.9$.
2. A variable u_i randomly drawn from a normal distribution, centered around 0 with variance equal to 1 [$u_i \sim N(0, 1)$].
3. A variable Y_i such that $Y_i = 0.5 + 1X_{1i} + 1X_{2i} + u_i$.

[7] Students facing OLS diagnostic procedures are often surprised that the first thing that we do after we estimate our theoretically specified model of interest is to estimate a large set of atheoretical auxiliary models to test the properties of our main model. We will see that, although these auxiliary models lead to the same types of output that we get from our main model, we are often interested in only one particular part of the results from the auxiliary model. With our "main" model of interest, we have learned that we should include every variable that our theories tell us should be included and exclude all other variables. In auxiliary models, we do not follow this rule. Instead, we are running these models to test whether certain properties have or have not been met in our original model.

We can see from the description of our simulated population that we have met all of the OLS assumptions, but that we have a high correlation between our two independent variables. Now we will conduct a series of random draws (samples) from this population and look at the results from the following regression models:

$$
\begin{aligned}
&\text{Model 1:} \quad Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,\\
&\text{Model 2:} \quad Y_i = \alpha + \beta_1 X_{1i} + u_i,\\
&\text{Model 3:} \quad Y_i = \alpha + \beta_2 X_{2i} + u_i.
\end{aligned}
$$

In each of these random draws, we increase the size of our sample starting with $n = 5$, then 10, and finally 25 cases. Results from models estimated with each sample of data are displayed in Table 11.10. In the first column of results ($n = 5$), we can see that both slope parameters are

Table 11.10 Random draws of increasing size from a population with substantial multicollinearity

Estimate	Sample: $n = 5$	Sample: $n = 10$	Sample: $n = 25$
Model 1:			
$\hat{\beta}_1$	0.546	0.882	1.012**
	(0.375)	(0.557)	(0.394)
$\hat{\beta}_2$	1.422*	1.450**	1.324***
	(0.375)	(0.557)	(0.394)
$\hat{\alpha}$	1.160**	0.912***	0.579***
	(0.146)	(0.230)	(0.168)
R^2	0.99	0.93	0.89
VIF_1	5.26	5.26	5.26
VIF_2	5.26	5.26	5.26
Model 2:			
$\hat{\beta}_1$	1.827**	2.187***	2.204***
	(0.382)	(0.319)	(0.207)
$\hat{\alpha}$	1.160**	0.912**	0.579***
	(0.342)	(0.302)	(0.202)
R^2	0.88	0.85	0.83
Model 3:			
$\hat{\beta}_2$	1.914***	2.244***	2.235***
	(0.192)	(0.264)	(0.192)
$\hat{\alpha}$	1.160***	0.912***	0.579***
	(0.171)	(0.251)	(0.188)
R^2	0.97	0.90	0.86

Notes: The dependent variable is $Y_i = 0.5 + 1X_{1i} + 1X_{2i} + u_i$.
Standard errors in parentheses.
Two-sided t-tests: $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

positive, as would be expected, but that the parameter estimate for X_1 is statistically insignificant and the parameter estimate for X_2 is on the borderline of statistical significance. The VIF statistics for both variables are equal to 5.26, indicating that the variance for each parameter estimate is substantially inflated by multicollinearity. The model's intercept is statistically significant and positive, but pretty far from what we know to be the true population value for this parameter. In Models 2 and 3 we get statistically significant positive parameter estimates for each variable, but both of these estimated slopes are almost twice as high as what we know to be the true population parameters. The 95 percent confidence interval for $\hat{\beta}_2$ does not include the true population parameter. This is a clear case of omitted-variables bias. When we draw a sample of 10 cases, we get closer to the true population parameters with $\hat{\beta}_1$ and $\hat{\alpha}$ in Model 1. The VIF statistics remain the same because we have not changed the underlying relationship between X_1 and X_2. This increase in sample size does not help us with the omitted-variables bias in Models 2 and 3. In fact, we can now reject the true population slope parameter for both models with substantial confidence. In our third sample with 25 cases, Model 1 is now very close to our true population model, in the sense of both the parameter values and that all of these parameter estimates are statistically significant. In Models 2 and 3, the omitted-variables bias is even more pronounced.

The findings in this simulation exercise mirror more general findings in the theoretical literature on OLS models. *Adding more data will alleviate multicollinearity, but not omitted-variables bias*. We now turn to an example of multicollinearity with real-world data.

YOUR TURN: Imagining a different simulation

How would the output in Table 11.10 be different if $r_{X_{1i},X_{2i}} = -0.9$?

11.5.4 Multicollinearity: a Real-World Example

In this section, we estimate a model of the thermometer scores for US voters for George W. Bush in 2004. Our model specification is the following:

$$\text{Bush Thermometer}_i = \alpha + \beta_1 \text{Income}_i + \beta_2 \text{Ideology}_i + \beta_3 \text{Education}_i \\ + \beta_4 \text{Party ID}_i + u_i.$$

Although we have distinct theories about the causal impact of each independent variable on people's feelings toward Bush, Table 11.11 indicates that some of these independent variables are substantially correlated with each other.

Table 11.11 Pairwise correlations between independent variables

	Bush Therm.	Income	Ideology	Education	Party ID
Bush Therm.	1.00	—	—	—	—
Income	0.09***	1.00	—	—	—
Ideology	0.56***	0.13***	1.00	—	—
Education	−0.07***	0.44***	−0.06*	1.00	—
Party ID	0.69***	0.15***	0.60***	0.06*	1.00

Notes: Cell entries are correlation coefficients.
Two-sided t-tests: ***$p < 0.01$; **$p < 0.05$; *$p < 0.10$.

Table 11.12 Model results from random draws of increasing size from the 2004 NES

Independent variable	Model 1	Model 2	Model 3
Income	0.77	0.72	0.11
	(0.90)	(0.51)	(0.15)
	{1.63}	{1.16}	{1.24}
Ideology	7.02	4.57*	4.26***
	(5.53)	(2.22)	(0.67)
	{3.50}	{1.78}	{1.58}
Education	−6.29	−2.50	−1.88***
	(3.32)	(1.83)	(0.55)
	{1.42}	{1.23}	{1.22}
Party ID	6.83	8.44***	10.00***
	(3.98)	(1.58)	(0.46)
	{3.05}	{1.70}	{1.56}
Intercept	21.92	12.03	13.73***
	(23.45)	(13.03)	(3.56)
R^2	0.71	0.56	0.57
n	20	74	821

Notes: The dependent variable is the respondent's thermometer score for George W. Bush.
Standard errors in parentheses; VIF statistics in braces.
Two-sided t-tests: ***$p < 0.01$; **$p < 0.05$; *$p < 0.10$.

In Table 11.12, we present estimates of our model using three different samples from the NES 2004 data. In Model 1, estimated with data from 20 randomly chosen respondents, we see that none of our independent variables are statistically significant despite the rather high R^2 statistic. The VIF statistics for Ideology and Party ID indicate that multicollinearity might be a problem. In Model 2, estimated with data from 74 randomly chosen respondents, Party ID is highly significant in the expected (positive) direction whereas Ideology is near the threshold of statistical significance. None of the VIF statistics for this model are stunningly high, though they are greater than 1.5 for Ideology, Education, and Party ID.[8] Finally, in Model 3, estimated with all 820 respondents for whom data on all of the variables were available, we see that Ideology, Party ID, and Education are all significant predictors of people's feelings toward Bush. The sample size is more than sufficient to overcome the VIF statistics for Party ID and Ideology. Of our independent variables, only Income remains statistically insignificant. Is this due to multicollinearity? After all, when we look at Table 11.11, we see that income has a highly significant positive correlation with Bush Thermometer scores. For the answer to this question, we need to go back to the lessons that we learned in Chapter 10: Once we control for the effects of Ideology, Party ID, and Education, the effect of income on people's feelings toward George W. Bush goes away.

11.5.5 Multicollinearity: What Should I Do?

In the introduction to this section on multicollinearity, we described it as a "common and vexing issue." The reason why multicollinearity is "vexing" is that there is no magical statistical cure for it. What is the best thing to do when you have multicollinearity? Easy (in theory): *collect more data*. But data are expensive to collect. If we had more data, we would use them and we wouldn't have hit this problem in the first place. So, if you do not have an easy way to increase your sample size, then multicollinearity ends up being something that you just have to live with. It is important to know that you have multicollinearity and to present your multicollinearity by reporting the results of VIF statistics or what happens to your model when you add and drop the "guilty" variables.

11.6 WRAPPING UP

The key to developing good models is having a good theory and then doing a lot of diagnostics to figure out what we have after estimating the model. What we've seen in this chapter is that there are additional (but not insurmountable!) obstacles to overcome when we consider that some of our theories involve noncontinuous independent variables. In the next chapter, we examine the research situations in which we encounter dummy dependent variables and a set of special circumstances that can arise when our data have been collected across time.

[8] When we work with real-world data, there tend to be many more changes as we move from sample to sample.

CONCEPTS INTRODUCED IN THIS CHAPTER

- auxiliary regression model – a regression model separate from the original theoretical model that is used to detect one or more statistical properties of the original model
- DFBETA score – a statistical measure for the calculation of the influence of an individual case on the value of a single parameter estimate
- dummying out – adding a dummy variable to a regression model to measure and isolate the effect of an influential observation
- dummy variable – a variable that takes on one of two values (usually one or zero)
- dummy-variable trap – perfect multicollinearity that results from the inclusion of dummy variables representing each possible value of a categorical variable
- high multicollinearity – in a multiple regression model, when two or more of the independent variables in the model are extremely highly correlated with one another, making it difficult to isolate the distinct effects of each variable
- influential case – in a regression model a case which has either a combination of large leverage and a large squared residual or a large DFBETA score
- interactive models – multiple regression models that contain at least one independent variable that we create by multiplying together two or more independent variables
- leverage – in a multiple regression model, the degree to which an individual case is unusual in terms of its value for a single independent variable, or its particular combination of values for two or more independent variables
- micronumerosity – a suggested synonym for multicollinearity
- reference category – in a multiple regression model, the value of a categorical independent variable for which we do not include a dummy variable
- variance inflation factor – a statistical measure to detect the contribution of each independent variable in a multiple regression model to overall multicollinearity

EXERCISES

1. Using the model presented in Table 11.4, how many days would you predict that it would take for a government to form if the government was made up of two different political parties with an ideological range of 2, if bargaining was taking place in the immediate aftermath of an election, and there was not a continuation rule? Show your work.

2. Using the model presented in Table 11.4, interpret the parameter estimate for the variable “Continuation Rule.”

3. Using the model presented in Table 11.4, interpret the parameter estimate for the variable “Number of Parties in the Government.”

4. Using the data set “nes2008.dta” (which is available on the textbook’s web site at www.cambridge.org/fpsr), investigate two possible causes of a respondent’s attitudes toward abortion (which you will, for the purposes of this exercise, need to treat as a continuous variable), using the respondent’s gender and the respondent’s level of education as your two key independent variables. First, construct an additive multiple regression model investigating the effects of gender and education on abortion attitudes. Next, construct an interactive multiple regression model that adds an interaction term for gender and education. Present the results of both models in a single table. Interpret, first, the additive regression model, and then interpret the interactive model. Does education have the same, a smaller, or larger effect on abortion attitudes for women than it does for men?

5. Using the data set “state_data.dta” (which is available on the textbook’s web site at www.cambridge.org/fpsr), estimate Model C in Table 10.7. Test for influential observations in the model using a leverage versus squared residual plot. Write about what this diagnostic test tells you.

6. Test for influential observations in the model that you estimated for Exercise 5 using DFBETA scores. Write about what this diagnostic test tells you.

7. Based on what you found in Exercises 5 and 6, how would you adjust the original model?

8. Test for multicollinearity in the model that you estimated for Exercise 5. Write about what you have found.