10 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.

10.1 EXTENSIONS OF OLS

In the previous two chapters we discussed in detail various aspects of the estimation and interpretation 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.

10.2 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.

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

During the 1996 U.S. 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 NES survey as our dependent variable. 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 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 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

```
.reg hillary_thermo income male female
```

Source	SS	df	MS		
Model	80916.663	2	40458.3315	Number of obs	= 1542
Residual	1266234.71	1539	822.764595	F(2, 1539)	= 49.17
				Prob > F	= 0.0000
Total	1347151.37	1541	874.205954	R–Squared	= 0.0601
				Adj R–Squared	= 0.0588
				Root MSE	= 28.684

hillary_th~o	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
income	–.8407732	.117856	– 7.13	0.000	–1.071948	–.6095978
male	(dropped)					
female	8.081448	1.495216	5.40	0.000	5.148572	11.01432
_cons	61.1804	2.220402	27.55	0.000	56.82507	65.53573

Figure 10.1. Stata output when we include both gender dummy variables in our model.

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 10.1 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 9 – "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

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

[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 10.1. 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	.06	.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: ***indicates $p < .01$; **indicates $p < .05$; *indicates $p < .10$.

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 10.1 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 8 and 9 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:

$$\begin{aligned}
\text{Model 1 for Men: } \hat{Y}_i &= 61.18 + (8.08 \times \text{Female}_i) - (0.84 \times \text{Income}_i), \\
\hat{Y}_i &= 61.18 + (8.08 \times 0) - (0.84 \times \text{Income}_i), \\
\hat{Y}_i &= 61.18 - (0.84 \times \text{Income}_i); \\
\text{Model 2 for Men: } \hat{Y}_i &= 69.26 - (8.08 \times \text{Male}_i) - (0.84 \times \text{Income}_i), \\
\hat{Y}_i &= 69.26 - (8.08 \times 1) - (0.84 \times \text{Income}_i), \\
\hat{Y}_i &= 61.18 - (0.84 \times \text{Income}_i).
\end{aligned}$$

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:

$$\begin{aligned}
\text{Model 1 for Women: } \hat{Y}_i &= 61.18 + (8.08 \times \text{Female}_i) - (0.84 \times \text{Income}_i), \\
\hat{Y}_i &= 61.18 + (8.08 \times 1) - (0.84 \times \text{Income}_i), \\
\hat{Y}_i &= 69.26 - (0.84 \times \text{Income}_i); \\
\text{Model 2 for Women: } \hat{Y}_i &= 69.26 - (8.08 \times \text{Male}_i) - (0.84 \times \text{Income}_i), \\
\hat{Y}_i &= 69.26 - (8.08 \times 0) - (0.84 \times \text{Income}_i), \\
\hat{Y}_i &= 69.26 - (0.84 \times \text{Income}_i).
\end{aligned}$$

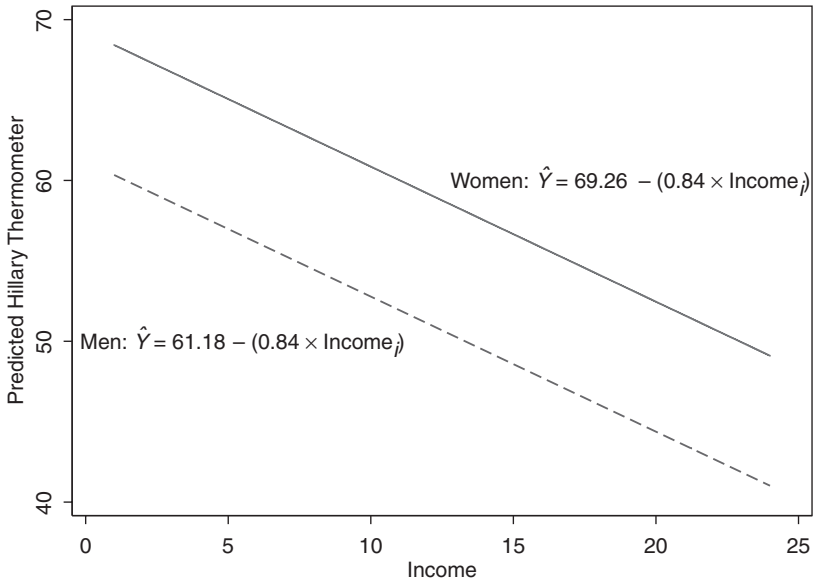

Figure 10.2. Regression lines from the model with a dummy variable for gender.

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 10.2. 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.

10.2.2 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 10.2.

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 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 run the regression

Table 10.2. Religious identification in the 1996 NES

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

and report parameter estimates to you, even though these estimates will be nonsensical.

In the previous subsection, 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 a dummy variable for all values 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 in reference to that value of the independent variable. So let's say that we choose to estimate the following model:

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

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 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

[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 10.3. 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	.06	.06	.06	.06	.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: ***indicates $p < .01$; **indicates $p < .05$; *indicates $p < .10$.

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 10.3 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 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 religion affiliation, controlling for income, is statistically significant? The answer is that it depends on which categories of religious affiliation we want to compare.

Table 10.4. 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	.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: ***indicates $p < .01$; **indicates $p < .05$; *indicates $p < .10$.

10.2.3 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 10.4 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 "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

[4] The model that we present in Table 10.4 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 10.5. Two overlapping dummy variables in models by Martin and Vanberg

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

Note: Numbers in cells represent the number of cases.

outgoing government had the first opportunity to form a new government. As Table 10.5 indicates, all four possible combinations of these two dummy variables occurred in the sample of cases on which the model presented in Table 10.4 was estimated.

So, how do we interpret these results? It's actually not as hard as it might first appear. Remember from Chapter 9 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.

10.3 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:

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

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

$$\begin{aligned}&\text{for Men (Female} = 0): \text{Hillary Thermometer}_i = \alpha \\ &\quad + \beta_1 \text{Women's Movement Thermometer}_i + u_i; \\ &\text{for Women (Female} = 1): \text{Hillary Thermometer}_i = \alpha \\ &\quad + \beta_1 \text{Women's Movement Thermometer}_i \\ &\quad + (\beta_2 + \beta_3)(\text{Women's Movement Thermometer}_i) + u_i.\end{aligned}$$

Table 10.6. 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	.27	.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: ***indicates $p < .01$; **indicates $p < .05$; *indicates $p < .10$.

And we can rewrite the formula for women as

$$\text{for Women (Female} = 1) : \text{Hillary Thermometer}_i = (\alpha + \beta_2) + (\beta_1 + \beta_3)(\text{Women's Movement Thermometer}_i) + u_i.$$

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 10.6 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, perhaps contrary to our expectations, the estimated effect of the Women's Movement Thermometer 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 10.6 is to look at them graphically in a figure such as Figure 10.3. 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

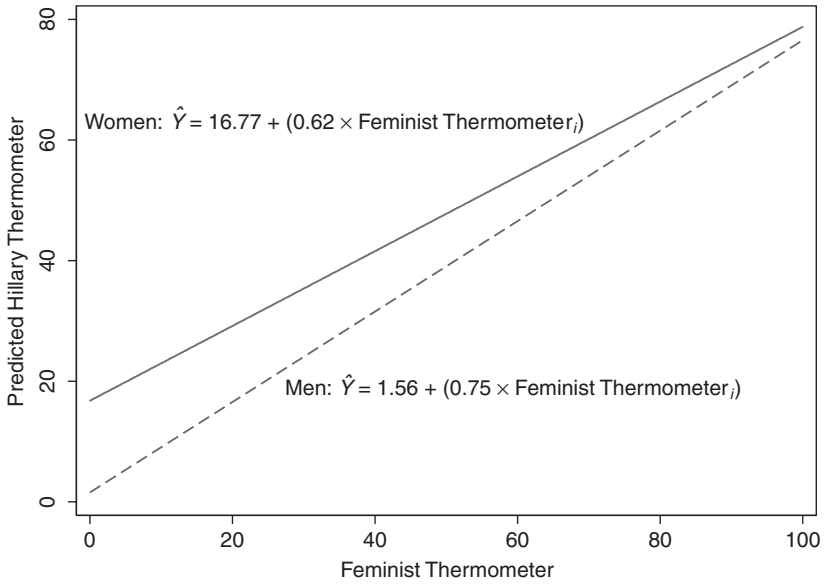

Figure 10.3. Regression lines from the interactive model.

toward Hillary Clinton, this gender gap narrows when we compare those individuals who feel more positively toward the feminist movement.

10.4 OUTLIERS AND INFLUENTIAL CASES IN OLS

In Section 5.10 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.

We should always strive to know our data well. This means looking at individual variables 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 outlier in all senses of this concept in the multivariate world. Nonetheless, we should look for outliers in the single-variable sense before we run our models and make sure that when we identify such cases that they are actual values and not values created by some type of data management mistake.

In the regression setting, individual cases can be outliers 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 outliers for a single case in OLS is often summarized as separate contributions to "influence" in the following formula:

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

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 subsections with a controversial real-world example.

10.4.1 Identifying Influential Cases

One of the most famous cases of outliers/influential cases in political data comes from the 2000 U.S. 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 outliers. 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 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

Table 10.7. Votes for Gore and Buchanan in Florida counties in the 2000 U.S. presidential election

Independent variable	Parameter estimate
Votes for Gore	0.004***
	(0.0005)
Intercept	80.63*
	(46.4)
R^2	.48
n	67

Notes: The dependent variable is the number of votes for Patrick Buchanan.
Standard errors in parentheses.
Two-sided t-tests: ***indicates $p < .01$; **indicates $p < .05$; *indicates $p < .10$.

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 10.7 that there was indeed a statistically significant positive relationship between Gore and Buchanan votes, and that this simple model accounts for 48% 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 10.4 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 this figure it is apparent that Pinellas, Hillsborough, and Orange

[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.

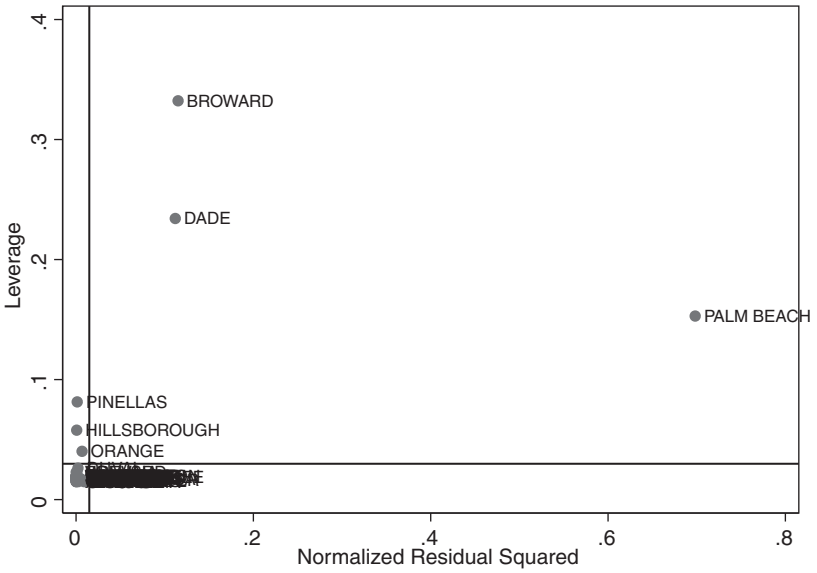

Figure 10.4. Stata lvr2plot for the model presented in Table 10.7.

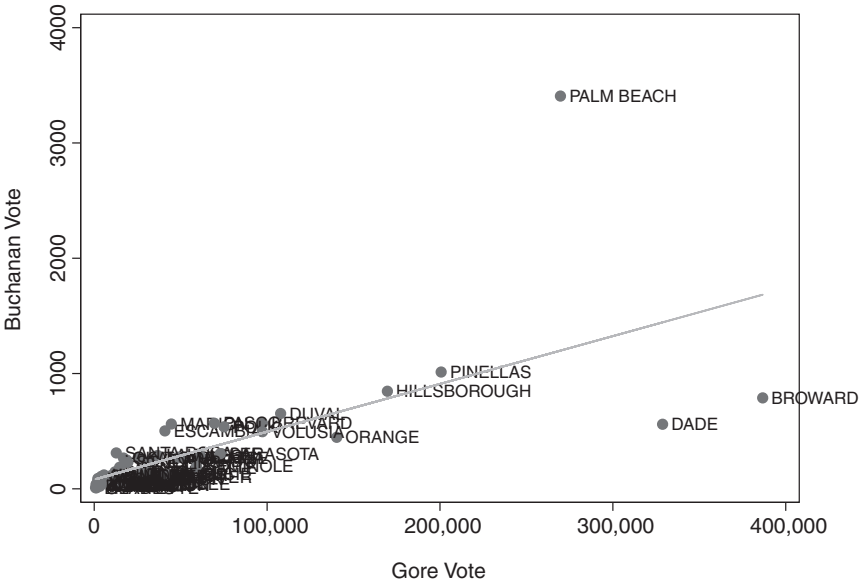

Figure 10.5. OLS line with scatter plot for Florida 2000.

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 10.4 and Table 10.7 from Figure 10.5, in which we plot the OLS regression line

Table 10.8 The five largest (absolute-value) DFBETA scores for β from the model presented in Table 10.7

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

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 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 10.8 displays the five largest absolute values of DFBETA for the slope parameter (β) from the model presented in Table 10.7. Not surprisingly, we see that omitting Palm Beach, Broward, or Dade has the largest impact on our estimate of the slope parameter.

10.4.2 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 10.9 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 10.7. 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.

Table 10.9. Votes for Gore and Buchanan in Florida counties in the 2000 U.S. presidential election

Independent variable	Model 1	Model 2	Model 3	Model 4	Model 5
Gore	0.004***	0.003***	0.003***	0.005***	0.005***
	(0.0005)	(0.0002)	(0.0002)	(0.0003)	(0.0003)
Palm Beach dummy	—	2606.3***	—	2095.5***	—
		(150.4)		(110.6)	
Broward dummy	—	—	—	−1066.0***	—
				(131.5)	
Dade dummy	—	—	—	−1025.6***	—
				(120.6)	
Intercept	80.6*	110.8***	110.8***	59.0***	59.0***
	(46.4)	(19.7)	(19.7)	(13.8)	(13.8)
R^2	.48	.91	.63	.96	.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: ***indicates $p < .01$; **indicates $p < .05$; *indicates $p < .10$.

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] 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 10.9, 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.

[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.

10.5 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 9 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.

10.5.1 How Does Multicollinearity Happen?

Multicollinearity is induced by a small number of degrees of freedom and/or high correlation between independent variables. Figure 10.6 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

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

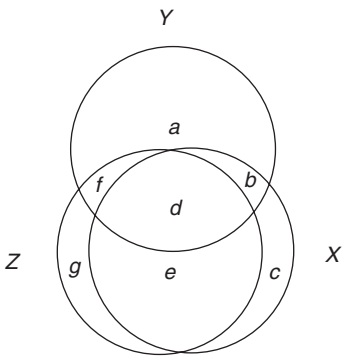

Figure 10.6. Venn diagram with multicollinearity.

Looking at the figure, 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 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 X and Z is also covariation 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 little leverage to understand 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 8 are valid and all of the the minimal mathematical requirements for OLS from Chapters 8 and 9 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 10.6 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.

10.5.2 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 10.6 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 multicollinerity. 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 10.6 with just one of the two independent variables, we would get a statistically significant relationship. But, as we know from

the discussions in Chapter 9, 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 9, 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 ($\frac{1}{VIF}$) 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.

10.5.3 Multicollinearity: A Simulated Example

Thus far we have made a few scattered references to simulation. In this subsection 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

[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.

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 $Y_i = 0.5 + 1X_{1i} + 1X_{2i} + u_i$.

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:

$$\text{Model 1: } Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,$$

$$\text{Model 2: } Y_i = \alpha + \beta_1 X_{1i} + u_i,$$

$$\text{Model 3: } Y_i = \alpha + \beta_2 X_{2i} + u_i.$$

In each of these random draws, we increase the size of our sample starting with 5, then 10, and finally 25 cases. Results from models estimated with each sample of data are displayed in Table 10.10. In the first column of results ($n = 5$), we can see that both slope parameters are 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% 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

Table 10.10. 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	.99	.93	.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	.88	.85	.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	.97	.90	.86

Notes: The dependent variable is $Y_i = .5 + 1X_{1i} + 1X_{2i} + u_i$.
Standard errors in parentheses.
Two-sided t-tests: ***indicates $p < .01$; **indicates $p < .05$; *indicates $p < .10$.

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.

10.5.4 Multicollinearity: A Real-World Example

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

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

Table 10.11. 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: ***indicates $p < .01$; **indicates $p < .05$; *indicates $p < .10$.

Table 10.12. 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	.71	.56	.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: *** indicates $p < .01$; ** indicates $p < .05$; * indicates $p < .10$.

Although we have distinct theories about the causal impact of each independent variable on people's feelings toward Bush, Table 10.11 indicates that some of these independent variables are substantially correlated with each other.

In Table 10.12, 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 10.11, 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 9: 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.

10.5.5 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.

10.6 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

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

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.

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.
- 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 10.4, how many days would you predict that it would take for a government to form if the government was made up