THE FUNDAMENTALS OF

Political Science Research

Paul M. Kellstedt
Texas A&M University

Guy D. Whitten
Texas A&M University

CAMBRIDGE
UNIVERSITY PRESS

11 Multiple Regression Models II: Crucial Extensions

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, dummy dependent variables, influential cases, multicollinearity, and models of time-series data.

11.1 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 have hit 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. Our third topic with dummy variables involves the interpretation of models in which our dependent variable is a dummy variable. 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. Finally, we conclude with a discussion of a series of problems specific to the analysis of time-series data.

11.2 BEING SMART WITH DUMMY INDEPENDENT VARIABLES IN OLS

In Chapter 6 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. We then consider more complicated examples in which we have an independent variable that is categorical with more than two categories.

11.2.1 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 a being a left-wing feminist. Given this, we theorize that there may have been a causal relationship between respondents' family incomes and their thermometer rating of Ms. Clinton – with wealthier individuals, holding all else constant, liking her less – as well as a relationship between respondents' 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              Number of obs =    1542
-------------+------------------------------           F(  2,  1539) =   49.17
       Model |  80916.663      2  40458.3315            Prob > F      =  0.0000
    Residual |  1266234.71   1539  822.764595           R-Squared     =  0.0601
-------------+------------------------------           Adj R-Squared =  0.0588
       Total |  1347151.37   1541  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    -1.071948   -.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 11.1. 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 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

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

[1] In this survey, respondents' 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)
n	1542	1542
R^2	.06	.06

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 of one 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 is 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 $Female_i = 0$ and $Male_i = 1$, the predicted values are calculated as follows:

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

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:

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

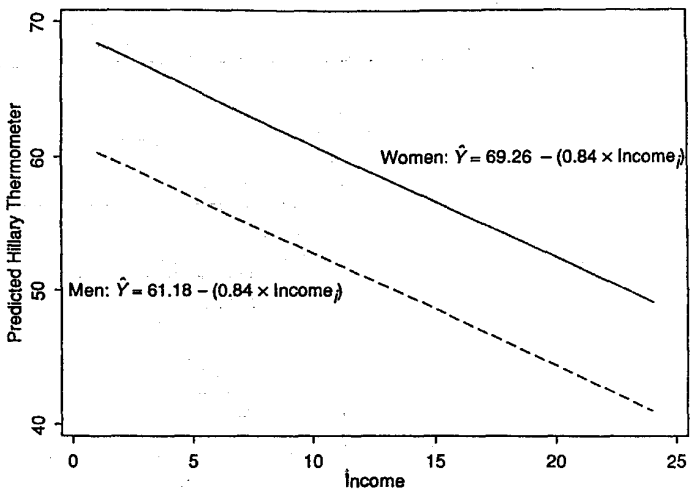

Figure 11.2. Regression lines from the interactive model.

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 the two lines.

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 respondents' 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 this variable. So running a regression model with the data as is would be meaningless. But beware: *Your statistics package does not know that this*

Table 11.2. 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

is a categorical variable. It will be more than happy to run the regression 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

[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)
n	1542	1542	1542	1542	1542
R^2	.06	.06	.06	.06	.06

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 $p < .10$.

as the estimated effect of being in the 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 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 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.

11.3 TESTING INTERACTIVE HYPOTHESES WITH DUMMY VARIABLES

All of the OLS models that we have examined so far have been 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 respondents' thermometer scores for Hillary Clinton. We begin with an additive model with the following specification:

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

In this model we are testing the theory that respondents' 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:

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

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

$$\text{for Men (Female} = 0): \text{Hillary Thermometer}_i = \alpha + \beta_1 \text{Women's Movement Thermometer}_i + u_i;$$

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

Table 11.4. 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)
n	1466	1466
R^2	.27	.27

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 11.4 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 11.4 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.

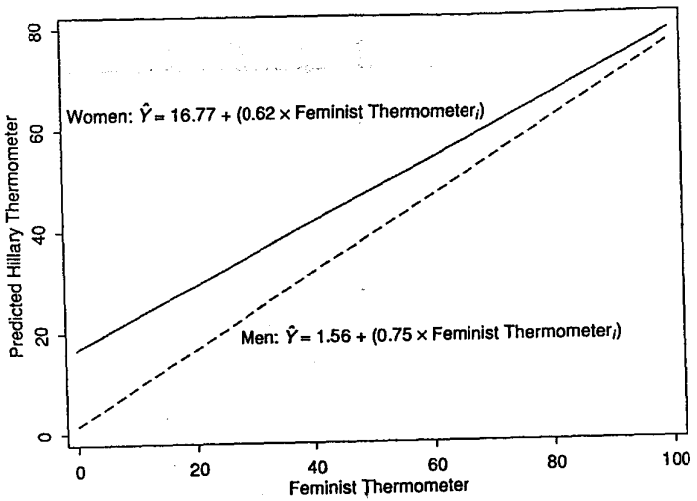

Figure 11.3. Regression lines from the interactive model.

11.4 DUMMY DEPENDENT VARIABLES

Thus far, our discussion of dummy variables has been limited to situations in which the variable in question is one of the independent variables in our model. The obstacles in those models are relatively straightforward. Things get a bit more complicated, however, when our dependent variable is a dummy variable.

Certainly, many of the dependent variables of theoretical interest to political scientists are not continuous. Very often, this means that we need to move to a statistical model other than OLS if we want to get reasonable estimates for our hypothesis testing. One exception to this is the linear probability model (LPM). The LPM is an OLS model in which the dependent variable is a dummy variable. It is called a "probability" model because we can interpret the $\hat{Y}$ values as "predicted probabilities." But, as we will see, it is not without problems. Because of these problems, most political scientists do not use the LPM. We provide a brief discussion of the popular alternatives to the LPM and then conclude this section with a discussion of goodness-of-fit measures when the dependent variable is a dummy variable.

11.4.1 The Linear Probability Model

As an example of a dummy dependent variable, we use the choice that most U.S. voters in the 2004 presidential election made between voting for the incumbent George W. Bush and his Democratic challenger John

Table 11.5. The effects of partisanship and performance evaluations on votes for Bush in 2004

Independent variable	Parameter estimate
Party Identification	0.09***
	(0.01)
Evaluation: War on Terror	0.08***
	(0.01)
Evaluation: Health of the Economy	0.08***
	(0.01)
Intercept	0.60***
	(0.01)
n	780
R^2	.73

Notes: The dependent variable is equal to one if the respondent voted for Bush and equal to zero if they voted for Kerry. Standard errors in parentheses. Two-sided t-tests: ***indicates $p < .01$; **indicates $p < .05$; *indicates $p < .10$.

Kerry.[4] Our dependent variable, which we will call "Bush," is equal to one for respondents who reported voting for Bush and equal to zero for respondents who reported voting for Kerry. For our model we theorize that the decision to vote for Bush or Kerry is a function of an individual's partisan identification (ranging from −3 for strong Democrats, to 0 for independents, to +3 for strong Republican identifiers) and their evaluation of the job that Bush did in handling the war on terror and the health of economy (both of these evaluations range from +2 for "approve strongly" to −2 for "disapprove strongly"). The formula for this model is:

$$\text{Bush}_i = \alpha + \beta_1 \text{Party ID}_i + \beta_2 \text{War Evaluation}_i + \beta_3 \text{Economic Evaluation}_i + u_i.$$

Table 11.5 presents the OLS results from this model. We can see from the table that all of the parameter estimates are statistically significant in the expected (positive) direction. Not surprisingly, we see that people

[4] There was only a handful of respondents to the NES who refused to reveal their vote to the interviewers or voted for a different candidate. But there were a large number of respondents who reported that they did not vote. By excluding all of these categories, we are defining the population about which we want to make inferences as those who voted for Kerry or Bush. Including respondents who voted for other candidates, refused to report their vote, or those who did not vote would amount to changing from a dichotomous categorical dependent variable to a multichotomous categorical dependent variable. The types of models used for this type of dependent variable are substantially more complicated.

who identified with the Republican Party and who had more approving evaluations of the president's handling of the war and the economy were more likely to vote for him. This model performs pretty well overall, with an R^2 statistic equal to .73.

To examine how the interpretation of this model is different from that of a regular OLS model, let's calculate some individual $\hat{Y}$ values. We know from Table 11.5 that the formula for $\hat{Y}$ is

$$\hat{Y}_i = 0.6 + 0.09 \times \text{Party ID}_i + 0.08 \times \text{War Evaluation}_i + 0.08 \times \text{Economic Evaluation}_i.$$

For a respondent who reported being a pure independent (Party ID = 0) with a somewhat approving evaluation of Bush's handling of the war on terror (War Evaluation = 1) and a somewhat disapproving evaluation of Bush's handling of the health of the economy (Economic Evaluation = −1), we would calculate $\hat{Y}_i$ as follows:

$$\hat{Y}_i = 0.6 + (0.09 \times 0) + (0.08 \times 1) + (0.08 \times -1) = 0.6.$$

One logical way to interpret this predicted value is to think of it as a predicted probability that the dummy dependent variable is equal to one. Using the example for which we just calculated $\hat{Y}_i$, we would predict that such an individual would have a 0.6 probability (or 60% chance) of voting for Bush in 2004. As you can imagine, if we change the values of our three independent variables around, the predicted probability of the individual voting for Bush changes correspondingly. This means that the LPM is a special case of OLS for which we can think of the predicted values of the dependent variable as predicted probabilities. From here on, we represent predicted probabilities for a particular case as "$\hat{P}_i$" or "$\hat{P}(Y_i = 1)$" and we can summarize this special property of the LPM as $\hat{P}_i = \hat{P}(Y_i = 1) = \hat{Y}_i$.

One of the problems with the LPM comes when we arrive at extreme values of the predicted probabilities. Consider, for instance, a respondent who reported being a strong Republican (Party ID = 3) with a strongly approving evaluation of Bush's handling of the war on terror (War Evaluation = 2) and a somewhat strongly approving evaluation of Bush's handling of the health of the economy (Economic Evaluation = 2). For this individual, we would calculate $\hat{P}_i$ as follows:

$$\hat{P}_i = \hat{Y}_i = 0.6 + (0.09 \times 3) + (0.08 \times 2) + (0.08 \times 2) = 1.19.$$

This means that we would predict that such an individual would have a 119% chance of voting for Bush in 2004. Such a predicted probability is, of course, nonsensical because probabilities cannot be smaller than zero or greater than one. So, one of the problems with the LPM is that it can

produce such values. In the greater scheme of things, though, this problem is not so severe, as we can make sensible interpretations of predicted values higher than one or lower than zero – these are cases for which we are very confident that probability is close to one (for $\hat{P}_i > 1$) or close to zero (for $\hat{P}_i < 0$).

To the extent that the LPM has potentially more serious problems, they come in two forms – heteroscedasticity and functional form. We discussed heteroscedasticity in Chapter 9 when we noted that any time that we estimate an OLS model we assume that there is homoscedasticity (or equal error variance). We can see that this assumption is particularly problematic with the LPM because the values of the dependent variable are all equal to zero or one, but the $\hat{Y}$ or predicted values range anywhere between zero and one (or even beyond these values). This means that the errors (or residual values) will tend to be largest for cases for which the predicted value is close to .5. Any nonuniform pattern of model error variance such as this is called heteroscedasticity, which means that the estimated standard errors may be too high or too low. We care about this because standard errors that are too high or too low will have bad effects on our hypothesis testing and thus ultimately on our conclusions about causal relationships.

The problem of functional form is related to the assumption of parametric linearity that we also discussed in Chapter 9. In the context of the LPM, this assumption amounts to saying that the impact of a one-unit change in an independent variable X is equal to the corresponding parameter estimate $\hat{\beta}$ regardless of the value of X or any other independent variable. This assumption may be particularly problematic for LPMs because the effect of a change in an independent variable may be greater for cases that would otherwise be at 0.5 than for those cases for which the predicted probability would otherwise be close to zero or one. Obviously the extent of both of these problems will vary across different models.

For these reasons, the typical political science solution to having a dummy dependent variable is to avoid using the LPM. Most applications that you will come across in political science research will use a binomial logit (BNL) or binomial probit (BNP) model instead of the LPM for models in which the dependent variable is a dummy variable. BNL and BNP models are similar to regression models in many ways, but they involve an additional step in interpreting them. In the next subsection we provide a brief overview of these types of models.

11.4.2 Binomial Logit and Binomial Probit

In cases in which their dependent variable is dichotomous, most political scientists use a BNL or a BNP model instead of a LPM. In this subsection we

provide a brief introduction to these two models, using the same example that we used for our LPM in the previous subsection. To understand these models, let's first rewrite our LPM from our preceding example in terms of a probability statement:

$$P_i = P(Y_i = 1) = \alpha + \beta_1 \times \text{Party ID}_i + \beta_2 \times \text{War Evaluation}_i + \beta_3 \times \text{Economic Evaluation}_i + u_i.$$

This is just a way of expressing the probability part of the LPM in a formula in which "$P(Y_i = 1)$" translates to "the probability that Y_i is equal to one," which in the case of our running example is the probability that the individual cast a vote for Bush. We then further collapse this to

$$P_i = P(Y_i = 1) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i,$$

and yet further to

$$P_i = P(Y_i = 1) = X_i\beta + u_i,$$

where we define $X_i\beta$ as the systematic component of Y such that $X_i\beta = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$.[5] The term u_i continues to represent the stochastic or random component of Y. So if we think about our predicted probability for a given case, we can write this as

$$\hat{Y}_i = \hat{P}_i = \hat{P}(Y_i = 1) = X_i\hat{\beta} = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}.$$

A BNL model with the same variables would be written as

$$P_i = P(Y_i = 1) = \Lambda(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i) = \Lambda(X_i\beta + u_i).$$

The predicted probabilities from this model would be written as

$$\hat{P}_i = \hat{P}(Y_i = 1) = \Lambda(\hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} = \Lambda(X_i\hat{\beta}).$$

A BNP with the same variables would be written as

$$P_i = P(Y_i = 1) = \Phi(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i) = \Phi(X_i\beta + u_i).$$

The predicted probabilities from this model would be written as

$$\hat{P}_i = \hat{P}(Y_i = 1) = \Phi(\hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} = \Phi(X_i\hat{\beta}).$$

The difference between the BNL model and the LPM is the Λ, and the difference between the BNP model and the LPM is the Φ. Λ and Φ are known as **link functions**. A link function *links* the linear component of a

[5] This shorthand comes from matrix algebra. Although matrix algebra is a very useful tool in statistics, it is not needed to master the material in this text.

Table 11.6. The effects of partisanship and performance evaluations on votes for Bush in 2004: Three different types of models

Variable	LPM	BNL	BNP
Party Identification	0.09***	0.82***	0.45***
	(0.01)	(0.09)	(0.04)
Evaluation: War on Terror	0.08***	0.60***	0.32***
	(0.01)	(0.09)	(0.05)
Evaluation: Health of the Economy	0.08***	0.59***	0.32***
	(0.01)	(0.10)	(0.06)
Intercept	0.60***	1.11***	0.58***
	(0.01)	(0.20)	(0.10)

Notes: The dependent variable is equal to one if the respondent voted for Bush and equal to zero if they voted for Kerry. Standard errors in parentheses.
Two-sided significance tests: ***indicates $p < .01$; **indicates $p < .05$; *indicates $p < .10$.

logit or probit model, $X_i\hat{\beta}$, to the quantity in which we are interested, the predicted probability that the dummy dependent variable equals one $\hat{P}(Y_i = 1)$ or $\hat{P}_i$. A major result of using these link functions is that the relationship between our independent and dependent variables is no longer assumed to be linear. In the case of a logit model, the link function, abbreviated as Λ, uses the cumulative logistic distribution function (and thus the name "logit") to link the linear component to the probability that $Y_i = 1$. In the case of the probit function, the link function abbreviated as Φ uses the cumulative normal distribution function to link the linear component to the predicted probability that $Y_i = 1$. Appendices C (for the BNL) and D (for the BNP) provide tables for converting $X_i\hat{\beta}$ values into predicted probabilities.

The best way to understand how the LPM, BNL, and BNP work similarly to and differently from each other is to look at them all with the same model and data. An example of this is presented in Table 11.6. From this table it is apparent that across the three models the parameter estimate for each independent variable has the same sign and significance level. But it is also apparent that the magnitude of these parameter estimates is different across the three models. This is mainly due to the difference of link functions. To better illustrate the differences between the three models presented in Table 11.6, we plotted the predicted probabilities from them in Figure 11.4. These predicted probabilities are for an individual who strongly approved of the Bush administration's handling of the war on terror but who strongly disapproved of the Bush administration's handling

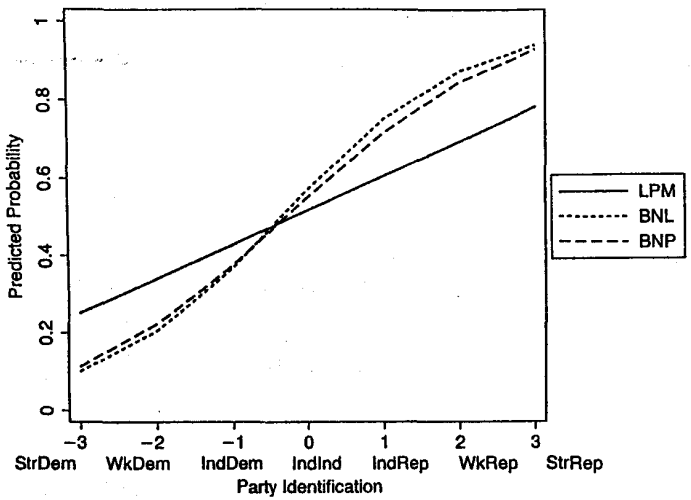

Figure 11.4. Three different models of Bush vote.

of the economy.[6] The horizontal axis in this figure is this individual's party identification ranging from strong Democratic Party identifiers on the left end to strong Republican Party identifiers on the right end. The vertical axis is the predicted probability of voting for Bush. We can see from this figure that the three models make very similar predictions. The main differences come as we move away from a predicted probability of 0.5.

The LPM line has, by definition, a constant slope across the entire range of X. The BNL and BNP lines of predicted probabilities change their slope such that they slope more and more gently as we move farther from predicted probabilities of 0.5. The differences between the BNL and BNP lines are trivial. This means that the effect of a movement in Party Identification on the predicted probability is constant for the LPM. But for the BNL and BNP, the effect of a movement in Party Identification *depends* on the value of the other variables in the model. It is important to realize that the differences between the LPM and the other two types of models are by construction instead of some novel finding. In other words, our choice of model determines the shape of our predicted probability line.

[6] These were the modal answers to the two evaluative questions that were included in the model presented in Table 11.6. It is fairly common practice to illustrate the estimated impact of a variable of interest from this type of model by holding all other variables constant at their mean or modal values and then varying that one variable to see how the predicted probabilities change.

Table 11.7. Classification table from LPM of the effects of partisanship and performance evaluations on votes for Bush in 2004

	Model-based expectations	
Actual Vote	Bush	Kerry
Bush	361	36
Kerry	28	355

Notes: Cell entries are the number of cases. Predictions are based on a cutoff of $\hat{Y} > 0.5$

11.4.3 Goodness-of-Fit with Dummy Dependent Variables

Although we can calculate an R^2 statistic when we estimate a linear probability model, R^2 doesn't quite capture what we are doing when want to assess the fit of such a model. What we are trying to assess is the ability of our model to separate our cases into those in which $Y = 1$ and those in which $Y = 0$. So it is helpful to think about this in terms of a 2×2 table of model-based expectations and actual values. To figure out the model's expected values, we need to choose a cutoff point at which we interpret the model as predicting that $Y = 1$. An obvious value to use for this cutoff point is $\hat{Y} > 0.5$. Table 11.7 shows the results of this in what we call a classification table. Classification tables compare model-based expectations with actual values of the dependent variable.

In this table, we can see the differences between the LPM's predictions and the actual votes reported by survey respondents to the 2004 NES. One fairly straightforward measure of the fit of this model is to look at the percentage of cases that were correctly classified through use the model. So if we add up the cases correctly classified and divide by the total number of cases we get

$$\text{correctly classified } LPM_{0.5} = \frac{361 + 355}{780} = \frac{716}{780} = 0.918$$

So our LPM managed to correctly classify 0.918 or 91.8% of the respondents and to erroneously classify the remaining 0.082 or 8.2%.

Although this seems like a pretty high classification rate, we don't really know what we should be comparing it with. One option is to compare our model's classification rate with the classification rate for a naive model (NM) that predicts that all cases will be in the modal category. In this case, the NM would predict that all respondents voted for Bush. So, if we calculate the correctly classified for the NM,

$$\text{correctly classified NM} = \frac{361 + 36}{780} = \frac{397}{780} = 0.509$$

This means that the NM correctly classified 0.509 or 50.9% of the respondents and erroneously classified the remaining 0.491 or 49.1%.

Turning now to the business of comparing the performance of our model with that of the NM, we can calculate the proportionate reduction of error when we move from the NM to our LPM with party identification and two performance evaluations as independent variables. The percentage erroneously classified in the naive model was 49.1 and the percentage erroneously classified in our LPM was 8.2. So we have reduced the error proportion by $49.1 - 8.2 = 40.9$. If we now divide this by the total error percentage of the naive model, we get $\frac{40.9}{49.1} = 0.833$. This means that we have a proportionate reduction of error equal to 0.833. Another way of saying this is that when we moved from the NM to our LPM we reduced the classification errors by 83.3%.

11.5 OUTLIERS AND INFLUENTIAL CASES IN OLS

In Section 6.4 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.

11.5.1 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 votes cast for Gore and those cast for Buchanan across the Florida counties.[7] 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

[7] 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.

Table 11.8. 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)
n	67
R^2	.48

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 $p < .10$.

was the case, we should see these counties appearing as outliers after we estimate our model.

We can see from Table 11.8 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 in the outlier/influence of particular cases. Figure 11.5 presents a Stata lvr2plot (short for

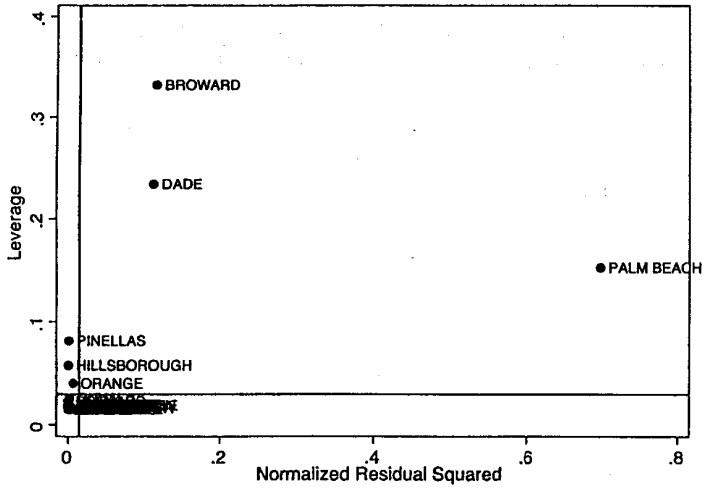

Figure 11.5. Stata lvr2plot for the model presented in Table 11.8.

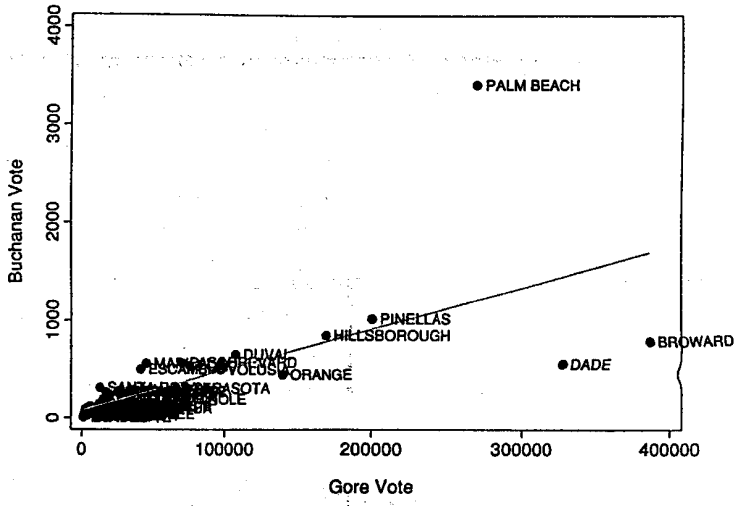

Figure 11.6. OLS line with scatter plot for Florida 2000.

"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 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 11.5 and Table 11.8 from Figure 11.6, 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 substantial outliers and thus 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

Table 11.9. The five largest (absolute-value) DFBETA scores for β from the model presented in Table 11.8

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

estimate without each case divided by the standard error of the original parameter estimate. Table 11.9 displays the five largest absolute values of DFBETA for the slope parameter (β) from the model presented in Table 11.8. Not surprisingly, we see that omitting Palm Beach, Broward, or Dade has the largest impact on our estimate of the slope parameter. By any measure, these cases exerted considerable influence on our model.

11.5.2 Dealing With Influential Cases

Now that we have discussed the identification of particularly influential/outlier 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 11.10 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 11.8. 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.[8] In Model 4 and Model 5,

[8] 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 11.10. 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)
n	67	67	66	67	64
R^2	.48	.91	.63	.96	.82

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 $p < .10$.

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 11.10, 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.

11.6 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

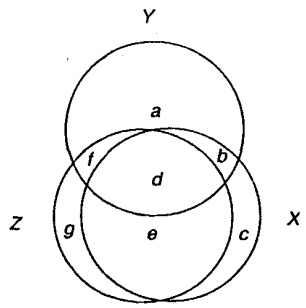

Figure 11.7. Venn diagram with multicollinearity.

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 less-than-perfect multicollinearity. As a result, when people refer to multicollinearity, they almost always mean "less-than-perfect 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.

11.6.1 How Does Multicollinearity Happen?

Multicollinearity is induced by a small number of degrees of freedom and/or high correlation between independent variables. Figure 11.7 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 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 9 are valid and all of the 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 11.7 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 estimate accurate standard errors.

11.6.2 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 11.7 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 11.7 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.[9] 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.

11.6.3 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 parameters for a regression model and draw sample data from this population, we gain insights into the ways in which statistical models work.

[9] 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.

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 11.11. 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 a sample of 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

Table 11.11. 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 $p < .10$.

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.

11.6.4 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

Table 11.12. 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 $p < .10$.

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 peoples' feelings toward Bush, Table 11.12 indicates that some of these independent variables are substantially correlated with each other.

In Table 11.13, 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.[10] 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 peoples' 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 11.12, 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

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

Table 11.13. 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)
n	20	74	821
R^2	.71	.56	.57

Notes: The dependent variable is the 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 $p < .10$.

for the effects of Ideology, Party ID, and Education, the effect of income on peoples' feelings toward George W. Bush goes away.

11.6.5 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 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.

11.7 BEING CAREFUL WITH TIME SERIES

In recent years there has been a massive proliferation of valuable time-series data in political science. Although this growth has led to exciting new research opportunities, it has also been the source of a fair amount of controversy. Swirling at the center of this controversy is the danger of spurious regressions that are due to trends in time-series data. As we will see, a failure to recognize this problem can lead to mistakes about inferring causality. In the remainder of this section we first introduce time-series notation, discuss the problems of spurious regressions, and then discuss the trade-offs involved with two possible solutions: the lagged dependent variable and the differenced dependent variable.

11.7.1 Time-Series Notation

In Chapter 4 we introduced the concept of a time-series observational study. Although we have seen some time-series data (such as the Ray Fair data set used in Chapters 8–10), we have not been using the mathematical notation specific to time- series data. Instead, we have been using a generic notation in which the subscript i represents an individual case. In time-series notation, individual cases are represented with the subscript t, and the numeric value of t represents the temporal order in which the cases occurred, and this ordering is very likely to matter.[11] Consider the following OLS population model written in the notation that we have worked with thus far:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i.$$

If the data of interest were time-series data, we would rewrite this model as

$$Y_t = \alpha + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t.$$

In most political science applications, time-series data occur at regular intervals. Common intervals for political science data are weeks, months, quarters, and years. In fact, these time intervals are important enough that they are usually front-and-center in the description of a data set. For instance, the data presented in Figure 2.1 would be described as a "monthly time series of presidential popularity."

Using this notation, we talk about the observations in the order in which they came. As such, it is often useful to talk about values of variables relative to their lagged values or lead values. Both lagged and lead values are expressions of values relative to a current time, which we call time t. A

[11] In cross-sectional data sets, it is almost always the case that the ordering of the cases is irrelevant to the analyses being conducted.

lagged value of a variable is the value of the variable from a previous time period. For instance, a lagged value from one period previous to the current time is referenced as being from time $t-1$. A lead value of a variable is the value of the variable from a future time period. For instance, a lead value from one period into the future from the current time is referenced as being from time $t+1$. Note that we would not want to specify a model with a leading value for an independent variable because this would amount to a theory that the future value of the independent variable exerted a causal influence on the past.

11.7.2 Memory and Lags in Time-Series Analysis

You might be wondering what, aside from changing a subscript from an i to a t, is so different about time-series modeling. We would like to bring special attention to one particular feature of time-series analysis that sets it apart from modeling cross-sectional data.

Consider the following simple model of presidential popularity, and assume that the data are in monthly form:

$$\text{Popularity}_t = \alpha + \beta_1 \text{Economy}_t + \beta_2 \text{Peace}_t + u_t,$$

where Economy and Peace refer to some measures of the health of the national economy and international peace, respectively. Now look at what the model assumes, quite explicitly. A president's popularity in any given month t is a function of that month's economy and that month's level of international peace (plus some random error term), *and nothing else, at any points in time*. What about last month's economic shocks, or the war that ended three months ago? They are nowhere to be found in this equation, which means quite literally that they can have no effect on a president's popularity ratings in this month. Every month – according to this model – the public starts from scratch evaluating the president, as if to say, on the first of the month: "Okay, let's just forget about last month. Instead, let's check this month's economic data, and also this month's international conflicts, and render a verdict on whether the president is doing a good job or not." There is no memory from month to month whatsoever. Every independent variable has an immediate impact, and that impact lasts exactly one month, after which the effect immediately dies out entirely.

This is preposterous, of course. The public does not erase its collective memory every month. Shifts in independent variables from many months in the past can have lingering effects into current evaluations of the president. In most cases, we imagine that the effects of shifts in independent variables eventually die out over a period of time, as new events become more salient

in the minds of the public, and, indeed, some collective "forgetting" occurs. But surely this does not happen in a single month.

And let's be clear what the problems are with a model like the preceding simple model of approval. If we are convinced that at least some past values of the economy still have effects today, and if at least some past values of international peace still have effects today, but we instead estimate only the contemporary effects (from period t), then we have committed omitted-variables bias – which, as we have emphasized over the last two chapters, is one of the most serious mistakes a social scientist can make. Failing to account for how past values of our independent variables might affect current values of our dependent variable is a serious issue in time-series observational studies, and nothing quite like this issue exists in the cross-sectional world. In time-series analysis, even if we know that Y is caused by X and Z, we still have to worry about how many past lags of X and Z might affect Y.

The clever reader might have a ready response to such a situation: Specify additional lags of our independent variables in our regression models:

$$\begin{aligned}\text{Popularity}_t = \alpha &+ \beta_1 \text{Economy}_t + \beta_2 \text{Economy}_{t-1} + \beta_3 \text{Economy}_{t-2} \\ &+ \beta_4 \text{Economy}_{t-3} + \beta_5 \text{Peace}_t + \beta_6 \text{Peace}_{t-1} + \beta_7 \text{Peace}_{t-2} \\ &+ \beta_8 \text{Peace}_{t-3} + u_t.\end{aligned}$$

This is, indeed, one possible solution to the question of how to incorporate the lingering effects of the past on the present. But the model is getting a bit unwieldy, with lots of parameters to estimate. More important, though, it leaves several questions unanswered:

1. How many lags of the independent variables should we include in our model? We have included lags from period t though $t-3$ in the preceding specification, but how do we know that this is the correct choice? From the outset of the book, we have emphasized that you should have *theoretical* reasons for including variables in your statistical models. But what theory tells with any specificity that we should include 3, 4, or 6 periods' worth of lags of our independent variables in our models?
2. If we do include several lags of all of our independent variables in our models, we will almost surely induce multicollinearity into them. That is, X_t, X_{t-1}, and X_{t-2} are likely to be highly correlated with one another. (Such is the nature of time series.) Those models, then, would have all of the problems associated with high multicollinearity

just identified – in particular, large standard errors and the adverse consequences on hypothesis testing.

Before showing two alternatives to saturating our models with lots of lags of variables, we need to confront a different problem in time-series analysis.

11.7.3 Trends and the Spurious Regression Problem

When discussing presidential popularity data, it's easy to see how a time series might have a "memory" – by which we mean that the current values of a series seem to be highly dependent of its past values.[12] Some series have memories of their pasts that are sufficiently long to induce statistical problems. In particular, we mention one, called the spurious regression problem.[13]

By way of example, consider the following facts: In post–World War II America, golf became an increasingly popular sport. As its popularity grew, perhaps predictably the number of golf courses in America grew to accommodate the demand for places to play. That growth continued steadily into the early 21st century. We can think of the number of golf courses in America as a time series, of course, presumably one on an annual metric. Over the same period of time, divorce rates in America grew and grew. Whereas divorce was formerly an uncommon practice, today it is commonplace in American society. We can think of family structure as a time series, too – in this case, the percentage of households in which a married couple is present.[14]

And both of these time series – likely for different reasons – have long memories. In the case of golf courses, the number of courses in year t obviously depends heavily on the number of courses in the previous year. In the case of divorce rates, the dependence on the past presumably stems from the lingering, multiperiod influence of the social forces that lead to divorce in the first place. Both the number of golf facilities in America and the percentage of families in which a married couple is present are shown

[12] In any time series representing some form of public opinion, the word "memory" is a particularly apt term, though its use applies to all other time series as well.

[13] The problem of spurious regressions was something that economists like John Maynard Keynes worried about long before it had been mathematically demonstrated by Granger and Newbold (1974) in the 1970s. Their main source of concern was the existence of general trends in a variable over time. To be clear, the word "trend" obviously has several popular meanings. In time-series analysis, though, we generally use the word trend to refer to a long-lasting movement in the history of a variable, not a temporary drift in one direction or another.

[14] For the purposes of this illustration, we are obscuring the difference between divorce and unmarried cohabitation.

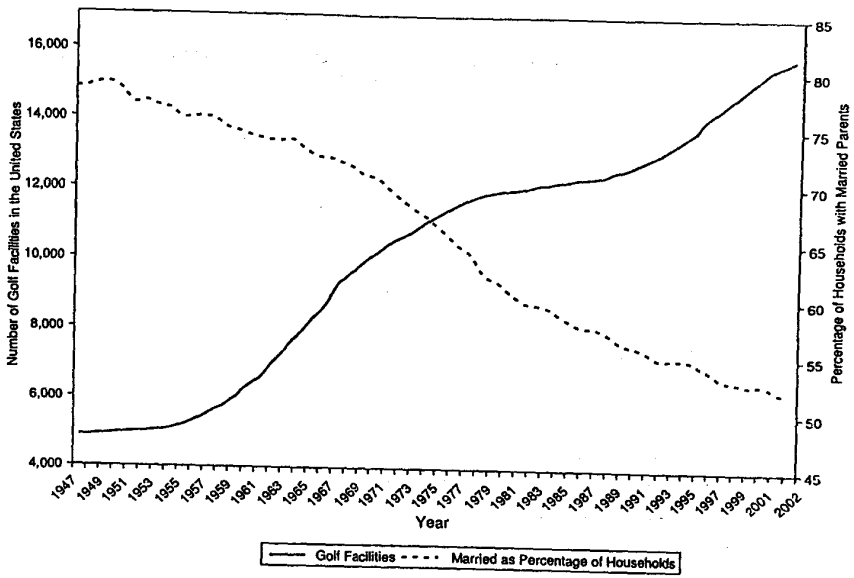

Figure 11.8. The growth of golf and the decline of the American family, 1947–2002.

in Figure 11.8.[15] And it's clear that, consistent with our description, both variables have trends. In the case of golf facilities, that trend is upward; for marriage, the trend is down.

What's the problem here? Any time one time series with a long memory is placed in a regression model with another series that also has a long memory, it can lead to falsely finding evidence of a causal connection between the two variables. This is known as the "spurious regression problem." If we take the demise of marriage as our dependent variable and use golf facilities as our independent variable, we would surely see that these two variables are related, statistically. In substantive terms, we might be tempted to jump to the conclusion that the growth of golf in America has led to the breakdown of the nuclear family. We show the results of that regression in Table 11.14. The dependent variable there is the percentage of households with a married couple, and the independent variable is the number of golf courses (in thousands). The results are exactly as feared. For every thousand golf facilities built in the United States, there are 2.53% fewer families with a married couple present. The R^2 statistic is quite high, suggesting that roughly 93% of the variance in divorce rates is explained by the growth of the golf industry.

[15] The National Golf Foundation kindly gave us the data on golf facilities. Data on family structure are from the Current Population Reports.

Table 11.14. Golf and the decline of the family, 1947–2002

Variable	Coefficient (Std. Err.)
Golf Facilities	−2.53*
	(0.09)
Constant	91.36*
	(1.00)
n	56
R^2	.93

Note: *indicates $p < .05$.

We're quite sure that some of you – presumably nongolfers – are nodding your heads and thinking, "But maybe golf *does* cause divorce rates to rise! Does the phrase 'golf widow' ring a bell?" But here's the problem with trending variables, and why it's such a potentially nasty problem in the social sciences. We could substitute *any* variable with a trend in it and come to the same "conclusion." To prove the point, let's take another example. Instead of examining the growth of golf, let's look at a different kind of growth – economic growth. In post-war America, GDP has grown steadily, with few interruptions in its upward trajectory. Figure 11.9 shows GDP, in annual terms, along with the now-familiar time series of the decline in marriage. Obviously, GDP is a long-memoried series, with a sharp upward trend, in which current values of the series depend extremely heavily on past values.

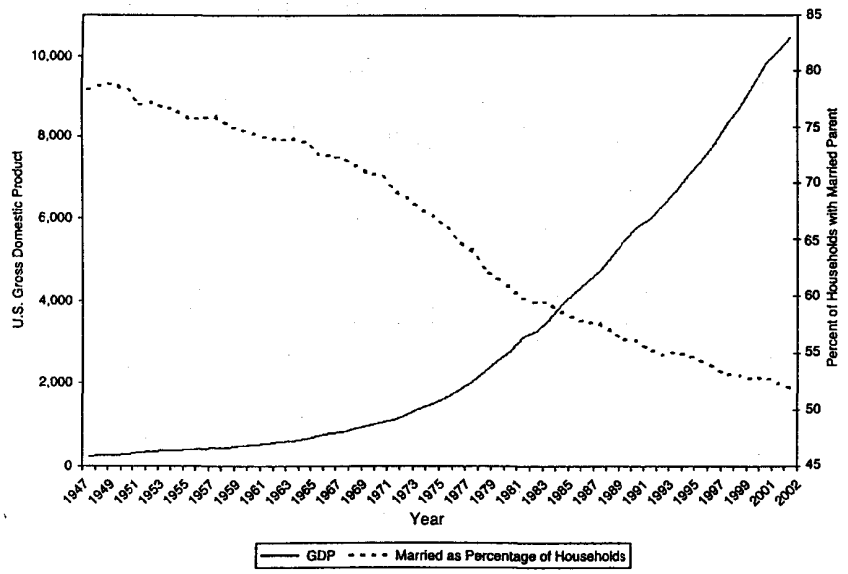

Figure 11.9. The growth of the U.S. economy and the decline of the family, 1947–2002.

Table 11.15. GDP and the decline of the family, 1947–2002

Variable	Coefficient (Std. Err.)
GDP (in trillions)	−2.71*
	(0.16)
Constant	74.00*
	(0.69)
n	56
R^2	.84

Note: *indicates $p < .05$.

The spurious regression problem has some bite here, as well. Using Divorce as our dependent variable and GDP as our independent variable, the regression results in Table 11.15 show a strong, negative, and statistically significant relationship between the two. This is not occurring because higher rates of economic output have led to the destruction of the American family. It is occurring because both variables have trends in them, and a regression involving two variables with trends – even if they are not truly associated – will produce spurious evidence of a relationship.

The two issues just mentioned – how to deal with lagged effects in a time series and whether or not the spurious regression problem is relevant – are tractable ones. Moreover, new solutions to these issues arise as the study of time-series analysis becomes more sophisticated. We subsequently present two potential solutions to both problems.

11.7.4 The Differenced Dependent Variable

One way to avoid the problems of spurious regressions is to use a differenced dependent variable. We calculate a differenced (or, equivalently, "first differenced") variable by subtracting the first lag of the variable (Y_{t-1}) from the current value Y_t. The resulting time series is typically represented as $\Delta Y_t = Y_t - Y_{t-1}$.

In fact, when time series have long memories, taking first differences of both independent and dependent variables can be done. In effect, instead of Y_t representing the *levels* of a variable, ΔY_t represents the period-to-period *changes* in the level of the variable. For many (but not all) variables with such long memories, taking first differences will eliminate the visual pattern of a variable that just seems to keep going up.

Figure 11.10 presents the first differences of the number of golf courses in the United States, as well as the first differences of the U.S. annual divorce rates. You will notice, of course, that the time series in these figures look drastically different from their counterparts in levels from Figure 11.8. In fact, the visual "evidence" of an association between the two variables that appeared in Figure 11.8 has now vanished. The misleading culprit? Trends in both time series.

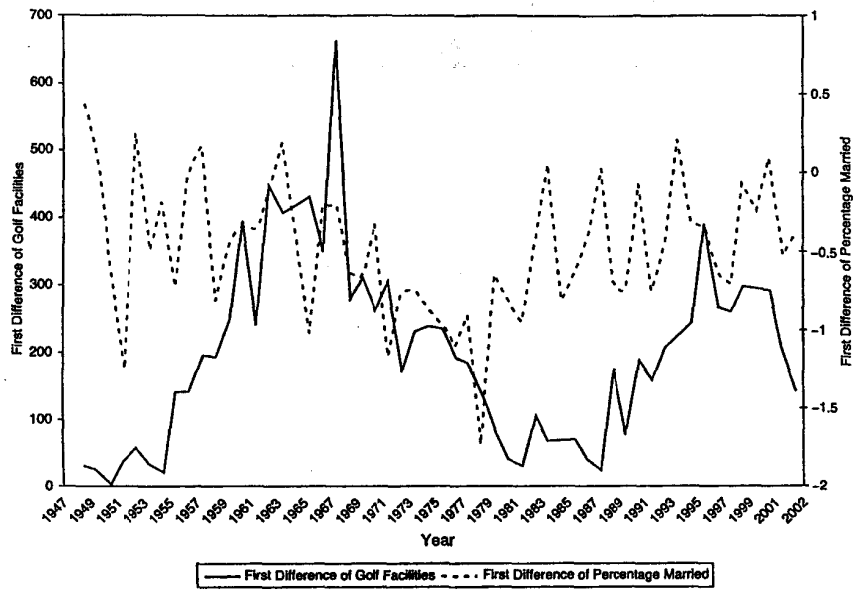

Figure 11.10. First differences of the number of golf courses and percentage of married families, 1947–2002.

Because, in these cases, taking first differences of the series removes the long memories from the series, these transformed time series will not be subject to the spurious regression problem. But we caution against thoughtless differencing of time series. In particular, taking first differences of time series can eliminate some (true) evidence of an association between time series in certain circumstances.

We recommend that, wherever possible, you use theoretical reasons to either difference a time series or to analyze it in levels. In effect, you should ask yourself if your theory about a causal connection between X and Y makes more sense in levels or in first differences. For example, if you are analyzing budgetary data from a government agency, does your theory specify particular things about the sheer amount of agency spending (in which case, you would analyze the data in levels), or does it specify particular things about what causes budgets to shift from year to year (in which case, you would analyze the data in first differences)?

It is also worth noting that taking first differences of your time series does not directly address the issue of the number of lags of independent variables to include in your models. For that, we turn to the lagged-dependent-variable specification.

11.7.5 The Lagged Dependent Variable

Consider for a moment a simple two-variable system with our familiar variables Y and X, except where, to allow for the possibility that previous lags of X might affect current levels of Y, we include a large number of lags of X in our model:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_k X_{t-k} + u_t.$$

This model is known as a **distributed lag model.** Notice the slight shift in notation here, in which we are subscripting our β coefficients by the number of periods that that variable is lagged from the current value; hence, the β for X_t is β_0 (because $t - 0 = t$). Under such a setup, the **cumulative impact** β of X on Y is

$$\beta = \beta_0 + \beta_1 + \beta_2 + \cdots + \beta_k = \sum_{i=0}^{k} \beta_i.$$

It is worth emphasizing that we are interested in that cumulative impact of X on Y, not merely the **instantaneous effect** of X_t on Y_t represented by the coefficient β_0.

But how can we capture the effects of X on Y without estimating such a cumbersome model like the preceding one? We have noted that a model like this would surely suffer from multicollinearity.

If we are willing to assume that the effect of X on Y is greatest initially and decays geometrically each period (eventually, after enough periods, becoming effectively 0), then a few steps of algebra would yield the following model that is mathematically identical to the preceding one.[16] That model looks like

$$Y_t = \lambda Y_{t-1} + \alpha + \beta_0 X_t + v_t.$$

This is known as the **Koyck transformation,** and is commonly referred to as the **lagged-dependent-variable** model, for reasons we hope are obvious. Compare the Koyck transformation with the preceding equivalent distributed lag model. Both have the same dependent variable, Y_t. Both have a variable representing the immediate impact of X_t on Y_t. But whereas the distributed lag model also has a slew of coefficients for variables representing all of the lags of 1 through k of X on Y_t, the lagged-dependent-variable model instead contains a single variable and coefficient, λY_{t-1}. Because, as we said, the two setups are equivalent, then this means that the lagged

[16] We realize that the model does not *look* mathematically identical, but it is. For ease of presentation, we skip the algebra necessary to demonstrate the equivalence.

dependent variable does *not* represent how Y_{t-1} somehow causes Y_t, but instead Y_{t-1} is a stand-in for the cumulative effects of all past lags of X (that is, lags 1 through k) on Y_t. We achieve all of that through estimating a single coefficient instead of a very large number of them.

The coefficient λ, then, represents the ways in which past values of X affect current values of Y, which nicely solves the problem outlined at the start of this section. Normally, the values of λ will range between 0 and 1.[17] You can readily see that if $\lambda = 0$ then there is literally no effect of past values of X on Y_t. Such values are uncommon in practice. As λ gets larger, that indicates that the effects of past lags of X on Y_t persist longer and longer into the future.

In these models, the cumulative effect of X on Y is conveniently described as

$$\beta = \frac{\beta_0}{1 - \lambda}.$$

Examining the formula, we easily see that, when $\lambda = 0$, the denominator is equal to 1 and the cumulative impact is exactly equal to the instantaneous impact. There is no lagged effect at all. When $\lambda = 1$, however, we run into problems; the denominator equals 0, so the quotient is undefined. But as λ approaches 1, you can see that the cumulative effect grows. Thus, as the values of the coefficient on the lagged dependent variable move from 0 toward 1, the cumulative impact of changes in X on Y grows.

This brief foray into time-series analysis obviously just scratches the surface. When reading research that uses time-series techniques, or especially when embarking on your own time-series analysis, it is important to be aware of both the issues of how the effects of shifts in independent variables can persist over several time periods, and also of the potential pitfalls of long-memoried trends.

11.8 WRAPPING UP

Even in its simplest varieties, OLS regression – and especially multiple OLS regression – can be complicated enough. What we've encountered in this chapter shows that there are additional (but not insurmountable!) obstacles to overcome when we consider that some of our theories involve noncontinuous variables and that there are some unique obstacles involving time-series data.

[17] In fact, values close to 1, and especially those greater than 1, indicate that there are problems with the model, most likely related to trends in the data.

Some of these techniques might seem intimidating at first, but we encourage you to press onward. One way to do this is to see how these techniques work in actual examples of political science research. In our final chapter, we examine three pieces of research that attempt to answer compelling theoretical questions.

CONCEPTS INTRODUCED IN THIS CHAPTER

additive models
auxiliary regression model
binomial logit
binomial probit
classification table
cumulative impact
DFBETA score
differenced (or "first differenced") dependent variable
distributed lag model
dummying out
dummy variables
dummy-variable trap
instantaneous effect
interactive models
Koyck transformation
lagged dependent variable
lagged values
lead values
leverage
linear probability model
link functions
micronumerosity
multicollinearity
predicted probability
proportionate reduction of error
reference category
spurious regression problem
variance inflation factor