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

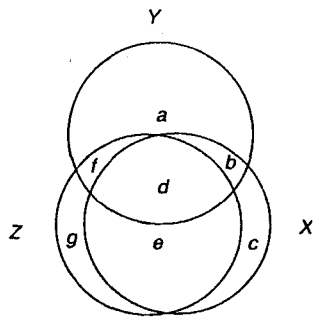

Figure 11.7. Venn diagram with multicollinearity.

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.