7 What Can Be Done About Multicollinearity?

Multicollinearity is something that nearly all users of multiple regression have heard about. Unfortunately, their knowledge of multicollinearity is often limited to two facts:

- It's bad.
- It has something to do with high correlations among the variables.

Beyond those truths, there is an enormous amount of confusion and mythology surrounding multicollinearity. This chapter will set you straight.

7.1. What Is Extreme Multicollinearity?

Multicollinearity comes in two forms: extreme and near-extreme. Extreme multicollinearity means that at least two of the independent variables in a regression equation are perfectly related by a linear function. Suppose you're trying to estimate the model

$$y = A + B_1x_1 + B_2x_2 + B_3x_3 + U.$$

Suppose that in your sample, it happens to be the case that

$$x_1 = 2 + 3x_2.$$

Notice that there is no error term in this equation. Then the correlation between x_1 and x_2 is 1.0 and we have a case of extreme multicollinearity.

The consequence of extreme multicollinearity is simple: It's impossible to get separate estimates for the coefficients B_1 and B_2. If you try to do it anyway, the computer will do one of two things:

- Print an error message but no results.
- Arbitrarily pick either x_1 or x_2, throw that variable out of the model, and estimate the model with the remaining variables. Usually, a warning message is also printed.

Box 7.1 is an example of the output from SPSS when the two independent variables are perfectly correlated.

BOX 7.1. Excluded Variable[b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	X2	a.	.	.	.	.000

a. Predictors in the Model: (Constant), X1
b. Dependent Variable: Time (months)

After reporting the results from the regression on x_1, we are told that x_2 has been excluded from the regression model. The tolerance is a useful statistic that we'll look at a bit later.

Box 7.2 is an example of a warning message produced by the SAS® System.

BOX 7.2

```
NOTE: Model is not full rank. Least-squares
solutions for the parameters are not unique. Some
statistics will be misleading. A reported DF of 0
or B means that the estimate is biased.
  The following parameters have been set to 0,
since the variables are a linear combination of
other variables as shown.

X1 = +2.0000 * INTERCEP +3.0000 * X2
```

The statement "Model is not full rank" is equivalent to saying that there is extreme multicollinearity. In this case, the variable x_2 is a perfect linear function of the variable x_1. The program excluded x_2

from the model because it happened to be the second variable listed on the model specification.

Why does this happen? Remember that multiple regression is trying to separate out the effects of two or more variables, even though they are correlated with each other. To do this, however, there must be some remaining variation on each x variable when the other x variables are held constant. If two variables are perfectly correlated, when you hold one constant, the other must be constant as well. Hence, it's impossible to separate their effects on the dependent variable. (It might be helpful to reread Section 6.5 and think about what would happen if x_1 and x_2 were perfectly correlated.)

Extreme multicollinearity is unusual in the social sciences, but it does happen. When it occurs, it's usually because of some artifact in the way the independent variables are constructed. For example, suppose you ask people to estimate the percentage of information about current events they get from (a) television, (b) radio, (c) print media, and (d) conversations with others. Respondents are told that the percentages must sum to 100. Let x_1 through x_4 be variables containing the four percentages. If you tried to use all four of these as independent variables in a multiple regression, you would have extreme multicollinearity. Because they always add up to 100, any one of them can be expressed as a linear function of the other three. For instance, we can write $x_4 = 100 - x_1 - x_2 - x_3$. This sort of problem often occurs with sets of dummy variables, as I'll explain in the next chapter.

What can be done about extreme multicollinearity? For the variables that are collinear with each other, the answer is nothing. There's simply no way to get distinct coefficient estimates for them. What is essential to remember is that *multicollinearity only affects the coefficient estimates for those variables that are collinear.* This is true for both extreme and near-extreme multicollinearity. If x_1 and x_2 are perfectly correlated but x_3 is uncorrelated with either of them, then there's no difficulty in getting estimates for the effect of x_3 controlling for the other two variables. The way to do it is what many regression programs do automatically: arbitrarily pick either x_1 or x_2 and remove it from the equation. Then the estimated coefficient for x_3 is OK. The estimated coefficient for, say, x_1 represents the combined effect of x_1 and x_2, so you must be cautious in interpreting it.

7.2. What Is Near-Extreme Multicollinearity?

The one good thing about extreme multicollinearity is that you can't miss it. By contrast, the much more common problem of near-extreme multicollinearity is more insidious. It can cause you to make seriously incorrect conclusions without you being aware of the problem.

Near-extreme multicollinearity means simply that there are strong (but not perfect) linear relationships among the independent variables. If the regression model has only two independent variables, near-extreme multicollinearity occurs if the two variables have a correlation that's close to 1 or –1. How close does it have to be? Like everything else in regression, it's a matter of degree. The closer you get to 1 or –1, the greater the associated problems.

It's a little more difficult to describe near-extreme multicollinearity when there are three or more variables that are collinear. Suppose you have three independent variables, x_1, x_2, and x_3. Pick one of them, say x_2, and regress it on x_1 and x_3. If the R^2 from that regression is near 1, then x_2 is collinear with x_1 and x_3. It's worth remembering that multicollinearity has nothing to do with the dependent variable—it's a characteristic of the relationships among the independent variables.

Near-extreme multicollinearity does not prevent calculation of the regression coefficients. It does make it more difficult to reliably estimate the coefficients of those variables that are collinear. This can cause a number of difficulties that are described in Section 7.4. First, however, we'll look at some methods for diagnosing the existence of near-extreme multicollinearity. In the remainder of this chapter, when I use the word "multicollinearity," you can assume I'm referring to the near-extreme case.

7.3. How Can Multicollinearity Be Diagnosed?

The old-fashioned way to check for multicollinearity is to examine the matrix of two-variable correlations among all the independent variables. Most statistical packages can produce such a matrix. If any of the correlations is very high (near 1 or –1), we conclude that multicollinearity is a problem.

The disadvantage of this method is that it's quite possible that *none* of the bivariate correlations may be very high, yet multicol-

linearity could still be serious. Consider the following correlation matrix for four independent variables:

	x_1	x_2	x_3	x_4
x_1	1.00			
x_2	.10	1.00		
x_3	.10	.10	1.00	
x_4	.60	.50	.60	1.00

Examined one by one, none of these correlations is so large as to cause major concern, but if we regress x_4 on the other three x variables, we get an R^2 of .81, high enough to qualify as serious multicollinearity by almost anyone's standards. This is, in fact, a good way of diagnosing multicollinearity: Regress each independent variable on all the other *independent* variables, and look for a high R^2 in any of these regressions. How high is high? Personally, I start to get concerned when any of these R^2s is above .60 or so.

Some regression programs will automatically produce these diagnostic R^2s if you request them. Actually, what they give you is usually something called a *tolerance*, which is 1 minus the R^2 for each independent variable. (Why they subtract the R^2s from 1, I don't know.) You want to watch out for low tolerances. Consistent with my standard for the R^2, I start to worry when any of the tolerances is below .40.

Another equivalent multicollinearity diagnostic that may be reported for each independent variable is something called the *variance inflation factor*. This is just the reciprocal of the tolerance (1/tolerance). I'll explain the reason for this name in the next section. Tolerances below .40 correspond to variance inflation factors above 2.50.

Here are the tolerances and variance inflation factors for the four variables in the correlation matrix above.

	Tolerance	*Variance Inflation Factor*
x_1	0.42	2.38
x_2	0.54	1.85
x_3	0.42	2.38
x_4	0.19	5.27

As we already saw, x_4 has serious problems with its low tolerance and high variance inflation factor (VIF). The other three variables have less serious problems, although they're quite close to my personal

criterion of .40 for the tolerance and 2.50 for the variance inflation factor.

In addition to these variable-by-variable measures of multicollinearity, some programs will give you more comprehensive statistics that help you determine which variables are linearly related to which other variables. These statistics may be useful in some cases, but they can also be more difficult to interpret than the tolerances or variance inflation factors.

7.4. What Are the Consequences of Multicollinearity?

Near-extreme multicollinearity is *not* a violation of any of the assumptions we discussed in Chapter 6. As long as the assumptions in Section 6.3 are satisfied, the least squares estimates are still BLUE (best linear unbiased estimates). So what's the problem? If an independent variable is highly collinear with other variables, the standard error of its coefficient will be large. This fact is captured by the variance inflation factor described in the last section. The square root of the variance inflation factor tells you how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other x variables in the equation. In the previous section, the variance inflation factor for x_4 was 5.27, which has a square root of 2.3. This means that the standard error for the coefficient of x_4 is 2.3 times as large as it would be if x_4 were uncorrelated with the other xs. It follows that the confidence interval around the coefficient will be more than twice as wide as if there were no multicollinearity, and the t statistic will be less than half as large.

So one big problem with multicollinearity is that it's harder to find statistically significant coefficients. Keep in mind that that statement applies only to the variables with high variance inflation factors. The variables with low variance inflation factors are unaffected. For the variables that are collinear, however, the inflation of the standard errors—although an accurate reflection of uncertainty—can produce quite misleading conclusions.

Here's a simple example. Suppose we interview a sample of 150 newlywed women and we ask them how many children they want (x_1) and their number of years of schooling (x_2). One year later, we

again ask them how many children they want (x_3). Five years later, we find out how many children they actually have (y). We then regress y on x_1, x_2, and x_3. The correlation matrix for the four variables is

	x_1	x_2	x_3	y
x_1	1.00			
x_2	–.15	1.00		
x_3	.90	–.20	1.00	
y	.35	–.23	.37	1.00

Notice the correlation of .90 between x_1 and x_3, the two measures of number of desired children.

If we regress y on all three x variables, we get the following standardized coefficients, t statistics, and variance inflation factors.

	Coefficient	*t Statistic*	*VIF*
x_1	0.12	0.67	5.29
x_2	–0.17	–2.15	1.05
x_3	0.23	1.33	5.39

Because the t statistic for schooling (x_2) is above 2 in magnitude, we conclude that it has a significant, negative impact on number of children. However, neither of the two measures of desired number of children has a significant effect on actual number of children.

But look what happens when we delete x_3 from the model:

	Coefficient	*t Statistic*	*VIF*
x_1	0.32	4.21	1.02
x_2	–0.18	–2.37	1.02

The coefficient for x_1 has nearly tripled, and the t statistic has increased to 4.21, which has a p value less than .0001. Similar results occur when x_1 is deleted from the model:

	Coefficient	*t Statistic*	*VIF*
x_2	–0.16	–2.11	1.04
x_3	0.34	4.38	1.04

What does this tell us? Clearly there is an important relationship between desired number of children and actual number of children, but that relationship is obscured when we put two highly correlated

measures of desired number of children in the same model. When we hold constant the desired number of children at time 1, there's so little remaining variation in desired number of children at time 2 that we can't get reliable estimates for that variable. The same thing happens when we control for the time 2 measure. Either of the reduced regressions therefore gives a more accurate picture than the regression that includes all three x variables.

There's something else to be learned from the regression with all three x variables. Notice that in the correlation matrix, the correlation between x_1 and y is .35, and the correlation between x_3 and y is .37. There is not much of a difference. When we put both variables in the same regression, however, the standardized coefficient for x_1 is .12 and the standardized coefficient for x_3 is .23. That is a big difference. This phenomenon is known as the *tipping effect*. The general principle is that when independent variables are highly correlated, small differences in their bivariate relationships with the dependent variable get magnified into large differences in the regression coefficients.

The difference between a correlation of .35 and a correlation of .37 is so small that it could easily arise from tiny changes in the data. Sampling errors, measurement errors, coding errors, missing data—any of these could easily produce a difference of that magnitude. Yet, as we've just seen, such small changes can lead to major differences in the magnitudes of the coefficients. This example points out the instability of regression coefficients under conditions of multicollinearity. In general, multicollinearity makes multiple regression much more sensitive to minor errors or departures from the assumptions of the model. In other words, the coefficients are less *robust*.

One manifestation of these problems is that the coefficients for variables that are collinear are often surprising or counterintuitive. Coefficients that you feel confident ought to be positive turn out to be negative. Standardized coefficients may be greater than 1.0 in magnitude. These are both common symptoms of multicollinearity. Even when such "strange" coefficients are statistically significant, you should be wary about placing too much importance on them.

As with many other problems we've discussed, multicollinearity is only a minor concern for models whose primary goal is prediction. Suppose you estimate a model with an R^2 of .85, which is pretty good for a predictive model. Then you add another independent variable that has a correlation of .90 with one of the variables already in the model. The R^2 will hardly change at all, and neither will the

actual predicted values. The standard errors of the predicted values may increase a bit, so there is some disadvantage of adding redundant variables to prediction models.

7.5. Are There Situations in Which Multicollinearity Is More Likely to Occur?

Certain kinds of data are particularly prone to multicollinearity. One is time-series data. In the classic time-series design, there is a single case that is observed at many points in time. For example, the case could be the U.S. economy measured at quarterly intervals over a period of 40 years. The variables might include GNP, unemployment rate, inflation rate, and so on. With time-series data, there is a tendency for variables to be highly correlated. One reason is that there are consistent long-term trends in many different variables.

Another common over-time design is the panel study. In this design, many cases are observed at two or more points in time. For example, a sample of 1,000 people may be interviewed annually over a period of 5 years. Although this kind of data is less prone to multicollinearity than the classic time-series design, most variables tend to be highly correlated with themselves at earlier or later points in time. We saw this in the example in Section 7.4, where desired number of children at time 1 was highly correlated with desired number of children at time 2.

A rather different kind of data is variously known as *aggregated, group-level,* or *ecological* data. With data of this sort, the units of analysis are groups of individuals, and the variables are summary measures of individual characteristics. Here's a real example. In a study designed to explain homicide rates among Black males (Phillips, 1997), data were collected for 222 metropolitan areas in the United States. In each city, the following variables were measured in 1990:

BHOM: Black homicide rate

BFEMHEAD: Percentage of Black households headed by females

BLFP: Percentage of Blacks in the labor force

BPOVERTY: Percentage of Black households below the poverty line

INEQ: Inequality of Black incomes

TABLE 7.1 Correlations Among Five Characteristics of U.S. Metropolitan Areas

Variable	*BHOM*	*BFEMHEAD*	*BLFP*	*BPOVERTY*	*INEQ*
BHOM	1.00				
BFEMHEAD	0.17	1.00			
BLFP	−0.20	−0.45	1.00		
BPOVERTY	0.20	0.56	−0.70	1.00	
INEQ	0.30	0.59	−0.51	0.66	1.00

Table 7.1 gives the correlations among these variables. Although the correlations between the homicide rate and the other four variables are only moderately high, the remaining correlations are all quite substantial.

Why are such high correlations common with aggregate data? Although there's no definitive answer to this question, the most plausible explanation is that random variation among individual people tends to average out when they are combined into groups. This phenomenon is a double-edged sword, however. You can often get very high R^2 for aggregate data, but it may be quite difficult to get reliable estimates of the coefficients because of multicollinearity.

Table 7.2 shows the results of regressing the homicide rate on the other four variables. Only one of the variables, Inequality of Black incomes, is statistically significant: Greater inequality is associated with a higher homicide rate. Note that the tolerances for all the variables are on the low side, with one of them (BPOVERTY) below .40. This should raise questions as to how much confidence to put in the results.

TABLE 7.2 Regression of Black Male Homicide Rate on Characteristics of Metropolitan Areas

Variable	*Coefficient*	*t*	*p*	*Tolerance*	*Variance Inflation Factor*
BFEMHEAD	−0.02	−0.15	0.88	.60	1.67
BLFP	−0.19	−1.05	0.29	.51	1.97
BPOVERTY	−0.08	−0.42	0.68	.37	2.67
INEQ	1.12	−3.14	0.002	.50	2.02

7.6. Are There Any Solutions?

There are a lot of options for dealing with near-extreme multicollinearity, but all of them are flawed in one way or another. The fundamental problem is that there isn't enough information in the data to separate out the effects of the collinear variables. No technical fix can adequately compensate for that lack of information.

In thinking about solutions, we need to distinguish two rather different situations:

- The collinear variables are conceptually distinct
- The collinear variables can be seen as alternative measures of the same conceptual variable

The case of alternative measures is much easier to deal with. For this case, there are four common solutions.

1. *Delete one or more variables from the model.* Recall the example in Section 7.4 in which we had two measures of desired number of children, taken 1 year apart, with a correlation of .90. A natural reaction to this example would be to say "Why put both variables in the model? They're obviously both measuring a stable phenomenon, so there's no point in trying to get separate estimates of each one controlling for the other." But which one do you eliminate? With a correlation that high, it really doesn't make much difference. The time 2 measure has a higher correlation with the dependent variable, so that's a point in its favor. On the other hand, the time 1 measure has the virtue of being unaffected by births in the first year.

2. *Combine the collinear variables into an index.* Deleting variables may be less attractive when the variables have lower correlations with one another. If the two measures of desired number of children had a correlation of .80 instead of .90, it might seem like something important is being lost by deleting one of them. One way to resolve that problem is to combine the two (or more) variables into a single variable, called an index. There are lots of different ways to do this, some simple, some complex. If the variables have the same units of measurement, a simple sum or average may suffice. If they have different units of measurement, it's better to average the *standardized scores* (subtract the mean and divide by the standard deviation).

3. *Estimate a latent variable model.* Instead of creating an index, some researchers take the ultimate step and estimate a latent variable model. Such models assume that there is a single, unobserved

variable that affects the two or more observed variables that are collinear. The models require specialized software and are rather complex, both conceptually and operationally (Hayduk, 1988). The main advantages are that (a) the method corrects for measurement error, thereby solving one of the problems we discussed in Chapter 6, and (b) the method impresses some people with its sophistication.

4. *Perform joint hypothesis tests.* There is another simple method that can be used to great advantage when you want to leave the original, collinear variables in the regression equation. As we saw in the examples in Section 7.4, the most serious danger of multicollinearity is concluding that none of the collinear variables has an effect on the dependent variable when, in fact, any one of them alone has a very strong effect. Instead of looking at the test statistics and p values for each variable, you can test the joint hypothesis that "none of the collinear variables has a coefficient that differs from zero." Many regression programs have options for testing hypotheses like this. In the example of desired number of children, the two collinear variables, x_1 and x_3, had individual p values of .44 and .13 when both variables were in the equation. When I ran a test of the hypothesis that both variables had a coefficient of 0, however, the p value was .007, indicating a clear rejection of the null hypothesis. In the Black homicide example (Table 7.2), there were three variables in the regression whose coefficients were not statistically significant. When I did a joint test for all three coefficients, the p value was .76, indicating that the conclusions were robust to the multicollinearity.

What about situations in which the collinear variables are conceptually distinct? Suppose, for example, that we want to estimate a model predicting children's academic performance. Two of the independent variables—hours per day doing homework and hours per day watching TV—have a correlation of –.80. What should we do? It's not surprising that these two variables are highly correlated, but one can imagine quite different causal mechanisms by which they affect academic performance. They're clearly not measuring exactly the same thing, so we probably wouldn't want to combine them into an index or estimate a latent variable model. Unfortunately, if the aim is to get distinct, reliable estimates of the coefficients for these two variables, there's not much that can be done with collinear data. The best I can advise is to experiment with deleting variables and performing joint hypothesis tests. At least this will help you avoid the error of concluding that neither of the variables

is important, but it doesn't accomplish the goal of getting distinct estimates.

The only real solution to the problem of multicollinearity is to get better data. Simply increasing the sample size can help a great deal. Although it may not remove the multicollinearity, a larger sample will reduce the inflated standard errors that stem from multicollinearity. Even better is to somehow get data in which the variables are not collinear. Instead of aggregate data, use individual-level data. Instead of time-series data, use cross-sectional data. Stratified sampling on the independent variables can also help reduce the multicollinearity. Of course, this is all easier said than done. Acquiring new data can be time-consuming and expensive, and there is no guarantee that the new data will be any less problematic than the old.

Chapter Highlights

1. Extreme multicollinearity occurs when an independent variable in a linear regression model is a perfect linear function of other independent variables. Regression estimates cannot be computed when there is extreme multicollinearity. Most regression programs exclude one or more of the variables to produce regression estimates.
2. Multicollinearity affects only the coefficient estimates for those variables that are collinear.
3. Near-extreme multicollinearity occurs when there are strong, but not perfect, linear relationships among the independent variables.
4. The best measure of near-extreme multicollinearity is the tolerance, a number associated with each independent variable. The tolerance is computed by regressing each independent variable on all the other independent variables, then subtracting the R^2 from that regression from 1. A low tolerance means serious multicollinearity.
5. Although multicollinearity does not violate any assumptions of the standard regression model, it does make it difficult to get reliable estimates of the coefficients of the variables that are collinear.

6. When two (or more) independent variables are highly collinear, it can appear that neither of them affects the dependent variable, but when either is excluded from the model, the remaining variable may have a highly significant effect.
7. Multicollinearity makes multiple regression much more sensitive to minor errors or departures from the assumptions of the model.
8. Multicollinearity is not so serious when the main goal of the regression analysis is to predict the dependent variable.
9. Multicollinearity is a common problem with time-series data or with aggregate data.
10. When the collinear variables can be thought of as alternative measures of the same underlying dimensions, there are four common solutions: (a) delete one or more variables, (b) combine the variables into an index, (c) estimate a latent variable model, or (d) perform joint hypothesis tests.

Questions to Think About

1. Professor Miller wants to estimate a regression model predicting attitude toward abortion, using data from the General Social Survey. This national survey is conducted annually, and Miller has a data set that combines results from several different years. He calculated each person's age by subtracting the year of birth from the year of the survey. His regression model includes three variables: respondent's age, respondent's year of birth, and calendar year of the survey. His regression program refuses to produce estimates. What's wrong here?
2. For the regression described in the preceding question, Professor Miller also wants to include years of schooling as an independent variable. Does he need to be concerned about collinearity problems affecting the estimates for this variable?
3. There are 20 other variables that Professor Miller wants to include in his regression analysis. He examines all the two-variable correlations (there are 120 of them) and finds that none is above .40. Can he be confident that multicollinearity is not a problem for these variables? Why or why not?

4. For a sample of recent college graduates, Dr. Harrison performs a regression analysis to study the effect of extracurricular activities in college on starting salary in the first job. Her independent variables include (x_1) number of campus organizations in which the student was a member, (x_2) number of organizations in which the student was an officer, and (x_3) number of hours per week devoted to extracurricular activities. She finds that none of these variables is statistically significant. Is she safe in concluding that extracurriculars have no impact on starting salaries? If not, what should she do?
5. A weather forecaster constructs a regression model to predict the amount of rainfall on a given day based on numerous atmospheric measurements made the previous day. The R^2 for this regression is .85, which is much better than previous models. When he examines multicollinearity diagnostics, however, he finds that several of the independent variables have tolerances less than .30. Should he be concerned about his model? If yes, what should he do?
6. An educational researcher wants to know how use of the Internet affects the grades of high school students. Her dependent variable is GPA. Her independent variables include three time measurements from the previous month: (x_1) hours spent on the Internet, (x_2) hours spent on other on-line services, and (x_3) hours of total computer usage. She finds that x_1 is statistically significant at the .05 level but x_2 and x_3 are not. The tolerance for x_1 is .38. Is she safe in concluding that total computer usage doesn't matter, but time on the Internet does matter?