ENTRY PREPARED FOR: *Encyclopedia of Social Measurement*

SUBMITTED: February 20, 2004

MULTICOLLINEARITY

D. Stephen Voss

University of Kentucky

I. Introduction

II. Full Multicollinearity

III. Is Partial Multicollinearity a Problem?

IV. Symptoms of Partial Multicollinearity

V. Methods of Dealing with Partial Multicollinearity

GLOSSARY

collinearity A special case of "multicollinearity" when one variable is a linear function of another. Social scientists often use the terms loosely, however, in which case it is interchangeable with multicollinearity.

full multicollinearity When two or more explanatory variables overlap completely, with one a perfect linear function of the others, such that the method of analysis cannot distinguish them from each other. This condition will prevent a multiple regression from estimating coefficients; the equation becomes unsolvable.

multiplicative model A model that presumes the dependent variable is a multiplicative function of the explanatory variables rather than an additive function. Researchers sometimes induce full

collinearity in these models when they attempt to include "interaction terms" or ratio variables.

partial multicollinearity When two or more explanatory variables overlap, such that they are correlated with each other in a sample, but still contain independent variation. This condition limits the extent to which analysis can distinguish their causal importance, but does not violate any assumptions required for regression.

perfect multicollinearity See "full multicollinearity."

MULTICOLLINEARITY (or, as it is sometimes abbreviated, collinearity) describes a condition that may appear when analysts simultaneously consider more than one explanation for a social outcome. It occurs when two or more of the explanatory variables in a sample overlap. Because of the overlap, methods of analysis cannot fully distinguish the explanatory factors from each other or isolate their independent influence. Social scientists usually apply the term when discussing multiple (linear) regression – in which case, it refers to a situation in which one independent variable is fully or partially a linear function of the others – but many forms of quantitative and qualitative analysis have their own version of "the multicollinearity problem" (King, Keohane, and Verba 1994, 122-24). Thus, although the following entry will explore multicollinearity formally using the notation of linear regression, it outlines the analytical issues more broadly as well.

I. INTRODUCTION

Multicollinearity stands out among the possible pitfalls of empirical analysis for the extent to which it is poorly understood by practitioners. Articles in social-science journals often expend an extensive amount of space to dismiss the presence of this condition, even though it poses little

threat to a properly interpreted analysis.

At its extreme, when explanatory variables overlap completely, multicollinearity violates the assumptions of the classical regression model (CRM). Full (or perfect) multicollinearity is easy to detect, though, because it prevents the estimation of coefficients altogether; an equation becomes unsolvable. This is not the sort of multicollinearity that customarily worries analysts. Full multicollinearity rarely appears in social science data unless the sample is exceedingly small (Berry and Feldman 1985, 38). Otherwise, it generally results from some kind of simple error in the data handling or model specification, one that is easy to diagnose and painless to address. When practitioners speculate about a possible "multicollinearity problem," therefore, they mean some sort of linear relationship among explanatory variables that falls short of complete overlap.

Partial multicollinearity – the use of overlapping variables that still exhibit independent variation – is ubiquitous in multiple regression. Two random variables will almost always correlate at some level in a sample, even if they share no fundamental relationship in the larger population. In other words, multicollinearity is a matter of degree; it is not a "problem" that does or does not appear (Harvey 1977). Furthermore, partial multicollinearity violates absolutely none of the assumptions of linear regression. It does not bias coefficient estimates. It does not result in inefficient use of the data available, nor does it cause falsely confident conclusions. The sole effect of this data limitation is that it makes conclusions more ambiguous or hesitant than they otherwise might have been. Despite the omnipresence and relative harmlessness of the multicollinearity problem, however, practitioners often write as though it represents an analytical flaw for which one must test and that one must solve if it appears. They may select solutions for "dealing with multicollinearity" that harm an analysis more than they help it. In short, practitioners take partial multicollinearity much more seriously than statistical

theory or the methodological literature would justify.

The remainder of this entry proceeds in four sections. Section II treats full multicollinearity in depth, both substantively and statistically. It provides the basic intuition for why collinear variables confound an analysis, using specific examples, then offers various mathematical illustrations of how full multicollinearity prevents the classical regression model from producing coefficient estimates. The third section then turns to partial multicollinearity. It distinguishes the effects of this data condition according to the purposes for which an analyst has included a particular variable. The fourth section briefly reviews the various symptoms of, and tests for, multicollinearity customarily used by social scientists. Then, Section V considers the array of methods for dealing with multicollinearity that analysts have available to them.

II. FULL MULTICOLLINEARITY

Full (or perfect) multicollinearity results when one explanatory variable contains no fluctuation independent of the movement in the others. Because the problem variable is indistinguishable from the remainder of the explanatory variables in such a situation, empirical analysis cannot parse it out. Conventional techniques break down in the face of such uninformative data, even when the empirical model is not actually a "linear" one. A maximum likelihood estimation, for example, becomes unsolvable if explanatory variables overlap so completely that they produce a flat likelihood curve and do not allow the method to converge on a set of coefficients. Comparative case analysis, similarly, cannot determine objectively which of the many traits that separate two cases actually account for differences in their outcomes. Additional assumptions or more-detailed analysis becomes necessary.

A. Intuitive Discussion

The symptoms of full multicollinearity differ depending on the analytical method, but the overarching problem remains the same: Two concepts that cannot be separated from each other within a sample cannot be distinguished from each other empirically in any analysis conducted solely on that sample. In a causal analysis, it becomes impossible to separate out the effects of the problem variable from the effects of the others. In an analysis oriented toward prediction, it becomes impossible to determine how much weight the problem variable should receive, compared to the other variables, when computing forecasts outside of the sample.

The presence of full multicollinearity does not mean that the explanatory concepts are theoretically indistinct in the population, or even that the conceptual differences would be indistinguishable in another data set produced using the same sampling method. It does mean that, without incorporating some kind of outside information, the current sample does not allow any way to avoid conflating the overlapping explanatory factors.

Let's start with a simple example. Suppose that a researcher wishes to understand why nations adopt strict censorship laws. One possibility might be to compare the United States, where expressive rights have become relatively permissive, with Canada, where public officials enjoy wide leeway when banning questionable forms of expression. Any number of important traits separate these two nations and therefore could explain the different realities. They achieved independence from Great Britain at a different time and in a different fashion. One elects an independent president whereas the other selects a prime minister in parliament. One contains two major political parties whereas the other sustains several. One allows the judiciary more flexibility and authority than the other does. Without incorporating some type of additional information, an analyst cannot say which feature of the two governments caused

policy outcomes to deviate, nor would a forecaster know how to predict censorship levels in a third country that shared some traits in common with the United States and others in common with Canada.

This problem becomes even clearer with a statistical analysis. Suppose that an engineering firm faces two simultaneous discrimination lawsuits. An African-American male files the first one, claiming that the company engages in wage discrimination against black employees. A white female files the second one, alleging that the company discriminates against women. An analyst collects a random sample of engineers in the firm, recording each person's race, gender, and salary. However, by chance, the sample ends up containing only two sorts of people: white men and black women. The analyst can compute average salaries for the two sorts of engineers and determine whether they differ more than should be true by chance alone. Indeed, the results might show that black women tend to suffer in their paychecks relative to white men, for one reason or another. This computation would be useless, however, for purposes of deciding which of the lawsuits had merit because the sample conflates the two explanatory factors of race and gender. It is impossible to say how much of the gap reflects racial differences between the two sets of engineers and how much of it represents gender differences. In these data, the two explanations for wage differences are indistinguishable.

These two examples may seem artificial. How likely is it that researchers would try to explain a complicated policy outcome using only two cases, without digging into the underlying policy-making process to parse out the different institutional explanations? How likely is it that a random sample would exclude white women and black men? Of course, neither situation will occur often.

On the other hand, these examples may not be as far-fetched as they seem. Full

multicollinearity rarely occurs in quantitative social science research. When it does appear, the problem almost always results from some kind of simple data handling error created by the analyst. For example, the analyst might accidentally copy from the same column in a data table twice, or might accidentally copy one variable over another within a software package. This sort of error would produce two identical variables with different names. An analyst might not realize that a model contains more variables than the sample contains observations, in essence mimicking the United States versus Canada confusion, or might divide up a sample and not realize that within a given subset of data a "variable" actually does not vary. These sorts of theoretically indefensible mistakes, resulting from poor data handling or careless model specification, produce most instances of full multicollinearity that practitioners will encounter. Fixing the mistakes automatically removes the estimation problem.

B. Formal Presentation

Full multicollinearity appears when one explanatory variable in a regression analysis is a direct linear function of the others. The underlying population relationship among the random variables might or might not be deterministic; the term "multicollinearity" only defines the condition in the sample.

In the classical regression model, full multicollinearity violates assumptions necessary for successful estimation. A regression model becomes unsolvable. The technique should not even be able to estimate coefficients, although statistical software packages might produce output if rounding errors create artificial variation sufficient to remove the perfect relationship among variables. More commonly, software packages either will refuse to give results for an unsolvable equation or will drop variables until the model becomes solvable.

There are various ways to illustrate why multivariate linear regression becomes impossible when an explanatory variable is statistically dependent on the others. For example, the notation partly depends on whether one approaches the CRM and its assumptions from the vantage of scalar algebra or matrix algebra.

1. Scalar Notation

The classical regression model selects slope coefficients according to the least squares criterion. It will report coefficient estimates that minimize the sum of squared prediction errors within the sample. For this reason, successful performance requires that some particular set of coefficients meet the least squares criterion. If two or more sets of coefficients could produce the same minimized sum of squared errors, then it becomes impossible to say which one is correct.

Full multicollinearity actually produces a situation in which an infinite number of coefficients would satisfy the least squares condition. Consider a simple analysis with only two explanatory variables in which one of them (X_2) is a direct linear function of the other (X_1). We can summarize this relationship as $X_2 = c + d\ X_1$, where c and d are constants. The usual regression equation poses that the outcome of interest – the dependent variable (y) – is a linear function of the explanatory variables and a random error term (e). The population model looks like this:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e \qquad (1)$$

where β_0 is the intercept, β_1 and β_2 are the slope coefficients for their respective explanatory variables. If we suppose that X_2 in the sample is a multiple of X_1 (that is, $X_2 = d\ X_1$), just to keep

things simple, then the regression would need to find the coefficient estimates b_1 and b_2 that produced the best predictions $\hat{y}$:

$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2$$

$$= b_0 + b_1 X_1 + b_2 (d X_1)$$

$$\hat{y} = b_0 + (b_1 + d\, b_2) X_1 \qquad (2)$$

An infinite number of coefficient pairs could produce the same values for $\hat{y}$, as long as any change in b_1 from one possible value to another (δb_1) is matched by corresponding change in b_2 to compensate for it ($\delta b_2 = -\delta b_1 / d$). Thus, an infinite number of (linearly related) coefficient pairs could produce the minimum sum of squared errors, $(y - \hat{y})^2$. Presume, for example, that X_1 is a percentage and X_2 is simply a proportion measure of the same variable, such that $d = 100$. For any possible solution, b_1 and b_2, another pair could produce the same predicted values as long as the increase/decrease in b_1 would be matched by a decrease/increase in b_2 by 1/100 of that amount.

A couple of other common model specification errors can illustrate how full multicollinearity might appear in an analysis and how it foils the estimation. Suppose, for example, that a researcher collects a sample of state-level electoral data for a single year. The dependent variable is the Republican proportion of a state's congressional delegation after a given year's election (Y). The researcher hypothesizes that Republican electoral success would be a function of the following explanatory variables:

1. The GOP presidential contender's success in that state in that year, through some sort of coattail effect (X_1),

2. The GOP presidential contender's success in that state in the previous election, some sort of signaling effect that influences the quality of candidates fielded by each party (X_2), and

3. The state's change in support for the GOP presidential candidate between the two elections, to capture some sort of bandwagon or trend effect ($X_3 = X_1 - X_2$).

Described in this way, those three hypotheses may all sound reasonable. In fact, though, the "bandwagon" effect is not distinguishable from the other two variables and in any sample it would be a perfect linear combination of the other two:

$$\begin{aligned} \hat{Y} &= b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 \\ &= b_0 + b_1 X_1 + b_2 X_2 + b_3 (X_1 - X_2) \\ \hat{Y} &= b_0 + (b_1 + b_3) X_1 + (b_2 - b_3) X_2 \end{aligned} \qquad (3)$$

For any estimate of b_3, another would produce the same prediction $\hat{Y}$ as long as either b_2 shifted in the same direction by the same amount or b_1 shifted in the opposite direction by that amount. The equation would permit an infinite number of possible solutions.

Theorists sometimes posit multiplicative rather than additive regression models, such as the Cobb-Douglas function in economics (Kmenta 1997, 510-12) or the Gravity Model in the study of international trade (Anderson 1979). Here's one multiplicative model:

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \cdots X_K^{\beta_K} e^{\varepsilon} \qquad (4)$$

where K represents the number of explanatory variables in a properly specified model and e

represents the constant 2.718 . . . This precise specification is convenient, presuming it is theoretically appropriate, because it means the model is linear with respect to the natural logs (ln) of the variables. Suppose, for example, a three-variable version of the model:

$$\ln(Y) = \ln(\beta_0 X_1^{\beta_1} X_2^{\beta_2} X_3^{\beta_3} e^{\varepsilon}) \tag{5}$$

$$\ln(Y) = \ln \beta_0 + \beta_1 \ln X_1 + \beta_1 \ln X_2 + \beta_3 \ln X_3 + \varepsilon$$

What practitioners may overlook is that, when using a multiplicative model of this sort, the way one thinks about model specification must change. In particular, common practices such as taking a ratio or creating an interaction term can create full multicollinearity. For example, suppose that one wishes to predict a nation-state's yearly imports. The three hypothesized explanatory variables might be features of each country's economy:

1. The gross national product, to serve as a measure of the economy's size (X_1),
2. The population, to serve as a measure of consumer needs (X_2), and
3. The per capita income, to serve as a measure of wealth ($X_3 = X_1 / X_2$).

These three explanations may make superficial sense, but they are not theoretically defensible in such a multiplicative model because $\ln(X_1 / X_2) = \ln (X_1) – \ln (X_2)$, which creates a situation directly parallel to equation (3) above. Attempting to create a multiplicative "interaction term" will have the same problem, if the model also includes the two components of the interaction, because $\ln(X_1 X_2) = \ln (X_1) + \ln (X_2)$.

It is worth considering the effect of full multicollinearity on the measure of uncertainty for multivariate regression coefficients as well. Of course, it may seem silly to investigate

standard errors for coefficients that cannot be estimated, but this perspective once again shows how the usual computations fall apart when variables are perfectly collinear. Scalar-algebra textbook discussions of multiple regression often present a formula for the standard error (s) of a coefficient (b) for a particular explanatory variable (X_j) along the following lines (as adapted from Berry and Feldman 1985, 13):

$$s_{b_j} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2 (1 - R_j^2)(n - k)}} \qquad (6)$$

Most of this notation is standard. The important portion to discuss here is R_j^2, which represents the coefficient of determination that would result if X_j were regressed on the other explanatory variables. In the case of full multicollinearity, $R_j^2 = 1$ because the other explanatory variables account for 100 percent of the variation in X_j, which means that $(1–R_j^2) = 0$, the entire denominator therefore becomes 0, and the ratio becomes undefined. Once again, the regression equations fall apart in the face of full multicollinearity.

2. Matrix Notation

Matrix-based regression customarily expresses the explanatory variables as a single matrix (X), that might take the following form:

$$X = \begin{bmatrix} 1 & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{22} & x_{23} & \cdots & x_{2k} \\ 1 & x_{32} & x_{33} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & x_{n3} & \cdots & x_{nk} \end{bmatrix} \qquad (7)$$

where n represents the number of observations and k represents the number of independent variables, inclusive of the constant. The first column contains a series of 1's because it corresponds to the intercept, or constant, term in the regression. The dependent variable y, meanwhile, would represent a single column of numbers (a vector) of order n by 1.

The classical regression model's solution for computing its vector of coefficients (b) necessitates the following step:

$$X'X\, b = X'\, y \qquad (8)$$

$$b = (X'\, X)^{-1}\, X'\, y \qquad (9)$$

where X' is the transpose of X. Going from equation (8) to equation (9) requires X'X to be invertible and $(X'\, X)^{-1}$ actually to exist. In other words, X' X (sometimes written as Q) has to be square and "nonsingular." It will always be square, since by definition the transpose X' will have as many rows as X has columns. When one column of X is a linear combination of the other columns of X, however, the matrix will be singular. One row/column of Q will be a multiple of another row/column, such that the formula Q b = X' y allows an infinite number of

solutions for b. The rank of X will be less than k, the determinant of Q will equal zero, and no inverse will exist. In short, the formula cannot produce coefficient estimates in the presence of full multicollinearity. The coefficient standard error equation ($\sigma^2 Q^{-1}$) also requires X' X to be invertible and therefore produces no results in these cases.

One example of how full multicollinearity often appears becomes easy to illustrate using this notation. Researchers often perform regression on secondary data sets that include categorical variables. For example, a survey of individuals might include a variable to represent each person's religion. If the options only include Protestant, Catholic, Jewish, and Other, the data in X might look like this:

$$X = \left|\begin{array}{cc} \text{Constant} & \text{Faith} \\ 1 & \text{Protestant} \\ 1 & \text{Protestant} \\ 1 & \text{Protestant} \\ 1 & \text{Catholic} \\ 1 & \text{Catholic} \\ 1 & \text{Catholic} \\ 1 & \text{Jewish} \\ 1 & \text{Jewish} \\ 1 & \text{Other} \end{array}\right| \qquad (10)$$

It should be obvious that the Faith variable cannot be included in a regression as is, because it is not an ordinal variable, let alone an interval variable. Even if the data processor entered these faiths numerically rather than using text strings, the numbers assigned would be entirely arbitrary. The customary solution is to create a series of binary variables, one for each religion, in which an individual receives a value of 1 if they profess the pertinent faith and a 0 otherwise.

One common mistake is for analysts to include all of these "dummy variables" computed from the categorical variable into a regression. The new data matrix would look like this:

$$X = \begin{vmatrix} \text{Constant} & \text{Protestant} & \text{Catholic} & \text{Jewish} & \text{Other} \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{vmatrix} \quad (11)$$

As long as one assumes that no individual professes more than one faith, each row will have a positive value in one and only one of the four religion columns – which means that the constant term is a perfect linear combination of the four dummy variables:

$$\text{Constant} = \text{Protestant} + \text{Catholic} + \text{Jewish} + \text{Other} \quad (12)$$

Including them all with the constant produces full multicollinearity. The analyst must choose between (1) dropping the constant, in which case each of the religion dummies generates an intercept term for people of that faith, or (2) dropping one of the religion dummies, in which case the coefficients for the remainder represent an intercept shift for those who do not appear in the baseline category (that is, the one excluded from the regression).

II. Is Partial Multicollinearity a Problem?

Whereas full multicollinearity rarely appears in social-science data unless sample sizes are tiny or the analyst makes some kind of mistake, partial multicollinearity is rarely absent. One explanatory variable almost always correlates with the other explanatory variables, because that will happen by chance in a sample most of the time even if the variables have no underlying relationship in the population. Sometimes the theory underlying a model directly mandates some degree of partial collinearity. A parabolic model, for example, includes both an explanatory variable and its square term; these must be correlated. A regression with nested binary variables, in which one singles out a subset of the other, also must exhibit multicollinearity. In short, multicollinearity is a matter of degree rather than one of kind (Harvey 1977).

A. Unreliable Coefficient Estimates

Partial multicollinearity violates none of the assumptions required to perform multiple regression, so it undermines none of the desirable features of the classical regression model (such as unbiasedness or efficiency). It also does not prevent accurate measures of uncertainty. Unfortunately, severe cases of multicollinearity can cause that uncertainty to be rather high.

Recall the intuitive example of full multicollinearity discussed above, in which an analyst attempting to test for wage discrimination against both blacks and women could not distinguish the two cases because the sample only included white men and black women. Suppose that the analyst collected one additional case, a black male. At this point, the two explanatory variables – race and gender – would no longer be perfectly collinear. It would be possible to compute average salaries by race and average salaries by gender without the two gaps necessarily being the same. In particular, when testing for racial discrimination the final engineer would appear

among the potentially disadvantaged class, whereas when testing for gender discrimination he would not.

The results of this analysis, however, would not be terribly reliable because they depend entirely on the salary of one randomly selected person. If he happened to collect a salary similar to the other men, the analysis would estimate that the company did not discriminate by race but seriously discriminated by gender. If he happened to collect a salary similar to that of the black women in the sample, by contrast, the analysis would estimate that the company did not discriminate by gender but seriously discriminated by race. And if his salary fell at the midpoint between the black women and the white men, the analysis would indicate relatively moderate levels of discrimination against both groups. Adding a single person might have permitted estimates, but it hardly made those estimates trustworthy.

More generally, a high degree of partial multicollinearity adds to the instability of estimates such that the few discordant cases – the limited independent variation across the overlapping explanatory variables – strongly shape the final results. Consider, for example, the results worked out in equation (2), but with X_2 only a partial linear function of X_1, $X_2 = c + d\ X_1 + u$, where c and d are again constants but u represents the remaining random variation in X_2, centered at 0. Dropping c again for convenience, the ultimate computation becomes:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

$$= b_0 + b_1 X_1 + b_2 (d X_1 + u)$$

$$\hat{Y} = b_0 + (b_1 + d\ b_2) X_1 + b_2 u \qquad (13)$$

In this case, multiple regression does allow a solution. Shift from one possible value of b_2 to another, and the sum of squared errors would change because of alterations in the b_2 u term.

However, the changes in ŷ resulting from different possible choices for b_2 will be relatively minor if u contains little variation (that is, if the variables are highly collinear). Adjustments in b_1 could compensate for most of the shift in b_2, allowing a wide range of (b_1, b_2) values to appear roughly equivalent. Furthermore, the estimated balance between b_1 and b_2 would depend heavily on the limited variation in u that happened to appear. If y happened to be particularly high when u crept upward, then the b_2 estimate would be positive and the b_1 estimate would compensate by anchoring predictions downward. If y happened to be notably low when u crept upward, by contrast, b_2 would be selected to drag the predictions down and b_1 would swing in the other direction. In other words, the coefficient estimates will be unstable and negatively correlated with each other.

Regression output will not hide the uncertainty created by a high degree of multicollinearity. Review equation (6), which presents the formula for a coefficient standard error in multiple regression. Roughly speaking, errors go down when X and y covary significantly (the numerator) and when the sample size is large relative to the number of explanatory variables (the rightmost term in the denominator). They also go down when the explanatory variable bounces around a lot (the leftmost term in the denominator), but only insofar as this variation is independent of the other explanatory variables – as indicated by the middle term $(1–R_j^2)$. When partial multicollinearity is high, the middle term $(1–R_j^2)$ for a variable drops, decreasing the denominator and so ultimately increasing the standard error. In short, the coefficient standard errors from multiple regression correctly indicate the uncertainty, and therefore unreliability, of coefficient estimates produced for variables that offer little to distinguish them from each other.

B. The Implications of Partial Multicollinearity

Explanatory variables subject to partial multicollinearity might appear in an analysis because they interest the researcher as potential causal explanations for the outcome being studied. They might appear in an analysis because the researcher believes they will help forecast outcomes beyond the sample. Or the multicollinear variables might appear simply to hold certain concepts constant – control variables that allow the researcher to isolate the independent effects of one or more other variables that do not exhibit such a high degree of collinearity. The implications of partial multicollinearity differ depending on whether a researcher intends to draw causal inferences and whether the coefficient on the variable in question represents a particular quantity of interest.

1. Statistical Control

Predictive models often contain a wide variety of explanatory variables, only some of which directly concern the researcher. The remainder, usually called "control variables," appear in an analysis simply to increase the accuracy of the theoretically important coefficient estimates. The researcher need not worry whether coefficient estimates for control variables are close to the truth, nor does an analyst necessarily mind if these coefficients are accompanied by high standard errors (that is, whether they "achieve statistical significance" or not). What matters is that these rival explanations appeared in the analysis at all, thereby protecting other coefficient estimates from omitted variable bias.

Partial multicollinearity among control variables is almost entirely harmless. It does not undercut their effectiveness at eliminating omitted variable bias. It does not produce any sort of bias. It does not reduce the fit of a regression. Coefficient standard errors properly report the

uncertainty attached to each estimate; there should be no opportunity to place more stock in a given coefficient than it deserves. About the only risk is if, aside from random noise, the control variables lack independent variation – overlapping so completely that they are redundant. This circumstance would create inefficiency in a regression model, which could be problematic in very small data sets. But such costs appear any time analysts include unnecessary variables in a model; they are not unique to cases of partial multicollinearity. Otherwise, unless the researcher places more confidence in coefficient estimates than warranted by their level of uncertainty – an indefensible flaw – partial multicollinearity in the control variables does not disrupt an analysis.

2. Optimizing Prediction and Forecasting

Researchers sometimes care more about the predictive power of a statistical model than they do about identifying causal effects. Forecasting models need not place as much emphasis on why one variable is correlated with another, or on the causal ordering among various independent variables, as long as the overall model generates accurate out-of-sample predictions.

Of course, this distinction between forecasting and causal analysis is more conceptual than real. A forecasting model based on causally unrelated, and therefore theoretically inappropriate, independent variables is not likely to perform well outside of the sample on which it is based. A successful causal model that appropriately captures the theoretical process underlying data generation is likely to be successful at forecasting. Nonetheless, to the extent forecasting and causal analysis represent different analytical projects, partial multicollinearity poses little risk to the forecasting side of the enterprise.

Multiple regression takes into account the joint variation in various independent variables when it minimizes the sum of squared errors. The technology will not be able to assign

responsibility directly to one variable or another, given their covariance, but it does adjust coefficient estimates and predicted values according to the relationship between that shared variance and the dependent variable. Therefore, the regression model milks multicollinear variables of any predictive power that they might bring to the task of forecasting. It uses variables in the sample as efficiently as possible.

Partial multicollinearity does carry some cost in a forecasting situation. Instability in the coefficient estimates for these variables naturally reduces confidence in predictions, increasing standard errors around them. An analyst naturally would like to reduce the imprecision caused by partial multicollinearity, if that is an option. Nevertheless, the problem is only as bad as the standard errors around the predictions. If these errors are already small enough that they produce predictions with an acceptable level of uncertainty, then partial multicollinearity may be ignored.

3. Determining Causation

Partial multicollinearity really only matters when it interferes with the precision of important coefficient estimates and therefore limits the researcher's ability to identify the causal process leading to a dependent variable. As indicated in equation 6, partial multicollinearity drives coefficient standard errors upward. A coefficient that might have passed conventional standards of statistical significance outside the presence of multicollinearity could flunk an hypothesis test because of inflated standard errors. An analyst therefore might fail to reject the null hypothesis only because an explanatory variable of interest overlaps with other variables in the sample.

Partial multicollinearity most resembles a pathology when variables unrelated or only thinly related in the larger population somehow end up highly correlated within one's sample, because it prevents a precision that otherwise should have been achievable. In essence, this

condition means that the analyst possesses poor data with insufficient information about a substantial portion of the larger population – much as is true when the sample simply contains too few observations (which, in Goldberger 1991, goes under the tongue-in-cheek name "micronumerosity"). Indeed, accidental cases of multicollinearity bear a particularly close relationship to the small-n problem because chance correlations among explanatory variables become increasingly improbable as the sample grows. Leaving aside small samples, therefore, partial multicollinearity usually will result when two or more variables actually have some sort of causal relationship among them.

The severity of partial multicollinearity depends on the overall nature of an analysis. For example, as equation 6 indicates, a model with impressive predictive power will produce small coefficient standard errors because the numerator in each case will be small. Large data sets also will produce small standard errors, because (n-k) will be large. Even in the presence of heavy multicollinearity, coefficient estimates may be precise enough to satisfy the analyst. Specifically, coefficient standard errors may be small enough for explanatory variables of interest to achieve statistical significance and establish a causal relationship, even if the analyst ignores partial multicollinearity entirely.

Even if some coefficients of interest fail to achieve statistical significance, partial multicollinearity still may pose no barrier to causal analysis – depending on the level of theoretical or causal precision that an analyst requires. For example, it is not uncommon for researchers to include multiple variables intended to capture the same basic theoretical concept: socioeconomic status, trust in government, social anomie, postmaterialist values, etc. These variables are not multicollinear by chance, through the collection of an unfortunate sample; they covary in the larger population. A package of variables selected to triangulate on a concept

naturally will be multicollinear to whatever extent they are tapping the same underlying social phenomenon. In fact, the better they perform as proxies for the underlying concept, the more they ought to covary and therefore the greater the multicollinearity – a paradoxical instance in which more multicollinearity is in fact a healthy sign. If the point of model specification was to establish the causal significance of a core concept, not the independent significance of each given proxy, then multicollinearity poses no problem at all. Individually, the variables might fail to pass a significance test, but the package of variables will increase the fit of the model measurably.

At some point the idea of "variable packages" becomes hazy. Two variables may not capture the exact same underlying concept, but researchers nevertheless may be aware that they are causally related to each other in the larger population, such that multicollinearity in the sample is no accident. This sort of multicollinearity, although theoretically meaningful, nonetheless can pose an obstacle to the analyst who wishes to distinguish two or more concepts statistically; it can hinder an analysis based on fine theoretical distinctions. The appearance of such multicollinearity may be helpful, because it offers a warning that concepts may not be as theoretically distinct as a modeler initially assumed, but it still risks leaving the analyst with regrettably hesitant causal conclusions. The analyst would only be able to generalize about the overlapping variables as a package, even if a project's needs demand otherwise.

Social-science practitioners sometimes express concern when *control* variables are partially collinear with variables of interest. This worry is symptomatic of the misunderstandings revolving around the "multicollinearity problem." Overlap between control variables and variables of interest in fact serves an important function; it is another instance when the appearance of multicollinearity should be reassuring rather than a cause for concern.

Control variables appear in a model primarily to ward off omitted variable bias. When an analyst omits relevant explanatory variables, the coefficient estimates expected for the remaining variables would take the following form in matrix algebra (Goldberger 1991, 189-190):

$$E(b_1) = \beta_1 + F\,\beta_2 \qquad (14)$$

where β_1 is a column vector containing the correct population parameter(s) for the included variables; β_2 is a column vector of population parameters for the omitted variable(s), representing their underlying relationship with the dependent variable; and F is a matrix representing the relationships among the omitted variables and those remaining in the estimated equation. If the term $F\,\beta_2$ represents a null vector – which it would if either F is a null matrix or β_2 is a null vector – then excluding the variables would not produce omitted variable bias. The expected result of an uncontrolled (that is, "restricted") regression would be the correct population parameters, β_1.

What does this mean for multicollinearity? The only time unbiasedness requires inclusion of a control variable is when one theorizes that F is not null – in other words, when that control term *should* be partially collinear with a variable of interest. Including the control variable implicitly acknowledges that variables are expected to be multicollinear. It is therefore odd, if not inexplicable, that practitioners would become concerned when the multicollinearity appears as theorized. Omitted variable bias undermines an important statistical property in an analysis: unbiasedness. In contrast, the inflated standard errors that result from multicollinearity are not an analytical flaw but a defense mechanism, an indication of the uncertainty attached to a coefficient of interest once one takes into account how much it overlaps with other plausible

explanations captured by the statistical controls. Multicollinearity is the solution to a problem, not the problem itself, in these instances.

C. Special Cases

Some forms of model specification necessarily induce partial multicollinearity, and so are worth a separate look: parabolic models, interaction terms, and models with nested binary variables. Parabolic models are one means of rigging linear regression to accommodate nonlinear systematic relationships between dependent and independent variables. If X_1 is the root explanatory variable, for example, the model might include $X_2 = X_1^2$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + e \qquad (15)$$

Inclusion of an explanatory variable in its squared form allows a bend in the systematic component of their relationship. Throwing in the cubed form of the variable allows for a second bend, and so on:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + ... + \beta_k X_1^k + e \qquad (16)$$

Because the exponential variables in this sort of model are simply manipulations of the core variable, they naturally covary (although the nature of the collinearity depends on the scale and direction of the source variable).

Does multicollinearity in a parabolic model prevent causal analysis? Not at all. As with using other "package" variables, applying multiple versions of the source variable does not

remove any explanatory power from it. Coefficients will minimize the sum of squared errors, taking into account the joint variance among the different manipulated forms of the variable as well as the variance independent to each of them. It is possible that an important variable will fail to achieve statistical significance in any of its exponential forms, simply because they overlap so much. But this is an important finding: It means that, even if the variable matters in general, the regression cannot determine with confidence whether the systematic component contains a significant bend. It is an hypothesis test for the functional form rather than for the variable itself. The biggest risk with including exponential versions of a variable is that they may be unnecessary, thereby reducing the efficiency of the estimation.

Another form of variable manipulation that necessarily produces multicollinearity is the multiplicative interaction term. For example, the systematic functional form might be:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2) + e \qquad (17)$$

The purpose of this sort of model is that it allows the coefficient on one variable – say, X_1 – to depend linearly on the value of the other variable (in this case X_2). As X_1 increases by one unit, y increases by $\beta_1 + \beta_3 X_2$. The multiplicative term necessarily correlations with its own components. To whatever extent this drives up coefficient standard errors, however, it is not pathological. As always, the joint variation still contributes to coefficient estimates and the model's fit. The inflated standard errors may cause some of the variables to fall short on an hypothesis test, but this is not a test of the overall package; it simply tests the functional form. If the interaction term fails to achieve significance, that means the data do not allow an analyst to conclude that the effect of one variable depends on the value of another.

A final special case that inherently produces multicollinearity is when the researcher uses nested binary variables. For example, suppose that a scholar wishes to gauge the effect of equal protection lawsuits on funding across school districts. The dependent variable might be a measure of inequality in the state's funding system. The explanatory variable of interest might be a representation of the lawsuit. Rather than include only a single variable, though, the analyst might opt for one binary variable representing whether a lawsuit had been filed and another representing whether the complainant won. Because it is necessary to file the lawsuit to win it, the two binary variables will be collinear; a zero on the first variable necessitates a zero on the other.

This collinearity would drive up their standard errors. As with the other special cases, though, this is healthy. The two variables as a package capture the effect of the lawsuit; their independent effects merely probe whether it is the filing or the winning that makes the difference. If the two variables exhibit a high degree of multicollinearity – say, for example, because most legal challenges ultimately prove successful – they may not return significant coefficients. But this is a useful and important result. It means that the act of filing and the act of winning a legal challenge cannot be distinguished theoretically within the bounds of the data.

IV. Symptoms of Partial Multicollinearity

Despite the relative innocuousness of partial multicollinearity, practitioners show a high degree of sensitivity to it. The methodological literature has responded to this concern by outlining a number of ways to "detect" multicollinearity. The result of this effort is a wide variety of diagnostic tools, more or less rigid in their application. These various diagnostics actually test for a variety of phenomena, no doubt in part because the ill-defined "multicollinearity problem"

itself is not a distinct infringement of any statistical principle and does not undermine any important statistical property. It is not entirely clear what one needs to test for.

The most informal, and probably the most useful, diagnostic tool is to compare the fit of a model with the uncertainty reported for its various coefficients. When a package of explanatory variables jointly assists in predicting the dependent variable, they will markedly improve the model's fit – but they may fail to achieve independent significance if they are highly multicollinear. The partial (or "multiple-partial") F-test for joint significance formalizes this comparison by determining whether inclusion of a package of variables – moving from a restricted to an unrestricted model – significantly increases the model's coefficient of determination (R^2):

$$F(k_U - k_R, n - k_U) = \frac{(R_U^2 - R_R^2)/(k_U - k_R)}{(1 - R_U^2)/(n - k_U)} \quad (18)$$

where n is the total number of observations, k indicates the number of explanatory variables (constant term included), subscript R indicates the relevant statistic from a model with the package of variables excluded (i.e., their coefficients "restricted" to be zero), and subscript U indicates a statistic from a model with the package of variables included (i.e., their coefficients "unrestricted" and so subject to estimation). A package of variables can achieve statistical significance in a partial F-test even if none of them does so individually on the conventional t-test.

One common practice in empirical journals is to include a covariance matrix, showing

the bivariate relationships among all the difference explanatory variables. As a diagnostic tool, this sort of table serves little purpose. It would not detect complex forms of multicollinearity, when one explanatory variable is a function of several of them. It provides no way to distinguish accidental and population-based forms of multicollinearity. And it does not address the substantive question of whether partial multicollinearity undermines the goals of a research project. An alternate but related solution is to present a series of partial regressions, in which each independent variable in turn is regressed on the others. This sort of diagnostic does not address the theoretical difficulties with the bivariate method, but at least it gives a clearer picture of how much independent variation each explanatory variable actually contains.

A number of specific tests for multicollinearity have been proposed (Kmenta 1986, 438-439). Some people have used the determinant of (X'X) from the matrix regression equation, for example. An even better solution (because it is bounded and unaffected by the dispersion of the explanatory variables) is simply to look at the coefficient of determination, as appears in equation 6, for each explanatory variable. Perhaps the most popular test, in recent times, is the Variance Inflation Factor (VIF). It simply represents the inverse of the middle term in equation 6:

$$VIF = 1 / (1 - R_j^2) \qquad (19)$$

The conventional wisdom or rule of thumb seems to be that a VIF of 10 or more signifies trouble, although others use more conservative standards (i.e., higher thresholds). All of these tests share the same basic shortcoming, though, which is they attempt to systemize what should be a substantive judgment, a matter of analysis and interpretation. Recent scholarship therefore

seeks to define the costs of including multicollinear variables in terms of the predictive impact of their presence (Greenberg and Parks 1997).

V. Methods of Dealing with Partial Multicollinearity

Practitioners generally perceive multicollinearity as troublesome. Once in a while – especially incidental multicollinearity that obscures causal effects – partial multicollinearity really does hinder a theoretically appropriate analysis. Analysts concerned about partial multicollinearity face numerous options.

A. Incorporate Additional Information

Multicollinearity becomes problematic when a sample is not adequately informative about differences among two or more variables, resulting in coefficient estimates that are insufficiently precise and coefficient standard errors that are unacceptably large. Because the cause is at root one of poor data, the most obvious way to remove high degrees of multicollinearity is to get better data.

As equation 6 indicates, data improvements can lower coefficient standard errors even if they do not address multicollinearity directly. The analyst could increase the number of observations (n), which would have the additional virtue of making incidental multicollinearity less likely. The analyst also could decrease the number of explanatory variables (k), if some of them are unnecessary, or add other theoretically relevant explanatory variables if they can significantly improve model specification (driving up the numerator). Such changes would compensate for a low $(1-R_j^2)$.

A researcher also could choose to oversample observations that would increase the

independent variation in an explanatory variable, increasing ($1-R_j^2$). For example, if a study uses both education and income and these are highly correlated, the researcher could affirmatively select for highly educated poor people and/or people who are wealthy but uneducated. Selecting on explanatory variables does not induce bias the way that selecting on a dependent variable will (King, Keohane, and Verba 1994, 137-149). This strategy does entail certain risks, however, because it produces an overall sample that is not representative of the larger population. It may decrease multicollinearity at the expense of inducing bias in other analysis carried out on the data (at least unless those analyses take into account the non-random selection through some sort of weighting or similar solution).

These solutions all required pulling new data directly into an ongoing analysis. An alternate way to incorporate additional data, though, is to borrow insights from previous research in the field. The analyst may be able to apportion the joint variance among variables, for example by assuming a causal order, or otherwise formalize the relationship among multicollinear variables. The analyst may be willing to constrain coefficients for one or more of the partially multicollinear variables, based on previous research. Obviously using prior knowledge is risky; it runs the risk of creating a scholarly literature that is less science and more echo chamber. Nevertheless, researchers trying to exploit imperfect data may decide that their best solution is to take advantage of prior knowledge about the variables being used.

Certain other highly technical solutions also exist. They also incorporate prior information, and therefore in a sense qualify as Bayesian estimators. Probably the most common example is ridge regression, such as "ordinary ridge regression" or ORR (Hoerl and Kennard 1970). This solution relies on the existence of an unmeasured positive constant k that one implicitly can insert into the standard regression equation to derive a new coefficient :

$$= (X'X + k\,I)^{-1}\,X'y \qquad (20)$$

where I is the identity matrix, such that k implicitly pumps up the diagonals of Q. This solution depends on the value of k, which is rarely given: the greater that number, the more it reduces variance but the more it biases coefficient estimates toward zero (Kmenta 1986, 440). Because determining k generally requires rather arbitrary deductions from the data, ridge regression is not used frequently.

B. Remove Additional Information

Adding data is a great idea, but an analyst may not have that luxury. Instead, researchers often react to a poor information source through the destructive strategy of throwing out even more information. Analysts occasionally toss out partially multicollinear variables, for example. More commonly, analysts will combine multicollinear variables into one composite measure. For example, they might perform a factor analysis to isolate the variance that variables share – using the resulting factor score as a new "index" variable. Such indices are reported on an arbitrary scale, and therefore are not interpretable in any direct substantive way. They also strip out any variation unique to the particular variables used to create the index, which may be substantively important. Finally, they might require questionable assumptions – such as the linearity assumption needed for factor analysis. Such indices do have an important virtue, though, which is that the relative weight of each variable in the index is derived statistically using information contained in the data.

Perhaps the most popular sort of index is the additive kind. Unlike indices derived from factor analysis or other statistical methods, the additive index is driven by assumption. Suppose, for example, that a researcher wishes to include "trust in other people" as an explanatory concept

in a causal model. The researcher has two survey questions representing interpersonal trust, one that asks whether people are "selfish" (variable X_1) and one that asks whether people "try to do the right thing" (variable X_2). Presumably these variables will contain some multicollinearity, because they are theoretically related, but also will exhibit some independent variation – because they get at the root explanatory concept of interest in different ways. The population model therefore would be:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e \qquad (21)$$

Recognizing that X_1 and X_2 both represent proxies for the same underlying concept, the analyst may create a new explanatory variable, an additive index $Z = (X_1 + X_2)$. The ostensible purpose of this additive index might be to avoid multicollinearity, or it could be an attempt to simplify presentation by reducing the overarching concept to a single variable with a single coefficient estimate and a single hypothesis test. The regression equation actually estimated becomes:

$$\begin{aligned} y &= b_0 + b_z Z + e \qquad (22) \\ &= b_0 + b_z (X_1 + X_2) + e \\ &= b_0 + b_z X_1 + b_z X_2 + e \end{aligned}$$

Using an additive index therefore produces coefficient estimates under the assumption that, in the population, $\beta_1 = \beta_2$ (a testable proposition that usually goes untested). If this constraint is inappropriate, the fit of the model will decline (imposing inefficiency) and implicit coefficient estimates will be biased – since neither $E(b_Z) = \beta_1$ nor $E(b_Z) = \beta_2$.

Practitioners sometimes justify additive indices as reasonable under the assumption that each proxy should receive the same weight (which factor analysis and related methods do not guarantee). Unless X_1 and X_2 are scaled so that they have the same variance, though, forcing two variables to receive the same implicit coefficient actually does not ensure that they receive the same weight. The variance of the additive index would be:

$$V(Z) = V(X_1 + X_2) = V(X_1) + V(X_2) + 2\ \mathrm{Cov}\ (X_1, X_2) \qquad (23)$$

Variation in the new variable therefore depends on which component initially contained more variance. The one with greater variance receives greater weight, both in the additive index before estimation and in the predicted values after estimation. Furthermore, as the rightmost term indicates, movement in the new variable will be overbalanced toward the joint variance between the two questions – washing out the independent variation in each proxy question that presumably justified its separate collection in the first place. In other words, a model constrained by use of an additive index assigns variables unequal influence, imposing a functional form both arbitrarily *and* atheoretically.

In sum, additive indices lack the virtue of statistically derived indices, but they possess all of the drawbacks: a meaningless scale and estimation that neglects the independent information provided by each proxy measure. The only real virtue of this solution to partial multicollinearity is simplicity: They are easy to compute and take up less space when reporting results.

C. Do Nothing

A final solution to the presence of multicollinearity is to do nothing about it at all. It does no real harm to a regression model aside from making some of the variables less precise, and the standard errors properly report this imprecision. The solution to do nothing may be appropriate when the multicollinearity appears for a theoretically meanginful reason, such as when numerous variables have been computed from the same source. It is also especially tempting in cases when the multicollinearity exists among control variables or within a forecasting model.

However, the option to do nothing is also a viable choice in causal models, as long as the researcher need not to parse out the independent effects of the multicollinear variables. The partial F-statistic in equation 18 allows a test of joint significance for any given package of variables. Although characterizing the effect of a package of variables may be trickier than reporting the effect of only one, it has the virtue of retaining both the scale and the independent variation of the source data.

Bibliography

Anderson, James E. 1979. "A Theoretical Foundation for the Gravity Equation." *American Economic Review* 69 (March): 106-116.

Berry, William D., and Stanley Feldman. 1985. *Multiple Regression in Practice.* Newbury Park, CA: Sage.

Fabrycy, Mark Z. 1975. "Multicollinearity Caused by Specification Errors." *Applied Statistics* 24(2): 250-254.

Goldberger, Arthur S. 1991. *A Course in Econometrics.* Cambridge, MA: Harvard University Press.

Greenberg, Edward, and Robert P. Parks. 1997. "A Predictive Approach to Model Selection and Multicollinearity." *Journal of Applied Econometrics* 12: 67-75.

Harvey, A.C. 1977. "Some Comments on Multicollinearity in Regression." *Applied Statistics* 26(2): 188-191.

Hoerl, A.E., and R.W. Kennard. 1970. "Ridge Regression: Biased Estimation for Non-Orthogonal Problems." *Technometrics* 12: 55-67.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry*. Princeton, NJ: Princeton University Press.

Kmenta, Jan. 1986. *Elements of Econometrics.* Ann Arbor, MI: University of Michigan Press.

Mansfield, Edward R., and Billy P. Helms. 1982. "Detecting Multicollinearity." *The American Statistician* 36(3): 158-160.