CHAPTER 37

The Problem of Multicollinearity

It will be recalled that one of the factors that affects the standard error of a partial regression coefficient is the degree to which that independent variable is correlated with the other independent variables in the regression equation. Other things being equal, an independent variable that is very highly correlated with one or more other independent variables will have a relatively large standard error. This implies that the partial regression coefficient is unstable and will vary greatly from one sample to the next. This is the situation known as multicollinearity. *Multicollinearity exists whenever an independent variable is highly correlated with one or more of the other independent variables in a multiple regression equation.* Multicollinearity is a problem because it undermines the statistical significance of an independent variable. Other things being equal, the larger the standard error of a regression coefficient, the less likely it is that this coefficient will be statistically significant.

Multicollinearity is one of the most vexing and intractable problems in all of regression analysis. Before examining this problem, we must distinguish between extreme multicollinearity and perfect multicollinearity. Perfect multicollinearity obtains whenever one independent variable is a perfect linear function of one or more of the other independent variables in a regression equation. This situation usually occurs whenever we construct a variable as a linear function of other variables. For example, we might construct a composite measure that is a simple additive function of several variables. A problem will arise, however, if we include this composite measure and the variables used

to construct it as independent variables in the same regression equation. Ordinary least-squares estimation procedures will fail in this situation because all of the variance in the composite measure can be explained by the variables used to construct it. In mathematical terms, it will not be possible to calculate the inverse of the matrix of the covariances among the independent variables. This problem can also arise in dummy variable regression analysis whenever we fail to exclude one of the categories of a categorical variable from the analysis. If we include binary variables for all of the categories of a categorical variable in the regression equation, the intercept will be a simple additive function of these binary variables. Once again, ordinary least-squares estimation procedures will fail

Even less than perfect multicollinearity is a problem in regression analysis. Indeed, extreme multicollinearity is often a more difficult problem because it frequently goes undetected. Extreme multicollinearity occurs whenever an independent variable is very highly correlated with one or more other independent variables. In this situation, the partial regression coefficient for the independent variable in question will have a comparatively large standard error. In practice, multicollinearity often takes the form of two very highly correlated independent variables. Whenever two independent variables are very highly correlated with one another, neither or them is likely to be statistically significant, even though they may both be highly correlated with the dependent variable. The partial regression coefficients for both of these independent variables will have relatively large standard errors precisely because they are highly correlated with one another. Indeed, we can often detect the presence of extreme multicollinearity in a multiple regression equation by simply examining the magnitude of the standard errors of the partial regression coefficients. It will be recalled that a regression coefficient must be larger than its standard error in order to be statistically significant. Indeed, at the conventional 5 percent probability level, a regression coefficient must be almost twice as large as its standard error. Therefore, whenever we find a partial regression coefficient whose standard error is relatively large, we must consider the possibility of multicollinearity.

If we suspect that an independent variable in a regression equation suffers from multicollinearity, it is necessary to identify the source of the problem. We can sometimes identify pairs of highly correlated independent variables by inspecting the matrix of correlations among the independent variables. However, this procedure will only identify multicollinearity produced by the correlation between pairs of inde-

pendent variables. It is possible for an independent variable to be highly correlated with several other independent variables without being highly correlated with any one of them. In order to examine this possibility, we must inspect the coefficients of determination between each independent variable and all of the other independent variables in a regression equation. If the computer programs used to perform the multiple regression analysis can provide the inverse of the matrix of the correlations among the independent variable, it is possible to obtain these coefficients of determination using the main diagonal elements of this inverse matrix. Specifically, each coefficient of determination is given by:

$$R^2_{i.jk} = 1 - \left(\frac{1}{r^{ii}}\right)$$

where r^{ii} is the diagonal element in row and column i of this inverse matrix. In any event, this procedure allows us to identify which independent variables are highly correlated with the other independent variables in the regression equation.

Although multicollinearity is a relatively easy problem to detect, it is not an easy problem to resolve. Indeed, several different statistical procedures have been devised for dealing with this problem. The most popular of these procedures is known as "ridge regression." In this procedure, the matrix of covariances among the independent variables is modified slightly in order to reduce the standard errors of the partial regression coefficients. Specifically, a relatively small positive value is added to the variance of each independent variable prior to calculating the least-squares estimates of the partial regression coefficients as follows :

$$\tilde{\mathbf{B}} = \tilde{\mathbf{C}}_{xx}^{-1}\mathbf{c}_{yx} = (\mathbf{X}'\mathbf{X} + \mathbf{I}k)^{-1}\mathbf{X}'\mathbf{y}$$

such that

$$\tilde{\mathbf{C}}^{-1} = \begin{bmatrix} s_1^2 + k & c_{12} & \cdots & c_{1j} \\ c_{21} & s_2^2 + k & \cdots & c_{2j} \\ \vdots & \vdots & & \vdots \\ c_{j1} & c_{j2} & \cdots & s_j^2 + k \end{bmatrix}$$

where k is the ridge regression constant that is added to the variance of each independent variable.

These ridge regression estimates of the partial regression coefficients will have smaller standard errors than their ordinary least-squares counterparts. However, the ridge regression estimates of the partial regression coefficients will also be biased. In short, ridge regression overcomes the problem of extreme multicollinearity by introducing a small amount of bias into the partial regression coefficients in order to reduce their standard errors. In practice, the main difficulty in using ridge regression involves the selection of an appropriate ridge regression constant to add to the main diagonal of the matrix of covariances among the independent variables. The objective is to find a constant that substantially reduces the standard errors of the partial regression coefficients without greatly increasing the bias of those coefficients. We can assess the degree of bias introduced into the partial regression coefficients associated with various ridge regression estimates by comparing them to the ordinary least-squares estimates of those same partial regression coefficients. After all, despite the fact that they have large standard errors in the presence of extreme multicollinearity, ordinary least-squares estimates of the partial regression coefficients are unbiased.

Obviously, ridge regression provides a less than ideal solution to the problem of multicollinearity. In fact, the simplest and most common solution to this problem is to avoid it altogether by respecifying the multiple regression equation. For example, we can often circumvent the problem of multicollinearity by deleting an independent variable from the analysis. This solution to the problem is most appropriate whenever multicollinearity is the result of two highly correlated independent variables. In this case, the deletion of one of the two variables from the analysis can often be justified on the grounds that these variables are simply redundant measures of the same underlying theoretical construct. In this regard, it must be noted that multicollinearity may affect the standard errors of some independent variables in a regression equation without affecting the standard errors of other independent variables in the same equation. For example, if we have a multiple regression equation with four independent variables, only two of which are highly correlated with one another, only the standard errors of the partial regression coefficients for the two highly correlated variables will be affected by multicollinearity. If we delete one of the two highly correlated independent variables from the regression equation, we can expect the partial regression coefficient of the other independent variable and its stan-

dard error to change substantially. However, to the extent that the other independent variables in this multiple regression equation are uncorrelated with the two highly correlated independent variables, the partial regression coefficients of the uncorrelated independent variables and their standard errors will remain relatively unchanged.

An alternative technique for circumventing the problem of multicollinearity by respecifying the multiple regression equation involves the creation of a composite measure. In this approach, two or more highly correlated independent variables are combined into a single composite measure. Once again, this technique presumes that the highly correlated independent variables are essentially redundant measures of the same underlying theoretical construct. For example, we might have a multiple regression equation in which both income and savings are independent variables. We can expect that these two independent variables will be highly correlated with one another. Rather than including both variables in the same multiple regression equation, and introducing multicollinearity into the analysis, we might combine them into a single measure composite representing economic capital. Indeed, there are reasons to believe that this composite measure would be a more reliable measure of economic capital than either income or savings separately. The main difficulty with this approach is that it presumes that the variables in the composite measure have the same unit of measurement. In this particular example, both income and savings are measured in the same unit of measurement: dollars. However, if two or more independent variables have different units of measurement, it is necessary to convert them to standard form before using them in a composite measure.