10

MULTICOLLINEARITY: WHAT HAPPENS IF THE REGRESSORS ARE CORRELATED?

> There is no pair of words that is more misused both in econometrics texts and in the applied literature than the pair "multi-collinearity problem." That many of our explanatory variables are highly collinear is a fact of life. And it is completely clear that there are experimental designs $\mathbf{X}'\mathbf{X}$ [i.e., data matrix] which would be much preferred to the designs the natural experiment has provided us [i.e., the sample at hand]. But a complaint about the apparent malevolence of nature is not at all constructive, and the *ad hoc* cures for a bad design, such as stepwise regression or ridge regression, can be disastrously inappropriate. Better that we should rightly accept the fact that our non-experiments [i.e., data not collected by designed experiments] are sometimes not very informative about parameters of interest.[1]

Assumption 10 of the *classical linear regression model* (CLRM) is that there is no **multicollinearity** among the regressors included in the regression model. In this chapter we take a critical look at this assumption by seeking answers to the following questions:

1. What is the nature of multicollinearity?
2. Is multicollinearity really a problem?
3. What are its practical consequences?
4. How does one detect it?
5. What remedial measures can be taken to alleviate the problem of multicollinearity?

[1]Edward E. Leamer, "Model Choice and Specification Analysis," in Zvi Griliches and Michael D. Intriligator, eds., *Handbook of Econometrics*, vol. I, North Holland Publishing Company, Amsterdam, 1983, pp. 300–301.

In this chapter we also discuss Assumption 7 of the CLRM, namely, that the number of observations in the sample must be greater than the number of regressors, and Assumption 8, which requires that there be sufficient variability in the values of the regressors, for they are intimately related to the assumption of no multicollinearity. Arthur Goldberger has christened Assumption 7 as the problem of **micronumerosity,**[2] which simply means small sample size.

10.1 THE NATURE OF MULTICOLLINEARITY

The term *multicollinearity* is due to Ragnar Frisch.[3] Originally it meant the existence of a "perfect," or exact, linear relationship among some or all explanatory variables of a regression model.[4] For the k-variable regression involving explanatory variable $X_1, X_2, \ldots, X_k$ (where $X_1 = 1$ for all observations to allow for the intercept term), an exact linear relationship is said to exist if the following condition is satisfied:

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_k X_k = 0 \qquad \textbf{(10.1.1)}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are constants such that not all of them are zero simultaneously.[5]

Today, however, the term multicollinearity is used in a broader sense to include the case of perfect multicollinearity, as shown by (10.1.1), as well as the case where the X variables are intercorrelated but not perfectly so, as follows[6]:

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_2 X_k + v_i = 0 \qquad \textbf{(10.1.2)}$$

where v_i is a stochastic error term.

To see the difference between *perfect* and *less than perfect* multicollinearity, assume, for example, that $\lambda_2 \neq 0$. Then, (10.1.1) can be written as

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - \frac{\lambda_k}{\lambda_2} X_{ki} \qquad \textbf{(10.1.3)}$$

[2]See his *A Course in Econometrics,* Harvard University Press, Cambridge, Mass., 1991, p. 249.

[3]Ragnar Frisch, *Statistical Confluence Analysis by Means of Complete Regression Systems*, Institute of Economics, Oslo University, publ. no. 5, 1934.

[4]Strictly speaking, *multicollinearity* refers to the existence of more than one exact linear relationship, and *collinearity* refers to the existence of a single linear relationship. But this distinction is rarely maintained in practice, and multicollinearity refers to both cases.

[5]The chances of one's obtaining a sample of values where the regressors are related in this fashion are indeed very small in practice except by design when, for example, the number of observations is smaller than the number of regressors or if one falls into the "dummy variable trap" as discussed in Chap. 9. See exercise 10.2.

[6]If there are only two explanatory variables, *intercorrelation* can be measured by the zero-order or simple correlation coefficient. But if there are more than two X variables, intercorrelation can be measured by the partial correlation coefficients or by the multiple correlation coefficient R of one X variable with all other X variables taken together.

which shows how X_2 is exactly linearly related to other variables or how it can be derived from a linear combination of other X variables. In this situation, the coefficient of correlation between the variable X_2 and the linear combination on the right side of (10.1.3) is bound to be unity.

Similarly, if $\lambda_2 \neq 0$, Eq. (10.1.2) can be written as

$$X_{2i} = -\frac{\lambda_1}{\lambda_2}X_{1i} - \frac{\lambda_3}{\lambda_2}X_{3i} - \cdots - \frac{\lambda_k}{\lambda_2}X_{ki} - \frac{1}{\lambda_2}v_i \qquad \textbf{(10.1.4)}$$

which shows that X_2 is not an exact linear combination of other X's because it is also determined by the stochastic error term v_i.

As a numerical example, consider the following hypothetical data:

X_2	X_3	X_3^*
10	50	52
15	75	75
18	90	97
24	120	129
30	150	152

It is apparent that $X_{3i} = 5X_{2i}$. Therefore, there is perfect collinearity between X_2 and X_3 since the coefficient of correlation r_{23} is unity. The variable X_3^* was created from X_3 by simply adding to it the following numbers, which were taken from a table of random numbers: 2, 0, 7, 9, 2. Now there is no longer perfect collinearity between X_2 and X_3^*. However, the two variables are highly correlated because calculations will show that the coefficient of correlation between them is 0.9959.

The preceding algebraic approach to multicollinearity can be portrayed succinctly by the Ballentine (recall Figure 3.9, reproduced in Figure 10.1). In this figure the circles Y, X_2, and X_3 represent, respectively, the variations in Y (the dependent variable) and X_2 and X_3 (the explanatory variables). The degree of collinearity can be measured by the extent of the overlap (shaded area) of the X_2 and X_3 circles. In Figure 10.1*a* there is no overlap between X_2 and X_3, and hence no collinearity. In Figure 10.1*b* through 10.1*e* there is a "low" to "high" degree of collinearity—the greater the overlap between X_2 and X_3 (i.e., the larger the shaded area), the higher the degree of collinearity. In the extreme, if X_2 and X_3 were to overlap completely (or if X_2 were completely inside X_3, or vice versa), collinearity would be perfect.

In passing, note that multicollinearity, as we have defined it, refers only to linear relationships among the X variables. It does not rule out nonlinear relationships among them. For example, consider the following regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i \qquad \textbf{(10.1.5)}$$

where, say, Y = total cost of production and X = output. The variables X_i^2 (output squared) and X_i^3 (output cubed) are obviously functionally related

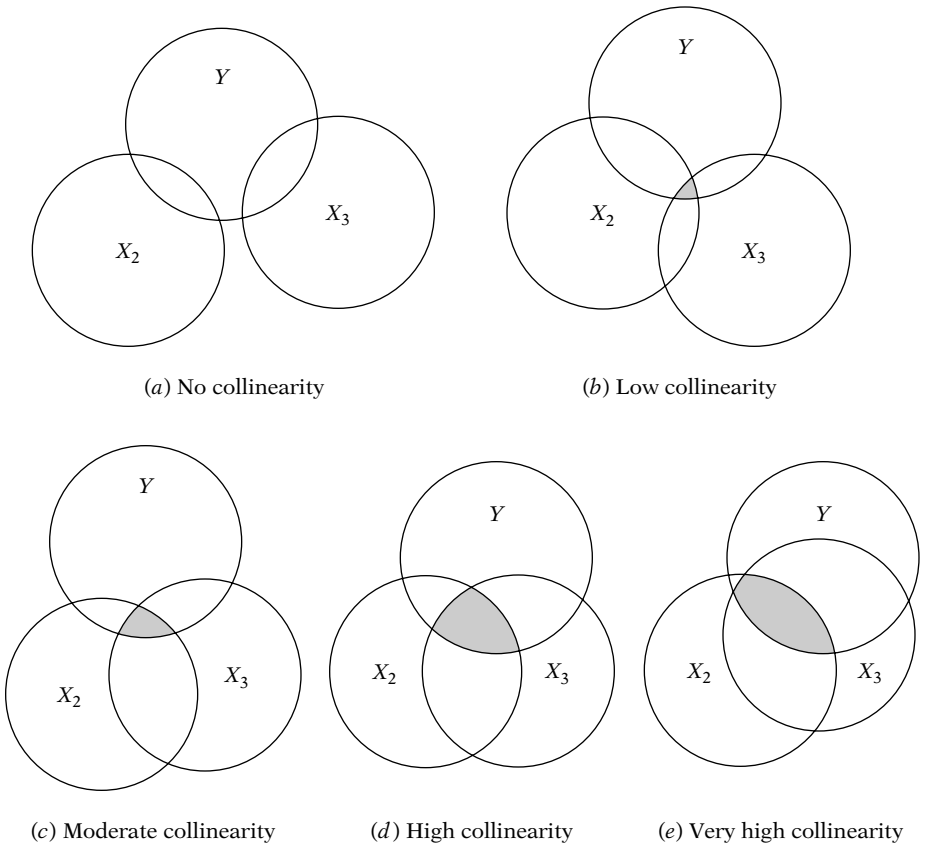

FIGURE 10.1 The Ballentine view of multicollinearity.

to X_i, but the relationship is nonlinear. Strictly, therefore, models such as (10.1.5) do not violate the assumption of no multicollinearity. However, in concrete applications, the conventionally measured correlation coefficient will show X_i, X_i^2, and X_i^3 to be highly correlated, which, as we shall show, will make it difficult to estimate the parameters of (10.1.5) with greater precision (i.e., with smaller standard errors).

Why does the classical linear regression model assume that there is no multicollinearity among the X's? The reasoning is this: **If multicollinearity is perfect in the sense of (10.1.1), the regression coefficients of the *X* variables are indeterminate and their standard errors are infinite. If multicollinearity is less than perfect, as in (10.1.2), the regression coefficients, although determinate, possess large standard errors (in relation to the coefficients themselves), which means the coefficients cannot be estimated with great precision or accuracy.** The proofs of these statements are given in the following sections.

There are several sources of multicollinearity. As Montgomery and Peck note, multicollinearity may be due to the following factors[7]:

1. *The data collection method employed,* for example, sampling over a limited range of the values taken by the regressors in the population.

2. *Constraints on the model or in the population being sampled.* For example, in the regression of electricity consumption on income (X_2) and house size (X_3) there is a physical constraint in the population in that families with higher incomes generally have larger homes than families with lower incomes.

3. *Model specification,* for example, adding polynomial terms to a regression model, especially when the range of the X variable is small.

4. *An overdetermined model.* This happens when the model has more explanatory variables than the number of observations. This could happen in medical research where there may be a small number of patients about whom information is collected on a large number of variables.

An additional reason for multicollinearity, especially in time series data, may be that the regressors included in the model share a *common trend,* that is, they all increase or decrease over time. Thus, in the regression of consumption expenditure on income, wealth, and population, the regressors income, wealth, and population may all be growing over time at more or less the same rate, leading to collinearity among these variables.

10.2 ESTIMATION IN THE PRESENCE OF PERFECT MULTICOLLINEARITY

It was stated previously that in the case of perfect multicollinearity the regression coefficients remain indeterminate and their standard errors are infinite. This fact can be demonstrated readily in terms of the three-variable regression model. Using the deviation form, where all the variables are expressed as deviations from their sample means, we can write the three-variable regression model as

$$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i \qquad \textbf{(10.2.1)}$$

Now from Chapter 7 we obtain

$$\hat{\beta}_2 = \frac{\left(\sum y_i x_{2i}\right)\left(\sum x_{3i}^2\right) - \left(\sum y_i x_{3i}\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i} x_{3i}\right)^2} \qquad \textbf{(7.4.7)}$$

[7]Douglas Montgomery and Elizabeth Peck, *Introduction to Linear Regression Analysis*, John Wiley & Sons, New York, 1982, pp. 289–290. See also R. L. Mason, R. F. Gunst, and J. T. Webster, "Regression Analysis and Problems of Multicollinearity," *Communications in Statistics A*, vol. 4, no. 3, 1975, pp. 277–292; R. F. Gunst, and R. L. Mason, "Advantages of Examining Multicollinearities in Regression Analysis," *Biometrics*, vol. 33, 1977, pp. 249–260.

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \tag{7.4.8}$$

Assume that $X_{3i} = \lambda X_{2i}$, where λ is a nonzero constant (e.g., 2, 4, 1.8, etc.). Substituting this into (7.4.7), we obtain

$$\begin{aligned} \hat{\beta}_2 &= \frac{(\sum y_i x_{2i})(\lambda^2 \sum x_{2i}^2) - (\lambda \sum y_i x_{2i})(\lambda \sum x_{2i}^2)}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2) - \lambda^2 (\sum x_{2i}^2)^2} \\ &= \frac{0}{0} \end{aligned} \tag{10.2.2}$$

which is an indeterminate expression. The reader can verify that $\hat{\beta}_3$ is also indeterminate.[8]

Why do we obtain the result shown in (10.2.2)? Recall the meaning of $\hat{\beta}_2$: It gives the rate of change in the average value of Y as X_2 changes by a unit, holding X_3 constant. But if X_3 and X_2 are perfectly collinear, there is no way X_3 can be kept constant: As X_2 changes, so does X_3 by the factor λ. What it means, then, is that there is no way of disentangling the separate influences of X_2 and X_3 from the given sample: For practical purposes X_2 and X_3 are indistinguishable. In applied econometrics this problem is most damaging since the entire intent is to separate the partial effects of each X upon the dependent variable.

To see this differently, let us substitute $X_{3i} = \lambda X_{2i}$ into (10.2.1) and obtain the following [see also (7.1.9)]:

$$\begin{aligned} y_i &= \hat{\beta}_2 x_{2i} + \hat{\beta}_3 (\lambda x_{2i}) + \hat{u}_i \\ &= (\hat{\beta}_2 + \lambda \hat{\beta}_3) x_{2i} + \hat{u}_i \\ &= \hat{\alpha} x_{2i} + \hat{u}_i \end{aligned} \tag{10.2.3}$$

where

$$\hat{\alpha} = (\hat{\beta}_2 + \lambda \hat{\beta}_3) \tag{10.2.4}$$

Applying the usual OLS formula to (10.2.3), we get

$$\hat{\alpha} = (\hat{\beta}_2 + \lambda \hat{\beta}_3) = \frac{\sum x_{2i} y_i}{\sum x_{2i}^2} \tag{10.2.5}$$

Therefore, although we can estimate α uniquely, there is no way to estimate β_2 and β_3 uniquely; mathematically

$$\hat{\alpha} = \hat{\beta}_2 + \lambda \hat{\beta}_3 \tag{10.2.6}$$

[8]Another way of seeing this is as follows: By definition, the coefficient of correlation between X_2 and X_3, r_{23}, is $\sum x_{2i} x_{3i} / \sqrt{\sum x_{2i}^2 \sum x_{3i}^2}$. If $r_{23}^2 = 1$, i.e., perfect collinearity between X_2 and X_3, the denominator of (7.4.7) will be zero, making estimation of β_2 (or of β_3) impossible.

gives us only one equation in two unknowns (note λ is given) and there is an infinity of solutions to (10.2.6) for given values of $\hat{\alpha}$ and λ. To put this idea in concrete terms, let $\hat{\alpha} = 0.8$ and $\lambda = 2$. Then we have

$$0.8 = \hat{\beta}_2 + 2\hat{\beta}_3 \tag{10.2.7}$$

or

$$\hat{\beta}_2 = 0.8 - 2\hat{\beta}_3 \tag{10.2.8}$$

Now choose a value of $\hat{\beta}_3$ arbitrarily, and we will have a solution for $\hat{\beta}_2$. Choose another value for $\hat{\beta}_3$, and we will have another solution for $\hat{\beta}_2$. No matter how hard we try, there is no unique value for $\hat{\beta}_2$.

The upshot of the preceding discussion is that in the case of perfect multicollinearity one cannot get a unique solution for the individual regression coefficients. But notice that one can get a unique solution for linear combinations of these coefficients. The linear combination $(\beta_2 + \lambda\beta_3)$ is uniquely estimated by α, given the value of λ.[9]

In passing, note that in the case of perfect multicollinearity the variances and standard errors of $\hat{\beta}_2$ and $\hat{\beta}_3$ individually are infinite. (See exercise 10.21.)

10.3 ESTIMATION IN THE PRESENCE OF "HIGH" BUT "IMPERFECT" MULTICOLLINEARITY

The perfect multicollinearity situation is a pathological extreme. Generally, there is no exact linear relationship among the X variables, especially in data involving economic time series. Thus, turning to the three-variable model in the deviation form given in (10.2.1), instead of exact multicollinearity, we may have

$$x_{3i} = \lambda x_{2i} + v_i \tag{10.3.1}$$

where $\lambda \neq 0$ and where v_i is a stochastic error term such that $\sum x_{2i}v_i = 0$. (Why?)

Incidentally, the Ballentines shown in Figure 10.1*b* to 10.1*e* represent cases of imperfect collinearity.

In this case, estimation of regression coefficients β_2 and β_3 may be possible. For example, substituting (10.3.1) into (7.4.7), we obtain

$$\hat{\beta}_2 = \frac{\sum(y_i x_{2i})\left(\lambda^2 \sum x_{2i}^2 + \sum v_i^2\right) - \left(\lambda \sum y_i x_{2i} + \sum y_i v_i\right)\left(\lambda \sum x_{2i}^2\right)}{\sum x_{2i}^2\left(\lambda^2 \sum x_{2i}^2 + \sum v_i^2\right) - \left(\lambda \sum x_{2i}^2\right)^2} \tag{10.3.2}$$

where use is made of $\sum x_{2i}v_i = 0$. A similar expression can be derived for $\hat{\beta}_3$.

[9] In econometric literature, a function such as $(\beta_2 + \lambda\beta_3)$ is known as an **estimable function.**

Now, unlike (10.2.2), there is no reason to believe a priori that (10.3.2) cannot be estimated. Of course, if v_i is sufficiently small, say, very close to zero, (10.3.1) will indicate almost perfect collinearity and we shall be back to the indeterminate case of (10.2.2).

10.4 MULTICOLLINEARITY: MUCH ADO ABOUT NOTHING? THEORETICAL CONSEQUENCES OF MULTICOLLINEARITY

Recall that if the assumptions of the classical model are satisfied, the OLS estimators of the regression estimators are BLUE (or BUE, if the normality assumption is added). Now it can be shown that even if multicollinearity is very high, as in the case of *near multicollinearity*, the OLS estimators still retain the property of BLUE.[10] Then what is the multicollinearity fuss all about? As Christopher Achen remarks (note also the Leamer quote at the beginning of this chapter):

> Beginning students of methodology occasionally worry that their independent variables are correlated—the so-called multicollinearity problem. But multicollinearity violates no regression assumptions. Unbiased, consistent estimates will occur, and their standard errors will be correctly estimated. The only effect of multicollinearity is to make it hard to get coefficient estimates with small standard error. But having a small number of observations also has that effect, as does having independent variables with small variances. (In fact, at a theoretical level, multicollinearity, few observations and small variances on the independent variables are essentially all the same problem.) Thus "What should I do about multicollinearity?" is a question like "What should I do if I don't have many observations?" No statistical answer can be given.[11]

To drive home the importance of sample size, Goldberger coined the term **micronumerosity,** to counter the exotic polysyllabic name *multicollinearity*. According to Goldberger, **exact micronumerosity** (the counterpart of exact multicollinearity) arises when n, the sample size, is zero, in which case any kind of estimation is impossible. *Near micronumerosity,* like near multicollinearity, arises when the number of observations barely exceeds the number of parameters to be estimated.

Leamer, Achen, and Goldberger are right in bemoaning the lack of attention given to the sample size problem and the undue attention to the multicollinearity problem. Unfortunately, in applied work involving secondary data (i.e., data collected by some agency, such as the GNP data collected by the government), an individual researcher may not be able to do much about the size of the sample data and may have to face "estimating problems

[10]Since near multicollinearity per se does not violate the other assumptions listed in Chap. 7, the OLS estimators are BLUE as indicated there.

[11]Christopher H. Achen, *Interpreting and Using Regression*, Sage Publications, Beverly Hills, Calif., 1982, pp. 82–83.

important enough to warrant our treating it [i.e., multicollinearity] as a violation of the CLR [classical linear regression] model."[12]

First, it is true that even in the case of near multicollinearity the OLS estimators are unbiased. But unbiasedness is a multisample or repeated sampling property. What it means is that, keeping the values of the X variables fixed, if one obtains repeated samples and computes the OLS estimators for each of these samples, the average of the sample values will converge to the true population values of the estimators as the number of samples increases. But this says nothing about the properties of estimators in any given sample.

Second, it is also true that collinearity does not destroy the property of minimum variance: In the class of all linear unbiased estimators, the OLS estimators have minimum variance; that is, they are efficient. But this does not mean that the variance of an OLS estimator will necessarily be small (in relation to the value of the estimator) in any given sample, as we shall demonstrate shortly.

Third, *multicollinearity is essentially a sample (regression) phenomenon* in the sense that even if the X variables are not linearly related in the population, they may be so related in the particular sample at hand: When we postulate the theoretical or population regression function (PRF), we believe that all the X variables included in the model have a separate or independent influence on the dependent variable Y. But it may happen that in any given sample that is used to test the PRF some or all of the X variables are so highly collinear that we cannot isolate their individual influence on Y. So to speak, our sample lets us down, although the theory says that all the X's are important. In short, our sample may not be "rich" enough to accommodate all X variables in the analysis.

As an illustration, reconsider the consumption–income example of Chapter 3. Economists theorize that, besides income, the wealth of the consumer is also an important determinant of consumption expenditure. Thus, we may write

$$\text{Consumption}_i = \beta_1 + \beta_2\,\text{Income}_i + \beta_3\,\text{Wealth}_i + u_i$$

Now it may happen that when we obtain data on income and wealth, the two variables may be highly, if not perfectly, correlated: Wealthier people generally tend to have higher incomes. Thus, although in theory income and wealth are logical candidates to explain the behavior of consumption expenditure, in practice (i.e., in the sample) it may be difficult to disentangle the separate influences of income and wealth on consumption expenditure.

Ideally, to assess the individual effects of wealth and income on consumption expenditure we need a sufficient number of sample observations of wealthy individuals with low income, and high-income individuals with

[12]Peter Kennedy, *A Guide to Econometrics*, 3d ed., The MIT Press, Cambridge, Mass., 1992, p. 177.

low wealth (recall Assumption 8). Although this may be possible in cross-sectional studies (by increasing the sample size), it is very difficult to achieve in aggregate time series work.

For all these reasons, the fact that the OLS estimators are BLUE despite multicollinearity is of little consolation in practice. We must see what happens or is likely to happen in any given sample, a topic discussed in the following section.

10.5 PRACTICAL CONSEQUENCES OF MULTICOLLINEARITY

In cases of near or high multicollinearity, one is likely to encounter the following consequences:

1. Although BLUE, the OLS estimators have large variances and covariances, making precise estimation difficult.

2. Because of consequence 1, the confidence intervals tend to be much wider, leading to the acceptance of the "zero null hypothesis" (i.e., the true population coefficient is zero) more readily.

3. Also because of consequence 1, the t ratio of one or more coefficients tends to be statistically insignificant.

4. Although the t ratio of one or more coefficients is statistically insignificant, R^2, the overall measure of goodness of fit, can be very high.

5. The OLS estimators and their standard errors can be sensitive to small changes in the data.

The preceding consequences can be demonstrated as follows.

Large Variances and Covariances of OLS Estimators

To see large variances and covariances, recall that for the model (10.2.1) the variances and covariances of $\hat{\beta}_2$ and $\hat{\beta}_3$ are given by

$$\text{var}\,(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)} \qquad \textbf{(7.4.12)}$$

$$\text{var}\,(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)} \qquad \textbf{(7.4.15)}$$

$$\text{cov}\,(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23}\sigma^2}{(1 - r_{23}^2)\sqrt{\sum x_{2i}^2 \sum x_{3i}^2}} \qquad \textbf{(7.4.17)}$$

where r_{23} is the coefficient of correlation between X_2 and X_3.

It is apparent from (7.4.12) and (7.4.15) that as r_{23} tends toward 1, that is, as collinearity increases, the variances of the two estimators increase and in the limit when $r_{23} = 1$, they are infinite. It is equally clear from (7.4.17) that as r_{23} increases toward 1, the covariance of the two estimators also increases in absolute value. [*Note:* $\text{cov}\,(\hat{\beta}_2, \hat{\beta}_3) \equiv \text{cov}\,(\hat{\beta}_3, \hat{\beta}_2)$.]

The speed with which variances and covariances increase can be seen with the **variance-inflating factor (VIF),** which is defined as

$$\text{VIF} = \frac{1}{(1 - r_{23}^2)} \tag{10.5.1}$$

VIF shows how the variance of an estimator is *inflated* by the presence of multicollinearity. As r_{23}^2 approaches 1, the VIF approaches infinity. That is, as the extent of collinearity increases, the variance of an estimator increases, and in the limit it can become infinite. As can be readily seen, if there is no collinearity between X_2 and X_3, VIF will be 1.

Using this definition, we can express (7.4.12) and (7.4.15) as

$$\text{var}\,(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2}\text{VIF} \tag{10.5.2}$$

$$\text{var}\,(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2}\text{VIF} \tag{10.5.3}$$

which show that the variances of $\hat{\beta}_2$ and $\hat{\beta}_3$ are directly proportional to the VIF.

To give some idea about how fast the variances and covariances increase as r_{23} increases, consider Table 10.1, which gives these variances and covariances for selected values of r_{23}. As this table shows, increases in r_{23}

TABLE 10.1 THE EFFECT OF INCREASING r_{23} ON VAR $(\hat{\beta}_2)$ AND COV $(\hat{\beta}_2, \hat{\beta}_3)$

Value of r_{23} (1)	VIF (2)	var $(\hat{\beta}_2)$ (3)*	$\frac{\text{var}\,(\hat{\beta}_2)(r_{23} \neq 0)}{\text{var}\,(\hat{\beta}_2)(r_{23} = 0)}$ (4)	cov $(\hat{\beta}_2, \hat{\beta}_3)$ (5)
0.00	1.00	$\frac{\sigma^2}{\sum x_{2i}^2} = A$	—	0
0.50	1.33	$1.33 \times A$	1.33	$0.67 \times B$
0.70	1.96	$1.96 \times A$	1.96	$1.37 \times B$
0.80	2.78	$2.78 \times A$	2.78	$2.22 \times B$
0.90	5.76	$5.26 \times A$	5.26	$4.73 \times B$
0.95	10.26	$10.26 \times A$	10.26	$9.74 \times B$
0.97	16.92	$16.92 \times A$	16.92	$16.41 \times B$
0.99	50.25	$50.25 \times A$	50.25	$49.75 \times B$
0.995	100.00	$100.00 \times A$	100.00	$99.50 \times B$
0.999	500.00	$500.00 \times A$	500.00	$499.50 \times B$

Note: $A = \frac{\sigma^2}{\sum x_{2i}^2}$

$B = \frac{-\sigma^2}{\sqrt{\sum x_{2i}^2 \sum x_{3i}^3}}$

$\times$ = times

*To find out the effect of increasing r_{23} on var $(\hat{\beta}_3)$, note that $A = \sigma^2 / \sum x_{3i}^2$ when $r_{23} = 0$, but the variance and covariance magnifying factors remain the same.

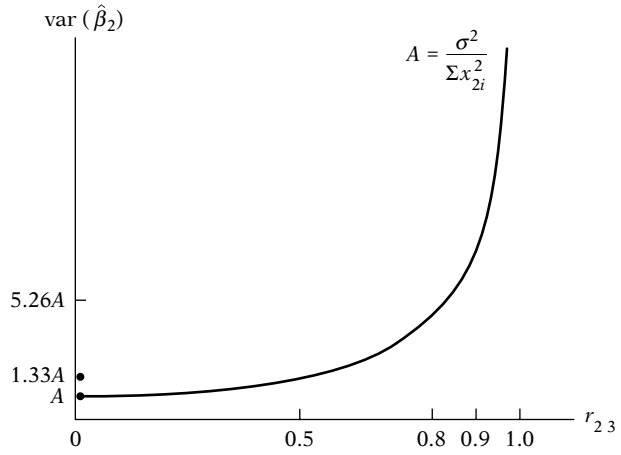

FIGURE 10.2 The behavior of var ($\hat{\beta}_2$) as a function of r_{23}.

have a dramatic effect on the estimated variances and covariances of the OLS estimators. When $r_{23} = 0.50$, the var ($\hat{\beta}_2$) is 1.33 times the variance when r_{23} is zero, but by the time r_{23} reaches 0.95 it is about 10 times as high as when there is no collinearity. And lo and behold, an increase of r_{23} from 0.95 to 0.995 makes the estimated variance 100 times that when collinearity is zero. The same dramatic effect is seen on the estimated covariance. All this can be seen in Figure 10.2.

The results just discussed can be easily extended to the k-variable model. In such a model, the variance of the kth coefficient, as noted in (7.5.6), can be expressed as:

$$\text{var}\,(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2}\left(\frac{1}{1 - R_j^2}\right) \qquad \textbf{(7.5.6)}$$

where $\hat{\beta}_j$ = (estimated) partial regression coefficient of regressor X_j
$R_j^2 = R^2$ in the regression of X_j on the remaining $(k - 2)$ regressions [*Note:* There are $(k - 1)$ regressors in the k-variable regression model.]
$\sum x_j^2 = \sum (X_j - \bar{X}_j)^2$

We can also write (7.5.6) as

$$\text{var}\,(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2}\text{VIF}_j \qquad \textbf{(10.5.4)}$$

As you can see from this expression, var ($\hat{\beta}_j$) is proportional to σ^2 and VIF but inversely proportional to $\sum x_j^2$. Thus, whether var ($\hat{\beta}_j$) is large or small

will depend on the three ingredients: (1) σ^2, (2) VIF, and (3) $\sum x_j^2$. The last one, which ties in with Assumption 8 of the classical model, states that the larger the variability in a regressor, the smaller the variance of the coefficient of that regressor, assuming the other two ingredients are constant, and therefore the greater the precision with which that coefficient can be estimated.

Before proceeding further, it may be noted that the inverse of the VIF is called **tolerance** (TOL). That is,

$$\text{TOL}_j = \frac{1}{\text{VIF}_j} = (1 - R_j^2) \tag{10.5.5}$$

When $R_j^2 = 1$ (i.e., perfect collinearity), $\text{TOL}_j = 0$ and when $R_j^2 = 0$ (i.e., no collinearity whatsoever), TOL_j is 1. Because of the intimate connection between VIF and TOL, one can use them interchangeably.

Wider Confidence Intervals

Because of the large standard errors, the confidence intervals for the relevant population parameters tend to be larger, as can be seen from Table 10.2. For example, when $r_{23} = 0.95$, the confidence interval for β_2 is larger than when $r_{23} = 0$ by a factor of $\sqrt{10.26}$, or about 3.

Therefore, in cases of high multicollinearity, the sample data may be compatible with a diverse set of hypotheses. Hence, the probability of accepting a false hypothesis (i.e., type II error) increases.

TABLE 10.2 THE EFFECT OF INCREASING COLLINEARITY ON THE 95% CONFIDENCE INTERVAL FOR β_2: $\hat{\beta}_2 \pm 1.96\ \text{se}(\hat{\beta}_2)$

Value of r_{23}	95% confidence interval for β_2
0.00	$\hat{\beta}_2 \pm 1.96\sqrt{\dfrac{\sigma^2}{\sum x_{2i}^2}}$
0.50	$\hat{\beta}_2 \pm 1.96\sqrt{(1.33)}\sqrt{\dfrac{\sigma^2}{\sum x_{2i}^2}}$
0.95	$\hat{\beta}_2 \pm 1.96\sqrt{(10.26)}\sqrt{\dfrac{\sigma^2}{\sum x_{2i}^2}}$
0.995	$\hat{\beta}_2 \pm 1.96\sqrt{(100)}\sqrt{\dfrac{\sigma^2}{\sum x_{2i}^2}}$
0.999	$\hat{\beta}_2 \pm 1.96\sqrt{(500)}\sqrt{\dfrac{\sigma^2}{\sum x_{2i}^2}}$

Note: We are using the normal distribution because σ^2 is assumed for convenience to be known. Hence the use of 1.96, the 95% confidence factor for the normal distribution.

The standard errors corresponding to the various r_{23} values are obtained from Table 10.1.

"Insignificant" *t* Ratios

Recall that to test the null hypothesis that, say, $\beta_2 = 0$, we use the t ratio, that is, $\hat{\beta}_2/\text{se}(\hat{\beta}_2)$, and compare the estimated t value with the critical t value from the t table. But as we have seen, in cases of high collinearity the estimated standard errors increase dramatically, thereby making the t values smaller. Therefore, in such cases, one will increasingly accept the null hypothesis that the relevant true population value is zero.[13]

A High R^2 but Few Significant *t* Ratios

Consider the k-variable linear regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i$$

In cases of high collinearity, it is possible to find, as we have just noted, that one or more of the partial slope coefficients are individually statistically insignificant on the basis of the t test. Yet the R^2 in such situations may be so high, say, in excess of 0.9, that on the basis of the F test one can convincingly reject the hypothesis that $\beta_2 = \beta_3 = \cdots = \beta_k = 0$. Indeed, this is one of the signals of multicollinearity—insignificant t values but a high overall R^2 (and a significant F value)!

We shall demonstrate this signal in the next section, but this outcome should not be surprising in view of our discussion on individual vs. joint testing in Chapter 8. As you may recall, the real problem here is the covariances between the estimators, which, as formula (7.4.17) indicates, are related to the correlations between the regressors.

Sensitivity of OLS Estimators and Their Standard Errors to Small Changes in Data

As long as multicollinearity is not perfect, estimation of the regression coefficients is possible but the estimates and their standard errors become very sensitive to even the slightest change in the data.

To see this, consider Table 10.3. Based on these data, we obtain the following multiple regression:

$$\begin{aligned} \hat{Y}_i &= 1.1939 + 0.4463X_{2i} + 0.0030X_{3i} \\ &\quad (0.7737) \quad (0.1848) \quad (0.0851) \\ t &= (1.5431) \quad (2.4151) \quad (0.0358) \qquad \textbf{(10.5.6)} \\ &\qquad R^2 = 0.8101 \qquad r_{23} = 0.5523 \\ &\qquad \text{cov}(\hat{\beta}_2, \hat{\beta}_3) = -0.00868 \qquad \text{df} = 2 \end{aligned}$$

[13] In terms of the confidence intervals, $\beta_2 = 0$ value will lie increasingly in the acceptance region as the degree of collinearity increases.

TABLE 10.3
HYPOTHETICAL DATA ON Y, X_2, AND X_3

Y	X_2	X_3
1	2	4
2	0	2
3	4	12
4	6	0
5	8	16

TABLE 10.4
HYPOTHETICAL DATA ON Y, X_2, AND X_3

Y	X_2	X_3
1	2	4
2	0	2
3	4	0
4	6	12
5	8	16

Regression (10.5.6) shows that none of the regression coefficients is individually significant at the conventional 1 or 5 percent levels of significance, although $\hat{\beta}_2$ is significant at the 10 percent level on the basis of a one-tail t test.

Now consider Table 10.4. The only difference between Tables 10.3 and 10.4 is that the third and fourth values of X_3 are interchanged. Using the data of Table 10.4, we now obtain

$$\begin{aligned} \hat{Y}_i &= 1.2108 + 0.4014X_{2i} + 0.0270X_{3i} \\ &\quad (0.7480) \quad (0.2721) \quad (0.1252) \\ t &= (1.6187) \quad (1.4752) \quad (0.2158) \qquad \textbf{(10.5.7)} \\ &\qquad R^2 = 0.8143 \qquad r_{23} = 0.8285 \\ &\qquad \operatorname{cov}(\hat{\beta}_2, \hat{\beta}_3) = -0.0282 \qquad \text{df} = 2 \end{aligned}$$

As a result of a slight change in the data, we see that $\hat{\beta}_2$, which was statistically significant before at the 10 percent level of significance, is no longer significant even at that level. Also note that in (10.5.6) $\operatorname{cov}(\hat{\beta}_2, \hat{\beta}_3) = -0.00868$ whereas in (10.5.7) it is -0.0282, a more than threefold increase. All these changes may be attributable to increased multicollinearity: In (10.5.6) $r_{23} = 0.5523$, whereas in (10.5.7) it is 0.8285. Similarly, the standard errors of $\hat{\beta}_2$ and $\hat{\beta}_3$ increase between the two regressions, a usual symptom of collinearity.

We noted earlier that in the presence of high collinearity one cannot estimate the individual regression coefficients precisely but that linear combinations of these coefficients may be estimated more precisely. This fact can be substantiated from the regressions (10.5.6) and (10.5.7). In the first regression the sum of the two partial slope coefficients is 0.4493 and in the second it is 0.4284, practically the same. Not only that, their standard errors are practically the same, 0.1550 vs. 0.1823.[14] Note, however, the coefficient of X_3 has changed dramatically, from 0.003 to 0.027.

[14]These standard errors are obtained from the formula

$$\operatorname{se}(\hat{\beta}_2 + \hat{\beta}_3) = \sqrt{\operatorname{var}(\hat{\beta}_2) + \operatorname{var}(\hat{\beta}_3) + 2\operatorname{cov}(\hat{\beta}_2, \hat{\beta}_3)}$$

Note that increasing collinearity increases the variances of $\hat{\beta}_2$ and $\hat{\beta}_3$, but these variances may be offset if there is high negative covariance between the two, as our results clearly point out.

Consequences of Micronumerosity

In a parody of the consequences of multicollinearity, and in a tongue-in-cheek manner, Goldberger cites exactly similar consequences of micronumerosity, that is, analysis based on small sample size.[15] The reader is advised to read Goldberger's analysis to see why he regards micronumerosity as being as important as multicollinearity.

10.6 AN ILLUSTRATIVE EXAMPLE: CONSUMPTION EXPENDITURE IN RELATION TO INCOME AND WEALTH

To illustrate the various points made thus far, let us reconsider the consumption–income example of Chapter 3. In Table 10.5 we reproduce the data of Table 3.2 and add to it data on wealth of the consumer. If we assume that consumption expenditure is linearly related to income and wealth, then, from Table 10.5 we obtain the following regression:

$$\begin{aligned} \hat{Y}_i &= 24.7747 + 0.9415X_{2i} - 0.0424X_{3i} \\ &\quad (6.7525) \quad (0.8229) \quad (0.0807) \\ t &= (3.6690) \quad (1.1442) \quad (-0.5261) \\ R^2 &= 0.9635 \qquad \bar{R}^2 = 0.9531 \qquad \text{df} = 7 \end{aligned} \tag{10.6.1}$$

Regression (10.6.1) shows that income and wealth together explain about 96 percent of the variation in consumption expenditure, and yet neither of the slope coefficients is individually statistically significant. Moreover, not only is the wealth variable statistically insignificant but also it has the wrong

TABLE 10.5 HYPOTHETICAL DATA ON CONSUMPTION EXPENDITURE Y, INCOME X_2, AND WEALTH X_3

Y, \$	X_2, \$	X_3, \$
70	80	810
65	100	1009
90	120	1273
95	140	1425
110	160	1633
115	180	1876
120	200	2052
140	220	2201
155	240	2435
150	260	2686

[15] Goldberger, op. cit., pp. 248–250.

TABLE 10.6 ANOVA TABLE FOR THE CONSUMPTION–INCOME–WEALTH EXAMPLE

Source of variation	SS	df	MSS
Due to regression	8,565.5541	2	4,282.7770
Due to residual	324.4459	7	46.3494

sign. A priori, one would expect a positive relationship between consumption and wealth. Although $\hat{\beta}_2$ and $\hat{\beta}_3$ are individually statistically insignificant, if we test the hypothesis that $\beta_2 = \beta_3 = 0$ simultaneously, this hypothesis can be rejected, as Table 10.6 shows. Under the usual assumption we obtain

$$F = \frac{4282.7770}{46.3494} = 92.4019 \qquad \textbf{(10.6.2)}$$

This F value is obviously highly significant.

It is interesting to look at this result geometrically. (See Figure 10.3.) Based on the regression (10.6.1), we have established the individual 95% confidence intervals for β_2 and β_3 following the usual procedure discussed in Chapter 8. As these intervals show, individually each of them includes the value of zero. Therefore, *individually* we can accept the hypothesis that the

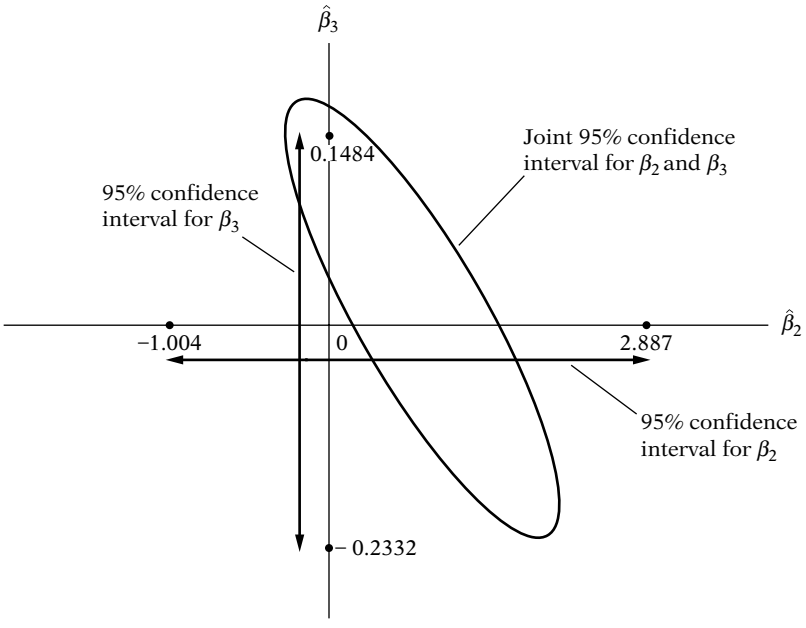

FIGURE 10.3 Individual confidence intervals for β_2 and β_3 and joint confidence interval (ellipse) for β_2 and β_3.

two partial slopes are zero. But, when we establish the joint confidence interval to test the hypothesis that $\beta_2 = \beta_3 = 0$, that hypothesis cannot be accepted since the joint confidence interval, actually an ellipse, does not include the origin.[16] As already pointed out, when collinearity is high, tests on individual regressors are not reliable; in such cases it is the overall F test that will show if Y is related to the various regressors.

Our example shows dramatically what multicollinearity does. The fact that the F test is significant but the t values of X_2 and X_3 are individually insignificant means that the two variables are so highly correlated that it is impossible to isolate the individual impact of either income or wealth on consumption. As a matter of fact, if we regress X_3 on X_2, we obtain

$$\begin{aligned} \hat{X}_{3i} &= \underset{}{7.5454} + \underset{}{10.1909X_{2i}} \\ &\quad\ (29.4758) \quad (0.1643) \\ t &= (0.2560) \quad (62.0405) \qquad R^2 = 0.9979 \end{aligned} \tag{10.6.3}$$

which shows that there is almost perfect collinearity between X_3 and X_2.

Now let us see what happens if we regress Y on X_2 only:

$$\begin{aligned} \hat{Y}_i &= 24.4545 + 0.5091X_{2i} \\ &\quad\ (6.4138) \quad (0.0357) \\ t &= (3.8128) \quad (14.2432) \qquad R^2 = 0.9621 \end{aligned} \tag{10.6.4}$$

In (10.6.1) the income variable was statistically insignificant, whereas now it is highly significant. If instead of regressing Y on X_2, we regress it on X_3, we obtain

$$\begin{aligned} \hat{Y}_i &= 24.411 + 0.0498X_{3i} \\ &\quad\ (6.874) \quad (0.0037) \\ t &= (3.551) \quad (13.29) \qquad R^2 = 0.9567 \end{aligned} \tag{10.6.5}$$

We see that wealth has now a significant impact on consumption expenditure, whereas in (10.6.1) it had no effect on consumption expenditure.

Regressions (10.6.4) and (10.6.5) show very clearly that in situations of extreme multicollinearity dropping the highly collinear variable will often make the other X variable statistically significant. This result would suggest that a way out of extreme collinearity is to drop the collinear variable, but we shall have more to say about it in Section 10.8.

[16] As noted in Sec. 5.3, the topic of joint confidence interval is rather involved. The interested reader may consult the reference cited there.

10.7 DETECTION OF MULTICOLLINEARITY

Having studied the nature and consequences of multicollinearity, the natural question is: How does one know that collinearity is present in any given situation, especially in models involving more than two explanatory variables? Here it is useful to bear in mind Kmenta's warning:

> **1.** Multicollinearity is a question of degree and not of kind. The meaningful distinction is not between the presence and the absence of multicollinearity, but between its various degrees.
>
> **2.** Since multicollinearity refers to the condition of the explanatory variables that are assumed to be nonstochastic, it is a feature of the sample and not of the population.
>
> Therefore, we do not "test for multicollinearity" but can, if we wish, measure its degree in any particular sample.[17]

Since multicollinearity is essentially a sample phenomenon, arising out of the largely nonexperimental data collected in most social sciences, we do not have one unique method of detecting it or measuring its strength. What we have are some rules of thumb, some informal and some formal, but rules of thumb all the same. We now consider some of these rules.

1. High R^2 but few significant t ratios. As noted, this is the "classic" symptom of multicollinearity. If R^2 is high, say, in excess of 0.8, the F test in most cases will reject the hypothesis that the partial slope coefficients are simultaneously equal to zero, but the individual t tests will show that none or very few of the partial slope coefficients are statistically different from zero. This fact was clearly demonstrated by our consumption–income–wealth example.

Although this diagnostic is sensible, its disadvantage is that "it is too strong in the sense that multicollinearity is considered as harmful only when all of the influences of the explanatory variables on Y cannot be disentangled."[18]

2. High pair-wise correlations among regressors. Another suggested rule of thumb is that if the pair-wise or zero-order correlation coefficient between two regressors is high, say, in excess of 0.8, then multicollinearity is a serious problem. The problem with this criterion is that, although high zero-order correlations may suggest collinearity, it is not necessary that they be high to have collinearity in any specific case. To put the matter somewhat technically, *high zero-order correlations are a sufficient but not a necessary condition for the existence of multicollinearity because it can exist even though the zero-order or simple correlations are comparatively low* (say, less than 0.50). To see this relationship, suppose we have a four-variable model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

[17] Jan Kmenta, *Elements of Econometrics,* 2d ed., Macmillan, New York, 1986, p. 431.
[18] Ibid., p. 439.

and suppose that

$$X_{4i} = \lambda_2 X_{2i} + \lambda_3 X_{3i}$$

where λ_2 and λ_3 are constants, not both zero. Obviously, X_4 is an exact linear combination of X_2 and X_3, giving $R^2_{4.23} = 1$, the coefficient of determination in the regression of X_4 on X_2 and X_3.

Now recalling the formula (7.11.5) from Chapter 7, we can write

$$R^2_{4.23} = \frac{r^2_{42} + r^2_{43} - 2r_{42}r_{43}r_{23}}{1 - r^2_{23}} \tag{10.7.1}$$

But since $R^2_{4.23} = 1$ because of perfect collinearity, we obtain

$$1 = \frac{r^2_{42} + r^2_{43} - 2r_{42}r_{43}r_{23}}{1 - r^2_{23}} \tag{10.7.2}$$

It is not difficult to see that (10.7.2) is satisfied by $r_{42} = 0.5, r_{43} = 0.5$, and $r_{23} = -0.5$, which are not very high values.

Therefore, in models involving more than two explanatory variables, the simple or zero-order correlation will not provide an infallible guide to the presence of multicollinearity. Of course, if there are only two explanatory variables, the zero-order correlations will suffice.

3. Examination of partial correlations. Because of the problem just mentioned in relying on zero-order correlations, Farrar and Glauber have suggested that one should look at the partial correlation coefficients.[19] Thus, in the regression of Y on X_2, X_3, and X_4, a finding that $R^2_{1.234}$ is very high but $r^2_{12.34}, r^2_{13.24}$, and $r^2_{14.23}$ are comparatively low may suggest that the variables X_2, X_3, and X_4 are highly intercorrelated and that at least one of these variables is superfluous.

Although a study of the partial correlations may be useful, there is no guarantee that they will provide an infallible guide to multicollinearity, for it may happen that both R^2 and all the partial correlations are sufficiently high. But more importantly, C. Robert Wichers has shown[20] that the Farrar-Glauber partial correlation test is ineffective in that a given partial correlation may be compatible with different multicollinearity patterns. The Farrar–Glauber test has also been severely criticized by T. Krishna Kumar[21] and John O'Hagan and Brendan McCabe.[22]

[19] D. E. Farrar and R. R. Glauber, "Multicollinearity in Regression Analysis: The Problem Revisited," *Review of Economics and Statistics,* vol. 49, 1967, pp. 92–107.

[20] "The Detection of Multicollinearity: A Comment," *Review of Economics and Statistics,* vol. 57, 1975, pp. 365–366.

[21] "Multicollinearity in Regression Analysis," *Review of Economics and Statistics,* vol. 57, 1975, pp. 366–368.

[22] "Tests for the Severity of Multicollinearity in Regression Analysis: A Comment," *Review of Economics and Statistics,* vol. 57, 1975, pp. 368–370.

4. Auxiliary regressions. Since multicollinearity arises because one or more of the regressors are exact or approximately linear combinations of the other regressors, one way of finding out which X variable is related to other X variables is to regress each X_i on the remaining X variables and compute the corresponding R^2, which we designate as R_i^2; each one of these regressions is called an **auxiliary regression,** auxiliary to the main regression of Y on the X's. Then, following the relationship between F and R^2 established in (8.5.11), the variable

$$F_i = \frac{R^2_{x_i \cdot x_2 x_3 \cdots x_k} / (k-2)}{\left(1 - R^2_{x_i \cdot x_2 x_3 \cdots x_k}\right) / (n-k+1)} \qquad \textbf{(10.7.3)}$$

follows the F distribution with $k-2$ and $n-k+1$ df. In Eq. (10.7.3) n stands for the sample size, k stands for the number of explanatory variables including the intercept term, and $R^2_{x_i \cdot x_2 x_3 \cdots x_k}$ is the coefficient of determination in the regression of variable X_i on the remaining X variables.[23]

If the computed F exceeds the critical F_i at the chosen level of significance, it is taken to mean that the particular X_i is collinear with other X's; if it does not exceed the critical F_i, we say that it is not collinear with other X's, in which case we may retain that variable in the model. If F_i is statistically significant, we will still have to decide whether the particular X_i should be dropped from the model. This question will be taken up in Section 10.8.

But this method is not without its drawbacks, for

> . . . if the multicollinearity involves only a few variables so that the auxiliary regressions do not suffer from extensive multicollinearity, the estimated coefficients may reveal the nature of the linear dependence among the regressors. Unfortunately, if there are several complex linear associations, this curve fitting exercise may not prove to be of much value as it will be difficult to identify the separate interrelationships.[24]

Instead of formally testing all auxiliary R^2 values, one may adopt **Klien's rule of thumb,** which suggests that multicollinearity may be a troublesome problem only if the R^2 obtained from an auxiliary regression is greater than the overall R^2, that is, that obtained from the regression of Y on all the regressors.[25] Of course, like all other rules of thumb, this one should be used judiciously.

5. Eigenvalues and condition index. If you examine the SAS output of the Cobb–Douglas production function given in Appendix 7A.5 you will see

[23]For example, $R^2_{x_2}$ can be obtained by regressing X_{2i} as follows: $X_{2i} = a_1 + a_3 X_{3i} + a_4 X_{4i} + \cdots + a_k X_{ki} + \hat{u}_i$.

[24]George G. Judge, R. Carter Hill, William E. Griffiths, Helmut Lütkepohl, and Tsoung-Chao Lee, *Introduction to the Theory and Practice of Econometrics*, John Wiley & Sons, New York, 1982, p. 621.

[25]Lawrence R. Klien, *An Introduction to Econometrics,* Prentice-Hall, Englewood Cliffs, N.J., 1962, p. 101.

that SAS uses *eigenvalues* and the *condition index* to diagnose multicollinearity. We will not discuss eigenvalues here, for that would take us into topics in matrix algebra that are beyond the scope of this book. From these eigenvalues, however, we can derive what is known as the **condition number *k*** defined as

$$k = \frac{\text{Maximum eigenvalue}}{\text{Minimum eigenvalue}}$$

and the **condition index (CI)** defined as

$$\text{CI} = \sqrt{\frac{\text{Maximum eigenvalue}}{\text{Minimum eigenvalue}}} = \sqrt{k}$$

Then we have this rule of thumb. If k is between 100 and 1000 there is moderate to strong multicollinearity and if it exceeds 1000 there is severe multicollinearity. Alternatively, if the CI ($= \sqrt{k}$) is between 10 and 30, there is moderate to strong multicollinearity and if it exceeds 30 there is severe multicollinearity.

For the illustrative example, $k = 3.0/0.00002422$ or about 123,864, and CI $= \sqrt{123,864} =$ about 352; both k and the CI therefore suggest severe multicollinearity. Of course, k and CI can be calculated between the maximum eigenvalue and any other eigenvalue, as is done in the printout. (*Note:* The printout does not explicitly compute k, but that is simply the square of CI.) Incidentally, note that a low eigenvalue (in relation to the maximum eigenvalue) is generally an indication of near-linear dependencies in the data.

Some authors believe that the condition index is the best available multicollinearity diagnostic. But this opinion is not shared widely. For us, then, the CI is just a rule of thumb, a bit more sophisticated perhaps. But for further details, the reader may consult the references.[26]

6. Tolerance and variance inflation factor. We have already introduced TOL and VIF. As R_j^2, the coefficient of determination in the regression of regressor X_j on the remaining regressors in the model, increases toward unity, that is, as the collinearity of X_j with the other regressors increases, VIF also increases and in the limit it can be infinite.

Some authors therefore use the VIF as an indicator of multicollinearity. The larger the value of VIF_j, the more "troublesome" or collinear the variable X_j. **As a rule of thumb,** if the VIF of a variable exceeds 10, which will happen if R_j^2 exceeds 0.90, that variable is said be highly collinear.[27]

Of course, one could use TOL_j as a measure of multicollinearity in view of its intimate connection with VIF_j. The closer is TOL_j to zero, the greater the degree of collinearity of that variable with the other regressors. On the

[26]See especially D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity,* John Wiley & Sons, New York, 1980, Chap. 3. However, this book is not for the beginner.

[27]See David G. Kleinbaum, Lawrence L. Kupper, and Keith E. Muller, *Applied Regression Analysis and other Multivariate Methods,* 2d ed., PWS-Kent, Boston, Mass., 1988, p. 210.

other hand, the closer TOL_j is to 1, the greater the evidence that X_j is not collinear with the other regressors.

VIF (or tolerance) as a measure of collinearity is not free of criticism. As (10.5.4) shows, var $(\hat{\beta}_j)$ depends on three factors: σ^2, $\sum x_j^2$, and VIF_j. A high VIF can be counterbalanced by a low σ^2 or a high $\sum x_j^2$. To put it differently, a high VIF is neither necessary nor sufficient to get high variances and high standard errors. Therefore, high multicollinearity, as measured by a high VIF, may not necessarily cause high standard errors. In all this discussion, the terms *high* and *low* are used in a relative sense.

To conclude our discussion of detecting multicollinearity, we stress that the various methods we have discussed are essentially in the nature of "fishing expeditions," for we cannot tell which of these methods will work in any particular application. Alas, not much can be done about it, for multicollinearity is specific to a given sample over which the researcher may not have much control, especially if the data are nonexperimental in nature—the usual fate of researchers in the social sciences.

Again as a parody of multicollinearity, Goldberger cites numerous ways of detecting micronumerosity, such as developing critical values of the sample size, n^*, such that micronumerosity is a problem only if the actual sample size, n, is smaller than n^*. The point of Goldberger's parody is to emphasize that small sample size and lack of variability in the explanatory variables may cause problems that are at least as serious as those due to multicollinearity.

10.8 REMEDIAL MEASURES

What can be done if multicollinearity is serious? We have two choices: (1) do nothing or (2) follow some rules of thumb.

Do Nothing

The "do nothing" school of thought is expressed by Blanchard as follows[28]:

> When students run their first ordinary least squares (OLS) regression, the first problem that they usually encounter is that of multicollinearity. Many of them conclude that there is something wrong with OLS; some resort to new and often creative techniques to get around the problem. But, we tell them, this is wrong. Multicollinearity is God's will, not a problem with OLS or statistical technique in general.

What Blanchard is saying is that multicollinearity is essentially a data deficiency problem (micronumerosity, again) and some times we have no choice over the data we have available for empirical analysis.

Also, it is not that all the coefficients in a regression model are statistically insignificant. Moreover, even if we cannot estimate one or more regression coefficients with greater precision, a linear combination of them (i.e., estimable function) can be estimated relatively efficiently. As we saw in

[28]Blanchard, O. J., Comment, *Journal of Business and Economic Statistics,* vol. 5, 1967, pp. 449–451. The quote is reproduced from Peter Kennedy, *A Guide to Econometrics,* 4th ed., MIT Press, Cambridge, Mass., 1998, p. 190.

(10.2.3), we can estimate α uniquely, even if we cannot estimate its two components given there individually. Sometimes this is the best we can do with a given set of data.[29]

Rule-of-Thumb Procedures

One can try the following rules of thumb to address the problem of multicollinearity, the success depending on the severity of the collinearity problem.

1. A priori information. Suppose we consider the model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

where $Y =$ consumption, $X_2 =$ income, and $X_3 =$ wealth. As noted before, income and wealth variables tend to be highly collinear. But suppose a priori we believe that $\beta_3 = 0.10\beta_2$; that is, the rate of change of consumption with respect to wealth is one-tenth the corresponding rate with respect to income. We can then run the following regression:

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_{2i} + 0.10\beta_2 X_{3i} + u_i \\ &= \beta_1 + \beta_2 X_i + u_i \end{aligned}$$

where $X_i = X_{2i} + 0.1X_{3i}$. Once we obtain $\hat{\beta}_2$, we can estimate $\hat{\beta}_3$ from the postulated relationship between β_2 and β_3.

How does one obtain a priori information? It could come from previous empirical work in which the collinearity problem happens to be less serious or from the relevant theory underlying the field of study. For example, in the Cobb–Douglas–type production function (7.9.1), if one expects constant returns to scale to prevail, then $(\beta_2 + \beta_3) = 1$, in which case we could run the regression (8.7.14), regressing the output-labor ratio on the capital-labor ratio. If there is collinearity between labor and capital, as generally is the case in most sample data, such a transformation may reduce or eliminate the collinearity problem. But a warning is in order here regarding imposing such a priori restrictions, "... since in general we will want to test economic theory's a priori predictions rather than simply impose them on data for which they may not be true."[30] However, we know from Section 8.7 how to test for the validity of such restrictions explicitly.

2. Combining cross-sectional and time series data. A variant of the extraneous or a priori information technique is the combination of cross-sectional and time-series data, known as *pooling the data.* Suppose we want

[29]For an interesting discussion on this, see Conlisk, J., "When Collinearity is Desirable," *Western Economic Journal,* vol. 9, 1971, pp. 393–407.

[30]Mark B. Stewart and Kenneth F. Wallis, *Introductory Econometrics,* 2d ed., John Wiley & Sons, A Halstead Press Book, New York, 1981, p. 154.

to study the demand for automobiles in the United States and assume we have time series data on the number of cars sold, average price of the car, and consumer income. Suppose also that

$$\ln Y_t = \beta_1 + \beta_2 \ln P_t + \beta_3 \ln I_t + u_t$$

where Y = number of cars sold, P = average price, I = income, and t = time. Out objective is to estimate the price elasticity β_2 and income elasticity β_3.

In time series data the price and income variables generally tend to be highly collinear. Therefore, if we run the preceding regression, we shall be faced with the usual multicollinearity problem. A way out of this has been suggested by Tobin.[31] He says that if we have cross-sectional data (for example, data generated by consumer panels, or budget studies conducted by various private and governmental agencies), we can obtain a fairly reliable estimate of the income elasticity β_3 because in such data, which are at a point in time, the prices do not vary much. Let the cross-sectionally estimated income elasticity be $\hat{\beta}_3$. Using this estimate, we may write the preceding time series regression as

$$Y_t^* = \beta_1 + \beta_2 \ln P_t + u_t$$

where $Y^* = \ln Y - \hat{\beta}_3 \ln I$, that is, Y^* represents that value of Y after removing from it the effect of income. We can now obtain an estimate of the price elasticity β_2 from the preceding regression.

Although it is an appealing technique, pooling the time series and cross-sectional data in the manner just suggested may create problems of interpretation, because we are assuming implicitly that the cross-sectionally estimated income elasticity is the same thing as that which would be obtained from a pure time series analysis.[32] Nonetheless, the technique has been used in many applications and is worthy of consideration in situations where the cross-sectional estimates do not vary substantially from one cross section to another. An example of this technique is provided in exercise 10.26.

3. Dropping a variable(s) and specification bias. When faced with severe multicollinearity, one of the "simplest" things to do is to drop one of the collinear variables. Thus, in our consumption–income–wealth illustration, when we drop the wealth variable, we obtain regression (10.6.4), which shows that, whereas in the original model the income variable was statistically insignificant, it is now "highly" significant.

But in dropping a variable from the model we may be committing a **specification bias** or **specification error.** Specification bias arises from

[31]J. Tobin, "A Statistical Demand Function for Food in the U.S.A.," *Journal of the Royal Statistical Society,* Ser. A, 1950, pp. 113–141.

[32]For a thorough discussion and application of the pooling technique, see Edwin Kuh, *Capital Stock Growth: A Micro-Econometric Approach*, North-Holland Publishing Company, Amsterdam, 1963, Chaps. 5 and 6.

incorrect specification of the model used in the analysis. Thus, if economic theory says that income and wealth should both be included in the model explaining the consumption expenditure, dropping the wealth variable would constitute specification bias.

Although we will discuss the topic of specification bias in Chapter 13, we caught a glimpse of it in Section 7.7. If, for example, the true model is

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

but we mistakenly fit the model

$$Y_i = b_1 + b_{12} X_{2i} + \hat{u}_i \tag{10.8.1}$$

then it can be shown that (see Appendix 13A.1)

$$E(b_{12}) = \beta_2 + \beta_3 b_{32} \tag{10.8.2}$$

where b_{32} = slope coefficient in the regression of X_3 on X_2. Therefore, it is obvious from (10.8.2) that b_{12} will be a biased estimate of β_2 as long as b_{32} is different from zero (it is assumed that β_3 is different from zero; otherwise there is no sense in including X_3 in the original model).[33] Of course, if b_{32} is zero, we have no multicollinearity problem to begin with. It is also clear from (10.8.2) that if both b_{32} and β_3 are positive (or both are negative), $E(b_{12})$ will be greater than β_2; hence, on the average b_{12} will overestimate β_2, leading to a positive bias. Similarly, if the product $b_{32}\beta_3$ is negative, on the average b_{12} will underestimate β_2, leading to a negative bias.

From the preceding discussion it is clear that dropping a variable from the model to alleviate the problem of multicollinearity may lead to the specification bias. Hence the remedy may be worse than the disease in some situations because, whereas multicollinearity may prevent precise estimation of the parameters of the model, omitting a variable may seriously mislead us as to the true values of the parameters. Recall that OLS estimators are BLUE despite near collinearity.

4. Transformation of variables. Suppose we have time series data on consumption expenditure, income, and wealth. One reason for high multicollinearity between income and wealth in such data is that over time both the variables tend to move in the same direction. One way of minimizing this dependence is to proceed as follows.

If the relation

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \tag{10.8.3}$$

[33] Note further that if b_{32} does not approach zero as the sample size is increased indefinitely, then b_{12} will be not only biased but also inconsistent.

holds at time t, it must also hold at time $t-1$ because the origin of time is arbitrary anyway. Therefore, we have

$$Y_{t-1} = \beta_1 + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + u_{t-1} \tag{10.8.4}$$

If we subtract (10.8.4) from (10.8.3), we obtain

$$Y_t - Y_{t-1} = \beta_2(X_{2t} - X_{2,t-1}) + \beta_3(X_{3t} - X_{3,t-1}) + v_t \tag{10.8.5}$$

where $v_t = u_t - u_{t-1}$. Equation (10.8.5) is known as the **first difference form** because we run the regression, not on the original variables, but on the differences of successive values of the variables.

The first difference regression model often reduces the severity of multicollinearity because, although the levels of X_2 and X_3 may be highly correlated, there is no a priori reason to believe that their differences will also be highly correlated.

As we shall see in the chapters on **time series econometrics,** an incidental advantage of the first-difference transformation is that it may make a nonstationary time series stationary. In those chapters we will see the importance of stationary time series. As noted in Chapter 1, loosely speaking, a time series, say, Y_t, is stationary if its mean and variance do not change systematically over time.

Another commonly used transformation in practice is the **ratio transformation.** Consider the model:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \tag{10.8.6}$$

where Y is consumption expenditure in real dollars, X_2 is GDP, and X_3 is total population. Since GDP and population grow over time, they are likely to be correlated. One "solution" to this problem is to express the model on a per capita basis, that is, by dividing (10.8.4) by X_3, to obtain:

$$\frac{Y_t}{X_{3t}} = \beta_1 \left(\frac{1}{X_{3t}}\right) + \beta_2 \left(\frac{X_{2t}}{X_{3t}}\right) + \beta_3 + \left(\frac{u_t}{X_{3t}}\right) \tag{10.8.7}$$

Such a transformation may reduce collinearity in the original variables.

But the first-difference or ratio transformations are not without problems. For instance, the error term v_t in (10.8.5) may not satisfy one of the assumptions of the classical linear regression model, namely, that the disturbances are serially uncorrelated. As we will see in Chapter 12, if the original disturbance term u_t is serially uncorrelated, the error term v_t obtained previously will in most cases be serially correlated. Therefore, the remedy may be worse than the disease. Moreover, there is a loss of one observation due to the differencing procedure, and therefore the degrees of freedom are

reduced by one. In a small sample, this could be a factor one would wish at least to take into consideration. Furthermore, the first-differencing procedure may not be appropriate in cross-sectional data where there is no logical ordering of the observations.

Similarly, in the ratio model (10.8.7), the error term

$$\left(\frac{u_t}{X_{3t}}\right)$$

will be heteroscedastic, if the original error term u_t is homoscedastic, as we shall see in Chapter 11. Again, the remedy may be worse than the disease of collinearity.

In short, one should be careful in using the first difference or ratio method of transforming the data to resolve the problem of multicollinearity.

5. Additional or new data. Since multicollinearity is a sample feature, it is possible that in another sample involving the same variables collinearity may not be so serious as in the first sample. Sometimes simply increasing the size of the sample (if possible) may attenuate the collinearity problem. For example, in the three-variable model we saw that

$$\operatorname{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2(1 - r_{23}^2)}$$

Now as the sample size increases, $\sum x_{2i}^2$ will generally increase. (Why?) Therefore, for any given r_{23}, the variance of $\hat{\beta}_2$ will decrease, thus decreasing the standard error, which will enable us to estimate β_2 more precisely.

As an illustration, consider the following regression of consumption expenditure Y on income X_2 and wealth X_3 based on 10 observations[34]:

$$\begin{aligned} \hat{Y}_i &= 24.377 \;+\; 0.8716X_{2i} - \;\; 0.0349X_{3i} \\ t &= (3.875) \quad (2.7726) \qquad (-1.1595) \qquad R^2 = 0.9682 \end{aligned} \tag{10.8.8}$$

The wealth coefficient in this regression not only has the wrong sign but is also statistically insignificant at the 5 percent level. But when the sample size was increased to 40 observations (micronumerosity?), the following results were obtained:

$$\begin{aligned} \hat{Y}_i &= \;\; 2.0907 \;+\; 0.7299X_{2i} + \;\; 0.0605X_{3i} \\ t &= (0.8713) \quad (6.0014) \qquad (2.0014) \qquad R^2 = 0.9672 \end{aligned} \tag{10.8.9}$$

Now the wealth coefficient not only has the correct sign but also is statistically significant at the 5 percent level.

[34] I am indebted to Albert Zucker for providing the results given in the following regressions.

Obtaining additional or "better" data is not always that easy, for as Judge et al. note:

> Unfortunately, economists seldom can obtain additional data without bearing large costs, much less choose the values of the explanatory variables they desire. In addition, when adding new variables in situations that are not controlled, we must be aware of adding observations that were generated by a process other than that associated with the original data set; that is, we must be sure that the economic structure associated with the new observations is the same as the original structure.[35]

6. Reducing collinearity in polynomial regressions. In Section 7.10 we discussed polynomial regression models. A special feature of these models is that the explanatory variable(s) appear with various powers. Thus, in the total cubic cost function involving the regression of total cost on output, (output)^2, and (output)^3, as in (7.10.4), the various output terms are going to be correlated, making it difficult to estimate the various slope coefficients precisely.[36] In practice though, it has been found that if the explanatory variable(s) are expressed in the deviation form (i.e., deviation from the mean value), multicollinearity is substantially reduced. But even then the problem may persist,[37] in which case one may want to consider techniques such as **orthogonal polynomials.**[38]

7. Other methods of remedying multicollinearity. Multivariate statistical techniques such as **factor analysis** and **principal components** or techniques such as **ridge regression** are often employed to "solve" the problem of multicollinearity. Unfortunately, these techniques are beyond the scope of this book, for they cannot be discussed competently without resorting to matrix algebra.[39]

10.9 IS MULTICOLLINEARITY NECESSARILY BAD? MAYBE NOT IF THE OBJECTIVE IS PREDICTION ONLY

It has been said that if the sole purpose of regression analysis is prediction or forecasting, then multicollinearity is not a serious problem because the higher the R^2, the better the prediction.[40] But this may be so ". . . as long as

[35]Judge et al., op. cit., p. 625. See also Sec. 10.9.

[36]As noted, since the relationship between X, X^2, and X^3 is nonlinear, polynomial regressions do not violate the assumption of no multicollinearity of the classical model, strictly speaking.

[37]See R. A. Bradley and S. S. Srivastava, "Correlation and Polynomial Regression," *American Statistician,* vol. 33, 1979, pp. 11–14.

[38]See Norman Draper and Harry Smith, *Applied Regression Analysis,* 2d ed., John Wiley & Sons, New York, 1981, pp. 266–274.

[39]A readable account of these techniques from an applied viewpoint can be found in Samprit Chatterjee and Bertram Price, *Regression Analysis by Example,* John Wiley & Sons, New York, 1977, Chaps. 7 and 8. See also H. D. Vinod, "A Survey of Ridge Regression and Related Techniques for Improvements over Ordinary Least Squares," *Review of Economics and Statistics,* vol. 60, February 1978, pp. 121–131.

[40]See R. C. Geary, "Some Results about Relations between Stochastic Variables: A Discussion Document," *Review of International Statistical Institute*, vol. 31, 1963, pp. 163–181.

the values of the explanatory variables for which predictions are desired obey the same near-exact linear dependencies as the original design [data] matrix X."[41] Thus, if in an estimated regression it was found that $X_2 = 2X_3$ approximately, then in a future sample used to forecast Y, X_2 should also be approximately equal to $2X_3$, a condition difficult to meet in practice (see footnote 35), in which case prediction will become increasingly uncertain.[42] Moreover, if the objective of the analysis is not only prediction but also reliable estimation of the parameters, serious multicollinearity will be a problem because we have seen that it leads to large standard errors of the estimators.

In one situation, however, multicollinearity may not pose a serious problem. This is the case when R^2 is high and the regression coefficients are individually significant as revealed by the higher t values. Yet, multicollinearity diagnostics, say, the condition index, indicate that there is serious collinearity in the data. When can such a situation arise? As Johnston notes:

> This can arise if individual coefficients happen to be numerically well in excess of the true value, so that the effect still shows up in spite of the inflated standard error and/or because the true value itself is so large that even an estimate on the downside still shows up as significant.[43]

10.10 AN EXTENDED EXAMPLE: THE LONGLEY DATA

We conclude this chapter by analyzing the data collected by Longley.[44] Although originally collected to assess the computational accuracy of least-squares estimates in several computer programs, the Longley data has become the workhorse to illustrate several econometric problems, including multicollinearity. The data are reproduced in Table 10.7. The data are time series for the years 1947–1962 and pertain to Y = number of people employed, in thousands; X_1 = GNP implicit price deflator; X_2 = GNP, millions of dollars; X_3 = number of people unemployed in thousands, X_4 = number of people in the armed forces, X_5 = noninstitutionalized population over 14 years of age; and X_6 = year, equal to 1 in 1947, 2 in 1948, and 16 in 1962.

---

[41] Judge et al., op. cit., p. 619. You will also find on this page proof of why, despite collinearity, one can obtain better mean predictions if the existing collinearity structure also continues in the future samples.

[42] For an excellent discussion, see E. Malinvaud, *Statistical Methods of Econometrics,* 2d ed., North-Holland Publishing Company, Amsterdam, 1970, pp. 220–221.

[43] J. Johnston, *Econometric Methods,* 3d ed., McGraw-Hill, New York, 1984, p. 249.

[44] Longley, J. "An Appraisal of Least-Squares Programs from the Point of the User," *Journal of the American Statistical Association,* vol. 62, 1967, pp. 819–841.

TABLE 10.7 LONGLEY DATA

Observation	y	X_1	X_2	X_3	X_4	X_5	Time
1947	60,323	830	234,289	2356	1590	107,608	1
1948	61,122	885	259,426	2325	1456	108,632	2
1949	60,171	882	258,054	3682	1616	109,773	3
1950	61,187	895	284,599	3351	1650	110,929	4
1951	63,221	962	328,975	2099	3099	112,075	5
1952	63,639	981	346,999	1932	3594	113,270	6
1953	64,989	990	365,385	1870	3547	115,094	7
1954	63,761	1000	363,112	3578	3350	116,219	8
1955	66,019	1012	397,469	2904	3048	117,388	9
1956	67,857	1046	419,180	2822	2857	118,734	10
1957	68,169	1084	442,769	2936	2798	120,445	11
1958	66,513	1108	444,546	4681	2637	121,950	12
1959	68,655	1126	482,704	3813	2552	123,366	13
1960	69,564	1142	502,601	3931	2514	125,368	14
1961	69,331	1157	518,173	4806	2572	127,852	15
1962	70,551	1169	554,894	4007	2827	130,081	16

Source: See footnote 44.

Assume that our objective is to predict Y on the basis of the six X variables. Using Eviews3, we obtain the following regression results:

Dependent Variable: *Y*
Sample: 1947–1962

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	-3482259.	890420.4	-3.910803	0.0036
X_1	15.06187	84.91493	0.177376	0.8631
X_2	-0.035819	0.033491	-1.069516	0.3127
X_3	-2.020230	0.488400	-4.136427	0.0025
X_4	-1.033227	0.214274	-4.821985	0.0009
X_5	-0.051104	0.226073	-0.226051	0.8262
X_6	1829.151	455.4785	4.015890	0.0030

R-squared	0.995479	Mean dependent var	65317.00
Adjusted R-squared	0.992465	S.D. dependent var	3511.968
S.E. of regression	304.8541	Akaike info criterion	14.57718
Sum squared resid	836424.1	Schwarz criterion	14.91519
Log likelihood	-109.6174	*F*-statistic	330.2853
Durbin-Watson stat	2.559488	Prob(*F*-statistic)	0.000000

A glance at these results would suggest that we have the collinearity problem, for the R^2 value is very high, but quite a few variables are statistically insignificant (X_1, X_2, and X_5), a classic symptom of multicollinearity. To shed more light on this, we show in Table 10.8 the intercorrelations among the six regressors.

TABLE 10.8 INTERCORRELATIONS

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1.000000	0.991589	0.620633	0.464744	0.979163	0.991149
X_2	0.991589	1.000000	0.604261	0.446437	0.991090	0.995273
X_3	0.620633	0.604261	1.000000	−0.177421	0.686552	0.668257
X_4	0.464744	0.446437	−0.177421	1.000000	0.364416	0.417245
X_5	0.979163	0.991090	0.686552	0.364416	1.000000	0.993953
X_6	0.991149	0.995273	0.668257	0.417245	0.993953	1.000000

This table gives what is called the **correlation matrix.** In this table the entries on the main diagonal (those running from the upper left-hand corner to the lower right-hand corner) give the correlation of one variable with itself, which is always 1 by definition, and the entries off the main diagonal are the pair-wise correlations among the X variables. If you take the first row of this table, this gives the correlation of X_1 with the other X variables. For example, 0.991589 is the correlation between X_1 and X_2, 0.620633 is the correlation between X_1 and X_3, and so on.

As you can see, several of these pair-wise correlations are quite high, suggesting that there may be a severe collinearity problem. Of course, remember the warning given earlier that such pair-wise correlations may be a sufficient but not a necessary condition for the existence of multicollinearity.

To shed further light on the nature of the multicollinearity problem, let us run the auxiliary regressions, that is the regression of each X variable on the remaining X variables. To save space, we will present only the R^2 values obtained from these regressions, which are given in Table 10.9. Since the R^2 values in the auxiliary regressions are very high (with the possible exception of the regression of X_4) on the remaining X variables, it seems that we do have a serious collinearity problem. The same information is obtained from the tolerance factors. As noted previously, the closer the tolerance factor is to zero, the greater is the evidence of collinearity.

Applying Klein's rule of thumb, we see that the R^2 values obtained from the auxiliary regressions exceed the overall R^2 value (that is the one obtained from the regression of Y on all the X variables) of 0.9954 in 3 out of

TABLE 10.9 R^2 VALUES FROM THE AUXILIARY REGRESSIONS

Dependent variable	R^2 value	Tolerance (TOL) $= 1 - R^2$
X_1	0.9926	0.0074
X_2	0.9994	0.0006
X_3	0.9702	0.0298
X_4	0.7213	0.2787
X_5	0.9970	0.0030
X_6	0.9986	0.0014

6 auxiliary regressions, again suggesting that indeed the Longley data are plagued by the multicollinearity problem. Incidentally, applying the F test given in (10.7.3) the reader should verify that the R^2 values given in the preceding tables are all statistically significantly different from zero.

We noted earlier that the OLS estimators and their standard errors are sensitive to small changes in the data. In exercise 10.32 the reader is asked to rerun the regression of Y on all the six X variables but drop the last data observations, that is, run the regression for the period 1947–1961. You will see how the regression results change by dropping just a single year's observations.

Now that we have established that we have the multicollinearity problem, what "remedial" actions can we take? Let us reconsider our original model. First of all, we could express GNP not in nominal terms, but in real terms, which we can do by dividing nominal GNP by the implicit price deflator. Second, since noninstitutional population over 14 years of age grows over time because of natural population growth, it will be highly correlated with time, the variable X_6 in our model. Therefore, instead of keeping both these variables, we will keep the variable X_5 and drop X_6. Third, there is no compelling reason to include X_3, the number of people unemployed; perhaps the unemployment rate would have been a better measure of labor market conditions. But we have no data on the latter. So, we will drop the variable X_3. Making these changes, we obtain the following regression results (RGNP = real GNP)[45]:

Dependent Variable: *Y*
Sample: 1947–1962

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	65720.37	10624.81	6.185558	0.0000
RGNP	9.736496	1.791552	5.434671	0.0002
X_4	-0.687966	0.322238	-2.134965	0.0541
X_5	-0.299537	0.141761	-2.112965	0.0562

R-squared	0.981404	Mean dependent var	65317.00
Adjusted *R*-squared	0.976755	S.D. dependent var	3511.968
S.E. of regression	535.4492	Akaike info criterion	15.61641
Sum squared resid	3440470.	Schwarz criterion	15.80955
Log likelihood	-120.9313	*F*-statistic	211.0972
Durbin-Watson stat	1.654069	Prob(*F*-statistic)	0.000000

Although the R^2 value has declined slightly compared with the original R^2, it is still very high. Now all the estimated coefficients are significant and the signs of the coefficients make economic sense.

[45]The coefficient of correlation between X_5 and X_6 is about 0.9939, a very high correlation indeed.

We leave it for the reader to devise alternative models and see how the results change. Also keep in mind the warning sounded earlier about using the ratio method of transforming the data to alleviate the problem of collinearity. We will revisit this question in Chapter 11.

10.11 SUMMARY AND CONCLUSIONS

1. One of the assumptions of the classical linear regression model is that there is no multicollinearity among the explanatory variables, the X's. Broadly interpreted, multicollinearity refers to the situation where there is either an exact or approximately exact linear relationship among the X variables.

2. The consequences of multicollinearity are as follows: If there is perfect collinearity among the X's, their regression coefficients are indeterminate and their standard errors are not defined. If collinearity is high but not perfect, estimation of regression coefficients is possible but their standard errors tend to be large. As a result, the population values of the coefficients cannot be estimated precisely. However, if the objective is to estimate linear combinations of these coefficients, *the estimable functions,* this can be done even in the presence of perfect multicollinearity.

3. Although there are no sure methods of detecting collinearity, there are several indicators of it, which are as follows:

(*a*) The clearest sign of multicollinearity is when R^2 is very high but none of the regression coefficients is statistically significant on the basis of the conventional t test. This case is, of course, extreme.

(*b*) In models involving just two explanatory variables, a fairly good idea of collinearity can be obtained by examining the zero-order, or simple, correlation coefficient between the two variables. If this correlation is high, multicollinearity is generally the culprit.

(*c*) However, the zero-order correlation coefficients can be misleading in models involving more than two X variables since it is possible to have low zero-order correlations and yet find high multicollinearity. In situations like these, one may need to examine the partial correlation coefficients.

(*d*) If R^2 is high but the partial correlations are low, multicollinearity is a possibility. Here one or more variables may be superfluous. But if R^2 is high and the partial correlations are also high, multicollinearity may not be readily detectable. Also, as pointed out by C. Robert, Krishna Kumar, John O'Hagan, and Brendan McCabe, there are some statistical problems with the partial correlation test suggested by Farrar and Glauber.

(*e*) Therefore, one may regress each of the X_i variables on the remaining X variables in the model and find out the corresponding coefficients of determination R_i^2. A high R_i^2 would suggest that X_i

is highly correlated with the rest of the X's. Thus, one may drop that X_i from the model, provided it does not lead to serious specification bias.

4. Detection of multicollinearity is half the battle. The other half is concerned with how to get rid of the problem. Again there are no sure methods, only a few rules of thumb. Some of these rules are as follows: (1) using extraneous or prior information, (2) combining cross-sectional and time series data, (3) omitting a highly collinear variable, (4) transforming data, and (5) obtaining additional or new data. Of course, which of these rules will work in practice will depend on the nature of the data and severity of the collinearity problem.

5. We noted the role of multicollinearity in prediction and pointed out that unless the collinearity structure continues in the future sample it is hazardous to use the estimated regression that has been plagued by multicollinearity for the purpose of forecasting.

6. Although multicollinearity has received extensive (some would say excessive) attention in the literature, an equally important problem encountered in empirical research is that of micronumerosity, smallness of sample size. According to Goldberger, "When a research article complains about multicollinearity, readers ought to see whether the complaints would be convincing if "micronumerosity" were substituted for "multicollinearity."[46] He suggests that the reader ought to decide how small n, the number of observations, is before deciding that one has a small-sample problem, just as one decides how high an R^2 value is in an auxiliary regression before declaring that the collinearity problem is very severe.

EXERCISES

Questions

10.1. In the k-variable linear regression model there are k normal equations to estimate the k unknowns. These normal equations are given in **Appendix C**. Assume that X_k is a perfect linear combination of the remaining X variables. How would you show that in this case it is impossible to estimate the k regression coefficients?

10.2. Consider the set of hypothetical data in Table 10.10. Suppose you want to fit the model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

to the data.

a. Can you estimate the three unknowns? Why or why not?

b. If not, what linear functions of these parameters, the estimable functions, can you estimate? Show the necessary calculations.

[46] Goldberger, op. cit., p. 250.