# Eco 213: Basic Data Analysis and Econometrics
# Lecture 6: Multicollinearity
# March 9, 2019

Dr. Garima Malik
Department of Economics
Shiv Nadar University

# Outline

- What is the nature of multicollinearity?
- Is multicollinearity really a problem?
- What are its practical consequences?
- How does one detect it?
- What remedial measures can be taken to alleviate the problem of multicollinearity?

# THE NATURE OF MULTICOLLINEARITY

- *Multicollinearity* originally meant the existence of a "perfect," or exact, linear relationship among some or all explanatory variables of a regression model. For the $k$-variable regression involving explanatory variable $X_1, X_2, \ldots, X_k$ an exact linear relationship is said to exist if the following condition is satisfied:
- $\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_k X_k = 0 \quad (10.1.1)$
- where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are constants such that not all of them are zero simultaneously.
- Today, however, the term multicollinearity is used to include the case where the $X$ variables are intercorrelated but not perfectly so, as follows:
- $\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_2 X_k + v_i = 0$ (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:

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

- which shows how $X_2$ is exactly *linearly related to other 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:

| $X2$ | $X3$ | $X^*3$ |
|---|---|---|
| 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 $X^*_3$. ($X_{3i} = 5X_{2i} + v_i$ ) 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 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 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.

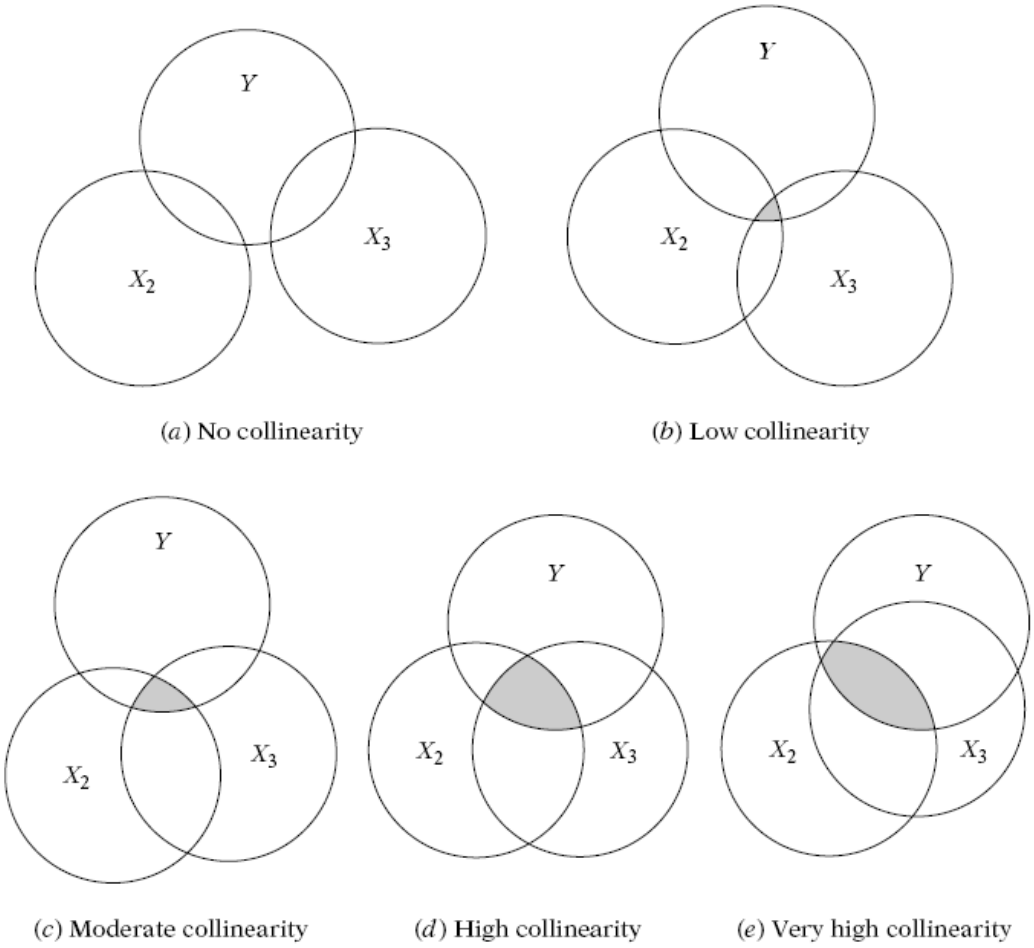


**FIGURE 10.1** The Ballentine view of multicollinearity.

- 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:
- $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X^2_i + \beta_3 X^3_i + u_i$ (10.1.5)
- where, say, $Y$ = total cost of production and $X$ = output. The variables $X^2_i$ (output squared) and $X^3i$ (output cubed) are obviously functionally related to $X_i$, but the relationship is nonlinear.

- 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,* the regression coefficients of the $X$ variables are indeterminate and their standard errors are infinite.
- *If multicollinearity is less than perfect*, the regression coefficients, although determinate, possess large standard errors which means the coefficients *cannot be estimated* with great precision or accuracy.

# Sources of Multicollinearity

- There are several sources of multicollinearity.
- 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 ($X_3$) (High $X_2$ always mean high $X_3$).
- 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.
- 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.

# ESTIMATION IN THE PRESENCE OF PERFECT MULTICOLLINEARITY

- In the case of perfect multicollinearity 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, we can write the three variable regression model as
- $y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i$ (10.2.1)

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

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

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

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\lambda^2 \sum x_{2i}^2) - (\lambda \sum y_i x_{2i})(\lambda \sum x_{2i}^2)}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2) - \lambda^2 (\sum x_{2i}^2)^2} \qquad \textbf{(10.2.2)}$$

$$= \frac{0}{0}$$

- which is an *indeterminate expression*. We can also verify that $\hat{\beta}_3$ is indeterminate.
- *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 $\lambda$. 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.

- To see this differently, let us substitute $X_{3i} = \lambda X_{2i}$ into (10.2.1) and obtain the following:
- $y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3(\lambda x_{2i}) + \hat{u}_i$
- $= (\hat{\beta}_2 + \lambda\hat{\beta}_3)x_{2i} + \hat{u}_i$ (10.2.3)
- $= \hat{\alpha} x_{2i} + \hat{u}_i$
- where
- $\hat{\alpha} = (\hat{\beta}_2 + \lambda\hat{\beta}_3)$ (10.2.4)
- Applying the usual OLS formula to (10.2.3), we get
- $\hat{\alpha} = (\hat{\beta}_2 + \lambda\hat{\beta}_3) = \Sigma x_{2i} y_i / \Sigma x^2_{2i}$ (10.2.5)
- Therefore, although we can estimate $\alpha$ uniquely, there is no way to estimate $\beta_2$ and $\beta_3$ uniquely; mathematically
- $\hat{\alpha} = \hat{\beta}_2 + \lambda\hat{\beta}_3$ (10.2.6)
- gives us only one equation in two unknowns (note $\lambda$ is given) and there is an infinity of solutions to (10.2.6) for given values of $\hat{\alpha}$ and $\lambda$.

- To put this idea in concrete terms, let $\hat{\alpha} = 0.8$ and $\lambda = 2$. Then we have
- $0.8 = \hat{\beta}_2 + 2\hat{\beta}_3$ (10.2.7)
- or
- $\hat{\beta}_2 = 0.8 - 2\hat{\beta}_3$ (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* $\hat{\beta}_2$. That is 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 $\alpha$, given the value of $\lambda$. 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.

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

- Generally, there is no exact linear relationship among the $X$ variables. Thus, turning to the three-variable model in the deviation form given in (10.2.1), instead of exact multicollinearity, we may have
- $x_{3i} = \lambda x_{2i} + v_i$ (10.3.1)
- where $\lambda \neq 0$ and where $v_i$ is a stochastic error term such that $x_{2i}v_i = 0$.
- In this case, estimation of regression coefficients $\beta_2$ and $\beta_3$ may be possible. For example, we obtain

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

$\beta$^3.

- 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).

## 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, $R^2$ *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

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

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

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

- 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.

- Wider Confidence Intervals:
- Because of the large standard errors, the confidence intervals for the relevant population parameters tend to be larger. For example, when $r_{23} = 0.95$, the confidence interval for $\beta_2$ is larger than when $r_{23} = 0$ by a factor of 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 increases.

**TABLE 10.2** THE EFFECT OF INCREASING COLLINEARITY ON THE 95% CONFIDENCE INTERVAL FOR $\beta_2$: $\hat{\beta}_2 \pm 1.96 \text{ se } (\hat{\beta}_2)$

| Value of $r_{23}$ | 95% confidence interval for $\beta_2$ |
|---|---|
| 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 $\sigma^2$ 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
- 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.

- 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 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)!*

- 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:

$$\hat{Y}_i = 1.1939 + 0.4463X_{2i} + 0.0030X_{3i} \qquad (10.5.6)$$
$$\text{(0.7737)} \quad \text{(0.1848)} \quad \text{(0.0851)}$$
$$t = (1.5431) \quad (2.4151) \quad (0.0358)$$

$R^2 = 0.8101$ $r_{23} = 0.5523$

- $\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = -0.00868$ df = 2

**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.
- Using the data of Table 10.4, we now obtain:

$$\hat{Y}_i = 1.2108 + 0.4014X_{2i} + 0.0270X_{3i}$$
$$\text{(0.7480)} \quad \text{(0.2721)} \quad \text{(0.1252)}$$
$$t = \text{(1.6187)} \quad \text{(1.4752)} \quad \text{(0.2158)}$$
$$\text{(10.5.7)}$$

$R^2 = 0.8143$ $\qquad$ $r_{23} = 0.8285$

$\text{cov}\,(\hat{\beta}_2, \hat{\beta}_3) = -0.0282$ $\qquad$ df = 2

- 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. Also note that in (10.5.6) $\text{cov}\,(\hat{\beta}_2, \hat{\beta}_3) = -0.00868$ 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.

**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 |

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

- Let us reconsider the consumption–income example in table 10.5. we obtain the following regression:

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

- 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. The wealth variable has the wrong sign. 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.

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

- Under the usual assumption we obtain:
- $F$ =4282.7770 / 46.3494 = 92.4019
  (10.6.2)
- This $F$ value is obviously highly significant. 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

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

- 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:

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

- 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

$$\hat{Y}_i = 24.411 + 0.0498X_{3i}$$
$$\text{se} = (6.874)\quad (0.0037) \qquad (10.6.5)$$
$$t = (3.551)\quad (13.29) \qquad R2 = 0.9567$$

- 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.

# DETECTION OF MULTICOLLINEARITY

- 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. Multicollinearity is a *feature of the sample and not of the population*. Therefore, we do not "*test for multicollinearity*" but we measure its *degree* in any particular sample.
- 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.

- *1. High $R^2$ but few significant t ratios.* 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.

- *2. High pair-wise correlations among regressors.* Another suggested rule of thumb is that if the *pair-wise or 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 correlations may suggest collinearity, it is not necessary that they be high to have collinearity in any specific case, *it can exist even though the simple correlations are comparatively low* (say, less than 0.50).

- *3. Auxiliary regressions.* 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 $R^2_i$; 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 $R^2$, the variable (10.7.3) follows the $F$ distribution with $k-2$ and $n-k+1$df.

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

- 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.

- 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.

- Klien's *rule of thumb*
- 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. Of course, like all other rules of thumb, this one should be used judiciously.

# 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.
- Why?
- Multicollinearity is essentially *a data deficiency problem* and some times *we have no choice over the data* we have available for empirical analysis.
- 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
- $y_i = \hat{\alpha} x_{2i} + \hat{u}_i$ (10.2.3)
- we can estimate $\alpha$ uniquely, even if we cannot estimate its two components individually. Sometimes this is the best we can do with a given set of data.

- Rule-of-Thumb Procedures
- 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. 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:
- $Y_i = \beta_1 + \beta_2 X_{2i} + 0.10\beta_2 X_{3i} + ui = \beta_1 + \beta_2 X_i + u_i$
- 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 $\beta_2$ and $\beta_3$. How does one obtain a priori information? It could *come from previous empirical work*.
- For example, in the Cobb–Douglas–type production function
- $Y_i = \beta_1 X_{2i}^{\beta 2} X_{3i}^{\beta 3} e^{ui}$ (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:

- $\ln (\text{GDP/Labor})_t = \beta_1 + \alpha \ln (\text{Capital/Labor})_t$
- 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."

- 2. Combining cross-sectional and time series data.
- A variant of the priori information technique is the combination of crosssectional and time-series data, known as *pooling the data.* Suppose we want to study the demand for automobiles in the US 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 $\beta_2$ and income elasticity $\beta_3$.
- In time series data *the price and income* variables generally tend to be *highly collinear*. A way out of this has been suggested by *Tobin* who says that if we have cross-sectional data, we can obtain a fairly reliable estimate of the income elasticity $\beta_3$ 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 $\beta_2$ from the preceding regression.

- Although it is an appealing technique, pooling the time series and crosssectional 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.

- 3. Dropping a variable(s) and specification bias.
- 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*.
- 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. 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
- $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$ (10.8.3)
- holds at time $t$, it must also hold at time $t - 1$ because the origin of time is arbitrary anyway. Therefore, we have
- $Y_{t-1} = \beta_1 + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + u_{t-1}$ (10.8.4)
- If we subtract (10.8.4) from (10.8.3), we obtain
- $Y_t - Y_{t-1} = \beta_2(X_{2t} - X_{2,t-1}) + \beta_3(X_{3t} - X_{3,t-1}) + v_t$ (10.8.5)
- where $v_t = u_t - u_{t-1}$. Equation (10.8.5) is known as *the first difference form*.

- 1.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*. an incidental advantage of the first-difference transformation is that it may make a *nonstationary* time series stationary. Loosely speaking, a time series, say, $Y_t$, is stationary if its mean and variance do not change systematically over time.

- 2.Another commonly used transformation *in practice is the ratio transformation*. Consider the model:
- $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$ (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 $X3$, to obtain:
- $Y_t/X_{3t} = \beta_1(1/X_{3t}) + \beta_2 (X_{2t}/X_{3t}) + \beta_3 + (u_t/X_{3t})$ *(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 *vt* in (10.8.5) may not satisfy one of the assumptions of the classical linear regression model, namely, that the *disturbances are serially uncorrelated*. 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. 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 $(u_t/X_{3t})$ will be *heteroscedastic*, if the original error term $u_t$ is homoscedastic. 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
- $var\ (\hat{\beta}_2) = \sigma^2 / \Sigma\ x^2_{2i}\ (1 - r^2_{23})$

  Now as the sample size increases, $\Sigma x^2_{2i}$ will generally increase. 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 $\beta_2$ 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:

  $\hat{Y}_i = 24.377 + 0.8716X_{2i} - 0.0349X_{3i}$
- $t$ = (3.875) (2.7726) (−1.1595) $R^2 = 0.9682$
  (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, the following results were obtained:
- $\hat{Y}_i = 2.0907 + 0.7299X2i + 0.0605X_{3i}$
- $t = (0.8713)\ \ (6.0014)\qquad (2.0014)\qquad\qquad R^2 = 0.9672$
  $\qquad\quad (10.8.9)$
- Now the wealth coefficient not only has the correct sign but also is statistically significant at the 5 percent level. Obtaining additional or "better" data is not always that easy.

- 6. 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. These cannot be discussed competently without resorting to matrix algebra.

## 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. But this may be so "... as long as the values of the explanatory variables for which predictions are desired obey the same near-exact linear dependencies as the original design [data] matrix $X$."

- 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, in which case prediction will become increasingly uncertain. 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.