EE 425: 1/2011
Multicollinearity

What happens when the regressors are linearly correlated?

1. Nature of multicollinearity
2. Consequences of multicollinearity
3. Detection of multicollinearity
4. Remedial measures

The Nature of Multicollinearity

- Perfect collinearity:- The regressors $X_1, \ldots X_k$ are said to have an exact linear relationship if there exists $\lambda_1, \ldots \lambda_k$, not all of them are zero, such that

  $\lambda_1 X_1 + \lambda_2 X_2 \ \ldots + \lambda_k X_k = 0$

It is possible to express one X as linear function of the remaining regressors

Example: If $\lambda_2 \neq 0$, then

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \ldots\ldots - \frac{\lambda_k}{\lambda_2} X_{ki}$$

The Nature of Multicollinearity

- Multicollinearity:- when the regressors $X_1, \ldots X_k$ are highly correlated, but less than perfect. It is possible to express one X as linear function of the remaining regressors plus some random terms.

Example: If $\lambda_2 \neq 0$, it is possible to express, say

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \ldots\ldots - \frac{\lambda_k}{\lambda_2} X_{ki} - \frac{1}{\lambda_2} v_i$$

Numerical Example

X_2	X_3	X^*_3
10	50	52 = 50+2
15	75	75 = 75 + 0
18	90	97= 90 + 7
24	120	129 = 120 + 9
30	150	150 = 150 + 0

$X_3 = 5X_{2i}$ ==> perfect collinearity

But $X^*_3 = 5X_{2i} + v_i$ ==> multicollinearity

Sources of multicollinearity

1. The method of data collection is inappropriate, ex. samples come from a relatively homogeneous population.
2. Constraints on the model, ex. Wealth and income are related in nature.
3. Model specification, ex. X, X^2 and X^3 are not linearly correlated, but are related.
4. An over-determined model
5. In time series data, several variables have common trend, ex. GDP, population etc.

Estimation in the presence of multicollinearity: With perfect collinearity, we cannot determine the OLS regression coefficients

Example: Let k=3 (including the intercept)

Model: If $X_{3i} = \lambda X_{2i}$ for all i

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

then $\sum x_{3i}^2 = \lambda^2 \sum x_{2i}^2$

$$\sum y_i x_{3i} = \lambda \sum y_i x_{2i}$$

$$\sum x_{2i} x_{3i} = \lambda \sum x_{2i}^2$$

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

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

$$= \left(\lambda^2 \sum x_{2i}^2\right) \frac{\sum y_i x_{2i} - \sum y_i x_{2i}}{\sum x_{2i}^2 - \sum x_{2i}^2} = \frac{0}{0}$$

OLS formula for var(B_2) and the consequence of perfect collinearity:

when $X_{3i} = \lambda X_{2i}$

$$\text{var}(\hat{\beta}_2) = \frac{\left(\sum x_{3i}^2\right)}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i} x_{3i}\right)^2} \sigma^2$$

$$= \frac{\left(\lambda^2 \sum x_{2i}^2\right)}{\left(\sum x_{2i}^2\right)\left(\lambda^2 \sum x_{2i}^2\right) - \left(\lambda \sum x_{2i}^2\right)^2} \sigma^2 = \frac{\sigma^2}{0}$$

or look at $r_{23}^2 = \dfrac{\left(\sum x_{2i} x_{3i}\right)^2}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right)}$

$$= \frac{\left(\lambda \sum x_{2i}^2\right)^2}{\left(\sum x_{2i}^2\right)\left(\lambda^2 \sum x_{2i}^2\right)} = 1$$

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

Consequences of Multicollinearity

- Practical consequences
- In the case of perfect collinearity, we cannot find OLS estimators.
- In the case of less than perfect collinearity (or some degree of collinearity), we can obtain OLS coefficients but their variances are large.
- Wider confidence intervals

With serious degree of multicollinearity, we may encounter conflicting test results: high R square and F-test is significant but some t statistics may be small and not statistically significant.

An illustrative example:

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

$\hat{Y}_i = 24.7747 + 0.9415X_{2i} - 0.0424X_{3i}$

t = (3.67) (1.1442) (-0.5261) $R^2 = 0.9635$

None of the parameter is significant, although R^2 is very high

Using ANOVA Table:

Source of Variation	SS	df	MSS
Due to regression	8,565.5541	2	4,282.777
Due to residual	324.4459	7	46.3494

$F = 4282.777/46.3494 = 92.4019$

If we use F-test, it is highly significant

Let's consider the following regressions

$\hat{Y}_i = 24.4545 + 0.5091X_{2i}$

$t = \quad (3.81) \quad (14.2432) \qquad R^2 = 0.9621$

$\hat{Y}_i = 24.411 + 0.04986X_{3i}$

$t = \quad (3.551 \quad (13.29) \qquad R^2 = 0.9567$

$\hat{X}_{3i} = 75454 + 10.1909X_{2i}$

$t = \quad (0.256) \quad (62.0405) \qquad R^2 = 0.9979$

Detection of Multicollinearity

- Multicollinearity is a question of degree , not of kind. Therefore we do not test for the existence of multicollinearity, but we can assess the degree of multicollinearity in any particular sample.
- How do we measure the degree of multicollinearity?

Detection of Multicollinearity

- High R square but few significant t ratios, this is the "classic" symptom of multicollinearity.
- High pair-wise correlations among regressors
- Auxiliary regressions
- Variance inflation factor (VIF): if X_j is highly correlated to other X's, then R_j square is close to 1 and VIF_j is close to infinity. As a rule of thumb, if VIF_j is greater than 10 there is a problem

Remedial Measures

- Do nothing: Blanchard:- "Multicollinearity is essentially a data problem and we can do nothing about it except to obtain more or new set of data."
- Do something
  - Use a prior information
  - Dropping a variable (or variables)
  - Transformation of variables
  - Find more information or data

Additional notes:

- If the objective is prediction only, the presence of multicollinearity may be tolerated.
- But if the purpose is to test the significance of an independent variable, the existence of multicollinearity makes it more difficult to find any statistically significant independent variable.
- There are other methods to solve multicollinearity, but they are beyond the scope of this course.