

Lecture 16: The problem of multicollinearity; standardised form; sources

Econometrics ■ Unit II: Multicollinearity and heteroskedasticity

Kapil Kumar

Department of Statistics
Chaudhary Charan Singh University, Meerut

B.Sc. (Hons) Statistics, Semester V
Academic Session 2026–2027

Unit II: Multicollinearity and heteroskedasticity

Lecture 16 begins Unit II: Multicollinearity and heteroskedasticity.

Today

- The problem of multicollinearity
 - Standardised form of the model
- Sources of multicollinearity

A basic assumption of the multiple regression model $y = X\beta + \varepsilon$ is that

$$\text{rank}(X) = k,$$

that is, the columns of X are linearly independent. Then $X'X$ is invertible and the OLS estimator $b = (X'X)^{-1}X'y$ exists.

The columns $X_1, X_2, \dots, X_k$ of X are *linearly dependent* if there are constants $c_1, c_2, \dots, c_k$, not all zero, such that

$$\sum_{j=1}^k c_j X_j = 0.$$

Definition 3.1

If the relation $\sum_j c_j X_j = 0$ holds *exactly*, we have **perfect (exact) multicollinearity**. Then $\text{rank}(X'X) < k$ and $(X'X)^{-1}$ does not exist.

If the relation holds *approximately*, that is, $\sum_j c_j X_j \approx 0$, we have **near multicollinearity**. Then $X'X$ is nearly singular; we say $X'X$ is *ill-conditioned*.

- In economic data, perfect multicollinearity is rare but near multicollinearity is common.
- The word “multicollinearity” usually refers to the near case: the explanatory variables are highly intercorrelated.

Standardised form of the model

Multicollinearity is easiest to study after *centring and scaling* the variables. For each explanatory variable put

$$S_{jj} = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2, \quad x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{\sqrt{S_{jj}}},$$

and similarly $y_i^* = (y_i - \bar{y}) / \sqrt{S_{yy}}$. Then each new column has mean zero and length one:

$$\sum_i x_{ij}^{*2} = \frac{\sum_i (x_{ij} - \bar{x}_j)^2}{S_{jj}} = \frac{S_{jj}}{S_{jj}} = 1.$$

The cross-product of two columns is

$$\sum_i x_{ij}^* x_{il}^* = \frac{\sum_i (x_{ij} - \bar{x}_j)(x_{il} - \bar{x}_l)}{\sqrt{S_{jj}S_{ll}}} = r_{jl},$$

the correlation coefficient between x_j and x_l .

Therefore:

- $X'X$ becomes the $k \times k$ **correlation matrix** of the explanatory variables, and
- $X'y$ becomes the $k \times 1$ vector of correlations r_{jy} between each explanatory variable and y .

Convention for this chapter

From now on in this chapter, X and y denote the centred and scaled data. Here k is the number of explanatory variables; there is no intercept in the standardised model. Also, σ^2 is the error variance on this scale.

Coefficients on the original scale are recovered from

$$\beta_j^{\text{orig}} = \beta_j \sqrt{S_{yy}/S_{jj}}, \quad \beta_0^{\text{orig}} = \bar{y} - \sum_j \beta_j^{\text{orig}} \bar{x}_j.$$

1. **Method of data collection.** Data may be collected only over a small part (subspace) of the possible values of the explanatory variables, where they happen to be nearly linearly related.
 - For example, if sampling covers only a narrow range of incomes, income and wealth may move almost together in the sample even if they do not in the population.
2. **Constraints on the model or the population.** The population itself may have a built-in relation.
 - For example, in electricity consumption data, families with higher incomes usually have larger houses, so income and house size are strongly related.

3. **Identities or definitional relations.** If data are collected on income, saving and expenditure, then $\text{income} = \text{saving} + \text{expenditure}$ exactly. Including all three gives perfect multicollinearity. Such a relation does not disappear when the sample size increases.
4. **Imprecise formulation of the model.** The model may be more complicated than necessary. For example, with three variables X_1, X_2, X_3 , adding the cross products X_1X_2, X_2X_3, X_1X_3 raises k from 3 to 6, and these products are usually highly correlated with the original variables. Polynomial terms (x, x^2, x^3) behave similarly.

5. **An over-determined model.** Too many explanatory variables may be included, sometimes more than the number of observations ($k > n$).
- Then $\text{rank}(X) \leq n < k$ and exact dependence is unavoidable.
 - This happens, for example, in medical studies with few patients but many measured variables.
 - With time-series data, a long period may be needed to get more observations, but economic behaviour may not remain the same over a long period.

Practice

- Exercise 3.1 (Exercises and Solutions booklet).

Next lecture

Lecture 17: Implications: perfect and near multicollinearity; the variance inflation factor.