

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

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

Previous lecture

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

Today

- Implications (consequences) of multicollinearity
 - Perfect multicollinearity
 - Near multicollinearity with two explanatory variables
 - Many explanatory variables: the variance inflation factor

Suppose $x_{2i} = 2x_{1i}$ for all i in

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i.$$

Substituting,

$$y_i = \beta_1 x_{1i} + 2\beta_2 x_{1i} + \varepsilon_i = (\beta_1 + 2\beta_2)x_{1i} + \varepsilon_i.$$

- The data can estimate only the combination $\theta = \beta_1 + 2\beta_2$.
- Any pair (β_1, β_2) with the same value of $\beta_1 + 2\beta_2$ gives exactly the same fitted values.
- So β_1 and β_2 cannot be estimated separately, and $(X'X)^{-1}$ does not exist.

Consider $y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ with $E(\varepsilon) = 0$, $\text{Var}(\varepsilon) = \sigma^2 I$, and x_1, x_2, y standardised. Let r be the correlation between x_1 and x_2 , and r_{1y}, r_{2y} their correlations with y . The normal equations $X'Xb = X'y$ are

$$\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} r_{1y} \\ r_{2y} \end{pmatrix}.$$

Step 1 (inverse). The determinant is $1 - r^2$. For $|r| < 1$,

$$(X'X)^{-1} = \frac{1}{1 - r^2} \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}.$$

Step 2 (estimates).

$$b_1 = \frac{r_{1y} - r r_{2y}}{1 - r^2}, \quad b_2 = \frac{r_{2y} - r r_{1y}}{1 - r^2}.$$

Step 3 (variances). Since $\text{Var}(b) = \sigma^2(X'X)^{-1}$,

$$\text{Var}(b_1) = \text{Var}(b_2) = \frac{\sigma^2}{1 - r^2}, \quad \text{Cov}(b_1, b_2) = -\frac{r\sigma^2}{1 - r^2}.$$

Interpretation.

- If $r = 0$ (uncorrelated regressors), $X'X = I$, $\text{rank}(X'X) = 2$ and $\text{Var}(b_1) = \text{Var}(b_2) = \sigma^2$.
- If $r = \pm 1$ (perfect correlation), $\text{rank}(X'X) = 1$ and the estimates are not defined.
- As $r \rightarrow \pm 1$, $\text{Var}(b_1) = \text{Var}(b_2) \rightarrow \infty$.

Two explanatory variables (cont.)

r	0	0.1	0.9	0.99
$\text{Var}(b_1) = \text{Var}(b_2)$	σ^2	$1.0101 \sigma^2$	$5.2632 \sigma^2$	$50.2513 \sigma^2$

- The standard errors rise sharply as $r \rightarrow \pm 1$, and break down at $r = \pm 1$ because $X'X$ becomes singular.
- If r is close to zero, the multicollinearity is called *non-harmful*; if r is close to ± 1 , it is called *harmful*.
- There is no sharp boundary between the two.

Result 3.2

Let $C = (X'X)^{-1}$ in the standardised model and let R_j^2 be the coefficient of determination from the regression of x_j on the other $k - 1$ explanatory variables. Then

$$C_{jj} = \frac{1}{1 - R_j^2}, \quad \text{so} \quad \text{Var}(b_j) = \sigma^2 C_{jj} = \frac{\sigma^2}{1 - R_j^2}.$$

Proof.

Reorder the columns so that x_j is first, and let $X_{(j)}$ contain the other columns. Write

$$X'X = \begin{pmatrix} 1 & a' \\ a & R_{(j)} \end{pmatrix}, \quad a = X'_{(j)}x_j, \quad R_{(j)} = X'_{(j)}X_{(j)}.$$

Proof (cont.)

Here a holds the correlations of x_j with the others and $R_{(j)}$ is their correlation matrix.

Step 1 (first column of the inverse). Put $c = 1/(1 - a'R_{(j)}^{-1}a)$ and consider the vector $u = (c, -cR_{(j)}^{-1}a)'$. Multiplying,

$$X'X u = \begin{pmatrix} c - c a' R_{(j)}^{-1} a \\ ca - c R_{(j)} R_{(j)}^{-1} a \end{pmatrix} = \begin{pmatrix} c(1 - a' R_{(j)}^{-1} a) \\ ca - ca \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

So u is the first column of $(X'X)^{-1}$, and

$$C_{jj} = c = 1/(1 - a'R_{(j)}^{-1}a).$$

Proof (cont.)

Step 2 (the auxiliary regression). Regress x_j on $X_{(j)}$. Its normal equations are $R_{(j)}g = a$, so $g = R_{(j)}^{-1}a$. The explained sum of squares is

$$(X_{(j)}g)'(X_{(j)}g) = g'R_{(j)}g = a'R_{(j)}^{-1}R_{(j)}R_{(j)}^{-1}a = a'R_{(j)}^{-1}a.$$

The total sum of squares of the standardised x_j is 1. Hence $R_j^2 = a'R_{(j)}^{-1}a$.

Step 3. Combining Steps 1 and 2, $C_{jj} = 1/(1 - R_j^2)$. □

- If x_j is nearly a linear combination of the other regressors, R_j^2 is close to 1 and $\text{Var}(b_j)$ becomes very large.
- On the original scale the same result reads

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{S_{jj}(1 - R_j^2)}.$$

- The covariance between b_i and b_j is also large when x_i and x_j are both involved in the near-dependence.

Practice

- Exercise 3.2 (Exercises and Solutions booklet).

Next lecture

Lecture 18: OLS estimates too large on average; practical consequences; detection by $|X'X|$.