

Lecture 21: Remedies: more data, dropping variables, prior information, generalized inverse, principal components

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 20: Diagnostics: condition indices and variance-decomposition proportions.

Today

- Remedies for multicollinearity
 - Obtain more data
 - Drop some of the collinear variables
 - Use relevant prior information
 - Use a generalized inverse
 - Principal component regression

Multicollinearity is basically a shortage of information in the sample. If new data are collected in a way that breaks the pattern of dependence, the variances come down.

But this is not always possible.

- The process may be over, or the cost may be too high.
- New data may not be comparable with old data.
- A longer time series may go too far into the past, when conditions were different.
- If the dependence is an identity (for example, $\text{income} = \text{saving} + \text{expenditure}$), more data will not help at all.

- Variables causing the dependence can be removed, for example those with the smallest t ratios.
- If only some parameters are of interest, dropping others may give estimators with smaller mean squared error.
- But if a dropped variable truly belongs to the model, the remaining estimates become biased (specification bias), and the model may predict worse.

Information from theory or from earlier studies can replace the missing sample information.

A known coefficient. If $\beta_2 = c$ is known in $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, then

$$y - cx_2 = \beta_0 + \beta_1 x_1 + \varepsilon,$$

and β_1 is estimated by regressing $y - cx_2$ on x_1 .

A known linear restriction. In a Cobb–Douglas production function

$$\log Q = \beta_0 + \beta_1 \log L + \beta_2 \log K + \varepsilon,$$

labour and capital are often highly correlated.

Under constant returns to scale, $\beta_1 + \beta_2 = 1$, so $\beta_2 = 1 - \beta_1$ and

$$\begin{aligned}\log Q &= \beta_0 + \beta_1 \log L + (1 - \beta_1) \log K + \varepsilon \\ \implies \log(Q/K) &= \beta_0 + \beta_1 \log(L/K) + \varepsilon.\end{aligned}$$

Only one slope is now estimated, so the collinearity between $\log L$ and $\log K$ no longer matters.

- More generally, exact linear restrictions are used in *restricted regression* and stochastic restrictions in *mixed regression*.
- The restriction must be correct; wrong prior information gives biased estimates. For example, estimates obtained for the U.K. economy may not be valid for India.

If $\text{rank}(X'X) < k$, a generalized inverse $(X'X)^-$ can be used:

$$\hat{\beta} = (X'X)^- X'y.$$

- The estimates are not unique unless a specific inverse, such as the Moore–Penrose inverse, is chosen, and different choices give different results.
- They do not identify the separate effects that the data cannot distinguish.

The idea is to replace the correlated regressors by uncorrelated linear combinations of them, and to drop the combinations that carry almost no information.

Step 1 (canonical form). Let $X'X = V\Lambda V'$ with $V'V = VV' = I$. Since $VV' = I$,

$$y = X\beta + \varepsilon = XVV'\beta + \varepsilon = Z\alpha + \varepsilon, \quad Z = XV, \quad \alpha = V'\beta.$$

The columns $Z_1, \dots, Z_k$ of Z are the **principal components**.

Step 2 (they are orthogonal).

$$Z'Z = V'X'XV = V'(V\Lambda V')V = (V'V)\Lambda(V'V) = \Lambda.$$

So $Z'_jZ_l = 0$ for $j \neq l$ and $Z'_jZ_j = \lambda_j$. The components are uncorrelated, and λ_j measures the variation in Z_j . Also $\sum_j \lambda_j = \text{tr}(X'X)$, so the components keep the total variation of the original regressors.

Step 3 (OLS on all components).

$$\hat{\alpha} = (Z'Z)^{-1}Z'y = \Lambda^{-1}Z'y,$$
$$\text{Var}(\hat{\alpha}) = \sigma^2\Lambda^{-1} = \sigma^2 \text{diag}\left(\frac{1}{\lambda_1}, \dots, \frac{1}{\lambda_k}\right).$$

A small λ_j means a large variance for $\hat{\alpha}_j$; this is where the problem of multicollinearity lies.

Step 4 (drop components with small eigenvalues). Arrange $\lambda_1 > \lambda_2 > \cdots > \lambda_k$. Discard the r components with the smallest eigenvalues, keeping $k - r$ components chosen so that, for example,

$$\frac{\sum_{j=1}^{k-r} \lambda_j}{\sum_{j=1}^k \lambda_j} \geq 0.90,$$

that is, the retained components explain at least 90% of the total variation. Divide $Z = (Z_{k-r}, Z_r)$ and $V = (V_{k-r}, V_r)$ in the same way.

Step 5 (estimate and transform back). Fit $y = Z_{k-r}\alpha_{k-r} + \epsilon^*$ by OLS:

$$\hat{\alpha}_{k-r} = (Z'_{k-r}Z_{k-r})^{-1}Z'_{k-r}y.$$

Since $\alpha = V'\beta$ implies $\beta = V\alpha$, the **principal component estimator** of β is

$$\hat{\beta}_{pc} = V_{k-r}\hat{\alpha}_{k-r}.$$

It is biased (the dropped components are set to zero) but has much smaller variance than OLS when the dropped eigenvalues are small.

Example 3.4 (Principal component regression)

Continue Example 3.3 with $X'y = (0.8, 0.7)'$. Since $\lambda_1/(\lambda_1 + \lambda_2) = 1.95/2 = 0.975 \geq 0.90$, keep only $Z_1 = Xv_1$, $v_1 = (1, 1)'/\sqrt{2}$. Then $Z_1'Z_1 = \lambda_1 = 1.95$ and

$$Z_1'y = v_1'X'y = \frac{0.8 + 0.7}{\sqrt{2}} = \frac{1.5}{\sqrt{2}}, \quad \hat{\alpha}_1 = \frac{1.5/\sqrt{2}}{1.95}.$$

Transforming back,

$$\hat{\beta}_{pc} = v_1\hat{\alpha}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \frac{1.5}{1.95\sqrt{2}} = \frac{1.5}{3.9} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0.3846 \\ 0.3846 \end{pmatrix}.$$

OLS gives $(1.3846, -0.6154)'$ (Example 3.5). The principal component estimates have the same sign and a sensible size.

Practice

- R practical P05: Diagnosing multicollinearity in the Longley data.

Next lecture

Lecture 22: Ridge regression: the estimator, constrained least squares, bias and covariance.