Lecture 56: Two-stage least squares: the method, derivation and consistency

Econometrics ▪ Unit IV: Simultaneous equations models

Kapil Kumar

Department of Statistics
Chaudhary Charan Singh University, Meerut

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

This lecture

Previous lecture

Lecture 55: Indirect least squares.

Today

- Two-stage least squares
  - The method
  - Derivation of the 2SLS formula
  - Consistency

Two-stage least squares

The trouble with OLS is that Y_1 is correlated with u. Two-stage least squares (2SLS) was developed by H. Theil and, independently, by Basmann (1957). Its idea is to replace Y_1 by its part explained by the exogenous variables, which is uncorrelated with u.

Stage 1. Regress each endogenous right-hand variable on *all* exogenous variables of the system by OLS:

$$\widehat{\Pi} = (X'X)^{-1}X'Y_1, \qquad \widehat{Y}_1 = X\widehat{\Pi} = X(X'X)^{-1}X'Y_1 = P_X Y_1,$$

where $P_X = X(X'X)^{-1}X'$.

Stage 2. Replace Y_1 by $\widehat{Y}_1$ in (8.1) and apply OLS: regress y_1 on $\widehat{Y}_1$ and X_1.

Derivation of the 2SLS formula

Step 1 (properties of P_X). P_X is symmetric, $P_X' = X(X'X)^{-1}X' = P_X$, and idempotent,

$$P_X^2 = X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}X' = P_X.$$

Also $P_X X = X(X'X)^{-1}X'X = X$. Since the columns of X_1 are among the columns of X, $P_X X_1 = X_1$.

Step 2 (the second-stage regressor matrix).

$$\widehat{D} = [\widehat{Y}_1 \;\; X_1] = [P_X Y_1 \;\; P_X X_1] = P_X D.$$

Derivation of the 2SLS formula (cont.)

Step 3 (OLS in the second stage).

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{2SLS} &= (\widehat{D}'\widehat{D})^{-1}\widehat{D}'y_1 = (D'P_X'P_XD)^{-1}D'P_X'y_1 \\ &= (D'P_XP_XD)^{-1}D'P_Xy_1 = (D'P_XD)^{-1}D'P_Xy_1.\end{aligned}$$

Therefore

$$\boxed{\widehat{\boldsymbol{\theta}}_{2SLS} = (D'P_XD)^{-1}D'P_Xy_1, \qquad P_X = X(X'X)^{-1}X'.} \tag{8.2}$$

The inverse exists when the equation satisfies the rank condition (in the sample), which requires the order condition.

Step 1. Substitute $y_1 = D\boldsymbol{\theta} + u$ in (8.2):

$$\widehat{\boldsymbol{\theta}}_{2SLS} = \boldsymbol{\theta} + (D'P_X D)^{-1} D'P_X u.$$

Step 2. Writing $P_X = X(X'X)^{-1}X'$ and dividing each factor by n,

$$\widehat{\boldsymbol{\theta}}_{2SLS} - \boldsymbol{\theta} = \Big[\frac{D'X}{n}\Big(\frac{X'X}{n}\Big)^{-1}\frac{X'D}{n}\Big]^{-1} \times \frac{D'X}{n}\Big(\frac{X'X}{n}\Big)^{-1}\frac{X'u}{n}.$$

Consistency of 2SLS (cont.)

Step 3. The exogenous variables are uncorrelated with u, so $\operatorname{plim} X'u/n = 0$. The other factors have finite limits (the bracket has a nonsingular limit when the equation is identified). Hence $\operatorname{plim} \widehat{\boldsymbol{\theta}}_{2SLS} = \boldsymbol{\theta}$: **2SLS is consistent**. It is generally biased in small samples.

Remember

$$\widehat{\boldsymbol{\theta}}_{2SLS} = (D' P_X D)^{-1} D' P_X y_1, \quad P_X = X(X'X)^{-1}X'.$$

After this lecture

Practice

Re-work the derivations and examples of this lecture without looking at the slides.

Next lecture

Lecture 57: 2SLS standard errors; an over-identified example; 2SLS, IV and ILS under exact identification.