Instrumental Variables (IV) estimator

Eco321: Econometrics

Donghwan Kim
Department of Economics
SUNY at Stony Brook

Spring 2005

Correlation between regressor and error term

Why does it matter?

The OLS estimator is biased and inconsistent (show it !!)

Correlation between regressor and error term

When does it happen?

- Measurement error in regressor
- Omitted variables
- Simultaneous equations model
  - $\longrightarrow$ See ch.14 in the textbook

Exogenous and Endogenous Variables

Exogenous Variables

- The variable X_i is said to be exogenous in the model if it is uncorrelated with error term

Endogenous Variables

- The variable X_i is said to be endogenous in the model if it is correlated with error term

Endogeneity Problem

Endogeneity Problem

- It occurs when a model has one or more endogenous explanatory variables

What is the solution?

- The solution to the endogeneity problem requires additional information which we call instrument variables or instruments.

Instrumental Variables

Model

$$y = X\beta + \epsilon$$

Divide X into two parts

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

- X_1: exogenous variables, $n \times k_1$ matrix
- X_2: endogenous variables, $n \times k_2$ matrix

Instrumental Variables

Properties for valid instrumental variables Z: $(n \times m)$ matrix

- Z must be uncorrelated with error term: $E(Z'\epsilon) = 0$
- Z must be correlated with X_2: $E(Z'X_2) \neq 0$

Instrumental Variables Z: $(n \times m)$ matrix

$$Z = [\ X_1 \quad Z_1\ ]$$

- X_1: Included exogenous variables, $n \times k_1$ matrix
- Z_1: Excluded exogenous variables, $n \times (m - k_1)$ matrix

Instrumental Variables

Identification

The parameters can be identified if

$$m - k_1 \geq k_2 \quad \text{that is,} \quad m \geq k$$

- If $m = k$, "just-identified" model
- If $m > k$, "over-identified" model

Instrumental Variables(IV) Estimator

The "just-identified" Model

$$\hat{\beta_{IV}} = (Z'X)^{-1}Z'y$$

The "over-identified" Model

$$\begin{aligned}\hat{\beta_{IV}} &= (\hat{X}'X)^{-1}\hat{X}'y \quad \text{where} \quad \hat{X} = P_z X = Z(Z'Z)^{-1}Z'X \\ &= (X'P_zX)^{-1}X'P_zy\end{aligned}$$

Digression

The Projection Matrix P_z

Assume X_i is each column of X

$$\begin{aligned} \hat{X}_i &= P_z X_i \\ &= Z(Z'Z)^{-1}Z'X \\ &= Z\hat{\delta}_{LS} \quad \text{from the model} \quad X = Z\delta + u \\ &= \text{The predicted value of } X_i \end{aligned}$$

Digression

When the matrices X'Z, Z'Z, and X'Z are a square matrix

$$\begin{aligned}\hat{\beta_{IV}} &= (X'P_zX)^{-1}X'P_zy \\ &= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y \\ &= (Z'X)^{-1}(Z'Z)(X'Z)^{-1}X'Z(Z'Z)^{-1}Z'y \\ &= (Z'X)^{-1}Z'y\end{aligned}$$

Two-Stage Least Squares (2SLS) Estimator

Step1: Regress each column of X_2 on Z

$$\hat{X} = [\ X_1 \quad \hat{X}_2\ ]$$

- $\hat{X}_2$ is the projected value of X_2
- You don't need to regress each column of X_1 on Z. (why?)

Step2: Regress y on $\hat{X}$

$$\begin{aligned}
\hat{\beta}_{2SLS} &= (\hat{X}'\hat{X})^{-1}\hat{X}'y \\
&= (X'P_z'P_zX)^{-1}X'P_z'y \\
&= (X'P_zX)^{-1}X'P_zy \\
&= \hat{\beta_{IV}}
\end{aligned}$$

IV estimator

Variance-covariance matrix of β_{IV}

$$Var(\hat{\beta_{IV}}) = \sigma^2 (X' P_z X)^{-1}$$

IV Estimator of σ^2

$$\hat{\sigma}_{IV} = \frac{1}{n}(y - X\hat{\beta}_{IV})'(y - X\hat{\beta}_{IV})$$