Random Regressors

Class 12

Wonmun Shin

(wonmun.shin@sejong.ac.kr)

Department of Economics, Sejong University

* This lecture note is written based on Professor Chang Sik Kim's lecture notes.

Random Regressors

Random Regressors

- Recall the classical assumptions in the simple regression model:
  - (A1) Independent variable (X) is not random but deterministic.
  - (A2) $E(e_i) = 0$
  - (A3) $Var(e_i) = \sigma^2$ *Homoskedasticity*
  - (A4) $Cov(e_i, e_j) = 0$ for $i \neq j$ *No Autocorrelation*
- We do not allow randomness in the regressors, which is obviously a restrictive assumption.
  - (A1) is clearly not realistic in many economic models, but we use this for mathematical simplicity.

Random Regressors [cont'd]

- However, even if we allow stochastic regressors, the properties of the OLS estimator still hold under slightly modified assumptions.
  - (A1)' $Cov(X_i, e_i) = 0$ *No correlation between X_i and e_i*
  - (A2)' $E(e_i \mid X_i) = 0$
  - (A3)' $Var(e_i \mid X_i) = \sigma^2$ *Homoskedasticity*
  - (A4)' $Cov(e_i, e_j \mid X_i, X_j) = 0$ for $i \neq j$ *No Autocorrelation*

Endogeneity Issue and IV Regression

Endogeneity Issue

- When X_i is random, it might be possible that X_i is correlated with the error term $(e_i) \rightarrow Cov\,(X_i,\, e_i) \neq 0$
- **Example:** earning (Y) and years of schooling (X)

$$Y_i = \beta_1 + \beta_2 X_i + e_i$$

  - e_i embodies all factors other than schooling that determines earnings, such as *ability*!
  - Ability affects earnings as well as years of schooling. $\rightarrow$ There is an association between X and e!

Endogeneity Issue [cont'd]

Q. What are the consequence of the correlation between X and e?

- Higher levels of X have two effect on Y!
  - Direct effect : $\beta_2 X_i$
    - Indirect effect : e affecting X, which in turn affects Y
  - The goal of regression is to estimate only the first effect, yielding an estimate of β_2.
  - But, the OLS estimate will combine two effects!

$$Y = \beta_1 + \beta_2 X + e(X)$$
$$\Rightarrow \quad \frac{dY}{dX} = \beta_2 + \frac{de}{dX}$$

  - The OLS estimate lets us know the total effect $\beta_2 + de/dX$ rather than β_2 alone.

A. The OLS estimator is **biased** and **inconsistent**!

Endogeneity Issue [cont'd]

- When X_i is endogenous, there may be an association between regressor and error. $\rightarrow Cov(X_i, e_i) \neq 0$
  - Source of **endogeneity**: measurement error in X, omitted-variable bias, reverse causality, $\cdots$.
- **Endogeneity Issue**
  - Without (A1)', the OLS estimator is **not unbiased** and **not consistent**.
  - That is, the OLS estimator does not converge to the true parameter even in a very large sample.
  - Moreover, none of usual testings and estimations are valid in this case.

Instrumental Variable (IV) Estimation

- Problems for OLS arises when X is random and correlated with the error, so that $Cov\left(X_i,\ e_i\right) \neq 0$.
- Suppose there exists another variable, say Z_i, which satisfies

$$Cov\left(Z_i,\ e_i\right) = 0 \quad (Exogeneity)$$

$$Cov(Z_i,\ X_i) \neq 0 \quad (Relevance)$$

- Z_i is called **instrumental variable (IV)**, or "instrument".
- **Example:** earning (Y) and years of schooling (X)
  - Card (1995) : Proximity to college or university (Z)
  - Angrist and Krueger (1991) : Month of birth (Z)

Instrumental Variable (IV) Estimation [cont'd]

- **IV estimator**

$$\hat{\beta}_2^{IV} = \frac{\Sigma\left(Z_i - \bar{Z}\right)\left(Y_i - \bar{Y}\right)}{\Sigma\left(Z_i - \bar{Z}\right)\left(X_i - \bar{X}\right)} = \frac{\Sigma z_i y_i}{\Sigma z_i x_i}$$

  - where $z_i = Z_i - \bar{Z}$, $x_i = X_i - \bar{X}$, and $y_i = Y_i - \bar{Y}$
  - IV estimators are **consistent** estimators (but, **not unbiased** and **not efficient**).

- **Example:** earning (Y) and years of schooling (X)
  - Suppose one unit change in Z is associated with 0.2 more years of schooling and with a \$500 increase in annual earnings.
  - Then, a one year increase in X is associated with $\$500/0.2 = \$2{,}500$ increase in Y.

$$\beta_2^{IV} = \frac{dY/dZ}{dX/dZ}$$

Instrumental Variable (IV) Estimation [cont'd]

$$\hat{\beta}_2^{IV} = \frac{\frac{\sum z_i y_i}{\sum z_i^2}}{\frac{\sum z_i x_i}{\sum z_i^2}} = \frac{\sum z_i y_i}{\sum z_i x_i}$$

- In practice, it is often estimated by the **two-stage least squares (2SLS)**.
  - [1st stage] Regress X on Z $\rightarrow$ Obtain $\hat{X}$
  - [2nd stage] Regress Y on $\hat{X}$ $\rightarrow$ Obtain $\hat{\beta}_2^{IV}$