Econometrics for Causal Inference

URP Last Part: Panel Data Basic of Basic

Sungkyunkwan University
- Machine Learning and Econometrics -

January 21, 2025

What We Have Covered

- What is Causal Inference
- Difference-in-Difference
- Regression Discontinuity

Introduction to Panel Data

- What and Why Panel Data Analysis
- Endogeneity Problem in Panel Data
- Solutions
  1. Fixed Effect
  2. Random Effect

What is Panel Data and Why is It Important

What is Panel Data

$$Y_{it}$$

- Individual i repeatedly appears in different time t
- In same time t, there are many individuals i

Two Forms of Panel Data Set

1. Balanced Panel
2. Unbalanced Panel

Why is Panel Data Important

- Large information
- Able to control unobservable individual characteristics
- We usually have "double-indexed" data

Endogeneity Problem in Panel Data

Assuming the true panel data model is

$$Y_{it} = \alpha + \beta X_{it} + \delta_t + \delta_i + \epsilon_{it}$$

- δ_i: unobservable time-invariant individual characteristics
- δ_t: unobservable common characteristics for all i in time t

  - If we introduce time dummies, the time effects can be easily controlled.
- Assume that there are really δ_i. However... if we ignore these factors and run a regression just using Y_{it} and X_{it}

  $\longrightarrow$ Error term$(u_{it}) = \delta_i + \epsilon_{it}$

  $\longrightarrow$ Then, it can be $E[X_{it} u_{it}] \neq 0$

Endogeneity Problem in Panel Data

Why is Endonegeity Important Problem?

- It leads to the bias and inconsistency of the estimator.
- In causal inference, if there is endogeneity problem, the identification fails.

$$\lim_{n \to \infty} \hat{\beta} \xrightarrow{p} \beta$$

Endogeneity Problem in Panel Data

Source of Endogeneity

- Reverse Causality
- Omitted Variable

$\longrightarrow$ Endogeneity problem in panel data and causal inference is relevant to above both

How can we solve it?

- Instrument Variable (IV) (All)
- **Fixed Effect, Random Effect Estimation (Panel)**
- Etc

Endogeneity Problem in Panel Data

Example: Endogeneity by Omitted Variable

Let's Assume "True" regression is

$$Y_i = \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i$$

If you regress equation ignoring X_2, that is, regress

$$Y_i = \beta_1 X_{1,i} + u_i$$

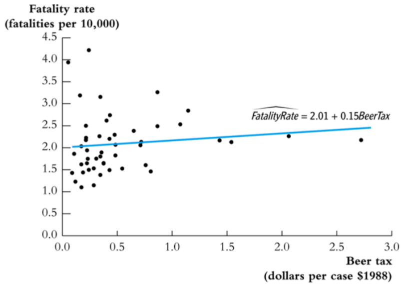
Then,

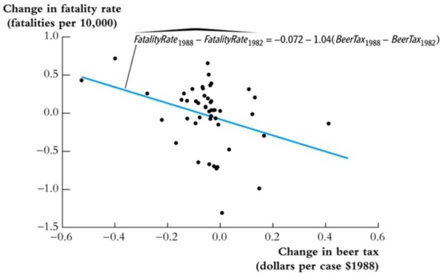
$$\hat{\beta}_1 = \frac{\Sigma X_{1,i} Y_i}{\Sigma {X_{1,i}}^2} = \beta_1 + \beta_2 \frac{\Sigma X_{1,i} X_{2,i}}{\Sigma {X_{1,i}}^2} + \frac{\Sigma X_{1,i} u_i}{\Sigma {X_{1,i}}^2}$$

Therefore,

$$E[\hat{\beta}_1] \neq \beta_1$$

Endogeneity Problem in Panel Data





Solution in Panel Data

$$Y_{it} = \alpha + \beta X_{it} + \delta_t + \delta_i + \epsilon_{it}$$

1. Fixed Effect
   - Assume that $Cov(X_{it}, \delta_i) \neq 0$
   - Within estimator, first difference estimator
2. Random Effect
   - Assume that $Cov(X_{it}, \delta_i) = 0$
   - GLS estimator

Solution 1: Fixed Effect

Assume that $Cov(X_{it}, \delta_i) \neq 0$

- Within Estimation

$$Y_{it} - \bar{Y}_i = \beta(X_{it} - \bar{X}) + \epsilon_{it} - \bar{\epsilon_{it}}$$

- First difference

$$Y_{it} - Y_{i,t-1} = \beta(X_{it} - X_{i,t-1}) + \epsilon_{it} - \epsilon_{it-1}$$

- By removing unobserved factor's effect with difference , we can get consistent estimator

Solution 2: Random Effect

Assume that $Cov(X_{it}, \delta_i) = 0$

$$Y_{it} = \alpha + \beta X_{it} + \delta_t + \delta_i + \epsilon_{it}$$

$$u_{it} = \delta_i + \epsilon_{it}$$

$$Y_{it} = \alpha + \beta X_{it} + \delta_t + u_{it}$$

Solution 2: Random Effect

GLS

$$Y_{it} = \alpha + \beta X_{it} + \delta_t + u_{it}$$

$$u_{it} = \delta_i + \epsilon_{it}$$

- Assume that $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$
- Also, assume that $\delta_i \sim N(0, \sigma_\delta^2)$ and $Cov(\epsilon, \delta_i) = 0$
- Then, we can find $Var(u_{it}) = Var(\delta_i + \epsilon_{it}) = \sigma_{u_{it}}^2$

$$Y_{it}/\sigma_{uit} = \alpha/\sigma_{uit} + \beta X_{it}/\sigma_{uit} + \delta_t/\sigma_{uit} + u_{it}/\sigma_{uit}$$

- Then, we can estimate the random effect model by GLS

FE vs RE

- If $Cov(\alpha_i, X_{it}) = 0$, FE and RE both consistent and RE is more efficient
- If $Cov(\alpha_i, X_{it}) \neq 0$, only FE is consistent
- Therefore we have to test

$$H_o : Cov(\alpha_i, X_{it}) = 0$$

- α_i is unobservable, so we only can compare $\hat{\beta}_{RE}$ and β_{FE}

  1. If both are similar, we cannot reject H_0 and use RE
  2. If both are not similar, we reject H_0 and use FE

Practical Packages in Programs

- Python: linearmodles (PanelOLS)
- R: PLM Package
- Stata: xtreg
- Julia: FixedEffects, FixedEffectModels