Panal Data 1: Framework

Instructor: Yuta Toyama

Last updated: 2021-07-09

Introduction

Introduction

- Panel data (パネルデータ)
  - combination of crosssection (クロスセクション) and time series (時系列) data

- Examples:
  1. Person i's income in year t.
  2. Vote share in county i for the presidential election year t.
  3. Country i's GDP in year t.

- Panel data is useful
  1. More variation (both cross-sectional and time series variation)
  2. Can deal with **time-invariant unobserved factors**.

Course Plan

- Framework
- Implementation in R
- **Difference-in-differences (DID, 差の差分法)**

Framework

Framework with Panel Data

- Consider the model

$$y_{it} = \beta' x_{it} + \epsilon_{it}, E[\epsilon_{it}|x_{it}] = 0$$

  where x_{it} is a k-dimensional vector

- If there is no correlation between x_{it} and ϵ_{it}, you can estimate the model by OLS **(pooled OLS)**

- A concern here is the omitted variable bias.

Introducing fixed effect (固定効果)

- Suppose that ϵ_{it} is decomposed as

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

  where α_i is called **unit fixed effect (固定効果)**, which is the time-invariant unobserved heterogeneity.

- With panel data, we can control for the unit fixed effects by incorporating the dummy variable for each unit i!

$$y_{it} = \beta' x_{it} + \gamma_2 D2_i + \cdots + \gamma_n Dn_i + u_{it}$$

  where Dl_i takes 1 if $l = i$.

Fixed Effect Model

- Model

$$y_{it} = \beta' x_{it} + \alpha_i + u_{it}$$

- Assumptions:
  1. u_{it} is uncorrelated with $(x_{i1}, \cdots, x_{iT})$, that is $E[u_{it}|x_{i1}, \cdots, x_{iT}] = 0$
  2. (Y_{it}, x_{it}) are independent across individual i.
  3. No outliers
  4. No perfect multicollinarity between explantory variables x_{it} and fixed effects α_i.

Assumption 1: Mean independence

- Assumption 1 is weaker than the assumption in OLS.
- Here, the time-invariant unobserved factor is captured by the fixed effect α_i.

Assumption 4: No Perfect Multicolinearity

- Consider the following model

$$wage_{it} = \beta_0 + \beta_1 experience_{it} + \beta_2 male_i + \beta_3 white_i + \alpha_i + u_{it}$$

  - $experience_{it}$ measures how many years worker i has worked before at time t.

- Multicollinearity issue because of $male_i$ and $white_i$.

- Intuitively, we cannot estimate the coefficient β_2 and β_3 because those **time-invariant variables are captured by the unit fixed effect α_i**.

Estimation

Estimation with Fixed Effects

- Can estimate the model by adding dummy variables for each individual.
  - **least square dummy variables (LSDV) estimator**.
  - Computationally demanding with many cross-sectional units
- We often use the following **within transformation**.

Estimation by within transformation

- Define the new variable $\tilde{Y}_{it}$ as

$$\tilde{Y}_{it} = Y_{it} - \bar{Y}_i$$

  where $\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$.

- Applying the within transformation, can eliminate the unit FE α_i

$$\tilde{Y}_{it} = \beta' \tilde{X}_{it} + \tilde{u}_{it}$$

- Apply the OLS estimator to the above equation!.

Importance of within variation in estimation

- The variation of the explanatory variable is key for precise estimation.
- Within transformation eliminates the time-invariant unobserved factor,
  - a large source of endogeneity in many situations.
- But, within transformation also absorbs the variation of X_{it}.
- Remember that

$$\tilde{X}_{it} = X_{it} - \bar{X}_i$$

  - The transformed variable $\tilde{X}_{it}$ has the variation over time t within unit i.
  - If X_{it} is fixed over time within unit i, $\tilde{X}_{it} = 0$, so that no variation.

Various Fixed Effects

- You can also add **time fixed effects (FE)**

$$y_{it} = \beta' x_{it} + \alpha_i + \gamma_t + u_{it}$$

- The regression above controls for both **time-invariant individual heterogeneity** and **(unobserved) aggregate year shock**.

- Panel data is useful to capture various unobserved shock by including fixed effects.

Cluster-Robust Standard Errors

- In OLS, we considered two types of error structures:
  1. Homoskedasticity $Var(u_i) = \sigma^2$
  2. Heteroskedasitcity $Var(u_i|x_i) = \sigma(x_i)$
- They assume the independence between observations, that is $Cov(u_i, u_{i'}) = 0$.
- In the panel data setting, we need to consider the **autocorrelation (自己相関)**.
  - the correlation between u_{it} and $u_{it'}$ across periods for each individual i.
- **Cluster-robust standard error (クラスターに頑健な標準誤差)** considers such autocorrelation.
  - The cluster is unit i. The errors within cluster are allowed to be correlated.