Panal Data 1: Framework

Instructor: Yuta Toyama

Last updated: 2020-03-30

Section 1

Introduction

Contents

- Framework
- Clustered Standard Errors
- Many FEs
- Implementation in R: felm command

Introduction

- Panel data has observations on n cross-sectional units at T time periods: $(X_{it}, Y_{it}$
- 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 because
  1. More variation (both cross-sectional and temporal variation)
  2. Can deal with time-invariant unobserved factors.
  3. (Not focus in this course) Dynamics of individual over time.

Overview

- 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 natural concern here is the omitted variable bias.
- We now consider that ϵ_{it} is written as

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

  where α_i is called **unit fixed effect**, 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 $l = i$.

  - Notice that we cannot do this for the cross-section data!
- We often write the model with unit FE as

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

Framework

- The fixed effects 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
- Let's discuss Assumptions 1, 2, and 4 in detail.

- Assumption 1 is weaker than the assumption in OLS, because the time-invariant factor α_i is captured by the fixed effect.
  - Example: Unobserved ability is caputured by α_i.
- Assumption 2 allows for serial correlation (i.e., $Cov(x_{it}, x_{it'}) \neq 0$) within individual i.
  - This is related to the cluster-robust standard error.
- Assumption 4 seems as usual, but it has an important role in panel data analysis.
- Consider the following regression with unit FE

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

  where $experience_{it}$ measures how many years worker i has worked before at time t.
  - In the regression above, we have 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 completely captured by the unit fixed effect α_i.

Estimation (within transformation)

- You can estimate the model by adding dummy variables for each individual. This is called **least square dummy variables (LSDV) estimator**.
- This is computationary demanding if we have many cross-sectional observations.
- We often use the following **within transformation**.
- Define the new variable $\tilde{Y}_{it}$ as

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

  where $\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$.
- Why is this useful? By applying the within transformation to the regression model, we can eliminate the unit fixed effect α_i

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

  Then apply the OLS estimator to the above equation!.

Importance of within variation

- As I talked before, the variation of the explanatory variable is key for precisely estimating the coefficients (once we control for the endogeneity).
- Within transformation eliminates the time-invariant unobserved factor, which is 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.

FE, FE, and FE

- In addition to unit FE, 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.

Panel + IV

- You can use IV regression with panel data. This is PS5.

Standard Errors

- In the cross-section data, we need to use the heteroskedasticity robust standard error.
  - Remember: Heteroskedasticity means $Var(u_i|x_i) = \sigma(x_i)$.
- In the panel data setting, we need to consider the **autocorrelation** of the variable, that is the correlation between u_{it} and $u_{it'}$ across periods for each individual i.
- The current standard is to use so-called **cluster-robust standard error**.
  - The cluster is unit i. The observations within cluster are allowed to be freely correlated.
  - Cluster-robust standard error takes care for such correlation.
- I will explain how to deal with this in R.