Recitation 7

Seung-hun Lee

Columbia University

Panel regression

Motivation

- **Panel data**: We observe multiple individuals for multiple periods of time.

$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + ... + \beta_k X_{k,it} + u_{it}$$

  $i = 1, 2, ..., N \to$ individuals, $t = 1, 2, ..., T \to$ time periods.
- Balanced: There are T datasets for each of the N individuals.
- Unbalalced: There are $t \leq T$ datapoints for some of the N individuals.

Advantages

- Panel data allows us to use more datasets.
- Panel data allows us to control for **unobserved heterogeneity** that are
  1. different accross N entities but always remain same for T periods in a given state (**cross section fixed effect**)
  2. different accross T times periods but remains the same for all N entities in a particular time period (**time fixed effects**)
  3. both of 1) and 2). (**two-way fixed effects**)

Panel regression

Key differences

- Suppose that $T = 2$ and we are interested in the relationship between vehicle related fatality rate (deaths per 10,000 people) and the beer tax. Suppose that we get these result for the two years

$$\begin{aligned}\hat{Y}_{i1} =& 2.01+ \quad 0.15X_{i1} \\ & (0.15) \quad\ (0.20) \\ \hat{Y}_{i2} =& 1.86+ \quad 0.44X_{i2} \\ & (0.11) \quad\ (0.20)\end{aligned}$$

- In such case, one might suspect that there is an omitted variable bias that affects these coefficients.
  - Omitted variable specific to the states (Strictness of the relevant law)
  - Time-trends? (Specific to each of years 1 and 2)

Panel regression

How can panel regression do better?

- Let Z_i denote the strictness of state laws on DUI that are unchanging.
- Now write

$$Y_{i1} = \beta_0 + \beta_1 X_{i1} + \beta_2 Z_i + u_{i1}$$
$$Y_{i2} = \beta_0 + \beta_1 X_{i2} + \beta_2 Z_i + u_{i2}$$

- Subtract the second equation from the first to get

$$(Y_{i2} - Y_{i1}) = \beta_1 (X_{i2} - X_{i1}) + \beta_2 (Z_i - Z_i) + u_{i2} - u_{i1}$$

With Z_i being the same for all periods, the above equation is reduced to

$$(Y_{i2} - Y_{i1}) = \beta_1 (X_{i2} - X_{i1}) + (u_{i2} - u_{i1})$$

- The Z_i variable has no role in this equation - because it is now gone.
- If we estimate this particular β_1, we can obtain much more accurate estimates of the effect of bear tax on fatality rate.

Panel regression

Specific methodologies (cross-sectional FE)

- There are two ways of estimating the data when $T \geq 3$
- Least square dummy variables: Include $N - 1$ individual dummies
- Within estimation: Subtract "demeaned" equation from the original
- Use:
$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it} \tag{1}$$
  where Z_i is the cross section fixed effect.
- Define $\alpha_i = \beta_0 + \beta_1 Z_i$. Then the above equation can be written as
$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it} \tag{2}$$
- α_i term can be thought of as an effect of being an entity i, which is **correlated with** X_{it}

Panel regression

LSDV method

- Define a new variable D_{ki} as follows

$$D_{ki} = \begin{cases} 1 & \text{If } i = k \\ 0 & \text{Otherwise} \end{cases}, \; k \in \{1, 2, ..., N\}$$

- Since we are going to include β_0, a common intercept, in our regression we need to remove one of the N (dummy variable trap)
- Then we can write

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_2 D_{2i} + ... + \delta_N D_{Ni} + u_{it} \tag{LSDV}$$

- This equation gives different intercepts for each i (can you see why?), while keeping the slope on X_{it} constant at β_1
- Control for unobserved cross section fixed effect by allowing the intercept to differ by each i

Panel regression

WE method

- Define $\bar{X}_i$, $\bar{Y}_i$ as sample mean of X_{it}, Y_{it} for given i over all possible t's.

$$\bar{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it}$$

  Consequently, $\bar{Y}_i$ can be written as

$$\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it} = \frac{1}{T}\sum_{t=1}^{T}(\beta_1 X_{it} + \alpha_i + u_{it}) = \beta_1 \bar{X}_i + \alpha_i + \bar{u}_i$$

- Subtract Y_{it} by $\bar{Y}_i$ to get

$$Y_{it} - \bar{Y}_i = \beta_1(X_{it} - \bar{X}_i) + (u_{it} - \bar{u}_i) \implies \tilde{Y}_{it} = \beta_1 \tilde{X}_{it} + \tilde{u}_{it}$$

- This process gets rid of α_i. Then, apply OLS estimation on this equation to get the within estimator

Panel regression

TWFE methods

- We have a DGP

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it}$$

- LSDV: With an overall constant β_0, we can put $N-1$ individual and $T-1$ time dummies
- WE: Demeaning should be done in the following method

$$Y_{it} - \bar{Y}_i - \bar{Y}_t + \bar{Y}$$

  This (and only this) would allow us to get rid of both the α_i individual fixed effect and the λ_t time effects

Panel regression

Least square assumptions for panels

- **P1** : $E[u_{it}|X_{i1},..,X_{iT},\alpha_i]=0$. It means that the conditional mean of the u_{it} term does not depend on any of the X_{it} values for entity i, whether in the future or in the past.
- **P2** : $(X_{i1},..,X_{iT},u_{i1},...u_{iT})$ is IID across $i=1,..,n$. **This does not rule out the correlation between u_{it},u_{ij} within entity i for different j and t**, allowing serial correlation within the same entity
- **P3** : (X_{it},u_{it}) have nonzero finite fourth moments (outliers are very unlikely) so that the panel estimators have a distribution
- **P4** : There is no perfect multicollinearity

$\rightarrow$ Because of P2, we need to use **clustered standard error** at a cross-sectional level.