Econometrics Lecture 12

Panel Data Analysis

Hang Miao

Rutgers University

April 13, 2021

Overview

1. Introduction
   - Definition
   - Example
2. Panel Data with Two Time Periods
3. (Entity) Fixed Effects Regression
   - Fixed Effects Form
   - $n-1$ Binary Regressors Form
   - Estimation Method: Overview
   - Estimation Method for $n-1$ Binary Regressors Form
   - Estimation Method for Fixed Effects Form: Entity-Demeaned OLS
4. Time Fixed Effects Regression
5. Both Entity Fixed and Time Fixed Effects Regression

Introduction

What is Panel Data

Definition (Panel Dataset)

A panel dataset contains observations on multiple entities (individuals, states, companies), where each entity is observed at two or more points in time(multiple time points).

Hypothetical examples:

- Data on 420 California school districts in 1999 and again in 2000, for 840 observations total.
- Data on 50 U.S. states, each state is observed in 3 years, for a total of 150 observations.
- Data on 1000 individuals, in four different months, for 4000 observations total.

Notations For Panel Data

Double Subscript Notation

A double subscript distinguishes entities and time periods

- i = entity (district), n = number of entities, so i = 1, $\cdots$,n
- t = time period (year), T = number of time periods so t = 1, $\cdots$,T
- the sample data could be denoted as $\{\boldsymbol{x}_{i,t}, y_{i,t}\}$

Triple Subscript Notation for x

A triple subscript distinguishes entities (district), features (STR, ElPct, PctLch) and time periods (years)

- i = entity (district), n = number of entities, so i = 1, $\cdots$,n
- j = features(STR, ELPct, PctLch)
- t = time period (year), T = number of time periods so t = 1, $\cdots$,T
- the sample data for i-th entity and j-th feacould at t-th period could be denoted as $\{x_{ij,t}\}$

Allias Name

- Another term for panel data is longitudinal data
- balanced panel: no missing observations, that is, all variables are observed for all entities (states) and all time periods (years)

Why Use Panel Data

With panel data we can control for factors that:

- vary across entities but do not vary over time
- could cause omitted variable bias if they are omitted
- are unobserved or unmeasured and therefore cannot be included in the regression using multiple regression

Key Idea:

If an omitted variable does not change over time, then any changes in Y over time cannot be caused by the omitted variable.

Example of a panel data set: Traffic deaths and alcohol taxes

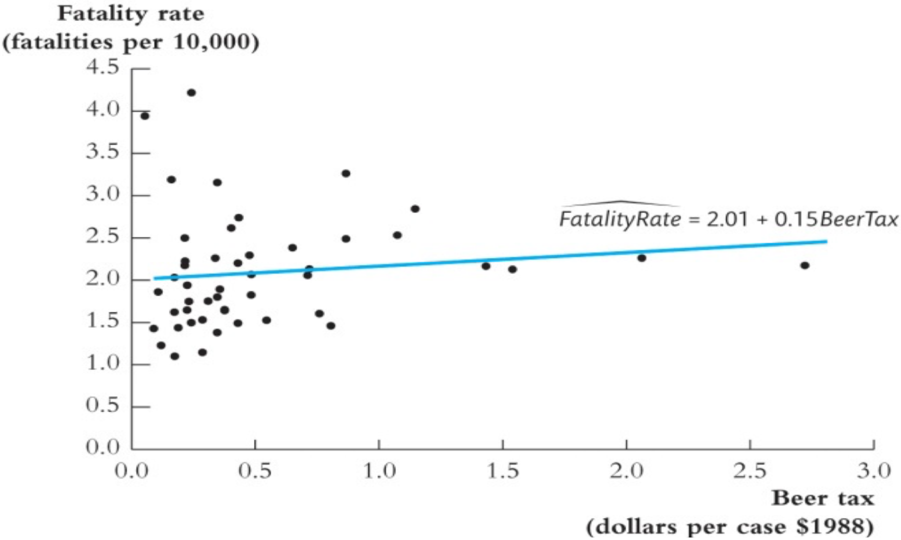
Observational unit: a year in a U.S. state

- 48 U.S. states, so n = # of entities = 48
- 7 years (1982,$\cdots$, 1988), so T = # of time periods = 7
- Balanced panel, so total # observations = 7 × 48 = 336

Variables

- Dependent Variable: Traffic fatality rate (traffic deaths in that state in that year, per 10,000 state residents)
- Independent Variable: Tax on a case of beer
- Other Independent Variable: legal driving age, drunk driving laws, etc.

Higher alcohol taxes, more traffic deaths?



$\widehat{FatalityRate} = 2.01 + 0.15 BeerTax$

Why might there be higher more traffic deaths in states that have higher alcohol taxes?

Other factors that determine traffic fatality rate:

- Quality (age) of automobiles
- Quality of roads
- Culture around drinking and driving
- Density of cars on the road

These Omitted Variable could cause Omitted Variable Bias

Traffic Density

- High traffic density means more traffic deaths
- (Western) states with lower traffic density have lower alcohol tax
- The two conditions for omitted variable bias are satisfied. Specifically, high taxes could reflect high traffic density (so the OLS coefficient would be biased positively high taxes, more deaths)

Cultural Attitudes

- Cultural attitudes towards drinking and driving arguably are a determinant of traffic deaths; and
- potentially are correlated with the beer tax.
- Then the two conditions for omitted variable bias are satisfied. Specifically, high taxes could pick up the effect of cultural attitudes towards drinking so the OLS coefficient would be biased

Panel Data with Two Time Periods

Model Setting

Underlying Model

$$\text{FatalityRate}_{i,t} = \beta_0 + \beta_1 \text{ BeerTax}_{i,t} + \beta_2 Z_i + u_{i,t}$$

- Where Z_i is a factor that does not change over time (traffic density, culture attitudes), at least during the years on which we have data.
- Suppose Z_i is not observed, so its omission could result in omitted variable bias.
- The effect of Z_i can be eliminated using T = 2 years.

The key idea

Any change in the fatality rate from one year t_1 to another year t_2 cannot be caused by Z_i , because Z_i (by assumption) does not related to time parameter t.

Two Time Periods Panel Data Estimation

Consider fatality rates in 1988 and 1982:

$$\begin{aligned} \text{FatalityRate}_{i,1982} &= \beta_0 + \beta_1 \text{ BeerTax}_{i,1982} + \beta_2 Z_i + u_{i,1982} && (1) \\ \text{FatalityRate}_{i,1988} &= \beta_0 + \beta_1 \text{ BeerTax}_{i,1988} + \beta_2 Z_i + u_{i,1988} && (2) \end{aligned}$$

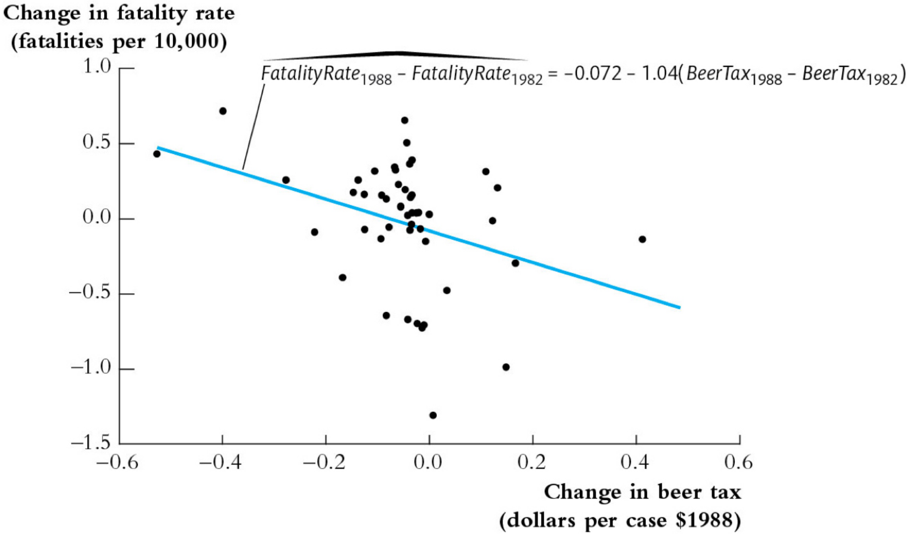
Subtracting equation $(1)_{1982} - (2)_{1988}$

$$\begin{aligned} \text{FatalityRate}_{i,1982} - \text{ FatalityRate}_{i,1988} = \beta_1 \big(\text{BeerTax}_{i,1982} - \text{BeerTax}_{i,1988}\big) \\ + \big(u_{i,1982} - u_{i,1988}\big) \end{aligned}$$

Consider fatality rates in 1988 and 1982:

- The new error term, $\big(u_{i,1982} - u_{i,1988}\big)$, is uncorrelated with either $\big(\text{BeerTax}_{i,1982} - \text{BeerTax}_{i,1988}\big)$.
- This difference equation can be estimated by OLS

Higher alcohol taxes, more traffic deaths?



(Entity) Fixed Effects Regression

Introduction

Model Setting

What if the panel data set has more than 2 time periods ($T > 2$)?

$$y_{i,t} = \beta_0 + \boldsymbol{\beta} \mathbf{x}_{i,t} + \beta_z Z_i + u_{i,t}, \quad i = 1, \cdots, n, \ t = 1, \cdots, T$$

Fixed Effects Form

$$y_{i,t} = \boldsymbol{\beta} \mathbf{x}_{i,t} + \alpha_i + u_{i,t}$$

$n - 1$ Binary Regressors Form

$$y_{i,t} = \boldsymbol{\beta} \mathbf{x}_{i,t} + \gamma_1 D_{i1} + \ldots + \gamma_{n-1} D_{i(n-1)} + u_{i,t}$$

where

$$D_{ij} = \begin{cases} 1 \text{ if } i = j \text{ (entity } \#j) \\ 0 \text{ otherwise} \end{cases}$$

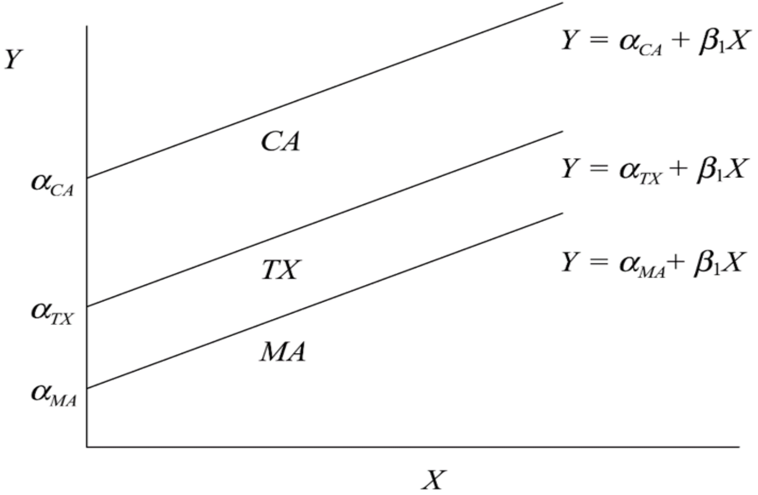
Example: Fixed Effects Form

We could first rewrite the previous example in fixed effects form. Suppose we have n = 3 states: California, Texas, and Massachusetts.

$$\begin{aligned} y_{CA,t} &= \beta_0 + \beta_1 x_{CA,t} + \beta_z Z_{CA} + u_{CA,t} \\ &= \alpha_{CA} + \beta_1 x_{CA,t} + u_{CA,t} \\ y_{TX,t} &= \beta_0 + \beta_1 x_{TX,t} + \beta_z Z_{TX} + u_{TX,t} \\ &= \alpha_{TX} + \beta_1 x_{TX,t} + u_{TX,t} \\ y_{MA,t} &= \beta_0 + \beta_1 x_{MA,t} + \beta_z Z_{MA} + u_{MA,t} \\ &= \alpha_{MA} + \beta_1 x_{MA,t} + u_{MA,t} \end{aligned}$$

- where $\alpha_i = \beta_0 + \beta_z Z_i$ (for example $\alpha_{CA} = \beta_0 + \beta_z Z_{CA}$) doesn't change over time
- α_i is the intercept for i state, and β_1 is the slope
- The intercept α_i is unique to each state, but the slope β_1 is the same in all the states: parallel lines.

The Regression Lines for each state as Binary Regressors


Y
α_{CA}
α_{TX}
α_{MA}
CA
TX
MA
$Y = \alpha_{CA} + \beta_1 X$
$Y = \alpha_{TX} + \beta_1 X$
$Y = \alpha_{MA} + \beta_1 X$
X

Example: $n-1$ Binary Regressors Form

The $n-1$ Binary Regressor form is:

$$y_{i,t} = \beta_0 + \beta_1 x_{i,t} + \gamma_{CA} D_{iCA} + \gamma_{TX} D_{iTX} + u_{i,t}$$

- where $D_{iCA} = 1$ if the state $i = CA$, otherwise $D_{iCA} = 0$
- where $D_{iTX} = 1$ if the state $i = TX$, otherwise $D_{iTX} = 0$
- leave out D_{iMA} (why?)

Estimation Method

Three Estimation Method

1. $n - 1$ binary regressors OLS regression
2. Entity-demeaned OLS regression
3. Changes specification, without an intercept (only works for T $=$ 2)

- These three methods produce identical estimates of the regression coefficients, and identical standard errors.
- We already did the changes specification (1988 minus 1982) but this only works for T $=$ 2 years
- Methods 1 and 2 work for general T
- Method 1 is only practical when n isnt too big

Estimation Method for $n-1$ Binary Regressors Form

$n-1$ Binary Regressors Form

$$y_{i,t} = \beta \boldsymbol{x}_{i,t} + \gamma_1 D_{i1} + \ldots + \gamma_{n-1} D_{i(n-1)} + u_{i,t}$$

where

$$D_{ij} = \begin{cases} 1 \text{ if } i = j \text{ (entity j)} \\ 0 \text{ otherwise} \end{cases}$$

- Estimate the $n-1$ Binary Regressors linear model by OLS
- Inference (hypothesis tests, confidence intervals) is as usual (using heteroskedasticity-robust standard errors)
- This is impractical when n is very large

Estimation Method for Fixed Effects Form: Entity-Demeaned OLS

Fixed Effects Form

$$y_{i,t} = \alpha_i + \boldsymbol{\beta} \boldsymbol{x}_{i,t} + u_{i,t}$$

Entity-Demeaned OLS

$$\frac{1}{T}\sum_{t=1}^{T} y_{i,t} = \alpha_i + \boldsymbol{\beta}\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_{i,t} + \frac{1}{T}\sum_{t=1}^{T} u_{i,t}$$

$$y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{i,t} = \boldsymbol{\beta}\left(\boldsymbol{x}_{i,t} - \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_{i,t}\right) + \left(u_{i,t} - \frac{1}{T}\sum_{t=1}^{T} u_{i,t}\right)$$

- Let $\tilde{y}_{i,t} = \left(y_{i,t} - \frac{1}{T}\sum_{t=1}^{T} y_{i,t}\right)$, $\tilde{\boldsymbol{x}}_{i,t} = \left(\boldsymbol{x}_{i,t} - \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_{i,t}\right)$
- Run OLS (regressing $\tilde{y}_{i,t}$ on $\tilde{\boldsymbol{x}}_{i,t}$) to estimate $\boldsymbol{\beta}$

Time Fixed Effects Regression

Introduction

An omitted variable might vary over time but not across states:

- Safer cars (air bags, etc.); changes in national laws
- These produce intercepts that change over time
- Let S_t denote the combined effect of variables which changes over time but not states (safer cars). The resulting population regression model is:

$$y_{i,t} = \beta_0 + \boldsymbol{\beta} \mathbf{x}_{i,t} + \beta_s S_t + u_{i,t}$$

Time Fixed Effects Form

$$y_{i,t} = \boldsymbol{\beta} \mathbf{x}_{i,t} + \lambda_t + u_{i,t}$$

$T - 1$ Binary Regressors Form

$$y_{i,t} = \boldsymbol{\beta} \mathbf{x}_{i,t} + \delta_1 B_{t1} + \ldots + \delta_{T-1} B_{t(T-1)} + u_{i,t}$$

where

$$B_{tj} = \begin{cases} 1 \text{ if } t = j \text{ (year } \#j) \\ 0 \text{ otherwise} \end{cases}$$

Estimation Method for $T-1$ Binary Regressors Form

$T-1$ Binary Regressors Form

$$y_{i,t} = \boldsymbol{\beta} \boldsymbol{x}_{i,t} + \delta_1 B_{t1} + \ldots + \delta_{T-1} B_{t(T-1)} + u_{i,t}$$

where

$$B_{tj} = \begin{cases} 1 \text{ if } t = j \text{ (period j)} \\ 0 \text{ otherwise} \end{cases}$$

- Estimate the above $T-1$ Binary Regressors linear model by OLS
- Inference (hypothesis tests, confidence intervals) is as usual (using heteroskedasticity-robust standard errors)
- This is impractical when T is very large

Estimation Method for Time Fixed Effects Form: Time-Demeaned OLS

Time Fixed Effects Form

$$y_{i,t} = \boldsymbol{\beta} \boldsymbol{x}_{i,t} + \lambda_t + u_{i,t}$$

Time-Demeaned OLS

$$\frac{1}{n}\sum_{i=1}^{n} y_{i,t} = \boldsymbol{\beta}\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_{i,t} + \lambda_t + \frac{1}{n}\sum_{i=1}^{n} u_{i,t}$$

$$y_{it} - \frac{1}{n}\sum_{i=1}^{n} y_{i,t} = \boldsymbol{\beta}\left(\boldsymbol{x}_{i,t} - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_{i,t}\right) + \left(u_{i,t} - \frac{1}{n}\sum_{i=1}^{n} u_{i,t}\right)$$

- Let $\tilde{y}_{i,t} = \left(y_{i,t} - \frac{1}{n}\sum_{i=1}^{n} y_{i,t}\right)$, $\tilde{\boldsymbol{x}}_{i,t} = \left(\boldsymbol{x}_{i,t} - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_{i,t}\right)$
- Run OLS (regressing $\tilde{y}_{i,t}$ on $\tilde{\boldsymbol{x}}_{i,t}$) to estimate $\boldsymbol{\beta}$

Both Entity Fixed and Time Fixed Effects Regression

Underlying Model Setting

$$y_{i,t} = \beta_0 + \boldsymbol{\beta} \boldsymbol{x}_{i,t} + \beta_z Z_i + \beta_s S_t + u_{i,t}$$

Both Entity and Time Fixed Effects Form

$$y_{i,t} = \boldsymbol{\beta} \boldsymbol{x}_{i,t} + \alpha_i + \lambda_t + u_{i,t}$$

$T-1$ Time Binary and $n-1$ Entity Binary Regressors Form

$$\begin{aligned} y_{i,t} = \boldsymbol{\beta} \boldsymbol{x}_{i,t} &+ \gamma_1 D_{i1} + \ldots + \gamma_{n-1} D_{i(n-1)} \\ &+ \delta_1 B_{t1} + \ldots + \delta_{T-1} B_{t(T-1)} + u_{i,t} \end{aligned}$$

where the binary variable D_{ij} and B_{tj} is defined as follow:

$$D_{ij} = \begin{cases} 1 \text{ if } i = j \text{ (entity j)} \\ 0 \text{ otherwise} \end{cases} , \quad B_{tj} = \begin{cases} 1 \text{ if } t = j \text{ (period j)} \\ 0 \text{ otherwise} \end{cases}$$

Estimation Method for Both Entity Fixed and Time Fixed Effects

When time period $T = 2$

- Computing the difference of dependant variable $y_{i,t_1} - y_{i,t_2}$ and independant variable $\boldsymbol{x}_{i,t_1} - \boldsymbol{x}_{i,t_2}$ with respect to the time dimension for each entity i.
- Let $\tilde{y}_i = y_{i,t_1} - y_{i,t_2}$, $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_{i,t_1} - \boldsymbol{x}_{i,t_2}$.
- Regress $\tilde{y}_i$ on $\tilde{\boldsymbol{x}}_i$ by OLS with an intercept included.
- Note that it is different from the two period estimation method introduce at the beginning which does not include an intercept. The reason is it only account for entity fixed effect while it account for both entity and time fixed effect here
- The interpretation for the intercept is the change of the time fixed effect $\lambda_{t_1} - \lambda_{t_2}$

Estimation Method for Both Entity Fixed and Time Fixed Effects

When time period $T > 2$

There are four equivalent ways to estimate underlying linear model when both entity and time fixed effects are incorporated:

- Entity Demeaning with $T - 1$ Time Binary Variable
- Time Demeaning with $n - 1$ Entity Binary Variable
- $T - 1$ Time Binary Variable and $n - 1$ Entity Binary Variable
- Entity Demeaning and Time Demeaning (or Time Demeaning and Entity Demeaning)

Estimation Method for Both Entity Fixed and Time Fixed Effects: Time Demeaning

$$y_{i,t} = \boldsymbol{\beta} \boldsymbol{x}_{i,t} + \alpha_i + \lambda_t + u_{i,t} \tag{1}$$

$$\frac{1}{n}\sum_{i=1}^{n} y_{i,t} = \boldsymbol{\beta}\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_{i,t} + \frac{1}{n}\sum_{i=1}^{n}\alpha_i + \lambda_t + \frac{1}{n}\sum_{i=1}^{n} u_{i,t} \tag{2}$$

Subtract (2) from (1) we have

$$\begin{aligned}\left(y_{i,t} - \frac{1}{n}\sum_{i=1}^{n} y_{i,t}\right) = \boldsymbol{\beta}\left(\boldsymbol{x}_{i,t} - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_{i,t}\right) + \left(\alpha_i - \frac{1}{n}\sum_{i=1}^{n}\alpha_i\right) \\ + \left(u_{i,t} - \frac{1}{n}\sum_{i=1}^{n} u_{i,t}\right)\end{aligned} \tag{3}$$

Estimation Method for Both Entity Fixed and Time Fixed Effects: Reduced to the Entity Fixed Effects Problem

Let

$$\tilde{y}_{i,t} = \left(y_{i,t} - \frac{1}{n} \sum_{i=1}^{n} y_{i,t} \right), \quad \tilde{\boldsymbol{x}}_{i,t} = \left(\boldsymbol{x}_{i,t} - \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_{i,t} \right)$$

$$\tilde{\alpha}_i = \left(\alpha_i - \frac{1}{n} \sum_{i=1}^{n} \alpha_i \right), \quad \tilde{u}_{i,t} = \left(u_{i,t} - \frac{1}{n} \sum_{i=1}^{n} u_{i,t} \right)$$

Equation (3) could be rewrite as follow:

$$\tilde{y}_{i,t} = \boldsymbol{\beta} \tilde{\boldsymbol{x}}_{i,t} + \tilde{\alpha}_i + \tilde{u}_{i,t} \tag{4}$$

Therefore, the Time Fixed Effect λ_t is eliminated by Time Demeaning process. Equation (4) is reduced to the original Entity Fixed Effect problem. We could adopt either "$n-1$ binary regressor method" or "Entity Demeaning method" to estimate $\boldsymbol{\beta}$

Estimation Method for Both Entity Fixed and Time Fixed Effects: Entity Demeaning after Time Demeaning

$$\tilde{y}_{i,t} = \boldsymbol{\beta}\tilde{\boldsymbol{x}}_{i,t} + \tilde{\alpha}_i + \tilde{u}_{i,t} \tag{5}$$

$$\frac{1}{T}\sum_{t=1}^{T}\tilde{y}_{i,t} = \boldsymbol{\beta}\frac{1}{T}\sum_{t=1}^{T}\tilde{\boldsymbol{x}}_{i,t} + \tilde{\alpha}_i + \frac{1}{T}\sum_{t=1}^{T}\tilde{u}_{i,t} \tag{6}$$

Subtract (6) from (5) we have

$$\left(\tilde{y}_{i,t} - \frac{1}{T}\sum_{t=1}^{T}\tilde{y}_{i,t}\right) = \boldsymbol{\beta}\left(\tilde{\boldsymbol{x}}_{i,t} - \frac{1}{T}\sum_{t=1}^{T}\tilde{\boldsymbol{x}}_{i,t}\right) + \left(\tilde{u}_{i,t} - \frac{1}{T}\sum_{t=1}^{T}\tilde{u}_{i,t}\right) \tag{7}$$

Estimation Method for Both Entity Fixed and Time Fixed Effects: Entity Demeaning after Time Demeaning

Let

$$\breve{y}_{i,t} = \left(\tilde{y}_{i,t} - \frac{1}{T} \sum_{t=1}^{T} \tilde{y}_{i,t} \right), \quad \breve{\boldsymbol{x}}_{i,t} = \left(\tilde{\boldsymbol{x}}_{i,t} - \frac{1}{T} \sum_{t=1}^{T} \tilde{\boldsymbol{x}}_{i,t} \right)$$

$$\breve{u}_{i,t} = \left(\tilde{u}_{i,t} - \frac{1}{T} \sum_{t=1}^{T} \tilde{u}_{i,t} \right)$$

Equation (7) could be rewrite as follow:

$$\breve{y}_{i,t} = \boldsymbol{\beta} \breve{\boldsymbol{x}}_{i,t} + \breve{u}_{i,t} \tag{8}$$

Therefore, the Entity Fixed Effect $\tilde{\alpha}_i$ is eliminated by Entity Demeaning process. Equation (8) is reduced to the Classical Linear Model. We could OLS to estimate $\boldsymbol{\beta}$

Estimation Method for Both Entity Fixed and Time Fixed Effects: $n-1$ Binary Regressor Method after Time Demeaning

Equation (4) could be rewrite in the form of $n-1$ Binary Regressor:

$n-1$ Binary Regressors Form

$$\tilde{y}_{i,t} = \boldsymbol{\beta}\tilde{\boldsymbol{x}}_{i,t} + \gamma_1\tilde{D}_{i1} + \ldots + \gamma_{n-1}\tilde{D}_{i(n-1)} + \tilde{u}_{i,t}$$

where

$$\tilde{D}_{ij} = \begin{cases} 1 \text{ if } i = j \text{ (entity j)} \\ 0 \text{ otherwise} \end{cases}$$

- Estimate the $n-1$ Binary Regressors linear model by OLS
- Inference (hypothesis tests, confidence intervals) is as usual (using heteroskedasticity-robust standard errors)
- This is impractical when n is very large