

Fixed Effects Regression

Prof. Tzu-Ting Yang

楊子霆

Institute of Economics, Academia Sinica

中央研究院經濟研究所

October 9, 2024

Observables and Unobservables Confounding Factors

- The main problem we face in estimating causal effect is that:
 - Each individual, firm, state, or country can select treatment
 - This choice could be correlated with factors that affect the outcomes of interest, which results in selection bias
- So far the key strategy to obtain causal effect was to control for **observed** confounding factors
- Yet, what if important confounding factors are unobserved?

Fixed Effects Regression: Main Idea

Fixed Effects Regression

- If unobserved confounding factors are **time-invariant**
 - If we have multiple time periods panel data or cross-sectional data
 - We can still obtain causal effect by estimating a regression that include many fixed effects in the model

Fixed Effects Regression

Example

- Suppose we are interested in the question whether joining union increase workers' earnings
- We might want to estimate the following regression:

$$Y_{it} = \delta + \alpha D_{it} + A_i' \gamma + X_{it}' \beta + \varepsilon_{it}$$

- Y_{it} is outcome variable: earnings
- D_{it} is treatment variable: union status
- X_{it} are observed time-varying covariates: experience, education
- A_i is unobserved but fixed confounder (time-invariant): ability or personality
- Assume $E[\varepsilon_{it} | A_i, X_{it}] = 0$

Fixed Effects Regression

Example

- This regression equation implies the following potential outcomes:

$$Y_{it}^0 = \delta + A_i' \gamma + X_{it}' \beta + \varepsilon_{it}$$

$$Y_{it}^1 = Y_{it}^0 + \alpha$$

Fixed Effects Regression

Example

- Because A_i is unobservable, we are not able to directly include it in the regression

$$Y_{it} = \delta + \alpha D_{it} + X'_{it}\beta + \underbrace{A'_i\gamma}_{u_{it}} + \varepsilon_{it}$$

- If A_i is correlated with union status D_{it}
 - There is a correlation of D_{it} with the error term u_{it}
 - This will lead to **omitted variable bias**

Fixed Effects Regression

Example

- Address this problem by including λ_i in the regression
 - $\lambda_i = \delta + A_i'\gamma$
 - That is, we can consider λ_i as individual-specific constant term
- We estimate the following regression with individual fixed effects

$$Y_{it} = \lambda_i + \alpha D_{it} + X_{it}'\beta + \varepsilon_{it} \quad (1)$$

- Therefore, D_{it} and the error term ε_{it} would be uncorrelated
- Then, OLS estimate of α is unbiased

Fixed Effects Regression

Estimation

- In practice, there are two ways of estimating this fixed effects model:
 1. Demeaning approach (sometimes called “within estimator”)
 2. Regression with ‘N-1 dummy variables’

Demeaning Approach

- 1 Calculate individual averages of the outcome variable and all covariates (over time)

$$\bar{Y}_i = \bar{\lambda}_i + \alpha \bar{D}_i + \bar{X}_i' \beta + \bar{\varepsilon}_i$$

- 2 Subtract these averages from regression equation (1):

$$Y_{it} - \bar{Y}_i = \alpha(D_{it} - \bar{D}_i) + (X_{it} - \bar{X}_i)' \beta + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

- Note that λ_i drops out since it is time-invariant
- Therefore the error ε_{it} and the treatment D_{it} would no longer be correlated.

Regression with ‘N-1 dummy variables’

$$Y_{it} = \delta + \sum_{i=2}^N \rho_i B_i + \alpha D_{it} + X'_{it} \beta + \varepsilon_{it}$$

- B_i is a dummy indicating individual i
- We only include $N - 1$ individual dummies to avoid collinearity
- We show that this representation is actually the same as a regression with fixed effects λ_i

$$Y_{it} = \lambda_i + \alpha D_{it} + X'_{it} \beta + \varepsilon_{it}$$

Regression with ‘N-1 dummy variables’

- Suppose we have three individuals in the sample so that we estimate the following regression:

$$Y_{it} = \delta + \beta_2 B_2 + \beta_3 B_3 + \alpha D_{it} + \varepsilon_{it},$$

- $B_2 = 1$ indicates this sample is individual 2, $B_2 = 0$ otherwise
- $B_3 = 1$ indicates this sample is individual 3, $B_3 = 0$ otherwise
- D_{it} is a continuous treatment variable (e.g. schooling years)

Regression with 'N-1 dummy variables'

- For individual 2, the regression can be:

$$\begin{aligned}Y_{2t} &= \delta + \beta_2 B_2 + \alpha D_{2t} + \varepsilon_{2t} \\ &= (\delta + \beta_2 B_2) + \alpha D_{2t} + \varepsilon_{2t}\end{aligned}$$

or

$$Y_{2t} = \lambda_2 + \alpha D_{2t} + \varepsilon_{2t}$$

- where $\lambda_2 = \delta + \beta_2 B_2$

Regression with 'N-1 dummy variables'

- For individual 3, the regression can be:

$$\begin{aligned}Y_{3t} &= \delta + \beta_3 B_3 + \alpha D_{3t} + \varepsilon_{3t} \\ &= (\delta + \beta_3 B_3) + \alpha D_{3t} + \varepsilon_{3t}\end{aligned}$$

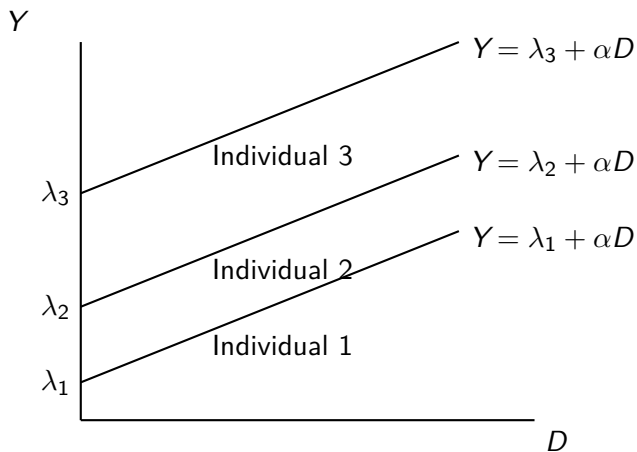
or

$$Y_{3t} = \lambda_3 + \alpha D_{3t} + \varepsilon_{3t}$$

- where $\lambda_3 = \delta + \beta_3 B_3$

Regression with 'N-1 dummy variables'

Graphical Representation



Regression with ‘N-1 dummy variables’

- Since regression with ‘N-1 dummy variables’ and regression with fixed effects are the same
- Thus, when there are not many groups (e.g. state, year, county), we usually control these fixed effects by simply including many dummy variables

Fixed Effects Regression

General Form

- We can include many types of fixed effects to control for all possible time-invariant confounding factors or common time factors

$$Y_{ist} = \lambda_i + \theta_t + \kappa_s + \alpha D_{ist} + X'_{ist}\beta + \varepsilon_{ist},$$

- λ_i is called a “individual fixed effect” or “individual effect”
 - It is the constant (fixed) effect of being in individual i
 - Example: ability or preference
- κ_s is called a “state fixed effect” or “state effect”
 - It is the constant (fixed) effect of being in state s
 - Example: culture or geographical features
- θ_t is called a “year fixed effect” or “year effect”
 - It is the constant (fixed) effect of being in year t
 - Example: business cycle or general time trend

Fixed Effects Regression: STATA Example

Fixed Effects Regression: STATA Example

Data and Code

- See **fixed_effects.do**
- Use `cps_2014_16.dta`

STATA Command: regress

Example:

```
1 regress incwage college i.statefip i.year, vce(robust  
   )
```

- You can simply use **regress** by including several sets of dummy variables to get fixed effects estimation

STATA Command: areg

Syntax:

```
1 areg depvar [indepvars] [if] [in] [weight], absorb(  
    varname) [options]
```

Example:

```
1 areg incwage college i.year, absorb(statefip) vce(  
    robust)
```

- **areg**: Implement regressions with one level of fixed effects
- **absorb(varname)**: Specifies the categorical variable, which is to be included in the regression as if it were specified by dummy variables
- Note that **areg** can only include one fixed effect using **absorb(varname)**
- For other types of fixed effects, you need to include dummy variables by yourself

STATA Command: reghdfe

- To include many levels of fixed effects, we can use this new command **reghdfe**

```
1  ssc install reghdfe
```

- For more details, please visit this website:
<http://scorreia.com/software/reghdfe/index.html>

STATA Command: reghdfe

Syntax:

```
1 reghdfe depvar [indepvars] [if] [in] [weight] ,  
   absorb(absvars) [options]
```

Example:

```
1 reghdfe incwage college, absorb(statefip year) vce(  
   robust)
```

- **reghdfe**: Implement regressions with many levels of fixed effects
- **absorb(varname)**: Specifies the categorical variable, which is to be included in the regression as if it were specified by dummy variables
- Note that **reghdfe** can include many level of fixed effects using **absorb(varname)**

Fixed Effects Regression: R Example

Fixed Effects Regression: R Example

Data and Code

- See **fixed_effects.R**
- Use `cps_2014_16.dta`

R Command: `lm()` for Fixed Effects

Example:

```
1 library(dplyr)
2
3 # Load data
4 data <- read_dta(paste0(rawdata, "/cps_2014_16.dta"))
5
6
7 # Fit the model with state and year fixed effects
  using dummy variables
8 model <- lm(incwage ~ college + factor(statefip) +
  factor(year), data = data)
9 summary(model)
```

- You can use `lm()` with dummy variables to estimate fixed effects
- Here, `factor()` is used to include categorical variables as fixed effects in the model

Create Sample for Analysis

Generating a Dummy Variable and Handling Missing Values:

```
1  # Generate a dummy variable indicating those with  
   bachelor's degree or above  
2  data <- data %>%  
3  mutate(college = ifelse(educ99 >= 15, 1, 0))  
4  
5  # Replace incwage value of 9999999 with NA to  
   indicate missing data  
6  data <- data %>%  
7  mutate(incwage = ifelse(incwage == 9999999, NA,  
   incwage))
```

- `mutate(college = ifelse(educ99 >= 15, 1, 0))`:
Creates a new variable called college that takes the value 1 if the education level is 15 or higher, otherwise 0.
- `mutate(incwage = ifelse(incwage == 9999999, NA, incwage))`: Replaces the value 9999999 in incwage with NA to indicate missing data.

R Command: plm() for Fixed Effects

Example:

```
1 library(plm)
2
3 # Convert to panel data
4 pdata <- pdata.frame(data, index = c("statefip", "
   year"))
5
6 # Fixed effects model using plm
7 model_fe <- plm(incwage ~ college, data = pdata,
   model = "within")
8 summary(model_fe)
```

- `plm()`: Implements regressions with one level of fixed effects (e.g., individual or time)
- `model = "within"`: Specifies that fixed effects are to be used
- `pdata.frame()`: Converts the data to a panel data frame, which is required by `plm`

R Command: lfe for High-Dimensional Fixed Effects

- To include multiple levels of fixed effects, we can use the `lfe` package

```
1 install.packages("lfe")
2 library(lfe)
3
4 # Fixed effects regression with multiple levels
   of fixed effects
5 model_hdfe <- felm(incwage ~ college | statefip +
   year, data = data)
6 summary(model_hdfe)
```

- `felm()`: Implements regressions with multiple levels of fixed effects
- Fixed effects are specified after the `|` symbol, allowing inclusion of many categorical variables