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

March 26, 2026

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?

Main Idea

Fixed Effects Regression

- If unobserved confounding factors are **time-invariant**
  - With panel data (multiple observations per unit over time), we can control for *all* time-invariant unobservables —even ones we cannot measure
  - The key idea: assign each individual (firm, state, ...) their own intercept λ_i that absorbs every time-invariant characteristic of unit i
- **Why "fixed effects"?**
  - λ_i is treated as a **fixed (non-random) parameter** to be estimated — one for every unit i
  - This contrasts with **random effects**, where $\lambda_i \sim N(0, \sigma^2_\lambda)$ is assumed to be a random draw *uncorrelated* with D_{it}
  - Because FE makes no distributional assumption on λ_i, it is consistent even when λ_i is correlated with the treatment

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:

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

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

- Key assumption: $E[\varepsilon_{it} \mid A_i, D_{it}, X_{it}] = 0$
  - After controlling for A_i and X_{it}, the treatment D_{it} is as good as randomly assigned

Fixed Effects Regression

Example

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

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

- 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

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

- 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):

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

- λ_i drops out because it is time-invariant ($\bar{\lambda}_i = \lambda_i$)
- **Intuition**: identification comes entirely from **within-unit variation** — how a worker's union status and earnings change *over time*, not from comparing different workers cross-sectionally
  - This is why the demeaned estimator is also called the **within estimator**
  - Cross-sectional differences between individuals (including A_i) are completely removed

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:

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

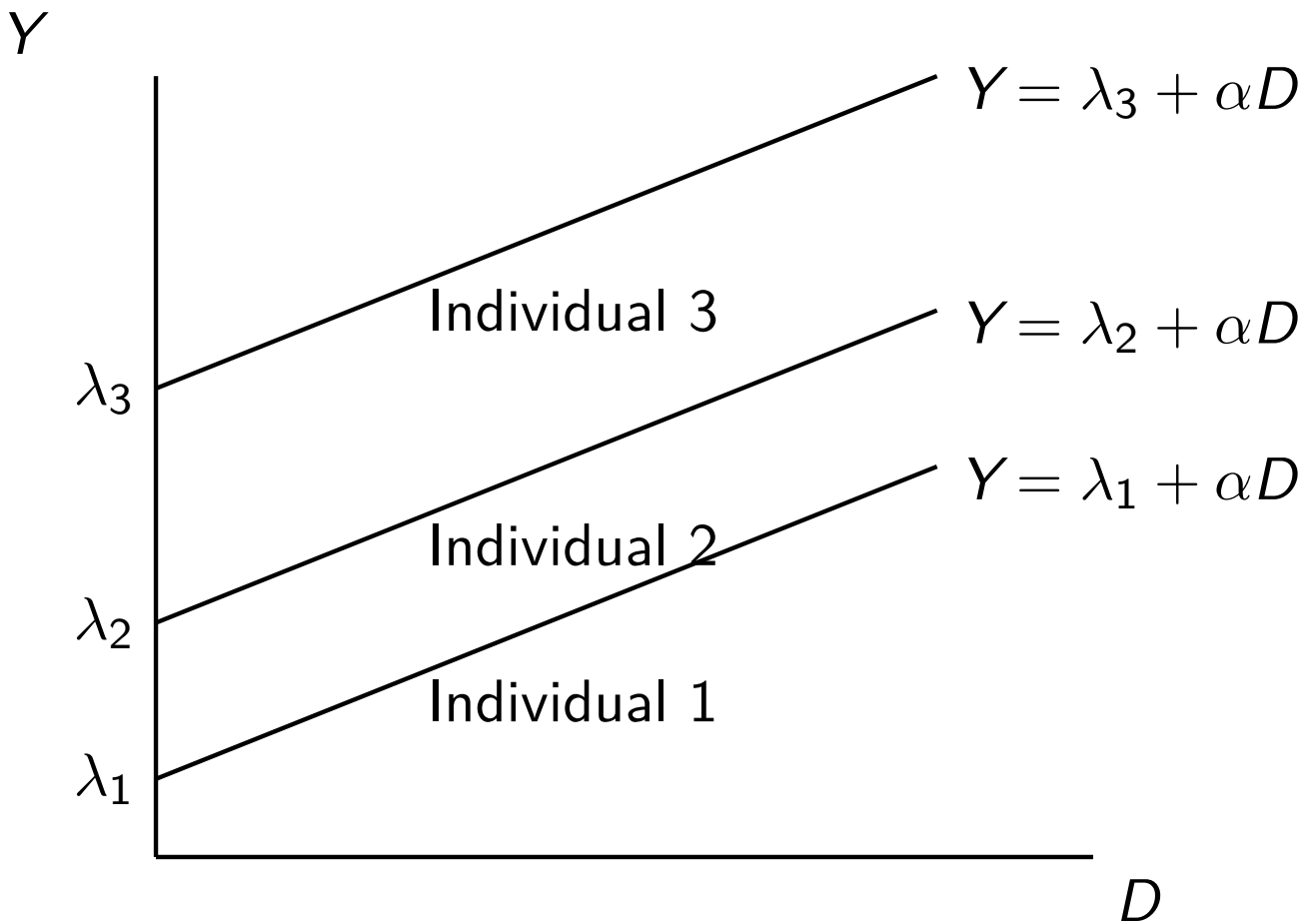
  or

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

- 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
- When there are not many groups (e.g. state, year), we usually control fixed effects by simply including 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

STATA Example

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:

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

Example:

```
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:

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

Example:

```
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)**

STATA Command: outreg2

Display your results

Syntax:

```
outreg2 using filename, [options]
```

Example:

```
reghdfe incwage college age age_sq i.sex i.race, absorb(
    statefip year) vce(robust)
outreg2 using results.csv, replace keep(college) ///
stats(coef se) addstat(Sample Size, e(N)) ///
addtext(Age, Yes, Sex, Yes, Race, Yes, State FE, Yes, Year
    FE, Yes)
```

- **outreg2**: Outputs regression results to a file (e.g., .csv, .tex)
- **using results.csv**: Specifies the output file name
- **keep(college)**: Includes only the coefficient for the variable `college`
- **stats(coef se)**: Displays coefficients and standard errors

STATA Command: outreg2

Display your results

Syntax:

```
outreg2 using filename, [options]
```

Example:

```
reghdfe incwage college age age_sq i.sex i.race, absorb(
    statefip year) vce(robust)
outreg2 using results.csv, replace keep(college) ///
stats(coef se) addstat(Sample Size, e(N)) ///
addtext(Age, Yes, Sex, Yes, Race, Yes, State FE, Yes, Year
    FE, Yes)
```

- **addstat()**: Adds custom statistics, such as sample size
- **addtext()**: Appends a row with information on controls and fixed effects

R Example

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

R Command: plm() for Fixed Effects

Example:

```
library(plm)

# Convert to panel data
pdata <- pdata.frame(data, index = c("statefip", "year"))

# Fixed effects model using plm
model_fe <- plm(incwage ~ college, data = pdata, model = "
    within")
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: fixest for High-Dimensional Fixed Effects

```r
library(fixest)

# Fixed effects regression with multiple controls and fixed
    effects
model_hdfe <- feols(incwage ~ college + age + age_sq +
factor(sex) + factor(race) |
statefip + year,
data = data, vcov = "hetero")

summary(model_hdfe)
```

- **feols()**: Performs linear regression with support for high-dimensional fixed effects
  - Fixed effects are specified after the | symbol, e.g., `statefip + year`
  - **vcov = "hetero"**: Computes heteroskedasticity-robust standard errors

R Command: modelsummary for Regression Output

```r
modelsummary(
models,
coef_omit = "^(?!college)",          # Keep only 'college'
statistic = "std.error",             # Show robust SEs
gof_map = c("nobs" = "Sample Size"),  # Rename N
notes = notes,                       # Add notes per model
output = file.path(workdata, "results.csv")  # Save output
)
```