Module 4: Difference-in-Differences and Effects of Medicaid Expansion

Part 2: Basics of Fixed Effects and Panel Data

Ian McCarthy | Emory University
Econ 470 & HLTH 470

Understanding Panel Data

Nature of the Data

- Repeated observations of the same units over time (balanced vs unbalanced)
- Identification due to variation **within unit**

Notation

- Unit $i = 1, \ldots, N$ over several periods $t = 1, \ldots, T$, which we denote y_{it}
- Treatment status D_{it}
- Regression model,
  $y_{it} = \delta D_{it} + \gamma_i + \gamma_t + \epsilon_{it}$ for $t = 1, \ldots, T$ and $i = 1, \ldots, N$

Benefits of Panel Data

- *May* overcome certain forms of omitted variable bias
- Allows for unobserved but time-invariant factor, γ_i, that affects both treatment and outcomes

Still assumes

- No time-varying confounders
- Past outcomes do not directly affect current outcomes
- Past outcomes do not affect treatment (reverse causality)

Some textbook settings

- Unobserved "ability" when studying schooling and wages
- Unobserved "quality" when studying physicians or hospitals

Panel Data and Regression

Fixed effects and regression

$y_{it} = \delta D_{it} + \gamma_i + \gamma_t + \epsilon_{it}$ for $t = 1, \ldots, T$ and $i = 1, \ldots, N$

- Allows correlation between γ_i and D_{it}
- Physically estimate γ_i in some cases via set of dummy variables
- More generally, "remove" γ_i via:
  - "within" estimator
  - first-difference estimator

Within Estimator

$y_{it} = \delta D_{it} + \gamma_i + \gamma_t + \epsilon_{it}$ for $t = 1, \ldots, T$ and $i = 1, \ldots, N$

- Most common approach (default in most statistical software)
- Equivalent to demeaned model,

$$y_{it} - \bar{y}_i = \delta(D_{it} - \bar{D}_i) + (\gamma_i - \bar{\gamma}_i) + (\gamma_t - \bar{\gamma}_t) + (\epsilon_{it} - \bar{\epsilon}_i)$$

- $\gamma_i - \bar{\gamma}_i = 0$ since γ_i is time-invariant
- Requires *strict exogeneity* assumption (error is uncorrelated with D_{it} for all time periods)

First-difference

$y_{it} = \delta D_{it} + \gamma_i + \gamma_t + \epsilon_{it}$ for $t = 1, \ldots, T$ and $i = 1, \ldots, N$

- Instead of subtracting the mean, subtract the prior period values
  $y_{it} - y_{i,t-1} = \delta(D_{it} - D_{i,t-1}) + (\gamma_i - \gamma_i) + (\gamma_t - \gamma_{t-1}) + (\epsilon_{it} - \epsilon_{i,t-1})$
- Requires exogeneity of ϵ_{it} and D_{it} only for time t and $t - 1$ (weaker assumption than within estimator)
- Sometimes useful to estimate both FE and FD just as a check

Keep in mind...

- Discussion only applies to linear case or very specific nonlinear models
- Fixed effects at lower "levels" accommodate fixed effects at higher levels (e.g., FEs for hospital combine to form FEs for zip code, etc.)
- Fixed effects can't solve reverse causality
- Fixed effects don't address unobserved, time-varying confounders
- Can't estimate effects on time-invariant variables
- May "absorb" a lot of the variation for variables that don't change much over time

Panel Data and Fixed Effects IRL

Within Estimator (Default) in practice

Stata

```
ssc install causaldata
causaldata gapminder.dta, use clear download
gen lgdp_pc=log(gdppercap)
tsset country year
xtreg lifeExp lgdp_pc, fe
```

R

```
library(fixest)
library(causaldata)
reg.dat <- causaldata::gapminder %>%
  mutate(lgdp_pc=log(gdpPercap))
feols(lifeExp~lgdp_pc | country, data=reg.dat)
```

Within Estimator (Default) in practice

	Default FE
Log GDP per Capita	9.769
	(0.702)

Within Estimator (Manually Demean) in practice

Stata

```
causaldata gapminder.dta, use clear download
gen lgdp_pc=log(gdppercap)
foreach x of varlist lifeExp lgdp_pc {
  egen mean_`x'=mean(`x')
  egen demean_`x'=`x'-mean_`x'
}
reg demean_lifeExp demean_lgdp_pc
```

R

```
library(causaldata)
reg.dat <- causaldata::gapminder %>%
  mutate(lgdp_pc=log(gdpPercap)) %>%
  group_by(country) %>%
  mutate(demean_lifeexp=lifeExp - mean(lifeExp, na.rm=TR
         demean_gdp=lgdp_pc - mean(lgdp_pc, na.rm=TRUE))
lm(demean_lifeexp~ 0 + demean_gdp, data=reg.dat)
```

Within Estimator (Manually Demean) in practice

	Default FE	Manual FE
Log GDP per Capita	9.769	9.769
	(0.702)	(0.701)

Note: `feols` defaults to clustering at level of FE, `lm` requires our input

First differencing (default) in practice

Stata

```
causaldata gapminder.dta, use clear download
gen lgdp_pc=log(gdppercap)
reg d.lifeExp d.lgdp_pc, noconstant
```

R

```
library(plm)
reg.dat <- causaldata::gapminder %>%
  mutate(lgdp_pc=log(gdpPercap))

plm(lifeExp ~ 0 + lgdp_pc, model="fd", individual="count
```

First differencing (manual) in practice

	Default FE	Manual FE	Default FD
Log GDP per Capita	9.769	9.769	5.290
	(0.702)	(0.284)	(0.291)

First differencing (manual) in practice

Stata

```
causaldata gapminder.dta, use clear download
gen lgdp_pc=log(gdppercap)
reg d.lifeExp d.lgdp_pc, noconstant
```

R

```
reg.dat <- causaldata::gapminder %>%
  mutate(lgdp_pc=log(gdpPercap)) %>%
  group_by(country) %>%
  arrange(country, year) %>%
  mutate(fd_lifeexp=lifeExp - lag(lifeExp),
         lgdp_pc=lgdp_pc - lag(lgdp_pc)) %>%
  na.omit()

lm(fd_lifeexp~ 0 + lgdp_pc , data=reg.dat)
```

First differencing (manual) in practice

	Default FE	Manual FE	Default FD	Manual FD
Log GDP per Capita	9.769	9.769	5.290	5.290
	(0.702)	(0.284)	(0.291)	(0.291)

FE and FD with same time period

	Default FE	Default FD	Manual FD
Log GDP per Capita	8.929	5.290	5.290
	(0.741)	(0.291)	(0.291)

Don't want to read too much into this, but...

- Likely strong serial correlation in this case (almost certainly)
- Mispecified model