Basics of Panel Data

Ian McCarthy | Emory University
Workshop on Causal Inference with Panel Data

Table of contents

What are Panel Data?

Nature of the Data

- Repeated observations of the same units over time

Notation

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

Estimating Regressions with Panel Data

Regression model

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

Fixed Effects

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

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

Within Estimator

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

- Most common approach (default in most statistical software)
- Equivalent to demeaned model,
  $y_{it} - \bar{y}_i = \delta(D_{it} - \bar{D}_i) + (u_i - \bar{u}_i) + (\epsilon_{it} - \bar{\epsilon}_i)$
- $u_i - \bar{u}_i = 0$ since u_i is time-invariant
- Requires *strict exogeneity* assumption (error is uncorrelated with D_{it} for all time periods)

First-difference

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

- Instead of subtracting the mean, subtract the prior period values
  $y_{it} - y_{i,t-1} = \delta(D_{it} - D_{i,t-1}) + (u_i - u_i) + (\epsilon_{it} - \epsilon_{i,t-1})$
- 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 can't solve reverse causality
- Fixed effects doesn'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

Seeing things in action

Within Estimator (Default)

Stata

```
ssc install bcuse
bcuse wagepan
tsset nr year
xtreg lwage exper expersq, fe
```

R

```
library(readstata13)
library(fixest)
wagepan ← read.dta13("http://fmwww.bc.edu/ec-p/data/woo
feols(lwage~exper + expersq | nr, data=wagepan)
```

Within Estimator (Manually Demean)

Stata

```
ssc install bcuse
bcuse wagepan
foreach x of varlist lwage exper expersq {
  egen mean_`x'=mean(`x')
  egen demean_`x'=`x'-mean_`x'
}
reg demean_lwage demean_exper demean_expersq
```

R

```
library(readstata13)
wagepan ← read.dta13("http://fmwww.bc.edu/ec-p/data/woo
wagepan ← wagepan %>%
  group_by(nr) %>%
  mutate(demean_lwage=lwage - mean(lwage),
         demean_exper=exper - mean(exper),
         demean_expersq=expersq - mean(expersq))
summary(lm(demean_lwage~demean_exper + demean_expersq, d
```

First differencing

Stata

```
ssc install bcuse
bcuse wagepan
reg d.lwage d.exper d.expersq, noconstant
```

R

```
library(readstata13)
wagepan ← read.dta13("http://fmwww.bc.edu/ec-p/data/woo
wagepan ← wagepan %>%
  group_by(nr) %>%
  arrange(year) %>%
  mutate(fd_lwage=lwage - lag(lwage),
         fd_exper=exper - lag(exper),
         fd_expersq=expersq - lag(expersq)) %>%
  na.omit()
summary(lm(fd_lwage~0 + fd_exper + fd_expersq, data=wage
```