Difference-in-differences

Tyler Ransom

Univ of Oklahoma

Apr 16, 2019

Today's plan

1. Difference-in-differences
   1.1 Refresher on Natural Experiments
   1.2 Policy Analysis with Pooled Cross Sections
   1.3 Policy Analysis with Panel Data
2. In-class activity: Practice with panel data in R

Refresher: Natural Experiments

Refresher: Natural experiments

- A **natural experiment** is a setting in which observational data is collected
- Treatment is either randomly assigned, or “as if” randomly assigned
- Treatment is **not** assigned by the researcher
- (otherwise it would be a randomized experiment)
- Natural experiments can be a boon to resolving $E(u|\mathbf{x}) \neq 0$
- NE’s always give rise to a control group and a treatment group

Natural experiments, IVs, and panel data

- We already talked about natural experiments and instrumental variables
- i.e. NE's can provide valid instruments of our endogenous x
- How valid the NE is will determine how valid the instrument is
- We can also make use of NEs when we have panel data
- Can correct for systematic differences between control, treatment groups

Policy Analysis with Pooled Cross Sections

Difference-in-differences

- Suppose we have a NE that gives a treatment group and a control group
- In our data, can create two dummies:
  1. dT equals 1 if treatment group, 0 otherwise
  2. $d2$ equals 1 if after policy, 0 if before
- We want to see how treatment affected y, so our equation is

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1 d2 \cdot dT + u$$

Difference-in-differences

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1 d2 \cdot dT + u$$

- dT corrects for systematic differences bet. treatment and control
- $d2$ corrects for systematic changes in the entire economy
- $d2 \cdot dT$ gives the **difference in differences**
- Can show that

$$\delta_1 = \left(\bar{y}_{2,T} - \bar{y}_{2,C}\right) - \left(\bar{y}_{1,T} - \bar{y}_{1,C}\right)$$

Interpretation of Difference-in-differences

- i.e. δ_1 measures change in y after policy in treatment and control
- holding fixed aggregate trends and persistent differences bet. two groups
- δ_1 sometimes called **average treatment effect**
- because it measures the effect of treatment on y (on average)

Parallel Trends Assumption

- Validity of the DiD approach relies on a key assumption:
- y can't change across T and C (bef. & after) for reasons **other than** the policy
- This is known as the **parallel trends assumption**
- If the parallel trends assumption fails, then u is correlated with $d2 \cdot dT$
- We're right back where we started (endogeneity problem)

More complicated models

- Nothing stops us from controlling for other observables

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1 d2 \cdot dT + \delta_2 \mathbf{x} + u$$

- $\mathbf{x}$ is anything that affects y that is unrelated to treatment
- We can also include more time periods (if we have them):

$$y = \beta_0 + \sum_{\tau} \delta_{0,\tau} d\tau + \beta_1 dT + \sum_{\tau} \delta_{1,\tau} d\tau \cdot dT + \delta_2 \mathbf{x} + u$$

- The latter model also known as an **event study**
- Can be used to test parallel trends assumption

Example: House Prices and Environmental Amenities

- Suppose a city decides to build a garbage incinerator
- Want to know if being close to incinerator causes property devaluation
- Data: `kielmc` in `wooldridge` package
- Treatment: *nearinc* (dummy for living near incinerator)
- Pre/Post: $y81$ (dummy for when incinerator was announced)

$$rprice = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 y81 \cdot nearinc + \delta_2 \mathbf{x} + u$$

Example: House Prices and Environmental Amenities

- We could look at difference in prices in 1981:

```
> df <- as_tibble(kielmc)
> df %>% filter(y81==1) %>% group_by(nearinc) %>%
  summarize(avg.price=mean(rprice))

  nearinc avg.price
        0   101308.
        1    70619.
```

- But this doesn't tell us causality

- Perhaps the city sited the incinerator in a "bad" neighborhood?

Example: House Prices and Environmental Amenities

- To see if the site was already in a bad neighborhood, look at pre-policy:

```
> df %>% filter(y81==0) %>% group_by(nearinc) %>%
  summarize(avg.price=mean(rprice))

  nearinc avg.price
        0    82517.
        1    63693.
```

- So, yes, even before incinerator was announced, property values were lower

Example: House Prices and Environmental Amenities

- Now let's estimate the DiD:

```
> df %<>% mutate(nearinc=as.factor(nearinc),
  y81=as.factor(y81))
> est.did <- lm(rprice ~ y81*nearinc, data=df)
```

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      82,517      2,727  30.260  < 2e-16 ***
y811             18,790      4,050   4.640 5.12e-06 ***
nearinc1        -18,824      4,875  -3.861 0.000137 ***
y811:nearinc1   -11,864      7,457  -1.591 0.112595
```

- ProTip: putting "*" between two factors adds their levels and interactions

Example: House Prices and Environmental Amenities

- Let's also include some other *x*'s (e.g. house characteristics)

- We might also want to use log price instead of price

```
> est.did.x <- lm(rprice ~ y81*nearinc + age +
                  I(age^2) + intst + land + area +
                  rooms + baths, data=df)

> est.did.xlp <- lm(lprice ~ y81*nearinc + age +
                    I(age^2) + intst + land + area +
                    rooms + baths, data=df)
```

Example: House Prices and Environmental Amenities

```
> stargazer(est.did,est.did.x,est.did.lp,est.did.xlp, keep.stat=c("N","rsq"), type="text")

=================================================================
                              Dependent variable:
              ---------------------------------------------------
                        rprice                     lprice
                   (1)             (2)          (3)        (4)
-----------------------------------------------------------------
y81           18,790.290***  13,928.480***  0.457***   0.403***
               (4,050.065)    (2,798.747)   (0.045)    (0.029)
nearinc       -18,824.370***   3,780.337    -0.340***   -0.035
               (4,875.322)    (4,453.415)   (0.055)    (0.046)
y81:nearinc    -11,863.900   -14,177.930*** -0.063     -0.093*
               (7,456.646)    (4,987.267)   (0.083)    (0.052)
Constant      82,517.230***    13,807.670   11.285***  10.371***
               (2,726.910)   (11,166.590)   (0.031)    (0.117)
-----------------------------------------------------------------
Observations       321            321          321        321
R2                0.174          0.660        0.409      0.789
=================================================================
Note:                                 *p<0.1; **p<0.05; ***p<0.01
```

Policy Analysis with Panel Data

Panel Data

- All of the previous discussion was about pooled cross-sectional data
- DiD can also be applied to panel data
- In this case, each unit can serve as its own control
- The regression equation is slightly altered:

$$y_{it} = \beta_0 + \delta_0 d2_t + \delta_1 d_{it} + a_i + u_{it}$$

- a_i takes the place of dT, d_{it} takes the place of the interaction
- This is more convincing, since we can control for unit-level unobservables

Example: Traffic laws & fatalities

- Let's examine the passage of two types of laws on traffic fatalities:
- open container laws and administrative per se laws (unit is US state)

```
> df.wide <- as_tibble(traffic1) %>%
             select(-starts_with("c"))

# Very similar code here as in original panel data slides
> df.long <- df.wide %>%
      gather(variable,value,-state) %>%
      mutate(year = parse_number(variable)) %>%
      mutate(variable =  gsub("\\d","",x = variable)) %>%
      spread(variable,value)
```

Example: Traffic laws & fatalities

- Now we estimate the model:

$$dthrte_{it} = \beta_0 + \delta_0 y90_t + \delta_1 open_{it} + \delta_2 admn_{it} + a_i + u_{it}$$

```
# First, create the y90 dummy
> df.long %<>% mutate(year = as.factor(year), y90 = as.nume

# Now estimate the fixed effects model
> est.did.fe <- plm(dthrte ~ year + open + admn,
                    index=c("state","year"),
                    model="within", data=df.long)

# Clustered SE's
> clustSE <- coef_test(est.did.fe, vcov = "CR2",
                 cluster = "individual", test="naive-t")
```

Example: Traffic laws & fatalities

```
> stargazer(est.did.fe, se=list(clustSE$SE), type="text")

========================================
                        dthrte
----------------------------------------
d2                     -0.497***
                        (0.045)
open                   -0.420**
                        (0.197)
admn                    -0.151
                        (0.155)
----------------------------------------
Observations              102
R2                       0.738
Adjusted R2              0.448
F Statistic     44.964*** (df = 3; 48)
```