Lab #19 - Differences-in-Differences

Econ 224

November 13th, 2018

Exercise II: Applied

The following questions are based on a dataset called `minwage.dta` that you can download from the Mastering 'Metrics website: click on "Instructor's Corner," then scroll down to the bottom of the page. This dataset contains information collected from fast food restaurants in New Jersey and eastern Pennsylvania during two interview waves, the first in March of 1992 and the second in November-December of the same year. Between these two interview waves – on April 1st to be precise – the New Jersey minimum wage increased by just under 19%, from \$4.25 to \$5.05 per hour. The minimum wage in Pennsylvania was unchanged during this period: \$4.25 per hour. The `minwage.dta` dataset is drawn from a famous but controversial study of the effects of minimum wages by Angrist & Kreuger. The study is so famous that there is even an oblique reference to it on the label of a certain brand of shampoo! (Sadly they do not provide the full citation.) Here is a description of the variables that you will need to carry out this exercise. When you see a pair of variables in the table below, e.g. `fte`/`fte2`, both measure the same thing but the one with the `2` is based on the *second* survey wave, while the one without the `2` is based on the *first* survey wave.

Name	Description
`state`	Dummy variable =1 for NJ, =0 for PA
`wage_st`/`wage_st2`	Starting wage in dollars/hour at the restaurant
`fte`/`fte2`	Full-time equiv. employment = #(Full time employees) + #(Part-time Employees)/2. Excludes managers.
`chain`	Categorical variable taking values in $\{1, 2, 3, 4\}$ to indicate the four chains in the dataset: Burger King, KFC, Roy Rogers, and Wendy's
`co_owned`	Dummy variable =1 if restaurant is company-owned, =0 if franchised
`sample`	Dummy variable =1 if wage and employment data are available for both survey waves at this restaurant

1. Preliminaries:
   (a) Download the data and load it in R using an appropriate package.
   (b) Restrict the sample to only those restaurants with `sample` equal to 1 to ensure that we are making an apples-to-apples comparison throughout the remainder of this lab.
   (c) Rename the column `state` to `treat`.
   (d) Create a *new* column called `state` that equals `PA` if `treat` is `0` and `NJ` if `treat` is `1`.
2. Baseline Diff-in-Diff Estimate: starting wages
   (a) Calculate the average wage in each survey wave separately for each state.
   (b) Calculate the within-state time-differences based on (a).
   (c) Calculate the between-state difference-in-differences based on (c).
   (d) Interpret your findings from (c). What do they tell us about the causal effect of increasing the minimum wage? What assumptions are required for this interpretation to be valid?
3. Baseline Diff-in-Diff Estimate: full time equivalent employment
   (a) Repeat question 2 but using full-time equivalent employment as the outcome variable rather than starting wages.

4. Reshape dataset for Diff-in-Diff Regression Estimation:
   (a) Create a tibble called `wave1` containing only the columns `state`, `treat`, `wage_st`, `fte`, `chain`, and `co_owned`. Add a column walled `post` to `wave1` that equals `0` for every observation.
   (b) Create a tibble called `wave2` containing only the columns `state`, `treat`, `wage_st2`, `fte2`, `chain`, and `co_owned`. Rename `wage_st2` to `wage_st` and `fte2` to `fte`. Then add a column walled `post` to `wave2` that equals `1` for every observation.
   (c) Create a tibble called `both_waves` by *stacking* `wave1` on top of `wave2`. You can do this using the `bind_rows` command from `dplyr`. (Read the help file for more details.)
5. Diff-in-Diff Regression Estimates:

Solutions

```
# 1 - Preliminaries
library(tidyverse)
library(haven)
ak <- read_dta('~/econ224/labs/minwage.dta')
ak <- ak %>% filter(sample == 1) %>%
  rename(treat = state) %>%
  mutate(state = case_when(treat == 0 ~ 'PA',
                           treat == 1 ~ 'NJ'))

# 2 - Baseline Diff-in-Diff: starting wages
DinD_wage <- ak %>% group_by(state) %>%
  summarize(mean_wage_st = mean(wage_st),
            mean_wage_st2 = mean(wage_st2)) %>%
  mutate(diff = mean_wage_st2 - mean_wage_st)
DinD_wage
```

```
# A tibble: 2 x 4
  state mean_wage_st mean_wage_st2    diff
  <chr>        <dbl>         <dbl>   <dbl>
1 NJ            4.61          5.08  0.469
2 PA            4.65          4.62 -0.0348
```

```
with(DinD_wage, diff[1] - diff[2])
```

```
[1] 0.5040066
```

```
# 3 - Baseline Diff-in-Diff: full-time equivalent employment
DinD_emp <- ak %>% group_by(state) %>%
  summarize(mean_fte = mean(fte),
            mean_fte2 = mean(fte2)) %>%
  mutate(diff = mean_fte2 - mean_fte)
DinD_emp
```

```
# A tibble: 2 x 4
  state mean_fte mean_fte2   diff
  <chr>    <dbl>     <dbl>  <dbl>
1 NJ        17.3      17.6  0.287
2 PA        20.1      18.1 -2.02
```

```r
with(DinD_emp, diff[1] - diff[2])
```

```
[1] 2.301994
```

```r
# Reshape dataset for Diff-in-Diff regression estimation
wave1 <- ak %>% select(state, treat, wage_st, fte, chain, co_owned) %>%
  mutate(post = 0)
wave2 <- ak %>% select(state, treat, wage_st2, fte2, chain, co_owned) %>%
  mutate(post = 1) %>%
  rename(wage_st = wage_st2, fte = fte2)
both_waves <- bind_rows(wave1, wave2)
both_waves
```

```
# A tibble: 702 x 7
   state treat wage_st   fte chain co_owned  post
   <chr> <dbl>   <dbl> <dbl> <dbl>    <dbl> <dbl>
 1 PA        0    5     30       4        1     0
 2 PA        0    5.5   19       4        1     0
 3 PA        0    5     67.5     1        0     0
 4 PA        0    5     18.5     1        0     0
 5 PA        0    5.25   6       2        1     0
 6 PA        0    5      7       2        1     0
 7 PA        0    5     12.5     3        1     0
 8 PA        0    5     55       1        0     0
 9 PA        0    5     21.5     1        0     0
10 PA        0    5.5   25.5     1        0     0
# ... with 692 more rows
```

```r
# Diff-in-Diff regression results: starting wages
# Should I add clustered standard errors?
reg_wage1 <- lm(wage_st ~ treat + post + treat:post, both_waves)
summary(reg_wage1)
```

```
Call:
lm(formula = wage_st ~ treat + post + treat:post, data = both_waves)

Residuals:
     Min       1Q   Median       3Q      Max
-0.40364 -0.11298 -0.03214  0.13702  1.63121

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.65364    0.03413 136.343   <2e-16 ***
treat       -0.04065    0.03788  -1.073    0.284
post        -0.03485    0.04827  -0.722    0.471
treat:post   0.50401    0.05357   9.409   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2773 on 698 degrees of freedom
Multiple R-squared:  0.4028,    Adjusted R-squared:  0.4002
F-statistic: 156.9 on 3 and 698 DF,  p-value: < 2.2e-16
```

```r
# Diff-in-Diff regression results: full-time equivalent employment
# Should I add clustered standard errors?
reg_emp1 <- lm(fte ~ treat + post + treat:post, both_waves)
summary(reg_emp1)
```

```
Call:
lm(formula = fte ~ treat + post + treat:post, data = both_waves)

Residuals:
    Min      1Q  Median      3Q     Max
-15.614  -6.275  -1.275   4.438  62.725

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.114      1.104  18.225   <2e-16 ***
treat         -2.838      1.225  -2.317   0.0208 *
post          -2.015      1.561  -1.291   0.1971
treat:post     2.302      1.732   1.329   0.1843
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.966 on 698 degrees of freedom
Multiple R-squared:  0.007971,  Adjusted R-squared:  0.003707
F-statistic: 1.869 on 3 and 698 DF,  p-value: 0.1334
```