Lab #19 - Minimum Wages and Unemployment

Econ 224

November 13th, 2018

Introduction

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

Exercises

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`.
   (e) Create a column called `low_wage` that takes the value `1` if `wage_st` is less than `5`.
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 `minwage` for Diff-in-Diff Regression Estimation:
    (a) Create a tibble called `wave1` containing only the columns `state`, `treat`, `wage_st`, `fte`, `chain`, `co_owned`, and `low_wage`. 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`, `co_owned`, and `low_wage`. 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:
    (a) Consider the following regression model using the variables `treat` and `post` constructed above:

$$Y_{i,s,t} = \beta_0 + \beta_1(\texttt{treat}_{i,s}) + \beta_2(\texttt{post}_t) + \beta_3(\texttt{treat}_{i,s} \times \texttt{post}_t) + \epsilon_{i,s,t}$$

    where i indexes *restaurants*, s indexes *states*, and t indexes *time periods*, i.e. the two survey waves. Explain the meaning of each of the four regression coefficients. Which one gives the Regression differences-in-differences effect?
    (b) Estimate the regression from part (a) based on `both_waves` using `wage_st` as the outcome variable. Summarize your results, including appropriate statistical inference. How do they compare to those that you calculated in question 2 above?
    (c) Estimate the regression from part (a) based on `both_waves` using `fte` as the outcome variable. Summarize your results, including appropriate statistical inference. How do they compare to those that you calculated in question 3 above?
    (d) An advantage of the regression-based formulation of differences-in-differences is that it allows us to control for other variables that might affect wages and employment. Repeat parts (b) and (c) adding `co_owned` and dummy variables for each of the four restaurant chains to your regression. Hint: rather than creating separate dummy variables from each of the values that `chain` can take, use `as.factor()` to convert `chain` to a factor. Then if you include `chain` in a regression, R will automatically create the dummy variables for you.
    (e) How do your results from part (d) compare with those of parts (b) and (c)?
6. Probing the Diff-in-Diff Assumption:
    (a) What assumption is required for the diff-in-diff approach to provide a valid causal estimate of the effects of New Jersey raising its minimum wage?

Solutions

```
# 1 - Preliminaries
library(tidyverse)
library(haven)
minwage <- read_dta('~/econ224/labs/minwage.dta')
minwage <- minwage %>% filter(sample == 1) %>%
  rename(treat = state) %>%
  mutate(state = case_when(treat == 0 ~ 'PA',
                           treat == 1 ~ 'NJ'),
         low_wage = 1 * (wage_st < 5))
```

```r
# 2 - Baseline Diff-in-Diff: starting wages
DinD_wage <- minwage %>% 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
```

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

```
[1] 0.5040066
```

```r
# 3 - Baseline Diff-in-Diff: full-time equivalent employment
DinD_emp <- minwage %>% 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
# 4 - Reshape dataset for Diff-in-Diff regression estimation
wave1 <- minwage %>%
  select(state, treat, wage_st, fte, chain, co_owned, low_wage) %>%
  mutate(post = 0)
wave2 <- minwage %>%
  select(state, treat, wage_st2, fte2, chain, co_owned, low_wage) %>%
  mutate(post = 1) %>%
  rename(wage_st = wage_st2, fte = fte2)
both_waves <- bind_rows(wave1, wave2)
```

Solution to 5(a)

Blah blah blah…

```r
# 5 - Diff-in-Diff regression results
library(stargazer)
reg_wage1 <- lm(wage_st ~ treat + post + treat:post, both_waves)
reg_emp1 <- lm(fte ~ treat + post + treat:post, both_waves)

stargazer(reg_wage1, reg_emp1, type = 'latex', header = FALSE, digits = 2,
          dep.var.labels = c('Starting Wage', 'Full-time Equiv. Employment'),
          omit.stat = c('f', 'ser', 'adj.rsq', 'rsq'))
```

Table 2:

	Dependent variable:	
	Starting Wage	Full-time Equiv. Employment
	(1)	(2)
treat	−0.04	−2.84**
	(0.04)	(1.22)
post	−0.03	−2.02
	(0.05)	(1.56)
treat:post	0.50***	2.30
	(0.05)	(1.73)
Constant	4.65***	20.11***
	(0.03)	(1.10)
Observations	702	702
Note:		*p<0.1; **p<0.05; ***p<0.01

```r
#---------- (d) control for co_owned and chain
both_waves <- both_waves %>% mutate(chain = as.factor(chain))
reg_wage2 <- lm(wage_st ~ treat + post + treat:post + co_owned + chain, both_waves)

reg_emp2 <- lm(fte ~ treat + post + treat:post + co_owned + chain, both_waves)

stargazer(reg_wage2, reg_emp2, type = 'latex', header = FALSE, digits = 2,
          dep.var.labels = c('Starting Wage', 'Full-time Equiv. Employment'),
          omit.stat = c('f', 'ser', 'adj.rsq', 'rsq'))
```

```r
# 6 - More stuff...
nj_only <- both_waves %>% filter(state == 'NJ')
pa_only <- both_waves %>% filter(state == 'PA')
summary(lm(wage_st ~ low_wage + post + low_wage:post, nj_only))
```

```
Call:
lm(formula = wage_st ~ low_wage + post + low_wage:post, data = nj_only)
```

Table 3:

	Dependent variable:	
	Starting Wage	Full-time Equiv. Employment
	(1)	(2)
treat	-0.04	-2.14^{*}
	(0.04)	(1.10)
post	-0.03	-2.02
	(0.05)	(1.40)
co_owned	0.07^{***}	-1.01
	(0.02)	(0.72)
chain2	0.02	-10.16^{***}
	(0.03)	(0.84)
chain3	0.05	-1.35
	(0.03)	(0.86)
chain4	0.12^{***}	-1.37
	(0.03)	(0.97)
treat:post	0.50^{***}	2.30
	(0.05)	(1.55)
Constant	4.59^{***}	22.56^{***}
	(0.04)	(1.05)
Observations	702	702
Note:		*p<0.1; **p<0.05; ***p<0.01

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.21224 -0.06318 -0.02402  0.03776  0.67598 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    5.113182   0.020228 252.779   <2e-16 ***
low_wage      -0.650944   0.023075 -28.209   <2e-16 ***
post          -0.004091   0.028607  -0.143    0.886    
low_wage:post  0.615872   0.032634  18.872   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1643 on 566 degrees of freedom
Multiple R-squared:  0.7759,    Adjusted R-squared:  0.7747 
F-statistic: 653.2 on 3 and 566 DF,  p-value: < 2.2e-16
```

```r
summary(lm(fte ~ low_wage + post + low_wage:post, nj_only))
```

```
Call:
lm(formula = fte ~ low_wage + post + low_wage:post, data = nj_only)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.989  -6.192  -0.759   4.241  63.241 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     18.989      1.062  17.886   <2e-16 ***
low_wage        -2.230      1.211  -1.841   0.0662 .  
post            -2.250      1.501  -1.499   0.1345    
low_wage:post    3.301      1.713   1.928   0.0544 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.625 on 566 degrees of freedom
Multiple R-squared:  0.007592,  Adjusted R-squared:  0.002332 
F-statistic: 1.443 on 3 and 566 DF,  p-value: 0.2292
```

```r
summary(lm(wage_st ~ low_wage + post + low_wage:post, pa_only))
```

```
Call:
lm(formula = wage_st ~ low_wage + post + low_wage:post, data = pa_only)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.55000 -0.18349 -0.06522  0.20000  1.72814 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    5.06522    0.05738  88.279  < 2e-16 ***
```

```
low_wage      -0.63173    0.07109  -8.887 4.94e-15 ***
post          -0.26522    0.08114  -3.268 0.001389 **
low_wage:post  0.35359    0.10053   3.517 0.000604 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2752 on 128 degrees of freedom
Multiple R-squared:  0.4255,    Adjusted R-squared:  0.4121
F-statistic: 31.61 on 3 and 128 DF,  p-value: 2.343e-15
```

```r
summary(lm(fte ~ low_wage + post + low_wage:post, pa_only))
```

```
Call:
lm(formula = fte ~ low_wage + post + low_wage:post, data = pa_only)

Residuals:
    Min      1Q  Median      3Q     Max
-15.302  -7.026  -2.022   4.751  46.804

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    20.6957     2.1515   9.619   <2e-16 ***
low_wage       -0.8933     2.6656  -0.335    0.738
post           -3.8478     3.0427  -1.265    0.208
low_wage:post   2.8129     3.7697   0.746    0.457
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.32 on 128 degrees of freedom
Multiple R-squared:  0.01455,   Adjusted R-squared:  -0.008549
F-statistic: 0.6299 on 3 and 128 DF,  p-value: 0.597
```