Differences in Differences

Dr. Soliman

2024-11-12

Task 1

1. Tabulate the number of stores by `state` and by survey wave (`observation`).

```
table(ck1994$state, ck1994$observation)
```

```
##
##                February 1992 November 1992
##   New Jersey             331           331
##   Pennsylvania            79            79
```

2. Create a full-time equivalent (FTE) employees variable called `empfte` equal to `empft` + 0.5*`emppt` + `nmgrs`. `empft` and `emppt` correspond respectively to the number of full-time and part-time employees. `nmgrs` corresponds to the number of managers. This is how Card and Krueger compute their full-time equivalent (FTE) employment variable (p.775 of the paper).

```
newdata <- ck1994 %>% mutate(empfte = empft + 0.5*emppt + nmgrs)
```

3. Compute the average number of FTE employment, average percentage of FT employees (out of the number of FTE employees), and average starting wage (`wage_st`) by state and by survey wave. Compare your results with *Table 2* of the paper.

```
averages <- newdata %>% group_by(state, observation) %>%
  summarise(wage_st_avg = mean(wage_st, na.rm = TRUE),
            empft_avg = mean(empft, na.rm = TRUE),
            empfte_avg = mean(empfte, na.rm = TRUE))
```

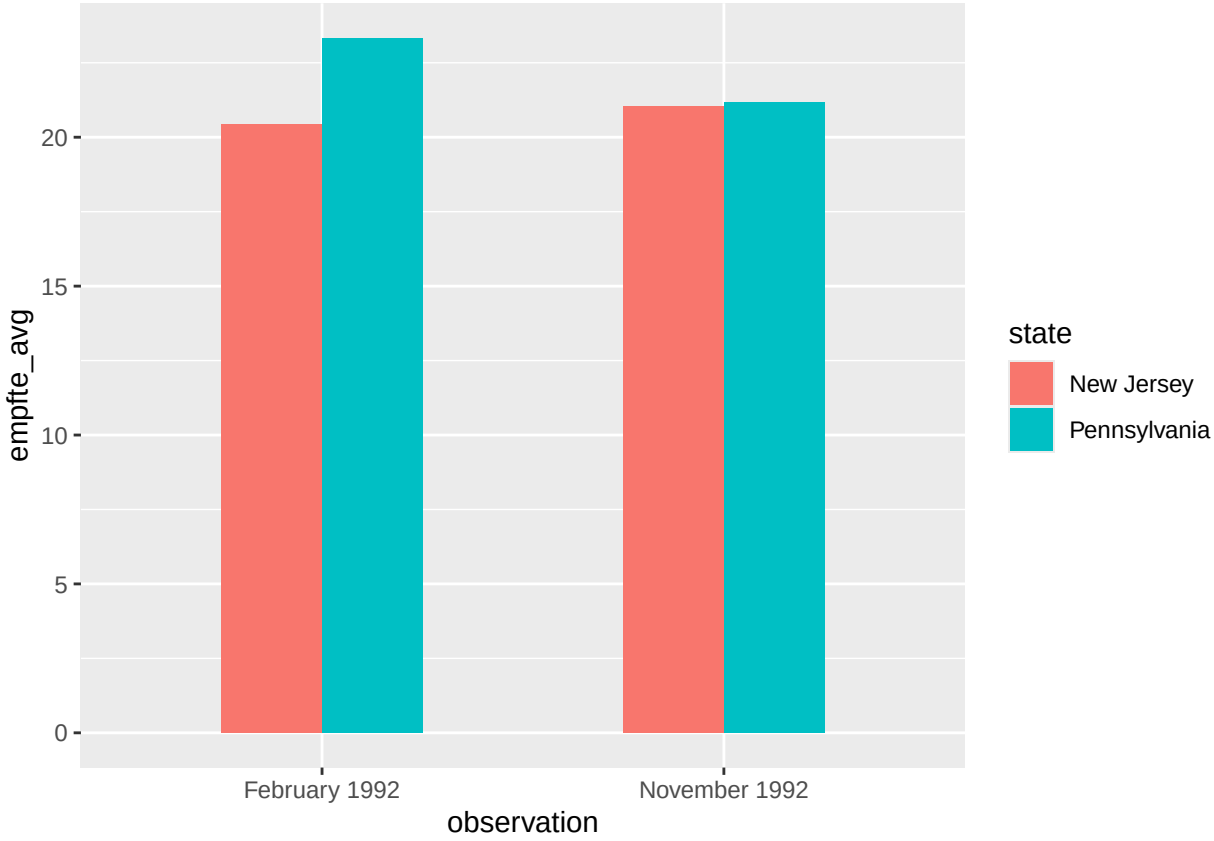
```
## `summarise()` has grouped output by 'state'. You can override using the
## `.groups` argument.
```

```
averages
```

```
## # A tibble: 4 x 5
## # Groups:   state [2]
##   state        observation   wage_st_avg empft_avg empfte_avg
##   <chr>        <chr>               <dbl>     <dbl>      <dbl>
## 1 New Jersey   February 1992        4.61      7.72       20.4
## 2 New Jersey   November 1992        5.08      8.45       21.0
## 3 Pennsylvania February 1992        4.63     10.2        23.3
## 4 Pennsylvania November 1992        4.62      7.56       21.2
```

4. Create a bar chart using the averages you created in the previous question with the y-axis of `empfte_avg` and x-axis of `observation`, and use `state` as the fill. Hint, use `geom_bar(stat="identity", width=.5, position = "dodge")` as the addition to your `ggplot` function.

```
ggplot(averages %>% ungroup(), aes(x = observation, y = empfte_avg, fill = state)) +
  geom_bar(stat="identity", width=.5, position = "dodge")
```



5. Calculate the difference in differences estimate.

```
# difference between New Jersey after and before minus
# the difference between Pennsylvania after and before
(averages$empfte_avg[2] - averages$empfte_avg[1]) -
  (averages$empfte_avg[4] - averages$empfte_avg[3])
```

```
## [1] 2.753606
```

Task 2

1. Create a dummy variable, `treat`, equal to 0 (or FALSE) if `state` is Pennsylvania and 1 (or TRUE) if New Jersey.

```
# these two are equivalent
analysisdata <- newdata %>% mutate(treat = ifelse(state == "New Jersey", 1, 0))
analysisdata <- newdata %>% mutate(treat = ifelse(state == "Pennsylvania", 0, 1))
```

2. Create a dummy variable, `post`, equal to 0 if `observation` is February 1992 and 1 otherwise.

```
analysisdata <- analysisdata %>% mutate(post = ifelse(observation == "February 1992", 0, 1))
```

3. Estimate the following regression model and interpret each coefficient.

$$empfte_{st} = \alpha + \beta treat_s + \gamma post_t + \delta(treat_s \times post_t) + \varepsilon_{st}$$

```
lm(empfte ~ treat + post + treat:post, analysisdata)
```

```
##
```

```
## Call:
## lm(formula = empfte ~ treat + post + treat:post, data = analysisdata)
##
## Coefficients:
## (Intercept)        treat         post   treat:post
##      23.331       -2.892       -2.166        2.754
```

```
# equivalent
lm(empfte ~ treat*post, analysisdata)
```

```
##
## Call:
## lm(formula = empfte ~ treat * post, data = analysisdata)
##
## Coefficients:
## (Intercept)        treat         post   treat:post
##      23.331       -2.892       -2.166        2.754
```

Note that these are all equivalent to what we calculated in the averages above.

- (Intercept) (23.331): This is the estimated average level of `empfte` in the control group before the treatment was introduced. This value represents the baseline level of the outcome variable for the control group in the pre-treatment period.
- treat (-2.892): This coefficient represents the difference in the outcome variable between the treatment and control groups in the pre-treatment period. In this case, the outcome level for the treatment group is estimated to be 2.892 units lower than for the control group in the pre-treatment period.
- post (-2.166): This coefficient represents the change in the outcome variable for the control group from the pre-treatment to the post-treatment period. Here, it suggests that for the control group, empfte decreased by an average of 2.166 units from the pre-treatment to the post-treatment period.
- treat:post (2.7543): This is the DID esimate, capturing the effect of the treatment on the outcome variable. Specifically, this coefficient estimates the difference in the change in the outcome variable between the treatment and control groups from the pre-treatment to the post-treatment period. Here, a positive value of 2.7543 indicates that the treatment is associated with an increase of about 2.75 units in empfte for the treatment group relative to what would have been expected based on the control group's trend.