Week 11: Causation

POP88162 Introduction to Quantitative Research Methods

Summarising Data

Read in the data for the minimum wage study by Card and Krueger (1994). You can find this dataset called `minwage.csv` on Blackboard.

```
minwage <- read.csv("../data/minwage.csv")
```

Let's start by conducting the usual checks of dataset's dimensionality, structure and distributions of variables.

```
str(minwage)
```

```
## 'data.frame':    358 obs. of  8 variables:
##  $ chain     : chr  "wendys" "wendys" "burgerking" "burgerking" ...
##  $ location  : chr  "PA" "PA" "PA" "PA" ...
##  $ wageBefore: num  5 5.5 5 5 5.25 5 5 5 5 5.5 ...
##  $ wageAfter : num  5.25 4.75 4.75 5 5 5 4.75 5 4.5 4.75 ...
##  $ fullBefore: num  20 6 50 10 2 2 2.5 40 8 10.5 ...
##  $ fullAfter : num  0 28 15 26 3 2 1 9 7 18 ...
##  $ partBefore: num  20 26 35 17 8 10 20 30 27 30 ...
##  $ partAfter : num  36 3 18 9 12 9 25 32 39 10 ...
```

```
summary(minwage)
```

```
##     chain             location           wageBefore      wageAfter
##  Length:358         Length:358         Min.   :4.250   Min.   :4.250
##  Class :character   Class :character   1st Qu.:4.250   1st Qu.:5.050
##  Mode  :character   Mode  :character   Median :4.500   Median :5.050
##                                        Mean   :4.618   Mean   :4.994
##                                        3rd Qu.:4.987   3rd Qu.:5.050
##                                        Max.   :5.750   Max.   :6.250
##   fullBefore       fullAfter        partBefore      partAfter
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 2.125   1st Qu.: 2.000   1st Qu.:11.00   1st Qu.:11.00
##  Median : 6.000   Median : 6.000   Median :16.25   Median :17.00
##  Mean   : 8.475   Mean   : 8.362   Mean   :18.75   Mean   :18.69
##  3rd Qu.:12.000   3rd Qu.:12.000   3rd Qu.:25.00   3rd Qu.:25.00
##  Max.   :60.000   Max.   :40.000   Max.   :60.00   Max.   :60.00
```

Subsetting Data

To simplify the ensuing analysis we will create two separate data frames: one, containing fast-food restaurants in New Jersey and another one with restaurants in Pennsylvania.

First, note how location is coded in the dataset. We have only one state name abbreviation for Pennsylvania (PA), but multiple ones for different parts of New Jersey (e.g. northNJ, shoreNJ, etc.)

```
table(minwage$location)
```

```
##
## centralNJ   northNJ        PA   shoreNJ   southNJ
##        45       146        67        33        67
```

To split up the data into two data data frames for each state we can use already familiar subsetting operations or, like here, a function `subset()`.

```r
minwageNJ <- subset(minwage, subset = (location != "PA"))
minwagePA <- subset(minwage, subset = (location == "PA"))
```

These two `subset()` function calls correspond to these subsetting operations:

```r
minwageNJ <- minwage[minwage$location != "PA",]
minwagePA <- minwage[minwage$location == "PA",]
```

As a first substantive data check let's start by examining what proportion of fast-food restaurants pay more than $5.05 before and after the introduction of new minimum wage set at this level in NJ.

```r
# NJ before
mean(minwageNJ$wageBefore < 5.05)
```

```
## [1] 0.9106529
```

```r
# NJ after
mean(minwageNJ$wageAfter < 5.05)
```

```
## [1] 0.003436426
```

```r
# PA before
mean(minwagePA$wageBefore < 5.05)
```

```
## [1] 0.9402985
```

```r
# PA after
mean(minwagePA$wageAfter < 5.05)
```

```
## [1] 0.9552239
```

Difference-in-means Analysis

Let's start our analysis by doing a simple difference in means comparison between NJ and PA after the introduction of new minimumwage in NJ.

First, we will create a new variable `fte_prop_after`, which will indicate the proportion of full-time employers after the change.

```r
minwageNJ$fte_prop_after <- minwageNJ$fullAfter/(minwageNJ$fullAfter + minwageNJ$partAfter)
minwagePA$fte_prop_after <- minwagePA$fullAfter/(minwagePA$fullAfter + minwagePA$partAfter)
```

We can now proceed to calculating the difference in means.

```r
mean(minwageNJ$fte_prop_after) - mean(minwagePA$fte_prop_after)
```

```
## [1] 0.04811886
```

And conducting a statistical test about this difference.

```r
t.test(minwageNJ$fte_prop_after, minwagePA$fte_prop_after)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  minwageNJ$fte_prop_after and minwagePA$fte_prop_after
## t = 1.4322, df = 99.761, p-value = 0.1552
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01854186  0.11477959
## sample estimates:
## mean of x mean of y
## 0.3204010 0.2722821
```

What is your substantive conclusion given this output?

Now, instead of testing this relationship using a t-test, fit a linear regression model to calculate the difference. Instead of working with two separate datasets for NJ and PA, for this task you might to want to modify the full `minwage` dataset.

Before-and-after Analysis

As we discussed in the lecture, rather than comparing post-change restaurants in NJ to their counterparts in PA, we might instead compare restaurants in NJ to themselves prior to the change in minimum wage.

First, let's create a new variable, which would capture the propotion of full-time employers in each fast-food restaurant in our dataset prior to the change.

```
minwageNJ$fte_prop_before <- minwageNJ$fullBefore/(minwageNJ$fullBefore + minwageNJ$partBefore)
```

We can now calculate the difference in means before and after the new law.

```
mean(minwageNJ$fte_prop_after) - mean(minwageNJ$fte_prop_before)
```

```
## [1] 0.02387474
```

And, as usually, run a statistical test on this difference.

```
t.test(minwageNJ$fte_prop_after, minwageNJ$fte_prop_before)
```

```
##
##  Welch Two Sample t-test
##
## data:  minwageNJ$fte_prop_after and minwageNJ$fte_prop_before
## t = 1.1952, df = 575.82, p-value = 0.2325
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01535869  0.06310817
## sample estimates:
## mean of x mean of y
## 0.3204010 0.2965262
```

Difference-in-difference (DiD) Analysis

Finally, let's conduct difference-in-difference analysis as discussed in the lecture. Recall that relative to before-and-after design it allows to address the confounding bias due to time trend and relative to simple difference-in-means between the two states it can (at least partially) address state-specific confounding.

Note that for DiD design we need two differences (hence, DiD!). First, we need to calculate before and after difference in one state, say, NJ.

```
NJdiff <- mean(minwageNJ$fte_prop_after) - mean(minwageNJ$fte_prop_before)
```

Next, we need to repeat the same for the other group, namely, fast-food restaurants in PA.

```r
minwagePA$fte_prop_before <- minwagePA$fullBefore/(minwagePA$fullBefore + minwagePA$partBefore)
PAdiff <- mean(minwagePA$fte_prop_after) - mean(minwagePA$fte_prop_before)
```

And, finally, we can calculate our difference-in-difference estimate.

```r
NJdiff - PAdiff
```

```
## [1] 0.06155831
```

Equivalently, we can also test the significance of this effect with a t-test.

```r
t.test(
  minwageNJ$fte_prop_after - minwageNJ$fte_prop_before,
  minwagePA$fte_prop_after - minwagePA$fte_prop_before
)
```

```
##
##  Welch Two Sample t-test
##
## data:  minwageNJ$fte_prop_after - minwageNJ$fte_prop_before and minwagePA$fte_prop_after - minwagePA$
## t = 1.3526, df = 90.777, p-value = 0.1796
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02884997  0.15196659
## sample estimates:
##   mean of x   mean of y
##  0.02387474 -0.03768357
```