Data Exploration Assignment - Econometrics

Erin Ankerson

## Packages

library(dplyr)  
library(vtable)  
library(rio)  
library(stringr)  
library(lubridate)  
library(fixest)  
library(ggplot2)

## Data Cleaning

### Scorecard Data

# Import College Scorecard data  
scorecard\_data <- import("Lab3\_Rawdata\\Most+Recent+Cohorts+(Scorecard+Elements).csv")  
 # Remove duplicates  
sc\_data\_cleaned <- scorecard\_data %>%  
 group\_by(INSTNM) %>%  
 filter(row\_number() == 1)  
 # Remove universities that do not predominantly grant bachelor's degrees  
sc\_data\_cleaned <- sc\_data\_cleaned %>%  
 filter(PREDDEG == '3')  
  
# Import ID Scorecard data to match colleges with data  
ID\_data <- import("Lab3\_Rawdata\\id\_name\_link.csv")  
  
# Check for any duplicate 'schnames' in ID\_data  
schname\_duplicates <- ID\_data %>%  
 group\_by(schname) %>%  
 summarize(count = n()) %>%  
 filter(count >=2)  
print(schname\_duplicates)

# A tibble: 32 × 2  
 schname count  
 <chr> <int>  
 1 american university of puerto rico 2  
 2 anderson university 2  
 3 aquinas college 2  
 4 augustana college 2  
 5 bethany college 2  
 6 bethel university 2  
 7 blue ridge community college 2  
 8 bradford school 2  
 9 bryan university 5  
10 california college san diego 3  
# ℹ 22 more rows

# Remove duplicates from the ID\_data  
ID\_data\_cleaned <- anti\_join(ID\_data, schname\_duplicates, by = 'schname')  
 # Check the data for the names  
ID\_dup\_check <- schname\_duplicates %>%  
 filter(schname %in% ID\_data\_cleaned)  
ID\_dup\_check

# A tibble: 0 × 2  
# ℹ 2 variables: schname <chr>, count <int>

### Trends Data

# Import "Trends Up To" Files and bind them into one dataset  
file\_names <- list.files(path = "C:\\Users\\Jean\\Documents\\2024 Winter Quarter\\Econometrics\\Projects and Data\\Lab3\_Rawdata\\trends\_up\_to", pattern = "\\.csv$", full.names = TRUE)  
  
trends\_up\_to\_data <- import\_list(file\_names, rbind = TRUE)  
  
trends\_clean <- trends\_up\_to\_data  
  
  
# Dates  
# Get the first 10 characters out of monthorweek variable  
trends\_clean$monthorweek <- str\_sub(trends\_clean$monthorweek, 1, 10)  
  
# Turn monthorweek string into usable date  
trends\_clean$monthorweek <- ymd(trends\_clean$monthorweek)  
class(trends\_clean$monthorweek)

[1] "Date"

# Aggregate further to round down to the first of each month  
trends\_clean$monthorweek <- floor\_date(trends\_clean$monthorweek, unit = c("month"))  
  
# Aggregate index variable  
#Standardize the index variable  
trends\_clean <- trends\_clean %>%  
 group\_by(schname, keyword) %>%  
 mutate(index = (index - mean(index, na.rm = TRUE)) / sd(index, na.rm = TRUE))  
## Now a one-unit change in the standardized index can be understood and interpreted as a one-standard-deviation change in search interest

I summarized the “monthorweek” variable to round down to the first of the month for each data point.

### Combine Data

Trends\_ID <- inner\_join(trends\_clean, ID\_data\_cleaned, by = c('schname' = 'schname'))  
  
data <- inner\_join(Trends\_ID, sc\_data\_cleaned, by = c('unitid' = 'UNITID'))  
  
# Note: The College Scorecard was released on September 12, 2015.  
# Differences-in-differences  
  
data$post\_treatment <- as.numeric(data$monthorweek >= "2015-10-01")  
  
# Convert md\_earn\_wne column to numeric  
data$`md\_earn\_wne\_p10-REPORTED-EARNINGS` <- as.numeric(as.character(data$`md\_earn\_wne\_p10-REPORTED-EARNINGS`))  
  
mean\_md\_earn\_wne <- mean(data$`md\_earn\_wne\_p10-REPORTED-EARNINGS`, na.rm = TRUE)  
  
# Treated Colleges = Colleges above the mean of the median earnings  
data$treated\_colleges <- ifelse(data$`md\_earn\_wne\_p10-REPORTED-EARNINGS` >= mean\_md\_earn\_wne, 1, 0)  
  
# Remove NA  
data <- na.omit(data)

The data starts in 2013, but according to my research, the College Scorecard was started on September 12, 2015. To take into account the major change after September 2015, I used difference-in-differences to create a post\_treatment variable.

In the post\_treatment variable, 1 = equal or greater than October 2015 and 0 = before October 2015.

To address “high-earning” vs. “low-earning” colleges, I found the mean of the variable “md\_earn\_wne\_p10-REPORTED-EARNINGS”. Anything above the mean is “high-earning” and anything below is “low-earning”. I saved this as a binary variable called “treated\_colleges” where 1 = high-earning and 0 = low-earning.

## Analysis

### Research Question:

Among colleges that predominantly grant bachelor’s degrees, did the release of the Scorecard shift student interest to high-earnings colleges relative to low-earnings ones?

### Regression

The regression I chose to use utilizes the standardized index for the Google Trends, as well as the post\_treatment variable to take into account before and after the College Scorecard was implemented and the treated\_colleges variable to take into account high vs. low earning colleges.

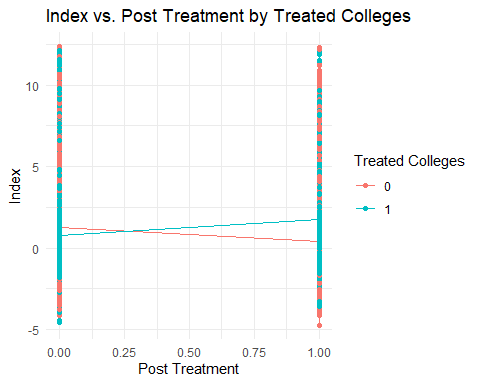
We are wondering if the release of the Scorecard shift student interest to high-earnings colleges relative to low-earnings ones, so the post\_treatment and treated\_colleges variables help take into account the before and after the Scorecard was implemented and what constitutes a high vs low earning college.

model <- feols(data = data, index ~ post\_treatment\*treated\_colleges)  
  
etable(model)

model  
Dependent Var.: index  
   
Constant 0.0360\*\*\* (0.0017)  
post\_treatment -0.2221\*\*\* (0.0041)  
treated\_colleges 0.0173\*\*\* (0.0026)  
post\_treatment x treated\_colleges -0.1073\*\*\* (0.0064)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type IID  
Observations 730,954  
R2 0.01006  
Adj. R2 0.01006  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Graph

# Graph of data  
  
ggplot(data, aes(x = post\_treatment, y = index, color = factor(treated\_colleges))) +  
 geom\_line() +  
 geom\_point() +  
 labs(x = "Post Treatment", y = "Index", color = "Treated Colleges", title = "Index vs. Post Treatment by Treated Colleges") +   
 theme\_minimal()



I graphed the model to see the effect of the Index vs. Post Treatment. The 0 represents before the implementation, and the 1 represents afterwards. The graph shows that the Index increased for high-earning colleges after the Scorecard was implemented and the Index decreased for low-earning colleges.

### Answering Research Question

The introduction of the College Scorecard increased the search activity on Google Trends for colleges with high-earning graduates by .0173 standard deviations relative to what it did for colleges with low-earning graduates, with a standard error of .0026. This result comes from the treated\_colleges coefficient in my regression.