Khanh Quach Data Exploration Project ECON 4110

# import the library will be used   
library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.1 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.0  
✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
✔ purrr 1.0.1   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(dplyr)  
library(lubridate)  
library(tidyr)  
library(purrr)  
library(stringr)  
library(fixest)  
library(rio)  
library(vtable)

Loading required package: kableExtra  
  
Attaching package: 'kableExtra'  
  
The following object is masked from 'package:dplyr':  
  
 group\_rows

# Getting date data  
ggtrend <- list.files(pattern = "trends\_up\_to\_", full.names = TRUE)  
# Read in the files and bind them together  
dataset <- import\_list(ggtrend, rbind = TRUE, fill = TRUE )

# Aggregating the Google Trends data   
dataset <- dataset %>%   
 mutate(week = str\_sub(monthorweek, start = 1, end = 10)) %>%   
 mutate(week = ymd(week)) %>%   
 mutate(month = floor\_date(week, unit = "month"))

dataset <- dataset %>%  
 group\_by(schname, keyword) %>%  
 mutate(std\_index = (index - mean(index))/sd(index))

# Reading the scorecard data and the id name data   
# Import the scorecard one.  
scorecard <- import("Most+Recent+Cohorts+(Scorecard+Elements).csv")  
id\_name\_link <- import("id\_name\_link.csv")

# Merge the scorecard data  
# First step: count and filter duplicate  
id\_name\_link <- id\_name\_link %>%  
 group\_by(schname) %>%  
 mutate(n = n()) %>%  
 filter(n == 1) # or drop can be fine.  
  
# Second step: unitid and or opeid columns to link with scorecard data.   
colnames(scorecard)[colnames(scorecard) == "UNITID"] = "unitid"  
  
# Join 2 data together  
  
id\_link <- inner\_join(id\_name\_link, scorecard, by ="unitid")  
gg\_link <-inner\_join(dataset, id\_link, by ="schname")

export(gg\_link, "finaldata.csv")

data\_to\_work <- import("finaldata.csv")

## The analysis

# Filter by the college == "3" since PREDEG is=3 in The scorecard dicitonary   
# mean that Predominatly bachelor - degree granting.   
merged\_data\_bachelors <- data\_to\_work %>%  
 filter(PREDDEG == 3)  
  
# Filter by the college == "3" since PREDEG is=3 in The scorecard dictionary  
merged\_data\_bachelors <- subset(data\_to\_work, PREDDEG == 3)  
  
week <- merged\_data\_bachelors %>%  
 group\_by(schname, monthorweek) %>%  
 mutate(week\_index = mean(index, na.rm = TRUE)) %>%  
 filter(!is.na(week\_index))

# make sure the column is numeric  
merged\_data\_bachelors$`md\_earn\_wne\_p10-REPORTED-EARNINGS` <- as.numeric(as.character(merged\_data\_bachelors$`md\_earn\_wne\_p10-REPORTED-EARNINGS`))

Warning: NAs introduced by coercion

# Calculate mean, standard deviation  
income.mean <- mean(na.omit(merged\_data\_bachelors$`md\_earn\_wne\_p10-REPORTED-EARNINGS`))  
income.sd <- sd(na.omit(merged\_data\_bachelors$`md\_earn\_wne\_p10-REPORTED-EARNINGS`))  
income.high <- income.mean + income.sd  
income.low <- income.mean - income.sd  
  
# Create binary variable for High/Low income  
merged\_data\_bachelors <- merged\_data\_bachelors %>%  
 mutate(Earnings = ifelse(`md\_earn\_wne\_p10-REPORTED-EARNINGS` >= income.mean, "High", "Low"))  
  
# Categorize incomes into High, Middle, and Low  
merged\_data\_bachelors <- merged\_data\_bachelors %>%  
 mutate(treated = case\_when(  
 `md\_earn\_wne\_p10-REPORTED-EARNINGS` >= income.high ~ "High",  
 `md\_earn\_wne\_p10-REPORTED-EARNINGS` <= income.low ~ "Low",  
 TRUE ~ "Middle Income" # this covers all other cases  
 ))

# create table to store variable that we will use   
merged\_data\_bachelors\_rec <- merged\_data\_bachelors %>%  
 select(unitid, schname, keyword, week, `md\_earn\_wne\_p10-REPORTED-EARNINGS`, Earnings, std\_index)

# Remove all of the missing values  
merged\_data\_bachelors\_rec <- drop\_na(merged\_data\_bachelors\_rec)

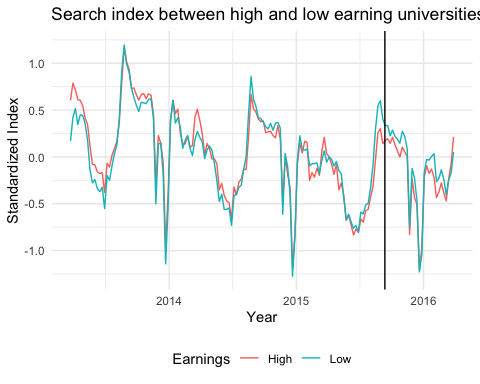
merged\_data\_bachelors\_rec <- merged\_data\_bachelors\_rec %>%   
 mutate(treated = `md\_earn\_wne\_p10-REPORTED-EARNINGS` >= income.high, post\_treatment = week >= as.Date("2015-09-12"))

#Build regression models  
# 1 regression   
reg <- feols(std\_index ~treated\* post\_treatment, data = merged\_data\_bachelors\_rec)  
etable(reg)

reg  
Dependent Var.: std\_index  
   
Constant 0.0363\*\*\* (0.0014)  
treatedTRUE -0.0076. (0.0041)  
post\_treatmentTRUE -0.1986\*\*\* (0.0032)  
treatedTRUE x post\_treatmentTRUE 0.0413\*\*\* (0.0097)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type IID  
Observations 741,598  
R2 0.00569  
Adj. R2 0.00569  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Dual line plot using un-dummy variable  
# One graph  
ggplot(merged\_data\_bachelors\_rec, aes(week, std\_index, color = Earnings)) +  
 stat\_summary(geom = 'line') +  
 labs(title = 'Search index between high and low earning universities post- and pre-treatment', x = 'Year', y = 'Standardized Index') +  
 geom\_vline(xintercept = as.Date ("2015-09-12")) +  
 theme\_minimal() + theme(legend.position = "bottom")

No summary function supplied, defaulting to `mean\_se()`



**Write up report:**

1. **Include at least one regression and one graph. ( Above )**
2. **Explain why you are performing the analysis you are performing and the choices you made in putting it together. Explain how your analysis addresses the research question**

**The goal of this analysis is to answer the research question:**

Did the introduction of the Scorecard cause a change in student interest, as reflected by Google searches for keywords related to high-earnings and low-earnings colleges, among institutions primarily conferring bachelor’s degrees?

**Things to Think about for the Analysis**

1. **Explain how your analysis addresses the research question**

The analysis addresses this question by using a Difference-in-Differences (DiD) regression model, a popular empirical strategy for estimating causal effects. It compares the changes over time in the dependent variable std\_index between a group that received the treatment high-earning colleges and a group that did not.

**Question 1:** There is a variable in the Scorecard with information about the median earnings of graduates ten years after graduation for each college. But how can we define "high-earning" and "low-earning" colleges? There's not a single answer - be ready to defend your choice.

After cleaning and merging all of the data, I found the information called  `md\_earn\_wne\_p10-REPORTED-EARNINGS`. This is a variable in the scorecard with information about the median earnings of graduates ten years after graduation for each college.

In order to define whether the average salary for the individual is considered high-income or low-income, I created a variable called "income. mean" to find the mean, and then I created a variable called sd to find the standard deviation of the salary for individual median earnings of graduates ten years after graduation for each college.

From this information, I decided to consider high-income earners as the ones who make more than $55,278 ( 43559 - mean + 11718 - sd). Similarly to that, I defined the lower-income variable and got the results of  $31,840

**Question 2:** What level should the data be at? You can leave the data as is, with one row per week per keyword. Or group\_by and summarize to put things to one week per college, or one month per college, or one month per keyword, etc. etc.

I decided to group\_by () and summarize to put things to one week per college.

As mentioned in the prompt given, " Google Trends indices are comparable only to themselves. That is, if the term "Boston university" increases from 15 to 16 from one week to the next, and the term "Seattle university" also went from 15 to 16, then we know that they both increased, but we don't know if the amount of increase was the same" Therefore in order to come out with the best result. I decided to filter by one week per college.

**Question 3** How should the regression model be designed to answer the question (transformations and functional form? Standard error adjustments? etc.), and how can we interpret the results once we have them?

std\_index = β0 + β1 \* income1 + β2 \* post\_treatment + β3 \* (income1 \* post\_treatment) + ε

Variable use to generate the regression: Week, id, median earnings -> low/ high. Low and High earnings, and std\_index ( standardize index)

In order to delve deeper into the research question, a line chart was generated. This chart comprises two distinct periods, demarcated by a vertical line representing the release of the Scorecard, which marks the beginning of the post-treatment phase.

Based on the calculated p-value, the null hypothesis suggesting no disparity between the treated group and the control group following the implementation of the treatment was rejected. Consequently, there exists statistically significant evidence indicating differing levels of search activity during the post-treatment period between the treated and control groups. Therefore, in terms of the treatment effect, which is typically represented by the interaction term in a Difference-in-Differences (DiD) setup, we reject the null hypothesis that the treatment has no effect.

**Any additional analyses you did that led you to design your main analysis that way (i.e. "I graphed Y vs. X and it looked nonlinear so I added a polynomial term" - you could even include this additional analysis if you like)**

Upon analyzing the line graph, it becomes evident that there is a noteworthy similarity in the trend observed between the high-earning and low-earning groups during the pre-treatment period. However, subsequent to the implementation of the treatment, there was a substantial surge in the search index for universities associated with “high earning,” while no noticeable change was observed in the search index for “low earning” universities.

**Explain what we should conclude, *in real world terms*, based on your results**

In real-world terms, this implies that students’ search patterns for colleges underwent a notable transformation following the release of the Scorecard. We observed a clear upswing in search activity for colleges associated with higher earnings, indicating a heightened attraction towards such institutions. Conversely, we did not observe a significant alteration in search activity for colleges associated with lower earnings.

These outcomes highlight the tangible impact of the Scorecard on students’ decision-making process when exploring potential colleges. The Scorecard’s provision of income-related information appears to have played a crucial role in steering students towards colleges with higher earning prospects.