My Le

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)

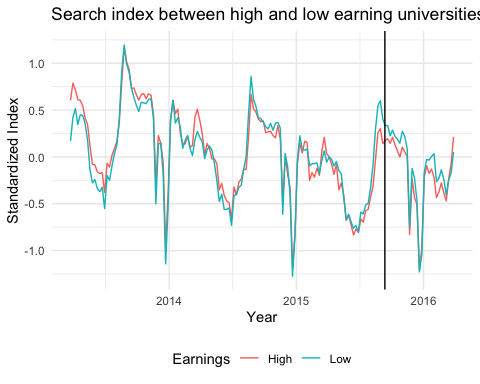
merged\_data\_bachelors\_rec <- merged\_data\_bachelors\_rec %>%   
 mutate(treated = `md\_earn\_wne\_p10-REPORTED-EARNINGS` >= income.high, post\_treatment = week >= as.Date("2015-09-12"))

#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()`



The Writeup

Research Question: The College Scorecard was released at the start of September 2015. 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 (as proxied by Google searches for keywords associated with those colleges)?

* Include at least one regression and one graph. The graph should be relevant to your analysis. Your graph might imply that you want to do something like use heteroskedasticity-robust standard errors, or a quadratic, or some sort of control variable. If it does, include that in your analysis.

My Answer: Above

* Explain why you are performing the analysis you are performing, and the choices you made in putting it together.

My Answer: The main reason I did this analysis and the options for putting it together is because this analysis was done to answer a key question for this project. In this way, I am able to evaluate the impact of the College Scorecard in shifting student interest to high-income colleges following a higher college graduation stock. The difference-in-differences (DID) regression model was chosen to explicitly compare pre- and post-search trends between the two groups upon release as a means of separating policy effects from time trends cause interference. The DID model provides a quasi-experimental framework for inferring causality from observational data with aggregation of data at the march level for each university. From here, I was able to focus on high-income and low-income organizations to directly address the project's research question.

* Explain how your analysis addresses the research question. (Including Things to Think for The Analysis - 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. 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 mont per keyword, etc. etc. 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?)

My Answer: My analysis addresses this research question using the Difference-in-Difference (DID) method, which is a common strategy in estimating causal effects. In this context, this DID method allows me to compare pre- and post-intervention outcomes in two separate groups that differ over time, which are the treatment group (the group admitted to high-income colleges) and the matched group (group not admitted to high-income colleges).

Question 1: The treatment here referred to is the publication of the College Scorecard. After cleaning the data, I classified "high-income" colleges and "low-income" schools. This is done using the variable md\_earn\_wne\_p10-REPORTED-EARNINGS from the College Scorecard, which is a scorecard variable that contains information about the average earnings of graduates ten years after graduation at each school. university. To determine the overall average salary, I drew a line between high and low earners. Using this approach, I calculated the overall mean of median income across all colleges, which is $43,559. In addition, I found the standard difference (sd) of these returns to be $11,718. Using these statistical data, colleges where graduates earn incomes one standard deviation above the average of $55,278. In contrast, colleges that graduate students earn less than the median minus one standard deviation of $31,841. The use of these specific numbers is an important step to better understand the issues that have been analyzed and adapted to effectively address the research question because it can provide a nuanced view. insight into how information about financial results affects student interest.

Question 2: My decision to aggregate the data was group\_by and summary to organize things into a week for each college, which addresses some important aspects of working with Google Trends data at the weekly degree. This can help me easily optimize the balance between level of detail and interpretability. Moreover, this approach also assists in minimizing the noise that weekly fluctuations can cause over short periods of time, which helps provide a clearer picture of trends. More specifically, by focusing on the weekly basis for each college, it can tell me that the increase in search volume from 15 to 16 for two different schools shows an increase in search volume from 15 to 16 for two different schools. increase student interest. However, the magnitude and significance of that increase may vary from school to school. So, deciding to group\_by each week for each college was a convenient approach so I could more directly evaluate the impact of the College Scorecard on the search behavior of prospective students.

Question 3: The designed regression model is std\_index = β0 + β1 \* income1 + β2 \* post\_treatment + β3 \* (income1 \* post\_treatment) + ε.

* "std\_index":
* "income1":
* "post\_treatment":
* "income1 \* post\_treatment":
* 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)

My Answer: In my analysis, I observed trends in the graph before and after the “treatment” period between the group of colleges with high-income graduates and the group of colleges with low income. This is considered a preliminary graphical analysis that helps me better understand its important role in shaping my analysis design. This way, I can look for standardized over time (X) and visually check the parallel trend assumption. The chart shows that before the release of the College Scorecard, search trends for both high- and low-income college groups followed a similar path. However, after implementing the College Scorecard, the chart shows a significant increase in search activity for high-income colleges, while there were no notable changes in the search index low-income colleges. This difference from previous parallel trends suggests that the College Scorecard has more impact on student interest in decision making for these two distinct groups. This additional graph trend analysis provides nuanced observations in a comprehensive way to ensure comprehensive coverage of the College Scorecard's impact on student search behaviors.

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

My Answer: In real world terms, I conclude based on my results that the analysis shows that students' college search patterns have been significantly affected following the release of the college scorecard. During my observations, I detected a clear increase in Google searches for high-paying colleges. In contrast, my observations show a significant shift in low-income college search practices. From this it can be seen that the introduction of the College Scorecard seems to have shifted student interest to colleges where graduates are statistically more likely to earn higher incomes. This suggests that students may prioritize financial outcomes in their educational choices when the information is readily available. From the results of this analysis, I can infer some key tangible impacts of the College Scorecard on students' decision making when researching colleges. One of the main impacts of this incident is that students have significantly increased awareness of the economic outcomes associated with different universities. The provision of this information plays an important role in students making decisions about their education because students actively seek to enter universities with high future earnings prospects. his hybrid. Another impact is policy-related, which means the College Scorecard can provide transparency in educational outcomes in universities by providing clear information. This can assist students in being able to easily consider different factors in their decision-making process. Briefly, the results of this analysis show behavioral changes among prospective students, especially those who prioritize economic outcomes when looking up information. Besides, this emphasizes on the value of data-driven approach in education to be able to deeply understand its broader significance in shaping the bios of today's educational landscape.