Data Exploration

Tina Chau

2023-05-21

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyverse)

## ── Attaching packages  
## ───────────────────────────────────────  
## tidyverse 1.3.2 ──

## ✔ ggplot2 3.4.0 ✔ purrr 1.0.0  
## ✔ tibble 3.1.8 ✔ stringr 1.5.0  
## ✔ tidyr 1.2.1 ✔ forcats 0.5.2  
## ✔ readr 2.1.3   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(rio)  
library(lubridate)

##   
## Attaching package: 'lubridate'  
##   
## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

library(readr)  
library(vtable)

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

library(stringr)  
library(fixest)  
library(ggplot2)

# read in trends data   
trends <- list.files('Lab3\_Rawdata',pattern = 'trends\_up\_to', full.names = TRUE)  
data <- import\_list(trends, rbind = TRUE, fill = TRUE)

# aggregate trends data   
data <- data %>% mutate(week = str\_sub(monthorweek, start = 1, end = 10))%>% mutate(week = ymd(week))%>% mutate(month = floor\_date(week, 'month'))

# standardize index by school name and keyword  
data <- data %>%  
 group\_by(schname, keyword) %>%  
 mutate(standardized\_index = (index - mean(index)) / sd(index))

# read in Scorecard data  
scorecard <- import('Lab3\_Rawdata/Most+Recent+Cohorts+(Scorecard+Elements).csv')  
  
names(scorecard)[1] <- 'unitid'  
names(scorecard)[2] <- 'opeid'

# read in id\_name\_link file  
id <- import('Lab3\_Rawdata/id\_name\_link.csv')  
id\_count <- id %>%  
 group\_by(schname) %>%  
 mutate(schname\_count = n()) %>%  
 filter(schname\_count == 1) %>%  
 ungroup()

# merge trends and id   
trend\_and\_id\_merged <- inner\_join(data, id, by = 'schname', copy = TRUE)

# merge trends and id with scorecard  
merged\_data <- inner\_join(trend\_and\_id\_merged, scorecard, by = 'unitid')

# merge trends and id with scorecard  
merged\_data <- inner\_join(trend\_and\_id\_merged, scorecard, by = 'unitid')  
  
  
merged\_bachelor\_data <- filter(merged\_data, PREDDEG == 3)

combined\_data <- merged\_bachelor\_data %>%  
 group\_by(schname, monthorweek) %>%  
 mutate(weekIndex = mean(index))  
  
combined\_data <- na.omit(combined\_data)

combined\_data$`md\_earn\_wne\_p10-REPORTED-EARNINGS` <- as.numeric(combined\_data$`md\_earn\_wne\_p10-REPORTED-EARNINGS`)

## Warning: NAs introduced by coercion

# find median earnings   
median\_earnings <- median(combined\_data$"md\_earn\_wne\_p10-REPORTED-EARNINGS", na.rm = TRUE)  
  
median\_earnings

## [1] 41900

# create binary variable both income and date   
combined\_data$income\_category <- ifelse(combined\_data$"md\_earn\_wne\_p10-REPORTED-EARNINGS" > 41900, 'High', 'Low')  
combined\_data <- combined\_data %>% filter(!is.na(income\_category))  
combined\_data$scorecard <- ifelse(combined\_data$month < as.Date("2015-09-01"), "Before", "After")

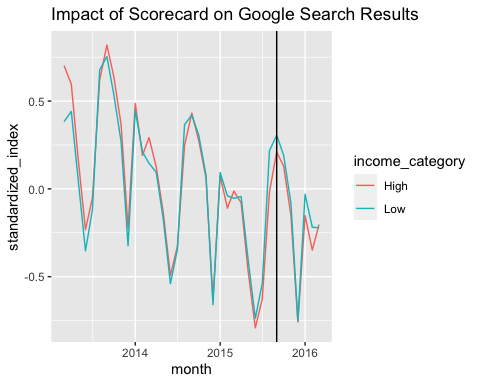
new\_combined <- combined\_data %>%   
 group\_by(income\_category, month) %>%   
 summarise(standardized\_index = mean(standardized\_index))

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

new\_combined

## # A tibble: 74 × 3  
## # Groups: income\_category [2]  
## income\_category month standardized\_index  
## <chr> <date> <dbl>  
## 1 High 2013-03-01 0.702   
## 2 High 2013-04-01 0.596   
## 3 High 2013-05-01 0.165   
## 4 High 2013-06-01 -0.232   
## 5 High 2013-07-01 -0.0504  
## 6 High 2013-08-01 0.621   
## 7 High 2013-09-01 0.819   
## 8 High 2013-10-01 0.638   
## 9 High 2013-11-01 0.367   
## 10 High 2013-12-01 -0.223   
## # … with 64 more rows

# create plot comparing high and low income vs. standardized index  
index\_graph <- ggplot(data = new\_combined, aes(x = month, y = standardized\_index, color = income\_category)) +  
 geom\_line() +  
 geom\_vline(xintercept = ymd('2015-09-01')) +  
 ggtitle('Impact of Scorecard on Google Search Results')  
  
print(index\_graph)



# regression   
regression <- feols(standardized\_index ~ scorecard + income\_category + month + scorecard \* income\_category, data = combined\_data)  
etable(regression)

## regression  
## Dependent Var.: standardized\_index  
##   
## Constant 12.86\*\*\* (0.0810)  
## scorecardBefore -0.2040\*\*\* (0.0048)  
## income\_categoryLow 0.0698\*\*\* (0.0051)  
## month -0.0008\*\*\* (4.82e-6)  
## scorecardBefore x income\_categoryLow -0.0856\*\*\* (0.0057)  
## \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
## S.E. type IID  
## Observations 764,103  
## R2 0.03790  
## Adj. R2 0.03789  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**The Writeup:**

The question I am trying to answer here is, “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?” To answer this question, I created a visual of the regression above to display the performed an analysis combining visualizations and regression modeling.

Firstly, I created a line plot using ggplot2 to visualize the relationship between the standardized index (a measure of student interest) and time (months). I chose to display two lines, one for high-income category and another for low-income category, using different colors. This allows for a comparison between the two income categories over time. I also included a vertical line at September 2015 to highlight the release of the Scorecard, as it is an important event for this analysis.

In addition to the visualization, I conducted a regression analysis using the feols function. The regression model examines the relationship between the standardized index (dependent variable) and several independent variables, including the Scorecard, income category, month, and the interaction between the Scorecard and income category. By including these variables in the model, I can assess their individual and combined effects on the standardized index. Then, to summarize the results of the regression, I used the etable function.

By combining these approaches, I was able to provide a comprehensive analysis that not only visually depicted the trends but also quantitatively assessed the relationship between the Scorecard, income category, and student interest. This allowed for a greater understanding of whether the Scorecard influenced student interest in high-earnings colleges compared to low-earnings ones.

Further looking into our regression results, we can see that the “Before” coefficient of -0.204 indicates that there is a decrease in the standardized index on average compared to the “After” group. This implies that there is potentially a decline in student interest prior to the scorecard release.

The income\_categoryLow group has a positive coefficient of 0.0698 which indicates that colleges in the low-income category have a higher standardized index compared to colleges in the high-income category.

The coefficient of -0.0856 for scorecardBefore x income\_categoryLowscorecard group indicates a negative relationship between scorecard and student interest for colleges in the low-income category.

And lastly, the -0.0008 coefficient for month indicates a one unit change in month is associate with -.0008 decrease in the standardized index of Google search.

Based on our results here, we can conclude that the release of the Scorecard had an impact on student interest in colleges. Prior to the Scorecard, there was a decrease in student interest on average. Colleges in the low-income category attracted greater student interest compared to those in the high-income category. And the Scorecard had a stronger negative effect on student interest in low-income colleges, suggesting potential disparities in perceived value and outcomes.