Data Exploration Project

Ian Ceazar Basco

## 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)?

## Libraries

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

## Data Cleaning

Reading and Aggregating the Google Trends data

googleTrends\_list <- list.files(path = "./Lab3\_Rawdata", pattern = "trends\_up\_to\_", full.names = TRUE)  
  
  
googleTrends\_rawData <- import\_list(googleTrends\_list, rbind = TRUE)  
  
googleTrends\_dataset <- googleTrends\_rawData %>%   
 mutate(dateWeek = ymd(str\_sub(monthorweek, end = 10))) %>%   
 mutate(dateMonth = floor\_date(dateWeek, 'month'))%>%   
 group\_by(schname, keyword) %>%  
 mutate(indexStandard = (index - mean(index, na.rm = TRUE)) / sd(index, na.rm = TRUE)) %>%   
 group\_by(schname, dateMonth) %>%   
 summarize(indexStandard = sum(indexStandard, na.rm = TRUE)) %>%   
 drop\_na()

Reading and Merging the Scorecard data and filtering the data with just the predominantly bachelor’s degree-granting schools.

scoreCard <- import('./Lab3\_Rawdata/Most+Recent+Cohorts+(Scorecard+Elements).csv')  
names(scoreCard) <- tolower(names(scoreCard))  
# print(scoreCard)  
  
idName <- import('./Lab3\_Rawdata/id\_name\_link.csv')  
idName <- idName %>%   
 group\_by(schname) %>%   
 mutate(n = n()) %>%   
 filter(n == 1)  
  
join\_idName <- inner\_join(idName, googleTrends\_dataset, join\_by(schname))  
join\_scoreCard <- inner\_join(join\_idName, scoreCard, join\_by(unitid, opeid))  
# summary(join\_scoreCard)  
  
mainData <- join\_scoreCard %>%   
 filter(preddeg == 3) %>% # Predominantly bachelor's-degree granting  
 mutate(earningsReported = as.integer(`md\_earn\_wne\_p10-reported-earnings`)) %>% #change to integer  
 drop\_na()  
  
# vtable(mainData)  
# summary(mainData)  
# summary(mainData$earningsReported) # mean = $42,808  
  
### use $42,808 as baseline for income.  
### Sept 01, 2015 as cutoff

## Assumptions

To answer the research question we have to:

**Define what are high and low-earning schools.**  
I decided to look at the summary of the variable that reports the median earnings of graduates for each college. We can see that we have a mean of $42,808 and I decided to use that as my baseline. I think that it will be a better baseline because it is higher than the median which I feel is a good separator between low vs high income colleges. I created the earningMedian variable that has ‘1’ if the earnings reported are higher or equal to the mean value and outputs ‘0’ otherwise.

mainData <- mainData %>%   
 mutate(earningMedian = case\_when(earningsReported >= 42808 ~ '1',  
 earningsReported < 42808 ~ '0')) %>%   
 mutate(releasedDate = case\_when(dateMonth >= ymd("2015-09-01") ~ '1',  
 dateMonth < ymd("2015-09-01") ~ '0'))  
summary(mainData$earningsReported)

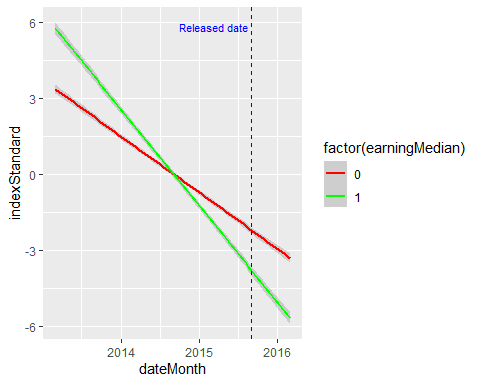
Min. 1st Qu. Median Mean 3rd Qu. Max.   
 17600 35800 41200 42808 48800 121500

**Decide what level should the data be at.**  
I decided to go with the school-month level because data cleaning would be simpler and easier when you need to figure out which data were from after or before the release of the Scorecard website. To achieve this, I created a new variable releasedDate which enters a ‘1’ if the date is after September 2015 and a ‘0’ if it is before.

## Visualization

I was interested to get a visual of what the relationship is between the standardized index vs. the year for high and low-earning schools. To my surprise, the indexes for high-earning (1) schools had a higher drop compared to low-earning (0) schools even after the release of the Scorecard website.

ggplot(mainData, aes(x= dateMonth , y = indexStandard, colour = factor(earningMedian))) +  
 geom\_smooth(method = 'lm') +  
 geom\_vline(xintercept = ymd("2015-09-01"), linetype = "dashed", color = "blue") +  
 annotate("text", x = ymd("2015-09-01"), y = 6, label = "Released date", vjust = 1, hjust = 1.05, color = "blue", size = 3) +  
 scale\_color\_manual(values = c("1" = "green",  
 "0" = "red"))



# high earning colleges seems to have a lower index than low earning  
# colleges after the 09/2015 released date.

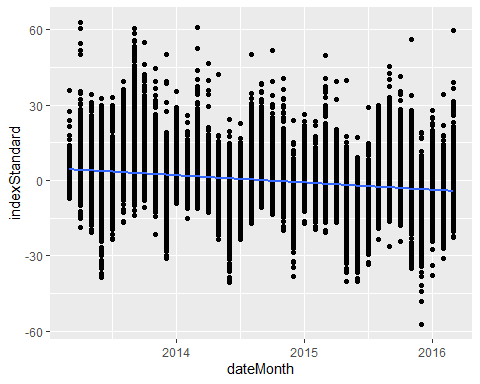
I wanted to confirm this by grouping the data by earningMedian and releasedDate and summarizing the index. The results confirmed that higher-earning (0.743) colleges had better indexes compared to low-earning (0.377) colleges BEFORE the launch of the Scorecard. The story is different post-launch as both categories experienced drops in their indexes. Particularly, higher-earning schools (-3.16)witnessed a more significant decline than lower-earning schools (-1.60).

mainData %>%   
 group\_by(earningMedian, releasedDate) %>%   
 summarize(mean(indexStandard))

# A tibble: 4 × 3  
# Groups: earningMedian [2]  
 earningMedian releasedDate `mean(indexStandard)`  
 <chr> <chr> <dbl>  
1 0 0 0.377  
2 0 1 -1.60   
3 1 0 0.743  
4 1 1 -3.16

I also wanted to see if we can visually see if heteroskedasticity is present. We can see from the below graph the uneven distribution of the scatterplot. This uneven distribution suggests that the variance of the dependent variable changes as the independent variable increases or decrease.

ggplot(mainData, aes(x = dateMonth, y = indexStandard)) +   
 geom\_point() +   
 geom\_smooth(method = 'lm')



These analyses give us an idea of what our regressions would look like.

## Regressions

With my first regression, I went with using the standardized index as my dependent variable and included an interactive term in my independent variable. I added the interactive term to explore how the relationship between two variables changes based on the values of the other variable.

regressionInt <- feols(indexStandard ~ i(month(dateMonth)) + releasedDate + earningMedian + releasedDate\*earningMedian, data = mainData, , vcov = 'hetero')  
# summary(regressionInt)  
etable(regressionInt)

regressionInt  
Dependent Var.: indexStandard  
   
Constant 3.032\*\*\* (0.1046)  
month(dateMonth) = 2 -2.481\*\*\* (0.1288)  
month(dateMonth) = 3 -1.362\*\*\* (0.1352)  
month(dateMonth) = 4 -0.8856\*\*\* (0.1457)  
month(dateMonth) = 5 -5.567\*\*\* (0.1558)  
month(dateMonth) = 6 -10.41\*\*\* (0.1692)  
month(dateMonth) = 7 -7.160\*\*\* (0.1499)  
month(dateMonth) = 8 1.661\*\*\* (0.1583)  
month(dateMonth) = 9 5.361\*\*\* (0.1688)  
month(dateMonth) = 10 2.649\*\*\* (0.1468)  
month(dateMonth) = 11 -0.5934\*\*\* (0.1504)  
month(dateMonth) = 12 -8.966\*\*\* (0.1474)  
releasedDate1 -3.861\*\*\* (0.1022)  
earningMedian1 0.3621\*\*\* (0.0749)  
releasedDate1 x earningMedian1 -1.924\*\*\* (0.1726)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type Heteroskedast.-rob.  
Observations 51,565  
R2 0.28824  
Adj. R2 0.28805  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Using etable we can see that all the independent variables are statistically significant. We can interpret the coefficient of earningMedian1, 0.3621, as the effect of high-income colleges on index scores before the release of the Scorecard. This means that higher-earning colleges have a 0.3621 higher index compared to low-earning colleges before the release of the Scorecard. We can interpret the interactive term releasedDate1 x earningMedian1 as the effect of the Scorecard release date on the index is 1.924 lower for higher-earning colleges than lower-earning colleges.

With a p-value of 2.2e-16, the wald test shows a result that is statistically significant at the 95% level.

wald(regressionInt)

Wald test, H0: joint nullity of month(dateMonth)::2, month(dateMonth)::3, month(dateMonth)::4, month(dateMonth)::5, month(dateMonth)::6, month(dateMonth)::7 and 8 others  
 stat = 1,296.9, p-value < 2.2e-16, on 14 and 51,550 DoF, VCOV: Heteroskedasticity-robust.

With my second regression, I decided to add a fixed effect to my equation. I used schname for my fixed effect term and compared both regressions using etable.

regressionFE <- feols(indexStandard ~ i(month(dateMonth)) + releasedDate + earningMedian + releasedDate\*earningMedian | schname, data = mainData, vcov = 'hetero')  
# summary(regressionFE)  
etable(regressionInt,regressionFE)

regressionInt regressionFE  
Dependent Var.: indexStandard indexStandard  
   
Constant 3.032\*\*\* (0.1046)   
month(dateMonth) = 2 -2.481\*\*\* (0.1288) -2.481\*\*\* (0.1305)  
month(dateMonth) = 3 -1.362\*\*\* (0.1352) -1.359\*\*\* (0.1372)  
month(dateMonth) = 4 -0.8856\*\*\* (0.1457) -0.8858\*\*\* (0.1477)  
month(dateMonth) = 5 -5.567\*\*\* (0.1558) -5.567\*\*\* (0.1580)  
month(dateMonth) = 6 -10.41\*\*\* (0.1692) -10.41\*\*\* (0.1716)  
month(dateMonth) = 7 -7.160\*\*\* (0.1499) -7.160\*\*\* (0.1520)  
month(dateMonth) = 8 1.661\*\*\* (0.1583) 1.661\*\*\* (0.1605)  
month(dateMonth) = 9 5.361\*\*\* (0.1688) 5.361\*\*\* (0.1711)  
month(dateMonth) = 10 2.649\*\*\* (0.1468) 2.649\*\*\* (0.1489)  
month(dateMonth) = 11 -0.5934\*\*\* (0.1504) -0.5934\*\*\* (0.1525)  
month(dateMonth) = 12 -8.966\*\*\* (0.1474) -8.966\*\*\* (0.1495)  
releasedDate1 -3.861\*\*\* (0.1022) -3.862\*\*\* (0.1036)  
earningMedian1 0.3621\*\*\* (0.0749)   
releasedDate1 x earningMedian1 -1.924\*\*\* (0.1726) -1.924\*\*\* (0.1750)  
Fixed-Effects: ------------------- -------------------  
schname No Yes  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type Heteroskedast.-rob. Heteroskedast.-rob.  
Observations 51,565 51,565  
R2 0.28824 0.28825  
Within R2 -- 0.28825  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The results that we get are similar to that of the first regression. However, I will still recommend using the model that included the fixed effects since it controls for variations within schools. Similar to the first regression, the wald test shows a result that is statistically significant at the 95% level.

wald(regressionFE)

Wald test, H0: joint nullity of month(dateMonth)::2, month(dateMonth)::3, month(dateMonth)::4, month(dateMonth)::5, month(dateMonth)::6, month(dateMonth)::7 and 7 others  
 stat = 1,356.9, p-value < 2.2e-16, on 13 and 50,148 DoF, VCOV: Heteroskedasticity-robust.

## Conclusion

Based on my calculations, the introduction of the College Scorecard decreased search activity on Google Trends for colleges with high-earning graduates by 1.924 units relative to what it did for colleges with low-earning graduates, with a standard error of 0.1750. This result comes from the releasedDate1 x earningMedian1 coefficient(s) in my regression.

etable(regressionFE)

regressionFE  
Dependent Var.: indexStandard  
   
month(dateMonth) = 2 -2.481\*\*\* (0.1305)  
month(dateMonth) = 3 -1.359\*\*\* (0.1372)  
month(dateMonth) = 4 -0.8858\*\*\* (0.1477)  
month(dateMonth) = 5 -5.567\*\*\* (0.1580)  
month(dateMonth) = 6 -10.41\*\*\* (0.1716)  
month(dateMonth) = 7 -7.160\*\*\* (0.1520)  
month(dateMonth) = 8 1.661\*\*\* (0.1605)  
month(dateMonth) = 9 5.361\*\*\* (0.1711)  
month(dateMonth) = 10 2.649\*\*\* (0.1489)  
month(dateMonth) = 11 -0.5934\*\*\* (0.1525)  
month(dateMonth) = 12 -8.966\*\*\* (0.1495)  
releasedDate1 -3.862\*\*\* (0.1036)  
releasedDate1 x earningMedian1 -1.924\*\*\* (0.1750)  
Fixed-Effects: -------------------  
schname Yes  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type Heteroskedast.-rob.  
Observations 51,565  
R2 0.28825  
Within R2 0.28825  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1