Data Exploration Challenge

Romane Beeharry

## Data Exploration Challenge

**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)?:

#Load the necessary libraries  
library(tidyverse)

Warning: package 'tidyverse' was built under R version 4.2.3

Warning: package 'ggplot2' was built under R version 4.2.2

Warning: package 'tibble' was built under R version 4.2.3

Warning: package 'tidyr' was built under R version 4.2.2

Warning: package 'readr' was built under R version 4.2.3

Warning: package 'purrr' was built under R version 4.2.2

Warning: package 'dplyr' was built under R version 4.2.3

Warning: package 'stringr' was built under R version 4.2.3

Warning: package 'forcats' was built under R version 4.2.3

Warning: package 'lubridate' was built under R version 4.2.3

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.4 ✔ readr 2.1.5  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ ggplot2 3.4.1 ✔ tibble 3.2.1  
✔ lubridate 1.9.3 ✔ 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(ggplot2)  
library(rio)

Warning: package 'rio' was built under R version 4.2.3

library(fixest)

Warning: package 'fixest' was built under R version 4.2.3

To observe the effect of the release of the Scorecard on student interest I will perform a set of regressions based on the Scorecard Trends Analysis data. The data contains the standardized indices by month for each school, which represents the search activity for a given month. The data is limited to colleges that predominantly grant bachelor’s degree.

It also contains information about the median earnings of individuals 10 years after having attended a school. This variable was converted into a binary variable to distinguish between high-earning and low-earning schools. The binary variable “HighEarning” is defined by a 1 if the earnings are greater than the mean average earning of $42300. Anything below the threshold is considered low income and defined by a 0.

Moreover, to account for the release of the scorecard, a binary variable “PostScorecard” was created by defining any date after September 2015 with one and any date before September 2015 with 0.

In all of the regressions, the dependent variable is StandardizedIndex, which is a measure of search activity. The independent variables include PostScorecard and HighEarning, as well as SchoolType.

reg\_data <- import("C:/Users/roman/OneDrive/Documents/OMSBA 5300 Econometrics/Data Exploration Challenge/Week 7 Data Exploration Challenge/Analysis Data/ScorecardTrendsAnalysis\_data.csv")  
  
reg\_data <- na.omit(reg\_data)

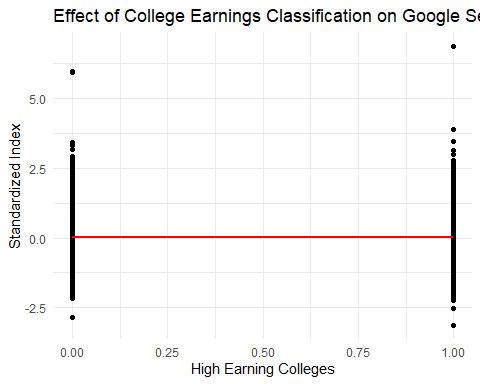
The first regression observes the effect of earnings on search interest before the release of the Scorecard. This regression is performed to observe if earning classification had any effect on the search activity before the scorecard was released.

reg <- feols(StandardizedIndex ~ HighEarning, data = reg\_data)  
etable(reg)

reg  
Dependent Var.: StandardizedIndex  
   
Constant 0.0044 (0.0035)  
HighEarning 0.0047 (0.0049)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type IID  
Observations 56,944  
R2 1.59e-5  
Adj. R2 -1.64e-6  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Create a ggplot scatter plot  
ggplot(reg\_data, aes(x = HighEarning, y = StandardizedIndex)) +  
 geom\_point() +   
 labs(  
 x = "High Earning Colleges",  
 y = "Standardized Index",  
 title = "Effect of College Earnings Classification on Google Search Activity"  
 ) +  
 geom\_smooth(method = "lm", se = FALSE, color = "red") +   
 theme\_minimal()

`geom\_smooth()` using formula = 'y ~ x'



In the table, we can see that the coefficient for “HighEarning” is 0.0047 with a standard error of 0.0049. This coefficient represents the average change in the standardized index for colleges classified as high-earning compared to those classified as low-earning. The coefficient is positive, indicating a positive relationship between high-earning status and the standardized index. For a unit increase in search activity, or standardized index, there is 0.0047 increase in interest for high-earning colleges. The result is not statistically significant at conventional levels (p > 0.05), which indicates that colleges being classified as high-earning or low earning does not have an effect on search activity before the release of the Scorecard. The graph shows that search activity for low-earning and high-earning colleges are very similar.

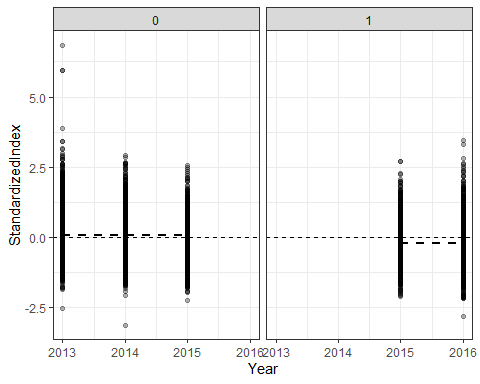
The second regression observes the effect of the release of the Scorecard on search activity.

reg\_post <- feols(StandardizedIndex ~ PostScorecard, data = reg\_data, vcov = 'hetero')  
etable(reg\_post)

reg\_post  
Dependent Var.: StandardizedIndex  
   
Constant 0.0585\*\*\* (0.0027)  
PostScorecard -0.2712\*\*\* (0.0063)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type Heteroskedast.-rob.  
Observations 56,944  
R2 0.03259  
Adj. R2 0.03257  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

reg\_data$Year <- year(reg\_data$Month)  
  
# Create the ggplot visualization   
ggplot(reg\_data, aes(Year, StandardizedIndex)) +  
 facet\_grid(~ PostScorecard) +   
 geom\_line(aes(y = fitted(reg\_post)), linetype = 'dashed', size = 0.8) +   
 geom\_point(alpha = 0.3) +   
 geom\_hline(yintercept = 0, linetype = "dashed") +   
 theme\_bw()

Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
ℹ Please use `linewidth` instead.



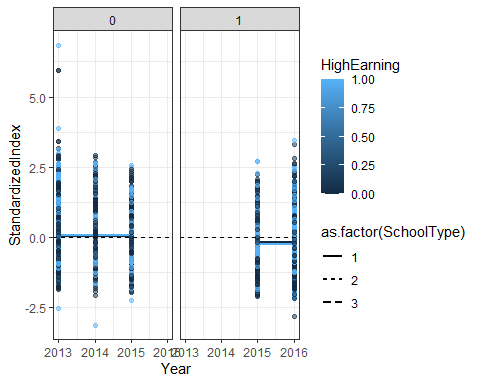
In the table, we can see that the coefficient for PostScorecard is -0.2712 with a standard error of 0.0063. This coefficient is associated with a decrease in search activity after the release of the College Scorecard. The negative coefficient suggests that, on average, the standardized index decreased by 0.2712 unit changes after the release of the Scorecard. The graph shows the search activity before the release of the scorecard and after the release of the scorecard over the years. The fitted line shows a slight decrease in search activity from 2015 to 2016.

Let’s see if there is any effects on search activity from PostScorecard release and Earning classification. Moreover, School Type is included as a fixed effect to control for different school types such as Public, Private Non-Profit and Private For-Profit.

reg\_earningpost <- feols(StandardizedIndex ~ PostScorecard + PostScorecard \* HighEarning | SchoolType, data = reg\_data, vcov = 'hetero')  
etable(reg\_earningpost)

reg\_earningpost  
Dependent Var.: StandardizedIndex  
   
PostScorecard -0.2197\*\*\* (0.0087)  
HighEarning 0.0241\*\*\* (0.0054)  
PostScorecard x HighEarning -0.1029\*\*\* (0.0126)  
Fixed-Effects: -------------------  
SchoolType Yes  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type Heteroskedast.-rob.  
Observations 56,944  
R2 0.03379  
Within R2 0.03378  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

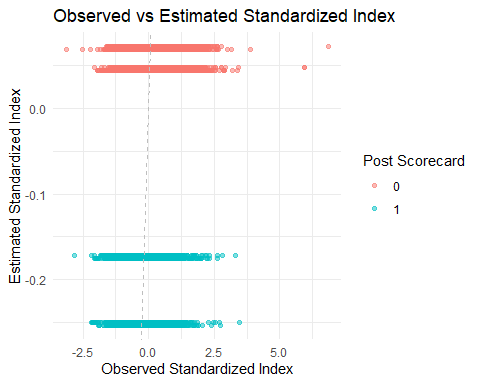
#Plot the regression  
ggplot(reg\_data, aes(Year, StandardizedIndex, group = interaction(SchoolType, HighEarning), col = HighEarning)) +  
 facet\_grid(~ PostScorecard) +   
 geom\_line(aes(y = fitted(reg\_earningpost), linetype = as.factor(SchoolType)), size = 0.8) +   
 geom\_point(alpha = 0.5) +   
 geom\_hline(yintercept = 0, linetype = "dashed") +   
 theme\_bw()



library(ggpubr)

Warning: package 'ggpubr' was built under R version 4.2.2

# Create a plot to visualize the regression results  
ggplot(reg\_data, aes(x = StandardizedIndex, y = fitted(reg\_earningpost), color = factor(PostScorecard))) +  
 geom\_point(alpha = 0.5) +   
 geom\_abline(intercept = 0, slope = 1, linetype = "dashed", color = "gray") +  
 labs(x = "Observed Standardized Index", y = "Estimated Standardized Index", color = "Post Scorecard") + # Add labels  
 ggtitle("Observed vs Estimated Standardized Index") +   
 theme\_minimal()



The above regression includes an interaction term PostScorecard\*HighEarning. This was used to capture the differential effect of the College Scorecard release on Google search activity between high-earning and low-earning colleges. The coefficient associated with this interaction term indicates whether the release of the College Scorecard had a significant impact on the difference in search activity between high-earning and low-earning colleges.

Based on the table, we can conclude that the introduction of the College Scorecard is associated with a decreased search activity on Google Trends for colleges with high-earning graduates by 0.1029 units relative to what it did for colleges with low-earning graduates, with a standard error of 0.0126. This result comes from the coefficient associated with the interaction term ‘PostScorecard \* HighEarning’ in my regression.

The first graph shows that the release of the scorecard had an effect of decreasing search activity on Google Trends for high-earning colleges of all types relative to low-earning colleges, as demonstrated for the years 2015 and 2016 in the Facet 1.

Overall, the introduction of the Scorecard had the effect of decreasing search activity on Google Trends for colleges with high-earning graduates relative to colleges with low-earning graduates.