Michael Vierela
OMSBA 5300
Seattle University
Data Exploration Assignment

Research Question

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?

Data Cleaning

I started the data cleaning process by first creating an object for a path to the location of data files, so it can easily be changed throughout all of the code with one change if someone wants to use the code from another location. I then read in all of the files using naming patterns to pull in all files and combined them all at once with import_list(), for efficiency.

After the Google Trends data was combined, the next step was to start manipulating the data to make it easier to group by date, school name, keywords, etc. I did this by creating a column for the first day of the week, converting the time frames into a string, pulling out the first 10 characters, populating the column with the new single dates, and converting them to a date format. Then I standardized the Google search indices by first grouping the school names and keywords and subtracting the index mean of those groupings from each index and dividing the difference by the standard deviation of those grouped indices. This makes the change in the indices more comparable because they're proportional to historical search activity. I wanted to work with data on a monthly basis, so I grouped the data by school names and created a new column for the average of the standardized indices for those groupings.

My next step was to combine the Google Trends data with the Scorecard data by merging them together with a data file linking them together by school name and UNITID, after removing duplicate school names from the linking data. I then filtered this data by variables related to the research question: colleges predominately granting bachelor's degrees (3 in the PREDDEG column), high-earnings colleges (those with median earnings of more than $75,000 in the md_earn_wne_p10-REPORTED-EARNINGS column), and low-earnings colleges (those with median earnings of less than $30,000 in the md_earn_wne_p10-REPORTED-EARNINGS column). I chose to use $75,000 and $30,000 as the categorical limits of high- and low-earnings colleges because the Scorecard data describe high-income families as earning more than $75,000 and low-income families as earning less than $30,000 and I wanted to keep it consistent.

After filtering the data to align with the research question, I decided to select the columns that I wanted to work with to solve this question; average standardized Google search indices (avgStdIndex) as the dependent variable; the month and year of the search indices (month_rounded) to analyze this data over time and create a dummy variable to indicate dates that occurred after the Scorecard data was released (afterRelease) to be used as an independent variable, and median earnings (md_earn_wne_p10-REPORTED-EARNINGS) to create a dummy variable for high- and low-earnings colleges (highEarnings) to be used as an independent variable.

After filtering out all of the NA, NULL, and suppressed data (labeled PrivacySuppressed), I created the dummy variables to indicate dates after the Scorecard was released (1 being after and 0 being before) and high- and low-earnings colleges (1 being high and 0 being low). I then used vtable() to make sure all of the data were formatted as the appropriate classes to use for analysis in code and noticed that these new dummy variables were numeric when they should have been categorical. I converted them to categorical using as.factor() and the data was cleaned and ready to use!

Graphical Analysis

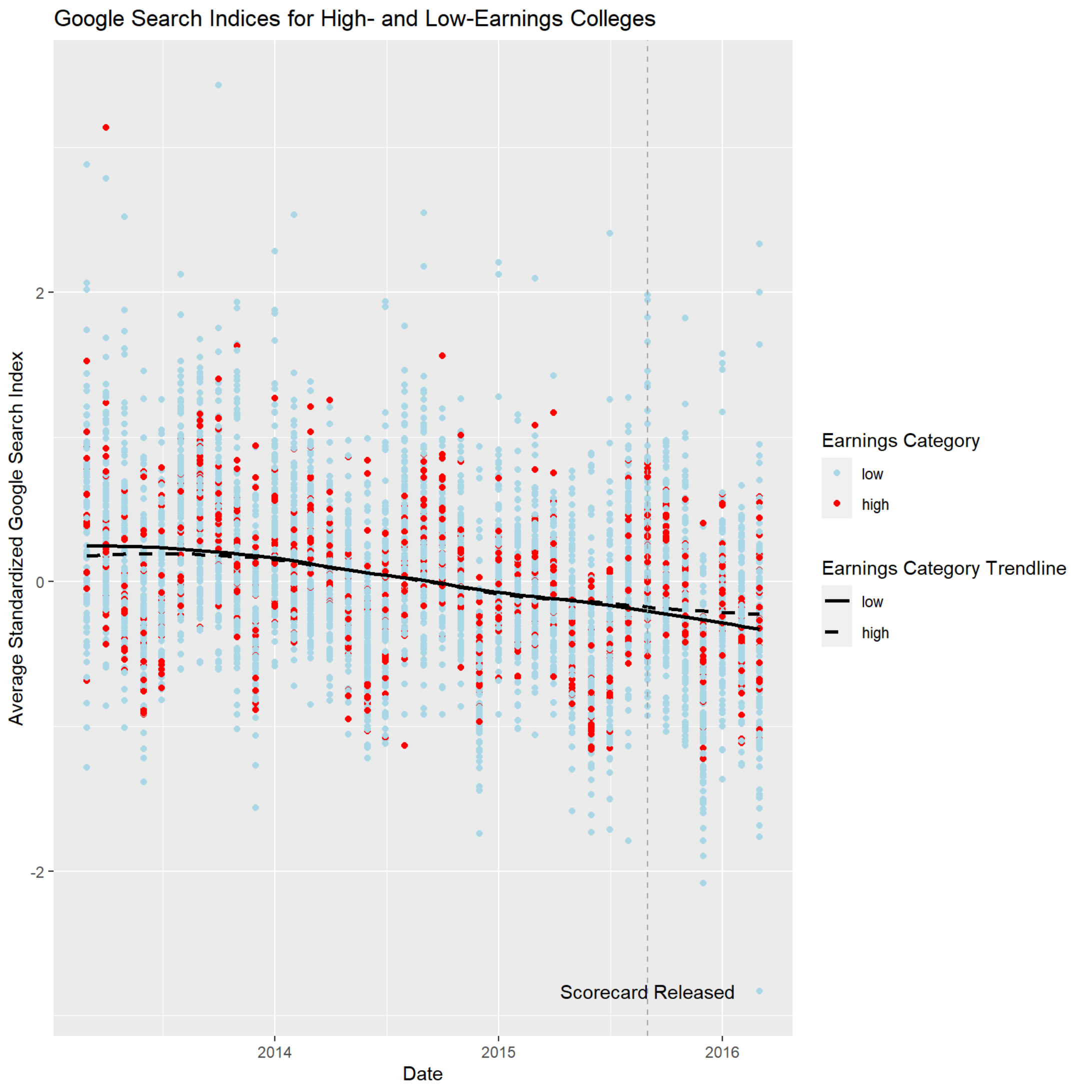


Figure 1

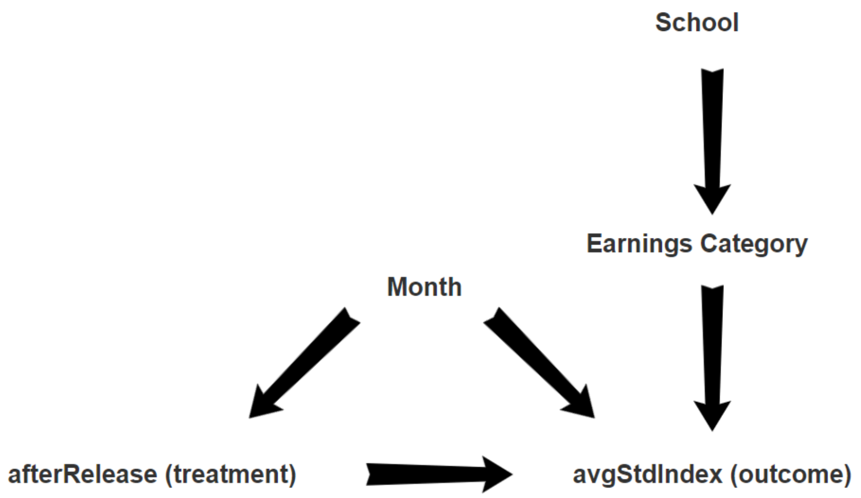
For my graphical analysis (Figure 1), I created a scatter plot and color-coded the points to distinguish between high- and low-earnings schools. I also plotted a regression line for each to show potential trends and patterns in the visual data. The scattered points for both high- and low-earnings colleges looked evenly distributed, with consistent high points and low points suggesting seasonality.

Further analysis of Figure 1 shows the regression lines of both declining prior to the Scorecard release date in September 2015, which is something to keep in mind for later analysis because it suggests that something else has been impacting search indices prior to the release of Scorecard data. The lines also start to diverge slightly after the release date of the Scorecard. I chose to use the 'loess' method instead of the 'lm' method in my plot because the research question is asking if searches were impacted by the release and I didn't want the visual representation to be lost within an presumed linear relationship.

Justification and Interpretation

After analyzing this plot, I wanted to create my own regression models, but first created a causal diagram (Figure 2) to determine if I should add any control variables to them.

Figure 2



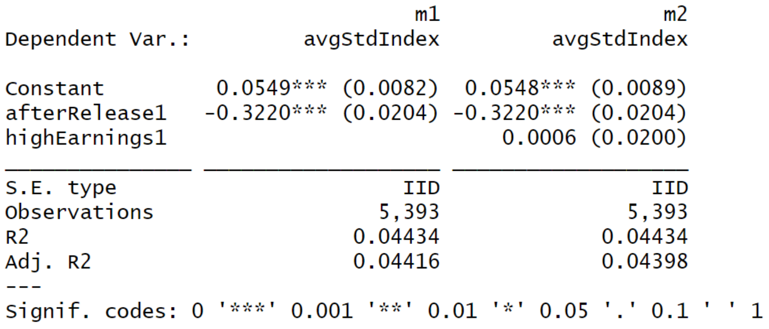
I started creating this causal diagram with the treatment variable (afterRelease) pointing toward search indices (outcome) because of an assumption that the release of Scorecard impacted student interest in colleges. I then set Month pointing toward the treatment and outcome because the date determines when the dummy variable is before or after the release date and seasonality of the months impact search indices for colleges in general because of college application timelines. With two variables left, I decided that neither the release date of Scorecard nor month of the year of college searches had any impact on the earnings of students 10 years after they graduated. I did, however, think that the earnings of students did impact whether other students were interested in certain colleges and that schools impacted the earning potential of students. After creating this diagram, I then evaluated potential paths:

Treatment → Outcome (good path) (front door path)
Treatment ← Month → Outcome (bad path) (back door path)
School → Earnings → Outcome (good path) (front door path)

Based on these paths, I knew that I had to control for Month because it impacted both the treatment variable and the outcome variable. I also decided to control for Earnings in order to answer the research question. Therefore, I created the following regression models (in R code):

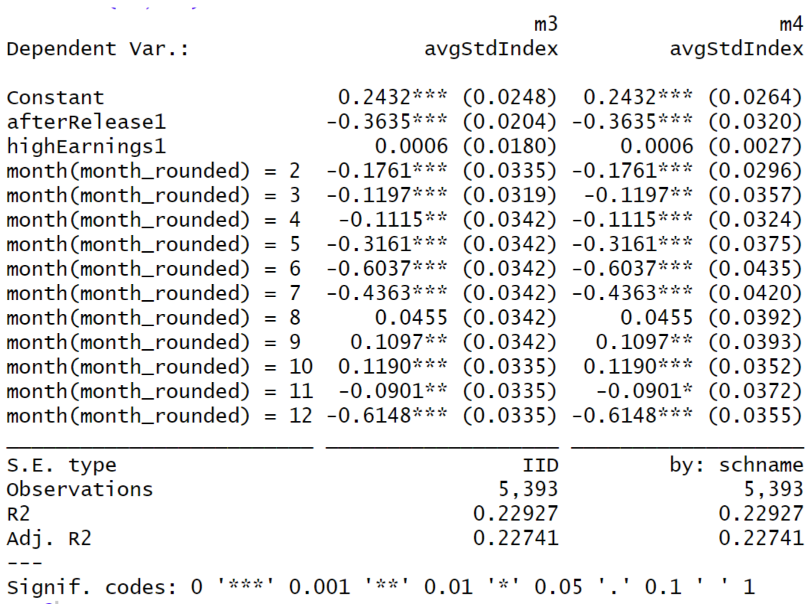
- m1 <- feols(avgStdIndex ~ afterRelease, data = gt_Scorecard_filtered)
- m2 <- feols(avgStdIndex ~ afterRelease + highEarnings, data = gt_Scorecard_filtered)
- m3 <- feols(avgStdIndex ~ afterRelease + highEarnings + i(month(month_rounded)), data = gt_Scorecard_filtered)
- m4 <- feols(avgStdIndex ~ afterRelease + highEarnings + i(month(month_rounded)), data = gt_Scorecard_filtered, vcov = ~schname)

I used Month as a categorical variable for model 3 and model 4 to accommodate for seasonality and also added School clustering to model 4 to include potential correlations between observations within the same school and to account for unobserved factors within the school level that might affect student interest (thinking back to the observed decline in college searches from Figure 1 - are schools providing less financial aid, or other factors?) Below are comparisons of each model:



```
                                 m1                  m2
Dependent Var.:         avgStdIndex         avgStdIndex

Constant         0.0549*** (0.0082)  0.0548*** (0.0089)
afterRelease1   -0.3220*** (0.0204) -0.3220*** (0.0204)
highEarnings1                           0.0006 (0.0200)
_______________ ___________________ ___________________
S.E. type                       IID                 IID
Observations                  5,393               5,393
R2                          0.04434             0.04434
Adj. R2                     0.04416             0.04398
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
                                          m3                m4
Dependent Var.:                  avgStdIndex       avgStdIndex

Constant                   0.2432*** (0.0248)  0.2432*** (0.0264)
afterRelease1             -0.3635*** (0.0204) -0.3635*** (0.0320)
highEarnings1                 0.0006 (0.0180)     0.0006 (0.0027)
month(month_rounded) = 2  -0.1761*** (0.0335) -0.1761*** (0.0296)
month(month_rounded) = 3  -0.1197*** (0.0319)  -0.1197** (0.0357)
month(month_rounded) = 4   -0.1115** (0.0342) -0.1115*** (0.0324)
month(month_rounded) = 5  -0.3161*** (0.0342) -0.3161*** (0.0375)
month(month_rounded) = 6  -0.6037*** (0.0342) -0.6037*** (0.0435)
month(month_rounded) = 7  -0.4363*** (0.0342) -0.4363*** (0.0420)
month(month_rounded) = 8      0.0455 (0.0342)     0.0455 (0.0392)
month(month_rounded) = 9    0.1097** (0.0342)   0.1097** (0.0393)
month(month_rounded) = 10  0.1190*** (0.0335)  0.1190*** (0.0352)
month(month_rounded) = 11  -0.0901** (0.0335)   -0.0901* (0.0372)
month(month_rounded) = 12 -0.6148*** (0.0335) -0.6148*** (0.0355)
_________________________ ___________________ ___________________
S.E. type                                 IID        by: schname
Observations                            5,393              5,393
R2                                    0.22927            0.22927
Adj. R2                               0.22741            0.22741
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As you'll notice, the coefficient for highEarnings in model 2 is extremely small, not statistically significant, and doesn't impact the constant or afterRelease coefficient much, if at all. The addition of the categorical month variables also impacts the constant and coefficient of afterRelease, and many of the months are statistically significant. The addition of the standard error adjustment for school clustering in model 4 doesn't impact any of the variables in model 3.

I also ran null hypothesis tests for highEarnings and a test of all three models containing the variable had p-values way above the .05 significance level, suggesting that we can't reject the null hypothesis and can likely remove the variable from the models without impacting them. Below are the results of the tests:

Wald test, H0: nullity of highEarnings1
stat = 8.219e-4, p-value = 0.97713, on 1 and 5,390 DoF, VCOV: IID.

Wald test, H0: nullity of highEarnings1
stat = 0.001074, p-value = 0.973854, on 1 and 5,379 DoF, VCOV: IID.

Wald test, H0: nullity of highEarnings1
stat = 0.047278, p-value = 0.827877, on 1 and 5,379 DoF, VCOV: Clustered (schname).

After analyzing graphs, building a causal diagram, constructing four models, and performing null hypothesis tests, I can confidently say that, although there's no evidence that the Scorecard shifts student interest to high-earnings colleges in relation to low-earnings colleges, there is evidence that the Scorecard has shifted student interest in other ways. Based on the variable coefficients, the introduction of the College Scorecard increased search activity on Google for colleges with high-earning graduates by only 0.0006 standard deviations relative to low-earning graduates with a standard error of 0.02. As mentioned previously, this coefficient is not statistically significant and does not pass a null hypothesis test and suggests there is likely no impact in student interest to high-earnings colleges. The afterRelease coefficient in model 3 and model 4, however, suggests that search activity after the release of the College Scorecard decreased search activity by 0.3635 standard deviations.