Data Exploration Project

Jomaica Alfiler

August 6, 2022

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)?

Data Preparation:

Data cleaning was required before running any analysis. The final goal data frame required Google trend data, school information, and College Scorecard data. Merging multiple files was required, as well as detecting and removing duplicate data.

Google trend data has been combined with an intermediary file containing school names and a key. This data frame was then joined to the College Scorecard data frame. During the process, duplicates were identified and removed, and the Scoreboard data was filtered for schools that primarily award Bachelor degrees.

Two variables were created from the index data. The first variable, "index," is a standardized index calculated using the keywords used to search for universities. The second variable, "mo index," is a monthly averaged index score for each university using the standardized index. Because index scores from Google Trends are relative, they cannot be compared to index scores from other entities. By standardizing these scores, we can compare index scores across schools.

There were two dummy variables created. The first variable, "post report," is used to identify data prior to the College Scoreboard's release. 0 represents data collected prior to 01-09-2015, and 1 represents data collected after that date. The second variable, "earning status," is based on the median reported earnings for students ten years after graduation for each school. Schools with values less than the median received a 0 while schools with values equal to or greater than the median received a 1. Due to a strong right hand skew brought on by high wages at predominantly medical schools, the median value was chosen.

Analysis:

In essence, the research question is asking what impact, if any, did the release of the College Scorecard have on Google search trends for reported alumni earnings. It would be logical to predict that schools with higher reported earnings would also have higher search index scores.

To answer this question, the variables must be determined. The variable of interest is the search index score. In this case, the index is the dependent variable since it is the variable that needs to be predicted. College Scorecard average 10 year earnings is the independent variable being used to predict the index score.

The research question is concerned with schools and not with keywords associated with them. Based on this, I believe a monthly index score should be calculated for each school based on the average index score for all associated keywords.

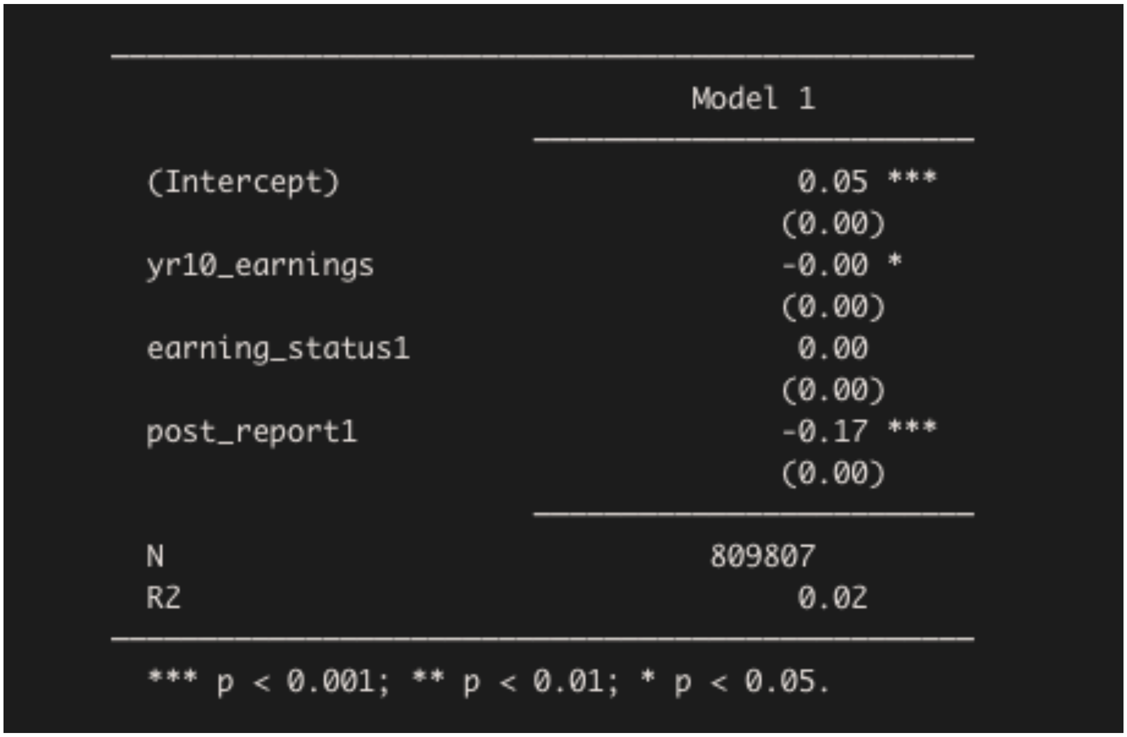
Also, we are interested in a possible relationship between low and high earning schools, so this variable will be treated as a dummy variable with 0 indicating a low earnings school and 1 indicating a high earning school.

In order to compare the index scores, we must also take into account the College Scorecard release date. This is included as a dummy variable with 0 being pre and 1 being post 01-09-2015.

As a result of this regression, we will be able to determine how reported earnings affect search index scores, and what the effect of high reported income is on the release of the College Scorecard will be.

This is the described regression:

monthly index = b0 + b1(earnings) + b2(earning status) + b3(scorecard release)



```
=================================================
                            Model 1
                        -------------------------
  (Intercept)                  0.05 ***
                              (0.00)
  yr10_earnings               -0.00 *
                              (0.00)
  earning_status1              0.00
                              (0.00)
  post_report1                -0.17 ***
                              (0.00)
                        -------------------------
  N                       809807
  R2                           0.02
=================================================
  *** p < 0.001; ** p < 0.01; * p < 0.05.
```

Figure 1: plot results of effect on `yr10_earnings`.

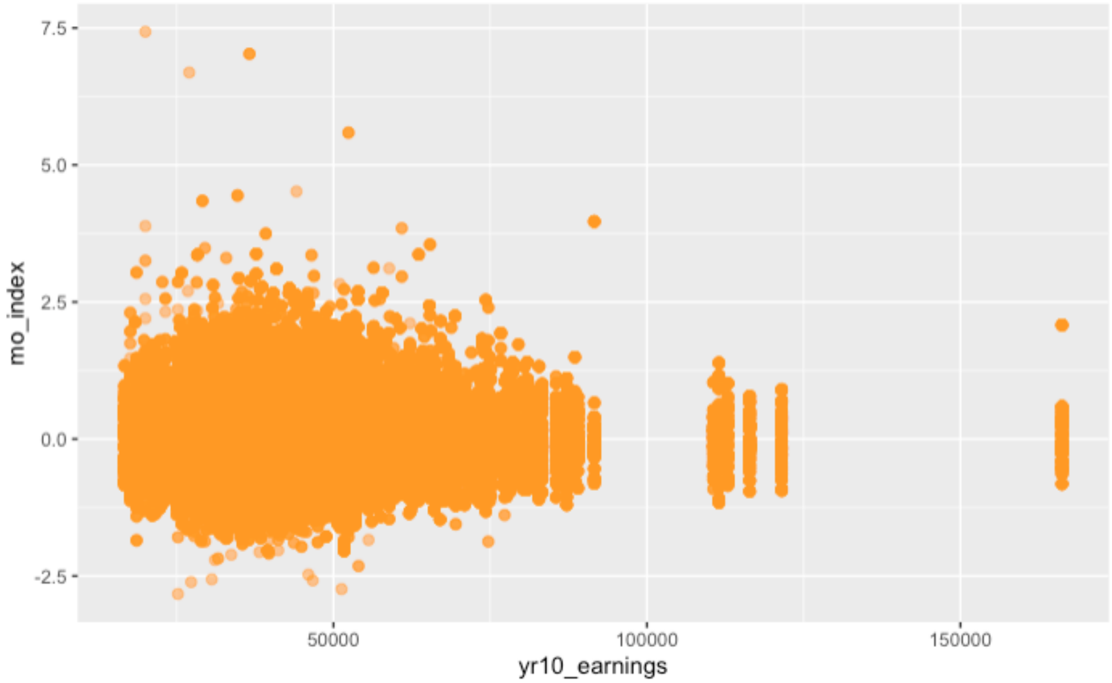


Figure 2: plot results of effect on `earning_status`.

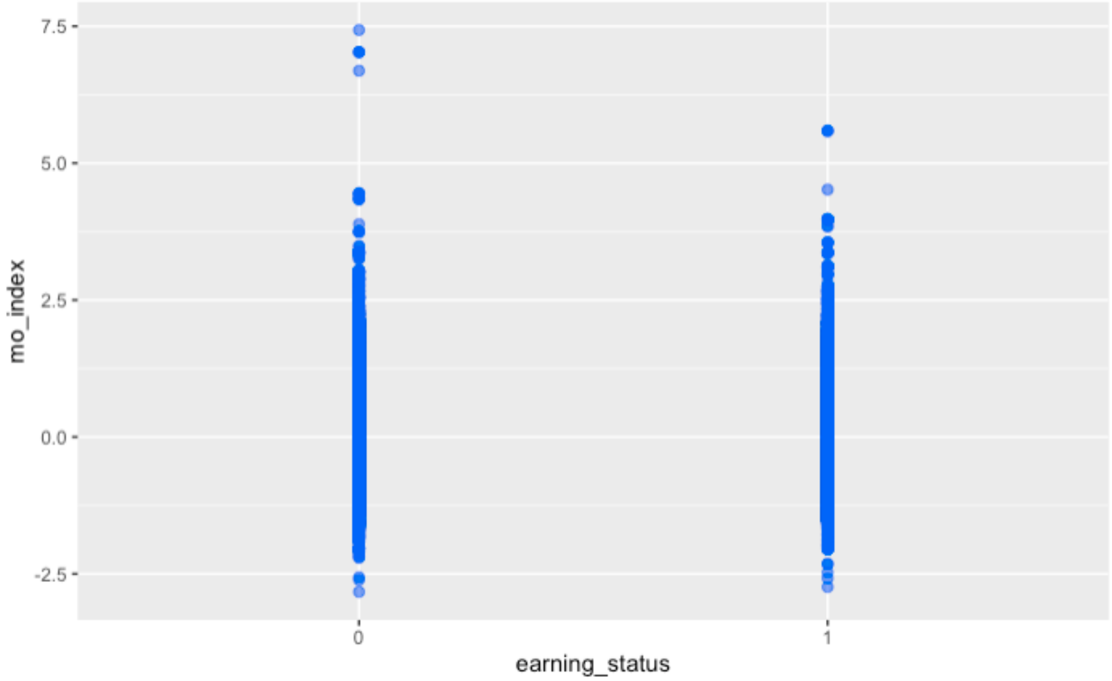
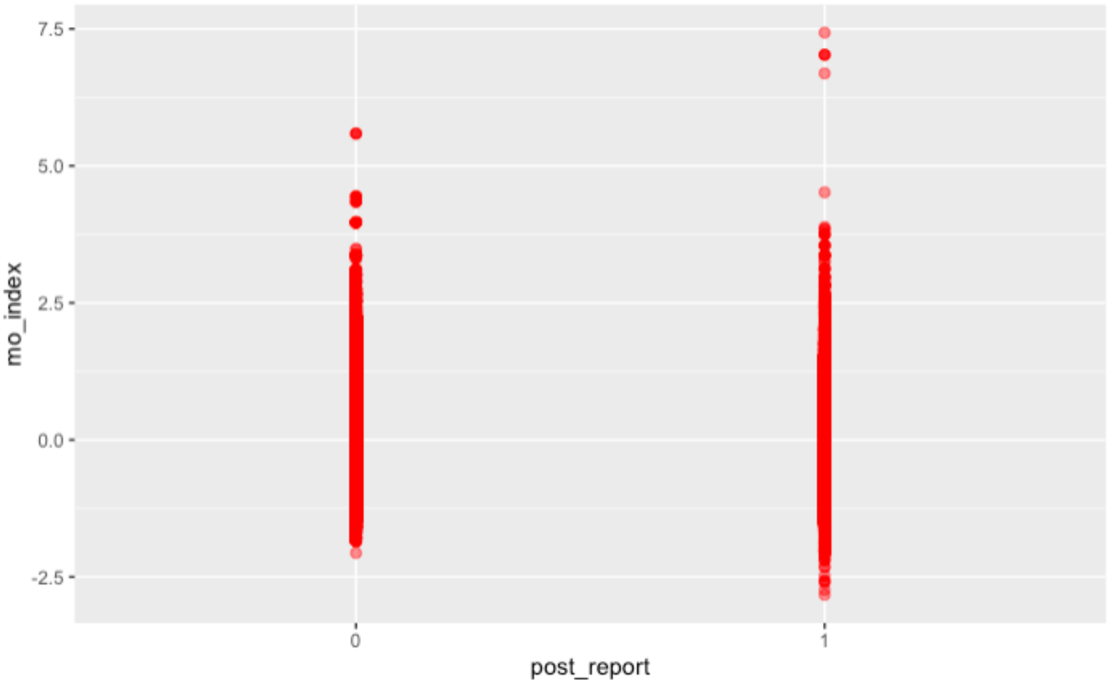


Figure 3: plot results of effect on `post_report`.



Results and Conclusion:

The monthly index scores and the reported average wages don't seem to be related in any way *(Shown in Figure 1)*. The data points do not appear to exhibit a linear relationship. What is visible is the right hand skew brought on by a few medical colleges.

Additionally, there is no connection between the high/low income categories and monthly index scores. It's interesting to see that a small number of outliers with high index scores are among the low earners.

A statistically significant inverse link exists between index scores and the release of the College Scorecard. Upon the release of the College Scorecard, index scores dropped by 0.17 units. The significance level for this relationship is more than 99%.

Additionally, there is no connection between the high/low income categories and monthly index scores. It's interesting to see that a small number of outliers with high index scores are among the low earners.

A statistically significant inverse link exists between index scores and the release of the College Scorecard. Upon the release of the College Scorecard, index scores dropped by 0.17 units. The significance level for this relationship is more than 99%.

Also, low earning is based on schools falling bellow the median reported earnings.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$11,800	$35,100	$40,700	$42,159	$47,800	$166,200

These findings are all surprising since they differ from what I had anticipated. I anticipated that students aiming to graduate from high-earning schools would result in an increase in index scores. Due to a potential rise in media interest and social awareness, I also anticipated an increase in index scores after the release of the scorecard. The College Scorecard's links to school websites and supplementary information could be one reason, eliminating the need for people to use search engines to get what they're looking for.