

Megan Mixon

Econ 5300

18 February, 2024

Data Translation Write-Up

In this analysis, the research question is: Among colleges that predominantly grant bachelor's degrees, did the release of the Scorecard information shift student interest to high-earnings colleges relative to low-earnings ones (as proxied by Google searches for keywords associated with those colleges)?

Though the Scorecard only provides information about the median earnings of graduates ten years after graduation for each college, we are going to define "high-earning" and "low-earning" colleges by setting a threshold to 75%. I chose to do this because I thought it would show the most accurate representation of the comparison between high and low earnings. I then created a dummy variable to determine if the median earnings of the colleges are high or low, based on the threshold.

To examine more thoroughly a significant change or shift in student interest to high-earnings colleges relative to low-earnings colleges, I created a variable for a before and after time period of when the Scorecard was released. Due to the Scorecard being released at the beginning of September 2015, I measured time as a binary variable, making the After_Scorecard release True and anything before the first of September 2015 False.

I also created a variable for cost to see the cost of each college and later compare it to median earnings. I felt it would be beneficial to see a representation of how much a college costs relative to the high and low earnings of the college's graduates. My thought was that it may shed some light on why a student chooses to go to a particular college or why their interest shifts to one college versus another college, even though this variable does not fully contribute to my answer for the research question.

For my first regression(model1), I chose to regress `standardized_index(Y)` on `After_ScoreCard(X)` and `median_earn` because I wanted to observe how the standardized index(the outcome) measures the shift in the student's interest on high-earnings colleges relative to low-earnings colleges after the scorecard was released. The `After_ScoreCard` is the independent variable in the regression because it shows what the effect of X on Y is relative to the high and low earnings of the colleges.

For my second regression(model2), I chose to regress `standardized_index(Y)` on `After_ScoreCard(X)` and `median_earn(Z)`, including an interaction term($X*Z$) because I believe

this would thoroughly answer the research question, showing whether the student's interest in higher earning colleges relative to lower earning colleges shifted at a particular time. I made After_ScoreCard and median_earn into an interaction term. Going off of the equation,

$$Y = B_0 + B_1X + B_2Z + B_3XZ$$

Z tells us how our prediction of Y changes when Z is increased by one unit. The coefficient of median_earn tells us how the prediction changes when median earnings of colleges increase by one unit, showing how much stronger the effect of X on Y gets when the median earnings increases by one unit. The interaction term is already a binary term, where After_ScoreCard = 1, and 0 otherwise. So to get the effect of X on Y for median_earn, we plug in median_earn = 1. As a result, we get the difference in the effect of X on Y between high and low earnings after the scorecard was released, thus showing how much stronger the effect of the scorecard was for searches for higher earning colleges relative to lower earning colleges after the scorecard was released.

In conclusion, the introduction of the College Scorecard decreased search activity on Google Trends for colleges with high-earning graduates by a 4.14 standard-deviation change in search interest relative to what it did for colleges with low-earning graduates, with a standard error of 2.54e-7. This result comes from the standardized_index regressed on After_ScoreCardTrue and median_earn with an interaction coefficient in my second regression. When examining the first regression model, we can see that the coefficients are negative, meaning that there is a 4.9 decrease in the standard deviations of searches for high earnings colleges relative to low earnings colleges after the scorecard was released. However, by using an interaction term, the second regression model presents the difference in the effect of X on Y between high and low earnings after the scorecard was released, showing that the release of the scorecard shifted student's interest, through search activity, away from high earning colleges relative to low earning colleges.

Based on my results, we should conclude that the reason why the release of the ScoreCard shifted student's interest away from high-earnings colleges is due to the amount of information the ScoreCard gave students, mainly regarding the cost of the college that then dissuaded students from further searching that college. Though a certain college had higher median earnings for graduates, once the ScoreCard showed that the cost of that college was very expensive, it likely made students uninterested in searching for that college. It's also plausible that the ScoreCard showed information about higher earning colleges having a lower acceptance rate, thus dissuading students from applying to those colleges.

By plotting the median earnings to unitid of colleges that predominantly grant bachelor's degrees, we can see a visual comparison of the high-earnings schools to low-earnings schools. This is relevant because it gives us a better idea of how colleges differ by graduate median

earnings across colleges. In this graph, I have the unitid on the x-axis and the median earnings on the y-axis. This graph shows the colleges with lower median earnings are more prevalent in the data than high median earning colleges. This may be the case because colleges with higher median earnings are more expensive than low median earning colleges; therefore, more students are interested in colleges with lower median earnings.

By plotting the median earnings to cost of colleges that predominantly grant bachelor's degrees, we can see a visual comparison of the median earnings of schools relative to the cost of attending that college. In this graph, I have the cost on the x-axis and the median earnings on the y-axis. This graph shows that the majority of median earnings are lower relative to lower costing colleges, with the exception of a few outliers. This may be because more students would rather attend colleges that cost less despite the median earnings a graduate makes from that particular college.

Note that when trying to graph `standardized_index` to `median_earnings`, we get a very noisy graph, as most of the variation has been taken out, resulting in a messy representation of median earnings to searches. Additionally, when trying to graph `standardized_index` on y-axis and `After_ScoreCard` on x-axis, it did not represent the data clearly as it presented two vertical lines with multiple points on them, due to time being a binary variable.