Data Exploration

Bao Do

2024-02-12

## Introduction:

The analysis aims to investigate whether the release of the College Scorecard in September 2015 influenced student interest, as proxied by Google search activity, in colleges with high-earning graduates relative to those with low-earning graduates among institutions that predominantly grant bachelor’s degrees. To address this research question, a differences-in-differences regression analysis with fixed effects was performed.

## Data Cleaning

### Read Google Trends Data

The initial step in this data cleaning process involves importing Google Trends data from multiple files and combining them into a unified dataset. Using the list of file names obtained from the specified directory, the data is read and concatenated.

#Load dataset   
directory\_path <- "C:/Users/dob22/OneDrive/Desktop/Spring 2024/Econometrics/Data\_Exploration\_Rawdata/Lab3\_Rawdata"  
  
#Read Google Trends Data  
file\_names <- list.files(directory\_path, pattern = "trends\_up\_to\_.+\\.csv", full.names = TRUE)  
  
#Read the files and bind them together  
trends\_data <- import\_list(file\_names, rbind = TRUE)

## Warning in (function (input = "", file = NULL, text = NULL, cmd = NULL, :  
## Stopped early on line 1562. Expected 6 fields but found 5. Consider fill=TRUE  
## and comment.char=. First discarded non-empty line: <<11,yti career institute -  
## york,yti.edu,2,>>

## Warning in (function (input = "", file = NULL, text = NULL, cmd = NULL, :  
## Stopped early on line 1095. Expected 6 fields but found 5. Consider fill=TRUE  
## and comment.char=. First discarded non-empty line: <<9,heidelberg  
## university,heidelberg.edu,2,>>

## Warning in (function (input = "", file = NULL, text = NULL, cmd = NULL, :  
## Stopped early on line 1094. Expected 6 fields but found 5. Consider fill=TRUE  
## and comment.char=. First discarded non-empty line: <<8,mount vernon nazarene  
## university,mvnu.edu,2,>>

## Warning in (function (input = "", file = NULL, text = NULL, cmd = NULL, :  
## Stopped early on line 3280. Expected 6 fields but found 5. Consider fill=TRUE  
## and comment.char=. First discarded non-empty line: <<41,potomac state college  
## of west virginia university,potomac state college of west virginia  
## university,1,>>

### Aggregating the Google Trends data

The next step involves aggregating the Google Trends data to a monthly level. This includes extracting the first ten characters from the ‘monthorweek’ variable, converting it into a date format, and then further aggregating the data by month.

#Get the first ten characters out of the monthorweek variable and aggregate the dates to months  
trends\_data$monthorweek <- str\_sub(trends\_data$monthorweek, 1, 10)  
  
trends\_data$date <- ymd(trends\_data$monthorweek)  
trends\_data$month <- floor\_date(trends\_data$date, unit = "month")

Subsequently, the index variable is standardized by school name and keyword, and the standardized index variable is aggregated to the keyword-month level.

#Standardize index variable by school name and keyword  
trends\_data <- trends\_data %>%  
 group\_by(schname, keyword) %>%  
 mutate(standardized\_index = (index - mean(index, na.rm = TRUE)) / sd(index, na.rm = TRUE))  
  
#Aggregate standardized index variable to the keyword-month level  
aggregated\_data <- trends\_data %>%  
 group\_by(keyword, month) %>%  
 summarize(mean\_standardized\_index = mean(standardized\_index, na.rm = TRUE))

## `summarise()` has grouped output by 'keyword'. You can override using the  
## `.groups` argument.

#print(aggregated\_data)

### Reading in the Scorecard data

Shifting to the Scorecard data, relevant files are read, and operations are performed to join and filter the data.

scorecard\_file <- "C:/Users/dob22/OneDrive/Desktop/Spring 2024/Econometrics/Data\_Exploration\_Rawdata/Lab3\_Rawdata/Most+Recent+Cohorts+(Scorecard+Elements).csv"  
dictionary\_file <- "C:/Users/dob22/OneDrive/Desktop/Spring 2024/Econometrics/Data\_Exploration\_Rawdata/Lab3\_Rawdata/CollegeScorecardDataDictionary-09-08-2015.csv"  
id\_name\_link\_file <- "C:/Users/dob22/OneDrive/Desktop/Spring 2024/Econometrics/Data\_Exploration\_Rawdata/Lab3\_Rawdata/id\_name\_link.csv"  
  
#Read in the Scorecard data  
scorecard <- import(scorecard\_file)  
  
#Read in the id\_name\_link file  
id\_name\_link <- import(id\_name\_link\_file)

### Merging the Scorecard data

#count how many times each school name shows up in id\_name\_link  
inl\_count <- id\_name\_link %>%  
 group\_by(schname) %>%  
 mutate(n = n()) %>%  
 filter(n == 1) %>%  
 select(-n)  
  
#Joining data  
joined\_data <- inner\_join(trends\_data, inl\_count, by = "schname")  
  
#Rename columns in joined\_data to match the case in scorecard  
colnames(joined\_data) <- toupper(colnames(joined\_data))  
  
#Join joined\_data and scorecard and delete duplicates  
final\_dataset <- inner\_join(joined\_data, scorecard, by = "OPEID")

## Warning in inner\_join(joined\_data, scorecard, by = "OPEID"): Detected an unexpected many-to-many relationship between `x` and `y`.  
## ℹ Row 45972 of `x` matches multiple rows in `y`.  
## ℹ Row 1073 of `y` matches multiple rows in `x`.  
## ℹ If a many-to-many relationship is expected, set `relationship =  
## "many-to-many"` to silence this warning.

final\_dataset <- final\_dataset %>%  
 group\_by(SCHNAME) %>%  
 filter(n\_distinct(SCHID) == 1) %>%  
 ungroup()

## Analysis

Before conducting analysis, the final dataset was filtered to only include institutions that predominantly grant bachelor’s degrees. Subsequently, the reported earnings column was examined to identify the highest and lowest values.

#Filter final dataset to bachelor's degree granting schools  
bachelors\_only <- filter(final\_dataset, PREDDEG == 3)  
  
#Convert md\_earn\_wne\_p10-REPORTED-EARNINGS to numeric and remove NULL values  
bachelors\_only$`md\_earn\_wne\_p10-REPORTED-EARNINGS` <- as.numeric(bachelors\_only$`md\_earn\_wne\_p10-REPORTED-EARNINGS`)

## Warning: NAs introduced by coercion

bachelors\_only <- bachelors\_only[!is.na(bachelors\_only$`md\_earn\_wne\_p10-REPORTED-EARNINGS`), ]  
  
# Find the highest value and lowest value  
highest\_value <- max(bachelors\_only$`md\_earn\_wne\_p10-REPORTED-EARNINGS`, na.rm = TRUE)  
print(highest\_value)

## [1] 74700

lowest\_value <- min(bachelors\_only$`md\_earn\_wne\_p10-REPORTED-EARNINGS`, na.rm = TRUE)  
print(lowest\_value)

## [1] 17600

High-earning and low-earning colleges were then defined by dividing the range of median earnings into three equal parts, with the top one-third representing high-earning colleges and the bottom one-third representing low-earning colleges. This section was done manually.

Range = Highest Value (74700) - Lowest Value (17600) = 57100 Width of each range = Range (57100) / Width of each range (3) = 19033 Low Earnings: 17600 to 36633 Medium Earnings: 36633 to 55666 High Earnings: 55666 to 74700

#Define high vs low earning colleges  
bachelors\_only$HIGH\_EARNINGS <- ifelse(bachelors\_only$`md\_earn\_wne\_p10-REPORTED-EARNINGS` > 55000, 1, 0)  
  
#sumtable(bachelors\_only)  
  
#Print to check the results  
print(unique(bachelors\_only$HIGH\_EARNINGS))

## [1] 0 1

The region variable was then created by mapping states to their overarching Census region. This grouping accounts for shared unobserved factors influencing colleges within geographic proximity and facilitates making broader conclusions.

#Aggregate up to the region-month level  
bachelors\_only <- bachelors\_only %>%  
 mutate(region = case\_when(  
 STABBR %in% c("CT", "ME", "MA", "NH", "RI", "VT") ~ "New England",  
 STABBR %in% c("DE", "DC", "MD", "NJ", "NY", "PA") ~ "Mid East",  
 STABBR %in% c("IL", "IN", "MI", "OH", "WI") ~ "Great Lakes",  
 STABBR %in% c("IA", "KS", "MN", "MO", "NE", "ND", "SD") ~ "Plains",  
 STABBR %in% c("AL", "AR", "FL", "GA", "KY", "LA", "MS", "NC", "SC", "TN", "VA", "WV") ~ "Southeast",  
 STABBR %in% c("AZ", "NM", "OK", "TX") ~ "Southwest",  
 STABBR %in% c("CO", "ID", "MT", "UT", "WY") ~ "Rocky Mountains",  
 STABBR %in% c("AK", "CA", "HI", "NV", "OR", "WA") ~ "Far West",  
 TRUE ~ "Other"  
 ))  
  
regional\_trends <- bachelors\_only %>%   
 group\_by(region, MONTH,`md\_earn\_wne\_p10-REPORTED-EARNINGS`,HIGH\_EARNINGS) %>%  
 summarise(MEAN\_STANDARDIZED\_INDEX = mean(STANDARDIZED\_INDEX))

## `summarise()` has grouped output by 'region', 'MONTH',  
## 'md\_earn\_wne\_p10-REPORTED-EARNINGS'. You can override using the `.groups`  
## argument.

The PRE\_POST\_POLICY variable was created to have an indicator separating the periods before and after September 2015 when the College Scorecard website launched. This variable takes the value “Pre” for months before September 2015 and “Post” for months after the launch date. This binary coding allows for a numerical comparison of the periods before and after the policy implementation, rather than relying solely on categorical labels.

Next, the code employs a differences-in-differences regression approach to assess the differential impact of the College Scorecard introduction on search activity between high-earning colleges (treatment group) and low-earning colleges (control group).

The first model, basic, is a simple regression of MEAN\_STANDARDIZED\_INDEX on md\_earn\_wne\_p10-REPORTED-EARNINGS. This baseline model serves as a reference point.The main model of interest, fe\_model, incorporates the difference-in-differences specification. It includes the interaction term between md\_earn\_wne\_p10-REPORTED-EARNINGS and PRE\_POST\_POLICY, as well as fixed effects for region and the HIGH\_EARNINGS indicator variable. The interaction term captures the differential impact of the College Scorecard introduction on search activity for colleges with varying levels of post-graduation earnings.

The coefficient on the interaction term md\_earn\_wne\_p10-REPORTED-EARNINGS \* PRE\_POST\_POLICY represents the difference-in-difference estimate, quantifying the change in search index trends for colleges with higher earnings relative to those with lower earnings, after the policy implementation compared to before. This is the key estimate of interest, as it isolates the effect of the College Scorecard release on search activity for high-earnings colleges relative to low-earnings colleges.

Finally, by including region fixed effects (factor(region)), the model controls for time-invariant regional factors that may influence search activity. The HIGH\_EARNINGS variable, which may be an indicator for colleges above a certain earnings threshold, is also included as a control variable.

#Create an indicator variable for the pre/post policy period  
regional\_trends <- regional\_trends %>%  
 mutate(PRE\_POST\_POLICY = ifelse(MONTH < as.Date("2015-09-01"), "Pre", "Post"))  
  
#Run differences-in-differences regression with fixed effects  
basic <- feols(MEAN\_STANDARDIZED\_INDEX ~ `md\_earn\_wne\_p10-REPORTED-EARNINGS`, data = regional\_trends)

## NOTE: 36 observations removed because of NA values (LHS: 36).

fe\_model <- feols(MEAN\_STANDARDIZED\_INDEX ~ `md\_earn\_wne\_p10-REPORTED-EARNINGS` \* PRE\_POST\_POLICY +   
 factor(region) + HIGH\_EARNINGS , data = regional\_trends)

## NOTE: 36 observations removed because of NA values (LHS: 36, RHS: 16).

#Print summary an etable results  
summary(fe\_model)

## OLS estimation, Dep. Var.: MEAN\_STANDARDIZED\_INDEX  
## Observations: 3,730   
## Standard-errors: IID   
## Estimate Std. Error  
## (Intercept) -0.23805460 0.10387707  
## `md\_earn\_wne\_p10-REPORTED-EARNINGS` -0.00000197 0.00000258  
## PRE\_POST\_POLICYPre 0.27061574 0.10635838  
## factor(region)Great Lakes -0.00002608 0.04302656  
## factor(region)Mid East 0.00620016 0.04997619  
## factor(region)New England 0.01991699 0.05801636  
## factor(region)Other 0.03815833 0.06738153  
## factor(region)Plains 0.00971634 0.04227658  
## factor(region)Rocky Mountains -0.00349260 0.06607236  
## factor(region)Southeast 0.00358101 0.03885235  
## factor(region)Southwest -0.00775350 0.06084509  
## HIGH\_EARNINGS -0.01184008 0.06582175  
## `md\_earn\_wne\_p10-REPORTED-EARNINGS`:PRE\_POST\_POLICYPre 0.00000312 0.00000274  
## t value Pr(>|t|)   
## (Intercept) -2.291695 0.021979 \*   
## `md\_earn\_wne\_p10-REPORTED-EARNINGS` -0.763609 0.445149   
## PRE\_POST\_POLICYPre 2.544376 0.010987 \*   
## factor(region)Great Lakes -0.000606 0.999516   
## factor(region)Mid East 0.124062 0.901273   
## factor(region)New England 0.343300 0.731392   
## factor(region)Other 0.566303 0.571222   
## factor(region)Plains 0.229828 0.818238   
## factor(region)Rocky Mountains -0.052860 0.957846   
## factor(region)Southeast 0.092170 0.926568   
## factor(region)Southwest -0.127430 0.898607   
## HIGH\_EARNINGS -0.179881 0.857256   
## `md\_earn\_wne\_p10-REPORTED-EARNINGS`:PRE\_POST\_POLICYPre 1.140582 0.254117   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## RMSE: 0.707116 Adj. R2: 0.040805

etable(basic,fe\_model)

## basic  
## Dependent Var.: MEAN\_STANDARDIZED\_INDEX  
##   
## Constant -0.0044 (0.0423)  
## `md\_earn\_wne\_p10-REPORTED-EARNINGS` 3.37e-7 (1.09e-6)  
## PRE\_POST\_POLICYPre   
## factor(region)GreatLakes   
## factor(region)MidEast   
## factor(region)NewEngland   
## factor(region)Other   
## factor(region)Plains   
## factor(region)RockyMountains   
## factor(region)Southeast   
## factor(region)Southwest   
## HIGH\_EARNINGS   
## `md\_earn\_wne\_p10-REPORTED-EARNINGS` x PRE\_POST\_POLICYPre   
## \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
## S.E. type IID  
## Observations 3,730  
## R2 2.57e-5  
## Adj. R2 -0.00024  
##   
## fe\_model  
## Dependent Var.: MEAN\_STANDARDIZED\_INDEX  
##   
## Constant -0.2381\* (0.1039)  
## `md\_earn\_wne\_p10-REPORTED-EARNINGS` -1.97e-6 (2.58e-6)  
## PRE\_POST\_POLICYPre 0.2706\* (0.1064)  
## factor(region)GreatLakes -2.61e-5 (0.0430)  
## factor(region)MidEast 0.0062 (0.0500)  
## factor(region)NewEngland 0.0199 (0.0580)  
## factor(region)Other 0.0382 (0.0674)  
## factor(region)Plains 0.0097 (0.0423)  
## factor(region)RockyMountains -0.0035 (0.0661)  
## factor(region)Southeast 0.0036 (0.0389)  
## factor(region)Southwest -0.0078 (0.0609)  
## HIGH\_EARNINGS -0.0118 (0.0658)  
## `md\_earn\_wne\_p10-REPORTED-EARNINGS` x PRE\_POST\_POLICYPre 3.12e-6 (2.74e-6)  
## \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
## S.E. type IID  
## Observations 3,730  
## R2 0.04389  
## Adj. R2 0.04081  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

In the basic model, neither the median 10 year earnings variable (md\_earn\_wne\_p10-REPORTED-EARNINGS) nor the constant are statistically significant. This basic specification also exhibits no explanatory power, with an R-squared of essentially 0.

The second model includes an indicator for the pre/post September 2015 period when the College Scorecard website launched, in addition to census region factors. Here, the policy variable (PRE\_POST\_POLICYPre) is positively associated with the outcome at the p<0.05 level, suggesting some increase in search activity after September 2015 on average. The expanded model also fits better, now explaining over 4% of variation. In addition, this second model incorporates the key difference-in-differences interaction between college earnings levels and pre/post indicators. However, the interaction term is not statistically significant. Furthermore, the regional factors yield small coefficients, implying limited explanatory region-specific variation relative to reference areas.

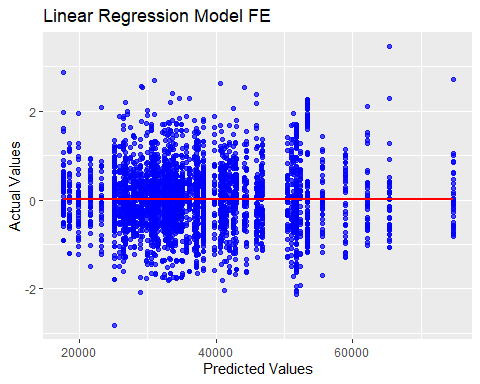
Overall, even with the added earnings indicator, the model indicates no significant evidence that the introduction of the College Scorecard website substantively reallocated attention towards high earning institutions among bachelor’s colleges. Neither the small double interaction term nor the individual earning predictor show significant relationships in shifting search trends.

#Plot the linear model  
ggplot(regional\_trends, aes(x = `md\_earn\_wne\_p10-REPORTED-EARNINGS`, y = MEAN\_STANDARDIZED\_INDEX)) +  
 geom\_point(color = "blue", alpha = 0.7) +  
 geom\_smooth(method = "lm", se = FALSE, color = "red") +  
 labs(title = "Linear Regression Model FE",  
 x = "Predicted Values",  
 y = "Actual Values")

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

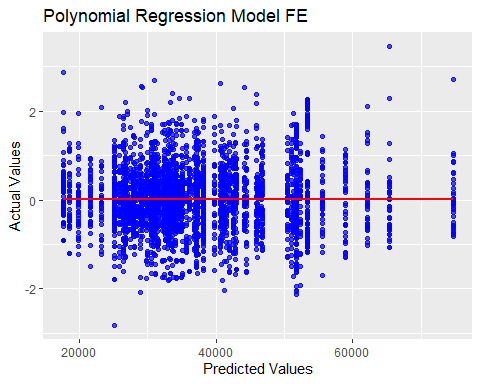
## Warning: Removed 36 rows containing non-finite values (`stat\_smooth()`).

## Warning: Removed 36 rows containing missing values (`geom\_point()`).



#Plot the polynomial regression model  
ggplot(regional\_trends, aes(x = `md\_earn\_wne\_p10-REPORTED-EARNINGS`, y = MEAN\_STANDARDIZED\_INDEX)) +  
 geom\_point(color = "blue", alpha = 0.7) +  
 geom\_smooth(method = "lm", formula = y ~ poly(x, degree = 3), se = FALSE, color = "red") +  
 labs(title = "Polynomial Regression Model FE",  
 x = "Predicted Values",  
 y = "Actual Values")

## Warning: Removed 36 rows containing non-finite values (`stat\_smooth()`).  
## Removed 36 rows containing missing values (`geom\_point()`).



Based on the scatterplots above, both the linear and polynomial regression visualizations appear nearly identical. Both graphs plot median 10 year college graduate earnings on the x-axis against average monthly search popularity on the y-axis, with each blue point representing an observation at the region-month level. The linear fit line minimizes squared errors assuming a straight-line association between earnings and searches. Meanwhile, the polynomial expands this to test for potential nonlinearities, fitting a more flexible cubic curve.

Additionally, both graphs exhibit a slight positive slope, indicating that the search index tends to increase with higher median graduate pay. However, this relationship is accompanied by considerable variability in the data points.Furthermore, the high degree of overlap implies adding nonlinear flexibility does not provide significantly improved model fit from a purely linear specification. The earnings-attention connection holds mostly steady rather than dramatically accelerating or declining at certain salary thresholds. As a result, neither model version indicates clear shifting attention toward high-paying schools after policy rollout either—the displayed graduating earnings connection itself remains marginal.

## Conclusion

The difference-in-differences model estimate suggests that after vs before the Scorecard rollout, the standardized search index increased by 0.00000312 units more for high earning colleges compared to low earning ones. However, this differential gain is negligible in practical terms and not statistically significant based on the standard error.

Similarly, the graphical analyses do not demonstrate strong evidence of diverging trends between high and low earning institutions across regions when examining search popularity before and after September 2015. While some areas saw slightly greater gains for higher earning schools post-policy, this pattern does not clearly or consistently emerge across locations. The association between graduate earnings and search interests remains positive but small in magnitude over time.

Therefore, based on the available data and methodology, there is insufficient support to conclude that the introduction of the College Scorecard website substantially shifted attention towards bachelor’s-level colleges with high-earning graduates relative to those with low-earning graduates.