Data Exploration

Anna Smalley

## Quarto

Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see <https://quarto.org>.

Installing libraries:

library(rio)  
library(fixest)  
library(ggplot2)  
library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.1 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.0  
✔ lubridate 1.9.2 ✔ tibble 3.2.1  
✔ purrr 1.0.1 ✔ tidyr 1.3.0  
── 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(vtable)

Loading required package: kableExtra  
  
Attaching package: 'kableExtra'  
  
The following object is masked from 'package:dplyr':  
  
 group\_rows

library(lubridate)  
library(dplyr)

Import data

Data<- import\_list(list.files(pattern = "trends\_up\_to", path = "~/Econ/Econometrics",full.names = TRUE), rbind = TRUE, fill = TRUE)

Getting date data

Data <- Data %>%  
 mutate(week = str\_sub(monthorweek, start = 1, end = 10)) %>%  
 mutate(week = ymd(week)) %>%  
 mutate(month = floor\_date(week, unit = "month"))

Aggregate

Data <- Data %>%  
 group\_by(schname, keyword) %>%  
 mutate(index = (index-mean(index)-sd(index)))

Import

Cohorts <- import('Most+Recent+Cohorts+(Scorecard+Elements).csv')

ID<- import('id\_name\_link.csv')

Clean ID data

ID <- ID %>%  
 group\_by(schname) %>%  
 mutate(n = n()) %>%  
 filter(n == 1)

Merge data

colnames(ID)[colnames(ID) == "unitid"] = "UNITID"  
  
ID<- inner\_join(ID, Cohorts, by = "UNITID" )  
  
Data <- inner\_join(Data,ID, by = "schname")

Remove colleges that don’t give college degrees

Data <- Data %>%  
 mutate(n = n()) %>%  
 filter(PREDDEG == 3)

**Create new binary variable that indicates whether the date (month) was before the scorecard came out or after.**

Data$Scorecard <- ifelse(lubridate::ymd(Data$month) >= lubridate::ymd("2015-09-01"), 1, 0)

**Condense data set**

Data <- Data %>%   
 select(schid, schname, keyword, keynum, monthorweek, index, month, UNITID, opeid, OPEID, INSTNM, 'md\_earn\_wne\_p10-REPORTED-EARNINGS', Scorecard )

**Change the data type of the median reported earnings.**

Data$'md\_earn\_wne\_p10-REPORTED-EARNINGS' <- as.numeric(Data$'md\_earn\_wne\_p10-REPORTED-EARNINGS')

Warning: NAs introduced by coercion

**Rename md\_earn\_wne\_p10-REPORTED-EARNINGS**

Data <- Data %>%  
 rename(md\_earnings = 'md\_earn\_wne\_p10-REPORTED-EARNINGS' )

**Overwrite md\_earnings to be binary of low income or high income**

Data$md\_earnings <- ifelse(Data$md\_earnings >= 55000, 1, 0)

**Create regression model:**

We are interested in an interaction model because we know that we need to shows how the effect of median earnings changes as the value of the Scorecard changes. So what this means in my model is that we want to use is:

Y(std. index of google search trends) = X (scorecard) \* Z (md\_earnings)

Regression1 <- feols(index ~ Scorecard\*md\_earnings, vcov = 'hetero', data=Data)

NOTE: 241,288 observations removed because of NA values (LHS: 196,629, RHS: 72,211).

etable(Regression1)

Regression1  
Dependent Var.: index  
   
Constant -12.84\*\*\* (0.0194)  
Scorecard -3.145\*\*\* (0.0472)  
md\_earnings 0.9888\*\*\* (0.0551)  
Scorecard x md\_earnings 0.7421\*\*\* (0.1328)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type Heteroskedas.-rob.  
Observations 741,598  
R2 0.00766  
Adj. R2 0.00765  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

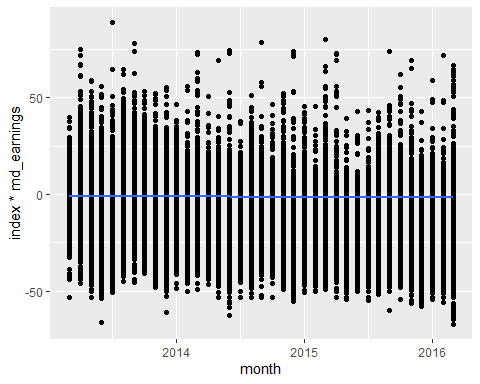
**Plot the Regression**

ggplot(data = Data, aes(x = month, y = index\*md\_earnings)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE)

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

Warning: Removed 241288 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 241288 rows containing missing values (`geom\_point()`).



**Results**

Here is the regression model:

Index (Std index of google trends scores) = Scorecard \* md\_earnings

Here is the regression results:

Dependent Var.: index  
   
Constant -12.84\*\*\* (0.0194)  
Scorecard -3.145\*\*\* (0.0472)  
md\_earnings 0.9888\*\*\* (0.0551)  
Scorecard x md\_earnings 0.7421\*\*\* (0.1328)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type Heteroskedas.-rob.  
Observations 741,598  
R2 0.00766  
Adj. R2 0.00765

Here is a graph that may be helpful in interpreting results:

This graph shows the month of the data on the x axis. Month is on the x-axis to show interest in high earning colleges over time considering that the College score card came out September 2015 On the Y axis is the index, which is representing interest in a college multiplied by the md\_earnings (this is a binary variable so only looks at high earning colleges). This line of best fit shows a decreasing slope hinting at a decreasing relationship between index and the scorecard variable.

This analysis will assit in understanding how, among colleges that give 4 year degrees, the interest in high or low earning colleges changed with the effect of the scorecard. There are many choices that I made along the way. The first step was cleaning the data. I removed colleges that don’t give 4 degrees since it was not of interest to the study. Then, I created two binary variables. The variable “Scorecard” become a binary variable that answered the question “is the data from before the score card was released or after?” The variable “md\_earnings” answered the question of whether the college was a high earning school or low earning school. I found a graphic on the National Center for Education Statistics that pointed to the average earnings for those with a bachelors degree between 2010-2020 was around $55,000 annually. The “md\_earnings” binary variabe was equal to one when the median earnings 10 years after graduation was at or above $55,000 and equal to 0 if it was below $55,000.

This analysis answers the research question with a model that looks like:

Y= B0 + B1X + B2Z + B3X\*Z + Error

Using the regression results we can see that the coefficient on the X variable, B1 is -3.145 which means that the scorecard decreased the interest in low earning colleges by -3.145. When we add B1 + B3 we get -2.4029. This number shows that the scorecard decreased interest in high earning colleges by -2.4029. Our B3 coefficient on its own represents the how the scorecard changes shifted interest in low earning schools and high earning schools. Since B3 is 0.7421 we can understand that there is a 0.7421 more interest in high earning schools with the release of the scorecard. Additionally the three stars next to the coefficients mean that they are significant at a p-value level of 0.001. So, this regression model shows that the college score card shifted interest to high earning schools but ultimately not by very much. This makes sense, when we talked about it in class, no one had heard of the college scorecard and so it hadn’t affected our college search very much. I also work at SU Admissions and I never get questions about incomes post-grad.