Data Exploration Project

Kaitlyn Au

ECON4110

### Data Cleaning:

# Load in libraries  
library(rio)  
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  
✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
✔ purrr 1.0.1   
── 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(lubridate)  
library(vtable)

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

library(fixest)  
library(ggplot2)  
library(lmtest)

Loading required package: zoo  
  
Attaching package: 'zoo'  
  
The following objects are masked from 'package:base':  
  
 as.Date, as.Date.numeric

library(multcomp)

Loading required package: mvtnorm  
Loading required package: survival  
Loading required package: TH.data  
Loading required package: MASS  
  
Attaching package: 'MASS'  
  
The following object is masked from 'package:dplyr':  
  
 select  
  
  
Attaching package: 'TH.data'  
  
The following object is masked from 'package:MASS':  
  
 geyser

# Read in Google Trends data files  
  
filelist <- list.files('Lab3\_Rawdata', pattern = 'trends', full.names = TRUE)  
my\_data <- import\_list(filelist, fill = TRUE, rbind = TRUE)

# Get first 10 characters out of the monthorweek variable  
my\_data$monthorweek <- ymd(str\_sub(my\_data$monthorweek, end = 10))

This next code chunk will standardize the *index* variable by school and keyword, aggregate the *monthorweek* variable to months. I also decided to create a dummy variable where the months before the release of the College Scorecard (September 2015) would be 0 and after the release would be 1:

my\_data <- my\_data %>%   
 group\_by(schname, keyword) %>%  
 mutate(index = (index - mean(index)) / sd(index), na.rm = TRUE) %>%  
 mutate(month = floor\_date(monthorweek, 'month')) %>%  
 mutate(after\_scorecard = ifelse(month >= "2015-09-01", 1, 0))

Then I filtered down the data to only include rows where the *index* variable is not NA and aggregate to the school–month level while also keeping the after\_scorecard variable in mind. Then I got the averages of the indexes for each group. I decided to go to the school-month level because I wanted to see the effects of the mean index for low vs. high earning colleges after the release of the college scorecard in a specific **month** (9/2015). Keeping it at the week-keyword level is unnecessary as we don’t need to see the specific changes week by week.

my\_data <- my\_data %>%  
 filter(index != is.na(index)) %>%  
 group\_by(schname, month, after\_scorecard) %>%  
 summarize(mean\_index = mean(index), na.rm = TRUE)

`summarise()` has grouped output by 'schname', 'month'. You can override using  
the `.groups` argument.

scorecard <- import("Lab3\_RawData/Most+Recent+Cohorts+(Scorecard+Elements).csv")  
id\_link <- import("Lab3\_RawData/id\_name\_link.csv")

vtable(scorecard)

scorecard

| Name | Class | Values |
| --- | --- | --- |
| UNITID | integer | Num: 100654 to 48285707 |
| OPEID | integer | Num: 100200 to 82098821 |
| opeid6 | integer | Num: 1002 to 42264 |
| INSTNM | character |  |
| CITY | character |  |
| STABBR | character |  |
| INSTURL | character |  |
| NPCURL | character |  |
| HCM2 | integer | Num: 0 to 1 |
| PREDDEG | integer | Num: 0 to 4 |
| CONTROL | integer | Num: 1 to 3 |
| LOCALE | character |  |
| HBCU | character |  |
| PBI | character |  |
| ANNHI | character |  |
| TRIBAL | character |  |
| AANAPII | character |  |
| HSI | character |  |
| NANTI | character |  |
| MENONLY | character |  |
| WOMENONLY | character |  |
| RELAFFIL | character |  |
| SATVR25 | character |  |
| SATVR75 | character |  |
| SATMT25 | character |  |
| SATMT75 | character |  |
| SATWR25 | character |  |
| SATWR75 | character |  |
| SATVRMID | character |  |
| SATMTMID | character |  |
| SATWRMID | character |  |
| ACTCM25 | character |  |
| ACTCM75 | character |  |
| ACTEN25 | character |  |
| ACTEN75 | character |  |
| ACTMT25 | character |  |
| ACTMT75 | character |  |
| ACTWR25 | character |  |
| ACTWR75 | character |  |
| ACTCMMID | character |  |
| ACTENMID | character |  |
| ACTMTMID | character |  |
| ACTWRMID | character |  |
| SAT\_AVG | character |  |
| SAT\_AVG\_ALL | character |  |
| PCIP01 | character |  |
| PCIP03 | character |  |
| PCIP04 | character |  |
| PCIP05 | character |  |
| PCIP09 | character |  |
| PCIP10 | character |  |
| PCIP11 | character |  |
| PCIP12 | character |  |
| PCIP13 | character |  |
| PCIP14 | character |  |
| PCIP15 | character |  |
| PCIP16 | character |  |
| PCIP19 | character |  |
| PCIP22 | character |  |
| PCIP23 | character |  |
| PCIP24 | character |  |
| PCIP25 | character |  |
| PCIP26 | character |  |
| PCIP27 | character |  |
| PCIP29 | character |  |
| PCIP30 | character |  |
| PCIP31 | character |  |
| PCIP38 | character |  |
| PCIP39 | character |  |
| PCIP40 | character |  |
| PCIP41 | character |  |
| PCIP42 | character |  |
| PCIP43 | character |  |
| PCIP44 | character |  |
| PCIP45 | character |  |
| PCIP46 | character |  |
| PCIP47 | character |  |
| PCIP48 | character |  |
| PCIP49 | character |  |
| PCIP50 | character |  |
| PCIP51 | character |  |
| PCIP52 | character |  |
| PCIP54 | character |  |
| DISTANCEONLY | character |  |
| UGDS | character |  |
| UGDS\_WHITE | character |  |
| UGDS\_BLACK | character |  |
| UGDS\_HISP | character |  |
| UGDS\_ASIAN | character |  |
| UGDS\_AIAN | character |  |
| UGDS\_NHPI | character |  |
| UGDS\_2MOR | character |  |
| UGDS\_NRA | character |  |
| UGDS\_UNKN | character |  |
| PPTUG\_EF | character |  |
| CURROPER | integer | Num: 0 to 1 |
| NPT4\_PUB-AVERAGE-ANNUAL-COST | character |  |
| NPT4\_PRIV | character |  |
| NPT41\_PUB | character |  |
| NPT42\_PUB | character |  |
| NPT43\_PUB | character |  |
| NPT44\_PUB | character |  |
| NPT45\_PUB | character |  |
| NPT41\_PRIV | character |  |
| NPT42\_PRIV | character |  |
| NPT43\_PRIV | character |  |
| NPT44\_PRIV | character |  |
| NPT45\_PRIV | character |  |
| PCTPELL | character |  |
| RET\_FT4 | character |  |
| RET\_FTL4 | character |  |
| RET\_PT4 | character |  |
| RET\_PTL4 | character |  |
| PCTFLOAN | character |  |
| UG25abv | character |  |
| GRAD\_DEBT\_MDN\_SUPP | character |  |
| GRAD\_DEBT\_MDN10YR\_SUPP | character |  |
| RPY\_3YR\_RT\_SUPP | character |  |
| C150\_4\_POOLED\_SUPP-REPORTED-GRAD-RATE | character |  |
| C200\_L4\_POOLED\_SUPP | character |  |
| md\_earn\_wne\_p10-REPORTED-EARNINGS | character |  |
| gt\_25k\_p6 | character |  |

# Filter scorecard to include only schools that primarily grant bachelor degrees (PREDDEG =3) & # only numerical values in the   
# earnings column:  
  
scorecard <- scorecard %>%  
 filter(PREDDEG == 3) %>%  
 filter(`md\_earn\_wne\_p10-REPORTED-EARNINGS` != "PrivacySuppressed") %>%  
 filter(`md\_earn\_wne\_p10-REPORTED-EARNINGS` != "NULL")  
  
scorecard$`md\_earn\_wne\_p10-REPORTED-EARNINGS` <- as.numeric(scorecard$`md\_earn\_wne\_p10-REPORTED-EARNINGS`)  
  
mean(scorecard$`md\_earn\_wne\_p10-REPORTED-EARNINGS`)

[1] 42116.98

scorecard <- scorecard %>%  
 mutate(earnings\_cat = ifelse(`md\_earn\_wne\_p10-REPORTED-EARNINGS` > 42117, 1, 0))  
  
scorecard <- scorecard %>%  
 dplyr::select(UNITID, OPEID,PREDDEG, earnings\_cat)

I decided to create a new dummy variable *earnings\_cat* that will categorize the mean earnings by high or low based off of the average value of the median reported earnings. The average mean reported earning was $42,116.98, but I decided to round it up because the numbers were originally reported as whole numbers. Any values of the reported earnings that were greater than $42,117 were given a 1, while those below the average got a 0. It might have made more sense to make the cutoffs at the first and third quartiles to truly get a high/low rating, but that that could have possibly limited our data too much. Using the mean as the cutoff point works well as colleges can be considered “high-earning” if they are above the average, and “low-earning” if they are below the average.

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

my\_data <- inner\_join(my\_data, id\_link, by = "schname") %>%  
 inner\_join(id\_link, scorecard, by = c("unitid", "opeid"))  
   
my\_data <- my\_data %>%  
 mutate(UNITID = unitid) %>%  
 mutate(OPEID = opeid)  
   
my\_data <- full\_join(my\_data, scorecard, by = c("UNITID","OPEID"))   
#did this full\_join so that the earnings\_cat column would be in the my\_data  
  
my\_data <- my\_data %>%  
 filter(earnings\_cat != is.na(earnings\_cat)| earnings\_cat == 0) %>%  
 filter(schname.x != is.na(schname.x))

## Data Analysis:

# Regression of mean\_index per month controlling for earnings category of the college (high/low)  
rmodel <- feols(mean\_index ~ month + earnings\_cat + after\_scorecard + earnings\_cat\*after\_scorecard,   
 data = my\_data, vcov = 'hetero')  
  
etable(rmodel)

rmodel  
Dependent Var.: mean\_index  
   
Constant 13.92\*\*\* (0.1752)  
month -0.0009\*\*\* (1.08e-5)  
earnings\_cat 0.0246\*\*\* (0.0052)  
after\_scorecard 0.3062\*\*\* (0.0100)  
earnings\_cat x after\_scorecard -0.1075\*\*\* (0.0123)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type Heteroskedasti.-rob.  
Observations 57,563  
R2 0.13090  
Adj. R2 0.13084  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

I decided to set up a regression of *mean\_index* on month while controlling for the earnings category of the college and whether the search occurred after the Scorecard release or not. I also included an interaction term with *earnings\_cat* and *after\_scorecard* that would allow me to see how the effect of *month* on *mean\_index* (avgerage student interest) varies depending on the earnings category of the college and whether the observation was before or after the introduction of the College Scorecard. This interaction effectively answers the research question: Among colleges that predominantly grant bachelor’s degrees, did the release of the College Scorecard shift student interest to high-earning colleges relative to low-earning ones?

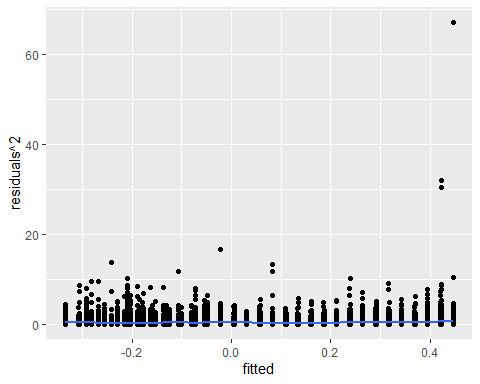
The scatterplot in the next code chunk is helping me visually inspect the data in my regression model for heteroskedasticity by plotting the residuals against the predicted values from my *rmodel* regression. I see that the residuals in this scatterplot have increased as the fitted values increase, which suggests there might be some heteroskedasticity, so I added *vcov = ‘hetero’* into my regression model (*rmodel*), which will adjust the standard errors to account for the potential heteroskedasticity I saw in the residuals. The results from a bptest(*rmodel*) in R would have been better to test for heteroskedasticity, but it wouldn’t work.

Looking at the results from the regression, all the variables are statistically significant at a high confidence level (0.001 = \*\*\*), suggesting that they are important predictors of the *mean\_index*. We can see that a one unit increase in *month*, decreases *mean\_index* by 0.0009 units, while holding all other variables constant. A one unit increase in the earnings category for colleges (1 = high-earning college), increases the *mean\_index* by 0.0246, holding all other variables constant. A one unit increase in after\_scorecard increases the mean\_index by 0.3062, holding all other variables constant. The interaction term suggests that the relationship between a one-unit increase in *earnings\_cat* and *mean\_index* differs before and after the College Scorecard was introduced. Before the introduction of the College Scorecard, the association between a one-unit increase in earnings\_cat (high-earning vs. low-earning colleges) and mean\_index was -0.0009. However, after the Scorecard, this association changes to (-0.0009-0.1075)= -0.1084. Therefore, the relationship between *earnings\_cat* and *mean\_index* became 0.1075 more negative after the Scorecard was introduced than before. The coefficient for the interaction term is also statistically significant at the 0.001 level (\*\*\*), indicating that the effect of *earnings\_cat* on *mean\_index* differs significantly before and after the introduction of the College Scorecard.

Overall, the introduction of the College Scorecard decreased search activity on Google Trends for colleges with high-earning graduates by 0.1075 units relative to what it did for colleges with low-earning graduates, with a standard error of 0.0123. This result comes from the *earnings\_cat\*after\_scorecard* coefficient in the regression.

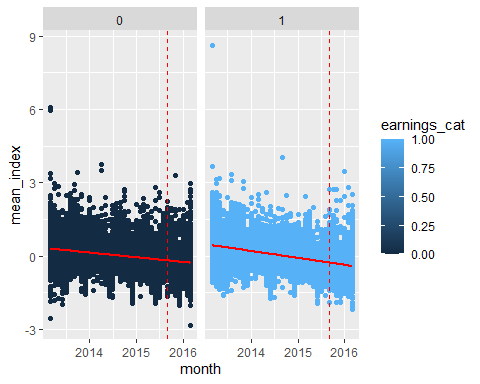
residuals <- residuals(rmodel)  
fitted\_values <- fitted(rmodel)  
ggplot(data.frame(residuals = residuals, fitted = fitted\_values),   
 aes(x = fitted, y = residuals^2)) +   
 geom\_point() +  
 geom\_smooth()

`geom\_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'



ggplot(my\_data, aes(x = month, y = mean\_index, vcov = 'hetero')) +  
 geom\_point(aes(color = earnings\_cat)) +  
 geom\_smooth(method = "lm", color = "red") +   
 geom\_vline(xintercept = as.Date("2015-09-01"), linetype = "dashed",   
 color = "red") +   
 facet\_wrap(~ earnings\_cat)

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



The graphs above are showing the change in mean\_index by month for the colleges in the data from March 2013-2016. The graph on the left in black shows the low earning colleges and the graph on the right in blue is the high-earning colleges. The red dashed line is marking the month when the College Scorecard was introduced.