Data Exploration Assignment

Avery Hultgren

library(rio)  
library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ forcats 1.0.0 ✔ readr 2.1.4  
✔ ggplot2 3.4.1 ✔ 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(fixest)  
library(ggplot2)  
library(lubridate)  
library(vtable)

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

**Data Cleaning**

gtrend <- import\_list(list.files(path = "~/Econometrics/Data\_Exploration\_Rawdata/Lab3\_Rawdata", full.names = TRUE),fill = TRUE, rbind = TRUE)

gtrend <- gtrend %>%   
 mutate(monthorweek = ymd(str\_sub(string = monthorweek, start = 1, end = 10)))

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

**“Now, a one-unit change in the standardized index can be understood and interpreted as a one-standard-deviation change in search interest”**

cohorts <- import('Most+Recent+Cohorts+(Scorecard+Elements).csv', fill = TRUE)  
  
colnames(cohorts)[colnames(cohorts) == "OPEID"] ="opeid"  
  
idlink <- import('id\_name\_link.csv', fill = TRUE)

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

**Join data sets together**

idlink <- inner\_join(idlink, cohorts, by = 'opeid')

Warning in inner\_join(idlink, cohorts, by = "opeid"): Detected an unexpected many-to-many relationship between `x` and `y`.  
ℹ Row 104 of `x` matches multiple rows in `y`.  
ℹ Row 5664 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.

gtrend <- inner\_join(gtrend,idlink, by = 'schname')

Warning in inner\_join(gtrend, idlink, by = "schname"): Detected an unexpected many-to-many relationship between `x` and `y`.  
ℹ Row 48882 of `x` matches multiple rows in `y`.  
ℹ Row 3568 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.

**Get rid of all colleges and universities that do not predominately give bachelor’s degrees**

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

**Data Exploration**

**Question:** *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)?*

**Work the data**

gtrend <- gtrend %>%  
 select(schname, keyword, schid,index,'md\_earn\_wne\_p10-REPORTED-EARNINGS',keynum,monthorweek,opeid,unitid)

**Split up schools via high and low earnings colleges**

gtrend <- gtrend %>%  
 rename(mdearn10 = `md\_earn\_wne\_p10-REPORTED-EARNINGS`)

gtrend$`mdearn10` <- as.numeric(gtrend$`mdearn10`)

Warning: NAs introduced by coercion

gtrend <- na.omit(gtrend)

**split data up before and after the scorecard was shown**

gtrend <- gtrend %>%   
 mutate(BoA = ifelse(ymd(monthorweek) >= ymd("2015-09-01"), 1, 0))

gtrend <- gtrend %>%  
 filter(BoA == 1)

**create new data sets of only high or lower earnings schools**

bigmon <- gtrend %>%  
 mutate(n = n()) %>%  
 filter(mdearn10 >= 65000)  
smolmon <- gtrend %>%  
 mutate(n = n()) %>%  
 filter(mdearn10 <= 30000)

**Run a regression of both the high and low earnings schools**

Sch\_Reg\_High <- feols(index ~ monthorweek + mdearn10, data = bigmon, vcov = 'hetero')

Sch\_Reg\_Low <- feols(index ~ monthorweek+mdearn10, data = smolmon, vcov = 'hetero')

etable(Sch\_Reg\_High,Sch\_Reg\_Low)

Sch\_Reg\_High Sch\_Reg\_Low  
Dependent Var.: index index  
   
Constant 39.88\*\*\* (3.499) 41.96\*\*\* (3.175)  
monthorweek -0.0024\*\*\* (0.0002) -0.0025\*\*\* (0.0002)  
mdearn10 -1.64e-6\* (6.74e-7) -4.11e-6 (3.09e-6)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type Heteroskedast.-rob. Heteroskedast.-rob.  
Observations 6,600 8,723  
R2 0.02113 0.02180  
Adj. R2 0.02083 0.02158  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

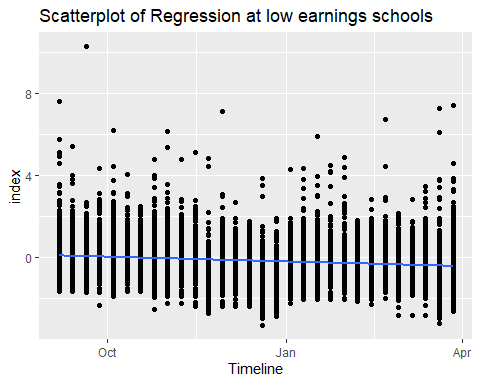
**Create graphs of both regressions**

plotHigh <- ggplot(data = bigmon, xlim = c(0, 15), ylim = c(0, 11), aes(x =monthorweek , y =(index))) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 labs(x = "Timeline", y = "index") +  
 ggtitle("Scatterplot of Regression at high earnings schools")

plotLow <- ggplot(data = smolmon, xlim = c(0, 15), ylim = c(0, 11), aes(x =(monthorweek) , y =(index))) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) +  
 labs(x = "Timeline", y = "index") +  
 ggtitle("Scatterplot of Regression at low earnings schools ")

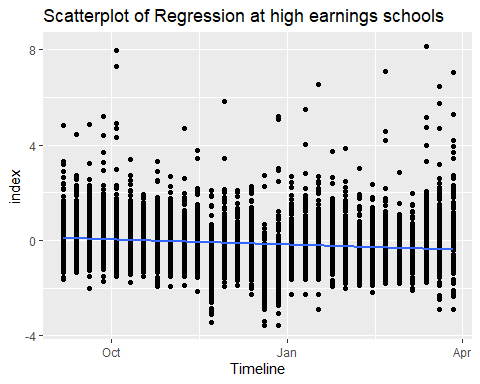
layout(matrix(c(1, 2)))  
  
print(plotLow)

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



print(plotHigh)

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



* **Explain why you are performing the analysis you are performing, and the choices you made in putting it together**

I chose to investigate the connection between the median earnings of students after 10 years and the index for the schools that were on the high and low end of the median. To do this I created two separate data sets, one with median earnings above and equal to $65,000/yr and the other with earnings below or equal to $30,000/yr. I then created a regression for each and regressed the index on to the median earnings of the two separate school groups, while controlling for the dates.

* **Explain how your analysis addresses the research question**

My analysis shows that after September 2015 there was little difference in the slope of the college indexes between high earning and low earning colleges.

* **Explain what we should conclude, in real world terms, based on your results**

Based on my results I believe that we should conclude that there was a insignificant amount of change in college searches due to the college scorecard being released by Google. This means that many people probably did not know about the scorecard or did not factor it into there college searches.