Data Exploration Project

In this project I worked on a real life, messy data set in order to answer the question of whether students' preference shifted after the data set publication or not. The project has several steps.

First, I try to combine different data files by universities' names or codes. According to the instructions, I dropped universities with the exact same name.

```
library(purrr)
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(readr)

#Generate Google trends data, drop duplicated values from id_name

list_names <- list.files(path = ".", pattern = "trends_up_to", full.names =
TRUE)

GT <- list_names %>%
  map(read_csv) %>%
  bind_rows()

id_name_link <- read_csv("id_name_link.csv") %>%
  group_by(schname) %>%
  mutate(N=n()) %>%
  filter(N==1)
```

Second, I generated a completely new data set named "new_index". In order to compare different universities, we need some normalization. Therefore, I grouped data by search keywords and standardized the index. After that, I computed the average weekly search index by school name. Since all indexes have the same unit now, it can be generalized through university names rather than keywords.

```
#Standardization of the index

GT <- GT %>%
  na.omit(GT) %>%
  mutate(new_date = as.Date(monthorweek)) %>%
```

```
  group_by(keyword) %>%
  mutate(new_index = (index-mean(index))/sd(index))

structural_date = as.Date("2015-01-01")

#Average index

new_GT_2015later <- GT %>%
  group_by(schname) %>%
  filter(new_date>=structural_date) %>%
  mutate(average_index_2015later = mean(new_index)) %>%
  distinct(schname, .keep_all= TRUE)

new_GT_2015before <- GT %>%
  group_by(schname) %>%
  filter(new_date<structural_date) %>%
  mutate(average_index_2015before = mean(new_index)) %>%
  distinct(schname, .keep_all= TRUE)

new_GT_2015later <- new_GT_2015later[c("schname", "average_index_2015later")]

new_GT_2015before <- new_GT_2015before[c("schname",
"average_index_2015before")]

#Join 2 data sets.
new_GT_2015before <- new_GT_2015before %>%
  left_join(new_GT_2015later, by = "schname", keep = TRUE) %>%
  left_join(id_name_link, by = c("schname.x" = "schname"), keep = TRUE) %>%
  na.omit()
```

Third, I imported the full data set with the most recent scorecard and cleaned the data set to prepare for the estimation. To put it in a different way, drop duplicated universities, make a new dummy variable which denotes ownership of the university, generate an earning variable and clean the necessary variables from NULL and categorical values. The percentile variable indicates the relative rank of the average earning of the particular university graduates. For example: if university x has 32% in their earnings, that indicates we can find 32% of total sample universities have below earnings from x university.

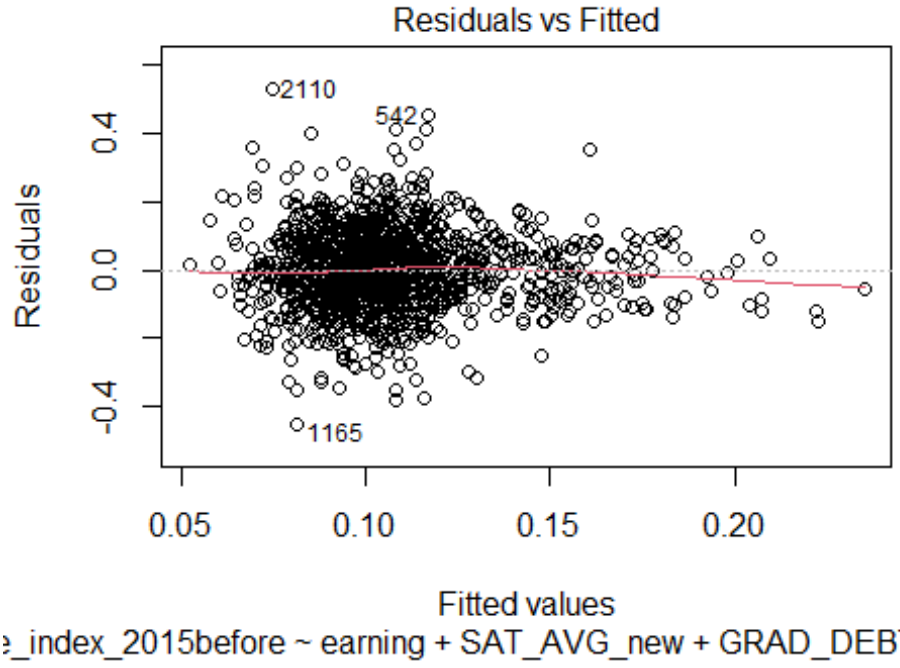
```
#Import the data

Full_data <- read_csv("Most+Recent+Cohorts+(Scorecard+Elements).csv")

Full_data <- Full_data %>%
  right_join(new_GT_2015before, by = c("UNITID" = "unitid"), keep = TRUE)

Full_data <- Full_data %>%
  mutate(own = replace(Full_data$CONTROL, Full_data$CONTROL == 2|3, 0)) %>%
  na.omit()
```

```
#Compute earning

earning <- rank(Full_data$`md_earn_wne_p10-REPORTED-
EARNINGS`)/nrow(Full_data)

#Clean chosen variables from NULLs, categorical values and transform into
numerical values.

Full_data <- cbind(Full_data, earning)
Full_data <- Full_data[grep("[[:digit:]]", Full_data$UGDS), ]
Full_data <- Full_data[grep("[[:digit:]]", Full_data$SAT_AVG), ]
Full_data <- Full_data[grep("[[:digit:]]", Full_data$GRAD_DEBT_MDN_SUPP), ]
Full_data <- Full_data[grep("[[:digit:]]", Full_data$PCTPELL), ]



Full_data <- Full_data %>%
  mutate(UGDS_new = as.numeric(UGDS)) %>%
  mutate(SAT_AVG_new = as.numeric(SAT_AVG)) %>%
  mutate(GRAD_DEBT_MDN_SUPP_new = as.numeric(GRAD_DEBT_MDN_SUPP))
```

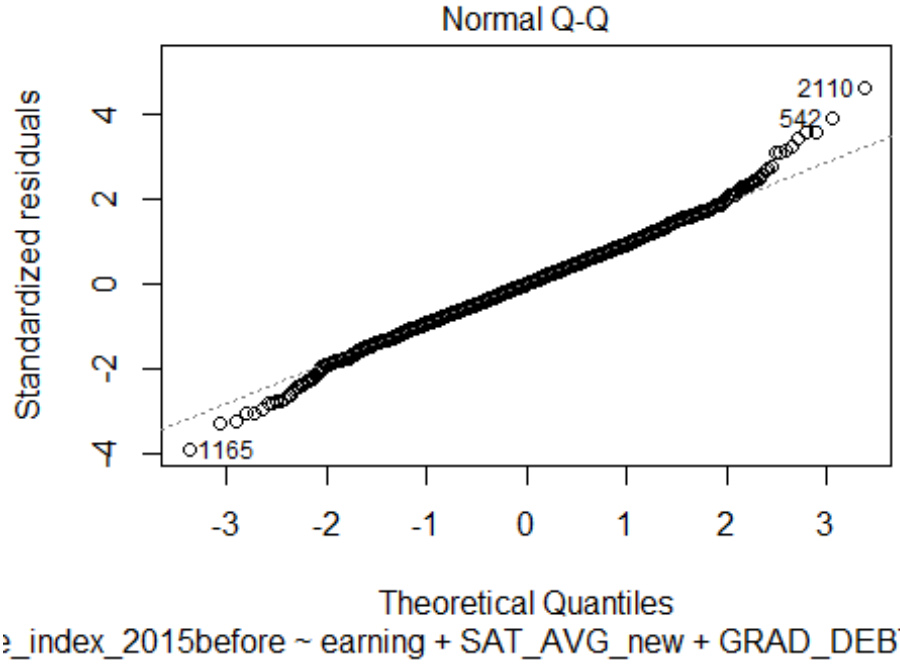
Lastly, I estimated two different linear OLS models to identify what determines university popularity/search among people. I chose earning, students average SAT score, graduates average debt, enrolled student. 1. If university students have a higher salary after graduation (10 years), it may positively impact their popularity. 2. Since SAT score is one way to measure a student's success and high-ranking universities have higher threshold for general scores, it may have positive correlation with popularity. 3. If the university students have higher debt than other universities, students may not choose that university. Therefore, I believe these 2 variables have negative correlation. 4. To represent university size, I used enrolled students' numbers. If a lot of students enroll and the school size is relatively big, it may positively affect their popularity. 5. I tried an ownership dummy before but it did not add any value to the estimation.

```
model1 <- lm(average_index_2015before ~ earning + SAT_AVG_new +
GRAD_DEBT_MDN_SUPP_new + UGDS_new, data = Full_data)

model2 <- lm(average_index_2015before ~ earning + SAT_AVG_new +
GRAD_DEBT_MDN_SUPP_new + UGDS_new, data = Full_data)

summary(model1)

##
## Call:
## lm(formula = average_index_2015before ~ earning + SAT_AVG_new +
##     GRAD_DEBT_MDN_SUPP_new + UGDS_new, data = Full_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45073 -0.07455 -0.00124  0.07274  0.52914
##
```

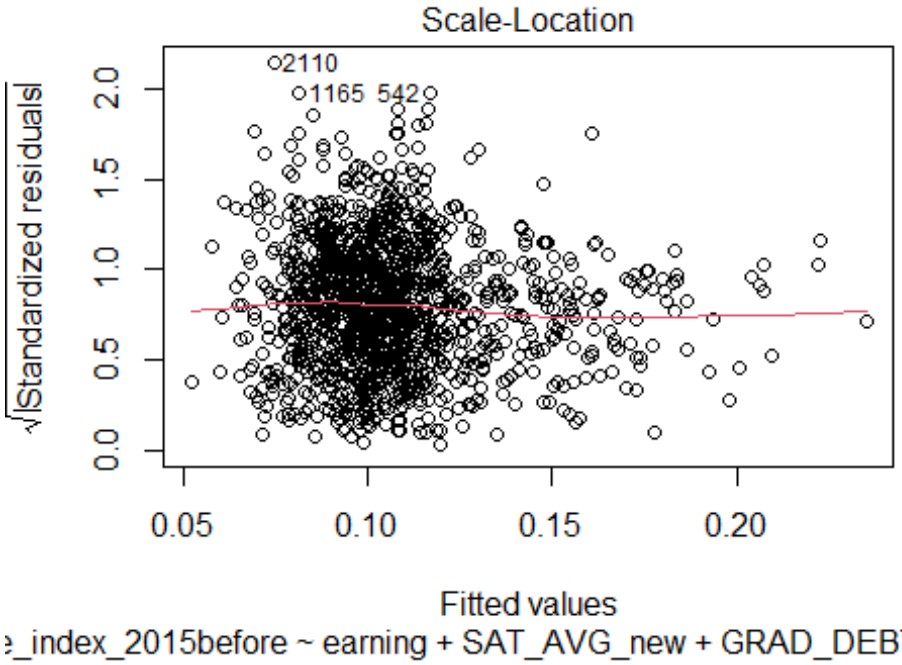
```
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.004e-01  3.453e-02   2.906  0.00372 **
## earning                  5.634e-02  1.870e-02   3.014  0.00263 **
## SAT_AVG_new             -5.470e-05  2.976e-05  -1.838  0.06623 .
## GRAD_DEBT_MDN_SUPP_new  5.569e-07  7.061e-07   0.789  0.43045
## UGDS_new                 2.936e-06  4.655e-07   6.308 3.84e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1152 on 1331 degrees of freedom
## Multiple R-squared:  0.04116,    Adjusted R-squared:  0.03828
## F-statistic: 14.28 on 4 and 1331 DF,  p-value: 2.016e-11

summary(model2)

##
## Call:
## lm(formula = average_index_2015before ~ earning + SAT_AVG_new +
##     GRAD_DEBT_MDN_SUPP_new + UGDS_new, data = Full_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45073 -0.07455 -0.00124  0.07274  0.52914
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.004e-01  3.453e-02   2.906  0.00372 **
## earning                  5.634e-02  1.870e-02   3.014  0.00263 **
## SAT_AVG_new             -5.470e-05  2.976e-05  -1.838  0.06623 .
## GRAD_DEBT_MDN_SUPP_new  5.569e-07  7.061e-07   0.789  0.43045
## UGDS_new                 2.936e-06  4.655e-07   6.308 3.84e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1152 on 1331 degrees of freedom
## Multiple R-squared:  0.04116,    Adjusted R-squared:  0.03828
## F-statistic: 14.28 on 4 and 1331 DF,  p-value: 2.016e-11

plot(model1)
```

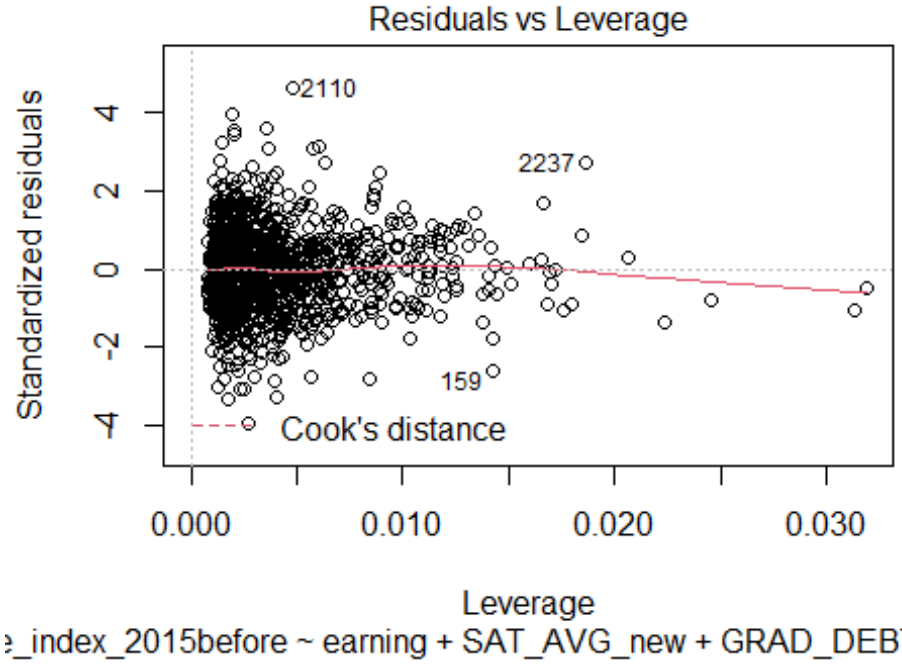


e_index_2015before ~ earning + SAT_AVG_new + GRAD_DEBT_MI

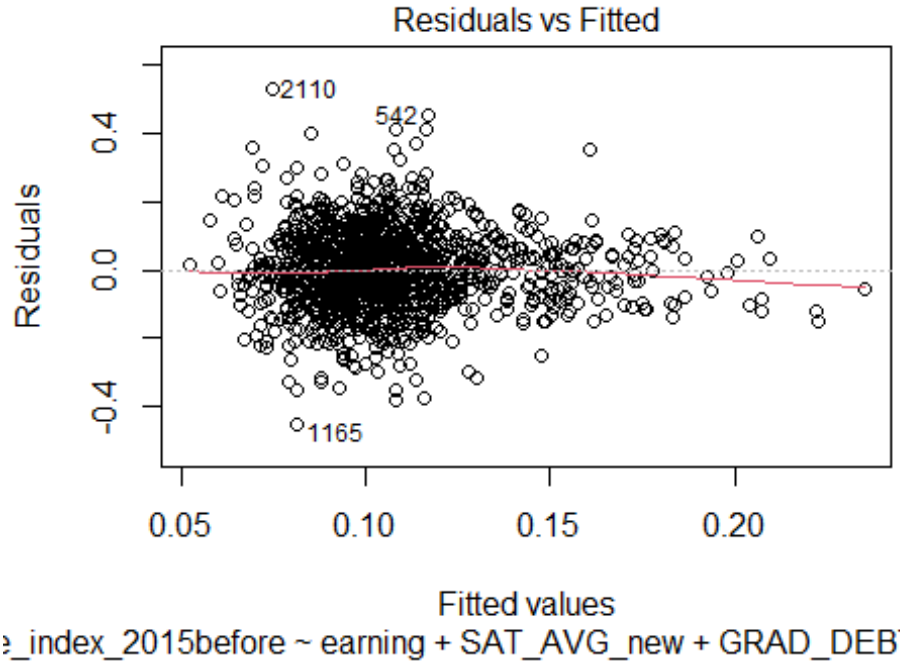


e_index_2015before ~ earning + SAT_AVG_new + GRAD_DEBT_MI

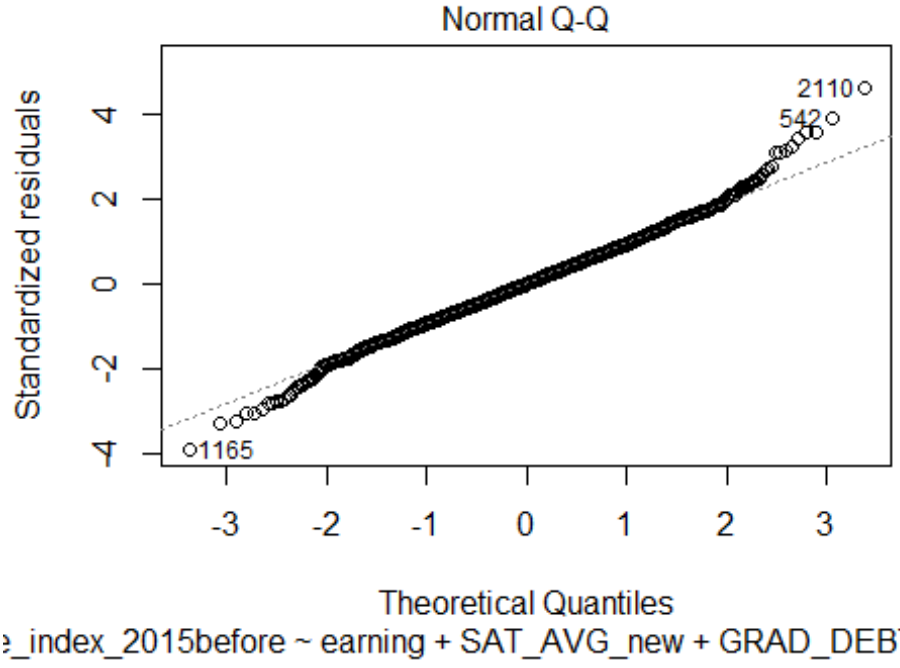

Scale-Location
2110
1165
542
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0.05
0.10
0.15
0.20
Fitted values
e_index_2015before ~ earning + SAT_AVG_new + GRAD_DEBT_MI


Residuals vs Leverage
2110
2237
159
Cook's distance
Standardized residuals
-4
-2
0
2
4
0.000
0.010
0.020
0.030
Leverage
e_index_2015before ~ earning + SAT_AVG_new + GRAD_DEBT_MI

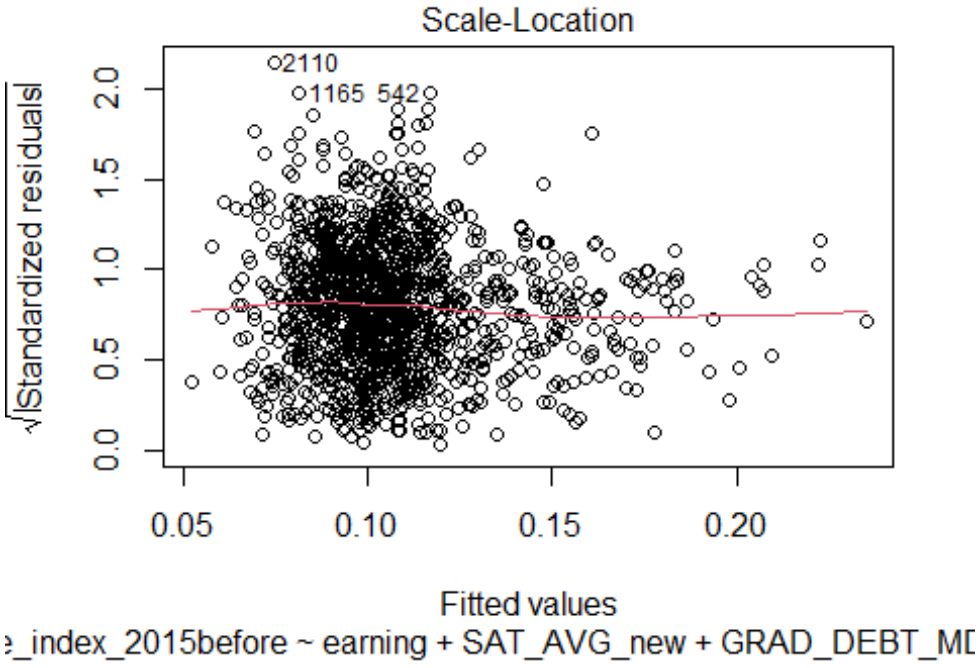
```
plot(model2)
```

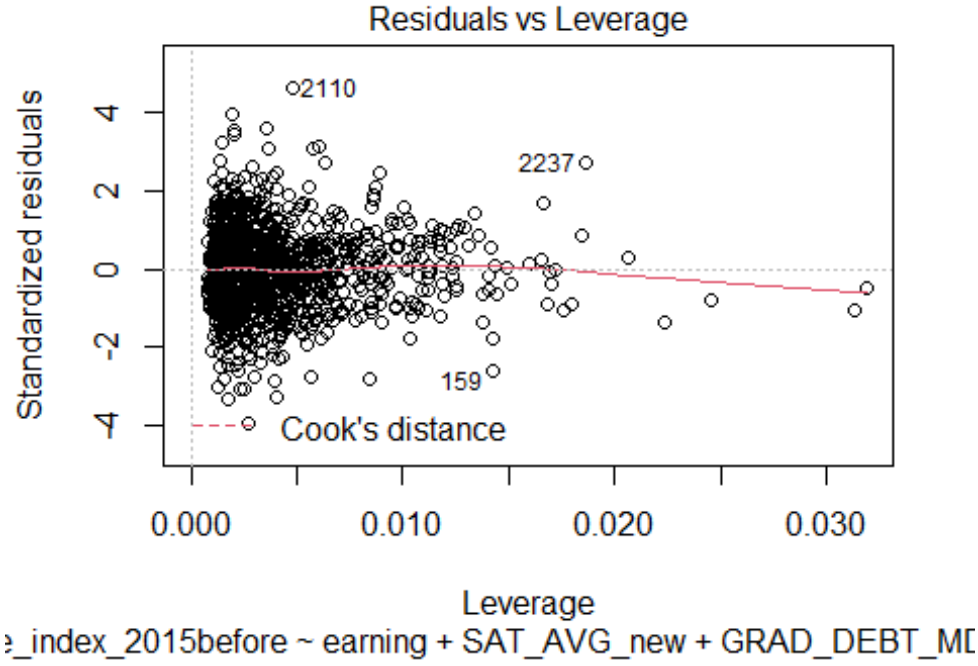


e_index_2015before ~ earning + SAT_AVG_new + GRAD_DEBT_MI



e_index_2015before ~ earning + SAT_AVG_new + GRAD_DEBT_MI





Here the confidence interval is 95%. From the project question and my estimation, I believe the 2015 report did not change students' preference toward low income to higher income university. Rather they have a positive attitude towards universities with students who earn higher even before the report. Interestingly, if students who graduated that

university's percentile increased by 10 percent, the standardized popularity index increased by 0.56 units. If the school is bigger, the number of enrolled students is higher, their popularity is also high. On the other hand, enrolled students' average SAT score negatively affects the popularity (90% confidence). Overall estimation was statistically significant for both models (F statistics) but R-square was low.