Equity and ACT in Colleges

Tanuj Guha

Commented [PI1]: Why is Mbonisi’s name not listed here?

2023-03-07

Abstract: In this project we are going to investigate the relationship between ACT scores, and several other factors that affect an individual’s college admissions. The intention of this project is to estimate how much of the ACT scores of admitted students can be explained by other factors such as English Proficiency, Tutoring, High School GPA, among other things.

Let us dive right into the project, and begin by importing the dataset.

Getting the data (appending legible names to dataset from dictionary)

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(readxl)
library(xlsx)
library(datadictionary)
data = read.csv("adm2021.csv")
dictData = read_xlsx("adm2021Dict.xlsx", sheet=2)
dict = data.frame(dictData$varname, dictData$varTitle)
names(data) <- dict$dictData.varTitle[match(names(data), dict$dictData.varnam
e)]
#write.xlsx(data, file = "appendedData.xlsx", sheetName="Colleges")
```

Creating the final data set with only the chosen variables.

```
colleges = select(data,'Unique identification number of the institution')

colleges <- cbind(colleges,select(data,'ACT Composite 75th percentile score')
,select(data,'Secondary school GPA'),select(data,'Completion of college-prepa
ratory program'),select(data,'TOEFL (Test of English as a Foreign Language'),
select(data,'Enrolled total'),select(data,'Percent of first-time degree/certi
ficate-seeking students submitting ACT scores'),select(data,'SAT Evidence-Bas
ed Reading and Writing 75th percentile score'),select(data,'SAT Math 75th per
centile score'))

colleges <- colleges %>%
        rename("ID" = "Unique identification number of the institution",
               "ACT" = "ACT Composite 75th percentile score",
               "GPA" = "Secondary school GPA",
               "CollegePrep" = "Completion of college-preparatory program",
               "TOEFL" = "TOEFL (Test of English as a Foreign Language",
               "NumberEnrolled" = "Enrolled total",
               "ACTPercentage" = "Percent of first-time degree/certificate-se
eking students submitting ACT scores",
                "SATWR" = "SAT Evidence-Based Reading and Writing 75th percent
ile score",
               "SATM"= "SAT Math 75th percentile score")

colleges = drop_na(colleges)
write.xlsx(colleges, file = "Colleges.xlsx", sheetName="Colleges")
```

This marks the end of house keeping data cleaning.

```
colleges = read_xlsx("colleges.xlsx")

## New names:
## • `` -> `...1`
```

Introduction: Conceptually speak, the ACT, or any other standardized Test for that matter, is touted as an indicator of a person's mastery of high school matter. This mastery is supposed to be a prerequisite for college coursework to further build upon. However, indicators are can often be misleading, and for an indicator to be robust, it has to be concise while being the *least* reductive.

Commented [PI2]:
speaking

Specifically, in this project, we are investigating the validity of ACT scores, and looking into factors that might contribute towards a higher score for a cohort of students. These cohorts are broken down in terms freshmen who enrolled in the same college.We expect the ACT scores to follow a normal distribution. WE also recognize that we haven't *exhaustively* identified all of the factors that determine the ACT scores, hence we expect the residuals to have a normal distribution, when plotted against predicted values.

Commented [PI3]:
Insert a space here, and after every period. There are other places you need to do this.

Data Collection:

This dataset has 7 predictor variables: __ Secondary school GPA (quant) __ Completion of college-preparatory program (category) __ TOEFL (Test of English as a Foreign Language (categorical) __ Enrolled total (quant) __ Percent of first-time degree/certificate-seeking students submitting ACT scores (quant) __ SAT Evidence-Based Reading and Writing 75th percentile score) (quant) __ SAT Math 75th percentile score (quant)

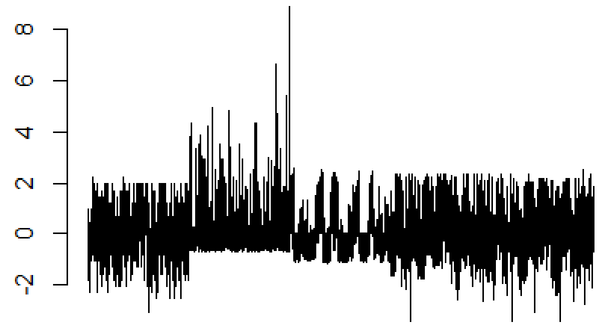
Commented [PI4]:
Need a blank line before and after this list in the R Markdown file to get the desired formatting.

Can you describe the categories more clearly? What are the categories and how is each category defined?

Each of these predictor variables explain, in some part, the 75th Percentile ACT scores of incoming students in a given college. For the sake of privacy, the colleges have been anonymized and are represented by proxies under column 'ID'.

Let us look at the distribution of each of the quantitative variables in this dataset.

```
numerical = cbind(colleges$ACT, colleges$NumberEnrolled, colleges$ACTPercenta
ge, colleges$SATM, colleges$SATWR)
barplot(scale(numerical), beside=T)
```



Commented [PI5]: This graph does not have labels, so I can't tell what I'm looking at. Your paragraph below helps to some degree, but it's not enough.

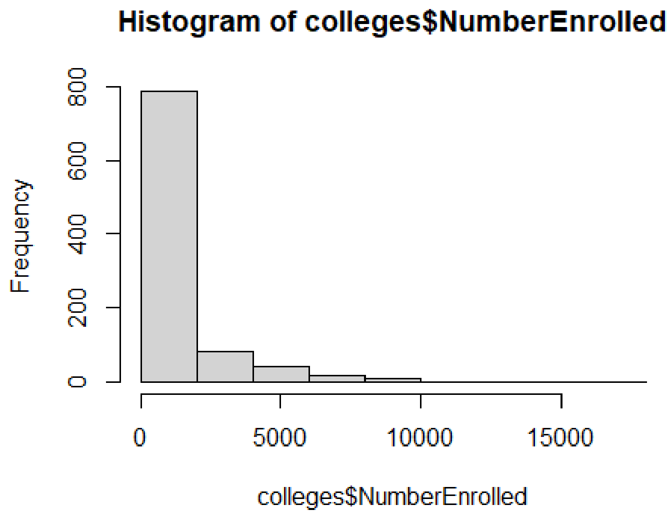
The idea for this graph is okay, but it needs better labelling, including the y-axis.

Note: un-labelled axes is one of my pet peeves. ☺

Some might accuse our approach as one that isn't assiduous, but I'd beg to disagree. The above graph visually indicated shows that, the only variable that needs transformation would be the 'NumberEnrolled' column. To arrive at the conclusion, we first created a dataset of all the numerical columns of from the colleges dataset. Then, we scaled the columns (scaling a column does not change its skewness). We chose to scale the columns because that way *all* of the columns can be represented in the same bar graph. Then, we plotted all of the columns side by side in a bar plot. We chose a singular bar plot for brevity's sake, and in our judgement, it was *enough* to indicate only those columns, where the skew was painfully obvious.

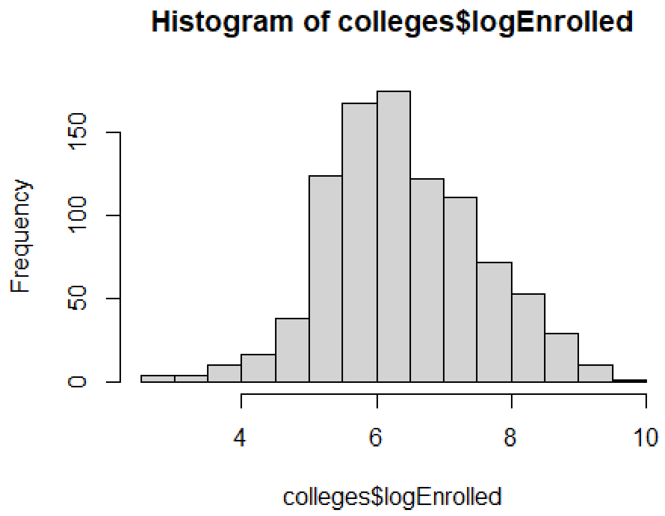
Now let us investigate 'NumberEnrolled' in greater detail.

```
hist(colleges$NumberEnrolled)
```



Yikes! I was going to do a Shapiro-Wilk test, but I think given the visual evidence, that would be a moot point. So, let us go right to the log-transformation.

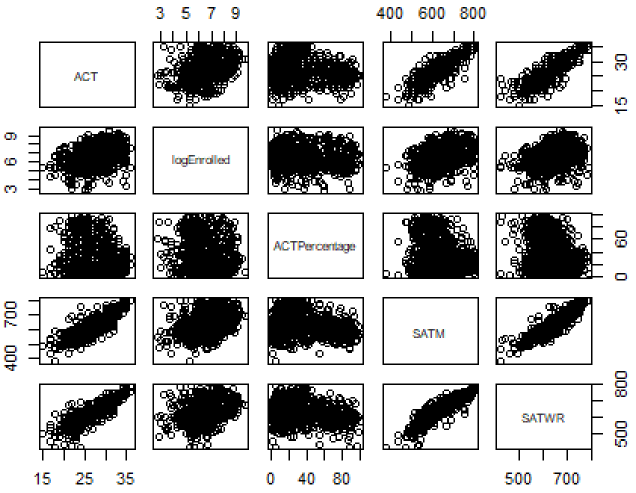
```
colleges$logEnrolled <- (log(colleges$NumberEnrolled))
hist(colleges$logEnrolled)
```



For our purposes, it (the log trnasformation) works! Let us now create a scatterplot matrix.

Commented [PI6]: The magic of the log transformation!

```
numerical = cbind(ACT = colleges$ACT, logEnrolled = colleges$logEnrolled, ACT
Percentage = colleges$ACTPercentage, SATM = colleges$SATM, SATWR = colleges$S
ATWR)
pairs(numerical)
```



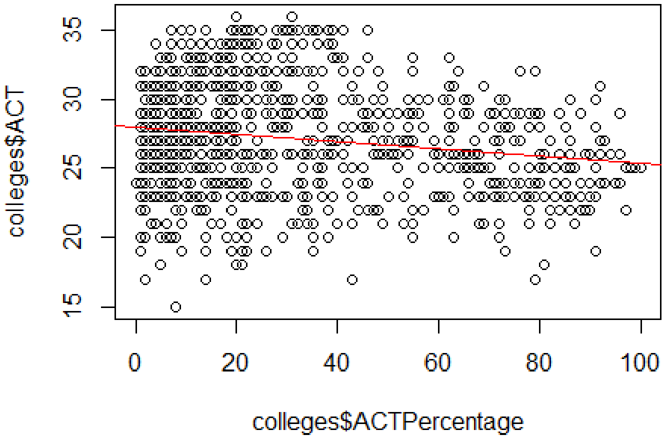
Visually, there seems to be a pretty strong corrleation between (75th percentile) ACT and SAT scores (dis aggregated by subjects). There also seems to be a relationship between 75th percentile ACT scores, and the percentage of students who choose to submit their ACT scores.

Commented [PI7]: I'm surprised you didn't pick SATM or SATWR, since they each have a higher correlation with ACT.

Of the predictor variables in the *colleges* dataset, we feel GPA and TOEFL should have been numerical. In the current case they are binned variables, describing the qualitative attributes *about* the test takes, over the actual representation of scores themselves. We also feel CollegePrep should have been binary.

Results: Let us do a simple linear regression. Let us predict the 75th Percentile of ACT scores, for a given cohort, by looking at what percentage of that cohort decide to submit scores.

```
model1 = lm(colleges$ACT ~ colleges$ACTPercentage)
summary(model1)

##
## Call:
## lm(formula = colleges$ACT ~ colleges$ACTPercentage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7862  -2.7862  -0.0996   2.9831   8.8212
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            27.997450   0.195552 143.171  < 2e-16 ***
## colleges$ACTPercentage -0.026408   0.004773  -5.533  4.1e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.937 on 935 degrees of freedom
## Multiple R-squared:  0.0317, Adjusted R-squared:  0.03066
## F-statistic: 30.61 on 1 and 935 DF,  p-value: 4.097e-08

plot(colleges$ACTPercentage, colleges$ACT)
abline(model1, col = "red")
```



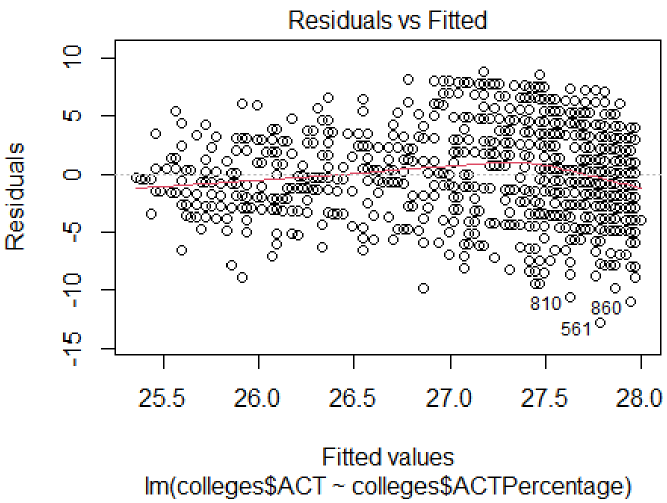
Given that 0.00000004097 is < 0.05. The linear model is significant in it's relationship. However, the relationship is **inverted** from what we had initially expected. We had expected that, for a given cohort, more people would choose to submit their scores, if their ACT scores were higher. However, the model shows that *infact the opposite is true*: if a greater percetnage of people choose to submit their ACT scores, then their 75th percentile ACT scores are lower!

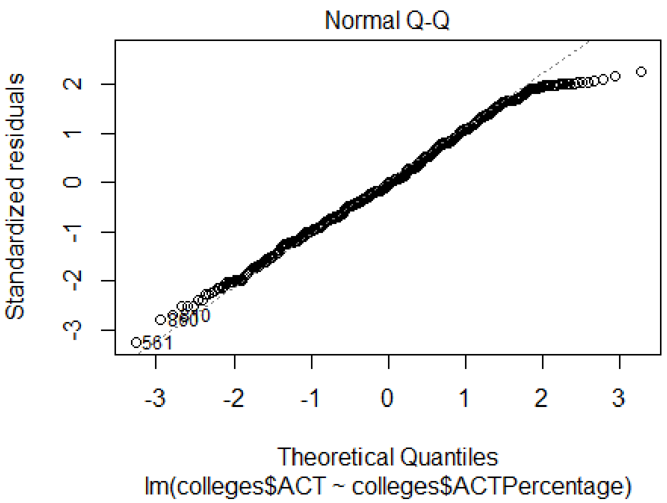
Commented [PI8]: in fact

(two words0

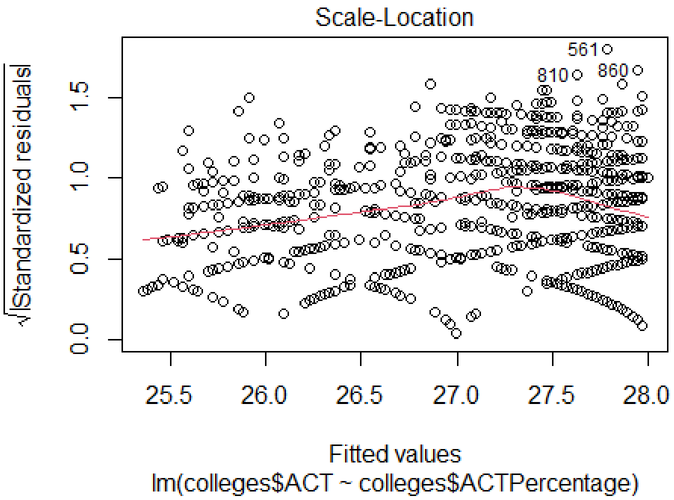
Commented [PI9]: Maybe that's because individuals with lower scores may choose to not submit. Thus, the schools with a lower percentage have higher average scores.

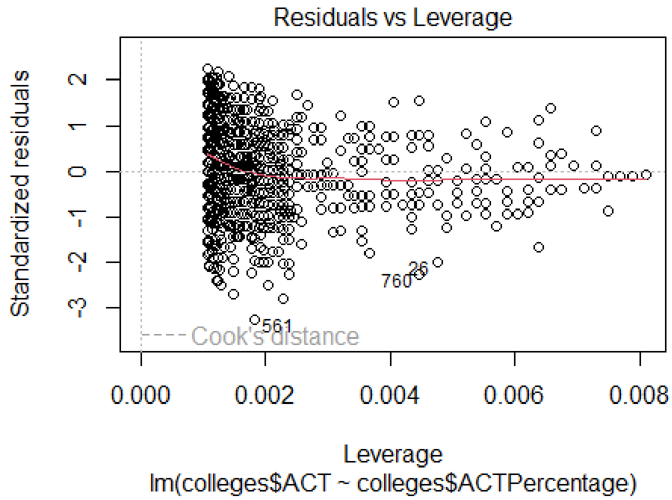
```
plot(model1)
```





Commented [PI10]: You did not include a summary or assessment of the first two residual plots.





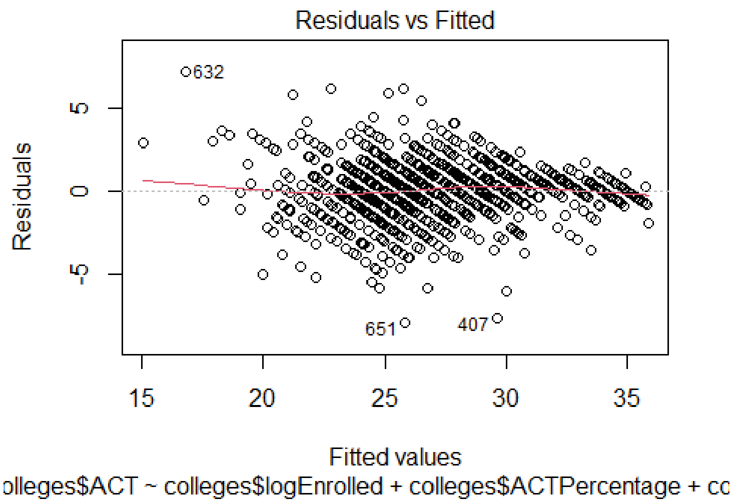
Now, let use all the numerical variables as predictors.

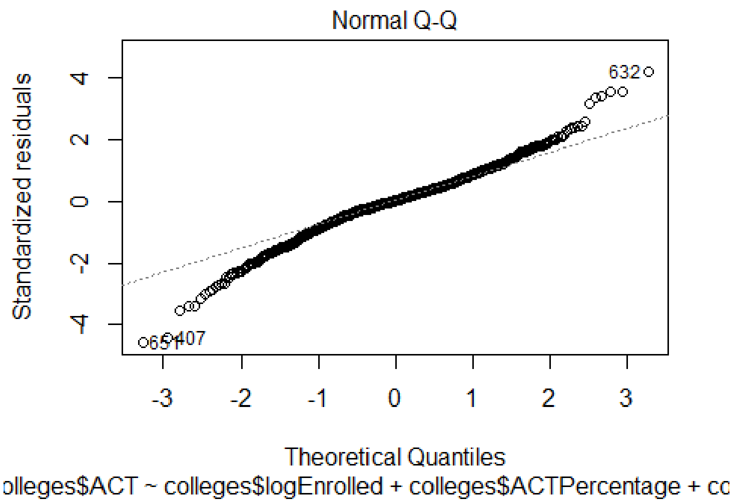
```
model2 = lm(colleges$ACT ~ colleges$logEnrolled + colleges$ACTPercentage + co
lleges$SATWR + colleges$SATM)
summary(model2)
```

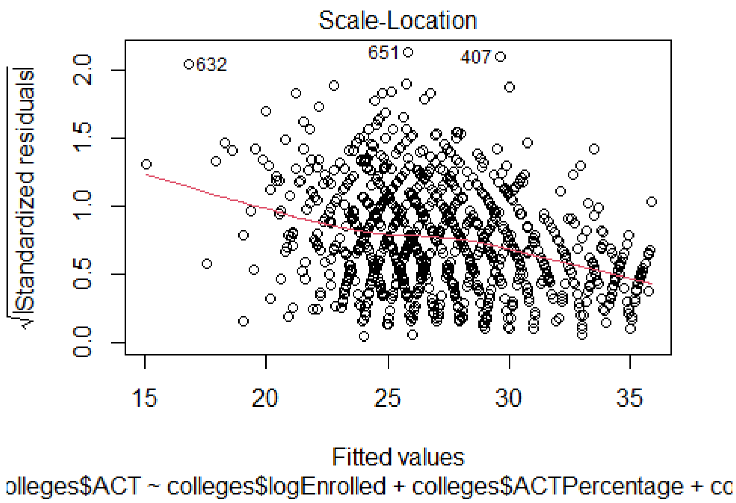
```
##
## Call:
## lm(formula = colleges$ACT ~ colleges$logEnrolled + colleges$ACTPercentage
+
##     colleges$SATWR + colleges$SATM)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8494 -0.8432  0.0414  0.9611  7.1673
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -6.915355   0.628696 -11.000  < 2e-16 ***
## colleges$logEnrolled    0.155204   0.053253   2.914  0.00365 **
## colleges$ACTPercentage -0.006243   0.002136  -2.923  0.00355 **
## colleges$SATWR          0.027678   0.002353  11.762  < 2e-16 ***
## colleges$SATM           0.025119   0.002045  12.285  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.726 on 932 degrees of freedom
## Multiple R-squared:  0.8144, Adjusted R-squared:  0.8136
## F-statistic:  1022 on 4 and 932 DF,  p-value: < 2.2e-16
```

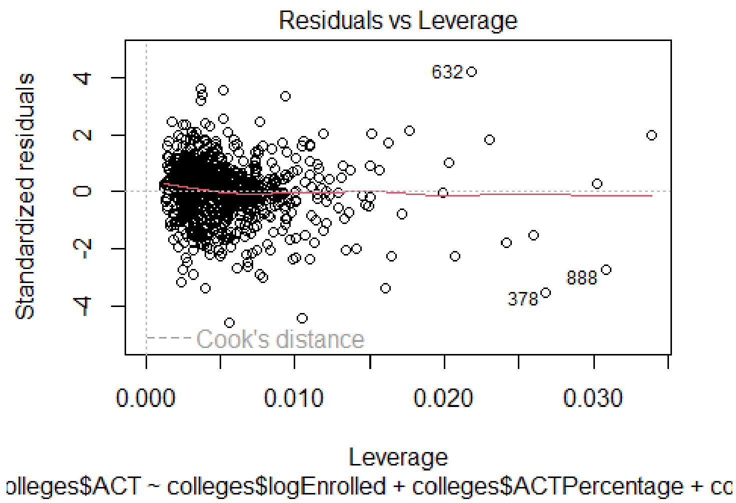
SAT Math and Writing/Reading were the most significant predictors in the model. They are also the most highly correlated with each other.

```
plot(model2)
```


Residuals vs Fitted
Residuals
5
0
-5
632
651
407
15
20
25
30
35
Fitted values
olleges$ACT ~ colleges$logEnrolled + colleges$ACTPercentage + cc


Normal Q-Q
Standardized residuals
4
2
0
-2
-4
632
651
407
-3
-2
-1
0
1
2
3
Theoretical Quantiles
olleges$ACT ~ colleges$logEnrolled + colleges$ACTPercentage + cc





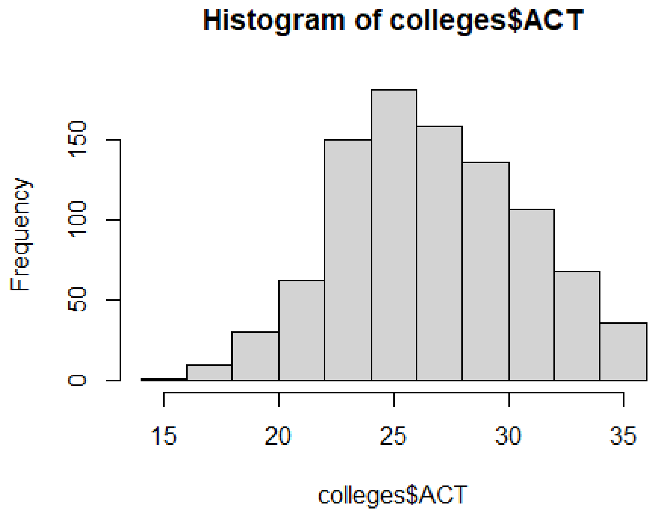
The residual variance seems pretty constant. Since none of the estimators are insignificant, we are going to stick with keeping them in the model.

Commented [PI11]: Yes, and you should have also commented on linearity in the first plot and normality in the second plot.

Commented [PI12]: Good decision.

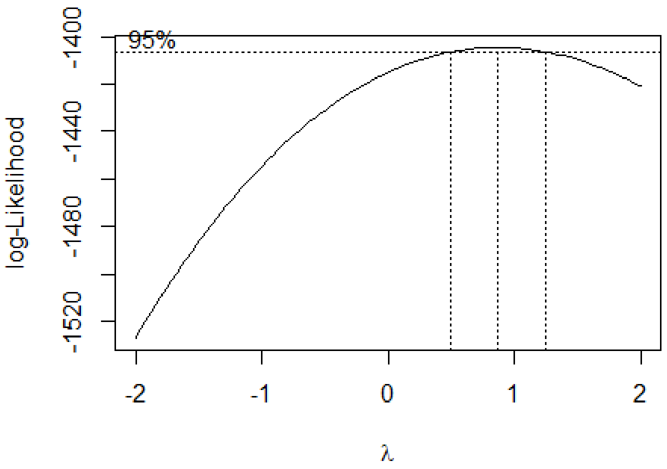
So far, in this first iteration, we have been pretty much focused on the predictor variables. We would now focus on the the dependent variable, the *75th ACT score*.

```
hist(colleges$ACT)
```



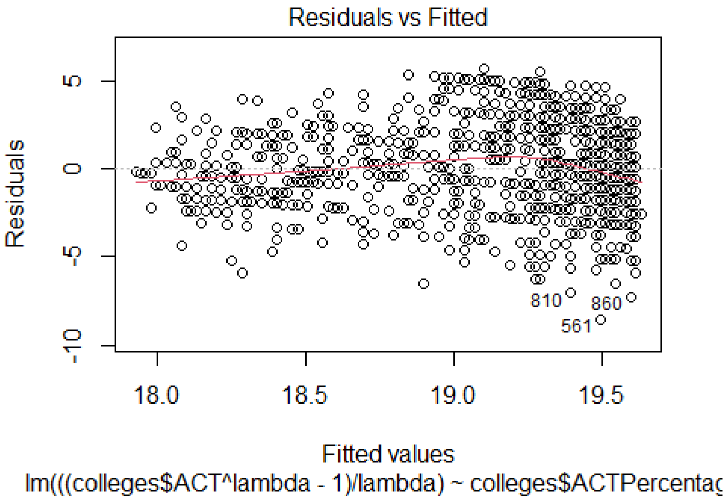
There is a slight *right skew*. Given the roughly apparent normality of the **ACT** variable. We will given the *kind of* shaky normality plot, let us do a Box Cox transformation.

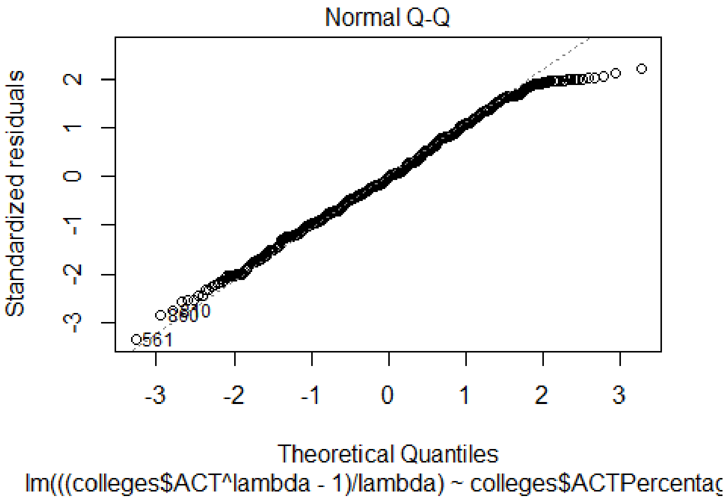
```
library(MASS)
bc = boxcox(model1)
```

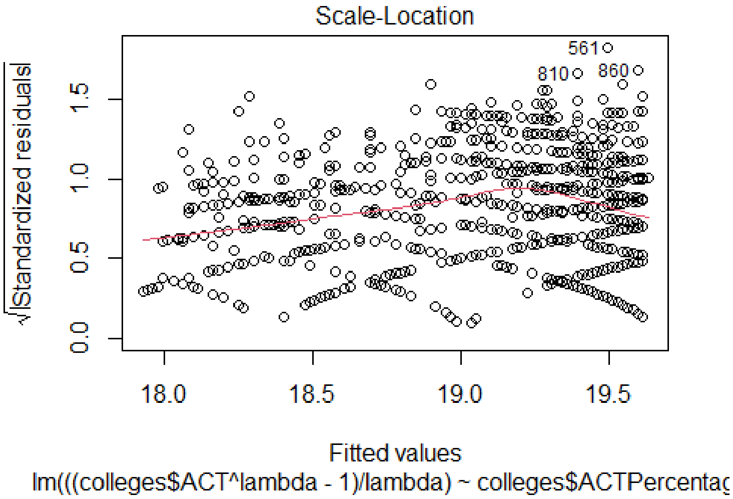


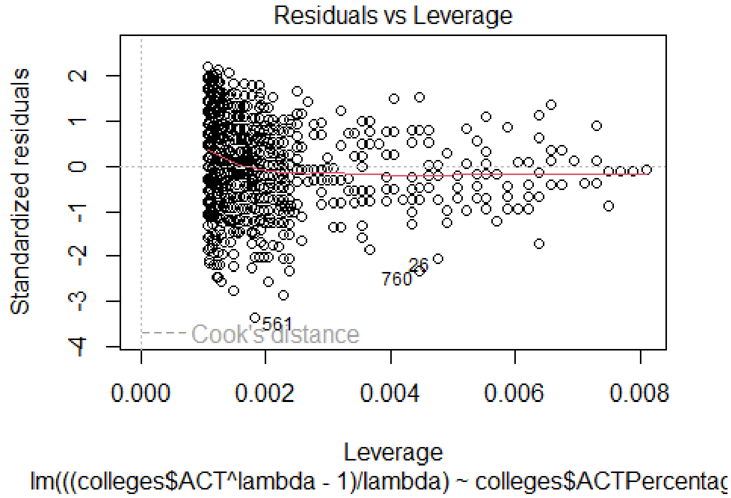
Commented [PI13]: The 95% confidence interval for the optimal lambda contains 1, so no transformation was needed.

```
lambda = bc$x[which.max(bc$y)]
model3 = lm(((colleges$ACT^lambda-1)/lambda) ~ colleges$ACTPercentage)
summary(model3)

##
## Call:
## lm(formula = ((colleges$ACT^lambda - 1)/lambda) ~ colleges$ACTPercentage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5449 -1.7865 -0.0396  1.9503  5.6335
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            19.630261   0.126897 154.694  < 2e-16 ***
## colleges$ACTPercentage -0.017040   0.003097  -5.501 4.87e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.555 on 935 degrees of freedom
## Multiple R-squared:  0.03135,    Adjusted R-squared:  0.03032
## F-statistic: 30.27 on 1 and 935 DF,  p-value: 4.865e-08

plot(model3)
```


Residuals vs Fitted
Residuals
5
0
-5
-10
18.0
18.5
19.0
19.5
810
860
561
Fitted values
lm(((colleges$ACT^lambda - 1)/lambda) ~ colleges$ACTPercentag


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
860
810
561
Theoretical Quantiles
lm(((colleges$ACT^lambda - 1)/lambda) ~ colleges$ACTPercentag





Analysis: The Box Cox transformation definitely helped. The Q-Q plot has improved, and so has the residuals-levarage graph.There is certainly a relationship between %of people submitting ACT scores, and the ACT scores of that cohort. Going forward we would like to investigate the interaction effect of different SAT Subjects. We would also like to label

Commented [PI14]: It made almost no difference compared to your first simple regression. I would expect this, since the optimal lambda was close to 1 (no transformation).

encode the differnt amount of test preparation, and observe the subsequent changes in a model that accounts for more variables,without multicollinearity within eaither of them.