# Final Project - STAT 4355.001

## *Expenditure Cost Efficiency on Public Education Analysis*

*Dat Tran*

May 8, 2020

## 1 Introduction

The data frame that I am going to explore is the data frame “sat”. This data frame originates from the package “faraway”, which is available inside the common R library. The “sat” data frame is a dataset about school expenditure and test scores from USA in 1994-95. It contains 50 rows and 7 columns, and initially the data in here were collected to study the relationship between expenditures on public education and test results. The value is labeled by the name of the state it originates from. The columns of the data are expend, ratio, salary, takers, verbal, math, and total. Expend column is the current expenditure per pupil in average daily attendance in public elementary and secondary school in 1994-95 (in thousands of dollars). Ratio column is the average pupil/teacher ratio of the following schools. Salary is the estimated average annual salary of teachers there. Takers are the percentage of all eligible students taking the SAT. Verbal, Math, and Total columns are average verbal, math, total SAT score, respectively, of the students in those following schools. For the purpose of this project, I use expend as the response variable (y), and it will be followed up by 5 regressors (x1, x2, x3, x4, x5) which are ratio, salary, takers, verbal, and math, respectively. I decide not to use the last variable, total, because it is a dependent variable to verbal and math already, as total is equal to verbal plus math.

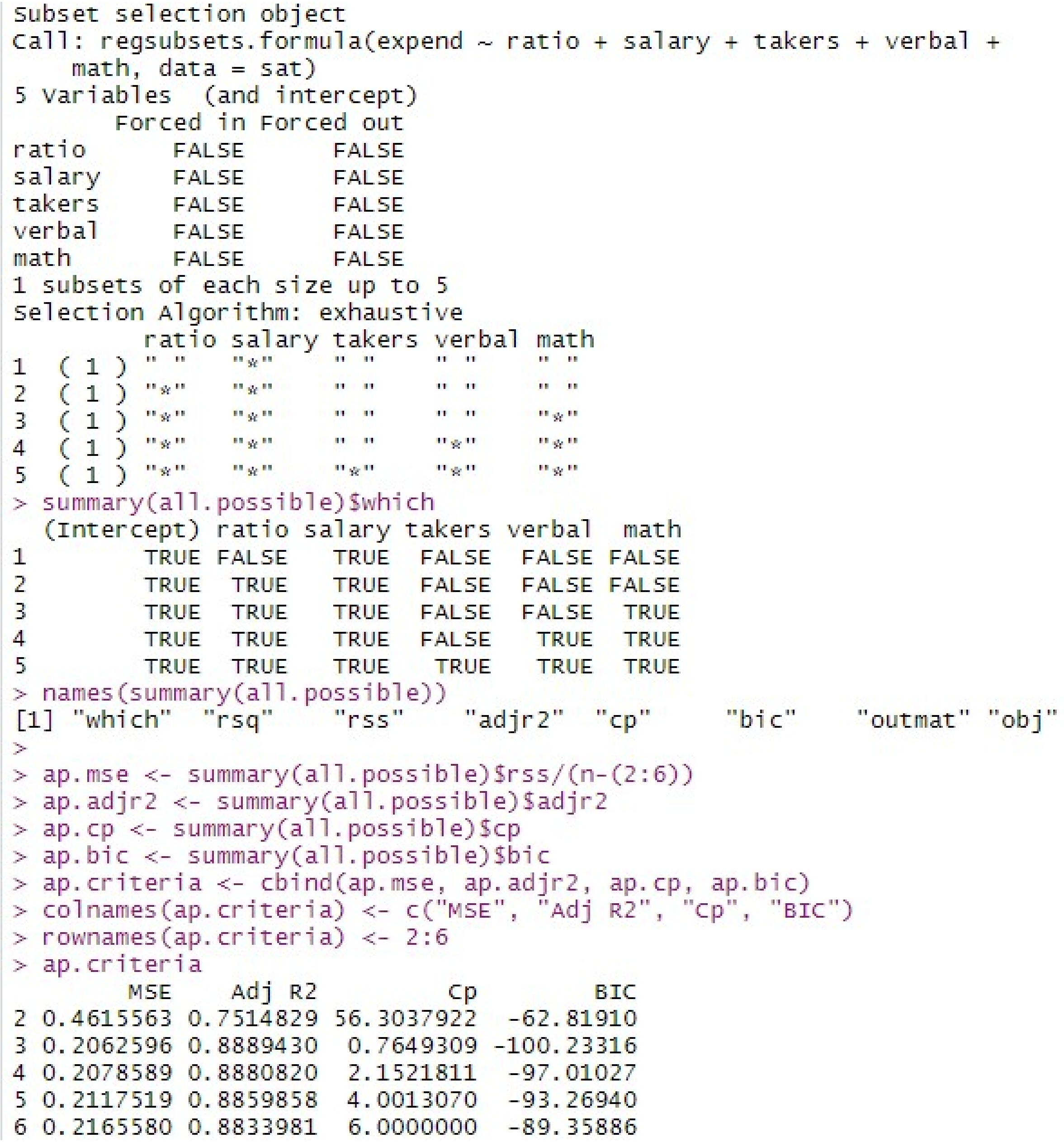
My goal with this data frame is to find out the connection between the expenditure cost and every other regressor variables. I want to know if the expenditure directly links with the ratio of teacher, and the salary they will receive, as well as the test result from the students themselves. This information is very crucial for both educators and parents, as well as their children. Educators can use this information to make necessary change to the educational systems, while parents can use this information to decide whether the cost they spend for their children worth it, and how much value their children will receive back from the system.

The best way for me to analyze this data frame is through linear regression modeling. My report will contain all 4 essential part of linear regression analysis: model fitting, residual analysis, model selection and transformation on data. For model selection, I will use all possible regressions, and backed up by a stepwise selection. And if the transformation is necessary, I will also make changes to variable in the data frame. At the end of my report, I will make recommendations based on what I have done to analyze the data frame, as well as my self-reflection on what went well in the project and what I need to improve to better my analysis and report.

# 2 Data Analysis with original data

## 2.1 All possible regressions with regsubsets

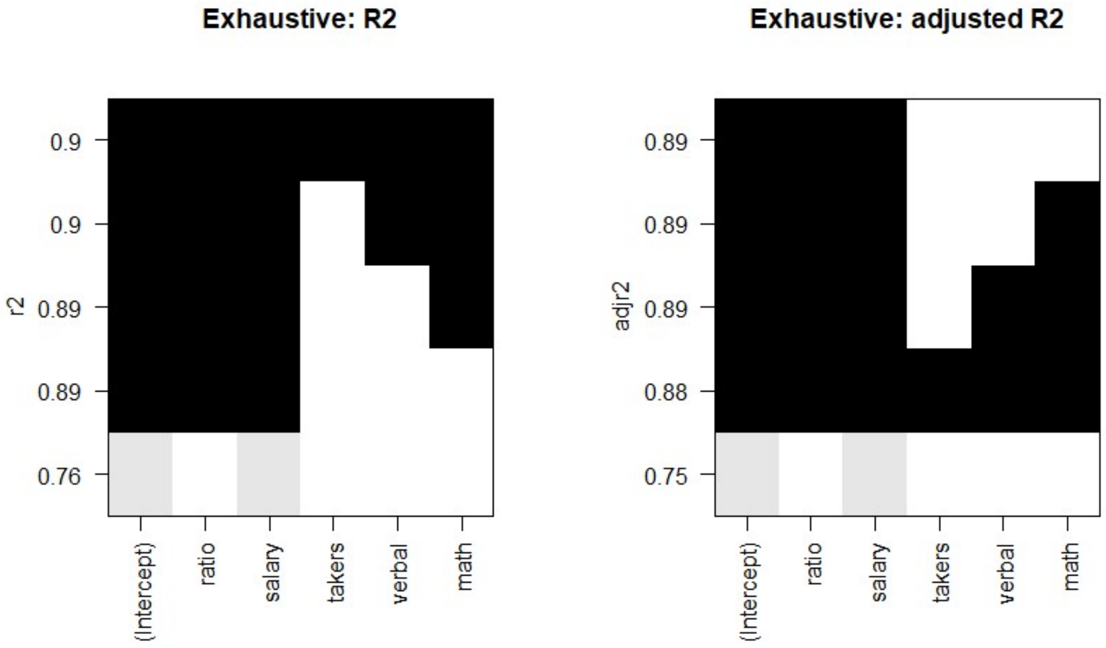
I began with loading the leaps package and run regsubsets command to be able to have a general view of the data. Then I created a table of selection criteria based on the result of regsubsets command on the data frame.



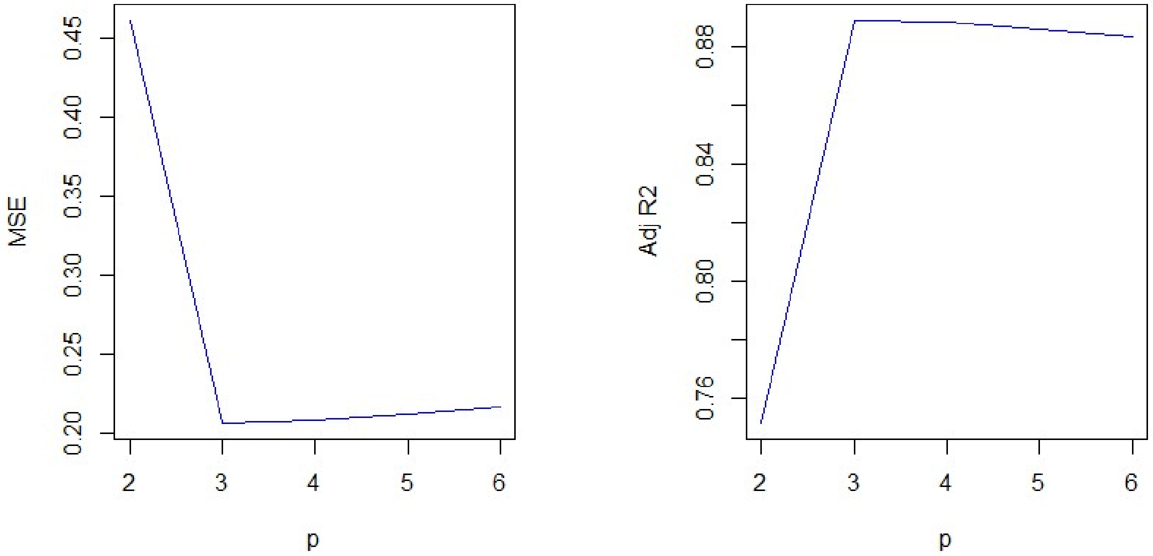
```
Subset selection object
Call: regsubsets.formula(expend ~ ratio + salary + takers + verbal +
    math, data = sat)
5 Variables  (and intercept)
       Forced in Forced out
ratio      FALSE      FALSE
salary     FALSE      FALSE
takers     FALSE      FALSE
verbal     FALSE      FALSE
math       FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: exhaustive
         ratio salary takers verbal math
1  ( 1 ) " "   "*"    " "    " "    " " 
2  ( 1 ) "*"   "*"    " "    " "    " " 
3  ( 1 ) "*"   "*"    " "    " "    "*" 
4  ( 1 ) "*"   "*"    " "    "*"    "*" 
5  ( 1 ) "*"   "*"    "*"    "*"    "*" 
> summary(all.possible)$which
  (Intercept) ratio salary takers verbal  math
1        TRUE FALSE   TRUE  FALSE  FALSE FALSE
2        TRUE  TRUE   TRUE  FALSE  FALSE FALSE
3        TRUE  TRUE   TRUE  FALSE  FALSE  TRUE
4        TRUE  TRUE   TRUE  FALSE   TRUE  TRUE
5        TRUE  TRUE   TRUE   TRUE   TRUE  TRUE
> names(summary(all.possible))
[1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"   
> 
> ap.mse <- summary(all.possible)$rss/(n-(2:6))
> ap.adjr2 <- summary(all.possible)$adjr2
> ap.cp <- summary(all.possible)$cp
> ap.bic <- summary(all.possible)$bic
> ap.criteria <- cbind(ap.mse, ap.adjr2, ap.cp, ap.bic)
> colnames(ap.criteria) <- c("MSE", "Adj R2", "Cp", "BIC")
> rownames(ap.criteria) <- 2:6
> ap.criteria
        MSE    Adj R2         Cp        BIC
2 0.4615563 0.7514829 56.3037922  -62.81910
3 0.2062596 0.8889430  0.7649309 -100.23316
4 0.2078589 0.8880820  2.1521811  -97.01027
5 0.2117519 0.8859858  4.0013070  -93.26940
6 0.2165580 0.8833981  6.0000000  -89.35886
```

In order to understand the selection criteria better, I decided to visualize them using plot. The plots covered R2, adjusted R2, MSE, Cp, and BIC.

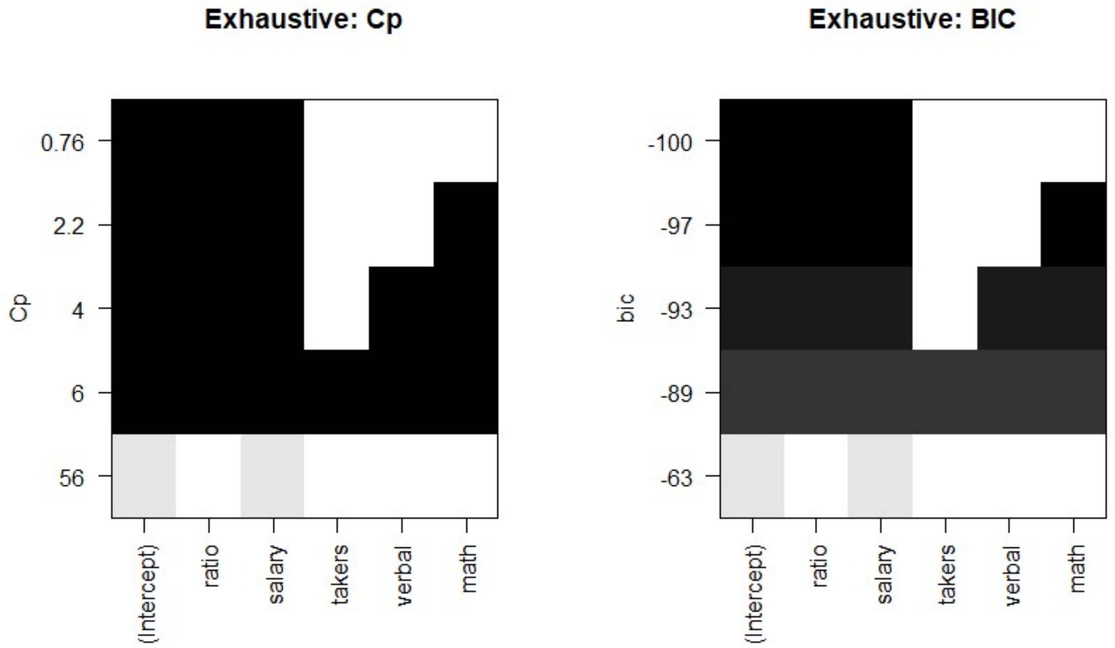
R2 and Adjusted R2 plots for model selection:



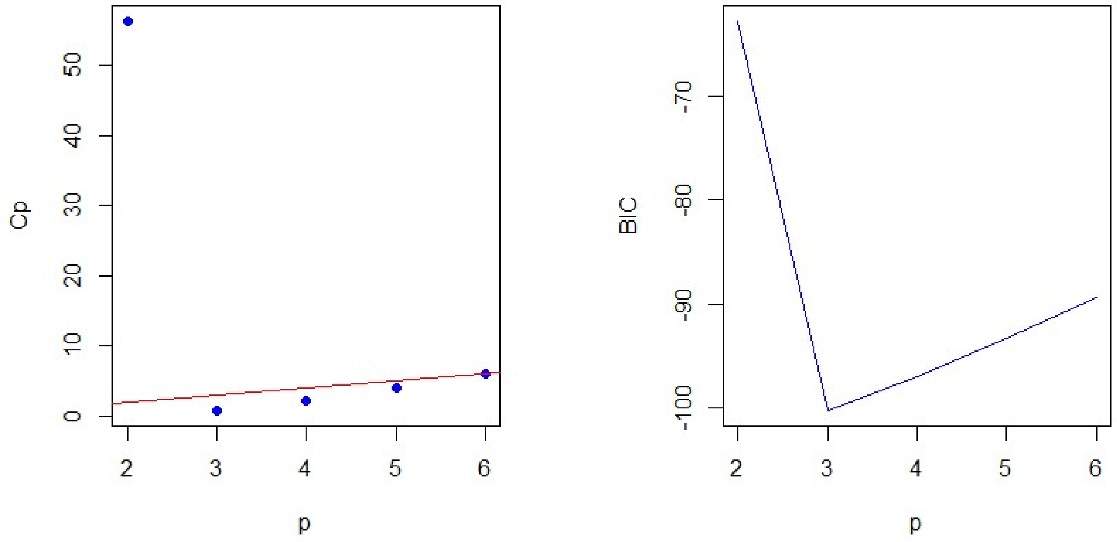
MSE and Adjusted R2 plots for regressor variable number selection (p):



Cp and BIC plots for model selection:

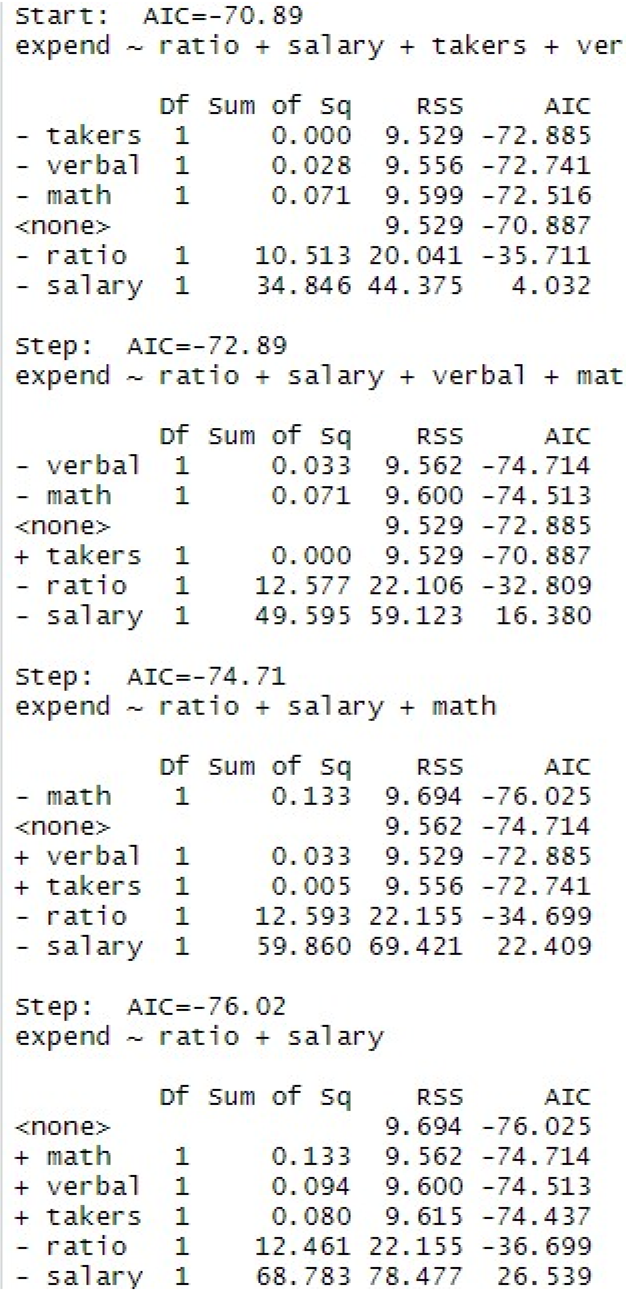


Cp and BIC plots for regressor variable number selection:



As we can see from the graphs above, in model selection graphs, all four graphs agree that ratio and salary seem to be the 2 variables that fit through all the selection, followed up by math and verbal, and finally taker. As for the regressor variable number selection (p), the best p that satisfies the most is p = 3, where there are 2 regressor variables. Based on what we have observed, I chose ratio and salary to fit in the model.

After using regsubsets to have a general idea of the data, as well as choosing the model I want to use, I used stepwise selection to check my resulted model again.

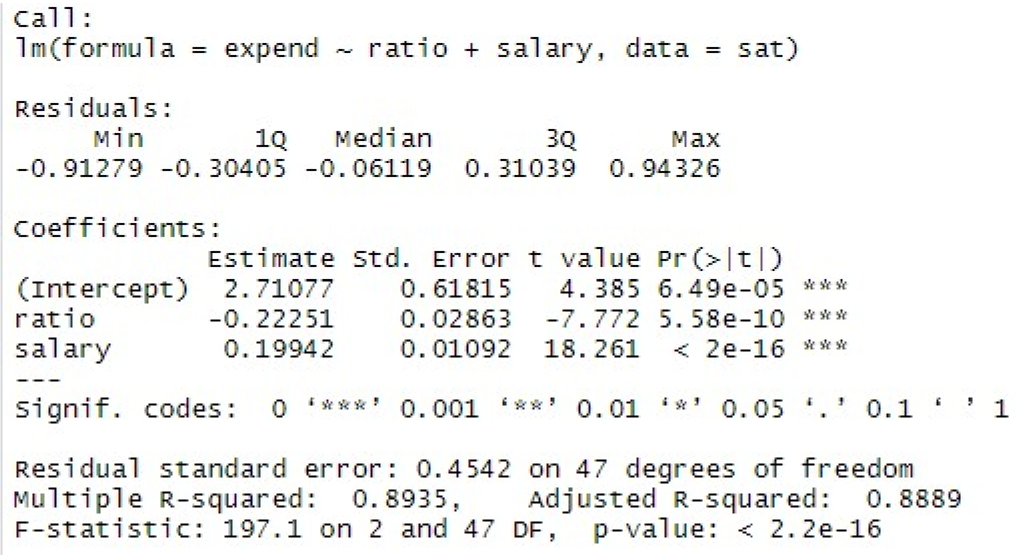


```
Start:  AIC=-70.89
expend ~ ratio + salary + takers + verbal + math

         Df Sum of Sq    RSS     AIC
- takers  1     0.000  9.529 -72.885
- verbal  1     0.028  9.556 -72.741
- math    1     0.071  9.599 -72.516
<none>                 9.529 -70.887
- ratio   1    10.513 20.041 -35.711
- salary  1    34.846 44.375   4.032

Step:  AIC=-72.89
expend ~ ratio + salary + verbal + math

         Df Sum of Sq    RSS     AIC
- verbal  1     0.033  9.562 -74.714
- math    1     0.071  9.600 -74.513
<none>                 9.529 -72.885
+ takers  1     0.000  9.529 -70.887
- ratio   1    12.577 22.106 -32.809
- salary  1    49.595 59.123  16.380

Step:  AIC=-74.71
expend ~ ratio + salary + math

         Df Sum of Sq    RSS     AIC
- math    1     0.133  9.694 -76.025
<none>                 9.562 -74.714
+ verbal  1     0.033  9.529 -72.885
+ takers  1     0.005  9.556 -72.741
- ratio   1    12.593 22.155 -34.699
- salary  1    59.860 69.421  22.409

Step:  AIC=-76.02
expend ~ ratio + salary

         Df Sum of Sq    RSS     AIC
<none>                 9.694 -76.025
+ math    1     0.133  9.562 -74.714
+ verbal  1     0.094  9.600 -74.513
+ takers  1     0.080  9.615 -74.437
- ratio   1    12.461 22.155 -36.699
- salary  1    68.783 78.477  26.539
```

The stepwise selection also agreed with choosing ratio and salary as 2 regressor variables for expend. We will move on model fitting and residual analysis for this model.

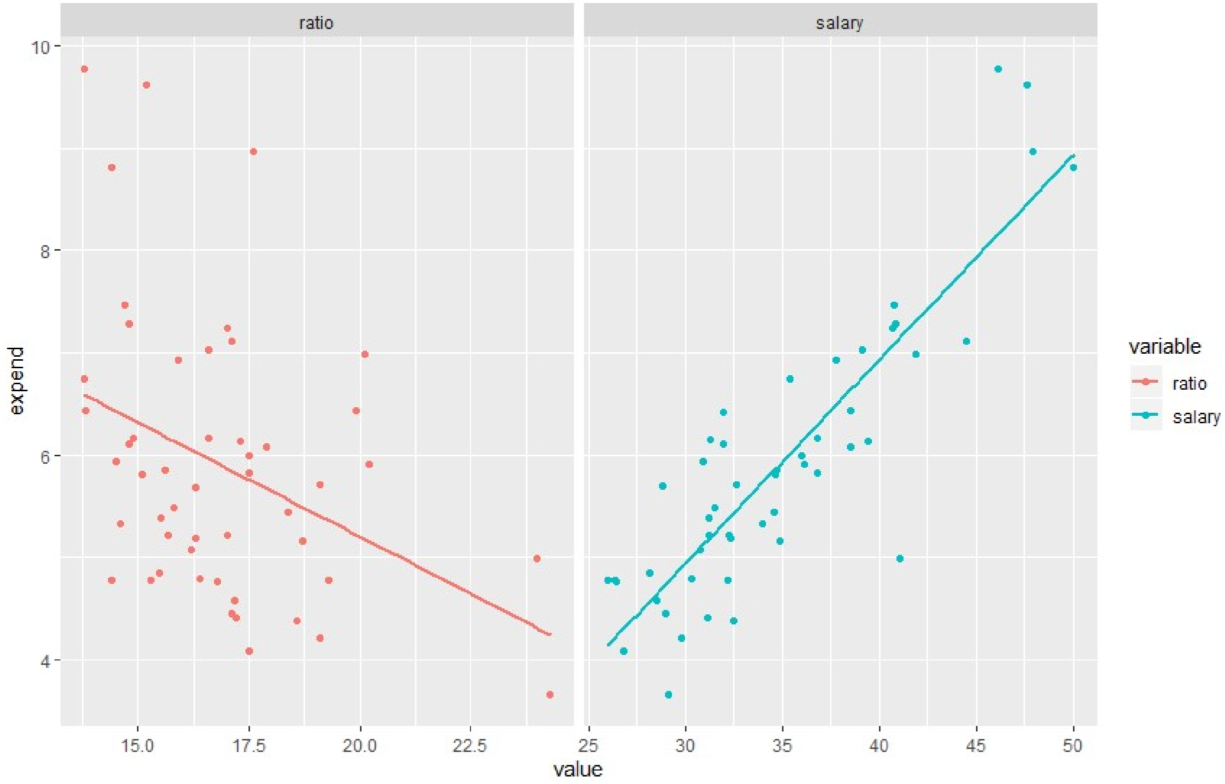
## 2.2 Model fitting and residual analysis

I began with fitting the model using expend as response variable, and ratio and salary as regressor variables.



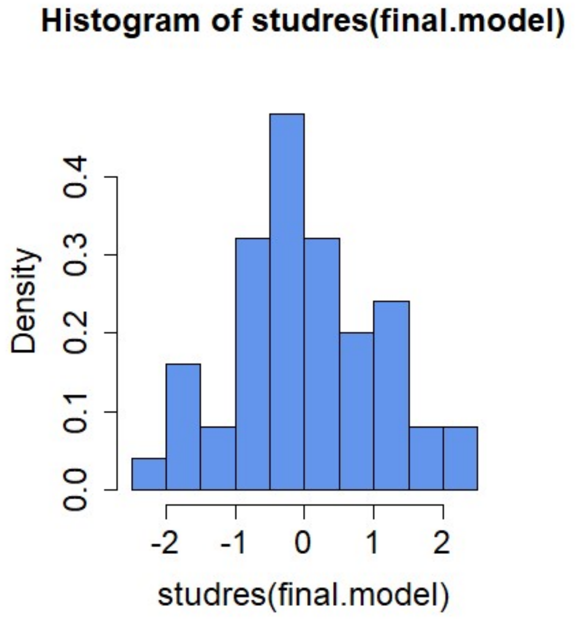
```
Call:
lm(formula = expend ~ ratio + salary, data = sat)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.91279 -0.30405 -0.06119  0.31039  0.94326 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.71077    0.61815   4.385 6.49e-05 ***
ratio       -0.22251    0.02863  -7.772 5.58e-10 ***
salary       0.19942    0.01092  18.261  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4542 on 47 degrees of freedom
Multiple R-squared:  0.8935,	Adjusted R-squared:  0.8889 
F-statistic: 197.1 on 2 and 47 DF,  p-value: < 2.2e-16
```

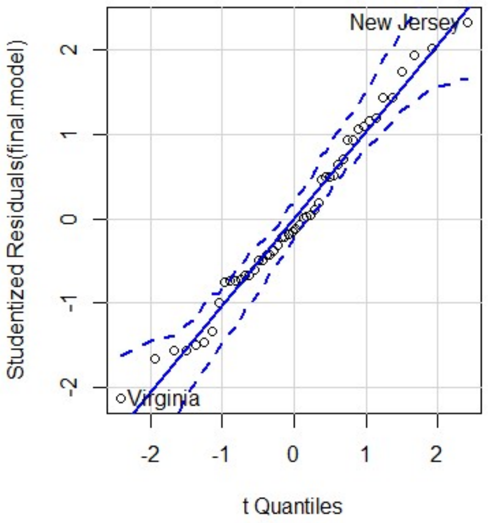
Then I plotted expend against ratio and salary, respectively



As we can see here, the linear regression model of expend against ratio did not fit really well to the best fit line, and transformation may be needed here. The model of expend against salary fit well with the best fit line, and only had 3 minor possible outliners.

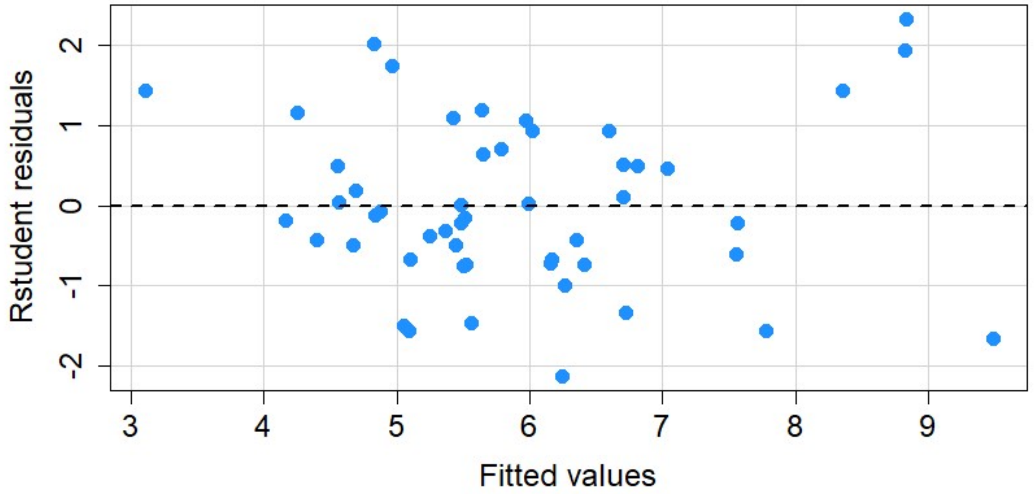
The next is the normal probability plot of the residuals.





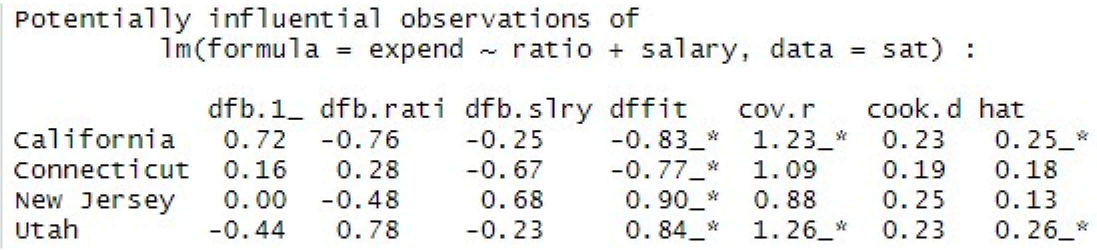
There are only minor problems with normality assumptions with the QQplot, with 2 possible outliners are Virginia (46) and New Jersey (30).

Next, I plotted the residual plot of the studentized residuals versus against the fitted values.

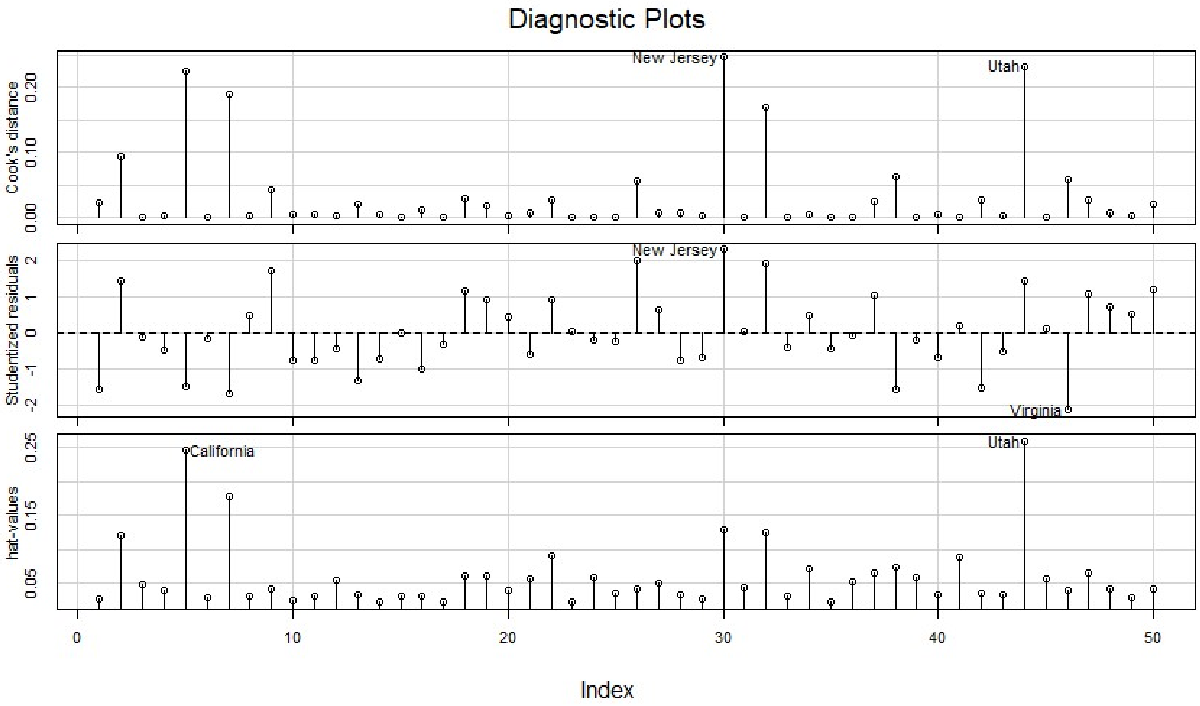


All values of the residual plot are between -2 and 2. The plot did not reveal any serious problem with inequality of variance. We did not see any potential outliners.

Finally, I performed influence analysis



```
Potentially influential observations of
	 lm(formula = expend ~ ratio + salary, data = sat) :

            dfb.1_ dfb.rati dfb.slry dffit   cov.r   cook.d hat  
California   0.72  -0.76    -0.25    -0.83_*  1.23_*  0.23   0.25_*
Connecticut  0.16   0.28    -0.67    -0.77_*  1.09    0.19   0.18  
New Jersey   0.00  -0.48     0.68     0.90_*  0.88    0.25   0.13  
Utah        -0.44   0.78    -0.23     0.84_*  1.26_*  0.23   0.26_*
```



Through influence analysis, we can see that all four potentially influential observations have influence on single fitted value. California and Utah potentially impact precision of estimation. They are also possible leverage points.

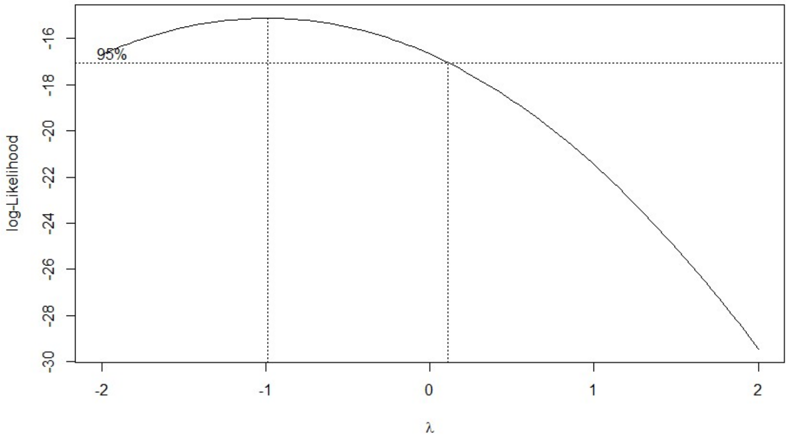
Throughout all possible regressions, model fitting and residual analysis, even though it may need a minor transformation for the linear regression model of expend against ratio, the model is acceptable. We can see that the more expenditure the schools have, they are likely to have lower pupil/teacher ratio, and the teacher will likely be paid higher. This will reduce the amount of workload each teacher has and increase their salary. As a result, teaching performance will most likely be increased as the expenditure goes up. However, this could only answer the first part of my question. I still need to answer how expenditure affects students' performance. In order to do this, transformation is needed. I will move on to next part, transformation and its data analysis.

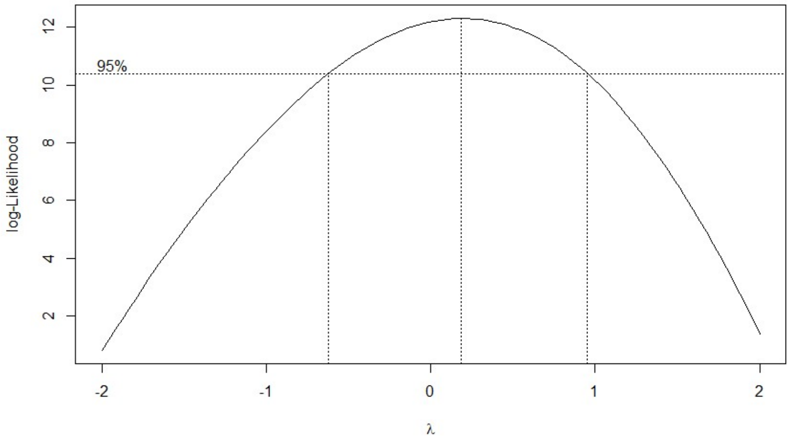
## 3 Transformation and Data Analysis with transformed data

### 3.1 Transformation with boxcox log-likelihood method

To address the problem that the model could not fit all variables in to get the answer we need, a transformation is needed. I began with calling the boxcox log-likelihood method for all 5 possible regressor variables (from top to bottom: ratio, salary, takers, verbal and math, respectively)

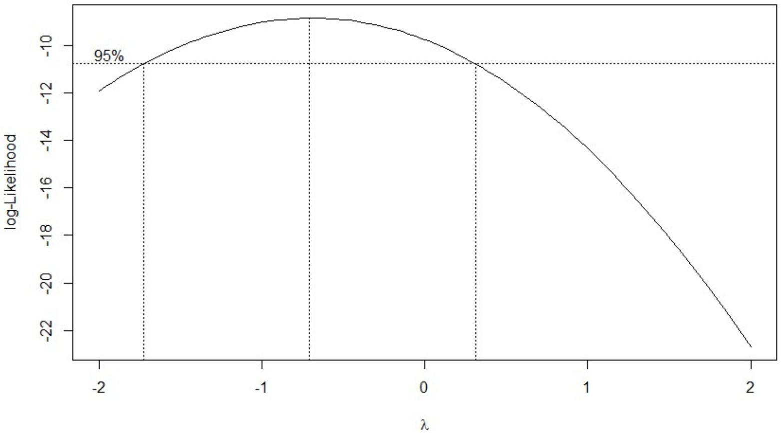
Ratio and Salary log-likelihood boxcox plots (Teaching performance/quality variables)

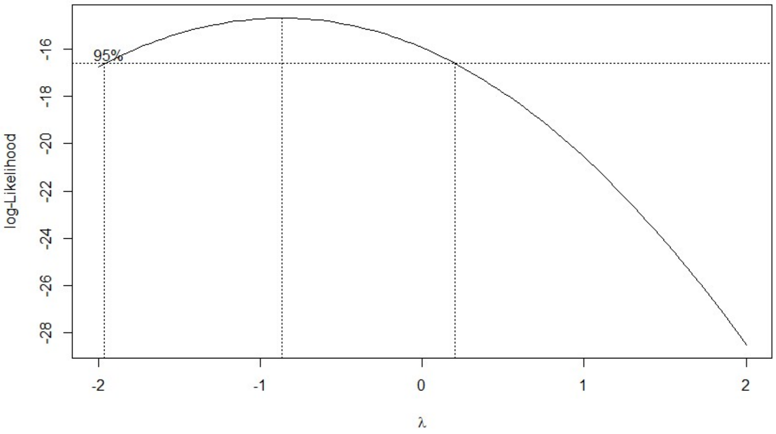


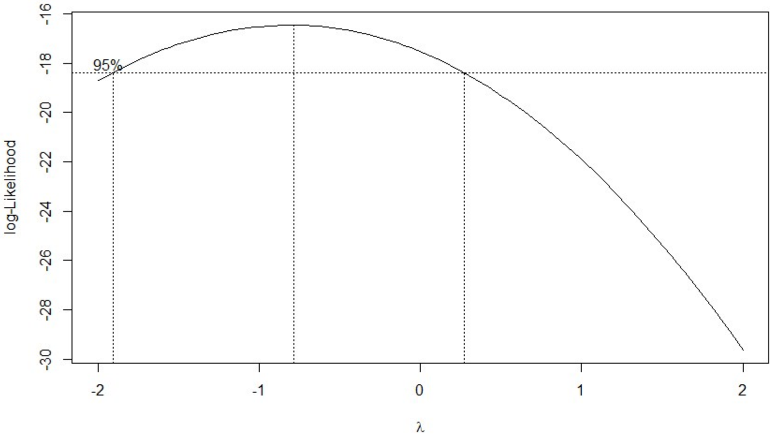


The log-likelihood of ratio has a range from -2 to 0, with the peak in -1, while the log-likelihood of salary stays between -0.75 to -1, peak in around 0.18.

Takers and Verbal and Math log-likelihood boxcox plots (Students' performance variables)







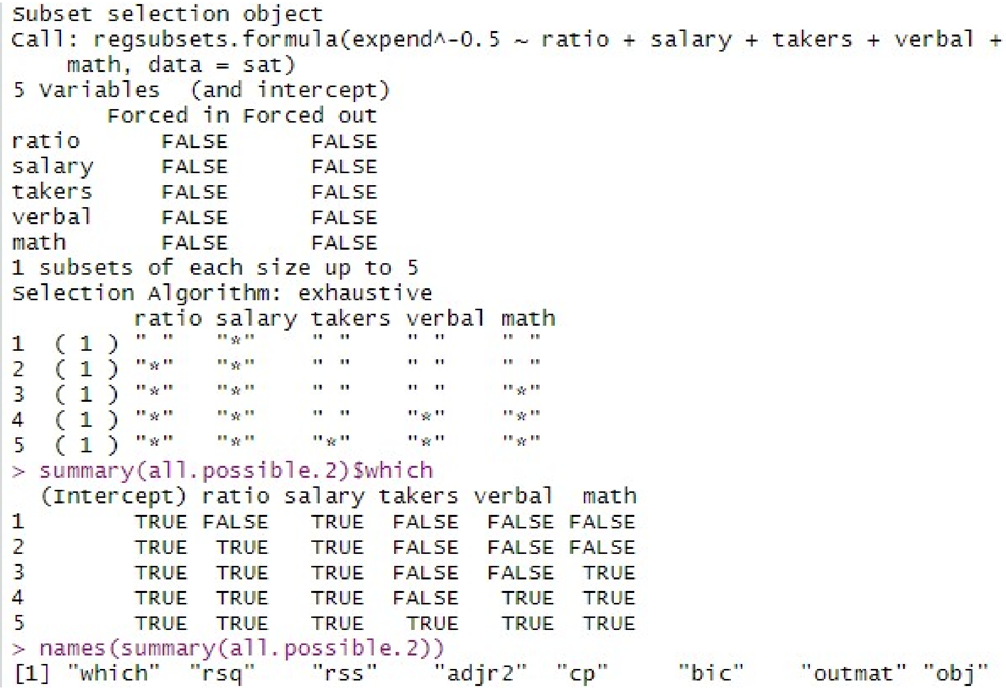
We can see that even though their intervals are different, the acceptable values always lie around -0.5. So I took -0.5 as the boxcox value for my transformation. The new model would be expend ^ -0.5 as response variable, and its possible regressor variables stayed the same.

After this I performed data analysis for the transformed data in the same way with the original data.

## 3.2 Data analysis with transformed data

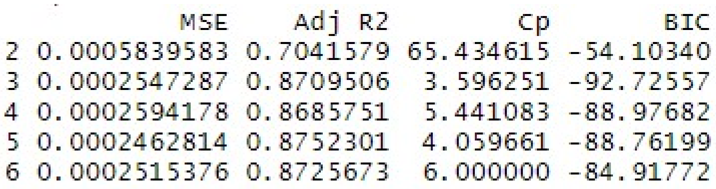
### 3.2.1 All possible regression with regsubsets

First, I ran regsubsets command on the newly transformed data frame to have a clear view of the data



```
Subset selection object
Call: regsubsets.formula(expend^-0.5 ~ ratio + salary + takers + verbal +
    math, data = sat)
5 Variables  (and intercept)
       Forced in Forced out
ratio      FALSE      FALSE
salary     FALSE      FALSE
takers     FALSE      FALSE
verbal     FALSE      FALSE
math       FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: exhaustive
         ratio salary takers verbal math
1  ( 1 ) " "   "*"    " "    " "    " " 
2  ( 1 ) "*"   "*"    " "    " "    " " 
3  ( 1 ) "*"   "*"    " "    " "    "*" 
4  ( 1 ) "*"   "*"    " "    "*"    "*" 
5  ( 1 ) "*"   "*"    "*"    "*"    "*" 
> summary(all.possible.2)$which
  (Intercept) ratio salary takers verbal  math
1        TRUE FALSE   TRUE  FALSE  FALSE FALSE
2        TRUE  TRUE   TRUE  FALSE  FALSE FALSE
3        TRUE  TRUE   TRUE  FALSE  FALSE  TRUE
4        TRUE  TRUE   TRUE  FALSE   TRUE  TRUE
5        TRUE  TRUE   TRUE   TRUE   TRUE  TRUE
> names(summary(all.possible.2))
[1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

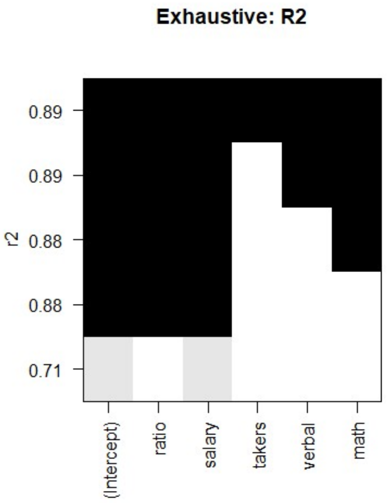
Next, I created the table of selection criteria based on the result of regsubsets command.

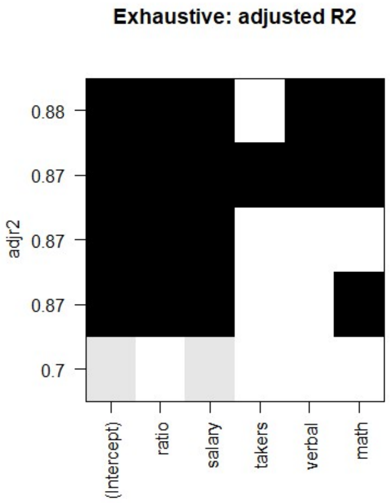


```
            MSE    Adj R2        Cp       BIC
2 0.0005839583 0.7041579 65.434615 -54.10340
3 0.0002547287 0.8709506  3.596251 -92.72557
4 0.0002594178 0.8685751  5.441083 -88.97682
5 0.0002462814 0.8752301  4.059661 -88.76199
6 0.0002515376 0.8725673  6.000000 -84.91772
```

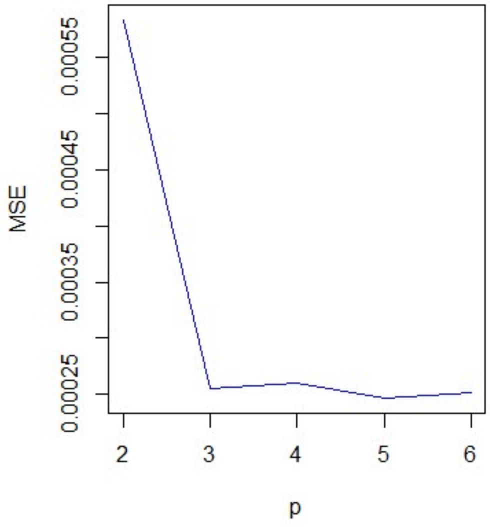
To compare with the original data frame, MSE is significantly smaller, Cp and BIC is noticeably smaller and Adjusted R2 stays around the same with the original set.

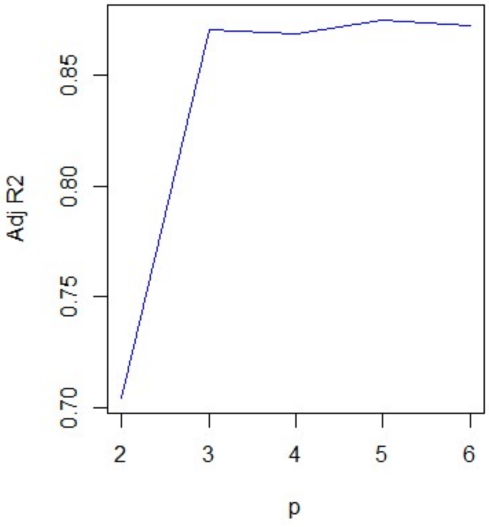
Then I visualized those index to understand them better, started with R2 and adjusted R2:



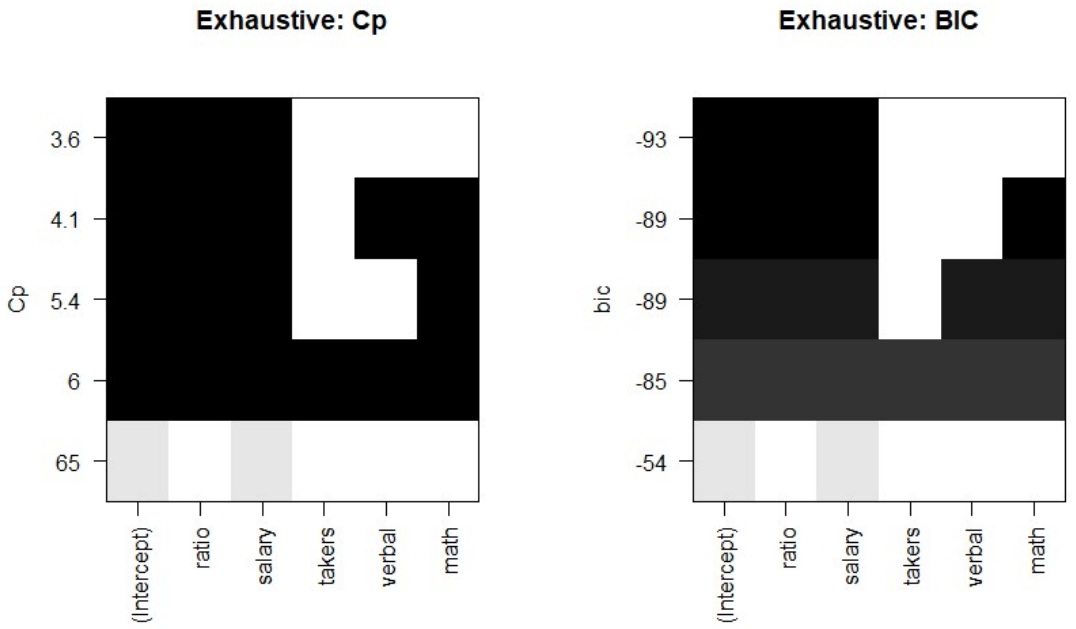


Then the p selection plots of MSE and Adjusted R2:

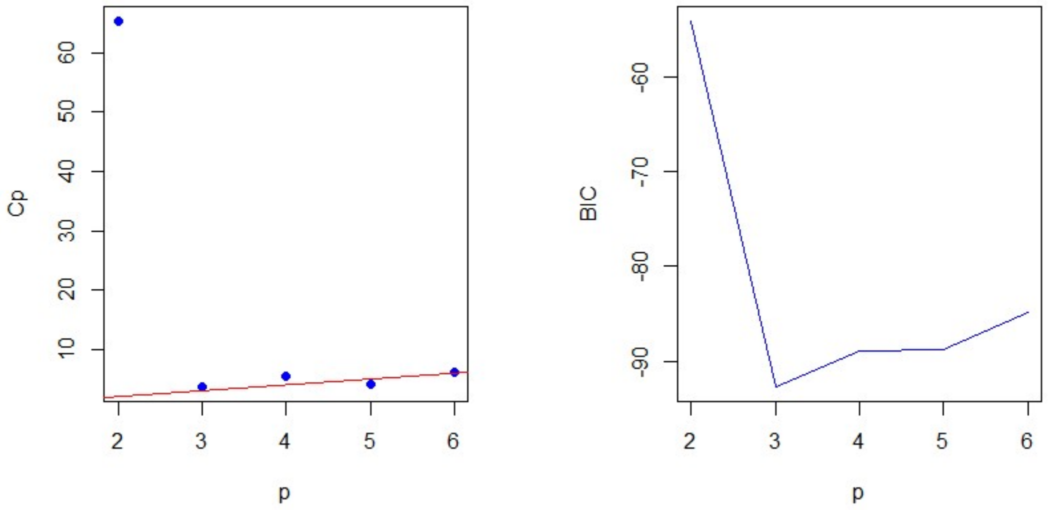




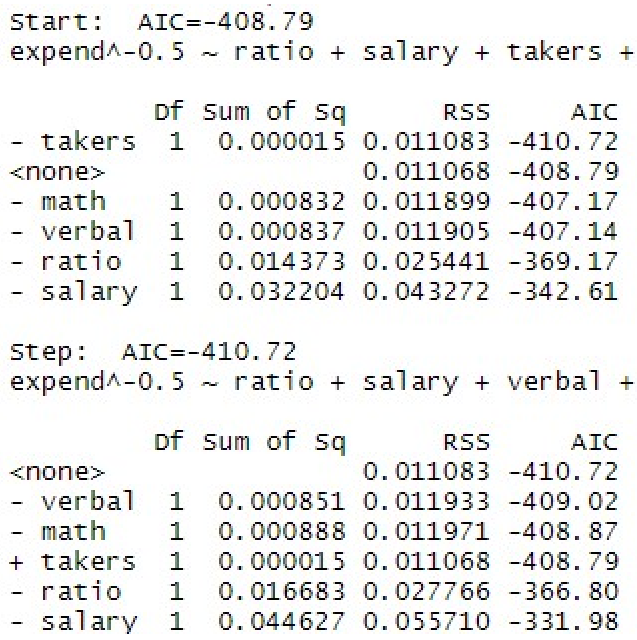
Next is model selection plots of Cp and BIC:



Last one is the p selection plots for Cp and BIC:



As we can see here, while takers still did not fit with the model, math and verbal did fit much better than they used to be in original data. Moreover, the p selection graphs also suggested us to choose a larger p. In this case I chose p = 5, which there are 4 regressor variables (ratio, salary, verbal, and math). To further support my decision, I performed stepwise analysis on the newly transformed model:

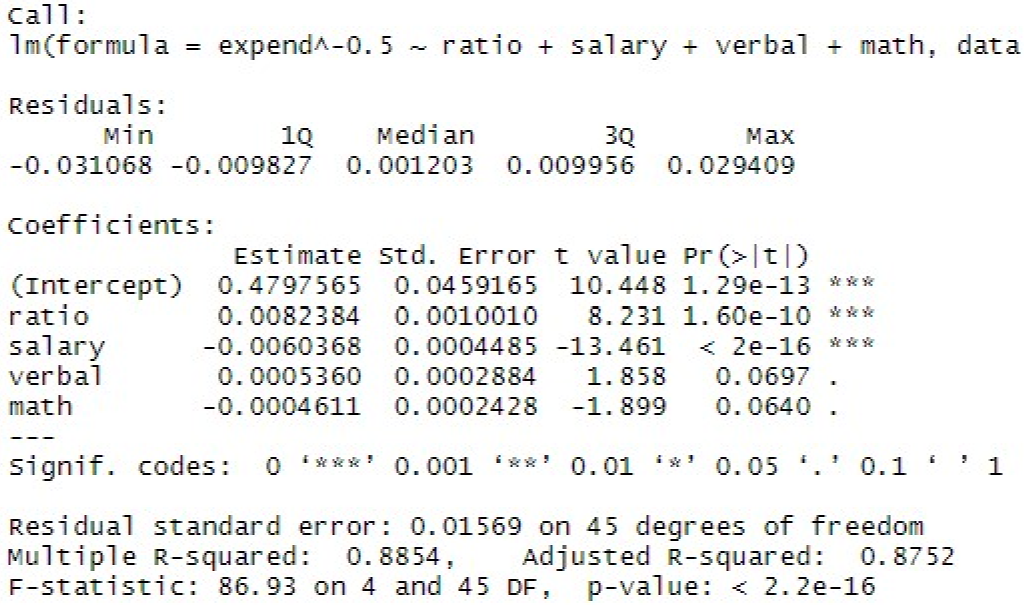


```
Start:  AIC=-408.79
expend^-0.5 ~ ratio + salary + takers + verbal + math

         Df Sum of Sq      RSS     AIC
- takers  1  0.000015 0.011083 -410.72
<none>                0.011068 -408.79
- math    1  0.000832 0.011899 -407.17
- verbal  1  0.000837 0.011905 -407.14
- ratio   1  0.014373 0.025441 -369.17
- salary  1  0.032204 0.043272 -342.61

Step:  AIC=-410.72
expend^-0.5 ~ ratio + salary + verbal + math

         Df Sum of Sq      RSS     AIC
<none>                0.011083 -410.72
- verbal  1  0.000851 0.011933 -409.02
- math    1  0.000888 0.011971 -408.87
+ takers  1  0.000015 0.011068 -408.79
- ratio   1  0.016683 0.027766 -366.80
- salary  1  0.044627 0.055710 -331.98
```

The stepwise selection also agreed with my 4 regressor variables. We will move on next part, model fitting and residual analysis.

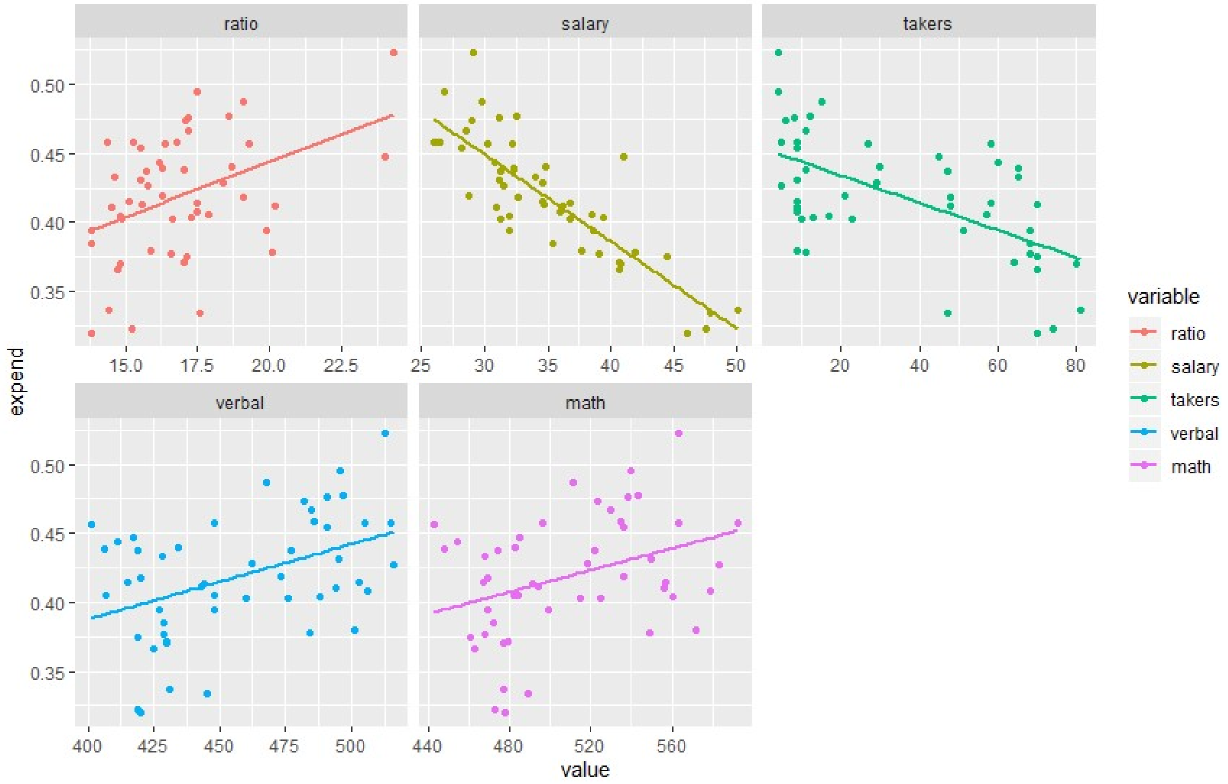
### 3.2.2 Model fitting and residual analysis

I began with fitting the model I selected from last part, using expend as response variable and ratio, salary, verbal, and math are regressor variables:



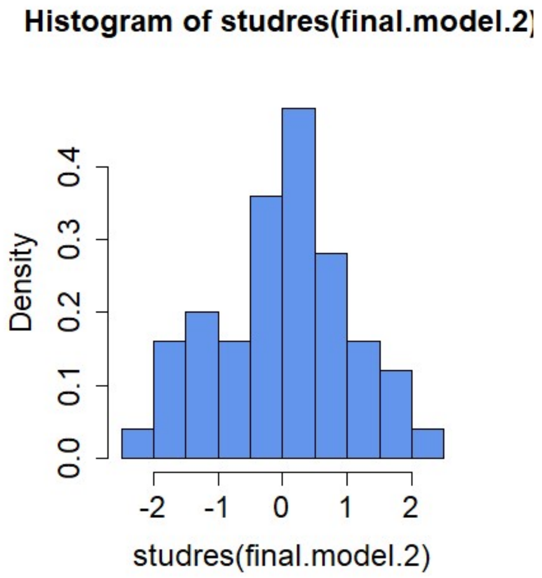
```
Call:
lm(formula = expend^-0.5 ~ ratio + salary + verbal + math, data = sat)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.031068 -0.009827  0.001203  0.009956  0.029409 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.4797565  0.0459165  10.448 1.29e-13 ***
ratio        0.0082384  0.0010010   8.231 1.60e-10 ***
salary      -0.0060368  0.0004485 -13.461  < 2e-16 ***
verbal       0.0005360  0.0002884   1.858   0.0697 .  
math        -0.0004611  0.0002428  -1.899   0.0640 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01569 on 45 degrees of freedom
Multiple R-squared:  0.8854,	Adjusted R-squared:  0.8752 
F-statistic: 86.93 on 4 and 45 DF,  p-value: < 2.2e-16
```

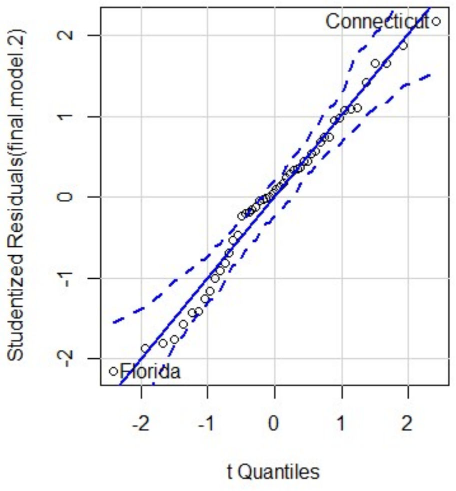
Then I plotted expend with every regressor variables:



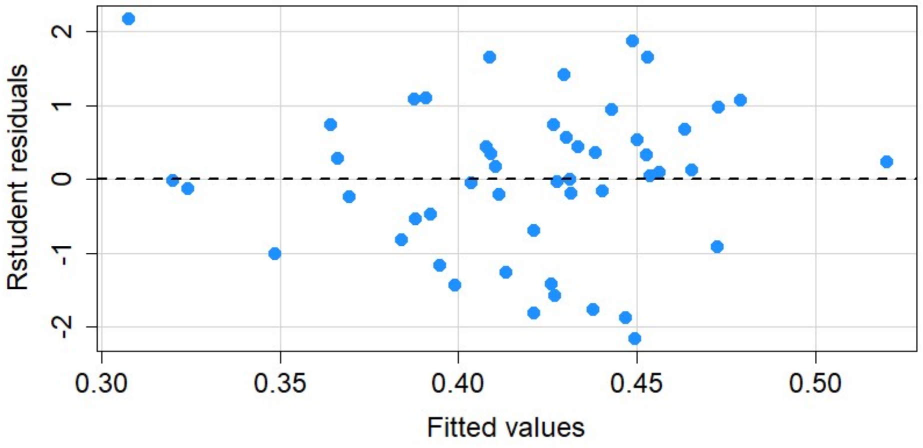
As we can see here, ratio and salary fitted much better in the model now. With the exception of takers, all 4 regressor variables fitted the model and did not need further transformation. However, there is an interesting finding here. The verbal and math score increase as expend ^-0.5 increases. This means that as expenditure cost increases, the verbal and math score of the student decreases, and the students' performance as well.

For the next part, I constructed a normal probability plot of the residuals and a residual plot:



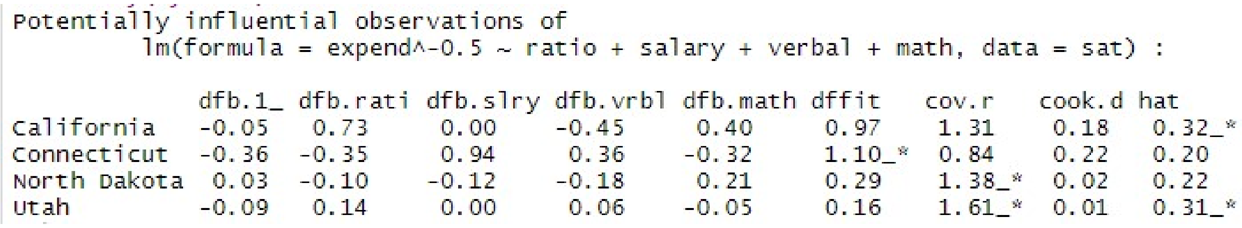


There is no problem with normality assumption in the QQplot here.

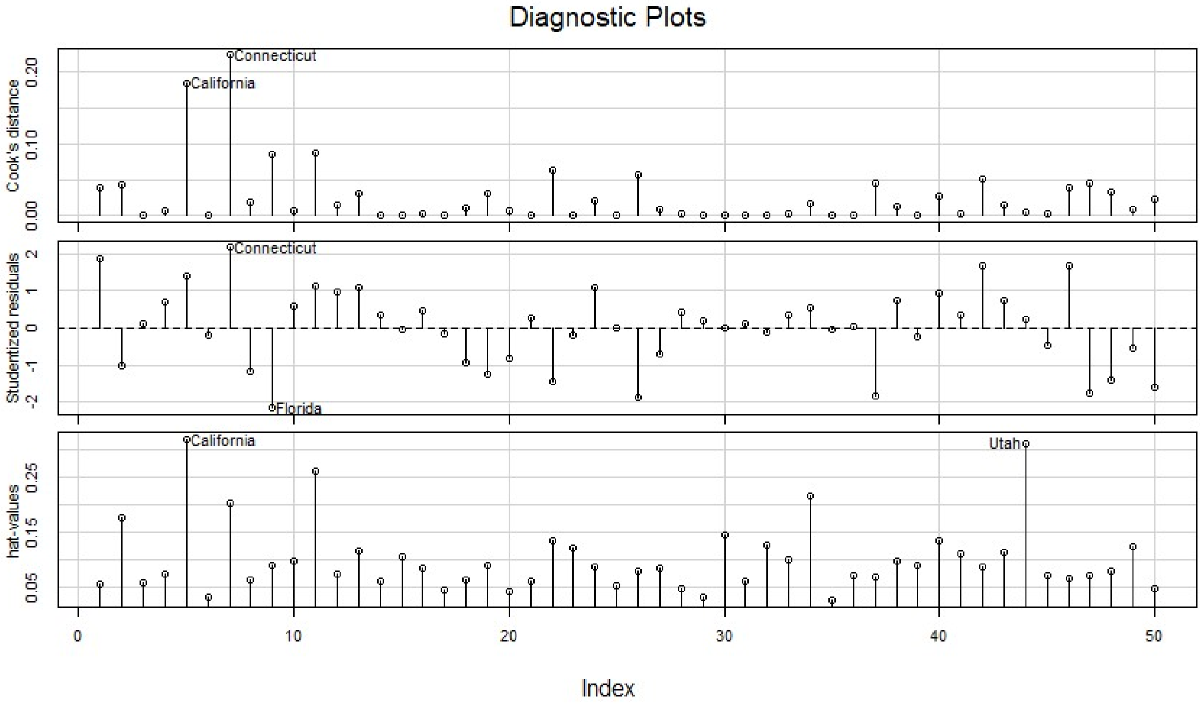


All the values of the residual plot stayed between -2 and 2. There seemed to be no serious problem with inequality of variance. We did not see any potential outliners.

Finally, I performed influence analysis:



```
Potentially influential observations of
	 lm(formula = expend^-0.5 ~ ratio + salary + verbal + math, data = sat) :

               dfb.1_ dfb.rati dfb.slry dfb.vrbl dfb.math dffit   cov.r   cook.d hat    
California     -0.05   0.73     0.00    -0.45     0.40     0.97    1.31    0.18   0.32_*
Connecticut    -0.36  -0.35     0.94     0.36    -0.32     1.10_*  0.84    0.22   0.20  
North Dakota    0.03  -0.10    -0.12    -0.18     0.21     0.29    1.38_*  0.02   0.22  
Utah           -0.09   0.14     0.00     0.06    -0.05     0.16    1.61_*  0.01   0.31_*
```



Through the influence analysis, we can see that Connecticut has influence on single fitted value. California and Utah is a possible leverage point. North Dakota and Utah potentially impact precision of estimation.

Overall, this model serves a lot more purpose that the previous one. It shows us not only the relationship between expenditure and teaching performance and quality, but also the relationship between that expend and the students' performance, even though it is quite surprising that the students are likely to perform worse when they go to more expensive schools. The data points of this model also fit the linear regression models better than the original one.

# 4 Conclusion and Self-reflection

## 4.1 Conclusion

Based on all the analysis we have done through 2 models, the original and the transformed model, we can conclude that there exists a clear relationship between the expenditure cost and the teaching performance and quality, as well as the students' performance. As the expenditure cost increases, the teaching performance and quality increases as well. However, this is not true with the students' performance. The data analysis shows that the students' performance, based on their test results, tend to decrease as the school they choose to attend has more expensive expenditure cost. This is a very interesting finding, and it serves as an alarm for both educators and parents. As for educators, they need to better their educational system and get rid of unnecessary costs that do not translate into students' performance. As for parents, they need to rethink again their strategy to make their children go into prestigious and expensive schools, without finding out other information related to that school.

We can get a lot information about the data frame itself after analyzing the data as well. In particular, the data frame does not consistently follow a trend in linear regression. This is shown in the first model that I have done based on the original data. That model, using expend as its response variable, can hardly fit any regressor other than ratio and salary, which translates into teaching performance and quality.

## 4.1 Self-reflection

About this project, for the most part it went well. I have succeeded in analyzing both the original data frame and the transformed data frame. This made the comparison between those two models and the two whole data frame much easier. I also found out the fitted transformation index for all of 5 possible regressor variables through boxcox log-likelihood method. Most of the analysis went well and I did not have any problem in executing and expressing them.

However, there were still parts that did not go well for this project. I could not fit the fifth possible regressor variables inside the transformed model, therefore I could not really see the relationship between expenditure cost and the number of student eligible to take the test. This can be improved by choosing the transformation index through different method instead, or additional ones, such as transforming the regressor variables themselves. This also leads to the fact that the relationship between expenditure and students' performance is not clear enough, possibly because of my choice of transformation index.

This project can go much further than the state it is right now. I want to discover the relationship between the expenditure cost and number of student eligible to take the test, the one thing that I am not able to do here. Also, as I went through the data frame, I noticed that the takers variable may relate to the verbal and math variables somehow. I want to make the takers variable into response variable and create a new model to see how far it can go. Lastly, I want to see if there is any relationship between the verbal and math scores. This can be done by using one of them as response variable and create a new model as well.

# 5 Appendix (R Code)

```
# STAT 4355.001 Course Project
# Dat Tran - dmt170030

# Preparation: Install library and read data
install.packages("nloptr")
install.packages("faraway")
library(faraway)
library(leaps)
library(MASS)
library(car)
library(ggplot2)
library(reshape2)
data(sat)
n <- nrow(sat)

## Part 2
# regsubset input
all.possible <- regsubsets(expend ~ ratio+salary+takers+verbal+math, data=sat)
summary(all.possible)
summary(all.possible)$which
names(summary(all.possible))

# Selection criteria
ap.mse <- summary(all.possible)$rss/(n-(2:6))
ap.adjr2 <- summary(all.possible)$adjr2
ap.cp <- summary(all.possible)$cp
ap.bic <- summary(all.possible)$bic
ap.criteria <- cbind(ap.mse, ap.adjr2, ap.cp, ap.bic)
colnames(ap.criteria) <- c("MSE", "Adj R2", "Cp", "BIC")
rownames(ap.criteria) <- 2:6
ap.criteria

# Visualizing selection criteria
par(mfrow=c(1,2))
plot(all.possible, scale="r2", main = "Exhaustive: R2")
plot(all.possible, scale="adjr2", main = "Exhaustive: adjusted R2")

par(mfrow=c(1,2))
plot(2:6, ap.mse, col = "blue", type = "l", xlab = "p", ylab = "MSE")
plot(2:6, ap.adjr2, col = "blue", type = "l", xlab = "p", ylab = "Adj R2")

par(mfrow=c(1,2))
plot(all.possible, scale="Cp", main = "Exhaustive: Cp")
plot(all.possible, main="Exhaustive: BIC")

par(mfrow=c(1,2))
plot(2:6, ap.cp, col = "blue", xlab = "p", ylab = "Cp", pch=16, cex=1)
abline(a=0,b=1, col = "red")
plot(2:6, ap.bic, col = "blue", type = "l", xlab = "p", ylab = "BIC")

# Stepwise selection
full <- lm(expend~ ratio+salary+takers+verbal+math, data=sat)
summary(full)
bwd.aic <- stepAIC(full, direction="both")
```

```
# Model fitting
final.model <- lm(expend ~ ratio + salary, data = sat)
summary(final.model)

# Linear regression plotting
sat2 <- melt(sat[, c(1, 2:3)], id.vars = "expend")
ggplot(sat2) +
  geom_jitter(aes(value,expend, colour=variable),) +
  geom_smooth(aes(value,expend, colour=variable), method=lm, se=FALSE) +
  facet_wrap(~variable, scales="free_x")

# Normal probability plot
par(mfrow=c(1,2))
hist(studres(final.model), breaks=10, freq=F, col="cornflowerblue",
     cex.axis=1.5, cex.lab=1.5, cex.main=1.5)
qqPlot(final.model)

# Residual Plot
par(mfrow=c(1,1))
residualPlot(final.model, type="rstudent", quadratic=F, col = "dodgerblue",
             pch=16, cex=1.5, cex.axis=1.5, cex.lab=1.5)

#Influence analysis
myInf <- influence.measures(final.model)
summary(myInf)

influenceIndexPlot(final.model, vars=c("Cook", "Studentized", "hat"))
```

```
## Part 3
# Transformation

par(mfrow=c(1,1))
boxcox(sat$expend ~ sat$ratio, lambda=seq(-2,2,1/10))
boxcox(sat$expend ~ sat$salary, lambda=seq(-2,2,1/10))
boxcox(sat$expend ~ sat$takers, lambda=seq(-2,2,1/10))
boxcox(sat$expend ~ sat$verbal, lambda=seq(-2,2,1/10))
boxcox(sat$expend ~ sat$math, lambda=seq(-2,2,1/10))

# Through this, can choose y'= y^-0.5 to test (expend)
# regsubset input
all.possible.2 <- regsubsets(expend^-0.5 ~ ratio+salary+takers+verbal+math, data=sat)
summary(all.possible.2)
summary(all.possible.2)$which
names(summary(all.possible.2))

# Selection criteria
ap.mse.2 <- summary(all.possible.2)$rss/(n-(2:6))
ap.adjr2.2 <- summary(all.possible.2)$adjr2
ap.cp.2 <- summary(all.possible.2)$cp
ap.bic.2 <- summary(all.possible.2)$bic
ap.criteria.2 <- cbind(ap.mse.2, ap.adjr2.2, ap.cp.2, ap.bic.2)
colnames(ap.criteria.2) <- c("MSE", "Adj R2", "Cp", "BIC")
rownames(ap.criteria.2) <- 2:6
ap.criteria.2
```

```
# Visualizing selection criteria
par(mfrow=c(1,2))
plot(all.possible.2, scale="r2", main = "Exhaustive: R2")
plot(all.possible.2, scale="adjr2", main = "Exhaustive: adjusted R2")

par(mfrow=c(1,2))
plot(2:6, ap.mse.2, col = "blue", type = "l", xlab = "p", ylab = "MSE")
plot(2:6, ap.adjr2.2, col = "blue", type = "l", xlab = "p", ylab = "Adj R2")

par(mfrow=c(1,2))
plot(all.possible.2, scale="Cp", main = "Exhaustive: Cp")
plot(all.possible.2, main="Exhaustive: BIC")

par(mfrow=c(1,2))
plot(2:6, ap.cp.2, col = "blue", xlab = "p", ylab = "Cp", pch=16, cex=1)
abline(a=0,b=1, col = "red")
plot(2:6, ap.bic.2, col = "blue", type = "l", xlab = "p", ylab = "BIC")

# Stepwise selection
full.2 <- lm(expend^-0.5~ ratio+salary+takers+verbal+math, data=sat)
summary(full.2)
bwd.aic <- stepAIC(full.2, direction="both")

# Model fitting
final.model.2 <- lm(expend^-0.5 ~ ratio + salary + verbal + math, data = sat)
summary(final.model.2)

# Linear regression plotting
```

```
sat2.2 <- melt(sat[, c(1, 2:6)], id.vars = "expend")
sat2.2$expend <- sat2.2$expend^-0.5
ggplot(sat2.2) +
  geom_jitter(aes(value,expend, colour=variable),) +
  geom_smooth(aes(value,expend, colour=variable), method=lm, se=FALSE) +
  facet_wrap(~variable, scales="free_x")

# Normal  probability plot
par(mfrow=c(1,2))
hist(studres(final.model.2), breaks=10, freq=F, col="cornflowerblue",
     cex.axis=1.5, cex.lab=1.5, cex.main=1.5)
qqPlot(final.model.2)

# Resisdual plot
par(mfrow=c(1,1))
residualPlot(final.model.2, type="rstudent", quadratic=F, col = "dodgerblue",
             pch=16, cex=1.5, cex.axis=1.5, cex.lab=1.5)

# Influence analysis
myInf.2 <- influence.measures(final.model.2)
summary(myInf.2)

influenceIndexPlot(final.model.2, vars=c("Cook", "Studentized", "hat"))
```