Modeling High School Dropout Rates with Best Subset Selection

Department of Mathematics and Statistics, San Diego State University

Abstract

High School dropout rates have been falling in the United States for the past few decades reaching a new low of roughly 16% of high school students failing to graduate on time, with the biggest drops among Latinos and African Americans. [i] *What are the most relevant factors in determining this fall in high school dropout rates? Looking at 135 public High Schools in the Northeastern US Metropolitan Area, I will seek to determine the relationship between ten potential predictors and the associated dropout rates. Considering all potential models using Best Subset Selection I will compare adjusted* R^2 , BIC *and Mallows* C_p *to determine the best model fit. I also will use cross validation to determine test error. Our final model minimizes* C_p*and* BIC *with predictors enrollment, average teacher salary, and average 10th grade English MCAS score with an adjusted* R^2*=.58*

1 Introduction

One of the biggest education successes over the past few decades has been the fall in High School dropout rates in across the country. The nations graduation rate rose to around 86 percent of students graduating on time, which is a near 7 percent increase since 2011.[ii] This increase is largely responsible due to the fall in dropout rates among Latinos and African Americans. [iii] Looking at ten potential variables related to high school dropout rate, what are the most important factors in determining high school dropout rates? I hypothesize that smaller school sizes, a lower student to teacher ratio and higher average standardized test scores are the most strongly associated with decreasing dropout rates. I will begin with some exploratory data analysis to better understand my predictor variables and ensure the assumptions of my regression model are met. Since we only have 10 potential predictors available to us, we can use best subset selection, which will consider all potential models with 1-10 predictor variables and compare each models adjusted R^2 , Mallows C_p , and BIC to determine the best model. I will also used cross validation to calculate test error for each 1-10 variable model. I will then provide test statistics and p-values for each variable and inferences from the final model as well as confirm homoscedasticity and normality of residuals. Finally, I will discuss some limitations of our study and potential future research.

[i] https://www.nytimes.com/2018/05/25/opinion/college-dropout.html

[ii] https://nces.ed.gov/ccd/tables/ACGR_RE_and_characteristics_2018-19.asp

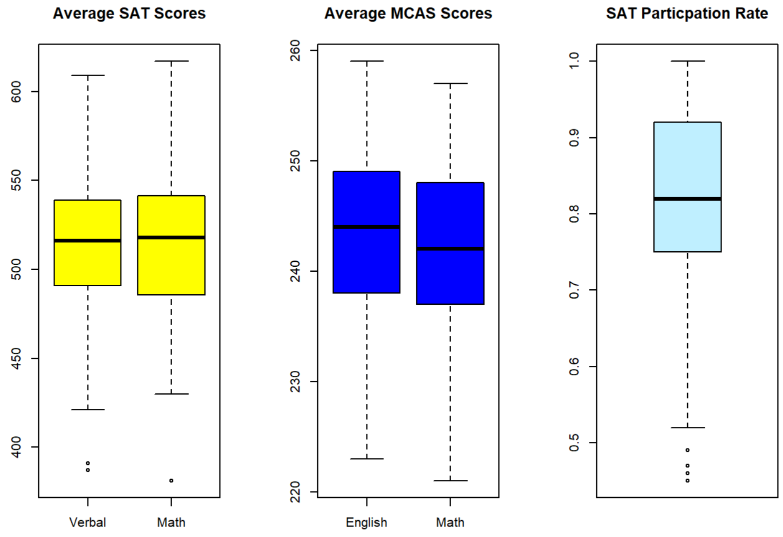
[iii] https://www.washingtonpost.com/news/education/wp/2017/12/04/u-s-high-school-graduation-rates-rise-to-new-high/

2 Exploratory Data Analysis

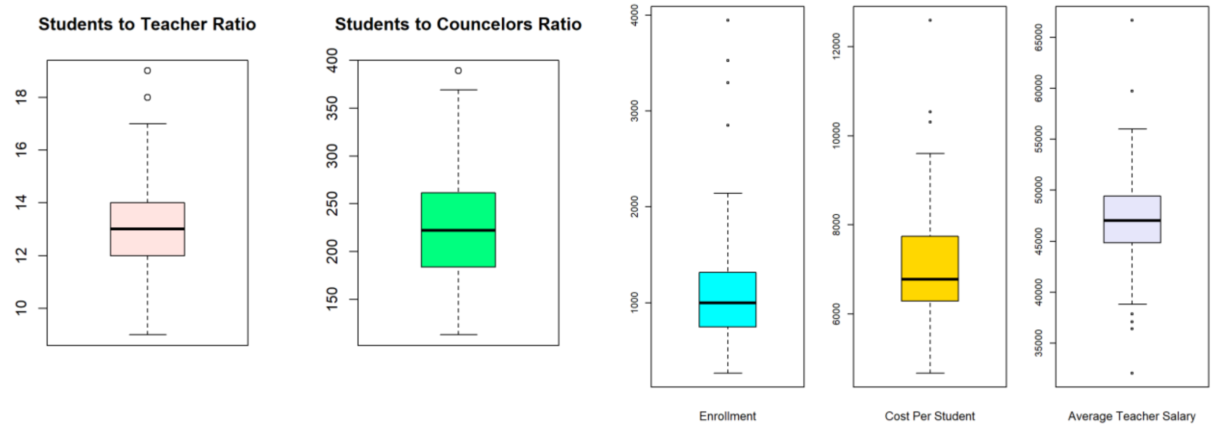
We have 10 potential predictors variables at our disposal for our regression model. Below is a table with each predictor listed and a short description of each.

PREDICTOR	**DESCRIPTION**
Enrollment	Number of students enrolled
Cost/Pupil	Cost per student
AveTeach$	Average teacher salary
SATV	Average SAT Verbal score
SATM	Average SAT Math score
SATPartRate	SAT participation rate
10GMCASEng	Average 10th grade MCAS English score
10GMCASMth	Average 10th grade MCAS Math score
S/TRatio	Student to teacher ratio
S/CounselRatio	Student to Counselor ratio

Our response variable DropoutRate, measures the average dropout rate per 100 students, each year. Note we have five predictors related to Standardized Exams, which could lead to potential multicollinearity issues (this will be addressed in later section). Let's look at some summary statistics for each predictor. First, here are boxplots of the standardized exam variables.

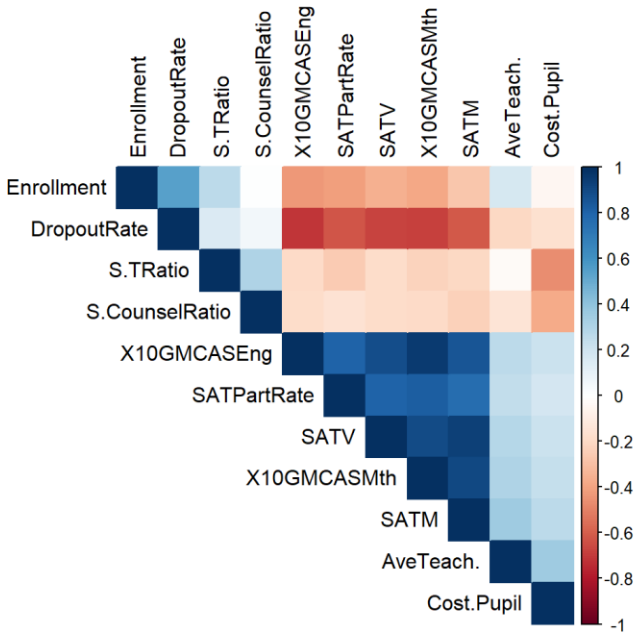


We seem to have a few outliers for the SAT exam, which could be high leverage point in our final model. The distribution of scores for each exam between the verbal and math scores are both similar with most SAT scores falling in the 490 – 540 range and GMACS score in the 240-250 range. There are a few low-end outliers for SAT participation rate, with the lowest being 45%. Now let's look at our student to teacher and student to counselor ratio box plots as well as the final three predictors.

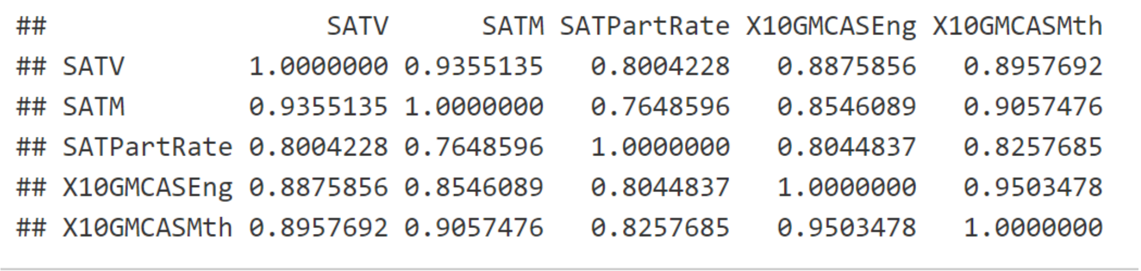


Our median student to teacher ratio is 13 to 1 with two outliers while the median student to counselor ratio is 222 to 1 with one outlier. Enrollment, cost per student, and average teacher salary all have a few outliers but similar averages and medians suggesting a roughly symmetric distribution.

Looking at the correlation plot between all predictors and response variable we can see some interesting relationships.

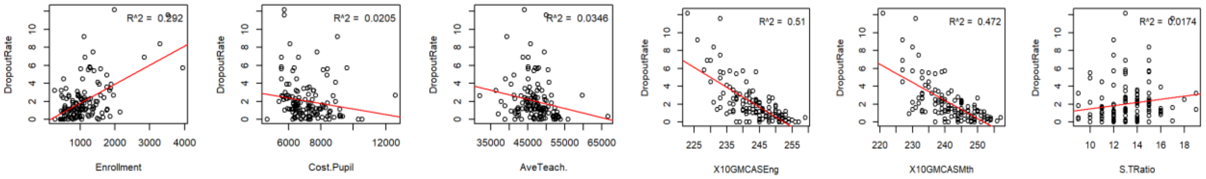


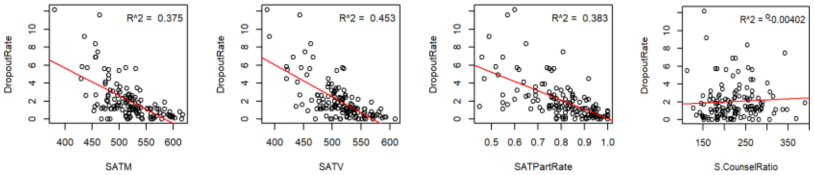
While most predictors are correlated with dropout rate as expected, the correlation between student/counselor ratio and dropout rate is about .06, a surprisingly weak relationship. Looking at the correlations between our standardized exam variables we can see that we will likely run into some multicollinearity issues if we were to include all of them in the final model.



```
##                  SATV      SATM SATPartRate X10GMCASEng X10GMCASMth
## SATV        1.0000000 0.9355135   0.8004228   0.8875856   0.8957692
## SATM        0.9355135 1.0000000   0.7648596   0.8546089   0.9057476
## SATPartRate 0.8004228 0.7648596   1.0000000   0.8044837   0.8257685
## X10GMCASEng 0.8875856 0.8546089   0.8044837   1.0000000   0.9503478
## X10GMCASMth 0.8957692 0.9057476   0.8257685   0.9503478   1.0000000
```

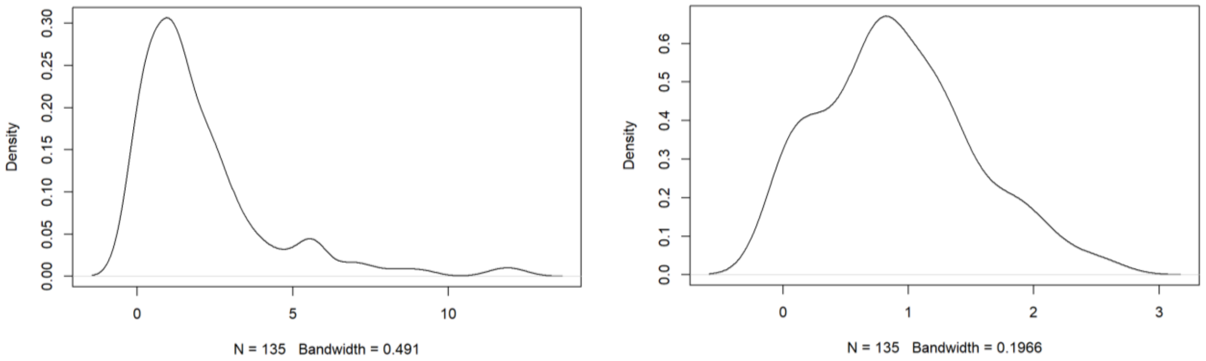
Let’s look at some scatterplots between each predictor and dropout rates.





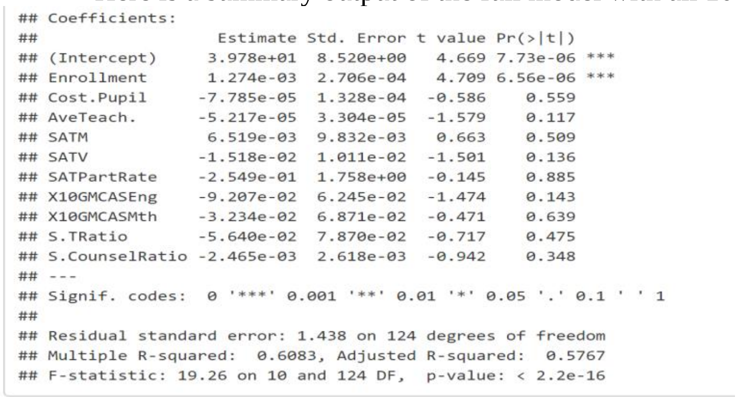
SAT verbal and MCAS math have the strongest relationship with dropout rate, both with an R^2 over .45. Student to teacher ratio could almost be thought of as categorial, as all points lie on integers between 9 and 19. There doesn’t seem to be any clear nonlinear relationships, although some have relatively small R^2 values - such as average teacher salary and cost per pupil.

Finally let’s look at the distribution of our response variable, dropout rate. Below is a kernel density estimate of Dropout Rate (left) and its log(1+x) transform (right).



Dropout rate appears to be right skewed, with a skewness value of 2.11, which could affect our coefficient estimates. Log transforming dropout rate gives us a much more symmetric bell-shaped distribution. When looking at MLR models with and without the log transform we see that the log transformed response reduces residual standard error from 1.438 to .395. Thus, from now on I will be using the log transformed response variable.

Here is a summary output of the full model with all 10 predictors. Note the only statistically

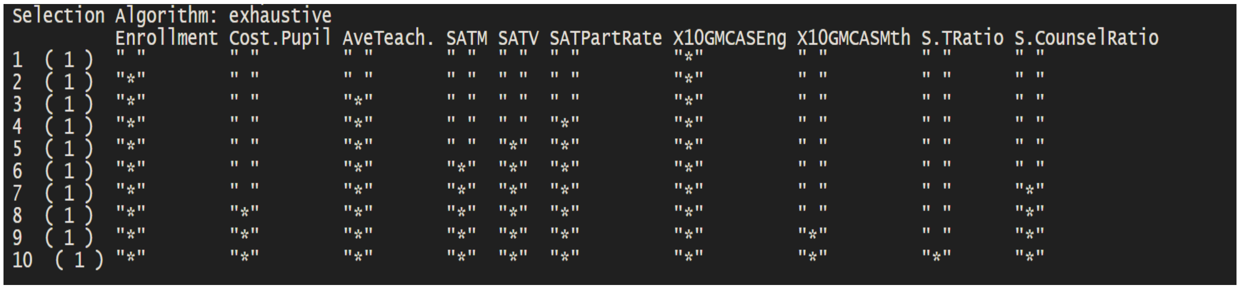


```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.978e+01  8.520e+00   4.669 7.73e-06 ***
## Enrollment       1.274e-03  2.706e-04   4.709 6.56e-06 ***
## Cost.Pupil      -7.785e-05  1.328e-04  -0.586    0.559
## AveTeach.       -5.217e-05  3.304e-05  -1.579    0.117
## SATM             6.519e-03  9.832e-03   0.663    0.509
## SATV            -1.518e-02  1.011e-02  -1.501    0.136
## SATPartRate     -2.549e-01  1.758e+00  -0.145    0.885
## X10GMCASEng     -9.207e-02  6.245e-02  -1.474    0.143
## X10GMCASMth     -3.234e-02  6.871e-02  -0.471    0.639
## S.TRatio        -5.640e-02  7.870e-02  -0.717    0.475
## S.CounselRatio -2.465e-03  2.618e-03  -0.942    0.348
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.438 on 124 degrees of freedom
## Multiple R-squared:  0.6083, Adjusted R-squared:  0.5767
## F-statistic: 19.26 on 10 and 124 DF,  p-value: < 2.2e-16
```

Significant variable is enrollment, and many of the predictors have very small coefficient estimates. How can we improve upon this model?

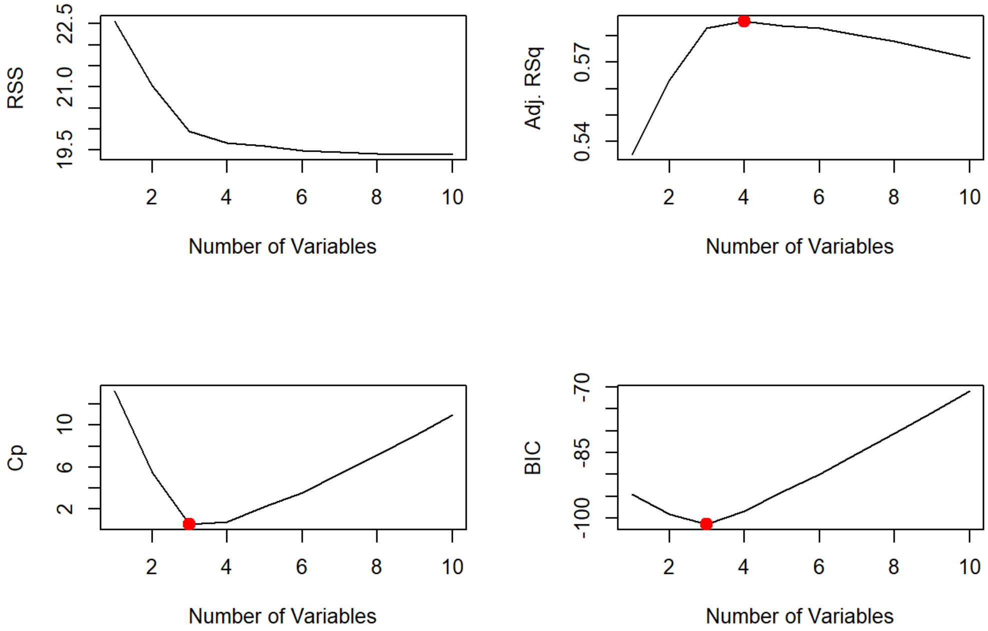
3 Best Subset Selection

For regression models with a small number of predictors it is possible to fit a separate least square regression model for each possible combination of the p predictors. Thus we can fit all models with one predictor, then all with two predictors, and so forth. For large p, fitting all 2^p would be too computationally intensive. Luckily our model only has 10 possible predictors and the *regsubssets* function in R can quickly give us the best model for each k = 1,2,…,10 predictors (here best is defined as having the smallest RSS, or equivalently largest R^2). From there we can select a single best model using cross validated prediction error, C_p, BIC, or adjusted R^2. Here is the summary output of our *regsubsets* function. (A * means that variable was included in the model.)



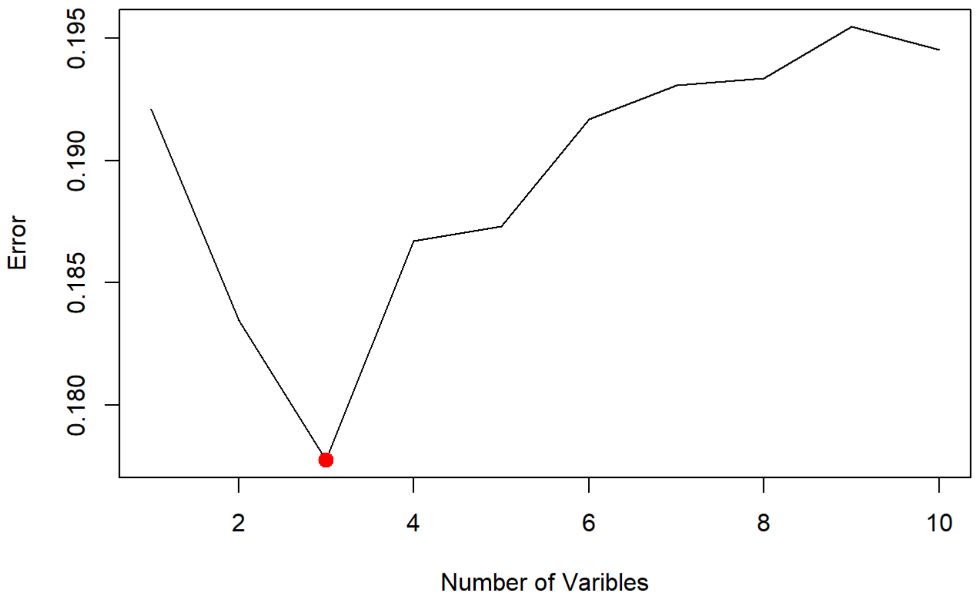
```
Selection Algorithm: exhaustive
          Enrollment Cost.Pupil AveTeach. SATM SATV SATPartRate X10GMCASEng X10GMCASMth S.TRatio S.CounselRatio
1  ( 1 )  " "        " "        " "       " "  " "  " "         "*"         " "         " "      " "
2  ( 1 )  "*"        " "        " "       " "  " "  " "         "*"         " "         " "      " "
3  ( 1 )  "*"        " "        "*"       " "  " "  " "         "*"         " "         " "      " "
4  ( 1 )  "*"        " "        "*"       " "  " "  "*"         "*"         " "         " "      " "
5  ( 1 )  "*"        " "        "*"       " "  "*"  "*"         "*"         " "         " "      " "
6  ( 1 )  "*"        " "        "*"       "*"  "*"  "*"         "*"         " "         " "      " "
7  ( 1 )  "*"        " "        "*"       "*"  "*"  "*"         "*"         " "         " "      "*"
8  ( 1 )  "*"        "*"        "*"       "*"  "*"  "*"         "*"         " "         " "      "*"
9  ( 1 )  "*"        "*"        "*"       "*"  "*"  "*"         "*"         "*"         " "      "*"
10  ( 1 ) "*"        "*"        "*"       "*"  "*"  "*"         "*"         "*"         "*"      "*"
```

The first variable to be included in enrollment, which was also the only statistically significant variable in our full model. In order to determine the best model, I plotted RSS, adjusted R^2, C_p and BIC for all 10 models (red dot denotes maximum or minimum point).



The four-variable model with enrollment, average teacher salary, SAT participation rate, and English GMCAS maximized our adjusted R^2 at around .58. C_p and BIC were both minimized using a three-variable model, removing SAT participation rate.

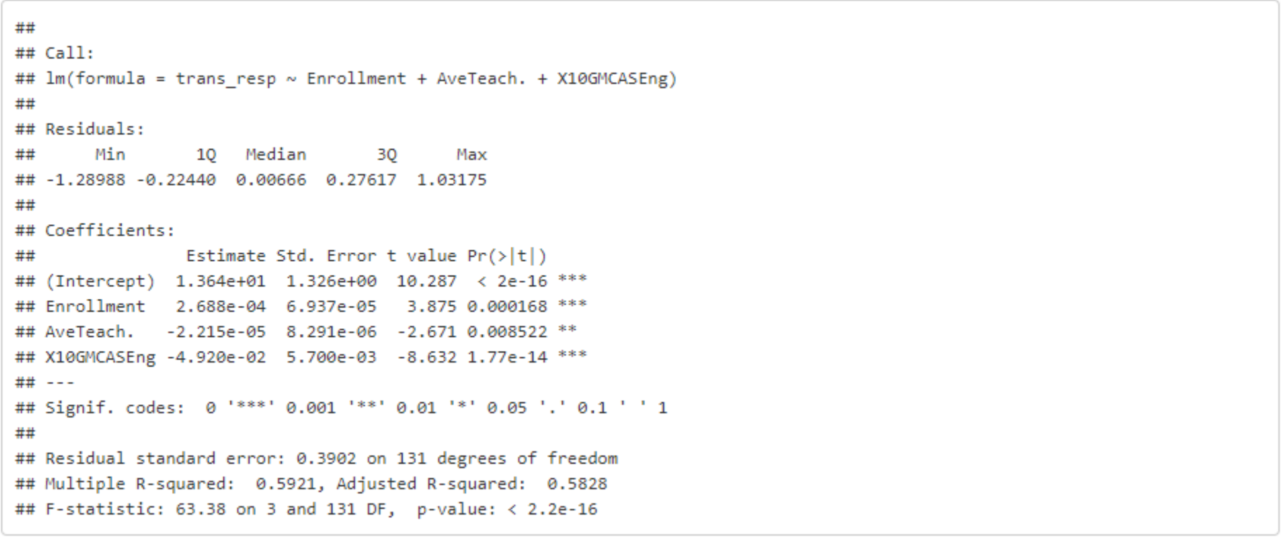
We can also apply a cross validation approach to get an estimate of test error by randomly splitting our data in half between a training set and a test set, then use regsubsets on our training data and compute the test MSE for model. Below is a plot of test error for each of the 10 models.



As we can see in the plot above, test error is minimized with our three-variable model. It should be noted that I ran this function a few times and received different values, so this result should be considered in the context of the other results regarding R^2, C_p, and BIC.

4 Final Model

Given the fact our three-variable model minimized C_p, BIC, test MSE, while achieving a very similar adjusted R^2 value to our four-variable model, suggests it is the best linear model for our data. Thus, our final model coefficient estimates, test statistics, and p values are as follows:



```
##
## Call:
## lm(formula = trans_resp ~ Enrollment + AveTeach. + X10GMCASEng)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28988 -0.22440  0.00666  0.27617  1.03175
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.364e+01  1.326e+00  10.287  < 2e-16 ***
## Enrollment   2.688e-04  6.937e-05   3.875 0.000168 ***
## AveTeach.   -2.215e-05  8.291e-06  -2.671 0.008522 **
## X10GMCASEng -4.920e-02  5.700e-03  -8.632 1.77e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3902 on 131 degrees of freedom
## Multiple R-squared:  0.5921, Adjusted R-squared:  0.5828
## F-statistic: 63.38 on 3 and 131 DF,  p-value: < 2.2e-16
```

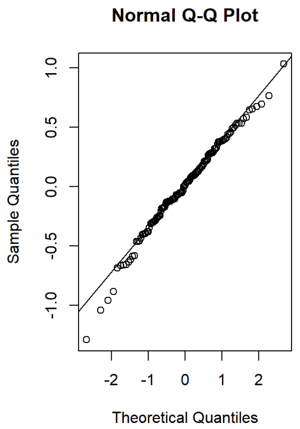
Note all predictors are statistically significant at the .05 level (equivalently we are 95% confident that these coefficient estimates are nonzero) and we have a small increase in adjusted R^2 from our full model. Residuals are also smaller and centered around zero. We can interpret the coefficient estimates as follows:

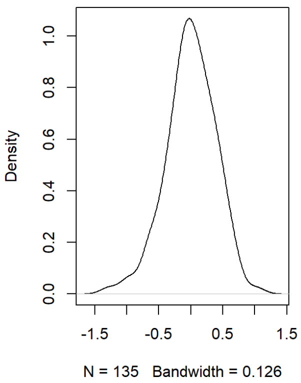
-For every 100-student increase in enrollment, dropout rates increase on average 2.72%, assuming all else fixed

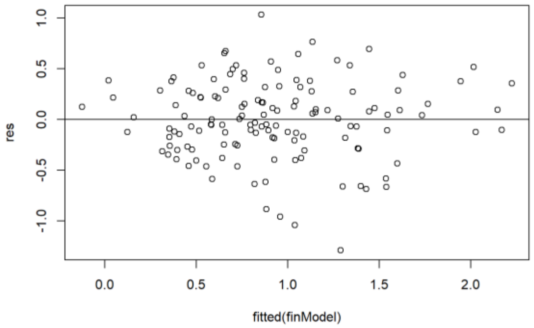
-For every 1000$ increase in average teacher pay, dropout rates fall on average 2.24% all else being equal

-A one-point increase in average GMCAS English scores in associated with a 5% decrease in dropout rates, all else fixed.

Here are a few diagnostics plots to ensure our assumptions about the residuals are correct. We assume the errors are normally distributed (QQ plot/density left) and centered around zero with constant variance (residual plot right)





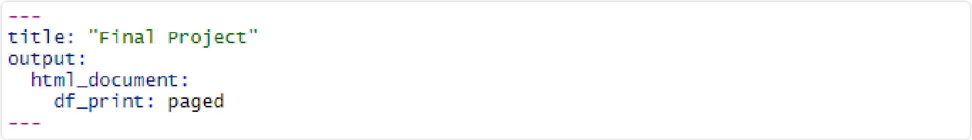


It appears all assumptions are met based on the plots above and we can draw valid statistical inference from our model.

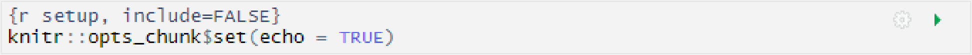
5 Conclusion

My original hypothesis was that enrollment, student to teacher ratio, and standardized test scores would be the most important factors associated with dropout rates. While student to teacher ratio played a much smaller role, both enrollment and English GMCAS scores appeared in the final model in addition to average teacher pay. It should be noted that some of these variables might just be indicators for other variables not included. For example, higher average teacher pay may be associated with being a school in a wealthier area, which in turn to play could play a role in dropout rates. Our model is limited to the ten predictor variables given, there is likely many more factors that can influence dropout rates. Enrollment does seem to have a few high leverage points which was not accounted for in our final model. We should also note that best model selection is looking only at potential linear models, there may be nonlinear relationships in our data although the scatterplots didn’t seem to suggest so. If the goal in better understand the fall in high school dropout rates across the country, future research should expand this study beyond the Northeastern Metropolitan area and include a more variables like GPA, economic status, demographic data, etc.

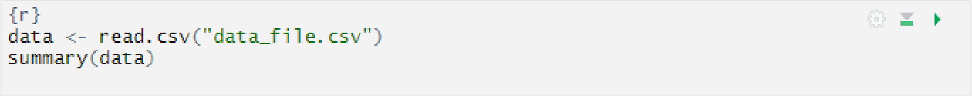
APPENDIX



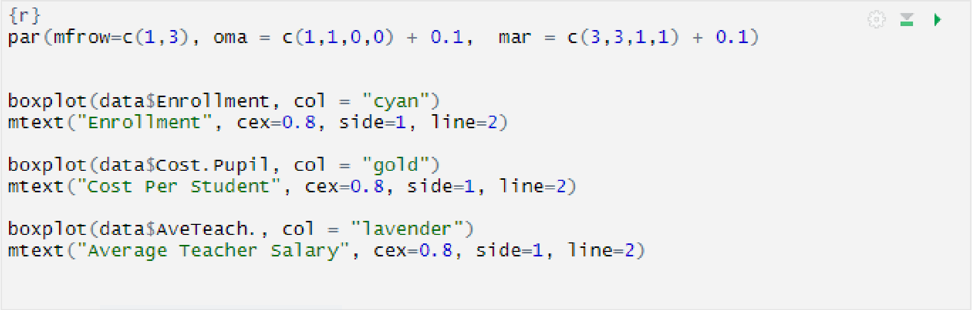
```
---
title: "Final Project"
output:
  html_document:
    df_print: paged
---
```



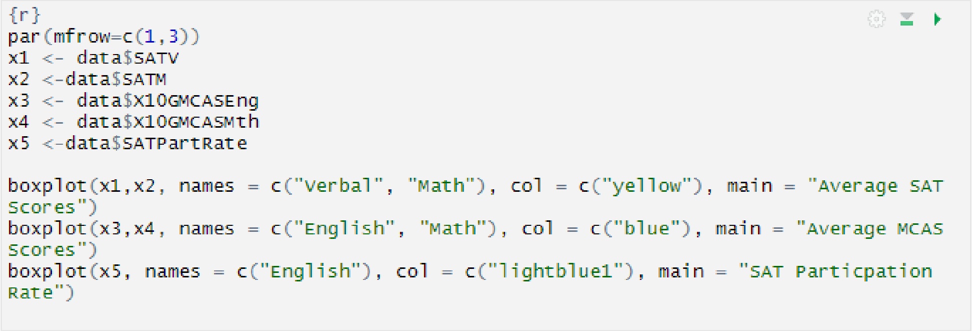
```
{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```



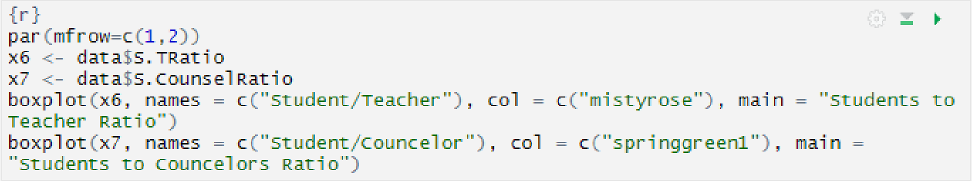
```
{r}
data <- read.csv("data_file.csv")
summary(data)
```



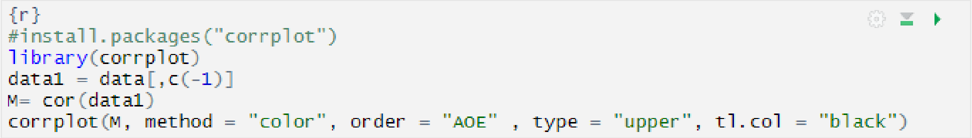
```
{r}
par(mfrow=c(1,3), oma = c(1,1,0,0) + 0.1,  mar = c(3,3,1,1) + 0.1)


boxplot(data$Enrollment, col = "cyan")
mtext("Enrollment", cex=0.8, side=1, line=2)

boxplot(data$Cost.Pupil, col = "gold")
mtext("Cost Per Student", cex=0.8, side=1, line=2)

boxplot(data$AveTeach., col = "lavender")
mtext("Average Teacher Salary", cex=0.8, side=1, line=2)
```



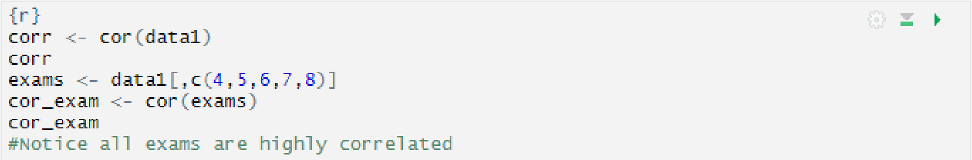
```
{r}
par(mfrow=c(1,3))
x1 <- data$SATV
x2 <-data$SATM
x3 <- data$X10GMCASEng
x4 <- data$X10GMCASMth
x5 <-data$SATPartRate

boxplot(x1,x2, names = c("Verbal", "Math"), col = c("yellow"), main = "Average SAT
Scores")
boxplot(x3,x4, names = c("English", "Math"), col = c("blue"), main = "Average MCAS
Scores")
boxplot(x5, names = c("English"), col = c("lightblue1"), main = "SAT Particpation
Rate")
```



```
{r}
par(mfrow=c(1,2))
x6 <- data$S.TRatio
x7 <- data$S.CounselRatio
boxplot(x6, names = c("Student/Teacher"), col = c("mistyrose"), main = "Students to
Teacher Ratio")
boxplot(x7, names = c("Student/Councelor"), col = c("springgreen1"), main =
"Students to Councelors Ratio")
```

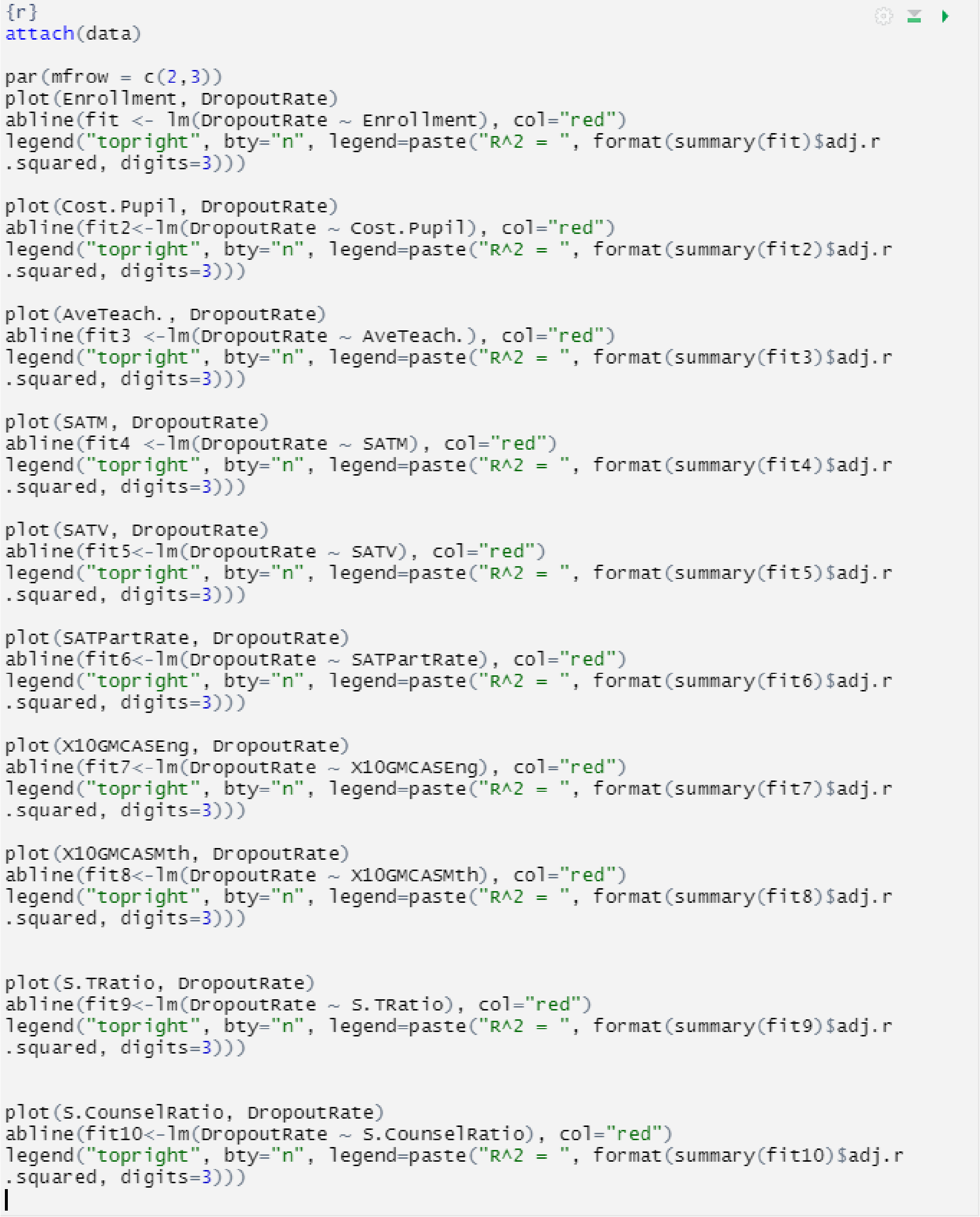


```
{r}
#install.packages("corrplot")
library(corrplot)
data1 = data[,c(-1)]
M= cor(data1)
corrplot(M, method = "color", order = "AOE" , type = "upper", tl.col = "black")
```

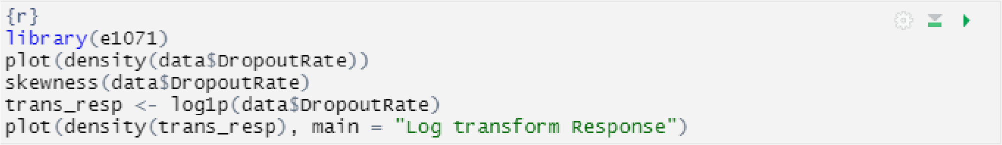


```
{r}
corr <- cor(data1)
corr
exams <- data1[,c(4,5,6,7,8)]
cor_exam <- cor(exams)
cor_exam
#Notice all exams are highly correlated
```

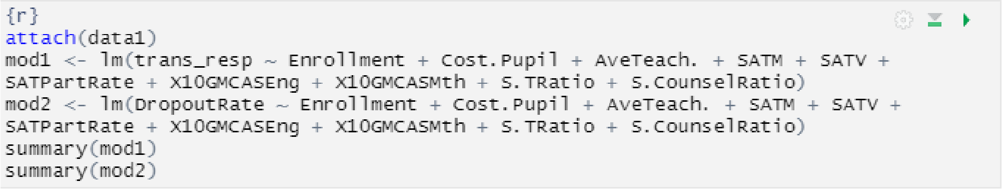
APPENDIX



```
{r}
attach(data)

par(mfrow = c(2,3))
plot(Enrollment, DropoutRate)
abline(fit <- lm(DropoutRate ~ Enrollment), col="red")
legend("topright", bty="n", legend=paste("R^2 = ", format(summary(fit)$adj.r
.squared, digits=3)))

plot(Cost.Pupil, DropoutRate)
abline(fit2<-lm(DropoutRate ~ Cost.Pupil), col="red")
legend("topright", bty="n", legend=paste("R^2 = ", format(summary(fit2)$adj.r
.squared, digits=3)))

plot(AveTeach. , DropoutRate)
abline(fit3 <-lm(DropoutRate ~ AveTeach.), col="red")
legend("topright", bty="n", legend=paste("R^2 = ", format(summary(fit3)$adj.r
.squared, digits=3)))

plot(SATM, DropoutRate)
abline(fit4 <-lm(DropoutRate ~ SATM), col="red")
legend("topright", bty="n", legend=paste("R^2 = ", format(summary(fit4)$adj.r
.squared, digits=3)))

plot(SATV, DropoutRate)
abline(fit5<-lm(DropoutRate ~ SATV), col="red")
legend("topright", bty="n", legend=paste("R^2 = ", format(summary(fit5)$adj.r
.squared, digits=3)))

plot(SATPartRate, DropoutRate)
abline(fit6<-lm(DropoutRate ~ SATPartRate), col="red")
legend("topright", bty="n", legend=paste("R^2 = ", format(summary(fit6)$adj.r
.squared, digits=3)))

plot(X10GMCASEng, DropoutRate)
abline(fit7<-lm(DropoutRate ~ X10GMCASEng), col="red")
legend("topright", bty="n", legend=paste("R^2 = ", format(summary(fit7)$adj.r
.squared, digits=3)))

plot(X10GMCASMth, DropoutRate)
abline(fit8<-lm(DropoutRate ~ X10GMCASMth), col="red")
legend("topright", bty="n", legend=paste("R^2 = ", format(summary(fit8)$adj.r
.squared, digits=3)))


plot(S.TRatio, DropoutRate)
abline(fit9<-lm(DropoutRate ~ S.TRatio), col="red")
legend("topright", bty="n", legend=paste("R^2 = ", format(summary(fit9)$adj.r
.squared, digits=3)))


plot(S.CounselRatio, DropoutRate)
abline(fit10<-lm(DropoutRate ~ S.CounselRatio), col="red")
legend("topright", bty="n", legend=paste("R^2 = ", format(summary(fit10)$adj.r
.squared, digits=3)))
```

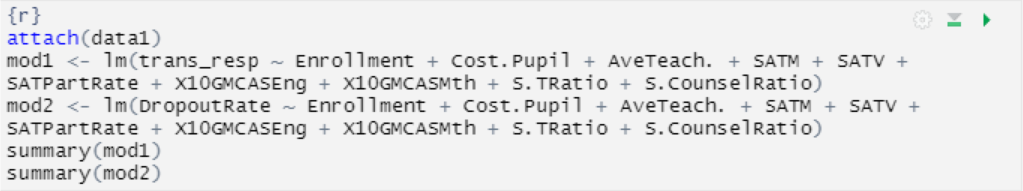


```
{r}
library(e1071)
plot(density(data$DropoutRate))
skewness(data$DropoutRate)
trans_resp <- log1p(data$DropoutRate)
plot(density(trans_resp), main = "Log transform Response")
```

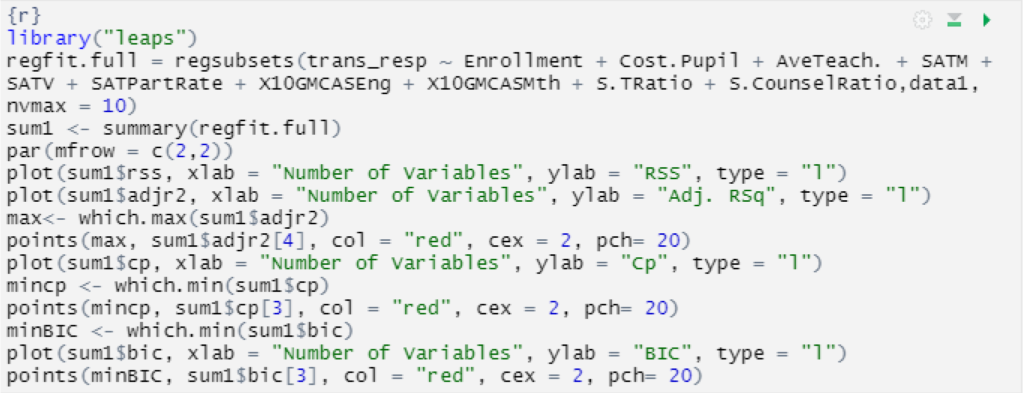


```
{r}
attach(data1)
mod1 <- lm(trans_resp ~ Enrollment + Cost.Pupil + AveTeach. + SATM + SATV +
SATPartRate + X10GMCASEng + X10GMCASMth + S.TRatio + S.CounselRatio)
mod2 <- lm(DropoutRate ~ Enrollment + Cost.Pupil + AveTeach. + SATM + SATV +
SATPartRate + X10GMCASEng + X10GMCASMth + S.TRatio + S.CounselRatio)
summary(mod1)
summary(mod2)
```

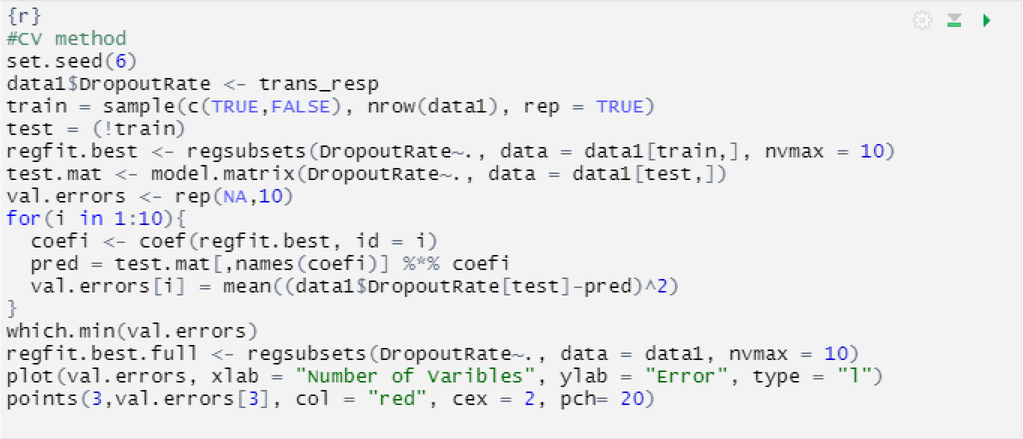
APPENDIX



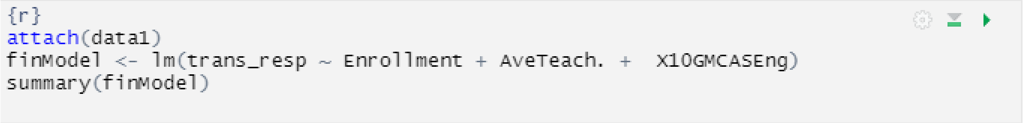
```
{r}
attach(data1)
mod1 <- lm(trans_resp ~ Enrollment + Cost.Pupil + AveTeach. + SATM + SATV +
SATPartRate + X10GMCASEng + X10GMCASMth + S.TRatio + S.CounselRatio)
mod2 <- lm(DropoutRate ~ Enrollment + Cost.Pupil + AveTeach. + SATM + SATV +
SATPartRate + X10GMCASEng + X10GMCASMth + S.TRatio + S.CounselRatio)
summary(mod1)
summary(mod2)
```



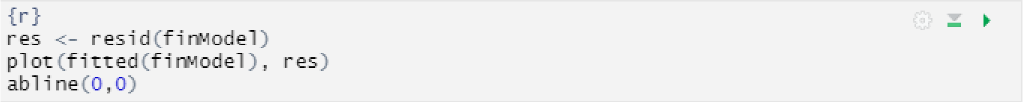
```
{r}
library("leaps")
regfit.full = regsubsets(trans_resp ~ Enrollment + Cost.Pupil + AveTeach. + SATM +
SATV + SATPartRate + X10GMCASEng + X10GMCASMth + S.TRatio + S.CounselRatio,data1,
nvmax = 10)
sum1 <- summary(regfit.full)
par(mfrow = c(2,2))
plot(sum1$rss, xlab = "Number of Variables", ylab = "RSS", type = "l")
plot(sum1$adjr2, xlab = "Number of Variables", ylab = "Adj. RSq", type = "l")
max<- which.max(sum1$adjr2)
points(max, sum1$adjr2[4], col = "red", cex = 2, pch= 20)
plot(sum1$cp, xlab = "Number of Variables", ylab = "Cp", type = "l")
mincp <- which.min(sum1$cp)
points(mincp, sum1$cp[3], col = "red", cex = 2, pch= 20)
minBIC <- which.min(sum1$bic)
plot(sum1$bic, xlab = "Number of Variables", ylab = "BIC", type = "l")
points(minBIC, sum1$bic[3], col = "red", cex = 2, pch= 20)
```



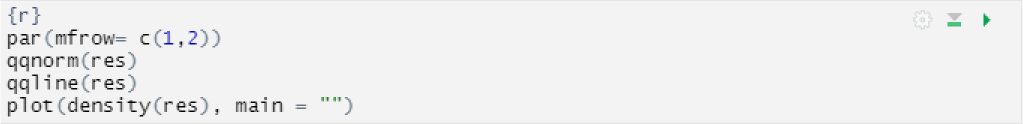
```
{r}
#CV method
set.seed(6)
data1$DropoutRate <- trans_resp
train = sample(c(TRUE,FALSE), nrow(data1), rep = TRUE)
test = (!train)
regfit.best <- regsubsets(DropoutRate~., data = data1[train,], nvmax = 10)
test.mat <- model.matrix(DropoutRate~., data = data1[test,])
val.errors <- rep(NA,10)
for(i in 1:10){
  coefi <- coef(regfit.best, id = i)
  pred = test.mat[,names(coefi)] %*% coefi
  val.errors[i] = mean((data1$DropoutRate[test]-pred)^2)
}
which.min(val.errors)
regfit.best.full <- regsubsets(DropoutRate~., data = data1, nvmax = 10)
plot(val.errors, xlab = "Number of Varibles", ylab = "Error", type = "l")
points(3,val.errors[3], col = "red", cex = 2, pch= 20)
```



```
{r}
attach(data1)
finModel <- lm(trans_resp ~ Enrollment + AveTeach. +  X10GMCASEng)
summary(finModel)
```



```
{r}
res <- resid(finModel)
plot(fitted(finModel), res)
abline(0,0)
```



```
{r}
par(mfrow= c(1,2))
qqnorm(res)
qqline(res)
plot(density(res), main = "")
```