Scores and Admissions

Mbonisi and Rick

2023-04-28

Table of Contents

# 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.

# Introduction

Conceptually speaking, the ACT, or any other standardized test for that matter, is touted as an indicator of a person’s mastery of high school material. This mastery is supposed to be a prerequisite for college coursework to build upon. However, indicators can often be misleading, and for an indicator to be robust, it has to be concise while being the least reductive. Specifically, in this project, we are investigating the validity of ACT scores and looking into factors that might contribute to a higher score for a cohort of students. These cohorts are 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 ACT scores. Hence, we expect the residuals to have a normal distribution when plotted against predicted values.

# Data Collection

Here we begin the data collection.

knitr::opts\_chunk$set(echo = TRUE)  
library(tidyverse)  
library(readxl)  
library(xlsx)  
library(datadictionary)  
library(olsrr)  
colleges = read\_xlsx("colleges.xlsx", sheet=1)

This marks the end of house keeping data acquistion.

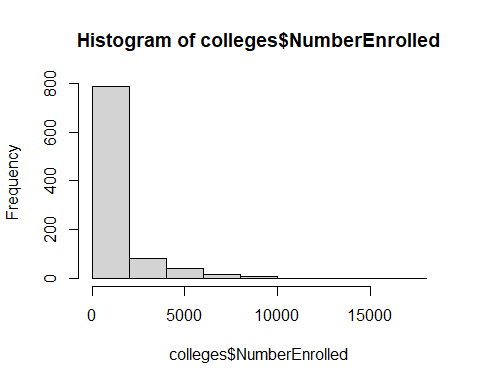
### This dataset has 7 predictor variables:

* **GPA** (quant) 75th percentile secondary school GPA of the cohort
* **TOEFL** (categorical) Categories signifying English Proficiency scores requirement from the college, ordered from the most lax to strictest requirements.
* **CollegePrep** (category) Completion of college-preparatory program
* **NumberEnrolled** (quant) The total number of students who enrolled in a college from that cohort
* **ACTPercentage** (quant) Percent of first-time degree/certificate-seeking students submitting ACT scores
* **SATWR** (quant) 75th percentile SAT Reading-Writing Score of the cohort
* **SATM** (quant) 75th percentile SAT Mathematics Score of the cohort

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

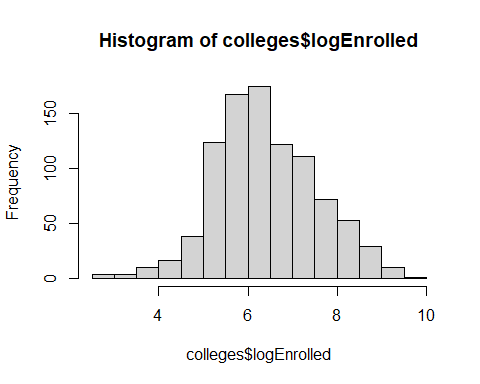
We know that variable ‘NumberEnrolled’ is skewed, as seen below.

hist(colleges$NumberEnrolled)

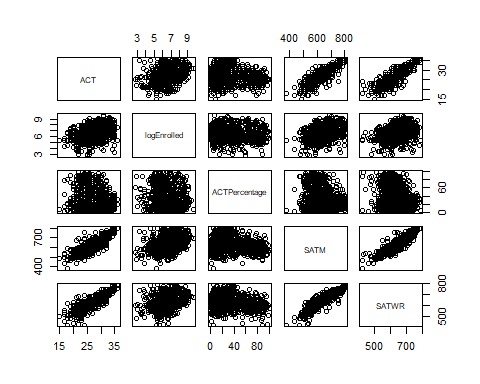


Therefore we will do a log transformation.

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

 Now, let us do a scatter plot matrix for a visual view of correlations among the variables.

numerical = cbind(ACT = colleges$ACT, logEnrolled = colleges$logEnrolled, ACTPercentage = colleges$ACTPercentage, SATM = colleges$SATM, SATWR = colleges$SATWR)  
pairs(numerical)



Multiple variable appear to be correlated to each other. 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.

# First order model

Here we begin the first order model.

model1 = lm(colleges$ACT ~ colleges$logEnrolled + colleges$ACTPercentage + colleges$SATWR + colleges$SATM + colleges$GPA + colleges$CollegePrep + colleges$TOEFL)  
summary(model1)

##   
## Call:  
## lm(formula = colleges$ACT ~ colleges$logEnrolled + colleges$ACTPercentage +   
## colleges$SATWR + colleges$SATM + colleges$GPA + colleges$CollegePrep +   
## colleges$TOEFL)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.8266 -0.8621 0.0407 0.9152 6.9328   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.559656 0.677665 -9.680 < 2e-16 \*\*\*  
## colleges$logEnrolled 0.146133 0.054269 2.693 0.00721 \*\*   
## colleges$ACTPercentage -0.006137 0.002136 -2.873 0.00415 \*\*   
## colleges$SATWR 0.027413 0.002357 11.633 < 2e-16 \*\*\*  
## colleges$SATM 0.024996 0.002054 12.168 < 2e-16 \*\*\*  
## colleges$GPA 0.038105 0.068181 0.559 0.57638   
## colleges$CollegePrep -0.093519 0.057593 -1.624 0.10476   
## colleges$TOEFL 0.047627 0.048139 0.989 0.32273   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.726 on 929 degrees of freedom  
## Multiple R-squared: 0.8151, Adjusted R-squared: 0.8137   
## F-statistic: 585.2 on 7 and 929 DF, p-value: < 2.2e-16

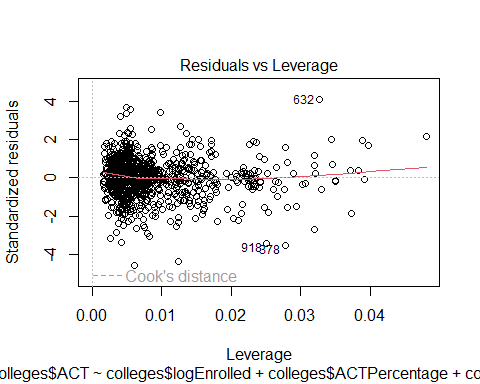
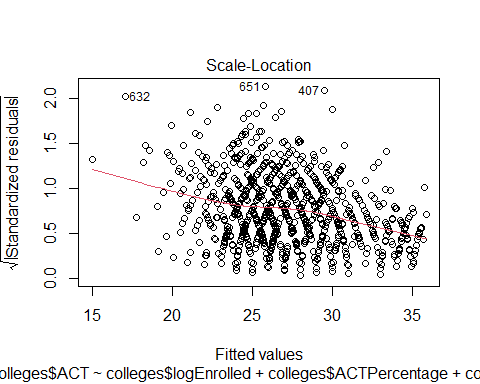
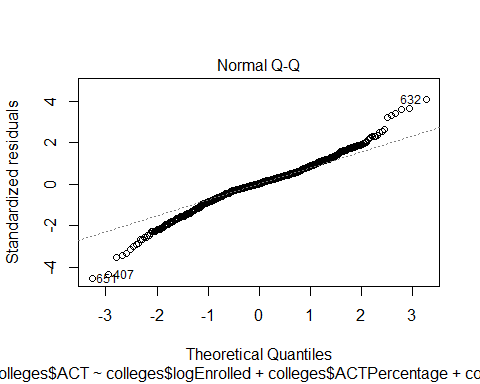
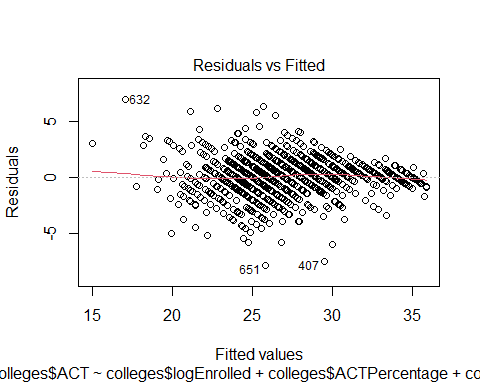
### First Order Model Summary

* The Significant predictor are: logEnrolled, ACTPercentage, SATWR and SATM.
* There SAT scores (dis aggregated by subjects) and the percentage of students who choose to submit their ACT scores appear to be correlated. This aligns with expectations that students with better scores usually choose to submit their scores.
* There is no evidence of curvature in the residuals.
* The residual variance appears constant.
* Appropriate Cook’s Distance for all entries (no outliers), according to the Std. Residuals / Leverage graph.

### First Order Model Visualization

Visualization of first order model.

plot(model1)



### Box-Cox Analysis

Finding Lambda Value:

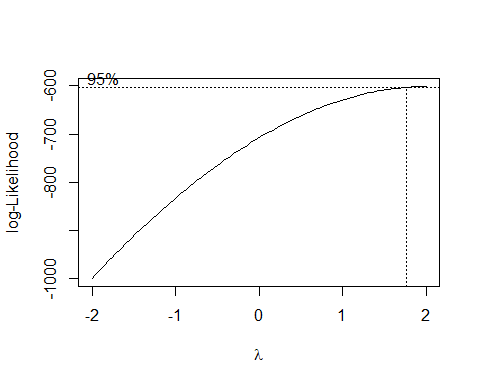
library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:olsrr':  
##   
## cement

## The following object is masked from 'package:dplyr':  
##   
## select

bc = boxcox(model1)



lambda = bc$x[which.max(bc$y)]  
cat(lambda)

## 2

Since Lambda = 2, therefore we run a quick fix.

colleges$ACT = ((colleges$ACT^lambda-1)/lambda)  
model1BC = lm(colleges$ACT ~ colleges$logEnrolled + colleges$ACTPercentage + colleges$SATWR + colleges$SATM + colleges$GPA + colleges$CollegePrep + colleges$TOEFL)  
summary(model1BC)

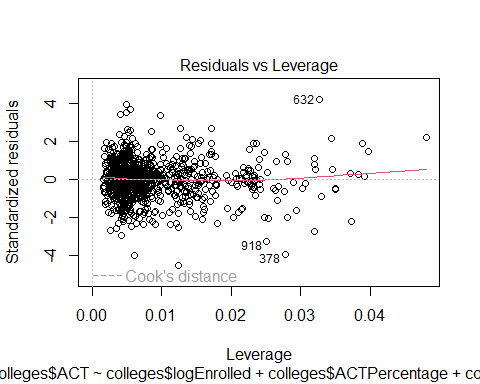
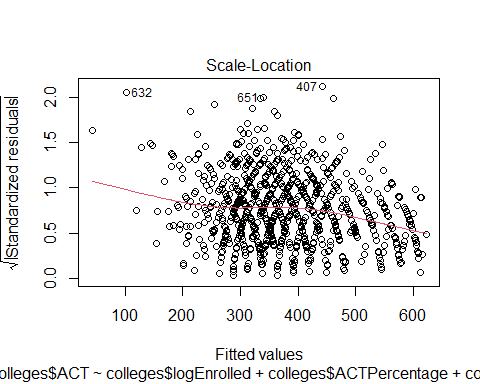
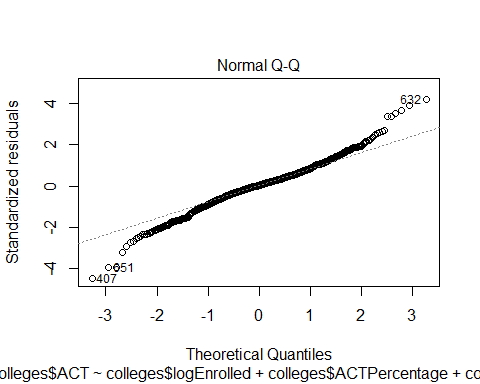
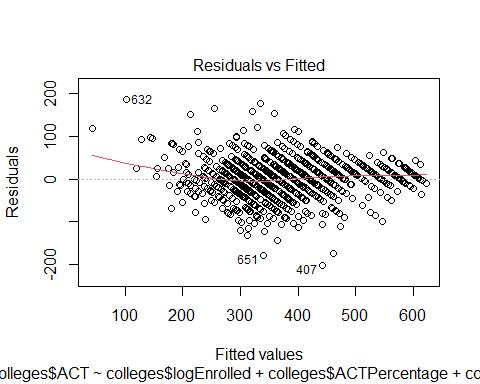
##   
## Call:  
## lm(formula = colleges$ACT ~ colleges$logEnrolled + colleges$ACTPercentage +   
## colleges$SATWR + colleges$SATM + colleges$GPA + colleges$CollegePrep +   
## colleges$TOEFL)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -200.809 -24.035 1.627 24.253 185.687   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -545.21233 17.68070 -30.837 < 2e-16 \*\*\*  
## colleges$logEnrolled 3.40279 1.41591 2.403 0.01645 \*   
## colleges$ACTPercentage -0.19497 0.05573 -3.499 0.00049 \*\*\*  
## colleges$SATWR 0.68324 0.06148 11.112 < 2e-16 \*\*\*  
## colleges$SATM 0.75212 0.05360 14.033 < 2e-16 \*\*\*  
## colleges$GPA 3.01139 1.77889 1.693 0.09082 .   
## colleges$CollegePrep -2.21597 1.50263 -1.475 0.14063   
## colleges$TOEFL 1.22971 1.25597 0.979 0.32779   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 45.03 on 929 degrees of freedom  
## Multiple R-squared: 0.8321, Adjusted R-squared: 0.8308   
## F-statistic: 657.5 on 7 and 929 DF, p-value: < 2.2e-16

Significant predictors: logEnrolled, ACTPercentage, SATWR and SATM, and *now* GPA!

### Residual Visualizations after Box-Cox Fix

Applying the ‘plot()’ function:

plot(model1BC)



Of the **significant** variables, only **ACTPercentage** seems to have a negative slope.

# Step-wise Regression

Here is the step wise regression on the first order model.

st = stepAIC(model1BC, direction="both")

## Start: AIC=7142.8  
## colleges$ACT ~ colleges$logEnrolled + colleges$ACTPercentage +   
## colleges$SATWR + colleges$SATM + colleges$GPA + colleges$CollegePrep +   
## colleges$TOEFL  
##   
## Df Sum of Sq RSS AIC  
## - colleges$TOEFL 1 1944 1885485 7141.8  
## <none> 1883542 7142.8  
## - colleges$CollegePrep 1 4409 1887951 7143.0  
## - colleges$GPA 1 5810 1889352 7143.7  
## - colleges$logEnrolled 1 11710 1895252 7146.6  
## - colleges$ACTPercentage 1 24817 1908359 7153.1  
## - colleges$SATWR 1 250366 2133907 7257.7  
## - colleges$SATM 1 399279 2282821 7321.0  
##   
## Step: AIC=7141.77  
## colleges$ACT ~ colleges$logEnrolled + colleges$ACTPercentage +   
## colleges$SATWR + colleges$SATM + colleges$GPA + colleges$CollegePrep  
##   
## Df Sum of Sq RSS AIC  
## - colleges$CollegePrep 1 4026 1889512 7141.8  
## <none> 1885485 7141.8  
## - colleges$GPA 1 5943 1891429 7142.7  
## + colleges$TOEFL 1 1944 1883542 7142.8  
## - colleges$logEnrolled 1 11472 1896958 7145.5  
## - colleges$ACTPercentage 1 24812 1910298 7152.0  
## - colleges$SATWR 1 251903 2137388 7257.3  
## - colleges$SATM 1 398509 2283995 7319.4  
##   
## Step: AIC=7141.77  
## colleges$ACT ~ colleges$logEnrolled + colleges$ACTPercentage +   
## colleges$SATWR + colleges$SATM + colleges$GPA  
##   
## Df Sum of Sq RSS AIC  
## <none> 1889512 7141.8  
## + colleges$CollegePrep 1 4026 1885485 7141.8  
## - colleges$GPA 1 5637 1895149 7142.6  
## + colleges$TOEFL 1 1561 1887951 7143.0  
## - colleges$logEnrolled 1 14114 1903626 7146.7  
## - colleges$ACTPercentage 1 25503 1915015 7152.3  
## - colleges$SATWR 1 255252 2144764 7258.5  
## - colleges$SATM 1 399823 2289335 7319.6

**lm(formula = collegeslogEnrolled + collegesSATWR + collegesGPA)**

# Centered Interaction Effects

First we create the centered dataset.

library(dplyr)  
colleges <- colleges %>% mutate\_if(is.character, as.numeric)  
CollegesM = colleges %>%   
 mutate\_at(.vars = colnames(colleges)[1:11],   
 .funs = list("scaled" = scale))

***Notice:*** The Above is an unorthodox, but arguably code-efficient, method of centering a dataset. The final dataset contains both ” **scaled** ” and non scaled version of all the original features. We simply choose the **scaled** version of numerical variables, and the original versions of categorical variables, when creating our model.

Creating interaction terms from **only the significant variables of the first order model**.

modelCen =lm(CollegesM$ACT ~ (CollegesM$logEnrolled\_scaled + CollegesM$ACTPercentage\_scaled +   
 CollegesM$SATWR\_scaled + CollegesM$SATM\_scaled + CollegesM$GPA\_scaled)^2)  
summary(modelCen)

##   
## Call:  
## lm(formula = CollegesM$ACT ~ (CollegesM$logEnrolled\_scaled +   
## CollegesM$ACTPercentage\_scaled + CollegesM$SATWR\_scaled +   
## CollegesM$SATM\_scaled + CollegesM$GPA\_scaled)^2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -185.476 -22.496 0.357 24.096 177.336   
##   
## Coefficients:  
## Estimate  
## (Intercept) 369.72789  
## CollegesM$logEnrolled\_scaled 5.07954  
## CollegesM$ACTPercentage\_scaled -8.48040  
## CollegesM$SATWR\_scaled 54.47629  
## CollegesM$SATM\_scaled 42.94485  
## CollegesM$GPA\_scaled -1.13563  
## CollegesM$logEnrolled\_scaled:CollegesM$ACTPercentage\_scaled 0.37052  
## CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled 10.95412  
## CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled -6.01962  
## CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled -3.40645  
## CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled -5.06717  
## CollegesM$ACTPercentage\_scaled:CollegesM$SATM\_scaled -4.54246  
## CollegesM$ACTPercentage\_scaled:CollegesM$GPA\_scaled 0.03517  
## CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled 4.64199  
## CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled -5.68932  
## CollegesM$SATM\_scaled:CollegesM$GPA\_scaled 6.21762  
## Std. Error t value  
## (Intercept) 1.81490 203.718  
## CollegesM$logEnrolled\_scaled 1.56424 3.247  
## CollegesM$ACTPercentage\_scaled 1.58523 -5.350  
## CollegesM$SATWR\_scaled 4.52844 12.030  
## CollegesM$SATM\_scaled 4.45560 9.638  
## CollegesM$GPA\_scaled 1.88807 -0.601  
## CollegesM$logEnrolled\_scaled:CollegesM$ACTPercentage\_scaled 1.62647 0.228  
## CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled 4.39449 2.493  
## CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled 4.03304 -1.493  
## CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled 1.51083 -2.255  
## CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled 3.24699 -1.561  
## CollegesM$ACTPercentage\_scaled:CollegesM$SATM\_scaled 3.67881 -1.235  
## CollegesM$ACTPercentage\_scaled:CollegesM$GPA\_scaled 1.97188 0.018  
## CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled 1.33021 3.490  
## CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled 4.37513 -1.300  
## CollegesM$SATM\_scaled:CollegesM$GPA\_scaled 4.34003 1.433  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## CollegesM$logEnrolled\_scaled 0.001207 \*\*   
## CollegesM$ACTPercentage\_scaled 1.11e-07 \*\*\*  
## CollegesM$SATWR\_scaled < 2e-16 \*\*\*  
## CollegesM$SATM\_scaled < 2e-16 \*\*\*  
## CollegesM$GPA\_scaled 0.547671   
## CollegesM$logEnrolled\_scaled:CollegesM$ACTPercentage\_scaled 0.819847   
## CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled 0.012853 \*   
## CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled 0.135891   
## CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled 0.024387 \*   
## CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled 0.118968   
## CollegesM$ACTPercentage\_scaled:CollegesM$SATM\_scaled 0.217233   
## CollegesM$ACTPercentage\_scaled:CollegesM$GPA\_scaled 0.985774   
## CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled 0.000507 \*\*\*  
## CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled 0.193797   
## CollegesM$SATM\_scaled:CollegesM$GPA\_scaled 0.152306   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 43.43 on 921 degrees of freedom  
## Multiple R-squared: 0.8451, Adjusted R-squared: 0.8426   
## F-statistic: 334.9 on 15 and 921 DF, p-value: < 2.2e-16

# Stepwise regression on centered interaction effects

Finally, creating a model with the centered variables, and their interaction terms, uding a stepwise approach.

stInt = stepAIC(modelCen, direction="both")

## Start: AIC=7083.18  
## CollegesM$ACT ~ (CollegesM$logEnrolled\_scaled + CollegesM$ACTPercentage\_scaled +   
## CollegesM$SATWR\_scaled + CollegesM$SATM\_scaled + CollegesM$GPA\_scaled)^2  
##   
## Df Sum of Sq  
## - CollegesM$ACTPercentage\_scaled:CollegesM$GPA\_scaled 1 0.6  
## - CollegesM$logEnrolled\_scaled:CollegesM$ACTPercentage\_scaled 1 97.9  
## - CollegesM$ACTPercentage\_scaled:CollegesM$SATM\_scaled 1 2876.3  
## - CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled 1 3190.1  
## <none>   
## - CollegesM$SATM\_scaled:CollegesM$GPA\_scaled 1 3871.9  
## - CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled 1 4202.8  
## - CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled 1 4594.5  
## - CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled 1 9590.4  
## - CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled 1 11722.0  
## - CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled 1 22973.8  
## RSS AIC  
## - CollegesM$ACTPercentage\_scaled:CollegesM$GPA\_scaled 1737500 7081.2  
## - CollegesM$logEnrolled\_scaled:CollegesM$ACTPercentage\_scaled 1737598 7081.2  
## - CollegesM$ACTPercentage\_scaled:CollegesM$SATM\_scaled 1740376 7082.7  
## - CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled 1740690 7082.9  
## <none> 1737500 7083.2  
## - CollegesM$SATM\_scaled:CollegesM$GPA\_scaled 1741372 7083.3  
## - CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled 1741702 7083.4  
## - CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled 1742094 7083.7  
## - CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled 1747090 7086.3  
## - CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled 1749222 7087.5  
## - CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled 1760473 7093.5  
##   
## Step: AIC=7081.18  
## CollegesM$ACT ~ CollegesM$logEnrolled\_scaled + CollegesM$ACTPercentage\_scaled +   
## CollegesM$SATWR\_scaled + CollegesM$SATM\_scaled + CollegesM$GPA\_scaled +   
## CollegesM$logEnrolled\_scaled:CollegesM$ACTPercentage\_scaled +   
## CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled + CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled +   
## CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled + CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled +   
## CollegesM$ACTPercentage\_scaled:CollegesM$SATM\_scaled + CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled +   
## CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled + CollegesM$SATM\_scaled:CollegesM$GPA\_scaled  
##   
## Df Sum of Sq  
## - CollegesM$logEnrolled\_scaled:CollegesM$ACTPercentage\_scaled 1 97.6  
## - CollegesM$ACTPercentage\_scaled:CollegesM$SATM\_scaled 1 2884.9  
## <none>   
## - CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled 1 3879.1  
## - CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled 1 4213.0  
## - CollegesM$SATM\_scaled:CollegesM$GPA\_scaled 1 4280.8  
## - CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled 1 4600.5  
## + CollegesM$ACTPercentage\_scaled:CollegesM$GPA\_scaled 1 0.6  
## - CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled 1 9612.0  
## - CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled 1 11759.7  
## - CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled 1 23114.8  
## RSS AIC  
## - CollegesM$logEnrolled\_scaled:CollegesM$ACTPercentage\_scaled 1737598 7079.2  
## - CollegesM$ACTPercentage\_scaled:CollegesM$SATM\_scaled 1740385 7080.7  
## <none> 1737500 7081.2  
## - CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled 1741379 7081.3  
## - CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled 1741713 7081.5  
## - CollegesM$SATM\_scaled:CollegesM$GPA\_scaled 1741781 7081.5  
## - CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled 1742101 7081.7  
## + CollegesM$ACTPercentage\_scaled:CollegesM$GPA\_scaled 1737500 7083.2  
## - CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled 1747112 7084.4  
## - CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled 1749260 7085.5  
## - CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled 1760615 7091.6  
##   
## Step: AIC=7079.24  
## CollegesM$ACT ~ CollegesM$logEnrolled\_scaled + CollegesM$ACTPercentage\_scaled +   
## CollegesM$SATWR\_scaled + CollegesM$SATM\_scaled + CollegesM$GPA\_scaled +   
## CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled + CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled +   
## CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled + CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled +   
## CollegesM$ACTPercentage\_scaled:CollegesM$SATM\_scaled + CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled +   
## CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled + CollegesM$SATM\_scaled:CollegesM$GPA\_scaled  
##   
## Df Sum of Sq  
## - CollegesM$ACTPercentage\_scaled:CollegesM$SATM\_scaled 1 2794.5  
## <none>   
## - CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled 1 3822.0  
## - CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled 1 4167.3  
## - CollegesM$SATM\_scaled:CollegesM$GPA\_scaled 1 4227.4  
## - CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled 1 4638.4  
## + CollegesM$logEnrolled\_scaled:CollegesM$ACTPercentage\_scaled 1 97.6  
## + CollegesM$ACTPercentage\_scaled:CollegesM$GPA\_scaled 1 0.3  
## - CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled 1 9572.1  
## - CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled 1 11665.1  
## - CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled 1 23255.3  
## RSS AIC  
## - CollegesM$ACTPercentage\_scaled:CollegesM$SATM\_scaled 1740392 7078.7  
## <none> 1737598 7079.2  
## - CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled 1741420 7079.3  
## - CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled 1741765 7079.5  
## - CollegesM$SATM\_scaled:CollegesM$GPA\_scaled 1741825 7079.5  
## - CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled 1742236 7079.7  
## + CollegesM$logEnrolled\_scaled:CollegesM$ACTPercentage\_scaled 1737500 7081.2  
## + CollegesM$ACTPercentage\_scaled:CollegesM$GPA\_scaled 1737598 7081.2  
## - CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled 1747170 7082.4  
## - CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled 1749263 7083.5  
## - CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled 1760853 7089.7  
##   
## Step: AIC=7078.74  
## CollegesM$ACT ~ CollegesM$logEnrolled\_scaled + CollegesM$ACTPercentage\_scaled +   
## CollegesM$SATWR\_scaled + CollegesM$SATM\_scaled + CollegesM$GPA\_scaled +   
## CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled + CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled +   
## CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled + CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled +   
## CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled + CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled +   
## CollegesM$SATM\_scaled:CollegesM$GPA\_scaled  
##   
## Df Sum of Sq  
## <none>   
## - CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled 1 4281  
## - CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled 1 4405  
## + CollegesM$ACTPercentage\_scaled:CollegesM$SATM\_scaled 1 2794  
## - CollegesM$SATM\_scaled:CollegesM$GPA\_scaled 1 4846  
## + CollegesM$ACTPercentage\_scaled:CollegesM$GPA\_scaled 1 9  
## + CollegesM$logEnrolled\_scaled:CollegesM$ACTPercentage\_scaled 1 7  
## - CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled 1 9691  
## - CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled 1 12322  
## - CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled 1 23474  
## - CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled 1 48521  
## RSS AIC  
## <none> 1740392 7078.7  
## - CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled 1744674 7079.0  
## - CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled 1744797 7079.1  
## + CollegesM$ACTPercentage\_scaled:CollegesM$SATM\_scaled 1737598 7079.2  
## - CollegesM$SATM\_scaled:CollegesM$GPA\_scaled 1745238 7079.3  
## + CollegesM$ACTPercentage\_scaled:CollegesM$GPA\_scaled 1740383 7080.7  
## + CollegesM$logEnrolled\_scaled:CollegesM$ACTPercentage\_scaled 1740385 7080.7  
## - CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled 1750083 7081.9  
## - CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled 1752714 7083.4  
## - CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled 1763866 7089.3  
## - CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled 1788913 7102.5

summary(stInt)

##   
## Call:  
## lm(formula = CollegesM$ACT ~ CollegesM$logEnrolled\_scaled + CollegesM$ACTPercentage\_scaled +   
## CollegesM$SATWR\_scaled + CollegesM$SATM\_scaled + CollegesM$GPA\_scaled +   
## CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled + CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled +   
## CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled + CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled +   
## CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled + CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled +   
## CollegesM$SATM\_scaled:CollegesM$GPA\_scaled)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -185.529 -22.282 0.383 24.052 176.260   
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) 369.596 1.810  
## CollegesM$logEnrolled\_scaled 5.048 1.557  
## CollegesM$ACTPercentage\_scaled -8.265 1.550  
## CollegesM$SATWR\_scaled 55.830 4.379  
## CollegesM$SATM\_scaled 41.895 4.358  
## CollegesM$GPA\_scaled -1.223 1.870  
## CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled 11.136 4.354  
## CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled -6.065 4.023  
## CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled -3.420 1.508  
## CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled -8.473 1.669  
## CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled 4.676 1.325  
## CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled -6.067 3.967  
## CollegesM$SATM\_scaled:CollegesM$GPA\_scaled 6.609 4.120  
## t value Pr(>|t|)   
## (Intercept) 204.246 < 2e-16 \*\*\*  
## CollegesM$logEnrolled\_scaled 3.242 0.001231 \*\*   
## CollegesM$ACTPercentage\_scaled -5.334 1.21e-07 \*\*\*  
## CollegesM$SATWR\_scaled 12.749 < 2e-16 \*\*\*  
## CollegesM$SATM\_scaled 9.614 < 2e-16 \*\*\*  
## CollegesM$GPA\_scaled -0.654 0.513271   
## CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled 2.558 0.010696 \*   
## CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled -1.508 0.131987   
## CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled -2.268 0.023541 \*   
## CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled -5.075 4.67e-07 \*\*\*  
## CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled 3.530 0.000436 \*\*\*  
## CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled -1.529 0.126543   
## CollegesM$SATM\_scaled:CollegesM$GPA\_scaled 1.604 0.109069   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 43.4 on 924 degrees of freedom  
## Multiple R-squared: 0.8448, Adjusted R-squared: 0.8428   
## F-statistic: 419.2 on 12 and 924 DF, p-value: < 2.2e-16

# Final Model

The final model is built using the variables chosen by stepwise regression on a set of predictor variables containing centered variables and their corresponding interaction effects.

FinalModel = lm(formula = CollegesM$ACT ~ CollegesM$logEnrolled\_scaled + CollegesM$ACTPercentage\_scaled +   
 CollegesM$SATWR\_scaled + CollegesM$SATM\_scaled + CollegesM$GPA\_scaled +   
 CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled + CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled +   
 CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled + CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled +   
 CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled + CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled +   
 CollegesM$SATM\_scaled:CollegesM$GPA\_scaled)  
summary(FinalModel)

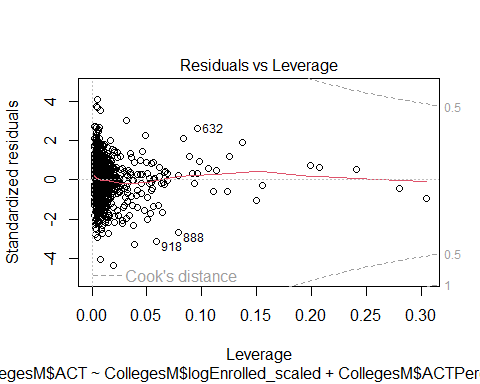
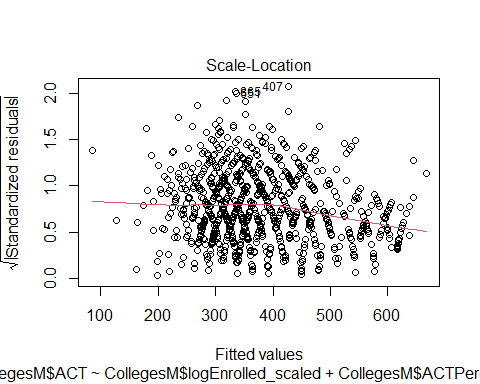
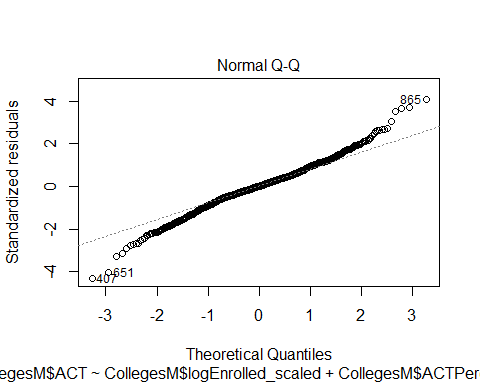
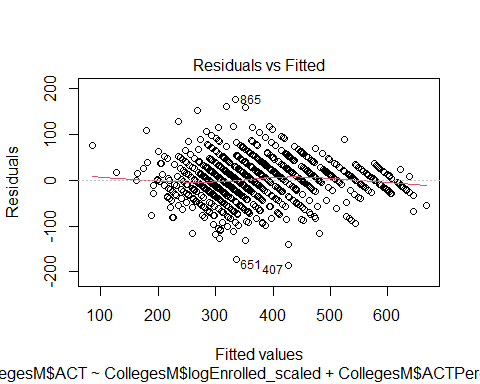
##   
## Call:  
## lm(formula = CollegesM$ACT ~ CollegesM$logEnrolled\_scaled + CollegesM$ACTPercentage\_scaled +   
## CollegesM$SATWR\_scaled + CollegesM$SATM\_scaled + CollegesM$GPA\_scaled +   
## CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled + CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled +   
## CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled + CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled +   
## CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled + CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled +   
## CollegesM$SATM\_scaled:CollegesM$GPA\_scaled)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -185.529 -22.282 0.383 24.052 176.260   
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) 369.596 1.810  
## CollegesM$logEnrolled\_scaled 5.048 1.557  
## CollegesM$ACTPercentage\_scaled -8.265 1.550  
## CollegesM$SATWR\_scaled 55.830 4.379  
## CollegesM$SATM\_scaled 41.895 4.358  
## CollegesM$GPA\_scaled -1.223 1.870  
## CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled 11.136 4.354  
## CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled -6.065 4.023  
## CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled -3.420 1.508  
## CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled -8.473 1.669  
## CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled 4.676 1.325  
## CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled -6.067 3.967  
## CollegesM$SATM\_scaled:CollegesM$GPA\_scaled 6.609 4.120  
## t value Pr(>|t|)   
## (Intercept) 204.246 < 2e-16 \*\*\*  
## CollegesM$logEnrolled\_scaled 3.242 0.001231 \*\*   
## CollegesM$ACTPercentage\_scaled -5.334 1.21e-07 \*\*\*  
## CollegesM$SATWR\_scaled 12.749 < 2e-16 \*\*\*  
## CollegesM$SATM\_scaled 9.614 < 2e-16 \*\*\*  
## CollegesM$GPA\_scaled -0.654 0.513271   
## CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled 2.558 0.010696 \*   
## CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled -1.508 0.131987   
## CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled -2.268 0.023541 \*   
## CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled -5.075 4.67e-07 \*\*\*  
## CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled 3.530 0.000436 \*\*\*  
## CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled -1.529 0.126543   
## CollegesM$SATM\_scaled:CollegesM$GPA\_scaled 1.604 0.109069   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 43.4 on 924 degrees of freedom  
## Multiple R-squared: 0.8448, Adjusted R-squared: 0.8428   
## F-statistic: 419.2 on 12 and 924 DF, p-value: < 2.2e-16

# Analysis of Final Model

### Plots of Final Model

Visualizing the fitted model.

plot(FinalModel)



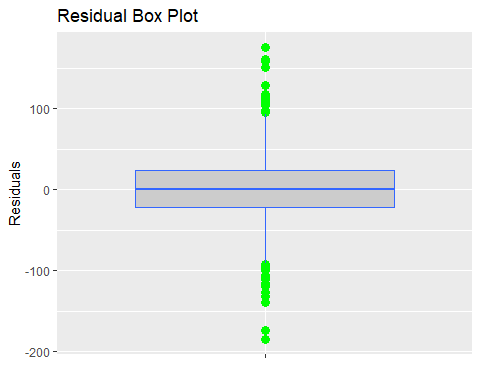
### Quick Visual Check of Model

* No homoskedasticity in residual vs fitted graph.(Non constant variance)
* No curvature effects
* No obvious outliers
* The residuals do appear to follow a normal distribution, albeit with fatter tails.

### Box Plot of residuals

Visualization of residual distribution.

ols\_plot\_resid\_box(FinalModel)



### Partial Regression plots

Added variable plots using avPLots:

library(car)

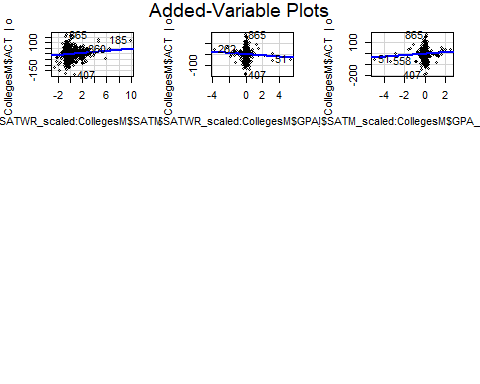
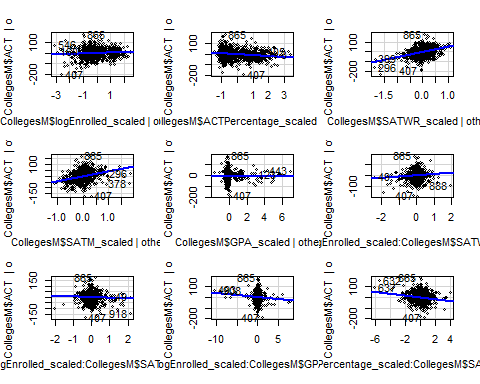
## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

## The following object is masked from 'package:purrr':  
##   
## some

avPlots(FinalModel)

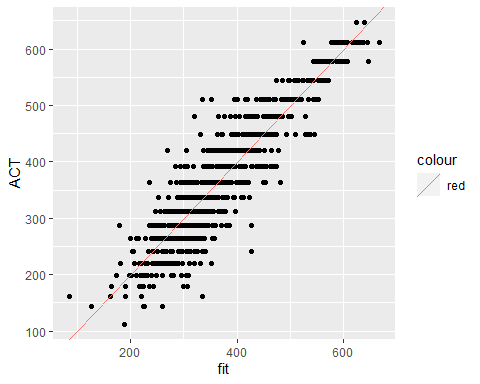


The strongest positive parameters seem to be SATM and SATWR.

### Response v Fitted plot.

Visualization with a 0 - 1, intercept slope line.

CollegesM$fit <- predict(FinalModel)  
ggplot(CollegesM) + geom\_point(aes(x = fit , y = ACT)) + geom\_abline(aes(intercept =0, slope=1, color="red"))

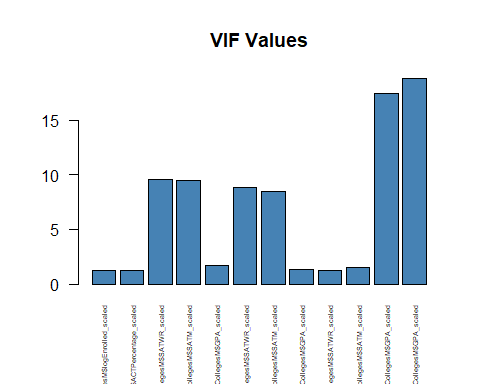


### Variance Inflaction Factors

vif\_values <- vif(FinalModel)

## there are higher-order terms (interactions) in this model  
## consider setting type = 'predictor'; see ?vif

#create horizontal bar chart to display each VIF value  
barplot(vif\_values, main = "VIF Values", horiz = F, col = "steelblue", las = 2, cex.names=0.45, mar =50)



vif\_values

## CollegesM$logEnrolled\_scaled   
## 1.205068   
## CollegesM$ACTPercentage\_scaled   
## 1.193169   
## CollegesM$SATWR\_scaled   
## 9.529892   
## CollegesM$SATM\_scaled   
## 9.437538   
## CollegesM$GPA\_scaled   
## 1.736888   
## CollegesM$logEnrolled\_scaled:CollegesM$SATWR\_scaled   
## 8.879179   
## CollegesM$logEnrolled\_scaled:CollegesM$SATM\_scaled   
## 8.434006   
## CollegesM$logEnrolled\_scaled:CollegesM$GPA\_scaled   
## 1.352211   
## CollegesM$ACTPercentage\_scaled:CollegesM$SATWR\_scaled   
## 1.240619   
## CollegesM$SATWR\_scaled:CollegesM$SATM\_scaled   
## 1.548713   
## CollegesM$SATWR\_scaled:CollegesM$GPA\_scaled   
## 17.465893   
## CollegesM$SATM\_scaled:CollegesM$GPA\_scaled   
## 18.795768

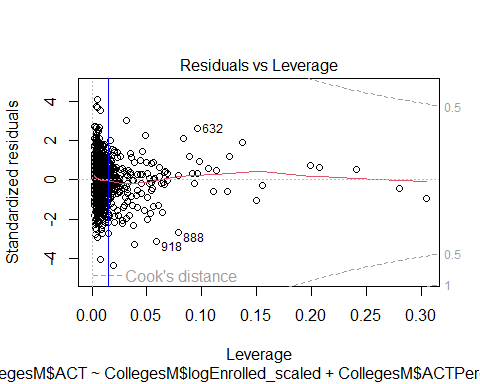
We see that **SATWR** and **SATM** are highly correlated to **GPA**.

### Leverage Values

We know the formula for high leverage cut off is 2(k+1)/n, where k is the number of features and the n is the number of observations. Given that our dataset has 937 observations, and 6 predictors, our leverage cut off is 0.014941302.

Finally plotting Standardized Residuals vs Leverage, with cutoff in the Leverage axis.

plot(FinalModel, which = 5) + abline(v=0.014941302, col="blue")



## integer(0)

**Needs Work** 5 D & E

### Interaction Plots

Our model included various interaction affects. Let us visualize them here.

par (mfrow=c(1,1))  
# Function to categorize a continuous variable into its quartiles  
categorize = function (x) {  
 quartiles = summary (x) [c(2, 3, 5)]  
 result = rep ("Q1", length (x))  
 result [(quartiles[1] < x) & (x <= quartiles [2])] = "Q2"  
 result [(quartiles[2] < x) & (x <= quartiles [3])] = "Q3"  
 result [quartiles[3] < x] = "Q4"  
 return (result)  
}  
  
# Interaction plots using the ggplot and dplyr packages  
  
library (ggplot2)  
library (dplyr)  
  
# Plot Log sales price vs Log sqfeet by quality level  
# Note the use of the "with" function to avoid having to specify  
# each column name with "realest$" in front of it  
  
with (CollegesM,  
 qplot (x= ACT, y= SATM\_scaled, color=as.factor (SATWR\_scaled)) +  
 geom\_smooth (method="lm"))

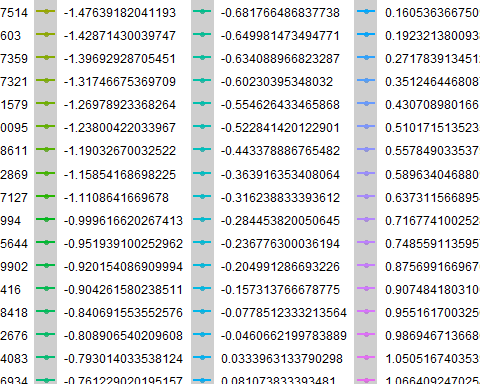
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.

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

## Warning in qt((1 - level)/2, df): NaNs produced

## Warning in qt((1 - level)/2, df): NaNs produced  
  
## Warning in qt((1 - level)/2, df): NaNs produced  
  
## Warning in qt((1 - level)/2, df): NaNs produced  
  
## Warning in qt((1 - level)/2, df): NaNs produced  
  
## Warning in qt((1 - level)/2, df): NaNs produced  
  
## Warning in qt((1 - level)/2, df): NaNs produced  
  
## Warning in qt((1 - level)/2, df): NaNs produced  
  
## Warning in qt((1 - level)/2, df): NaNs produced  
  
## Warning in qt((1 - level)/2, df): NaNs produced  
  
## Warning in qt((1 - level)/2, df): NaNs produced  
  
## Warning in qt((1 - level)/2, df): NaNs produced  
  
## Warning in qt((1 - level)/2, df): NaNs produced

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning  
## -Inf  
  
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning  
## -Inf

 ```