R Notebook

# Set global figure size  
knitr::opts\_chunk$set(fig.width=6, fig.height=3.5)

## Abstract

For this project we are going to look at the chance of getting into graduate school for undergraduate students. The data that we have contains the student’s ID numbers that were generated instead of using the students real name to protect the students identity. We also have student’s different test scores, cgpa, the rating of the university/college they attend and the chances of students. Our explanatory variables are TOEFL score, University Rating, CGPA and Research. The response variable is the chance of getting into graduate school.

## Introduction

TOEFL score is a discrete, quantitative variable. This is an exam that each student takes as part and uses the test scores as part of their application for graduate school admission. The highest score is the better score. This score usually shows how a student will perform if he gets admitted into graduate school. University Ratings is a variable that describes how good the University that a student attended for their undergraduate. This is more like a dummy variable to describe how good a University is based on a range of attributes. This variable ranges from 0 to 5 0 being bad and the best grade a university can get is 5. We wanted to see if the school you attended might affect your chance of getting into graduate school. CGPA is the cumulative GPA. It is a continuous variable ranging from 0 upto 10. Shows your academic performance as an undergraduate. The research variable is a dummy variable that shows whether a student has done a research paper or not. 1 is for the ones that have done the research paper and the 0 for those who did not. We will be using these explanatory variables to try to explain the chances of getting into graduate school which is another continuous variable variable. This is a probability of a student getting into graduate school.

## Exploratory Analysis

# Read in the data  
  
Admissions = read.csv("/Users/thand/Downloads/Math 327 Project Draft/Admission\_Predict (2).csv")  
  
# Keep the complete cases, noting how many observations there are before and after.  
  
dim (Admissions)

## [1] 400 9

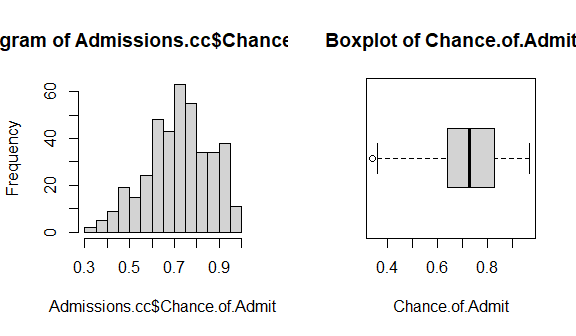
Admissions.cc = Admissions [complete.cases (Admissions), ]  
dim (Admissions.cc)

## [1] 400 9

There are 400 observations in the data set. First we performed some exporatory analysis to understand the distribution of each variable and their pairwise correlations. We also checked for obvious outliers and considered whether to transform any obviously skewed variables.

We will examine the association of each predictor with the Chance of admittance to determine the strength and direction of those associations. Then we will fit a first-order model with all of the predictors and use residual analysis and coefficient tests to determine how to proceed. We want to find a model that fits as well as possible using as few predictors as possible.

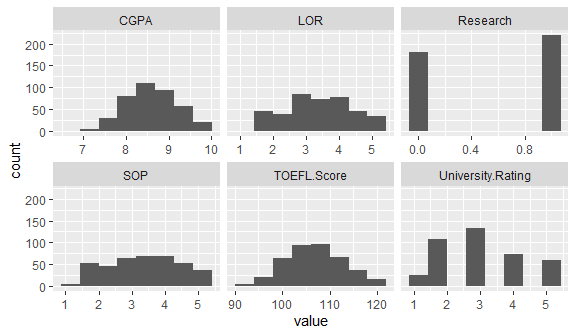
# Histogram and boxplot of the response variable  
  
par (mfrow = c(1, 2))  
hist (Admissions.cc$Chance.of.Admit)  
boxplot (Admissions.cc$Chance.of.Admit, horizontal = T, xlab="Chance.of.Admit", main="Boxplot of Chance.of.Admit")



The distribution of chances of admittance is moderately right-skewed. We will use residual analysis to guide the need for a transformation of chances of admittance.

Distributions of the quantitative predictor variables:

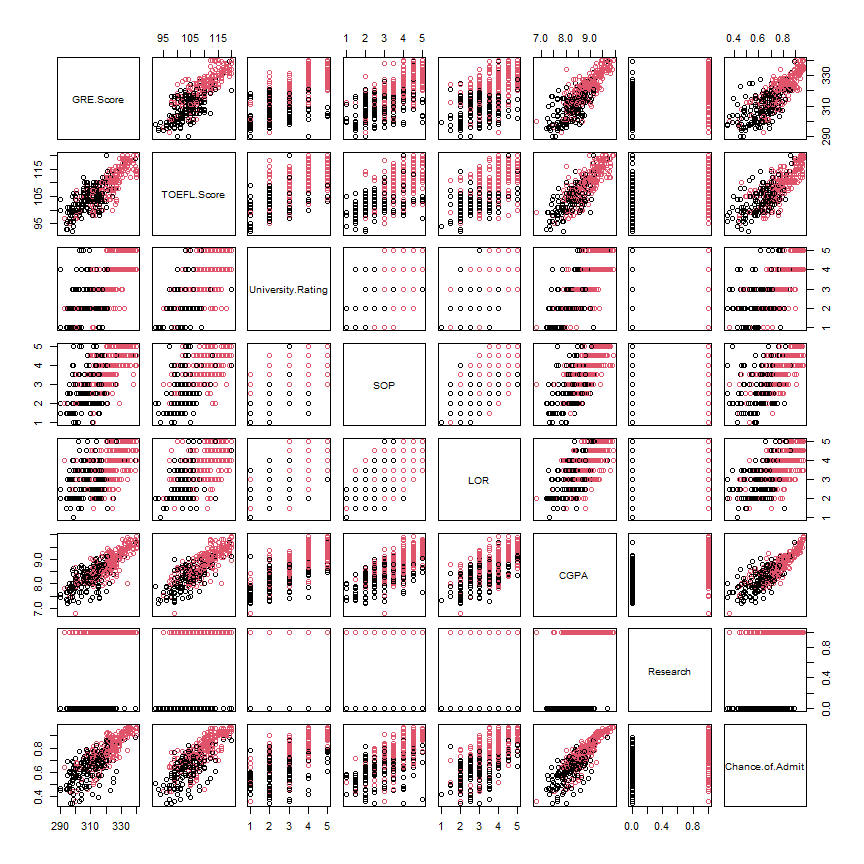
# ggplot2 and tidyr packages required  
library (ggplot2)  
library (tidyr)  
  
ggplot(gather(Admissions.cc [, 3:8]), aes(value)) +   
 geom\_histogram(bins = 8) +   
 facet\_wrap(~key, scales = 'free\_x')



None of the predictor variables are extremely right or left skewed. Therefore the necessary transformation can only be done after the box-cox analysis suggestion.

##Pair-wise correlation between variables

# Scatterplot matrix of columns 2 through 9  
plot (Admissions.cc[,2:9], col=Admissions.cc$Research+1)



cormat = cor (Admissions.cc [,2:8])  
round (cormat, 2)

## GRE.Score TOEFL.Score University.Rating SOP LOR CGPA  
## GRE.Score 1.00 0.84 0.67 0.61 0.56 0.83  
## TOEFL.Score 0.84 1.00 0.70 0.66 0.57 0.83  
## University.Rating 0.67 0.70 1.00 0.73 0.66 0.75  
## SOP 0.61 0.66 0.73 1.00 0.73 0.72  
## LOR 0.56 0.57 0.66 0.73 1.00 0.67  
## CGPA 0.83 0.83 0.75 0.72 0.67 1.00  
## Research 0.58 0.49 0.45 0.44 0.40 0.52  
## Research  
## GRE.Score 0.58  
## TOEFL.Score 0.49  
## University.Rating 0.45  
## SOP 0.44  
## LOR 0.40  
## CGPA 0.52  
## Research 1.00

The pairs plot shows that there is no correlation between Chance of Admittance and Research and University Rating. A log transformation may resolve that. This will be examined further via residual analysis. The four predictors, GRE Score, TOEFL Score, LOR and CPGA are correlated with each other.

## First Order Model

Next, we fit a first-order linear model will all of the quantitative predictors predictors.

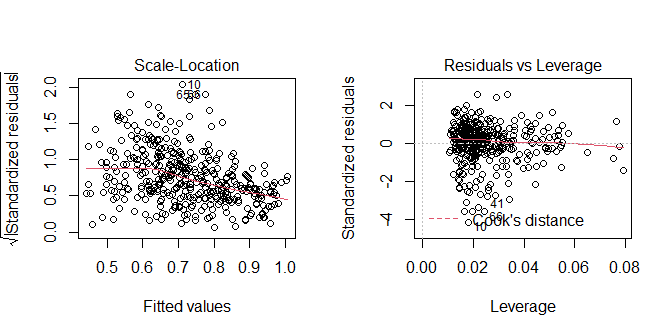
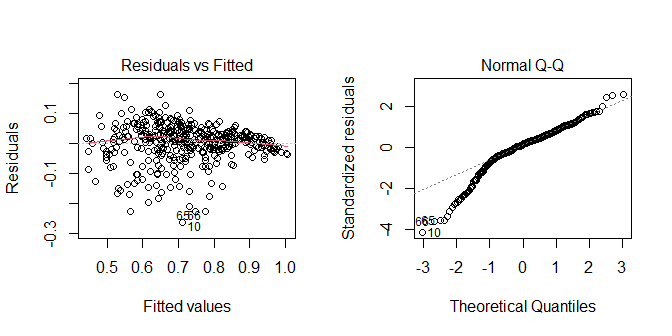
fit1 = lm (Chance.of.Admit ~ GRE.Score + TOEFL.Score + as.factor(University.Rating) + CGPA + as.factor(Research) + LOR , data=Admissions.cc)  
summary (fit1)

##   
## Call:  
## lm(formula = Chance.of.Admit ~ GRE.Score + TOEFL.Score + as.factor(University.Rating) +   
## CGPA + as.factor(Research) + LOR, data = Admissions.cc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.261454 -0.023215 0.008857 0.038054 0.161654   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.2404429 0.1267914 -9.783 < 2e-16 \*\*\*  
## GRE.Score 0.0017333 0.0005963 2.907 0.00386 \*\*   
## TOEFL.Score 0.0029999 0.0010837 2.768 0.00590 \*\*   
## as.factor(University.Rating)2 -0.0156008 0.0146433 -1.065 0.28736   
## as.factor(University.Rating)3 -0.0110124 0.0157121 -0.701 0.48379   
## as.factor(University.Rating)4 -0.0097924 0.0189673 -0.516 0.60595   
## as.factor(University.Rating)5 0.0075639 0.0208072 0.364 0.71641   
## CGPA 0.1178841 0.0120800 9.759 < 2e-16 \*\*\*  
## as.factor(Research)1 0.0240763 0.0079394 3.033 0.00259 \*\*   
## LOR 0.0217535 0.0050836 4.279 2.36e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.06368 on 390 degrees of freedom  
## Multiple R-squared: 0.8051, Adjusted R-squared: 0.8006   
## F-statistic: 179 on 9 and 390 DF, p-value: < 2.2e-16

Using all predictor variables, the model explains 80% of the variation in Chance of Admittance . The residual standard error is 0.064 (Chance of Admittance). From our table of results, we can see that every predictor variable is statistically significant in our model. We will the try to move around these variables to see if it will improve the significance of these predictor variables. The most significant predictors are CPGA and LOR.

## Residual Analysis - First-order model

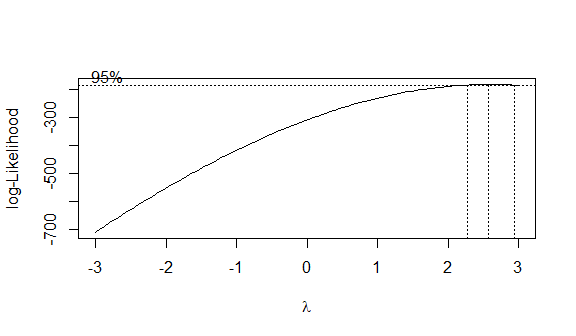
# Note: the fig.height option in the line above and the mfrow option in the line  
# below allow two plots wide and the possibility of a page break between the first  
# two and the last two residual plots.  
par (mfrow=c(1,2))  
plot (fit1)



The residual vs fitted value plot shows linearity but constant variance. The normal plot shows a a left-skewness by the deviation of the points from the mean on the on the lower end. Non-linearity present in the Q-Q plot indicates non-normal distribution for the residuals. The Scale-Location plot also indicates decreasing in residual variance with increasing fitted values.The Residuals vs. Leverage plot does not indicate any obvious concerns.

##Box-cox analysis

library ("MASS")  
boxcox (fit1, lambda = seq (-3, 3, by=0.1))



The Box-Cox analysis indicates that an optimal transformation would have a power just above 2. Since 2.5 is close to the 95% confidence interval for , we will use a squared transformation on the response variable for ease of interpretation.

##Second Model

fit2 = lm ((Chance.of.Admit)^2 ~ GRE.Score + TOEFL.Score + as.factor(University.Rating) + CGPA + as.factor(Research) + LOR, data=Admissions.cc)  
summary (fit2)

##   
## Call:  
## lm(formula = (Chance.of.Admit)^2 ~ GRE.Score + TOEFL.Score +   
## as.factor(University.Rating) + CGPA + as.factor(Research) +   
## LOR, data = Admissions.cc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.32039 -0.03630 0.01353 0.05113 0.21592   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.2150569 0.1619832 -13.675 < 2e-16 \*\*\*  
## GRE.Score 0.0024947 0.0007618 3.275 0.001151 \*\*   
## TOEFL.Score 0.0045277 0.0013845 3.270 0.001170 \*\*   
## as.factor(University.Rating)2 -0.0334076 0.0187076 -1.786 0.074912 .   
## as.factor(University.Rating)3 -0.0352642 0.0200730 -1.757 0.079737 .   
## as.factor(University.Rating)4 -0.0162274 0.0242318 -0.670 0.503462   
## as.factor(University.Rating)5 0.0237187 0.0265823 0.892 0.372796   
## CGPA 0.1613476 0.0154329 10.455 < 2e-16 \*\*\*  
## as.factor(Research)1 0.0370839 0.0101430 3.656 0.000291 \*\*\*  
## LOR 0.0277324 0.0064946 4.270 2.46e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08136 on 390 degrees of freedom  
## Multiple R-squared: 0.8398, Adjusted R-squared: 0.8361   
## F-statistic: 227.2 on 9 and 390 DF, p-value: < 2.2e-16

Using the transformed model of the response variable raise to the power 2, the model explains 84% of the variation in Chance of Admittance as compared to the 80% in the first model . The residual standard error is 0.084 (Chance of Admittance). Compared to model from using fit 1, the adjusted r-squared increased from 80% to 84%.

## Stepwise regression

Using the first model, with the transformation model that increased the by 5% increase in the adjusted r-squared.

fit2aic = step (fit2, direction = 'both')

## Start: AIC=-1997.26  
## (Chance.of.Admit)^2 ~ GRE.Score + TOEFL.Score + as.factor(University.Rating) +   
## CGPA + as.factor(Research) + LOR  
##   
## Df Sum of Sq RSS AIC  
## <none> 2.5814 -1997.3  
## - TOEFL.Score 1 0.07079 2.6522 -1988.4  
## - GRE.Score 1 0.07099 2.6524 -1988.4  
## - as.factor(Research) 1 0.08847 2.6698 -1985.8  
## - as.factor(University.Rating) 4 0.14037 2.7217 -1984.1  
## - LOR 1 0.12069 2.7021 -1981.0  
## - CGPA 1 0.72346 3.3048 -1900.4

fit1aic.int = step(fit1, direction = 'both', criterion = "BIC")

## Start: AIC=-2193.22  
## Chance.of.Admit ~ GRE.Score + TOEFL.Score + as.factor(University.Rating) +   
## CGPA + as.factor(Research) + LOR  
##   
## Df Sum of Sq RSS AIC  
## - as.factor(University.Rating) 4 0.01926 1.6008 -2196.4  
## <none> 1.5816 -2193.2  
## - TOEFL.Score 1 0.03108 1.6126 -2187.4  
## - GRE.Score 1 0.03427 1.6158 -2186.6  
## - as.factor(Research) 1 0.03729 1.6189 -2185.9  
## - LOR 1 0.07426 1.6558 -2176.9  
## - CGPA 1 0.38619 1.9678 -2107.8  
##   
## Step: AIC=-2196.38  
## Chance.of.Admit ~ GRE.Score + TOEFL.Score + CGPA + as.factor(Research) +   
## LOR  
##   
## Df Sum of Sq RSS AIC  
## <none> 1.6008 -2196.4  
## + as.factor(University.Rating) 4 0.01926 1.5816 -2193.2  
## - TOEFL.Score 1 0.03292 1.6338 -2190.2  
## - GRE.Score 1 0.03638 1.6372 -2189.4  
## - as.factor(Research) 1 0.03912 1.6400 -2188.7  
## - LOR 1 0.09133 1.6922 -2176.2  
## - CGPA 1 0.43201 2.0328 -2102.8

From the above step-wise regression, the university rating is the least significant predictor with the least AIC value of -2196.38.

## Centering Variables

GRE.centered = scale(Admissions.cc$GRE.Score, center = TRUE, scale = FALSE)  
TOEFL.centered = scale(Admissions.cc$TOEFL.Score, center = TRUE, scale =   
 FALSE)  
CGPA.centered = scale(Admissions.cc$CGPA, center = TRUE, scale = FALSE)  
LOR.centered = scale(Admissions.cc$LOR, center = TRUE, scale = FALSE)  
fit1aic.centered = lm(Chance.of.Admit ~ GRE.centered + TOEFL.centered +  
 as.factor(University.Rating) +   
 CGPA.centered +   
 as.factor(Research) +   
 LOR.centered, data=Admissions.cc)  
summary(fit1aic.centered)

##   
## Call:  
## lm(formula = Chance.of.Admit ~ GRE.centered + TOEFL.centered +   
## as.factor(University.Rating) + CGPA.centered + as.factor(Research) +   
## LOR.centered, data = Admissions.cc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.261454 -0.023215 0.008857 0.038054 0.161654   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.7196801 0.0155306 46.340 < 2e-16 \*\*\*  
## GRE.centered 0.0017333 0.0005963 2.907 0.00386 \*\*   
## TOEFL.centered 0.0029999 0.0010837 2.768 0.00590 \*\*   
## as.factor(University.Rating)2 -0.0156008 0.0146433 -1.065 0.28736   
## as.factor(University.Rating)3 -0.0110124 0.0157121 -0.701 0.48379   
## as.factor(University.Rating)4 -0.0097924 0.0189673 -0.516 0.60595   
## as.factor(University.Rating)5 0.0075639 0.0208072 0.364 0.71641   
## CGPA.centered 0.1178841 0.0120800 9.759 < 2e-16 \*\*\*  
## as.factor(Research)1 0.0240763 0.0079394 3.033 0.00259 \*\*   
## LOR.centered 0.0217535 0.0050836 4.279 2.36e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.06368 on 390 degrees of freedom  
## Multiple R-squared: 0.8051, Adjusted R-squared: 0.8006   
## F-statistic: 179 on 9 and 390 DF, p-value: < 2.2e-16

The adjusted r-squared is still at 80% and the residual standard error is at 0.064.

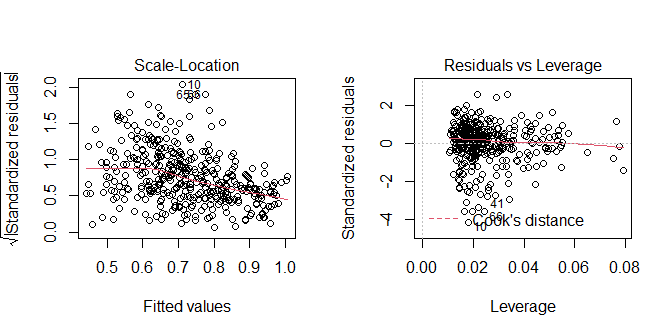
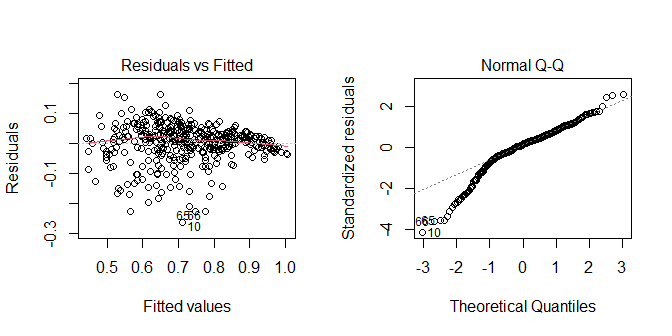
##Interaction effects of model from stepwise regression

#number 3b  
fit1sbc.int = lm(Chance.of.Admit ~ (GRE.Score + TOEFL.Score +   
 as.factor(University.Rating) +   
 CGPA + as.factor(Research) + LOR)^2,   
 data = Admissions.cc)  
summary(fit1sbc.int)

##   
## Call:  
## lm(formula = Chance.of.Admit ~ (GRE.Score + TOEFL.Score + as.factor(University.Rating) +   
## CGPA + as.factor(Research) + LOR)^2, data = Admissions.cc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.256532 -0.019277 0.005867 0.033754 0.151838   
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) -7.7536538 3.2882728  
## GRE.Score 0.0222912 0.0157450  
## TOEFL.Score 0.0531671 0.0386044  
## as.factor(University.Rating)2 -0.7788090 0.6102569  
## as.factor(University.Rating)3 -1.0526597 0.6604353  
## as.factor(University.Rating)4 -1.9616565 0.7839831  
## as.factor(University.Rating)5 -1.5548524 0.9000024  
## CGPA 0.2880069 0.4189018  
## as.factor(Research)1 -0.4531646 0.3369521  
## LOR 0.3722956 0.2219801  
## GRE.Score:TOEFL.Score -0.0001627 0.0001470  
## GRE.Score:as.factor(University.Rating)2 0.0009134 0.0024657  
## GRE.Score:as.factor(University.Rating)3 0.0024373 0.0026436  
## GRE.Score:as.factor(University.Rating)4 0.0029442 0.0032635  
## GRE.Score:as.factor(University.Rating)5 0.0021333 0.0038362  
## GRE.Score:CGPA -0.0005806 0.0015311  
## GRE.Score:as.factor(Research)1 0.0006907 0.0014683  
## GRE.Score:LOR -0.0003141 0.0010713  
## TOEFL.Score:as.factor(University.Rating)2 0.0041048 0.0050607  
## TOEFL.Score:as.factor(University.Rating)3 0.0025447 0.0054055  
## TOEFL.Score:as.factor(University.Rating)4 0.0101662 0.0064280  
## TOEFL.Score:as.factor(University.Rating)5 0.0082497 0.0072501  
## TOEFL.Score:CGPA 0.0001987 0.0032794  
## TOEFL.Score:as.factor(Research)1 0.0007723 0.0027416  
## TOEFL.Score:LOR -0.0014090 0.0018193  
## as.factor(University.Rating)2:CGPA 0.0216275 0.0487088  
## as.factor(University.Rating)3:CGPA 0.0101048 0.0498313  
## as.factor(University.Rating)4:CGPA -0.0078036 0.0632152  
## as.factor(University.Rating)5:CGPA 0.0048418 0.0797958  
## as.factor(University.Rating)2:as.factor(Research)1 0.0122482 0.0396567  
## as.factor(University.Rating)3:as.factor(Research)1 0.0149344 0.0419442  
## as.factor(University.Rating)4:as.factor(Research)1 0.0137739 0.0518872  
## as.factor(University.Rating)5:as.factor(Research)1 0.0282066 0.0595279  
## as.factor(University.Rating)2:LOR -0.0453352 0.0278184  
## as.factor(University.Rating)3:LOR -0.0256214 0.0291618  
## as.factor(University.Rating)4:LOR -0.0053572 0.0328801  
## as.factor(University.Rating)5:LOR -0.0136935 0.0371659  
## CGPA:as.factor(Research)1 0.0199675 0.0307482  
## CGPA:LOR -0.0088660 0.0187896  
## as.factor(Research)1:LOR -0.0015708 0.0134673  
## t value Pr(>|t|)   
## (Intercept) -2.358 0.0189 \*  
## GRE.Score 1.416 0.1577   
## TOEFL.Score 1.377 0.1693   
## as.factor(University.Rating)2 -1.276 0.2027   
## as.factor(University.Rating)3 -1.594 0.1118   
## as.factor(University.Rating)4 -2.502 0.0128 \*  
## as.factor(University.Rating)5 -1.728 0.0849 .  
## CGPA 0.688 0.4922   
## as.factor(Research)1 -1.345 0.1795   
## LOR 1.677 0.0944 .  
## GRE.Score:TOEFL.Score -1.107 0.2689   
## GRE.Score:as.factor(University.Rating)2 0.370 0.7113   
## GRE.Score:as.factor(University.Rating)3 0.922 0.3572   
## GRE.Score:as.factor(University.Rating)4 0.902 0.3676   
## GRE.Score:as.factor(University.Rating)5 0.556 0.5785   
## GRE.Score:CGPA -0.379 0.7047   
## GRE.Score:as.factor(Research)1 0.470 0.6384   
## GRE.Score:LOR -0.293 0.7695   
## TOEFL.Score:as.factor(University.Rating)2 0.811 0.4178   
## TOEFL.Score:as.factor(University.Rating)3 0.471 0.6381   
## TOEFL.Score:as.factor(University.Rating)4 1.582 0.1146   
## TOEFL.Score:as.factor(University.Rating)5 1.138 0.2559   
## TOEFL.Score:CGPA 0.061 0.9517   
## TOEFL.Score:as.factor(Research)1 0.282 0.7783   
## TOEFL.Score:LOR -0.774 0.4392   
## as.factor(University.Rating)2:CGPA 0.444 0.6573   
## as.factor(University.Rating)3:CGPA 0.203 0.8394   
## as.factor(University.Rating)4:CGPA -0.123 0.9018   
## as.factor(University.Rating)5:CGPA 0.061 0.9516   
## as.factor(University.Rating)2:as.factor(Research)1 0.309 0.7576   
## as.factor(University.Rating)3:as.factor(Research)1 0.356 0.7220   
## as.factor(University.Rating)4:as.factor(Research)1 0.265 0.7908   
## as.factor(University.Rating)5:as.factor(Research)1 0.474 0.6359   
## as.factor(University.Rating)2:LOR -1.630 0.1040   
## as.factor(University.Rating)3:LOR -0.879 0.3802   
## as.factor(University.Rating)4:LOR -0.163 0.8707   
## as.factor(University.Rating)5:LOR -0.368 0.7128   
## CGPA:as.factor(Research)1 0.649 0.5165   
## CGPA:LOR -0.472 0.6373   
## as.factor(Research)1:LOR -0.117 0.9072   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.06374 on 360 degrees of freedom  
## Multiple R-squared: 0.8198, Adjusted R-squared: 0.8002   
## F-statistic: 41.99 on 39 and 360 DF, p-value: < 2.2e-16

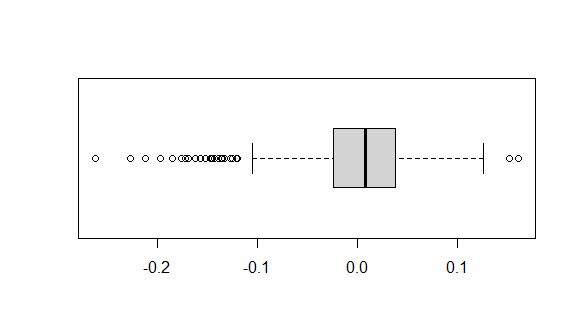
##Residual diagnostics

par (mfrow = c(1,2))  
plot (fit1aic.centered)



##Box plot of residuals

boxplot (fit1aic.centered$residuals, horizontal=T)



##Residuals vs. time or sequence of data collection, if appropriate

##Report and interpret the Variance Inflation Factors

library(car)

## Loading required package: carData

car::vif (fit1aic.centered)

## GVIF Df GVIF^(1/(2\*Df))  
## GRE.centered 4.605020 1 2.145931  
## TOEFL.centered 4.256686 1 2.063174  
## as.factor(University.Rating) 2.743878 4 1.134477  
## CGPA.centered 5.105494 1 2.259534  
## as.factor(Research) 1.540329 1 1.241100  
## LOR.centered 2.052606 1 1.432692

The residual value decreased to 0.047 but that also in turn reduced the adjusted r-squared value, therefore, the first model without the log valu os the better model to go with in with instance.

Next steps will include (1) evaluating interaction effects, (2) applying one or more model selection methods, (3) selecting a final model, and (4) evaluating some predictions from the final model.