Predictors Of Graduate School Acceptance Probability

Siyabonga Mabuza & Chilekwa Lingoma

10/29/2021

Table of Contents

[Abstract 1](#_Toc93102005)

[Introduction 2](#_Toc93102006)

[Exploratory Analysis 2](#_Toc93102007)

[The Distributions of the quantitative predictor variables: 3](#_Toc93102008)

[Pairwise Correlations. 4](#_Toc93102009)

[Model 1 6](#_Toc93102010)

[First Order Model With All Predictors. 6](#_Toc93102011)

[Residual Anaysis- First-order Model. 7](#_Toc93102012)

[Box-Cox Anaysis 9](#_Toc93102013)

[Model 2 - Cube of Chance of Admission vs. All Predictors 9](#_Toc93102014)

[Model 3- Remove Non-Significant Predictors. 12](#_Toc93102015)

[Interactions 13](#_Toc93102016)

[Model 4- Step Wise Regression With Centered Interaction Effects. 15](#_Toc93102017)

[Final Model 20](#_Toc93102018)

[Residual and Influence Anaysis 20](#_Toc93102019)

[Interaction Plot 26](#_Toc93102020)

[Conclusions. 28](#_Toc93102021)

# Abstract

This report examines graduate school acceptance probability as a function of several predictor variables such as College GPA, Graduate Record Examination, TOEFL, Statement of Purpose, and Letter Recommendation. The aim of this report is to build a linear regression model which can predict the graduate school acceptance probability with a suitable amount of precision to understand to the extent of each predictor’s relation to graduate school acceptance probability.

The prediction interval we would expect from our final model is 94.5%. Thus, we would expect and accuracy of approximately 95%. This model suggests that CGPA is the most significant predictor towards graduate school admission.

# Introduction

Upon completing their college education, college graduates often consider graduate school as a means to further their education. This report examines data on 400 college graduates from India. The primary goal of this report is to determine what factors are the most predictive of US graduate school acceptance probability (Chance of Admission) measured from a scale of 0-1. Moreover, this report aims to understand how the predictors are related to graduate school chance of admission. The predictor variables include:

• Graduate Record Examination (GRE) Score (out of 340)

• TOEFL Score (out of 120)

• Statement Of Purpose (SOP) (ranging from 1-5)

• Letter Of Recommendation (LOR) (ranging from 1-5)

• College GPA (CGPA) (out of 9.92)

Firstly, a basic level of investigative analysis was undertaken to understand the distribution of each variable and their pairwise correlations. Furthermore, we also checked for obvious outliers and considered whether to transform any obviously skewed variables.

We will examine the association of each predictor variable to the chance of admission to determine the strength and direction of each association. A first-order model will then be fitted with all of the predictor variables coupled with residual analysis and coefficient tests. This will dictate how we proceed with our analysis. The goal is to find a model with an optimum level of fitness while using the fewest number of predictors as possible.

The investigation will then potentially determine which predictor has the highest significance to US graduate school acceptance for any Indian college student wishing to further their education in the United States. Our hypothesis is: College GPA is the predictor with the largest significance to US graduate chance of admission. Generally, the higher the College GPA the higher the chance of admission. What concerns us with this hypothesis is the subjectivity associated with how each graduate institution values College GPA. Some institutions may only value College GPA while others may value College GPA as well as other predictors (TOEFL Score, Letter of Recommendation etc.) Some of the other predictors may even be more significant to chance of admission than College GPA. This reports seeks to examine if this is true or not.

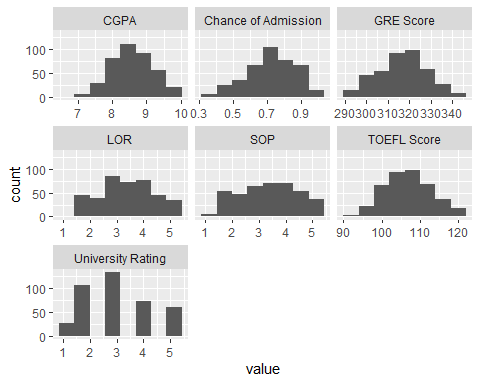
# Exploratory Analysis

library(readxl)  
US\_graduate\_scaled <- read\_excel("C:/Users/Siyabonga Mabuza/Desktop/Admissions dataset.xlsx")  
head(US\_graduate\_scaled)

## # A tibble: 6 x 8  
## `Serial No.` `GRE Score` `TOEFL Score` `University Rating` SOP LOR CGPA  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 337 118 4 4.5 4.5 9.65  
## 2 2 324 107 4 4 4.5 8.87  
## 3 3 316 104 3 3 3.5 8   
## 4 4 322 110 3 3.5 2.5 8.67  
## 5 5 314 103 2 2 3 8.21  
## 6 6 330 115 5 4.5 3 9.34  
## # ... with 1 more variable: Chance of Admission <dbl>

# The Distributions of the quantitative predictor variables:

library(ggplot2)  
library(tidyr)  
  
ggplot(gather(US\_graduate\_scaled [, 2:8]), aes(value)) +  
 geom\_histogram(bins = 8) +  
 facet\_wrap(~key, scales = "free\_x")

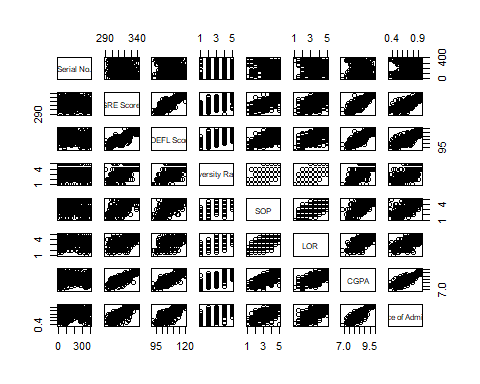


None of the predictor variables seem to be extremely right or left skewed. Thus, no transformations are necessary.

# Pairwise Correlations.

We will now examine pairwise correlations.

plot (US\_graduate\_scaled)



cormat = cor (US\_graduate\_scaled [,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  
## Chance of Admission 0.80 0.79 0.71 0.68 0.67 0.87  
## Chance of Admission  
## GRE Score 0.80  
## TOEFL Score 0.79  
## University Rating 0.71  
## SOP 0.68  
## LOR 0.67  
## CGPA 0.87  
## Chance of Admission 1.00

The pairwise correlations plot shows the linear relationships between Chance of Admission and GRE, CGPA, TOEFL, SOP, University Rating, and LOR. College GPA has the highest correlation with chance of admission (r=0.87). GRE score (r=0.80) and TOEFL score (r=0.79) also have relatively high correlations with chance of admission.

Three predictors, CGPA, GRE score, and TOEFL scores are highly correlated with each other (r=0.83 to 0.84)

# Model 1

The first model will include all of the individual predictors.

# First Order Model With All Predictors.

A first-order model will now be fitted with all of the individual predictors.

fit1 = lm (`Chance of Admission` ~ CGPA + LOR + SOP + `GRE Score`+ `TOEFL Score`+ `University Rating` , data=US\_graduate\_scaled)  
summary (fit1)

##   
## Call:  
## lm(formula = `Chance of Admission` ~ CGPA + LOR + SOP + `GRE Score` +   
## `TOEFL Score` + `University Rating`, data = US\_graduate\_scaled)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.279178 -0.023112 0.009864 0.035841 0.159383   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.4138594 0.1154455 -12.247 < 2e-16 \*\*\*  
## CGPA 0.1198749 0.0123470 9.709 < 2e-16 \*\*\*  
## LOR 0.0227486 0.0055995 4.063 5.86e-05 \*\*\*  
## SOP -0.0019614 0.0056041 -0.350 0.7265   
## `GRE Score` 0.0022761 0.0005779 3.938 9.70e-05 \*\*\*  
## `TOEFL Score` 0.0027534 0.0010999 2.503 0.0127 \*   
## `University Rating` 0.0060620 0.0048204 1.258 0.2093   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.06447 on 393 degrees of freedom  
## Multiple R-squared: 0.7987, Adjusted R-squared: 0.7956   
## F-statistic: 259.9 on 6 and 393 DF, p-value: < 2.2e-16

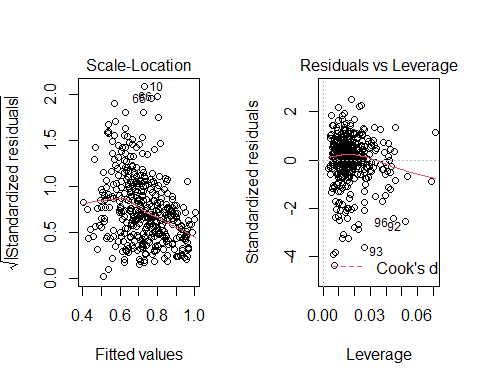
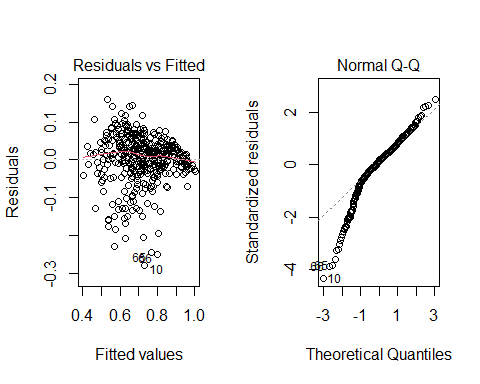
anova (fit1)

## Analysis of Variance Table  
##   
## Response: Chance of Admission  
## Df Sum Sq Mean Sq F value Pr(>F)   
## CGPA 1 6.1885 6.1885 1488.9854 < 2.2e-16 \*\*\*  
## LOR 1 0.1054 0.1054 25.3696 7.228e-07 \*\*\*  
## SOP 1 0.0024 0.0024 0.5726 0.449668   
## `GRE Score` 1 0.1487 0.1487 35.7751 4.983e-09 \*\*\*  
## `TOEFL Score` 1 0.0297 0.0297 7.1407 0.007849 \*\*   
## `University Rating` 1 0.0066 0.0066 1.5815 0.209288   
## Residuals 393 1.6334 0.0042   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The analysis of the variance table suggests that all but SOP are significant predictors. CGPA and GRE scores are the most significant predictors. Between the significant predictors TOEFL seems to have the lowest significance. The coefficient tests suggest that all predictors except SOP and University Rating are significant. CGPA has the highest significance, followed by GRE. SOP has the lowest significance followed by the TOEFL score. The adjusted R-squared is 0.7956, this indicates that 79.56% of the variability in chance of admission is explained by this model. The residual standard error is 0.06447 probability this is okay relative to the range of chance of admission values (0.34 to 0.97 probability)

# Residual Anaysis- First-order Model.

par (mfrow=c(1,2))  
plot(fit1)



• The Residuals vs Fitted plot shows good elements of linearity and constant variance. However, there is a slight curvature and the residual variance seems to be slightly decreasing with increasing fitted values. Thus, linearity and variance need can be improved.

• The Normal Q-Q plot shows that residuals relatively follow a normal distribution.There is some left skewness, this can be improved with a transformation.

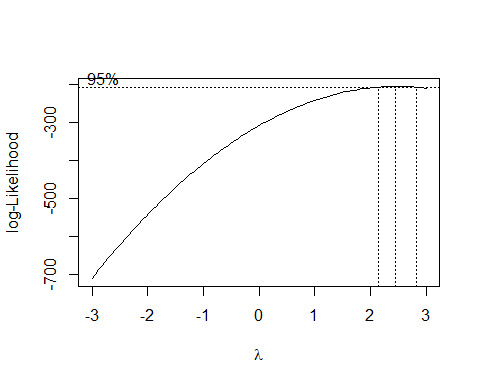
• The Scale-Location plot shows a slight amount of decreasing residual variance with increasing fitted values. This can be improved with a transformation.

• The Residuals vs Leverage plot shows that there are some outliers and possible high leverage values.This can be improved with a transformation.

Due to the aforementioned concerns we will run a Box-Cox analysis to find a power transformation that potentially addresses our concerns with the residuals.

# Box-Cox Anaysis

MASS::boxcox (fit1, lambda=c(-3, 3, by=0.1))



The Box-cox indicates that an optimal transformation would have a power close to 2.5, therefore, we will use a chance of admission^2.5 transformation.

# Model 2 - Cube of Chance of Admission vs. All Predictors

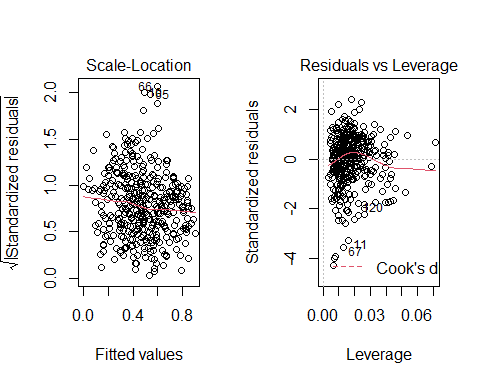
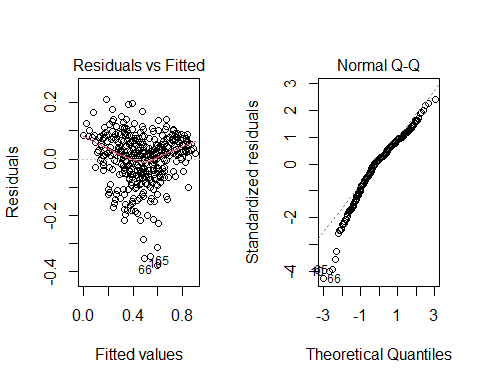
The power to 2.5 transformation will now be re-fitted into the first-order model with all of the predictor variables.

fit2 = lm((`Chance of Admission`)^2.5 ~ (CGPA) + (LOR) + (SOP) + (`GRE Score`) +  
(`TOEFL Score`)+ (`University Rating`), data=US\_graduate\_scaled)  
summary(fit2)

##   
## Call:  
## lm(formula = (`Chance of Admission`)^2.5 ~ (CGPA) + (LOR) + (SOP) +   
## (`GRE Score`) + (`TOEFL Score`) + (`University Rating`),   
## data = US\_graduate\_scaled)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.37658 -0.04377 0.01473 0.06023 0.21058   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.774e+00 1.582e-01 -17.536 < 2e-16 \*\*\*  
## CGPA 1.721e-01 1.692e-02 10.172 < 2e-16 \*\*\*  
## LOR 2.753e-02 7.674e-03 3.587 0.000376 \*\*\*  
## SOP -2.453e-05 7.680e-03 -0.003 0.997453   
## `GRE Score` 3.636e-03 7.920e-04 4.591 5.94e-06 \*\*\*  
## `TOEFL Score` 4.394e-03 1.507e-03 2.915 0.003761 \*\*   
## `University Rating` 1.744e-02 6.606e-03 2.639 0.008639 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08835 on 393 degrees of freedom  
## Multiple R-squared: 0.8331, Adjusted R-squared: 0.8306   
## F-statistic: 327 on 6 and 393 DF, p-value: < 2.2e-16

This model explains 83.06% of the variation in chance of admission^2.5. The residual standard error is 0.08835 probability^2.5.

par (mfrow=c(1,2))  
plot(fit2)



The residuals for model 2 compared to model 1 look slightly better, there is a more uniform spread. The normal Q-Q graph in particular is slightly better in this new model compared to model 1 but there is some left skewness, the residuals now have an improved normal distribution. However, Rows 10, 65, 66 are still obvious outliers.

CGPA seems to be the most significant predictor in this model followed by GRE Score and TOEFL Score. SOP has no statistical significance (P=0.997). Generally, all of the predictors have positive slopes. Thus, an increase in each predictor does increase the chance of admission, the extent of the increase in chance of admission depends on which predictor is increased.

confint(fit2)

## 2.5 % 97.5 %  
## (Intercept) -3.085406111 -2.463305995  
## CGPA 0.138855685 0.205389599  
## LOR 0.012443102 0.042617294  
## SOP -0.015123931 0.015074874  
## `GRE Score` 0.002079262 0.005193444  
## `TOEFL Score` 0.001430361 0.007357340  
## `University Rating` 0.004447631 0.030423089

• Average chance of admission^2.5 increases between 0.139 and 0.205 per CGPA with 95% confidence with other predictors held constant.

• Average chance of admission^2.5 increases between 0.012 and 0.0426 per LOR with 95% confidence with other predictors held constant.

• Average chance of admission^2.5 increases between -0.015 and 0.015 per SOP with 95% confidence with other predictors held constant.

• Average chance of admission^2.5 increases between 0.002 and 0.005 per GRE Score with 95% confidence with other predictors held constant.

•Average chance of admission^2.5 increases between 0.001 and 0.007 per TOEFL Score with 95% confidence with other predictors held constant.

•Average chance of admission^2.5 increases between 0.004 and 0.030 per University Rating with 95% confidence with other predictors held constant.

# Model 3- Remove Non-Significant Predictors.

Stepwise regression will now be applied to the chance of admission^2.5 model using the AIC criterion.

Fit2sqaic = step(fit2, direction = 'both')

## Start: AIC=-1934.21  
## (`Chance of Admission`)^2.5 ~ (CGPA) + (LOR) + (SOP) + (`GRE Score`) +   
## (`TOEFL Score`) + (`University Rating`)  
##   
## Df Sum of Sq RSS AIC  
## - SOP 1 0.00000 3.0677 -1936.2  
## <none> 3.0677 -1934.2  
## - `University Rating` 1 0.05437 3.1221 -1929.2  
## - `TOEFL Score` 1 0.06633 3.1341 -1927.7  
## - LOR 1 0.10046 3.1682 -1923.3  
## - `GRE Score` 1 0.16455 3.2323 -1915.3  
## - CGPA 1 0.80770 3.8754 -1842.7  
##   
## Step: AIC=-1936.21  
## (`Chance of Admission`)^2.5 ~ CGPA + LOR + `GRE Score` + `TOEFL Score` +   
## `University Rating`  
##   
## Df Sum of Sq RSS AIC  
## <none> 3.0677 -1936.2  
## + SOP 1 0.00000 3.0677 -1934.2  
## - `University Rating` 1 0.06000 3.1277 -1930.5  
## - `TOEFL Score` 1 0.06730 3.1350 -1929.5  
## - LOR 1 0.11999 3.1877 -1922.9  
## - `GRE Score` 1 0.16509 3.2328 -1917.2  
## - CGPA 1 0.82489 3.8926 -1843.0

The results indicate that SOP can be removed from the model.

# Interactions

New centered predictors will now be created for each predictor.

CGPA.c= scale(US\_graduate\_scaled$CGPA, center = TRUE, scale = FALSE)  
LOR.c= scale(US\_graduate\_scaled$LOR, center = TRUE, scale = FALSE)  
`GRE Score.c`= scale(US\_graduate\_scaled$`GRE Score`, center = TRUE, scale = FALSE)  
`TOEFL Score.c`= scale(US\_graduate\_scaled$`TOEFL Score`, center = TRUE, scale = FALSE)  
`University Rating.c`= scale(US\_graduate\_scaled$`University Rating`, center = TRUE, scale = FALSE)

Now, we will add all possible two-way interaction effects to the stepwise regression model obtained above.

fit2sbc.int = lm ((`Chance of Admission`)^2.5 ~ (CGPA.c+ LOR.c + `GRE Score.c`+ `TOEFL Score.c`+ `University Rating.c`)^2, data=US\_graduate\_scaled)  
summary (fit2sbc.int)

##   
## Call:  
## lm(formula = (`Chance of Admission`)^2.5 ~ (CGPA.c + LOR.c +   
## `GRE Score.c` + `TOEFL Score.c` + `University Rating.c`)^2,   
## data = US\_graduate\_scaled)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.36407 -0.03930 0.00844 0.05209 0.20056   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 4.539e-01 5.772e-03 78.646 < 2e-16  
## CGPA.c 1.725e-01 1.675e-02 10.299 < 2e-16  
## LOR.c 2.943e-02 6.755e-03 4.356 1.7e-05  
## `GRE Score.c` 3.156e-03 8.192e-04 3.853 0.000137  
## `TOEFL Score.c` 5.017e-03 1.486e-03 3.375 0.000813  
## `University Rating.c` 1.497e-02 6.163e-03 2.429 0.015579  
## CGPA.c:LOR.c -2.362e-03 2.198e-02 -0.107 0.914454  
## CGPA.c:`GRE Score.c` 8.832e-04 1.805e-03 0.489 0.624951  
## CGPA.c:`TOEFL Score.c` 3.966e-03 3.994e-03 0.993 0.321425  
## CGPA.c:`University Rating.c` -9.433e-04 2.104e-02 -0.045 0.964267  
## LOR.c:`GRE Score.c` -1.236e-04 1.178e-03 -0.105 0.916509  
## LOR.c:`TOEFL Score.c` -2.422e-03 2.190e-03 -1.106 0.269372  
## LOR.c:`University Rating.c` 1.505e-02 8.697e-03 1.731 0.084333  
## `GRE Score.c`:`TOEFL Score.c` -9.482e-05 1.657e-04 -0.572 0.567613  
## `GRE Score.c`:`University Rating.c` 5.732e-04 9.663e-04 0.593 0.553430  
## `TOEFL Score.c`:`University Rating.c` 1.554e-03 1.763e-03 0.881 0.378715  
##   
## (Intercept) \*\*\*  
## CGPA.c \*\*\*  
## LOR.c \*\*\*  
## `GRE Score.c` \*\*\*  
## `TOEFL Score.c` \*\*\*  
## `University Rating.c` \*   
## CGPA.c:LOR.c   
## CGPA.c:`GRE Score.c`   
## CGPA.c:`TOEFL Score.c`   
## CGPA.c:`University Rating.c`   
## LOR.c:`GRE Score.c`   
## LOR.c:`TOEFL Score.c`   
## LOR.c:`University Rating.c` .   
## `GRE Score.c`:`TOEFL Score.c`   
## `GRE Score.c`:`University Rating.c`   
## `TOEFL Score.c`:`University Rating.c`   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08399 on 384 degrees of freedom  
## Multiple R-squared: 0.8527, Adjusted R-squared: 0.8469   
## F-statistic: 148.1 on 15 and 384 DF, p-value: < 2.2e-16

Model 3 has a slightly higher adjusted R squared at 0.8469 compared to model 2 at 0.83. Furthermore, the residual standard error is lower (0.0840 vs. 0.0922).

# Model 4- Step Wise Regression With Centered Interaction Effects.

Now we will apply step wise regression to the model with the centered interaction effects from step 3.

fit2sbc.int2= step (fit2sbc.int, direction='both')

## Start: AIC=-1966.01  
## (`Chance of Admission`)^2.5 ~ (CGPA.c + LOR.c + `GRE Score.c` +   
## `TOEFL Score.c` + `University Rating.c`)^2  
##   
## Df Sum of Sq RSS AIC  
## - CGPA.c:`University Rating.c` 1 0.0000142 2.7086 -1968.0  
## - LOR.c:`GRE Score.c` 1 0.0000776 2.7087 -1968.0  
## - CGPA.c:LOR.c 1 0.0000815 2.7087 -1968.0  
## - CGPA.c:`GRE Score.c` 1 0.0016883 2.7103 -1967.8  
## - `GRE Score.c`:`TOEFL Score.c` 1 0.0023084 2.7109 -1967.7  
## - `GRE Score.c`:`University Rating.c` 1 0.0024817 2.7111 -1967.7  
## - `TOEFL Score.c`:`University Rating.c` 1 0.0054783 2.7141 -1967.2  
## - CGPA.c:`TOEFL Score.c` 1 0.0069527 2.7156 -1967.0  
## - LOR.c:`TOEFL Score.c` 1 0.0086300 2.7172 -1966.7  
## <none> 2.7086 -1966.0  
## - LOR.c:`University Rating.c` 1 0.0211246 2.7297 -1964.9  
##   
## Step: AIC=-1968.01  
## (`Chance of Admission`)^2.5 ~ CGPA.c + LOR.c + `GRE Score.c` +   
## `TOEFL Score.c` + `University Rating.c` + CGPA.c:LOR.c +   
## CGPA.c:`GRE Score.c` + CGPA.c:`TOEFL Score.c` + LOR.c:`GRE Score.c` +   
## LOR.c:`TOEFL Score.c` + LOR.c:`University Rating.c` + `GRE Score.c`:`TOEFL Score.c` +   
## `GRE Score.c`:`University Rating.c` + `TOEFL Score.c`:`University Rating.c`  
##   
## Df Sum of Sq RSS AIC  
## - LOR.c:`GRE Score.c` 1 0.0000659 2.7087 -1970.0  
## - CGPA.c:LOR.c 1 0.0001189 2.7087 -1970.0  
## - CGPA.c:`GRE Score.c` 1 0.0016944 2.7103 -1969.8  
## - `GRE Score.c`:`TOEFL Score.c` 1 0.0028100 2.7114 -1969.6  
## - `GRE Score.c`:`University Rating.c` 1 0.0031491 2.7118 -1969.5  
## - `TOEFL Score.c`:`University Rating.c` 1 0.0058880 2.7145 -1969.1  
## - CGPA.c:`TOEFL Score.c` 1 0.0079117 2.7165 -1968.8  
## - LOR.c:`TOEFL Score.c` 1 0.0088270 2.7174 -1968.7  
## <none> 2.7086 -1968.0  
## - LOR.c:`University Rating.c` 1 0.0236901 2.7323 -1966.5  
## + CGPA.c:`University Rating.c` 1 0.0000142 2.7086 -1966.0  
##   
## Step: AIC=-1970  
## (`Chance of Admission`)^2.5 ~ CGPA.c + LOR.c + `GRE Score.c` +   
## `TOEFL Score.c` + `University Rating.c` + CGPA.c:LOR.c +   
## CGPA.c:`GRE Score.c` + CGPA.c:`TOEFL Score.c` + LOR.c:`TOEFL Score.c` +   
## LOR.c:`University Rating.c` + `GRE Score.c`:`TOEFL Score.c` +   
## `GRE Score.c`:`University Rating.c` + `TOEFL Score.c`:`University Rating.c`  
##   
## Df Sum of Sq RSS AIC  
## - CGPA.c:LOR.c 1 0.0002071 2.7089 -1972.0  
## - CGPA.c:`GRE Score.c` 1 0.0016823 2.7104 -1971.8  
## - `GRE Score.c`:`TOEFL Score.c` 1 0.0030372 2.7117 -1971.5  
## - `GRE Score.c`:`University Rating.c` 1 0.0031272 2.7118 -1971.5  
## - `TOEFL Score.c`:`University Rating.c` 1 0.0066252 2.7153 -1971.0  
## - CGPA.c:`TOEFL Score.c` 1 0.0099963 2.7187 -1970.5  
## <none> 2.7087 -1970.0  
## - LOR.c:`TOEFL Score.c` 1 0.0150725 2.7238 -1969.8  
## - LOR.c:`University Rating.c` 1 0.0236822 2.7324 -1968.5  
## + LOR.c:`GRE Score.c` 1 0.0000659 2.7086 -1968.0  
## + CGPA.c:`University Rating.c` 1 0.0000024 2.7087 -1968.0  
##   
## Step: AIC=-1971.97  
## (`Chance of Admission`)^2.5 ~ CGPA.c + LOR.c + `GRE Score.c` +   
## `TOEFL Score.c` + `University Rating.c` + CGPA.c:`GRE Score.c` +   
## CGPA.c:`TOEFL Score.c` + LOR.c:`TOEFL Score.c` + LOR.c:`University Rating.c` +   
## `GRE Score.c`:`TOEFL Score.c` + `GRE Score.c`:`University Rating.c` +   
## `TOEFL Score.c`:`University Rating.c`  
##   
## Df Sum of Sq RSS AIC  
## - CGPA.c:`GRE Score.c` 1 0.0014807 2.7104 -1973.8  
## - `GRE Score.c`:`TOEFL Score.c` 1 0.0028331 2.7117 -1973.5  
## - `GRE Score.c`:`University Rating.c` 1 0.0030861 2.7120 -1973.5  
## - `TOEFL Score.c`:`University Rating.c` 1 0.0072328 2.7161 -1972.9  
## - CGPA.c:`TOEFL Score.c` 1 0.0097988 2.7187 -1972.5  
## <none> 2.7089 -1972.0  
## - LOR.c:`TOEFL Score.c` 1 0.0240665 2.7330 -1970.4  
## + CGPA.c:LOR.c 1 0.0002071 2.7087 -1970.0  
## + LOR.c:`GRE Score.c` 1 0.0001541 2.7087 -1970.0  
## + CGPA.c:`University Rating.c` 1 0.0000280 2.7089 -1970.0  
## - LOR.c:`University Rating.c` 1 0.0287533 2.7376 -1969.8  
##   
## Step: AIC=-1973.75  
## (`Chance of Admission`)^2.5 ~ CGPA.c + LOR.c + `GRE Score.c` +   
## `TOEFL Score.c` + `University Rating.c` + CGPA.c:`TOEFL Score.c` +   
## LOR.c:`TOEFL Score.c` + LOR.c:`University Rating.c` + `GRE Score.c`:`TOEFL Score.c` +   
## `GRE Score.c`:`University Rating.c` + `TOEFL Score.c`:`University Rating.c`  
##   
## Df Sum of Sq RSS AIC  
## - `GRE Score.c`:`TOEFL Score.c` 1 0.0020460 2.7124 -1975.5  
## - `TOEFL Score.c`:`University Rating.c` 1 0.0057581 2.7161 -1974.9  
## - `GRE Score.c`:`University Rating.c` 1 0.0083686 2.7187 -1974.5  
## <none> 2.7104 -1973.8  
## - LOR.c:`TOEFL Score.c` 1 0.0253033 2.7357 -1972.0  
## + CGPA.c:`GRE Score.c` 1 0.0014807 2.7089 -1972.0  
## - CGPA.c:`TOEFL Score.c` 1 0.0264554 2.7368 -1971.9  
## + LOR.c:`GRE Score.c` 1 0.0000326 2.7103 -1971.8  
## + CGPA.c:LOR.c 1 0.0000055 2.7104 -1971.8  
## + CGPA.c:`University Rating.c` 1 0.0000045 2.7104 -1971.8  
## - LOR.c:`University Rating.c` 1 0.0297486 2.7401 -1971.4  
##   
## Step: AIC=-1975.45  
## (`Chance of Admission`)^2.5 ~ CGPA.c + LOR.c + `GRE Score.c` +   
## `TOEFL Score.c` + `University Rating.c` + CGPA.c:`TOEFL Score.c` +   
## LOR.c:`TOEFL Score.c` + LOR.c:`University Rating.c` + `GRE Score.c`:`University Rating.c` +   
## `TOEFL Score.c`:`University Rating.c`  
##   
## Df Sum of Sq RSS AIC  
## - `TOEFL Score.c`:`University Rating.c` 1 0.005907 2.7183 -1976.6  
## - `GRE Score.c`:`University Rating.c` 1 0.006470 2.7189 -1976.5  
## <none> 2.7124 -1975.5  
## + `GRE Score.c`:`TOEFL Score.c` 1 0.002046 2.7104 -1973.8  
## + CGPA.c:`GRE Score.c` 1 0.000694 2.7117 -1973.5  
## + CGPA.c:`University Rating.c` 1 0.000646 2.7118 -1973.5  
## + CGPA.c:LOR.c 1 0.000023 2.7124 -1973.5  
## + LOR.c:`GRE Score.c` 1 0.000008 2.7124 -1973.5  
## - LOR.c:`TOEFL Score.c` 1 0.030050 2.7425 -1973.0  
## - CGPA.c:`TOEFL Score.c` 1 0.030846 2.7433 -1972.9  
## - LOR.c:`University Rating.c` 1 0.039311 2.7517 -1971.7  
##   
## Step: AIC=-1976.58  
## (`Chance of Admission`)^2.5 ~ CGPA.c + LOR.c + `GRE Score.c` +   
## `TOEFL Score.c` + `University Rating.c` + CGPA.c:`TOEFL Score.c` +   
## LOR.c:`TOEFL Score.c` + LOR.c:`University Rating.c` + `GRE Score.c`:`University Rating.c`  
##   
## Df Sum of Sq RSS AIC  
## <none> 2.7183 -1976.6  
## + `TOEFL Score.c`:`University Rating.c` 1 0.005907 2.7124 -1975.5  
## + `GRE Score.c`:`TOEFL Score.c` 1 0.002195 2.7161 -1974.9  
## + CGPA.c:`University Rating.c` 1 0.001883 2.7164 -1974.9  
## + LOR.c:`GRE Score.c` 1 0.001538 2.7168 -1974.8  
## - `GRE Score.c`:`University Rating.c` 1 0.025939 2.7443 -1974.8  
## - LOR.c:`TOEFL Score.c` 1 0.026390 2.7447 -1974.7  
## + CGPA.c:LOR.c 1 0.000408 2.7179 -1974.6  
## + CGPA.c:`GRE Score.c` 1 0.000166 2.7182 -1974.6  
## - LOR.c:`University Rating.c` 1 0.044733 2.7631 -1972.0  
## - CGPA.c:`TOEFL Score.c` 1 0.046009 2.7643 -1971.9

summary (fit2sbc.int2)

##   
## Call:  
## lm(formula = (`Chance of Admission`)^2.5 ~ CGPA.c + LOR.c + `GRE Score.c` +   
## `TOEFL Score.c` + `University Rating.c` + CGPA.c:`TOEFL Score.c` +   
## LOR.c:`TOEFL Score.c` + LOR.c:`University Rating.c` + `GRE Score.c`:`University Rating.c`,   
## data = US\_graduate\_scaled)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.36226 -0.03848 0.00905 0.05212 0.20071   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.4534098 0.0055214 82.118 < 2e-16 \*\*\*  
## CGPA.c 0.1719316 0.0158683 10.835 < 2e-16 \*\*\*  
## LOR.c 0.0295591 0.0066554 4.441 1.17e-05 \*\*\*  
## `GRE Score.c` 0.0032824 0.0007518 4.366 1.62e-05 \*\*\*  
## `TOEFL Score.c` 0.0047348 0.0014227 3.328 0.000958 \*\*\*  
## `University Rating.c` 0.0152626 0.0059563 2.562 0.010769 \*   
## CGPA.c:`TOEFL Score.c` 0.0046717 0.0018183 2.569 0.010563 \*   
## LOR.c:`TOEFL Score.c` -0.0026918 0.0013834 -1.946 0.052395 .   
## LOR.c:`University Rating.c` 0.0165361 0.0065274 2.533 0.011689 \*   
## `GRE Score.c`:`University Rating.c` 0.0009386 0.0004865 1.929 0.054441 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08349 on 390 degrees of freedom  
## Multiple R-squared: 0.8521, Adjusted R-squared: 0.8487   
## F-statistic: 249.7 on 9 and 390 DF, p-value: < 2.2e-16

four interactions were retained using this criterion. CGPA.c:TOEFL Score.c, LOR.c:TOEFL Score.c, LOR.c:TOEFL Score.c, and LOR.c:University Rating.c. Meaning that the slope between chance of admission ^2.5 and CPA is different for different levels of TOEFL Score. Furthermore, the slope between chance of admission ^2.5 and LOR and is different for different levels of TOEFL Score and University Rating.

summary(fit2sbc.int2)

##   
## Call:  
## lm(formula = (`Chance of Admission`)^2.5 ~ CGPA.c + LOR.c + `GRE Score.c` +   
## `TOEFL Score.c` + `University Rating.c` + CGPA.c:`TOEFL Score.c` +   
## LOR.c:`TOEFL Score.c` + LOR.c:`University Rating.c` + `GRE Score.c`:`University Rating.c`,   
## data = US\_graduate\_scaled)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.36226 -0.03848 0.00905 0.05212 0.20071   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.4534098 0.0055214 82.118 < 2e-16 \*\*\*  
## CGPA.c 0.1719316 0.0158683 10.835 < 2e-16 \*\*\*  
## LOR.c 0.0295591 0.0066554 4.441 1.17e-05 \*\*\*  
## `GRE Score.c` 0.0032824 0.0007518 4.366 1.62e-05 \*\*\*  
## `TOEFL Score.c` 0.0047348 0.0014227 3.328 0.000958 \*\*\*  
## `University Rating.c` 0.0152626 0.0059563 2.562 0.010769 \*   
## CGPA.c:`TOEFL Score.c` 0.0046717 0.0018183 2.569 0.010563 \*   
## LOR.c:`TOEFL Score.c` -0.0026918 0.0013834 -1.946 0.052395 .   
## LOR.c:`University Rating.c` 0.0165361 0.0065274 2.533 0.011689 \*   
## `GRE Score.c`:`University Rating.c` 0.0009386 0.0004865 1.929 0.054441 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08349 on 390 degrees of freedom  
## Multiple R-squared: 0.8521, Adjusted R-squared: 0.8487   
## F-statistic: 249.7 on 9 and 390 DF, p-value: < 2.2e-16

The adjusted R-square and residual standard error basically remained the same. Adjusted R-squared (0.8487 vs. 0.8469). Residual standard error (0.08349 vs. 0.08399)

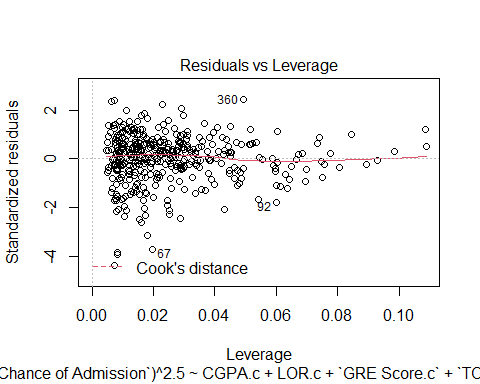
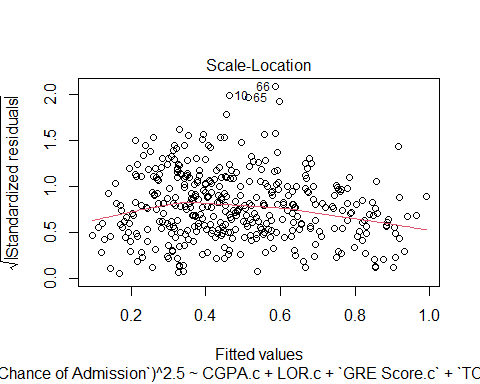
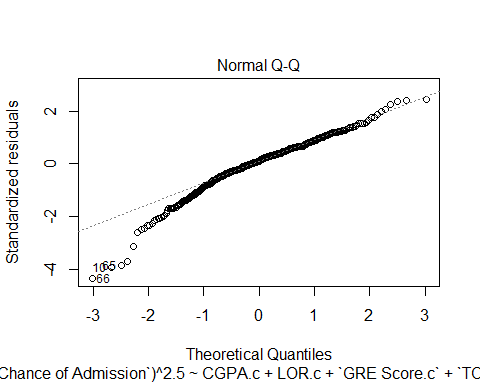
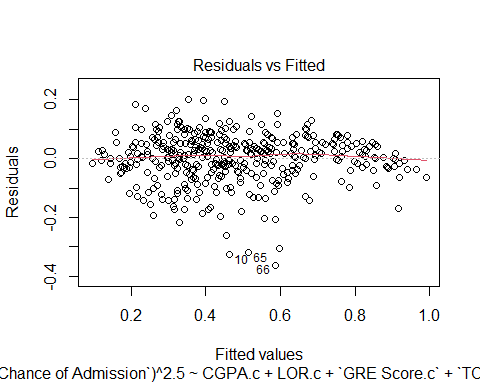
# Final Model

The model above appears to be our final model.

# Residual and Influence Anaysis

Next, we will examine the residuals vs fitted plot, the QQ plot, the square root of absolute standardized residuals vs fitted plot and the standardized residuals vs leverage, with Cook’s distance to understand the residuals of our final model.

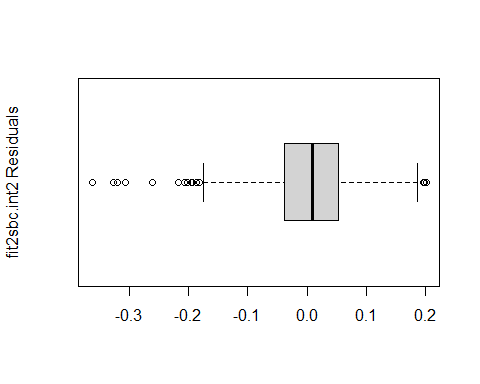
plot(fit2sbc.int2)



The residuals vs fitted plot shows good constant variance. The Normal Q-Q plot shows a mostly normal distribution and is only slightly left skewed. Square root of absolute standardized residuals vs fitted values plot equally looks good, with a slightly wider spread of residuals it does not show much evidence of issues within the data. The residuals vs leverage plot similarly shows good constant variance with some outliers but no points sitting outside cooks distance. However, there seems to be some high leverage values Thus, the data looks good enough for us to proceed. The students with serial numbers 10,65,66,67 are the obvious outliers illustrated by the plots.

Now we fit a box plot of the residuals.

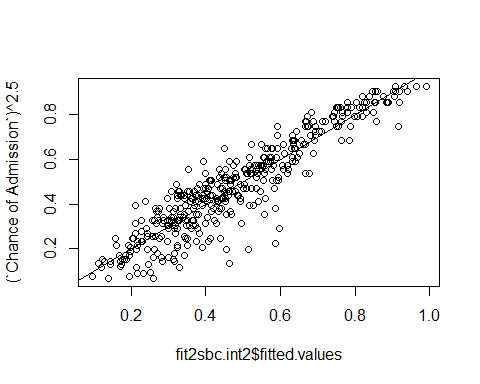
boxplot(fit2sbc.int2$residuals, ylab="fit2sbc.int2 Residuals", horizontal = T)



The box plot is asymmetric with several outliers which is not shocking because the residuals vs leverage plot showed that there are some outliers to be expected.

Next, we will plot chances of admission vs. fitted values.

plot ((`Chance of Admission`)^2.5 ~ fit2sbc.int2$fitted.values, data=US\_graduate\_scaled)   
abline(0,1)



The plot of cube chance of admission vs. fitted values shows relatively a good fit. There seems to be numerous outliers.

Next, we will examine variance inflation factors (vif).

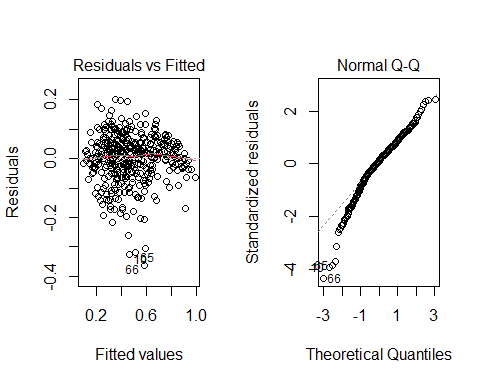
car:: vif (fit2sbc.int2)

## CGPA.c LOR.c   
## 5.125682 2.046921   
## `GRE Score.c` `TOEFL Score.c`   
## 4.259434 4.268423   
## `University Rating.c` CGPA.c:`TOEFL Score.c`   
## 2.656682 3.027404   
## LOR.c:`TOEFL Score.c` LOR.c:`University Rating.c`   
## 3.693898 2.622501   
## `GRE Score.c`:`University Rating.c`   
## 2.327916

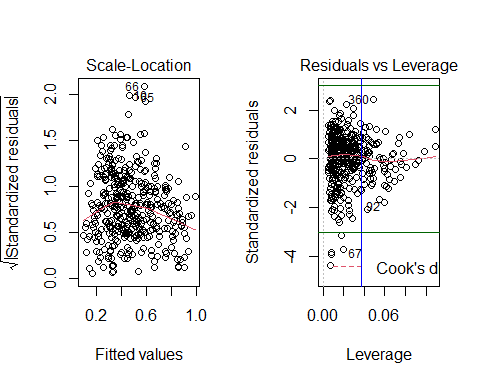
The VIF values all seem okay except CGPA.c which also happens to be our most significant predictor. CGPA may be strongly impacting the other individual variables and their predictive power, however, because it is our most significant predictor, we will not drop it. GRE Score.c and 4.268423 also have relatively high VIFs but they are below the 5 limit.

We will calculate the high leverage cutoff using the equation 3(k+1)/n. Our cut off is 3(4+1)/400 = 0.0375. Any values greater than 0.0375 are high leverage points.

par (mfrow=c(1,2))  
plot(fit2sbc.int2)



abline(v=0.0375, col="blue")  
abline(h =3, col = "darkgreen")  
abline(h =-3, col = "darkgreen")



There is a significant amount of high leverage values. There do not seem to be to be any points with high Cook’s distance. There are five outliers as illustrated by the horizontal lines in the residuals vs leverage plot.

# Interaction Plot

Below are the plots of two of our interaction effects. (There are four in total we will show only two)

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 [which ((quartiles[1] < x) & (x <= quartiles [2]))] = "Q2"  
 result [which ((quartiles[2] < x) & (x <= quartiles [3]))] = "Q3"  
 result [which (quartiles[3] < x)] = "Q4"  
 return (result)  
}

library (ggplot2)  
library (dplyr)

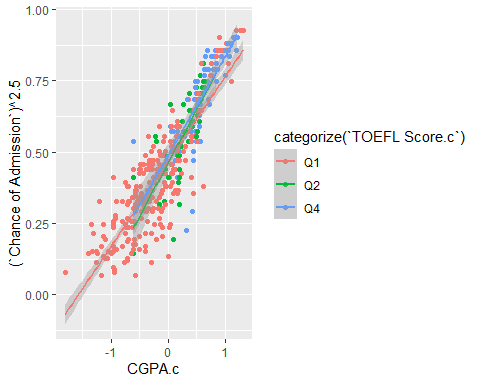
##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

with (US\_graduate\_scaled,  
 qplot (x=CGPA.c, y=(`Chance of Admission`)^2.5, color=categorize(`TOEFL Score.c`)) +  
 geom\_smooth (method="lm"))

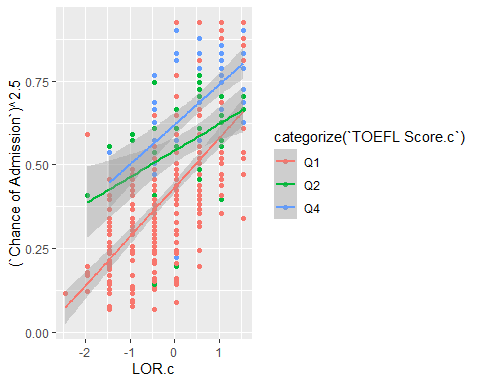
## `geom\_smooth()` using formula 'y ~ x'



All three quartiles maintain a similar graidiant therefore, we can see from this plot that the higher the TOEFL Score, and CGPA the higher the chances of admittance.

library (ggplot2)  
library (dplyr)  
  
with (US\_graduate\_scaled,  
 qplot (x=LOR.c, y=(`Chance of Admission`)^2.5, color=categorize(`TOEFL Score.c`)) +  
 geom\_smooth (method="lm"))

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



The plot indicates that with a higher LOR and TOEFL score ones chance of admittance is much higher. Quartile four (categorized in blue) indicates that with higher TOEFL Score and LOR ones chance of admission is greater than 0.50. There does not seem to be a trend in the slopes.

# Conclusions.

Our final model is:

Predicted cube chance of admission = 0.453 + 0.171 (CGPA) + 0.029 (LOR) + 0.003 (GRE SCORE) + 0.004 (TOEFL SCORE) + 0.015(University Rating) + 0.005 (CGPA)(TOEFL SCORE) - 0.002 (LOR)(TOEFL SCORE) + 0.004 (LOR)(UNIVERSITY RATING). 0.001(GRE SCORE)((UNIVERSITY RATING))

Since there are interaction effects between CGPA and TOEFL SCORE; LOR and TOEFL SCORE; LOR and UNIVERSITY RATING, we cannot numerically interpret a single slope for chance of admission^2.5 vs CGPA and LOR. Instead, we note from the interaction plots above that the relationship between chance of admission and CGPA is stronger when an applicant has a higher TOEFL SCORE. Furthermore, the relationship between chance of admission and LOR is stronger when an applicant has a higher TOEFL SCORE or University Rating.

The final model has adjusted R Squared of 0.8487, which means that 84.87% of the variation on chance of admission^2.5 is explained by the model. The residual standard error is 0.08349 probability^2.5.

To illustrate use of the model for prediction, we will show some example students applying to college and their predicted chance of admission:

preds = predict (fit2sbc.int2, interval='prediction')

## Warning in predict.lm(fit2sbc.int2, interval = "prediction"): predictions on current data refer to \_future\_ responses

US\_graduate\_scaled$pred.chance =(preds [,1])  
US\_graduate\_scaled$pred.lower =(preds [,2])  
US\_graduate\_scaled$pred.upper =(preds [,3])  
US\_graduate\_scaled$chance.of.admission.new=US\_graduate\_scaled$`Chance of Admission`^2.5  
  
US\_graduate\_scaled [c(1, 399, 78, 150, 250, 311),c("CGPA", "LOR", "GRE Score", "TOEFL Score","University Rating", "chance.of.admission.new", "pred.chance","pred.lower", "pred.upper" )]

## # A tibble: 6 x 9  
## CGPA LOR `GRE Score` `TOEFL Score` `University Rating` chance.of.admission~  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 9.65 4.5 337 118 4 0.812  
## 2 8.78 4 312 103 3 0.367  
## 3 8.22 2 301 99 2 0.328  
## 4 8.26 3 311 106 2 0.555  
## 5 8.83 4 321 111 3 0.520  
## 6 8.74 3.5 320 104 3 0.504  
## # ... with 3 more variables: pred.chance <dbl>, pred.lower <dbl>,  
## # pred.upper <dbl>

US\_graduate\_scaled$in.interval = ifelse (US\_graduate\_scaled$pred.lower <= US\_graduate\_scaled$chance.of.admission.new &  
 US\_graduate\_scaled$chance.of.admission.new <= US\_graduate\_scaled$pred.upper,  
 1, 0)  
mean (US\_graduate\_scaled$in.interval)

## [1] 0.945

All of the six example students whose predictions are shown above have prediction intervals that contain the observed chance of admission. Furthermore, the prediction interval we would expect from this model is 94.5%.

This model suggests that CGPA is the most significant predictor towards graduate school admission, thus, confirming our hypothesis.