Project 1

Matthias Rathbun

7/18/2022

rm(list=ls())

# Set up the work space

Reduce the number of displayed digits.

options(  
 digits = 4  
)

This project analyzes the data set Grad\_Admission. Variables description are as follows:

ID: Unique identification code for each student

GRE: GRE Scores (out of 340)

TOEFL: TOEFL Scores (out of 120)

Urate: University Rating (out of 5)

SOP: Statement of Purpose Strength (out of 5)

LOR: Letter of Recommendation Strength (out of 5)

CGPA: Undergraduate GPA (out of 4)

Chance:Chance of Admission (ranging from 0 to 100)

# 1. Load the data

Grad <- read.csv(  
 file = "Grad\_Admission.csv",  
 header = TRUE  
)

# 2. Print the dataset (Limit your print to include only the first and last ten observations)

head(Grad[,-1],10)

## GRE TOEFL Urate SOP LOR CGPA Chance  
## 1 337 118 4 4.5 4.5 3.86 92  
## 2 324 107 4 4.0 4.5 3.55 76  
## 3 316 104 3 3.0 3.5 3.20 72  
## 4 322 110 3 3.5 2.5 3.47 80  
## 5 314 103 2 2.0 3.0 3.28 65  
## 6 330 115 5 4.5 3.0 3.74 90  
## 7 321 109 3 3.0 4.0 3.28 75  
## 8 308 101 2 3.0 4.0 3.16 68  
## 9 302 102 1 2.0 1.5 3.20 50  
## 10 323 108 3 3.5 3.0 3.44 45

tail(Grad,10)

## ID GRE TOEFL Urate SOP LOR CGPA Chance  
## 391 391 314 102 2 2.0 2.5 3.30 64  
## 392 392 318 106 3 2.0 3.0 3.46 71  
## 393 393 326 112 4 4.0 3.5 3.65 84  
## 394 394 317 104 2 3.0 3.0 3.50 77  
## 395 395 329 111 4 4.5 4.0 3.69 89  
## 396 396 324 110 3 3.5 3.5 3.62 82  
## 397 397 325 107 3 3.0 3.5 3.64 84  
## 398 398 330 116 4 5.0 4.5 3.78 91  
## 399 399 312 103 3 3.5 4.0 3.51 67  
## 400 400 333 117 4 5.0 4.0 3.86 95

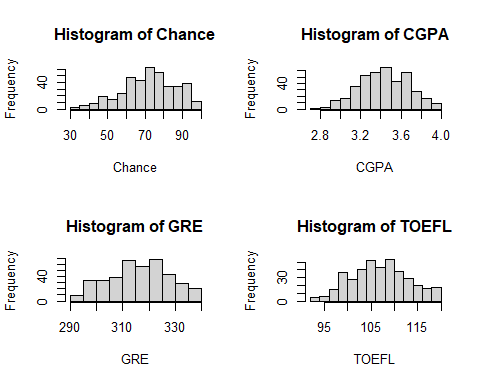
# 3. Print a table of the n/mean/standard deviation/min/max of three of the Features (GRE, TOEFL, CGPA).

summary(  
 object = Grad[c(2,3,7)]  
)

## GRE TOEFL CGPA   
## Min. :290 Min. : 92 Min. :2.72   
## 1st Qu.:308 1st Qu.:103 1st Qu.:3.27   
## Median :317 Median :107 Median :3.44   
## Mean :317 Mean :107 Mean :3.44   
## 3rd Qu.:325 3rd Qu.:112 3rd Qu.:3.62   
## Max. :340 Max. :120 Max. :3.97

# 4. Make histograms for “Chance”, “CGPA”, “GRE” and “TOEFL”

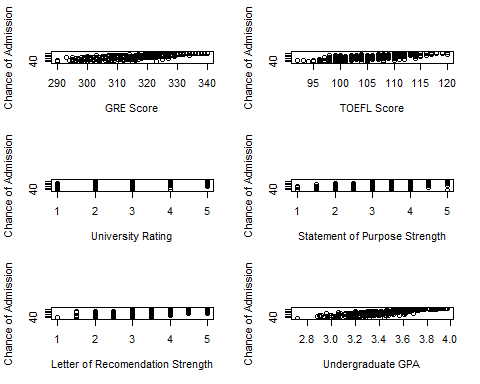
par(mfrow=c(2,2))  
hist(Grad[,8],xlab = "Chance", main = "Histogram of Chance")  
hist(Grad[,7],xlab = "CGPA", main = "Histogram of CGPA")  
hist(Grad[,2],xlab = "GRE", main = "Histogram of GRE")  
hist(Grad[,3],xlab = "TOEFL", main = "Histogram of TOEFL")



It seems that the probability distributions of Chance and CGPA are left skewed and they are not normal. The probability distributions of GRE and TOEFL are approximately normal.

# 5. Consider “Chance” as the response variable and print scatter plots of each of the other six variables against it.

par(mfrow=c(3,2))  
plot(x = Grad$GRE, y = Grad$Chance, xlab="GRE Score", ylab="Chance of Admission")  
plot(x = Grad$TOEFL, y = Grad$Chance, xlab="TOEFL Score", ylab="Chance of Admission")  
plot(x = Grad$Urate, y = Grad$Chance, xlab="University Rating", ylab="Chance of Admission")  
plot(x = Grad$SOP, y = Grad$Chance, xlab="Statement of Purpose Strength", ylab="Chance of Admission")  
plot(x = Grad$LOR, y = Grad$Chance, xlab="Letter of Recomendation Strength", ylab="Chance of Admission")  
plot(x = Grad$CGPA, y = Grad$Chance, xlab="Undergraduate GPA", ylab="Chance of Admission")



# 6. Draw your initial conclusions about the relationship between independent variables and the response variable based on the scatterplot.

There seems to be a strong positive linear relationship between Chance and the following features: GRE, TOEFL, and CGPA. The seems to be some relationship between SOP, LOR, Urate with respect to Chance. Since there are few possible rating options for those 3 features, it can be harder to tell if there is a relationship, but there seems to be a slight positive trend. These 3 features will not be useful by themselves, but can be useful to bolster a model that already contains GRE, TOEFL, or CGPA.

# 7. Confirm the validity of five major linear regression assumptions and comment on them.

test\_model <- lm(Chance~GRE+TOEFL+Urate+SOP+LOR+CGPA, data = Grad)

## 7.1. Existence

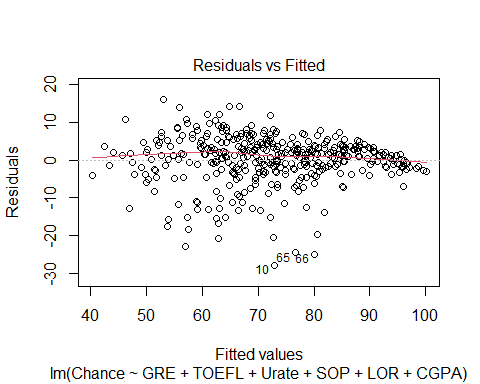
This assumption is always true for any regression model. Since a model can be made from the data, this assumption is true

## 7.2. Independence

Data is not a time-series. Each entry is independent of each other. Thus this assumption is true.

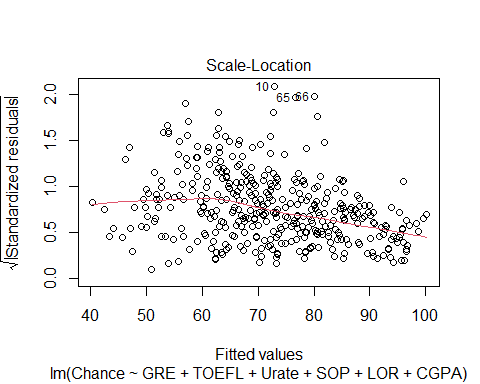
## 7.3. Linearity

plot(test\_model,1)

 Residuals are around a horizontal line without distinct patterns. This indicates Linearity. This assumption holds true.

## 7.4. Homoscedasticity

plot(test\_model, 3)



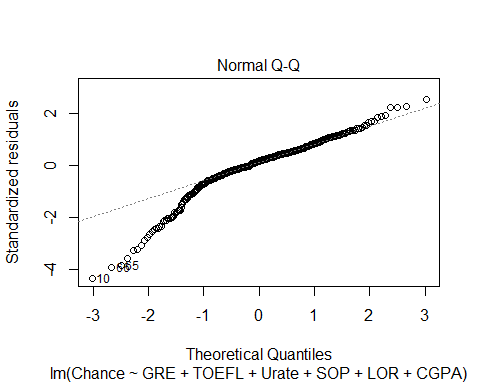
lmtest::bptest(test\_model)

##   
## studentized Breusch-Pagan test  
##   
## data: test\_model  
## BP = 20, df = 6, p-value = 0.003

While there seems to be less variance where Chance of Admission increases, the Breusch-Pagan Test at passes. This confirms the Homoscedasticity assumption.

## 7.5.

plot(test\_model, 2)



sresid <- MASS::studres(test\_model) #using MASS package function to transform data easily  
shapiro.test(sresid)

##   
## Shapiro-Wilk normality test  
##   
## data: sresid  
## W = 0.92, p-value = 9e-14

Most of the points fall along the reference line In the QQ plot. The left endpoints do deviate, suggesting the distribution is left-skewed. The model passes the Shapiro-Wilk test at . This confirms the Normality assumption of the model.

# 8. Choose the best three independent variables based on your immediate insight into the relationship and list them.

The best three independent variables based on the scatter plots in section 5 are GRE, TOEFL, CGPA. The model is Y: Chance X1: GRE X2: TOEFL X3: CGPA E: Error

# 9. Build up a table, including the correlation between all independent variables and the response variable.

model\_corr\_matrix <- cor(Grad[,-1],use = "pairwise.complete.obs")  
model\_corr\_matrix

## GRE TOEFL Urate SOP LOR CGPA Chance  
## GRE 1.0000 0.8360 0.6690 0.6128 0.5576 0.8332 0.8026  
## TOEFL 0.8360 1.0000 0.6956 0.6580 0.5677 0.8283 0.7916  
## Urate 0.6690 0.6956 1.0000 0.7345 0.6601 0.7469 0.7113  
## SOP 0.6128 0.6580 0.7345 1.0000 0.7296 0.7184 0.6757  
## LOR 0.5576 0.5677 0.6601 0.7296 1.0000 0.6702 0.6699  
## CGPA 0.8332 0.8283 0.7469 0.7184 0.6702 1.0000 0.8734  
## Chance 0.8026 0.7916 0.7113 0.6757 0.6699 0.8734 1.0000

All the features have a positive correlation with admission chance. It can be deduced that it will be worth testing all of these features in the model coefficient they all display correlation with the response variable. The relationship between correlation coefficient and slope is that their signs match. When Correlation coefficient is positive, slope is positive, and when it is negative, slope is negative.

# 10. Report the result of hypothesis testing for the correlation coefficient of each independent variable.

## 10.1. GRE vs Chance

cor.test(Grad$GRE, Grad$Chance, method = "pearson")

##   
## Pearson's product-moment correlation  
##   
## data: Grad$GRE and Grad$Chance  
## t = 27, df = 398, p-value <2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.7647 0.8350  
## sample estimates:  
## cor   
## 0.8026

## 10.2.TOEFL vs Chance

cor.test(Grad$TOEFL, Grad$Chance, method = "pearson")

##   
## Pearson's product-moment correlation  
##   
## data: Grad$TOEFL and Grad$Chance  
## t = 26, df = 398, p-value <2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.7519 0.8256  
## sample estimates:  
## cor   
## 0.7916

## 10.3. Urate vs Chance

cor.test(Grad$Urate, Grad$Chance, method = "pearson")

##   
## Pearson's product-moment correlation  
##   
## data: Grad$Urate and Grad$Chance  
## t = 20, df = 398, p-value <2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.6592 0.7565  
## sample estimates:  
## cor   
## 0.7113

## 10.4. SOP vs Chance

cor.test(Grad$SOP, Grad$Chance, method = "pearson")

##   
## Pearson's product-moment correlation  
##   
## data: Grad$SOP and Grad$Chance  
## t = 18, df = 398, p-value <2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.6187 0.7257  
## sample estimates:  
## cor   
## 0.6757

## 10.5. LOR vs Chance

cor.test(Grad$LOR, Grad$Chance, method = "pearson")

##   
## Pearson's product-moment correlation  
##   
## data: Grad$LOR and Grad$Chance  
## t = 18, df = 398, p-value <2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.6120 0.7206  
## sample estimates:  
## cor   
## 0.6699

## 10.6. CGPA vs Chance

cor.test(Grad$CGPA, Grad$Chance, method = "pearson")

##   
## Pearson's product-moment correlation  
##   
## data: Grad$CGPA and Grad$Chance  
## t = 36, df = 398, p-value <2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.8479 0.8948  
## sample estimates:  
## cor   
## 0.8734

# 11. Build up the univariate models

## 11.1. Chance regression on GRE

lm\_1 <- lm(  
 formula = Chance ~ GRE,  
 data = Grad  
)  
anova(lm\_1)

## Analysis of Variance Table  
##   
## Response: Chance  
## Df Sum Sq Mean Sq F value Pr(>F)   
## GRE 1 52273 52273 721 <2e-16 \*\*\*  
## Residuals 398 28873 73   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(lm\_1)

##   
## Call:  
## lm(formula = Chance ~ GRE, data = Grad)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -33.61 -4.60 0.41 5.64 18.34   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -243.6084 11.7814 -20.7 <2e-16 \*\*\*  
## GRE 0.9976 0.0372 26.8 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.52 on 398 degrees of freedom  
## Multiple R-squared: 0.644, Adjusted R-squared: 0.643   
## F-statistic: 721 on 1 and 398 DF, p-value: <2e-16

The model is:

## 11.2. Chance regression on TOEFL

lm\_2 <- lm(  
 formula = Chance ~ TOEFL,  
 data = Grad  
)  
anova(lm\_2)

## Analysis of Variance Table  
##   
## Response: Chance  
## Df Sum Sq Mean Sq F value Pr(>F)   
## TOEFL 1 50848 50848 668 <2e-16 \*\*\*  
## Residuals 398 30298 76   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(lm\_2)

##   
## Call:  
## lm(formula = Chance ~ TOEFL, data = Grad)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -31.25 -5.13 1.33 5.45 21.07   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -127.340 7.742 -16.4 <2e-16 \*\*\*  
## TOEFL 1.860 0.072 25.8 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.73 on 398 degrees of freedom  
## Multiple R-squared: 0.627, Adjusted R-squared: 0.626   
## F-statistic: 668 on 1 and 398 DF, p-value: <2e-16

The model is :

## 11.3. Chance regression on Urate

lm\_3 <- lm(  
 formula = Chance ~ Urate,  
 data = Grad  
)  
anova(lm\_3)

## Analysis of Variance Table  
##   
## Response: Chance  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Urate 1 41050 41050 407 <2e-16 \*\*\*  
## Residuals 398 40096 101   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(lm\_3)

##   
## Call:  
## lm(formula = Chance ~ Urate, data = Grad)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -38.53 -4.56 1.47 6.34 27.21   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 45.054 1.446 31.1 <2e-16 \*\*\*  
## Urate 8.868 0.439 20.2 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10 on 398 degrees of freedom  
## Multiple R-squared: 0.506, Adjusted R-squared: 0.505   
## F-statistic: 407 on 1 and 398 DF, p-value: <2e-16

The model is :

## 11.4. Chance regression on SOP

lm\_4 <- lm(  
 formula = Chance ~ SOP,  
 data = Grad  
)  
anova(lm\_4)

## Analysis of Variance Table  
##   
## Response: Chance  
## Df Sum Sq Mean Sq F value Pr(>F)   
## SOP 1 37053 37053 334 <2e-16 \*\*\*  
## Residuals 398 44094 111   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(lm\_4)

##   
## Call:  
## lm(formula = Chance ~ SOP, data = Grad)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -49.75 -5.39 1.82 7.04 22.39   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.894 1.856 21.5 <2e-16 \*\*\*  
## SOP 9.571 0.523 18.3 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.5 on 398 degrees of freedom  
## Multiple R-squared: 0.457, Adjusted R-squared: 0.455   
## F-statistic: 334 on 1 and 398 DF, p-value: <2e-16

The model is :

## 11.5. Chance regression on LOR

lm\_5 <- lm(  
 formula = Chance ~ LOR,  
 data = Grad  
)  
anova(lm\_5)

## Analysis of Variance Table  
##   
## Response: Chance  
## Df Sum Sq Mean Sq F value Pr(>F)   
## LOR 1 36414 36414 324 <2e-16 \*\*\*  
## Residuals 398 44732 112   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(lm\_5)

##   
## Call:  
## lm(formula = Chance ~ LOR, data = Grad)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -34.94 -6.26 0.06 7.39 29.33   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 35.726 2.107 16.9 <2e-16 \*\*\*  
## LOR 10.633 0.591 18.0 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.6 on 398 degrees of freedom  
## Multiple R-squared: 0.449, Adjusted R-squared: 0.447   
## F-statistic: 324 on 1 and 398 DF, p-value: <2e-16

The model is :

## 11.6. Chance Regression on CGPA

lm\_6 <- lm(  
 formula = Chance ~ CGPA,  
 data = Grad  
)  
anova(lm\_6)

## Analysis of Variance Table  
##   
## Response: Chance  
## Df Sum Sq Mean Sq F value Pr(>F)   
## CGPA 1 61898 61898 1280 <2e-16 \*\*\*  
## Residuals 398 19248 48   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(lm\_6)

##   
## Call:  
## lm(formula = Chance ~ CGPA, data = Grad)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -27.460 -2.959 0.973 4.180 18.076   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -107.22 5.03 -21.3 <2e-16 \*\*\*  
## CGPA 52.23 1.46 35.8 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.95 on 398 degrees of freedom  
## Multiple R-squared: 0.763, Adjusted R-squared: 0.762   
## F-statistic: 1.28e+03 on 1 and 398 DF, p-value: <2e-16

The model is :

# 12. Report the ANOVA table for two variables with the highest R-square value? What conclusion is achievable looking at these tables?

## 12.1. ANOVA for CGPA model

anova(lm\_6)

## Analysis of Variance Table  
##   
## Response: Chance  
## Df Sum Sq Mean Sq F value Pr(>F)   
## CGPA 1 61898 61898 1280 <2e-16 \*\*\*  
## Residuals 398 19248 48   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## 12.2. ANOVA for GRE model

anova(lm\_1)

## Analysis of Variance Table  
##   
## Response: Chance  
## Df Sum Sq Mean Sq F value Pr(>F)   
## GRE 1 52273 52273 721 <2e-16 \*\*\*  
## Residuals 398 28873 73   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Both of these variables are significant by themselves. But the CGPA has much less error than the GRE model. This means that CGPA explains more of the variance in Chance than GRE.

# 13. Build up a model, including all six variables available in the dataset.

full\_model <- lm(Chance~GRE+TOEFL+Urate+SOP+LOR+CGPA, data = Grad)  
anova(full\_model)

## Analysis of Variance Table  
##   
## Response: Chance  
## Df Sum Sq Mean Sq F value Pr(>F)   
## GRE 1 52273 52273 1258.2 < 2e-16 \*\*\*  
## TOEFL 1 3921 3921 94.4 < 2e-16 \*\*\*  
## Urate 1 2370 2370 57.0 3.0e-13 \*\*\*  
## SOP 1 757 757 18.2 2.5e-05 \*\*\*  
## LOR 1 1574 1574 37.9 1.9e-09 \*\*\*  
## CGPA 1 3923 3923 94.4 < 2e-16 \*\*\*  
## Residuals 393 16328 42   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(full\_model)

##   
## Call:  
## lm(formula = Chance ~ GRE + TOEFL + Urate + SOP + LOR + CGPA,   
## data = Grad)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -27.915 -2.381 0.935 3.577 16.057   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -141.4185 11.5420 -12.25 < 2e-16 \*\*\*  
## GRE 0.2270 0.0578 3.93 0.0001 \*\*\*  
## TOEFL 0.2762 0.1099 2.51 0.0124 \*   
## Urate 0.5995 0.4821 1.24 0.2144   
## SOP -0.2000 0.5604 -0.36 0.7213   
## LOR 2.2784 0.5598 4.07 5.7e-05 \*\*\*  
## CGPA 30.0138 3.0886 9.72 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.45 on 393 degrees of freedom  
## Multiple R-squared: 0.799, Adjusted R-squared: 0.796   
## F-statistic: 260 on 6 and 393 DF, p-value: <2e-16

The model is : Urate and SOP are not Significant at

# 14. Remove all non-significant variables from the model and rebuild the model.

best\_model <- lm(Chance~GRE+TOEFL+LOR+CGPA, data = Grad)  
anova(best\_model)

## Analysis of Variance Table  
##   
## Response: Chance  
## Df Sum Sq Mean Sq F value Pr(>F)   
## GRE 1 52273 52273 1259.6 <2e-16 \*\*\*  
## TOEFL 1 3921 3921 94.5 <2e-16 \*\*\*  
## LOR 1 4090 4090 98.6 <2e-16 \*\*\*  
## CGPA 1 4469 4469 107.7 <2e-16 \*\*\*  
## Residuals 395 16392 41   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(best\_model)

##   
## Call:  
## lm(formula = Chance ~ GRE + TOEFL + LOR + CGPA, data = Grad)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -27.969 -2.288 0.932 3.651 16.170   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -146.2511 10.5731 -13.83 < 2e-16 \*\*\*  
## GRE 0.2311 0.0576 4.01 7.2e-05 \*\*\*  
## TOEFL 0.2929 0.1076 2.72 0.0068 \*\*   
## LOR 2.3960 0.4839 4.95 1.1e-06 \*\*\*  
## CGPA 30.7403 2.9622 10.38 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.44 on 395 degrees of freedom  
## Multiple R-squared: 0.798, Adjusted R-squared: 0.796   
## F-statistic: 390 on 4 and 395 DF, p-value: <2e-16

The model is :

The is not much of a difference of between the anova tables outside of there being less variables. Also mean square error is lower in the model from section 14.

# 15. Build up confidence bands and prediction bands for all records.

test\_df = Grad[, c(2:7)]  
conf\_bands <- predict(best\_model, newdata = test\_df, interval = "confidence")  
write.csv(conf\_bands,"confidenceBands.csv")  
head(conf\_bands, 10)

## fit lwr upr  
## 1 95.64 94.30 96.98  
## 2 79.88 78.59 81.17  
## 3 64.00 62.59 65.41  
## 4 73.05 71.79 74.30  
## 5 64.51 63.57 65.44  
## 6 85.86 84.30 87.42  
## 7 70.28 68.63 71.93  
## 8 61.24 59.66 62.83  
## 9 55.39 53.72 57.05  
## 10 72.97 71.97 73.97

pred\_bands <- predict(best\_model, newdata = test\_df, interval = "prediction")  
write.csv(pred\_bands,"predictionBands.csv")  
head(pred\_bands, 10)

## fit lwr upr  
## 1 95.64 82.90 108.37  
## 2 79.88 67.15 92.61  
## 3 64.00 51.26 76.74  
## 4 73.05 60.32 85.78  
## 5 64.51 51.81 77.21  
## 6 85.86 73.10 98.62  
## 7 70.28 57.51 83.05  
## 8 61.24 48.48 74.00  
## 9 55.39 42.61 68.16  
## 10 72.97 60.26 85.67

# 16. Write the appropriate equation to predict the admission chance with variables included in final model from section 14.

The Intercept has no meaning in this case. With 0 for GRE, TOEFL, LOR, CGPA, one would have a negative chance of admission. This number is outside the valid range of chance thus no meaningful information can be gathered from it.

The meaning of the slope is that for every 1 point increase in GRE, admission chance increases by . For every 1 point increase in TOEFL, admission chance increases by . For every 1 point increase in LOR, admission chance increases by . For every 1 point in CGPA, admission chance increases by .

# 17. What conclusion can you arrive at from this exploration in terms of the suitability of descriptive statistics and regression in data exploration? What is the recommendation that you would provide future data explorations to include as a result necessarily?

Exploratory Data Analysis is a necessary step before developing models for a data set. Descriptive statistics of the data set and Visualizations made from them along with basic regression were able to paint a picture of the data. It allows for deduction of which type of model to train and which features will have the most weight. Future data explorations should scale the data to lie between 0 and 1. Along with data scaling, more models should be explored. There are more powerful linear regression models. Such models include Batch Gradient Descent, Stochastic Gradient Descent, Ridge Regression, Lasso Regression, ElasticNet regression. Polynomial and interaction feature transformation can be used as well to increase the number of features in the model potentially increasing accuracy. A train/test split of the data set should also be utilized in order to prevent over fitting of such models.