Regression Project for QMB 6304

Mandeep Singh

**Part 1: Preprocessing Part**

#Preprocessing  
#1.  
rm(list=ls())  
library(car)

## Loading required package: carData

library(rio)  
library(readxl)  
library(plyr)  
library(moments)  
library(corrplot)

## corrplot 0.84 loaded

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

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

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

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

Cabs=import("Project\_6304 Regression Project Data.csv")  
colnames(Cabs)=tolower(make.names(colnames(Cabs)))

*The required libraries are loaded and then data is imported.*

#2.  
set.seed(13394871)  
cab= Cabs[sample(1:nrow(Cabs),100,replace = FALSE),]

*Seed is set as my U number and a random sample of 100 rows is taken and is stored in the variable cab.*

#3.  
str(cab)

## 'data.frame': 100 obs. of 9 variables:  
## $ taxi\_id : int 2534 6612 4265 6299 8129 8098 7557 4836 5741 3885 ...  
## $ trip\_seconds: int 240 300 2820 1200 540 420 720 1680 300 180 ...  
## $ trip\_miles : num 0.7 1.6 1.1 33.5 0 1.2 1.5 16.2 0.9 0.3 ...  
## $ fare : num 5.25 7.25 48.75 41.75 10.25 ...  
## $ tips : num 2 0 10 9.15 0 2 2 0 0 0 ...  
## $ tolls : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ extras : num 2 1.5 4 4 0 0 0 0 0 0 ...  
## $ trip\_total : num 9.25 8.75 62.75 54.9 10.25 ...  
## $ payment\_type: chr "Credit Card" "Cash" "Credit Card" "Credit Card" ...

clean=subset(cab, trip\_seconds!=0 & trip\_miles != 0.00)  
cleanest= subset(clean, select = -c(tolls) )

*The abberancies which can be clearly seen in the data are the cases in which trip\_seconds and trip\_miles are 0 and other variables have some non zero value which is not feasible. Also, both these variables are independent variables which when 0(abberant case) can decrease the fit of the model. So, I have removed the cases in which either one or both of them are 0. Also, there is a column of tolls which is of no significance because it has the 0 value throughout the cases. After this cleaning process is done, we are left with 77 cases and 8 variables.*

**Part 2: Analysis Part**

#Analysis  
#1.  
#trip\_seconds  
summary(cleanest$trip\_seconds)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 60 360 480 734 900 3300

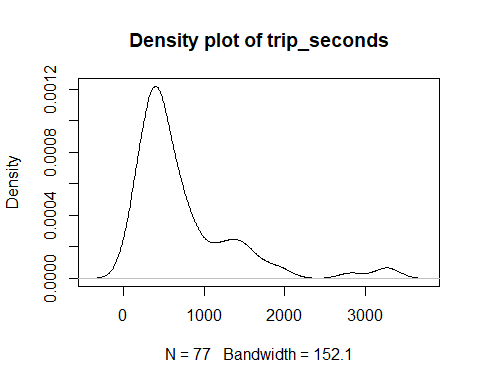
skewness(cleanest$trip\_seconds)

## [1] 2.121026

kurtosis(cleanest$trip\_seconds)

## [1] 7.89086

p1 = density(cleanest$trip\_seconds)  
plot(p1,pch=19,main="Density plot of trip\_seconds")



*Distribution of trip\_seconds is skewed right which can be seen in the density plot and through the skewness value. Kurtosis shows tail heavy data.*

#trip\_miles  
summary(cleanest$trip\_miles)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.100 0.900 1.500 3.728 3.600 33.500

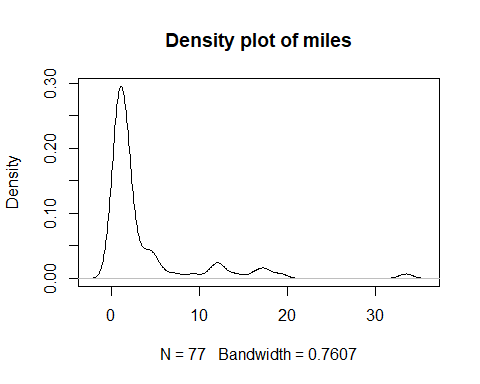
skewness(cleanest$trip\_miles)

## [1] 2.827447

kurtosis(cleanest$trip\_miles)

## [1] 12.21238

p2 = density(cleanest$trip\_miles)  
plot(p2,pch=19,main="Density plot of miles")



*Distribution of trip\_miles is skewed right which can be seen in the density plot and through the skewness value. Kurtosis shows tail heavy data.*

#fare  
summary(cleanest$fare)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.00 6.25 8.05 13.57 14.00 51.25

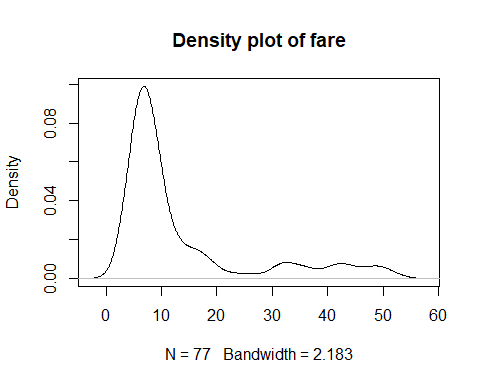
skewness(cleanest$fare)

## [1] 1.741947

kurtosis(cleanest$fare)

## [1] 4.739556

p3 = density(cleanest$fare)  
plot(p3,pch=19,main="Density plot of fare")



*Distribution of fare is skewed right which can be seen in the density plot and through the skewness value. Kurtosis shows tail heavy data.*

#tips  
summary(cleanest$tips)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 1.538 2.000 11.900

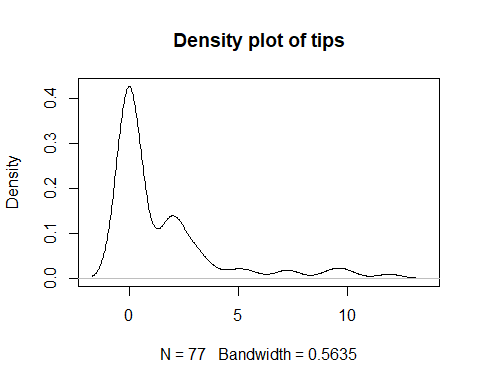
skewness(cleanest$tips)

## [1] 2.181401

kurtosis(cleanest$tips)

## [1] 7.391835

p4 = density(cleanest$tips)  
plot(p4,pch=19,main="Density plot of tips")



*Distribution of tips is skewed right which can be seen in the density plot and through the skewness value. Kurtosis shows tail heavy data.*

#extras  
summary(cleanest$extras)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 0.987 1.500 7.000

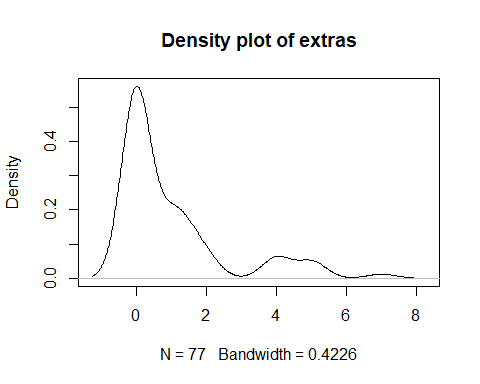
skewness(cleanest$extras)

## [1] 1.843986

kurtosis(cleanest$extras)

## [1] 5.649827

p5 = density(cleanest$extras)  
plot(p5,pch=19,main="Density plot of extras")



*Distribution of extras is skewed right which can be seen in the density plot and through the skewness value. Kurtosis shows tail heavy data.*

#trip\_total  
summary(cleanest$trip\_total)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.00 7.00 10.00 16.09 15.90 62.75

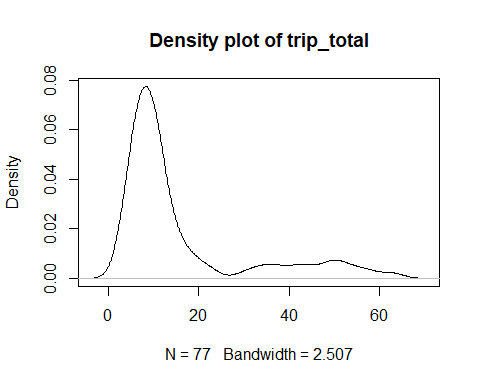
skewness(cleanest$trip\_total)

## [1] 1.698711

kurtosis(cleanest$trip\_total)

## [1] 4.593383

p6 = density(cleanest$trip\_total)  
plot(p6,pch=19,main="Density plot of trip\_total")



*Distribution of trip\_total is skewed right which can be seen in the density plot and through the skewness value. Kurtosis shows tail heavy data.*

#2.  
#Creating factor variable and then the table.  
cleanest$payment\_type=as.factor(cleanest$payment\_type)  
str(cleanest$payment\_type)

## Factor w/ 2 levels "Cash","Credit Card": 2 1 2 2 2 2 1 1 1 2 ...

summary(cleanest$payment\_type)

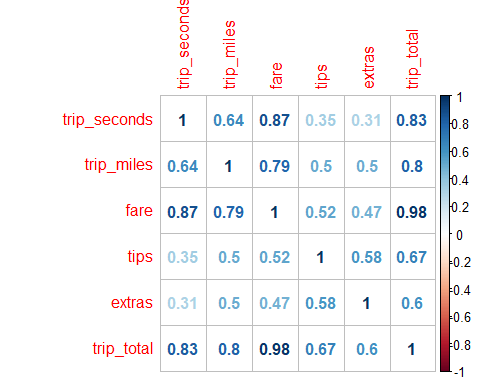
## Cash Credit Card   
## 44 33

levels(cleanest$payment\_type)

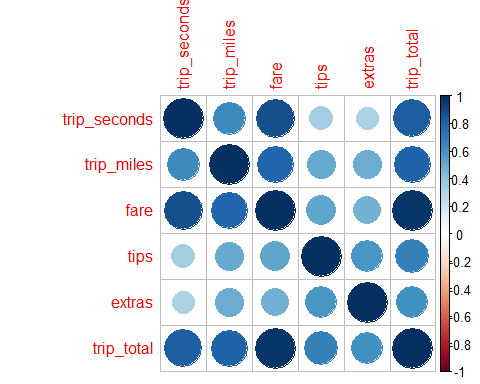
## [1] "Cash" "Credit Card"

*Using the payment\_type factor variable, table of ‘number of cases in each level’ is created.*

#3.  
corr.data = cleanest[,2:7]  
cm=cor(corr.data)  
#Number correlation plot matrix  
corrplot(cm,method = "number")



#Circle correlation plot matrix  
corrplot(cm,method = "circle")



*Here is the correlation matrix using all continuous variables except taxi\_id. Two correlation plots have been shown, one in number format and the other in circle format. A correlation plot represents correlation between different variables and correlation means the relationship between two variables. The correlation matrix of circle shows how much two variables are correlated by depicting the size of the circle and its color. A bigger circle and a darker blue color represent more correlation between the two variables. For example, trip\_total is highly correlated with fare. Similar is the case with number correlation matrix. A bigger number represent more correlation between the two variables. The number ranges from 0 to 1 with 0 representing no correlation and 1 representing highest correlation. Pictorial representation of correlation matrix is relatively easy and can be easily understood by non-statistics people as well.*

#4.  
#Applying regression model  
regression.output=lm(fare~trip\_seconds+trip\_miles+payment\_type,data=cleanest)  
summary(regression.output)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + payment\_type,   
## data = cleanest)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.5987 -1.4778 -0.4316 0.9293 28.7918   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.152740 0.937828 1.229 0.223   
## trip\_seconds 0.012009 0.001088 11.034 < 2e-16 \*\*\*  
## trip\_miles 0.844037 0.127062 6.643 4.75e-09 \*\*\*  
## payment\_typeCredit Card 1.055979 1.126798 0.937 0.352   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.806 on 73 degrees of freedom  
## Multiple R-squared: 0.8594, Adjusted R-squared: 0.8536   
## F-statistic: 148.7 on 3 and 73 DF, p-value: < 2.2e-16

#Getting Confidence interval  
confint(regression.output)

## 2.5 % 97.5 %  
## (Intercept) -0.716347895 3.02182730  
## trip\_seconds 0.009839634 0.01417785  
## trip\_miles 0.590803324 1.09727089  
## payment\_typeCredit Card -1.189725912 3.30168447

*This regression model uses fare as the dependent variable and trip\_seconds, trip\_miles and payment\_type are independent variable. trip\_seconds and trip\_miles have significant p values out of the independent variables. R-squared value is of 0.8594 and Adjusted R-squared value is of 0.8536 which signifies that 85.36% of variance is explained by these independent varibles.* *Impact of each significant independent variable: With an increment of 1 second in trip\_second, we can expect an increment of $0.012 or With an increment of 100 seconds in trip\_second, we can expect an increment of $1.2.*  *With an increment of 1 mile in trip\_mile, we can expect an increment of $0.84 or With an increment of 100 mile in trip\_mile, we can expect an increment of $84.* *Others are insignificant variables because of their high p value so their impact is not taken into consideration.* *The confidence interval of 95% can be seen above ranging from 2.5% to 97.5% for all the variables.*

#5.  
#model1  
regression.output1=lm(fare~trip\_seconds+poly(trip\_seconds,2)+trip\_miles+poly(trip\_miles,2)+ trip\_seconds:trip\_miles+payment\_type,data=cleanest)  
summary(regression.output1)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + poly(trip\_seconds, 2) + trip\_miles +   
## poly(trip\_miles, 2) + trip\_seconds:trip\_miles + payment\_type,   
## data = cleanest)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.0183 -1.4847 -0.2116 1.1397 26.9831   
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.1749377 1.2815183 0.137 0.89181   
## trip\_seconds 0.0138424 0.0016180 8.555 1.74e-12 \*\*\*  
## poly(trip\_seconds, 2)1 NA NA NA NA   
## poly(trip\_seconds, 2)2 -1.4886491 8.4112235 -0.177 0.86003   
## trip\_miles 1.2852564 0.4176238 3.078 0.00298 \*\*   
## poly(trip\_miles, 2)1 NA NA NA NA   
## poly(trip\_miles, 2)2 -1.9173953 5.9922383 -0.320 0.74994   
## payment\_typeCredit Card 0.2098885 1.1682411 0.180 0.85794   
## trip\_seconds:trip\_miles -0.0003243 0.0002463 -1.317 0.19213   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.724 on 70 degrees of freedom  
## Multiple R-squared: 0.8697, Adjusted R-squared: 0.8585   
## F-statistic: 77.85 on 6 and 70 DF, p-value: < 2.2e-16

#model2  
regression.output2=lm(fare~trip\_seconds+poly(trip\_seconds,2)+trip\_miles+payment\_type,data=cleanest)  
summary(regression.output2)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + poly(trip\_seconds, 2) + trip\_miles +   
## payment\_type, data = cleanest)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.7969 -1.6282 -0.1298 1.2569 26.5794   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.328881 0.924905 1.437 0.1551   
## trip\_seconds 0.012405 0.001087 11.407 < 2e-16 \*\*\*  
## poly(trip\_seconds, 2)1 NA NA NA NA   
## poly(trip\_seconds, 2)2 -9.876157 5.079031 -1.944 0.0557 .   
## trip\_miles 0.777168 0.129363 6.008 7.02e-08 \*\*\*  
## payment\_typeCredit Card 0.547914 1.136375 0.482 0.6312   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.717 on 72 degrees of freedom  
## Multiple R-squared: 0.8664, Adjusted R-squared: 0.859   
## F-statistic: 116.7 on 4 and 72 DF, p-value: < 2.2e-16

#model3  
regression.output3=lm(fare~trip\_seconds+trip\_miles+poly(trip\_miles,2)+payment\_type,data=cleanest)  
summary(regression.output3)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + poly(trip\_miles,   
## 2) + payment\_type, data = cleanest)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.5819 -1.4674 -0.4144 0.8902 29.0028   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.253886 0.993334 1.262 0.211   
## trip\_seconds 0.011816 0.001244 9.496 2.51e-14 \*\*\*  
## trip\_miles 0.858651 0.135480 6.338 1.79e-08 \*\*\*  
## poly(trip\_miles, 2)1 NA NA NA NA   
## poly(trip\_miles, 2)2 -1.793360 5.501856 -0.326 0.745   
## payment\_typeCredit Card 1.022684 1.138352 0.898 0.372   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.835 on 72 degrees of freedom  
## Multiple R-squared: 0.8596, Adjusted R-squared: 0.8518   
## F-statistic: 110.2 on 4 and 72 DF, p-value: < 2.2e-16

#model4  
regression.output4=lm(fare~trip\_seconds+trip\_miles+trip\_seconds:trip\_miles+payment\_type,  
 data=cleanest)  
summary(regression.output4)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + trip\_seconds:trip\_miles +   
## payment\_type, data = cleanest)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.9592 -1.4639 -0.2541 1.1775 26.9298   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.0651979 1.0496000 -0.062 0.9506   
## trip\_seconds 0.0141762 0.0014071 10.075 2.15e-15 \*\*\*  
## trip\_miles 1.3210052 0.2387973 5.532 4.83e-07 \*\*\*  
## payment\_typeCredit Card 0.2525071 1.1469435 0.220 0.8264   
## trip\_seconds:trip\_miles -0.0003551 0.0001522 -2.333 0.0225 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.666 on 72 degrees of freedom  
## Multiple R-squared: 0.8693, Adjusted R-squared: 0.862   
## F-statistic: 119.7 on 4 and 72 DF, p-value: < 2.2e-16

#model5  
regression.output5=lm(fare~trip\_seconds+trip\_miles+trip\_seconds:trip\_miles  
 +poly(trip\_seconds,2):poly(trip\_seconds,2)  
 +payment\_type,data=cleanest)  
summary(regression.output5)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + trip\_seconds:trip\_miles +   
## poly(trip\_seconds, 2):poly(trip\_seconds, 2) + payment\_type,   
## data = cleanest)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.1200 -1.4919 -0.2578 1.2932 26.6450   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.1877500 1.2727701 0.148 0.88315   
## trip\_seconds 0.0139181 0.0015905 8.751 6.78e-13 \*\*\*  
## trip\_miles 1.2225173 0.3663965 3.337 0.00135 \*\*   
## poly(trip\_seconds, 2)1 NA NA NA NA   
## poly(trip\_seconds, 2)2 -2.6724055 7.5060760 -0.356 0.72287   
## payment\_typeCredit Card 0.2504555 1.1539770 0.217 0.82880   
## trip\_seconds:trip\_miles -0.0002952 0.0002274 -1.298 0.19839   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.695 on 71 degrees of freedom  
## Multiple R-squared: 0.8695, Adjusted R-squared: 0.8603   
## F-statistic: 94.6 on 5 and 71 DF, p-value: < 2.2e-16

#model6  
regression.output6=lm(fare~trip\_seconds+poly(trip\_seconds,2)  
 +trip\_miles+poly(trip\_miles,2)  
 +trip\_seconds:trip\_miles+poly(trip\_seconds,2):poly(trip\_seconds,2)  
 +payment\_type,data=cleanest)  
summary(regression.output6)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + poly(trip\_seconds, 2) + trip\_miles +   
## poly(trip\_miles, 2) + trip\_seconds:trip\_miles + poly(trip\_seconds,   
## 2):poly(trip\_seconds, 2) + payment\_type, data = cleanest)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.0183 -1.4847 -0.2116 1.1397 26.9831   
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.1749377 1.2815183 0.137 0.89181   
## trip\_seconds 0.0138424 0.0016180 8.555 1.74e-12 \*\*\*  
## poly(trip\_seconds, 2)1 NA NA NA NA   
## poly(trip\_seconds, 2)2 -1.4886491 8.4112235 -0.177 0.86003   
## trip\_miles 1.2852564 0.4176238 3.078 0.00298 \*\*   
## poly(trip\_miles, 2)1 NA NA NA NA   
## poly(trip\_miles, 2)2 -1.9173953 5.9922383 -0.320 0.74994   
## payment\_typeCredit Card 0.2098885 1.1682411 0.180 0.85794   
## trip\_seconds:trip\_miles -0.0003243 0.0002463 -1.317 0.19213   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.724 on 70 degrees of freedom  
## Multiple R-squared: 0.8697, Adjusted R-squared: 0.8585   
## F-statistic: 77.85 on 6 and 70 DF, p-value: < 2.2e-16

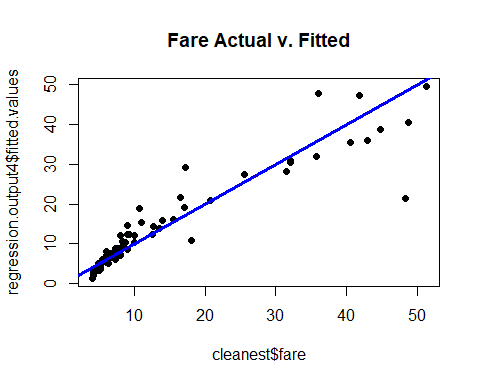
*To achieve a better fit for the model, I have tried various variable transforms and relevant interactions. I have added squares of trip\_seconds and trip\_miles and interaction of both the independent variables. After doing the possible combinations, it can be concluded that model 4 has the best fit because of its most number variables with significant p values(3) and highest adjusted r squared value.*

#6.  
#Applying the best regression model  
regression.output4=lm(fare~trip\_seconds+trip\_miles+trip\_seconds:trip\_miles+payment\_type,  
 data=cleanest)  
summary(regression.output4)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + trip\_seconds:trip\_miles +   
## payment\_type, data = cleanest)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.9592 -1.4639 -0.2541 1.1775 26.9298   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.0651979 1.0496000 -0.062 0.9506   
## trip\_seconds 0.0141762 0.0014071 10.075 2.15e-15 \*\*\*  
## trip\_miles 1.3210052 0.2387973 5.532 4.83e-07 \*\*\*  
## payment\_typeCredit Card 0.2525071 1.1469435 0.220 0.8264   
## trip\_seconds:trip\_miles -0.0003551 0.0001522 -2.333 0.0225 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.666 on 72 degrees of freedom  
## Multiple R-squared: 0.8693, Adjusted R-squared: 0.862   
## F-statistic: 119.7 on 4 and 72 DF, p-value: < 2.2e-16

*The above regression model provides us with the best fit because of its best adjusted r-squared values among others and most number of significant p values.* *It takes in independent variables as trip\_seconds, trip\_miles, interaction between these two i.e. trip\_seconds:trip\_miles and payment\_type. It has an adjusted R-squared value of 0.862.*

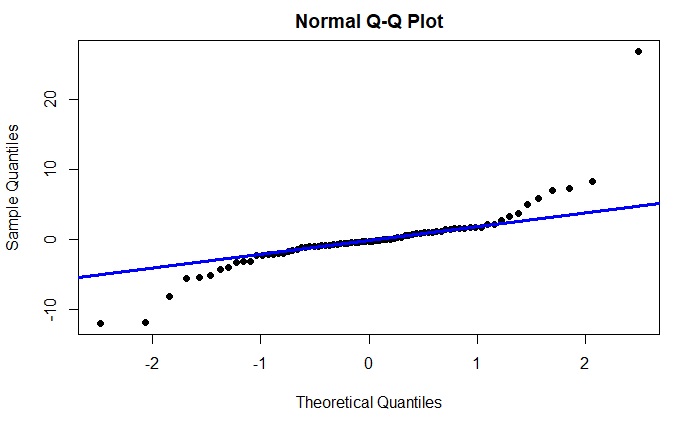
#Linearity assumption check  
plot(cleanest$fare,regression.output4$fitted.values,  
 pch=19,main="Fare Actual v. Fitted")+abline(0,1,lwd=3,col="blue")



## integer(0)

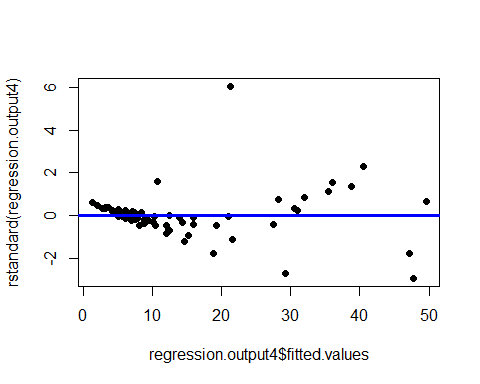
*It can be inferred from the above graph that linearity assumption does not hold true because considerable number of points lies away from the line. Weak linearity can be observed.*

#Normality assumption check  
qqnorm(regression.output4$residuals,pch=19)



*It can be inferred from the above graph that normality assumption holds true because considerable number of points lies on the line. There are some points on both the ends but those are not in large number.*

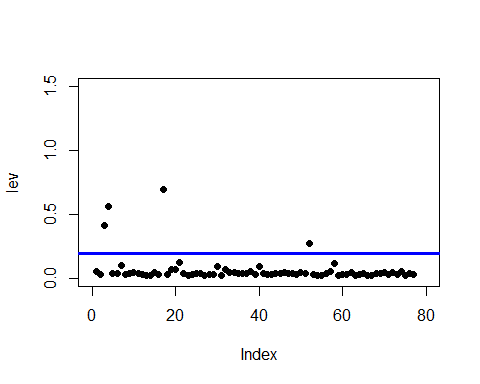
#Equality of Variances assumption check  
plot(regression.output4$fitted.values,rstandard(regression.output4),pch=19)+abline(0,0,col="blue",lwd=3)



## integer(0)

*It can be inferred that there is a pattern which leads to the conclusion that Equality of Variances assumption is not held true in this case. There are some points which are not aligning with the pattern but frequency of these points is very less.*

#7.  
#Identifying leverage points  
lev=hat(model.matrix(regression.output4))  
plot(lev,pch=19,,xlim=c(0,80),ylim=c(0,1.5)) + abline(3\*mean(lev),0,col="blue",lwd=3)



## integer(0)

#Pointing out leverage points  
cleanest[lev>(3\*mean(lev)),]

## taxi\_id trip\_seconds trip\_miles fare tips extras trip\_total  
## 81620 4265 2820 1.1 48.75 10.00 4 62.75  
## 770215 6299 1200 33.5 41.75 9.15 4 54.90  
## 383092 4986 3300 19.2 51.25 0.00 0 51.25  
## 1396789 3463 3240 11.3 36.00 0.00 0 36.00  
## payment\_type  
## 81620 Credit Card  
## 770215 Credit Card  
## 383092 Cash  
## 1396789 Cash

#Removing leverage points  
sans\_lev=filter(cleanest,taxi\_id!=4265)  
sans\_lev=filter(sans\_lev,taxi\_id!=6299)  
sans\_lev=filter(sans\_lev,taxi\_id!=4986)  
sans\_lev=filter(sans\_lev,taxi\_id!=3463)  
  
#Applying the same regression model on data without leverage points  
regression.output.sans\_lev=lm(fare~trip\_seconds+trip\_miles+trip\_seconds:trip\_miles+payment\_type,  
 data=sans\_lev)  
summary(regression.output.sans\_lev)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + trip\_seconds:trip\_miles +   
## payment\_type, data = sans\_lev)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.8252 -0.8877 -0.2491 0.6507 28.9950   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.4403982 1.1649465 1.236 0.221   
## trip\_seconds 0.0114216 0.0019023 6.004 8.36e-08 \*\*\*  
## trip\_miles 0.8459481 0.5945595 1.423 0.159   
## payment\_typeCredit Card -0.0532403 1.0666829 -0.050 0.960   
## trip\_seconds:trip\_miles 0.0002039 0.0003542 0.576 0.567   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.202 on 68 degrees of freedom  
## Multiple R-squared: 0.8465, Adjusted R-squared: 0.8375   
## F-statistic: 93.78 on 4 and 68 DF, p-value: < 2.2e-16

*After identifying and removing the inappropriately high leverage points, the same regression model was ran on the new data(without leverage points) and the quality of fit in this final regression model was worse than before. The number of variables with significant p-values is just 1 now i.e. trip\_seconds. Also the adjusted R-squared value is now reduced to 0.8375 from 0.862. The quality of fit in this regression model has gone down.*

#8.  
#Setting a different seed and taking a random sample.  
set.seed(13394876)  
cab\_8= Cabs[sample(1:nrow(Cabs),100,replace = FALSE),]  
#Applying the same cleaning procedure  
clean8=subset(cab\_8, trip\_seconds!=0 & trip\_miles != 0.00)  
cleanest8= subset(clean8, select = -c(tolls) )  
#Applying the same regression model on this new data  
regression.outputfinal=lm(fare~trip\_seconds+trip\_miles+trip\_seconds:trip\_miles+payment\_type,  
 data=cleanest8)  
summary(regression.outputfinal)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + trip\_seconds:trip\_miles +   
## payment\_type, data = cleanest8)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.8520 -1.9963 -0.5190 0.9595 24.2540   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -9.719e-01 1.287e+00 -0.755 0.45273   
## trip\_seconds 1.353e-02 1.476e-03 9.169 1.67e-13 \*\*\*  
## trip\_miles 1.359e+00 2.335e-01 5.822 1.74e-07 \*\*\*  
## payment\_typeCredit Card 2.154e+00 1.298e+00 1.660 0.10151   
## trip\_seconds:trip\_miles -2.738e-04 6.937e-05 -3.947 0.00019 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.295 on 68 degrees of freedom  
## Multiple R-squared: 0.845, Adjusted R-squared: 0.8359   
## F-statistic: 92.69 on 4 and 68 DF, p-value: < 2.2e-16

*After applying the regression model on the new data, it can be observed that the new model fits worse than the previous model. The adjusted r square got down from 0.862 to 0.8359. In this model, there are 3 variables with significant p values i.e. trip\_seconds, trip\_miles and trip\_seconds:trip\_miles.*