Regression Project QMB-6304 Analytical Methods for Business Solution

SAI SURAJ ARGULA

10/9/2019

Write a simple R script to execute the following data preprocessing and statistical analysis. Show your R code, analytical output, and interpretations.

Preprocessing

1. Load the file “6304 Regression Project Data.csv” into R. This file contains information on 1,705,805 taxi cab trips in the City of Chicago during 2016. The data was taken from kaggle.com and modified. Variables in this data set are:
2. taxi\_id: a unique identifier for each individual taxi cab.
3. trip\_seconds: the number of seconds elapsed during the trip.
4. trip\_miles: the number of miles logged during the trip.
5. fare: the base fare charged to the customer for the trip.
6. tips: the tip given by the customer to the driver for the trip.
7. tolls: any surcharges for road or bridge tolls incurred during the trip.
8. extras: charges for any incidentals requested by the customer.
9. trip\_total: the total charge to the customer for the trip.
10. payment\_type: the method of payment used by the customer. This includes cash, credit card, and several other methods of payment jointly classed as “other”.

preprocessing

step1-loading data

#removing data sets from the current environment, loading libraries, setting current Worfking directory, import dataset, change col names to lower, check for number of coloumns and rows  
rm(list=ls())   
library(rio)   
library(moments)   
getwd( )

## [1] "C:/Users/MY/Desktop/suraj/QMB6304.002F19 Analytical Methods for Bus/Project"

setwd("C:\\Users\\MY\\Desktop\\suraj\\QMB6304.002F19 Analytical Methods for Bus\\Project")   
projectdata=import("6304 Regression Project Data.csv")  
colnames(projectdata)=tolower(make.names(colnames(projectdata)))  
nrow(projectdata)

## [1] 1705421

ncol(projectdata)

## [1] 9

There are 1705421 rows and 9 coloumns in the dataset.

1. Using the numerical portion of your U number as a random number seed and the random selection method presented in class, take a random sample of 100 taxi trips from this master data set.

#step2-Taking sample of 100 taxi trips from the master dataset with random seed of Numerical portion of UNumber into dataframe sample.  
set.seed(02248584)   
sample=projectdata[sample(1:nrow(projectdata),100,replace=FALSE),]   
#View(sample)  
summary(sample) #summary of sample dataset

## taxi\_id trip\_seconds trip\_miles fare   
## Min. : 6 Min. : 0.0 Min. : 0.000 Min. : 3.250   
## 1st Qu.:2290 1st Qu.: 240.0 1st Qu.: 0.000 1st Qu.: 5.938   
## Median :5040 Median : 540.0 Median : 1.250 Median : 9.250   
## Mean :4653 Mean : 727.2 Mean : 3.669 Mean :15.312   
## 3rd Qu.:6814 3rd Qu.: 975.0 3rd Qu.: 3.850 3rd Qu.:16.500   
## Max. :8597 Max. :3780.0 Max. :36.600 Max. :85.000   
## tips tolls extras trip\_total   
## Min. : 0.000 Min. :0 Min. : 0.00 Min. : 3.250   
## 1st Qu.: 0.000 1st Qu.:0 1st Qu.: 0.00 1st Qu.: 7.188   
## Median : 0.000 Median :0 Median : 0.00 Median : 10.250   
## Mean : 1.553 Mean :0 Mean : 1.48 Mean : 18.344   
## 3rd Qu.: 2.000 3rd Qu.:0 3rd Qu.: 1.00 3rd Qu.: 19.297   
## Max. :11.500 Max. :0 Max. :48.00 Max. :133.000   
## payment\_type   
## Length:100   
## Class :character   
## Mode :character   
##   
##   
##

attach(sample) #attaching sample for further analysis

1. Using your judgment and the R tools you know, cleanse your random sample data of aberrant cases. Such cleansing cases is somewhat subjective, so explain your process and reasoning for identifying aberrancies in the data and removing them. State how many cases you are left with in your random sample after cleansing. This will be your primary data set for analysis.

Solution

#Cleaning sample data

Assumption 1:- There are trips with 0 trip seconds and 0 trip miles in the sample dataset.This might be because the customer have booked the cab and not boarded instead cancelled the cab. But there are some values on trip and tips maybe because of the cancelletion fee or tips given to driver for making him wait. This might not be a good data to consider for calculating regression model on fare can largely influence fit of model, removing these data from the datset.

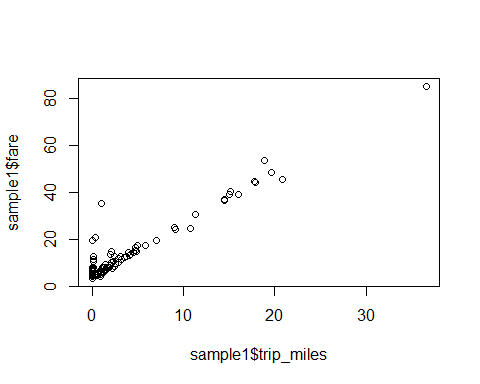
Taking new reduced sample as sample1

#removing rows from sample trip seconds=0 and trip miles= 0 and assigning to new dataframe sample1  
sample1=sample[!(sample$trip\_seconds==0 & sample$trip\_miles==0),]

New Sample conatins 88 sample observations left

Assumption 2:- There are trips with 0 trip miles but has some trip seconds trip miles in the sample dataset.This might be because the customer have booked and boarded cab but cancelled the cab before the start which might resulted in cancellation fee or tips given to driver. Since there are 15 points with this data,this might not be a good data to consider for calculating regression model on fare can largely influence fit of model, removing these data from the datset. Also below the graph between miles and fare shows the points with 0 miles are not corresonponding to the trend of other points and might influence the model.

plot(sample1$trip\_miles,sample1$fare)



#plot(sample1$trip\_seconds,sample1$fare)  
  
sample2=sample1[!(sample1$trip\_miles==0),] #removing rows from sample1 having trip seconds and miles are 0 and assigning to new dataframe sample2

New Sample2 conatins 73 observations left

Main Analysis

1. Using your cleansed sample data, provide summaries and density plots of each of the continuous variables in your data set with the exception of taxi\_id. Explain any apparent differences in the statistical distributions of these variables in your sample data.

continous variable in the data are- trip seconds, trip miles,trip fare, total fare, extra fare and tips Variable:- trip\_seconds

#tripseconds  
str(sample2$trip\_seconds)

## int [1:73] 2340 300 300 1500 300 480 2280 540 2280 600 ...

skewness(sample2$trip\_seconds)

## [1] 1.555917

kurtosis(sample2$trip\_seconds)

## [1] 5.64362

summary(sample2$trip\_seconds)

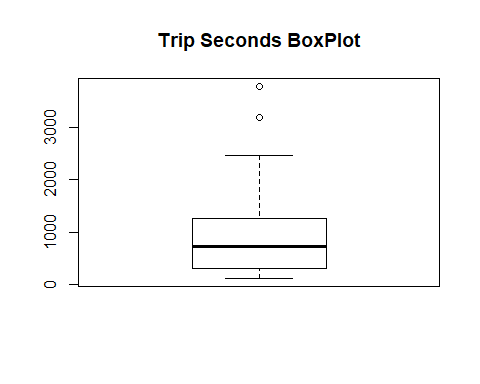
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 120.0 300.0 720.0 924.7 1260.0 3780.0

sd(sample2$trip\_seconds)

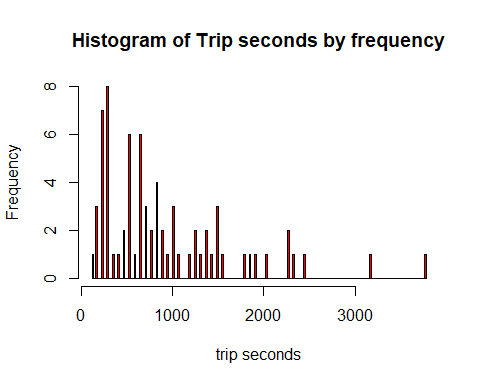
## [1] 736.2632

Skewness is greater than 1 suggesting data distribution(trip seconds) is highly skewed and a positive skewness indicates that the size of the right-handed tail is larger than the left-handed tail. Kurtosis is greater than 0. Suggesting data distribution(trip seconds) has long tails i.e, leptokurtic distribution mean of data is 924.7 seconds sd of data is 736.26 seconds

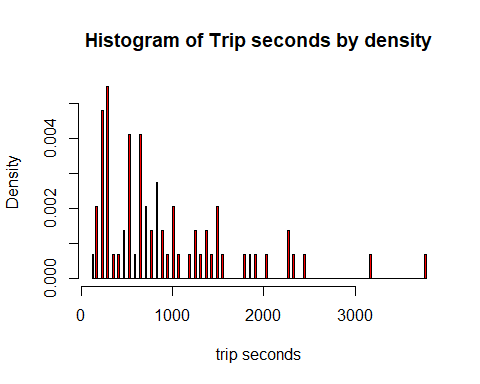
boxplot(sample2$trip\_seconds, main="Trip Seconds BoxPlot")#



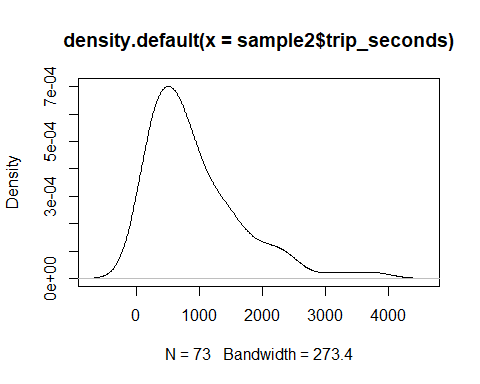
hist(sample2$trip\_seconds,breaks=200,col="red",main="Histogram of Trip seconds by frequency",xlab = "trip seconds") #frequency



hist(sample2$trip\_seconds,freq=FALSE,breaks=200,col="red",main="Histogram of Trip seconds by density",xlab = "trip seconds")#density histogram



plot(density(sample2$trip\_seconds))



From the above histograms and density graphs, we could see most of the trips second in the sample has trip seconds between 0 to 2000 seconds a very few above 2000 seconds. Data is skewed towards left and has long tails towards right. Further from the box plot of the trip\_seconds we can see there are few points which are outliers.

Variable:- trip\_miles

#trip miles  
str(sample2$trip\_miles)

## num [1:73] 17.8 2.9 1 2 1 1.53 36.6 1.5 15.2 1.4 ...

skewness(sample2$trip\_miles)

## [1] 2.242384

kurtosis(sample2$trip\_miles)

## [1] 8.728509

summary(sample2$trip\_miles)

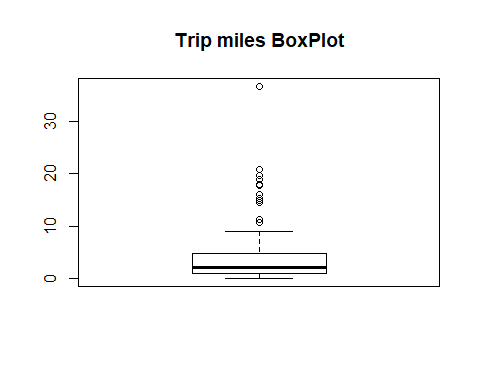
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.100 1.000 2.100 5.027 4.900 36.600

sd(sample2$trip\_miles)

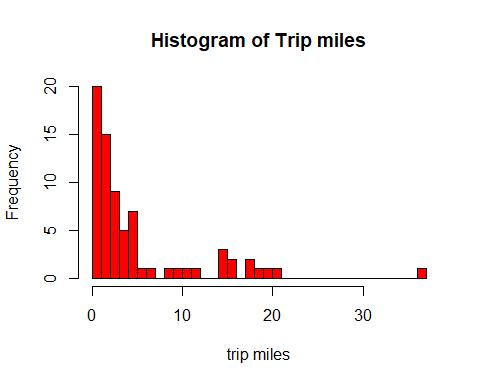
## [1] 6.710693

Skewness is greater than 1 suggesting data distribution(trip miles) is highly skewed and a positive skewness indicates that the size of the right-handed tail is larger than the left-handed tail. Kurtosis is greater than 0. Suggesting data distribution(trip miles) has long tails i.e, leptokurtic distribution mean of data is 2.10 sd of data is 6.71

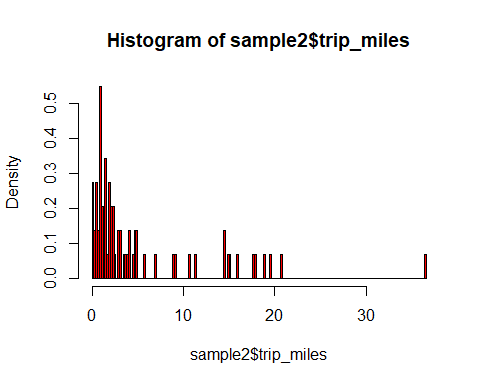
boxplot(sample2$trip\_miles, main="Trip miles BoxPlot")



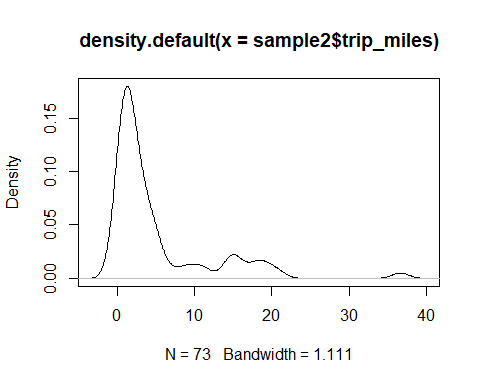
hist(sample2$trip\_miles,breaks=30,col="red",main="Histogram of Trip miles",xlab = "trip miles") #frequency



hist(sample2$trip\_miles,freq=FALSE,breaks=200,col="red")#density histogram



plot(density(sample2$trip\_miles))



From the above histograms and density graphs, we could see most of the trips miles in the sample has trip miles between 0 to 10 miles a very few above 10 miles. Data is skewed towards left and has long tails towards right. Further from the box plot of the trip\_miles we can see there are few points which are outliers.

#trip fare  
str(sample2$fare)

## num [1:73] 44.5 10 6 13.5 5.65 ...

skewness(sample2$fare)

## [1] 2.079706

kurtosis(sample2$fare)

## [1] 7.968656

summary(sample2$fare)

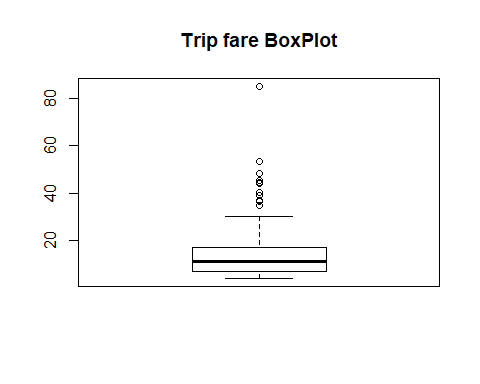
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.25 7.25 11.25 16.50 17.25 85.00

sd(sample2$fare)

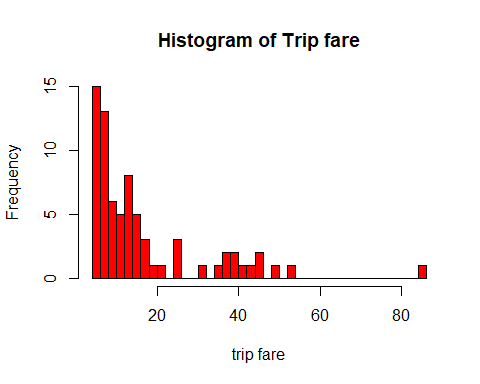
## [1] 14.98017

Skewness is greater than 1 suggesting data distribution(trip fare) is highly skewed. Kurtosis is greater than 0. Suggesting data distribution(trip fare) has long tails i.e, leptokurtic distribution mean of data is 16.50 sd of data is 14.98

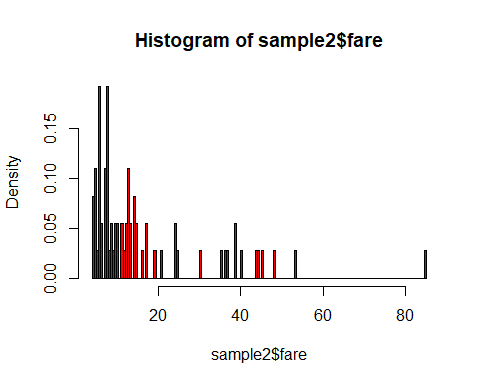
boxplot(sample2$fare, main="Trip fare BoxPlot")



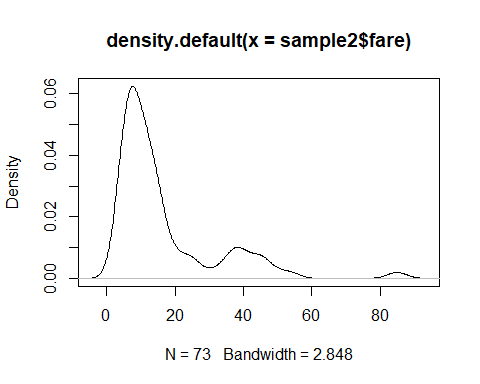
hist(sample2$fare,breaks=30,col="red",main="Histogram of Trip fare",xlab = "trip fare") #frequency



hist(sample2$fare,freq=FALSE,breaks=200,col="red")#density histogram



plot(density(sample2$fare))



From the above histograms and density graphs, we could see most of the trips in the sample has trip fare between 0 to 20 dollars a very few above 20 dollars. Data is skewed towards left and has long tails towards right. Further from the box plot of the trip\_seconds we can see there are few points which are outliers.

#trip tips  
str(sample2$tips)

## num [1:73] 9.7 2 0 3 0 0 0 2 8.45 2 ...

skewness(sample2$tips)

## [1] 1.912684

kurtosis(sample2$tips)

## [1] 5.91274

summary(sample2$tips)

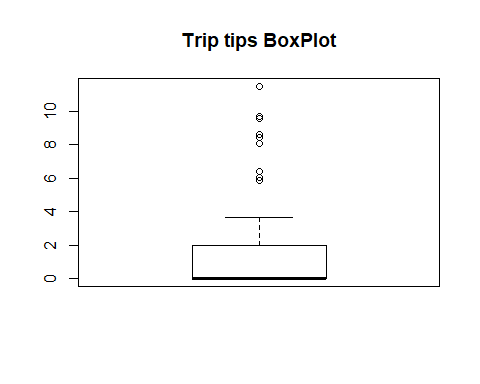
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 1.73 2.00 11.50

sd(sample2$tips)

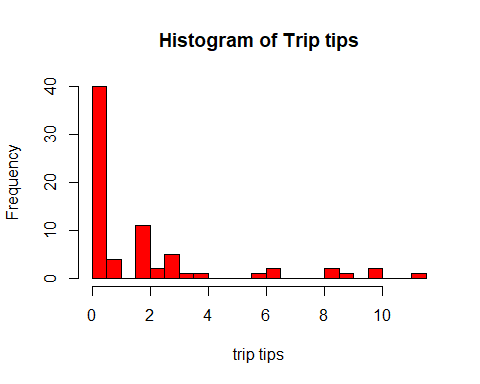
## [1] 2.761035

Skewness is greater than 1 suggesting data distribution(trip tips) is highly skewed and a positive skewness indicates that the size of the right-handed tail is larger than the left-handed tail. Kurtosis is greater than 0. Suggesting data distribution(trip tips) has long tails i.e, leptokurtic distribution mean of data is 1.73 sd of data is 2.76 median of the data is zero. suggesting more than half of the data points are 0’s.

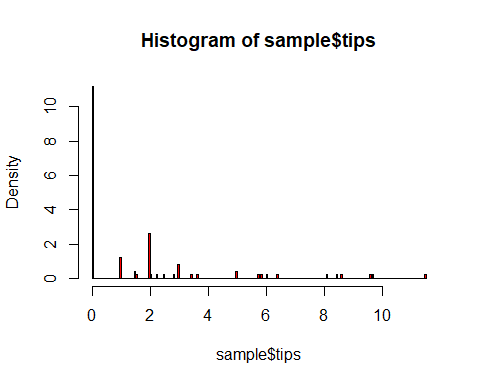
boxplot(sample2$tips, main="Trip tips BoxPlot")



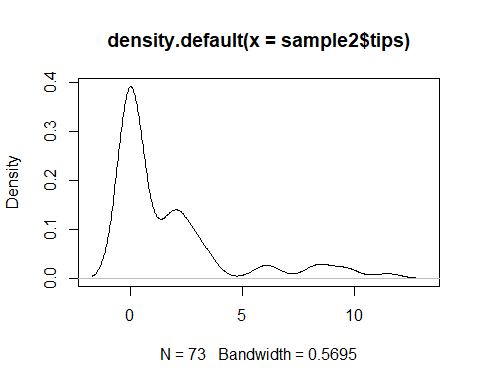
hist(sample2$tips,breaks=30,col="red",main="Histogram of Trip tips",xlab = "trip tips") #frequency



hist(sample$tips,freq=FALSE,breaks=200,col="red")#density histogram



plot(density(sample2$tips))



From the above histograms and density graphs, we could see most of the trips in the sample has trip fare between 0 to 2 dollars a few above 2 dollars. Data is skewed towards left and has long tails towards right. Further from the box plot of the trip\_seconds we can see there are few points which are outliers.

Trip toll- There are no tolls in the sample data

Extra’s-

#extras  
str(sample2$extras)

## num [1:73] 4 0 1 0 1 5 48 1 2 1 ...

skewness(sample2$extras)

## [1] 6.259183

kurtosis(sample2$extras)

## [1] 44.25381

summary(sample2$extras)

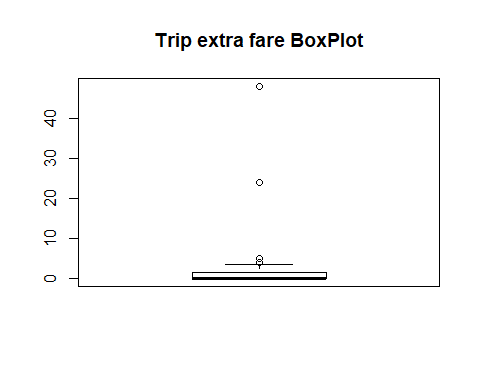
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 1.89 1.50 48.00

sd(sample2$extras)

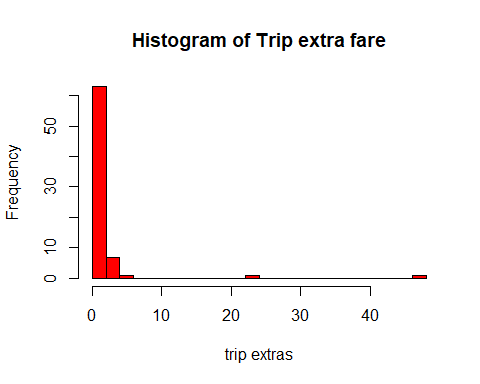
## [1] 6.23846

Skewness is greater than 1 suggesting data distribution(trip extra fare) is highly skewed and a positive skewness indicates that the size of the right-handed tail is larger than the left-handed tail. Kurtosis is greater than 0. Suggesting data distribution(trip extra fare) has long tails i.e, leptokurtic distribution mean of data is 1.89 dollars sd of data is 6.23 dollars median of the data is zero. suggesting more than half of the data points are 0’s.

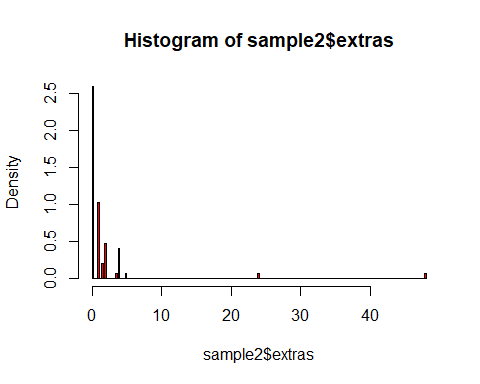
boxplot(sample2$extra, main="Trip extra fare BoxPlot")



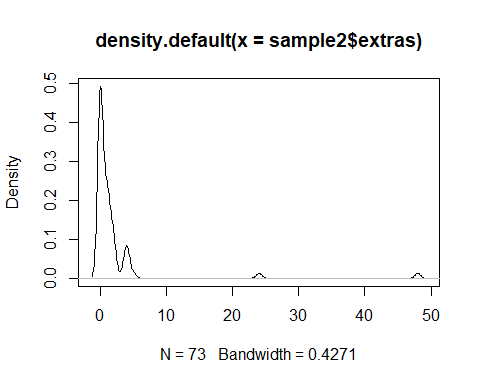
hist(sample2$extras,breaks=30,col="red",main="Histogram of Trip extra fare",xlab = "trip extras") #frequency



hist(sample2$extras,freq=FALSE,breaks=200,col="red")#density histogram



plot(density(sample2$extras))



From the above histograms and density graphs, we could see most of the trips in the sample has trip extra fare between 0 and upto 5 dollars a very few above5 dollars. Data is skewed towards left and has long tails towards right. Further from the box plot of the trip\_seconds we can see there are few points which are outliers.

#triptotal  
str(sample2$trip\_total)

## num [1:73] 58.2 12 7 16.5 6.65 ...

skewness(sample2$trip\_total)

## [1] 2.81657

kurtosis(sample2$trip\_total)

## [1] 13.22872

summary(sample$trip\_total)

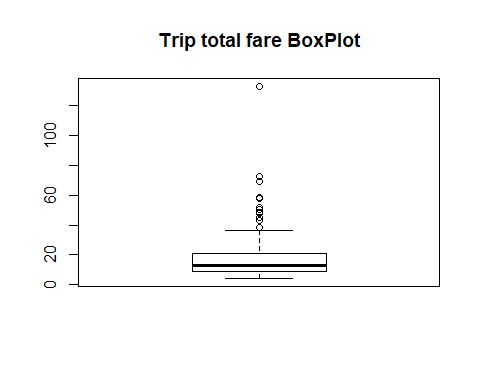
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.250 7.188 10.250 18.344 19.297 133.000

sd(sample2$trip\_total)

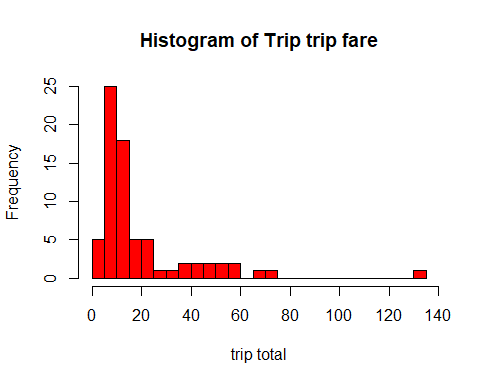
## [1] 21.0481

Skewness is greater than 1 suggesting data distribution(trip total) is highly skewed and a positive skewness indicates that the size of the right-handed tail is larger than the left-handed tail. Kurtosis is greater than 0. Suggesting data distribution(trip total) has long tails i.e, leptokurtic distribution mean of data is 18.34 dollars sd of data is 21.04 dollars

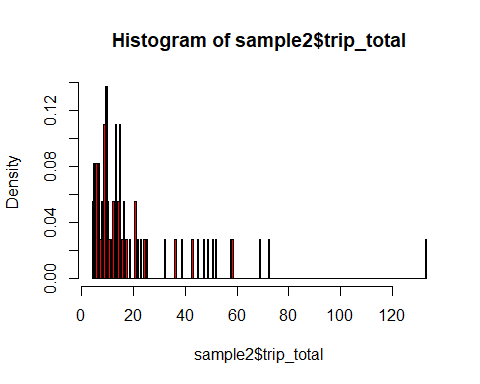
boxplot(sample2$trip\_total, main="Trip total fare BoxPlot")



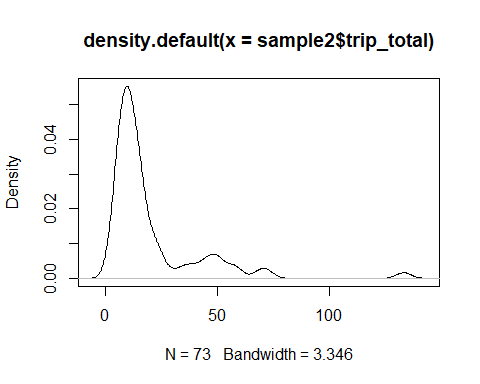
hist(sample2$trip\_total,breaks=30,col="red",main="Histogram of Trip trip fare",xlab = "trip total") #frequency



hist(sample2$trip\_total,freq=FALSE,breaks=200,col="red")#density histogram



plot(density(sample2$trip\_total))



From the above histograms and density graphs, we could see most of the trips total fare in the sample has trip seconds between 0 to 20 dollars a very few above 20 dollars. Data is skewed towards left and has long tails towards right. Further from the box plot of the trip\_seconds we can see there are few points which are outliers.

1. Using the payment\_type factor variable and your cleansed sample data, provide a table of the number of cases in each level of payment\_type.

#payment type  
class(sample2$payment\_type) # checking the class type of the variable Payment type

## [1] "character"

sample3=sample2 #moving into other sample  
sample3$payment\_type=as.factor(sample3$payment\_type) #converting payment type into factors  
#sample3$payment\_type  
class(sample3$payment\_type) #now checing class of variable payment type

## [1] "factor"

levels(sample3$payment\_type) # levels of factors in payment type

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

str(sample3$payment\_type) #structure of payment type

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

library(plyr) #loading library Plyr for the use of count  
count(sample3$payment\_type) #count table of number of cases in each level

## x freq  
## 1 Cash 38  
## 2 Credit Card 35

There are two types of payments credit card and cash with 38 and 35 cases respectively.

1. Construct an easily read and easily understood correlation matrix using all continuous variables except taxi\_id. Give a brief interpretation of the matrix understandable by a non-statistician.

#analysis-3  
#Copy the continuous variables to a new data object.  
sample4=subset(sample3,select=c("trip\_seconds","trip\_miles","fare","tips","tolls","extras","trip\_total"))  
#removing tolls as it doesn't have any values  
sample5=subset(sample4,select=c("trip\_seconds","trip\_miles","fare","tips","extras","trip\_total"))  
#Correlation analysis of the continuous variables.  
library(corrplot)

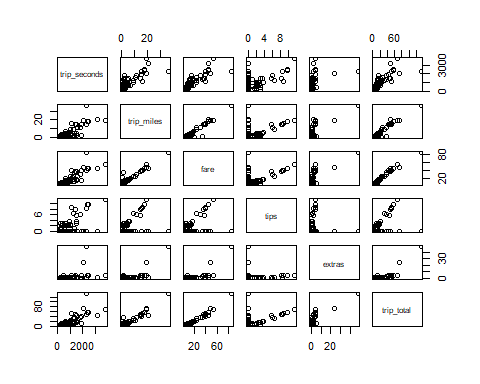
## corrplot 0.84 loaded

cor(sample5)

## trip\_seconds trip\_miles fare tips extras  
## trip\_seconds 1.0000000 0.7869539 0.8494909 0.544838822 0.370712968  
## trip\_miles 0.7869539 1.0000000 0.9583781 0.493658943 0.701551813  
## fare 0.8494909 0.9583781 1.0000000 0.498510854 0.699167358  
## tips 0.5448388 0.4936589 0.4985109 1.000000000 0.004620328  
## extras 0.3707130 0.7015518 0.6991674 0.004620328 1.000000000  
## trip\_total 0.7859384 0.9547785 0.9843310 0.487342480 0.794601777  
## trip\_total  
## trip\_seconds 0.7859384  
## trip\_miles 0.9547785  
## fare 0.9843310  
## tips 0.4873425  
## extras 0.7946018  
## trip\_total 1.0000000

Above is the correlation matrix of continous variables in the sample. \* We could see a high correlation between trip fare and trip miles and also trip fare and trip seconds. \* High correlation between trip miles,trip seconds and trip total.

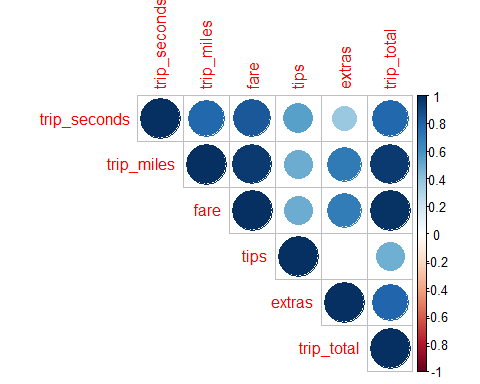
plot(sample5)



Above is the correlation Plots of continous variables in the sample. We could see the some relationships between some of the variables, for example trip miles vs fare.

Visualising correlation matrix-

#corrplot(sample)  
corofsample5=cor(sample5)  
#corrplot(corofsample5,method="circle")  
#corrplot(corofsample5,method="pie")  
#corrplot(corofsample5,method="ellipse")  
#corrplot(corofsample5,method="color")  
#corrplot(corofsample5,method="number")  
#corrplot(corofsample5,method="square")  
corrplot(corofsample5,method="circle",type="upper")



#corrplot(corofsample5,method="circle",type="lower")

Above correlation plot shows relation between different continous variables. Alsoevry correlation is positive suggesting increaing trend between different variables.

#Correlation matrix with p values.  
library(Hmisc)

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:plyr':  
##   
## is.discrete, summarize

## The following objects are masked from 'package:base':  
##   
## format.pval, units

corofsample5withp=rcorr(as.matrix(corofsample5))  
corofsample5withp

## trip\_seconds trip\_miles fare tips extras trip\_total  
## trip\_seconds 1.00 0.58 0.67 0.27 -0.17 0.45  
## trip\_miles 0.58 1.00 0.98 -0.35 0.62 0.97  
## fare 0.67 0.98 1.00 -0.33 0.57 0.96  
## tips 0.27 -0.35 -0.33 1.00 -0.91 -0.52  
## extras -0.17 0.62 0.57 -0.91 1.00 0.76  
## trip\_total 0.45 0.97 0.96 -0.52 0.76 1.00  
##   
## n= 6   
##   
##   
## P  
## trip\_seconds trip\_miles fare tips extras trip\_total  
## trip\_seconds 0.2288 0.1472 0.6049 0.7511 0.3680   
## trip\_miles 0.2288 0.0005 0.4924 0.1938 0.0013   
## fare 0.1472 0.0005 0.5233 0.2354 0.0019   
## tips 0.6049 0.4924 0.5233 0.0109 0.2913   
## extras 0.7511 0.1938 0.2354 0.0109 0.0793   
## trip\_total 0.3680 0.0013 0.0019 0.2913 0.0793

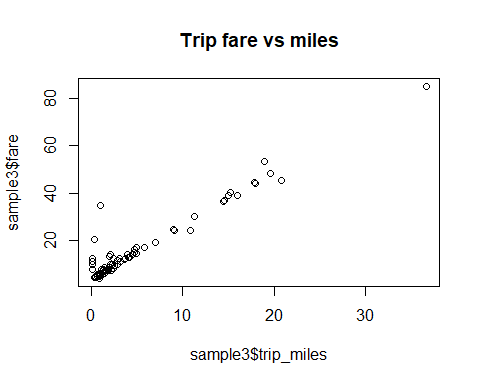
Above is the correlation matrix of continous variables in the sample with p values and showing both relation and significance of relation.

1. Using fare as the dependent variable, build a regression model using trip\_seconds, trip\_miles, and payment\_type as potential independent variables. Evaluate the quality of fit of the model to your cleansed data. Explain the impact each independent variable in your model on the dependent variable, considering the 95% confidence interval on the beta coefficients.

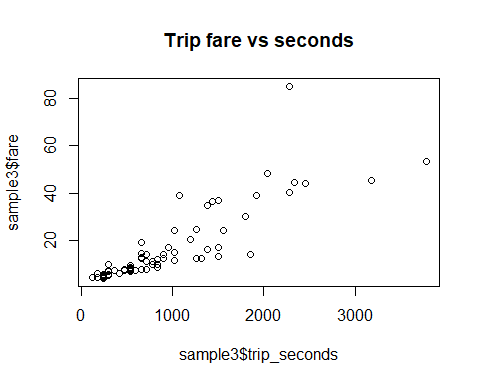
#analysis4-regression model b/w  
attach(sample3)

## The following objects are masked from sample:  
##   
## extras, fare, payment\_type, taxi\_id, tips, tolls, trip\_miles,  
## trip\_seconds, trip\_total

#farelr=lm(sample3$fare~sample3$trip\_miles+sample3$trip\_seconds+sample3$payment\_type,data=sample3)  
plot(sample3$trip\_miles,sample3$fare, main = "Trip fare vs miles")



plot(sample3$trip\_seconds,sample3$fare, main = "Trip fare vs seconds")



#plot(payment\_type,fare) #box plot

From the above two graphs, we could see a visible relation(linear) between independent and dependent variables.

Simple Regression

farelr=lm(fare~trip\_miles+trip\_seconds+payment\_type,data=sample3)  
summary(farelr)

##   
## Call:  
## lm(formula = fare ~ trip\_miles + trip\_seconds + payment\_type,   
## data = sample3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.8508 -1.1588 -0.5382 0.0602 22.6552   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.5604047 0.7900618 4.506 2.62e-05 \*\*\*  
## trip\_miles 1.6967706 0.1040080 16.314 < 2e-16 \*\*\*  
## trip\_seconds 0.0051722 0.0009515 5.436 7.74e-07 \*\*\*  
## payment\_typeCredit Card -0.7855587 0.8612779 -0.912 0.365   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.653 on 69 degrees of freedom  
## Multiple R-squared: 0.943, Adjusted R-squared: 0.9406   
## F-statistic: 380.7 on 3 and 69 DF, p-value: < 2.2e-16

R-square value or coefficient of determination for the above model is 94.3%.

Intercept- Intercept suggests that there is a default base fare of $3.56 associated with booking a taxi. P value is less than 0.05, suggesting there is significant evidence to reject the null hypothesis that this Beta coefficient(Intercept) is Zero and accept the alternate hypothesis that beta coefficient is not Zero.

Trip mile- Beta value here suggest that for every increase in a mile of trip, the fare of the trip can increase by $1.69. P value is less than 0.05, suggesting there is significant evidence to reject the null hypothesis that Beta coefficient on trip mile variable is Zero and accept the alternate hypothesis that beta coefficient is not Zero.

Trip second- Beta value here suggest that for every increase in time of a second on trip, the fare of the trip can increase by $0.005. P value is less than 0.05, suggesting there is significant evidence to reject the null hypothesis that Beta coefficient on trip seconds variable is Zero and accept the alternate hypothesis that beta coefficient is not Zero.

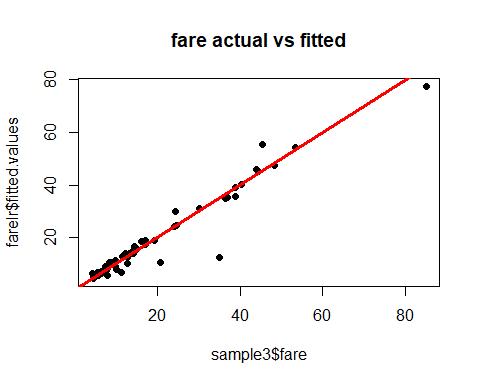
Payment type(credit card)- Beta value here suggest that if the payment for the taxi trip is made by credit card there is a reduction of $0.785 on fare rather than payment by cash. P value is greater than 0.05, suggesting there is no significant evidence to reject the null hypothesis that Beta coefficient on trip fare payment type by credit card variable is Zero and accept the null hypothesis that beta coefficient is not Zero with a confidence interval of 95%.

confint(farelr,level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) 1.98427450 5.136534964  
## trip\_miles 1.48928041 1.904260863  
## trip\_seconds 0.00327391 0.007070445  
## payment\_typeCredit Card -2.50376109 0.932643704

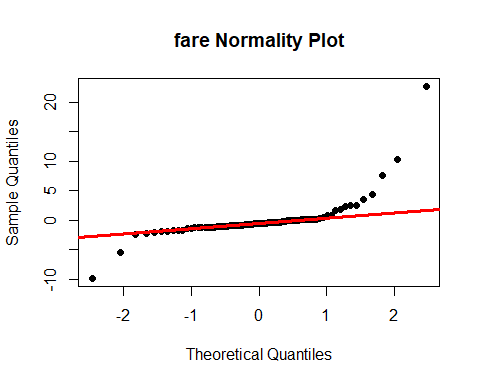
Though the Beta coefficients in the above model have a particular values confidence interval suggests variation. 95% Confidence Interval for Beta coefficient or Intercept is from 1.984 to 5.136. 95% Confidence Interval for Beta coefficient on trip miles is from 1.489 to 1.904. 95% confidence interval for Beta coefficient on trip seconds is from 0.003 to 0.007. 95% confidence interval for Beta coefficient on Payment type is from 0-2.50 to 0.932.

#checking for LINE assumptions on this model  
#linearty  
{plot(sample3$fare,farelr$fitted.values,pch=19,main="fare actual vs fitted")  
abline(0,1,col="red",lwd=3)}



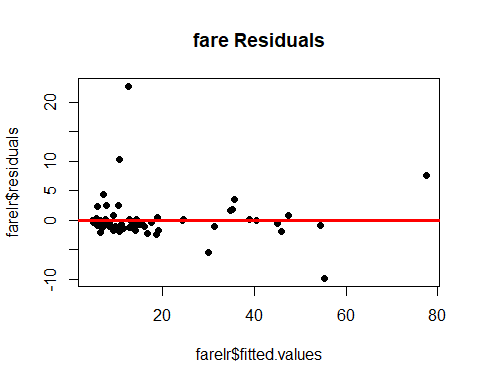
From the above graph, fitted vs actuals, most of the values fit to the line, the model is farely linear with some outliers.

#normality  
{qqnorm(farelr$residuals,pch=19,main="fare Normality Plot")  
qqline(farelr$residuals,col="red",lwd=3)}

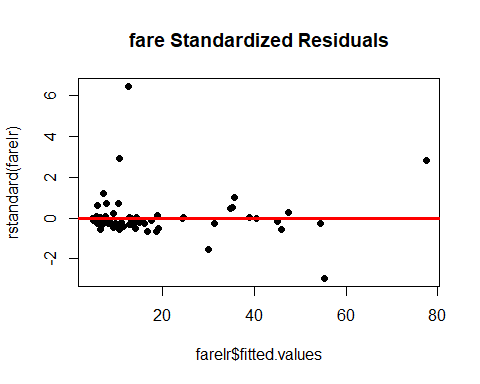


From the below graph normality assumption is accurate for points between -2 to +1 standard deviations and other points are not normally distributed and not on line. These might be due to high leverage points and outliers in the input dataset.

#equality of variance  
{plot(farelr$fitted.values,farelr$residuals,pch=19,main="fare Residuals")  
abline(0,0,col="red",lwd=3)}

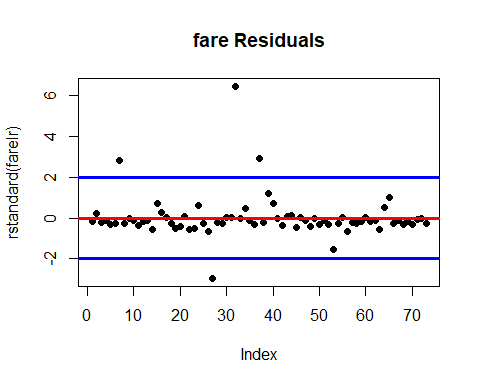


#standardised  
{plot(farelr$fitted.values,rstandard(farelr),pch=19,main="fare Standardized Residuals")  
abline(0,0,col="red",lwd=3)}



From the above graph the data shows heteroscedasticity with most of the data below line in range below $20, might be due to outliers and leverage points. Also there are some ouliers beyond -2sd and 2sd. This assumption is failed.

{plot(rstandard(farelr),pch=19,main="fare Residuals")  
abline(0,0,col="red",lwd=3)  
abline(0,0,col="red",lwd=3)  
abline(2,0,col='blue',lwd=3)  
abline(-2,0,col='blue',lwd=3)}



The residual plot shows a fairly random pattern with some points being outliers.

1. Investigate relevant interactions and common independent variable transforms to determine if adding these to your model will result in a better model fit. Depending on your random data selection you may find it necessary to do some additional cleansing of your data in order to get a better model fit for the majority of data points.

Independent variable transformation to higher order

#introducing higher order terms of trip miles and trip seconds on previously built model and removing payment card asit's insignificant in previous model  
farehm=lm(fare~poly(trip\_miles,1)+poly(trip\_seconds,1),data=sample3)  
summary(farehm)

##   
## Call:  
## lm(formula = fare ~ poly(trip\_miles, 1) + poly(trip\_seconds,   
## 1), data = sample3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.3405 -1.0899 -0.5648 -0.2065 23.0797   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.495 0.427 38.632 < 2e-16 \*\*\*  
## poly(trip\_miles, 1) 96.782 5.913 16.369 < 2e-16 \*\*\*  
## poly(trip\_seconds, 1) 31.816 5.913 5.381 9.31e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.648 on 70 degrees of freedom  
## Multiple R-squared: 0.9423, Adjusted R-squared: 0.9407   
## F-statistic: 572 on 2 and 70 DF, p-value: < 2.2e-16

library(car)

## Loading required package: carData

vif(farehm)

## poly(trip\_miles, 1) poly(trip\_seconds, 1)   
## 2.626716 2.626716

farehm2=lm(fare~poly(trip\_miles,2)+poly(trip\_seconds,1),data=sample3)  
summary(farehm2)

##   
## Call:  
## lm(formula = fare ~ poly(trip\_miles, 2) + poly(trip\_seconds,   
## 1), data = sample3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.6649 -1.3212 -0.7479 0.4885 21.7269   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.4952 0.4047 40.762 < 2e-16 \*\*\*  
## poly(trip\_miles, 2)1 91.9541 5.8318 15.768 < 2e-16 \*\*\*  
## poly(trip\_miles, 2)2 11.0050 3.6822 2.989 0.00388 \*\*   
## poly(trip\_seconds, 1) 37.9517 5.9678 6.359 1.89e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.457 on 69 degrees of freedom  
## Multiple R-squared: 0.9489, Adjusted R-squared: 0.9467   
## F-statistic: 427.5 on 3 and 69 DF, p-value: < 2.2e-16

vif(farehm2)

## GVIF Df GVIF^(1/(2\*Df))  
## poly(trip\_miles, 2) 2.979238 2 1.313791  
## poly(trip\_seconds, 1) 2.979238 1 1.726047

farehm3=lm(fare~poly(trip\_miles,3)+poly(trip\_seconds,1),data=sample3)  
summary(farehm3)

##   
## Call:  
## lm(formula = fare ~ poly(trip\_miles, 3) + poly(trip\_seconds,   
## 1), data = sample3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.6031 -1.3167 -0.7296 0.4564 21.7239   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.4952 0.4076 40.467 < 2e-16 \*\*\*  
## poly(trip\_miles, 3)1 91.8625 6.0329 15.227 < 2e-16 \*\*\*  
## poly(trip\_miles, 3)2 11.0297 3.7275 2.959 0.00424 \*\*   
## poly(trip\_miles, 3)3 0.2418 3.6266 0.067 0.94703   
## poly(trip\_seconds, 1) 38.0681 6.2597 6.081 6.12e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.483 on 68 degrees of freedom  
## Multiple R-squared: 0.949, Adjusted R-squared: 0.9459   
## F-statistic: 316 on 4 and 68 DF, p-value: < 2.2e-16

vif(farehm3)

## GVIF Df GVIF^(1/(2\*Df))  
## poly(trip\_miles, 3) 3.230546 3 1.215848  
## poly(trip\_seconds, 1) 3.230546 1 1.797372

#p value is very much greater than 0.05  
farehm4=lm(fare~poly(trip\_miles,2)+poly(trip\_seconds,2),data=sample3)  
summary(farehm4)

##   
## Call:  
## lm(formula = fare ~ poly(trip\_miles, 2) + poly(trip\_seconds,   
## 2), data = sample3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.8588 -1.4047 -0.1887 0.6845 20.3020   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.4952 0.3839 42.970 < 2e-16 \*\*\*  
## poly(trip\_miles, 2)1 91.2303 5.5377 16.474 < 2e-16 \*\*\*  
## poly(trip\_miles, 2)2 12.7441 3.5426 3.597 0.000605 \*\*\*  
## poly(trip\_seconds, 2)1 38.8904 5.6702 6.859 2.55e-09 \*\*\*  
## poly(trip\_seconds, 2)2 -9.7976 3.3264 -2.945 0.004413 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.28 on 68 degrees of freedom  
## Multiple R-squared: 0.9547, Adjusted R-squared: 0.9521   
## F-statistic: 358.5 on 4 and 68 DF, p-value: < 2.2e-16

vif(farehm4)

## GVIF Df GVIF^(1/(2\*Df))  
## poly(trip\_miles, 2) 3.064386 2 1.323079  
## poly(trip\_seconds, 2) 3.064386 2 1.323079

farehm5=lm(fare~poly(trip\_miles,2)+poly(trip\_seconds,3),data=sample3)  
summary(farehm5)

##   
## Call:  
## lm(formula = fare ~ poly(trip\_miles, 2) + poly(trip\_seconds,   
## 3), data = sample3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.3688 -1.6022 -0.1798 0.8380 20.4205   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.4952 0.3838 42.973 < 2e-16 \*\*\*  
## poly(trip\_miles, 2)1 93.3356 5.9199 15.766 < 2e-16 \*\*\*  
## poly(trip\_miles, 2)2 13.1898 3.5699 3.695 0.000445 \*\*\*  
## poly(trip\_seconds, 3)1 37.3282 5.8787 6.350 2.17e-08 \*\*\*  
## poly(trip\_seconds, 3)2 -9.8708 3.3269 -2.967 0.004167 \*\*   
## poly(trip\_seconds, 3)3 3.6137 3.5940 1.005 0.318282   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.28 on 67 degrees of freedom  
## Multiple R-squared: 0.9554, Adjusted R-squared: 0.9521   
## F-statistic: 287 on 5 and 67 DF, p-value: < 2.2e-16

farehm6=lm(fare~poly(trip\_miles,3)+poly(trip\_seconds,3),data=sample3)  
summary(farehm6)

##   
## Call:  
## lm(formula = fare ~ poly(trip\_miles, 3) + poly(trip\_seconds,   
## 3), data = sample3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.9908 -1.3086 -0.1030 0.8313 19.9723   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.4952 0.3715 44.400 < 2e-16 \*\*\*  
## poly(trip\_miles, 3)1 99.0752 6.2285 15.907 < 2e-16 \*\*\*  
## poly(trip\_miles, 3)2 13.5189 3.4580 3.909 0.000221 \*\*\*  
## poly(trip\_miles, 3)3 -9.9566 4.2369 -2.350 0.021774 \*   
## poly(trip\_seconds, 3)1 31.2723 6.2462 5.007 4.38e-06 \*\*\*  
## poly(trip\_seconds, 3)2 -14.7665 3.8352 -3.850 0.000269 \*\*\*  
## poly(trip\_seconds, 3)3 7.6016 3.8704 1.964 0.053741 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.174 on 66 degrees of freedom  
## Multiple R-squared: 0.9588, Adjusted R-squared: 0.9551   
## F-statistic: 256.3 on 6 and 66 DF, p-value: < 2.2e-16

By introducing higher order terms on both trip miles and trip seconds, R2 square value has not significantly changing, Also Residual SE has also not significantly changing.

#Sticking to basic model  
farelr2=lm(fare~poly(trip\_miles,1)+poly(trip\_seconds,1),data=sample3)  
summary(farelr2)

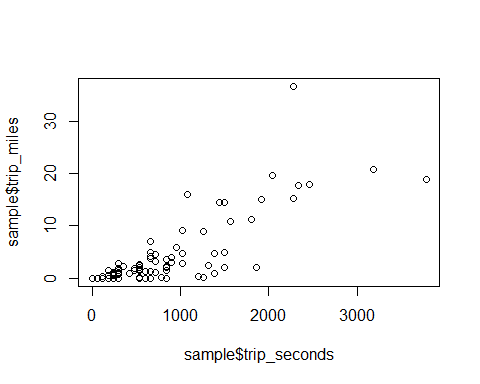
##   
## Call:  
## lm(formula = fare ~ poly(trip\_miles, 1) + poly(trip\_seconds,   
## 1), data = sample3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.3405 -1.0899 -0.5648 -0.2065 23.0797   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.495 0.427 38.632 < 2e-16 \*\*\*  
## poly(trip\_miles, 1) 96.782 5.913 16.369 < 2e-16 \*\*\*  
## poly(trip\_seconds, 1) 31.816 5.913 5.381 9.31e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.648 on 70 degrees of freedom  
## Multiple R-squared: 0.9423, Adjusted R-squared: 0.9407   
## F-statistic: 572 on 2 and 70 DF, p-value: < 2.2e-16

Sticking to basic model.

Introducing Interaction term-

It is abvious that Independent variables trip time and trip mile are interrelated and trip seconds increases as increase in trip miles with progress of trip.

#and interaction between trip miles and trip seconds  
plot(sample$trip\_seconds,sample$trip\_miles)



#there is a relation b/w trip seconds and trip miles i.e interaction

we could see there is some relationship between Trip time(seconds) and trip fare.Adding this Interaction to our linear model.

#adding interaction  
farelr3=lm(fare~poly(trip\_miles,1)+poly(trip\_seconds,1)+trip\_seconds:trip\_miles,data=sample3)  
summary(farelr3)

##   
## Call:  
## lm(formula = fare ~ poly(trip\_miles, 1) + poly(trip\_seconds,   
## 1) + trip\_seconds:trip\_miles, data = sample3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.9034 -1.1834 -0.5331 0.1620 22.7553   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.739e+01 8.767e-01 19.833 < 2e-16 \*\*\*  
## poly(trip\_miles, 1) 1.075e+02 1.093e+01 9.832 9.32e-15 \*\*\*  
## poly(trip\_seconds, 1) 3.642e+01 7.097e+00 5.131 2.53e-06 \*\*\*  
## trip\_seconds:trip\_miles -1.053e-04 9.034e-05 -1.165 0.248   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.639 on 69 degrees of freedom  
## Multiple R-squared: 0.9435, Adjusted R-squared: 0.941   
## F-statistic: 383.7 on 3 and 69 DF, p-value: < 2.2e-16

#sticking to the basic model  
farelr2=lm(fare~poly(trip\_miles,1)+poly(trip\_seconds,1),data=sample3)  
summary(farelr2)

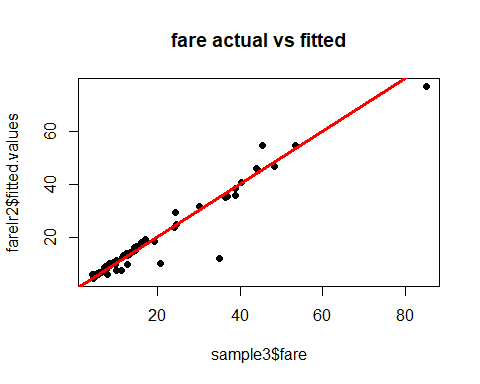
##   
## Call:  
## lm(formula = fare ~ poly(trip\_miles, 1) + poly(trip\_seconds,   
## 1), data = sample3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.3405 -1.0899 -0.5648 -0.2065 23.0797   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.495 0.427 38.632 < 2e-16 \*\*\*  
## poly(trip\_miles, 1) 96.782 5.913 16.369 < 2e-16 \*\*\*  
## poly(trip\_seconds, 1) 31.816 5.913 5.381 9.31e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.648 on 70 degrees of freedom  
## Multiple R-squared: 0.9423, Adjusted R-squared: 0.9407   
## F-statistic: 572 on 2 and 70 DF, p-value: < 2.2e-16

vif(farelr2)

## poly(trip\_miles, 1) poly(trip\_seconds, 1)   
## 2.626716 2.626716

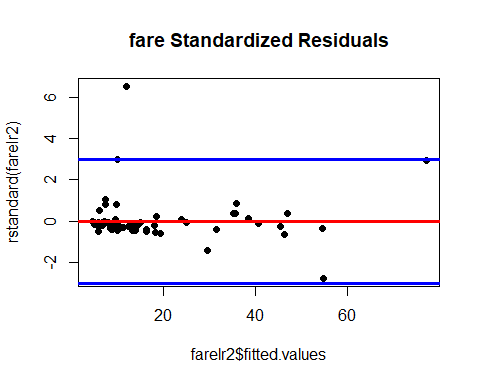
The graph between trip miles and trip seconds suggest there is an interaction between these two variable and however, P value of the interaction term is greater than 0.05 and suggests that the interaction term is not statistically significant. Also the difference’s between R2 value and residual standard error, without interaction(R2-0.9423,RSE-3.648) and with interaction(R2-0.9435,RSE-3.6399) are not significant to add interaction into model. Also Vif of model is less than 10,suggesting the variance of a beta coefficient is not being inflated by multicollinearity.

#actual vs fitted of thismodel  
{plot(sample3$fare,farelr2$fitted.values,pch=19,main="fare actual vs fitted")  
abline(0,1,col="red",lwd=3)}

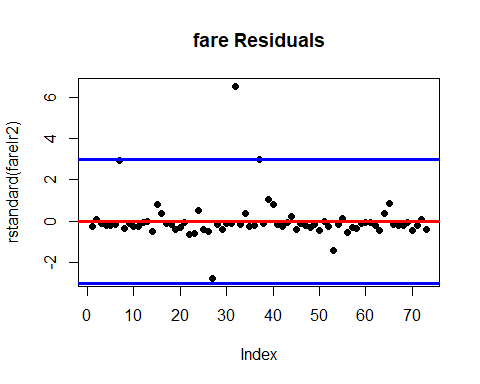


From the above graph, fitted vs actuals, most of the values fit to the line, the model is farely linear with some outliers.

#standardised  
{plot(farelr2$fitted.values,rstandard(farelr2),pch=19,main="fare Standardized Residuals")  
abline(0,0,col="red",lwd=3)  
abline(3,0,col='blue',lwd=3)  
abline(-3,0,col='blue',lwd=3)}



#standardised residual plot with index  
{plot(rstandard(farelr2),pch=19,main="fare Residuals")  
abline(0,0,col="red",lwd=3)  
abline(3,0,col='blue',lwd=3)  
abline(-3,0,col='blue',lwd=3)}



From the above graphs, We could see there are some outliers. Identifying and removing those points increased our R2 value and decreased Residual standard error significantly.

#Identifying are removing outliers  
sample6=sample3[]  
sample6$res=rstandard(farelr2)  
sd2=3\*sd(sample6$res)  
sample7=sample6[!abs(sample6$res)>3,]  
#new regression on new sample  
farelr4=lm(fare~poly(trip\_miles,1)+poly(trip\_seconds,1),data=sample7)  
summary(farelr4)

##   
## Call:  
## lm(formula = fare ~ poly(trip\_miles, 1) + poly(trip\_seconds,   
## 1), data = sample7)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.8588 -0.5139 -0.2096 0.2622 5.1806   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.1739 0.2081 77.723 < 2e-16 \*\*\*  
## poly(trip\_miles, 1) 108.5573 2.9470 36.837 < 2e-16 \*\*\*  
## poly(trip\_seconds, 1) 19.5501 2.9470 6.634 6.43e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.753 on 68 degrees of freedom  
## Multiple R-squared: 0.9868, Adjusted R-squared: 0.9864   
## F-statistic: 2533 on 2 and 68 DF, p-value: < 2.2e-16

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| model | miles order, Pvalue(<0.05) | Seconds order,P value<0.05 | Interaction | R2 value | RSE | Significance of Multicollinearity(considering VIf<5) |
| farehm | 1,Significant | 1,Significant | No | 0.9423 | 3.648 | no |
| farehm2 | 2,Significant | 1,Significant | No | 0.9489 | 3.457 | no |
| farehm3 | 3,Non significant on cube | 1,Significant | No | 0.949 | 3.483 | no |
| farehm4 | 2,Significant | 2,Significant | No | 0.9547 | 3.28 | no |
| farehm5 | 2,Significant | 3,Non significant on cube | No | 0.9554 | 3.28 | no |
| farehm6 | 3,Significant | 3,Significant | No | 0.959 | 3.198 | no |
| farehm7 | 2,Significant | 1,Significant | yes | 0.9588 | 3.174 | yes |
| farehm8 | 2,Significant | 2,Non significant | yes | 0.958 | 3.182 | no |
| farehm9 | 1,Significant | 2,Significant | yes | 0.9479 | 3.52 | yes |
| farehm10 | 1,Significant | 1,Significant | yes,non Significant | 0.9435 | 3.639 | yes |
| farehm11 | 3,significant | 3,not significant | yes | 94.2 | 2.974 | no |

1. Of the various combinations you ran in Step 5, report the model which provides what you deem as the “best fit” to your sample data. Explain why you selected this particular model and show the standard R regression output for the model. Evaluate and explain your model’s conformity to the LINE assumptions of regression.

#sticking to the basic model  
farelr4=lm(fare~poly(trip\_miles,1)+poly(trip\_seconds,1),data=sample7)  
summary(farelr4)

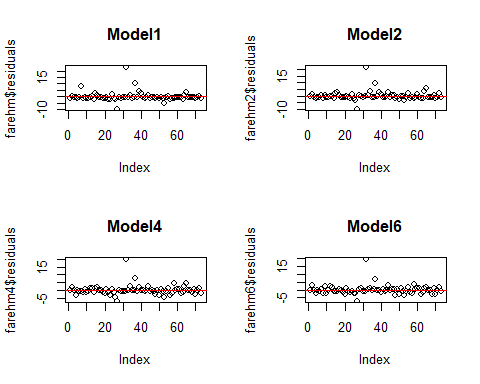
##   
## Call:  
## lm(formula = fare ~ poly(trip\_miles, 1) + poly(trip\_seconds,   
## 1), data = sample7)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.8588 -0.5139 -0.2096 0.2622 5.1806   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.1739 0.2081 77.723 < 2e-16 \*\*\*  
## poly(trip\_miles, 1) 108.5573 2.9470 36.837 < 2e-16 \*\*\*  
## poly(trip\_seconds, 1) 19.5501 2.9470 6.634 6.43e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.753 on 68 degrees of freedom  
## Multiple R-squared: 0.9868, Adjusted R-squared: 0.9864   
## F-statistic: 2533 on 2 and 68 DF, p-value: < 2.2e-16

vif(farelr4)

## poly(trip\_miles, 1) poly(trip\_seconds, 1)   
## 2.824678 2.824678

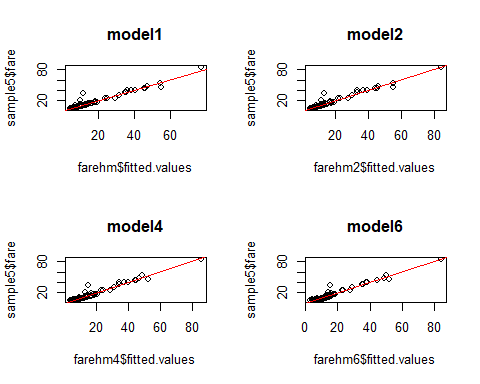
I have choosen to stick to this model, because by introducing higher order terms and interaction on both trip miles and trip seconds, R2 square value has not significantly changed and Residual SE values are also not significantly changed.Also,by introducing those terms there is a chance of model overfitting and Vif of model is less than 10, suggesting the variance of a beta coefficient is not being inflated by multicollinearity.

{par(mfrow=c(2,2))  
plot(farehm$residuals,main="Model1")  
abline(0,0,col="red")  
plot(farehm2$residuals,main="Model2")  
abline(0,0,col="red")  
plot(farehm4$residuals,main="Model4")  
abline(0,0,col="red")  
plot(farehm6$residuals,main="Model6")  
abline(0,0,col="red")  
par(mfrow=c(1,1))}



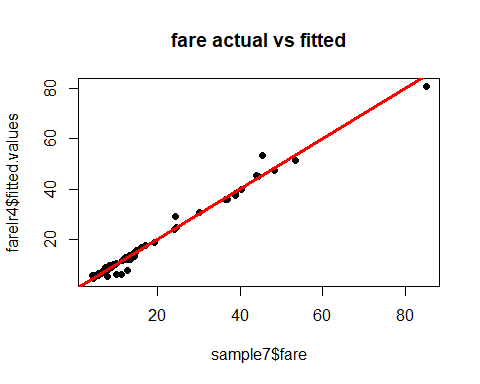
residual plots for model1s.

{par(mfrow=c(2,2))  
plot(farehm$fitted.values,sample5$fare,main="model1")  
abline(0,1,col="red")  
plot(farehm2$fitted.values,sample5$fare,main="model2")  
abline(0,1,col="red")  
plot(farehm4$fitted.values,sample5$fare,main="model4")  
abline(0,1,col="red")  
plot(farehm6$fitted.values,sample5$fare,main="model6")  
abline(0,1,col="red")  
par(mfrow=c(1,1))}



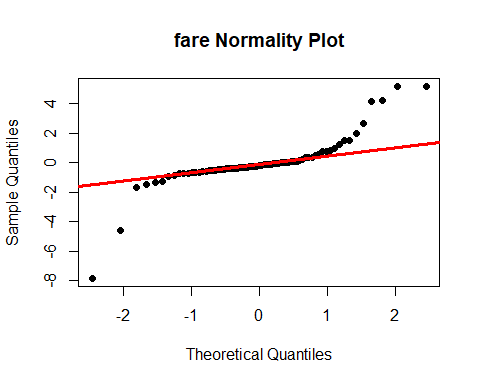
Actual vs Fitted for all the Variables are fairly fitted.

#checking for LINE assumptions on this model  
#linearty  
{plot(sample7$fare,farelr4$fitted.values,pch=19,main="fare actual vs fitted")  
abline(0,1,col="red",lwd=3)}



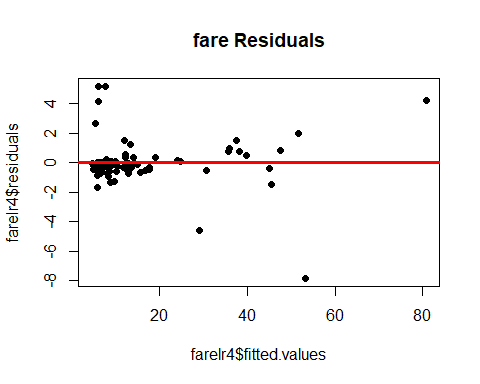
From the above graph, fitted vs actuals, most of the values fit to the line, the model is farely linear with some outliers.

#normality  
{qqnorm(farelr4$residuals,pch=19,main="fare Normality Plot")  
qqline(farelr4$residuals,col="red",lwd=3)}

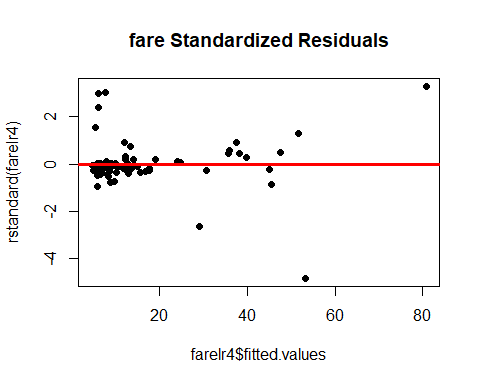


From the above graph normality assumption is accurate for points between -1 to +1 standard deviations and other points are not normally distributed and not on line. These might be due to high leverage points and outliers in the input dataset.

#equality of variance  
{plot(farelr4$fitted.values,farelr4$residuals,pch=19,main="fare Residuals")  
abline(0,0,col="red",lwd=3)}

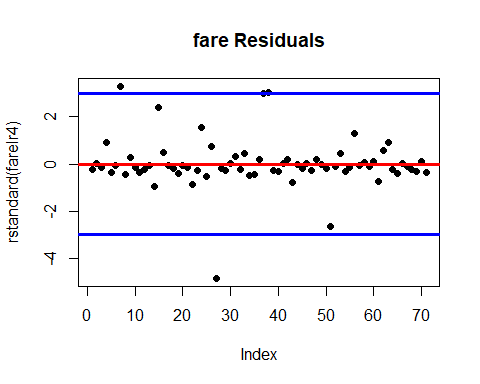


#standardised  
{plot(farelr4$fitted.values,rstandard(farelr4),pch=19,main="fare Standardized Residuals")  
abline(0,0,col="red",lwd=3)}



From the above graph the data shows heteroscedasticity with most of the data below line in range below $20, might be due to leverage points. Also there are some ouliers beyond -2sd and 2sd. This assumption is failed.

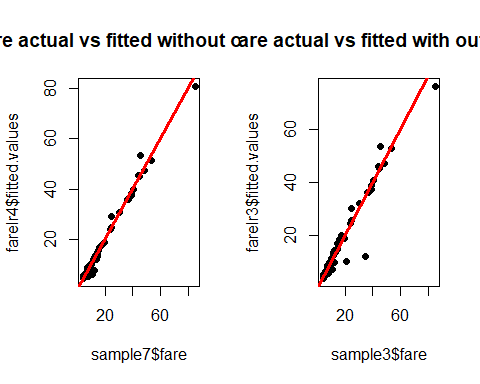
{plot(rstandard(farelr4),pch=19,main="fare Residuals")  
abline(0,0,col="red",lwd=3)  
abline(0,0,col="red",lwd=3)  
abline(3,0,col='blue',lwd=3)  
abline(-3,0,col='blue',lwd=3)}



The residual plot shows a fairly random pattern with some points being outliers. Independence assumption can be ruled out, since residuals are not related to each other.

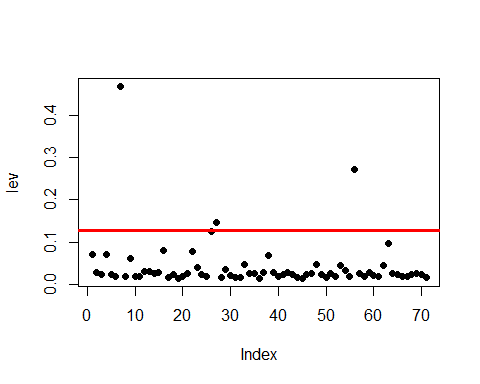
We can see the from the above graph there are some outliers in the dataset with standard deviation of that point being greater than 2sd line, This points are influencing our regression model with a large Resisual square value, inorder to reduce residual standard error, we need to remove these ouliers from the sample dataset we considered.

{par(mfrow=c(1,2))  
{plot(sample7$fare,farelr4$fitted.values,pch=19,main="fare actual vs fitted without outliers")  
abline(0,1,col="red",lwd=3)}  
{plot(sample3$fare,farelr3$fitted.values,pch=19,main="fare actual vs fitted with outliers")  
abline(0,1,col="red",lwd=3)}  
par(mfrow=c(1,1))}



1. Investigate and remove any data points deemed to have an inappropriately high leverage in determining the plot of the model. Rerun your model without these points and evaluate the quality of fit in this final regression model.

#Leverage of Points  
lev=hat(model.matrix(farelr4))  
{plot(lev,pch=19)  
abline(3\*mean(lev),0,col="red",lwd=3)}



reduced.sample=sample7[lev>(3\*mean(lev)),]  
View(reduced.sample)  
#index(sample$taxi\_id=reduced.sample$taxi\_id)  
sample8=sample7[-which(sample6$taxi\_id %in% reduced.sample$taxi\_id),]  
  
farelr5=lm(fare~poly(trip\_miles,1)+poly(trip\_seconds,1),data=sample8)  
summary(farelr5)

##   
## Call:  
## lm(formula = fare ~ poly(trip\_miles, 1) + poly(trip\_seconds,   
## 1), data = sample8)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.6620 -0.5128 -0.1874 0.1867 4.6469   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14.605 0.160 91.28 < 2e-16 \*\*\*  
## poly(trip\_miles, 1) 78.097 2.289 34.12 < 2e-16 \*\*\*  
## poly(trip\_seconds, 1) 24.573 2.289 10.74 4.91e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.319 on 65 degrees of freedom  
## Multiple R-squared: 0.9886, Adjusted R-squared: 0.9883   
## F-statistic: 2826 on 2 and 65 DF, p-value: < 2.2e-16

From the above graph, there are 3 leverage points,those being Identified and removed. R2 value is now 98.86 with ResidualStandard error being 1.319.

1. Return to the full data set of 1.7 million cases. Pull another sample of n=100 cases. (Be sure to use a new random number seed of the numerical portion of your U number plus 5.) To this data set apply the same cleansing procedures you used on your original sample data set. Referring to the model you developed in Step 6 above, apply that model to the new random set of data and evaluate how well the model fits this second data set.

set.seed(02248589) #022485894+5=02248589  
newsample=projectdata[sample(1:nrow(projectdata),100,replace=FALSE),]   
#View(sample)  
summary(newsample) #summary of sample dataset

## taxi\_id trip\_seconds trip\_miles fare   
## Min. : 64 Min. : 0.0 Min. : 0.000 Min. : 3.250   
## 1st Qu.:1600 1st Qu.: 240.0 1st Qu.: 0.000 1st Qu.: 5.750   
## Median :3974 Median : 450.0 Median : 1.000 Median : 7.625   
## Mean :3874 Mean : 574.2 Mean : 2.408 Mean :12.469   
## 3rd Qu.:5808 3rd Qu.: 735.0 3rd Qu.: 2.050 3rd Qu.:12.000   
## Max. :8734 Max. :2340.0 Max. :19.100 Max. :96.750   
## tips tolls extras trip\_total   
## Min. : 0.000 Min. :0 Min. : 0.00 Min. : 3.25   
## 1st Qu.: 0.000 1st Qu.:0 1st Qu.: 0.00 1st Qu.: 6.50   
## Median : 0.000 Median :0 Median : 0.00 Median : 9.05   
## Mean : 1.821 Mean :0 Mean : 0.98 Mean : 15.27   
## 3rd Qu.: 2.000 3rd Qu.:0 3rd Qu.: 1.00 3rd Qu.: 13.25   
## Max. :47.000 Max. :0 Max. :24.50 Max. :106.75   
## payment\_type   
## Length:100   
## Class :character   
## Mode :character   
##   
##   
##

#cleaning Step1  
newsample1=newsample[!(newsample$trip\_seconds==0 & newsample$trip\_miles==0),]  
  
#cleaning step2  
newsample2=newsample1[!(newsample1$trip\_miles==0),]  
predictedFare = predict(farelr5,newdata = newsample2)  
cor(predictedFare,newsample2$fare)

## [1] 0.9026485

The model developed predicted the new sample with an accuracy of 90.2%.

#RSE=sum((predictedFare - newsample2$fare)^2)  
#install.packages("Metrics")  
library(Metrics)

MSEP=mse(predictedFare,newsample2$fare)  
MSEP

## [1] 18.69661

MSEA=mse(farelr5$fitted.values,sample8$fare)  
MSEA

## [1] 1.66406

MAEP=mae(predictedFare,newsample2$fare)  
MAEP

## [1] 1.636319

MAEA=mae(farelr5$fitted.values,sample8$fare)  
MAEA

## [1] 0.78421

rmseP=MSEP^1/2  
rmseP

## [1] 9.348303

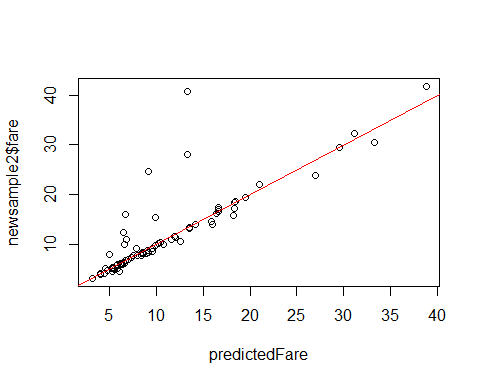
rmseA=MSEA^1/2  
rmseA

## [1] 0.8320299

#There are huge differences in Mean absolute Errors, Mean squared errors and Root Mean squared error of predicted values of sample vs errors in actual model we built. Lets plot the graph to see the huge differences.

#rsquared(predictedFare,newsample2$fare)#farelr6=lm(fare~poly(trip\_miles,1)+poly(trip\_seconds,1),data=newsample)  
#summary(farelr6)

{plot(predictedFare,newsample2$fare)  
abline(0,1,col="red")}



#The huge differences in the errors is due to the outliers in the new sample datset.to reduce error metrics we need to remove outliers

farelr6=lm(fare~poly(trip\_miles,1)+poly(trip\_seconds,1),data=newsample2)  
summary(farelr6)

##   
## Call:  
## lm(formula = fare ~ poly(trip\_miles, 1) + poly(trip\_seconds,   
## 1), data = newsample2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.4937 -2.0272 -0.5526 0.7881 25.5310   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.2826 0.5282 25.146 < 2e-16 \*\*\*  
## poly(trip\_miles, 1) 47.5375 6.6293 7.171 6.52e-10 \*\*\*  
## poly(trip\_seconds, 1) 55.0884 6.6293 8.310 5.42e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.482 on 69 degrees of freedom  
## Multiple R-squared: 0.8685, Adjusted R-squared: 0.8647   
## F-statistic: 227.8 on 2 and 69 DF, p-value: < 2.2e-16

|  |
| --- |
| Also, R square value for the predicted sample is 86.85 with a residual standard error of predicted sample is 4.482. |