Regression Project

Yagna Dheepika Venkitasamy

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:
   1. taxi\_id: a unique identifier for each individual taxi cab.
   2. trip\_seconds: the number of seconds elapsed during the trip.
   3. trip\_miles: the number of miles logged during the trip.
   4. fare: the base fare charged to the customer for the trip.
   5. tips: the tip given by the customer to the driver for the trip.
   6. tolls: any surcharges for road or bridge tolls incurred during the trip.
   7. extras: charges for any incidentals requested by the customer.
   8. trip\_total: the total charge to the customer for the trip.
   9. 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”.
2. 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.
3. 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.

Pre-processing:

rm(list=ls())  
library(readxl)  
library(moments)  
library(rio)

## Warning: package 'rio' was built under R version 3.5.3

library(data.table)  
library(car)

## Loading required package: carData

library(carData)  
master.data = read.csv("6304 Regression Project Data.csv")  
attach(master.data)  
set.seed(24889542)  
my.data = master.data[sample(1:nrow(master.data),100,replace=FALSE),]  
attach(my.data)

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

summary(my.data)

## taxi\_id trip\_seconds trip\_miles fare   
## Min. : 6 Min. : 0 Min. : 0.00 Min. : 3.250   
## 1st Qu.:1765 1st Qu.: 240 1st Qu.: 0.00 1st Qu.: 6.000   
## Median :4072 Median : 420 Median : 1.10 Median : 7.625   
## Mean :4037 Mean : 594 Mean : 3.27 Mean :14.318   
## 3rd Qu.:6374 3rd Qu.: 735 3rd Qu.: 2.55 3rd Qu.:15.625   
## Max. :8604 Max. :3780 Max. :33.00 Max. :80.750   
## tips tolls extras trip\_total   
## Min. : 0.000 Min. :0 Min. : 0.000 Min. : 3.250   
## 1st Qu.: 0.000 1st Qu.:0 1st Qu.: 0.000 1st Qu.: 6.963   
## Median : 0.000 Median :0 Median : 0.000 Median : 9.500   
## Mean : 1.913 Mean :0 Mean : 1.485 Mean : 17.715   
## 3rd Qu.: 2.000 3rd Qu.:0 3rd Qu.: 1.000 3rd Qu.: 18.525   
## Max. :25.150 Max. :0 Max. :45.000 Max. :150.900   
## payment\_type  
## Cash :50   
## Credit Card:50   
## Other : 0   
##   
##   
##

Data cleaning Steps: Removing the max outlier in trip total variable

max(my.data$trip\_total)

## [1] 150.9

which(my.data$trip\_total ==150.9)

## [1] 38

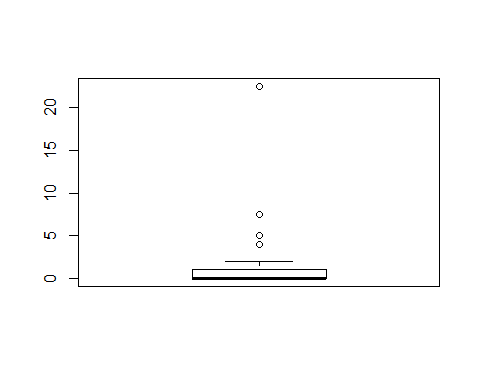
my.data = my.data[-38,]

Removing the data rows that has trip seconds as 0 and trip total>0

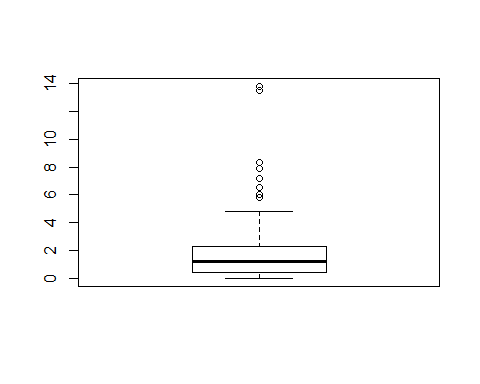
final.data = subset(my.data, trip\_seconds !=0 )

Removing the outlier in the extras and trip miles variables

outliers\_extras = boxplot(final.data$extras)$out



final.data = final.data[-which(final.data$extras %in% outliers\_extras),]  
outliers\_miles = boxplot(final.data$trip\_miles)$out



final.data = final.data[-which(final.data$trip\_miles %in% outliers\_miles),]

* The data sampled is cleaned with the underlying assumption that if a passenger gets into the cab, he would travel a few miles for a few seconds, he would pay the fare with tips and applicable extras as total trip fare.
* If the dataset contains rows with trip miles as 0 and trip seconds as 0 but if he has paid some fare, which doesn’t make any sense. Hence I removed the such rows which had the aberrancies. There were 15 of such rows.
* If the samples contained exorbitantly high values in the variables such as extras and miles and trip total, I removed them. There were 9 + 8 + 1= 18 such samples.
* I ended up with having 67 data samples to run my analysis on.

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.

summary(final.data [,c(2:9)])

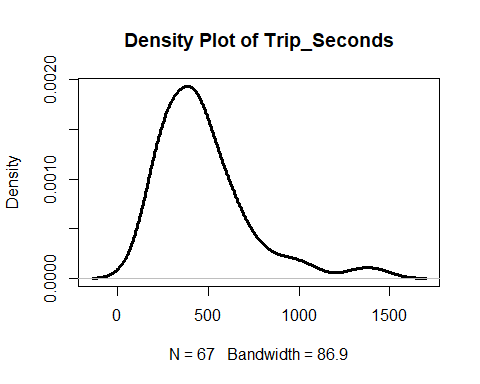
## trip\_seconds trip\_miles fare tips   
## Min. : 120.0 Min. :0.000 Min. : 3.500 Min. :0.0000   
## 1st Qu.: 300.0 1st Qu.:0.400 1st Qu.: 5.750 1st Qu.:0.0000   
## Median : 420.0 Median :1.000 Median : 7.000 Median :0.0000   
## Mean : 472.8 Mean :1.199 Mean : 7.828 Mean :0.7455   
## 3rd Qu.: 600.0 3rd Qu.:1.550 3rd Qu.: 8.375 3rd Qu.:2.0000   
## Max. :1440.0 Max. :4.830 Max. :20.500 Max. :3.5000

## tolls extras trip\_total payment\_type  
## Min. :0 Min. :0.0000 Min. : 3.500 Cash :41   
## 1st Qu.:0 1st Qu.:0.0000 1st Qu.: 6.375 Credit Card:26   
## Median :0 Median :0.0000 Median : 7.750 Other : 0   
## Mean :0 Mean :0.3209 Mean : 8.894   
## 3rd Qu.:0 3rd Qu.:1.0000 3rd Qu.:10.000   
## Max. :0 Max. :2.0000 Max. :20.500

## Most of the variables in the sample data are right skewed due to the presence of huge number of outliers in the trip seconds variable. This is apparent in the values of mean and median where all the variables have Mean > Median which is the case for right skewed distribution. Tolls appears to be normally distributed as all the values are 0 in my sample.

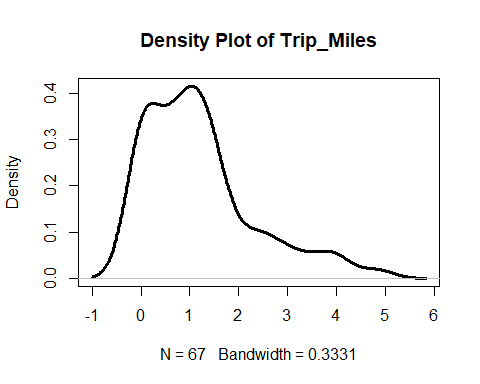
##Now let’s see the density plots for each variable:

par(mfrow=c(1, 1))  
plot(density(final.data$trip\_seconds),lwd=3, main="Density Plot of Trip\_Seconds")

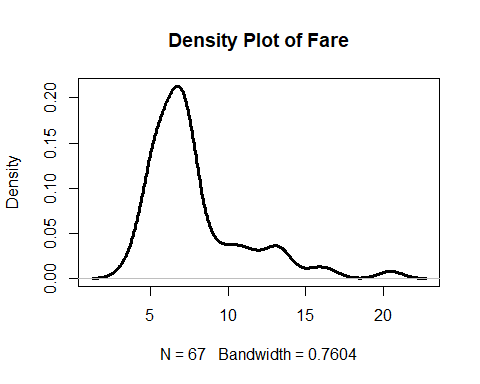


We can see that mean of the trip seconds variable is at 472.8 seconds where most of the values are concentrated as in the peak of the curve. There is also presence of high number of outliers which is clear from the extended tail beyond 1000 seconds.

plot(density(final.data$trip\_miles),lwd=3, main="Density Plot of Trip\_Miles")

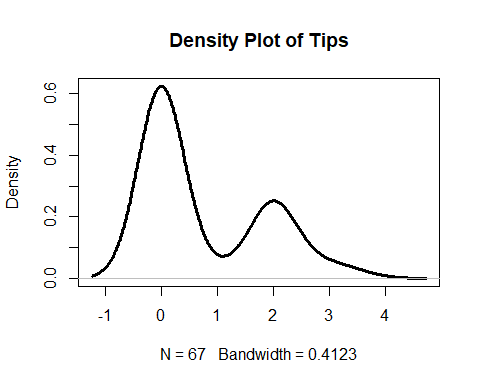


The density plot of trip miles variables seems to be bimodal distribution which had 2 peaks around 0.5 and 1.2. The mean of the distribution is at 1.199 miles and median of 1. The distribution is right skewed too as mean > median and outlier at 4.8 miles. plot(density(final.data$fare),lwd=3, main="Density Plot of Fare")

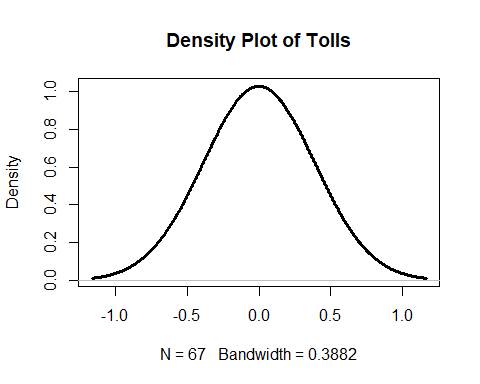


The density plot of fare variable seems to be unimodal with the peak at $7.8. The mean of the distribution is at $7.8 and median at $7. The distribution is right skewed too as mean > median and outlier at $20.5.

plot(density(final.data$tips),lwd=3, main="Density Plot of Tips")

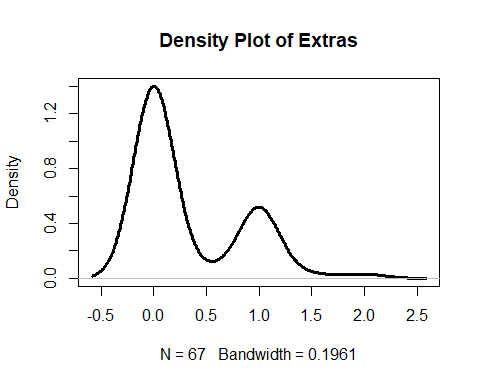


The density plot of tips variable seems to be bimodal distribution which had 2 peaks around 0 and 2. The mean of the distribution is at 1.199 miles and median of 1. The distribution is right skewed too as mean > median and outlier at 4.8 miles. plot(density(final.data$tolls),lwd=3, main="Density Plot of Tolls")



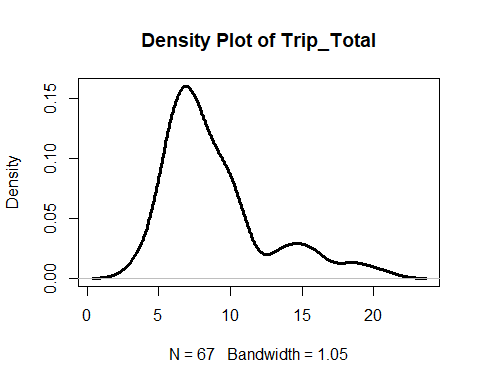
We can see that the density plot of tolls variable is normally distributed only because all the values in my sample for this variable is 0.

plot(density(final.data$extras),lwd=3, main="Density Plot of Extras")



The density plot of extras variables seems to be bimodal distribution which had 2 peaks around $0 and $1. The mean of the distribution is at 1.199 $0.32 and median of 0. The distribution is right skewed too as mean > median and outlier at $2.

plot(density(final.data$trip\_total),lwd=3, main="Density Plot of Trip\_Total")



The density plot of total fare variable seems to be unimodal distribution that has its peak at $9. The mean of the distribution is at $8.894 and median of $7.75. The distribution is right skewed too as mean > median and outlier at $20.50.

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.

library(plyr)  
count(final.data, vars = c("payment\_type"))

## payment\_type freq  
## 1 Cash 41  
## 2 Credit Card 26

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.

cor(final.data[,c(2:5,7,8)])

## trip\_seconds trip\_miles fare tips extras trip\_total  
## trip\_seconds 1.0000000 0.51820523 0.8275828 0.27086442 0.17824083 0.8304162  
## trip\_miles 0.5182052 1.00000000 0.5600508 0.03357544 0.35342341 0.5506228  
## fare 0.8275828 0.56005078 1.0000000 0.13269347 0.20316516 0.9450726  
## tips 0.2708644 0.03357544 0.1326935 1.00000000 0.09095017 0.4202991  
## extras 0.1782408 0.35342341 0.2031652 0.09095017 1.00000000 0.3434427  
## trip\_total 0.8304162 0.55062278 0.9450726 0.42029912 0.34344273 1.0000000  
##

library(corrplot)

## corrplot 0.84 loaded

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

xx = rcorr(as.matrix(final.data[,c(2:5,7,8)]))  
xx

## trip\_seconds trip\_miles fare tips extras trip\_total  
## trip\_seconds 1.00 0.52 0.83 0.27 0.18 0.83  
## trip\_miles 0.52 1.00 0.56 0.03 0.35 0.55  
## fare 0.83 0.56 1.00 0.13 0.20 0.95  
## tips 0.27 0.03 0.13 1.00 0.09 0.42  
## extras 0.18 0.35 0.20 0.09 1.00 0.34  
## trip\_total 0.83 0.55 0.95 0.42 0.34 1.00  
##   
## n= 67   
##   
##   
## P  
## trip\_seconds trip\_miles fare tips extras trip\_total  
## trip\_seconds 0.0000 0.0000 0.0266 0.1490 0.0000   
## trip\_miles 0.0000 0.0000 0.7874 0.0033 0.0000   
## fare 0.0000 0.0000 0.2844 0.0992 0.0000   
## tips 0.0266 0.7874 0.2844 0.4642 0.0004   
## extras 0.1490 0.0033 0.0992 0.4642 0.0044   
## trip\_total 0.0000 0.0000 0.0000 0.0004 0.0044

## We use this correlation matrix as a reference to correlation co-efficient between two variables in our data. It measures the strength and direction of the linear relationship between two variables in the scatterplot. The values in the matrix are always between 0 and 1. For a strong correlation, the value should be close 1.

## The trip seconds indicate the duration of the trip. If the trip second is large, customer will be charged more. From the above matrix it’s clear that there is strong positive correlation between fare and trip seconds, trip total and trip seconds and fare and trip total.

## The trip miles is the distance travelled by the customer. Longer the distance is, he will end up paying higher fare. Hence there is a positive correlation between trip miles and fare and trip miles and trip total.

## As the customer travels long distance, the duration also goes up, which is apparent in the positive correlation between trip miles and trip seconds.

## Trip total is calculated by base fare with tips and extras. Hence there should be a strong positive correlation between fare and trip total as indicated in the matrix.

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.

regout=lm(trip\_total~trip\_seconds+trip\_miles+payment\_type,data=final.data)  
summary(regout)

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + payment\_type,   
## data = final.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.6851 -1.0743 -0.1547 0.5169 8.8223   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.821692 0.458569 6.153 5.81e-08 \*\*\*  
## trip\_seconds 0.009428 0.001001 9.422 1.21e-13 \*\*\*  
## trip\_miles 0.677189 0.223161 3.035 0.0035 \*\*   
## payment\_typeCredit Card 2.067538 0.449961 4.595 2.13e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.742 on 63 degrees of freedom  
## Multiple R-squared: 0.7823, Adjusted R-squared: 0.772   
## F-statistic: 75.47 on 3 and 63 DF, p-value: < 2.2e-16

## Regression Equation be like:

## Trip\_Total = $2.821 + 0.009 \* Trip\_seconds + 0.677 \* Trip\_miles + 2.067 \* Credit\_Card

##Considering the statistically significant variables in our model:

According to the P-Value, all the independent variables are statistically significant.

Interpretation :

Intercept : The slope of the Intercept is the expected trip fare when the values of trip seconds, trip miles and payment type are at 0 respectively. Which is not possible practically. when all independent variables in the model are 0 the trip total would be equal to slope of intercept which is $2.821. Here intercept is a significant term as the p value is less than 0.05. Hence we could reject the null hypothesis.

Trip\_Seconds: Trip Seconds has a positive impact on trip total as the coefficient is positive. With each second increase in duration there is an expected increase of $0.009 in the total fare for the customer. Here trip seconds is statistically significant as the p-value is less than 0.05.

Trip\_Miles: Even trip miles has positive impact on total fare. With each mile increase in the trip distance there is an expected increase of $0.677 in total fare. Even this is a statistically significant term as the p value is less than 0.05

Payment Type: Payment type is statistically significant and has positive correlation on the trip fare. When the customer uses credit card as the payment type, there is an expected increase of $2.067 in the total trip fare.

Adjusted R^2: The model explains about 77.2%. which is a good value. There is still 23% variance or unexpectedness using this model.

confint(regout)

## 2.5 % 97.5 %  
## (Intercept) 1.905314973 3.73806908  
## trip\_seconds 0.007428655 0.01142817  
## trip\_miles 0.231237439 1.12314083  
## payment\_typeCredit Card 1.168361439 2.96671403

Interpretation:

Trip\_Seconds : We can conclude 95% of times the trip fare increase is between $0.007 and $0.011 with an increase in one second of trip duration.

Trip\_Miles : We can conclude that 95% of the times the trip fare increase is between $0.23 to $ 1.123 with one-mile increase in the trip distance.

Payment\_Type: We can conclude that 95% of the times the trip fare increase is between $1.168 to $2.966 with the change in payment type to credit card.

Intercept: We can conclude that 95% of times increase in total trip fare is between $1.90 and $3.73. When all independent variables in the model are 0 the total fare would be equal to slope of intercept which is $2.82.

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.

model.all.out=lm(trip\_total~trip\_seconds+trip\_miles+payment\_type+fare+extras,data=final.data)  
summary(model.all.out)

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + payment\_type +   
## fare + extras, data = final.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.00095 -0.08597 0.05021 0.12673 1.54986   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.0880496 0.1564316 0.563 0.57559   
## trip\_seconds 0.0012566 0.0004038 3.112 0.00283 \*\*   
## trip\_miles 0.0507979 0.0653651 0.777 0.44008   
## payment\_typeCredit Card 1.8879247 0.1214621 15.543 < 2e-16 \*\*\*  
## fare 0.9079686 0.0332154 27.336 < 2e-16 \*\*\*  
## extras 0.9693627 0.1224463 7.917 6.12e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4667 on 61 degrees of freedom  
## Multiple R-squared: 0.9849, Adjusted R-squared: 0.9836   
## F-statistic: 794.3 on 5 and 61 DF, p-value: < 2.2e-16

## Using the kitchen sink model for all continuous variables, by taking trip\_total as dependent variable and all other variables as independent, we can see that the adjusted R^2 is at 98.36%. This model has payment type, fare and extras as statistically significant and trip seconds as somewhat significant but the intercept and the trip miles as non-significant as the p-value > 0.05. There is only 1.64% of unexpectedness using this model. There is no considerable difference in using this model over the previous one.

model.fare.squared.out=lm(trip\_total~trip\_seconds+trip\_miles+payment\_type+fare+extras+I(fare^2), data =final.data)  
summary(model.fare.squared.out)

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + payment\_type +   
## fare + extras + I(fare^2), data = final.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.00923 -0.11360 -0.00686 0.12373 1.31700   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.9480496 0.4740702 -2.000 0.0501 .   
## trip\_seconds 0.0007695 0.0004437 1.734 0.0880 .   
## trip\_miles -0.0108236 0.0685881 -0.158 0.8751   
## payment\_typeCredit Card 1.9049926 0.1176125 16.197 < 2e-16 \*\*\*  
## fare 1.1736046 0.1195857 9.814 4.33e-14 \*\*\*  
## extras 0.9984195 0.1189995 8.390 1.06e-11 \*\*\*  
## I(fare^2) -0.0105762 0.0045865 -2.306 0.0246 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.451 on 60 degrees of freedom  
## Multiple R-squared: 0.9861, Adjusted R-squared: 0.9847   
## F-statistic: 709.6 on 6 and 60 DF, p-value: < 2.2e-16

## By taking trip\_total as dependent variable and all other variables as independent along with squared value of fare, we can see that the adjusted R^2 is at 98.47%. This model has payment type, fare and extras as statistically significant and all other variables as non-significant as the p-value > 0.05. There is only 1.53% of unexpectedness using this model.

model.miles.squared.out=lm(trip\_total~trip\_seconds+trip\_miles+payment\_type+fare+extras+I(trip\_miles^2), data =final.data)  
summary(model.miles.squared.out)

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + payment\_type +   
## fare + extras + I(trip\_miles^2), data = final.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.96099 -0.09185 0.01605 0.18000 1.49403   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.3904520 0.2049579 1.905 0.06157 .   
## trip\_seconds 0.0013046 0.0003924 3.325 0.00151 \*\*   
## trip\_miles -0.2287855 0.1422788 -1.608 0.11308   
## payment\_typeCredit Card 1.9362075 0.1198656 16.153 < 2e-16 \*\*\*  
## fare 0.8794406 0.0347444 25.312 < 2e-16 \*\*\*  
## extras 0.8922010 0.1238769 7.202 1.13e-09 \*\*\*  
## I(trip\_miles^2) 0.0884314 0.0402858 2.195 0.03204 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4527 on 60 degrees of freedom  
## Multiple R-squared: 0.986, Adjusted R-squared: 0.9846   
## F-statistic: 704.1 on 6 and 60 DF, p-value: < 2.2e-16

## By taking trip\_total as dependent variable and all other variables as independent along with squared value of trip miles, we can see that the adjusted R^2 is at 98.46%. This model has payment type, fare and extras as statistically significant and trip seconds as somewhat significant and all other variables as non-significant as the p-value > 0.05. There is only 1.54% of unexpectedness using this model. Not much of a difference from using the previous model.

model.seconds.squared.out=lm(trip\_total~trip\_seconds+trip\_miles+payment\_type+fare+extras+I(trip\_seconds^2), data =final.data)  
summary(model.seconds.squared.out)

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + payment\_type +   
## fare + extras + I(trip\_seconds^2), data = final.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.00611 -0.08898 0.05517 0.12846 1.54493   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.020e-02 2.369e-01 0.296 0.768   
## trip\_seconds 1.346e-03 9.765e-04 1.379 0.173   
## trip\_miles 5.063e-02 6.592e-02 0.768 0.445   
## payment\_typeCredit Card 1.889e+00 1.228e-01 15.381 < 2e-16 \*\*\*  
## fare 9.069e-01 3.502e-02 25.894 < 2e-16 \*\*\*  
## extras 9.704e-01 1.239e-01 7.834 9.43e-11 \*\*\*  
## I(trip\_seconds^2) -5.845e-08 5.790e-07 -0.101 0.920   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4705 on 60 degrees of freedom  
## Multiple R-squared: 0.9849, Adjusted R-squared: 0.9834   
## F-statistic: 651.1 on 6 and 60 DF, p-value: < 2.2e-16

## By taking trip\_total as dependent variable and all other variables as independent along with squared value of trip seconds, we can see that the adjusted R^2 is at 98.34%. This model has payment type, fare and extras as statistically significant and all other variables as non-significant as the p-value > 0.05. There is only 1.66% of unexpectedness using this model. Not much of a difference from using the previous model.

model.secondsmiles.squared.out=lm(trip\_total~trip\_seconds+trip\_miles+payment\_type+fare+extras+I(trip\_seconds^2)+I(trip\_miles^2), data =final.data)  
summary(model.secondsmiles.squared.out)

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + payment\_type +   
## fare + extras + I(trip\_seconds^2) + I(trip\_miles^2), data = final.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.98419 -0.09715 0.00827 0.17602 1.46803   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.167e-01 2.546e-01 1.244 0.2184   
## trip\_seconds 1.738e-03 9.617e-04 1.807 0.0759 .   
## trip\_miles -2.409e-01 1.453e-01 -1.658 0.1026   
## payment\_typeCredit Card 1.943e+00 1.213e-01 16.012 < 2e-16 \*\*\*  
## fare 8.733e-01 3.710e-02 23.536 < 2e-16 \*\*\*  
## extras 8.941e-01 1.247e-01 7.168 1.4e-09 \*\*\*  
## I(trip\_seconds^2) -2.812e-07 5.694e-07 -0.494 0.6232   
## I(trip\_miles^2) 9.199e-02 4.118e-02 2.234 0.0293 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4556 on 59 degrees of freedom  
## Multiple R-squared: 0.9861, Adjusted R-squared: 0.9844   
## F-statistic: 596 on 7 and 59 DF, p-value: < 2.2e-16

## By taking trip\_total as dependent variable and all other variables as independent along with squared value of trip seconds and trip miles, we can see that the adjusted R^2 is at 98.44%. This model has payment type, fare and extras as statistically significant and all other variables as non-significant as the p-value > 0.05. There is only 1.56% of unexpectedness using this model. Not much of a difference from using the previous model.

model.faresecondsmiles.squared.out=lm(trip\_total~trip\_seconds+trip\_miles+payment\_type+fare+extras+I(fare^2)+I(trip\_seconds^2)+I(trip\_miles^2), data =final.data)  
summary(model.faresecondsmiles.squared.out)

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + payment\_type +   
## fare + extras + I(fare^2) + I(trip\_seconds^2) + I(trip\_miles^2),   
## data = final.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.87117 -0.14744 0.04535 0.14684 1.18080   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.072e+00 4.508e-01 -2.378 0.020701 \*   
## trip\_seconds -1.138e-03 1.187e-03 -0.959 0.341724   
## trip\_miles -4.382e-01 1.434e-01 -3.055 0.003394 \*\*   
## payment\_typeCredit Card 1.968e+00 1.109e-01 17.749 < 2e-16 \*\*\*  
## fare 1.371e+00 1.424e-01 9.623 1.27e-13 \*\*\*  
## extras 8.992e-01 1.138e-01 7.904 8.87e-11 \*\*\*  
## I(fare^2) -1.924e-02 5.352e-03 -3.594 0.000672 \*\*\*  
## I(trip\_seconds^2) 1.027e-06 6.341e-07 1.619 0.110885   
## I(trip\_miles^2) 1.201e-01 3.837e-02 3.131 0.002726 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4156 on 58 degrees of freedom  
## Multiple R-squared: 0.9886, Adjusted R-squared: 0.987   
## F-statistic: 628.4 on 8 and 58 DF, p-value: < 2.2e-16

## By taking trip\_total as dependent variable and all other variables as independent along with squared value of trip seconds, fare and trip miles, we can see that the adjusted R^2 is at 98.7%. This model has most of the variables as statistically significant and only trip seconds and trip seconds squared variables as non-significant as the p-value > 0.05. There is only 1.3% of unexpectedness using this model. One of the highest R^2 values gotten from all the models. This is considered as the best fit model for my analysis.

model.extras.squared.out=lm(trip\_total~trip\_seconds+trip\_miles+payment\_type+extras+I(extras^2)+tips, data =final.data)  
summary(model.extras.squared.out)

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + payment\_type +   
## extras + I(extras^2) + tips, data = final.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.0920 -0.7639 -0.2327 0.4973 8.5002   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.7794751 0.4655042 5.971 1.37e-07 \*\*\*  
## trip\_seconds 0.0095249 0.0009924 9.598 9.88e-14 \*\*\*  
## trip\_miles 0.4569040 0.2337964 1.954 0.0553 .   
## payment\_typeCredit Card 2.3624001 0.9341659 2.529 0.0141 \*   
## extras -0.2212931 1.2092541 -0.183 0.8554   
## I(extras^2) 1.0234454 0.9326750 1.097 0.2769   
## tips -0.1951697 0.4380063 -0.446 0.6575   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.693 on 60 degrees of freedom  
## Multiple R-squared: 0.8043, Adjusted R-squared: 0.7847   
## F-statistic: 41.09 on 6 and 60 DF, p-value: < 2.2e-16

## By taking trip\_total as dependent variable and all other variables as independent along with squared value of extras, we can see that the adjusted R^2 has dropped down to 78.47%. This model has trip seconds and intercept as statistically significant and all other variables as non-significant as the p-value > 0.05. There is 22% of unexpectedness using this model. Eliminating this model from best fit consideration due to high p value. The removal of fare variable has considerably impacted the model which has rendered most of the variables as non-significant. This implies that trip total is greatly dependent on fare variable along with extras and trip\_seconds.

model.secondsmilesextras.squared.out=lm(trip\_total~trip\_seconds+trip\_miles+payment\_type+extras+I(extras^2)+fare+I(trip\_seconds^2)+I(trip\_miles^2), data =final.data)  
summary(model.secondsmilesextras.squared.out)

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + payment\_type +   
## extras + I(extras^2) + fare + I(trip\_seconds^2) + I(trip\_miles^2),   
## data = final.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.95831 -0.10097 0.01004 0.18121 1.47246   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.248e-01 2.547e-01 1.276 0.20721   
## trip\_seconds 1.801e-03 9.635e-04 1.869 0.06666 .   
## trip\_miles -2.938e-01 1.544e-01 -1.903 0.06201 .   
## payment\_typeCredit Card 1.940e+00 1.213e-01 15.992 < 2e-16 \*\*\*  
## extras 1.221e+00 3.466e-01 3.523 0.00084 \*\*\*  
## I(extras^2) -2.850e-01 2.820e-01 -1.011 0.31629   
## fare 8.712e-01 3.716e-02 23.445 < 2e-16 \*\*\*  
## I(trip\_seconds^2) -3.354e-07 5.718e-07 -0.587 0.55978   
## I(trip\_miles^2) 1.123e-01 4.579e-02 2.451 0.01726 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4555 on 58 degrees of freedom  
## Multiple R-squared: 0.9863, Adjusted R-squared: 0.9844   
## F-statistic: 521.8 on 8 and 58 DF, p-value: < 2.2e-16

## By taking trip\_total as dependent variable and all other variables as independent along with squared value of trip seconds, miles and extras, we can see that the adjusted R^2 is at 98.44%. This model has payment type, fare and extras as statistically significant and all other variables as non-significant as the p-value > 0.05. There is only 1.56% of unexpectedness using this model. Not much of a difference from using the previous models.

model.noextras.out=lm(trip\_total~trip\_seconds+trip\_miles+payment\_type +fare,data=final.data)  
summary(model.noextras.out)

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + payment\_type +   
## fare, data = final.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.3952 -0.3734 -0.1216 0.5363 1.3850   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.183072 0.220285 0.831 0.4091   
## trip\_seconds 0.001145 0.000570 2.009 0.0489 \*   
## trip\_miles 0.211090 0.087778 2.405 0.0192 \*   
## payment\_typeCredit Card 2.004128 0.170290 11.769 <2e-16 \*\*\*  
## fare 0.911987 0.046906 19.443 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6591 on 62 degrees of freedom  
## Multiple R-squared: 0.9693, Adjusted R-squared: 0.9674   
## F-statistic: 489.9 on 4 and 62 DF, p-value: < 2.2e-16

## By taking trip\_total as dependent variable and all other variables as independent except extras, we can see that the adjusted R^2 is at 96.74%. This model has payment type and fare as statistically significant and all other variables as non-significant as the p-value > 0.05. There is only 3.26% of unexpectedness using this model. Not considering this model for best fit selection as we have better models available.

model.all.tips.out=lm(trip\_total~trip\_seconds+trip\_miles+payment\_type +fare+extras+tips, data=final.data)  
summary(model.all.tips.out)

## Warning in summary.lm(model.all.tips.out): essentially perfect fit: summary  
## may be unreliable

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + payment\_type +   
## fare + extras + tips, data = final.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.271e-15 -8.343e-16 5.280e-17 4.013e-16 1.245e-14   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.703e-15 7.700e-16 7.407e+00 5.06e-10 \*\*\*  
## trip\_seconds 4.753e-18 2.134e-18 2.227e+00 0.029688 \*   
## trip\_miles 1.203e-15 3.225e-16 3.731e+00 0.000426 \*\*\*  
## payment\_typeCredit Card 2.638e-15 1.328e-15 1.987e+00 0.051535 .   
## fare 1.000e+00 1.730e-16 5.780e+15 < 2e-16 \*\*\*  
## extras 1.000e+00 6.014e-16 1.663e+15 < 2e-16 \*\*\*  
## tips 1.000e+00 6.286e-16 1.591e+15 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.291e-15 on 60 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 2.788e+31 on 6 and 60 DF, p-value: < 2.2e-16

## By taking trip\_total as dependent variable and all other variables as independent along with tips, we can see that the adjusted R^2 is at 100%. This model most of the variables as statistically significant and trip seconds and payment type as non-significant as the p-value > 0.05. This model is highly unreliable as the adjusted R^2 is at 100%.

model.secondsmilesextrasfare.squared.out=lm(trip\_total~trip\_seconds+trip\_miles+payment\_type+extras+I(extras^2)+fare+I(trip\_seconds^2)+I(trip\_miles^2)+I(fare^2), data =final.data)  
summary(model.secondsmilesextras.squared.out)   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + payment\_type +   
## extras + I(extras^2) + fare + I(trip\_seconds^2) + I(trip\_miles^2),   
## data = final.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.95831 -0.10097 0.01004 0.18121 1.47246   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.248e-01 2.547e-01 1.276 0.20721   
## trip\_seconds 1.801e-03 9.635e-04 1.869 0.06666 .   
## trip\_miles -2.938e-01 1.544e-01 -1.903 0.06201 .   
## payment\_typeCredit Card 1.940e+00 1.213e-01 15.992 < 2e-16 \*\*\*  
## extras 1.221e+00 3.466e-01 3.523 0.00084 \*\*\*  
## I(extras^2) -2.850e-01 2.820e-01 -1.011 0.31629   
## fare 8.712e-01 3.716e-02 23.445 < 2e-16 \*\*\*  
## I(trip\_seconds^2) -3.354e-07 5.718e-07 -0.587 0.55978   
## I(trip\_miles^2) 1.123e-01 4.579e-02 2.451 0.01726 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4555 on 58 degrees of freedom  
## Multiple R-squared: 0.9863, Adjusted R-squared: 0.9844   
## F-statistic: 521.8 on 8 and 58 DF, p-value: < 2.2e-16

## By taking trip\_total as dependent variable and all other variables as independent along with squared value of trip seconds, trip miles and extras we can see that the adjusted R^2 is at 98.44%. This model has payment type, fare and extras as statistically significant and all other variables as non-significant as the p-value > 0.05. There is only 1.56% of unexpectedness using this model. Not much of a difference from using the previous models.

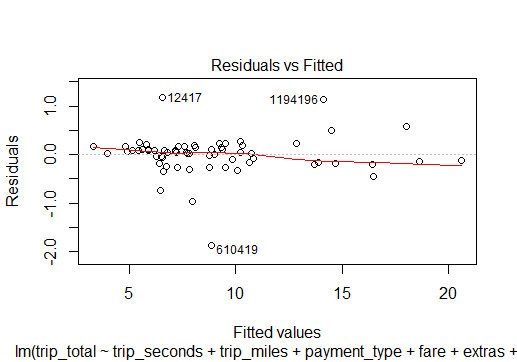
1. Of the various combinations you ran in Step 5, report the model which provides what you deem as the “best fit” to your sample data. Explain why you selected this particular model and show the standard R regression output for the model. Evaluate and explain your model’s conformity to the LINE assumptions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

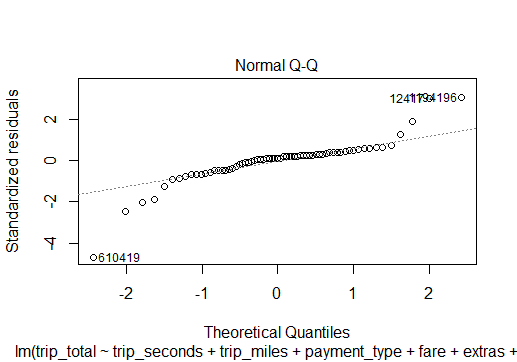
model.faresecondsmiles.squared.out=lm(trip\_total~trip\_seconds+trip\_miles+payment\_type+fare+extras+I(fare^2)+I(trip\_seconds^2)+I(trip\_miles^2), data =final.data)  
summary(model.faresecondsmiles.squared.out)

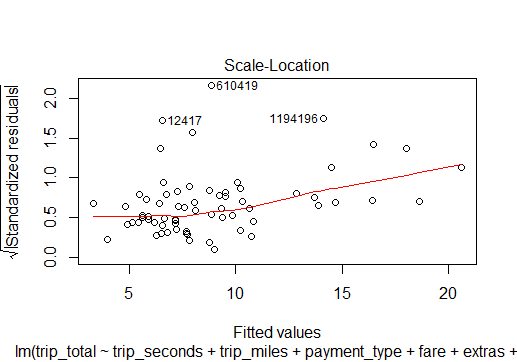
##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + payment\_type +   
## fare + extras + I(fare^2) + I(trip\_seconds^2) + I(trip\_miles^2),   
## data = final.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.87117 -0.14744 0.04535 0.14684 1.18080   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.072e+00 4.508e-01 -2.378 0.020701 \*   
## trip\_seconds -1.138e-03 1.187e-03 -0.959 0.341724   
## trip\_miles -4.382e-01 1.434e-01 -3.055 0.003394 \*\*   
## payment\_typeCredit Card 1.968e+00 1.109e-01 17.749 < 2e-16 \*\*\*  
## fare 1.371e+00 1.424e-01 9.623 1.27e-13 \*\*\*  
## extras 8.992e-01 1.138e-01 7.904 8.87e-11 \*\*\*  
## I(fare^2) -1.924e-02 5.352e-03 -3.594 0.000672 \*\*\*  
## I(trip\_seconds^2) 1.027e-06 6.341e-07 1.619 0.110885   
## I(trip\_miles^2) 1.201e-01 3.837e-02 3.131 0.002726 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4156 on 58 degrees of freedom  
## Multiple R-squared: 0.9886, Adjusted R-squared: 0.987   
## F-statistic: 628.4 on 8 and 58 DF, p-value: < 2.2e-16

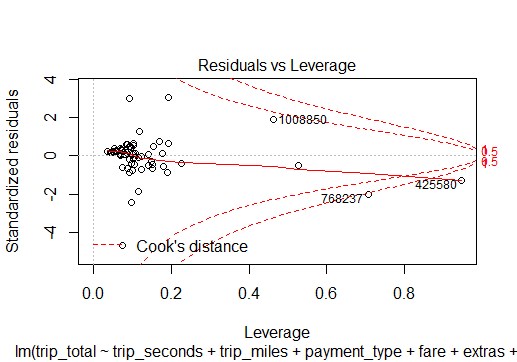
## By taking trip\_total as dependent variable and all other variables as independent along with squared value of trip seconds, fare and trip miles, we can see that the adjusted R^2 is at 98.7%. This model has most of the variables as statistically significant and only intercept, trip seconds and trip seconds squared variables as non-significant as the p-value > 0.05. There is only 1.3% of unexpectedness using this model. One of the highest R^2 values gotten from all the models. This is considered as the best fit model for my analysis.

plot(model.faresecondsmiles.squared.out)

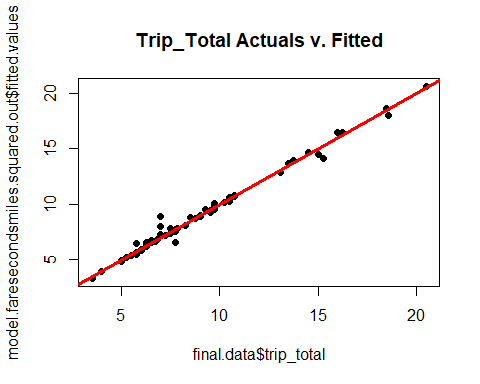




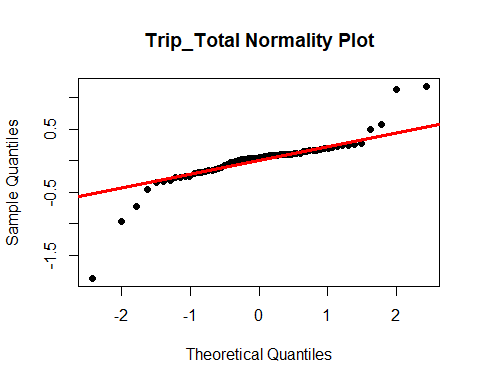




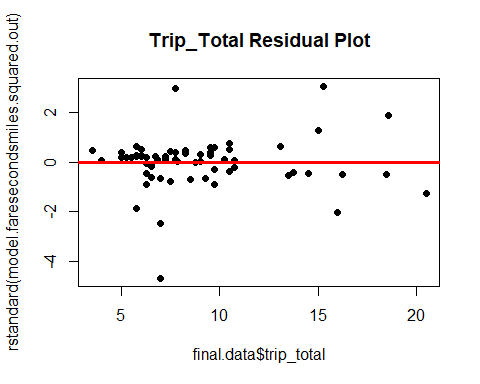
par(mfrow=c(1,1))  
#Linearity  
plot(final.data$trip\_total,model.faresecondsmiles.squared.out$fitted.values,  
 pch=19,main="Trip\_Total Actuals v. Fitted")  
abline(0,1,col="red",lwd=3)



#Normality  
qqnorm(model.faresecondsmiles.squared.out$residuals,pch=19,  
 main="Trip\_Total Normality Plot")  
qqline(model.faresecondsmiles.squared.out$residuals,lwd=3,col="red")



#Equality of Variances  
plot(final.data$trip\_total,rstandard(model.faresecondsmiles.squared.out),  
 pch=19,main="Trip\_Total Residual Plot")  
abline(0,0,col="red",lwd=3)



LINE Assumptions:

Linearity: From the residuals vs fitted values plot, we can see that the observations are somewhat equally distributed around 0, except for a few extreme values (outliers). This model can be called fairly linear. So, this assumption is satisfied.

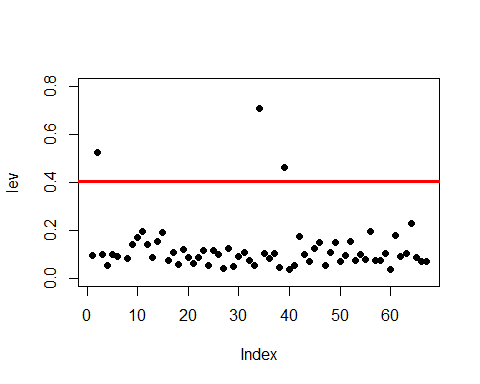
Independence: This is not time series data, so independence assumption is satisfied.

Normality: From the QQ plot, we can see that most of the observations do not fall on the QQ line or are around it. There are a few extreme values (outliers) that deviate away from it. Hence, this assumption is not satisfied, and the model cannot be called normal. Its clear that the data is skewed.

Equality of Variances: From the standard residuals vs fitted values plot, we can see that the observations are scattered all over and are not equally distributed around zero, thus resulting in heteroscedasticity. So, this assumption fails.

7. 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.

lev=hat(model.matrix(model.faresecondsmiles.squared.out))  
plot(lev,pch=19, ylim=c(0,.8))  
abline(3\*mean(lev),0,col="red",lwd=3)



mean(3\*mean(lev))

## [1] 0.4029851

final.data[lev>(3\*mean(lev)),]

## taxi\_id trip\_seconds trip\_miles fare tips tolls extras trip\_total  
## 1054727 6524 960 4.83 16.5 0.0 0 2 18.5  
## 425580 4111 720 0.00 20.5 0.0 0 0 20.5  
## 768237 8028 1440 1.80 13.0 2.0 0 1 16.0  
## 1008850 1669 1320 4.10 15.5 3.1 0 0 18.6  
## payment\_type  
## 1054727 Cash  
## 425580 Credit Card  
## 768237 Credit Card  
## 1008850 Credit Card

##There are the above mentioned four high leverage points that are significantly affecting the model’s regression plot. I try to remove these points and rerun the model to see if anything change.

which(final.data$taxi\_id == 4111)

## [1] 7

which(final.data$taxi\_id == 6524)

## [1] 2

which(final.data$taxi\_id == 8028)

## [1] 34

which(final.data$taxi\_id == 1669)

## [1] 39

red.date = final.data[-34,]  
red.data2 = red.date[-7,]  
red.data3 = red.data2[-2,]  
red.data4 = red.data3[-39,]  
  
model.faresecondsmiles.squared.out=lm(trip\_total~trip\_seconds+trip\_miles+payment\_type+fare+extras+I(fare^2)+I(trip\_seconds^2)+I(trip\_miles^2), data = red.data4)  
summary(model.faresecondsmiles.squared.out)

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + payment\_type +   
## fare + extras + I(fare^2) + I(trip\_seconds^2) + I(trip\_miles^2),   
## data = red.data4)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.83770 -0.07813 0.07327 0.13604 1.19139   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.862e-01 7.508e-01 -1.047 0.300   
## trip\_seconds -1.735e-03 1.729e-03 -1.004 0.320   
## trip\_miles -2.859e-01 1.878e-01 -1.522 0.134   
## payment\_typeCredit Card 1.948e+00 1.133e-01 17.201 < 2e-16 \*\*\*  
## fare 1.291e+00 2.747e-01 4.701 1.83e-05 \*\*\*  
## extras 9.763e-01 1.233e-01 7.919 1.33e-10 \*\*\*  
## I(fare^2) -1.246e-02 1.562e-02 -0.798 0.429   
## I(trip\_seconds^2) 1.566e-06 1.340e-06 1.169 0.248   
## I(trip\_miles^2) 5.388e-02 6.327e-02 0.852 0.398   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4138 on 54 degrees of freedom  
## Multiple R-squared: 0.9842, Adjusted R-squared: 0.9819   
## F-statistic: 420.7 on 8 and 54 DF, p-value: < 2.2e-16

## We can see that the adjusted R^2 has gone down marginally by the removal of these high leverage points. The new R^2 is now at 98.19% and this model still has only 1.81% of unexpectedness. The model has only the payment type, fare and extras as significant and all other variables as non-significant as the p value is >0.05.

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(248895425)  
my.2.data = master.data[sample(1:nrow(master.data),100,replace=FALSE),]  
attach(my.2.data)

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

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

summary(my.2.data)

## taxi\_id trip\_seconds trip\_miles fare   
## Min. : 158 Min. : 0.0 Min. : 0.000 Min. : 3.450   
## 1st Qu.:2350 1st Qu.: 300.0 1st Qu.: 0.100 1st Qu.: 5.938   
## Median :5130 Median : 480.0 Median : 1.060 Median : 8.500   
## Mean :4629 Mean : 792.6 Mean : 3.354 Mean :15.129   
## 3rd Qu.:6666 3rd Qu.: 900.0 3rd Qu.: 2.600 3rd Qu.:13.375   
## Max. :8675 Max. :7500.0 Max. :21.400 Max. :70.250   
## tips tolls extras trip\_total   
## Min. : 0.000 Min. :0 Min. : 0.000 Min. : 3.450   
## 1st Qu.: 0.000 1st Qu.:0 1st Qu.: 0.000 1st Qu.: 6.938   
## Median : 0.000 Median :0 Median : 0.000 Median : 9.375   
## Mean : 2.043 Mean :0 Mean : 1.225 Mean :18.397   
## 3rd Qu.: 2.555 3rd Qu.:0 3rd Qu.: 1.000 3rd Qu.:16.712   
## Max. :13.000 Max. :0 Max. :35.000 Max. :84.750   
## payment\_type  
## Cash :50   
## Credit Card:49   
## Other : 1   
##   
##   
##

Data cleaning Steps: Removing the max outlier in trip total variable

max(my.2.data$trip\_total)

## [1] 84.75

which(my.2.data$trip\_total == 84.75)

## [1] 51

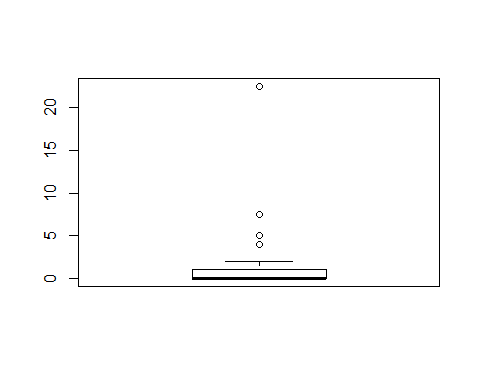
my.2.data = my.data[-51,]

Removing the data rows that has trip seconds as 0 and trip total>0

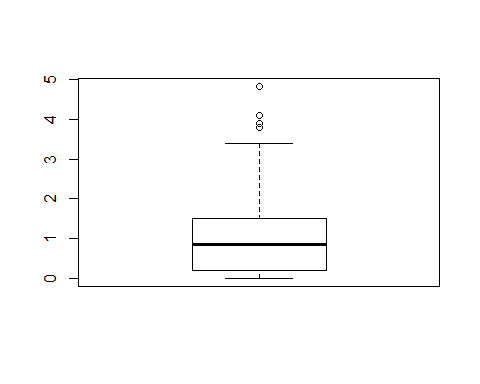
final.2.data = subset(my.2.data, trip\_seconds !=0 )

Removing the outlier in the extras and trip miles variables

outliers\_extras2 = boxplot(final.2.data$extras)$out



final.2.data = final.data[-which(final.2.data$extras %in% outliers\_extras2),]  
outliers\_miles2 = boxplot(final.2.data$trip\_miles)$out



* The second data sampled is cleaned with the underlying assumption that if a passenger gets into the cab, he would travel a few miles for a few seconds, he would pay the fare with tips and applicable extras as total trip fare.
* If the dataset contains rows with trip miles as 0 and trip seconds as 0 but if he has paid some fare, which doesn’t make any sense. Hence I removed the such rows which had the aberrancies. There were 16 of such rows.
* If the samples contained exorbitantly high values in the variables such as extras and miles and trip total, I removed them. There were 9 + 4 + 1= 14 such samples.
* I ended up with having 63 data samples to run my best fit model on.

final.2.data = final.data[-which(final.2.data$trip\_miles %in% outliers\_miles2),]  
  
summary(final.2.data)

## taxi\_id trip\_seconds trip\_miles fare   
## Min. : 6 Min. : 120.0 Min. :0.00 Min. : 3.500   
## 1st Qu.:1674 1st Qu.: 300.0 1st Qu.:0.40 1st Qu.: 5.750   
## Median :3474 Median : 420.0 Median :1.00 Median : 7.000   
## Mean :3599 Mean : 449.5 Mean :1.17 Mean : 7.638   
## 3rd Qu.:5611 3rd Qu.: 540.0 3rd Qu.:1.50 3rd Qu.: 7.925   
## Max. :8409 Max. :1320.0 Max. :4.10 Max. :20.500   
## tips tolls extras trip\_total   
## Min. :0.0000 Min. :0 Min. :0.0000 Min. : 3.500   
## 1st Qu.:0.0000 1st Qu.:0 1st Qu.:0.0000 1st Qu.: 6.250   
## Median :0.0000 Median :0 Median :0.0000 Median : 7.750   
## Mean :0.7294 Mean :0 Mean :0.2937 Mean : 8.661   
## 3rd Qu.:2.0000 3rd Qu.:0 3rd Qu.:1.0000 3rd Qu.: 9.750   
## Max. :3.5000 Max. :0 Max. :1.5000 Max. :20.500   
## payment\_type  
## Cash :39   
## Credit Card:24   
## Other : 0   
##   
##   
##

model.faresecondsmiles.squared.out2=lm(trip\_total~trip\_seconds+trip\_miles+payment\_type+fare+extras+I(fare^2)+I(trip\_seconds^2)+I(trip\_miles^2), data = final.2.data)  
summary(model.faresecondsmiles.squared.out2)

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + payment\_type +   
## fare + extras + I(fare^2) + I(trip\_seconds^2) + I(trip\_miles^2),   
## data = final.2.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.84007 -0.07057 0.07465 0.13935 1.21311   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.139e+00 4.700e-01 -2.422 0.018801 \*   
## trip\_seconds -2.134e-03 1.407e-03 -1.517 0.135140   
## trip\_miles -3.826e-01 1.731e-01 -2.210 0.031338 \*   
## payment\_typeCredit Card 1.947e+00 1.138e-01 17.112 < 2e-16 \*\*\*  
## fare 1.427e+00 1.526e-01 9.354 6.88e-13 \*\*\*  
## extras 9.737e-01 1.234e-01 7.892 1.47e-10 \*\*\*  
## I(fare^2) -2.155e-02 5.719e-03 -3.768 0.000408 \*\*\*  
## I(trip\_seconds^2) 2.057e-06 9.124e-07 2.255 0.028207 \*   
## I(trip\_miles^2) 8.678e-02 5.447e-02 1.593 0.116981   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4135 on 54 degrees of freedom  
## Multiple R-squared: 0.9873, Adjusted R-squared: 0.9854   
## F-statistic: 524.2 on 8 and 54 DF, p-value: < 2.2e-16

## We can see that the adjusted R^2 has gone down marginally by selecting a new set of samples with the new seed. The new R^2 is 98.54% as compared to 98.7% and this model has Payment type, fare and extras as statistically significant and other variables as non-significant as the p-value > 0.05.