QMB Regression Project

Divyang Chetanbhai Mehta – U 5159 2551

10/18/2019

# IMPORTING LIBRARIES

rm(list=ls())  
library(readxl)  
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:base':  
##   
## format.pval, units

library(corrplot)

## corrplot 0.84 loaded

library(e1071)

##   
## Attaching package: 'e1071'

## The following object is masked from 'package:Hmisc':  
##   
## impute

library(ggfortify)  
  
# Here I have loaded three libraries namely, readxl, Hmisc and corrplot   
# All three of the libraries are being used in the assignment

# 

# PREPROCESSING 1

infotaxi = read.csv(file="6304 Regression Project Data.csv", header=TRUE)  
colnames(infotaxi)=tolower(make.names(colnames(infotaxi)))  
attach(infotaxi)  
  
# OBSERVATION OF PREPROSSESING 1-  
# There are approximately 1.7 million records of taxi data.  
# I have imported the data along with the header names of each column.

# PREPROCESSING 2

set.seed(51592551)  
sampletaxi = infotaxi[sample(1:nrow(infotaxi),100,replace=FALSE),]  
  
# OBSERVATIONS OF PREPROCESSING 2:  
# I have set the seed value as the numeric part of the U Number as suggested  
# I have also created a random data set of 100 data points out of 1.7 million

# PREPROCESSING 3

secmileszero = subset(sampletaxi, sampletaxi$trip\_seconds != 0 & sampletaxi$trip\_miles != 0)  
secmileszero$trip\_minutes = secmileszero$trip\_seconds/60  
reordercol <- secmileszero[,c(1,2,10,3,4,5,8,9)]  
sum(is.na(sampletaxi))

## [1] 0

sum(is.na(secmileszero))

## [1] 0

# OBSERVATION OF PROCESSING 3:

# Taken the random data sample 'infotaxi' for data cleaning. Left with 81 records out of 100 after removing all the records that have trip\_seconds = zero and trip\_miles = zero

# Added a column named 'trip\_minutes' for better understanding of trip duration

# Reordered the columns as per requirement

# Checked if the UNCLEANED sample has any NULL values - output is 0 i.e. there are no NULL values in the cleansed sample data

# Checked if the CLEANED sample has any NULL values - output is 0 i.e. there are no NULL values in the cleansed sample data

# There are no null values in the Uncleaned as well as cleaned data sample set

# ANALYSIS 1

# Deciding the continuous variables  
sum(sampletaxi$extras == 0)

## [1] 74

#Summary of reordered cleansed data set  
summary(reordercol)

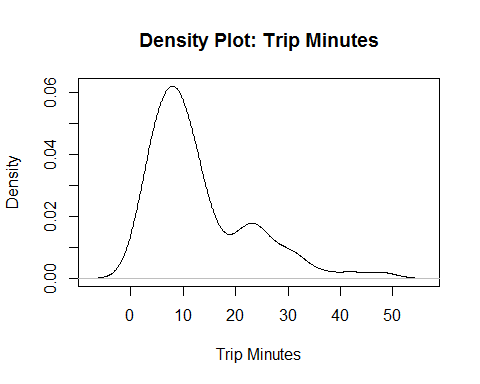
## taxi\_id trip\_seconds trip\_minutes trip\_miles   
## Min. : 173 Min. : 60 Min. : 1 Min. : 0.100   
## 1st Qu.:2005 1st Qu.: 420 1st Qu.: 7 1st Qu.: 1.100   
## Median :3775 Median : 600 Median :10 Median : 1.900   
## Mean :4006 Mean : 780 Mean :13 Mean : 4.395   
## 3rd Qu.:6044 3rd Qu.:1020 3rd Qu.:17 3rd Qu.: 4.500   
## Max. :8736 Max. :2880 Max. :48 Max. :20.700   
## fare tips trip\_total payment\_type  
## Min. : 3.75 Min. : 0.000 Min. : 3.75 Cash :45   
## 1st Qu.: 6.75 1st Qu.: 0.000 1st Qu.: 7.75 Credit Card:36   
## Median : 9.25 Median : 0.000 Median :10.75 Other : 0   
## Mean :15.07 Mean : 1.785 Mean :17.41   
## 3rd Qu.:17.75 3rd Qu.: 2.050 3rd Qu.:18.25   
## Max. :52.75 Max. :12.250 Max. :61.25

# summary of the cleaned data  
# i.e. reordercol showing the minimum, maximum, mean, median, 1st quantile and 3rd quantile

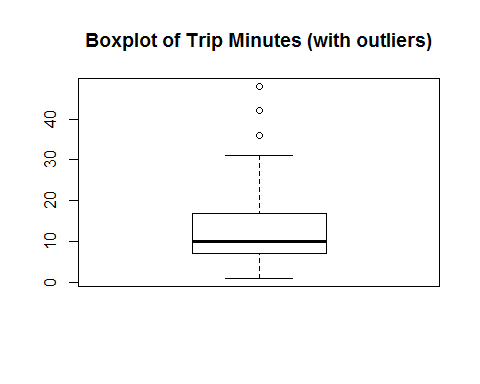
# the Density Plots and Boxplots of different continuous variables have been plotted along with the interquantile range, summary, skewness and kurtosis of the attributes. Also, the points before Q1 and points after Q3 are calculated and it's values has been shown.

## # Continuous Variable 01: trip minutes (for better understanding)

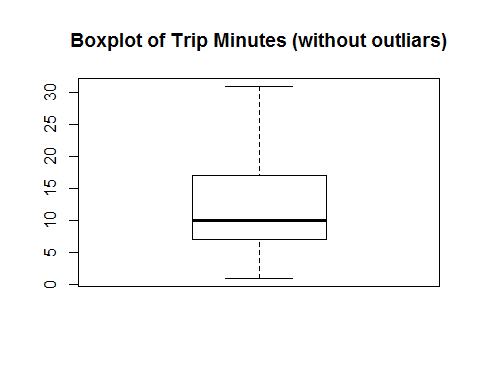
plot(density(reordercol$trip\_minutes),main="Density Plot: Trip Minutes", xlab = "Trip Minutes")



boxplot(reordercol$trip\_minutes, data=reordercol, main = "Boxplot of Trip Minutes (with outliers)")



boxplot(reordercol$trip\_minutes, data=reordercol, outline = FALSE, main = "Boxplot of Trip Minutes (without outliars)")



quantile(reordercol$trip\_minutes, probs = c(0.25, 0.75))

## 25% 75%   
## 7 17

iqrmin = IQR(reordercol$trip\_minutes)  
iqrmin

## [1] 10

minbefore = (1.5\*(iqrmin) - quantile(reordercol$trip\_minutes, 0.25))  
minbefore

## 25%   
## 8

which(reordercol$trip\_minutes < minbefore)

## [1] 1 4 5 8 10 16 17 19 23 27 29 38 39 40 42 44 49 50 52 54 57 58 61  
## [24] 66 70 73

minafter = (1.5\*(iqrmin) + quantile(reordercol$trip\_minutes, 0.75))  
minafter

## 75%   
## 32

which(reordercol$trip\_minutes > minafter)

## [1] 41 63 65

summary(reordercol$trip\_minutes)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1 7 10 13 17 48

skewness(reordercol$trip\_minutes)

## [1] 1.372383

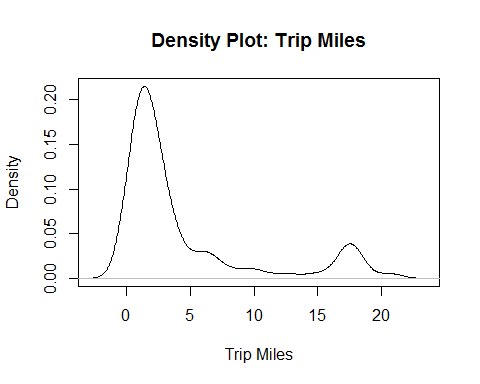
kurtosis(reordercol$trip\_minutes)

## [1] 1.614715

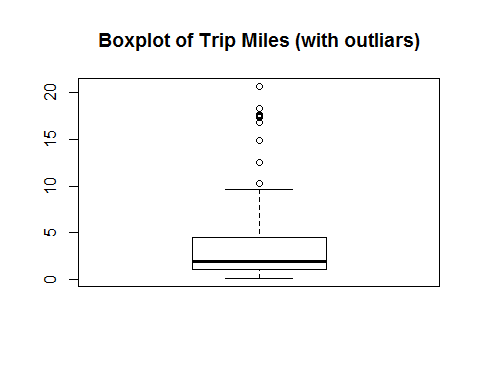
### # OBSERVATION CV 01:

# The density plot of trip minutes indicates that the density is highest at trip minutes equal to 10 minutes (600 seconds)  
# The Boxplot of trip minutes (with outliers) shows three outliers after trip minutes equal to 32 minutes (1980 seconds) with maximum at 48 min, minimum at 1 min with the median value equal to 10 min  
# The Boxplot of trip minutes (without outliers) shows maximuum at 32 min, minimum at 1 min with the median value equal to 10 min  
# the interquantile range is 10 for trip minutes with Q1 at 7 min and Q3 at 17 min  
# In this plot, there are total 29 outliers and they lie before 8 min (26 records) and after 32 minutes (3 records)  
# the skewness of 'trip\_minutes' is 1.3982 i.e. the 'trip\_minutes' is skewed positively high  
# the kurtosis of 'trip\_minutes' is 4.7308 i.e. the 'trip\_minutes' is platykurtic in nature

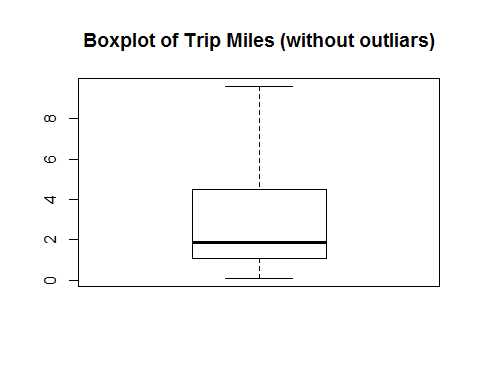
## # Continuous Variable 02: trip miles plot(density(reordercol$trip\_miles),main="Density Plot: Trip Miles", xlab = "Trip Miles")



boxplot(reordercol$trip\_miles, data=reordercol, main = "Boxplot of Trip Miles (with outliars)")



boxplot(reordercol$trip\_miles, data=reordercol, outline = FALSE, main = "Boxplot of Trip Miles (without outliars)")



quantile(reordercol$trip\_miles, probs = c(0.25, 0.75))

## 25% 75%   
## 1.1 4.5

iqrmile = IQR(reordercol$trip\_miles)  
iqrmile

## [1] 3.4

milebefore = (1.5\*(iqrmile) - quantile(reordercol$trip\_miles, 0.25))  
milebefore

## 25%   
## 4

which(reordercol$trip\_miles < milebefore)

## [1] 1 3 4 5 6 8 9 10 11 12 13 16 17 18 19 20 21 23 24 25 27 28 29  
## [24] 31 32 33 34 35 38 39 40 42 44 45 47 48 49 50 51 52 53 54 56 57 58 59  
## [47] 60 61 62 66 67 68 69 70 72 73 74 75 77 81

mileafter = (1.5\*(iqrmile) + quantile(reordercol$trip\_miles, 0.75))  
mileafter

## 75%   
## 9.6

which(reordercol$trip\_miles > mileafter)

## [1] 7 14 15 37 41 43 55 63 64 65 71 79

summary(reordercol$trip\_miles)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.100 1.100 1.900 4.395 4.500 20.700

skewness(reordercol$trip\_miles)

## [1] 1.678078

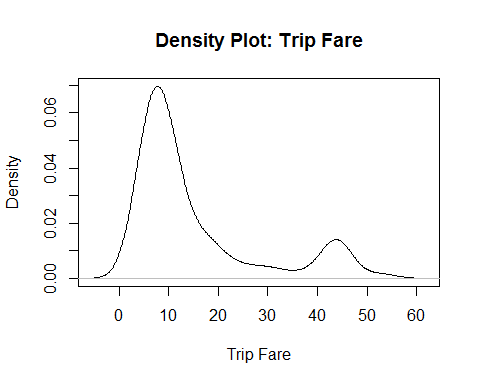
kurtosis(reordercol$trip\_miles)

## [1] 1.437553

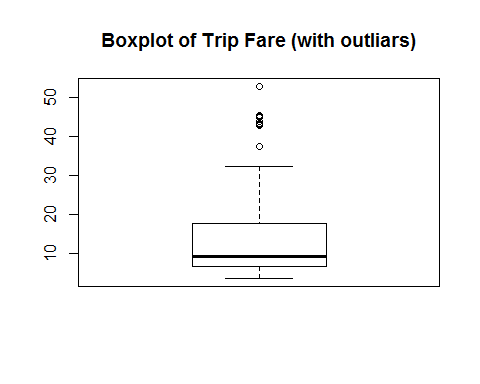
### # OBSERVATION CV 02:

# The density plot of trip miles indicates that the density is highest at trip miles equal to approximately 2 miles and then it increases again at 17 miles.  
# The Boxplot of trip miles (with outliers) shows a few outliers after trip miles equal to 9.6 miles with maximum at 20.7 miles and minimum at 0.1 miles.  
# The Boxplot of trip miles (without outliers) shows maximuum at 11 miles, minimum at 0.1 min with the median value equal to 1.9 min  
# the interquantile range is 3.4 for trip miles with Q1 at 1.1 miles and Q3 at 4.5 miles  
# In this plot, there are total 72 outliers and they lie before 4 miles (60 records) and after 9.6 miles (12 record)  
# the skewness of 'trip\_miles' is 1.7096 i.e. the 'trip\_minutes' is skewed positively high  
# the kurtosis of 'trip\_miles' is 4.5492 i.e. the 'trip\_minutes' is platykurtic in nature

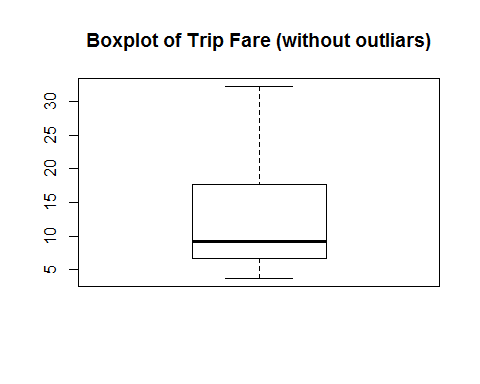
## # Continuous Variable 03: trip fare plot(density(reordercol$fare),main="Density Plot: Trip Fare", xlab = "Trip Fare")



boxplot(reordercol$fare, data=reordercol, main = "Boxplot of Trip Fare (with outliars)")



boxplot(reordercol$fare, data=reordercol, outline = FALSE, main = "Boxplot of Trip Fare (without outliars)")



quantile(reordercol$fare, probs = c(0.25, 0.75))

## 25% 75%   
## 6.75 17.75

iqrfare = IQR(reordercol$fare)  
iqrfare

## [1] 11

farebefore = (1.5\*(iqrfare) - quantile(reordercol$fare, 0.25))  
farebefore

## 25%   
## 9.75

which(reordercol$fare < farebefore)

## [1] 1 4 5 6 8 9 10 13 16 17 19 23 24 25 27 29 32 33 34 38 39 40 42  
## [24] 44 45 49 50 51 52 53 54 56 57 58 61 66 68 69 70 72 73 74 77

fareafter = (1.5\*(iqrfare) + quantile(reordercol$fare, 0.75))  
fareafter

## 75%   
## 34.25

which(reordercol$fare > fareafter)

## [1] 7 15 37 41 43 55 63 65 71 79 81

summary(reordercol$fare)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.75 6.75 9.25 15.07 17.75 52.75

skewness(reordercol$fare)

## [1] 1.496163

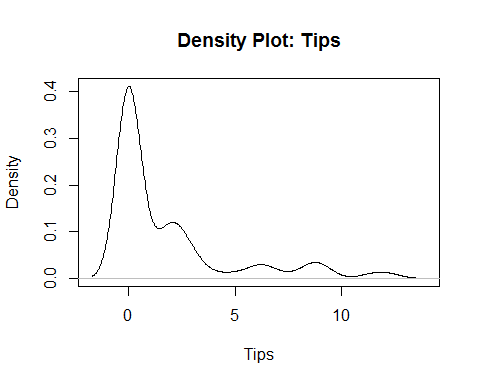
kurtosis(reordercol$fare)

## [1] 0.8837799

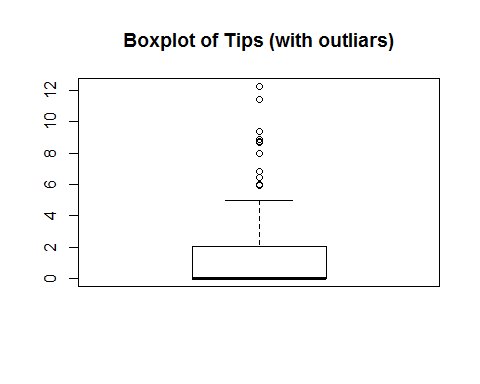
### # OBSERVATION CV 03:

# The density plot of trip fare indicates that the density is highest at trip fare equal to $10 and then it increases again at $43  
# The Boxplot of trip fare (with outliers) shows a few outliers after trip fare equal to $34.5 with maximum at $52.75 and minimum at $3.75  
# The Boxplot of trip fare (without outliers) shows maximuum at $34.25, minimum at $3.75 with the median value equal to $9.25  
# the interquantile range is 11 for trip miles with Q1 at $6.75 and Q3 at $17.75  
# There are total 54 outliers and they lie before $9.75 (43 records) and after $34.25 (11 records)  
# the skewness of 'fare' is 1.5243 i.e. the 'trip\_minutes' is skewed positively high  
# the kurtosis of 'fare' is 3.9815 i.e. the 'trip\_minutes' is platykurtic in nature

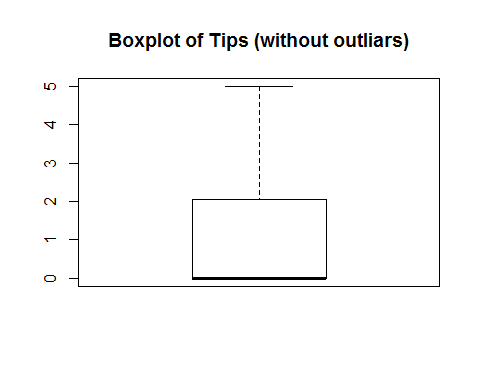
## # Continuous Variable 04: tips plot(density(reordercol$tips),main="Density Plot: Tips", xlab = "Tips")



boxplot(reordercol$tips, data=reordercol, main = "Boxplot of Tips (with outliars)")



boxplot(reordercol$tips, data=reordercol, outline = FALSE, main = "Boxplot of Tips (without outliars)")



quantile(reordercol$tips, probs = c(0.25, 0.75))

## 25% 75%   
## 0.00 2.05

iqrtip = IQR(reordercol$tips)  
iqrtip

## [1] 2.05

tipbefore = (1.5\*(iqrtip) - quantile(reordercol$tips, 0.25))  
tipbefore

## 25%   
## 3.075

which(reordercol$tip < iqrtip)

## [1] 1 2 4 6 7 8 9 10 11 13 15 16 17 18 19 21 22 23 24 25 26 27 28  
## [24] 29 31 32 33 35 36 38 39 40 42 44 45 46 47 48 49 50 51 52 53 54 57 58  
## [47] 61 62 65 66 68 69 70 72 73 74 75 76 77 80

tipafter = (1.5\*(iqrtip) + quantile(reordercol$tips, 0.75))  
tipafter

## 75%   
## 5.125

which(reordercol$tip > iqrtip)

## [1] 3 5 14 20 30 34 37 41 43 55 56 59 60 63 64 67 71 78 79 81

summary(reordercol$tips)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 1.785 2.050 12.250

skewness(reordercol$tips)

## [1] 1.849414

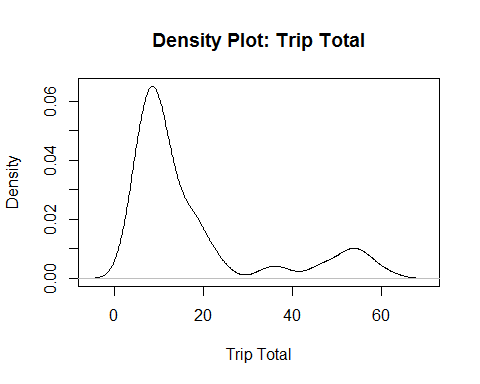
kurtosis(reordercol$tips)

## [1] 2.592417

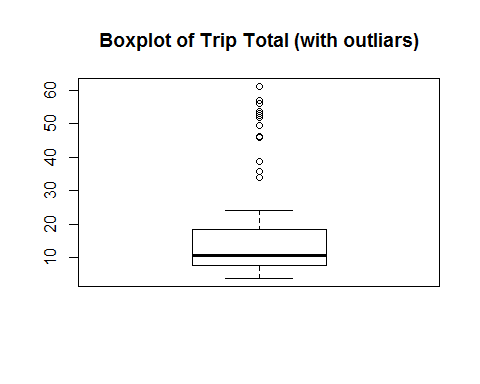
### # OBSERVATION CV 04:

# The density plot of tips indicates that the density is highest at tips equal to $0 and density decreases with increase in amount of tips  
# The Boxplot of tips (with outliers) shows a few outliers after tip equal to $5.125 with maximum at $12.25 and minimum at $0  
# The Boxplot of trip fare (without outliers) shows maximuum at $5.125, minimum at $0 with the median value equal to $0  
# the interquantile range is 2.05 for trip miles with Q1 at $0 and Q3 at $2.05  
# There are total 80 outliers and they lie before $3.075 (60 records) and after $5.125 (20 records)  
# the skewness of 'tips' is 1.8842 i.e. the 'trip\_minutes' is skewed positively high  
# the kurtosis of 'tips' is 5.7331 i.e. the 'trip\_minutes' is platykurtic in nature

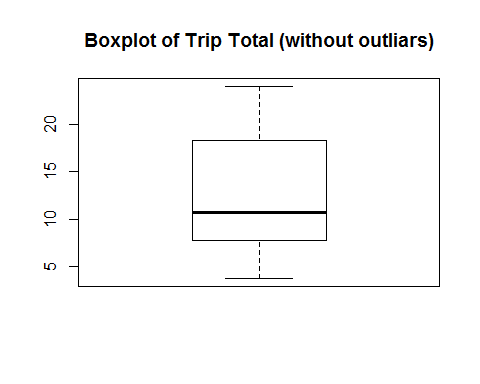
## # Continuous Variable 05 - trip total plot(density(reordercol$trip\_total),main="Density Plot: Trip Total", xlab = "Trip Total")



boxplot(reordercol$trip\_total, data=reordercol, main = "Boxplot of Trip Total (with outliars)")



boxplot(reordercol$trip\_total, data=reordercol, outline = FALSE, main = "Boxplot of Trip Total (without outliars)")



quantile(reordercol$trip\_total, probs = c(0.25, 0.75))

## 25% 75%   
## 7.75 18.25

iqrtriptotal = IQR(reordercol$trip\_total)  
iqrtriptotal

## [1] 10.5

triptotalbefore = (1.5\*(iqrtriptotal) - quantile(reordercol$trip\_total, 0.25))  
triptotalbefore

## 25%   
## 8

which(reordercol$trip\_total < iqrtriptotal)

## [1] 1 4 5 6 8 9 10 13 16 17 19 21 23 24 25 27 29 32 38 39 40 42 45  
## [24] 49 50 51 52 53 54 57 58 61 66 68 69 70 72 73 77

triptotalafter = (1.5\*(iqrtriptotal) + quantile(reordercol$tips, 0.75))  
triptotalafter

## 75%   
## 17.8

which(reordercol$trip\_total > iqrtriptotal)

## [1] 2 3 7 11 12 14 15 18 20 22 26 28 30 31 33 34 35 36 37 41 43 44 46  
## [24] 47 48 55 56 59 60 62 63 64 65 67 71 74 75 76 78 79 80 81

summary(reordercol$trip\_total)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.75 7.75 10.75 17.41 18.25 61.25

skewness(reordercol$trip\_total)

## [1] 1.56633

kurtosis(reordercol$trip\_total)

## [1] 1.065356

### # OBSERVATION CV 05 –

# The density plot of trip total indicates that the density is highest at total equal to ~ $10 and density decreases with increase in amount of total fare. Although, there is a bump in the plot at total fare equal to ~ $57  
# The Boxplot of trip total (with outliers) shows a few outliers after trip total equal to $23 with maximum at $61.25 and minimum at $3.75  
# The Boxplot of trip fare (without outliers) shows maximuum at $23, minimum at $3.75 with the median value equal to $10.75  
# the interquantile range is 10.5 for trip miles with Q1 at $7.75 and Q3 at $18.25  
# the skewness of 'trip\_total' is 1.5958 i.e. the 'trip\_minutes' is skewed positively high  
# the kurtosis of 'trip\_total' is 4.1676 i.e. the 'trip\_minutes' is platykurtic in nature

# 

# ANALYSIS 2

#creating a factor of payment type and getting the count of different types of payment

factdata = as.factor(reordercol$payment\_type)  
table(factdata)

## factdata  
## Cash Credit Card Other   
## 45 36 0

## # OBSERVATIONS ANALYSIS 2:

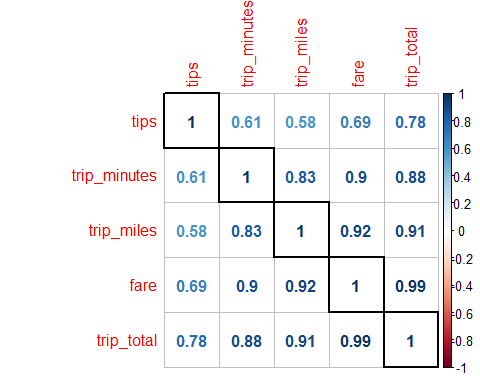
# A table of payment type has been generated using the factor of the payment type attribute from the sample data set  
# the table shows that out of 81 records, 45 have been paid through cash, 36 have been paid through credit cards and  
# none of them have been paid via any other mode of payment

# 

# ANALYSIS 3

# Correlation matrix using all the independent variable

updatereordercol <- secmileszero[,c(10,3,4,5,8,9)]  
correlation = cor(updatereordercol[sapply(updatereordercol, is.numeric)])  
corrplot(correlation, method ="number", order = "hclust", addrect = 5)



## # OBSERVATIONS ANALYSIS 3:

# All the above correlations are positive in nature

# With increase in fare, the total trip fair also increases irrespective of the fact if the tip is given or not, hence the almost exact correlation value between 'total\_trip' and 'fare' = 0.99 justifies the positive correlation

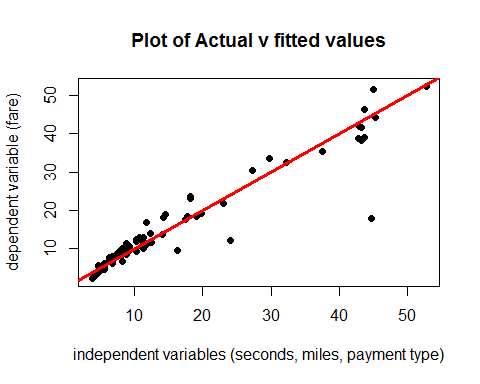
# As the distance covered increases, the fare for the trip also increases, this justifies the positive correlation = 0.92 between 'trip\_miles' and 'fare'

# As the distance covered increases, the total trip fare also increases, this justifies the positive correlation = 0.91 between 'trip\_total' and 'trip\_miles'

# With increase in the duration of the trip i.e. 'trip\_minutes', the 'total\_trip' fare also increases and there is a positive correlation between them with the value of 0.88

# ANALYSIS 4

multireg = lm(reordercol$fare~reordercol$trip\_seconds + reordercol$trip\_miles + reordercol$payment\_type)  
plot(reordercol$fare, multireg$fitted.values, pch = 19, main = "Plot of Actual v fitted values", xlab = "independent variables (seconds, miles, payment type)", ylab = "dependent variable (fare)")  
abline(0,1,col="red",lwd=3)



# Above is the plot of linear model where 'fare' is the dependent variable and the independent variables are 'trip\_seconds', 'trip\_miles' and 'payment\_type'  
# The plot shows how the data points are heteroscadastic in nature i.e. as the independent variables increase

summary(multireg)

##   
## Call:  
## lm(formula = reordercol$fare ~ reordercol$trip\_seconds + reordercol$trip\_miles +   
## reordercol$payment\_type)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.6488 -1.4015 -0.5917 0.7323 26.8123   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.490843 0.846260 1.762 0.0821  
## reordercol$trip\_seconds 0.009188 0.001405 6.539 6.12e-09  
## reordercol$trip\_miles 1.339591 0.146601 9.138 6.51e-14  
## reordercol$payment\_typeCredit Card 1.191125 0.927014 1.285 0.2027  
##   
## (Intercept) .   
## reordercol$trip\_seconds \*\*\*  
## reordercol$trip\_miles \*\*\*  
## reordercol$payment\_typeCredit Card   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.017 on 77 degrees of freedom  
## Multiple R-squared: 0.9077, Adjusted R-squared: 0.9041   
## F-statistic: 252.3 on 3 and 77 DF, p-value: < 2.2e-16

# The fit of the model is ~ 0.91 i.e. 91% which is a decent fit for the given model.  
# the intercept for 'trip\_seconds' is 0.0092 seconds and the p value is significant  
# the intercept for 'trip\_miles' is 1.3396 miles and the p value is significant  
# the intercept for 'payment\_type' is 1.19 and the p value is 0.2027  
# the p values for very low for both 'trip\_seconds' and 'trip\_miles' where as it is 0.2027 for 'payment\_type'

confint(multireg, level = 0.9)

## 5 % 95 %  
## (Intercept) 0.081917199 2.89976819  
## reordercol$trip\_seconds 0.006848651 0.01152753  
## reordercol$trip\_miles 1.095517237 1.58366443  
## reordercol$payment\_typeCredit Card -0.352245750 2.73449574

# the confidence interval for the model is 2.8998  
# the confidence interval for the attribute 'trip\_seconds' is (0.0115 - 0.0068) = 0.0047 seconds  
# the confidence interval for the attribute 'trip\_miles' is (1.5837 - 1.0955) = 0.4822 miles

## # OBSERVATION ANALYSIS 4:

# The tests for Linearity, Normality and Equality of Variances has been performed and the observations have been written above.

# 

# ANALYSIS 5: Investigating relevant interactions and common independent variables transforms

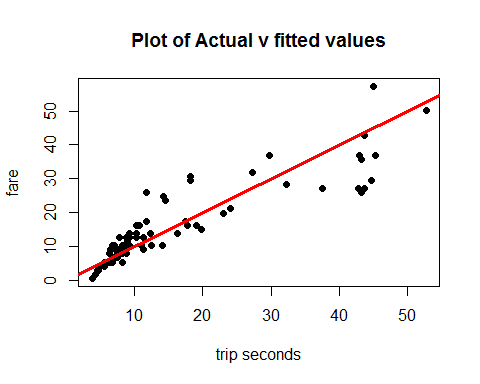
# Investigating relevant interactions and common independent variables transforms

# The columns that are not needed in further analysis have been removed  
# i.e. the attributes 'taxi\_id', 'trip\_minutes' and 'payment\_type' have been removed from the data set  
cleanreorder = reordercol[,c(2,4,5,6,7,8)]

## # A5.01 only trip seconds sec = lm(cleanreorder$fare~cleanreorder$trip\_seconds) summary(sec)

##   
## Call:  
## lm(formula = cleanreorder$fare ~ cleanreorder$trip\_seconds)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.1829 -2.4553 0.1641 2.2512 17.3171   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.60938 1.08493 -0.562 0.576   
## cleanreorder$trip\_seconds 0.02011 0.00112 17.953 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.79 on 79 degrees of freedom  
## Multiple R-squared: 0.8031, Adjusted R-squared: 0.8006   
## F-statistic: 322.3 on 1 and 79 DF, p-value: < 2.2e-16

plot(cleanreorder$fare, sec$fitted.values, pch = 19, main = "Plot of Actual v fitted values", xlab = "fitted values: trip\_seconds", ylab = "fare")  
abline(0,1,col="red",lwd=3)



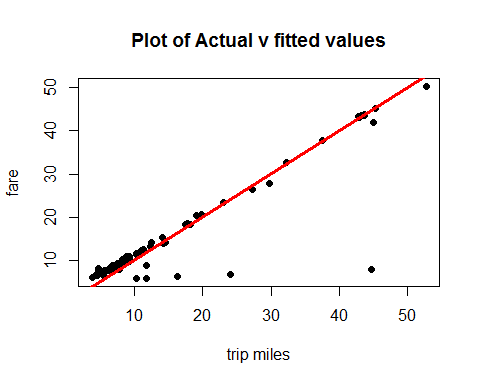
### # OBSERVATIONS A5.01:

# R-square: 0.8031 (Not Very High)  
# p-value: 0.576 which is much higher than 0.05. Hence, we fail to reject the Null Hypothesis.  
# p-value of trip\_seconds: 2e-16 which is significantly lower than 0.05. Hence we reject the Null Hypothesis.  
# line equation: fare = -0.60938 + 0.02011 \* trip\_seconds  
# the line equation indicates that with unit increase in 'trip\_seconds' the fare increases by $0.02011

## # A5.02 only trip miles miles = lm(cleanreorder$fare~cleanreorder$trip\_miles) summary(miles)

##   
## Call:  
## lm(formula = cleanreorder$fare ~ cleanreorder$trip\_miles)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.513 -1.562 -1.110 -0.305 36.803   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.5671 0.7129 7.809 2.05e-11 \*\*\*  
## cleanreorder$trip\_miles 2.1632 0.1013 21.362 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.013 on 79 degrees of freedom  
## Multiple R-squared: 0.8524, Adjusted R-squared: 0.8506   
## F-statistic: 456.4 on 1 and 79 DF, p-value: < 2.2e-16

plot(cleanreorder$fare, miles$fitted.values, pch = 19, main = "Plot of Actual v fitted values", xlab = "fitted values: trip miles", ylab = "fare")  
abline(0,1,col="red",lwd=3)



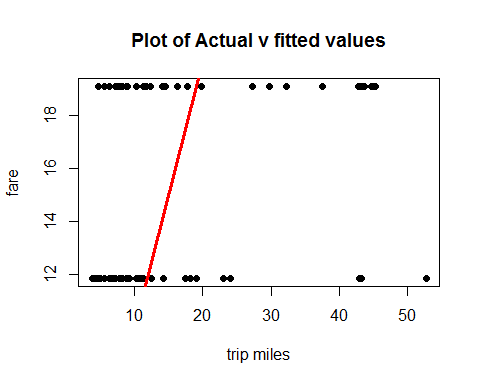
### # OBSERVATIONS A5.02:

# R-square: 0.8524 (Not Very High)  
# p-value: 2.05e-11. significantly lower than 0.05. Hence, we reject the Null Hypothesis.  
# p-value of trip\_miles: 2e-16. significantly lower than 0.05. Hence, we reject the Null Hypothesis.  
# line equation: fare = 5.5671 + 2.1632 \* trip\_miles  
# the line equation indicates that with unit increase in 'trip\_miles' the fare increases by $2.1632

## # A5.03 only trip payment pay = lm(cleanreorder$fare~cleanreorder$payment\_type) summary(pay)

##   
## Call:  
## lm(formula = cleanreorder$fare ~ cleanreorder$payment\_type)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.333 -7.368 -4.583 2.382 40.882   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 11.868 1.868 6.352 1.26e-08  
## cleanreorder$payment\_typeCredit Card 7.216 2.803 2.575 0.0119  
##   
## (Intercept) \*\*\*  
## cleanreorder$payment\_typeCredit Card \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.53 on 79 degrees of freedom  
## Multiple R-squared: 0.07741, Adjusted R-squared: 0.06573   
## F-statistic: 6.629 on 1 and 79 DF, p-value: 0.0119

plot(cleanreorder$fare, pay$fitted.values, pch = 19, main = "Plot of Actual v fitted values", xlab = "trip miles", ylab = "fare")  
abline(0,1,col="red",lwd=3)



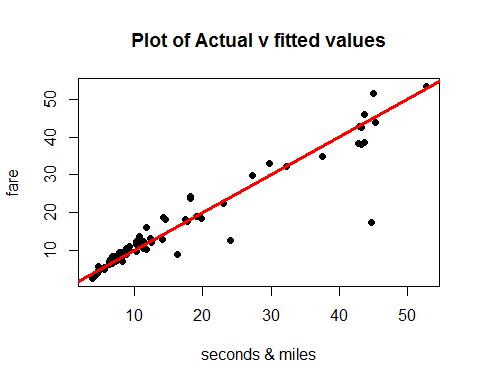
### # OBSERVATIONS A5.03:

# R-square: 0.07741 (very low)  
# p-value: 1.26e-08 which is significantly lower than 0.05. Hence, we reject the Null Hypothesis.  
# p-value of payment\_type: 0.0119 which close to 0.05. Hence, conclusion of rejecting the null hypothesis or faiting to reject the Null Hypothesis depends on the reader.  
# line equation: fare = 11.868 + 7.216 \* payment\_type

## # A5.04 Seconds and Miles milsec = lm(cleanreorder$fare~cleanreorder$trip\_seconds + cleanreorder$trip\_miles) summary(milsec)

##   
## Call:  
## lm(formula = cleanreorder$fare ~ cleanreorder$trip\_seconds +   
## cleanreorder$trip\_miles)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.4484 -1.2909 -0.2054 0.4241 27.4125   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.851575 0.801659 2.310 0.0236 \*   
## cleanreorder$trip\_seconds 0.009332 0.001407 6.635 3.87e-09 \*\*\*  
## cleanreorder$trip\_miles 1.352386 0.146871 9.208 4.26e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.033 on 78 degrees of freedom  
## Multiple R-squared: 0.9057, Adjusted R-squared: 0.9033   
## F-statistic: 374.5 on 2 and 78 DF, p-value: < 2.2e-16

plot(cleanreorder$fare, milsec$fitted.values, pch = 19, main = "Plot of Actual v fitted values", xlab = "seconds & miles", ylab = "fare")  
abline(0,1,col="red",lwd=3)



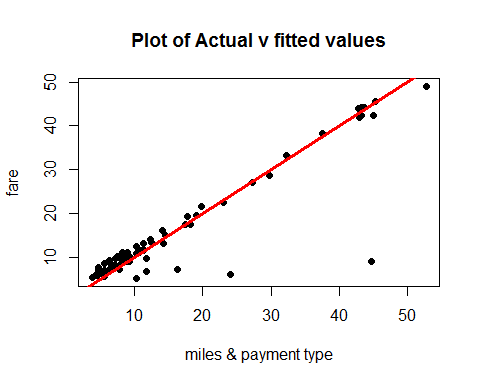
### # OBSERVATIONS A5.04:

# R-square: 0.9057 (Decently High)  
# p-value: 0.0236 which is close to 0.05 and not very significant. Hence, we fail to reject the Null Hypothesis  
# p-value of trip\_seconds: 3.87e-09 which is significantly lower than 0.05. Hence, we reject the Null Hypothesis for 'trip\_seconds'  
# p-value of trip\_miles: 4.26e-14 which is significantly lower than 0.05. Hence, we reject the Null Hypothesis for 'trip\_miles'  
# line equation: fare = 1.8516 + 0.0093 \* trip\_seconds + 1.3524 \* trip\_miles  
# the line equation indicates that with unit increase in 'trip\_miles' the fare increases by $1.3524 and with a unit increase in 'trip\_seconds', the fare increases by $0.0093

## # A5.05 miles and payment types paymil = lm(cleanreorder$fare~cleanreorder$trip\_miles + cleanreorder$payment\_type) summary(paymil)

##   
## Call:  
## lm(formula = cleanreorder$fare ~ cleanreorder$trip\_miles + cleanreorder$payment\_type)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.069 -1.851 -0.835 -0.272 35.756   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 4.9792 0.8140 6.117 3.56e-08  
## cleanreorder$trip\_miles 2.1276 0.1034 20.568 < 2e-16  
## cleanreorder$payment\_typeCredit Card 1.6750 1.1450 1.463 0.148  
##   
## (Intercept) \*\*\*  
## cleanreorder$trip\_miles \*\*\*  
## cleanreorder$payment\_typeCredit Card   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.977 on 78 degrees of freedom  
## Multiple R-squared: 0.8564, Adjusted R-squared: 0.8527   
## F-statistic: 232.5 on 2 and 78 DF, p-value: < 2.2e-16

plot(cleanreorder$fare, paymil$fitted.values, pch = 19, main = "Plot of Actual v fitted values", xlab = "miles & payment type", ylab = "fare")  
abline(0,1,col="red",lwd=3)



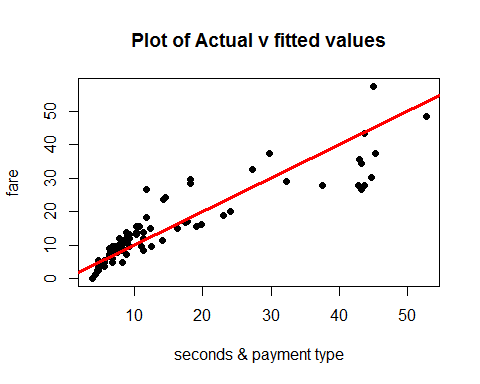
### # OBSERVATIONS A5.05:

# R-square: 0.8564 (Not Very High)  
# p-value: 3.56e-08 which is significanly lower than 0.05. Hence, reject the Null Hypothesis.  
# p-value of trip\_miles: 2e-16 which is significantly less than 0.05. Hence, we reject the Null Hypothesis for 'trip\_miles'  
# p-value of payment\_type: 0.148 which is higher than 0.05. Hence, we fail to reject the Null Hypothesis for 'payment\_method'  
# line equation: fare = 4.9792 + 2.1276 \* trip\_miles + 1.3524 \* payment\_type  
# the line equation indicates that with unit increase in 'trip\_miles' the fare increases by $2.1276 and depending on the type of payment increase in 'payment\_type', the fare increases by $0.0093

## # A5.06 payment types and trip seconds paysec = lm(cleanreorder$fare~cleanreorder$trip\_seconds + cleanreorder$payment\_type) summary(paysec)

##   
## Call:  
## lm(formula = cleanreorder$fare ~ cleanreorder$trip\_seconds +   
## cleanreorder$payment\_type)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.9672 -2.6987 -0.3013 2.6822 16.5328   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -1.109845 1.143208 -0.971 0.335  
## cleanreorder$trip\_seconds 0.019743 0.001148 17.200 <2e-16  
## cleanreorder$payment\_typeCredit Card 1.766533 1.326686 1.332 0.187  
##   
## (Intercept)   
## cleanreorder$trip\_seconds \*\*\*  
## cleanreorder$payment\_typeCredit Card   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.761 on 78 degrees of freedom  
## Multiple R-squared: 0.8075, Adjusted R-squared: 0.8026   
## F-statistic: 163.6 on 2 and 78 DF, p-value: < 2.2e-16

plot(cleanreorder$fare, paysec$fitted.values, pch = 19, main = "Plot of Actual v fitted values", xlab = "seconds & payment type", ylab = "fare")  
abline(0,1,col="red",lwd=3)



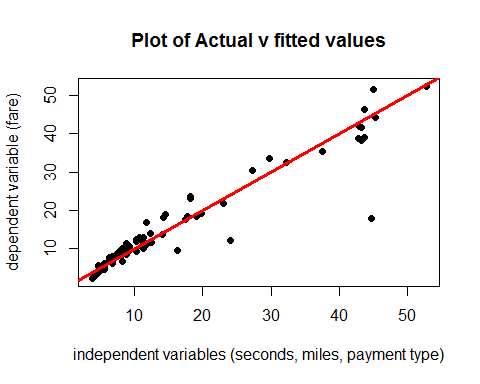
### # OBSERVATIONS A5.06:

# R-square: 0.8075 (Not Very High)  
# p-value: 0.335 which is higher than 0.05. Hence, we fail to reject the Null Hypothesis.  
# p-value of trip\_miles: 2e-16 which is significantly less than 0.05. Hence, we reject the Null Hypothesis for 'trip\_miles'  
# p-value of payment\_type: 0.187 which is higher than 0.05. Hence, we fail to reject the Null Hypothesis for 'payment\_type'  
# line equation: fare = -1.1098 + 0.0197 \* trip\_seconds + 1.7665 \* payment\_type  
# the line equation indicates that with unit increase in 'trip\_seconds' the fare increases by $0.0197 and depending on the type of payment increase in 'payment\_type', the fare increases by $1.7665

## # A5.07 all three: trip seconds, trip miles and payment type secmilpay = lm(cleanreorder$fare~reordercol$trip\_seconds + reordercol$trip\_miles + reordercol$payment\_type) summary(secmilpay)

##   
## Call:  
## lm(formula = cleanreorder$fare ~ reordercol$trip\_seconds + reordercol$trip\_miles +   
## reordercol$payment\_type)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.6488 -1.4015 -0.5917 0.7323 26.8123   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.490843 0.846260 1.762 0.0821  
## reordercol$trip\_seconds 0.009188 0.001405 6.539 6.12e-09  
## reordercol$trip\_miles 1.339591 0.146601 9.138 6.51e-14  
## reordercol$payment\_typeCredit Card 1.191125 0.927014 1.285 0.2027  
##   
## (Intercept) .   
## reordercol$trip\_seconds \*\*\*  
## reordercol$trip\_miles \*\*\*  
## reordercol$payment\_typeCredit Card   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.017 on 77 degrees of freedom  
## Multiple R-squared: 0.9077, Adjusted R-squared: 0.9041   
## F-statistic: 252.3 on 3 and 77 DF, p-value: < 2.2e-16

plot(reordercol$fare, secmilpay$fitted.values, pch = 19, main = "Plot of Actual v fitted values", xlab = "independent variables (seconds, miles, payment type)", ylab = "dependent variable (fare)")  
abline(0,1,col="red",lwd=3)



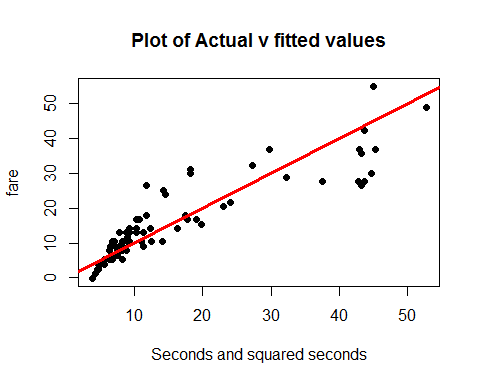
### # OBSERVATIONS A5.07:

# R-square: 0.9077 (Decently High)  
# p-value: 0.0821 which is higher than 0.05. Hence, we fail to reject the Null Hypothesis.  
# p-value of trip\_seconds: 6.12e-19 which is significantly less than 0.05. Hence, we reject the Null Hypothesis for 'trip\_seconds'  
# p-value of trip\_miles: 6.51e-14 which is significantly less than 0.05. Hence, we reject the Null Hypothesis for 'trip\_miles'  
# p-value of payment\_type: 0.2027 which higher than 0.05. Hence, we fail to reject the Null Hypothesis for 'payment\_type'  
# line equation: fare = 1.4908 + 0.0092 \* trip\_seconds + 1.3396 \* trip\_miles + 1.1911 \* payment\_type  
# the line equation indicates that with unit increase in 'trip\_seconds' the fare increases by $0.0092, for unit increase in 'trip\_miles' the fare increases by $1.3396 and depending on the type of payment increase in 'payment\_type', the fare increases by $1.1911

## # A5.08 sec and squared sec secsq = cleanreorder$trip\_seconds ^ 2 secsecsq = lm(cleanreorder$fare~ cleanreorder$trip\_seconds + secsq) summary(secsecsq)

##   
## Call:  
## lm(formula = cleanreorder$fare ~ cleanreorder$trip\_seconds +   
## secsq)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.646 -2.606 0.358 2.173 16.854   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.498e+00 1.718e+00 -0.872 0.386   
## cleanreorder$trip\_seconds 2.243e-02 3.648e-03 6.149 3.11e-08 \*\*\*  
## secsq -9.822e-07 1.468e-06 -0.669 0.506   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.81 on 78 degrees of freedom  
## Multiple R-squared: 0.8043, Adjusted R-squared: 0.7992   
## F-statistic: 160.2 on 2 and 78 DF, p-value: < 2.2e-16

plot(cleanreorder$fare, secsecsq$fitted.values, pch = 19, main = "Plot of Actual v fitted values", xlab = "Seconds and squared seconds", ylab = "fare")  
abline(0,1,col="red",lwd=3)



### # OBSERVATIONS A5.08:

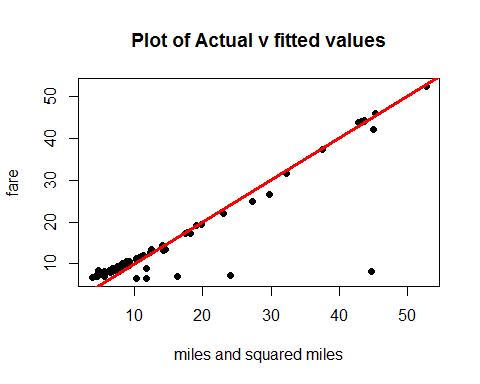
# R-square: 0.8043 (Not Very High)  
# p-value: 0.386 which is higher than 0.05. Hence, we fail to reject the Null Hypothesis.  
# p-value of trip\_seconds: 3.11e-08 which is significantly less than 0.05. Hence, we reject the Null Hypothesis for 'trip\_seconds'  
# p-value of squared trip\_seconds: 0.506 which is higher than 0.05. Hence, we fail to reject the Null Hypothesis for 'squared trip\_seconds'  
# line equation: fare = -1.498 + 2.243e-02 \* trip\_seconds - 9.822e-07 \* secsq  
# the line equation indicates that with unit increase in 'trip\_seconds' the fare increases by $2.243e-02 and with unit increase in 'secsq' the fare increases by -9.822e-07

## # A5.09 mile and squared miles

milsq = cleanreorder$trip\_miles ^ 2  
milmilsq = lm(cleanreorder$fare ~ cleanreorder$trip\_miles + milsq)  
summary(milmilsq)

##   
## Call:  
## lm(formula = cleanreorder$fare ~ cleanreorder$trip\_miles + milsq)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.770 -1.772 -0.913 -0.271 36.505   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.34841 1.00269 6.331 1.43e-08 \*\*\*  
## cleanreorder$trip\_miles 1.69552 0.43457 3.902 0.000201 \*\*\*  
## milsq 0.02570 0.02323 1.107 0.271904   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.005 on 78 degrees of freedom  
## Multiple R-squared: 0.8547, Adjusted R-squared: 0.851   
## F-statistic: 229.4 on 2 and 78 DF, p-value: < 2.2e-16

plot(cleanreorder$fare, milmilsq$fitted.values, pch = 19, main = "Plot of Actual v fitted values", xlab = "miles and squared miles", ylab = "fare")  
abline(0,1,col="red",lwd=3)



### # OBSERVATIONS A5.09:

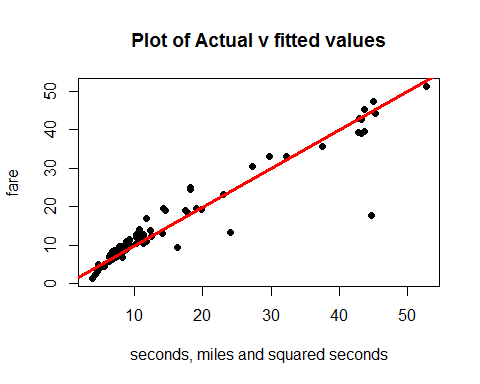
# R-square: 0.8547 (Decently High)  
# p-value: 1.43e-08 which is significantly lower than 0.05. Hence, we reject the Null Hypothesis.  
# p-value of trip\_miles: 0.0002 which is significantly less than 0.05. Hence, we reject the Null Hypothesis for 'trip\_miles'  
# p-value of squared trip\_miles: 0.2719 which is higher than 0.05. Hence, we fail to reject the Null Hypothesis for 'squared trip\_miles'  
# line equation: fare = 6.3484 + 1.6955 \* trip\_miles + 0.0257 \* milsq  
# the line equation indicates that with unit increase in 'trip\_miles' the fare increases by $1.6955 and with unit increase in 'milsq' the fare increases by $0.0257

## # A5.10 sec and mile and squared sec

secsq = cleanreorder$trip\_seconds ^ 2  
milsq = cleanreorder$trip\_miles ^ 2  
secmilsecsq = lm(cleanreorder$fare ~ cleanreorder$trip\_seconds + cleanreorder$trip\_miles + secsq)  
summary(secmilsecsq)

##   
## Call:  
## lm(formula = cleanreorder$fare ~ cleanreorder$trip\_seconds +   
## cleanreorder$trip\_miles + secsq)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.6364 -1.5092 -0.2654 0.7095 26.9195   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.785e-01 1.197e+00 0.316 0.753   
## cleanreorder$trip\_seconds 1.312e-02 2.693e-03 4.873 5.77e-06 \*\*\*  
## cleanreorder$trip\_miles 1.370e+00 1.457e-01 9.402 2.02e-14 \*\*\*  
## secsq -1.662e-06 1.011e-06 -1.644 0.104   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.99 on 77 degrees of freedom  
## Multiple R-squared: 0.9089, Adjusted R-squared: 0.9053   
## F-statistic: 256 on 3 and 77 DF, p-value: < 2.2e-16

plot(cleanreorder$fare, secmilsecsq$fitted.values, pch = 19, main = "Plot of Actual v fitted values", xlab = "seconds, miles and squared seconds", ylab = "fare")  
abline(0,1,col="red",lwd=3)



### # OBSERVATIONS A5.10:

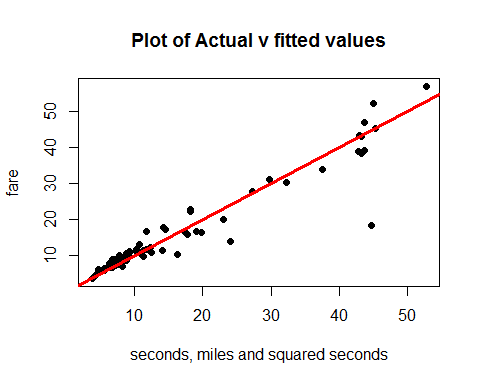
# R-square: 9089 (High)  
# p-value: 0.753 which is much higher than 0.05. Hence, we fail to reject the Null Hypothesis.  
# p-value of trip\_seconds: 5.77e-06 which is significantly less than 0.05. Hence, we reject the Null Hypothesis for 'trip\_seconds'  
# p-value of trip\_miles: 2.02e-14 which is significantly lower than 0.05. Hence, we reject the Null Hypothesis for 'trip\_miles'  
# p-value of squared trip\_seconds: -1.662e-06 which is significantly lower than 0.05. Hence, we reject the Null Hypothesis for 'squared trip\_seconds'  
# line equation: fare = 3.785e-0 + 1.312e-02 \* trip\_seconds + 1.370 \* trip\_miles + (-1.662e-06) \* secsq  
# the line equation indicates that with unit increase in 'trip\_seconds' the fare increases by $1.312e-02, with unit increase in 'trip\_miles' the fare increases by $1.370 and with unit increase in 'milsq' the fare increases by -$1.662e-06 or decreases by $1.662e-06

## # A5.11 sec and mile and squared mile

secsq = cleanreorder$trip\_seconds ^ 2  
milsq = cleanreorder$trip\_miles ^ 2  
secmilsecsq = lm(cleanreorder$fare ~ cleanreorder$trip\_seconds + cleanreorder$trip\_miles + milsq)  
summary(secmilsecsq)

##   
## Call:  
## lm(formula = cleanreorder$fare ~ cleanreorder$trip\_seconds +   
## cleanreorder$trip\_miles + milsq)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.2158 -1.4205 -0.4943 -0.0013 26.4511   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.998548 0.915990 3.274 0.00159 \*\*   
## cleanreorder$trip\_seconds 0.009785 0.001380 7.093 5.54e-10 \*\*\*  
## cleanreorder$trip\_miles 0.518704 0.378463 1.371 0.17450   
## milsq 0.043662 0.018359 2.378 0.01987 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.918 on 77 degrees of freedom  
## Multiple R-squared: 0.9121, Adjusted R-squared: 0.9087   
## F-statistic: 266.4 on 3 and 77 DF, p-value: < 2.2e-16

plot(cleanreorder$fare, secmilsecsq$fitted.values, pch = 19, main = "Plot of Actual v fitted values", xlab = "seconds, miles and squared seconds", ylab = "fare")  
abline(0,1,col="red",lwd=3)



### # OBSERVATIONS A5.11:

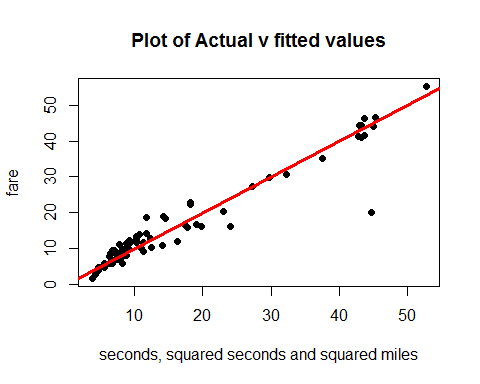
# R-square: 0.9121 (High)  
# p-value: 0.00159 which is lower than 0.05. Hence, we reject the Null Hypothesis.  
# p-value of trip\_seconds: 5.54e-10 which is significantly less than 0.05. Hence, we reject the Null Hypothesis for 'trip\_seconds'  
# p-value of trip\_miles: 0.1745 which is more than 0.05. Hence, we fail to reject the Null Hypothesis for 'trip\_miles'  
# p-value of squared trip\_miles: 0.0199 which is higher than 0.05. Hence, we fail to reject the Null Hypothesis for 'squared trip\_miles'  
# line equation: fare = 2.9985 + 0.0098 \* trip\_seconds + 0.5187 \* trip\_miles + 0.0437 \* milsq  
# the line equation indicates that with unit increase in 'trip\_seconds' the fare increases by $0.0098, with unit increase in 'trip\_miles' the fare increases by $0.5187 and with unit increase in 'milsq' the fare increases by $0.0437

## # A5.12 sec and miles and sec squared and mile squared

secsq = cleanreorder$trip\_seconds ^ 2  
milsq = cleanreorder$trip\_miles ^ 2  
secmilsecsqmilsq = lm(cleanreorder$fare ~ cleanreorder$trip\_seconds + cleanreorder$trip\_miles + secsq + milsq)  
summary(secmilsecsqmilsq)

##   
## Call:  
## lm(formula = cleanreorder$fare ~ cleanreorder$trip\_seconds +   
## cleanreorder$trip\_miles + secsq + milsq)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.9192 -1.5708 -0.2235 0.9589 24.6673   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.285e-01 1.103e+00 0.570 0.57038   
## cleanreorder$trip\_seconds 1.839e-02 2.828e-03 6.504 7.43e-09 \*\*\*  
## cleanreorder$trip\_miles -5.779e-02 3.926e-01 -0.147 0.88335   
## secsq -3.628e-06 1.060e-06 -3.422 0.00100 \*\*   
## milsq 7.581e-02 1.960e-02 3.868 0.00023 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.671 on 76 degrees of freedom  
## Multiple R-squared: 0.9239, Adjusted R-squared: 0.9199   
## F-statistic: 230.5 on 4 and 76 DF, p-value: < 2.2e-16

plot(cleanreorder$fare, secmilsecsqmilsq$fitted.values, pch = 19, main = "Plot of Actual v fitted values", xlab = "seconds, squared seconds and squared miles", ylab = "fare")  
abline(0,1,col="red",lwd=3)



### # OBSERVATIONS A5.12:

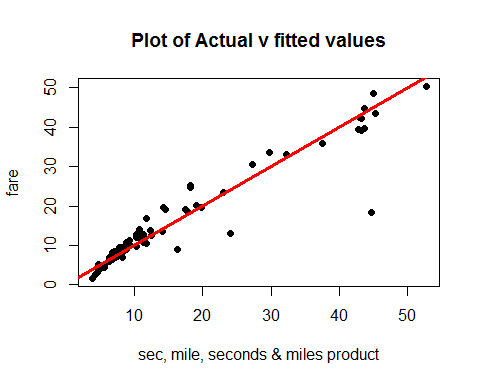
# R-square: 0.9239 (Highest among all 12 models)  
# p-value: 7.43e-09 which is significantly lower than 0.05. Hence, we reject the Null Hypothesis.  
# p-value of trip\_seconds: -5.5779e-02 which significantly lower than 0.05. Hence, we reject the Null Hypothesis for 'trip\_seconds'.  
# p-value of trip\_miles: -3.628e-06 which is higher than 0.05. Hence, we reject the Null Hypothesis for 'trip\_miles.  
# p-value of squared trip\_seconds: -3.628e-06 which is significantly lower than 0.05. Hence, we reject the Null Hypothesis for 'squared trip\_seconds'  
# p-value of squared trip\_miles: 7.518e-02 which which is significantly lower than 0.05. Hence, we reject the Null Hypothesis for 'squared trip\_miles'  
# line equation: fare = 6.285e-01 + 1.839e-02 \* trip\_seconds + (-5.779e-02) \* trip\_miles + (-3.628e-06) \* secsq + 7.581e-02 \* milsq  
# the line equation indicates that with unit increase in 'trip\_seconds' the fare increases by $1.839e-02, with unit increase in 'trip\_miles' the fare increases by -$5.779e-02, with unit increase in 'secsq' the fare increases by -$3.628e-06 and with unit increase in 'milsq' the fare increases by $7.581e-02

## # A5.13 Interaction 1 - sec \* mil, mile, sec

secondmile = cleanreorder$trip\_seconds \* cleanreorder$trip\_miles  
secmillm = lm(cleanreorder$fare ~ cleanreorder$trip\_seconds + cleanreorder$trip\_miles + secondmile)  
summary(secmillm)

##   
## Call:  
## lm(formula = cleanreorder$fare ~ cleanreorder$trip\_seconds +   
## cleanreorder$trip\_miles + secondmile)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.0096 -1.3996 -0.4159 0.6657 26.4638   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.6821903 1.0582833 0.645 0.5211   
## cleanreorder$trip\_seconds 0.0107561 0.0016319 6.591 4.89e-09 \*\*\*  
## cleanreorder$trip\_miles 1.6994449 0.2537634 6.697 3.10e-09 \*\*\*  
## secondmile -0.0002422 0.0001452 -1.668 0.0994 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.988 on 77 degrees of freedom  
## Multiple R-squared: 0.909, Adjusted R-squared: 0.9054   
## F-statistic: 256.3 on 3 and 77 DF, p-value: < 2.2e-16

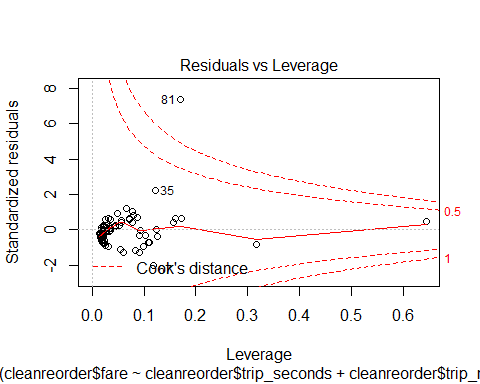
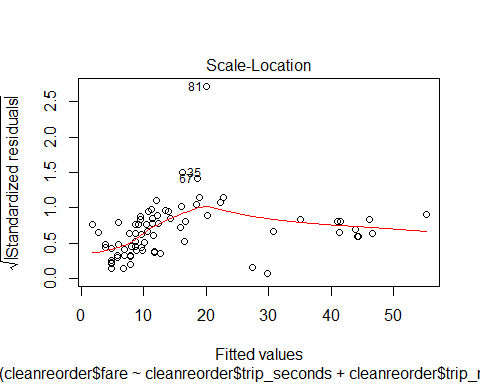
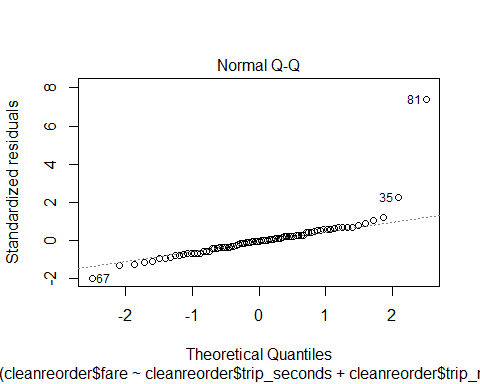
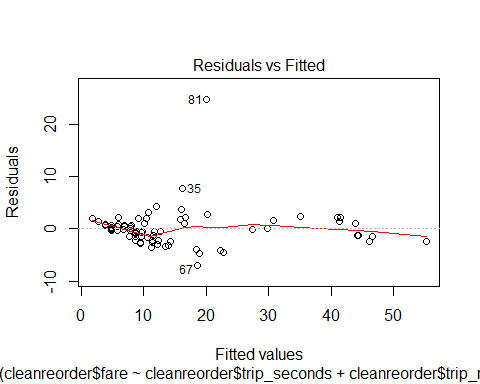
plot(cleanreorder$fare, secmillm$fitted.values, pch = 19, main = "Plot of Actual v fitted values", xlab = "sec, mile, seconds & miles product", ylab = "fare")  
abline(0,1,col="red",lwd=3)



### # OBSERVATIONS A5.13:

# R-square: 0.909 (not very high)  
# p-value: 0.5211 which is much higher than 0.05. Hence, we fail to reject the Null Hypothesis.  
# p-value for trip\_seconds: 4.89e-09 which is significantly lower than 0.05. Hence, we reject the Null Hypothesis for 'trip\_seconds'.  
# p-value for trip\_miles: 3.10e-09 which is significantly lower than 0.05. Hence, we reject the Null Hypothesis for 'trip\_miles'.  
# p-value for secondmile: 2e-16 which is significantly low than 0.05. Hence, we reject the Null Hypothesis.  
# line equation: fare = 0.6822 + 0.0108 \* trip\_seconds + 1.6994 \* trip\_miles + 1.027e-03 \* secondmile  
# the line equation indicates that with unit increase in 'trip\_seconds' the fare increases by $0.0108, with unit increase in 'trip\_miles' the fare increases by $1.6994 and with unit increase in 'secondmile' the fare increases by $1.027e-03

plot(secmilsecsqmilsq)



# The R-squared value for the best fit model is 0.9239 or 92.39%. Next, the plot of residuals vs leverage  
# helps identifying the influencial data points of the model. The above plot indicates how one point is outside the   
# cooks distance, but it is right outside the cook's distance line (dashed line) and it's value is 81.

## # OBSERVATIONS OF ANALYSIS 5 -

# Various combinations of given variables have been taken for creating the linear models for all of them in order  
# to find the best fit model for the given sameple data set. It is found that the lowest R squared values among the  
# combinations # is 0.9239 for a model with dependent variable as 'fare' and independent variables as 'trip\_seconds',   
# 'trip\_miles', # 'square of trip\_seconds' and 'square of trip miles'. Let us call it the BEST FIT MODEL for this project.  
# The plot of the best fit model indicates that the p value of the model is not significant.

# 

# ANALYSIS 6

# The best fit model among the 12 models above is the one where the independet variables are sec and miles and sec squared and mile squared and taking it's linear model.

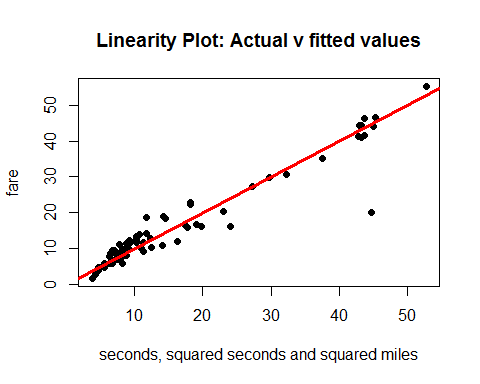
secsq = reordercol$trip\_seconds ^ 2  
milsq = reordercol$trip\_miles ^ 2  
secmilsecsqmilsq = lm(reordercol$fare ~ reordercol$trip\_seconds + reordercol$trip\_miles + secsq + milsq)  
summary(secmilsecsqmilsq)

##   
## Call:  
## lm(formula = reordercol$fare ~ reordercol$trip\_seconds + reordercol$trip\_miles +   
## secsq + milsq)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.9192 -1.5708 -0.2235 0.9589 24.6673   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.285e-01 1.103e+00 0.570 0.57038   
## reordercol$trip\_seconds 1.839e-02 2.828e-03 6.504 7.43e-09 \*\*\*  
## reordercol$trip\_miles -5.779e-02 3.926e-01 -0.147 0.88335   
## secsq -3.628e-06 1.060e-06 -3.422 0.00100 \*\*   
## milsq 7.581e-02 1.960e-02 3.868 0.00023 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.671 on 76 degrees of freedom  
## Multiple R-squared: 0.9239, Adjusted R-squared: 0.9199   
## F-statistic: 230.5 on 4 and 76 DF, p-value: < 2.2e-16

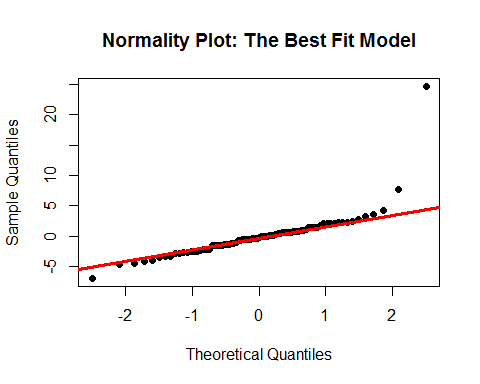
# Linear model of The Best Fit model shows that the R-Squared is 0.9239 and the p-value is ~ 0.5704 which is much higher  
# than 0.05. So, we reject

# Testing the above model on LINE assumptions   
# Linearity Test for The Best Fit Model

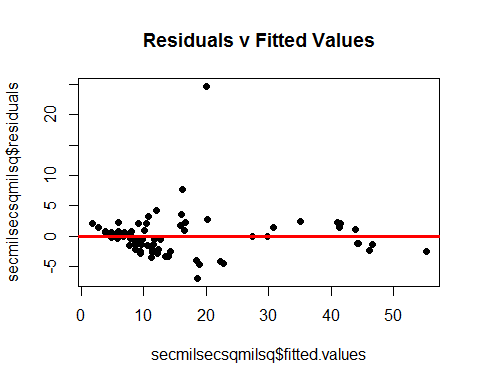
plot(reordercol$fare, secmilsecsqmilsq$fitted.values, pch = 19, main = "Linearity Plot: Actual v fitted values", xlab = "seconds, squared seconds and squared miles", ylab = "fare")  
abline(0,1,col="red",lwd=3)



# Normality Test for The Best Fit Model  
qqnorm(secmilsecsqmilsq$residuals,pch=19,main="Normality Plot: The Best Fit Model")  
qqline(secmilsecsqmilsq$residuals,col="red",lwd=3)



# Equality of Variances Test for The Best Fit Model  
plot(secmilsecsqmilsq$fitted.values,secmilsecsqmilsq$residuals,pch=19,main="Residuals v Fitted Values")  
abline(0,0,col="red",lwd=3)



## # OBSERVATIONS OF ANALYSIS 6:

# LINEARITY TEST: the plot of linearity for the Best Fit Model shows that dataset 'secmilsecsqmilsq' is LINEAR in nature the data points are clustered before independent variables value ~ 12 and after that, the data points start scattering with one outlier below the absolute line.

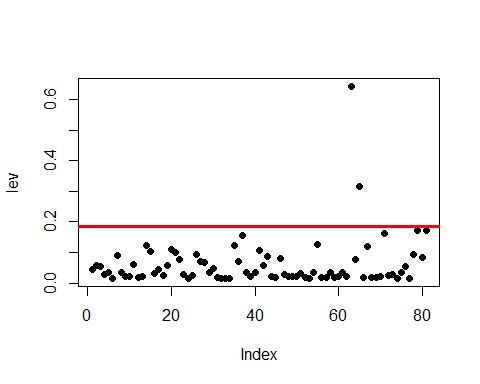
# NORMALITY TEST: the normality plot for the Best Fit Model shows that the dataset 'secmilsecsqmilsq' is NORMAL in nature the data points are along the absolute line for most of the part except for a few outliers that are diverging from the line with one point very far from the line.

# EQUALITY OF VARIANCES TEST: the equality of variances plot for the Best Fit Model shows that the dataset 'secmilsecsqmilsq' is NOT EQUALLY VARIANT in nature as many of the points are scattered above and below the absolute line with one data point being too far away from the absolute line.

# 

# ANALYSIS 7

lev=hat(model.matrix(secmilsecsqmilsq))  
plot(lev,pch=19)  
abline(3\*mean(lev),0,col="red",lwd=3)



levpoints = which(lev > 3\*mean(lev))

reordercol[63,]

## taxi\_id trip\_seconds trip\_minutes trip\_miles fare tips trip\_total  
## 947402 4436 2880 48 16.8 45 12.25 61.25  
## payment\_type  
## 947402 Credit Card

reordercol[65,]

## taxi\_id trip\_seconds trip\_minutes trip\_miles fare tips trip\_total  
## 24543 4719 2520 42 20.7 52.75 0 53.75  
## payment\_type  
## 24543 Cash

removelev = reordercol[-c(levpoints),]  
cleanremovelev = removelev[c(2,4,5,6,7,8)]  
summary(reordercol)

## taxi\_id trip\_seconds trip\_minutes trip\_miles   
## Min. : 173 Min. : 60 Min. : 1 Min. : 0.100   
## 1st Qu.:2005 1st Qu.: 420 1st Qu.: 7 1st Qu.: 1.100   
## Median :3775 Median : 600 Median :10 Median : 1.900   
## Mean :4006 Mean : 780 Mean :13 Mean : 4.395   
## 3rd Qu.:6044 3rd Qu.:1020 3rd Qu.:17 3rd Qu.: 4.500   
## Max. :8736 Max. :2880 Max. :48 Max. :20.700   
## fare tips trip\_total payment\_type  
## Min. : 3.75 Min. : 0.000 Min. : 3.75 Cash :45   
## 1st Qu.: 6.75 1st Qu.: 0.000 1st Qu.: 7.75 Credit Card:36   
## Median : 9.25 Median : 0.000 Median :10.75 Other : 0   
## Mean :15.07 Mean : 1.785 Mean :17.41   
## 3rd Qu.:17.75 3rd Qu.: 2.050 3rd Qu.:18.25   
## Max. :52.75 Max. :12.250 Max. :61.25

summary(cleanremovelev)

## trip\_seconds trip\_miles fare tips   
## Min. : 60.0 Min. : 0.100 Min. : 3.75 Min. : 0.000   
## 1st Qu.: 390.0 1st Qu.: 1.100 1st Qu.: 6.75 1st Qu.: 0.000   
## Median : 540.0 Median : 1.900 Median : 9.00 Median : 0.000   
## Mean : 731.4 Mean : 4.032 Mean :14.22 Mean : 1.675   
## 3rd Qu.: 900.0 3rd Qu.: 3.900 3rd Qu.:16.88 3rd Qu.: 2.025   
## Max. :2160.0 Max. :18.300 Max. :45.25 Max. :11.400   
## trip\_total payment\_type  
## Min. : 3.75 Cash :44   
## 1st Qu.: 7.50 Credit Card:35   
## Median :10.75 Other : 0   
## Mean :16.40   
## 3rd Qu.:17.88   
## Max. :57.15

# After investigating the newly cleaned data set, the plot of the 'lev' shows that there are two points above the leverage line which inlfuences the data set. Both of the points with high leverage has been removed and saved into 'cleanremovelev'. The summary of 'cleanremovelev' shows the minimum, maximum, mean, median, Q1 and Q3 for all the independent variables.

# The Summary comparison of previously cleaned data set ('reordercol') and currently cleaned data set ('cleanremovelev') shows changes the summary of the independent variables

summarycomparison <- data.frame(  
 summary = c("MINIMUM", "MAXIMUM", "MEAN", "MEDIAN", "Q1", "Q3"),  
 old\_seconds = c(60.0, 2880.1, 780.0, 600.0, 420.0, 1020.0),  
 new\_seconds = c(60.0, 2160.0, 731.4, 540.0, 390.0, 900.0),  
 old\_miles = c(0.100, 20.700, 4.395, 1.900, 1.100, 4.500),  
 new\_miles = c(0.100, 18.300, 4.032, 1.900, 1.100, 3.900),  
 old\_fare = c(3.75, 52.75, 15.07, 9.25, 6.75, 17.75),  
 new\_fare = c(3.75, 45.25, 14.22, 9.00, 6.75, 16.88),  
 old\_tips = c(0.000, 12.250, 1.785, 0.000, 0.000, 2.050),  
 new\_tips = c(0.000, 11.400, 1.675, 2.025, 0.000, 2.025),  
 old\_totaltrips = c(3.75, 61.25, 17.41, 10.75, 7.75, 18.25),  
 new\_totaltrips = c(3.75, 57.15, 16.40, 10.75, 7.50, 17.88)  
)  
summarycomparison

## summary old\_seconds new\_seconds old\_miles new\_miles old\_fare new\_fare  
## 1 MINIMUM 60.0 60.0 0.100 0.100 3.75 3.75  
## 2 MAXIMUM 2880.1 2160.0 20.700 18.300 52.75 45.25  
## 3 MEAN 780.0 731.4 4.395 4.032 15.07 14.22  
## 4 MEDIAN 600.0 540.0 1.900 1.900 9.25 9.00  
## 5 Q1 420.0 390.0 1.100 1.100 6.75 6.75  
## 6 Q3 1020.0 900.0 4.500 3.900 17.75 16.88  
## old\_tips new\_tips old\_totaltrips new\_totaltrips  
## 1 0.000 0.000 3.75 3.75  
## 2 12.250 11.400 61.25 57.15  
## 3 1.785 1.675 17.41 16.40  
## 4 0.000 2.025 10.75 10.75  
## 5 0.000 0.000 7.75 7.50  
## 6 2.050 2.025 18.25 17.88

# The above command shows the comparison of summary of the two data sets out of which the first data set (with prefix 'old') is the data set which was cleaned before Analysis 1 and the other data set is the one (with prefix 'new') which was cleaned by removing the leverage points. The table shows the changes in seconds, miles, fare, tips and total trip fare.

paymentcomparison <- data.frame(  
 paymenttype = c("CASH", "CREDIT CARD", "OTHER"),  
 old\_payment = c(45, 36, 0),  
 new\_payment = c(44, 35, 0)  
)  
paymentcomparison

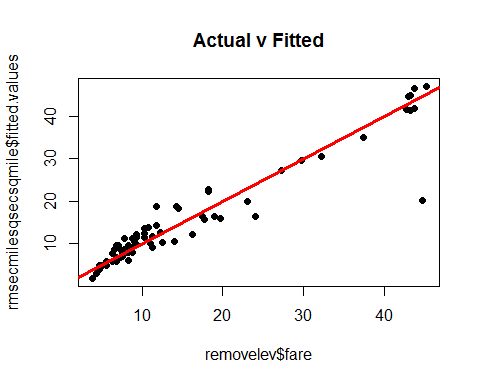
## paymenttype old\_payment new\_payment  
## 1 CASH 45 44  
## 2 CREDIT CARD 36 35  
## 3 OTHER 0 0

# the comparison for payment shows that after removing the leverage points, the counts for both CASH and CREDIT CARD decreased by 1 while the counts for other remained the same i.e. zero

levsqmil = removelev$trip\_miles ^ 2  
levsqsec = removelev$trip\_seconds ^ 2  
rmsecmilesqsecsqmile = lm(removelev$fare ~ removelev$trip\_seconds + removelev$trip\_miles + levsqmil + levsqsec)  
summary(rmsecmilesqsecsqmile)

##   
## Call:  
## lm(formula = removelev$fare ~ removelev$trip\_seconds + removelev$trip\_miles +   
## levsqmil + levsqsec)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.0134 -1.6142 -0.2415 1.0997 24.5065   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.557e-01 1.321e+00 0.572 0.56890   
## removelev$trip\_seconds 1.837e-02 3.802e-03 4.830 7.17e-06 \*\*\*  
## removelev$trip\_miles -1.542e-01 4.159e-01 -0.371 0.71188   
## levsqmil 8.153e-02 2.154e-02 3.785 0.00031 \*\*\*  
## levsqsec -3.552e-06 1.873e-06 -1.896 0.06185 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.704 on 74 degrees of freedom  
## Multiple R-squared: 0.9084, Adjusted R-squared: 0.9034   
## F-statistic: 183.4 on 4 and 74 DF, p-value: < 2.2e-16

plot(removelev$fare, rmsecmilesqsecsqmile$fitted.values, pch = 19, main = "Actual v Fitted")  
abline(0,1,col="red",lwd=3)



## # OBSERVATIONS OF ANALYSIS 7:

# the linear model plot after removing the leverage points shows that, even though points are the data points are somewhat linear in nature with one outlier below the absolute line.

# Also, the summary of linear model after removing the leverage points shows that the R-Sqaured value decreased from 0.9239 to 0.9084 because of removing the leverage points which indicates that the leverage points were influencing the model and removing them has resulted into lower accuracy.

# The p-value is almost similar for both the data sets i.e. ~ 0.56 which is greater than 0.05 so we fail to reject the Null Hypothesis

# 

# ANALYSIS 8

set.seed(51592556) # 2551 + 5 = 2556  
newsampletaxi = infotaxi[sample(1:nrow(infotaxi),100,replace=FALSE),]

# Applying same cleansing for the newly created sample data set 'newsampletaxi'  
newsecmileszero = subset(newsampletaxi, newsampletaxi$trip\_seconds != 0 & newsampletaxi$trip\_miles != 0)  
newsecmileszero$trip\_minutes = newsecmileszero$trip\_seconds/60  
newreordercol <- newsecmileszero[,c(1,2,10,3,4,5,8,9)]  
sum(is.na(newsampletaxi))

## [1] 0

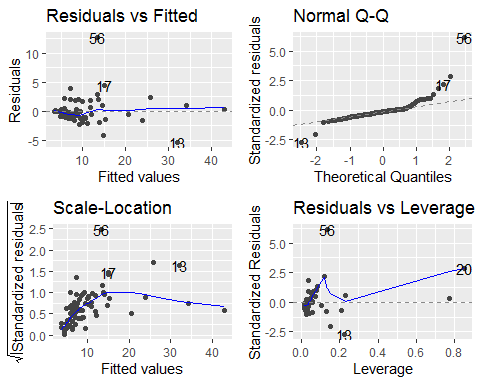
sum(is.na(newsecmileszero))

## [1] 0

# Using the Best Fit Model on newly cleansed data set 'newreordercol' - sec and miles and sec squared and mile squared  
newsecsq = newreordercol$trip\_seconds ^ 2  
newmilsq = newreordercol$trip\_miles ^ 2  
newsecmilsecsqmilsq = lm(newreordercol$fare ~ newreordercol$trip\_seconds + newreordercol$trip\_miles + newsecsq + newmilsq)  
summary(newsecmilsecsqmilsq)

##   
## Call:  
## lm(formula = newreordercol$fare ~ newreordercol$trip\_seconds +   
## newreordercol$trip\_miles + newsecsq + newmilsq)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.2918 -0.9470 -0.2675 0.2747 12.6491   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.892e+00 6.109e-01 4.733 1.18e-05 \*\*\*  
## newreordercol$trip\_seconds 1.077e-02 1.801e-03 5.982 9.54e-08 \*\*\*  
## newreordercol$trip\_miles 1.515e-01 2.886e-01 0.525 0.6013   
## newsecsq -1.110e-06 6.093e-07 -1.822 0.0729 .   
## newmilsq 8.190e-02 1.749e-02 4.682 1.43e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.221 on 67 degrees of freedom  
## Multiple R-squared: 0.9153, Adjusted R-squared: 0.9103   
## F-statistic: 181 on 4 and 67 DF, p-value: < 2.2e-16

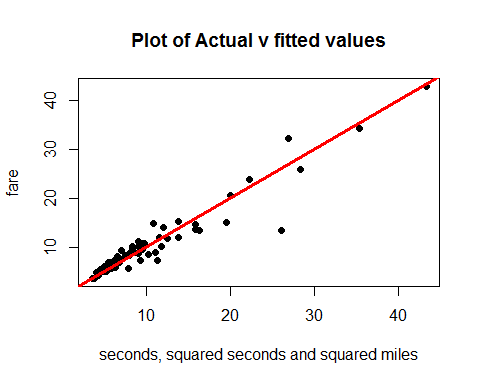
autoplot(newsecmilsecsqmilsq)



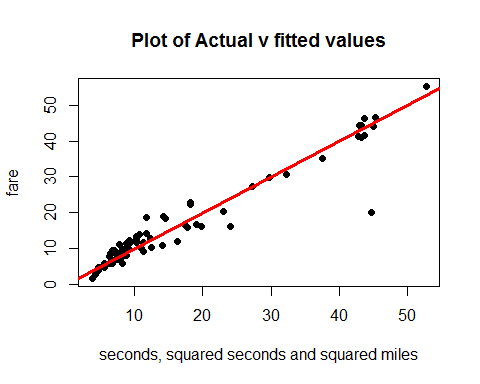
## #LINE assumptions

### # Linearity

plot(newreordercol$fare, newsecmilsecsqmilsq$fitted.values, pch = 19, main = "Plot of Actual v fitted values", xlab = "seconds, squared seconds and squared miles", ylab = "fare")  
abline(0,1,col="red",lwd=3)

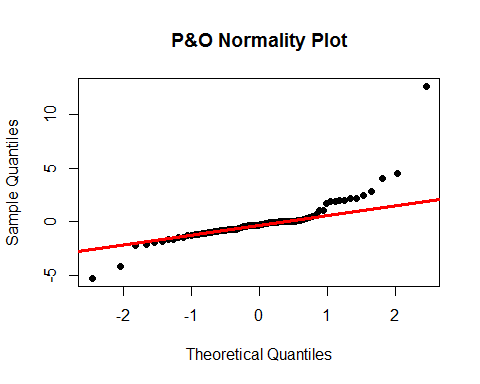


plot(reordercol$fare, secmilsecsqmilsq$fitted.values, pch = 19, main = "Plot of Actual v fitted values", xlab = "seconds, squared seconds and squared miles", ylab = "fare")  
abline(0,1,col="red",lwd=3)



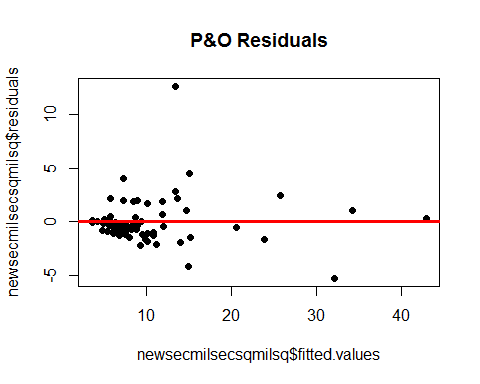
### # Normality

qqnorm(newsecmilsecsqmilsq$residuals,pch=19,main="P&O Normality Plot")  
qqline(newsecmilsecsqmilsq$residuals,col="red",lwd=3)



### # Equality of Variances

plot(newsecmilsecsqmilsq$fitted.values,newsecmilsecsqmilsq$residuals,pch=19,main="P&O Residuals")  
abline(0,0,col="red",lwd=3)



## # OBSERVATIONS OF ANALYSIS 8

# Summary - The summary of newly cleansed data suggests that the R-Sqaured decreased from 0.9239 to 0.9153  
# Also, the p-value is significantly low as compared to the previous linear model. Hence, we reject the Null Hypothesis

# Line equation: fare = 2.892 + 1.077e-02 \* trip\_seconds + 1.515e-01 \* trip\_miles + (-1.110e-06) \* newsecsq + 8.190e-02 \* newmilsq

# Autoplot - residuals v fitted: it shows that the plot is linear in nature and three points that are away from the line which indicates that the model is a good fit for the data set as there is no Non-linearity

# Autoplot - Normal Q-Q: it suggests that there no normality in the data points as the plot shows that the data is getting along with the absolute line for a good part and then the points begin to scatter. Also, there are a few outliers towards both the tails

# Autoplot - Scale-location: The residuals are clustered before 10 but, begin spreading more after the residual value 10 which causes the absolute line to curve in order to fit the points

# Autoplot - Residuals - Leverage: there are three points in the plot of residuals v leverage namely, 13, 20 and 56 These points are influencial points because the plot does not even show the cook's distance line and all the points are well inside the cook's distnace line.

# LINEARITY TEST: the plot of linearity for the NEW Best Fit Model shows that dataset 'newsecmilsecsqmilsq' is ALSO LINEAR in nature. The data points are clustered before the independent variables value ~ 15 and after that, the data points start scattering with outlier below the absolue line.

# NORMALITY TEST: the normality plot for the NEW Best Fit Model shows that the dataset 'newsecmilsecsqmilsq' is NORMAL in nature the data points are along the absolute line for most of the part except for an outlier near the lower tail and two outliers near the upper tail.

# EQUALITY OF VARIANCES - the equality of variances plot for the NEW Best Fit Model shows that the dataset 'newsecmilsecsqmilsq' is NOT EQUALLY VARIANT in nature as many of the points are scattered above and below the absolute line with one data point being too far away from the absolute line and the data points are not scattered equally.