QMB-Regression-Project

Vidhi Pankaj Jadav

2019-10-20

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

##Preprocessing 1

#Loading the CSV file  
taxis = read.csv("6304 Regression Project Data.csv", header = TRUE)  
attach(taxis)  
ncol(taxis)

## [1] 9

nrow(taxis)

## [1] 1705421

##Preprocessing 2

#100 random samples by taking U number as a random number seed  
#Sample data  
set.seed(74470957)  
sample.taxis = taxis[sample(1:nrow(taxis),100, replace=F),]  
ncol(sample.taxis)

## [1] 9

nrow(sample.taxis)

## [1] 100

#We are now left with 100 rows.

##Preprocessing 3  
  
#Cleansing the random data set  
  
sum(abs(sample.taxis$tolls))

## [1] 0

#The absolute values of tolls column adds up to 0.

#Removing column tolls since it has no impact as each row of tolls has 0 value and hence 0 sum.

sample.taxis$tolls = NULL  
  
#Removing all rows where trip seconds and trip miles are 0. Logically speaking, if we are not traveling any distance, there is no time involved and vice versa, unless we consider the traffic signal factor in between the trip.

filter.sample= subset (sample.taxis, sample.taxis$trip\_seconds!=0 & sample.taxis$trip\_miles!=0.00)

#Converting the trip\_seconds column to trip\_minutes, because generally we talk in minutes to reach a destination rather than addressing it in second.

filter.sample$trip\_minutes = filter.sample$trip\_seconds/60  
  
#Removing trip\_seconds column from the filter.sample.

filter.sample$trip\_seconds=NULL

#Replacing the trip\_seconds column with trip\_minutes by placing it at the same position.

order\_sample = c("taxi\_id","trip\_minutes","trip\_miles","fare","tips","extras","trip\_total","payment\_type")  
final\_sample=filter.sample[,order\_sample]

#Displaying information about the structure of the final data sample  
  
str(final\_sample)

## 'data.frame': 80 obs. of 8 variables:  
## $ taxi\_id : int 4819 7002 3558 8116 4131 417 5960 3454 1594 3478 ...  
## $ trip\_minutes: num 14 3 9 4 46 2 7 8 8 23 ...  
## $ trip\_miles : num 2.5 0.5 1.6 0.5 13 ...  
## $ fare : num 10.5 4.5 8 4.65 35 4.25 6.75 7.5 12.5 38.5 ...  
## $ tips : num 2.1 0 0 2 7 2 0 2 0 0 ...  
## $ extras : num 0 0 0 0 0 0 0 0 1 12 ...  
## $ trip\_total : num 12.6 4.5 8 6.65 42 6.25 6.75 9.5 13.5 50.5 ...  
## $ payment\_type: Factor w/ 3 levels "Cash","Credit Card",..: 2 1 1 2 2 2 1 2 1 1 ...

#Attaching the final sample

attach(final\_sample)

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

nrow(final\_sample)

## [1] 80

ncol(final\_sample)

## [1] 8

#The final\_sample after all the cleansing leaves me with total of 80 observations out of 100 and 8 variables out of 9.

##Analysis 1

#Summary  
summary\_sample = final\_sample[,c(2,3,4,5,6,7)]  
summary(summary\_sample)

## trip\_minutes trip\_miles fare tips   
## Min. : 1.0 Min. : 0.100 Min. : 4.000 Min. :0.000   
## 1st Qu.: 5.0 1st Qu.: 0.675 1st Qu.: 5.750 1st Qu.:0.000   
## Median : 9.5 Median : 1.150 Median : 8.125 Median :0.000   
## Mean :11.6 Mean : 3.228 Mean :12.593 Mean :1.575   
## 3rd Qu.:13.0 3rd Qu.: 2.525 3rd Qu.:12.062 3rd Qu.:2.000   
## Max. :50.0 Max. :19.700 Max. :50.000 Max. :8.750   
## extras trip\_total   
## Min. : 0.000 Min. : 4.00   
## 1st Qu.: 0.000 1st Qu.: 7.00   
## Median : 0.000 Median : 9.75   
## Mean : 0.825 Mean :14.99   
## 3rd Qu.: 1.000 3rd Qu.:13.31   
## Max. :18.000 Max. :52.50

"The summary displays-  
minimum value,   
maximum value,  
1st Quartile (lowest 25% of data)'  
mean,   
meadian (2nd Quartle),   
3rd Quartile (highest 25% of data)  
for each attribute in the sample."

## [1] "The summary displays-\nminimum value, \nmaximum value,\n1st Quartile (lowest 25% of data)'\nmean, \nmeadian (2nd Quartle), \n3rd Quartile (highest 25% of data)\nfor each attribute in the sample."

par(mfrow=c(2,3))  
  
#Column - trip\_minutes  
skewness(final\_sample$trip\_minutes)

## [1] 1.980442

kurtosis(final\_sample$trip\_minutes)

## [1] 7.213707

plot(density(final\_sample$trip\_minutes),lwd=3,col="blue",main="Density plot for trip in minutes")  
  
#Column - trip\_miles  
skewness(final\_sample$trip\_miles)

## [1] 2.235974

kurtosis(final\_sample$trip\_miles)

## [1] 6.606866

plot(density(final\_sample$trip\_miles),lwd=3,col="blue",main="Density plot for trip in miles")  
  
#Column - fare  
skewness(final\_sample$fare)

## [1] 1.917533

kurtosis(final\_sample$fare)

## [1] 5.414419

plot(density(final\_sample$fare),lwd=3,col="blue",main="Density plot for trip fare")  
  
#Column - tips  
skewness(final\_sample$tips)

## [1] 1.786593

kurtosis(final\_sample$tips)

## [1] 5.42166

plot(density(final\_sample$tips),lwd=3,col="blue",main="Density plot for Tips")  
  
#Column - extras  
skewness(final\_sample$extras)

## [1] 5.336911

kurtosis(final\_sample$extras)

## [1] 33.87773

plot(density(final\_sample$extras),lwd=3,col="blue",main="Density plot for trip in Extras")  
  
#Column - trip\_total  
skewness(final\_sample$trip\_total)

## [1] 1.821679

kurtosis(final\_sample$trip\_total)

## [1] 4.766076

plot(density(final\_sample$trip\_total),lwd=3,col="blue",main="Density plot for Trip Total")

#Observation: It is evident from the density plots of all the 6 continuous variables that the distributions are positively skewed(right-skewed) with the extras having the most leptokurtic curve with its very pointy peak and flattened tails followed by the trip\_minutes, trip\_miles, tips, fare and trip\_total.   
#As seen from the summary of data, the mean is greater than the median (mean>median) for each attribute plots which also signifies the data is right-skewed in all the plots.  
  
  
par(mfrow=c(1,1))

##Analysis 2

#Column - payment\_type  
library(plyr)  
count(final\_sample,c("payment\_type"))

## payment\_type freq  
## 1 Cash 39  
## 2 Credit Card 41

table(final\_sample$payment\_type)

##   
## Cash Credit Card Other   
## 39 41 0

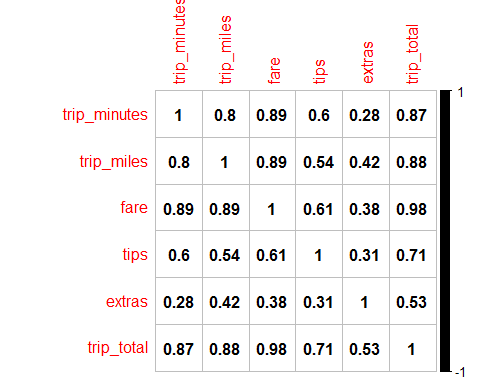
#In payment\_type factor variable the number of cases where customer pays cash is 39 and where amount is paid by credit card is 41.

##Analysis 3

#Correlation matrix  
my\_cor\_col = final\_sample[,c(2,3,4,5,6,7)]  
my\_cor\_matrix=cor(my\_cor\_col)  
library(corrplot)

## corrplot 0.84 loaded

corrplot(my\_cor\_matrix, col="black", method='number')



#Conclusion  
#The correlation matrix between the continuous independent variables shows-  
#There is a perfect correlation(100% correlation) of each attribute with itself depicting a value of 1.  
#Highest correlation of 98%(.98) between trip total and fare. This means as fare increases trip total increases.  
#Another strong correlation of 89%(.89) is between fare and trip\_miles,trip\_minutes. As time and distance of trip increases the fare also rises.  
#There is strong correlation of 88%(.88) between trip miles and trip total.  
#Strong correlation is also seen between trip toal and trip minutes which is 87%.  
#The correlation between trip miles and minutes is preety good.  
#A good correlation of 71% is seen between tips and trip total.  
#Tips and trip\_minutes are fairly correlated with 60% correlation.  
#Tips and fare; and exras and trip\_total are moderately correlated with a linear relationship of 54% and 53% respectively..  
#Other with a corelation less than 50 as weakly correlated."

##Analysis 4  
  
#Regression model  
taxi.out = lm(fare ~ trip\_minutes+trip\_miles+payment\_type, data=final\_sample)  
summary(taxi.out)

##   
## Call:  
## lm(formula = fare ~ trip\_minutes + trip\_miles + payment\_type,   
## data = final\_sample)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.3202 -1.4126 -0.5866 0.3675 23.0104   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.80205 0.79085 2.279 0.0255 \*   
## trip\_minutes 0.58788 0.07807 7.531 8.67e-11 \*\*\*  
## trip\_miles 1.15305 0.14996 7.689 4.32e-11 \*\*\*  
## payment\_typeCredit Card 0.48591 0.91791 0.529 0.5981   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.959 on 76 degrees of freedom  
## Multiple R-squared: 0.8875, Adjusted R-squared: 0.883   
## F-statistic: 199.8 on 3 and 76 DF, p-value: < 2.2e-16

#Observation:  
#The beta coefficients of trip\_minutes and trip\_miles impact the fare greatly as it has very significant p-values. The beta coeeficient of intercept has a significant p-value and fare is dependent on it while the p\_value of payment\_type is not significant and does not impact fare.We look at the adjusted R-squarred value as we have more than 1 independent variable.  
#The fit of my model has good accuracy of 88.3% with the adjusted R-squarred value of .883.  
  
  
#Impact of each continuous variable on the dependent variable  
cor(final\_sample$fare,final\_sample$trip\_minutes)

## [1] 0.8940891

#Trip minutes highly impacts the fare with a corelation of 89.4%.  
cor(final\_sample$fare,final\_sample$trip\_miles)

## [1] 0.8912985

#Trip miles highly impacts the fare with a corelation of 89.12%.  
cor(final\_sample$fare,final\_sample$tips)

## [1] 0.6124126

#Tips fairly impacts the fare with a corelation of 61.24%.  
cor(final\_sample$fare,final\_sample$extras)

## [1] 0.3820868

#Extras has poor corelation with fare at 38.2%  
cor(final\_sample$fare,final\_sample$trip\_total)

## [1] 0.9763247

#Trip total strongly impacts the fare with 97.63%.  
  
confint(taxi.out,level=.9)

## 5 % 95 %  
## (Intercept) 0.4851597 3.1189308  
## trip\_minutes 0.4578898 0.7178741  
## trip\_miles 0.9033504 1.4027498  
## payment\_typeCredit Card -1.0425468 2.0143603

"The 95% confidence interval(C.I.) range for intercept has a wide range that is between .48 and 3.11  
 The 95% C.I. range of trip\_minutes has a moderate range between .45 and .71 considering 80 samples.  
 The 95% C.I. range of trip\_miles has a fairly tight range between .9 and 1.4 which means 5 differnet values will be repeated for ??2 coefficient amongst 80 sample point.  
 The 95% C.I. range of payment\_type is considered to be wide between -1.04 to 2.01"

## [1] "The 95% confidence interval(C.I.) range for intercept has a wide range that is between .48 and 3.11\n The 95% C.I. range of trip\_minutes has a moderate range between .45 and .71 considering 80 samples.\n The 95% C.I. range of trip\_miles has a fairly tight range between .9 and 1.4 which means 5 differnet values will be repeated for ??2 coefficient amongst 80 sample point.\n The 95% C.I. range of payment\_type is considered to be wide between -1.04 to 2.01"

##Analysis 5  
  
trip\_minutes2=trip\_minutes^2  
trip\_miles2=trip\_miles^2

int=trip\_minutes\*trip\_miles

#Added squared term of trip\_minutes  
taxi.out1 = lm(fare ~ trip\_minutes+trip\_minutes2+trip\_miles+payment\_type, data=final\_sample)  
summary(taxi.out1)

##   
## Call:  
## lm(formula = fare ~ trip\_minutes + trip\_minutes2 + trip\_miles +   
## payment\_type, data = final\_sample)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.6409 -1.9604 -0.5488 1.1247 22.2476   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.033278 1.071881 0.031 0.9753   
## trip\_minutes 0.898339 0.151568 5.927 8.78e-08 \*\*\*  
## trip\_minutes2 -0.007268 0.003073 -2.365 0.0206 \*   
## trip\_miles 1.157316 0.145630 7.947 1.51e-11 \*\*\*  
## payment\_typeCredit Card 0.122432 0.904503 0.135 0.8927   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.845 on 75 degrees of freedom  
## Multiple R-squared: 0.8953, Adjusted R-squared: 0.8897   
## F-statistic: 160.3 on 4 and 75 DF, p-value: < 2.2e-16

#Accuracy of this model is 88.97%.

#Added squared term of trip\_miles  
taxi.out2 = lm(fare ~ trip\_minutes+trip\_miles+trip\_miles2+payment\_type, data=final\_sample)  
summary(taxi.out2)

##   
## Call:  
## lm(formula = fare ~ trip\_minutes + trip\_miles + trip\_miles2 +   
## payment\_type, data = final\_sample)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.8949 -1.2266 -0.7627 0.2564 22.5643   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.67816 0.90125 2.972 0.00398 \*\*   
## trip\_minutes 0.59745 0.07688 7.771 3.25e-11 \*\*\*  
## trip\_miles 0.36651 0.43531 0.842 0.40249   
## trip\_miles2 0.04407 0.02295 1.920 0.05863 .   
## payment\_typeCredit Card 0.52095 0.90228 0.577 0.56542   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.891 on 75 degrees of freedom  
## Multiple R-squared: 0.8928, Adjusted R-squared: 0.887   
## F-statistic: 156.1 on 4 and 75 DF, p-value: < 2.2e-16

#Accuracy of this model is 88.7%.

#Added squared term of trip\_minutes and trip\_miles  
taxi.out3 = lm(fare ~ trip\_minutes+trip\_minutes2+trip\_miles+trip\_miles2+payment\_type, data=final\_sample)  
summary(taxi.out3)

##   
## Call:  
## lm(formula = fare ~ trip\_minutes + trip\_minutes2 + trip\_miles +   
## trip\_miles2 + payment\_type, data = final\_sample)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.3105 -1.5640 -0.2247 1.1741 20.8475   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.471062 1.002290 0.470 0.639747   
## trip\_minutes 1.136494 0.155369 7.315 2.53e-10 \*\*\*  
## trip\_minutes2 -0.012412 0.003188 -3.894 0.000215 \*\*\*  
## trip\_miles -0.356460 0.440319 -0.810 0.420794   
## trip\_miles2 0.084995 0.023529 3.612 0.000550 \*\*\*  
## payment\_typeCredit Card -0.067212 0.841213 -0.080 0.936533   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.569 on 74 degrees of freedom  
## Multiple R-squared: 0.911, Adjusted R-squared: 0.905   
## F-statistic: 151.5 on 5 and 74 DF, p-value: < 2.2e-16

#Accuracy of this model is 90.5%.

#Added squared term of trip\_minutes and trip\_miles and interaction between trip\_minutes and trip\_miles by taking product.  
taxi.out4 = lm(fare ~ trip\_minutes+trip\_minutes2+trip\_miles+trip\_miles2+payment\_type+int, data=final\_sample)  
summary(taxi.out4)

##   
## Call:  
## lm(formula = fare ~ trip\_minutes + trip\_minutes2 + trip\_miles +   
## trip\_miles2 + payment\_type + int, data = final\_sample)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.0325 -0.6266 0.0136 0.5707 13.2853   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.619033 0.987224 2.653 0.00979 \*\*   
## trip\_minutes 0.436959 0.199649 2.189 0.03182 \*   
## trip\_minutes2 0.020163 0.007343 2.746 0.00760 \*\*   
## trip\_miles 0.958721 0.473869 2.023 0.04672 \*   
## trip\_miles2 0.146283 0.024286 6.023 6.34e-08 \*\*\*  
## payment\_typeCredit Card -0.138787 0.738630 -0.188 0.85148   
## int -0.103437 0.021558 -4.798 8.26e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.133 on 73 degrees of freedom  
## Multiple R-squared: 0.9323, Adjusted R-squared: 0.9268   
## F-statistic: 167.6 on 6 and 73 DF, p-value: < 2.2e-16

#Accuracy of this model is 92.68%.

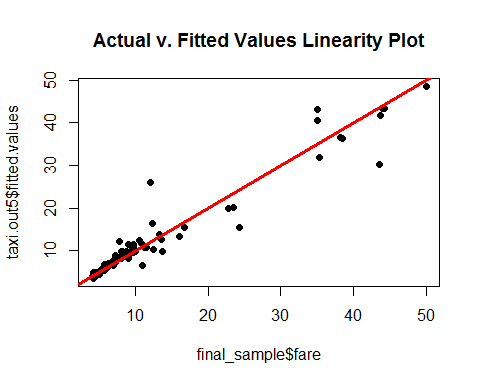
#Removed payment\_type from the above model.  
taxi.out5 = lm(fare ~ trip\_minutes+trip\_minutes2+trip\_miles+trip\_miles2+int, data=final\_sample)  
summary(taxi.out5)

##   
## Call:  
## lm(formula = fare ~ trip\_minutes + trip\_minutes2 + trip\_miles +   
## trip\_miles2 + int, data = final\_sample)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.0297 -0.6029 -0.0277 0.5405 13.2948   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.589219 0.968017 2.675 0.00920 \*\*   
## trip\_minutes 0.430588 0.195462 2.203 0.03071 \*   
## trip\_minutes2 0.020232 0.007286 2.777 0.00695 \*\*   
## trip\_miles 0.965178 0.469530 2.056 0.04335 \*   
## trip\_miles2 0.145993 0.024078 6.063 5.17e-08 \*\*\*  
## int -0.103355 0.021413 -4.827 7.27e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.112 on 74 degrees of freedom  
## Multiple R-squared: 0.9323, Adjusted R-squared: 0.9277   
## F-statistic: 203.8 on 5 and 74 DF, p-value: < 2.2e-16

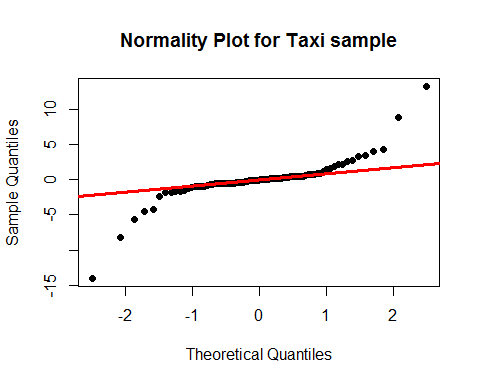
#Most significant fit with accuracy of 92.77%.

##Analysis 6

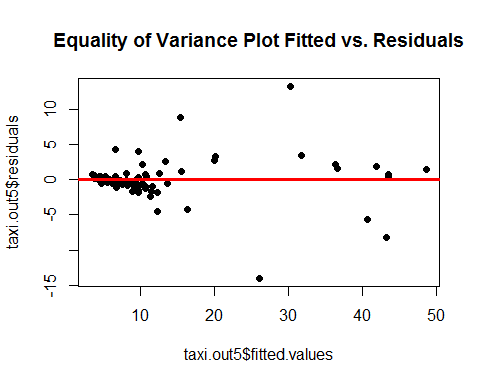
#According to the interactions considered in the step 5 analysis by squaring and multiplying tip\_minutes and trip\_miles,   
#We have a significant p-value of the intercept which means that we reject the null hypothesis and the beta0 coefficient has a value of 2.58(beta0=2.58).  
#The p-value of trip\_minutes is significant. With a per minute increase in trip\_minutes, the fare will increase by 43 cents (beta1=.43).   
#The p-value of trip\_minutes2 is significant. With a per minute increase in trip\_minutes2, the fare will increase by 2 cents(beta2=.02).   
#The p-value of trip miles is fairly significant. As the distance of the trip increases per mile, the fare increases by 96 cents(beta3=0.96).  
#The p-value of trip miles2 is very significant.. As the trip\_minutes2 increases per mile, the fare increases by 14 cents(beta4=0.14).  
#The p-value of int is very significant. As the int increases, the fare decreases by 10 cents(beta5=-0.10).  
#The p-value of payment\_type is no significan and hence, we fail to reject the null hypothesis. The mode of payment does not impact the cost of fare.  
#The accuracy of the model has improved by adding interactions from 88.3% to 92.77% with adjusted r squarred value of .9277.  
#This taxi.out5 model considers the trip minutes and its squarred term, trip\_miles and its squarred term and a product of trip\_minutes and trip\_miles."  
  
#Linearity  
plot(final\_sample$fare,taxi.out5$fitted.values,pch=19,main="Actual v. Fitted Values Linearity Plot")  
abline(0,1,col="red",lwd=3)



#Observation: The graph shows a linear distrbution as the lower and upper ends are close to normal distribution while some points in between show some deviation which are above and below the absolute line.  
  
#Normality  
qqnorm(taxi.out5$residuals,pch=19,main="Normality Plot for Taxi sample")  
qqline(taxi.out5$residuals,col="red",lwd=3)



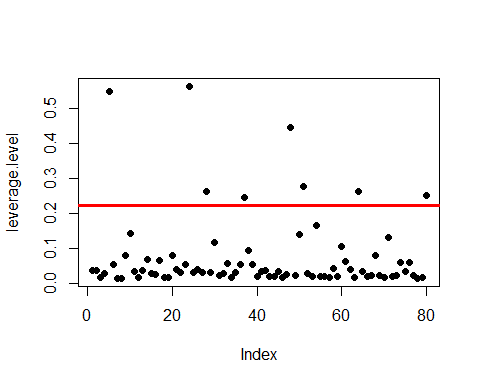
#Observation: The graph shows a large mid portion aligning with the absolute line and the tail to the left is below the line, and the one to the right is above.  
#The plot shows a normal distribution and at both the ends some deviation is prominent.  
  
  
#Equality of Variances  
plot(taxi.out5$fitted.values,taxi.out5$residuals,pch=19,main="Equality of Variance Plot Fitted vs. Residuals")  
abline(0,0,col="red",lwd=3)



#Observation:The nature of this plot show heteroscedasticity which means as the fitted values increases, the range of the residuals also increases.  
#The equality of variance plot shows most of the data to the left around the absolute line, the mid portion shows scattered data and the right side shows some data points around the line.   
#This is an ambiguous plot and does not conform with equality of variances.  
  
#Since we are not dealing with time series data we do not need to check for independence.

##Analysis 7

#Determining the high leverg epoints  
leverage.level=hat(model.matrix(taxi.out5))  
leverageplot=plot(leverage.level,pch=19)  
abline(3\*mean(leverage.level),0,col="red",lwd=3)



#There are 8 high leverage points in my sample.  
  
#A dataframe of the leverage points is created below  
lev.point = final\_sample[leverage.level>(3\*mean(leverage.level)),]  
leverage.point = final\_sample[leverage.level>(3\*mean(leverage.level)),1]  
  
  
#Removing the 8 high leverage points from the sample and creating a new sample.  
final\_sample1=final\_sample[-which(final\_sample$taxi\_id %in% leverage.point),]  
attach(final\_sample1)

## The following objects are masked from final\_sample:  
##   
## extras, fare, payment\_type, taxi\_id, tips, trip\_miles,  
## trip\_minutes, trip\_total  
##   
## The following objects are masked from taxis:  
##   
## extras, fare, payment\_type, taxi\_id, tips, trip\_miles,  
## trip\_total

taxi.out6 = lm(fare ~ trip\_minutes+I(trip\_minutes^2)+trip\_miles+I(trip\_miles^2)+trip\_minutes:trip\_miles, data=final\_sample1)  
summary(taxi.out6)

##   
## Call:  
## lm(formula = fare ~ trip\_minutes + I(trip\_minutes^2) + trip\_miles +   
## I(trip\_miles^2) + trip\_minutes:trip\_miles, data = final\_sample1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.0149 -0.4396 -0.1190 0.1639 7.6709   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.670817 0.618520 5.935 1.20e-07 \*\*\*  
## trip\_minutes -0.031651 0.175765 -0.180 0.858   
## I(trip\_minutes^2) 0.048830 0.009611 5.081 3.31e-06 \*\*\*  
## trip\_miles 1.928607 0.418599 4.607 1.92e-05 \*\*\*  
## I(trip\_miles^2) 0.218071 0.025190 8.657 1.77e-12 \*\*\*  
## trip\_minutes:trip\_miles -0.198156 0.029684 -6.676 6.09e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.681 on 66 degrees of freedom  
## Multiple R-squared: 0.9492, Adjusted R-squared: 0.9454   
## F-statistic: 246.7 on 5 and 66 DF, p-value: < 2.2e-16

#Interpretation: After applying the best fit model as in analysis 5 and removing the high leverage points-  
#The model has all p-values to be significant except the trip\_minutes.  
#Since the squared term of trip\_minutes relies on the unsquared term it is necessary for it to be in the model.  
#This model shows an improved accuracy 0f 94.54% which is approximately 2% higher than the previous model with 88.3% and has an adjusted R squared value of .9454.  
  
  
##Analysis8

#Sample data  
#Performing the similar analysis as done in the preprocessing part at th begining.   
  
#100 random samples by adding 5 to the U number as a random number seed  
#Sample data  
set.seed(74470962)  
new.sample.taxis = taxis[sample(1:nrow(taxis),100, replace=F),]  
  
##Preprocessing 3  
#Cleansing the random data set  
  
sum(new.sample.taxis$tolls)

## [1] 0

#The tolls column adds up to 0.  
new.sample.taxis$tolls = NULL  
#Removing Column tolls since it has no impact as each row of tolls has 0 value.   
  
new.filter.sample= subset (new.sample.taxis, new.sample.taxis$trip\_seconds!=0 & new.sample.taxis$trip\_miles!=0.00)  
#Removing all rows where trip secconds and trip miles are 0.   
  
new.filter.sample$trip\_minutes = new.filter.sample$trip\_seconds/60  
#Converting the trip\_seconds column to trip\_minutes.  
  
new.filter.sample$trip\_seconds=NULL  
#Removing trip\_seconds column from the filter.sample.  
  
new\_order\_sample = c("taxi\_id","trip\_minutes","trip\_miles","fare","tips","extras","trip\_total","payment\_type")  
new\_final\_sample=new.filter.sample[,new\_order\_sample]  
#Replacing the trip\_seconds column with trip\_minutes by placing it at the same position.  
  
str(new\_final\_sample)

## 'data.frame': 72 obs. of 8 variables:  
## $ taxi\_id : int 8368 4951 2336 2625 3454 1829 4269 2916 3015 978 ...  
## $ trip\_minutes: num 16 13 19 9 574 5 10 23 8 3 ...  
## $ trip\_miles : num 0.1 3.4 0.1 3.2 1.3 1 0.1 13.2 1.1 0.5 ...  
## $ fare : num 10.75 12.25 12.75 9.65 6.5 ...  
## $ tips : num 0 1 0 0 0 2 0 0 0 3 ...  
## $ extras : num 1 1 1 0 0 0 0 5 0 0 ...  
## $ trip\_total : num 11.75 14.25 13.75 9.65 6.5 ...  
## $ payment\_type: Factor w/ 3 levels "Cash","Credit Card",..: 1 2 1 1 1 2 1 1 1 2 ...

#Displaying information about the structure of the final data sample  
  
attach(new\_final\_sample)

## The following objects are masked from final\_sample1:  
##   
## extras, fare, payment\_type, taxi\_id, tips, trip\_miles,  
## trip\_minutes, trip\_total

## The following objects are masked from final\_sample:  
##   
## extras, fare, payment\_type, taxi\_id, tips, trip\_miles,  
## trip\_minutes, trip\_total

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

#Attaching the final sample  
  
#Linear Regression Model  
taxi.out7 = lm(fare ~ trip\_minutes+I(trip\_minutes^2)+trip\_miles+I(trip\_miles^2)+trip\_minutes:trip\_miles, data=new\_final\_sample)  
summary(taxi.out7)

##   
## Call:  
## lm(formula = fare ~ trip\_minutes + I(trip\_minutes^2) + trip\_miles +   
## I(trip\_miles^2) + trip\_minutes:trip\_miles, data = new\_final\_sample)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.4169 -1.0452 -0.1823 0.7739 19.2274   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.0880541 0.7554373 1.440 0.154510   
## trip\_minutes 0.8198251 0.0834422 9.825 1.53e-14 \*\*\*  
## I(trip\_minutes^2) -0.0012802 0.0001256 -10.192 3.50e-15 \*\*\*  
## trip\_miles 0.7658157 0.2572744 2.977 0.004071 \*\*   
## I(trip\_miles^2) 0.1200427 0.0225171 5.331 1.27e-06 \*\*\*  
## trip\_minutes:trip\_miles -0.0597365 0.0146822 -4.069 0.000129 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.819 on 66 degrees of freedom  
## Multiple R-squared: 0.9497, Adjusted R-squared: 0.9459   
## F-statistic: 249.2 on 5 and 66 DF, p-value: < 2.2e-16

#Observation: The above regression model with a new sample shows all p-values to be significant except the beta0 intercept value.  
#The accuracy of the model is very good at 94.59% with an adjusted R squared value of .9459.11"