Regression Project for QMB-6304

Saahil Rasheed

20/10/2019

#U85555983

**#Pre-processing**

#Loading all the libraries here to keep the rest of the code clean.

rm(list=ls())  
library(rio)  
library(tidyverse)

## -- Attaching packages -------------------------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1 v purrr 0.3.2  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 0.8.3 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts ----------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(corrplot)

## corrplot 0.84 loaded

library(Hmisc)

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':  
##   
## src, summarize

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

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

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

## The following object is masked from 'package:purrr':  
##   
## some

library(moments)

#Loading the Chicago cabs data set and storing it in a data frame called “taxi”.

taxi=import("6304 Regression Project Data.csv")  
colnames(taxi)=tolower(make.names(colnames(taxi)))  
attach(taxi)  
names(taxi)

## [1] "taxi\_id" "trip\_seconds" "trip\_miles" "fare"   
## [5] "tips" "tolls" "extras" "trip\_total"   
## [9] "payment\_type"

Setting seed = U Number) and take a random sample of 100

set.seed(85555983)  
my.taxi=taxi[sample(1:nrow(taxi),100,replace=FALSE),]

Cleaning the dataset by removing records where Trip Seconds and Trip Miles are 0. I did this because when both of the above conditions were true, some records had a fare price. I expect the fare to be nill when a taxi trip has lasted for 0 seconds or if the trip lasted for 1 mile.

my.taxi = subset(my.taxi,trip\_seconds!=0 & trip\_miles!=0)  
attach(my.taxi)

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

#my.taxi = filter(my.taxi, trip\_seconds!=0)

**#Analysis**

**#Analysis Part 1**

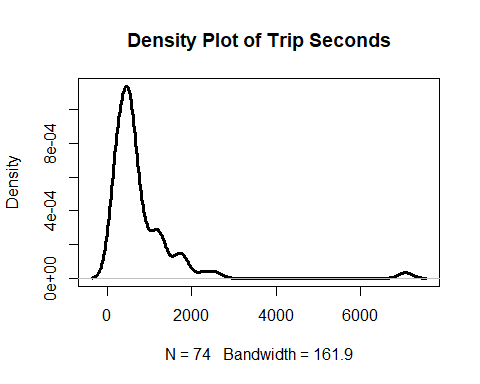
Summaries of continuous variables and density plots along with statistical analysis for each.

**##Trip Seconds:**

summary(trip\_seconds)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 120.0 315.0 540.0 775.1 885.0 7080.0

plot(density(trip\_seconds),lwd=3,main="Density Plot of Trip Seconds")



sd(trip\_seconds)

## [1] 906.6199

quantile(trip\_seconds,probs=seq(0,1,.25))

## 0% 25% 50% 75% 100%   
## 120 315 540 885 7080

skewness(trip\_seconds)

## [1] 4.834741

kurtosis(trip\_seconds)

## [1] 32.94752

**Interpretation:**

Trip Seconds has a highly right-skewed distribution with a mean of 775 trip seconds. This is proved by:

* Skewness= 4.83, which is much greater than 0. (Skewness of 0 means normal distribution)
* Mean (775) > Median (540) [Right Skewed if Mean is greater than Median]

The Kurtosis is 32.94 which signifies that most of the data is concentrated on the tails. In Trip Seconds, most of the data is concentrated on the lower tail from 0 to 2000 trip seconds. (left-hand side of the density plot)

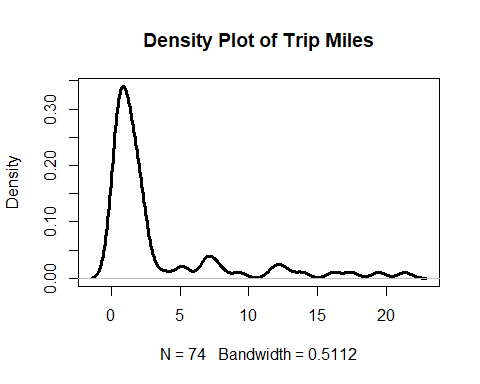
The Standard Deviation is 906.62. It shows the deviation of Trip Seconds from the mean value of Trip Seconds. This is a high value indicating that the values are diffused or spread over a wide range. (Wide Confidence Interval) This is also proved by the quantile and density plot which shows the range of all the values in this variable.

**##Trip Miles:**

summary(trip\_miles)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.100 0.700 1.400 3.361 2.500 21.400

plot(density(trip\_miles),lwd=3,main="Density Plot of Trip Miles")



#standard Deviation of Trip Miles:  
sd(trip\_miles)

## [1] 4.87322

quantile(trip\_miles,probs=seq(0,1,.25))

## 0% 25% 50% 75% 100%   
## 0.1 0.7 1.4 2.5 21.4

#Skewness of Trip Miles:  
skewness(trip\_miles)

## [1] 2.144396

#Kurtosis of Trip Miles  
kurtosis(trip\_miles)

## [1] 6.815007

**Interpretation of Trip Miles:**

* Trip Miles has a right-skewed distribution with a mean of 3.3 trip miles. This is proved by: Skewness= 2.14, which is greater than 0. (Skewness of 0 means normal distribution)
* Mean (3.3) > Median (1.4)

The Kurtosis is 6.81 which signifies that most of the data is concentrated on the tails (Kurtosis greater than 3). In Trip Miles, most of the data is concentrated on the lower tail from 0 to 4 miles.

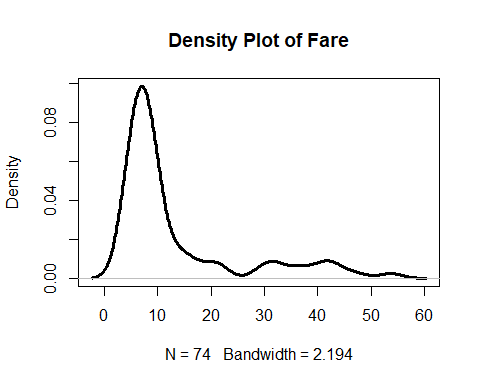
The Standard Deviation is 4.87 (miles) It shows the deviation of Trip Miles from the mean value of Trip Miles This is a high value indicating that the values are diffused or spread over a wide range. (Wide Confidence Interval). Because of this wide Confidence Interval, we are not confident when estimating the value of the mean. This is also proved by the quantile and density plot which shows the range of all the values in this variable.

**##Fare**

summary(fare)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.25 6.25 8.25 13.44 13.97 53.75

plot(density(fare),lwd=3,main="Density Plot of Fare")



sd(fare)

## [1] 12.20098

quantile(fare,probs=seq(0,1,.25))

## 0% 25% 50% 75% 100%   
## 4.250 6.250 8.250 13.975 53.750

skewness(fare)

## [1] 1.734626

kurtosis(fare)

## [1] 4.823407

**Interpretation of Fare:**

Fare has a right-skewed distribution with a mean of 13.44 dollars (assuming fare to be in dollars). This right-skewdness is proved by:

* Skewness= 1.74, which is greater than 0. (Skewness of 0 means normal distribution)
* Mean (13.44) > Median (8.25)

The Kurtosis is 4.82 which signifies that most of the data is concentrated on the tails (Kurtosis greater than 3). In Fare, most of the data is concentrated on the lower tail from 0 to 20 dollars.

The Standard Deviation is 12.2 (dollars). It shows the deviation of Fare from the mean value of Fare ($13.44) This is a high value indicating that the values are diffused or spread over a wide range. This is also proved by the quantile and density plot which shows the range of all the values in this variable.

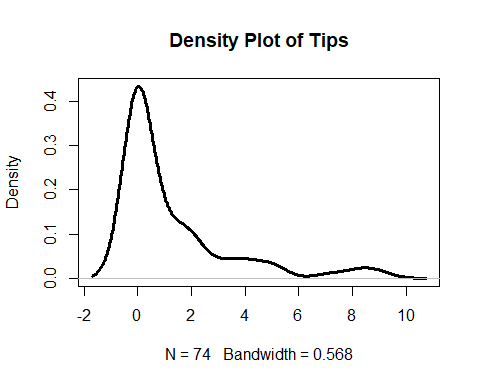
The high standard deviation also means that Fare has a Wide Confidence Interval. Because of this wide Confidence Interval, we are not very confident when estimating the value of the mean.

**##Tips**

summary(tips)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 1.285 2.000 9.050

plot(density(tips),lwd=3,main="Density Plot of Tips")



sd(tips)

## [1] 2.173938

quantile(tips,probs=seq(0,1,.25))

## 0% 25% 50% 75% 100%   
## 0.00 0.00 0.00 2.00 9.05

skewness(tips)

## [1] 2.03054

kurtosis(tips)

## [1] 6.618769

**Interpretation of Tips:**

Tips is right-skewed with a mean of 1.28 dollars. This right-skewdness is proved by:

* Skewness= 2.03, which is greater than 0. (Skewness of 0 means normal distribution)
* Mean (1.28) > Median (0.0)

The Kurtosis is 6.61 which signifies that most of the data is concentrated on the tails (Kurtosis greater than 3). In Fare, most of the data is concentrated on the lower tail from 0 to 2 dollars.

The Standard Deviation is 2.17 (dollars). It shows the deviation of Tips from the mean value of Tips ($1.28) This is a high value indicating that the values are diffused or spread over a wide range. This is also proved by the quantile and density plot which shows the range of all the values in this variable.

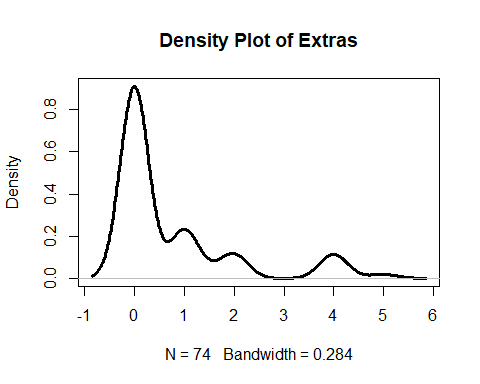
This high standard deviation also means that Tips has a Wide Confidence Interval. Because of this wide Confidence Interval, we are not very confident when estimating the value of the mean of Tips (Tips mean can vary).

**##Extras**

summary(extras)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.7365 1.0000 5.0000

plot(density(extras),lwd=3,main="Density Plot of Extras")



sd(extras)

## [1] 1.277366

quantile(extras,probs=seq(0,1,.25))

## 0% 25% 50% 75% 100%   
## 0 0 0 1 5

skewness(extras)

## [1] 1.850491

kurtosis(extras)

## [1] 5.42884

**Interpretation of Extras:**

Extras has a right-skewed distribution with a mean of 0.73 dollars. This right-skewdness is proved by:

* Skewness= 1.85, which is slightly greater than 0. (Skewness of 0 means normal distribution)
* Mean (0.73) > Median (0.0)

The Kurtosis is 5.42 which signifies that most of the data is concentrated on the tails (Kurtosis greater than 3). In Fare, most of the data is concentrated on the lower tail from 0 to 1 dollar.

The Standard Deviation is 1.27 (dollars). It shows the deviation of Extras from the mean value of Extras ($0.73) This is a relatively high value indicating that the values are diffused or spread over a wide range. This is also proved by the quantile and density plot which shows the range of all the values in this variable.

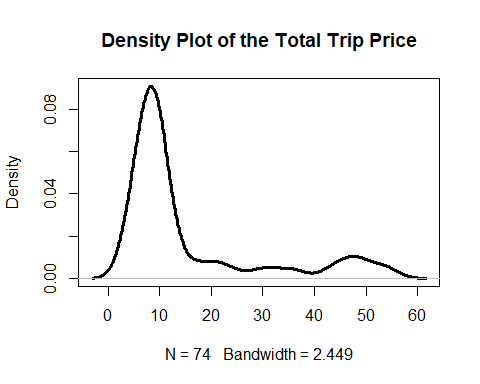
This high standard deviation also means that Extras has a Wide Confidence Interval. Because of this wide Confidence Interval, we are not very confident when estimating the value of the mean of Extras (Extras mean can vary).

**##Trip Total**

summary(trip\_total)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.25 7.30 9.25 15.46 15.93 54.30

plot(density(trip\_total),lwd=3,main="Density Plot of the Total Trip Price")



sd(trip\_total)

## [1] 14.12863

quantile(trip\_total,probs=seq(0,1,.25))

## 0% 25% 50% 75% 100%   
## 4.250 7.300 9.250 15.925 54.300

skewness(trip\_total)

## [1] 1.651892

kurtosis(trip\_total)

## [1] 4.297826

**Interpretation of Trip Total:**

Trip Total has a right-skewed distribution with a mean of 15.46 dollars. This right-skewdness is proved by:

* Skewness= 1.65, which is slightly greater than 0. (Skewness of 0 means normal distribution)
* Mean (15.46) > Median (9.25)

The Kurtosis is 4.29 which signifies that most of the data is concentrated on the tails (Kurtosis greater than 3). In Fare, most of the data is concentrated on the lower tail from 0 to 20 dollars.

The Standard Deviation is 14.12 (dollars). It shows the deviation of Trip Total from the mean value of Trip Total ($15.46) This is a relatively high value indicating that the values are diffused or spread over a wide range. This is also proved by the quantile and density plot which shows the range of all the values in this variable.

This high standard deviation also means that Trip Total has a Wide Confidence Interval. Because of this wide Confidence Interval, we are not very confident when estimating the value of the mean of Total Trip price (Total Trip mean can vary).

**#Analysis Part 2**

Making Payment Type (payment\_type) a factor variable:

my.taxi$payment\_type=as.factor(my.taxi$payment\_type)  
str(my.taxi)

## 'data.frame': 74 obs. of 9 variables:  
## $ taxi\_id : int 1419 2318 3869 6819 4444 605 6523 5556 7674 3404 ...  
## $ trip\_seconds: int 480 1320 360 840 1740 540 600 480 180 600 ...  
## $ trip\_miles : num 1.9 12.2 0.8 2.9 1 1.2 2.3 1.8 0.4 0.1 ...  
## $ fare : num 8.25 31.75 5.5 10.75 43.5 ...  
## $ tips : num 0 7.15 0 0 5 0 0 0 0 0 ...  
## $ tolls : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ extras : num 0 4 0 0 0 0 1 0 0 1 ...  
## $ trip\_total : num 8.25 42.9 5.5 10.75 48.5 ...  
## $ payment\_type: Factor w/ 3 levels "Cash","Credit Card",..: 1 2 1 1 2 1 1 1 1 1 ...

Creating a table which displays the number of cases in each level of Payment Type: The levels are: Cash (Base Class), Credit Card and Other

table(payment\_type)

## payment\_type  
## Cash Credit Card Other   
## 40 33 1

**#Analysis Part 3**

Copying the continuous variables to a new data object called corr\_matrix.

corr\_matrix=subset(my.taxi,select=c("trip\_seconds","trip\_miles","fare","tips","extras","trip\_total"))

Correlation Matrix with p values:

cm=rcorr(as.matrix(corr\_matrix))  
cm

## trip\_seconds trip\_miles fare tips extras trip\_total  
## trip\_seconds 1.00 0.69 0.76 0.21 0.19 0.70  
## trip\_miles 0.69 1.00 0.82 0.22 0.38 0.78  
## fare 0.76 0.82 1.00 0.50 0.48 0.98  
## tips 0.21 0.22 0.50 1.00 0.53 0.63  
## extras 0.19 0.38 0.48 0.53 1.00 0.59  
## trip\_total 0.70 0.78 0.98 0.63 0.59 1.00  
##   
## n= 74   
##   
##   
## P  
## trip\_seconds trip\_miles fare tips extras trip\_total  
## trip\_seconds 0.0000 0.0000 0.0776 0.1114 0.0000   
## trip\_miles 0.0000 0.0000 0.0628 0.0008 0.0000   
## fare 0.0000 0.0000 0.0000 0.0000 0.0000   
## tips 0.0776 0.0628 0.0000 0.0000 0.0000   
## extras 0.1114 0.0008 0.0000 0.0000 0.0000   
## trip\_total 0.0000 0.0000 0.0000 0.0000 0.0000

**Interpretation of Correlation Matrix:**

The Correlation Matrix gives the correlation (Multiple R or Correlation Coefficients) between each pair of the continuous independent variables that we have taken in the analysis.

There are two matrices created. The First matrix shows the Correlation Coefficient values. The second matrix shows the p-values for each Correlation coefficient. This signifies the the amount of confidence we have in determining whether the values in the first matrix are equal to or not equal to zero. (Hypothesis Test)

In Layman’s terms, the Correlation Matrix is used to display a matrix with a set of values, where each value is the measure of how closely two variables move in relation to one another. In other words,it means how much a single variable is affected by change in another variable.

For example, [to explain this to a Non-statistician]

i) Correlation between Trip Miles and Fare = 0.82. This is a positive and pretty strong correlation (approaching 1) This means that with a positive change (increase) in Trip Mile, there is a high probability of Fare price increasing as well. (Also means if there a decrease in Trip Mile, it is highly likely that fare will decrease as well)

1. Correlation between Extras and Trip Seconds = 0.19. This is a low positive correlation (approaching 0) and this means that if there is a change in Trip Seconds, there is a low probability of a positive change in Extras and vice versa. In other words, it is most likely there will be no change in the latter variable for any change in the former variable.

[A correlation coefficient of 0 shows no relationship between the change in two variables.]

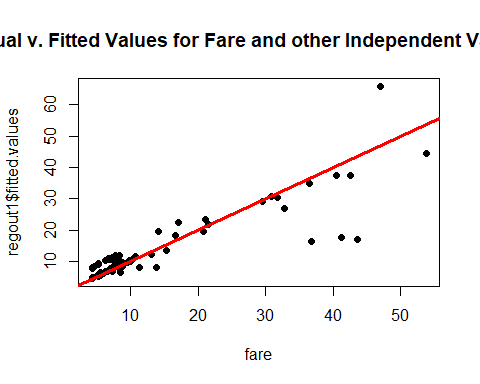
**#Analysis Part 4**

Regression Model using trip\_seconds, trip\_miles and payment\_type (Fare as dependent variable and 95% Confidence interval on beta coefficients):

regout1=lm(fare~trip\_seconds+trip\_miles+payment\_type,data=my.taxi)  
summary(regout1)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + payment\_type,   
## data = my.taxi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.9336 -3.5011 -0.4310 0.3342 26.6267   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.367845 1.141603 2.950 0.00434 \*\*   
## trip\_seconds 0.005026 0.001116 4.504 2.64e-05 \*\*\*  
## trip\_miles 1.383725 0.204402 6.770 3.47e-09 \*\*\*  
## payment\_typeCredit Card 3.376683 1.447118 2.333 0.02255 \*   
## payment\_typeOther 1.546784 6.387848 0.242 0.80939   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.132 on 69 degrees of freedom  
## Multiple R-squared: 0.7613, Adjusted R-squared: 0.7474   
## F-statistic: 55 on 4 and 69 DF, p-value: < 2.2e-16

plot(fare,regout1$fitted.values,pch=19,main="Actual v. Fitted Values for Fare and other Independent Variables")  
abline(0,1,col="red",lwd=3)



regout1.coefficients=cbind("Beta Coefficients"=coef(regout1),confint(regout1))  
regout1.coefficients

## Beta Coefficients 2.5 % 97.5 %  
## (Intercept) 3.367845419 1.090410365 5.645280473  
## trip\_seconds 0.005025864 0.002799801 0.007251926  
## trip\_miles 1.383724937 0.975954537 1.791495336  
## payment\_typeCredit Card 3.376683497 0.489761482 6.263605511  
## payment\_typeOther 1.546784186 -11.196623245 14.290191617

**Interpretation:**

Adjusted R-square value of 0.7474 or 74%. This means that the model explains roughly 74% of the variation in the fare price.

The Regression Equation for the model created above is, Fare(Y) = 3.36 + 0.005026 m1+ 1.383 m2+ 3.376 m3+ 1.546 m4, Where m1,m2,m3,m4 are the slopes of trip\_seconds, trip\_miles,Credit Card and Other variables respectively.

• **Beta Coefficient 1 or trip seconds**: It can be said that with every second increase in the “trip seconds” variable(m1), the fare(y) for taxi is expected to increase by 0.005 dollars. (Assuming it to be in $ here) -The p value for this Beta Coefficient is 2.64e-05 (less than 5%), which means that this coefficient is significant, and we can be confident that the actual value for trip seconds isn’t 0. (Rejecting the Null Hypothesis)

• **Beta Coefficient 2 or trip miles:** It can be said that with every single increase (+1) in the “trip miles” variable, the fare is expected to increase by $1.38 - The p value for this Beta Coefficient is 3.47e-09 which is much lesser than 5%. In other words, this coefficient is significant, and we can be confident that the actual value for trip miles isn’t 0. (Rejecting the Null Hypothesis)

• **Beta Coefficient 3 or Payment Type:Credit Card**: The base value for factor variable, Payment Type is “Cash”. It can be said that with every dollar increase (+1) in Credit Card price, the fare is expected to increase by $3.37 –The p value for this Beta Coefficient is 0.02255 (less than 5%), which means that this coefficient is significant, and we can be confident that the actual value isn’t 0. (Rejecting the Null Hypothesis)

• **Beta Coefficient 4 or Payment Type:Other**: It can be said that with every dollar increase (+1) in payment by Other means, the fare is expected to increase by $1.54 –The p value for this Beta Coefficient is 0.80939 (more than 5%), which means that this coefficient is not significant, and we cannot be confident that the actual value isn’t 0. (Failing to reject the Null Hypothesis)

• **Beta Coefficient 0 or Payment Type: Cash (For the intercept):** β0 value is 3.367845, which basically means that if the taxi trip has no trip seconds, no trip miles, no payment through Credit Card or “Other” means, the fare is expected to be $3.367 to be paid by Cash (Base level of factor variable). In other words, a person pays $3.36 for a taxi trip that did not happen.

regout1.coefficients=cbind("Beta Coefficients"=coef(regout1),confint(regout1))  
regout1.coefficients

## Beta Coefficients 2.5 % 97.5 %  
## (Intercept) 3.367845419 1.090410365 5.645280473  
## trip\_seconds 0.005025864 0.002799801 0.007251926  
## trip\_miles 1.383724937 0.975954537 1.791495336  
## payment\_typeCredit Card 3.376683497 0.489761482 6.263605511  
## payment\_typeOther 1.546784186 -11.196623245 14.290191617

**Confidence Intervals:**

The ranges for the Confidence Intervals (C.I) of Trip Seconds, Payment type:Other, Payment type:Credit Card and Trip Miles is pretty wide. Thus, we cannot say with confidence that the beta coefficient values for each of those variables are as stated, because of the wide range of the C.I’s. [even though the p-values are significant for all of them.

The Confidence Interval for Trip Seconds is tight. This means that we can estimate with confidence the beta coefficient value for Trip Seconds.

(It’s always good to cross-check with Confidence Intervals)

**#Analysis Part 5**

Investigating different Regression Models, trying to improve the model in Analysis Part 4. I explored through ten different regression models, trying out different combinations of independent variables that I took in Analysis Part 5.

Interactions between Trip seconds and Trip Miles is denoted by “trip\_seconds:trip\_miles”.

**##Regression Model 2**: All independent variables and Interaction between “Trip seconds” and “Trip Miles”

regout2=lm(fare~trip\_seconds+trip\_miles+payment\_type+trip\_seconds:trip\_miles,data=my.taxi)  
summary(regout2)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + payment\_type +   
## trip\_seconds:trip\_miles, data = my.taxi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.8854 -2.0658 -0.2259 1.8553 14.0247   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.694e+00 1.014e+00 -2.656 0.00983 \*\*   
## trip\_seconds 1.740e-02 1.545e-03 11.263 < 2e-16 \*\*\*  
## trip\_miles 1.980e+00 1.523e-01 13.003 < 2e-16 \*\*\*  
## payment\_typeCredit Card 1.281e+00 1.001e+00 1.280 0.20499   
## payment\_typeOther -1.093e+01 4.513e+00 -2.421 0.01816 \*   
## trip\_seconds:trip\_miles -7.897e-04 8.614e-05 -9.168 1.68e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.131 on 68 degrees of freedom  
## Multiple R-squared: 0.8932, Adjusted R-squared: 0.8854   
## F-statistic: 113.8 on 5 and 68 DF, p-value: < 2.2e-16

**Interpretation:** Improved Adjusted R-square value of 0.88 or 88%. This means that the model explains roughly 88% of the variation in the fare price.

This is a big improvement from the previous model and I’ve decided to keep the interaction variable due to the fact that it is significant (except for payment\_type Credit Card level.)

**##Regression Model 3:** Keeping the second model and applying square transform on Trip Seconds. All independent variables+Interaction+Trip Seconds Square

regout3=lm(fare~trip\_seconds+trip\_miles+payment\_type+I(trip\_seconds^2)+trip\_seconds:trip\_miles,data=my.taxi)  
summary(regout3)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + payment\_type +   
## I(trip\_seconds^2) + trip\_seconds:trip\_miles, data = my.taxi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.7865 -2.1191 -0.3323 2.2702 13.7197   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.887e+00 9.777e-01 -2.953 0.00434 \*\*   
## trip\_seconds 2.085e-02 2.002e-03 10.414 1.21e-15 \*\*\*  
## trip\_miles 9.670e-01 4.207e-01 2.299 0.02465 \*   
## payment\_typeCredit Card 1.355e+00 9.628e-01 1.407 0.16393   
## payment\_typeOther -1.297e+01 4.410e+00 -2.940 0.00450 \*\*   
## I(trip\_seconds^2) -1.873e-06 7.290e-07 -2.569 0.01242 \*   
## trip\_seconds:trip\_miles -1.553e-04 2.604e-04 -0.596 0.55302   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.97 on 67 degrees of freedom  
## Multiple R-squared: 0.9028, Adjusted R-squared: 0.8941   
## F-statistic: 103.7 on 6 and 67 DF, p-value: < 2.2e-16

**Interpretation:**

Slightly improved Adjusted R-square value of 0.89 or 89%. This means that the model explains roughly 88% of the variation in the fare price.

This is a small improvement from the previous model but the interaction variable along with Credit Card level are insignificant in this model

**##Regression Model 4:** Previous Model and applying Squared Transform on Trip Miles All independent variables+Interaction+Trip Seconds Squared+Trip Miles Squared

regout4=lm(fare~trip\_seconds+trip\_miles+payment\_type+I(trip\_seconds^2)+I(trip\_miles^2)+trip\_seconds:trip\_miles,data=my.taxi)  
summary(regout4)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + payment\_type +   
## I(trip\_seconds^2) + I(trip\_miles^2) + trip\_seconds:trip\_miles,   
## data = my.taxi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.2517 -1.3540 -0.2031 1.6967 8.5710   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.468e-01 6.702e-01 -0.965 0.3380   
## trip\_seconds 1.642e-02 1.367e-03 12.007 < 2e-16 \*\*\*  
## trip\_miles 7.269e-01 2.721e-01 2.672 0.0095 \*\*   
## payment\_typeCredit Card 5.934e-01 6.250e-01 0.949 0.3459   
## payment\_typeOther -1.246e+00 3.083e+00 -0.404 0.6875   
## I(trip\_seconds^2) 3.132e-06 6.948e-07 4.508 2.75e-05 \*\*\*  
## I(trip\_miles^2) 2.060e-01 2.108e-02 9.773 1.88e-14 \*\*\*  
## trip\_seconds:trip\_miles -2.298e-03 2.761e-04 -8.324 6.96e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.557 on 66 degrees of freedom  
## Multiple R-squared: 0.9603, Adjusted R-squared: 0.9561   
## F-statistic: 227.9 on 7 and 66 DF, p-value: < 2.2e-16

**Interpretation:**

a BIG improvement in Adjusted R-square value of 0.9561 or 95% from 89%. This means that the model explains roughly 95% of the variation in the fare price.

This is a big improvement from the previous model with all independent variables significant except for Credit Card level and Other level along with the Intercept (Cash). All the levels in payment\_type are insignificant.

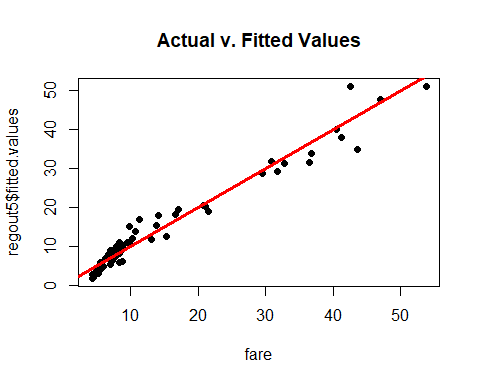
**##Regression Model 5**: Previous Model but excluding the Payment Type

All independent variables+Interaction+Trip Seconds Squared+Trip Miles Squared-Payment Type

regout5=lm(fare~trip\_seconds+trip\_miles+I(trip\_seconds^2)+I(trip\_miles^2)+trip\_seconds:trip\_miles,data=my.taxi)  
summary(regout5)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + I(trip\_seconds^2) +   
## I(trip\_miles^2) + trip\_seconds:trip\_miles, data = my.taxi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.5870 -1.3672 -0.3692 1.5115 8.7270   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.506e-01 5.967e-01 -0.588 0.55870   
## trip\_seconds 1.619e-02 1.210e-03 13.387 < 2e-16 \*\*\*  
## trip\_miles 7.586e-01 2.685e-01 2.825 0.00619 \*\*   
## I(trip\_seconds^2) 3.348e-06 6.245e-07 5.360 1.07e-06 \*\*\*  
## I(trip\_miles^2) 2.128e-01 1.900e-02 11.198 < 2e-16 \*\*\*  
## trip\_seconds:trip\_miles -2.391e-03 2.512e-04 -9.516 3.99e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.541 on 68 degrees of freedom  
## Multiple R-squared: 0.9596, Adjusted R-squared: 0.9566   
## F-statistic: 322.9 on 5 and 68 DF, p-value: < 2.2e-16

plot(fare,regout5$fitted.values,pch=19,main="Actual v. Fitted Values")  
abline(0,1,col="red",lwd=3)



**Interpretation:**

A small improvement in Adjusted R-square value of 0.9566 or 95.66% from 95.61%. This means that the model explains roughly 95% of the variation in the fare price.

This is a big and important improvement from the previous model because we have removed payment type from consideration in our analysis. By doing this, we are reducing the risk of overfitting the model. All the variables except the intercept are significant. This is currently our best model which gets the best fit to the sample.

**##Regression Model 6:** Investigating further, by removing Trip Miles Squared Transform. TS+TM++Interaction+Trip Seconds Squared

regout6=lm(fare~trip\_seconds+trip\_miles+I(trip\_seconds^2)+trip\_seconds:trip\_miles,data=my.taxi)  
summary(regout6)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + I(trip\_seconds^2) +   
## trip\_seconds:trip\_miles, data = my.taxi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.5868 -1.7197 -0.5414 1.7678 15.9626   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.751e+00 9.768e-01 -1.792 0.07749 .   
## trip\_seconds 1.885e-02 1.986e-03 9.493 3.82e-14 \*\*\*  
## trip\_miles 1.214e+00 4.443e-01 2.733 0.00795 \*\*   
## I(trip\_seconds^2) -1.406e-06 7.670e-07 -1.833 0.07116 .   
## trip\_seconds:trip\_miles -2.702e-04 2.764e-04 -0.978 0.33170   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.255 on 69 degrees of freedom  
## Multiple R-squared: 0.8851, Adjusted R-squared: 0.8784   
## F-statistic: 132.8 on 4 and 69 DF, p-value: < 2.2e-16

**Interpretation:**

By investigating further, there is a decrease in Adjusted R-square value which is 0.87 or 87%. This means that the model explains roughly 87% of the variation in the fare price.

**##Regression Model 7:** Exploring further by removing payment type, TM2 and TM TS+Interaction+Trip Seconds Squared+Trip Miles Squared-Payment Type-tm2-tm

regout7=lm(fare~trip\_seconds+I(trip\_seconds^2)+trip\_seconds:trip\_miles,data=my.taxi)  
summary(regout7)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + I(trip\_seconds^2) + trip\_seconds:trip\_miles,   
## data = my.taxi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.1513 -2.3135 -0.6952 2.0144 15.7577   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.481e+00 1.016e+00 -1.458 0.149   
## trip\_seconds 2.223e-02 1.625e-03 13.673 < 2e-16 \*\*\*  
## I(trip\_seconds^2) -3.372e-06 2.778e-07 -12.137 < 2e-16 \*\*\*  
## trip\_seconds:trip\_miles 4.371e-04 1.016e-04 4.300 5.43e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.447 on 70 degrees of freedom  
## Multiple R-squared: 0.8726, Adjusted R-squared: 0.8672   
## F-statistic: 159.8 on 3 and 70 DF, p-value: < 2.2e-16

**Interpretation:**

By investigating further, there is a decrease in Adjusted R-square value which is 0.86 or 86%. This means that the model explains roughly 86% of the variation in the fare price.

**##Regression Model 8**: tm+ts+Trip Seconds Squared+Trip Miles Squared

regout8=lm(fare~trip\_seconds+trip\_miles+I(trip\_seconds^2),data=my.taxi)  
summary(regout8)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + I(trip\_seconds^2),   
## data = my.taxi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.9449 -1.8303 -0.5218 1.8459 16.3733   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.561e+00 9.570e-01 -1.631 0.107   
## trip\_seconds 1.971e-02 1.778e-03 11.085 < 2e-16 \*\*\*  
## trip\_miles 8.079e-01 1.563e-01 5.170 2.12e-06 \*\*\*  
## I(trip\_seconds^2) -2.121e-06 2.304e-07 -9.205 1.11e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.253 on 70 degrees of freedom  
## Multiple R-squared: 0.8835, Adjusted R-squared: 0.8785   
## F-statistic: 176.9 on 3 and 70 DF, p-value: < 2.2e-16

**Interpretation:**

Adjusted R-square value which is 0.8785 or 87%. This means that the model explains roughly 87% of the variation in the fare price.

**##Regression Model 9.** Investigating Further. Trip Seconds+Trip Miles+Payment Type+TM2+Interaction

regout9=lm(fare~trip\_seconds+trip\_miles+payment\_type+I(trip\_miles^2)+trip\_seconds:trip\_miles,data=my.taxi)  
summary(regout9)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + payment\_type +   
## I(trip\_miles^2) + trip\_seconds:trip\_miles, data = my.taxi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.1475 -1.6643 -0.1937 1.6954 10.0886   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.556e+00 7.254e-01 -2.145 0.0356 \*   
## trip\_seconds 2.056e-02 1.149e-03 17.896 < 2e-16 \*\*\*  
## trip\_miles 3.469e-02 2.549e-01 0.136 0.8922   
## payment\_typeCredit Card 9.087e-01 7.050e-01 1.289 0.2018   
## payment\_typeOther -6.789e+00 3.209e+00 -2.116 0.0381 \*   
## I(trip\_miles^2) 1.360e-01 1.617e-02 8.409 4.41e-12 \*\*\*  
## trip\_seconds:trip\_miles -1.085e-03 6.999e-05 -15.505 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.903 on 67 degrees of freedom  
## Multiple R-squared: 0.9481, Adjusted R-squared: 0.9434   
## F-statistic: 203.8 on 6 and 67 DF, p-value: < 2.2e-16

**Interpretation:**

Adjusted R-square value which is 0.9434 or 94.34%. This means that the model explains roughly 94% of the variation in the fare price.

This a good model but falls slightly short of Regression Model 5 which has an adjusted R-square of 95.66%.

**##Regression Model 10.** Trying to reduce the risk of overfitting. Removing payment type from Regression Model 9.

regout10=lm(fare~trip\_seconds+trip\_miles+I(trip\_miles^2)+trip\_seconds:trip\_miles,data=my.taxi)  
summary(regout10)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + I(trip\_miles^2) +   
## trip\_seconds:trip\_miles, data = my.taxi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.220 -1.815 -0.502 1.699 11.078   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8.629e-01 6.973e-01 -1.237 0.220   
## trip\_seconds 2.017e-02 1.131e-03 17.829 < 2e-16 \*\*\*  
## trip\_miles -7.187e-02 2.596e-01 -0.277 0.783   
## I(trip\_miles^2) 1.435e-01 1.650e-02 8.699 1.06e-12 \*\*\*  
## trip\_seconds:trip\_miles -1.082e-03 6.981e-05 -15.493 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.009 on 69 degrees of freedom  
## Multiple R-squared: 0.9425, Adjusted R-squared: 0.9392   
## F-statistic: 282.8 on 4 and 69 DF, p-value: < 2.2e-16

**Interpretation:**

Adjusted R-square value which is 0.9392 or almost 94%. This means that the model explains roughly 94% of the variation in the fare price.

This is a very good model and it reduces the risk of overfitting by taking out payment type from consideration.

**##Conductive Stepwise Regression.**

regout11=step(lm(fare~trip\_miles+trip\_seconds,data=my.taxi),direction = "both")

## Start: AIC=274.84  
## fare ~ trip\_miles + trip\_seconds  
##   
## Df Sum of Sq RSS AIC  
## <none> 2799.2 274.85  
## - trip\_seconds 1 756.22 3555.5 290.54  
## - trip\_miles 1 1835.97 4635.2 310.17

summary(regout11)

##   
## Call:  
## lm(formula = fare ~ trip\_miles + trip\_seconds, data = my.taxi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.3553 -2.0260 -1.7191 0.1527 28.6823   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.853480 0.968360 5.012 3.81e-06 \*\*\*  
## trip\_miles 1.422958 0.208521 6.824 2.46e-09 \*\*\*  
## trip\_seconds 0.004909 0.001121 4.380 4.02e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.279 on 71 degrees of freedom  
## Multiple R-squared: 0.7424, Adjusted R-squared: 0.7352   
## F-statistic: 102.3 on 2 and 71 DF, p-value: < 2.2e-16

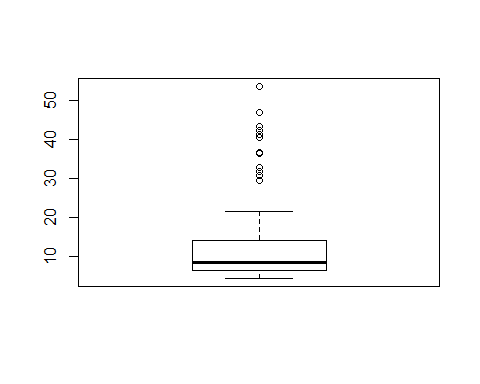
**Interpretation:** Low adjusted R-square value of 0.73.

Previous models are better.

**##CHECKING FOR OUTLIERS**

##Constructing Boxplot to check for any outliers in fare.

boxplot(fare)



outliers = filter(my.taxi,fare>40)  
outliers

## taxi\_id trip\_seconds trip\_miles fare tips tolls extras trip\_total  
## 1 4444 1740 1.0 43.50 5.00 0 0 48.50  
## 2 537 2280 21.4 53.75 0.00 0 0 53.75  
## 3 4100 1860 1.0 41.25 9.05 0 4 54.30  
## 4 6942 1620 16.2 40.50 8.50 0 2 51.00  
## 5 8760 7080 19.5 47.00 0.00 0 0 47.00  
## 6 7833 1260 17.5 42.50 0.00 0 4 46.50  
## payment\_type  
## 1 Credit Card  
## 2 Cash  
## 3 Credit Card  
## 4 Credit Card  
## 5 Cash  
## 6 Credit Card

**Interpretation:**

There are quite a few outliers in the fare variable, and I filtered them out (fare > $40) using the filter function. I have created a temporary data frame called “outliers” to refer to these outliers.

**##Removing two of these outliers:**

my.taxi\_clean = filter(my.taxi,taxi\_id!=4444 & taxi\_id!=4100)

I removed these two because both of them had Trip Mile of of 1 (which is very low) but a very high fare price. This seemed aberrant to me because I don’t know a lot of people who’d actually pay 40 bucks for a 1 mile trip.

In other words, these were clearly outliers in my sample.

**##Rerunning Regression Model 5** (which was the best Regression Model) ##NOW CALLED REGRESSION MODEL 5 BEST.

The best model that I got in my analysis after removing some outliers and running Regression Model 5 again.

regout5\_best=lm(fare~trip\_seconds+trip\_miles+I(trip\_seconds^2)+I(trip\_miles^2)+trip\_seconds:trip\_miles,data=my.taxi\_clean)  
summary(regout5\_best)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + I(trip\_seconds^2) +   
## I(trip\_miles^2) + trip\_seconds:trip\_miles, data = my.taxi\_clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.8048 -0.9746 -0.0781 1.0109 9.4043   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.240e-01 5.051e-01 1.433 0.156472   
## trip\_seconds 1.320e-02 1.075e-03 12.275 < 2e-16 \*\*\*  
## trip\_miles 8.874e-01 2.141e-01 4.146 9.87e-05 \*\*\*  
## I(trip\_seconds^2) 2.079e-06 5.393e-07 3.854 0.000266 \*\*\*  
## I(trip\_miles^2) 1.588e-01 1.751e-02 9.067 3.31e-13 \*\*\*  
## trip\_seconds:trip\_miles -1.655e-03 2.337e-04 -7.083 1.15e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.017 on 66 degrees of freedom  
## Multiple R-squared: 0.9706, Adjusted R-squared: 0.9684   
## F-statistic: 436.1 on 5 and 66 DF, p-value: < 2.2e-16

**Interpretation:**

Adjusted R square of 0.9684 or almost 97%. This means that the model explains roughly 97% of the variation in the fare price.

This is the best fit to the sample my.taxi that I have found in my analysis.

**#Analysis Part 6**

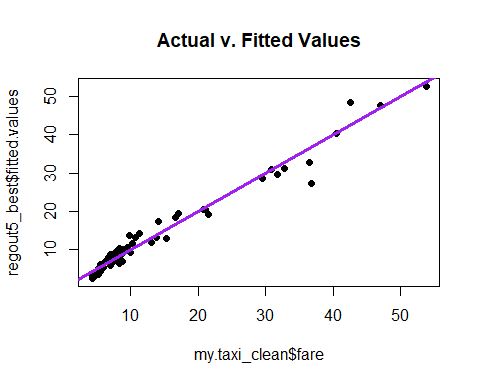
In my analysis, out of all the Regression Models I created in Analysis Part 5, the model which gave me the “best fit” to my sample data is Regression Model 5 - BEST.

Here is the summary report again with the plot and confidence intervals for each beta coefficient:

regout5\_best=lm(fare~trip\_seconds+trip\_miles+I(trip\_seconds^2)+I(trip\_miles^2)+trip\_seconds:trip\_miles,data=my.taxi\_clean)  
summary(regout5\_best)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + I(trip\_seconds^2) +   
## I(trip\_miles^2) + trip\_seconds:trip\_miles, data = my.taxi\_clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.8048 -0.9746 -0.0781 1.0109 9.4043   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.240e-01 5.051e-01 1.433 0.156472   
## trip\_seconds 1.320e-02 1.075e-03 12.275 < 2e-16 \*\*\*  
## trip\_miles 8.874e-01 2.141e-01 4.146 9.87e-05 \*\*\*  
## I(trip\_seconds^2) 2.079e-06 5.393e-07 3.854 0.000266 \*\*\*  
## I(trip\_miles^2) 1.588e-01 1.751e-02 9.067 3.31e-13 \*\*\*  
## trip\_seconds:trip\_miles -1.655e-03 2.337e-04 -7.083 1.15e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.017 on 66 degrees of freedom  
## Multiple R-squared: 0.9706, Adjusted R-squared: 0.9684   
## F-statistic: 436.1 on 5 and 66 DF, p-value: < 2.2e-16

plot(my.taxi\_clean$fare,regout5\_best$fitted.values,pch=19,main="Actual v. Fitted Values")  
abline(0,1,col="purple",lwd=3)



regout5\_bb.=cbind("Beta Coefficients"=coef(regout5\_best),confint(regout5\_best))  
regout5\_bb.

## Beta Coefficients 2.5 % 97.5 %  
## (Intercept) 7.239828e-01 -2.844632e-01 1.732429e+00  
## trip\_seconds 1.319559e-02 1.104925e-02 1.534194e-02  
## trip\_miles 8.874268e-01 4.600378e-01 1.314816e+00  
## I(trip\_seconds^2) 2.078529e-06 1.001771e-06 3.155286e-06  
## I(trip\_miles^2) 1.587809e-01 1.238175e-01 1.937443e-01  
## trip\_seconds:trip\_miles -1.655426e-03 -2.122078e-03 -1.188774e-03

**Interpretation:**

In my analysis, this is the “best” fit to my sample data because:

Highest Adjusted R-square value of 0.9684 or 96.84%. This means that the model explains roughly 97% (rounding up) of the variation in the fare price.

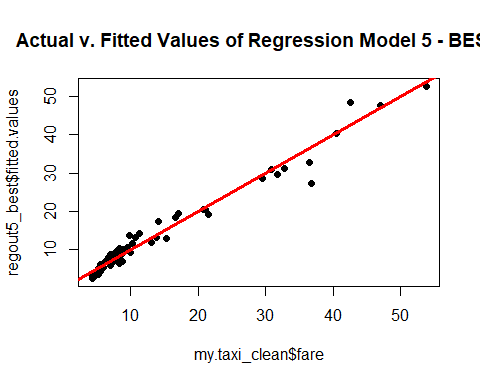
We have removed payment type from consideration. By doing this, we are reducing the risk of overfitting the model. All the variables except the intercept are significant.

From the Actual vs Fitted Values plot, the regression line fits most of the data points but for one or two observations.

**##Checking the satisfaction of LINE Assumptions:**

**##Linearity Assumption**

plot(my.taxi\_clean$fare,regout5\_best$fitted.values,pch=19,main="Actual v. Fitted Values of Regression Model 5 - BEST")  
abline(0,1,col="red",lwd=3)



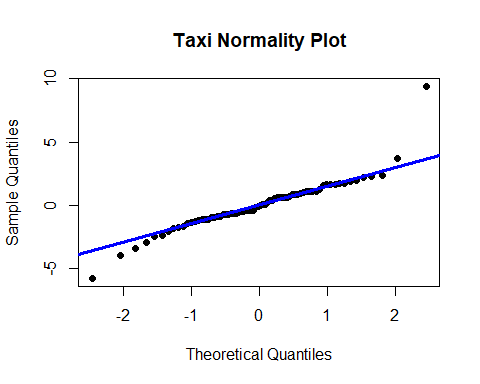
**Interpretation:**

In the above plot, the actual data points from the cleansed my.taxi\_clean sample are plotted against the model’s fitted values. Most of the data values follow a linear direction. One or two data points in the middle are do not follow a linear path but these are negligible, considering most of the observations follow a linear path.

**Conclusion**: The model is in conformance with the Linearity Assumption.

**##Normality Assumption**

qqnorm(regout5\_best$residuals,pch=19,  
 main="Taxi Normality Plot")  
qqline(regout5\_best$residuals,lwd=3,col="blue")



**Interpretation:**

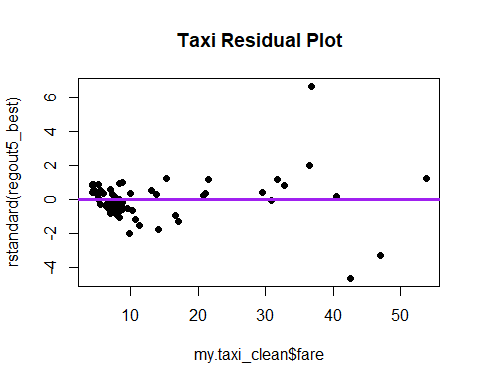
In the above graph, the model’s residuals or error values are plotted to check for the Normality assumption.

It can be seen that most of the points follow the normality distribution but a few data points at both the tail ends seem to taper off away from the normality line. However, these are trivial or negligible considering most of the observations follow the normal line.

It can be concluded that the Normality assumption is **satisfied** by the model.

**##Equality of Variances Assumption:**

plot(my.taxi\_clean$fare,rstandard(regout5\_best),pch=19,main="Taxi Residual Plot")  
abline(0,0,col="purple",lwd=3)

 Interpretation:

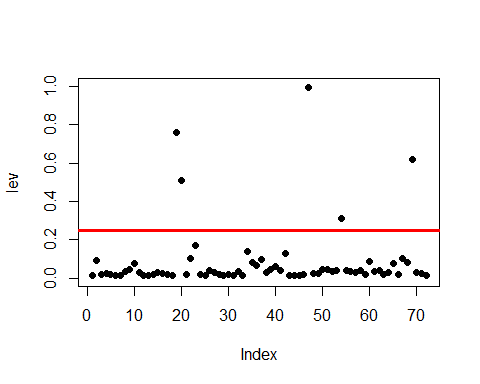
For evaluating the Equality of Variances assumption, the actual charges for the cleansed sample (my.taxi\_clean) are plotted against the standardized residuals from the model.

It can be seen that most of the data point are 1 or 2 standard deviations away from the mean (0) but some data points are more than 2 standard deviations from the mean. However, these outliers are low compared to most of the observations.

The Equality of Variances Assumption is **satisfied** by the model.

**#Analysis Part 7**

lev=hat(model.matrix(regout5\_best))  
plot(lev,pch=19,ylim=c(0,1.0))  
abline(3\*mean(lev),0,col="red",lwd=3)



high\_lev = my.taxi\_clean[lev>(3\*mean(lev)),]

Interpretation: To identify the high leverage points from my model, I plotted a red line 3 times the mean of lev (data points in matrix) to check for any aberrances. From my graph, there are FIVE high leverage data points in the model.

I stored these high leverage points in a data fram called high\_lev.

**#Removing the high leverage points**

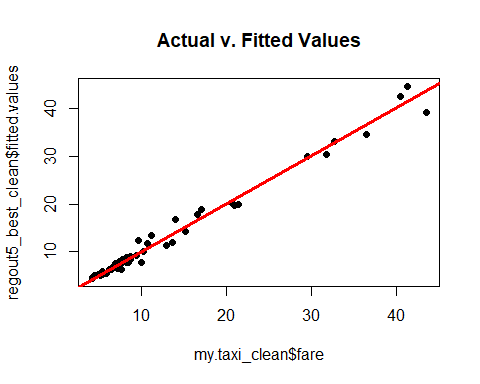
my.taxi\_clean = filter(my.taxi,taxi\_id!=537 & taxi\_id!=3624 & taxi\_id!=8760 & taxi\_id!=4373 & taxi\_id!=7833)

**#Rerunning the Regression 5 BEST MODEL after removing these outliers.**

regout5\_best\_clean=lm(fare~trip\_seconds+trip\_miles+I(trip\_seconds^2)+I(trip\_miles^2)+trip\_seconds:trip\_miles,data=my.taxi\_clean)  
summary(regout5\_best\_clean)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + I(trip\_seconds^2) +   
## I(trip\_miles^2) + trip\_seconds:trip\_miles, data = my.taxi\_clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.2923 -0.3717 -0.0319 0.3980 4.3440   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.056e+00 4.479e-01 9.055 5.21e-13 \*\*\*  
## trip\_seconds -2.122e-03 1.456e-03 -1.458 0.15   
## trip\_miles 1.953e+00 1.919e-01 10.178 6.27e-15 \*\*\*  
## I(trip\_seconds^2) 1.396e-05 8.050e-07 17.342 < 2e-16 \*\*\*  
## I(trip\_miles^2) 2.236e-01 1.728e-02 12.935 < 2e-16 \*\*\*  
## trip\_seconds:trip\_miles -3.246e-03 2.092e-04 -15.518 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.174 on 63 degrees of freedom  
## Multiple R-squared: 0.9858, Adjusted R-squared: 0.9847   
## F-statistic: 877.2 on 5 and 63 DF, p-value: < 2.2e-16

plot(my.taxi\_clean$fare,regout5\_best\_clean$fitted.values,pch=19,main="Actual v. Fitted Values")  
abline(0,1,col="red",lwd=3)



**Interpretation:**

After removing the outliers and re-running the model, I got the BEST adjusted R-square value of 0.9846 or 98%. This means that the model explains roughly 98% of the variation in the fare price.

**#Analysis Part 8**

Pulling a new sample from the original data set “taxi”. (Taking seed as U Number plus 5)

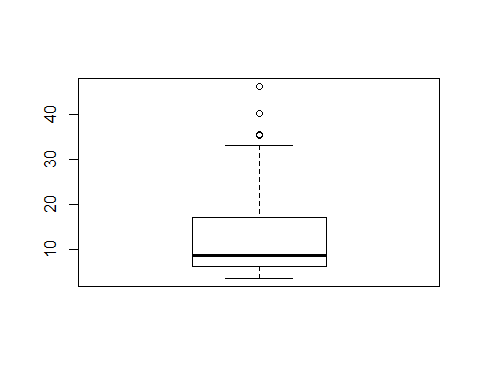
Storing this in a new data frame called “my.taxi2”

set.seed(85555988)  
my.taxi2=taxi[sample(1:nrow(taxi),100,replace=FALSE),]

##Data Cleansing

my.taxi2 = subset(my.taxi2,trip\_seconds!=0 & trip\_miles!=0)

boxplot(my.taxi2$fare)



outliers\_new = filter(my.taxi2,fare>40)  
outliers\_new

## taxi\_id trip\_seconds trip\_miles fare tips tolls extras trip\_total  
## 1 8445 3240 15.0 40.25 8.05 0 0 48.30  
## 2 2366 2400 17.8 46.25 8.00 0 0 54.25  
## payment\_type  
## 1 Credit Card  
## 2 Credit Card

##Removing two of these outliers:

my.taxi2\_clean = filter(my.taxi2,taxi\_id!=8445 & taxi\_id!=2366)

##Conducting Regression Analysis using the best model that I found in Analysis Part 6:

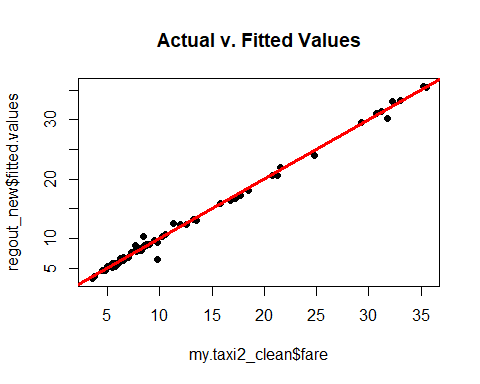
regout\_new=lm(fare~trip\_seconds+trip\_miles+I(trip\_seconds^2)+I(trip\_miles^2)+trip\_seconds:trip\_miles,data=my.taxi2\_clean)  
summary(regout\_new)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + I(trip\_seconds^2) +   
## I(trip\_miles^2) + trip\_seconds:trip\_miles, data = my.taxi2\_clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.8114 -0.2499 -0.0289 0.2221 3.3493   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.797e+00 2.607e-01 10.729 7.49e-16 \*\*\*  
## trip\_seconds 5.127e-03 7.947e-04 6.451 1.79e-08 \*\*\*  
## trip\_miles 1.667e+00 1.245e-01 13.384 < 2e-16 \*\*\*  
## I(trip\_seconds^2) -4.689e-07 4.006e-07 -1.171 0.246173   
## I(trip\_miles^2) 3.460e-02 9.390e-03 3.684 0.000479 \*\*\*  
## trip\_seconds:trip\_miles -1.645e-04 1.122e-04 -1.466 0.147614   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6406 on 63 degrees of freedom  
## Multiple R-squared: 0.995, Adjusted R-squared: 0.9947   
## F-statistic: 2530 on 5 and 63 DF, p-value: < 2.2e-16

regout\_new.coefficients=cbind("Beta Coefficients"=coef(regout\_new),confint(regout\_new))  
regout\_new.coefficients

## Beta Coefficients 2.5 % 97.5 %  
## (Intercept) 2.797248e+00 2.276233e+00 3.318263e+00  
## trip\_seconds 5.126876e-03 3.538816e-03 6.714937e-03  
## trip\_miles 1.666562e+00 1.417733e+00 1.915391e+00  
## I(trip\_seconds^2) -4.689385e-07 -1.269474e-06 3.315973e-07  
## I(trip\_miles^2) 3.459691e-02 1.583195e-02 5.336187e-02  
## trip\_seconds:trip\_miles -1.644666e-04 -3.886504e-04 5.971721e-05

plot(my.taxi2\_clean$fare,regout\_new$fitted.values,pch=19,main="Actual v. Fitted Values")  
abline(0,1,col="red",lwd=3)



**Interpretation:**

After cleansing and running the best model that I found in Analysis Part 6, I found that the model does extremely well on this new sample.

* I got an adjusted R-square value of 0.9947 or 99%. This means that the model explains roughly 99% of the variation in the fare price.
* Additionally, the Confidence Intervals for most of the beta co-efficients is tight.This means that we can estimate with confidence the values of the beta co-efficients.

From the Actual vs Fitted Values plot, the regression line fits the data points almost perfectly. I couldn’t ask for better from this model.

Thank You,

Saahil Rasheed