QMB Project

Ankita Patidar

10/20/2019

# Pre processing

library(readxl)  
library(dplyr)  
library(ggplot2)  
library(moments)  
library(Hmisc)  
library(corrplot)  
library(e1071)  
taxi = read.csv("6304 Regression Project Data.csv")  
colnames(taxi)= tolower(make.names(colnames(taxi)))  
attach(taxi)  
#Checking dimensions of original data  
dim(taxi)

## [1] 1705421 9

set.seed(94212287)  
taxi\_sample = taxi[sample(1:nrow(taxi),100, replace = FALSE),]  
str(taxi\_sample)

## 'data.frame': 100 obs. of 9 variables:  
## $ taxi\_id : int 329 8193 3225 8138 5134 7459 3332 1946 1254 3499 ...  
## $ trip\_seconds: int 420 420 0 420 180 1920 660 1200 0 720 ...  
## $ trip\_miles : num 0.9 0.1 0 1.9 0.7 17.6 0.1 7.2 0 4.7 ...  
## $ fare : num 6.25 10.25 13.25 7.75 5 ...  
## $ tips : num 0 2.45 2.65 3 2 8.7 2 0 2.35 2.29 ...  
## $ tolls : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ extras : num 0 2 0 1 0 0 0 0 0 1 ...  
## $ trip\_total : num 6.25 14.7 15.9 11.75 7 ...  
## $ payment\_type: Factor w/ 3 levels "Cash","Credit Card",..: 1 2 2 2 2 2 2 1 2 2 ...

#Checking for number of rows and columns for sampled data  
dim(taxi\_sample)

## [1] 100 9

summary(taxi\_sample)

## taxi\_id trip\_seconds trip\_miles fare   
## Min. : 4 Min. : 0.0 Min. : 0.000 Min. : 0.000   
## 1st Qu.:2265 1st Qu.: 240.0 1st Qu.: 0.100 1st Qu.: 6.188   
## Median :4099 Median : 480.0 Median : 0.800 Median : 8.250   
## Mean :4324 Mean : 552.6 Mean : 1.819 Mean :11.440   
## 3rd Qu.:6754 3rd Qu.: 780.0 3rd Qu.: 2.100 3rd Qu.:12.125   
## Max. :8561 Max. :2280.0 Max. :17.600 Max. :50.000   
## tips tolls extras trip\_total   
## Min. : 0.000 Min. :0 Min. :0.000 Min. : 0.00   
## 1st Qu.: 0.000 1st Qu.:0 1st Qu.:0.000 1st Qu.: 7.00   
## Median : 0.000 Median :0 Median :0.000 Median : 9.25   
## Mean : 1.311 Mean :0 Mean :0.485 Mean :13.24   
## 3rd Qu.: 2.000 3rd Qu.:0 3rd Qu.:0.625 3rd Qu.:13.44   
## Max. :13.000 Max. :0 Max. :5.500 Max. :56.10   
## payment\_type  
## Cash :48   
## Credit Card:50   
## Other : 2   
##   
##   
##

1.Trip\_seconds, Trip\_miles, fare and trip\_total have values zero which seems incorrect as a right cannot be of 0 seconds and 0$. 2.Tolls have all values 0, so this cannot be used for our analysis.

which(taxi\_sample$trip\_seconds == 0.00 | taxi\_sample$trip\_miles == 0)

## [1] 3 9 11 12 13 16 19 31 35 42 47 52 53 58 62 67 72 74 76 79 90 95 96  
## [24] 99

which(taxi\_sample$fare == 0 | taxi\_sample$trip\_total==0)

## [1] 19

#These values are irrelevant, so we are removing these values  
taxi\_sample=subset(taxi\_sample,select=-c(tolls))  
new\_sample = filter(taxi\_sample, trip\_seconds != 0 & trip\_miles != 0.00 & fare != 0 & trip\_total!=0)  
dim(new\_sample)

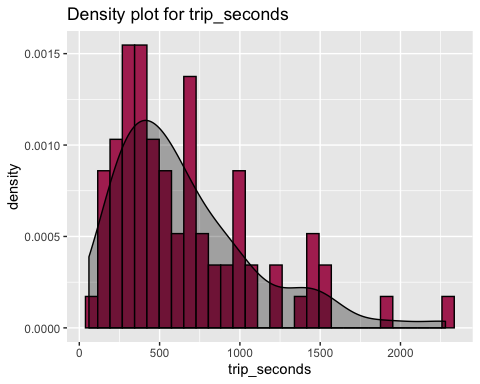
## [1] 76 8

# Analysis 1 : Density Plots for each continuous variables.

summary(new\_sample$trip\_seconds)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 60.0 300.0 540.0 655.3 855.0 2280.0

# Density plot for trip\_seconds  
ggplot(new\_sample, aes(x=trip\_seconds)) +   
 geom\_histogram(aes(y=..density..), colour="black", fill="maroon")+   
 geom\_density(alpha=.3, fill="black") + ggtitle("Density plot for trip\_seconds")



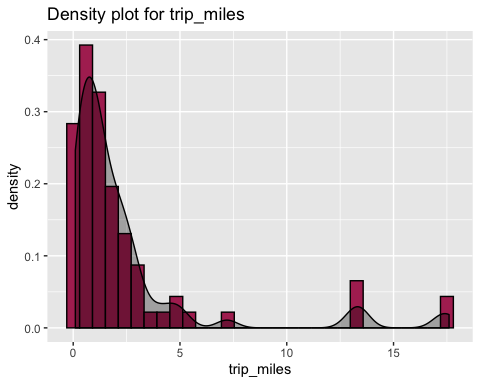
s=skewness(new\_sample$trip\_seconds)  
k=kurtosis(new\_sample$trip\_seconds)  
print(paste0('According to above analysis and density plot, the skewness =',s,', we can depict that the distribution of trip\_seconds is right skewed.The kurtosis value is ',k,', which means trip\_seconds values are not normally distributed.'))

## [1] "According to above analysis and density plot, the skewness =1.34191013830994, we can depict that the distribution of trip\_seconds is right skewed.The kurtosis value is 1.75613712284494, which means trip\_seconds values are not normally distributed."

summary(new\_sample$trip\_miles)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.100 0.600 1.200 2.393 2.325 17.600

s=skewness(new\_sample$trip\_miles)  
k=kurtosis(new\_sample$trip\_miles)  
ggplot(new\_sample, aes(x=trip\_miles)) +   
 geom\_histogram(aes(y=..density..), colour="black", fill="maroon")+  
 geom\_density(alpha=.3, fill="black") + ggtitle("Density plot for trip\_miles")



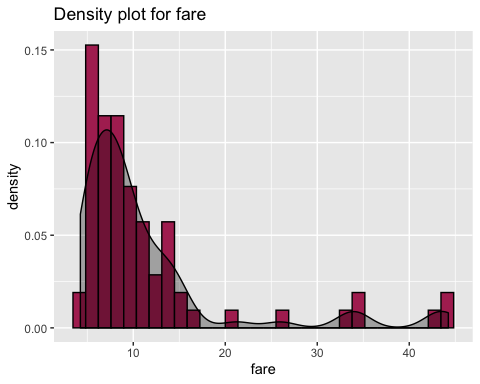
print(paste0('According to above analysis and density plot, the skewness =',s,', we can depict that the distribution of trip\_seconds is right skewed.The kurtosis value is ',k,', which means trip\_seconds values are not normally distributed.'))

## [1] "According to above analysis and density plot, the skewness =2.87354336510321, we can depict that the distribution of trip\_seconds is right skewed.The kurtosis value is 7.85009444159161, which means trip\_seconds values are not normally distributed."

summary(new\_sample$fare)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.250 6.250 8.375 11.343 12.125 44.250

s=skewness(new\_sample$fare)  
k=kurtosis(new\_sample$fare)  
ggplot(new\_sample, aes(x=fare)) +   
 geom\_histogram(aes(y=..density..), colour="black", fill="maroon")+  
 geom\_density(alpha=.3, fill="black") + ggtitle("Density plot for fare")



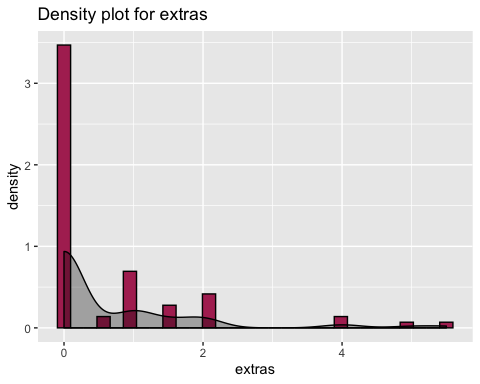
print(paste0('According to above analysis and density plot, the skewness =',s,', we can depict that the distribution of trip\_seconds is right skewed.The kurtosis value is ',k,', which means trip\_seconds values are not normally distributed.'))

## [1] "According to above analysis and density plot, the skewness =2.37501820081551, we can depict that the distribution of trip\_seconds is right skewed.The kurtosis value is 5.07794027220122, which means trip\_seconds values are not normally distributed."

summary(new\_sample$extras)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 0.625 1.000 5.500

s=skewness(new\_sample$extras)  
k=kurtosis(new\_sample$extras)  
ggplot(new\_sample, aes(x=extras)) +   
 geom\_histogram(aes(y=..density..), colour="black", fill="maroon")+  
 geom\_density(alpha=.3, fill="black") + ggtitle("Density plot for extras")



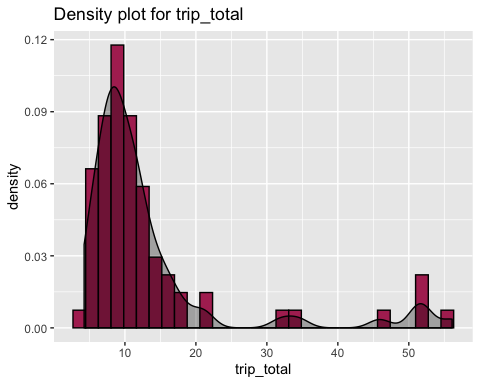
print(paste0('According to above analysis and density plot, the skewness =',s,', we can depict that the distribution of trip\_seconds is right skewed.The kurtosis value is ',k,', which means trip\_seconds values are not normally distributed.'))

## [1] "According to above analysis and density plot, the skewness =2.38122019583194, we can depict that the distribution of trip\_seconds is right skewed.The kurtosis value is 5.85726549299181, which means trip\_seconds values are not normally distributed."

summary(new\_sample$trip\_total)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.250 7.875 9.750 13.377 13.250 56.100

s=skewness(new\_sample$trip\_total)  
k=kurtosis(new\_sample$trip\_total)  
ggplot(new\_sample, aes(x=trip\_total)) +   
 geom\_histogram(aes(y=..density..), colour="black", fill="maroon")+  
 geom\_density(alpha=.3, fill="black") + ggtitle("Density plot for trip\_total")



print(paste0('According to above analysis and density plot, the skewness =',s,', we can depict that the distribution of trip\_seconds is right skewed.The kurtosis value is ',k,', which means trip\_seconds values are not normally distributed.'))

## [1] "According to above analysis and density plot, the skewness =2.52898265311067, we can depict that the distribution of trip\_seconds is right skewed.The kurtosis value is 5.68279762174453, which means trip\_seconds values are not normally distributed."

# Analysis 2

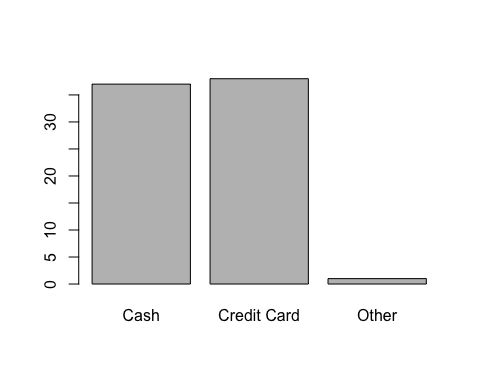
str(new\_sample)

## 'data.frame': 76 obs. of 8 variables:  
## $ taxi\_id : int 329 8193 8138 5134 7459 3332 1946 3499 2450 4692 ...  
## $ trip\_seconds: int 420 420 420 180 1920 660 1200 720 960 240 ...  
## $ trip\_miles : num 0.9 0.1 1.9 0.7 17.6 0.1 7.2 4.7 2 1.2 ...  
## $ fare : num 6.25 10.25 7.75 5 43.5 ...  
## $ tips : num 0 2.45 3 2 8.7 2 0 2.29 1.75 2 ...  
## $ extras : num 0 2 1 0 0 0 0 1 0 1 ...  
## $ trip\_total : num 6.25 14.7 11.75 7 52.2 ...  
## $ payment\_type: Factor w/ 3 levels "Cash","Credit Card",..: 1 2 2 2 2 2 1 2 2 2 ...

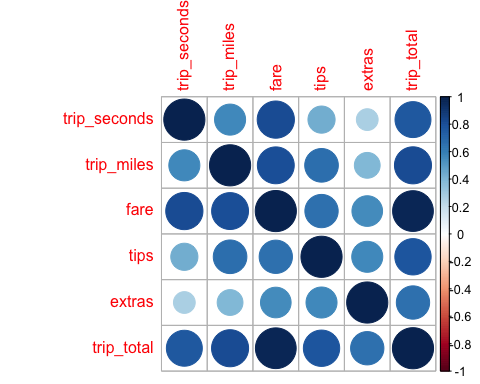
# payment\_type is a factor variable with 3 levels, "Cash", "Credit Card" and "Other"  
count(new\_sample, payment\_type)

## # A tibble: 3 x 2  
## payment\_type n  
## <fct> <int>  
## 1 Cash 37  
## 2 Credit Card 38  
## 3 Other 1

plot(new\_sample$payment\_type)

 # Analysis 3

# Creating a new object for continuous variable only except taxi\_id of our new\_sample data set.  
contvar = new\_sample[,c(2:7)]  
xx=cor(contvar)  
corrplot(xx,method="circle")



corrplot(xx,method="number")



#Correlation matrix with p-value  
rcorr(as.matrix(contvar))

## trip\_seconds trip\_miles fare tips extras trip\_total  
## trip\_seconds 1.00 0.57 0.82 0.43 0.28 0.76  
## trip\_miles 0.57 1.00 0.80 0.67 0.39 0.81  
## fare 0.82 0.80 1.00 0.65 0.55 0.98  
## tips 0.43 0.67 0.65 1.00 0.57 0.78  
## extras 0.28 0.39 0.55 0.57 1.00 0.65  
## trip\_total 0.76 0.81 0.98 0.78 0.65 1.00  
##   
## n= 76   
##   
##   
## P  
## trip\_seconds trip\_miles fare tips extras trip\_total  
## trip\_seconds 0.0000 0.0000 0.0000 0.0148 0.0000   
## trip\_miles 0.0000 0.0000 0.0000 0.0005 0.0000   
## fare 0.0000 0.0000 0.0000 0.0000 0.0000   
## tips 0.0000 0.0000 0.0000 0.0000 0.0000   
## extras 0.0148 0.0005 0.0000 0.0000 0.0000   
## trip\_total 0.0000 0.0000 0.0000 0.0000 0.0000

The above correlation matrix shows correlation coefficients values between 2 variables which signifies relation between those 2 variables. The range of corelation coefficient is between -1 and +1.

The value approaching towards +1 is highly positively correlated while value approaching towards -1 is highly negatively correlated, In our analysis all the variables are positively correlated.

Fare is highly correlated with trip\_seconds, trip\_miles and trip\_total as their values are close to +1

Fare is comparatively less correlated to extras as it has value close to 0.

Similarly, other values in number method or size and colour intensity of circle in circle method represents the correlation among the variables based on size or intensity of color.

# Analysis 4

## Base Model Regression

regout = lm(fare ~ trip\_seconds + trip\_miles + payment\_type, data = new\_sample)  
summary(regout)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + payment\_type,   
## data = new\_sample)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.6199 -1.3910 -0.1344 0.8135 25.0584   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.489129 0.843873 0.580 0.5640   
## trip\_seconds 0.010929 0.001171 9.335 5.66e-14 \*\*\*  
## trip\_miles 1.214163 0.140873 8.619 1.19e-12 \*\*\*  
## payment\_typeCredit Card 1.629157 0.849872 1.917 0.0593 .   
## payment\_typeOther -2.103466 3.703946 -0.568 0.5719   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.643 on 71 degrees of freedom  
## Multiple R-squared: 0.8461, Adjusted R-squared: 0.8374   
## F-statistic: 97.56 on 4 and 71 DF, p-value: < 2.2e-16

#In the above regression model, fare is a dependent variable and trip\_seconds, trip\_miles and payment\_type are independent variable.  
# The Adjusted R-squared value is .8461, which means that our model is able to explain 84.61% variance in dependent variable by independent variables  
# The p- value for trip\_seconds and trip\_miles is significant and hence we are good to reject the null hypothesis that Beta coefficients value is 0.  
# The p-value for payment\_type Credit card and Other is insignificant and thus we fail to reject the null hypothesis.  
# From the above analysis it can be inferred that, the Beta coefficient for trip\_seconds indicates that with every 1 second increase in trip\_seconds, fare increases by .011 keeping others constant.In other words, with 10 minutes increase in time elapsed for the trip, fare increases by 6.6$ when other value doesn't change.   
# With 1 mile increase in miles, fare increase by 1.21$ when other value doesn't change.  
#The Beta coefficient for intercept explains fare when all other variables are zero,which in practical doesn't make sense and it is insignificant in our model.  
#payment\_type Credit Card suggests 1.629$ more than Cash payment and it is insignificant as its p value is greater than .05.  
#payment\_type Other suggests 2$ less than Cash payment and it is insignificant as its p value is greater than .05.

cbind("beta"=coef(regout),confint(regout))

## beta 2.5 % 97.5 %  
## (Intercept) 0.4891291 -1.193505032 2.17176321  
## trip\_seconds 0.0109290 0.008594584 0.01326341  
## trip\_miles 1.2141630 0.933270764 1.49505524  
## payment\_typeCredit Card 1.6291569 -0.065439616 3.32375343  
## payment\_typeOther -2.1034657 -9.488924946 5.28199346

#The output above shows the confidence interval for the variables.  
#From above analysis, we can say that we are 95% confident that the beta coefficient value for trip\_seconds will be in the interval of .0085 to 0.0132. And it is true as, 0.011 lies in this range.  
#Similary, it can be inferred for other variables as well.

# Analysis 5

Time spend in trip can be related to miles driven, so we can use interaction on this 2 variables.

regout1 = lm(fare ~ trip\_seconds+ trip\_miles +trip\_seconds\*trip\_miles+ payment\_type , data = new\_sample)  
# The adjusted R square is .8462, summary of regout1 model is as below:  
summary(regout1)

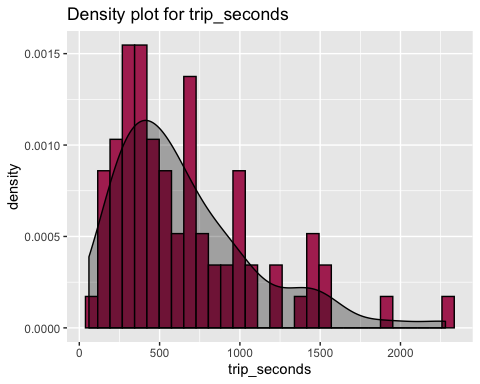
##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + trip\_seconds \*   
## trip\_miles + payment\_type, data = new\_sample)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.6042 -1.3818 -0.1452 0.7421 25.0737   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.145e-01 1.035e+00 0.594 0.5547   
## trip\_seconds 1.083e-02 1.260e-03 8.601 1.43e-12 \*\*\*  
## trip\_miles 1.124e+00 4.491e-01 2.502 0.0147 \*   
## payment\_typeCredit Card 1.624e+00 8.559e-01 1.898 0.0618 .   
## payment\_typeOther -2.052e+00 3.737e+00 -0.549 0.5846   
## trip\_seconds:trip\_miles 6.259e-05 2.952e-04 0.212 0.8327   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.667 on 70 degrees of freedom  
## Multiple R-squared: 0.8462, Adjusted R-squared: 0.8352   
## F-statistic: 77.01 on 5 and 70 DF, p-value: < 2.2e-16

cbind("beta"=coef(regout1),confint(regout1))

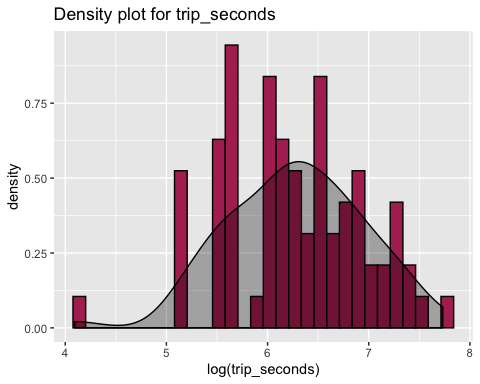
## beta 2.5 % 97.5 %  
## (Intercept) 6.144863e-01 -1.4500228554 2.6789955057  
## trip\_seconds 1.083475e-02 0.0083222332 0.0133472642  
## trip\_miles 1.123824e+00 0.2280712859 2.0195767989  
## payment\_typeCredit Card 1.624476e+00 -0.0826250291 3.3315767788  
## payment\_typeOther -2.052168e+00 -9.5052897800 5.4009545266  
## trip\_seconds:trip\_miles 6.258542e-05 -0.0005262231 0.0006513939

From earlier graph, distribution of trip\_seconds is right skewed, which can be fixed by using log transformation.

ggplot(new\_sample, aes(x=trip\_seconds)) +   
 geom\_histogram(aes(y=..density..), colour="black", fill="maroon")+  
 geom\_density(alpha=.3, fill="black") + ggtitle("Density plot for trip\_seconds")



ggplot(new\_sample, aes(x=log(trip\_seconds))) +   
 geom\_histogram(aes(y=..density..), colour="black", fill="maroon")+  
 geom\_density(alpha=.3, fill="black") + ggtitle("Density plot for trip\_seconds")



# Transformations applied on the base model to find better model fit  
  
# Log transformation on trip\_seconds  
regout2 = lm(fare ~ log(trip\_seconds)+ trip\_miles + trip\_seconds\*trip\_miles+payment\_type , data = new\_sample)  
# The adjusted R square is .8482, summary of regout2 model is as below:  
summary(regout2)

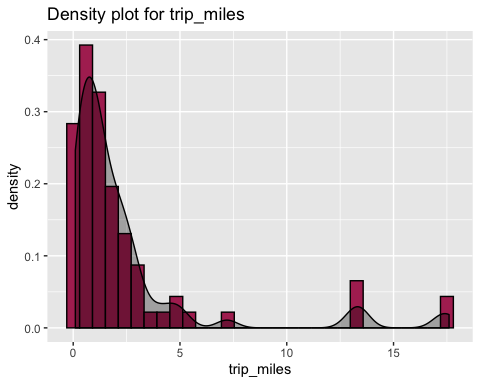
##   
## Call:  
## lm(formula = fare ~ log(trip\_seconds) + trip\_miles + trip\_seconds \*   
## trip\_miles + payment\_type, data = new\_sample)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.7906 -1.2426 -0.3067 0.6648 24.8626   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.5174336 9.2736730 1.026 0.30834   
## log(trip\_seconds) -1.6970837 1.7566941 -0.966 0.33738   
## trip\_miles 1.2755864 0.4760089 2.680 0.00921 \*\*   
## trip\_seconds 0.0134780 0.0030124 4.474 2.95e-05 \*\*\*  
## payment\_typeCredit Card 1.4818642 0.8689701 1.705 0.09263 .   
## payment\_typeOther -1.9225441 3.7411442 -0.514 0.60897   
## trip\_miles:trip\_seconds -0.0000514 0.0003181 -0.162 0.87209   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.669 on 69 degrees of freedom  
## Multiple R-squared: 0.8482, Adjusted R-squared: 0.835   
## F-statistic: 64.27 on 6 and 69 DF, p-value: < 2.2e-16

cbind("beta"=coef(regout2),confint(regout2))

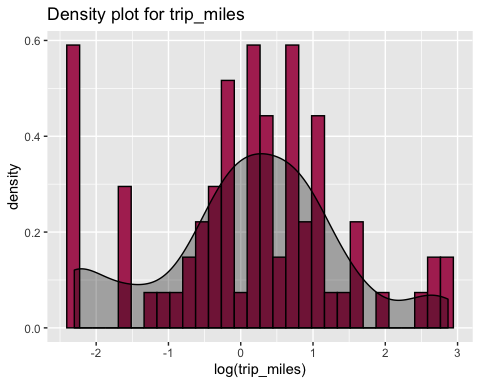
## beta 2.5 % 97.5 %  
## (Intercept) 9.517434e+00 -8.9830378543 2.801791e+01  
## log(trip\_seconds) -1.697084e+00 -5.2015925192 1.807425e+00  
## trip\_miles 1.275586e+00 0.3259746705 2.225198e+00  
## trip\_seconds 1.347795e-02 0.0074684077 1.948750e-02  
## payment\_typeCredit Card 1.481864e+00 -0.2516836697 3.215412e+00  
## payment\_typeOther -1.922544e+00 -9.3859225350 5.540834e+00  
## trip\_miles:trip\_seconds -5.139805e-05 -0.0006859104 5.831143e-04

From earlier graph, distribution of trip\_miles is right skewed, which can be fixed by using log transformation.

ggplot(new\_sample, aes(x=trip\_miles)) +   
 geom\_histogram(aes(y=..density..), colour="black", fill="maroon")+  
 geom\_density(alpha=.3, fill="black") + ggtitle("Density plot for trip\_miles")



ggplot(new\_sample, aes(x=log(trip\_miles))) +   
 geom\_histogram(aes(y=..density..), colour="black", fill="maroon")+  
 geom\_density(alpha=.3, fill="black") + ggtitle("Density plot for trip\_miles")



regout3 = lm(fare ~ log(trip\_seconds)+ log(trip\_miles) + trip\_seconds\*trip\_miles+ payment\_type , data = new\_sample)  
# The adjusted R square is .8616, summary of regout3 model is as below:  
summary(regout3)

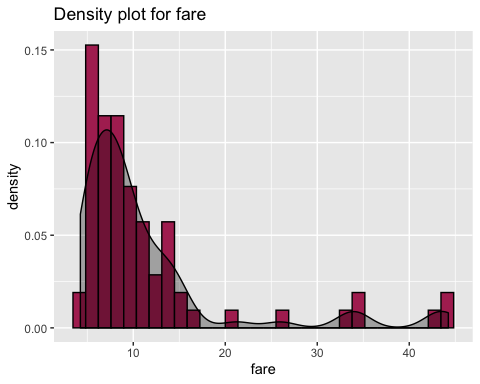
##   
## Call:  
## lm(formula = fare ~ log(trip\_seconds) + log(trip\_miles) + trip\_seconds \*   
## trip\_miles + payment\_type, data = new\_sample)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.1204 -0.9400 -0.2178 0.6125 25.8847   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.6928252 8.9218345 1.086 0.281131   
## log(trip\_seconds) -1.9674198 1.6932921 -1.162 0.249342   
## log(trip\_miles) -1.4762549 0.5766522 -2.560 0.012693 \*   
## trip\_seconds 0.0138718 0.0029021 4.780 9.76e-06 \*\*\*  
## trip\_miles 2.3534755 0.6220793 3.783 0.000329 \*\*\*  
## payment\_typeCredit Card 1.3390544 0.8378363 1.598 0.114629   
## payment\_typeOther -0.9823075 3.6177918 -0.272 0.786813   
## trip\_seconds:trip\_miles -0.0004939 0.0003514 -1.405 0.164475   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.53 on 68 degrees of freedom  
## Multiple R-squared: 0.8616, Adjusted R-squared: 0.8473   
## F-statistic: 60.46 on 7 and 68 DF, p-value: < 2.2e-16

cbind("beta"=coef(regout3),confint(regout3))

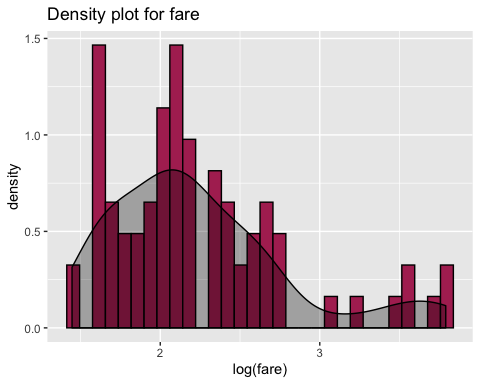
## beta 2.5 % 97.5 %  
## (Intercept) 9.6928252055 -8.110418415 27.496068826  
## log(trip\_seconds) -1.9674198403 -5.346331608 1.411491928  
## log(trip\_miles) -1.4762549093 -2.626946428 -0.325563390  
## trip\_seconds 0.0138718386 0.008080802 0.019662876  
## trip\_miles 2.3534754741 1.112135627 3.594815321  
## payment\_typeCredit Card 1.3390543811 -0.332821936 3.010930698  
## payment\_typeOther -0.9823075185 -8.201498653 6.236883616  
## trip\_seconds:trip\_miles -0.0004938671 -0.001195122 0.000207388

From earlier graph, distribution of fare is right skewed, which can be fixed by using log transformation.

ggplot(new\_sample, aes(x=fare)) +   
 geom\_histogram(aes(y=..density..), colour="black", fill="maroon")+  
 geom\_density(alpha=.3, fill="black") + ggtitle("Density plot for fare")



ggplot(new\_sample, aes(x=log(fare))) +   
 geom\_histogram(aes(y=..density..), colour="black", fill="maroon")+  
 geom\_density(alpha=.3, fill="black") + ggtitle("Density plot for fare")



regout4 = lm(log(fare) ~ log(trip\_seconds)+ log(trip\_miles) + trip\_seconds\*trip\_miles+ payment\_type , data = new\_sample)  
# The adjusted R square is .9212, summary of regout4 model is as below:  
summary(regout4)

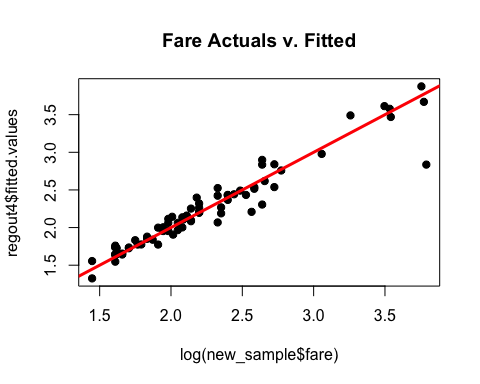
##   
## Call:  
## lm(formula = log(fare) ~ log(trip\_seconds) + log(trip\_miles) +   
## trip\_seconds \* trip\_miles + payment\_type, data = new\_sample)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.25994 -0.08446 -0.01917 0.04912 0.95338   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.139e-01 4.204e-01 0.271 0.787329   
## log(trip\_seconds) 2.234e-01 7.979e-02 2.799 0.006660 \*\*   
## log(trip\_miles) -9.583e-02 2.717e-02 -3.526 0.000759 \*\*\*  
## trip\_seconds 6.366e-04 1.368e-04 4.655 1.55e-05 \*\*\*  
## trip\_miles 1.841e-01 2.931e-02 6.280 2.74e-08 \*\*\*  
## payment\_typeCredit Card 1.017e-01 3.948e-02 2.576 0.012162 \*   
## payment\_typeOther 8.437e-03 1.705e-01 0.049 0.960673   
## trip\_seconds:trip\_miles -7.166e-05 1.656e-05 -4.327 5.08e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1663 on 68 degrees of freedom  
## Multiple R-squared: 0.9212, Adjusted R-squared: 0.9131   
## F-statistic: 113.6 on 7 and 68 DF, p-value: < 2.2e-16

cbind("beta"=coef(regout4),confint(regout4))

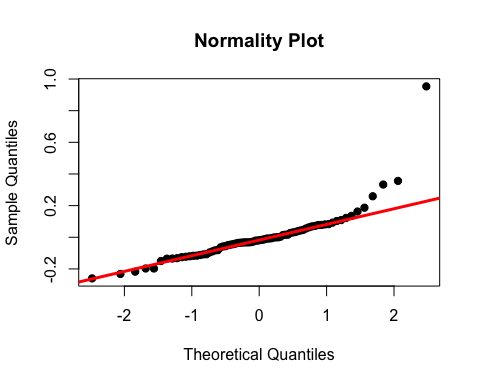
## beta 2.5 % 97.5 %  
## (Intercept) 1.138725e-01 -0.7250849777 9.528301e-01  
## log(trip\_seconds) 2.233648e-01 0.0641374416 3.825922e-01  
## log(trip\_miles) -9.582624e-02 -0.1500512681 -4.160122e-02  
## trip\_seconds 6.365495e-04 0.0003636535 9.094455e-04  
## trip\_miles 1.840865e-01 0.1255897304 2.425832e-01  
## payment\_typeCredit Card 1.017188e-01 0.0229334937 1.805040e-01  
## payment\_typeOther 8.437474e-03 -0.3317586651 3.486336e-01  
## trip\_seconds:trip\_miles -7.165719e-05 -0.0001047030 -3.861134e-05

# Analysis 6

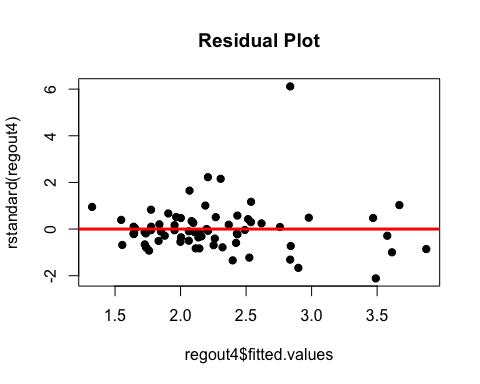
#Linearity  
plot(log(new\_sample$fare),regout4$fitted.values,pch=19,main="Fare Actuals v. Fitted")  
abline(0,1,col="red",lwd=3)



# There is a linear relationshp between the actual and predicted values of fare, as most of the data points fall on the regression line.  
#Normality  
qqnorm(regout4$residuals,pch=19,main="Normality Plot")  
qqline(regout4$residuals,lwd=3,col="red")



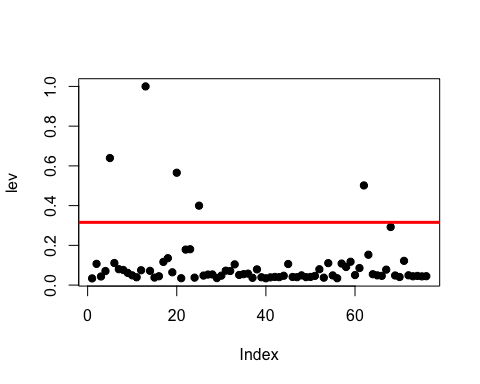
# We can see that few points at extreme ends are away from the line but with available data points of 76 we can say that normality assumption holds true as maximum residuals lie on the line.  
#Equality of Variances  
plot(regout4$fitted.values,rstandard(regout4),pch=19,main="Residual Plot")  
abline(0,0,col="red",lwd=3)



# Equality of variance is true for our model as there is no pattern seen in the above residual plot.  
  
# Our model confirms the LINE assumtions.

# Analysis 7

# Identifying high leverage points  
  
lev=hat(model.matrix(regout4))  
plot(lev,pch=19)  
abline(3\*mean(lev),0,col="red",lwd=3)



levpt =new\_sample[lev>(3\*mean(lev)),]  
levpt

## taxi\_id trip\_seconds trip\_miles fare tips extras trip\_total  
## 5 7459 1920 17.6 43.50 8.7 0.0 52.20  
## 13 8384 900 2.7 11.50 0.0 0.0 11.50  
## 20 6078 2280 0.4 26.00 4.0 2.0 32.00  
## 25 3808 60 0.5 4.25 3.0 0.0 7.25  
## 62 5389 960 13.0 33.00 13.0 5.5 51.50  
## payment\_type  
## 5 Credit Card  
## 13 Other  
## 20 Credit Card  
## 25 Credit Card  
## 62 Credit Card

# Removing high leverage points from the sampled data  
no\_levpt\_sample = new\_sample[-c(5,13,20,25,62),]  
regout5 = lm(log(fare) ~ log(trip\_seconds)+ log(trip\_miles) + trip\_seconds\*trip\_miles+ payment\_type , data = no\_levpt\_sample)  
# The adjusted R square is .91, summary of regout model is as below:  
summary(regout5)

##   
## Call:  
## lm(formula = log(fare) ~ log(trip\_seconds) + log(trip\_miles) +   
## trip\_seconds \* trip\_miles + payment\_type, data = no\_levpt\_sample)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.37683 -0.06126 -0.00970 0.04739 0.86029   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.904e-01 6.630e-01 1.343 0.18402   
## log(trip\_seconds) 4.455e-02 1.303e-01 0.342 0.73347   
## log(trip\_miles) -1.402e-01 3.276e-02 -4.279 6.40e-05 \*\*\*  
## trip\_seconds 1.066e-03 2.418e-04 4.409 4.05e-05 \*\*\*  
## trip\_miles 2.917e-01 5.313e-02 5.491 7.35e-07 \*\*\*  
## payment\_typeCredit Card 1.029e-01 3.808e-02 2.702 0.00881 \*\*   
## trip\_seconds:trip\_miles -1.457e-04 3.292e-05 -4.427 3.81e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1595 on 64 degrees of freedom  
## Multiple R-squared: 0.9099, Adjusted R-squared: 0.9014   
## F-statistic: 107.7 on 6 and 64 DF, p-value: < 2.2e-16

#As we have removed levarage points which was impacting ability of model to explain dependend variables, model will be able to predict better.  
cbind("beta"=coef(regout5),confint(regout5))

## beta 2.5 % 97.5 %  
## (Intercept) 0.8904019607 -0.4340951522 2.214899e+00  
## log(trip\_seconds) 0.0445494902 -0.2156768180 3.047758e-01  
## log(trip\_miles) -0.1401605788 -0.2055998291 -7.472133e-02  
## trip\_seconds 0.0010663799 0.0005832520 1.549508e-03  
## trip\_miles 0.2917266845 0.1855931753 3.978602e-01  
## payment\_typeCredit Card 0.1029128776 0.0268309982 1.789948e-01  
## trip\_seconds:trip\_miles -0.0001457422 -0.0002115153 -7.996917e-05

# Analysis 8   
  
# Set a new sample seed and a sample of 100 records  
set.seed(83121176)  
taxi\_sample2 = taxi[sample(1:nrow(taxi),100, replace = FALSE),]  
which(taxi\_sample$tolls !=0)

## integer(0)

taxi\_sample2 = subset(taxi\_sample2,select=-c(tolls))  
new\_sample2= filter(taxi\_sample2, trip\_seconds != 0 & trip\_miles != 0.00 & fare != 0 & trip\_total!=0)  
summary(new\_sample2)

## taxi\_id trip\_seconds trip\_miles fare   
## Min. : 30 Min. : 120.0 Min. : 0.100 Min. : 4.250   
## 1st Qu.:1635 1st Qu.: 375.0 1st Qu.: 0.700 1st Qu.: 6.500   
## Median :3642 Median : 540.0 Median : 1.300 Median : 8.375   
## Mean :3743 Mean : 700.3 Mean : 2.994 Mean :13.009   
## 3rd Qu.:5797 3rd Qu.: 885.0 3rd Qu.: 2.775 3rd Qu.:13.875   
## Max. :8553 Max. :2520.0 Max. :21.600 Max. :52.750   
## tips extras trip\_total payment\_type  
## Min. : 0.000 Min. : 0.0000 Min. : 4.75 Cash :40   
## 1st Qu.: 0.000 1st Qu.: 0.0000 1st Qu.: 8.00 Credit Card:30   
## Median : 0.000 Median : 0.0000 Median : 9.35 Other : 0   
## Mean : 1.674 Mean : 0.8429 Mean : 15.53   
## 3rd Qu.: 2.000 3rd Qu.: 0.0000 3rd Qu.: 14.25   
## Max. :21.190 Max. :32.0000 Max. :105.94

# Best fit Model on new sampled data  
regout6 = lm(log(fare) ~ log(trip\_seconds)+ log(trip\_miles) + trip\_seconds\*trip\_miles+ payment\_type , data = new\_sample2)  
summary(regout6)

##   
## Call:  
## lm(formula = log(fare) ~ log(trip\_seconds) + log(trip\_miles) +   
## trip\_seconds \* trip\_miles + payment\_type, data = new\_sample2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.36230 -0.10174 -0.03912 0.04217 0.55556   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.187e-01 7.645e-01 0.286 0.775789   
## log(trip\_seconds) 1.989e-01 1.431e-01 1.390 0.169414   
## log(trip\_miles) -1.069e-01 3.505e-02 -3.049 0.003357 \*\*   
## trip\_seconds 8.573e-04 2.175e-04 3.941 0.000206 \*\*\*  
## trip\_miles 1.592e-01 2.416e-02 6.590 1.03e-08 \*\*\*  
## payment\_typeCredit Card 7.057e-02 4.728e-02 1.493 0.140473   
## trip\_seconds:trip\_miles -6.293e-05 1.332e-05 -4.725 1.33e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1933 on 63 degrees of freedom  
## Multiple R-squared: 0.914, Adjusted R-squared: 0.9058   
## F-statistic: 111.5 on 6 and 63 DF, p-value: < 2.2e-16

The summary shown above shows the adjusted R square value .914 which is close to what was for first sample(.92).

Also, the p-value for all the variables is significant except Intercept, payment\_type as Credit Card and log(trip\_second).

The significant p value depicts that these variables are significant explaining variance on dependent variable and adding value to our model.

Median Resisdual for our model is close to 0.

So our final model is able to predict 92% percent of variation of trip fare.