Module 6 Assignment

Umakanth Sai Balguri - (U01403700)



Screenshot from Canvas

# Preprocessing

### Loading Library files

rm(list=ls())  
library(rmarkdown)  
library(rio)  
library(moments)  
library(corrplot)

## corrplot 0.84 loaded

library(Hmisc)

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##   
## Attaching package: 'Hmisc'

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

library(carData)  
library(car)

### Loading the file into R

trips=import("6304 Assignment 6 Data.xlsx",  
 sheet="Sheet 1")  
colnames(trips)=tolower(make.names(colnames(trips)))  
trips$payment\_type=as.factor(trips$payment\_type)

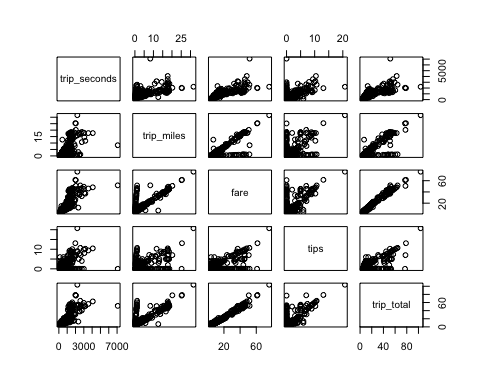
### Selecting Random Samples based on Seed number = U-number and taking a sample of 500 observations for primary data set and 800 observations for secondary data set respectively

set.seed(01403700)  
trips.d1=trips[sample(1:nrow(trips),500,replace=FALSE),]  
set.seed(01403700)  
trips.d2=trips[sample(1:nrow(trips),800,replace=FALSE),]

# Analysis

### Response to Q1

cont.trips.d1=subset(trips.d1,select=c("trip\_seconds","trip\_miles",  
 "fare","tips", "trip\_total"))  
plot(cont.trips.d1)

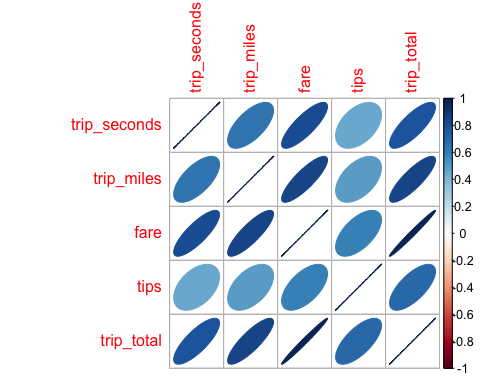


### Response to Q2

xy=cor(cont.trips.d1)  
corrplot(xy,method="number")



corrplot(xy,method="ellipse")



### Response to Q3

regout=lm(trip\_total~trip\_seconds+trip\_miles+fare+tips+payment\_type,data=trips.d1)  
summary(regout)

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + fare +   
## tips + payment\_type, data = trips.d1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.3314 -0.4459 -0.2354 0.5711 15.4353   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.1438126 0.1123544 -1.280 0.20115   
## trip\_seconds -0.0007280 0.0001551 -4.692 3.5e-06 \*\*\*  
## trip\_miles 0.0626923 0.0228822 2.740 0.00637 \*\*   
## fare 1.1023829 0.0132468 83.219 < 2e-16 \*\*\*  
## tips 0.9624307 0.0422058 22.803 < 2e-16 \*\*\*  
## payment\_typeCredit Card -0.0056824 0.1675087 -0.034 0.97295   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.281 on 494 degrees of freedom  
## Multiple R-squared: 0.992, Adjusted R-squared: 0.992   
## F-statistic: 1.23e+04 on 5 and 494 DF, p-value: < 2.2e-16

Based on the above analysis, the impact on Price due to each varible is as follows:

For every 1000 seconds increase in trip\_seconds, the trip\_total decreases by 72.8 cents

For every 10 miles increase in trip\_miles, the trip\_total increases by 62.6 cents

For every 1 dollar increase in fare, the trip\_total increases by 1.10 dollars

For every 1 dollar increase in tips, the trip\_total increases by 96.2 cents

While the payment type is credit card the trip\_total decreases by 0.56 cents

Of All the variables, payment\_type has a p-value>0.05. Hence it is non significant. All the other variables i.e trip\_seconds, trip\_miles fare, tips are highly significant as they have p-value <0.05 (trip\_seconds, fare, tips are highly significant as p-value<0.001)

### Response to Q4

Based on the above co-relation analysis it is evident that all the variables except the payment\_type are of significan importance. Hence, I have created a regression model dropping the payment type.

regout.2=lm(trip\_total~trip\_seconds+trip\_miles+fare+tips,data=trips.d1)  
summary(regout.2)

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + fare +   
## tips, data = trips.d1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.3333 -0.4447 -0.2372 0.5681 15.4350   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.145936 0.093210 -1.566 0.11806   
## trip\_seconds -0.000728 0.000155 -4.697 3.42e-06 \*\*\*  
## trip\_miles 0.062636 0.022800 2.747 0.00623 \*\*   
## fare 1.102483 0.012903 85.443 < 2e-16 \*\*\*  
## tips 0.961417 0.029772 32.292 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.28 on 495 degrees of freedom  
## Multiple R-squared: 0.992, Adjusted R-squared: 0.992   
## F-statistic: 1.541e+04 on 4 and 495 DF, p-value: < 2.2e-16

For every 1000 seconds increase in trip\_seconds, the trip\_total decreases by 72.8 cents

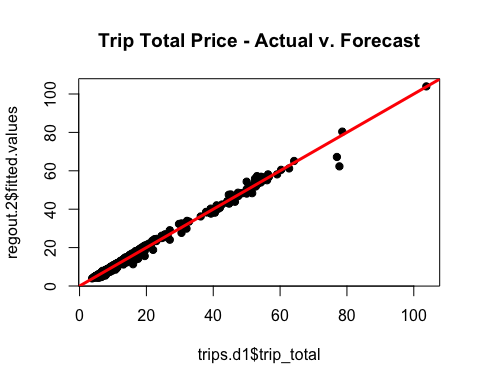
For every 10 miles increase in trip\_miles, the trip\_total increases by 62.6 cents

For every 1 dollar increase in fare, the trip\_total increases by 1.10 dollars

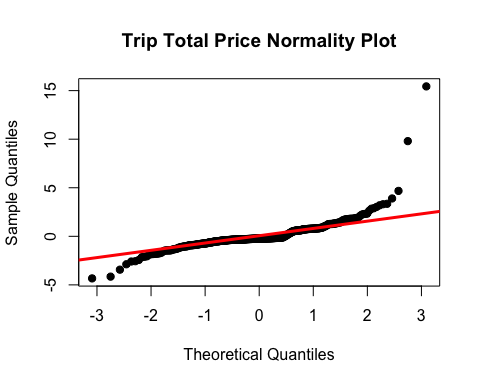
For every 1 dollar increase in tips, the trip\_total increases by 96.1 cents

### Response to Q5

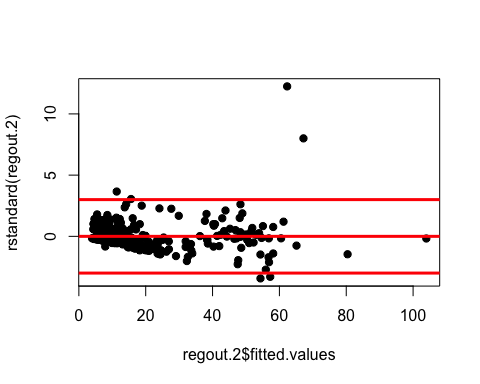
plot(trips.d1$trip\_total,regout.2$fitted.values,pch=19,main="Trip Total Price - Actual v. Forecast")  
abline(0,1,lwd=3,col="red")



qqnorm(regout.2$residuals,pch=19,main="Trip Total Price Normality Plot")  
qqline(regout.2$residuals,lwd=3,col="red")



plot(regout.2$fitted.values,rstandard(regout.2),pch=19)  
abline(0,0,col="red",lwd=3)  
abline(3,0,col="red",lwd=3)  
abline(-3,0,col="red",lwd=3)



Linearity= The suggested model highly observes linearity.

Normality= The suggested model observes normality.

Equality= Except for 5 outliers, the model observes equality.

Overall, the model’s good fit is re-confirmed by the linear model (LINE) analysis.

### Response to Q6

vif(regout.2)

## trip\_seconds trip\_miles fare tips   
## 3.069627 3.546227 6.903901 1.551693

As all the values are less than 8, it can be intrepreted that multicollinearity is not present in the suggested model. This is based on the Variance Inflation Factors (VIF) function output.

### Response to Q7

Creating same model for secondary dataset:

regout.3=lm(trip\_total~trip\_seconds+trip\_miles+fare+tips,data=trips.d2)  
summary(regout.3)

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + fare +   
## tips, data = trips.d2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.2871 -0.4536 -0.2455 0.5850 21.4158   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.1966073 0.0850725 -2.311 0.0211 \*   
## trip\_seconds -0.0007085 0.0001530 -4.631 4.25e-06 \*\*\*  
## trip\_miles 0.0227875 0.0106090 2.148 0.0320 \*   
## fare 1.1058642 0.0098042 112.795 < 2e-16 \*\*\*  
## tips 1.0452184 0.0264231 39.557 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.503 on 795 degrees of freedom  
## Multiple R-squared: 0.9897, Adjusted R-squared: 0.9896   
## F-statistic: 1.903e+04 on 4 and 795 DF, p-value: < 2.2e-16

Based on the P-values, and R squared value, the proposed model is a best fit even for secondary data set. One minor point to note is that the p value for trip\_miles has increased to 0.032. This is not an issue as it is still less than 0.05. (For the previous data set it was less than 0.01)