Assignment 6

Alka Santosh Naik

2023-10-10

#Alka Santosh Naik  
rm(list=ls())  
library(rio)  
library(corrplot)

## corrplot 0.92 loaded

library(car)

## Loading required package: carData

library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

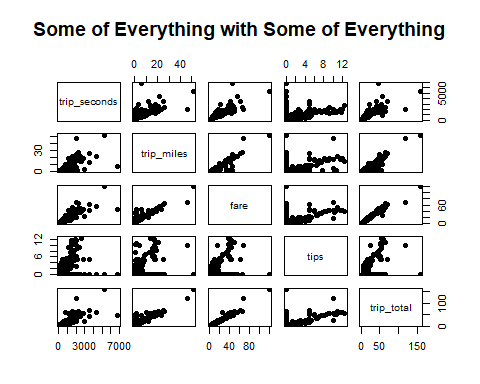
datataxi=import("6304 Module 6 Assignment Data.xlsx")  
colnames(datataxi)=tolower(make.names(colnames(datataxi)))  
set.seed(16999752)  
datataxi$payment\_type=as.factor(datataxi$payment\_type)  
str(datataxi)

## 'data.frame': 10000 obs. of 7 variables:  
## $ taxi\_id : num 3869 2521 5025 3478 6759 ...  
## $ trip\_seconds: num 1380 960 900 240 600 360 180 900 480 960 ...  
## $ trip\_miles : num 6.3 4.3 0.1 1.26 0.1 1.2 0.5 6.9 0.6 5.3 ...  
## $ fare : num 18.25 14.25 11.25 6.5 8.05 ...  
## $ tips : num 0 2.85 0 2 2 0 3 0 3 0 ...  
## $ trip\_total : num 18.2 17.1 11.2 9.5 10.1 ...  
## $ payment\_type: Factor w/ 2 levels "Cash","Credit Card": 2 2 1 2 2 1 2 1 2 1 ...

sample.datataxi1=datataxi[sample(1:nrow(datataxi),500),]  
attach(sample.datataxi1)  
sample.datataxi2=datataxi[sample(1:nrow(datataxi),800),]

Analysis: structure of datataxi.

#1  
some.of.taxi=subset(sample.datataxi1,select=c("trip\_seconds","trip\_miles","fare","tips","trip\_total"))  
plot(some.of.taxi,pch=19,main="Some of Everything with Some of Everything")

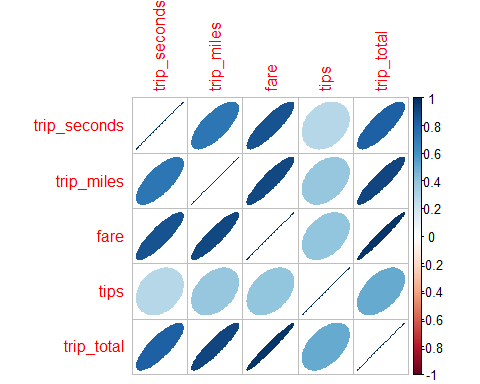


Analysis: scatterplot matrix of the continuous variables only

#2  
cor(some.of.taxi)

## trip\_seconds trip\_miles fare tips trip\_total  
## trip\_seconds 1.0000000 0.7258840 0.8654178 0.2803217 0.8149266  
## trip\_miles 0.7258840 1.0000000 0.9013324 0.3842031 0.9107644  
## fare 0.8654178 0.9013324 1.0000000 0.3989867 0.9776728  
## tips 0.2803217 0.3842031 0.3989867 1.0000000 0.5075140  
## trip\_total 0.8149266 0.9107644 0.9776728 0.5075140 1.0000000

taxicorr=cor(some.of.taxi)  
corrplot(taxicorr,method="ellipse")



Analysis: two correlation matrices of the continuous variables.

#3  
taxi.out1=lm(trip\_total~.-taxi\_id,data=sample.datataxi1)  
summary(taxi.out1)

##   
## Call:  
## lm(formula = trip\_total ~ . - taxi\_id, data = sample.datataxi1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.7740 -0.4185 -0.1221 0.4789 27.6144   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.1417014 0.2028358 -0.699 0.485   
## trip\_seconds -0.0014943 0.0003325 -4.494 8.70e-06 \*\*\*  
## trip\_miles 0.3393174 0.0455745 7.445 4.34e-13 \*\*\*  
## fare 1.0829407 0.0288974 37.475 < 2e-16 \*\*\*  
## tips 0.8982120 0.0755182 11.894 < 2e-16 \*\*\*  
## payment\_typeCredit Card 0.0916929 0.3084377 0.297 0.766   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.349 on 494 degrees of freedom  
## Multiple R-squared: 0.9768, Adjusted R-squared: 0.9766   
## F-statistic: 4169 on 5 and 494 DF, p-value: < 2.2e-16

Analysis: From looking at the P values we can say trip\_seconds, trip\_miles,fare,tips are highly significant and we don’t find any significance for payment type = card.

#4  
  
taxi.relevel=sample.datataxi1  
taxi.relevel$payment\_type=as.factor(taxi.relevel$payment\_type)  
taxi.relevel$payment\_type=relevel(taxi.relevel$payment\_type,"Cash")  
taxi.out2=lm(trip\_total~trip\_seconds+trip\_miles+fare+tips+payment\_type,data=taxi.relevel)  
summary(taxi.out2)

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + fare +   
## tips + payment\_type, data = taxi.relevel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.7740 -0.4185 -0.1221 0.4789 27.6144   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.1417014 0.2028358 -0.699 0.485   
## trip\_seconds -0.0014943 0.0003325 -4.494 8.70e-06 \*\*\*  
## trip\_miles 0.3393174 0.0455745 7.445 4.34e-13 \*\*\*  
## fare 1.0829407 0.0288974 37.475 < 2e-16 \*\*\*  
## tips 0.8982120 0.0755182 11.894 < 2e-16 \*\*\*  
## payment\_typeCredit Card 0.0916929 0.3084377 0.297 0.766   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.349 on 494 degrees of freedom  
## Multiple R-squared: 0.9768, Adjusted R-squared: 0.9766   
## F-statistic: 4169 on 5 and 494 DF, p-value: < 2.2e-16

#excluding payment type   
taxi.out3=lm(trip\_total~trip\_seconds+trip\_miles+fare+tips,data=taxi.relevel)  
summary(taxi.out3)

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + fare +   
## tips, data = taxi.relevel)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.7331 -0.4378 -0.1224 0.4625 27.6156   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.1127997 0.1778547 -0.634 0.526   
## trip\_seconds -0.0014823 0.0003297 -4.495 8.65e-06 \*\*\*  
## trip\_miles 0.3399284 0.0454862 7.473 3.57e-13 \*\*\*  
## fare 1.0812137 0.0282813 38.231 < 2e-16 \*\*\*  
## tips 0.9143232 0.0525435 17.401 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.347 on 495 degrees of freedom  
## Multiple R-squared: 0.9768, Adjusted R-squared: 0.9767   
## F-statistic: 5220 on 4 and 495 DF, p-value: < 2.2e-16

Analysis:

As payment method is not significant we drop the independent variable and see a small improvement in the residual standard errors.

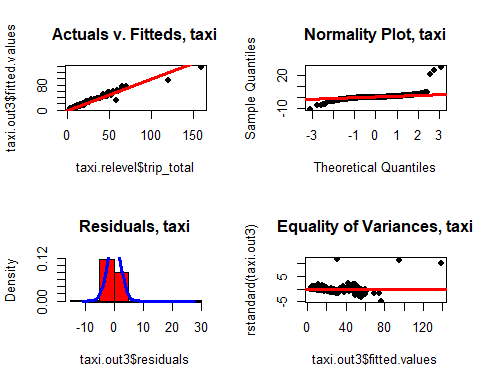
Trip second decreases the price by 0.0014

Trip mile increases the price by 0.3399

Fare increase the price by 1

Tip increase the price by 0.9.

#5  
  
# Assumptions of Regression  
par(mfrow=c(2,2))  
# Linearity  
plot(taxi.relevel$trip\_total,taxi.out3$fitted.values,  
 pch=19,main="Actuals v. Fitteds, taxi")  
abline(0,1,col="red",lwd=3)  
# Normality  
qqnorm(taxi.out3$residuals,pch=19,  
 main="Normality Plot, taxi")  
qqline(taxi.out3$residuals,lwd=3,col="red")  
hist(taxi.out3$residuals,col="red",  
 main="Residuals, taxi",  
 probability=TRUE)  
curve(dnorm(x,mean(taxi.out3$residuals),  
 sd(taxi.out3$residuals)),  
 from=min(taxi.out3$residuals),  
 to=max(taxi.out3$residuals),  
 lwd=3,col="blue",add=TRUE)  
# Equality of Variances  
plot(taxi.out3$fitted.values,rstandard(taxi.out3),  
 pch=19,main="Equality of Variances, taxi")  
abline(0,0,lwd=3,col="red")

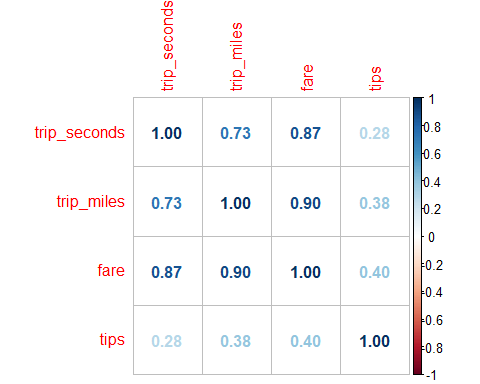


par(mfrow=c(1,1))

Analysis:

From “actual v. fitted, taxi” graph we can say that the data follows linearity, from the normality plot we can say the data is partially normal but from the “residual taxi” data we can confirm the data is normally distributed. “equality of variances taxi” graph we cannot confirms equality of variances as we see some pattern in the data with few outliers.

#6  
taxi.relevel2=subset(taxi.relevel,select=c("trip\_seconds","trip\_miles","fare","tips"))  
taxicorr2=cor(taxi.relevel2)  
corrplot(taxicorr2,method="number")



cor(taxi.relevel2)

## trip\_seconds trip\_miles fare tips  
## trip\_seconds 1.0000000 0.7258840 0.8654178 0.2803217  
## trip\_miles 0.7258840 1.0000000 0.9013324 0.3842031  
## fare 0.8654178 0.9013324 1.0000000 0.3989867  
## tips 0.2803217 0.3842031 0.3989867 1.0000000

Analysis: We see high multicollinearity between the following

1)trip seconds with trip miles and fare

2)trip miles with fare

#7  
taxi.out4=lm(trip\_total~trip\_seconds+trip\_miles+fare+tips,data=sample.datataxi2)  
summary(taxi.out4)

##   
## Call:  
## lm(formula = trip\_total ~ trip\_seconds + trip\_miles + fare +   
## tips, data = sample.datataxi2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.467 -0.749 0.013 0.354 34.165   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.3109807 0.1349537 -2.304 0.0215 \*   
## trip\_seconds -0.0012313 0.0002535 -4.856 1.44e-06 \*\*\*  
## trip\_miles 0.0383385 0.0199353 1.923 0.0548 .   
## fare 1.1027647 0.0164808 66.912 < 2e-16 \*\*\*  
## tips 1.4343083 0.0394141 36.391 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.417 on 795 degrees of freedom  
## Multiple R-squared: 0.9773, Adjusted R-squared: 0.9772   
## F-statistic: 8570 on 4 and 795 DF, p-value: < 2.2e-16

Analysis:

We see improved r square value and decreased residual standard error with increase in number of data.