Regression Analysis Project

Brian Mwangi Kimotho

2024-06-16

Importing the dataset:

library(rio)  
taxi = import("C:/Users/bryan/OneDrive/Documents/USF/Courses/Weekend projects/Taxi Dataset.xlsx")

Converting payment\_type variable from a character variable to a factor variable:

attach(taxi)  
colnames(taxi)=tolower(make.names(colnames(taxi)))  
taxi$payment\_type= as.factor(taxi$payment\_type)  
str(taxi)

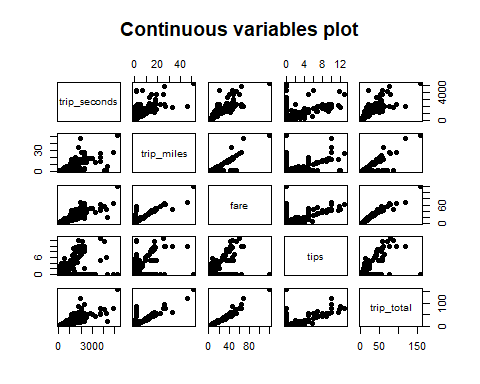
## 'data.frame': 10000 obs. of 7 variables:  
## $ taxi\_id : num 3869 2521 5025 3478 6759 ...  
## $ trip\_seconds: num 1380 960 900 240 600 360 180 900 480 960 ...  
## $ trip\_miles : num 6.3 4.3 0.1 1.26 0.1 1.2 0.5 6.9 0.6 5.3 ...  
## $ fare : num 18.25 14.25 11.25 6.5 8.05 ...  
## $ tips : num 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 ...

For purposes of clear comparison analysis, I decide to create two subsets, a primary dataset and a secondary dataset, from the main dataset. Primary dataset has population size n= 500 and secondary dataset n=800.

set.seed(84234288)  
taxi\_sample1=taxi[sample(1:nrow(taxi),500),]  
set.seed(84234288)  
taxi\_sample2=taxi[sample(1:nrow(taxi),800),]

To identify any early correlations, I created a scatterplot matrix of the continuous variables.

taxi\_sample1cont=subset(taxi\_sample1,select=c("trip\_seconds","trip\_miles",  
 "fare","tips", "trip\_total"))  
plot(taxi\_sample1cont,pch=19,  
 main="Continuous variables plot")



Similarly, I used numbers and ellipses to give further details on the correlations between the continuous variables.

library(corrplot)

## corrplot 0.92 loaded

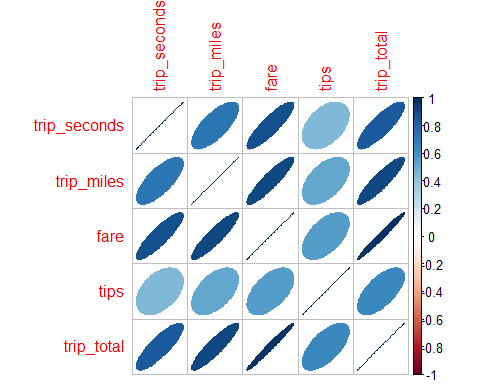
taxi1\_corrrplot=cor(taxi\_sample1cont)  
taxi1\_corrrplot

## trip\_seconds trip\_miles fare tips trip\_total  
## trip\_seconds 1.0000000 0.7286814 0.8714835 0.4430580 0.8350605  
## trip\_miles 0.7286814 1.0000000 0.9019654 0.5143915 0.9011624  
## fare 0.8714835 0.9019654 1.0000000 0.5516603 0.9815618  
## tips 0.4430580 0.5143915 0.5516603 1.0000000 0.6430320  
## trip\_total 0.8350605 0.9011624 0.9815618 0.6430320 1.0000000

corrplot(taxi1\_corrrplot,method="number")



corrplot(taxi1\_corrrplot,method="ellipse")



I used my primary data set to conduct a full regression analysis using trip\_total as the dependent and all other variables (except for taxi\_id) as independents. After the regression, I observed the impact each variable had on price, if any, and whether or not the variable’s impact is significant.

taxi\_sample1model=lm(trip\_total~.-taxi\_id, data=taxi\_sample1)  
summary(taxi\_sample1model)

##   
## Call:  
## lm(formula = trip\_total ~ . - taxi\_id, data = taxi\_sample1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.374 -0.489 -0.050 0.672 26.423   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.5352242 0.2038425 -2.626 0.008916 \*\*   
## trip\_seconds -0.0011423 0.0003254 -3.510 0.000488 \*\*\*  
## trip\_miles 0.1648593 0.0443461 3.718 0.000224 \*\*\*  
## fare 1.1460115 0.0287737 39.828 < 2e-16 \*\*\*  
## tips 1.0352748 0.0753407 13.741 < 2e-16 \*\*\*  
## payment\_typeCredit Card -0.2437133 0.3090529 -0.789 0.430735   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.42 on 494 degrees of freedom  
## Multiple R-squared: 0.9798, Adjusted R-squared: 0.9796   
## F-statistic: 4792 on 5 and 494 DF, p-value: < 2.2e-16

trip\_seconds, trip\_miles, fare, tips variables all have a significant impact on the model due to a p-value of less than 0.05. The payment\_type variable has a lower significance, as shown by the high p-value of 0.43

Furthermore, using my primary dataset, I tried to find a regression model with the best fit to y=trip\_total using any or all of the remaining variables (except for taxi\_id), using my correlation analysis as a guide for this.

taxi\_sample1model\_partial= lm(trip\_total~.-taxi\_id-payment\_type, data= taxi\_sample1)  
summary(taxi\_sample1model\_partial)

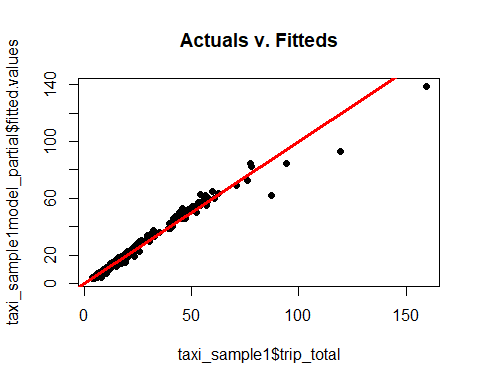
##   
## Call:  
## lm(formula = trip\_total ~ . - taxi\_id - payment\_type, data = taxi\_sample1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.4472 -0.4882 -0.0457 0.6822 26.4672   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.6179695 0.1746948 -3.537 0.000442 \*\*\*  
## trip\_seconds -0.0011634 0.0003242 -3.589 0.000365 \*\*\*  
## trip\_miles 0.1644933 0.0443268 3.711 0.000230 \*\*\*  
## fare 1.1499375 0.0283289 40.592 < 2e-16 \*\*\*  
## tips 0.9937820 0.0539022 18.437 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.419 on 495 degrees of freedom  
## Multiple R-squared: 0.9798, Adjusted R-squared: 0.9796   
## F-statistic: 5994 on 4 and 495 DF, p-value: < 2.2e-16

I chose to use the continuous variables in my model, and excluded taxi\_id since it is used as an identifier for the taxi, and also excluded payment\_type variable as it had little sigificance on my model based on my previous full regression, having a p-value of 0.43.

Each 1000 seconds change in the trip led to a reduction in price by 1.16 dollars Each 100 miles change in the trip led to an increase in price by 16.5 dollars Each fare change in the trip led to an increase in price by 1.15 dollars Each tips change in the trip led to an increase in price by 0.99 dollars

We must consider whether our regression analysis followed the LINE assumptions principles. #Assumptions of Regression #Linearity

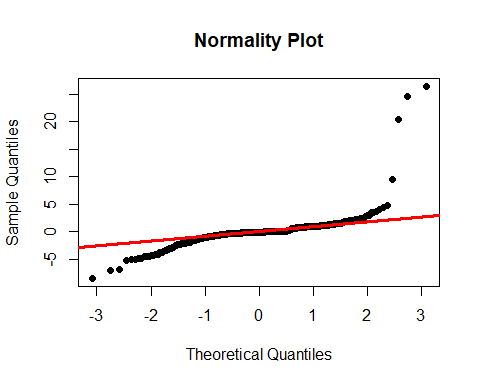
plot(taxi\_sample1$trip\_total,taxi\_sample1model\_partial$fitted.values,  
 pch=19,main="Actuals v. Fitteds")  
abline(0,1,col="red",lwd=3)



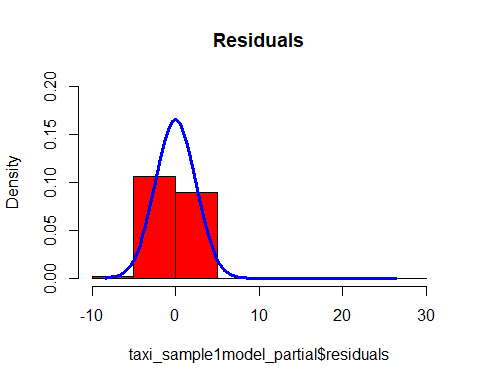
The model exhibits a linear relationship as shown by the plot.

#Normality

qqnorm(taxi\_sample1model\_partial$residuals,pch=19,  
 main="Normality Plot")  
qqline(taxi\_sample1model\_partial$residuals,lwd=3,col="red")



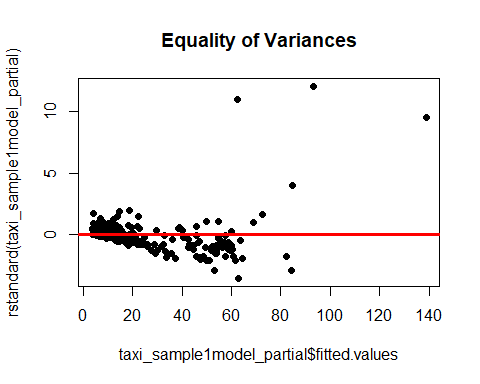
hist(taxi\_sample1model\_partial$residuals,ylim = c(0, 0.20), col="red",  
 main="Residuals",  
 probability=TRUE)  
curve(dnorm(x,mean(taxi\_sample1model\_partial$residuals),  
 sd(taxi\_sample1model\_partial$residuals)),  
 from=min(taxi\_sample1model\_partial$residuals),  
 to=max(taxi\_sample1model\_partial$residuals),  
 lwd=3,col="blue",add=TRUE)



From both the qqplot and histogram, majority of the points exhibit normal distrivution, with a few outliers

# Equality of Variances

plot(taxi\_sample1model\_partial$fitted.values,rstandard(taxi\_sample1model\_partial),  
 pch=19,main="Equality of Variances")  
abline(0,0,lwd=3,col="red")



Ther is no outright pattern, thus the model generally exhibits an equality of variances.

I also decided to use VIF to assess whether multicollinearity was present in the model.

library(car)

## Loading required package: carData

vif(taxi\_sample1model\_partial)

## trip\_seconds trip\_miles fare tips   
## 4.517889 5.790824 11.877216 1.450524

Based on VIF output, the fare variable has a high value of 11.87, thus might already be explained by other variables in the model. The trip\_seconds, trip\_miles and tips variables exhibit little multicollinearity, based on their lower than 5 VIF values, thus can be used in the model.

Using our secondary dataset, I parameterized a model with the same set of independent variables as used in the earlier model, so as to test whether this model is also an acceptable fit to the different set of data.

taxi\_sample2model\_partial= lm(trip\_total~.-taxi\_id-payment\_type, data= taxi\_sample2)  
summary(taxi\_sample2model\_partial)

##   
## Call:  
## lm(formula = trip\_total ~ . - taxi\_id - payment\_type, data = taxi\_sample2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.7819 -0.4749 -0.0623 0.5548 27.3999   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.6065931 0.1418519 -4.276 2.13e-05 \*\*\*  
## trip\_seconds -0.0016554 0.0002636 -6.279 5.61e-10 \*\*\*  
## trip\_miles 0.1217634 0.0288190 4.225 2.66e-05 \*\*\*  
## fare 1.1892261 0.0207471 57.320 < 2e-16 \*\*\*  
## tips 0.9494493 0.0442616 21.451 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.545 on 795 degrees of freedom  
## Multiple R-squared: 0.9764, Adjusted R-squared: 0.9763   
## F-statistic: 8229 on 4 and 795 DF, p-value: < 2.2e-16

The new model based on the second sample is also an acceptable fit to the different set of data. This is shown by the small P-values of all the independent variable coefficients. Similarly, the model explains about 98% of the data, as shown by the R-squared and adjusted R squared.