**Regression Project**

**Rashesh Kothari**

The following analysis has been done on the data of taxi cab trips in the City of Chicago during 2016.

**rm**(list=**ls**())  
**library**(rio)  
**library**(dplyr)

**library**(sqldf)

**library**(corrplot)

*#Read in data.*  
uberdata = **read.csv**("6304 Regression Project Data.csv")…………… imports the csv file “6304 Regression Project Data.csv”  
**colnames**(uberdata)=**tolower**(**make.names**(**colnames**(uberdata)))  
**attach**(uberdata)  
  
*#setting a seed for sampling the data*  
**set.seed**(98368729) …………… Sets a seed as numeric part of my UID i.e 98368729  
*#taking a sample out of a dataset*  
ubersamp =uberdata[**sample**(1**:nrow**(uberdata),100,replace=FALSE),]…Takes the sample of size 100 from the imported file

*#removing the discrepancies*  
ubersample = ubersamp[ubersamp**$**trip\_seconds **!=** 0 **&** ubersamp**$**trip\_miles **!=** 0,]

Since there are few records where seconds and miles are 0, taking them would reduce the quality of data, hence it is a good practice to remove them

*#check if there's any null values present in the sample.*  
**sum**(**is.na**(ubersample))

## [1] 0

*#Summary to check*  
**summary**(ubersample)

##     taxi\_id      trip\_seconds      trip\_miles          fare          
##  Min.   : 235   Min.   :  60.0   Min.   : 0.040   Min.   :  3.450    
##  1st Qu.:1986   1st Qu.: 300.0   1st Qu.: 0.700   1st Qu.:  5.750    
##  Median :3955   Median : 540.0   Median : 1.550   Median :  8.625    
##  Mean   :4183   Mean   : 726.7   Mean   : 4.115   Mean   : 14.019    
##  3rd Qu.:6232   3rd Qu.: 795.0   3rd Qu.: 3.250   3rd Qu.: 13.562    
##  Max.   :8728   Max.   :3660.0   Max.   :45.000   Max.   :107.500    
##       tips            tolls       extras         trip\_total       
##  Min.   : 0.000   Min.   :0   Min.   : 0.000   Min.   :  3.450    
##  1st Qu.: 0.000   1st Qu.:0   1st Qu.: 0.000   1st Qu.:  7.438    
##  Median : 0.000   Median :0   Median : 0.000   Median : 10.250    
##  Mean   : 2.110   Mean   :0   Mean   : 1.549   Mean   : 17.678    
##  3rd Qu.: 2.375   3rd Qu.:0   3rd Qu.: 1.000   3rd Qu.: 14.750    
##  Max.   :21.800   Max.   :0   Max.   :32.000   Max.   :130.800    
##       payment\_type  
##  Cash       :34     
##  Credit Card:37     
##  Other      : 1   

Since tolls are 0, I am not considering the column in analysis  
ubersample = **select**(ubersample, **-c**(tolls,taxi\_id))……………Removing Tolls and Taxi\_Id  
*#replaceing NA with 0, if any*  
ubersample[**is.na**(ubersample)] = 0  
*#converting trip seconds to trip minutes and renaming the column*  
ubersample**$**trip\_seconds= (ubersample**$**trip\_seconds**/**60)  
**colnames**(ubersample)[**colnames**(ubersample)**==**"trip\_seconds"] <- "trip\_minutes"  
**attach**(ubersample)

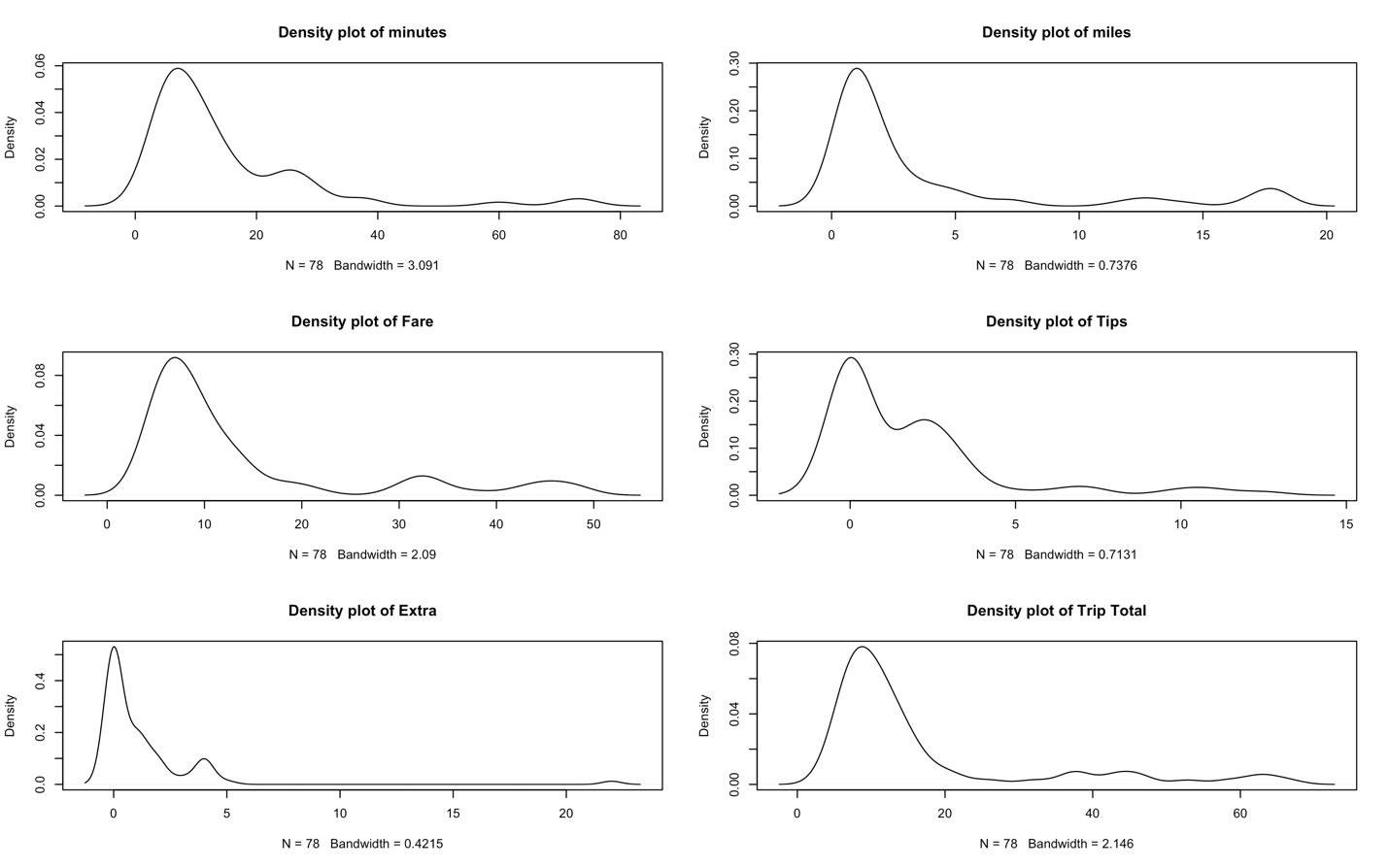
|  |
| --- |
| 1. Using your cleansed sample data, provide summaries and density plots of each of the continuous variables in your data set with the exception of taxi\_id. Explain any apparent differences in the statistical distributions of these variables in your sample data. |

*#Analysis 1: Plotting the density graph*  
**par**(mfrow=**c**(3,2))   
**plot**(**density**(trip\_minutes),main = "Density plot of minutes" , data=ubersample)

**plot**(**density**(fare),main = "Density plot of fare", data=ubersample)

**plot**(**density**(tips),main = "Density plot of tips",  data=ubersample)

**plot**(**density**(ubersample**$**extra),main = "Density plot of extra")  
**plot**(**density**(trip\_total),main = "Density plot of trip\_total", data=ubersample)



|  |
| --- |
| Conclusion:   1. **Density plot of Minutes**: We can see that distribution of data is right-skewed distribution. i.e. Majority of records are present in the **range of 0 to 40mins**. There are very few rides which has trip time more than 40mins. This would help the business when trying a new promotional codes, they could target the users who has time>40, so that the number of people would increase and thus the revenue will increase. 2. **Density plot of Miles:** Majority of records are present in the range of **0 to 6 miles.** There are few rides which are very long rides (between 17 to 20 miles) It seems that people are highly inclined towards having short trips. 3. **Density plot of Fare:** Fare ranges between 0 to $20. There are bumps near $30-$50 which shows that there are few rides which has a fare > $30. It seems that Fare has some relationship with minutes and miles. Because both minutes and miles has a bump towards the tail 4. **Density plot of Tips:** People tend to tip between $0 to $4. However, there are few generous people who tend to give more than $5. 5. **Density of Extra:** Extras are very less, just between $0 to $3. There’s an asymptote after $5. 6. **Density of Trip total:** The rightly-skewed graph shows that the majority of trip total is between $0 to $20. However, there are two bumps (similar to minutes and miles) between $40 to $70. |

*#Plotting boxplots*  
**par**(mfrow=**c**(3,2))

*#Minutes*  
**boxplot**(trip\_minutes, main="Trip Minutes", data = ubersample)  
**which**(ubersample**$**trip\_minutes **>** 27.5)

## [1]  3 11 22 27 40

*#(3rd qu - 1st qud )\* 1.5 + 3rd qu.*  
*#Miles*  
**boxplot**(ubersample**$**trip\_miles, main="Trip Miles")  
**which**(ubersample**$**trip\_miles **>** 5.7)

##  [1]  3  7 11 22 26 27 32 40 55 66

*#Fare*  
**boxplot**(ubersample**$**fare , main="Trip Fare")  
**which**(ubersample**$**fare **>** 5.7)

##  [1]  1  2  3  5  7  8 10 11 12 13 14 15 16 17 18 20 21 22 24 25 26 27 28  
## [24] 29 31 32 33 34 36 37 39 40 42 44 45 46 48 49 51 52 53 54 55 56 57 59  
## [47] 60 62 63 65 66 67 68 70 71 72

*#Tips*  
**boxplot**(tips, main="Tips", data=ubersample)  
*#Extra*  
**boxplot**(extras, main="Extras", data=ubersample)  
*#Total*  
**boxplot**(trip\_total, main="Trip Total", data=ubersample)

#removing the outliers

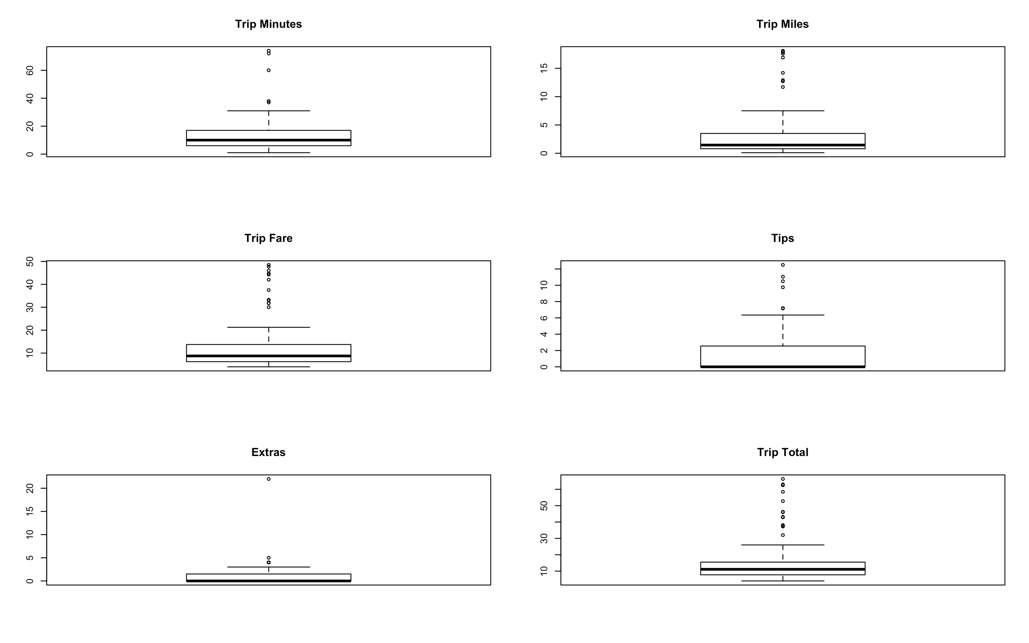
ubersample1 = ubersample[which(ubersample$trip\_miles < 5.7) ,]

ubersample2 = ubersample1[which(ubersample1$trip\_minutes < 27.5),]

ubersample3 = ubersample2 [which(ubersample2$tips < 6),]

ubersample4 = ubersample3[which(ubersample3$extras < 4),]

|  |
| --- |
| After removing the outliers from the sample, we’ve 62 records.  Outliers = (3rd quartile - 1st quartile )\* 1.5 + 3rd quartile |



|  |
| --- |
| From the boxplots we can see that there are few outliers which lies beyond the 3rd quartile in each of the factors. |

|  |
| --- |
| 1. Using the payment\_type factor variable and your cleansed sample data, provide a table of the number of cases in each level of payment\_type. |

*#Analysis 2*  
**sqldf**("select payment\_type,count(payment\_type) as counts  
     from ubersample group by payment\_type")

##   payment\_type counts  
## 1         Cash     34  
## 2  Credit Card     37  
## 3        Other      1

|  |  |
| --- | --- |
| Payment Type | Counts |
| Cash | 34 |
| Credit Card | 37 |
| Other | 1 |

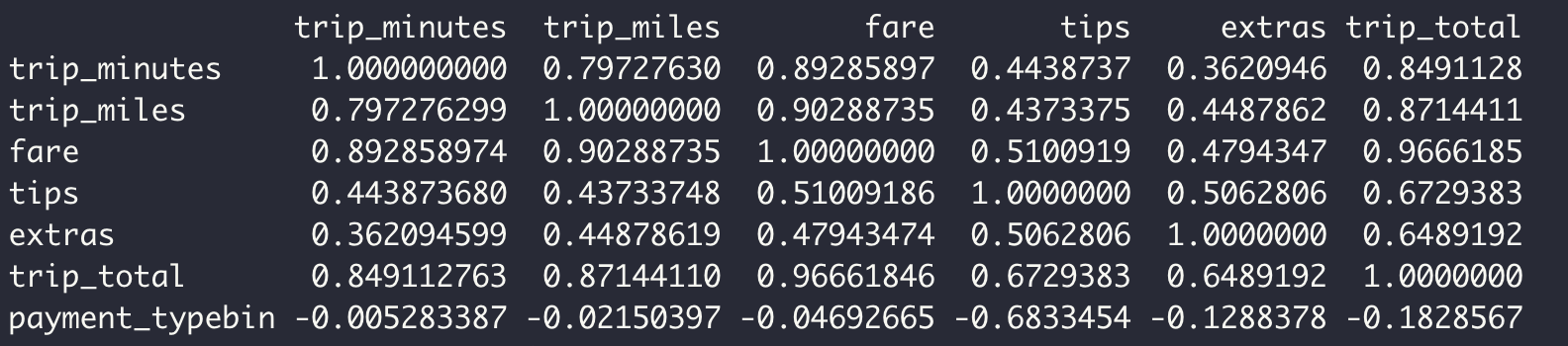
|  |
| --- |
| 1. Construct an easily read and easily understood correlation matrix using all continuous variables except taxi\_id. Give a brief interpretation of the matrix understandable by a non-statistician. |

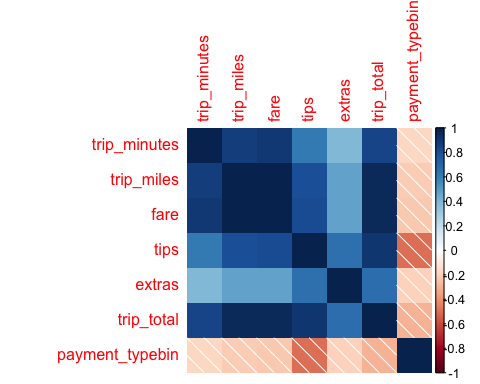
*#Analysis 3*  
*#Making sure that payment type has no NA values*  
ubersample[**is.na**(payment\_type)] = 0  
*#Converting the text to binary*

|  |
| --- |
| Since, the correlation works only for numeric values, I have converted the payment type string to a binary. |

**for**(i **in** 1**:length**(ubersample**$**payment\_type)) {  
 **if**(ubersample**$**payment\_type[i]**==** 'Cash') {  
   ubersample**$**payment\_typebin[i]=1  
 }   
 **if**(ubersample**$**payment\_type[i]**==** 'Credit Card') {  
   ubersample**$**payment\_typebin[i]=0  
 }   
 **if**(ubersample**$**payment\_type[i]**==** 'Others')  {  
   ubersample**$**payment\_typebin[i]  = -1  
 }  
}  
**par**(mfrow=**c**(1,1))  
correlation = **cor**(ubersample[**sapply**(ubersample, is.numeric)])

correlation  
**corrplot**(correlation, method = "shade")





|  |
| --- |
| From the correlation matrix we can conclude the following observations:   1. There’s a high correlation between trip\_minutes and trip\_miles : 0.79 2. Similarly, correlation(trip\_minutes , fare) is 0.89 3. Correlation between Miles and Fare is the hight at 0.90 4. Highest correlation is observed between the fare and trip\_total at 0.96 5. We can see that there’s almost no relation for the Payment Type. It is independent of any other factors.   We can see that the fare/trip\_total is dependent on very few factors: Trip\_minutes and trip\_miles. |

|  |
| --- |
| 1. Using fare as the dependent variable, build a regression model using trip\_seconds, trip\_miles, and payment\_type as potential independent variables. Evaluate the quality of fit of the model to your cleansed data. Explain the impact each independent variable in your model on the dependent variable, considering the 95% confidence interval on the beta coefficients. |

*#Analysis 4*  
*#Using Fare as the dependent variable, building a regression model*  
fare\_minutes\_miles\_payment = **lm**(ubersample**$**fare**~**  
                  ubersample**$**trip\_minutes  
                **+** ubersample**$**trip\_miles  
                **+** ubersample**$**payment\_typebin)  
**summary**(fare\_minutes\_miles\_payment)

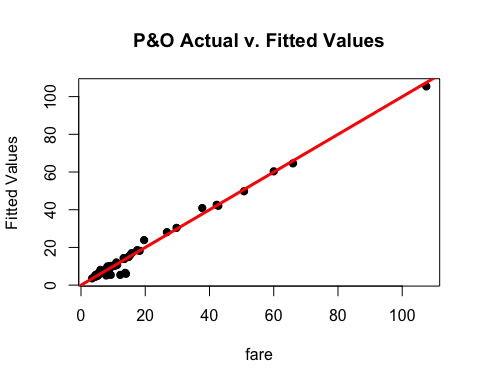
##   
## Call:  
## lm(formula = ubersample$fare ~ ubersample$trip\_minutes + ubersample$trip\_miles +   
##     ubersample$payment\_typebin)  
##   
## Residuals:  
##     Min      1Q  Median      3Q     Max   
## -4.2454 -0.7818 -0.3569 -0.0248  7.9821   
##   
## Coefficients:  
##                            Estimate Std. Error t value Pr(>|t|)      
## (Intercept)                 3.80038    0.44572   8.526 2.43e-12 \*\*\*  
## ubersample$trip\_minutes     0.17747    0.04119   4.308 5.43e-05 \*\*\*  
## ubersample$trip\_miles       2.01767    0.06457  31.248  < 2e-16 \*\*\*  
## ubersample$payment\_typebin -0.49314    0.46771  -1.054    0.295      
## ---  
## Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.932 on 68 degrees of freedom  
## Multiple R-squared:  0.9874, Adjusted R-squared:  0.9869   
## F-statistic:  1781 on 3 and 68 DF,  p-value: < 2.2e-16

*#Chacking the confidence interval*  
**confint**(fare\_minutes\_miles\_payment)

##                                  2.5 %    97.5 %  
## (Intercept)                 2.91096058 4.6897904  
## ubersample$trip\_minutes     0.09526829 0.2596704  
## ubersample$trip\_miles       1.88882677 2.1465216  
## ubersample$payment\_typebin -1.42643702 0.4401647

|  |
| --- |
| Looking at the p-values, of payment\_type, it is more than 0.05. So, we fail to reject the null hypothesis.  Forming a linear model for Fare using trip\_minutes, trip\_miles, and payment typebin, we get the equation as follows:  F(fare) = 3.88 + 0.177(trip\_minutes) + 2.01(trip\_miles)   1. Every minute spend in the cab will increase the fare by $0.177 and it ranges from $0.09 to $0.25 2. For every mile, the fare will increase by $2.01 and it ranges from $1.88 to $2.14   Moreover, the R-squared term is 0.98. Which says that this model fits our data perfectly. The Confidence internal has the range very tight, so this will help us in predicting the fare more accurately.  Looking at the actual vs predicted values, most of the points lies on the regression line and thus we can say that there’s a linear relationship between fare and minutes and miles. |

*#Plotting a Actual vs Fitted Values graph*  
**plot**(ubersample**$**fare,fare\_minutes\_miles\_payment**$**fitted.values,pch=19,main="P&O Actual v. Fitted Values",  
    xlab = "fare", ylab = "Fitted Values")  
**abline**(0,1,col="red",lwd=3)



|  |
| --- |
| 1. Investigate relevant interactions and common independent variable transforms to determine if adding these to your model will result in a better model fit. Depending on your random data selection you may find it necessary to do some additional cleansing of your data in order to get a better model fit for the majority of data points. |

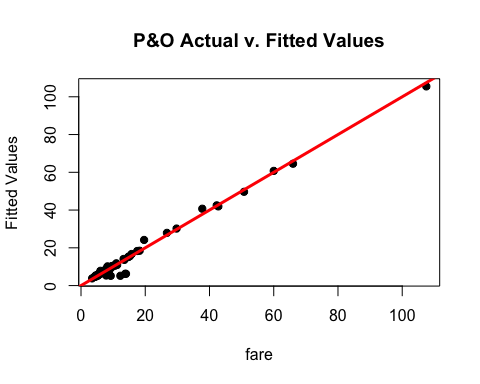
*#Analysis  5*

|  |
| --- |
| Since the correlation between fare and the payment type is almost negligible, I let’s see if removing the term has any effect on the linear regression model. |

*#Removing Payment\_Type*  
fare\_minutes\_miles = **lm**(ubersample**$**fare**~**  
                   ubersample**$**trip\_minutes**+**  
                   ubersample**$**trip\_miles)  
**summary**(fare\_minutes\_miles)

##   
## Call:  
## lm(formula = ubersample$fare ~ ubersample$trip\_minutes + ubersample$trip\_miles)  
##   
## Residuals:  
##     Min      1Q  Median      3Q     Max   
## -4.5245 -0.6130 -0.4483 -0.1363  7.7586   
##   
## Coefficients:  
##                         Estimate Std. Error t value Pr(>|t|)      
## (Intercept)              3.54557    0.37481   9.460 4.38e-14 \*\*\*  
## ubersample$trip\_minutes  0.17619    0.04121   4.275 6.00e-05 \*\*\*  
## ubersample$trip\_miles    2.02677    0.06404  31.647  < 2e-16 \*\*\*  
## ---  
## Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.934 on 69 degrees of freedom  
## Multiple R-squared:  0.9872, Adjusted R-squared:  0.9869   
## F-statistic:  2667 on 2 and 69 DF,  p-value: < 2.2e-16

**plot**(ubersample**$**fare,fare\_minutes\_miles**$**fitted.values,pch=19,main="P&O Actual v. Fitted Values",  
    xlab = "fare", ylab = "Fitted Values")  
**abline**(0,1,col="red",lwd=3)

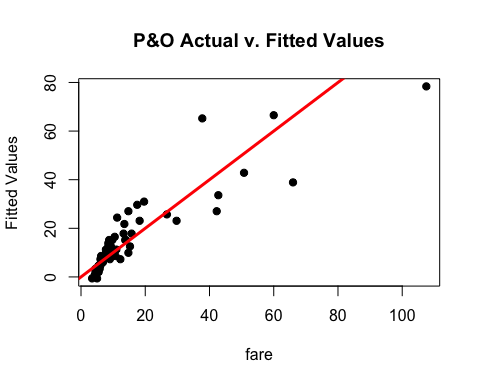


|  |
| --- |
| We can see that there’s no difference in any of the terms:   1. Equation remains the same as: F(fare) = 3.88 + 0.177(trip\_minutes) + 2.01(trip\_miles) 2. Multiple R-squared:  0.9872 Is also same.   So, instead of building a model based on payment type, this can be used alternatively. |

***#Removing Miles***  
fare\_minutes = **lm**(ubersample**$**fare**~**  
                   ubersample**$**trip\_minutes)  
**summary**(fare\_minutes)

##   
## Call:  
## lm(formula = ubersample$fare ~ ubersample$trip\_minutes)  
##   
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -27.4649  -3.2567   0.4593   2.5297  29.1205   
##   
## Coefficients:  
##                         Estimate Std. Error t value Pr(>|t|)      
## (Intercept)             -1.92447    1.30062   -1.48    0.143      
## ubersample$trip\_minutes  1.31646    0.07821   16.83   <2e-16 \*\*\*  
## ---  
## Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.563 on 70 degrees of freedom  
## Multiple R-squared:  0.8019, Adjusted R-squared:  0.799   
## F-statistic: 283.3 on 1 and 70 DF,  p-value: < 2.2e-16

**plot**(ubersample**$**fare,fare\_minutes**$**fitted.values,pch=19,main="P&O Actual v. Fitted Values",  
    xlab = "fare", ylab = "Fitted Values")  
**abline**(0,1,col="red",lwd=3)

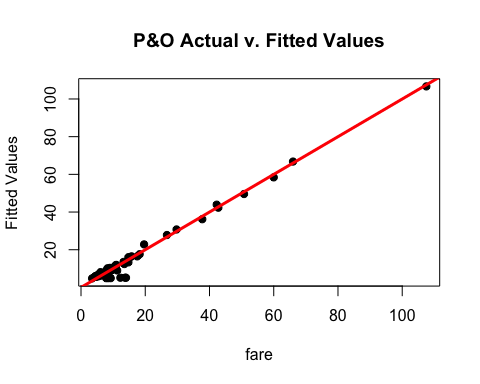


|  |
| --- |
| Building a model just by using minutes, it reduces the accuracy significantly. Multiple R-squared value is just about 80%.  Points are closer to the regression line below 30  However between 30-65 few points are away from the regression line |

*#Removing Minutes, Keeping Miles*  
fare\_miles = **lm**(ubersample**$**fare**~**  
                     ubersample**$**trip\_miles)  
**summary**(fare\_miles)

##   
## Call:  
## lm(formula = ubersample$fare ~ ubersample$trip\_miles)  
##   
## Residuals:  
##     Min      1Q  Median      3Q     Max   
## -3.1837 -1.2105 -0.6961  0.4470  8.8525   
##   
## Coefficients:  
##                       Estimate Std. Error t value Pr(>|t|)      
## (Intercept)            4.69431    0.29182   16.09   <2e-16 \*\*\*  
## ubersample$trip\_miles  2.26618    0.03471   65.30   <2e-16 \*\*\*  
## ---  
## Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.159 on 70 degrees of freedom  
## Multiple R-squared:  0.9838, Adjusted R-squared:  0.9836   
## F-statistic:  4263 on 1 and 70 DF,  p-value: < 2.2e-16

**plot**(ubersample**$**fare,fare\_miles**$**fitted.values,pch=19,main="P&O Actual v. Fitted Values",  
    xlab = "fare", ylab = "Fitted Values")  
**abline**(0,1,col="red",lwd=3)

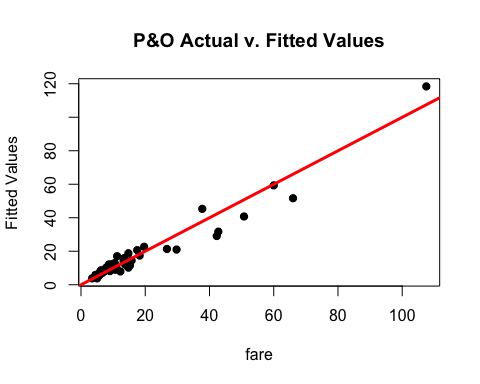


|  |
| --- |
| Building a model just by using miles, it keeps the accuracy almost similar to the model where we’ve used minutes + miles . Multiple R-squared value is just about 98.38%. |

**par**(mfrow=**c**(1,1))   
*#Squaring the Miles*  
sq.miles = ubersample**$**trip\_miles **\*\*** 2   
fare\_miles2\_minutes = **lm**(ubersample**$**fare**~**  
                    ubersample**$**trip\_minutes  
                  **+** sq.miles)  
  
**summary**(fare\_miles2\_minutes)

##   
## Call:  
## lm(formula = ubersample$fare ~ ubersample$trip\_minutes + sq.miles)  
##   
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -10.9119  -1.6734  -0.6481   0.6240  14.3575   
##   
## Coefficients:  
##                         Estimate Std. Error t value Pr(>|t|)      
## (Intercept)             3.097718   0.787227   3.935 0.000196 \*\*\*  
## ubersample$trip\_minutes 0.690857   0.062631  11.031  < 2e-16 \*\*\*  
## sq.miles                0.036134   0.002705  13.361  < 2e-16 \*\*\*  
## ---  
## Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.022 on 69 degrees of freedom  
## Multiple R-squared:  0.9448, Adjusted R-squared:  0.9432   
## F-statistic: 590.1 on 2 and 69 DF,  p-value: < 2.2e-16

**plot**(ubersample**$**fare,fare\_miles2\_minutes**$**fitted.values,pch=19,main="P&O Actual v. Fitted Values",  
    xlab = "fare", ylab = "Fitted Values")  
**abline**(0,1,col="red",lwd=3)   *#...compare it with normal minute+miles graph*



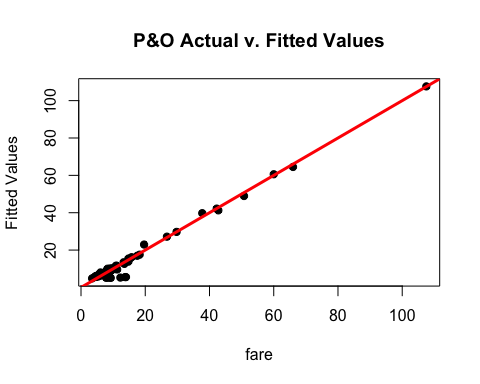
|  |
| --- |
| If we’re too see the effect of squared terms on the model, we take the squared terms of miles and built the following model:  F(fare) = 3.09 +(0.690857)Minutes + 0.036134 Miles^2  Using this model reduces the accuracy significantly ~94%  Points are closer to the regression line below 30 however, between 35-60 points are little far away from the regression line |

*#Squaring the Minutes*

sq.minutes = ubersample**$**trip\_minutes **\*\*** 2   
fare\_miles\_minutes2 = **lm**(ubersample**$**fare**~**  
                          ubersample**$**trip\_miles  
                        **+** sq.minutes)  
  
**summary**(fare\_miles\_minutes2)

##   
## Call:  
## lm(formula = ubersample$fare ~ ubersample$trip\_miles + sq.minutes)  
##   
## Residuals:  
##     Min      1Q  Median      3Q     Max   
## -3.3205 -1.1278 -0.5272  0.2517  8.4492   
##   
## Coefficients:  
##                        Estimate Std. Error t value Pr(>|t|)      
## (Intercept)           4.7550094  0.2800079   16.98  < 2e-16 \*\*\*  
## ubersample$trip\_miles 2.1021439  0.0684564   30.71  < 2e-16 \*\*\*  
## sq.minutes            0.0022214  0.0008107    2.74  0.00782 \*\*   
## ---  
## Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.065 on 69 degrees of freedom  
## Multiple R-squared:  0.9854, Adjusted R-squared:  0.985   
## F-statistic:  2334 on 2 and 69 DF,  p-value: < 2.2e-16

**plot**(ubersample**$**fare,fare\_miles\_minutes2**$**fitted.values,pch=19,main="P&O Actual v. Fitted Values",  
    xlab = "fare", ylab = "Fitted Values")  
**abline**(0,1,col="red",lwd=3)



|  |
| --- |
| If we’re too see the effect of squared terms on the model, we take the squared terms of minutes and built the following model:  F(fare) = 4.7 + (0.0022)Minutes^2 + 2.1 Miles  Using this model keeps the accuracy to about ~98%  But looking at the beta value, it is much smaller for the minutes. |

|  |
| --- |
| Since, fare\_minutes\_miles provides the best accuracy (98.72%) and it consists of simple terms of minutes and miles, it is the ideal model. |

Below is the comparison table to analyze the output more clearly:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Trip\_Minutes | Trip\_miles | Trip\_Minutes ^ 2 | Trip\_miles^2 | R-squared |
| Y | Y |  |  | 0.9872 |
| Y |  |  |  | 0.8019 |
|  | Y |  |  | 0.9838 |
| Y |  |  | Y | 0.9448 |
|  | Y | Y |  | 0.9854 |

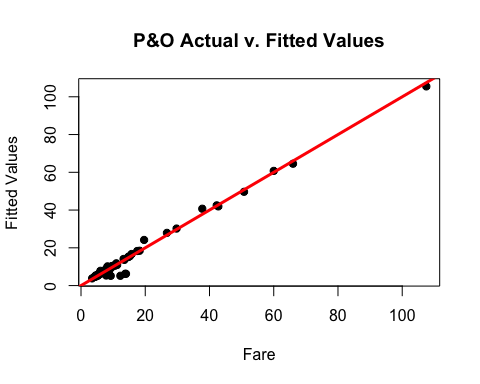
I’ve run the same linear regression models after removing the outliers: Below is the caparison table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Trip\_Minutes | Trip\_miles | Trip\_Minutes ^ 2 | Trip\_miles^2 | R-squared |
| Y | Y |  |  | 0.8783 |
| Y |  |  |  | 0.6940 |
|  | Y |  |  | 0.6737 |
| Y |  |  | Y | 0.8954 |
|  | Y | Y |  | 0.8392 |

|  |
| --- |
| Removing the outliers from the dataset in fact reduces the accuracy to 87%. It is unusual but rare in some cases where the effect of removing the outliers decreases the least squared values. It this case, it might be possible that due to small sample size, it is affecting the R-squared value. |

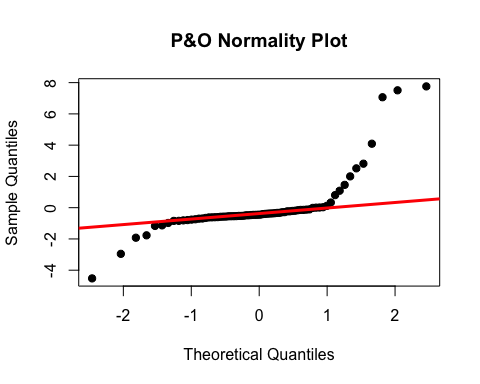
|  |
| --- |
| 1. Of the various combinations you ran in Step 5, report the model which provides what you deem as the “best fit” to your sample data. Explain why you selected this particular model and show the standard R regression output for the model. Evaluate and explain your model’s conformity to the LINE assumptions of regression |

*#Analysis 6*  
*#LINE assumptions*  
*#Linearity*  
**plot**(ubersample**$**fare,fare\_minutes\_miles**$**fitted.values,pch=19,main="P&O Actual v. Fitted Values",  
    xlab = "Fare", ylab= "Fitted Values")  
**abline**(0,1,col="red",lwd=3)



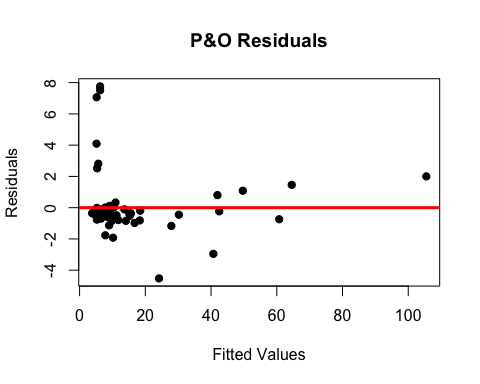
|  |
| --- |
| Plotting a graph between actual values and predicted values of linear regression, there seems to have a linear relation between fare, minutes and miles. |

*#Normality*  
**qqnorm**(fare\_minutes\_miles**$**residuals,pch=19,main="P&O Normality Plot")  
**qqline**(fare\_minutes\_miles**$**residuals,col="red",lwd=3)



|  |
| --- |
| Normality is to check how the residuals are normally distributed. It’s good when the residuals follows the diagonal line, but in this case we can see that few points after 1 are not following the line. So, we can say that errors are not normally distributed. |

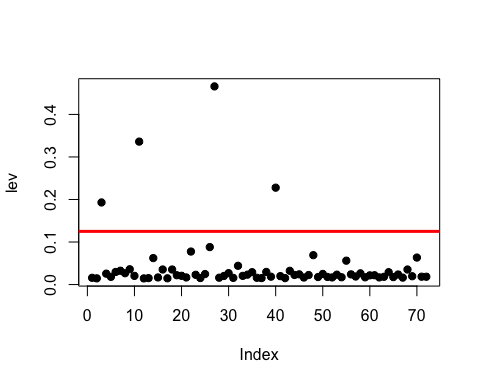
*#Equality of Variances*  
**plot**(fare\_minutes\_miles**$**fitted.values,fare\_minutes\_miles**$**residuals,pch=19,main="P&O Residuals",  
    xlab="Fitted Values", ylab  = "Residuals")  
**abline**(0,0,col="red",lwd=3)



|  |
| --- |
| This plot shows the homogeneity of variance of the residuals. Here I don’t think there’s any pattern among the residuals. Hence the data obeys homoscedasticity. |

|  |
| --- |
| 1. Investigate and remove any data points deemed to have an inappropriately high leverage in determining the plot of the model. Rerun your model without these points and evaluate the quality of fit in this final regression model. |

*#Analysis 7*  
*#Checking the leverage points*  
lev=**hat**(**model.matrix**(fare\_minutes\_miles))  
**plot**(lev,pch=19)  
**abline**(3**\*mean**(lev),0,col="red",lwd=3)



|  |
| --- |
| Here, we can see that there are 4 points above the leverage line. |

**which** (lev**>**3**\*mean**(lev))

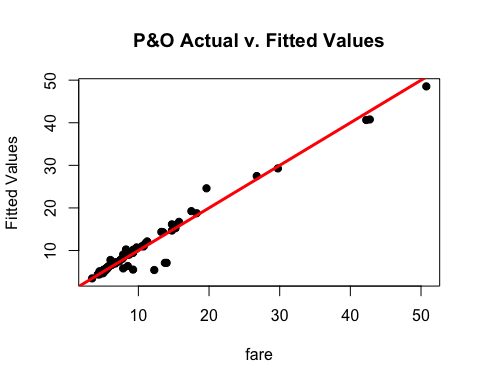
## [1]  3 11 27 40

*#removing the records for which the leverage is high*  
sample\_without\_leverage\_points=  ubersample[**which** (lev**<**3**\*mean**(lev)) ,]  
sample\_without\_leverage\_points\_lm  =  **lm**(sample\_without\_leverage\_points**$**fare**~**  
                                       sample\_without\_leverage\_points**$**trip\_minutes**+**  
                                       sample\_without\_leverage\_points**$**trip\_miles)  
**summary**(sample\_without\_leverage\_points\_lm)

##   
## Call:  
## lm(formula = sample\_without\_leverage\_points$fare ~ sample\_without\_leverage\_points$trip\_minutes +   
##     sample\_without\_leverage\_points$trip\_miles)  
##   
## Residuals:  
##     Min      1Q  Median      3Q     Max   
## -4.9663 -0.6882 -0.3682 -0.0717  6.9011   
##   
## Coefficients:  
##                                             Estimate Std. Error t value  
## (Intercept)                                  3.10131    0.43478   7.133  
## sample\_without\_leverage\_points$trip\_minutes  0.27959    0.05473   5.109  
## sample\_without\_leverage\_points$trip\_miles    1.81440    0.09532  19.035  
##                                             Pr(>|t|)      
## (Intercept)                                 1.01e-09 \*\*\*  
## sample\_without\_leverage\_points$trip\_minutes 3.06e-06 \*\*\*  
## sample\_without\_leverage\_points$trip\_miles    < 2e-16 \*\*\*  
## ---  
## Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.862 on 65 degrees of freedom  
## Multiple R-squared:  0.9583, Adjusted R-squared:  0.9571   
## F-statistic: 747.5 on 2 and 65 DF,  p-value: < 2.2e-16

**plot**(sample\_without\_leverage\_points**$**fare,sample\_without\_leverage\_points\_lm**$**fitted.values,pch=19,main="P&O Actual v. Fitted Values",  
    xlab = "fare", ylab = "Fitted Values")  
**abline**(0,1,col="red",lwd=3)

|  |
| --- |
| Removing the high leveraged points from the dataset in fact reduces the accuracy to 95%. It is unusual but rare in some cases where the effect of removing the outliers decreases the least squared values. It this case, it might be possible that due to small sample size, it is affecting the R-squared value. |



|  |
| --- |
| 1. Return to the full data set of 1.7 million cases. Pull another sample of n=100 cases. (Be sure to use a new random number seed of the numerical portion of your U number plus 5.) To this data set apply the same cleansing procedures you used on your original sample data set. Referring to the model you developed in Step 6 above, apply that model to the new random set of data and evaluate how well the model fits this second data set. |

*#Analysis 8*  
*#Setting random seed as  Uid + 5*  
**set.seed**(98368734)  
ubernewsample =uberdata[**sample**(1**:nrow**(uberdata),100,replace=FALSE),]  
ubernewsample = ubernewsample[ubernewsample**$**trip\_seconds **!=** 0 **&** ubernewsample**$**trip\_miles **!=** 0,]  
  
*#Doing the same preprocessing on new Sample*  
*#check if there's any null values present in the sample.*  
**sum**(**is.na**(ubernewsample))

## [1] 0

*#Summary to check*  
**summary**(ubernewsample)

##     taxi\_id      trip\_seconds      trip\_miles          fare        
##  Min.   :   4   Min.   :  60.0   Min.   : 0.100   Min.   : 4.00    
##  1st Qu.:2458   1st Qu.: 360.0   1st Qu.: 0.800   1st Qu.: 6.00    
##  Median :4528   Median : 510.0   Median : 1.550   Median : 8.00    
##  Mean   :4477   Mean   : 778.5   Mean   : 3.465   Mean   :12.42    
##  3rd Qu.:6405   3rd Qu.: 900.0   3rd Qu.: 3.300   3rd Qu.:12.44    
##  Max.   :8696   Max.   :4920.0   Max.   :18.600   Max.   :52.75    
##       tips            tolls       extras        trip\_total      
##  Min.   : 0.000   Min.   :0   Min.   :0.000   Min.   : 4.250    
##  1st Qu.: 0.000   1st Qu.:0   1st Qu.:0.000   1st Qu.: 6.812    
##  Median : 2.000   Median :0   Median :0.000   Median : 9.500    
##  Mean   : 1.953   Mean   :0   Mean   :0.609   Mean   :14.980    
##  3rd Qu.: 2.150   3rd Qu.:0   3rd Qu.:1.000   3rd Qu.:15.275    
##  Max.   :11.550   Max.   :0   Max.   :5.000   Max.   :69.300    
##       payment\_type  
##  Cash       :33     
##  Credit Card:44     
##  Other      : 1     
##                     
##                     
##

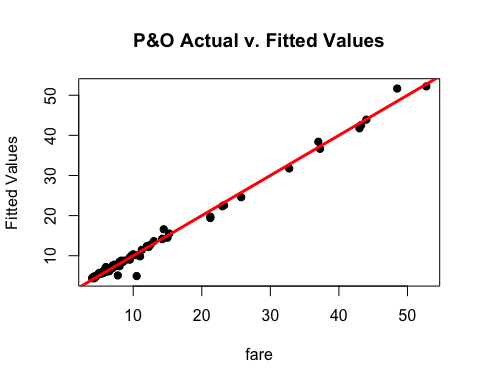
*#Since tolls are 0, I am not considering the column in  analysis*  
ubernewsample = **select**(ubernewsample, **-c**(tolls,taxi\_id))  
*#replaceing NA with 0, if any*  
ubernewsample[**is.na**(ubernewsample)] = 0  
*#converting trip seconds to trip minutes*  
ubernewsample**$**trip\_seconds= (ubernewsample**$**trip\_seconds**/**60)  
**colnames**(ubernewsample)[**colnames**(ubernewsample)**==**"trip\_seconds"] <- "trip\_minutes"  
**attach**(ubernewsample)

*#Fare as a function of  minutes and miles*  
sample\_with\_new\_data\_lm = **lm**(ubernewsample**$**fare**~**  
                      ubernewsample**$**trip\_minutes**+**  
                      ubernewsample**$**trip\_miles)  
**summary**(sample\_with\_new\_data\_lm)

##   
## Call:  
## lm(formula = ubernewsample$fare ~ ubernewsample$trip\_minutes +   
##     ubernewsample$trip\_miles)  
##   
## Residuals:  
##     Min      1Q  Median      3Q     Max   
## -3.1717 -0.3851 -0.1618  0.1532  5.5678   
##   
## Coefficients:  
##                            Estimate Std. Error t value Pr(>|t|)      
## (Intercept)                 3.47842    0.16026   21.70   <2e-16 \*\*\*  
## ubernewsample$trip\_minutes  0.15682    0.01466   10.70   <2e-16 \*\*\*  
## ubernewsample$trip\_miles    1.99241    0.04127   48.27   <2e-16 \*\*\*  
## ---  
## Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.012 on 75 degrees of freedom  
## Multiple R-squared:  0.9922, Adjusted R-squared:  0.992   
## F-statistic:  4777 on 2 and 75 DF,  p-value: < 2.2e-16

|  |
| --- |
| Building a model based on the new sample, having minutes and miles, in fact has the higher accuracy: 99.2%.  The equation is a follows:  F(fare) = 3.47 + (0.15)Minutes + (1.99) Miles  This means that, increase in every minute will increase the fare by $0.15 and similarly every mile will add $1.99 to the fare. |

**plot**(ubernewsample**$**fare,sample\_with\_new\_data\_lm**$**fitted.values,pch=19,main="P&O Actual v. Fitted Values",  
    xlab = "fare", ylab = "Fitted Values")  
**abline**(0,1,col="red",lwd=3)



|  |
| --- |
| Plotting a actual vs predicted values, we can see that almost all the points lie on the regression line. Thus, our model can predict the fare based on the Minutes and Miles of the ride. |