Regression Project QMB-6304

Darya Gahramanova

Preprocessing

```
library(rio)
setwd("~/USF/Fall 2019/QMB-6304")
taxi <- read.csv(file="6304 Regression Project Data.csv", header=TRUE, sep=",
")
colnames(taxi)=tolower(make.names(colnames(taxi)))
set.seed(68884865)
my.sample = taxi[sample(1:nrow(taxi),100,replace=FALSE),]
summary(my.sample)
```

```
##     taxi_id      trip_seconds    trip_miles         fare
##  Min.   :  12   Min.   :   0   Min.   : 0.000   Min.   : 3.250
##  1st Qu.:1894   1st Qu.: 300   1st Qu.: 0.075   1st Qu.: 5.938
##  Median :3891   Median : 450   Median : 0.900   Median : 7.750
##  Mean   :4104   Mean   : 645   Mean   : 2.300   Mean   :11.935
##  3rd Qu.:6114   3rd Qu.: 780   3rd Qu.: 1.925   3rd Qu.:10.750
##  Max.   :8646   Max.   :3300   Max.   :32.300   Max.   :80.000
##       tips             tolls       extras         trip_total
##  Min.   : 0.000   Min.   :0   Min.   :0.0000   Min.   : 3.250
##  1st Qu.: 0.000   1st Qu.:0   1st Qu.:0.0000   1st Qu.: 6.725
##  Median : 0.000   Median :0   Median :0.0000   Median : 8.625
##  Mean   : 1.257   Mean   :0   Mean   :0.7525   Mean   :13.945
##  3rd Qu.: 2.000   3rd Qu.:0   3rd Qu.:1.0000   3rd Qu.:12.412
##  Max.   :10.050   Max.   :0   Max.   :7.0000   Max.   :87.000
##       payment_type
##  Cash       :61
##  Credit Card:39
##  Other      : 0
##
##
##
```

"my.sample" object contains 100 random observations of the whole population. Summary output shows that some "trip_seconds" and "trip_miles" variables equal zero, which doesn't make sense in this case. These observations should be excluded from the analysis to get more accurate predictions. Tolls column doesn't have any values other than 0 and will not be considered for the analysis.

First, let's create a "my.taxi" object that will only contain those variables in "my.sample" where "trip_miles" is greater than 0.

```
my.taxi = subset(my.sample,trip_miles>0)
summary(my.taxi)
```

```
##     taxi_id      trip_seconds      trip_miles          fare
##  Min.   :  12   Min.   : 180.0   Min.   : 0.100   Min.   : 4.05
##  1st Qu.:1706   1st Qu.: 360.0   1st Qu.: 0.800   1st Qu.: 6.25
##  Median :3878   Median : 540.0   Median : 1.500   Median : 8.25
##  Mean   :4082   Mean   : 750.4   Mean   : 3.067   Mean   :12.52
##  3rd Qu.:6133   3rd Qu.: 900.0   3rd Qu.: 2.200   3rd Qu.:10.88
##  Max.   :8646   Max.   :3300.0   Max.   :32.300   Max.   :80.00
##       tips             tolls       extras         trip_total
##  Min.   : 0.000   Min.   :0   Min.   :0.0000   Min.   : 4.05
##  1st Qu.: 0.000   1st Qu.:0   1st Qu.:0.0000   1st Qu.: 7.15
##  Median : 0.000   Median :0   Median :0.0000   Median :10.00
##  Mean   : 1.317   Mean   :0   Mean   :0.7833   Mean   :14.62
##  3rd Qu.: 2.000   3rd Qu.:0   3rd Qu.:1.0000   3rd Qu.:12.90
##  Max.   :10.050   Max.   :0   Max.   :7.0000   Max.   :87.00
##       payment_type
##  Cash       :42
##  Credit Card:33
##  Other      : 0
##
##
##
```

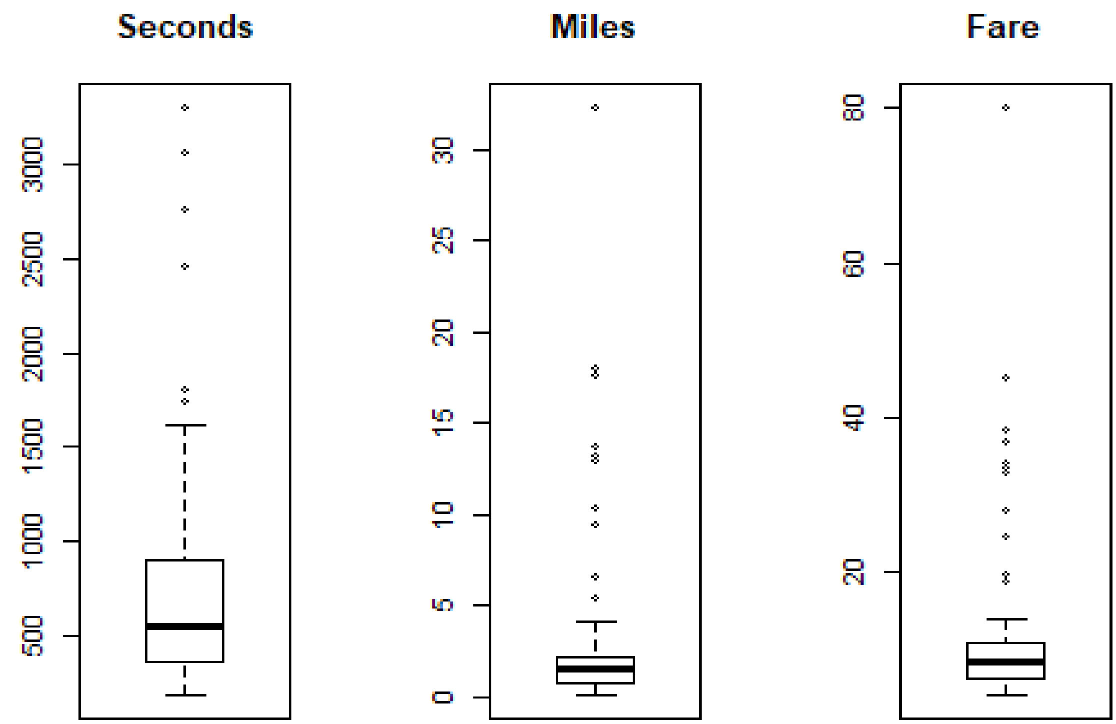
"my.taxi" sample now containts 75 observations. New summary output shows that null "trip_seconds" have been eliminated as well. Looking at the mean, quantile, and max values of the variables it can be inferred that there are outliers with very high values. Let's focus on fare, seconds, and miles. We will look at their skewness and kurtosis levels.

```r
library(moments)
skew=data.frame(Kurtosis=c(kurtosis(my.taxi$fare),kurtosis(my.taxi$trip_secon
ds),kurtosis(my.taxi$trip_miles)), Skewness=c(skewness(my.taxi$fare),skewness
(my.taxi$trip_seconds),skewness(my.taxi$trip_miles)), row.names=c("Fare","Sec
onds","Miles") )
skew
```

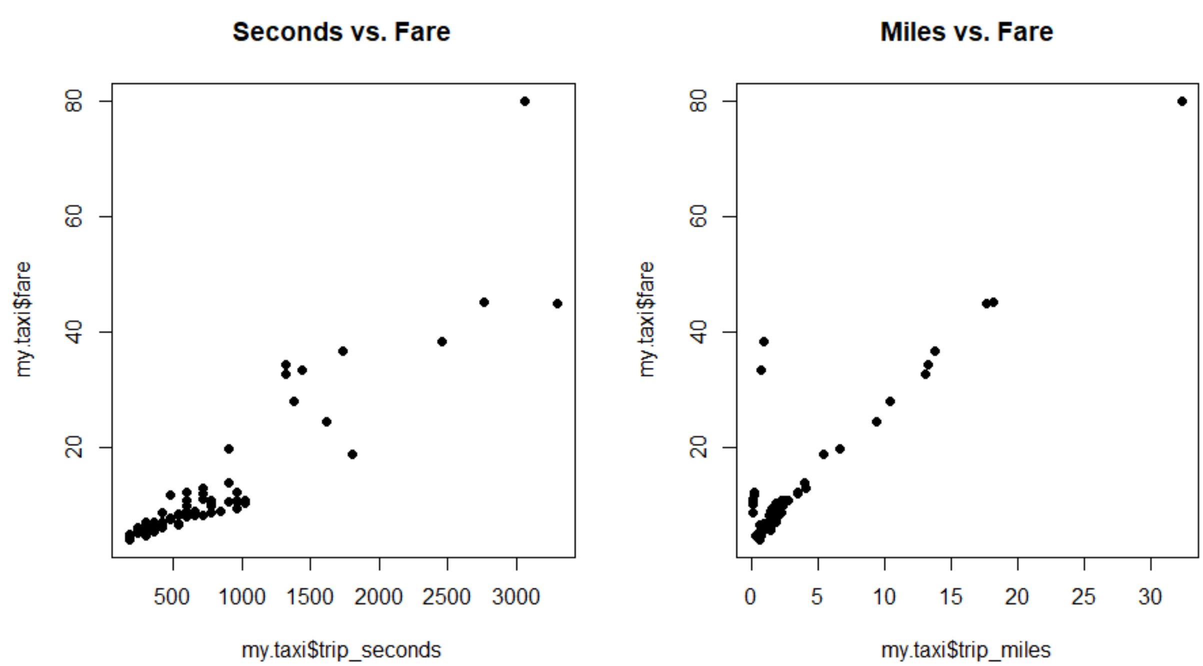
```
##          Kurtosis Skewness
## Fare    14.182610 3.073608
## Seconds  8.316967 2.251955
## Miles   16.964188 3.491394
```

The output shows that these variables are highly skewed left, and the kurtosis is high, meaning that the majority of data points correspond to lower values of fare, seconds, and miles. The following plots confirm this pattern.

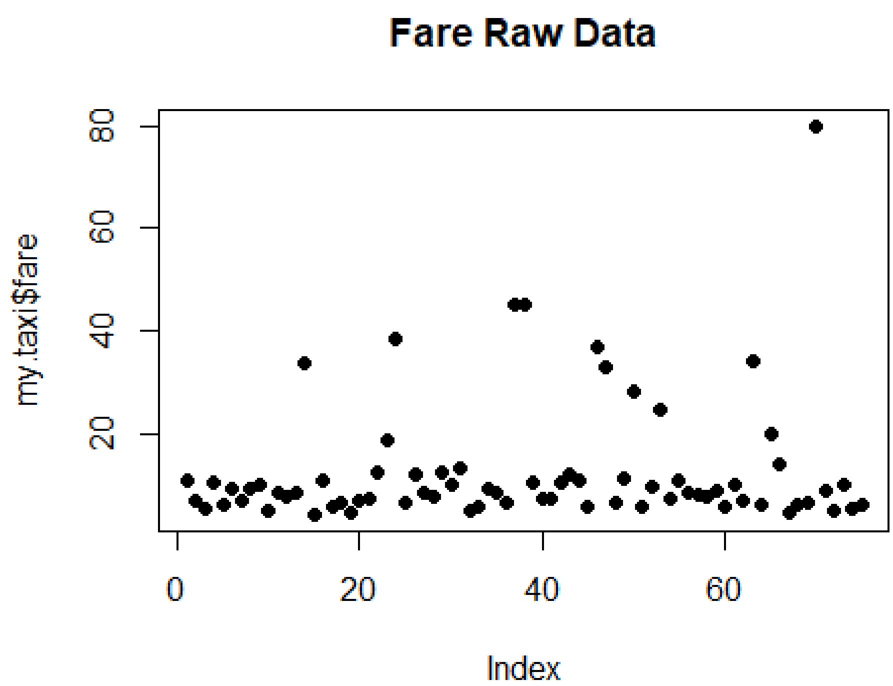
```r
par(mfrow=c(1,3))
boxplot(my.taxi$trip_seconds, main="Seconds")
boxplot(my.taxi$trip_miles, main="Miles")
boxplot(my.taxi$fare, main="Fare")
```



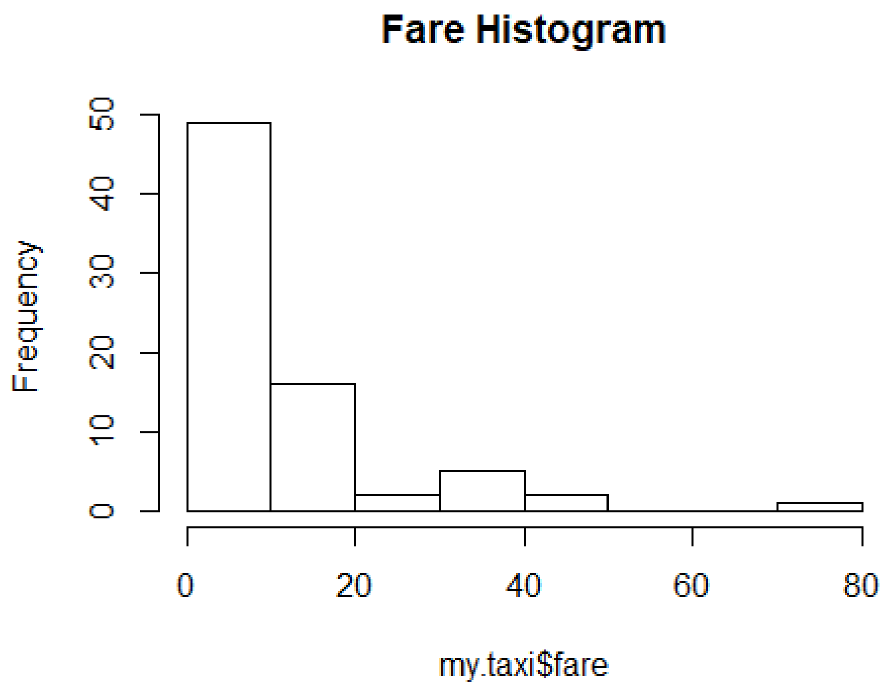

```r
par(mfrow=c(1,2))
plot(my.taxi$trip_seconds,my.taxi$fare,pch=19, main="Seconds vs. Fare")
plot(my.taxi$trip_miles,my.taxi$fare,pch=19, main="Miles vs. Fare")
```



```r
par(mfrow=c(1,1))
plot(my.taxi$fare,pch=19, main="Fare Raw Data")
```




```r
hist(my.taxi$fare, main="Fare Histogram")
```



These plots show that although there might be a linear pattern between fare, seconds, and miles, we have some data points that are much higher the average of the rest of the data. This might mislead our judgement regarding the existence of linear relationships between the majority of data points. Moreover, a couple of observations correspond to very low traveled miles, but high fares.

Let's look at these outliers.

```
my.taxi[which(my.taxi$fare>3*mean(my.taxi$fare)),]

##         taxi_id trip_seconds trip_miles  fare  tips tolls extras
## 213582     5394         2460        0.9 38.25  0.00     0      1
## 49675      7458         3300       17.6 45.00 10.00     0      5
## 443883      759         2760       18.1 45.25 10.05     0      5
## 1555145    4572         3060       32.3 80.00  0.00     0      7
##         trip_total payment_type
## 213582       39.25         Cash
## 49675        60.00  Credit Card
## 443883       60.30  Credit Card
## 1555145      87.00         Cash

my.taxi[which(my.taxi$fare>30 & my.taxi$trip_miles<1),]

##        taxi_id trip_seconds trip_miles  fare tips tolls extras trip_total
## 104845     817         1440        0.7 33.50    5     0      4      42.50
## 213582    5394         2460        0.9 38.25    0     0      1      39.25
##        payment_type
```

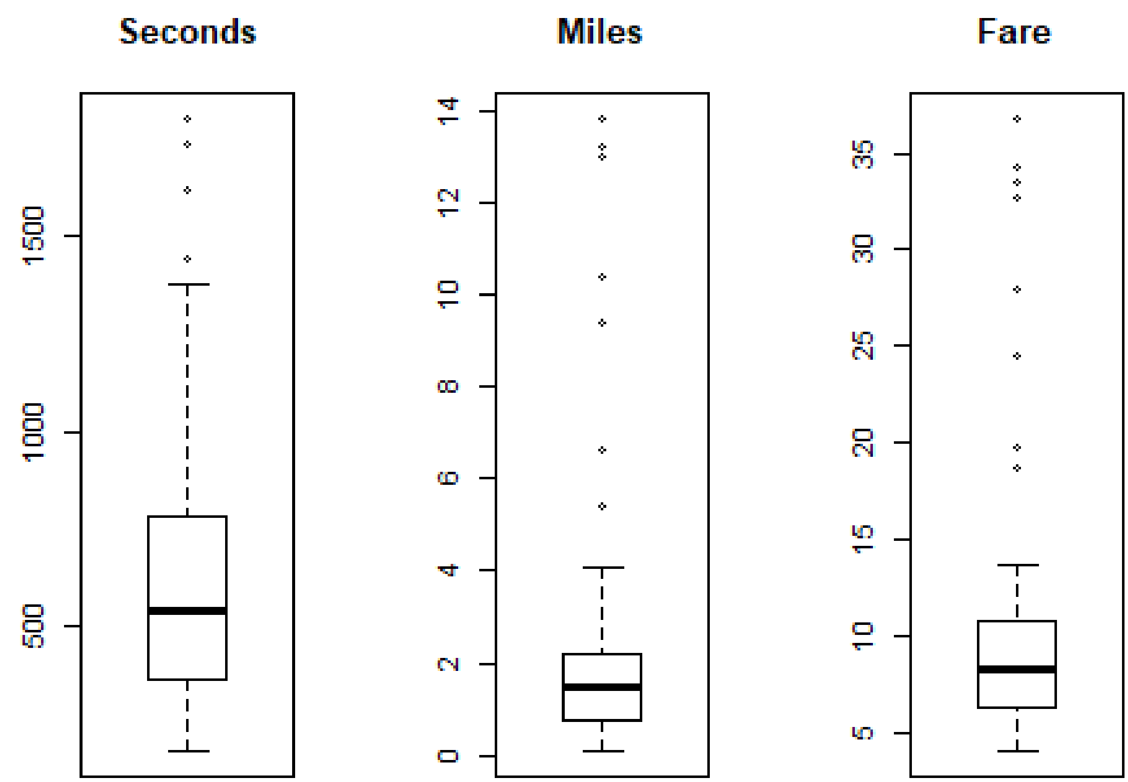
```
## 104845  Credit Card
## 213582         Cash
```

First table represents instances with extremely high fare rates (more than three times the mean of the sample). With the exception of taxi #5394, where the fare is $38.25 for only 0.9 miles traveled, these extreme cases don't appear to be faulty - fares are high for high time and distance traveled. However, these 4 outliers will strongly affect our regression model and drive the trend line towards them. We will get rid of these points in a new "mytaxi" object. Second table demonstrates those weird cases when a fare is very high for small distance traveled. However, time traveled for those points is reasonably high. Assuming that these might be the cases when traffic was terrible, we will not get rid of these cases for the analysis. An interaction between seconds and miles traveled might thus be an important predictor in the regression model.

```
mytaxi = my.taxi[my.taxi$fare<3*mean(my.taxi$fare),]
summary(mytaxi)
```

```
##     taxi_id      trip_seconds      trip_miles          fare
##  Min.   :  12   Min.   : 180.0   Min.   : 0.100   Min.   : 4.05
##  1st Qu.:1706   1st Qu.: 360.0   1st Qu.: 0.750   1st Qu.: 6.25
##  Median :3803   Median : 540.0   Median : 1.500   Median : 8.25
##  Mean   :4056   Mean   : 629.6   Mean   : 2.269   Mean   :10.29
##  3rd Qu.:6133   3rd Qu.: 780.0   3rd Qu.: 2.200   3rd Qu.:10.75
##  Max.   :8646   Max.   :1800.0   Max.   :13.800   Max.   :36.75
##       tips            tolls       extras          trip_total
##  Min.   :0.000   Min.   :0   Min.   :0.0000   Min.   : 4.050
##  1st Qu.:0.000   1st Qu.:0   1st Qu.:0.0000   1st Qu.: 7.025
##  Median :0.000   Median :0   Median :0.0000   Median : 9.000
##  Mean   :1.108   Mean   :0   Mean   :0.5739   Mean   :11.973
##  3rd Qu.:2.000   3rd Qu.:0   3rd Qu.:1.0000   3rd Qu.:12.125
##  Max.   :7.650   Max.   :0   Max.   :7.0000   Max.   :45.900
##       payment_type
##  Cash       :40
##  Credit Card:31
##  Other      : 0
##
##
##
```

```
par(mfrow=c(1,3))
boxplot(mytaxi$trip_seconds, main="Seconds")
boxplot(mytaxi$trip_miles, main="Miles")
boxplot(mytaxi$fare, main="Fare")
```




```
par(mfrow=c(1,1))
```

“mytaxi” sample now has 71 observations. Looking at the new boxplots, we can tell that we got rid of the rides with fares higher than $40. There are still many outliers, but this is due to the nature of this data, we don’t have to eliminate all of them at this point.

Analysis

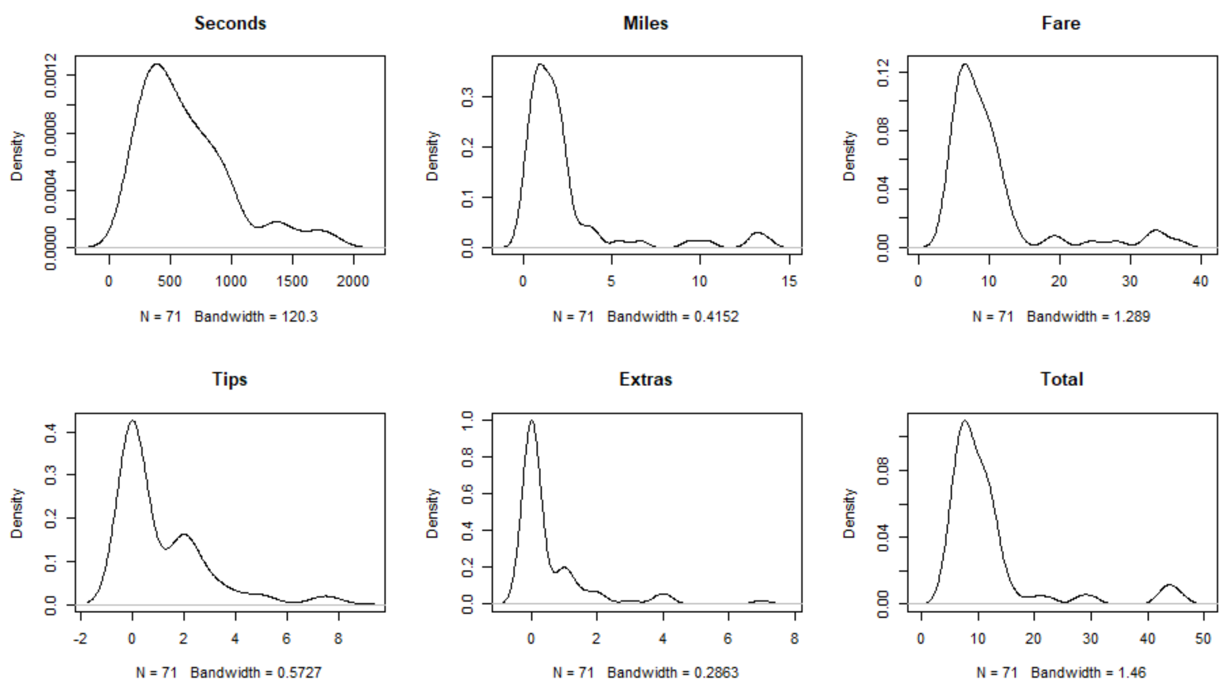
1

Here is an overview of distribution of the continious variables of the cleansed sample.

```
cont.taxi=subset(mytaxi,select=c("trip_seconds","trip_miles","fare","tips","e
xtras","trip_total"))
summary(cont.taxi)

##   trip_seconds      trip_miles         fare            tips
##  Min.   : 180.0   Min.   : 0.100   Min.   : 4.05   Min.   :0.000
##  1st Qu.: 360.0   1st Qu.: 0.750   1st Qu.: 6.25   1st Qu.:0.000
##  Median : 540.0   Median : 1.500   Median : 8.25   Median :0.000
##  Mean   : 629.6   Mean   : 2.269   Mean   :10.29   Mean   :1.108
##  3rd Qu.: 780.0   3rd Qu.: 2.200   3rd Qu.:10.75   3rd Qu.:2.000
##  Max.   :1800.0   Max.   :13.800   Max.   :36.75   Max.   :7.650
##      extras         trip_total
##  Min.   :0.0000   Min.   : 4.050
```

```
##   1st Qu.:0.0000   1st Qu.: 7.025
##   Median :0.0000   Median : 9.000
##   Mean   :0.5739   Mean   :11.973
##   3rd Qu.:1.0000   3rd Qu.:12.125
##   Max.   :7.0000   Max.   :45.900
```

```r
par(mfrow=c(2,3))
plot(density(mytaxi$trip_seconds), main="Seconds")
plot(density(mytaxi$trip_miles), main="Miles")
plot(density(mytaxi$fare), main="Fare")
plot(density(mytaxi$tips), main="Tips")
plot(density(mytaxi$extras), main="Extras")
plot(density(mytaxi$trip_total), main="Total")
```




```r
par(mfrow=c(1,1))
```

Cleansed data is still highly skewed left even after removing the main outliers, as fares for the majority of trips are falling in the $0-$20 range. Trip seconds is the most normally distributed variable among the continious ones. Trip miles and fare rates seem to have more variability and outliers. Tips and extras vary even more significantly. Medians of these variables = 0, meaning that most of the trips did not have any tips/extras. Randomness of these variables proves that they are not good predictors of the price for the trip. Trip total is simply calculated by adding tips and extras to the fare, so it is not considered for the analysis either.

Let's now look at the contents of the "payment_type" factor variable.

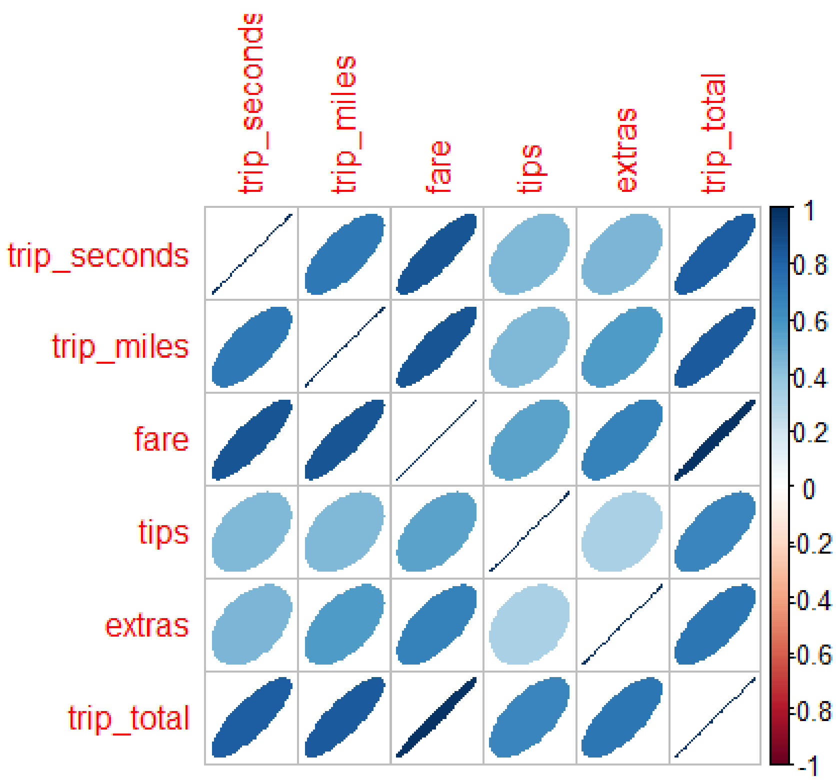
```
pmt = as.data.frame(table(mytaxi$payment_type))
names(pmt) = c("Payment Type","Number of Cases")
pmt

##   Payment Type Number of Cases
## 1         Cash              40
## 2  Credit Card              31
## 3        Other               0
```

In our sample there are only 2 types of payment: cash and credit card, with number of cases 40 and 31 respectively.

3

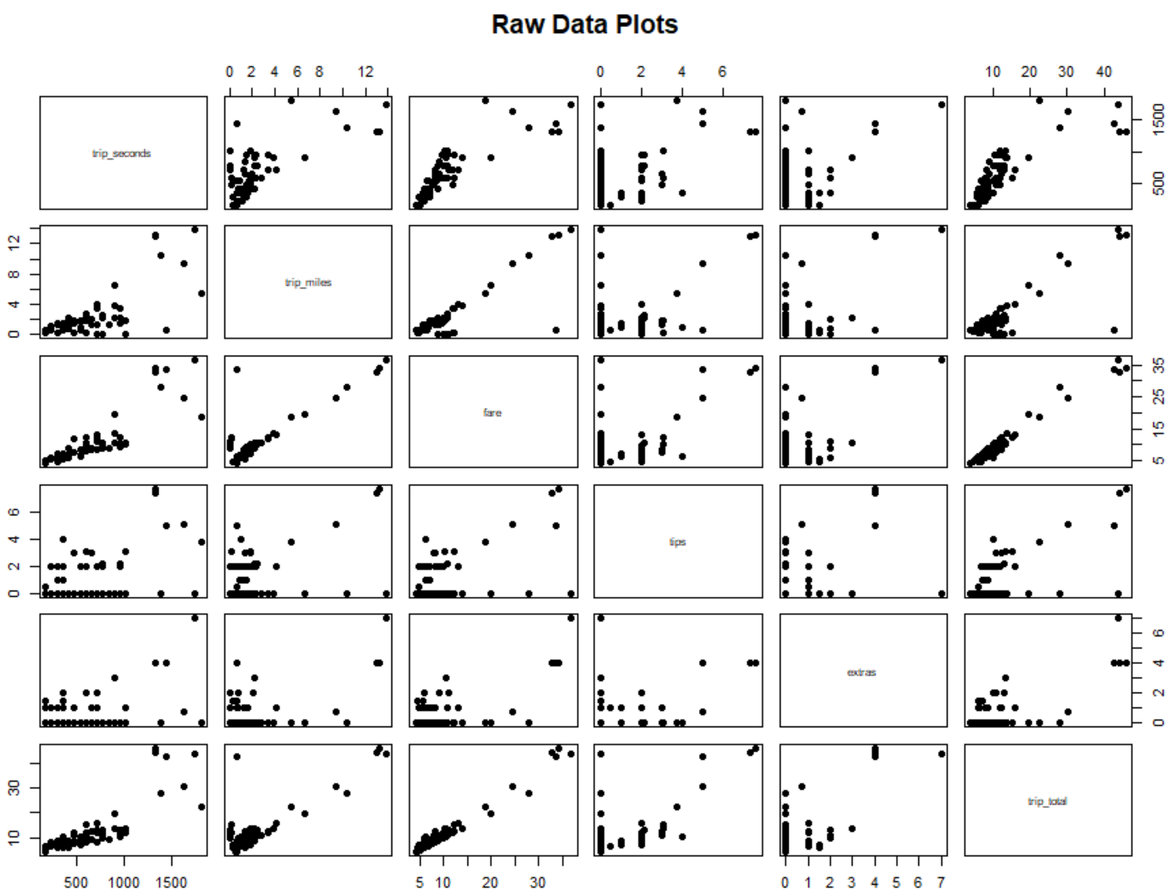
Now let's look at the correlations between the continious variables.

```
library(corrplot)

corrplot(cor(cont.taxi),method="ellipse")
library(Hmisc)
```




```
rcorr(as.matrix(cont.taxi))

##              trip_seconds trip_miles fare tips extras trip_total
## trip_seconds         1.00       0.71 0.85 0.45   0.45       0.82
## trip_miles           0.71       1.00 0.86 0.45   0.56       0.84
## fare                 0.85       0.86 1.00 0.53   0.67       0.98
```

```
## tips                    0.45        0.45 0.53 1.00    0.33        0.65
## extras                  0.45        0.56 0.67 0.33    1.00        0.73
## trip_total              0.82        0.84 0.98 0.65    0.73        1.00
##
## n= 71
##
##
## P
##              trip_seconds trip_miles fare   tips   extras trip_total
## trip_seconds              0.0000     0.0000 0.0000 0.0000 0.0000
## trip_miles   0.0000                  0.0000 0.0000 0.0000 0.0000
## fare         0.0000       0.0000            0.0000 0.0000 0.0000
## tips         0.0000       0.0000     0.0000        0.0053 0.0000
## extras       0.0000       0.0000     0.0000 0.0053        0.0000
## trip_total   0.0000       0.0000     0.0000 0.0000 0.0000

plot(cont.taxi,pch=19, main="Raw Data Plots")
```



Both trip seconds and miles separately indicate correlation coefficients of over 0.85 with fares, supported by low p-values. This means that some linear relationship exists between each of these two independent variables and our target. Unfortunately, there is also a correlation of 0.71 between seconds and miles themselves, which might inflate variance in our model. Plots also demonstrate linear relationships between these variables. Tips and extras, on the other hand, don't indicate much of a linear relationship with fares. As mentioned earlier, these variables appear to be pretty random and should not be considered for model. Trip total is calculated using fares, so they have very high correlation, and there is no need to include both in the model.

4

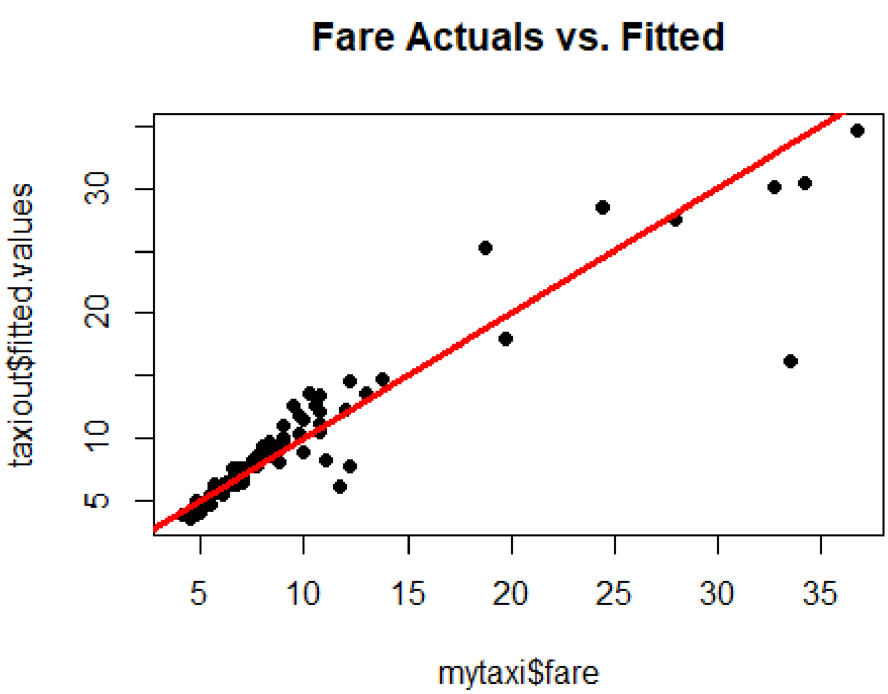
Let's run a multiple linear regression on trip miles, seconds, payment type as independent variables, and fares as a dependent variable. We will also estimate confidence intervals for each of the coefficients and put them all in a table.

```
taxiout = lm(fare~trip_seconds+trip_miles+payment_type, data=mytaxi)
summary(taxiout)

##
## Call:
## lm(formula = fare ~ trip_seconds + trip_miles + payment_type,
##     data = mytaxi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5145 -0.9351 -0.1832  0.4948 17.3851
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.463881   0.705912   2.074    0.042 *
## trip_seconds            0.009310   0.001262   7.376 3.21e-10 ***
## trip_miles              1.233608   0.162279   7.602 1.26e-10 ***
## payment_typeCredit Card 0.380759   0.685456   0.555    0.580
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.814 on 67 degrees of freedom
## Multiple R-squared:  0.8562, Adjusted R-squared:  0.8498
## F-statistic:   133 on 3 and 67 DF,  p-value: < 2.2e-16

coefs=cbind("Beta Coefficients"=coef(taxiout),confint(taxiout))
coefs

##                         Beta Coefficients        2.5 %     97.5 %
## (Intercept)                   1.463880791  0.054873960 2.87288762
## trip_seconds                  0.009310218  0.006790704 0.01182973
## trip_miles                    1.233607754  0.909698454 1.55751705
## payment_typeCredit Card       0.380758761 -0.987417692 1.74893521
```

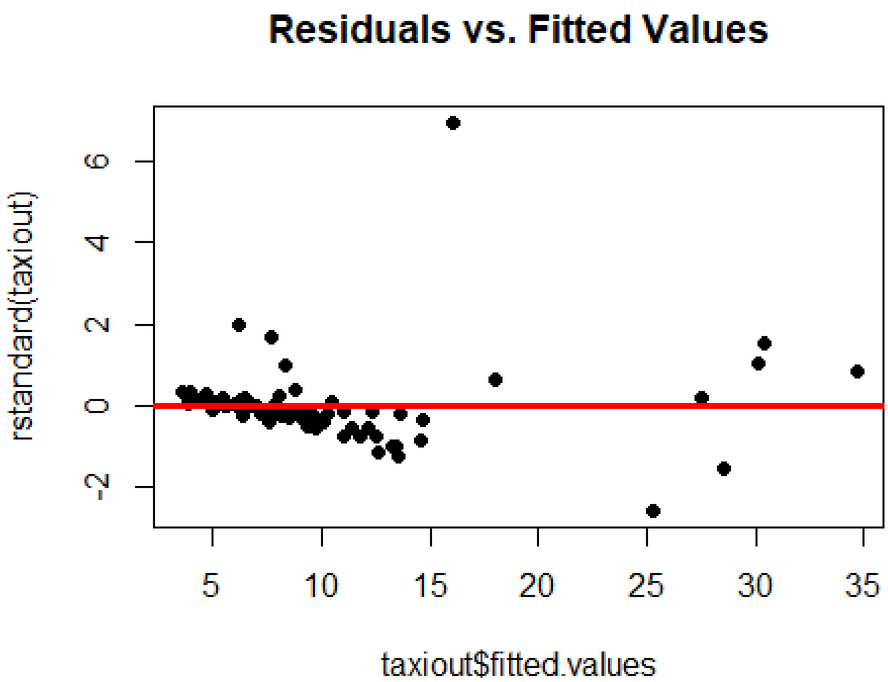
First of, the intercept value corresponds to the dollar amount that would be charged even if trip time and distance were equal zero. Although this would not occur in real life, this indicates that there is probably some minimum starting amount embedded in trip fares. P-value is less than 5%, and the range of intercept falls between \$0.05 and \$2.87, meaning that this minimal fare value would be positive at 95% confidence level. The model is highly confident that trip time has a significant effect on fare rate. Every minute of the trip increases fare by approximately $0.56. This value ranges from $0.41 to $0.71 at 95% confidence level. Although this range is somewhat wide, it doesn't cross zero, so the relationship definitely exists. Similarly, distance has significant effect on fare. Every mile driven increases fare by $1.23 - but can be anywhere between $0.91 and $1.56. Again, the range is pretty wide, but the relationship definitely exists. Wide ranges are expected due to the small sample size and some extreme variables. The model found no relationship between payment type and fare. P-value is high and the confidence interval crosses 0. This variable has no benefit in predicting ride prices. Adjusted R-squared indicates that trip seconds, miles, and payment type together help to explain 85% of fare. This is a considerably high value, but the model can still be improved by dumping the insignificant variable, testing variable transformations, and re-running without points with high leverage.

We can look at the quality of fit of this model by comparing predictions to actual fare values using correlation and residual plots.

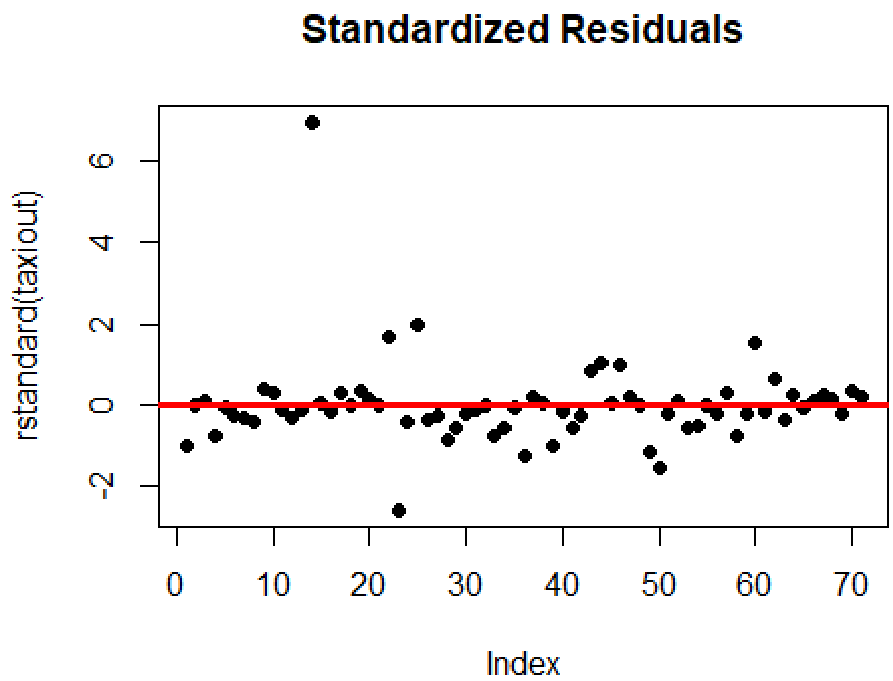
```
cor(mytaxi$fare,taxiout$fitted.values)

## [1] 0.9253202

plot(mytaxi$fare,taxiout$fitted.values,pch=19,main="Fare Actuals vs. Fitted")
abline(0,1,col="red",lwd=3)
```



```
plot(taxiout$fitted.values,rstandard(taxiout),pch=19,main="Residuals vs. Fitt
ed Values")
abline(0,0,col="red",lwd=3)
```




```
plot(rstandard(taxiout),pch=19,main="Standardized Residuals")
abline(0,0,col="red",lwd=3)
```



Again, at first glance, correlation and fit plots indicate a promising linear relationship and good fit. But we can see the outliers that might strongly affect the model, and we should test the interactions between variables.

First, we should see whether our variables are dependent on each other. Higher trip distance usually corresponds to higher trip time, and this relationship could inflate variance.

```
library(car)

vif(taxiout)

## trip_seconds   trip_miles payment_type
##     2.061171     2.043375     1.036166
```

These vif values indicate that the relationships between the independent variables are not that strong, so we can proceed with a model that includes both trip seconds and miles.

5

Let's explore different combinations of the same continious variables. First, we will throw in everything, including squared terms and interaction.

```
taxiout2 = lm(fare~trip_seconds+I(trip_seconds^2)+trip_miles+I(trip_miles^2)+
trip_miles:trip_seconds+payment_type, data=mytaxi)
summary(taxiout2)
```

```
##
## Call:
## lm(formula = fare ~ trip_seconds + I(trip_seconds^2) + trip_miles +
##     I(trip_miles^2) + trip_miles:trip_seconds + payment_type,
##     data = mytaxi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4919 -0.5245 -0.1175  0.2841  8.8370
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.989e+00  9.042e-01   4.412 4.01e-05 ***
## trip_seconds           -1.423e-04  3.034e-03  -0.047 0.962738
## I(trip_seconds^2)       1.115e-05  2.356e-06   4.733 1.26e-05 ***
## trip_miles              1.885e+00  5.117e-01   3.684 0.000475 ***
## I(trip_miles^2)         2.854e-01  3.612e-02   7.901 4.82e-11 ***
## payment_typeCredit Card -1.847e-01  5.056e-01  -0.365 0.716133
## trip_seconds:trip_miles -3.492e-03  5.779e-04  -6.041 8.61e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.037 on 64 degrees of freedom
## Multiple R-squared:  0.9281, Adjusted R-squared:  0.9213
## F-statistic: 137.6 on 6 and 64 DF,  p-value: < 2.2e-16
```

```r
cor(mytaxi$fare,taxiout2$fitted.values)
```

```
## [1] 0.9633652
```

These variable transformations bumped R-squared up to 0.92. Squared seconds and miles are indicated as having a significant effect. Interaction term between them is also significant, as discussed earlier. Correlation of the model fit with raw data increased from 0.92 to 0.96. However, seconds and miles were significant even without squared terms. What if we remove them?

```r
taxiout2.1 = lm(fare~trip_seconds+I(trip_seconds^2)+trip_miles+trip_miles:tri
p_seconds+payment_type, data=mytaxi)
summary(taxiout2.1)$r.squared
```

```
## [1] 0.8579127
```

```r
taxiout2.2 = lm(fare~trip_seconds+I(trip_miles^2)+trip_miles+trip_miles:trip_
seconds+payment_type, data=mytaxi)
summary(taxiout2.2)$r.squared
```

```
## [1] 0.9028954
```

```r
taxiout2.3 = lm(fare~trip_seconds+trip_miles+trip_miles:trip_seconds+payment_
type, data=mytaxi)
summary(taxiout2.3)$r.squared
```

```
## [1] 0.8567058
```

R-squared values decreased again. It's better to keep squared values of seconds and miles in the model. We will remove payment type and re-run regression:

```
taxiout3 = lm(fare~trip_seconds+I(trip_seconds^2)+trip_miles+I(trip_miles^2)+
trip_miles:trip_seconds, data = mytaxi)
summary(taxiout3)
```

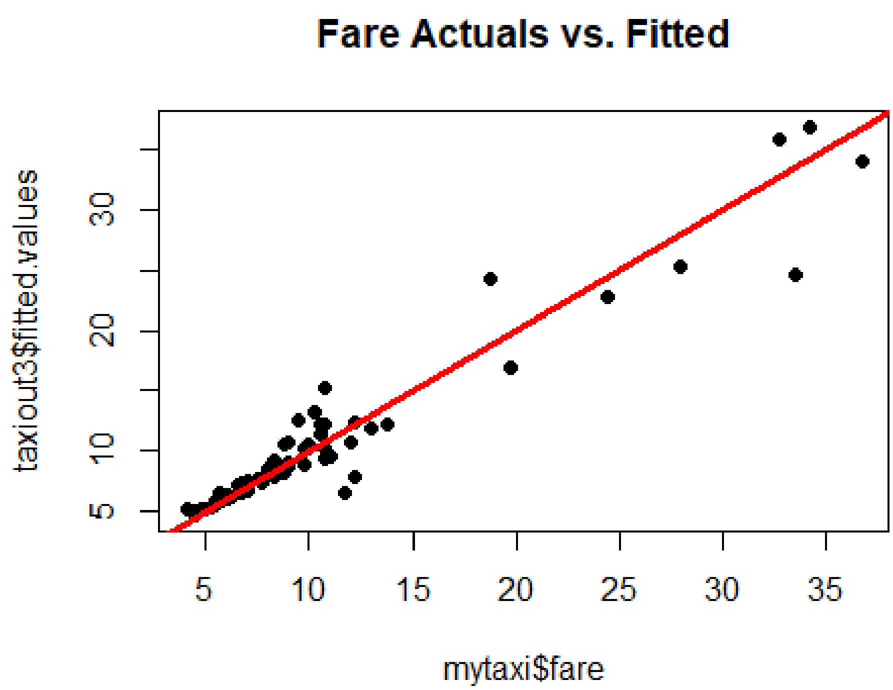
```
##
## Call:
## lm(formula = fare ~ trip_seconds + I(trip_seconds^2) + trip_miles +
##     I(trip_miles^2) + trip_miles:trip_seconds, data = mytaxi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4965 -0.4795 -0.0406  0.3322  8.8441
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.897e+00  8.629e-01   4.517 2.71e-05 ***
## trip_seconds            1.188e-05  2.984e-03   0.004 0.996836
## I(trip_seconds^2)       1.099e-05  2.297e-06   4.783 1.03e-05 ***
## trip_miles              1.862e+00  5.044e-01   3.691 0.000459 ***
## I(trip_miles^2)         2.838e-01  3.561e-02   7.969 3.32e-11 ***
## trip_seconds:trip_miles -3.457e-03  5.663e-04  -6.105 6.39e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.023 on 65 degrees of freedom
## Multiple R-squared:  0.9279, Adjusted R-squared:  0.9224
## F-statistic: 167.4 on 5 and 65 DF,  p-value: < 2.2e-16
```

```
cor(mytaxi$fare,taxiout3$fitted.values)
```

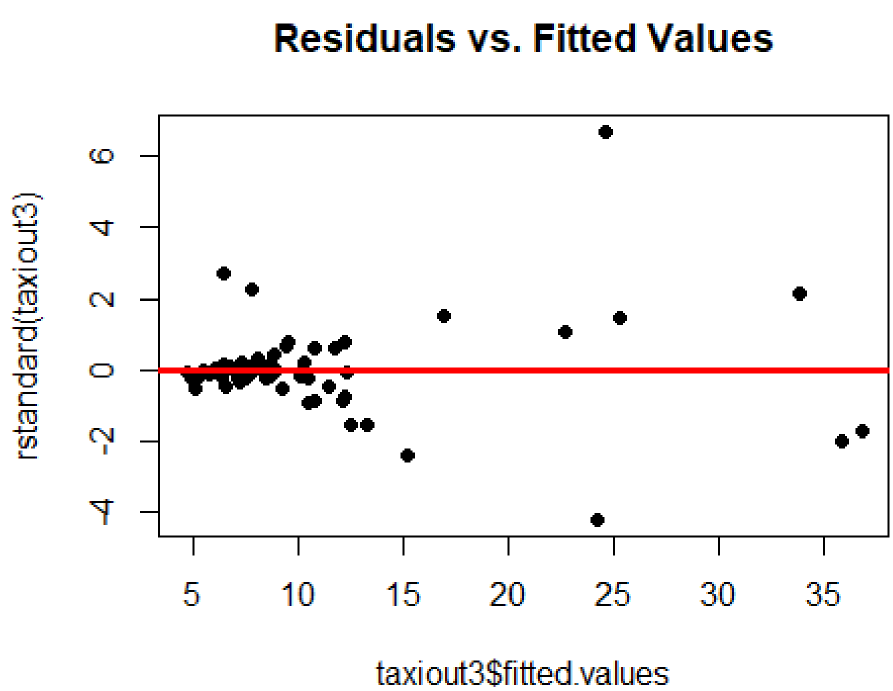
```
## [1] 0.9632874
```

R-squared increased a little more. Let's look at the fit:

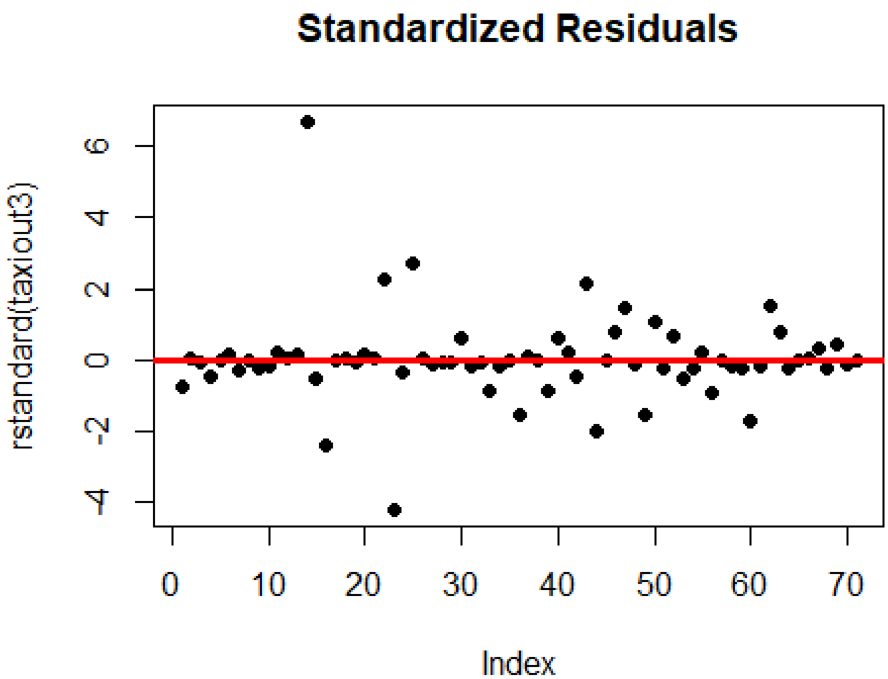
```
plot(mytaxi$fare,taxiout3$fitted.values,pch=19,main="Fare Actuals vs. Fitted"
)
abline(0,1,col="red",lwd=3)
```




```r
plot(taxiout3$fitted.values,rstandard(taxiout3),pch=19,main="Residuals vs. Fi
tted Values")
abline(0,0,col="red",lwd=3)
```



```r
plot(rstandard(taxiout3),pch=19,main="Standardized Residuals")
abline(0,0,col="red",lwd=3)
```



Now fares with lower values fit better along the line. Let's examine the outlier - should we get rid of it?

```r
mytaxi[which(rstandard(taxiout3)>5),]

##         taxi_id trip_seconds trip_miles fare tips tolls extras trip_total
## 104845     817         1440        0.7 33.5    5     0      4       42.5
##         payment_type
## 104845  Credit Card
```

This point corresponds to an under 1 mile trip with a fare of $33.5. Although trip time is high, something weird happened there. Either the car was very slow, or the record is wrong. In any case, this outlier negatively affects the model fit - let's get rid of it and re-run the model.

```r
mytaxi2 = mytaxi[-which(rstandard(taxiout3)>5),]
taxiout4 = lm(fare~trip_seconds+I(trip_seconds^2)+trip_miles+I(trip_miles^2)+
trip_miles:trip_seconds, data = mytaxi2)
summary(taxiout4)

##
## Call:
## lm(formula = fare ~ trip_seconds + I(trip_seconds^2) + trip_miles +
##     I(trip_miles^2) + trip_miles:trip_seconds, data = mytaxi2)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4622 -0.6479 -0.1958  0.2152  4.8174
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.640e+00  4.966e-01   5.317 1.43e-06 ***
## trip_seconds             9.661e-03  1.864e-03   5.183 2.38e-06 ***
## I(trip_seconds^2)       -1.940e-06  1.688e-06  -1.149   0.2547
## trip_miles               5.838e-01  3.033e-01   1.925   0.0587 .
## I(trip_miles^2)          1.045e-01  2.507e-02   4.168 9.41e-05 ***
## trip_seconds:trip_miles -1.983e-04  4.200e-04  -0.472   0.6385
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.138 on 64 degrees of freedom
## Multiple R-squared:  0.9737, Adjusted R-squared:  0.9716
## F-statistic: 473.1 on 5 and 64 DF,  p-value: < 2.2e-16
```

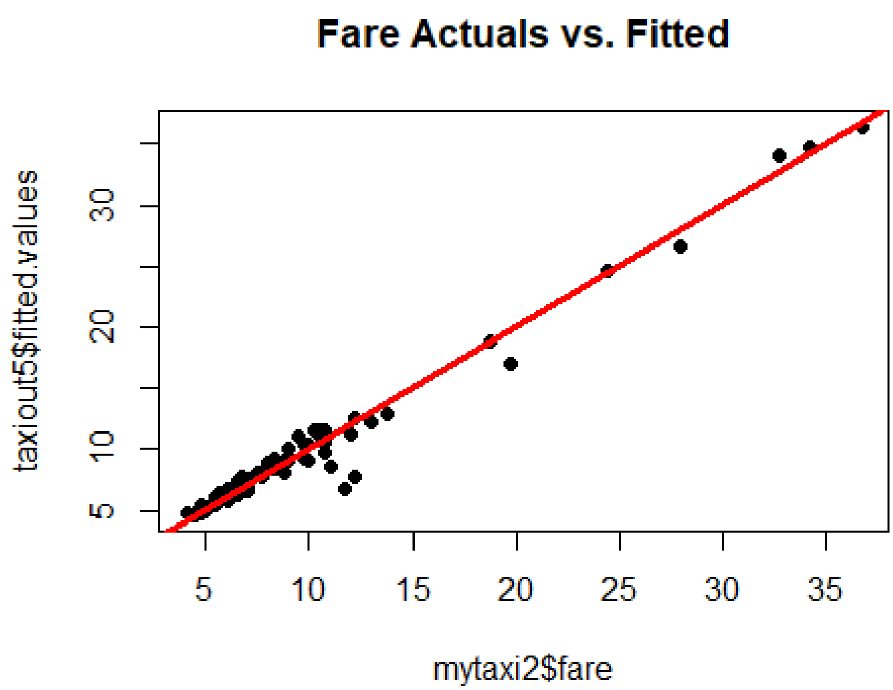
Now the interaction term and squared seconds became insignificant. Let's leave the interaction term in, as it helps with weird cases, but get rid of the squared seconds:

```
taxiout5 = lm(fare~trip_seconds+trip_miles+I(trip_miles^2)+trip_miles:trip_se
conds, data = mytaxi2)
summary(taxiout5)
```

```
##
## Call:
## lm(formula = fare ~ trip_seconds + trip_miles + I(trip_miles^2) +
##     trip_miles:trip_seconds, data = mytaxi2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4971 -0.5834 -0.2160  0.0981  4.9673
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.9754408  0.4030219   7.383 3.64e-10 ***
## trip_seconds             0.0077030  0.0007581  10.161 4.70e-15 ***
## trip_miles               0.8230986  0.2210983   3.723 0.000414 ***
## I(trip_miles^2)          0.1239538  0.0185060   6.698 5.91e-09 ***
## trip_seconds:trip_miles -0.0006228  0.0002005  -3.106 0.002809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.14 on 65 degrees of freedom
## Multiple R-squared:  0.9731, Adjusted R-squared:  0.9715
## F-statistic: 588.2 on 4 and 65 DF,  p-value: < 2.2e-16
```

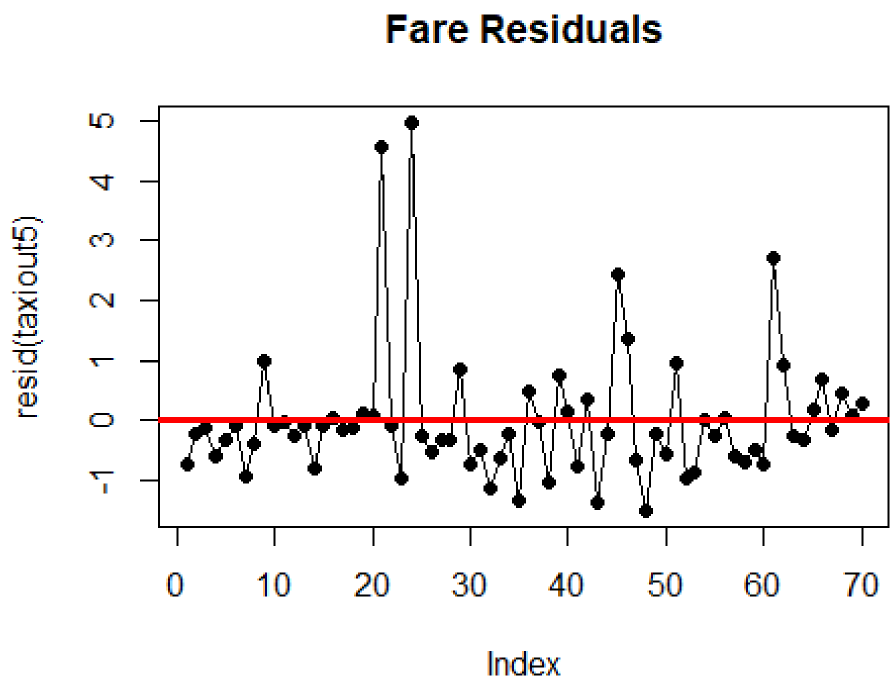
R-squared is now 0.97. This will be the final model: *fare ~ trip_seconds + trip_miles + trip_miles^2 + trip_miles:trip_seconds*. It provides the best fit without unnecessary variables, and the interaction term is kept in case if another sample from this population has weird outliers. Now let's test conformity of this model with the LINE assumptions.

```
plot(mytaxi2$fare,taxiout5$fitted.values, pch=19,main="Fare Actuals vs. Fitte
d")
abline(0,1,col="red",lwd=3)
```



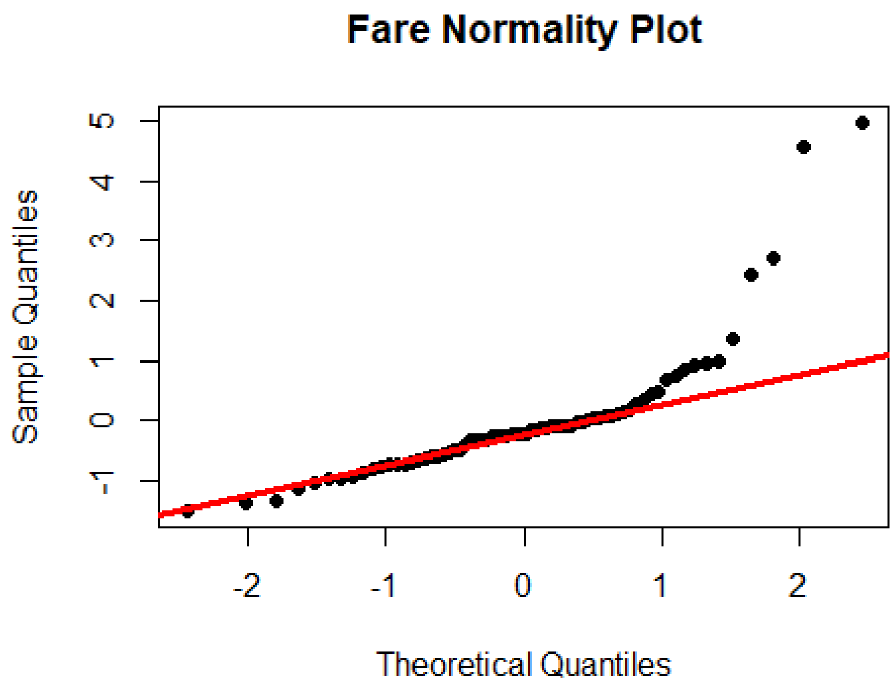
With a couple of exceptions, both lower and higher values of fares fit along the trend line. This model conforms to the linearity assumption.

```
plot(resid(taxiout5),pch=19,main="Fare Residuals",type="o")
abline(0,0,col="red",lwd=3)
```



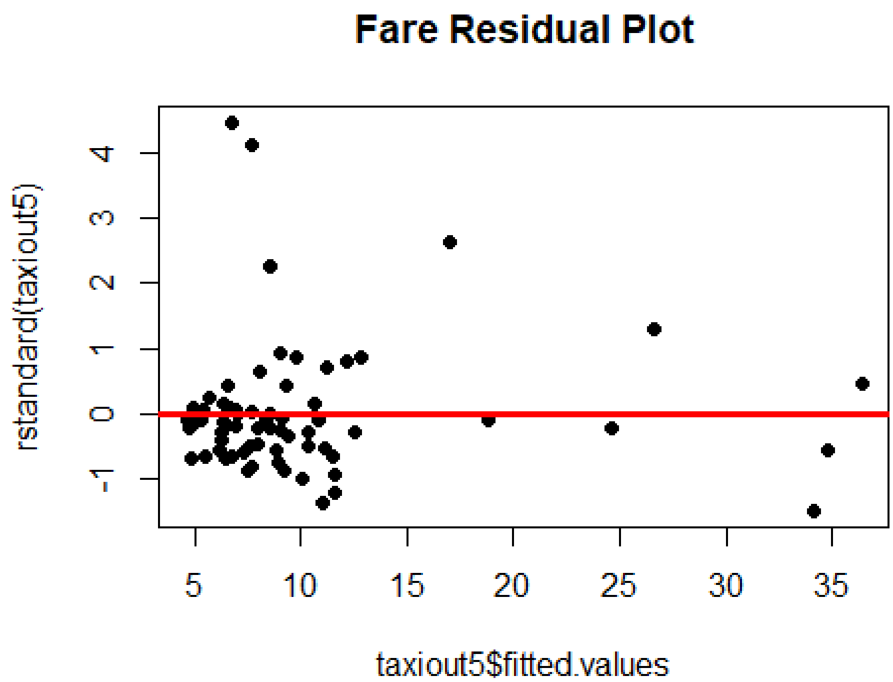
Independence assumption is not violated either; we did not expect any cyclical relationships in this case anyway.

```
qqnorm(taxiout5$residuals,pch=19,main="Fare Normality Plot")
qqline(taxiout5$residuals,lwd=3,col="red")
```



Normality assumption is somewhat violated. Although majority of points fits along the normality line, the residuals of several odd cases don't seem to fit normal distribution.

```
plot(taxiout5$fitted.values,rstandard(taxiout5),pch=19,main="Fare Residual Pl
ot")
abline(0,0,col="red",lwd=3)
```

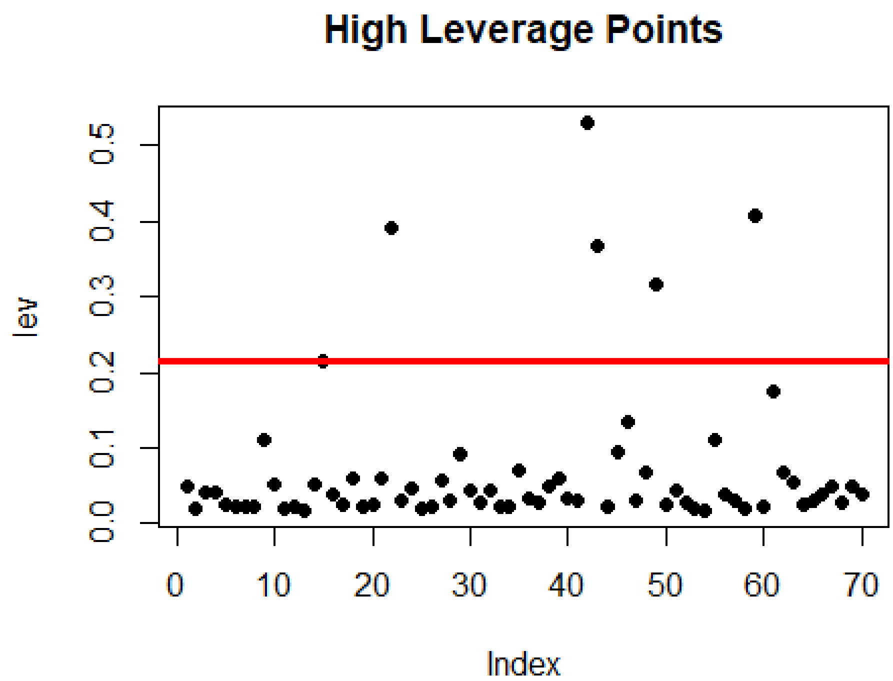


As expected and inferred from previous visualizations of data, equality of variances assumption is violated. Most of the fare values are less than $20 and have small residuals, but the higher the fare, the more spread out the residuals are, the harder it is to make predictions. Some odd points might have high fares but small distances traveled. This could not be avoided in this type of data, unless focusing on smaller fare rides only, but in real life these extreme cases may take place.

7

Let's see our high leverage points – instances that strongly affect our model.

```
lev=hat(model.matrix(taxiout5))
plot(lev,pch=19,main="High Leverage Points")
abline(3*mean(lev),0,col="red",lwd=3)
```




```
mytaxi2[lev>(3*mean(lev)),]

##         taxi_id trip_seconds trip_miles  fare tips tolls extras trip_total
## 1326778    1419         1800        5.4 18.75 3.75     0   0.00      22.50
## 1411841    6301         1740       13.8 36.75 0.00     0   7.00      43.75
## 765588     1581         1320       13.0 32.75 7.35     0   4.00      44.10
## 1195854     175         1620        9.4 24.45 5.04     0   0.75      30.24
## 1328101     819         1320       13.2 34.25 7.65     0   4.00      45.90
##         payment_type
## 1326778  Credit Card
## 1411841         Cash
## 765588   Credit Card
## 1195854  Credit Card
## 1328101  Credit Card

outliers = which(lev>(3*mean(lev)))
```

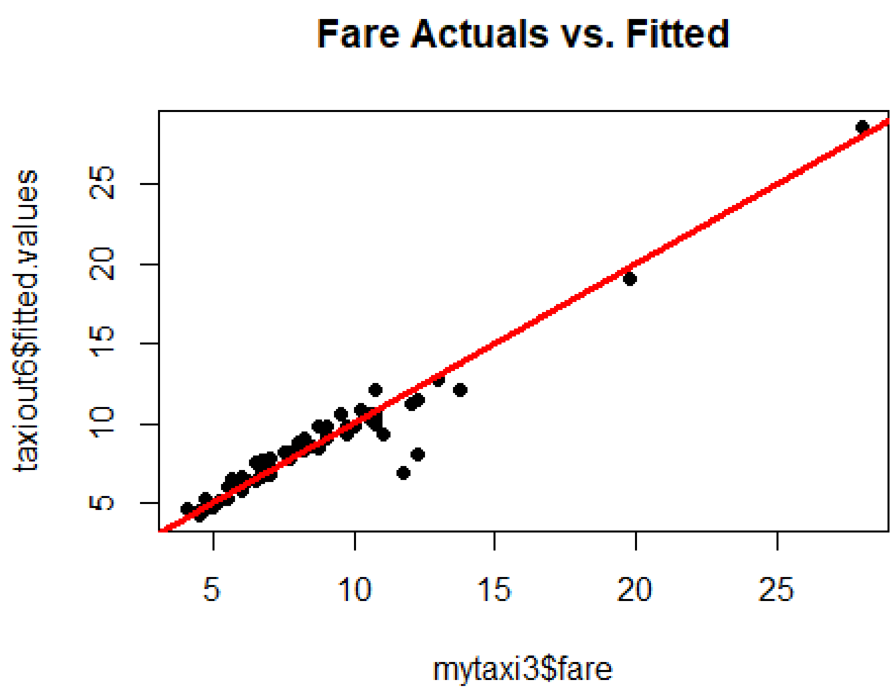
These correspond to the cases that have very high fare values but do not necesseraly have very high residuals. No extremely weird scenarios here: high time and distance values, and hence high fares. Let's try to run the model without those points and see how the model fits the majority of data with lower fares.

```
mytaxi3 = mytaxi2[-outliers,]
taxiout6 = lm(fare~trip_seconds*trip_miles+I(trip_miles^2)+trip_miles:trip_se
conds, data = mytaxi3)
summary(taxiout6)
```

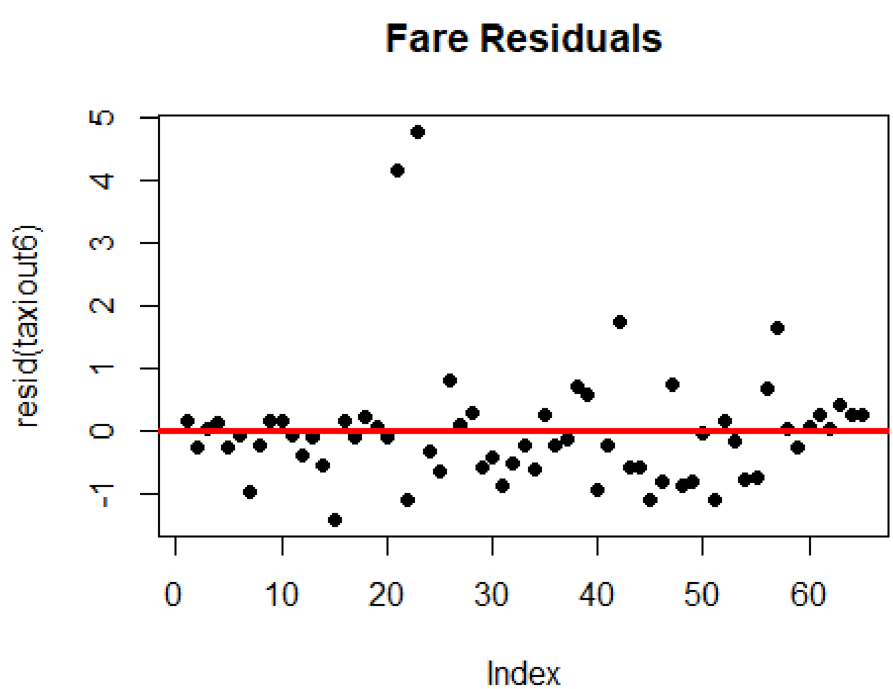
```
##
## Call:
## lm(formula = fare ~ trip_seconds * trip_miles + I(trip_miles^2) +
##     trip_miles:trip_seconds, data = mytaxi3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4189 -0.5885 -0.0903  0.1651  4.7908
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.1612465  0.4794549   4.508 3.10e-05 ***
## trip_seconds            0.0099497  0.0009435  10.546 2.75e-15 ***
## trip_miles              1.3611743  0.3270861   4.162 0.000102 ***
## I(trip_miles^2)         0.3516687  0.0601717   5.844 2.23e-07 ***
## trip_seconds:trip_miles -0.0027516  0.0006582  -4.181 9.59e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.032 on 60 degrees of freedom
## Multiple R-squared:  0.9277, Adjusted R-squared:  0.9229
## F-statistic: 192.5 on 4 and 60 DF,  p-value: < 2.2e-16
```

R-squared decreased to 0.92, as the regression line was previously trying to fit those high leverage points. Let's see the residual plots again.

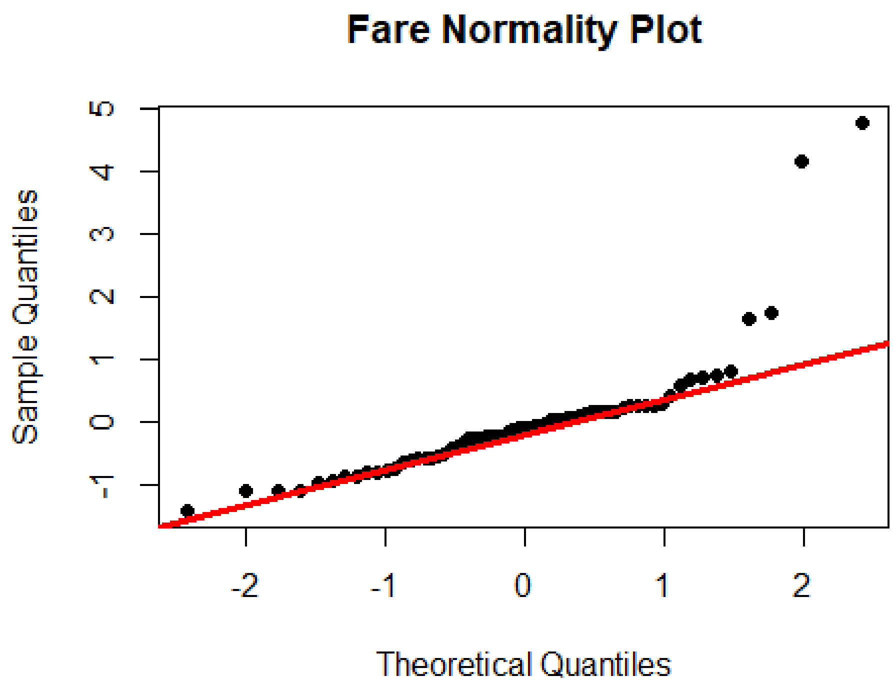
```
# Linearity
plot(mytaxi3$fare,taxiout6$fitted.values, pch=19,main="Fare Actuals vs. Fitte
d")
abline(0,1,col="red",lwd=3)
```



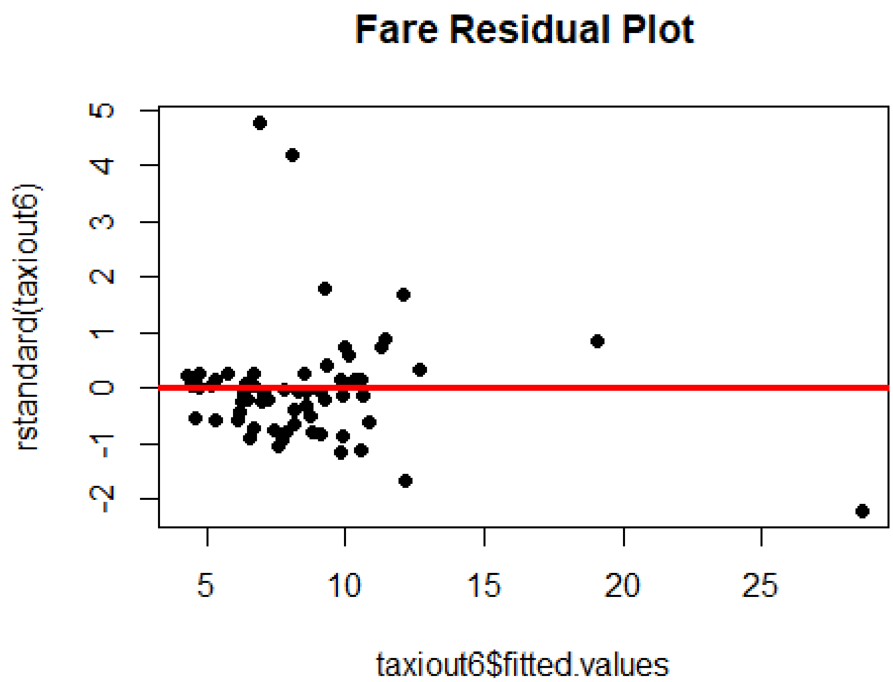

```r
# Independence
plot(resid(taxiout6),pch=19,main="Fare Residuals")
abline(0,0,col="red",lwd=3)
```



```r
# Normality
qqnorm(taxiout6$residuals,pch=19,main="Fare Normality Plot")
qqline(taxiout6$residuals,lwd=3,col="red")
```




```r
# Equality of variances
plot(taxiout6$fitted.values,rstandard(taxiout6),pch=19,main="Fare Residual Pl
ot")
abline(0,0,col="red",lwd=3)
```



Although we didn't get of the couple of weird residuals higher than 4, removing high leverage points helped to get a better fit for the smaller data points. Normality plot also looks a little better.

8

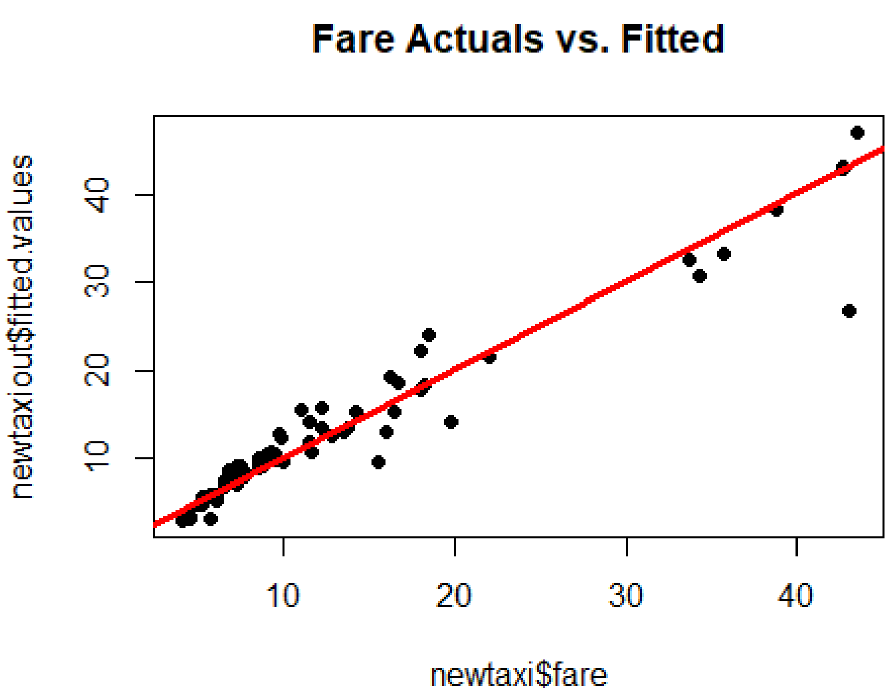
We will now get another sample of 100 observations and test the model on it. Data will be cleansed using the same technique: first, we'll get rid of the trip miles equal to zero. Second, we will remove extreme data points that are more than 3 times higher than the mean. The model used is *fare ~ trip_seconds + trip_miles + trip_miles^2 + trip_miles:trip_seconds:*

```
set.seed(68884870)
new.sample = taxi[sample(1:nrow(taxi),100,replace=FALSE),]
newtaxi = subset(new.sample,trip_miles>0)
newtaxi = newtaxi[newtaxi$fare<3*mean(newtaxi$fare),]

newtaxiout = lm(fare ~ trip_seconds + trip_miles + I(trip_miles^2) + trip_sec
onds:trip_miles, data = newtaxi)
summary(newtaxiout)

##
## Call:
## lm(formula = fare ~ trip_seconds + trip_miles + I(trip_miles^2) +
##     trip_seconds:trip_miles, data = newtaxi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

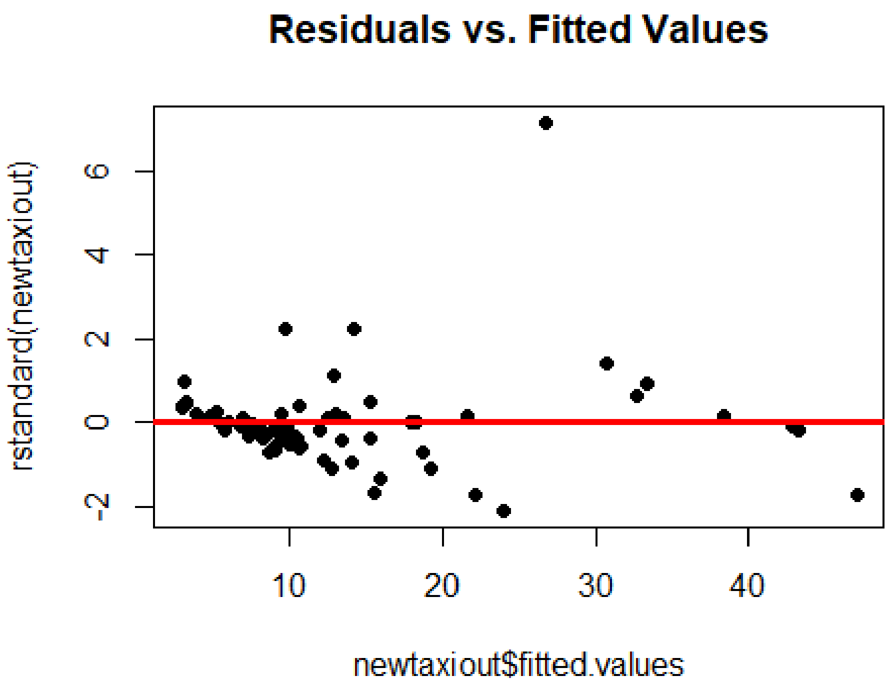
```
## -5.4895 -1.0016 -0.1982  0.8911 16.2376
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.1679534  0.7077641   1.650  0.10338
## trip_seconds             0.0130416  0.0011025  11.829  < 2e-16 ***
## trip_miles               0.9621691  0.2836709   3.392  0.00115 **
## I(trip_miles^2)          0.1126427  0.0184906   6.092 5.39e-08 ***
## trip_seconds:trip_miles -0.0010222  0.0001612  -6.340 1.95e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.769 on 70 degrees of freedom
## Multiple R-squared:  0.9315, Adjusted R-squared:  0.9276
## F-statistic: 237.9 on 4 and 70 DF,  p-value: < 2.2e-16
```

We are getting an R-squared 0.93, and all of the variables are estimated to be significant. This proves that there was a little bit of overfitting due to the high leverage points. Still, the model seems to fit well on a new sample.

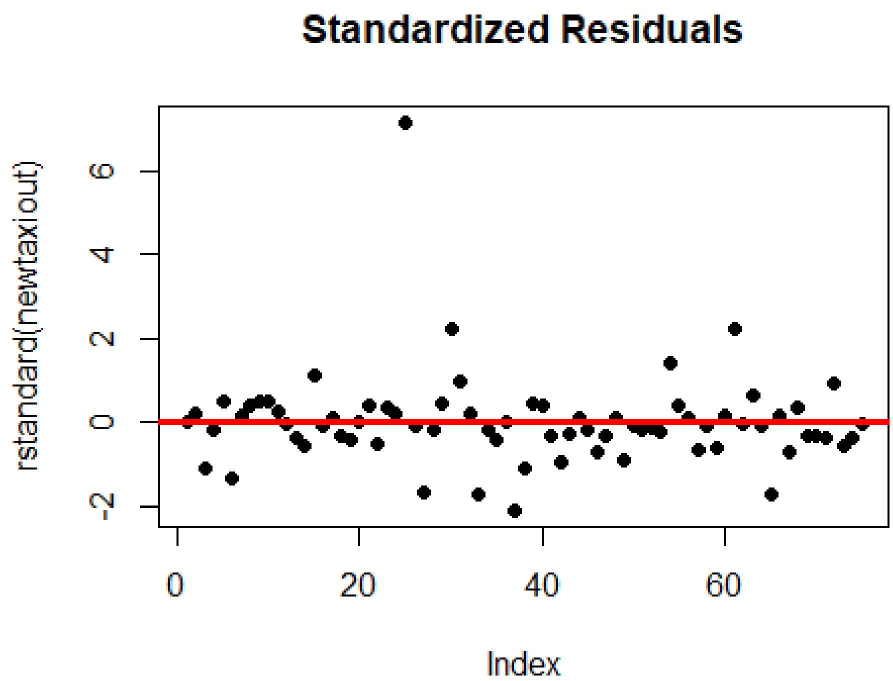
```
cor(newtaxi$fare,newtaxiout$fitted.values)

## [1] 0.9651302

plot(newtaxi$fare,newtaxiout$fitted.values,pch=19,main="Fare Actuals vs. Fitt
ed")
abline(0,1,col="red",lwd=3)
```



```
plot(newtaxiout$fitted.values,rstandard(newtaxiout),pch=19,main="Residuals vs
. Fitted Values")
abline(0,0,col="red",lwd=3)
```




```
plot(rstandard(newtaxiout),pch=19,main="Standardized Residuals")
abline(0,0,col="red",lwd=3)
```



A similar trend with residuals is observed with this sample. Although the model fits majority of data poins, some odd cases continue to take place. This proves that the outliers are quite common in this dataset and we should not get rid of all of them.

Last, let's look at the main outlier in the new sample.

```
newtaxi[which(rstandard(newtaxiout)>4),]
```

```
##         taxi_id trip_seconds trip_miles fare tips tolls extras trip_total
## 1374363    7140         2040          1   43  9.7     0    5.5       58.2
##         payment_type
## 1374363  Credit Card
```

Again, high fares with very low miles driven indicate that there is either an error in the dataset or the trip was extremely slow. If the point is to predict any possible trip from this data, we cannot get rid of all of the outliers. If the goal is to predict trips that are, say, cheaper than $20, a stricter cleansing should be applied to this dataset, and the result will be a better fit with smaller residuals.