Deliverable 3

Numeric and Binary targets Forecasting Models

Júlia Gasull i Claudia Sánchez

December 13, 2020

Contents

1 First setups

```
if(!is.null(dev.list())) dev.off()  # Clear plots
rm(list=ls())                       # Clean workspace
```

1.1 Load Required Packages for this deliverable

We load the necessary packages and set working directory

```
setwd("~/Github Repositories/FIB-ADEI-LAB/deliverable3")
filepath<-"~/Github Repositories/FIB-ADEI-LAB/deliverable3"
#setwd("C:/Users/Claudia Sánchez/Desktop/FIB/TARDOR 2020-2021/ADEI/DELIVERABLE1/FIB-ADEI-LAB/deliverable
#filepath<-"C:/Users/Claudia Sánchez/Desktop/FIB/TARDOR 2020-2021/ADEI/DELIVERABLE1/FIB-ADEI-LAB/deliver
```

```r
# Load Required Packages
options(contrasts=c("contr.treatment","contr.treatment"))
requiredPackages <- c("missMDA","chemometrics","mvoutlier","effects","FactoMineR","car","lmtest","ggplot

missingPackages <- requiredPackages[!(requiredPackages %in% installed.packages()[,"Package"])]
if(length(missingPackages)) install.packages(missingPackages)
lapply(requiredPackages, require, character.only = TRUE)
```

1.2 Load processed data from last deliverable

```r
load(paste0(filepath,"/Taxi5000_del2.RData"))
summary(df)
```

```
##               VendorID         RateCodeID   Pickup_longitude Pickup_latitude
##  f.Vendor-Mobile  : 973   Rate-1    :4496   Min.   :-74.02   Min.   :40.58
##  f.Vendor-VeriFone:3650   Rate-Other: 127   1st Qu.:-73.96   1st Qu.:40.70
##                                             Median :-73.94   Median :40.75
##                                             Mean   :-73.93   Mean   :40.75
##                                             3rd Qu.:-73.92   3rd Qu.:40.80
##                                             Max.   :-73.80   Max.   :40.86
##  Dropoff_longitude Dropoff_latitude Passenger_count Trip_distance
##  Min.   :-74.02    Min.   :40.58    Min.   :1.000   Min.   : 0.010
##  1st Qu.:-73.96    1st Qu.:40.70    1st Qu.:1.000   1st Qu.: 1.010
##  Median :-73.94    Median :40.75    Median :1.000   Median : 1.760
##  Mean   :-73.93    Mean   :40.75    Mean   :1.371   Mean   : 2.724
##  3rd Qu.:-73.91    3rd Qu.:40.79    3rd Qu.:1.000   3rd Qu.: 3.400
##  Max.   :-73.80    Max.   :40.86    Max.   :6.000   Max.   :30.000
##   Fare_amount        Extra        MTA_tax      Tip_amount      Tolls_amount
##  Min.   : 1.00   Min.   :0.0000   No : 119   Min.   : 0.000   Min.   :0.0000
##  1st Qu.: 6.00   1st Qu.:0.0000   Yes:4504   1st Qu.: 0.000   1st Qu.:0.0000
##  Median : 9.00   Median :0.5000              Median : 0.000   Median :0.0000
##  Mean   :11.61   Mean   :0.3523              Mean   : 1.022   Mean   :0.0477
##  3rd Qu.:14.50   3rd Qu.:0.5000              3rd Qu.: 1.700   3rd Qu.:0.0000
##  Max.   :60.00   Max.   :1.0000              Max.   :17.000   Max.   :5.5400
##  improvement_surcharge  Total_amount         Payment_type        Trip_type
##  No : 118              Min.   :  0.00   Credit card:2096   Street-Hail:4511
##  Yes:4505              1st Qu.:  7.80   Cash       :2497   Dispatch    : 112
##                        Median : 10.80   No paid    :  30
##                        Mean   : 13.93
##                        3rd Qu.: 17.00
##                        Max.   :128.76
##       hour                  period         tlenkm         traveltime
##  Min.   : 0.0   Period night    :1642   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 9.0   Period morning  : 542   1st Qu.: 1.609   1st Qu.: 5.767
##  Median :15.0   Period valley   :1260   Median : 2.800   Median : 9.550
##  Mean   :13.4   Period afternoon:1179   Mean   : 4.349   Mean   :12.487
##  3rd Qu.:19.0                          3rd Qu.: 5.472   3rd Qu.:16.125
##  Max.   :23.0                          Max.   :48.280   Max.   :60.000
##      espeed         pickup            dropoff           Trip_distance_range
##  Min.   : 3.00   Length:4623        Length:4623        Long_dist  : 665
##  1st Qu.:14.83   Class :character   Class :character   Medium_dist: 986
##  Median :18.56   Mode  :character   Mode  :character   Short_dist :2972
##  Mean   :20.34
##  3rd Qu.:23.58
##  Max.   :55.00
##  paidTolls   TipIsGiven passenger_groups   hcpck        claKM
##  No  :4576   No :2882   Couple: 343      kHP-1:1930   kKM-3: 844
##  Yes :  40   Yes:1741   Group : 395      kHP-2:1634   kKM-5:2353
##  NA's:   7              Single:3885      kHP-3: 262   kKM-2: 486
##                                          kHP-4: 758   kKM-1: 831
##                                          kHP-5:  39   kKM-4: 109
##
```

```
##       f.cost          f.tt      hcpckMCA  hcpckMCA_hcpck  hcpckMCA_claKM
##  (11,18] :1188   (10,15]: 913   1:  30   kHPmca-4:1952   kHPmca-2:1088
##  (18,30] : 724   (15,20]: 549   2:1088   kHPmca-3:1433   kHPmca-3:1433
##  (30,50] : 221   (20,50]: 694   3:1433   kHPmca-2:1088   kHPmca-1:  30
##  (50,129):  63   (5,10] :1511   4:1952   kHPmca-1:  30   kHPmca-4:1952
##  (8,11]  :1151   [0,5]  : 894   5: 120   kHPmca-5: 120   kHPmca-5: 120
##  [0,8]   :1276   NA's   :  62
```

2 Refactor

```
names(df)
```

```
##  [1] "VendorID"              "RateCodeID"            "Pickup_longitude"
##  [4] "Pickup_latitude"       "Dropoff_longitude"     "Dropoff_latitude"
##  [7] "Passenger_count"       "Trip_distance"         "Fare_amount"
## [10] "Extra"                 "MTA_tax"               "Tip_amount"
## [13] "Tolls_amount"          "improvement_surcharge" "Total_amount"
## [16] "Payment_type"          "Trip_type"             "hour"
## [19] "period"                "tlenkm"                "traveltime"
## [22] "espeed"                "pickup"                "dropoff"
## [25] "Trip_distance_range"   "paidTolls"             "TipIsGiven"
## [28] "passenger_groups"      "hcpck"                 "claKM"
## [31] "f.cost"                "f.tt"                  "hcpckMCA"
## [34] "hcpckMCA_hcpck"        "hcpckMCA_claKM"
```

```
names(df)[names(df) == "VendorID"] <- "f.vendor_id"
names(df)[names(df) == "RateCodeID"] <- "f.code_rate_id"
names(df)[names(df) == "Pickup_longitude"] <- "q.pickup_longitude"
names(df)[names(df) == "Pickup_latitude"] <- "q.pickup_latitude"
names(df)[names(df) == "Dropoff_longitude"] <- "q.dropoff_longitude"
names(df)[names(df) == "Dropoff_latitude"] <- "q.dropoff_latitude"
names(df)[names(df) == "Passenger_count"] <- "q.passenger_count"
names(df)[names(df) == "Trip_distance"] <- "q.trip_distance"
names(df)[names(df) == "Fare_amount"] <- "q.fare_amount"
names(df)[names(df) == "Extra"] <- "q.extra"
names(df)[names(df) == "MTA_tax"] <- "f.mta_tax"
names(df)[names(df) == "Tip_amount"] <- "q.tip_amount"
names(df)[names(df) == "Tolls_amount"] <- "q.tolls_amount"
names(df)[names(df) == "improvement_surcharge"] <- "f.improvement_surcharge"
names(df)[names(df) == "Total_amount"] <- "target.total_amount"
names(df)[names(df) == "Payment_type"] <- "f.payment_type"
names(df)[names(df) == "Trip_type"] <- "f.trip_type"
names(df)[names(df) == "hour"] <- "q.hour"
names(df)[names(df) == "period"] <- "f.period"
names(df)[names(df) == "tlenkm"] <- "q.tlenkm"
names(df)[names(df) == "traveltime"] <- "q.traveltime"
names(df)[names(df) == "espeed"] <- "q.espeed"
names(df)[names(df) == "pickup"] <- "qual.pickup"
names(df)[names(df) == "dropoff"] <- "qual.dropoff"
names(df)[names(df) == "Trip_distance_range"] <- "f.trip_distance_range"
names(df)[names(df) == "paidTolls"] <- "f.paid_tolls"
names(df)[names(df) == "TipIsGiven"] <- "target.tip_is_given"
names(df)[names(df) == "passenger_groups"] <- "f.passenger_groups"
#names(df)[names(df) == "f.cost"] <- ""
#names(df)[names(df) == "f.tt"] <- ""

df$hcpck <- NULL
df$claKM <- NULL
df$hcpckMCA <- NULL
df$hcpckMCA_hcpck <- NULL
df$hcpckMCA_claKM <- NULL

names(df)
```

```
##  [1] "f.vendor_id"             "f.code_rate_id"
##  [3] "q.pickup_longitude"      "q.pickup_latitude"
##  [5] "q.dropoff_longitude"     "q.dropoff_latitude"
##  [7] "q.passenger_count"       "q.trip_distance"
##  [9] "q.fare_amount"           "q.extra"
## [11] "f.mta_tax"               "q.tip_amount"
## [13] "q.tolls_amount"          "f.improvement_surcharge"
## [15] "target.total_amount"     "f.payment_type"
## [17] "f.trip_type"             "q.hour"
## [19] "f.period"                "q.tlenkm"
## [21] "q.traveltime"            "q.espeed"
## [23] "qual.pickup"             "qual.dropoff"
## [25] "f.trip_distance_range"   "f.paid_tolls"
## [27] "target.tip_is_given"     "f.passenger_groups"
## [29] "f.cost"                  "f.tt"
```

Remove total amount equal to 0

```
df<-df[!(df$target.total_amount=="0"),]
```

3 Create factors needed for this deliverable (according to teacher's video recording)

We must create: f.cost, f.dist, f.tt and f.hour. We already have f.cost and f.tt, so we will only have to create f.dist and f.hour:

3.1 f.dist

```
df$f.dist[df$q.trip_distance<=1.6] = "(0, 1.6]"
df$f.dist[(df$q.trip_distance>1.6) & (df$q.trip_distance<=3)] = "(1.6, 3]"
df$f.dist[(df$q.trip_distance>3) & (df$q.trip_distance<=5.5)] = "(3, 5.5]"
df$f.dist[(df$q.trip_distance>5.5) & (df$q.trip_distance<=30)] = "(5.5, 30]"
df$f.dist<-factor(df$f.dist)
```

3.2 f.hour

```
df$f.hour[(df$q.hour>=17) & (df$q.hour<18)] = "17"
df$f.hour[(df$q.hour>=18) & (df$q.hour<19)] = "18"
df$f.hour[(df$q.hour>=19) & (df$q.hour<20)] = "19"
df$f.hour[(df$q.hour>=20) & (df$q.hour<21)] = "20"
df$f.hour[(df$q.hour>=21) & (df$q.hour<22)] = "21"
df$f.hour[(df$q.hour>=22) & (df$q.hour<23)] = "22"
df$f.hour[(df$q.hour<17)] = "other"
df$f.hour[(df$q.hour>=23)] = "other"
df$f.hour<-factor(df$f.hour)
```

4 Listing out variables

```
vars_con<-names(df)[c(3:10,12:13,15,18,20:22)];
vars_dis<-names(df)[c(1:2,16,19,27:32)];
vars_res<-names(df)[c(15,27)];
vars_cexp<-vars_con[c(5:10,12:15)];
```

5 Useful information

5.1 Y (Numeric Target).

This variable will be the target for linear model building (connected to blocks Statistical Modeling I and II).

6 Explanatory variables numeric only

Before we begin to see correlations with our target, we should consider the normality of this.

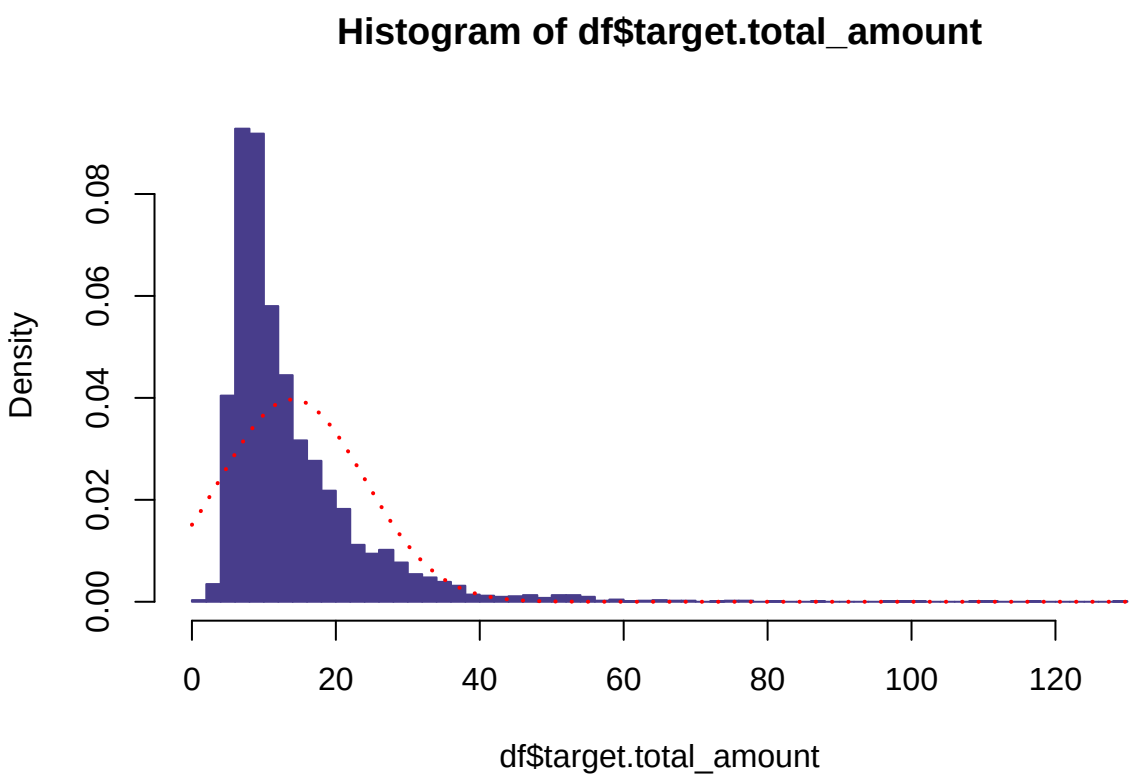
6.1 Normality

```
hist(df$target.total_amount,50,freq=F,col="darkslateblue",border = "darkslateblue")
mm<-mean(df$target.total_amount);ss<-sd(df$target.total_amount);mm;ss
```

```
## [1] 13.96265
```

```
## [1] 10.03383
```

```
curve(dnorm(x,mean=mm,sd=ss),col="red",lwd=2,lty=3, add=T)
```




```
shapiro.test(df$target.total_amount)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$target.total_amount
## W = 0.73071, p-value < 2.2e-16
```

We see that the target total_amount is not normally distributed for the following reasons:

- graph: there is no symmetry in the plot
- shapiro: we see that the p-value is too large to accept the assumption that target.total_amount is normally distributed

6.1.1 Symmetry

```
skewness(df$target.total_amount)
```

```
## [1] 3.176789
```

Normal data should have 0 skewness: we see that our data is right skewed (3.18).

6.1.2 Kurtosis

```
kurtosis(df$target.total_amount)  # Normal data should 3 5.35 >> 3
```

```
## [1] 21.09556
```

Normal data should be 3. We have 21.1, so, in this case, our data is not normal.

6.2 Initial model: take the most correlated variables

```
# we use spearman method since out target is not normally distributed
round(cor(df[,c("target.total_amount",vars_cexp)], method="spearman"),dig=2)
```

```
##                     target.total_amount q.passenger_count q.trip_distance
## target.total_amount                1.00              0.01            0.93
## q.passenger_count                  0.01              1.00            0.01
## q.trip_distance                    0.93              0.01            1.00
## q.fare_amount                      0.97              0.01            0.95
## q.extra                            0.03              0.05           -0.05
## q.tip_amount                       0.41             -0.01            0.26
## q.tolls_amount                     0.15              0.01            0.14
## q.hour                            -0.01              0.01           -0.05
## q.tlenkm                           0.91              0.00            0.98
## q.traveltime                       0.90             -0.01            0.87
## q.espeed                           0.29              0.02            0.46
##                     q.fare_amount q.extra q.tip_amount q.tolls_amount q.hour
## target.total_amount          0.97    0.03         0.41           0.15  -0.01
## q.passenger_count            0.01    0.05        -0.01           0.01   0.01
## q.trip_distance              0.95   -0.05         0.26           0.14  -0.05
## q.fare_amount                1.00   -0.06         0.25           0.14  -0.04
## q.extra                     -0.06    1.00         0.02          -0.02   0.32
## q.tip_amount                 0.25    0.02         1.00           0.11   0.02
## q.tolls_amount               0.14   -0.02         0.11           1.00  -0.01
## q.hour                      -0.04    0.32         0.02          -0.01   1.00
## q.tlenkm                     0.93   -0.03         0.25           0.14  -0.04
## q.traveltime                 0.93   -0.03         0.22           0.11  -0.02
## q.espeed                     0.28   -0.01         0.14           0.12  -0.07
##                     q.tlenkm q.traveltime q.espeed
## target.total_amount     0.91         0.90     0.29
## q.passenger_count       0.00        -0.01     0.02
## q.trip_distance         0.98         0.87     0.46
## q.fare_amount           0.93         0.93     0.28
## q.extra                -0.03        -0.03    -0.01
## q.tip_amount            0.25         0.22     0.14
## q.tolls_amount          0.14         0.11     0.12
## q.hour                 -0.04        -0.02    -0.07
## q.tlenkm                1.00         0.88     0.45
## q.traveltime            0.88         1.00     0.05
## q.espeed                0.45         0.05     1.00
```

We see that the diagonal is full of '1', since this command gives us the correlation between the same variable. Apart from this diagonal, however, there are more high correlations. Let's see which ones are correlated with our target:

- q.fare_amount: 0.97
- q.trip_distance: 0.93
- q.tlenkm: 0.91 (like trip_distance)
- q.traveltime: 0.90
- q.tip_amount: 0.41 (not much, but must be taken into account)
- q.espeed: 0.29 (not much, but must be taken into account)
- q.tolls_amount: 0.15 (not much, but must be taken into account)
- we can see that some of them are not correlated:
  - q.extra (0.03)
  - q.passenger_count (0.01)
  - q.hour (-0.01)

After seeing the correlation, to make an initial model, we should select the ones that are most correlated, which are:

- q.fare_amount
- q.trip_distance (we are not taking tlenkm because of redundance)
- q.traveltime

- q.tip_amount
- q.espeed
- q.tolls_amount

6.3 Second model: take the entire dataset with a condes

```
res.con <- condes(df,num.var=which(names(df)=="target.total_amount"))
```

```
res.con$quanti
```

```
##                      correlation       p.value
## q.fare_amount         0.94425003  0.000000e+00
## q.trip_distance       0.89702734  0.000000e+00
## q.tlenkm              0.88362042  0.000000e+00
## q.traveltime          0.76448863  0.000000e+00
## q.tip_amount          0.56622837  0.000000e+00
## q.espeed              0.39683909 9.313540e-174
## q.tolls_amount        0.25751662  9.659999e-71
## q.hour               -0.03110910  3.465376e-02
## q.pickup_longitude   -0.04064371  5.775239e-03
## q.dropoff_longitude -0.06391905  1.401371e-05
## q.pickup_latitude    -0.12322848  4.560732e-17
## q.dropoff_latitude   -0.14812217  4.926074e-24
```

Com hem pogut veure abans, les variables més correlacionades són:

- q.fare_amount: 0.94
  - it is normal for the rate to go up when the price goes up
- q.trip_distance: 0.90
  - the more distance, the more time, and therefore the more price
- q.tlenkm: 0.88
  - just like the previous one
- q.traveltime: 0.76
  - the longer, the more price
- q.tip_amount: 0.57
  - not so much related, but we can keep in mind that people tend to give a percentage of the total price
- q.espeed: 0.40
- q.tolls_amount: 0.26

```
res.con$quali
```

```
##                                  R2       p.value
## f.trip_distance_range   0.560261998  0.000000e+00
## f.cost                  0.908376615  0.000000e+00
## f.tt                    0.549724557  0.000000e+00
## f.dist                  0.636791987  0.000000e+00
## f.paid_tolls            0.109693994 5.477144e-117
## target.tip_is_given     0.057803014  1.250800e-61
## f.payment_type          0.052910669  4.024719e-55
## f.code_rate_id          0.018930689  6.290954e-21
## f.mta_tax               0.005160632  1.044478e-06
## f.trip_type             0.003203349  1.204051e-04
## f.improvement_surcharge 0.002760154  3.583467e-04
## qual.dropoff            0.008369578  2.171667e-02
```

To talk about factor variables, we need to visualize res.con$quali. So let's see:

- f.trip_distance_range
  - we see that they are totally related, just as we see with que.trip_distance, since the longer distance, the longer time, and therefore the more price
- f.cost
  - is equivalent to our target
- f.tt
  - he longer time, the more price
- f.dist
  - just like with f.trip_distance_range

- f.paid_tolls
  - if you pay more, it means that the trip has lasted longer, and therefore has been longer, and is more likely to have gone through more tolls
- target.tip_is_given
  - just like before, but we can keep in mind that people tend to give a percentage of the total price

6.4 Third model: if few explanatory variables are available: take all of them

```
vars_cexp
```

```
##  [1] "q.passenger_count" "q.trip_distance"   "q.fare_amount"
##  [4] "q.extra"           "q.tip_amount"      "q.tolls_amount"
##  [7] "q.hour"            "q.tlenkm"          "q.traveltime"
## [10] "q.espeed"
```

```
cor(df$q.trip_distance,df$q.tlenkm)
```

```
## [1] 0.993941
```

To give an example, we see that the two distances we have, trip_distance and tlenkm, are closely related, since they represent the same.

6.4.1 Model 1

```
model_1 <- lm(
  target.total_amount~.,
  data=df[,c("target.total_amount",vars_cexp)]
)
summary(model_1)
```

```
##
## Call:
## lm(formula = target.total_amount ~ ., data = df[, c("target.total_amount",
##     vars_cexp)])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.835 -0.192 -0.056  0.069 95.131
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.162e+00  1.899e-01  11.389  < 2e-16 ***
## q.passenger_count  9.589e-03  3.679e-02   0.261    0.794
## q.trip_distance    7.341e-01  1.468e-01   5.002 5.88e-07 ***
## q.fare_amount      8.926e-01  1.485e-02  60.101  < 2e-16 ***
## q.extra            1.082e+00  1.074e-01  10.074  < 2e-16 ***
## q.tip_amount       1.042e+00  2.316e-02  44.981  < 2e-16 ***
## q.tolls_amount     1.050e+00  7.780e-02  13.498  < 2e-16 ***
## q.hour             1.824e-05  5.816e-03   0.003    0.997
## q.tlenkm           9.138e-03  8.326e-02   0.110    0.913
## q.traveltime      -5.666e-02  8.703e-03  -6.510 8.29e-11 ***
## q.espeed          -6.943e-02  7.309e-03  -9.500  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.584 on 4600 degrees of freedom
## Multiple R-squared:  0.9338, Adjusted R-squared:  0.9337
## F-statistic:  6489 on 10 and 4600 DF,  p-value: < 2.2e-16
```

Model_1 explains 93.38% of the variability of the target. We also see, according to the F-statistic, that it should be rejected.

We cannot use variables that are so correlated at the same time to act as explanatory variables. Therefore, we need to make a model in which we do not have these correlations.

But first, let's see which of them are that correlated:

```r
vif(model_1)  # Check association between explanatory vars
```

```
## q.passenger_count   q.trip_distance     q.fare_amount           q.extra
##          1.004232        115.121825         10.382784          1.071486
##      q.tip_amount    q.tolls_amount            q.hour          q.tlenkm
##          1.247112          1.069572          1.073491         97.049752
##      q.traveltime          q.espeed
##          5.259497          2.799567
```

When the variance inflation factor is greater than 5, we need to consider whether or not we keep a variable.

- q.trip_distance: 115.12
- q.tlenkm: 97.05
- q.fare_amount: 10.38
- q.traveltime: 5.26

In this case we have to choose how far we stay. Since we work better with km than with miles (or inches, or whatever it is), we could choose the variable q.tlenkm.

```r
Anova(model_1)
```

```
## Anova Table (Type II tests)
##
## Response: target.total_amount
##                     Sum Sq   Df   F value    Pr(>F)
## q.passenger_count      0.5    1    0.0679    0.7944
## q.trip_distance      167.1    1   25.0233 5.875e-07 ***
## q.fare_amount      24124.8    1 3612.1640 < 2.2e-16 ***
## q.extra              677.9    1  101.4942 < 2.2e-16 ***
## q.tip_amount       13513.3    1 2023.3250 < 2.2e-16 ***
## q.tolls_amount      1216.8    1  182.1841 < 2.2e-16 ***
## q.hour                 0.0    1    0.0000    0.9975
## q.tlenkm               0.1    1    0.0120    0.9126
## q.traveltime         283.1    1   42.3864 8.292e-11 ***
## q.espeed             602.7    1   90.2445 < 2.2e-16 ***
## Residuals          30722.3 4600
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This table shows us whether or not it is necessary to have certain variables in our model. We can remove a variable from our model with the following:

6.4.2 Model 1 with BIC

```r
model_1_bic <- step( model_1, k=log(nrow(df)) )
```

```
## Start:  AIC=8837.76
## target.total_amount ~ q.passenger_count + q.trip_distance + q.fare_amount +
##     q.extra + q.tip_amount + q.tolls_amount + q.hour + q.tlenkm +
##     q.traveltime + q.espeed
##
##                      Df Sum of Sq   RSS     AIC
## - q.hour              1       0.0 30722  8829.3
## - q.tlenkm            1       0.1 30722  8829.3
## - q.passenger_count   1       0.5 30723  8829.4
## <none>                            30722  8837.8
## - q.trip_distance     1     167.1 30889  8854.3
## - q.traveltime        1     283.1 31005  8871.6
## - q.espeed            1     602.7 31325  8918.9
## - q.extra             1     677.9 31400  8930.0
## - q.tolls_amount      1    1216.8 31939  9008.4
## - q.tip_amount        1   13513.3 44236 10510.2
## - q.fare_amount       1   24124.8 54847 11501.7
##
## Step:  AIC=8829.33
## target.total_amount ~ q.passenger_count + q.trip_distance + q.fare_amount +
```

```
##     q.extra + q.tip_amount + q.tolls_amount + q.tlenkm + q.traveltime +
##     q.espeed
##
##                       Df Sum of Sq   RSS     AIC
## - q.tlenkm             1       0.1 30722  8820.9
## - q.passenger_count    1       0.5 30723  8821.0
## <none>                             30722  8829.3
## - q.trip_distance      1     167.3 30890  8845.9
## - q.traveltime         1     283.5 31006  8863.2
## - q.espeed             1     606.2 31329  8911.0
## - q.extra              1     717.2 31440  8927.3
## - q.tolls_amount       1    1216.8 31939  9000.0
## - q.tip_amount         1   13520.5 44243 10502.5
## - q.fare_amount        1   24147.0 54869 11495.1
##
## Step:  AIC=8820.9
## target.total_amount ~ q.passenger_count + q.trip_distance + q.fare_amount +
##     q.extra + q.tip_amount + q.tolls_amount + q.traveltime +
##     q.espeed
##
##                       Df Sum of Sq   RSS     AIC
## - q.passenger_count    1       0.5 30723  8812.5
## <none>                             30722  8820.9
## - q.traveltime         1     307.4 31030  8858.4
## - q.espeed             1     610.5 31333  8903.2
## - q.extra              1     719.3 31442  8919.2
## - q.tolls_amount       1    1223.7 31946  8992.6
## - q.trip_distance      1    1814.8 32537  9077.1
## - q.tip_amount         1   13546.6 44269 10496.8
## - q.fare_amount        1   27138.0 57860 11731.4
##
## Step:  AIC=8812.53
## target.total_amount ~ q.trip_distance + q.fare_amount + q.extra +
##     q.tip_amount + q.tolls_amount + q.traveltime + q.espeed
##
##                     Df Sum of Sq   RSS     AIC
## <none>                           30723  8812.5
## - q.traveltime       1     307.4 31030  8850.0
## - q.espeed           1     610.2 31333  8894.8
## - q.extra            1     723.1 31446  8911.4
## - q.tolls_amount     1    1224.6 31947  8984.3
## - q.trip_distance    1    1815.2 32538  9068.8
## - q.tip_amount       1   13548.8 44272 10488.7
## - q.fare_amount      1   27138.0 57861 11723.0
```

The BIC has been eliminating the variables it has considered, without worsening the AIC. However, since it does not take into account either correlations or concepts, it is probably not optimal.

Let's see how it turned out:

```
vif(model_1_bic)
```

```
## q.trip_distance   q.fare_amount         q.extra    q.tip_amount  q.tolls_amount
##       11.047912        9.218707        1.008119        1.243277        1.064477
##    q.traveltime        q.espeed
##        4.795267        2.757712
```

Note that trip_distance still has a vif greater than 5 (11.047912), and so does fare_amount (9.218707).

```
summary(model_1_bic)
```

```
##
## Call:
## lm(formula = target.total_amount ~ q.trip_distance + q.fare_amount +
##     q.extra + q.tip_amount + q.tolls_amount + q.traveltime +
##     q.espeed, data = df[, c("target.total_amount", vars_cexp)])
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.849 -0.191 -0.057  0.070 95.124
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.173916   0.165191  13.160  < 2e-16 ***
## q.trip_distance  0.749490   0.045447  16.491  < 2e-16 ***
## q.fare_amount    0.892050   0.013990  63.764  < 2e-16 ***
## q.extra          1.084258   0.104168  10.409  < 2e-16 ***
## q.tip_amount     1.041425   0.023115  45.055  < 2e-16 ***
## q.tolls_amount   1.051037   0.077595  13.545  < 2e-16 ***
## q.traveltime    -0.056379   0.008307  -6.787 1.29e-11 ***
## q.espeed        -0.069341   0.007252  -9.562  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.584 on 4603 degrees of freedom
## Multiple R-squared:  0.9338, Adjusted R-squared:  0.9337
## F-statistic:  9276 on 7 and 4603 DF,  p-value: < 2.2e-16
```

However, we see that it continues to explain much of the variability of our target (93.38%).

Therefore, we will try to make a model manually based on what model_1_bic has shown us and our knowledge of the data:

6.4.3 Model 2

```
model_2 <- lm(
  target.total_amount~
    q.passenger_count +
    q.fare_amount +
    q.extra +
    q.tip_amount +
    q.tolls_amount +
    q.hour +
    q.tlenkm +
    q.traveltime +
    q.espeed
  ,
  data=df[,c("target.total_amount",vars_cexp)]
)
summary(model_2)
```

```
##
## Call:
## lm(formula = target.total_amount ~ q.passenger_count + q.fare_amount +
##     q.extra + q.tip_amount + q.tolls_amount + q.hour + q.tlenkm +
##     q.traveltime + q.espeed, data = df[, c("target.total_amount",
##     vars_cexp)])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.884 -0.193 -0.062  0.057 95.280
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.0366538  0.1886809  10.794  < 2e-16 ***
## q.passenger_count 0.0077488  0.0368870   0.210    0.834
## q.fare_amount     0.9304944  0.0128068  72.656  < 2e-16 ***
## q.extra           1.0675657  0.1076659   9.916  < 2e-16 ***
## q.tip_amount      1.0486524  0.0231757  45.248  < 2e-16 ***
## q.tolls_amount    1.0284515  0.0778861  13.205  < 2e-16 ***
## q.hour           -0.0008473  0.0058287  -0.145    0.884
## q.tlenkm          0.4051201  0.0258591  15.666  < 2e-16 ***
```

```
## q.traveltime      -0.0644403  0.0085849  -7.506 7.27e-14 ***
## q.espeed          -0.0658133  0.0072921  -9.025  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.591 on 4601 degrees of freedom
## Multiple R-squared:  0.9334, Adjusted R-squared:  0.9333
## F-statistic:  7170 on 9 and 4601 DF,  p-value: < 2.2e-16
```

We see that the explainability is now 93.34%.

```
vif(model_2)  # Check association between explanatory vars
```

```
## q.passenger_count     q.fare_amount           q.extra      q.tip_amount
##          1.004132          7.680532          1.070685          1.242570
##    q.tolls_amount            q.hour          q.tlenkm      q.traveltime
##          1.066240          1.072540          9.313988          5.091473
##          q.espeed
##          2.772111
```

Even so, owning one is still beyond the reach of the average person.

We try to make a new model without the distance:

6.4.4 Model 3

```
model_3 <- lm(
  target.total_amount~
    q.passenger_count +
    q.fare_amount +
    q.extra +
    q.tip_amount +
    q.tolls_amount +
    q.hour +
    q.traveltime +
    q.espeed
  ,
  data=df[,c("target.total_amount",vars_cexp)]
)
summary(model_3)
```

```
##
## Call:
## lm(formula = target.total_amount ~ q.passenger_count + q.fare_amount +
##     q.extra + q.tip_amount + q.tolls_amount + q.hour + q.traveltime +
##     q.espeed, data = df[, c("target.total_amount", vars_cexp)])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.322 -0.251  0.000  0.117 95.540
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.2903616  0.1562258   1.859   0.0631 .
## q.passenger_count  0.0132996  0.0378522   0.351   0.7253
## q.fare_amount      1.0440693  0.0108341  96.369   <2e-16 ***
## q.extra            1.1208455  0.1104332  10.150   <2e-16 ***
## q.tip_amount       1.0607708  0.0237700  44.627   <2e-16 ***
## q.tolls_amount     1.0842604  0.0798441  13.580   <2e-16 ***
## q.hour            -0.0001983  0.0059813  -0.033   0.9736
## q.traveltime      -0.0089434  0.0080250  -1.114   0.2651
## q.espeed           0.0052878  0.0058573   0.903   0.3667
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.659 on 4602 degrees of freedom
```

```
## Multiple R-squared:  0.9299, Adjusted R-squared:  0.9298
## F-statistic:  7630 on 8 and 4602 DF,  p-value: < 2.2e-16
```

We see that the explainability is now 92.99%.

```
vif(model_3)  # Check association between explanatory vars
```

```
## q.passenger_count     q.fare_amount           q.extra      q.tip_amount
##          1.004039          5.219389          1.069616          1.241186
##    q.tolls_amount            q.hour      q.traveltime          q.espeed
##          1.064009          1.072486          4.224578          1.698328
```

The live ones are fine now. Still, we've pulled the distance, which conceptually we can't afford. Therefore, we will try to remove another variable with a high vif (q.fare_amount), instead of q.tlenkm:

6.4.5 Model 4

```
model_4 <- lm(
  target.total_amount~
    q.passenger_count +
    q.extra +
    q.tip_amount +
    q.tolls_amount +
    q.hour +
    q.tlenkm +
    q.traveltime +
    q.espeed
  ,
  data=df[,c("target.total_amount",vars_cexp)]
)
summary(model_4)
```

```
##
## Call:
## lm(formula = target.total_amount ~ q.passenger_count + q.extra +
##     q.tip_amount + q.tolls_amount + q.hour + q.tlenkm + q.traveltime +
##     q.espeed, data = df[, c("target.total_amount", vars_cexp)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.629  -0.639  -0.298   0.154  96.125
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.516476   0.271899  16.611  < 2e-16 ***
## q.passenger_count  0.005450   0.054048   0.101 0.919678
## q.extra            0.524497   0.157374   3.333 0.000867 ***
## q.tip_amount       1.238137   0.033742  36.694  < 2e-16 ***
## q.tolls_amount     1.327020   0.113962  11.644  < 2e-16 ***
## q.hour             0.006590   0.008539   0.772 0.440317
## q.tlenkm           1.468677   0.031234  47.021  < 2e-16 ***
## q.traveltime       0.189991   0.011485  16.543  < 2e-16 ***
## q.espeed          -0.045716   0.010677  -4.282 1.89e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.796 on 4602 degrees of freedom
## Multiple R-squared:  0.8571, Adjusted R-squared:  0.8568
## F-statistic:  3450 on 8 and 4602 DF,  p-value: < 2.2e-16
```

We see that the explainability is now 85.71%.

```
vif(model_4)  # Check association between explanatory vars
```

```
## q.passenger_count           q.extra      q.tip_amount    q.tolls_amount
##          1.004131          1.065525          1.226836          1.063271
##            q.hour          q.tlenkm      q.traveltime          q.espeed
```

```
##          1.072210          6.329422          4.244321          2.768123
```

Despite having high vifs, we still have high explicability of the variability of our target and, given that the variable we have taken out we can remove with time and distance from the trip, we do not need it.

So we continue to stay with this variable and make new models. We apply BIC to help us a little:

```
model_4_bic <- step( model_4, k=log(nrow(df)) )
```

```
## Start:  AIC=12369.79
## target.total_amount ~ q.passenger_count + q.extra + q.tip_amount +
##     q.tolls_amount + q.hour + q.tlenkm + q.traveltime + q.espeed
##
##                      Df Sum of Sq   RSS   AIC
## - q.passenger_count  1         0 66331 12361
## - q.hour             1         9 66339 12362
## <none>                           66330 12370
## - q.extra            1       160 66491 12372
## - q.espeed           1       264 66595 12380
## - q.tolls_amount     1      1954 68285 12495
## - q.traveltime       1      3944 70275 12628
## - q.tip_amount       1     19407 85738 13545
## - q.tlenkm           1     31868 98198 14170
##
## Step:  AIC=12361.36
## target.total_amount ~ q.extra + q.tip_amount + q.tolls_amount +
##     q.hour + q.tlenkm + q.traveltime + q.espeed
##
##                   Df Sum of Sq   RSS   AIC
## - q.hour           1         9 66339 12354
## <none>                         66331 12361
## - q.extra          1       161 66492 12364
## - q.espeed         1       264 66595 12371
## - q.tolls_amount   1      1955 68286 12487
## - q.traveltime     1      3944 70275 12619
## - q.tip_amount     1     19413 85743 13537
## - q.tlenkm         1     31873 98204 14162
##
## Step:  AIC=12353.52
## target.total_amount ~ q.extra + q.tip_amount + q.tolls_amount +
##     q.tlenkm + q.traveltime + q.espeed
##
##                   Df Sum of Sq   RSS   AIC
## <none>                         66339 12354
## - q.extra          1       189 66528 12358
## - q.espeed         1       273 66613 12364
## - q.tolls_amount   1      1957 68296 12479
## - q.traveltime     1      3937 70276 12611
## - q.tip_amount     1     19443 85782 13530
## - q.tlenkm         1     31908 98247 14156
```

Following BIC, we have to eliminate variables until the vif's are less than 5. Therefore, the model that meets this is:

6.4.6 Model 5

```
model_5 <- lm(
  target.total_amount~
    q.passenger_count +
    q.extra +
    q.tip_amount +
    q.tolls_amount +
    q.tlenkm +
    q.traveltime
  ,
  data=df
```

```
)
summary(model_5)
```

```
##
## Call:
## lm(formula = target.total_amount ~ q.passenger_count + q.extra +
##     q.tip_amount + q.tolls_amount + q.tlenkm + q.traveltime,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.989  -0.648  -0.286   0.180  96.220
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.702670   0.127289  29.089  < 2e-16 ***
## q.passenger_count 0.002893   0.054139   0.053 0.957387
## q.extra           0.572837   0.153213   3.739 0.000187 ***
## q.tip_amount      1.235198   0.033787  36.558  < 2e-16 ***
## q.tolls_amount    1.326554   0.114175  11.619  < 2e-16 ***
## q.tlenkm          1.362728   0.019479  69.960  < 2e-16 ***
## q.traveltime      0.223825   0.008417  26.593  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.804 on 4604 degrees of freedom
## Multiple R-squared:  0.8565, Adjusted R-squared:  0.8563
## F-statistic:  4579 on 6 and 4604 DF,  p-value: < 2.2e-16
```

We see that the explainability is now 85.65%

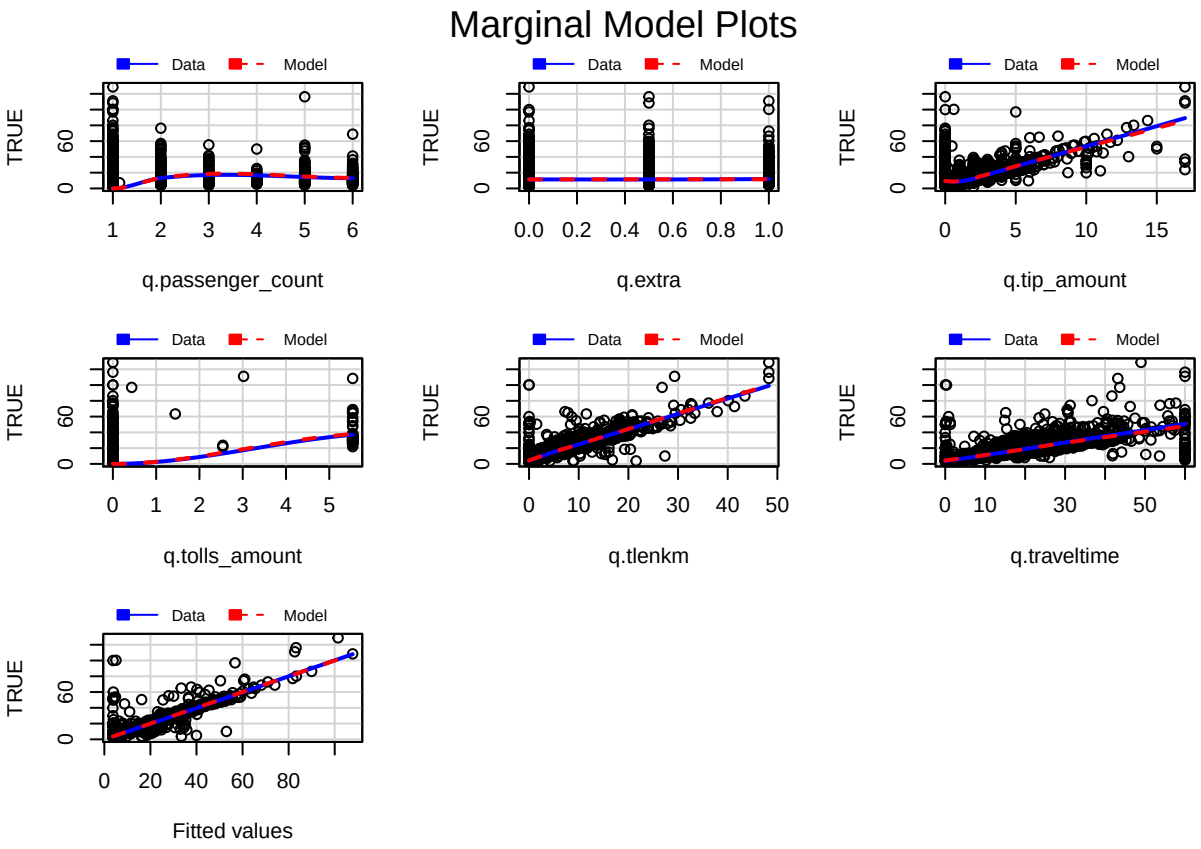
```
vif(model_5)  # Check association between explanatory vars
```

```
## q.passenger_count           q.extra      q.tip_amount    q.tolls_amount
##          1.003693          1.006079          1.225445          1.063204
##          q.tlenkm      q.traveltime
##          2.452236          2.270921
```

There is no vif that exceeds 5.

Let's now discriminate the variables independently:

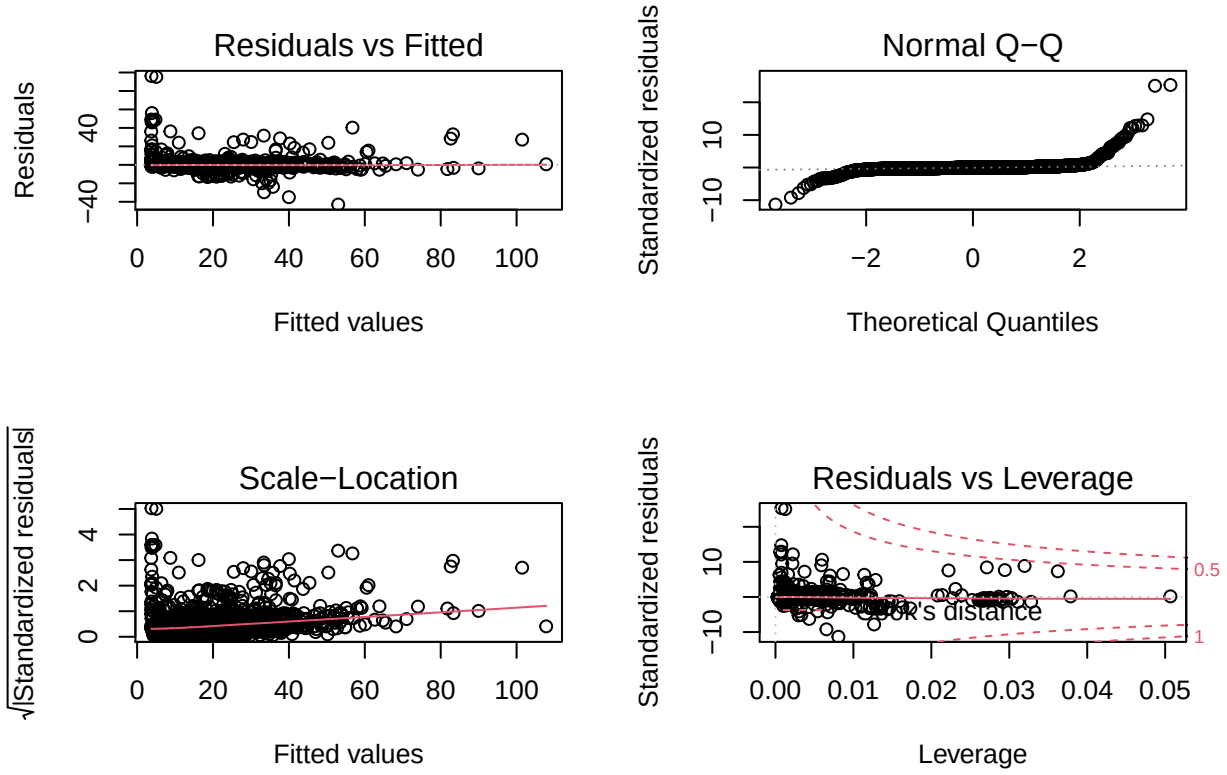
```
marginalModelPlots(model_5)
```



We see that there is not much mismatch of the marginal variables. If there were any, we would have to transform our explanatory variables.

6.5 Diagnostics

```
par(mfrow=c(2,2))
plot(model_5, id.n=0 )
```




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

Looking at the results, we can say that: * There is no normalcy * And, in terms of the Residual vs Leverage graph, our variables are within the R model, but it's not very reliable, so it doesn't help us much.

All this is due to the fact that our target variable was no longer normally distributed. To solve this, we apply the logarithm:

PER PREGUNTAAAAAAR———————————————————————————

```
model_6 <- lm(
  log(target.total_amount)~
    q.passenger_count +
    q.extra +
    q.tip_amount +
    q.tolls_amount +
    q.tlenkm +
    q.traveltime
  ,
  data=df
)
summary(model_6)
```

```
##
## Call:
## lm(formula = log(target.total_amount) ~ q.passenger_count + q.extra +
##     q.tip_amount + q.tolls_amount + q.tlenkm + q.traveltime,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.48656 -0.11197  0.03604  0.14401  2.73740
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.8619225  0.0084440 220.502  < 2e-16 ***
## q.passenger_count -0.0014410  0.0035914  -0.401    0.688
## q.extra            0.0695052  0.0101637   6.839 9.05e-12 ***
## q.tip_amount       0.0625017  0.0022413  27.886  < 2e-16 ***
## q.tolls_amount     0.0307963  0.0075741   4.066 4.86e-05 ***
## q.tlenkm           0.0550963  0.0012922  42.639  < 2e-16 ***
## q.traveltime       0.0218598  0.0005583  39.151  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2523 on 4604 degrees of freedom
## Multiple R-squared:  0.7957, Adjusted R-squared:  0.7954
## F-statistic:  2989 on 6 and 4604 DF,  p-value: < 2.2e-16
```

We see that when doing the logarithm, the coefficient of determination is getting lower and lower, now it is 79.57%. We have seen that it has gotten worse than the previous model. Therefore, we discard it. We will work with model_5.

However, let's look at a comparison of the latest models used:

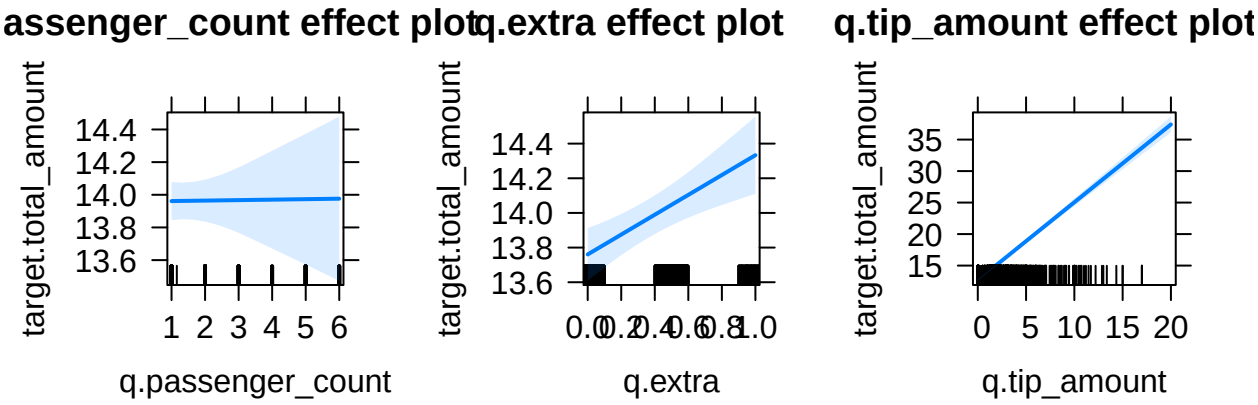
```
BIC(model_4, model_5, model_6)
```

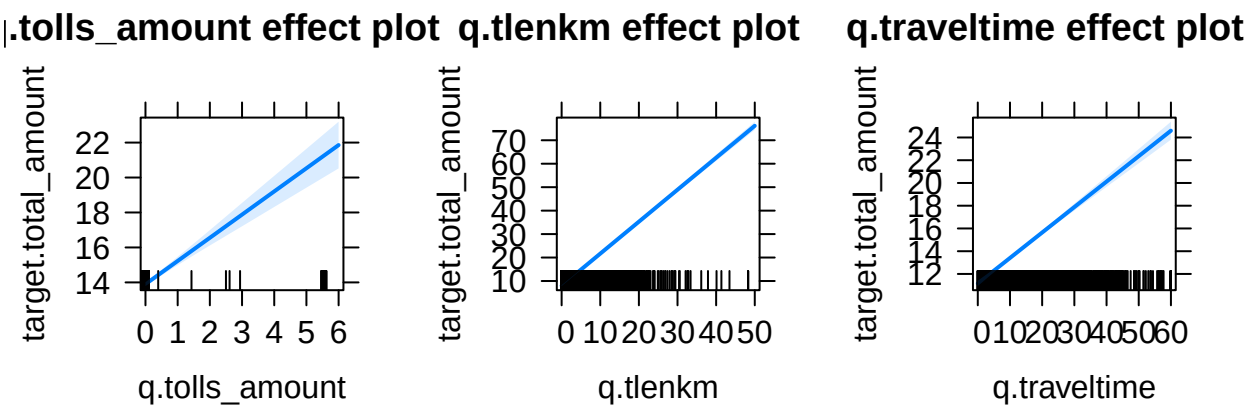
```
##         df        BIC
## model_4 10 25463.6783
## model_5  8 25466.3721
## model_6  8   447.0392
```

We can see that model_5 is much better than model_4, and given the coefficient of determination of model_6, it is clear that our winner is model_5.

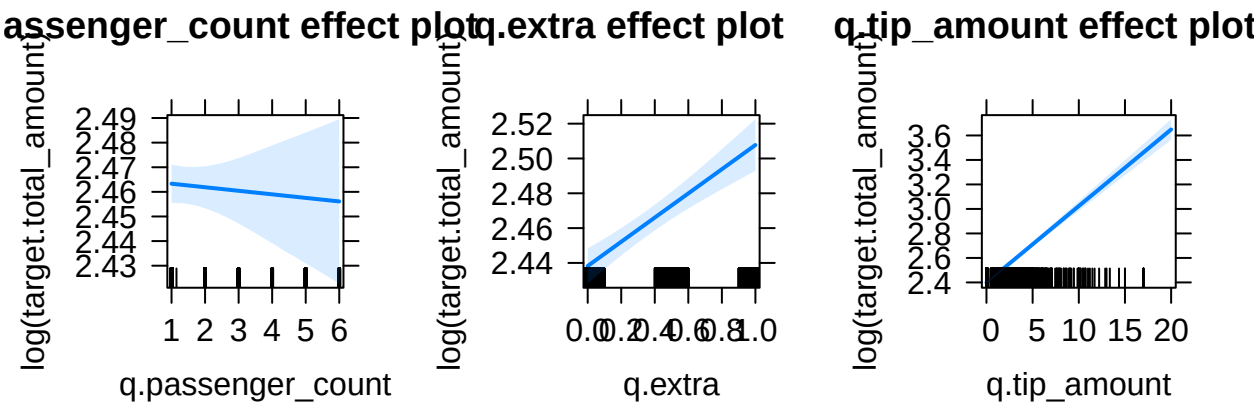
So, let's look at the effects of this model:

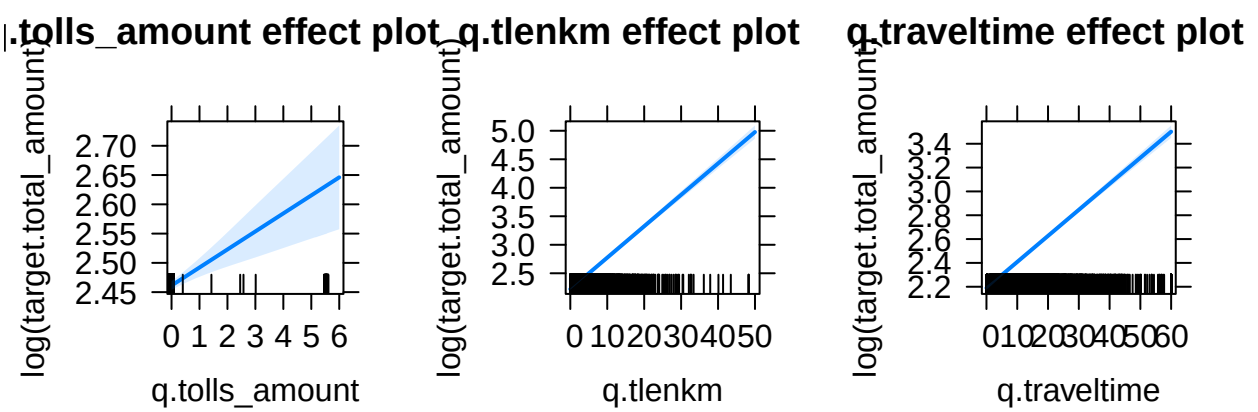
```
library(effects)
plot(allEffects(model_5))
```






```
plot(allEffects(model_6))
```

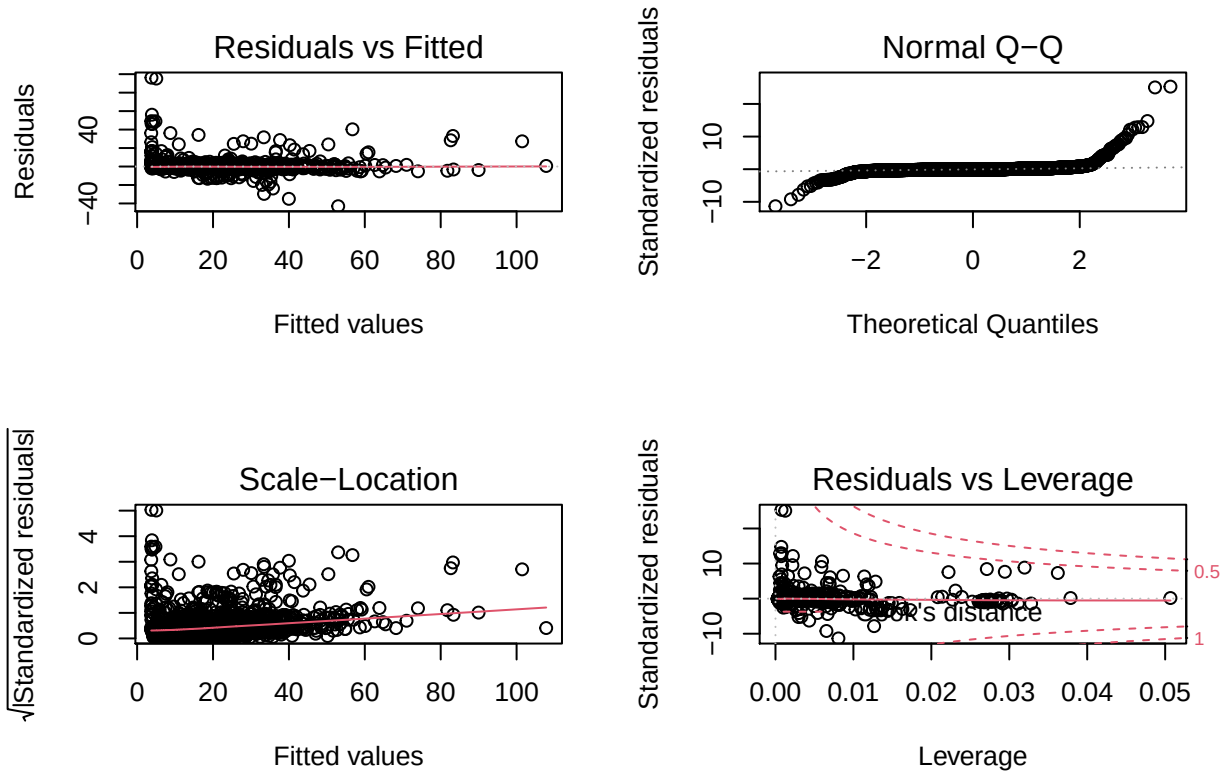




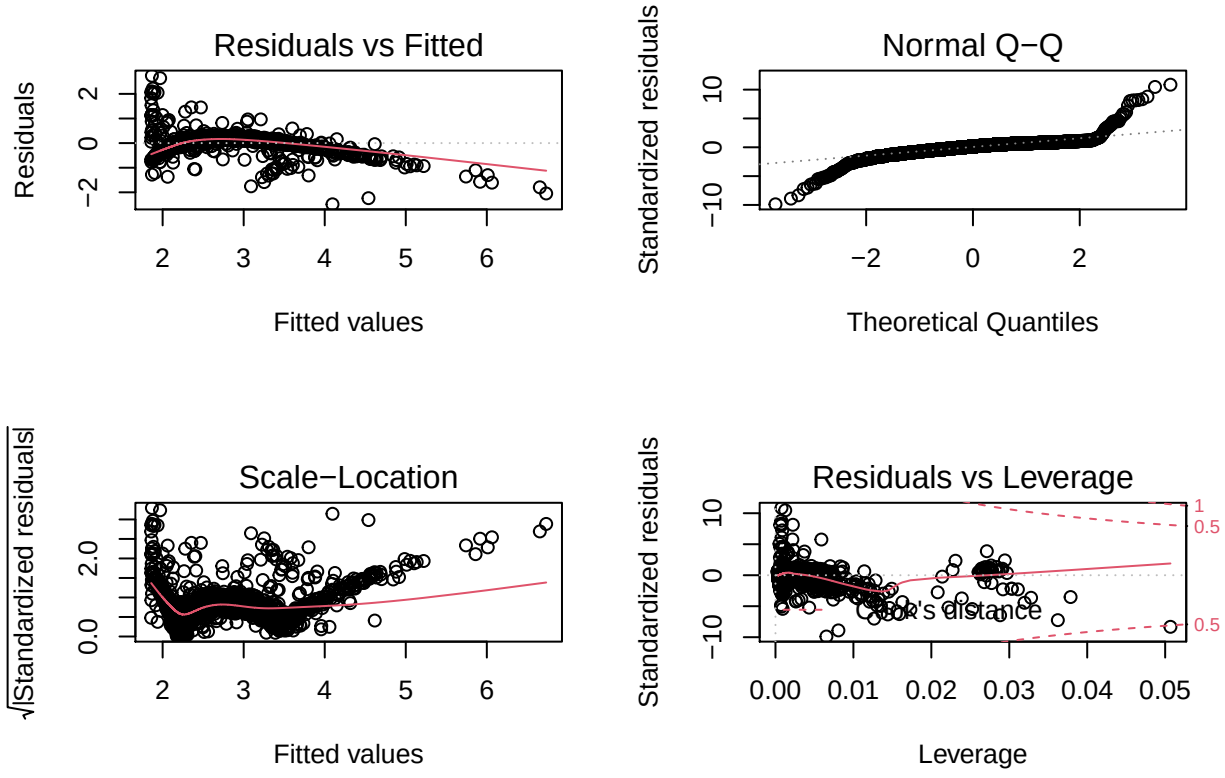
We see that our model defines the following:

- q.passenger_count does not depend on target.total_amount
- q.extra grows if target.total_amount grows
- q.tip_amount grows if target.total_amount grows
- q.tolls_amount grows if target.total_amount grows
- q.tlenkm grows if target.total_amount grows
- q.traveltime grows if target.total_amount grows

```
par(mfrow=c(2,2))
plot(model_5, id.n=0 )
```




```
plot(model_6, id.n=0 )
```




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

We see that the waste is not yet entirely optimal.

```
# summary(resid(m24))
# sel1<-Boxplot(rstudent(m24));sel1 # sel1 already contains row numbers
# # ll1<-which(row.names(df) %in% names(rstudent(m24)[sel1]));ll1
# sel2<-which(hatvalues(m24)>6*length(m24$coefficients)/nrow(df));sel2;length(sel2) # sel2 contains row
```

```
# ll2<-which(row.names(df) %in% names(hatvalues(m24)[sel2]));ll2
# sel3<-which(abs(cooks.distance(m24))>4/(nrow(df)-length(m24$coefficients)));sel3;length(sel3)
# ll3<-which(row.names(df) %in% names(cooks.distance(m24)[sel3]));ll3
# # sel4<-Boxplot(cooks.distance(m24));sel4  # sel4 already contains row numbers
# sel3<-which((cooks.distance(m24))>0.1);sel3;length(sel3)# sel3 contains row names
# ll3<-which(row.names(df) %in% names(cooks.distance(m24)[sel3]));ll3
#
# df[sel3,1:23]
#
# influencePlot(m24,id=list(method="noteworthy", n=5, cex=0.5))
# with(df,tapply(Total_amount,RateCodeID,summary))
```

6.6 Multivariant Analysis conducted in previous deliverables has to be used to select the initial model. Students have some degrees in freedom in model building, but the following conditions are requested:

- At least two numerical variables have to be considered as explicative variables for initial steps in model building, called covariates. Non-linear models have to be checked for consistency.
- Select the most significant factors found in Multivariant Data Analysis as initial model factors. Put some reasonable limits to initial model complexity.
- **You have to consider at least one interaction between a couple of factors and one interaction between factor and covariate.**
- Diagnostics of the final model have to be undertaken. Lack of fit observations and influence data have to be selected and discussed (connections to multidimensional outliers in Multivariant Data Analysis is highly valuable)

6.7 Outcome/Target : A binary response variable (Binary Target) will be the response variable for Binary Regression Models included in Statistical Modeling Part III.

- Explicative Variables for modeling purposes are those available in dataset, exceptions will be indicated, if any.
- Multivariant Analysis conducted in previous deliverables has to be used to select the initial model. Students have some degrees in freedom in model building, but the following conditions are requested:
  - Split the sample in work and test samples (consisting on a 80-20 split). Working data frame has to be used for model building purposes.
  - At least two numerical variables have to be considered as explicative variables for initial steps in model building.
  - Select the most significant factors according to feature selection as initial model factors. Put some reasonable limits to initial model complexity.
  - **You have to consider at least one interaction between a couple of factors and one interaction between factor and covariate.**
  - Diagnostics of the final model have to be undertaken. Lack of fit observations and influence data have to be selected and discussed (connections to multidimensional outliers in Multivariant Data Analysis is highly valuable).
  - You have to predict Y (Binary Target) in the Working Data Frame vs the rest according to the best validated model that you can find and make a confusion matrix.
  - Make a confusion matrix in the Testing Data Frame for **Y (Binary Target)** according to the best validated model found.

6.8 Confusion Matrix:

When referring to the performance of a classification model, we are interested in the model's ability to correctly predict or separate the classes. When looking at the errors made by a classification model, the confusion matrix gives the full picture. Consider e.g. a three class problem with the classes A, and B. The confusion matrix shows how the predictions are made by the model. The rows correspond to the known class of the data, i.e. the labels in the data. The columns correspond to the predictions made by the model. The value of each of element in the matrix is the number of predictions made with the class corresponding to the column for examples with the correct value as represented by the row. Thus, the diagonal elements show the number of correct classifications made for each class, and the off-diagonal elements show the errors made.