Estadística avanzada para ciencia de datos

Regresión - Ejercicios

Jakub Maciążek

Exercise I

ToDo:

- Import ToyotaCorolla dataset from the course website.
- Build linear models m1, m2 using Multiple Linear Regression.
- Analyze the goodness of fit and model quality, explaining the results.

Data import and analysys

```
#water_df <- read.csv("S:/0_Universidad_de_Malaga/MI_Ingenieria_y_ciencia_de_datos/Estatistica_avanzada_

#head(water_df)

ToyotaCorolla <- read.csv("S:/0_Universidad_de_Malaga/MI_Ingenieria_y_ciencia_de_datos/Estatistica_avanz
head(ToyotaCorolla)
##   Price Age    KM FuelType HP MetColor Automatic   CC Doors Weight
## 1 13500  23 46986   Diesel 90        1         0 2000     3   1165
## 2 13750  23 72937   Diesel 90        1         0 2000     3   1165
## 3 13950  24 41711   Diesel 90        1         0 2000     3   1165
## 4 14950  26 48000   Diesel 90        0         0 2000     3   1165
## 5 13750  30 38500   Diesel 90        0         0 2000     3   1170
## 6 12950  32 61000   Diesel 90        0         0 2000     3   1170
```

From the analysis of the data set, it can be seen it contains both continuous and categorical variables:

Continuous:

- price
- age
- KM (run kilometers)
- weight

Categorical:

- FuelType (Diesel, Petrol)
- MetColor (True, False)
- Automatic (True, False)
- Doors (Nr of doors)

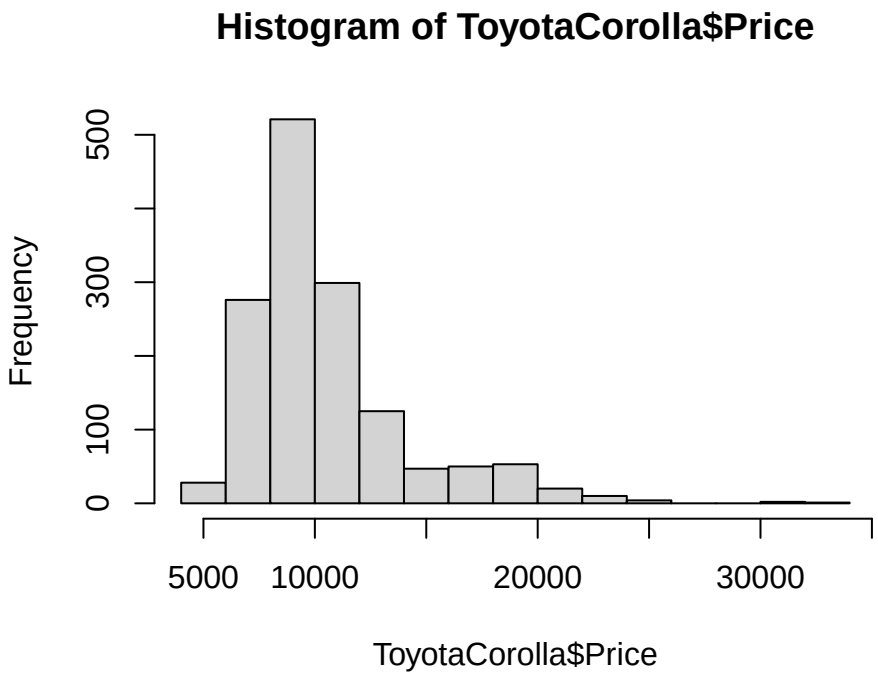
Variables that theoretically could be continuous, but in practice come from a short range of values, and therefore should be considered as categorical.

- HP (Horse power)
- CC (cylinder capacity)

Regression requires numerical inputs, therefore categorical variables need to be converted into binary or numerical values. In this case all of the variables, except the FuelType, meet this criterion. Therefore FuelType will be converted under the hood by R to dummy binary variables.

Dependent variable data normality - For given example, it makes sens to build a model predicting car price based on other variables. This makes "price" variable a dependent one, which requires it to be normally distributed and continuous for linear model to work.

```
shapiro.test(ToyotaCorolla$Price)
##
##  Shapiro-Wilk normality test
##
## data:  ToyotaCorolla$Price
## W = 0.85228, p-value < 2.2e-16
hist(ToyotaCorolla$Price)
```



To test for normality, histogram graph was presented and Shapiro-Wilk test was conducted, which resulted in p-value smaller than 2.2e-16.

From the histogram it can be seen, that data is highly positively skewed. Also, the result of the test is smaller than 0.05, meaning that distribution of variable differs from Normal distribution. Therefore, to receive more accurate prediction it would be useful to either use Generalized Linear Model from for exp. "gamma" family, or to transform the price with logarithmic function, before using standard linear model.

However, knowing the model might be inaccurate due to above reasons, standard linear regression was conducted, as required by the Task.

Build of linear model

```
#m1 <- lm(Price ~ Age + KM + FuelType + HP + MetColor + Automatic + Doors + Weight , data = ToyotaCorol
#summary(m1)


m1 <- lm(Price ~. , data = ToyotaCorolla)

summary(m1)
##
## Call:
## lm(formula = Price ~ ., data = ToyotaCorolla)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10642.3   -737.7      3.1    731.3   6451.5
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3.801e+03  1.304e+03  -2.915 0.003613 **
## Age            -1.220e+02  2.602e+00 -46.889  < 2e-16 ***
## KM             -1.621e-02  1.313e-03 -12.347  < 2e-16 ***
## FuelTypeDiesel  3.390e+03  5.188e+02   6.535 8.86e-11 ***
## FuelTypePetrol  1.121e+03  3.324e+02   3.372 0.000767 ***
## HP              6.081e+01  5.756e+00  10.565  < 2e-16 ***
## MetColor        5.716e+01  7.494e+01   0.763 0.445738
## Automatic       3.303e+02  1.571e+02   2.102 0.035708 *
## CC             -4.174e+00  5.453e-01  -7.656 3.53e-14 ***
## Doors          -7.776e+00  4.006e+01  -0.194 0.846129
## Weight          2.001e+01  1.203e+00  16.629  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1316 on 1425 degrees of freedom
## Multiple R-squared:  0.8693, Adjusted R-squared:  0.8684
## F-statistic:   948 on 10 and 1425 DF,  p-value: < 2.2e-16
```

Starting analysis with examination of the **F-statistic** and its corresponding **p-value**. It can be observed that its value equals 948 and is far greater than 1, and p-value is essentialy 0, which means that at least one of the predictor variables is significantly related to the outcome variable. To check which predictors are significant, coefficient table was examined.

```
summary(m1)$coefficient
##                     Estimate   Std. Error     t value      Pr(>|t|)
## (Intercept)    -3.801361e+03 1.304082e+03  -2.9149719  3.612782e-03
## Age            -1.220145e+02 2.602185e+00 -46.8892412 4.353493e-291
## KM             -1.620832e-02 1.312771e-03 -12.3466468  2.406197e-33
## FuelTypeDiesel  3.390077e+03 5.187954e+02   6.5345159  8.860931e-11
## FuelTypePetrol  1.120676e+03 3.323653e+02   3.3718209  7.667291e-04
```

```
## HP              6.081328e+01 5.755864e+00  10.5654473  3.575005e-25
## MetColor        5.715977e+01 7.493902e+01   0.7627505  4.457384e-01
## Automatic       3.302509e+02 1.570956e+02   2.1022288  3.570833e-02
## CC             -4.174372e+00 5.452599e-01  -7.6557477  3.525245e-14
## Doors          -7.776268e+00 4.006426e+01  -0.1940949  8.461293e-01
## Weight          2.000936e+01 1.203309e+00  16.6286120  6.939602e-57
```

Based on the **t-statistic**, it can be seen, that some variables are insignificant, as their Pr(>|t|) is higher than 0.05, meaning there is no significant association between the predictor and the outcome variable.

In order to make a better model, new one was calculated, involving only significant variables (meaning all except MetColor and Doors).

```
m2 <- lm(Price ~. -Doors -MetColor  , data = ToyotaCorolla)

summary(m2)
##
## Call:
## lm(formula = Price ~ . - Doors - MetColor, data = ToyotaCorolla)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10593.9   -726.9     -2.3    720.1   6459.4
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3.718e+03  1.261e+03  -2.948  0.00325 **
## Age            -1.221e+02  2.596e+00 -47.041  < 2e-16 ***
## KM             -1.625e-02  1.309e-03 -12.416  < 2e-16 ***
## FuelTypeDiesel  3.388e+03  5.090e+02   6.655 4.03e-11 ***
## FuelTypePetrol  1.112e+03  3.317e+02   3.353  0.00082 ***
## HP              6.089e+01  5.639e+00  10.799  < 2e-16 ***
## Automatic       3.305e+02  1.562e+02   2.116  0.03452 *
## CC             -4.168e+00  5.369e-01  -7.763 1.57e-14 ***
## Weight          1.994e+01  1.126e+00  17.709  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1315 on 1427 degrees of freedom
## Multiple R-squared:  0.8693, Adjusted R-squared:  0.8685
## F-statistic:  1186 on 8 and 1427 DF,  p-value: < 2.2e-16
```

For new model all the predictors are significant, and received model has a following equation:

$Price =$
$-3718 - 122.1 * Age - 0.01625 * KM + 3388 * FuelTypeDiesel + 1112 * FuelTypePetrol$
$+ 60.89 * HP + 330.5 * Automatic - 4.168 * CC + 19.94 * Weight$

Based on DF information, it can be said that it was computed based on 1427 independent pieces of information.

Model goodness and results analysys

Example model explanation

For the second model m2, following equation was obtained: Price = -3718 -122.1 * Age - 0.01625 * KM + 3388 * FuelTypeDiesel + 1112 * FuelTypePetrol + 60.89 * HP + 330.5 * Automatic - 4.168 * CC + 19.94 * Weight. This means that:

- the value of intercept equals -3718. That is the estimated value of a car that would have value 0 for all of the parameters. In this case it is purely theoretical and has no meaningful interpretation.
- For each year of car use, its value decreases by 122.1. It also decreases by ~0.016 for every driven kilometer. For new cars, price would be lower by 4.168 for every cubic centimetres capacity of the engine.
- Value of the car increases by 3388 for Diesels, and by 1112 for Petrol ones, meaning Diesel ones are more expensive by default. Moreover value of the car increases by 60.89 for every HP of engine, and by 19.94 for every kg of weight. Also, car that is Automatic increases value by 330.5.

Goodness of fit - Adj. R2

To asses model accuracy, R-squared value can be used. However, it will always increase when more variables are added to the model, even if those variables are only weakly associated with the response.

In this case the models to compare have different number of predictors, therefore another measure, Adjusted R Square, is used to compare their goodness, as it is adjusted to the number of used parameters.

For model m1, value of Adjusted R Squared equals 0.8684, which means that the model explains ~87% of variance when predicting value of the car. Model m2 is slightly better, as for the same parameter has value 0.8685 while using less predictors.

Large proportion of the variability in the response has been explained by the regression.

Goodness of fit - RSE

Other measure of the quality of the model is RSE (Residual Standard Error). It is the absolute measure of standard deviation of residual, meaning the average distance between prediction and actual value. Divided by mean value of predicted variable, error rate (relative error) is received.

For m1 and m2 those are respectively 0.1226107 and 0.1225507, which means that both models have ~13% prediction error rate, with m2 being slightly more accurate again.

```
rse1 <- sigma(m1)/mean(ToyotaCorolla$Price)
rse1
## [1] 0.1226107

rse2 <- sigma(m2)/mean(ToyotaCorolla$Price)
rse2
## [1] 0.1225507
```

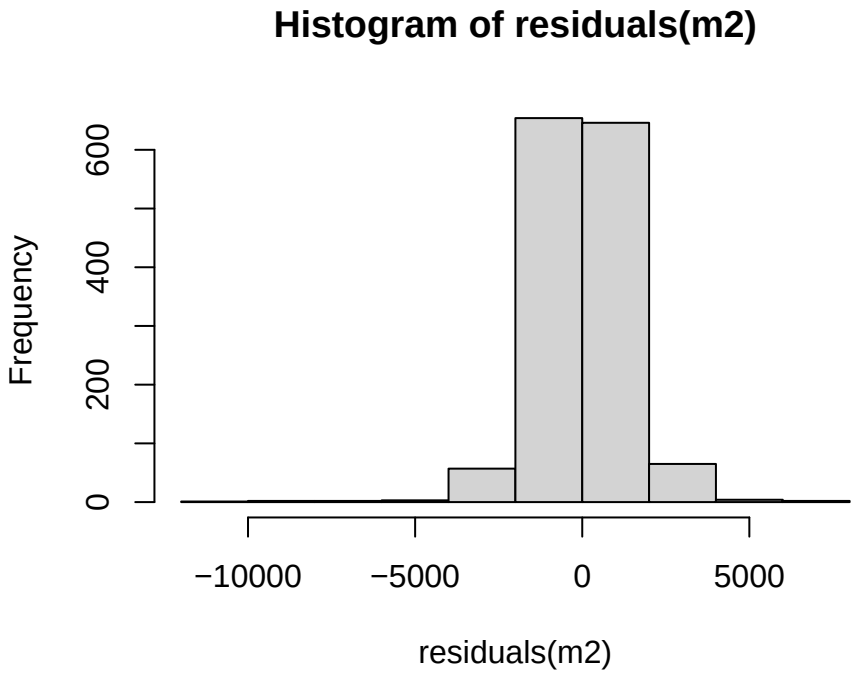
Model quality - residuals analysis

In order for the model not to be biased, residual need to have a zero average. In case of model m2, the average is equal to 6.925008e-14, which is essentialy zero.

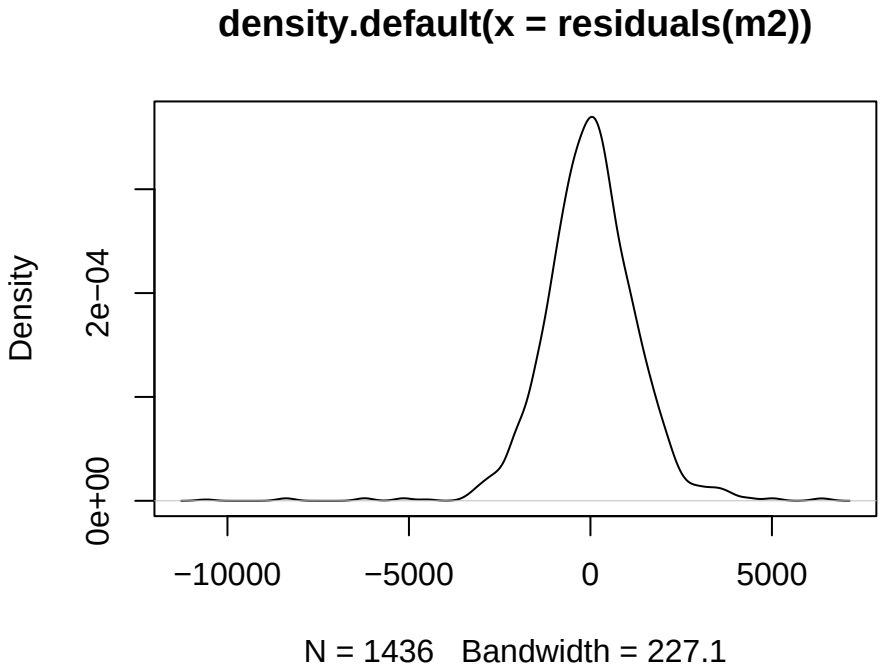
```
mean(residuals(m2))
## [1] 6.925008e-14
```

Errors must also be normally distributed. As can be seen by the histogram and density plot below, they are normally-like distributed.

```
hist(residuals(m2))
```




```
plot(density(residuals(m2)))
```

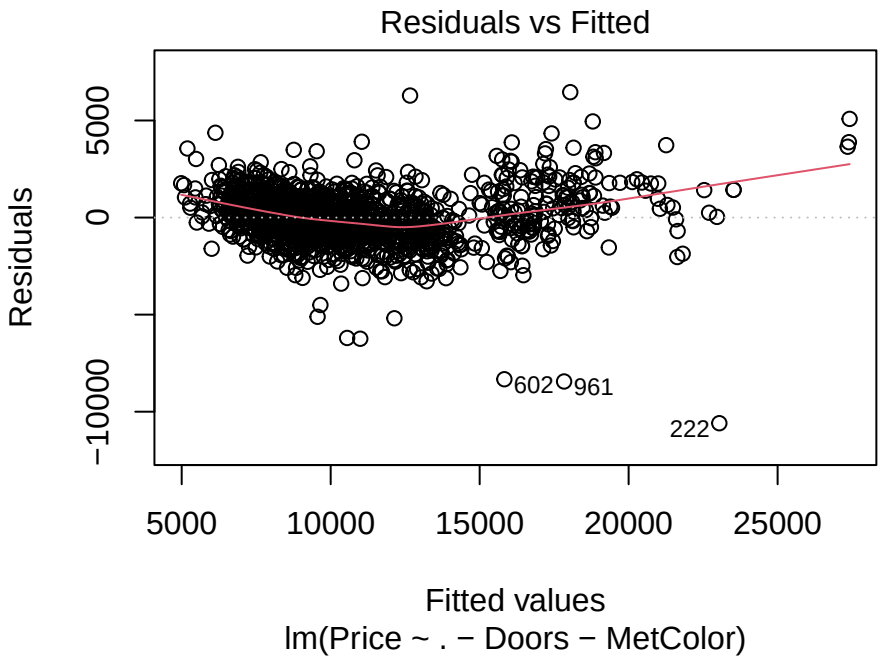


Model quality - plots analysis: visual verification

1) Residuals vs fitted values

On the plot below, small number of outliers can be seen, however for most of the fitted values, their residuals oscillate symmetrically around the curve close to 0.

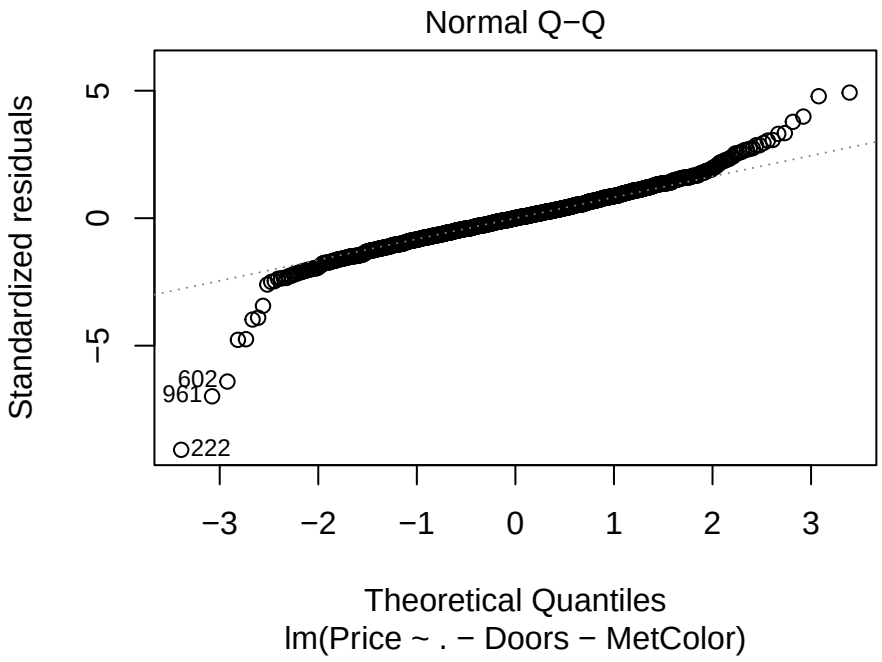
```
plot(m2, which = 1)
```



2) Normal Q-Q

This plot illustrates whether errors are normally distributed (in which case they should form a straight line). In this case it is true for most of the values except a couple of outliers. Shape of the curve suggests that errors distributions has *fatter tails*, meaning compared to the normal distribution there is more data located at the extremes of the distribution and less data in the center of the distribution. In terms of quantiles this means that the first quantile is much less than the first theoretical quantile and the last quantile is greater than the last theoretical quantile.

```
plot(m2, which = 2)
```

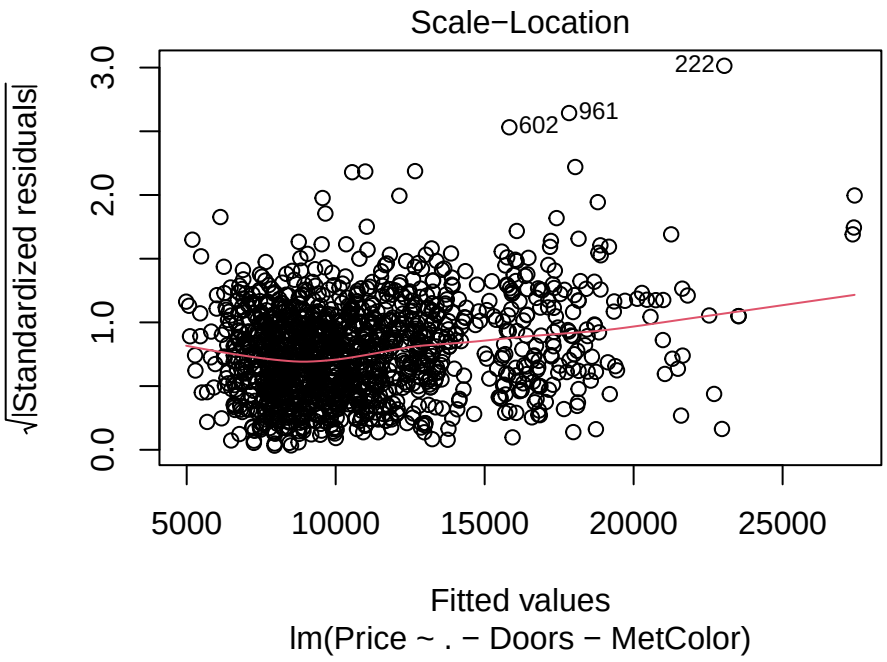


3) Scale-Location

Based on the scale-location plot, we can verify that red line is roughly horizontal, meaning the assumption of homoscedasticity is likely satisfied for a given regression model (spread od residuals is roughly equal for all fitted values).

There is no pattern among residuals, they are randomly scattered around the red line with roughly equal variability at all fitted values. (Although more points can be seen to the left of the plot, it corresponds to the skew in Price data, not residuals themselves.)

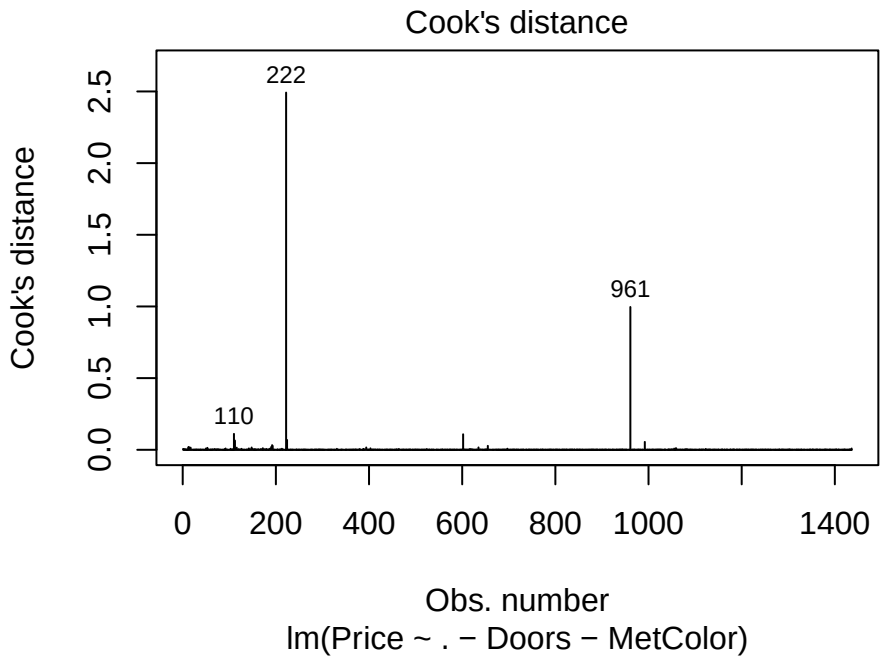
```
plot(m2, which = 3)
```



4) Cook's distance

The last plot shows which points have the greatest influence on the regression (leverage points). In this case those are point 110, 222 and 961.

```
plot(m2, which = 4)
```

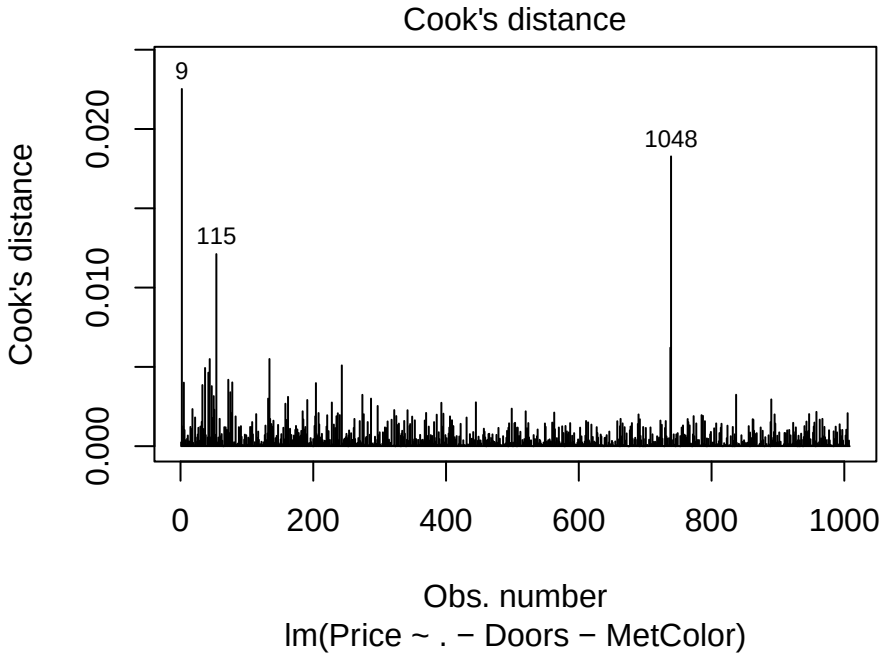


They have great influence on the model, can be detected as outliers, therefore removing these points should increase quality of the model.

```
# https://www.statology.org/how-to-identify-influential-data-points-using-cooks-distance/

#find Cook's distance for each observation in the dataset
cooksD <- cooks.distance(m2)

#identify influential points (traditional threshold 4/n, here value from the graph used -> biggest 3 ou
influential_obs <- as.numeric(names(cooksD)[(cooksD >= cooksD[109])])

#define new data frame with influential points removed
outliers_removed <- ToyotaCorolla[-influential_obs, ]

head(outliers_removed)
##    Price Age    KM FuelType  HP MetColor Automatic   CC Doors Weight
## 7  16900  27 94612   Diesel  90        1         0 2000     3   1245
## 9  21500  27 19700   Petrol 192        0         0 1800     3   1185
## 22 16950  29 43905   Petrol 110        0         1 1600     3   1170
## 23 15950  28 56349   Petrol 110        1         0 1600     3   1120
## 24 16950  28 32220   Petrol 110        1         0 1600     3   1120
## 25 16250  29 25813   Petrol 110        1         0 1600     3   1120

m2_or <- lm(Price ~. -Doors -MetColor , data = outliers_removed)

summary(m2_or)
##
## Call:
## lm(formula = Price ~ . - Doors - MetColor, data = outliers_removed)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1785.39  -479.02    -3.58   469.35  1917.79
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -5.616e+03  1.186e+03  -4.734 2.52e-06 ***
## Age            -1.154e+02  1.823e+00 -63.297  < 2e-16 ***
## KM             -1.535e-02  8.923e-04 -17.204  < 2e-16 ***
## FuelTypeDiesel  3.635e+03  6.195e+02   5.868 6.00e-09 ***
## FuelTypePetrol  1.313e+03  4.010e+02   3.274 0.001095 **
## HP              5.943e+01  7.401e+00   8.031 2.71e-15 ***
## Automatic       4.284e+02  1.247e+02   3.435 0.000617 ***
## CC             -4.693e+00  5.732e-01  -8.187 8.14e-16 ***
## Weight          2.194e+01  1.084e+00  20.238  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 679 on 999 degrees of freedom
## Multiple R-squared:  0.9371, Adjusted R-squared:  0.9366
## F-statistic:  1860 on 8 and 999 DF,  p-value: < 2.2e-16
```

```
plot(m2_or, which = 4)
```




```
rse2_or <- sigma(m2_or)/mean(outliers_removed$Price)
rse2_or
## [1] 0.06642365
```

As can be see, removing that 3 values had increased Adjusted R Squared from 0.8685 to 0.9366, which means model explains now about 6% more od the variance in predicted values. Also error rate has decreased from ~13% to ~7%.

This is a huge improvement, that as can be seen by new Cook's distance graph might be even increased by removing more outliers. However for now, this one will be used in further analysis replacing m2.

```
m2 <- m2_or
```

Exercise II

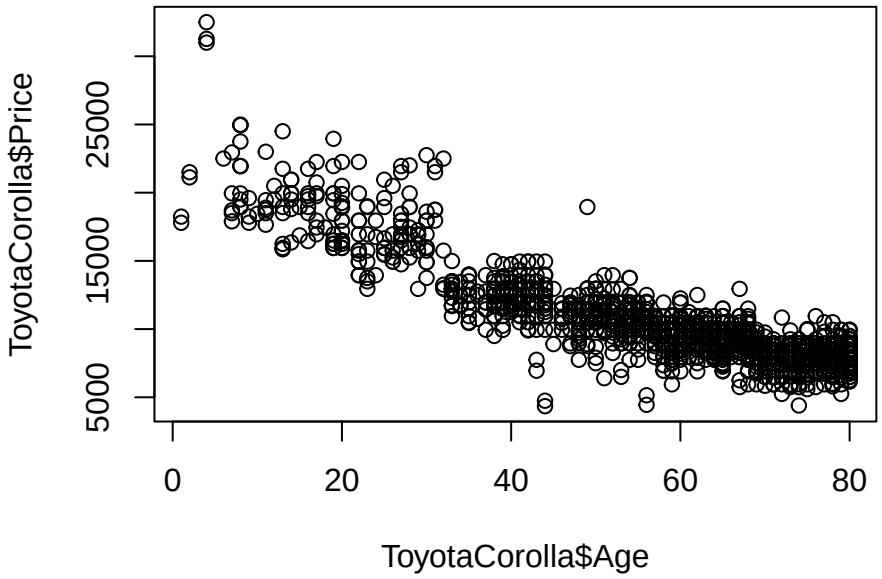
ToDo:

- With the ToyotaCorolla dataset, build two polynomial and orthogonal models, and call them m3 and m4.
- Repeat the analysis made for m1 and m2. Also, analyze the significancy of the regression coefficients.

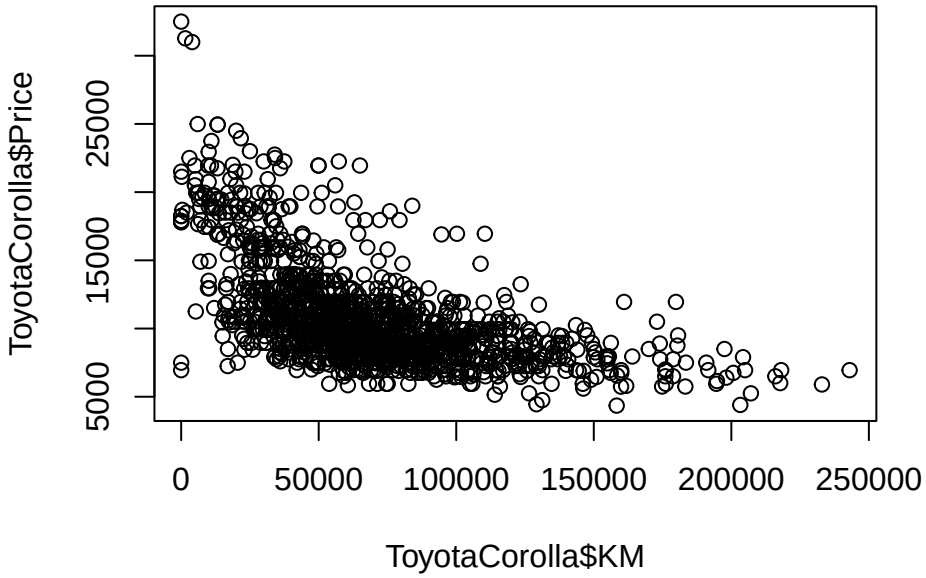
Selection of nonlinear dependent variables

In order to find non-linear dependent variables, multiple scatter plots were generated.

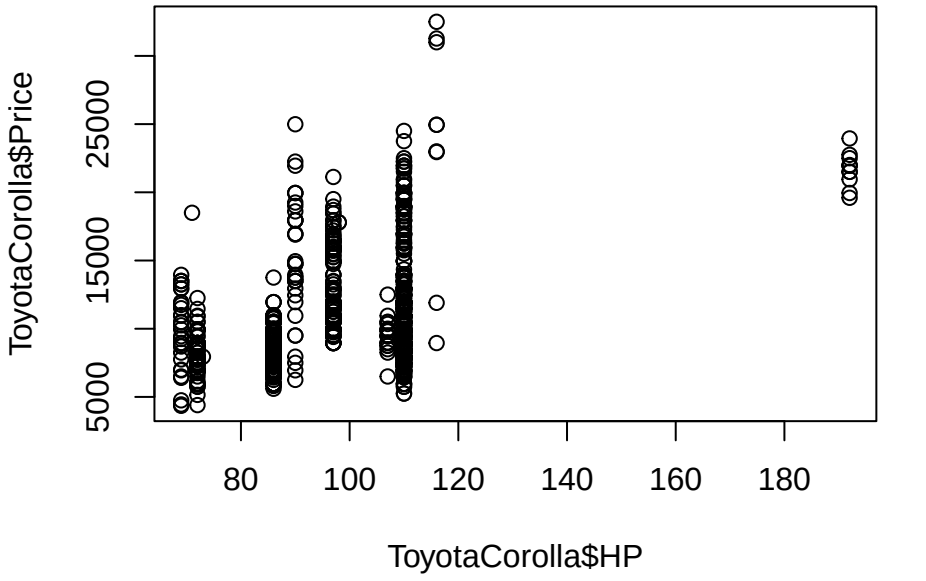
```
plot(ToyotaCorolla$Age, ToyotaCorolla$Price)
```



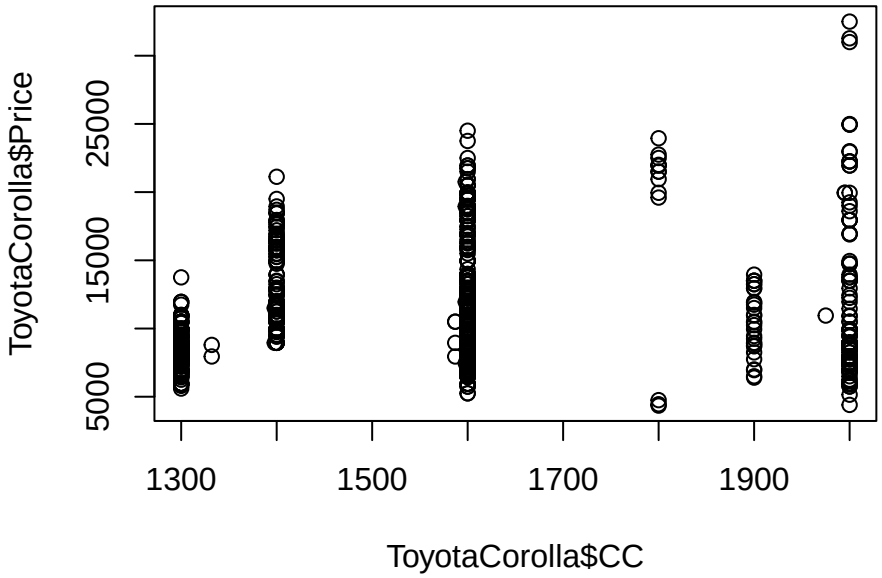
```
plot(ToyotaCorolla$KM, ToyotaCorolla$Price)
```



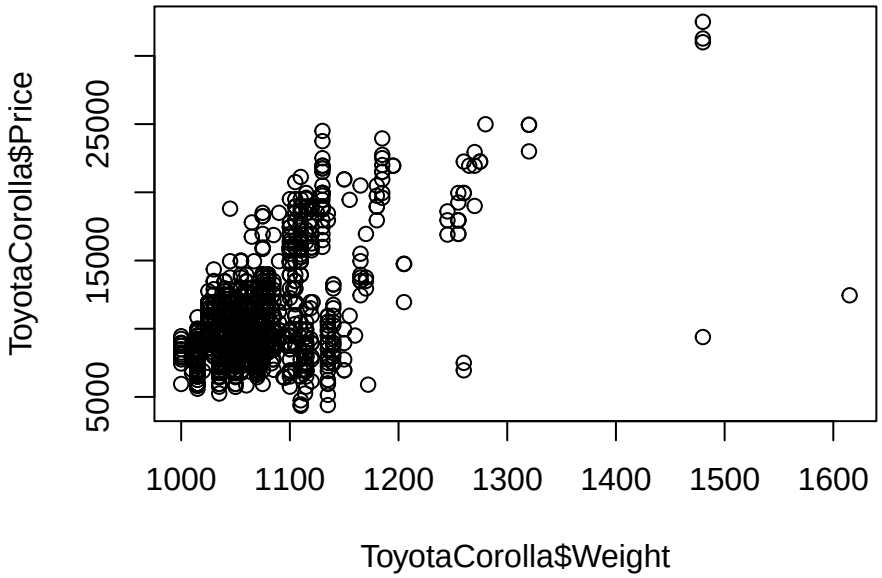

```
plot(ToyotaCorolla$HP, ToyotaCorolla$Price)
```



```
plot(ToyotaCorolla$CC, ToyotaCorolla$Price)
```




```
plot(ToyotaCorolla$Weight, ToyotaCorolla$Price)
```



Based on this plots following relationships can be observed:

- Age has rather linear influence on the price, slightly exponential
- KM have exponential effect on Price
- Both HP and CC are more categorical, and for each category Price comes from specific range. To fit a curve through those lines polynomial could be used. For HP of 3rd degree and for CC of 3rd or 4th degree.
- For weight, it is hard to observe any pattern, especially due to outliers of very hight weight.

Based on above observations, polynomial aspect will be added to Age, 3rd degree to HP and 3rd to CC.

Polynomial model creation and analysis

After examining multiple models, best one was recived for removing of Weigt term and addition of terms: I(CC^2), I(CC^3), I(Age^2), I(Weight^2).

```
m3 <- update(m2,  ~. +I(KM^2) +I(KM^3))

summary(m3)
##
## Call:
## lm(formula = Price ~ Age + KM + FuelType + HP + Automatic + CC +
##     Weight + I(KM^2) + I(KM^3), data = outliers_removed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1749.48  -493.75    -6.98   452.71  1950.19
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -5.072e+03  1.195e+03  -4.243 2.41e-05 ***
## Age            -1.145e+02  1.932e+00 -59.273  < 2e-16 ***
## KM             -2.969e-02  5.474e-03  -5.424 7.32e-08 ***
## FuelTypeDiesel  3.425e+03  6.285e+02   5.449 6.39e-08 ***
## FuelTypePetrol  1.155e+03  4.089e+02   2.825 0.004824 **
## HP              5.820e+01  7.439e+00   7.823 1.31e-14 ***
## Automatic       4.330e+02  1.243e+02   3.484 0.000515 ***
## CC             -4.562e+00  5.753e-01  -7.929 5.88e-15 ***
## Weight          2.177e+01  1.082e+00  20.111  < 2e-16 ***
## I(KM^2)         1.756e-07  6.026e-08   2.915 0.003640 **
## I(KM^3)        -6.077e-13  1.996e-13  -3.044 0.002392 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 676.5 on 997 degrees of freedom
## Multiple R-squared:  0.9377, Adjusted R-squared:  0.9371
## F-statistic:  1500 on 10 and 997 DF,  p-value: < 2.2e-16
rse3 <- sigma(m3)/mean(outliers_removed$Price)
rse3
## [1] 0.0661817
```

```
m3 <- update(m2,  ~. +I(Age^2))

summary(m3)
##
## Call:
## lm(formula = Price ~ Age + KM + FuelType + HP + Automatic + CC +
##     Weight + I(Age^2), data = outliers_removed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1712.15  -496.96   -22.55   452.48  1976.96
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3.201e+03  1.318e+03  -2.429 0.015327 *
## Age            -1.458e+02  7.663e+00 -19.019  < 2e-16 ***
## KM             -1.517e-02  8.865e-04 -17.115  < 2e-16 ***
## FuelTypeDiesel  3.611e+03  6.148e+02   5.873 5.81e-09 ***
## FuelTypePetrol  1.246e+03  3.983e+02   3.129 0.001807 **
## HP              5.937e+01  7.344e+00   8.085 1.79e-15 ***
## Automatic       4.523e+02  1.239e+02   3.651 0.000275 ***
## CC             -4.446e+00  5.720e-01  -7.773 1.90e-14 ***
## Weight          2.006e+01  1.171e+00  17.126  < 2e-16 ***
## I(Age^2)        2.825e-01  6.934e-02   4.074 4.99e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 673.7 on 998 degrees of freedom
## Multiple R-squared:  0.9381, Adjusted R-squared:  0.9376
## F-statistic:  1681 on 9 and 998 DF,  p-value: < 2.2e-16
rse3 <- sigma(m3)/mean(outliers_removed$Price)
rse3
## [1] 0.06591118
```

```
m3 <- update(m2,  ~. +I(CC^2) +I(CC^3))

summary(m3)
##
## Call:
## lm(formula = Price ~ Age + KM + FuelType + HP + Automatic + CC +
##     Weight + I(CC^2) + I(CC^3), data = outliers_removed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1819.15  -468.26   -16.91   455.41  1890.13
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.036e+05  2.415e+04   4.289 1.97e-05 ***
## Age            -1.182e+02  1.943e+00 -60.826  < 2e-16 ***
## KM             -1.525e-02  8.854e-04 -17.223  < 2e-16 ***
## FuelTypeDiesel  3.255e+03  8.439e+02   3.858 0.000122 ***
## FuelTypePetrol  1.351e+03  3.975e+02   3.399 0.000702 ***
```

```
## HP              5.889e+01  7.438e+00   7.918 6.41e-15 ***
## Automatic       3.719e+02  1.243e+02   2.993 0.002830 **
## CC             -2.100e+02  4.554e+01  -4.612 4.51e-06 ***
## Weight          2.238e+01  1.083e+00  20.676  < 2e-16 ***
## I(CC^2)         1.264e-01  2.823e-02   4.477 8.43e-06 ***
## I(CC^3)        -2.553e-05  5.780e-06  -4.417 1.11e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 672.8 on 997 degrees of freedom
## Multiple R-squared:  0.9384, Adjusted R-squared:  0.9378
## F-statistic:  1518 on 10 and 997 DF,  p-value: < 2.2e-16
rse3 <- sigma(m3)/mean(outliers_removed$Price)
rse3
## [1] 0.06581593
```

```
m3 <- update(m2,  ~. +I(CC^2) +I(CC^3) + I(Age^2))

summary(m3)
##
## Call:
## lm(formula = Price ~ Age + KM + FuelType + HP + Automatic + CC +
##     Weight + I(CC^2) + I(CC^3) + I(Age^2), data = outliers_removed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1728.02  -470.72   -23.91   452.42  1950.22
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.071e+05  2.397e+04   4.466 8.88e-06 ***
## Age            -1.489e+02  7.622e+00 -19.530  < 2e-16 ***
## KM             -1.506e-02  8.794e-04 -17.129  < 2e-16 ***
## FuelTypeDiesel  3.296e+03  8.371e+02   3.938 8.81e-05 ***
## FuelTypePetrol  1.285e+03  3.946e+02   3.255  0.00117 **
## HP              5.898e+01  7.378e+00   7.993 3.62e-15 ***
## Automatic       3.960e+02  1.234e+02   3.209  0.00137 **
## CC             -2.120e+02  4.517e+01  -4.693 3.07e-06 ***
## Weight          2.049e+01  1.166e+00  17.570  < 2e-16 ***
## I(CC^2)         1.280e-01  2.801e-02   4.569 5.51e-06 ***
## I(CC^3)        -2.589e-05  5.734e-06  -4.516 7.06e-06 ***
## I(Age^2)        2.856e-01  6.871e-02   4.156 3.51e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 667.3 on 996 degrees of freedom
## Multiple R-squared:  0.9394, Adjusted R-squared:  0.9388
## F-statistic:  1404 on 11 and 996 DF,  p-value: < 2.2e-16
rse3 <- sigma(m3)/mean(outliers_removed$Price)
rse3
## [1] 0.06528527
```

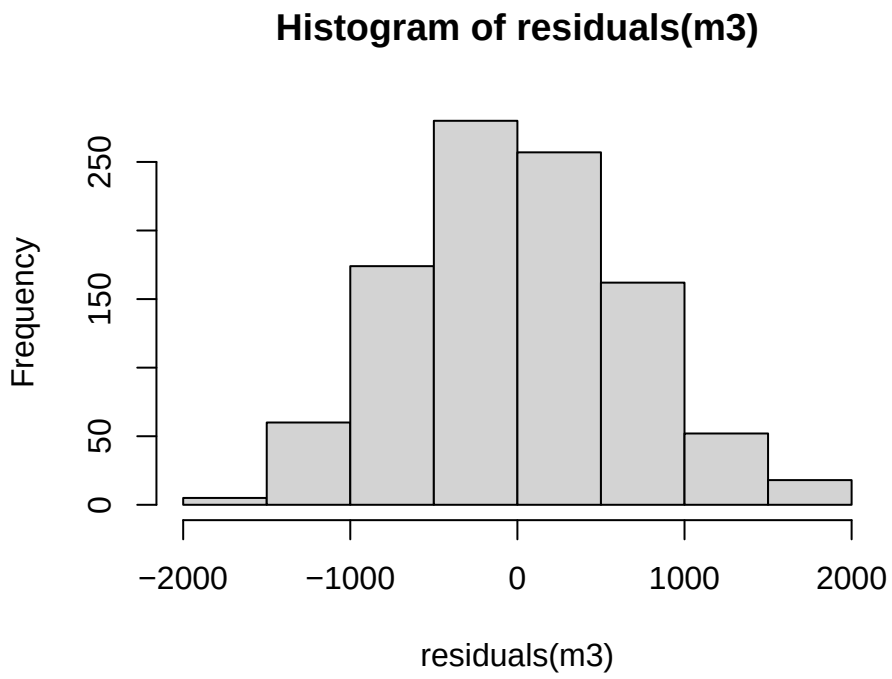
```
m3 <- update(m3,  ~. -Weight +I(Weight^2))

summary(m3)
##
## Call:
## lm(formula = Price ~ Age + KM + FuelType + HP + Automatic + CC +
##     I(CC^2) + I(CC^3) + I(Age^2) + I(Weight^2), data = outliers_removed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1726.22  -470.42   -25.75   446.42  1947.97
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.144e+05  2.396e+04   4.773 2.09e-06 ***
## Age            -1.478e+02  7.634e+00 -19.367  < 2e-16 ***
## KM             -1.502e-02  8.783e-04 -17.101  < 2e-16 ***
## FuelTypeDiesel  3.034e+03  8.384e+02   3.619  0.00031 ***
## FuelTypePetrol  1.293e+03  3.942e+02   3.280  0.00107 **
## HP              5.367e+01  7.470e+00   7.185 1.31e-12 ***
## Automatic       3.980e+02  1.232e+02   3.231  0.00128 **
## CC             -2.056e+02  4.509e+01  -4.560 5.76e-06 ***
## I(CC^2)         1.248e-01  2.796e-02   4.465 8.91e-06 ***
## I(CC^3)        -2.539e-05  5.724e-06  -4.436 1.02e-05 ***
## I(Age^2)        2.763e-01  6.880e-02   4.015 6.38e-05 ***
## I(Weight^2)     9.501e-03  5.381e-04  17.656  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 666.6 on 996 degrees of freedom
## Multiple R-squared:  0.9396, Adjusted R-squared:  0.9389
## F-statistic:  1408 on 11 and 996 DF,  p-value: < 2.2e-16
rse3 <- sigma(m3)/mean(outliers_removed$Price)
rse3
## [1] 0.0652092
```

Model m3 is based on m2, meaning data with removed outliers, and it's formula is as follows:

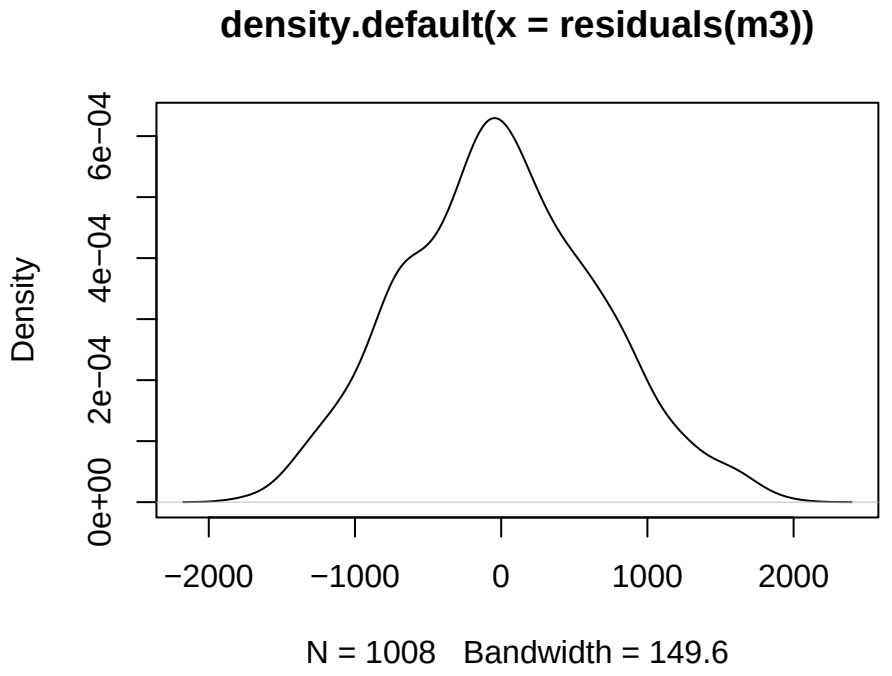
$$Price = Age + KM + FuelType + HP + Automatic + CC + I(CC^2) + I(CC^3) + I(Age^2) + I(Weight^2)$$

Model has **Adj. R^2** value of 0.9389, meaning it predicts ~**94%** od variability in outcome data, and has only ~**6.5% error rate**.

```
mean(residuals(m3))
## [1] 1.298148e-14
hist(residuals(m3))
```




```
plot(density(residuals(m3)))
```

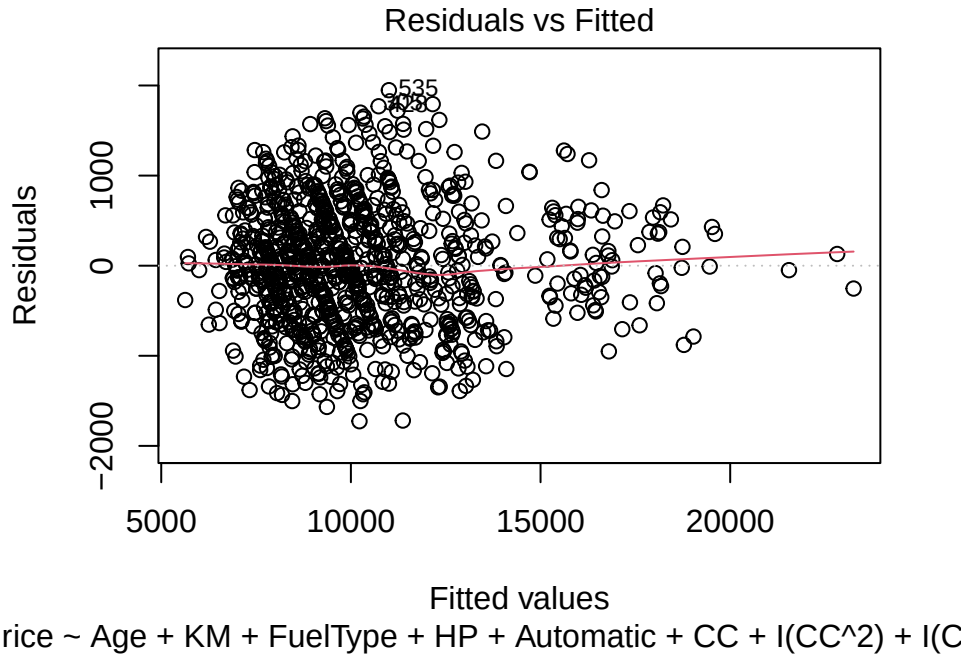


There is no bias in the model as errors mean equals 0 and they are normally distributed.

1) Residuals vs fitted values

In comparison to m2, number of outliers is smaller and residuals even better oscillate symmetrically around the 0 curve.

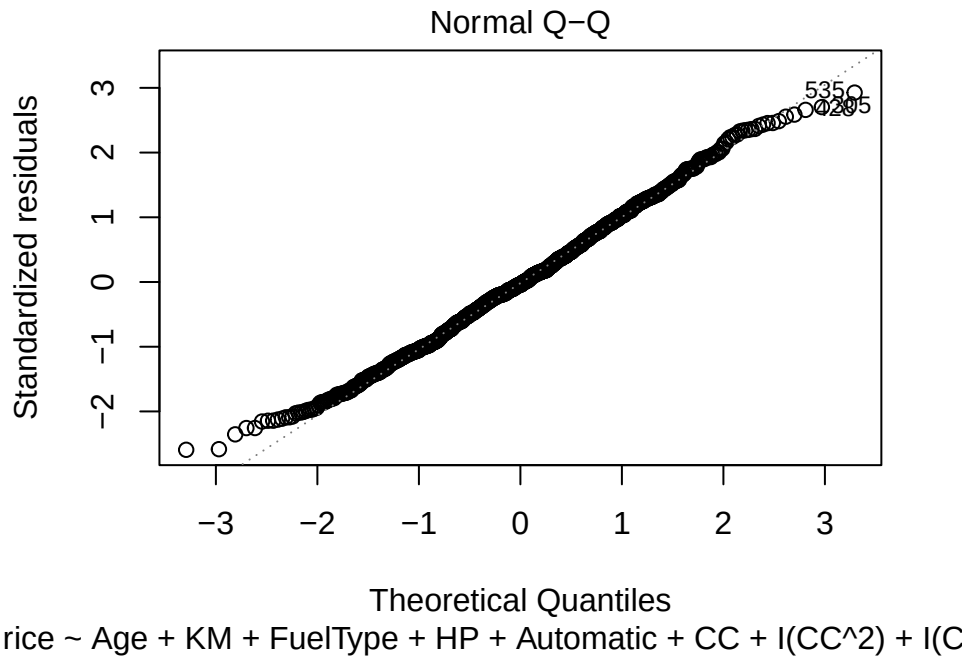
```
plot(m3, which = 1)
```



2) Normal Q-Q

Comparing to model m2, errors are better distributed, closer to normal distribution. There was a change in a way how they differ from gaussian distribution. Now distribution has *thin tails*, meaning the first quantiles are occurring at larger than expected values and the last quantiles are occurring at less than expected values.

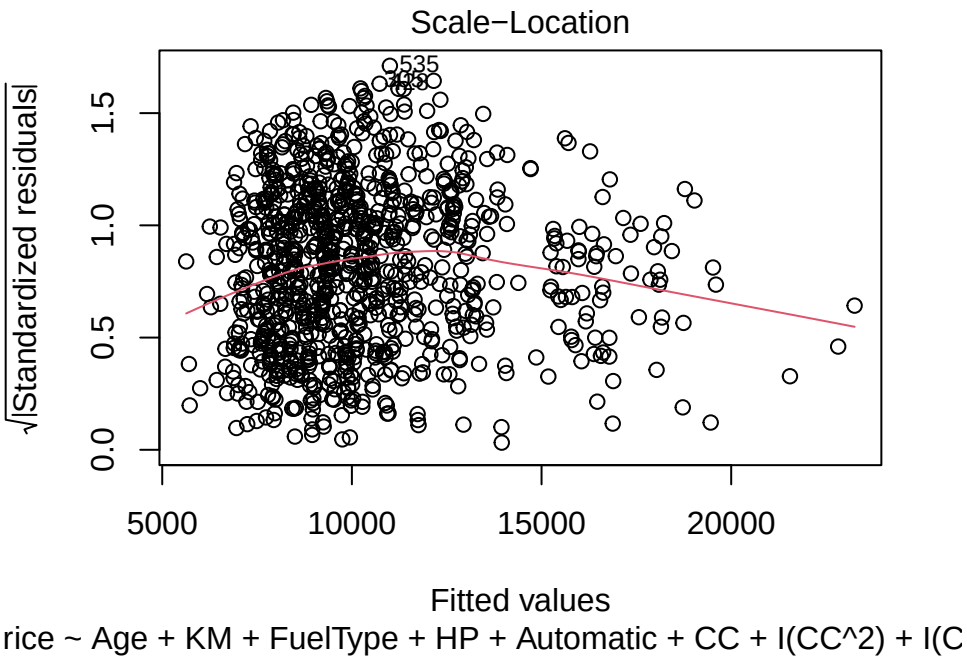
```
plot(m3, which = 2)
```



3) Scale-Location

Residuals are more randomly scattered in comparison to model m2, however red line seems to be less horizontal, meaning less satisfied assumption of homoscedasticity. However, this preception might be the result of more zoomed plot than before.

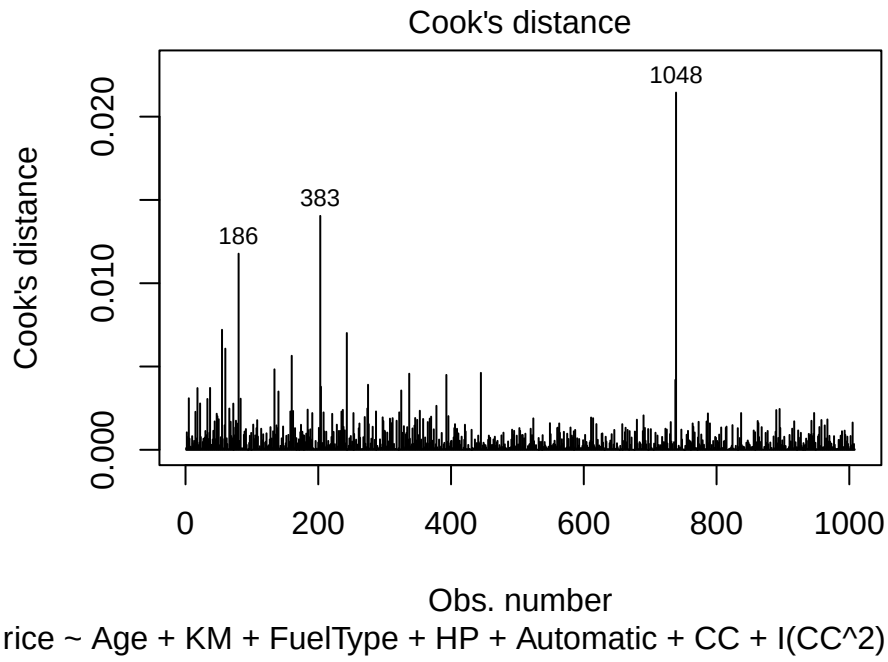
```
plot(m3, which = 3)
```



4) Cook's distance

New leverage points of great influence were detected: 186, 383, 1048. Their removal may improve the quality of prediction.

```
plot(m3, which = 4)
```



Polynomial orthogonal model

Model m4 was constructed in a similar way to model m3, but for orthogonal polynomials. The only found meaningful combination was for the modification of m2 model, with following formula:

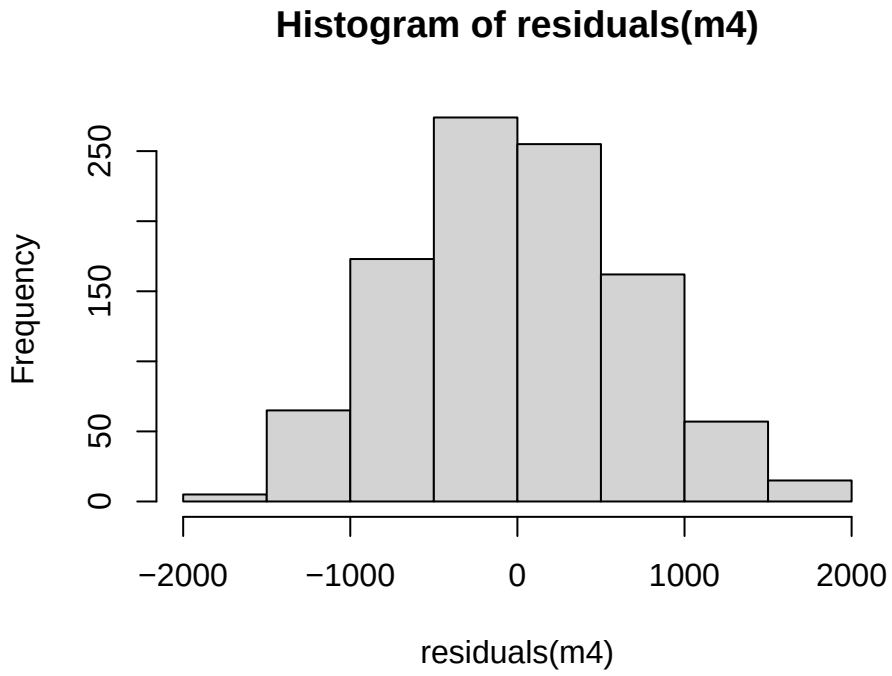
$Price = Age + KM + FuelType + HP + Automatic + CC + poly(Weight, 2)$

```
m4 <- update(m2,  ~. -Weight +poly(Weight, 2))

summary(m4)
##
## Call:
## lm(formula = Price ~ Age + KM + FuelType + HP + Automatic + CC +
##     poly(Weight, 2), data = outliers_removed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1775.34  -479.59   -15.57   466.71  1911.11
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.722e+04  5.329e+02  32.312  < 2e-16 ***
## Age              -1.160e+02  1.833e+00 -63.277  < 2e-16 ***
## KM               -1.523e-02  8.927e-04 -17.063  < 2e-16 ***
## FuelTypeDiesel    2.543e+03  7.455e+02   3.411 0.000672 ***
## FuelTypePetrol    1.325e+03  4.002e+02   3.310 0.000968 ***
## HP                4.575e+01  9.025e+00   5.070 4.75e-07 ***
## Automatic         4.304e+02  1.245e+02   3.458 0.000567 ***
```

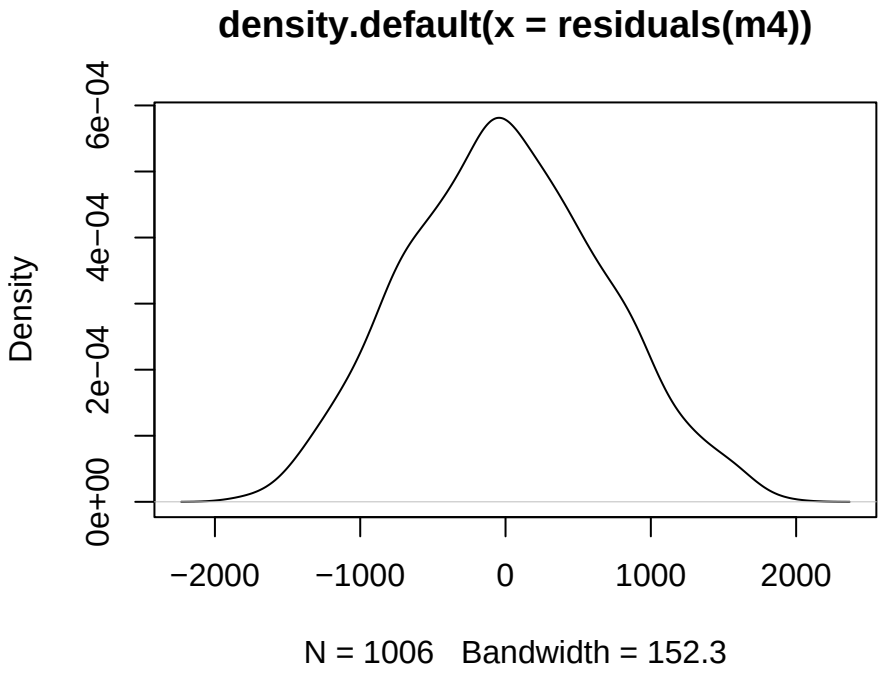
```
## CC                -3.430e+00  7.463e-01  -4.596 4.86e-06 ***
## poly(Weight, 2)1   2.634e+04  1.296e+03  20.331  < 2e-16 ***
## poly(Weight, 2)2   2.520e+03  9.597e+02   2.626 0.008774 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 677.6 on 996 degrees of freedom
## Multiple R-squared:  0.9374, Adjusted R-squared:  0.9368
## F-statistic:  1656 on 9 and 996 DF,  p-value: < 2.2e-16
rse4 <- sigma(m4)/mean(outliers_removed$Price)
rse4
## [1] 0.06628571
```

New model m4 has **Adj. R^2** value of 0.9368, meaning it predicts ~**94%** od variability in outcome data, and has ~**6.6% error rate**, which means that it is less accurate than m3.

```
mean(residuals(m4))
## [1] 2.340049e-14
hist(residuals(m4))
```




```
plot(density(residuals(m4)))
```

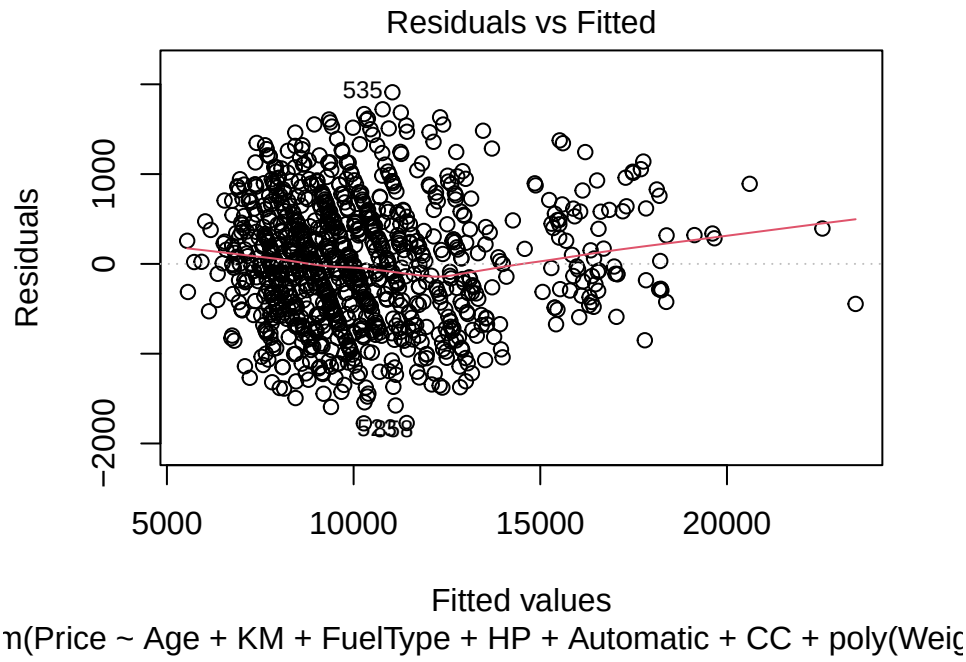


Again, there is no bias in the model as errors mean equals 0 and residuals are normally distributed.

1) Residuals vs fitted values

Like in m3, number of outliers is smaller than m2 and residuals better oscillate symmetrically around the 0 curve, although worse than in m3.

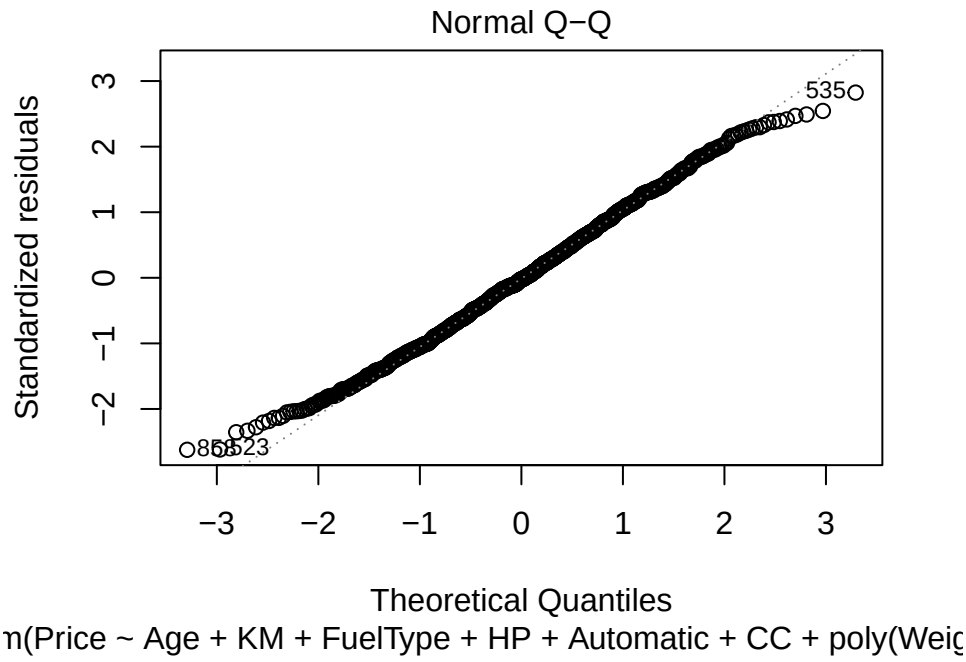
```
plot(m4, which = 1)
```



2) Normal Q-Q

Same as in m3, errors are almost normally distributed, with the distribution being more "pointy" than normal one.

```
plot(m4, which = 2)
```

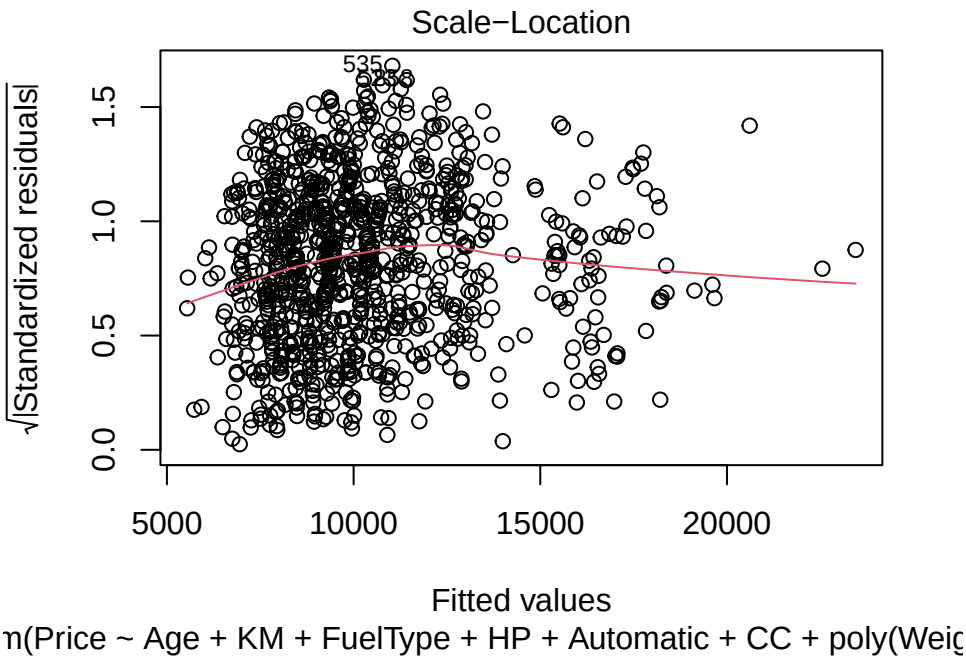


m(Price ~ Age + KM + FuelType + HP + Automatic + CC + poly(Weig

3) Scale-Location

Residuals are more randomly scattered in comparison to model m2, and red line seems to be more horizontal than m3, meaning assumption of homoscedasticity may be better satisfied.

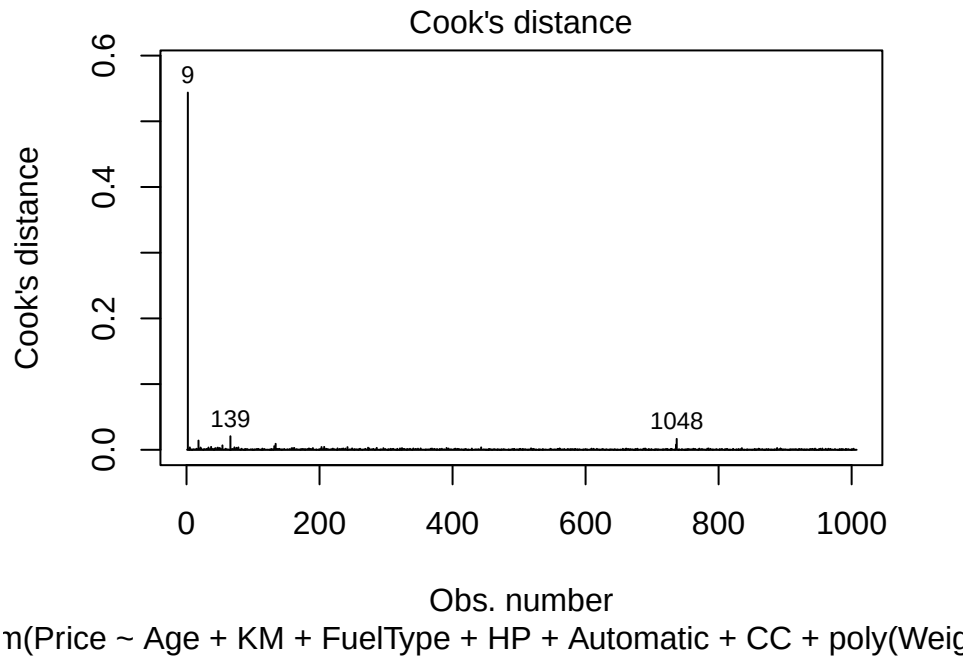
```
plot(m4, which = 3)
```



4) Cook's distance

New leverage points of great influence were detected, most importantly 9th one. It is very distinct from other ones in this distribution, and also in comparison to m3 cook's distances. This may mean that it is highly influenced by the weight.

```
plot(m4, which = 4)
```



Exercise III

ToDo:

- Compare all the models built for ToyotaCorolla using ANOVA and explain the results.

anova(m1, m2, m3, m4)

Anova method can be used when one model extends another one. In the process of constructing the models, not only new parameters were added, but also data set changed. Therefore, below simplified versions of previous models were retrained for the purpose of comparison.

Of course, this does not reflect comparisons between original models. However, in reality, comparison wouldn't be necessary because in this case model with smaller number of variables is better. Therefore the only importance is to compare m2 with m3 and m4, but this requires adding weight to m3 and m4 or removing it from m2.

```
m1t <- lm(Price ~. , data = outliers_removed)

m2t <- lm(Price ~. -Doors -MetColor  , data = outliers_removed)

m3t <- update(m2t,  ~. +I(CC^2) +I(CC^3) + I(Age^2) +I(Weight^2))

m4t <- update(m2t,  ~. +poly(Weight, 2))
```

```
anova(m2t, m1t)
## Analysis of Variance Table
##
```

```
## Model 1: Price ~ (Age + KM + FuelType + HP + MetColor + Automatic + CC +
##     Doors + Weight) - Doors - MetColor
## Model 2: Price ~ Age + KM + FuelType + HP + MetColor + Automatic + CC +
##     Doors + Weight
##   Res.Df       RSS Df Sum of Sq      F   Pr(>F)
## 1    997 460411214
## 2    995 455221337  2   5189877 5.6719 0.003554 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m2t, m3t)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ (Age + KM + FuelType + HP + MetColor + Automatic + CC +
##     Doors + Weight) - Doors - MetColor
## Model 2: Price ~ Age + KM + FuelType + HP + Automatic + CC + Weight +
##     I(CC^2) + I(CC^3) + I(Age^2) + I(Weight^2)
##   Res.Df       RSS Df Sum of Sq      F   Pr(>F)
## 1    997 460411214
## 2    993 442310362  4  18100853 10.159 4.61e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m2t, m4t)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ (Age + KM + FuelType + HP + MetColor + Automatic + CC +
##     Doors + Weight) - Doors - MetColor
## Model 2: Price ~ Age + KM + FuelType + HP + Automatic + CC + Weight +
##     poly(Weight, 2)
##   Res.Df       RSS Df Sum of Sq      F   Pr(>F)
## 1    997 460411214
## 2    996 457245658  1   3165556 6.8954 0.008774 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The parameter Pr(>F) is the probability, that rejecting null hypothesis (the most complex model does not fit better than the simplest model) could be an error. Therefore if the resulting p-value is sufficiently low (usually less than 0.05), we conclude that the more complex model is significantly better than the simpler model, and thus favor the more complex model.

In this case, p-value in all tests was lower than 0.05, meaning m2t model was significantly worse from all other models.

We cannot compare models m1t, m3t and m4t, because in order to use ANOVA, the second model must be an extended model of the first model, which is not the case here.