Regression Models Course Project

Anikó Medgyesi

2017 január 8

MTCars Dataset Analysis

The exercise is about the analysis of MTCars dataset from Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

1.Is an automatic or manual transmission better for MPG?

2.Quantify the MPG difference between automatic and manual transmissions?

Dataset description

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

A data frame with 32 observations on 11 variables.

[, 1] mpg Miles/(US) gallon [, 2] cyl Number of cylinders [, 3] disp Displacement (cu.in.) [, 4] hp Gross horsepower [, 5] drat Rear axle ratio [, 6] wt Weight (1000 lbs) [, 7] qsec 1/4 mile time [, 8] vs V/S [, 9] am Transmission (0 = automatic, 1 = manual) [,10] gear Number of forward gears [,11] carb Number of carburetors

Some of the variables should be considered as factors.

```
data(mtcars)
mtcars$am <- factor(mtcars$am, levels=c(0,1), labels=c('Automatic', 'Manual'))
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
 mtcars$gear <- factor(mtcars$gear)
 mtcars$carb <- factor(mtcars$carb)
```

Connection between mpg and transmission.

Based on exploratory data analysis it is seems to be that With manual transmission it is possible to reach a better mpg performance. See appendix.

Let us check with a linear regression model:

```
fit1 <- lm(mpg ~ am, mtcars)
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amManual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

We can see on the paremeters of linear modell that there is indeed a relationship between the am and mpg, with a switch to manual transmission it is likely that the mpg will be increased an avareage 7,24 mpg with p-value less then 0.001. The R-squared value 0.3598, so the the modell explains only the 36% of variance.

What other parameters should be considered in the model?

```
analysis <- aov(mpg ~ ., data = mtcars)
summary(analysis)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          2  824.8   412.4  51.377 1.94e-07 ***
## disp         1   57.6    57.6   7.181   0.0171 *
## hp           1   18.5    18.5   2.305   0.1497
## drat         1   11.9    11.9   1.484   0.2419
## wt           1   55.8    55.8   6.950   0.0187 *
## qsec         1    1.5     1.5   0.190   0.6692
## vs           1    0.3     0.3   0.038   0.8488
## am           1   16.6    16.6   2.064   0.1714
## gear         2    5.0     2.5   0.313   0.7361
## carb         5   13.6     2.7   0.339   0.8814
## Residuals   15  120.4     8.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The analysis shows that we should integrate into the model the Weight(wt), Displacement(disp), and Number of cilinders(cyl) The new modell:

```
fit2 <- lm(mpg ~ am + cyl + disp + wt, data = mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4.5029 -1.2829 -0.4825  1.4954  5.7889
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.816067   2.914272  11.604 8.79e-12 ***
## amManual     0.141212   1.326751   0.106  0.91605
## cyl6        -4.304782   1.492355  -2.885  0.00777 **
## cyl8        -6.318406   2.647658  -2.386  0.02458 *
## disp         0.001632   0.013757   0.119  0.90647
## wt          -3.249176   1.249098  -2.601  0.01513 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.652 on 26 degrees of freedom
## Multiple R-squared:  0.8376, Adjusted R-squared:  0.8064
## F-statistic: 26.82 on 5 and 26 DF,  p-value: 1.73e-09
```

The extended modell explains the 86% of the variance.

The comparision of the two models:

```
anova(fit1,fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + disp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 182.87  4    538.03 19.124 1.927e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This results in a p-value of 8.637e-08, so we can accept, that the fit2 modell is better then fit1. So the manual use of gears means about 1.81 MPG increase. We can trust in our regression models, because our resdidials are approximately normally distributed. (see appendix)

Appendix

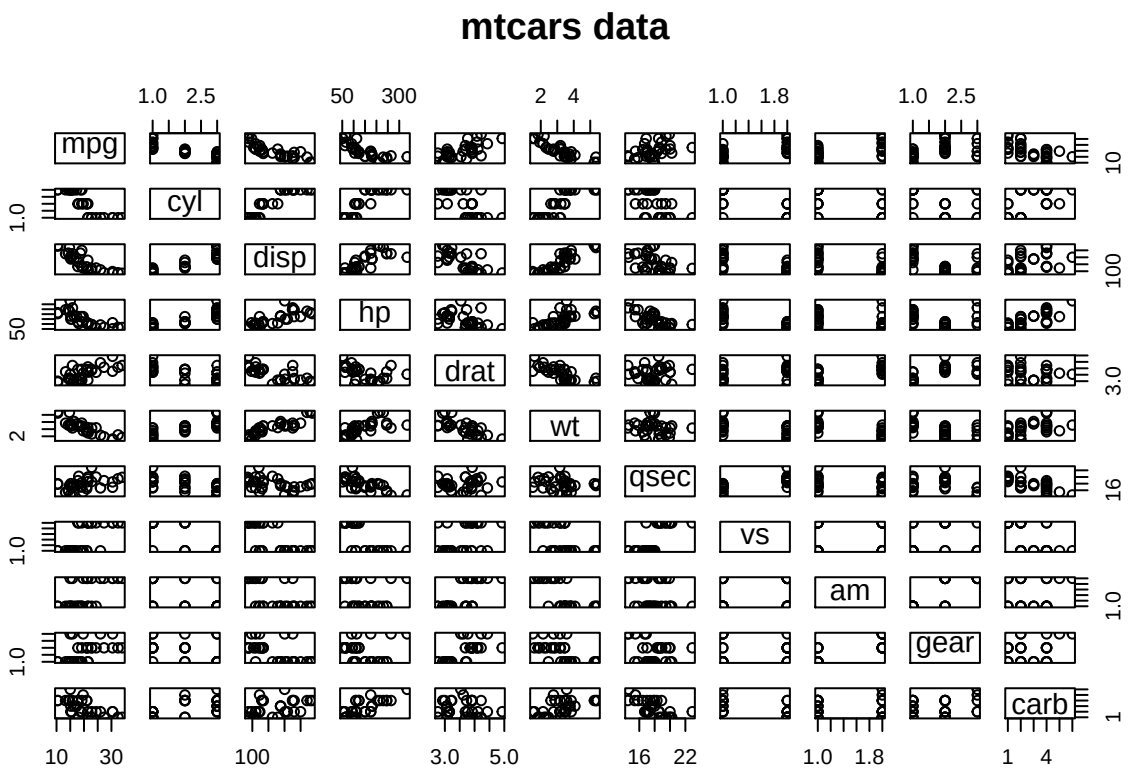
Visualisation for mtcars data:

```
summary(mtcars)
```

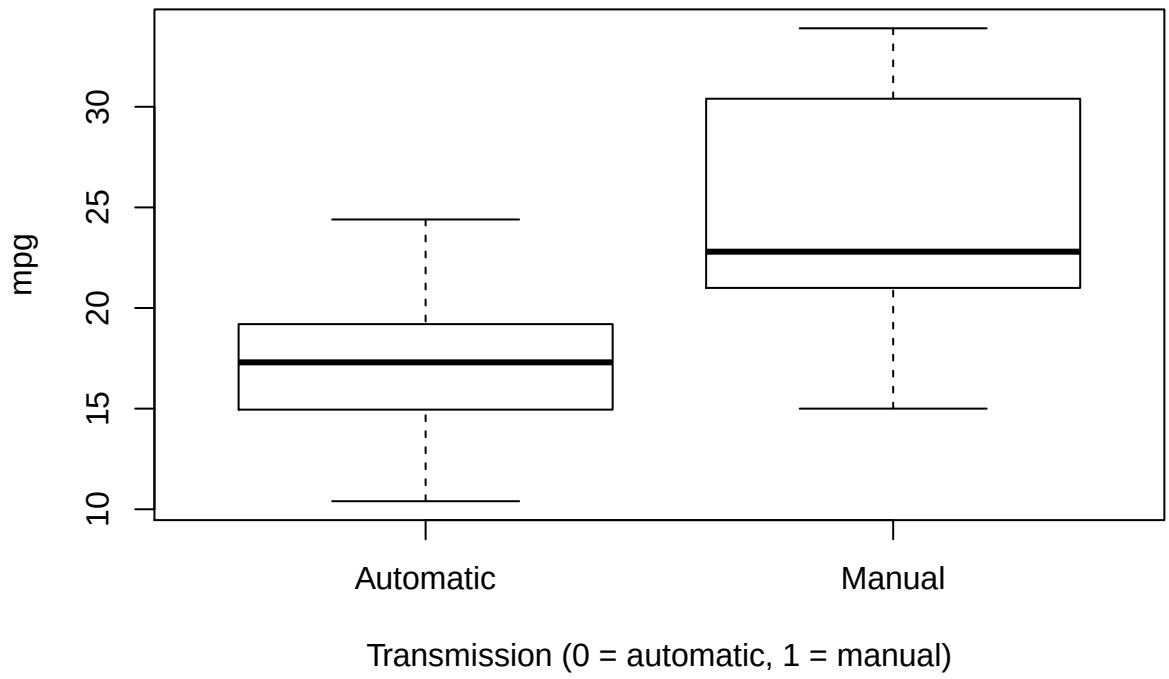
```
##       mpg        cyl         disp             hp             drat
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930
##        wt             qsec       vs              am     gear   carb
##  Min.   :1.513   Min.   :14.50   0:18   Automatic:19   3:15   1: 7
```

```
##  1st Qu.:2.581   1st Qu.:16.89   1:14   Manual   :13   4:12   2:10
##  Median :3.325   Median :17.71                         5: 5   3: 3
##  Mean   :3.217   Mean   :17.85                                4:10
##  3rd Qu.:3.610   3rd Qu.:18.90                                6: 1
##  Max.   :5.424   Max.   :22.90                                8: 1
```

```r
require(graphics)
pairs(mtcars, main = "mtcars data")
```

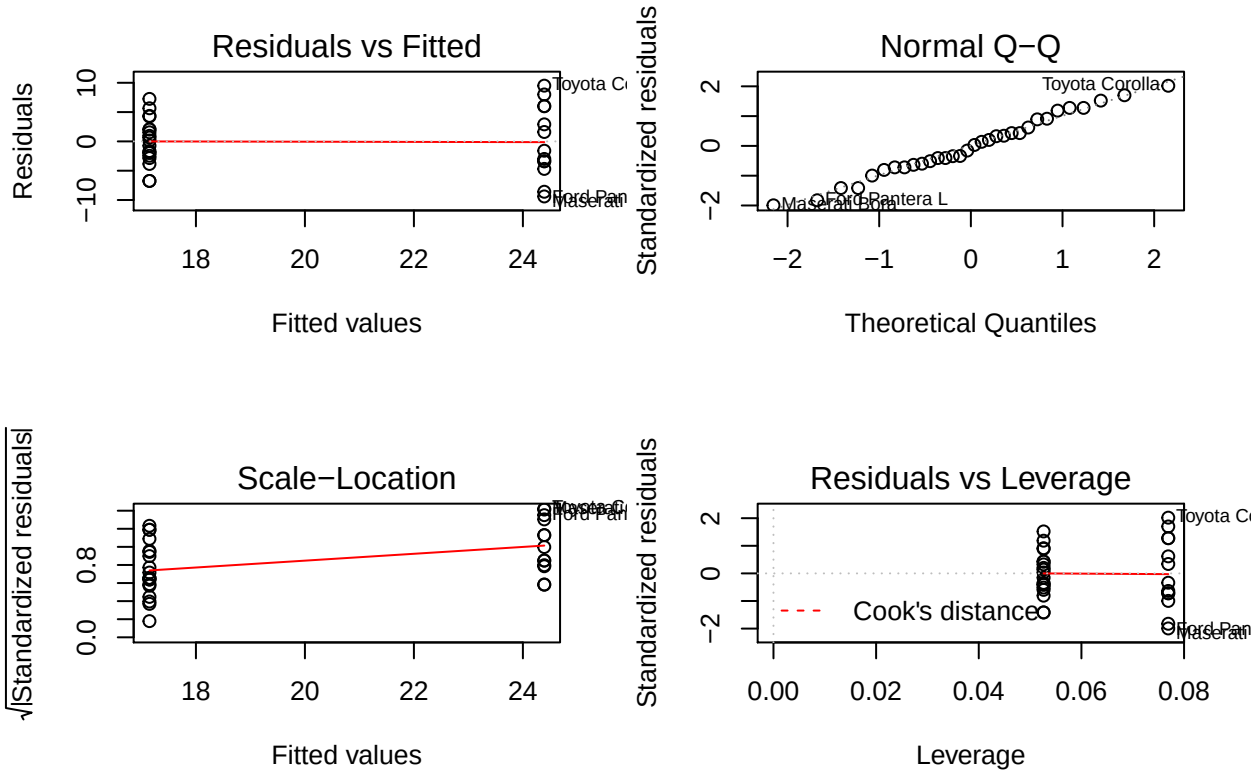


Visualisation the Miles/gallon(mpg) based on Transmission(am).



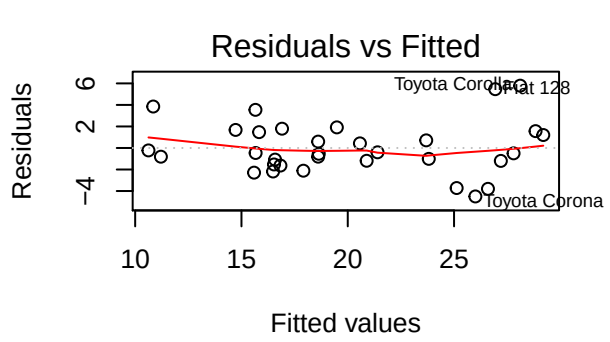
Check residuals of fit1 if normally distributed

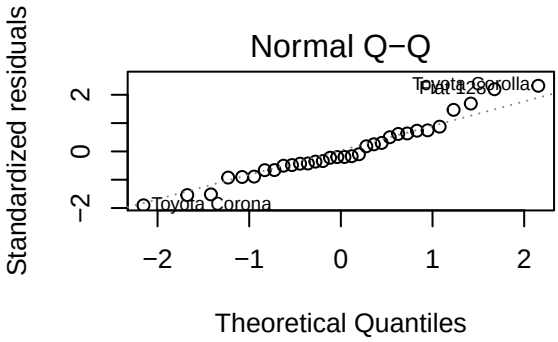
```
par(mfrow = c(2,2))
plot(fit1)
```

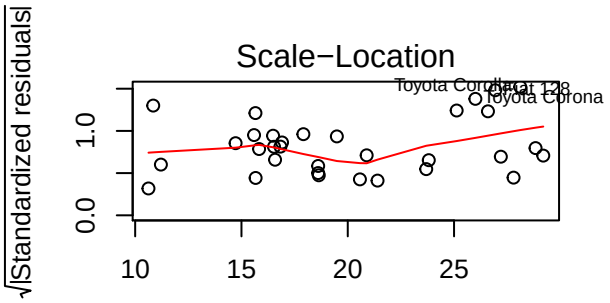


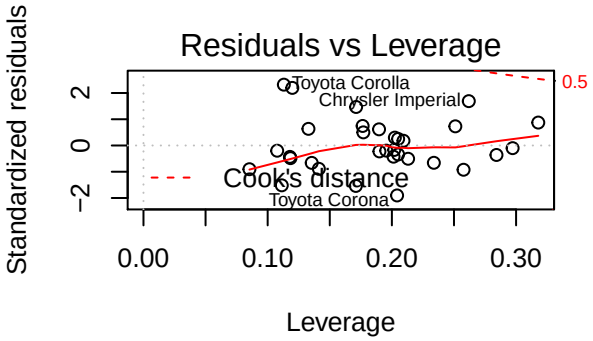
Check residuals of fit2 if normally distributed

```
par(mfrow = c(2,2))
plot(fit2)
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Toyota Corona


Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Toyota Corona


Scale−Location
√|Standardized residuals|
Fitted values
Toyota Corolla
Fiat 128
Toyota Corona


Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
0.5