Regression Models Course Project

Elizaveta Oginskaya

June 5, 2016

Executive Summary

In this paper a set of car parameters was analysed to find whether MPG depends on transmission type:

- **Is an automatic or manual transmission better for MPG?** Yes, with 95% confidence level we can say that in general MPG is better for automatic transmission.
- **Quantify the MPG difference between automatic and manual transmissions.** Taking into consideration *only* transmission type automatic transmission allows to increase MPG by **7.24** miles per gallon. But there are two other significant parameters: *weight (wt)* and *1/4 mile time (qsec)*. Taking them into account as constant automatic transmission raise MPG only by average **4.30** miles per gallon.

As the report should only include 2 pages of results and annexes can only contain figures, please look at to .rmd file if you have any questions to the code.

Exploring and Adjusting Data

The datasets contained 32 observations of 11 parameters of different car models (for more information on data please visit http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html) *(Annex 1)*

The original dataset has all variables of numeric class. Before the main part of analysis variables `cyl`, `gear` and `carb` were converted into factors (the code is hidden, but available in *.rmd* file, look it up also for working dataset structure).

Statistical Inference

To answer the first question, whether the MPG parameter depends on transmission type or not the t-test is taken, where the null hypothesis assume that the difference in MPG mean for cars with automatic transmission (factor = 1) and cars with manual transmission (factor = 0) is insignificant.

The 95% confidence interval doesn't contain 0 and the p-value is very small, so the null hypothesis should be rejected in favor of the statement that means of two groups of cars are different.

Regression Model

Simple Regression Model

To evaluate the difference in MPG for automatic and manual transmissions the simple linear regression model was build. It assumes that MPG depends only on transmission type.

```r
fit1 <- lm(mpg ~ am, data = mtcars)
summary(fit1)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04
```

The model shows that the average MPG of cars with manual transmission equals **17.15** and automatic transmission increase MPG by **7.24** to **24.39**. P-values for the both coefficients are relatively small so the both coefficients are significant. Adjusted R-squared value is only 0.338, which means that the model explains only **33.8%** of mpg variation. *(Annex 2)*

Finding Optimal Regression Model

For the search of the optimal model three more models were build: Regression model including all 10 regressors explains more variance of MPG - **77.9%**, but many of its coefficients have relatively large values and appears insignificant (see *.rmd* file)

The *stepwise* algorithm by *AIC* was implied to find an optimal fit model. The algorithm found the following best fit model:

```r
summary(fits)$call
```

```
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
```

All coefficients of the model except for the intercept turned out to be significant. The adjusted R-squared value for the models equals **0.834**.

On the base of previous model the last model was built by excluding the intercept:

```r
fitn <- lm(mpg ~ wt + qsec + am - 1, data = mtcars)
summary(fitn)$coef
```

```
##       Estimate Std. Error   t value     Pr(>|t|)
## wt   -3.185455  0.4827586 -6.598442 3.128844e-07
## qsec  1.599823  0.1021276 15.664944 1.091522e-15
## am    4.299519  1.0241147  4.198279 2.329423e-04
```

This model has all coefficients significant and explains **98.6%** of MPG variance.

Nested models

```r
as.matrix(anova(fit1, fitn))
```

```
##   Res.Df      RSS Df Sum of Sq       F       Pr(>F)
## 1     30 720.8966 NA        NA      NA           NA
## 2     29 180.8323  1  540.0643 86.6099 3.284554e-10
```

The ANOVA test shows that all three regressors are significant

Residuals Diagnostics

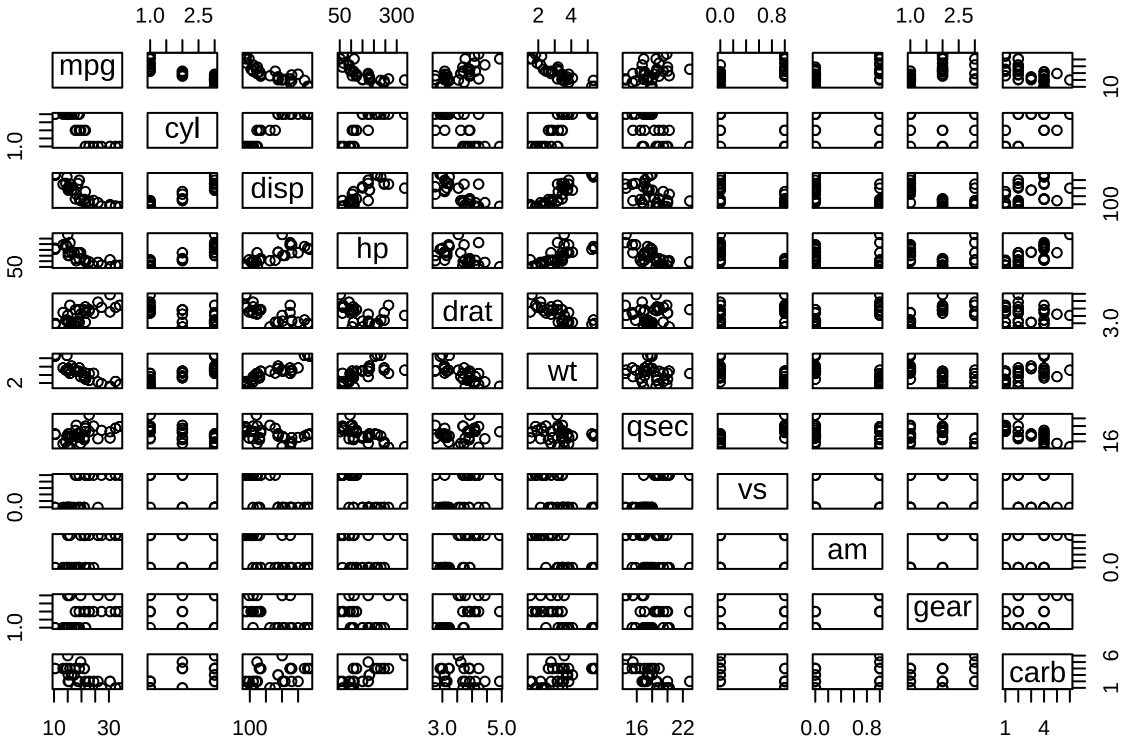
The residuals plots (Annex) show that:

- Residuals vs. Fitted plot: points ate randomly distributed, so residuals are IID.
- Normal Q-Q plot: the points mostly follow on the line, so residuals are normally distributed.
- The Scale-Location plot: points distributed in a constant pattern, the variance of residuals is constant.
- Residuals VS Leverage: some distinct points indicates few outliers

Annexes

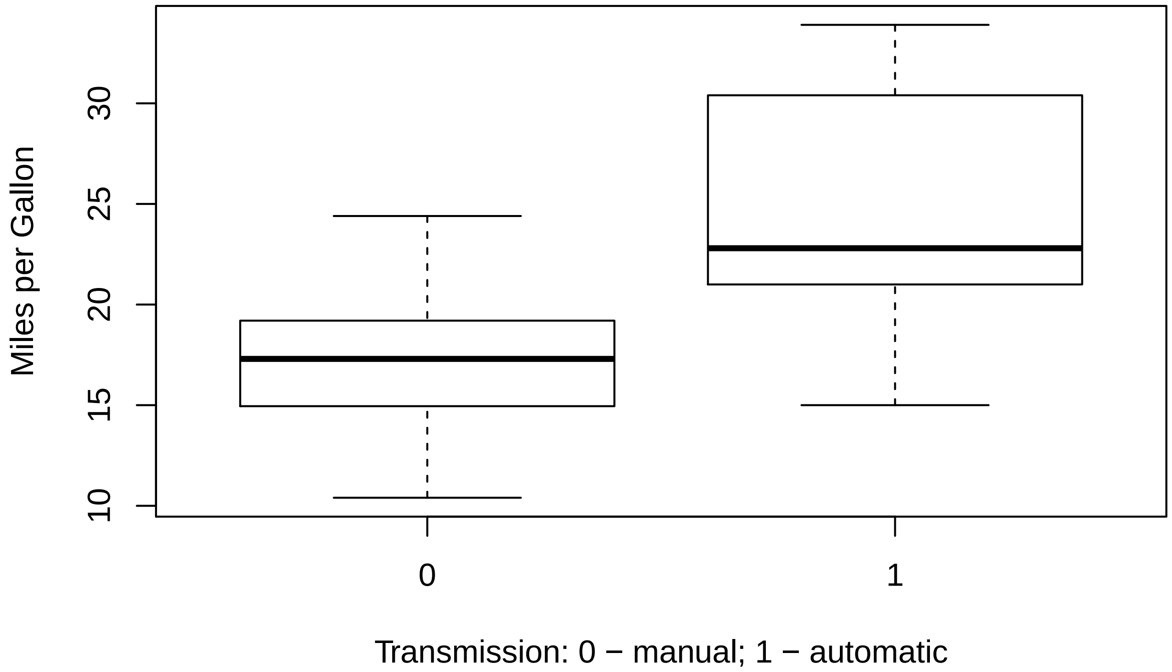
Annex 1: Data Pairs

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



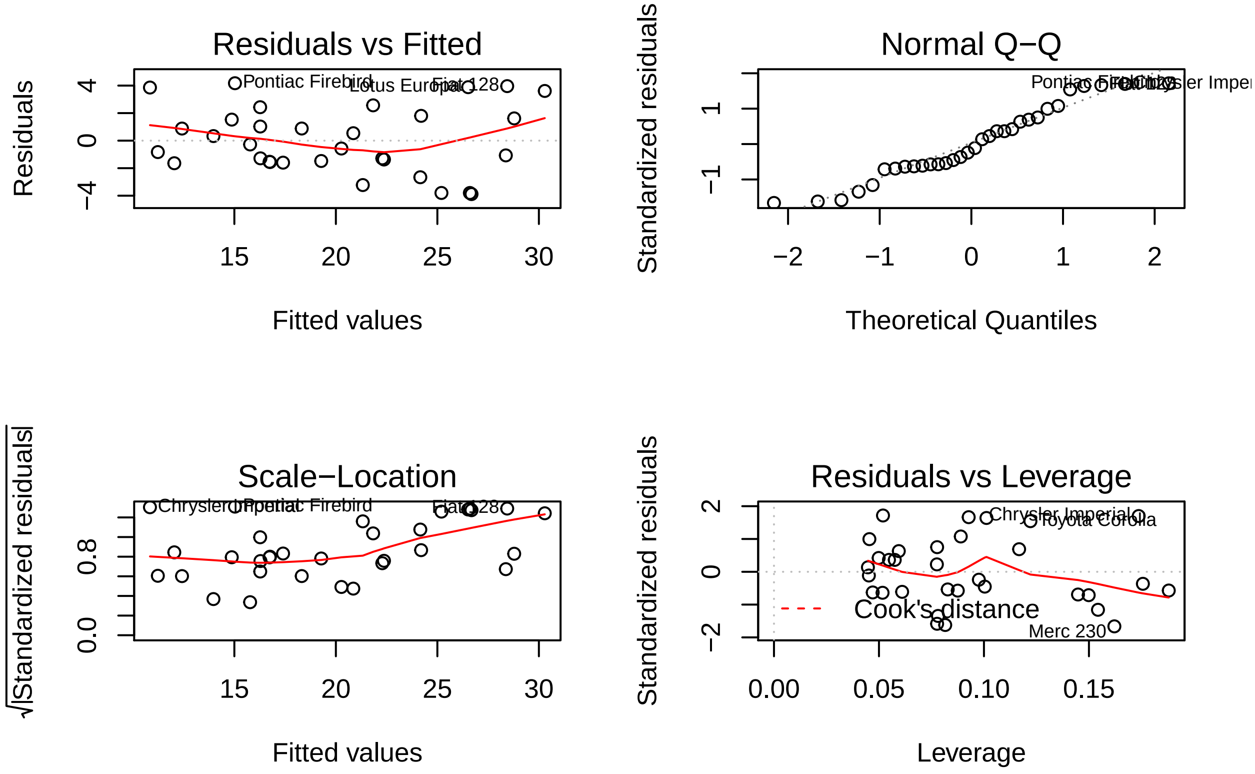
Annex 2: MPG by transmission type:

```
par(mfrow = c(1, 1))
boxplot(mpg ~ am, data = mtcars, xlab = 'Transmission: 0 - manual; 1 - automatic', ylab = 'Miles per Gal
```



Annex 3: Residuals Diagnostics

```r
par(mfrow = c(2, 2))
plot(fitn)
```


Residuals vs Fitted
Residuals
Fitted values
−4
0
4
15
20
25
30
Pontiac Firebird
Lotus Europa
Fiat 128
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−1
1
−2
−1
0
1
2
Pontiac Firebird
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
0.0
0.8
15
20
25
30
Chrysler Imperial
Pontiac Firebird
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
−2
0
2
0.00
0.05
0.10
0.15
Chrysler Imperial
Toyota Corolla
Merc 230
Cook's distance