# MPG Dependence on Transmission Type

*by Milica Micic*

## Executive Summary

We analyse the data that is extracted from the 1974 *Motor Trend US* magazine, containing fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). To address the question which transmission is better (automatic or manual) in respect to the miles per gallon ratio, we look into different models to find the one that describes this dependence the best. We find with 78% confidence that models with manual transmission always provide better fuel consumption by 1.81 mpg. With 95% confidence we could also claim that models with manual transmission have better mpg ratio by 1.8, but not always given that the 95% confidence interval [-1.06, 4.68] contains zero.

All hidden codes for producing results presented in this report can be found in the Rmd file available on my GitHub account: https://github.com/afmicka/Regression_Models

## Exploratory Data Analysis

In order to explore various relationships between variables, we show in Figure 1 (see Appendix) the pairs plot for `mtcars` dataset. It suggests that most of the variables have a strong correlation with fuel consumption (`mpg`). Latter we address the issue of which correlations are significant. With our main interest being the effect of car transmission type on fuel consumption, we show in Figure 2 (see Appendix) distribution of miles per gallon (`mpg`) for both manual and automatic transmission. It clearly shows that manual transmission provides greater mileage per gallon.

## Regression Models

In this section, we start building linear regression models based on the different variables and try to find the best model fit. Given that our main interest is the fuel consumption dependence on transmission type, we start from a simple model that shows how `mpg` changes only between automatic and manual transmission disregarding all other variables in the dataset.

```
simpleModel <- lm(mpg ~ am, data = mtcars)
```

This model tells us that mean `mpg` value is 17.15 (intercept in Table 1) for automatic transmission and that the average mileage increases by 7.24 (amManual coefficient) with manual transmission, giving the mean `mpg` value of 24.39 for manual transmission.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **amManual** | 7.245 | 1.764 | 4.106 | 0.000285 |
| **(Intercept)** | 17.15 | 1.125 | 15.25 | 1.134e-15 |

Table 1: Fitting linear model: mpg ~ am

In order to select the best model, we need to find out which variables have the biggest impact on fuel consumption, beside transmission type. We first create the model that uses all variables as predictors

```r
fullModel <- lm(mpg ~ ., data = mtcars)
```

and use "Backward step-wise regression", starting from this model, to remove those variables that are not statistically significant and create the model that best fits the data.

```r
bestModel <- step(fullModel)
```

The model for 'best fit' tells us that the number of cylinders (`cyl`), weight (`wt`), horsepower (`hp`) and transmission (`am`) are the relevant variables to fit `mpg`, with 84% of the variability explained by this model. It also suggests that manual transmission provides greater `mpg` ratio by 1.8 in comparison to automatic transmission, keeping all other relevant variables constant. Comparing these three models - best to simple and full to best - with `anova()` function:

```r
anova(simpleModel, bestModel, fullModel)
```

we find that there is a very significant difference between the best and the simplest model (see Table 2). p-value is very small meaning that variables added in the best model are important. On the other hand, high p-value for F-statistics in the third row of the table shows that adding more variables to the best model does not change anything and it would be redundant to use model with all variables as predictors.

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|---|
| **2** | 26 | 151 | 4 | 569.9 | 17.75 | 1.476e-05 |
| **3** | 15 | 120.4 | 11 | 30.62 | 0.3468 | 0.9588 |

Table 2: Model comparison between simple and best model (first row), and best and full model (second row)

## Residuals and Diagnostics

In Figure 3 (see Appendix) we show the diagnostic of residuals for our regression (best) model. Residuals versus fitted plot and scale-location plot show no pattern confirming the variable independence in our model. Normal Q-Q plot indicate that residuals are approximately normally distributed. Residuals versus leverage plot indicates that there might be some leverage points influencing coefficients. Looking into the absolute `dfbeta` values that are indicators of the change in the slope coefficients when the observation is removed, we find that there are no values greater than 1. This means that there is no car model whose exclusion would greatly influence our regression model, including the ones labeled in Figure 3 that have the greatest `dfbeta` values. The points with highest hatvalues are Maserati Bora (hatvalue = 0.471), Lincoln Continental (hatvalue = 0.294), and Toyota Corona (hatvalue = 0.278). However, these values are smaller than 1, suggesting that there is no reason for removing these observations from the dataset.

## Statistical Uncertainty

The uncertainty in the conclusion that MPG is better with manual transmission can be quantified by calculating the 95 % confidence interval for our intercept of 1.81 MPG. We get that this interval around 1.8 MPG is [-1.06, 4.68]. However, it contains zero suggesting that it is possible that in some cases automatic transmission has better `mpg` than manual. In order to avoid this uncertainty, we look for the interval that does not contain zero ensuring that the manual transmission would always have better `mpg` than automatic, but with smaller confidence percentage. We find this to be 78% confidence interval [0.05, 3.56].

## Appendix

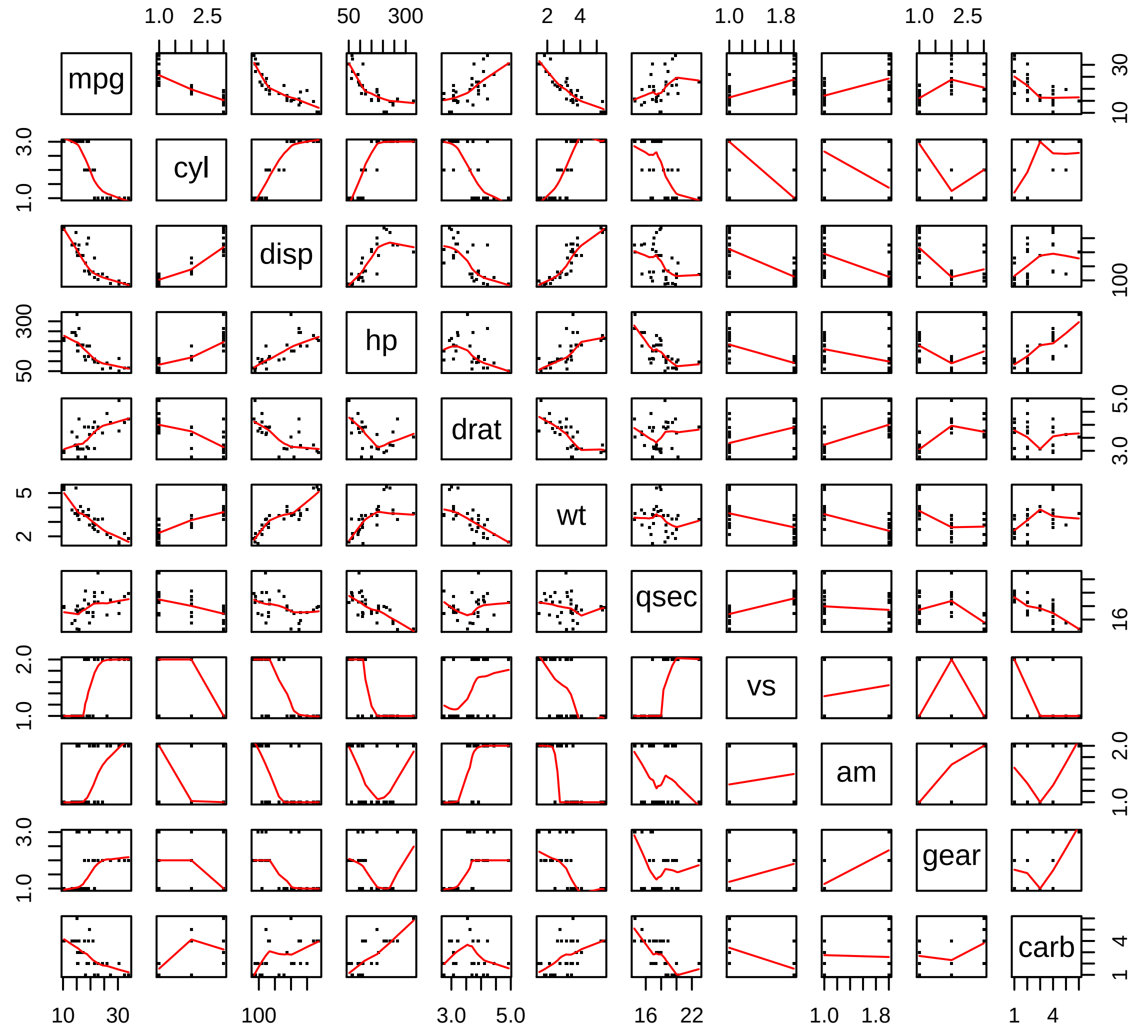


Figure 1: Pairs plot for 'mtcars' dataset.

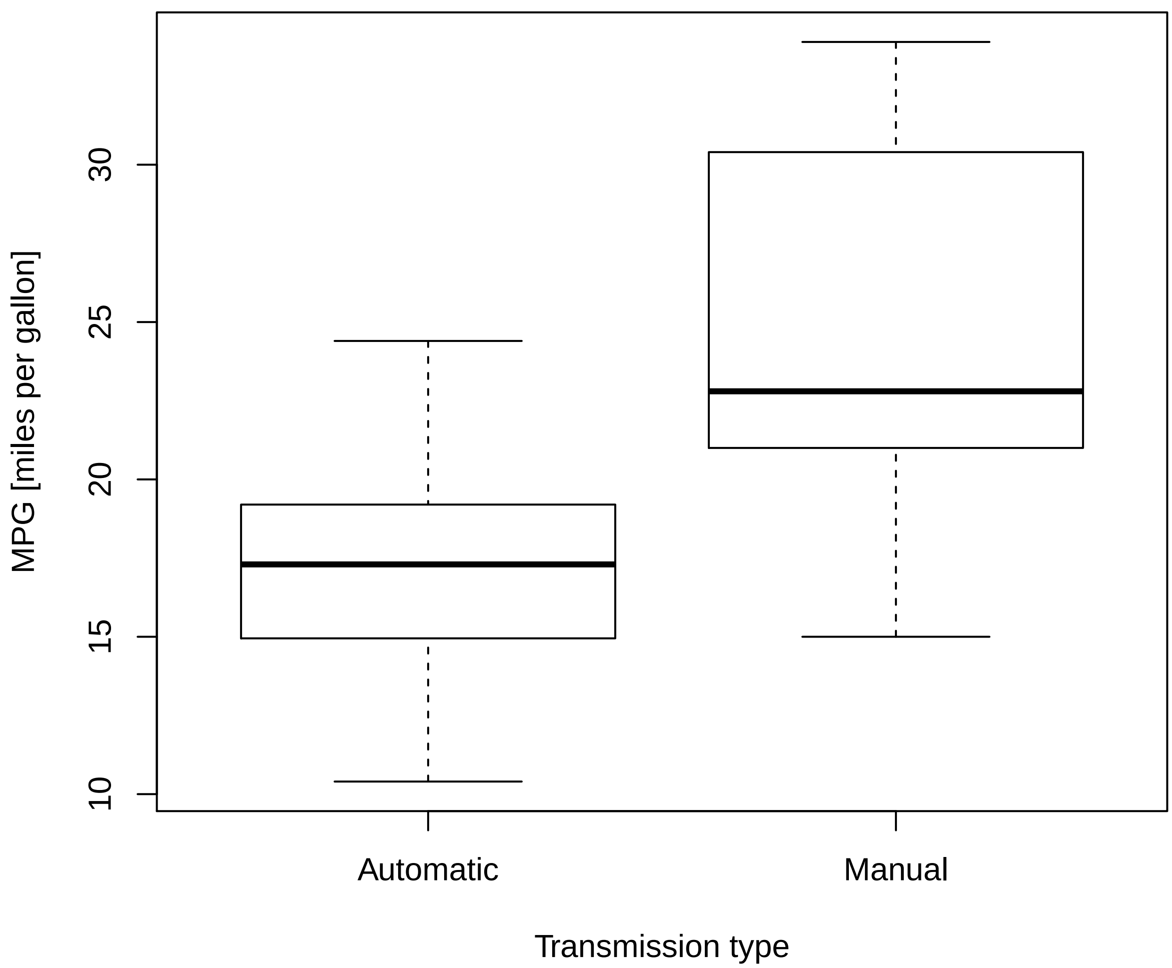


Figure 2: The distirbution of miles per gallon (mpg) for automatic (left) and manual (right) transmission.

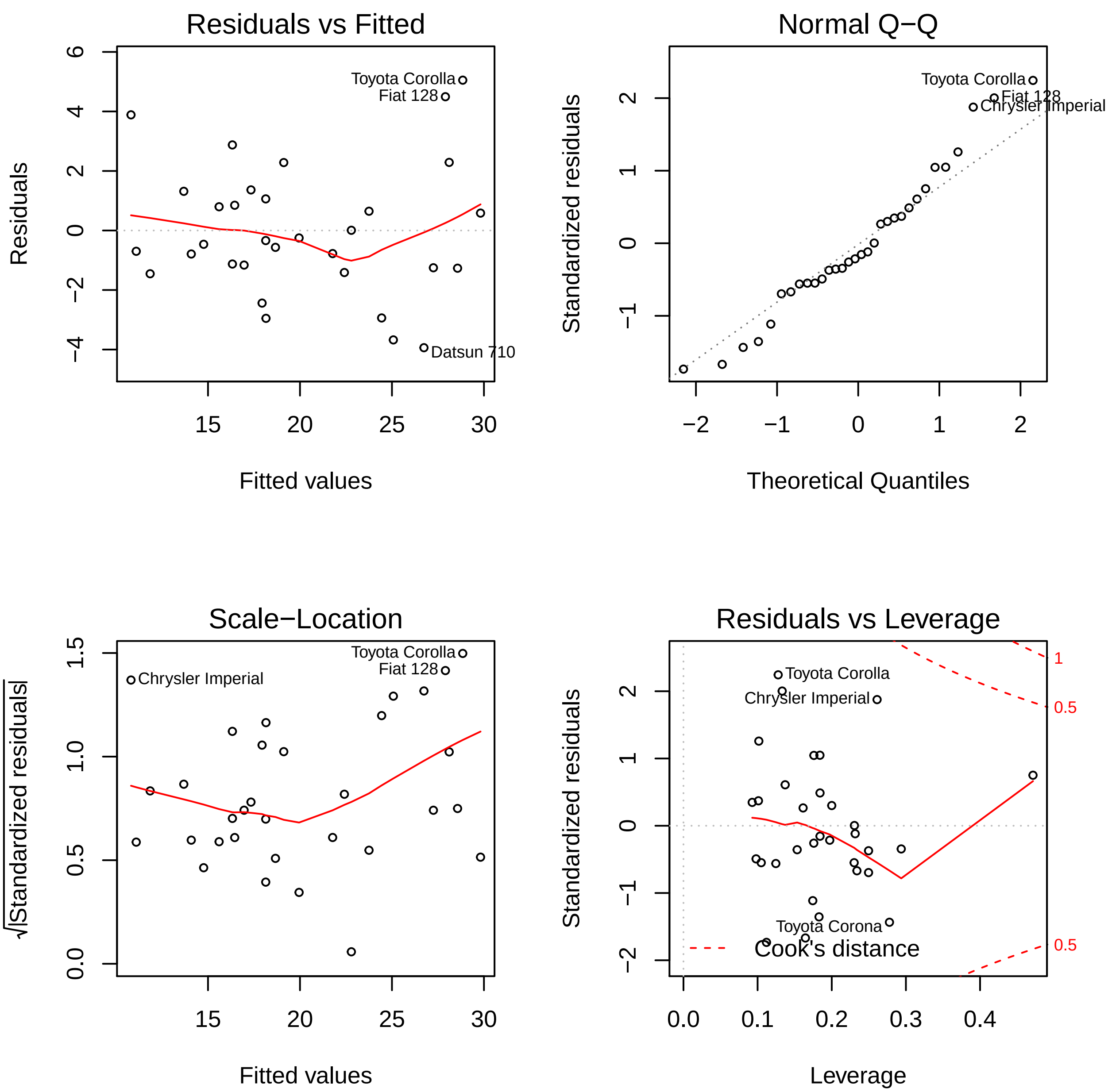


Figure 3: The diagnostic of residuals