Influence of the Type of Transmission on MPG

Miroslav Micic

January 22, 2015

Executive Summary

We analyse data in "mtcars" data set in order to establish if automatic or manual transmission is better for MPG. We look into two models: the simplest model where we look into correlation between MPG and type of transmission when all other variables are disregarded; and the best fit model where all important (relevant) variables are taken into account. We find that in both models manual transmission has better MPG than automatic transmission. In the simplest model this difference is 7.24 MPG. When all relevant variables are included, manual transmission is still better by 1.81 MPG. The 95 % confidence interval is [-1.06, 4.68] MPG so MPG might not always be larger for manual transmission. Confidence interval which guarantees this is [0.05, 3.56] with 78 % confidence. In conclusion, we claim that there is 78 % chance that manual transmission is always better than automatic by MPG in the interval [0.05, 3.56] and the most likely MPG difference of 1.8.

All codes are hidden in the report and presented in Rmd file available on github (see last section of the report).

Exploratory Data Analysis

First, we load in "mtcars" data set. After visually inspecting data with head() function we decide to transform variables "cyl", "vs", "am", "gear", and "carb" into factors. Before any detailed analysis, we perform a visual inspection of possible correlations between the variables in the "mtcars" data set. Figure 1 in the Appendix compares all pairs of variables. It seems that there is a definite correlation between "mpg" and "cyl", "disp", "hp", "drat", "wt" variables. Figure 2 in the Appendix shows boxplot of "mpg" for both manual and automatic transmission. We examine this later in more details.

Multiple Models

Simplest Model:

We start with the simplest model where "mpg" is a function of "am" only. Here we look how "mpg" changes between automatic and manual transmission disregarding all other variables in the data set.

In this context, first intercept (first estimate in the summary of lm() function) is the mean mpg for automatic transmission and second intercept is the increase in the mean mpg with manual transmission. This tells us that mean "mpg" = 17.15 for automatic and 24.39 for manual transmission. Manual transmission is better by 7.24 MPG.

Best model:

We use R function step() to perform variable selection and find the set of variables that best fit the data. We call these variables "relevant variables".

The best fit is: mpg ~ wt + cyl + hp + am

Estimate in the last row of the summary of our best fit shows manual transmission increase in "mpg" using automatic transmission as a reference set, and keeping all other relevant variables constant. Its value is 1.8 MPG. Hence, manual transmission has better MPG than the automatic transmission by 1.8 MPG.

Complete model:

We also fit a model (complete model) where all variables in the mtcars data set are taken into account.

Comparing simplest, best, and complete models:

We use anova () function to compare our three models in the following order: best model is compared to the simplest model and complete model is compared to the best model.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
30	720.9	NA	NA	NA	NA
26	151	4	569.9	17.75	1.476e-05
15	120.4	11	30.62	0.3468	0.9588

Table 1: Comparison, best to simplest model in the second row, and complete to best model in the last row.

Table 1 shows that there is a very significant difference between best and simplest model (P-value for F-statistics in the second row of the table 1). This means that variables added in the best model are important. On the other hand, P-value for F-statistics in the third row shows that there is almost no difference between complete and best model. Therefore, it would be redundant to use complete model since adding more variables does not not change anything.

Residuals and Model Diagnostics

The diagnostic of residuals for the best fit is presented in Figure 3 in the Appendix. Residuals versus fitted plot and scale-location plot show that there is no pattern which means that our fit is good. The normal Q-Q plot shows that residuals are approximately normally distributed. Residuals versus leverage plot shows that there might be some influence of particular points on the coefficients. We investigate this further by looking into dfbetas and hatvalues. The change in the coefficients if some point is taken out or not is represented by dfbetas. We sum dfbetas for each type of car to get the cars with the largest accumulated influence on the fit coefficients.

These are: Toyota Corona with dfbeta = 3.097, Chrysler Imperial with dfbeta = 2.896, and Volvo 142E with dfbeta = 1.930.

Cars with the largest hatvalues are: Maserati Bora with hatvalue = 0.471, Lincoln Continental with hatvalue = 0.294, and Toyota Corona with hatvalue = 0.278. These cars are the biggest outliers. Toyota Corona has large values of both dfbeta and hatvalue parameters. In conclusion, Toyota Corona is the car that deviates from the best fit more than any other car.

Quantify the Uncertainty by Statistical Inference

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 do this by adding and subtracting the standard deviation of the intercept multiplied by the appropriate t-quantile.

The 95 % confidence interval around 1.8 MPG is [-1.06, 4.68] MPG. This means that there is 95 % chance (uncertainty in our conclusion) that the difference in MPG between manual and automatic transmission is in the interval [-1.06, 4.68] MPG. One can notice that interval contains zero which means that it is possible that in some cases automatic transmission has better MPG than manual.

It is much more useful to look for the confidence interval around 1.8 MPG which does not contain zero. In this interval, manual transmission would always have better MPG than automatic, but it would have smaller confidence percentage.

We find that the interval which guarantees that manual transmission always has better MPG than automatic is [0.05, 3.56]. The confidence of this interval is 78 %. In conclusion, we claim that there is 78 % chance that manual transmission is always better than automatic by MPG in the interval [0.05, 3.56] and the most likely MPG difference of 1.8.

Proof that the report was done in Rmd (knitr)

All of the codes in the report are hidden and can be found in the Rmd code. Rmd code generating this report can be found at the github with the following link: https://github.com/zanlik1977/Regression_Models_Project

Apendix

Figure 1

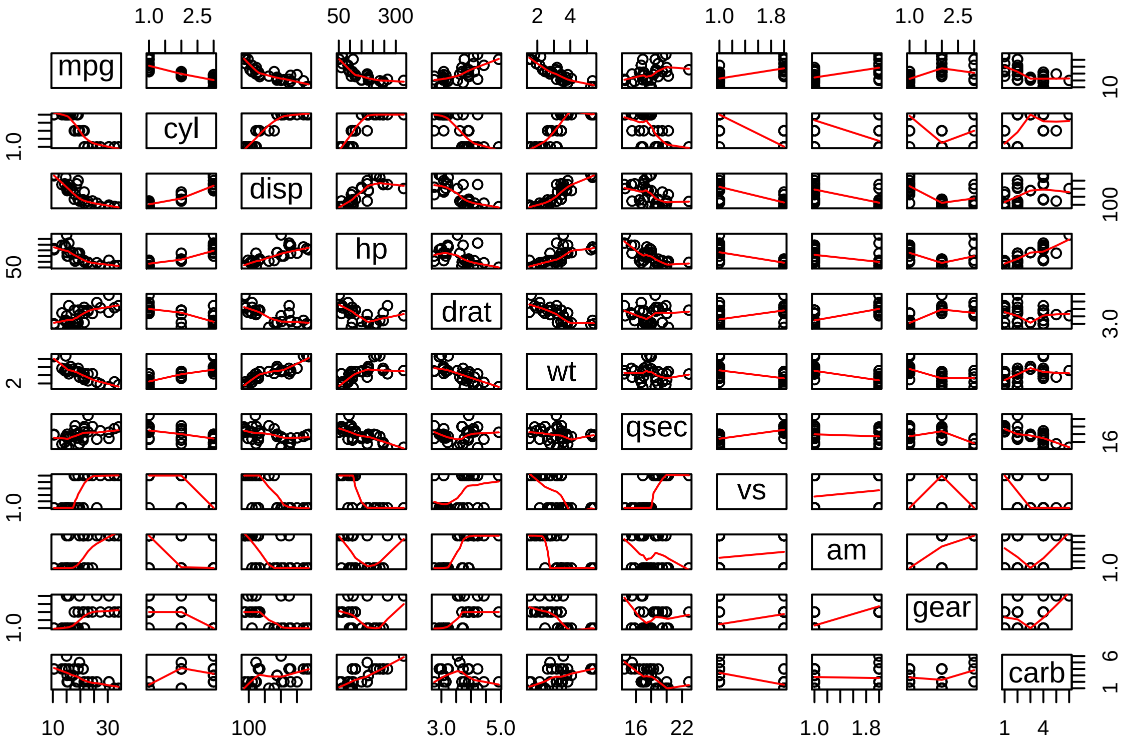


Figure 2

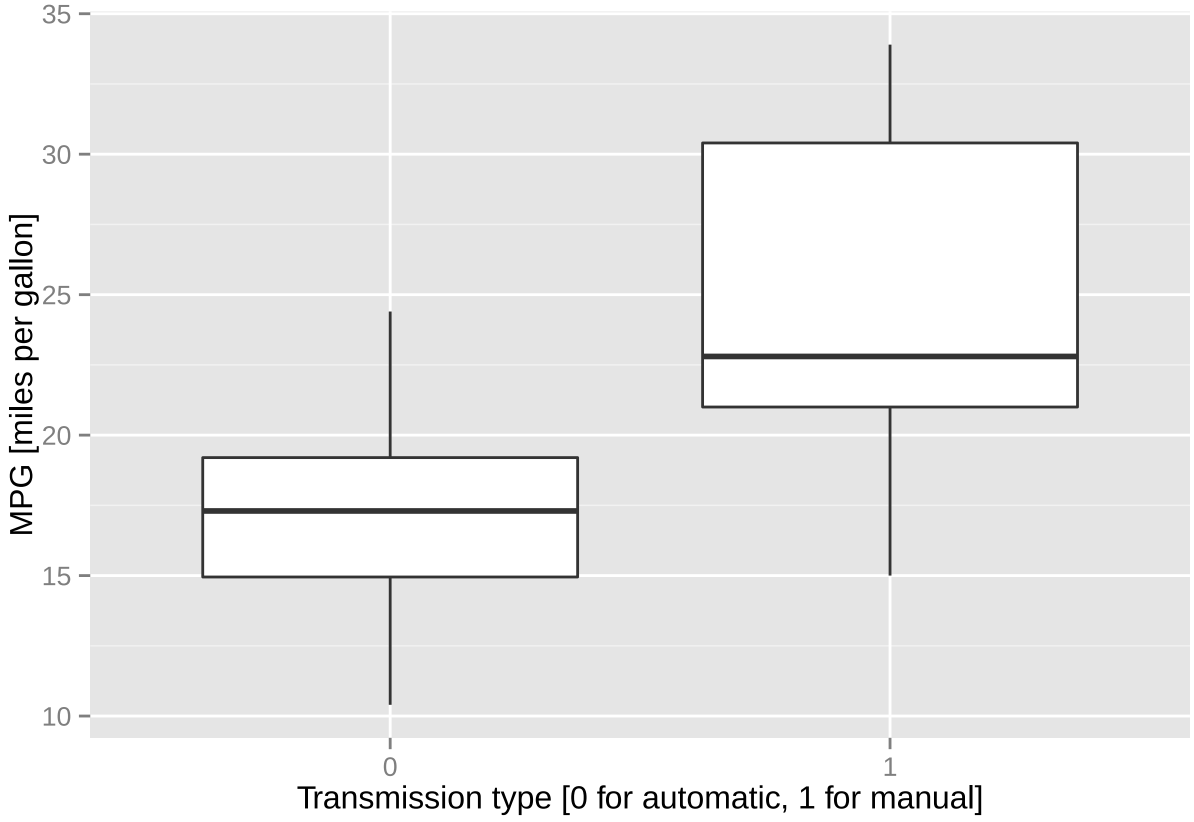
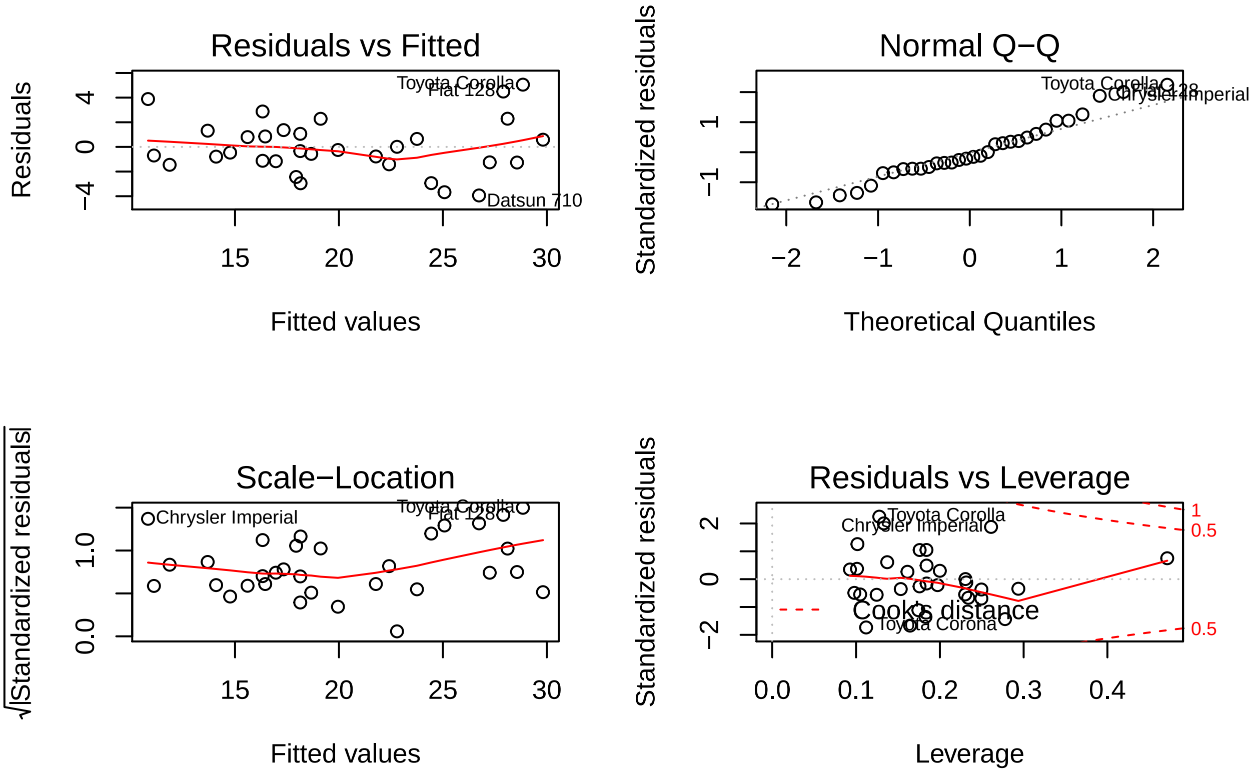


Figure 3


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5