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.

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.

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

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am1          7.244939   1.764422  4.106127 2.850207e-04
```

In this context, first estimate is the mean mpg for automatic transmission and second estimate 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.

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”.

```r
null=lm(mpg~1, data=mtcars)
full=lm(mpg~., data=mtcars)
fit_best <- step(null, scope=list(lower=null, upper=full), direction="forward")
```

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

```r
summary(fit_best)$coeff
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## am1          1.80921138 1.39630450  1.295714 2.064597e-01
```

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

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 influence:

```
##     Toyota Corona Chrysler Imperial        Volvo 142E
##          3.096553          2.895524          1.929774
```

Cars with largest hatvalues are:

```
##       Maserati Bora Lincoln Continental       Toyota Corona
##           0.4713671           0.2936819           0.2777872
```

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 t-quantile:

```
sumCoef <- summary(fit_best)$coefficients
sumCoef[6,1] + c(-1,1) * qt(.975, df=fit_best$df) * sumCoef[6,2]
```

```
## [1] -1.060934  4.679356
```

The uncertainty in our conclusion that MPG for manual transmission is better by 1.81 MPG is in the interval [-1.06, 4.68] MPG.

```
sumCoef[6,1] + c(-1,1) * qt(.89, df=fit_best$df) * sumCoef[6,2]
```

```
## [1] 0.05433494 3.56408783
```

This is the interval that quarantees that manual transmission has better MPG with 78 % confidence.

Proof that the report was done in Rmd (knitr)

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

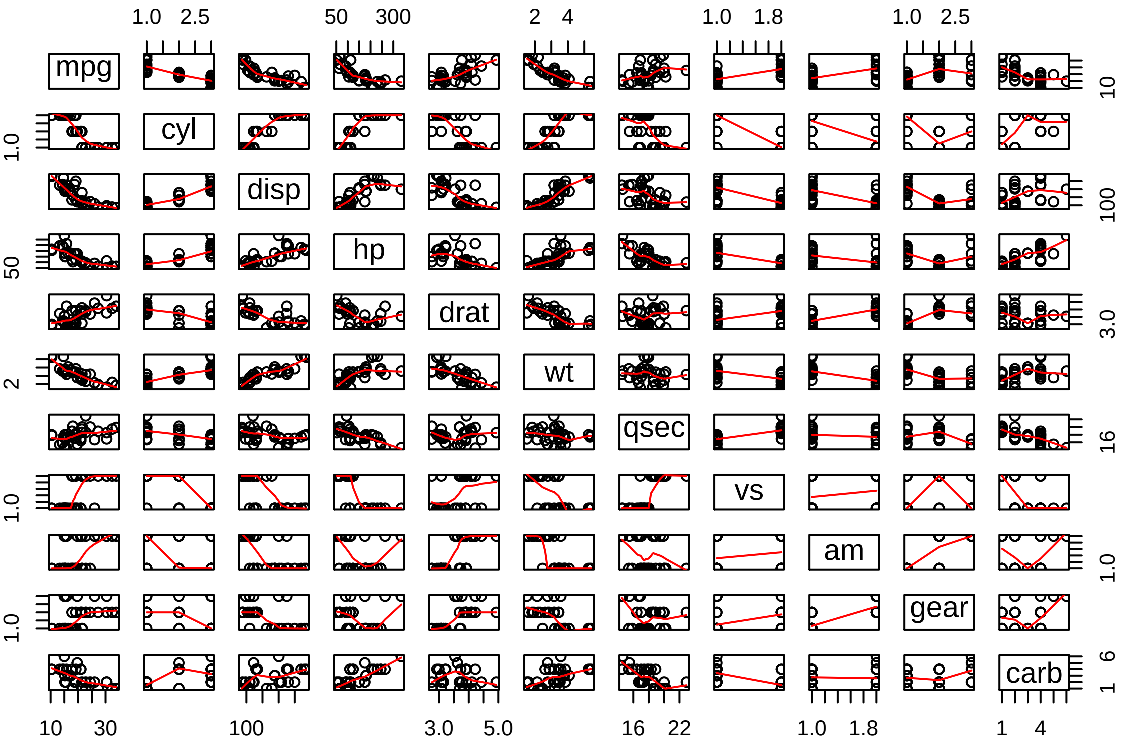


Figure 2

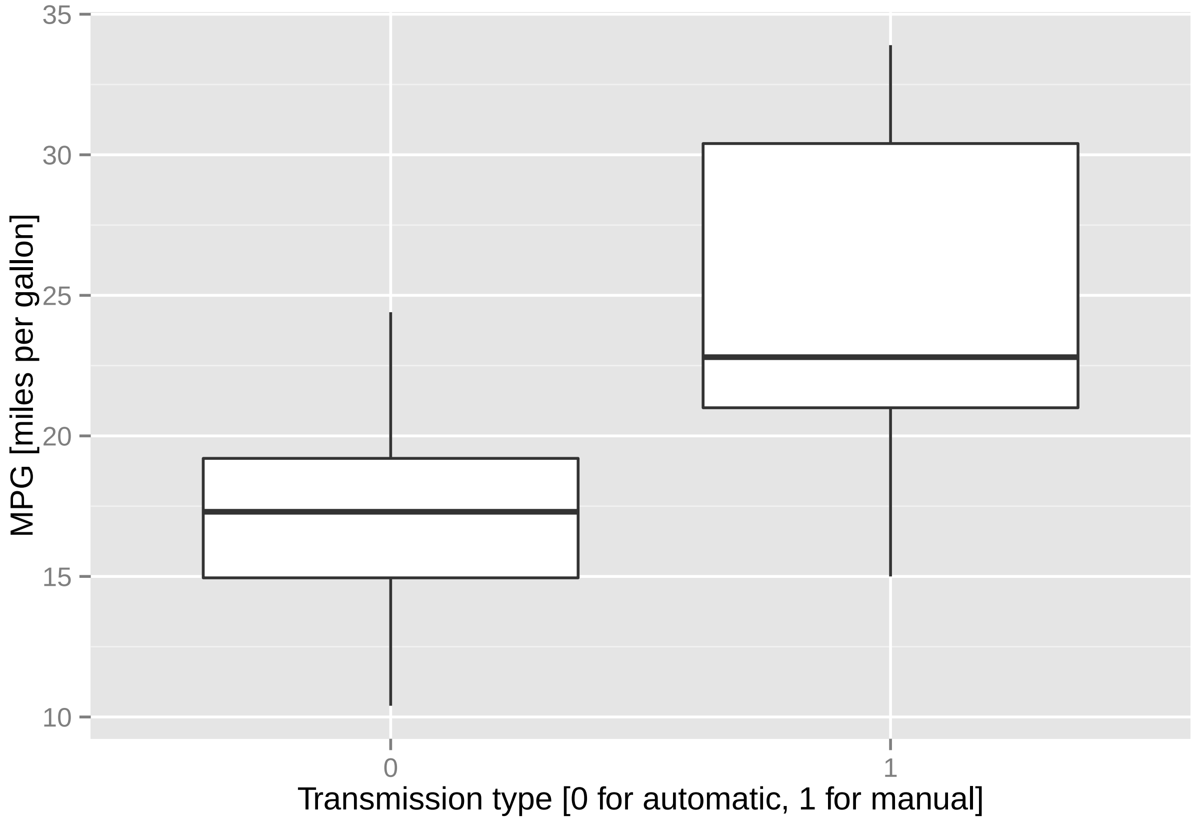
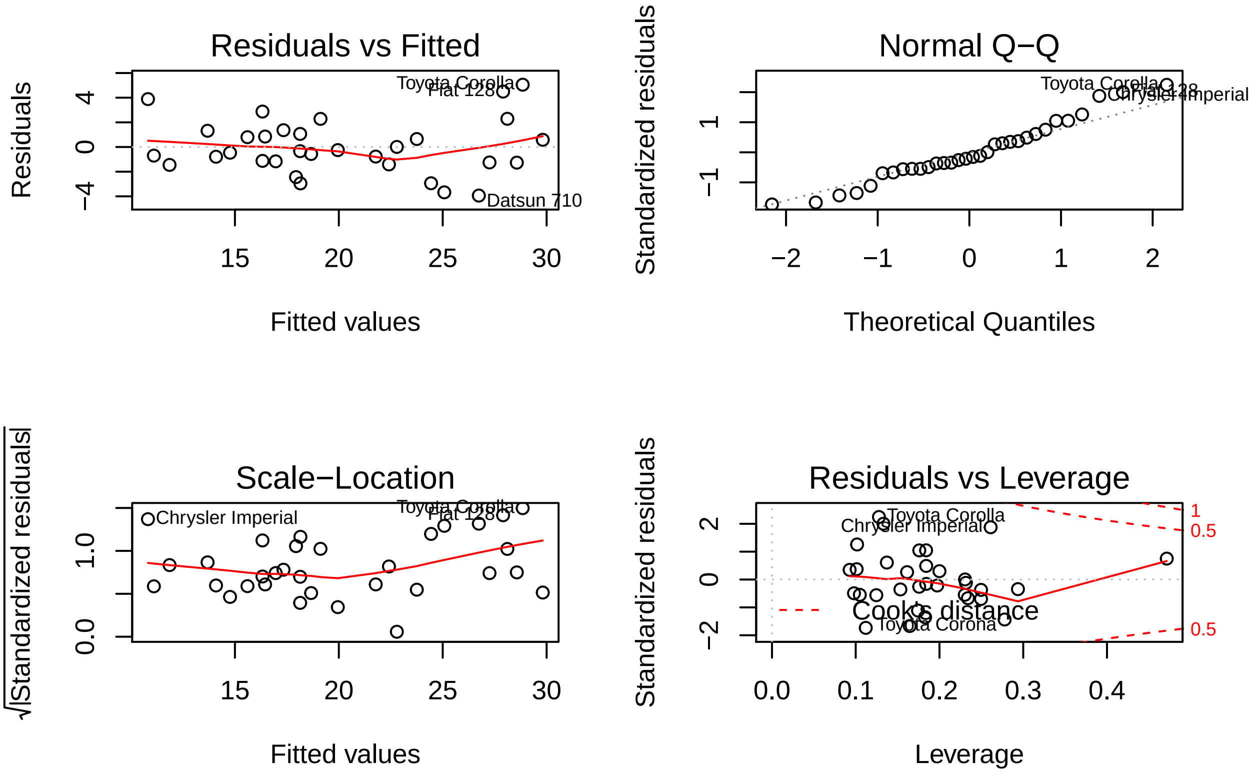


Figure 3


Residuals vs Fitted
Residuals
Fitted values
−4
0
4
15
20
25
30
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−1
1
−2
−1
0
1
2
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
0.0
1.0
15
20
25
30
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
−2
0
2
0.0
0.1
0.2
0.3
0.4
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5