Exploring the relationship between miles per gallon and transmission type

Evgeniy Zabrodskiy

18 December 2015

Executive summary

The analysis has shown that miles per gallon does not depend on the transmission type.
The difference that can be observed without taking into account other predictors is explained by parameters which have quite obvious relation with the mpg outcome, such as the weight of the car (the more the weight of the car - the more gas is needed per mile), the horse power (the more powerful the engine - the more gas it uses per mile).

Exploratory analysis

All the figures related to exploratory analysis are provided in the Appendix (dataset structure, first rows of the data and some plots).
We may assume that number of cylinders (cyl), horse power (hp) and weight (wt) are highly correlated with each other and with the miles per gallon (mpg). There are other parameters which presumably have little effect on the outcome due to lack of correlation. To prove this point, we can look at correlations of parameters using `ggcorr` function (the plot is provided in Appendix).

Model selection

We'll fit several models, subsequently adding predictors and test the models using the analysis of variance (anova).
Since our main questions are about the relationship of MPG and transmission type, the first model will only include one predictor: *factor(am)*. The model will show the mean MPG for cars with automatic transmission (Intercept) and the difference of the mean MPG for cars with manual transmission (factor(am)1 coefficient) from the atomatic ones.

```
## lm(formula = mpg ~ factor(am), data = mtcars)
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

Both coefficients are strongly significant (p-values $<< \alpha = 0.05$). However, we intentionally omitted all the other predictors and the model has high bias (R-squared is equal to 0.3597989 which is quite low).
Next several models are nested, i.e. each subsequent model contains additional predictor. If new predictor does not improve the model (i.e. p-value for the F-statistic in anova test was bigger than 0.05) than we fit new model with another predictor instead. As a result we find a model which is cosidered the best.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + hp
```

```
## Model 3: mpg ~ factor(am) + hp + wt
## Model 4: mpg ~ factor(am) + hp + wt + factor(cyl)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 245.44  1    475.46 81.8529 1.634e-09 ***
## 3     28 180.29  1     65.15 11.2157  0.002484 **
## 4     26 151.03  2     29.27  2.5191  0.099998 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Predictors *disp*, *drat*, *qsec*, *vs*, *gear* and *carb* did not improve the model. Predictor *factor(cyl)* was useful at some point but trying to use it in different order did not lead to significant improvement compared to the model with *hp* and *wt* only. It is seen from the parameters of the model 4: the p-value is bigger than 0.05. So, **we choose previous model 3 for further analysis**. It is worth mentioning that variance inflation increased after adding *wt* in model 3 due to its high correllation with *hp*.

Model diagnostics

We need to check the normality of the residuals in order to be sure that `anova` results can be trusted. We'll use `shapiro.test` on resuduals of the chosen model.

```
##
##  Shapiro-Wilk normality test
##
## data:  fit3$residuals
## W = 0.9453, p-value = 0.1059
```

The p-value is bigger than 0.05 which means that we fail to reject the null-hypothesis (the data is normally distributed), thus the model can be accepted.
Multiple R-squared: 0.8398903 which means that most of the variability is explained by the model.
Looking at model diagnostics plots (provided in Appendix), we can see that there are a few cars that have parameters with high leverage and influence but even removing them didn't change the answers for the main questions of the analysis.

Questions and Answers

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 34.00287512 2.642659337 12.866916 2.824030e-13
## factor(am)1  2.08371013 1.376420152  1.513862 1.412682e-01
## hp          -0.03747873 0.009605422 -3.901830 5.464023e-04
## wt          -2.87857541 0.904970538 -3.180850 3.574031e-03
```

From the coefficients we can see that for manual transmission (factor(am)1), the estimated change of *mpg* is 2.0837101 compared to automatic transmission with *hp* and *wt* fixed. However, **looking at p-value = 0.1412682 which is bigger than 0.05, or looking at 95% confidence interval for the estimate -0.7357587, 4.903179 which contains 0, the change in *mpg* is not statistically significant**.
Based on this result, the answers for the questions of the analysis are provided below:

1. Is an automatic or manual transmission better for MPG
   There is no difference.

2. Quantify the MPG difference between automatic and manual transmissions
   The MPG difference is not statistically significant.

Appendix

Exloratory summaries

Dataset structure:

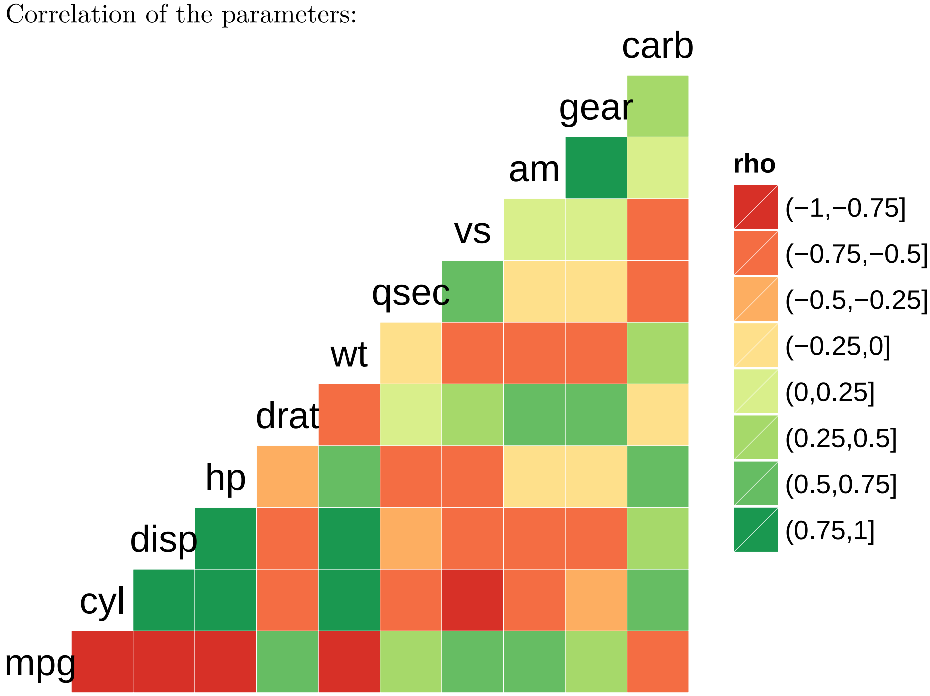
```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

First few rows of the dataset:

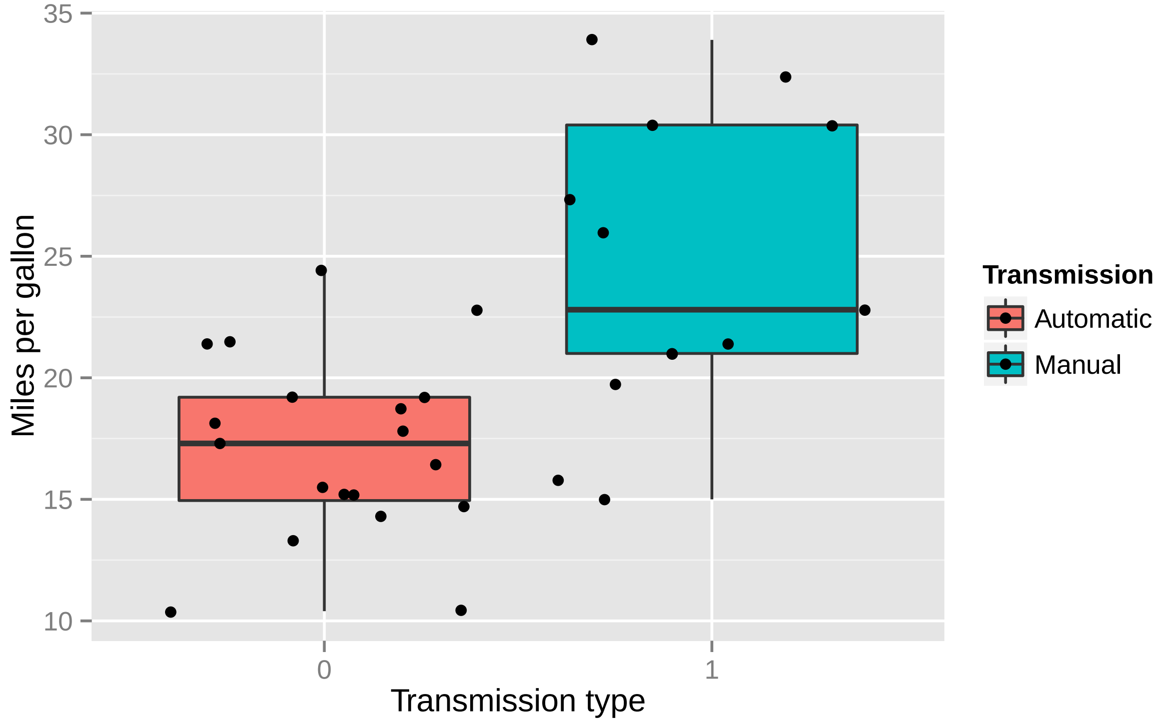
```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Exploratory plots

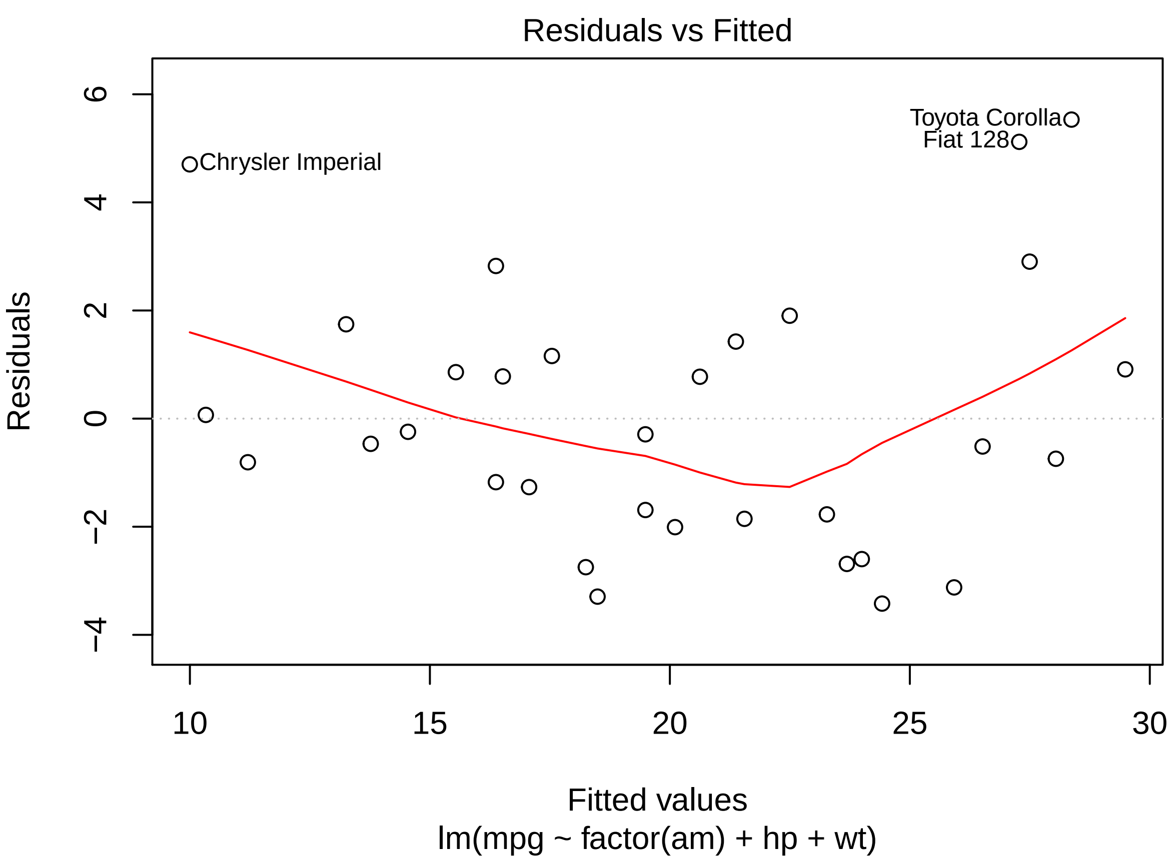
Correlation of the parameters:

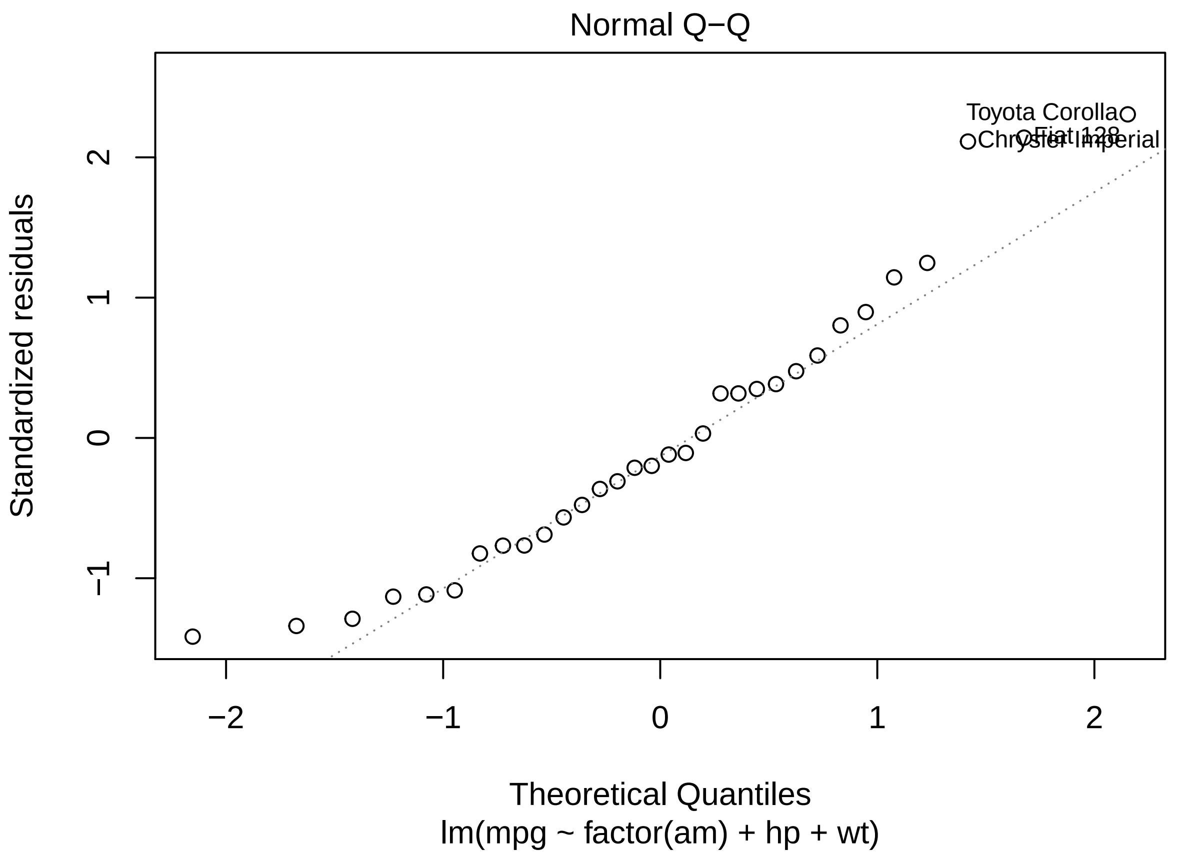


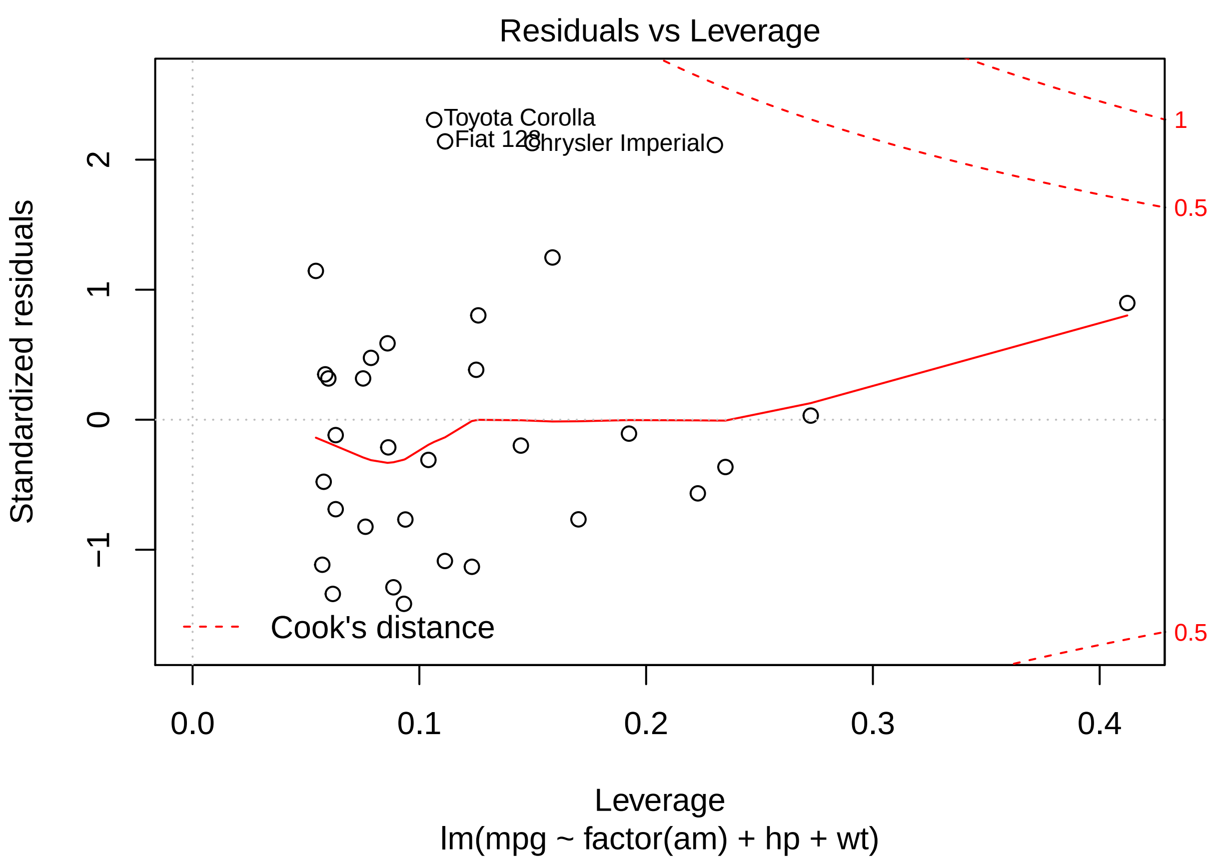
Miles per gallon for each transmission type:



Model diagnostics plots:




Normal Q−Q
Toyota Corolla
Fiat 128
Chrysler Imperial
Standardized residuals
Theoretical Quantiles
lm(mpg ~ factor(am) + hp + wt)


Residuals vs Leverage
Toyota Corolla
Fiat 128
Chrysler Imperial
Cook's distance
1
0.5
0.5
Standardized residuals
Leverage
lm(mpg ~ factor(am) + hp + wt)