Regression Models - Course Project

Dennis van den Berg

21/09/2015

Abstract

This paper investigates the relationship of automatic versus manual transmission in cars on petrol mileage. After exploratory analysis of the dataset containing 11 variables measured for 32 different car models, a parsimoneous regression model was selected using several Anova nested model searches, which suggested an optimal regression model that also adjusts for car weight and horsepower. We found a higher petrol mileage (+2.1 +/- 1.4 mpg) for manual transmission compared to automatic, but this result was not significant (p=0.07 whereas alpha=0.05). Furthermore a significantly lower petrol mileage for increased weight (-2.9 +/- 0.9 mpg per 1000 lb) and also for increased horsepower (-0.037 +/- 0.010 mpg per hp) was observed. Residual plots and a Shapiro-Wilk normality test suggest there could still be some pattern left in the residuals (likely caused by 3 outliers in the data) and might be explained using a more complex regression model.

Introduction

For Motor Trend magazine, we analyzed petrol mileage data of cars in miles per gallon (mpg) and its relationship to a set of car properties. Specifically, our goal was to answer these 2 questions:

1. "Is an automatic or manual transmission better for mpg"
2. "Quantify the mpg difference between automatic and manual transmissions"

For this purpose we used the mtcars dataset of Henderson and Velleman, available from Base R. It contains 32 observations on 11 variables:

```
mpg   Miles/(US) gallon           qsec  1/4 mile time
cyl   Number of cylinders         vs    V/S
disp  Displacement (cu.in.)       am    Transmission (0 = automatic, 1 = manual)
hp    Gross horsepower            gear  Number of forward gears
drat  Rear axle ratio             carb  Number of carburetors
wt    Weight (lb/1000)
```

Exploratory Analysis

Not correcting for other variables shows that manual transmission (am=1) gives a significantly (p=0.000285) higher petrol mileage (+7.245 +/- 1.764 mpg) than automatic (am=0). This can be seen in the `mpg ~ am` plot in Appendix A. Note that throughout our analysis we will use significance threshold alpha=0.05 in order to determine significance.

However, from the other plots in the appendix we can see that petrol mileage is not solely related to transmission type, but that there are heavy correlations with weight (wt), horsepower (hp) and number of cylinders (cyl), for instance. Furthermore we observe that transmission type seems to be heavily correlated with these variables as well. This suggests a necessity to adjust for additional variables in a regression model.

Model Selection

We did several nested model searches (Anova), each of which compared progressively more complex models with additional regressors (in different orders):

```
mpg ~ factor(am)
..
mpg ~ factor(am) + cyl + disp + hp + drat + wt + qsec + vs + gear + carb
```

These led us to select `wt` as a necessary regressor (with p-value Pr(>F) *always* smaller than 0.05), `cyl`, `hp`, `carb`, `gear` and `vs` as potential regressors (p-value smaller than 0.05 only in *some* ordered searches), and to discard `qsec`, `drat` and `disp` as potentially relevant ones (p-value *never* smaller than 0.05).

A second round of Anova nested searches was performed by starting from the minimal model `mpg ~ factor(am) + wt` and adding the potential regressors in several different orders again. This led us to choose the following as the best (parsimonious) model for our investigation:

```
mpg ~ factor(am) + wt + hp
```

We have to note that adding any of the `hp`, `cyl`, `vs` and `carb` variables as regressor to `mpg ~ factor(am) + wt` seems to be an improvement. We choose `hp` because it gives the *most significant* improvement. Furthermore, adding any of the remaining variables to `mpg ~ factor(am) + wt + hp`, as well as replacing `factor(am)` by an interaction term `factor(am)*gear` did not improve the model.

Results

```
##                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
## wt          -2.87857541 0.904970538 -3.180850 3.574031e-03
## hp          -0.03747873 0.009605422 -3.901830 5.464023e-04
```

Performing a linear regression with the above model we find that weight (wt) and horsepower (hp) are significant regressors (with p-values 0.0036 and 0.00055 much lower than our significance level alpha=0.05). We find negative coefficients for these regressors, pointing out a significantly lower petrol mileage for increased weight (-2.9 +/- 0.9 mpg per 1000 lb) and also for increased horsepower (-0.037 +/- 0.010 mpg per hp).

The positive coefficient for transmission (am) means more petrol mileage (**+2.1 +/- 1.4 mpg**) for manual transmission compared to automatic, making manual transmission the better option for all other variables kept equal. However, due to a high p-value (p=0.14/2=0.07 for a one-sided t-test, which is above treshold alpha=0.05) we fail to reject the null hypothesis that the transmission coefficient is equal or less than 0, so we conclude that **manual transmission does not *significantly* influence petrol mileage**, when adjusted for weight and horsepower.

Discussion

Performing a Shapiro-Wilk normality test on the residuals of the investigated model failed to reject normality (albeit with a fairly low p-value of 0.11), suggesting that the Anova analysis used during the model selection was valid. Several residual plots (see appendix A) show that there is still some pattern left in the residuals, however. This could explain the fairly low p-value of the Shapiro-Wilk normality test.

Note that we could have used other or additional variables as regressors, but that this model gave best results in our Anova search with a minimal number of regressors, provided that we included transmission. Choosing a slightly different set of regressors (for examplen using `cyl`, `vs` or `carb` instead of `hp`) might change the regression coefficient for transmission type (am). It is interesting to note that we investigated these and found that all of these alternatives led to *positive though insignificant* regression coefficients for transmission type. This suggests robustness of our findings.

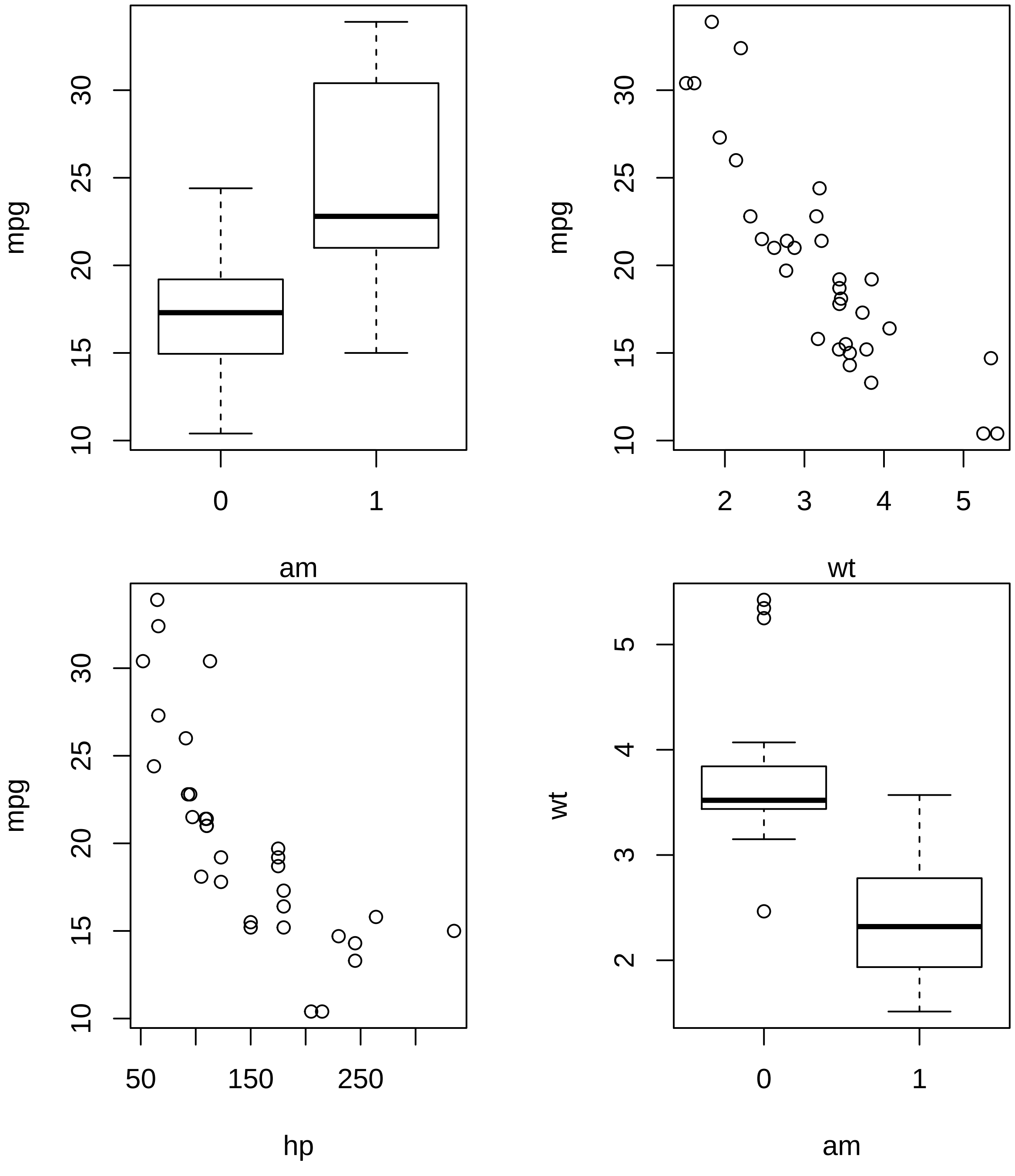
In a "Residuals vs Fitted" plot three points in particular could be considered potential outliers (Chrystler Imperial, and Toyota Corolla and Fiat 128). All of these are located in the extreme regions of the weight/horsepower variables (either in high horsepower / high weight region, or in the low horsepower / low weight region). It is possible that a model with non-linear regressors would be better suited in fitting these extremes. We did not investigate this however.

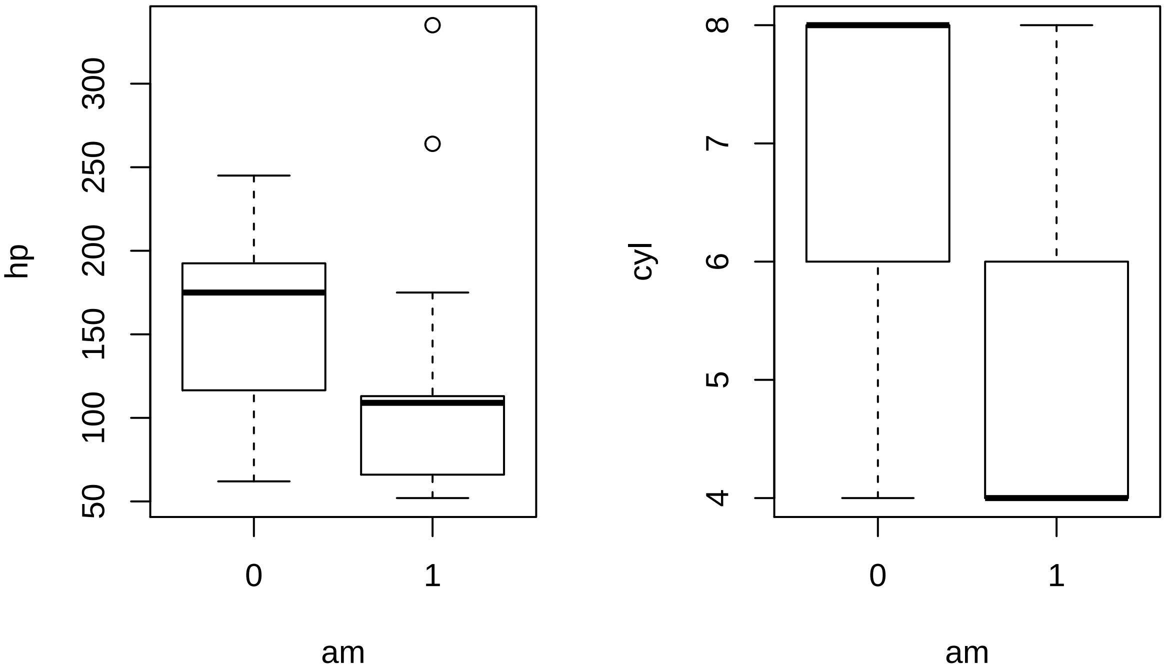
Sources

- Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391–411.

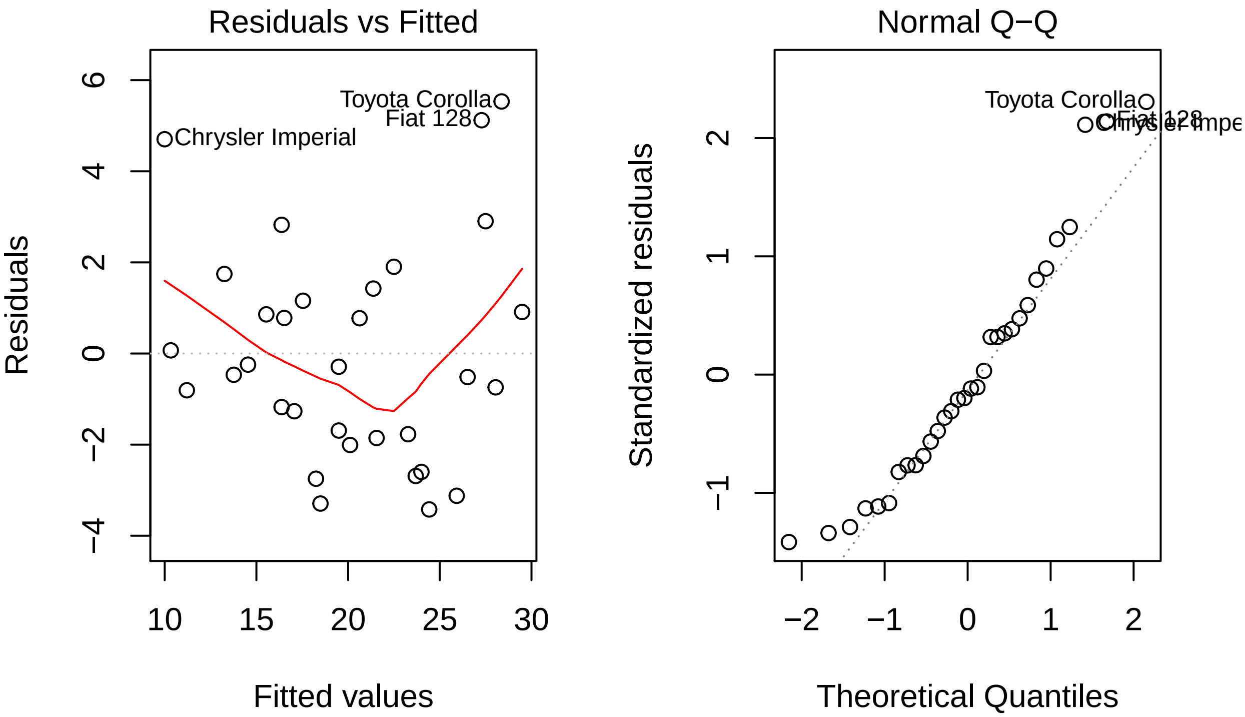
Appendix A: Figures

Exploratory plots from mtcars dataset





Residual plots and diagnostics on linear model


Residuals vs Fitted
Residuals
Toyota Corolla
Fiat 128
Chrysler Imperial
Fitted values
Normal Q−Q
Standardized residuals
Toyota Corolla
Fiat 128
Chrysler Impe
Theoretical Quantiles