Motor Trend

Ariane

Sunday, October 26, 2014

Summary

We wonder whether an automatic or a manual transmission is better for the MPG of a car. From the data set mtcars, we cannot answer this question. Indeed, cars with a manual transmission seem to have a higher MPG. However, the probability that this hypothesis is false is rather high.

Exploratory analysis

We have a data set of a collection of cars. There are eleven variables:

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

We are interested in the effect of an automatic or a manual transmission on the MPG (miles per gallon). We first draw a panel of the 6 continuous variables (mpg, disp, hp, drat, wt, and qsec) to check if they are correlated (see Figure 1 in Appendix). We also compute the correlations between them:

```
##          mpg    disp      hp    drat      wt    qsec
## mpg   1.0000 -0.8476 -0.7762  0.6812 -0.8677  0.4187
## disp -0.8476  1.0000  0.7909 -0.7102  0.8880 -0.4337
## hp   -0.7762  0.7909  1.0000 -0.4488  0.6587 -0.7082
## drat  0.6812 -0.7102 -0.4488  1.0000 -0.7124  0.0912
## wt   -0.8677  0.8880  0.6587 -0.7124  1.0000 -0.1747
## qsec  0.4187 -0.4337 -0.7082  0.0912 -0.1747  1.0000
```

The variables disp, hp, drat, and wt are highly correlated between each other and with the variable mpg. The variable qsec is less correlated with mpg. The variable wt is the more correlated with mpg. Therefore, we will keep only this one in our regression model.

We then do a box-and-whisker plot of the data for each factor variable (cyl, vs, am, gear, and carb) to check if these variables have an effect on mpg (see Figure 2 in Appendix). The variables cyl, vs, and am seem to have more effect on the mpg than the variables gear and carb. Therefore, we will only keep the three first variables in our regression model.

Regression models

We first test a model with only wt as continuous variable, and cyl, vs and am as factor variables.

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  32.1811     3.7286  8.6309 4.140e-09
## wt           -3.1061     0.9203 -3.3751 2.326e-03
## cyl6         -3.7329     1.6376 -2.2795 3.108e-02
## cyl8         -4.7483     2.6572 -1.7869 8.561e-02
## vs1           1.2418     1.9042  0.6521 5.200e-01
## am1           0.6232     1.5012  0.4151 6.814e-01
```

We test a second model with all the variables and improve it using the R function "step".

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 33.70832    2.60489 12.9404 7.733e-13
## cyl6        -3.03134    1.40728 -2.1540 4.068e-02
## cyl8        -2.16368    2.28425 -0.9472 3.523e-01
## hp          -0.03211    0.01369 -2.3450 2.693e-02
## wt          -2.49683    0.88559 -2.8194 9.081e-03
## am1          1.80921    1.39630  1.2957 2.065e-01
```

The R^2 associated with the first model is equal to 0.8401 and the R^2 associated with the second model is equal to 0.8659. These are rather high values, therefore the models seem to fit well the data, the second model being a little better than the first model.

We make a plot of the residuals versus the fitted values for the two models (Figures 3 and 4 in Appendix) and we cannot see evidence of a correlation between the residuals and the fitted values. Therefore, we may think that we did not forget to take into account a variable that should have been included in the model.

We compare the two models using the R function "anova".

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + cyl + vs + am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df RSS Df Sum of Sq F Pr(>F)
## 1     26 180
## 2     26 151  0        29
```

The second model has a lower RSS. Therefore, it allows fitting better the data. For the two models, the regression coefficient associated to am = 1 (manual transmission) is positive. It means that the mpg is higher for manual transmission than for automatic transmission. However, the probability that the regression coefficient associated to am = 1 is negative is equal to 0.3407 for the first model, and to 0.1032 for the second model. This is a rather high probability. Therefore, we cannot eliminate the hypothesis that the coefficient is negative, that is that automatic transmission is better than manual transmission.

The data in the mtcars data set are therefore not sufficient to answer the question whether an automatic or a manual transmission is better for MPG.

Appendix

Figure 1. Motor cars

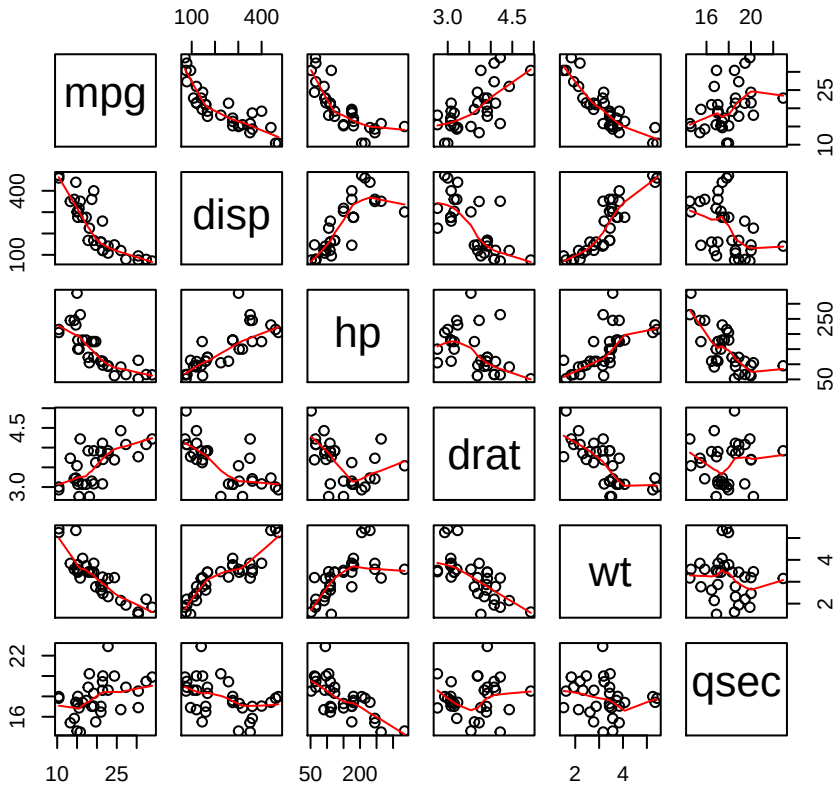

Figure 2. MPG

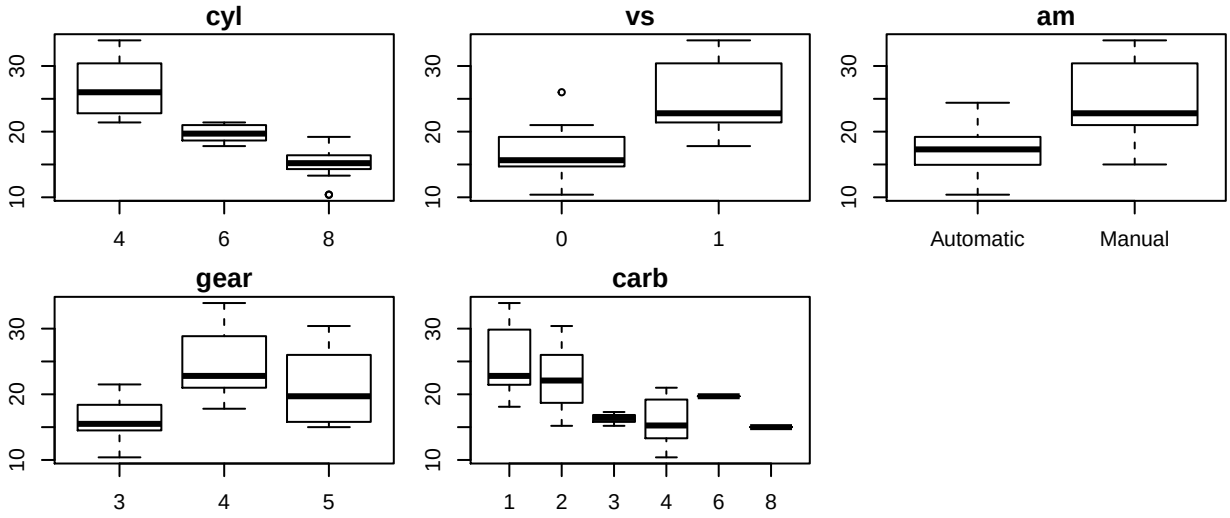

Figure 3. Model 1

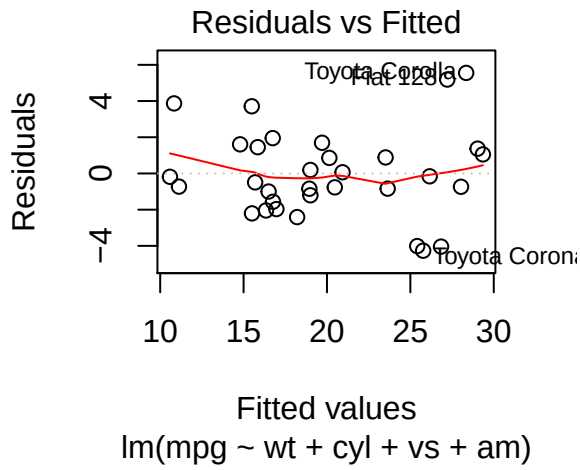

Figure 3. Model 1

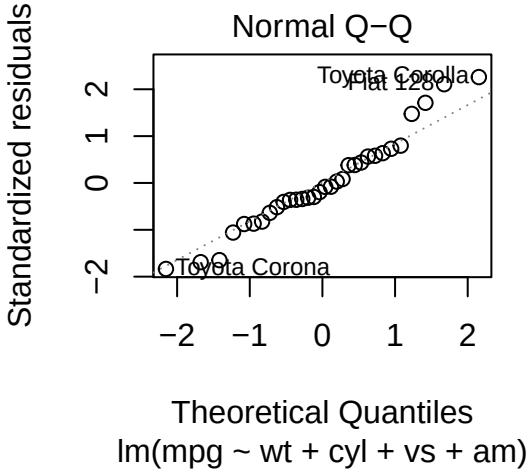

Figure 3. Model 1

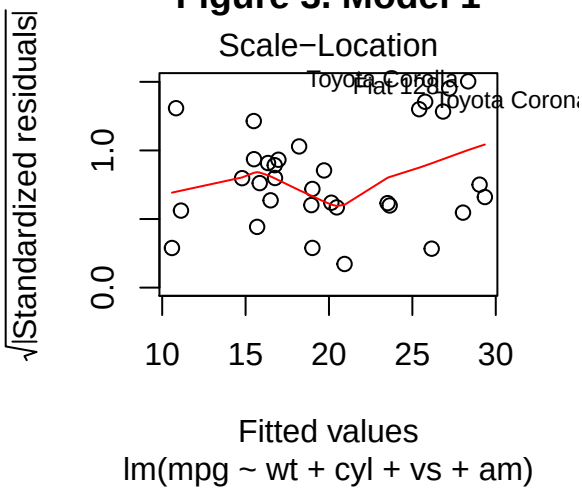

Figure 3. Model 1

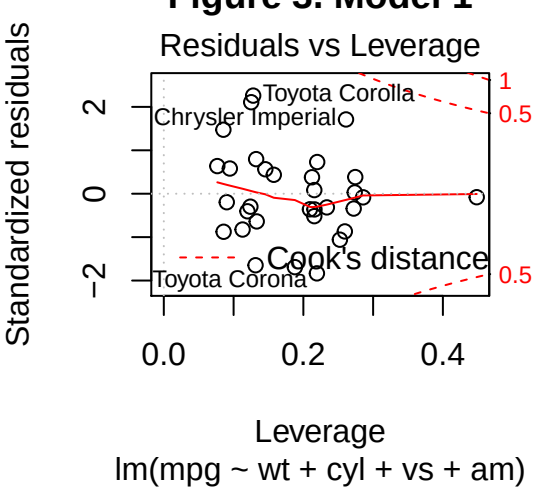

Figure 4. Model 2

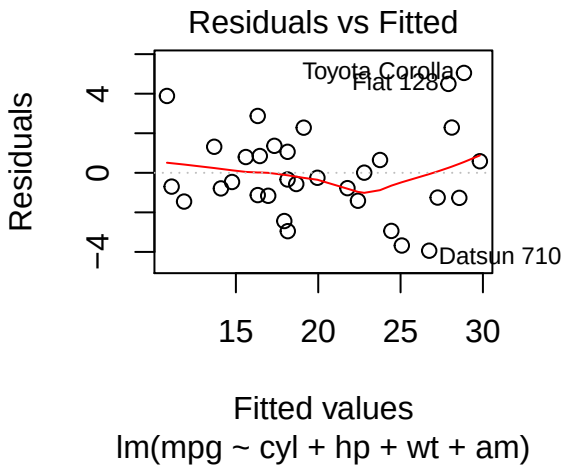

Figure 4. Model 2

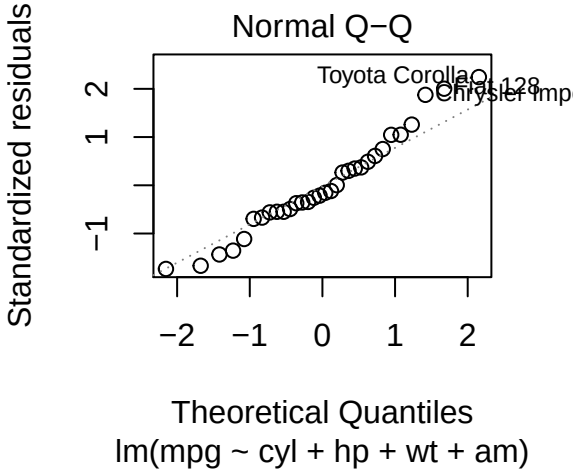

Figure 4. Model 2

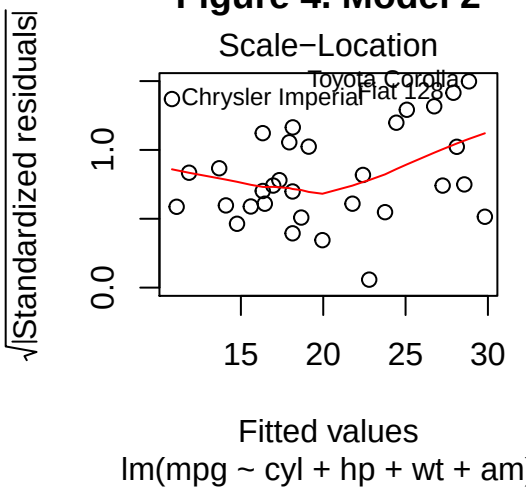

Figure 4. Model 2