Peer-graded Assignment: Regression models

Steven Michiels

4/20/2020

Executive summary

Questions to be answered: 1) "Is an automatic or manual transmission better for MPG" and 2) "Quantify the MPG difference between automatic and manual transmissions".

A two-sided t-test showed that the mpg for manual tranmission is signicantly (p=0.001) higher than for automatic transmission. A linear model with the transmission type (am) as only predictor, however, showed that the transmission type only explains roughly 1/3 of the variability in mp.

A three-predictor linear model including the weight, the qsec and the am achieves to explain aournd 85% of the variability in mpg, and weight appears to be the most important predictor. When accounting for the weight and for the qsec, the linear coefficient of am is 2.9+/-1.4, which means that moving from automatic to manual transmission increased the mpg with a value of 2.9 +/- 1.4.

Exploratory data analyses

Boxplots and violinplots of the mpg in function of the transmission type suggest that there may indeed be a significant association between the transmission type and the mpg. See the appendix for these plots.

We make **subgroups** of the data per transmission type and perform a **two-sided t-test for the null-hypothesis that the mean mpg is not different between the subgroups**. At a significance level of .05, we reject this null-hypothesis with a p-value of .1%, meaning there is a **significant association between the transmission type and the mpg**.

Model building

One-variable model: transmission type

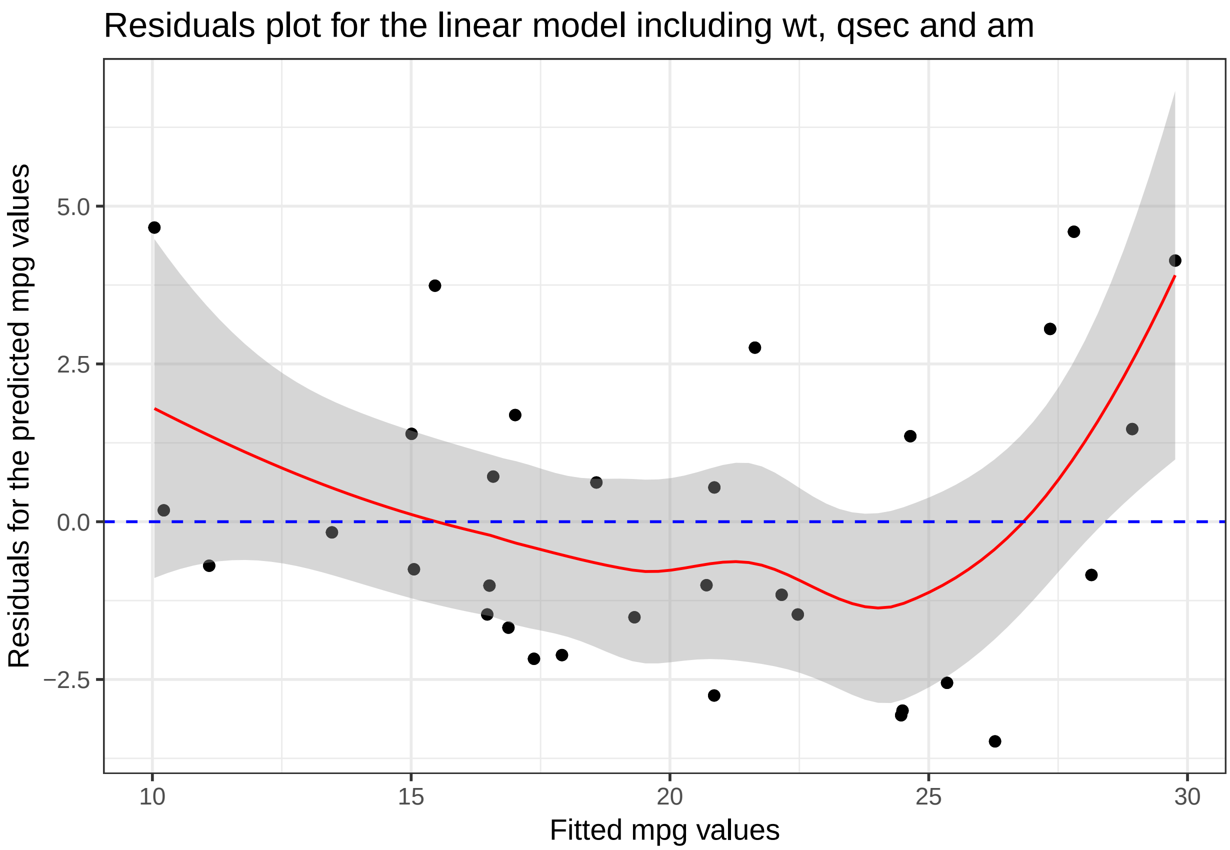
We create a linear regression with the transmission type as the predictor for the mpg. We find a **linear coefficient of 7.25 +/-1.8**, indicating that a change from automatic to manual transmission increases the number of miles per gallon with 7.25 +/-1.8. We find an **R^2 value and an adjusted R^2-value of only .36 and .34 respectively**, however, which means that **only roughly a third of the variability in mpg is explained by the transmission type**.

We make a **correlation plot** to see which other variables may affect the mpg as well. Besides the transmission type, **the weight and the number of cylinders, for instance, are candidate predictors** as well. Note that we'll have to be careful, as **considerable collinearity** between the independent variables exist.

Best subset three-variables model: wt, qsec and am

We make a **new linear model** using **best subset analysis** (function regsubsets) and take the number of variables with the **lowest BIC-value**. This yields **three variables** to be included for the new model: **the weigth, the time in seconds for 1/4 mile and the transmission type**. The linear coefficients are -3.9+/-.7 (wt), 1.2+/-.3 (qsec) and 2.9 +/- 1.4 (am). When accounting for other significant variables, the **effect of the transmission type** thus is **reduced**. The effect of the transmission type moreover has a **large standard error**. The **largest part of the variability (R^2=75%) of the variability in mpg is explained by the weight.**

The **R^2-value and the adjusted R^2 values for this model are .83 and .85**, respectively, which explains quite a bit of the variability for the mpg. When we plot the **residuals**, however, we do see a pattern in the residuals that are **indicating non-linearity or interaction terms** in the data.

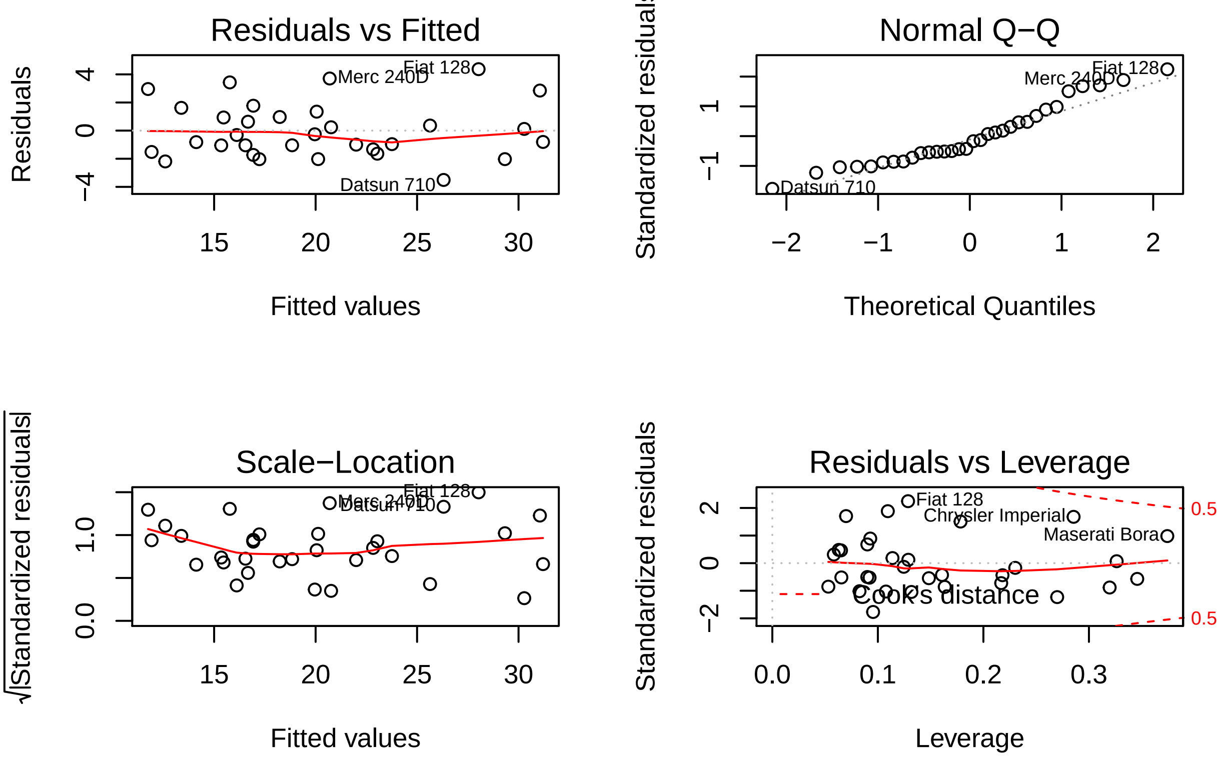


Tuned model

We **redo the best subset analysis**, but now **including interaction terms** for the weight, the number of seconds for 1/4 mile and the transmission type. The lowest BIC is now reached again for three variables: qsec, am and am/wt. The **R^2-value and the adjusted R^2 values for this model now reach .9 and .88**, respectively, which is better than all the previous models. The linear coefficients are: 2.68+/-1.3 (qsec), 14.0+/-3.4 (am) and -4.1+/-1.2 (am/wt). When we plot the **residuals**,we see that the pattern has disappeared.

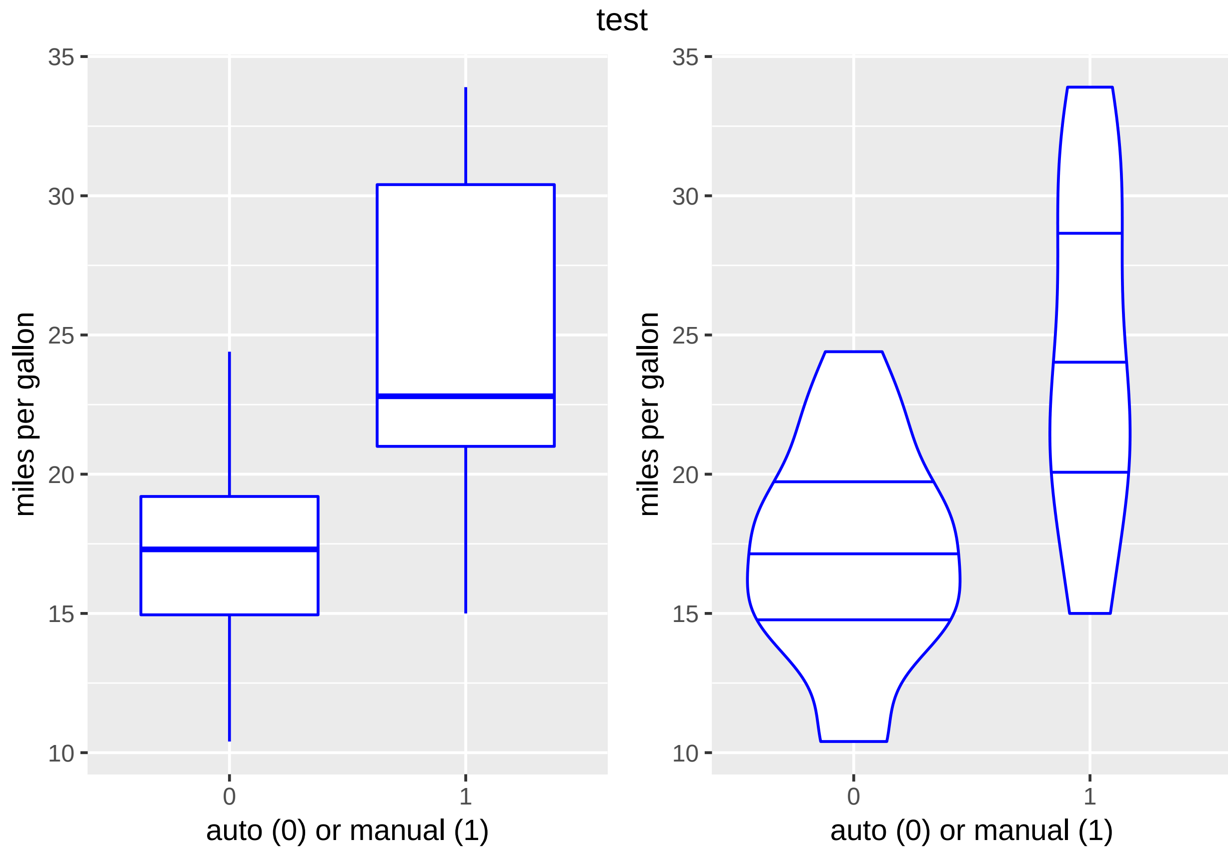
qsec 2.6831 1.3002 2.064 0.049171 *
am 14.0026 3.3918 4.128 0.000334 * **wt 6.6931 7.4051 0.904 0.374379**
am:wt -4.1411 1.1815 -3.505 0.001675 qsec:wt -0.5401 0.4137 -1.306 0.203141

```
par(mfrow=c(2,2))
plot(modSel2)
```

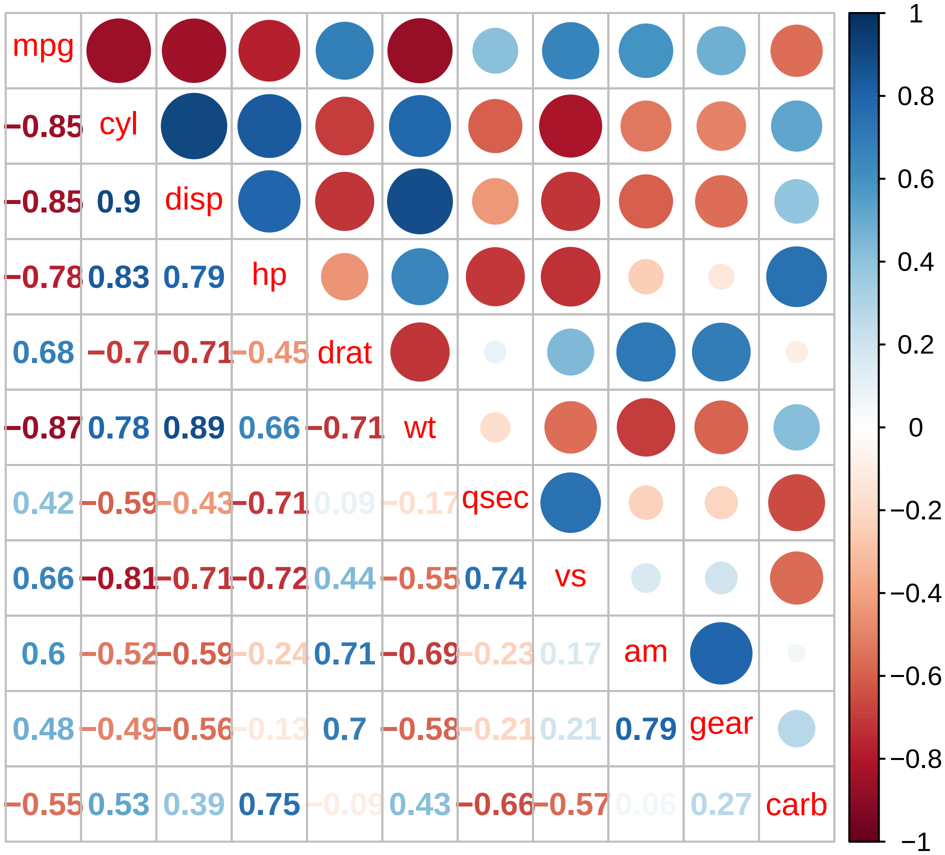

Residuals vs Fitted
Residuals
Fitted values
Merc 240D
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Chrysler Imperial
Maserati Bora
Cook's distance
0.5

Appendix

Boxplots and violinplots of the mpg in function of the transmission type suggest that there may indeed be a significant association between the transmission type and the mpg.

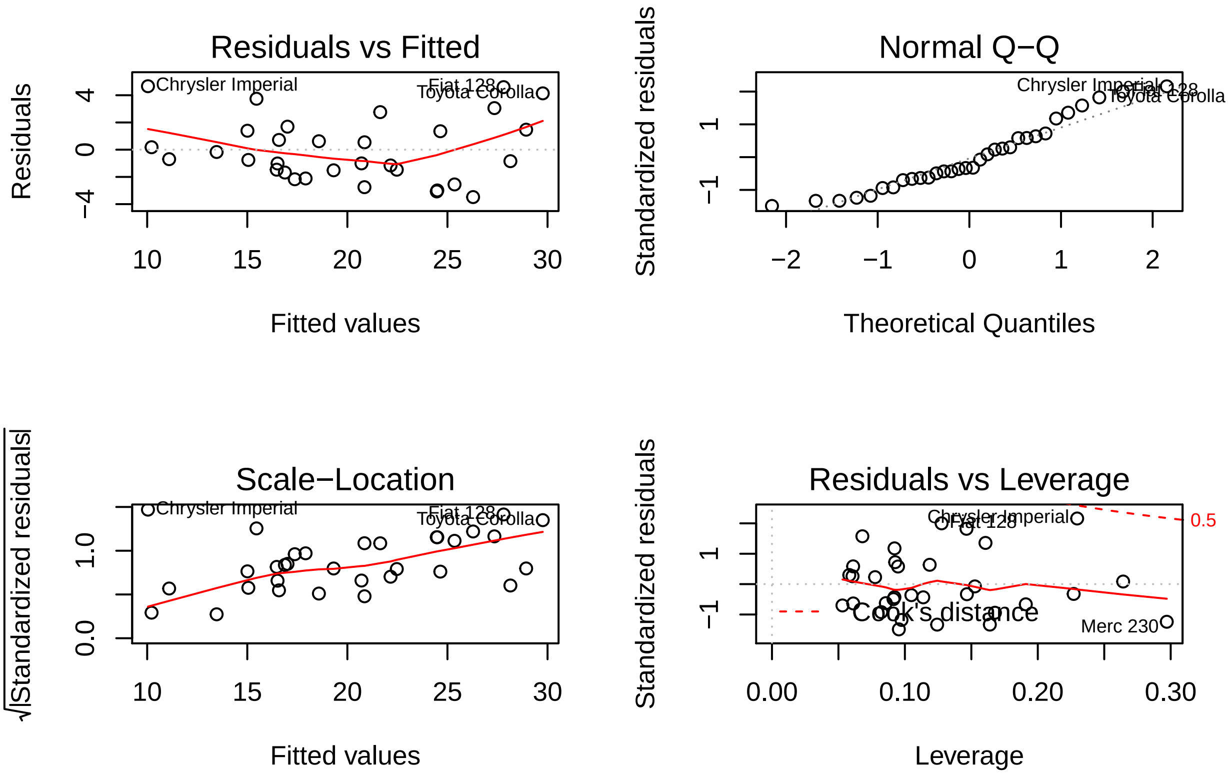


Correlationplot to identify candidate predictors



Based on the best subset analysis, we include three variables in the model: wt, qsec and am. We create model plots for this model.

```
## (Intercept)          wt        qsec          am
##    9.617781   -3.916504    1.225886    2.935837
```



Summary plots for the tuned model that includes qsec, am and an interaction term wt/am

```
par(mfrow=c(2,2))
plot(modSel2)
```