Course project - Regression. Analysis of transmission effects on mileage.

Lars Bungum

8 January 2017

```
knitr::opts_chunk$set(echo = TRUE)
```

Executive summary

This report has explored the mtcars dataset to assess the effect of transmission types on mileage. Treating the variable as a binary factor with weight as an independent coefficient that correlates a lot with many other factors, shows that there is a different relationship in the two groups. For manual cars, the effect of reduced weight on mileage is much more profound.

Hence, we can conclude that reducing the weight of the vehicle is very important if you opt for a manual drive, if high mileage is your objective, for the sake of the environment, say. For automatic cars the effect is not as big, and there might be other factors to look for when choosing an environmentally friendly car, such as materials, production facilities, etc.

Report structure

This report is intended to analyze the mtcars dataset in order to investigate the influence on automatic and manual gears ("am") on the mileage per gallon (MPG).

The report will begin with some exploratory data analysis, in order to select the best model to explain the dataset. It continues with some analysis, and ends with a summary. Finally, some plots are provided in the appendix.

All figures and tables are found in the Appendix, at the end of the document.

Exploratory data analysis

Statistics on the dataset

Table 1 lists 11 variables, of which mpg is to be treated as the dependent one. Some of them may be interpreted as categorical, notably the "am", signalling automatic/manual transmission, the amount of cylinders, the number of gears (3,4 or 5) and the number and possibly the amount of carburators.

Fitting a model based on only am, and all variables

```
fitam <- lm(mpg ~ am , data=mtcars)
fitall <- lm(mpg ~. , data=mtcars)
```

Table 2 summarizes the coefficients of this model. It shows that the number of variables introduce so high variance, that there is no statistically significant linear relationship between any coefficient when including all variables. Hence, a better model to fit the data is necessrary.

Analyzing the variance inflation factors and correlation between the coefficients.

Looking at the square root of the inflation factors (getting it measured in effect on the standard deviation), suggests that there is a significant correlation between some of these coefficient.

```
library(car)
sqrt(vif(fitall))
```

```
##      cyl     disp       hp     drat       wt     qsec       vs       am
## 3.920948 4.649757 3.135608 1.837014 3.894212 2.743712 2.228424 2.156035
##     gear     carb
## 2.314617 2.812249
```

Choosing an appropriate model

Because of correlation between coefficients, we retain wt, qsec and cyl and compare it with a model without cyl.

```
fitaqc <- lm(mpg ~ factor(am)+qsec+cyl , data = mtcars)
fit <- lm(mpg ~ factor(am)+wt+qsec , data = mtcars)
```

Figures 1 and 2 show residual plots of the two models. There seems to be a certain pattern in the residuals of the model with am+qsec+cyl (spread among high and low values), and wt is reintroduced to the model.

Table 3 summarizes the model. It can be seen that the R-squared value explains 80% of the variation, similar to the model using all the factors.

Anova comparison of the candidate models

Table 4 models fitam, fitall and fit with anova suggests that there is no statistically significant difference between the two last models, and hence that the parsimonious one of them is just as good. However, introducing the regressors disp and cyl does have a significant effect.

Statistical analysis of the influence of transmission type

Both for the models fitall and fit, there is a positive relationship between the transmission type (am) and mpg. When looking for confounding factors, the relationship is not reversed, but changed in magnitude. For the final model of choice. The summaries of all the above models, there is a clear positive relationship between manual transmission (am=1) and mpg albeit at different significance levels.

Training a simple model with a binary variable separating data points

We know that mpg is negatively related to the weight of the vehicle. Figure 2 plots the deta points with a color that signals whether or not it is using manual (dark blue) or automatic (light blue) transmission.

By training another linear model that treats mpg as a dependent variable and wt as the independent, the asterisk * "multiplies" this regression by calculating the slopes for the manual and automatic transmission groups, respectively.

```
fitmult <- lm(mpg ~ wt*factor(am)  , data = mtcars)
```

Table 5 summarizes the models. It provides two slopes, both for the group with automatic and for manual transmission. Plotting the two regression lines illustrates how the relation between weight and mpg differs between the two groups.

Figure 2 furthermore shows that the mpg is reduced when the weight of the vehicle at a significantly different pace between the groups. For the manual cars, the mpg goes down a lot faster as the weight goes up, while for the automatically transmitted cars, the mpg is not reduced as quickly as the weight increases.

The summary of the fitmult model shows that all the coefficients are statistically significant, including the relation between the weight and the factor. The effect of reducing weight on mileage is clearly different between the two groups of transmission systems. This makes sense, as manual gears gives the driver the opportunity to adjust driving style to conditions, which is likely to result in more fuel efficient than the algorithms available in the 1970s would provide. With recent technology I am not so sure!

Appendix

Table 1

```
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

Table 2

```
summary(fitall)$coef
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## am           2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

Table 3

```
summary(fit)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## factor(am)1  2.935837  1.4109045  2.080819 4.671551e-02
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
```

Table 4

```
anova(fitam, fitall, fit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 3: mpg ~ factor(am) + wt + qsec
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     21 147.49  9    573.40 9.0711 1.779e-05 ***
## 3     28 169.29 -7    -21.79 0.4432    0.8636
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 5

```
summary(fitmult)$coef
```

```
##                  Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    31.416055  3.0201093 10.402291 4.001043e-11
## wt             -3.785908  0.7856478 -4.818836 4.551182e-05
## factor(am)1    14.878423  4.2640422  3.489277 1.621034e-03
## wt:factor(am)1 -5.298360  1.4446993 -3.667449 1.017148e-03
```

Figure 1

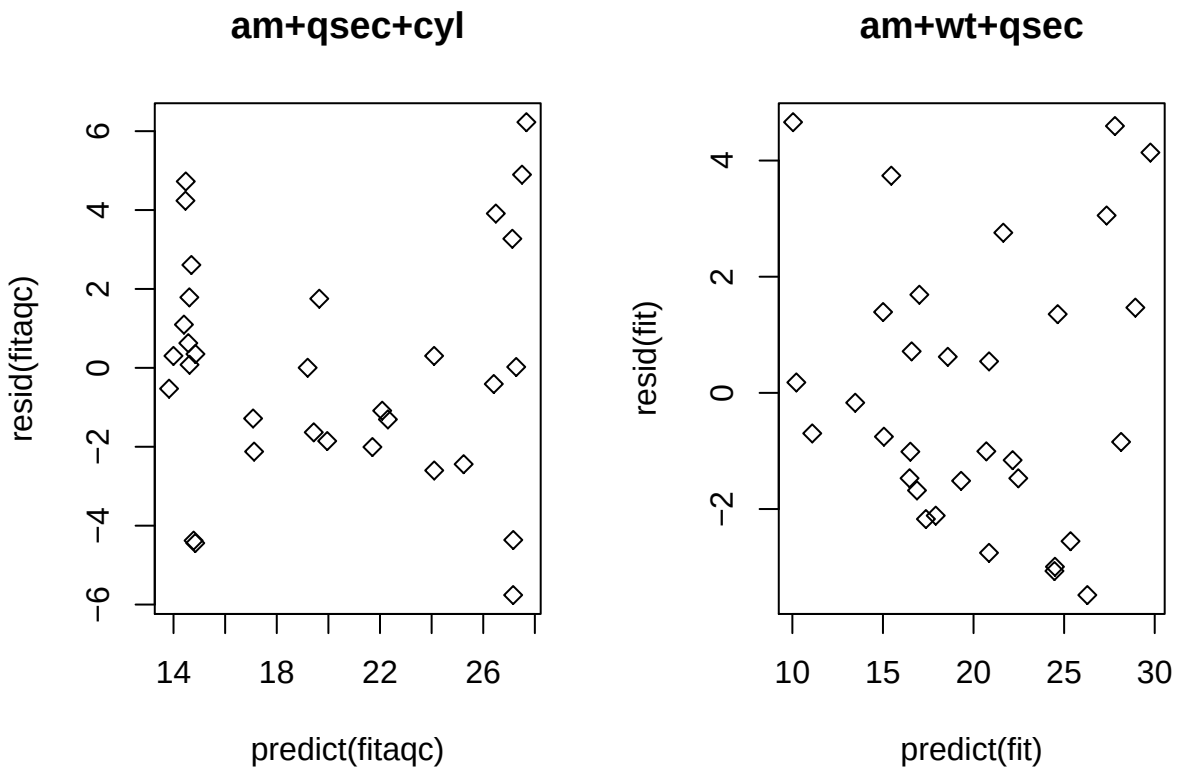

Figure 2