Regression Models - Course Project

Giovanni Picardi

30 gennaio 2015

Executive summary

Dataset mtcars has been explored trying to investigate how the variable `am`, an indicator variable of manual transmission, is related to `mpg` (miles per US gallon). Considering `am` alone shows a higher mean `mpg` for manual than for automatic transmission, but other variables have a stronger correlation with the outcome: weight `wt`, displacement `disp` and gross horsepower `hp`. Fitting and analyzing linear models with `am` and one of these variables (centered) as predictors, possibly considering interaction, leads to the conlcusion that no absolute difference of `mpg` can be attributed to transmission type for the average values of the other predictor (intercept), but only a different slope favouring automatic transmission, although in both cases transmission type has a too strong correlation with the other predictor, making the conclusion too tied to the linear model hypothesis. On the other hand a simple model considering `am` and `hp`, with transmission type equally distributed in the range of the `hp` values in the dataset, reveals a higher expected `mpg` for manual transmission for a given `hp`.

Exploratory data analysis

The datasets has no missing values, all of the 11 variables are numeric and they are related to a range of different motorcars (32 models). A comparison of boxplots of the variable `mpg` for automatic and manual transmission (**Figure 1**) shows that the mean mileage per gallon is higher for manual transmission (hypothesis tested in the following section). The outcome `mpg` shows the following correlation coefficients with the other variables:

```
##     wt    cyl   disp     hp   drat     vs     am   carb   gear   qsec
## -0.868 -0.852 -0.848 -0.776  0.681  0.664  0.600 -0.551  0.480  0.419
```

The relationship between pairs of variables, outcome inlcuded, can be visually examined with `pairs(mtcars)` (**Figure 2**): the more "linearly" related to `mpg` are the weight `wt`, the displacement `disp` and the horse power `hp`. The relationship between these four variables with the outcome `mpg` is detailed in **Figure 3**, highlighting the different transmission types.

Modelling

Fitting a linear model for `mpg` with the single predictor `am` (i.e.: `mpg ~ am`) allows to compare the mean `mpg` for automatic and manual transmission:

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 7.2449392713 1.7644216316 4.1061269831 0.0002850207
```

and to say with appropriate confidence that manual transmission cars are expected to show a mean mileage 7.24 miles per US gallon higher than automatic transmission cars.

Linear models where `am` appears toghether with one of `wt`, `disp` and `hp`, **centered on their respective means**, were analyzed to verify the effects of other variables and the explicative power of `am` on `mpg`.

The ANOVA for the nested linear models `mpg ~ am` and `mpg ~ am + wt` would suggest that the inclusion of `wt` is highly desirable, but the `am` coefficient shows a too high p-value, so no effect on expected `mpg` for mean `wt` can be attributed to `am` with appropriate confidence; moreover the ANOVA for the nested models `mpg ~ wt` and `mpg ~ wt + am` would on the contrary suggest that `wt` alone is sufficient: this is in agreement with the distribution of the different type of transmission with respect to weight (see **Figure 3**). The inclusion of the interaction term between `wt` and `am` whould leads to a more explicative model `mpg ~ wt*am` whose coefficients are

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 19.235844  0.7356848 26.146856 3.243563e-21
## wt          -3.785908  0.7856478 -4.818836 4.551182e-05
## amManual    -2.167728  1.4188862 -1.527767 1.377893e-01
## wt:amManual -5.298360  1.4446993 -3.667449 1.017148e-03
```

that explains a lot of variance (adjusted $R^2 = 0.8151$), and states, again, that no absolute difference is to be expected in `mpg` between automatic and manual transmission at the mean weight (intercept), but that the expected decrease in `mpg` for a 1000 lb increment in weight is 5.3 miles/gallon larger in magnitude for manual than for automatic transmission. Diagnostics for the model are in **Figure 4**.

Repeating the same analysis with `disp` instead of `wt` leads to very similar conclusions: no effect on expected `mpg` for mean `disp` can be attributed to `am` with appropriate confidence and `disp` alone seems sufficient in explaining `mpg` (see again **Figure 3**). The inclusion of the interaction term leads to a more explicative model `mpg ~ disp*am` whose coefficients are

```
##                   Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)    18.79292502 0.763132062 24.6260457 1.623095e-20
## disp           -0.02758360 0.006218951 -4.4354101 1.295371e-04
## amManual        0.45175784 1.391508909  0.3246532 7.478567e-01
## disp:amManual -0.03145482 0.011457373 -2.7453781 1.043728e-02
```

with adjusted $R^2 = 0.7674$, that states, again, that no absolute difference is to be expected in `mpg` between automatic and manual transmission at the mean displacement (intercept), but that the expected decrease in `mpg` for a 1 cu. in. increment in displacement is 0.03 miles/gallon larger in magnitude for manual than for automatic transmission. Diagnostics for the model are in **Figure 5**.

The ANOVA for the models `mpg ~ am` and `mpg ~ am+hp`, compared to the ANOVA for the models `mpg ~ hp` and `mpg ~ hp+am` shows that the model with two predictor is always preferable to the ones with a single predictor. The fitting of the model `mpg ~ am+hp` leads to the coefficients

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 17.9468091 0.675884466 26.553072 6.711058e-22
## amManual     5.2770853 1.079540576  4.888270 3.460318e-05
## hp          -0.0588878 0.007856745 -7.495191 2.920375e-08
```

with adjusted $R^2 = 0.767$, that allows to draw the conclusion that manual transmission gives rise, for fixed `hp`, to an expected `mpg` 5.28 miles/gallon larger than automatic transmission. Diagnostics for the model are in **Figure 6**, and show no pattern in residuals and the best Q-Q plot: this is the soundest model.

Results

Manual transmission is better on average for MPG: the expected MPG increment is 7.24 miles/gallon, with 95% confidence interval (3.64, 10.85). Manual transmission is better on average for fixed gross horse power: the expected MPG increment is 5.28 miles/gallon, with 95% confidence interval (3.07, 7.48).

Appendix

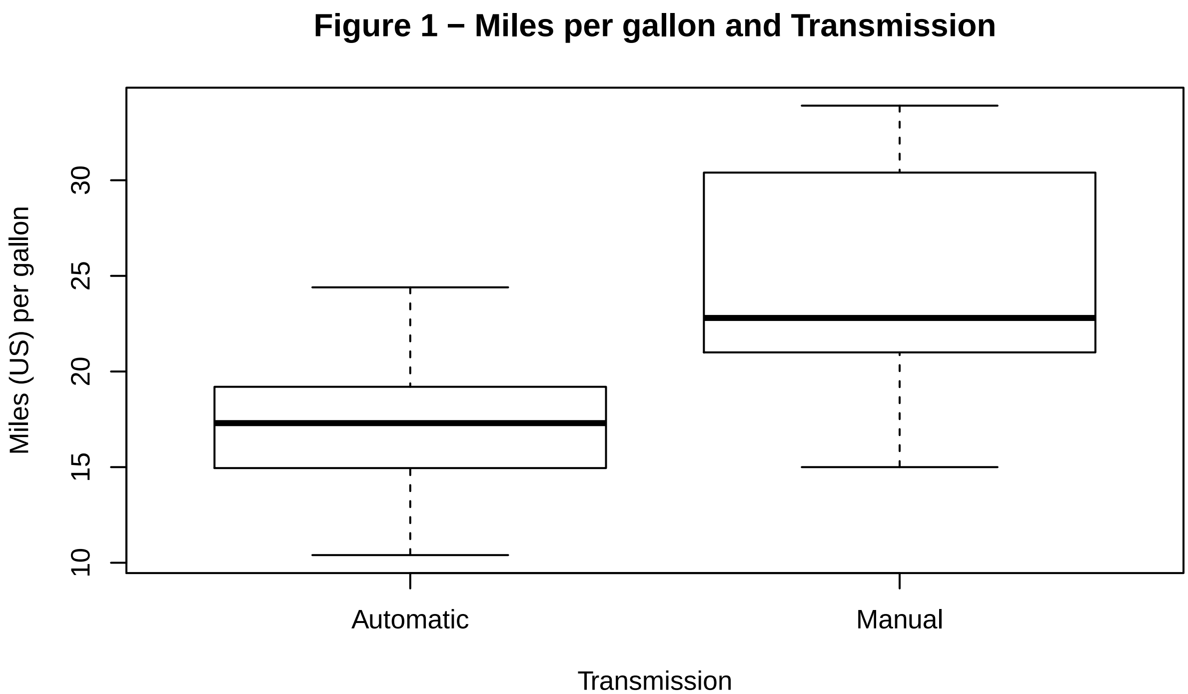

Figure 1 − Miles per gallon and Transmission
Miles (US) per gallon
10
15
20
25
30
Automatic
Manual
Transmission

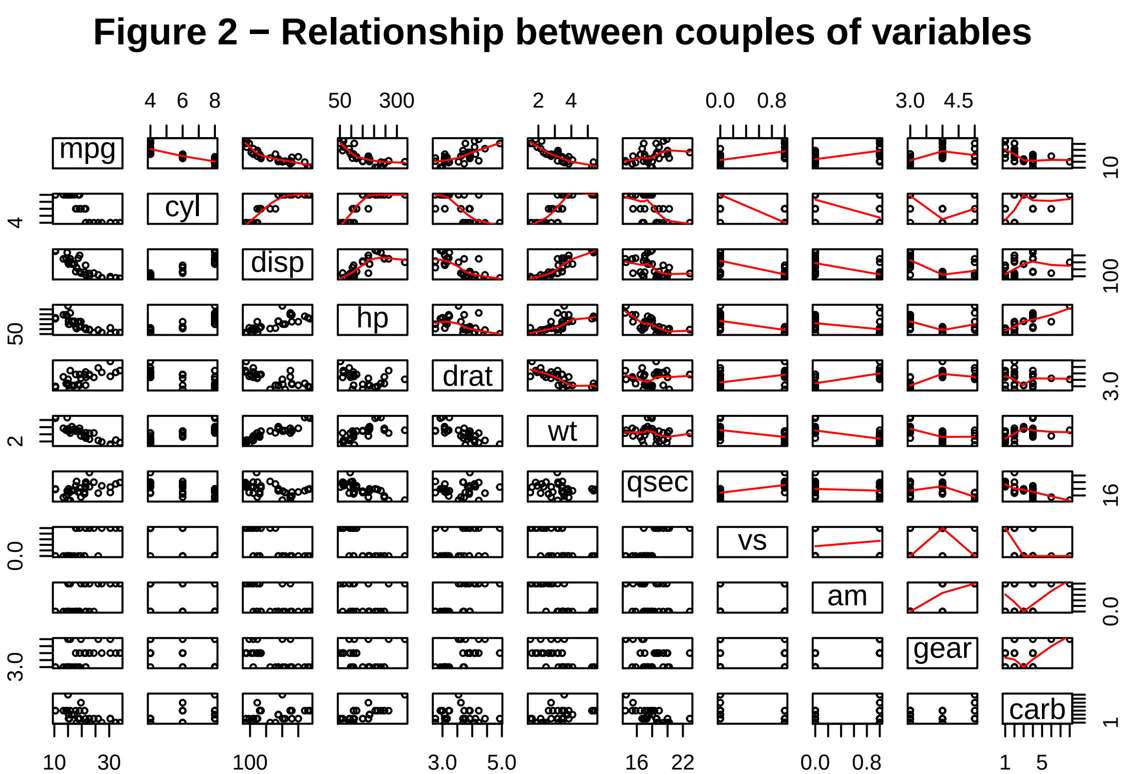

Figure 2 − Relationship between couples of variables
mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb
4 6 8
50 300
2 4
0.0 0.8
3.0 4.5
10 30
100
3.0 5.0
16 22
0.0 0.8
1 5

Figure 3 − Main variables, type of transmission and outcome

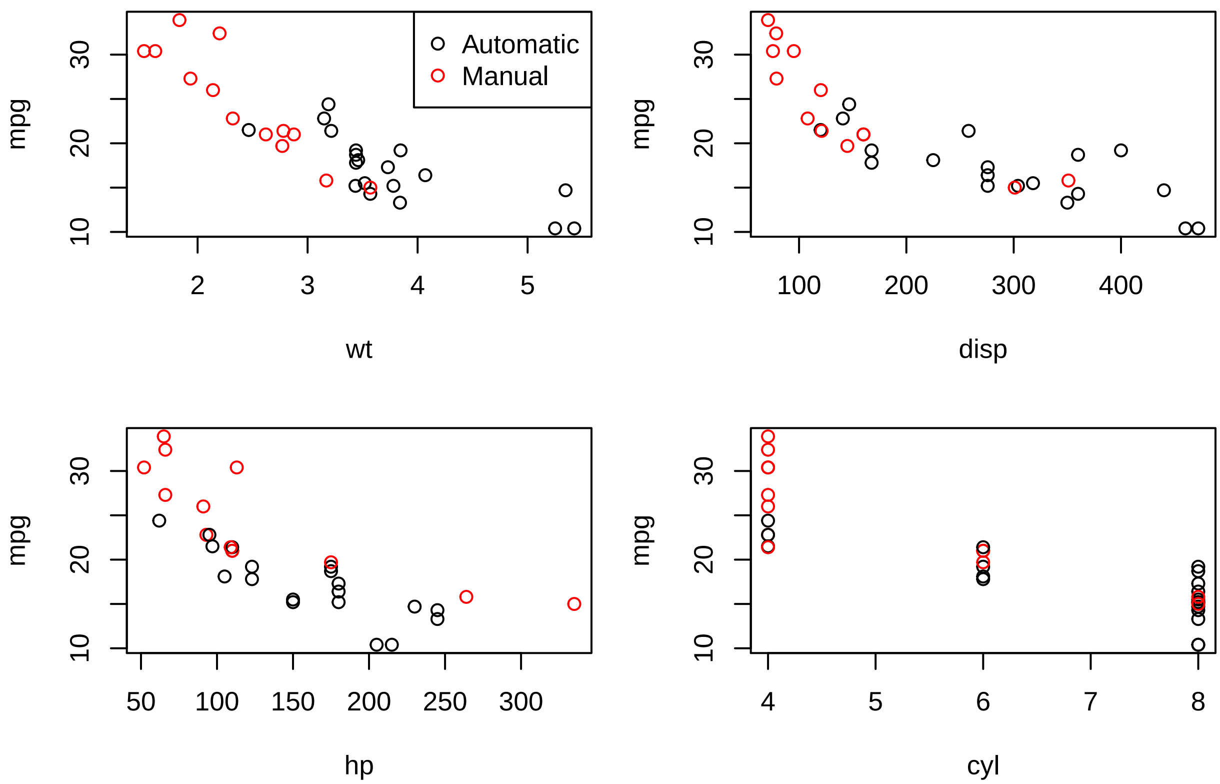

Figure 4 − Diagnostics for model mpg ~ wt*am

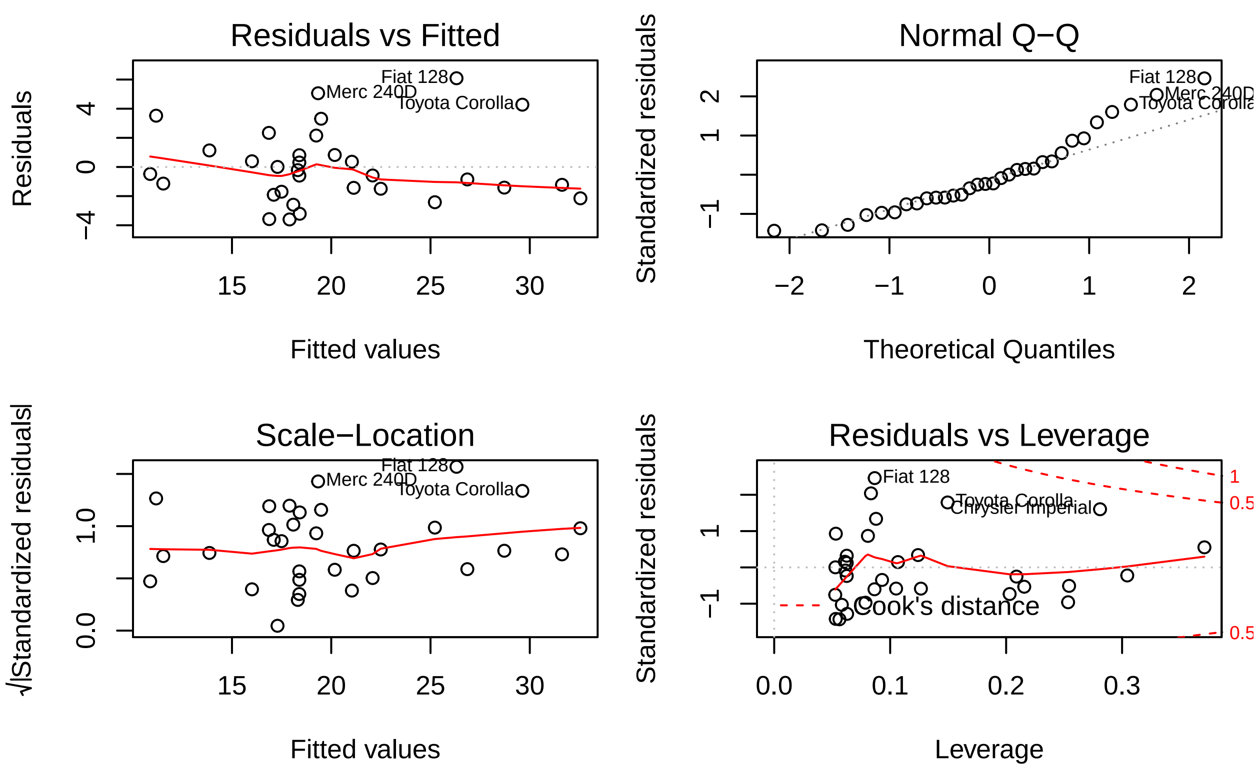

Figure 5 − Diagnostics for model mpg ~ disp*am

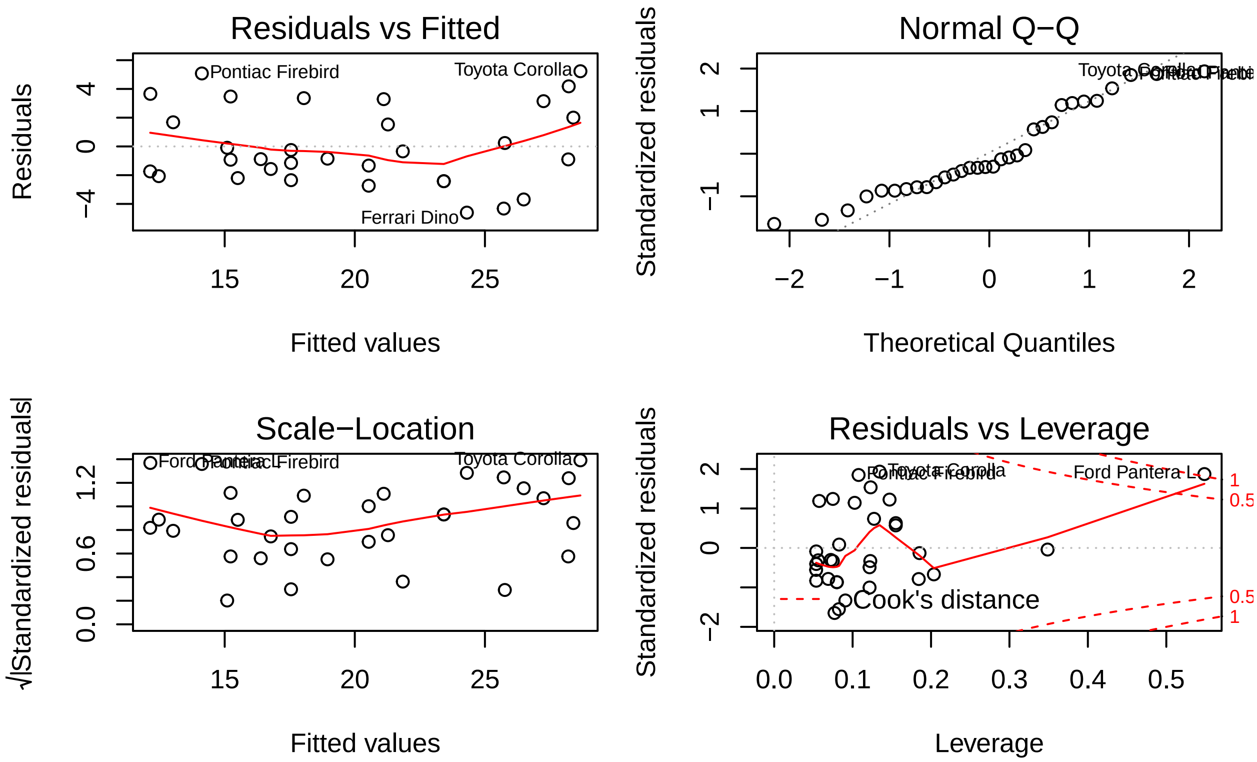

Figure 6 − Diagnostics for model mpg ~ hp+am