Regression Models - Course Project

Giovanni Picardi

23 dicembre 2015

Executive summary

Dataset mtcars has been explored trying to find one or more predictors for the outcome `mpg` (miles per US gallon) and to investigate how the variable `am`, an indicator variable of manual transmission, is related to `mpg`. An exploratory analysis of the dataset showed that manual transmission is associated to a higher mean `mpg` with respect to automatic transmission, as confirmed via T-test, but also that other variables are in a much stronger relationship with the outcome. A few linear models have been analysed and compared in terms of R^2, confidence interval of the `am` coefficient and residuals. Finally an approximate quantitative impact of transmission type on `mpg` has been estimated, based on the `am` coefficient of the selected model.

Exploratory data analysis

The datasets has no missing values, all of the 11 variables are numeric and they are related to a wide range of different motorcars (32 models). A comparison of boxplots of the variable `mpg` for automatic and manual transmission (Figure 1 in Appendix) shows that the mean mileage per gallon is higher for manual transmission. Although `mpg` distributions are far from being normal (Figue 2 in Appendix), a T-test can help in testing this hypothesis or, more precisely, in rejecting the hypothesis that `mpg` means for automatic and manual transmission are equal, giving a cautionary confidence interval:

```r
t.test(mtcars$mpg[mtcars$am == 0], mtcars$mpg[mtcars$am == 1])$conf
```

```
## [1] -11.280194  -3.209684
## attr(,"conf.level")
## [1] 0.95
```

The negative extremes of the 95% confidence interval confirm the hypothesis that the `mpg` mean for manual transmission is higher.

The relationship between pairs of variables, outcome inlcuded, can be visually examined with `pairs(mtcars)` (Figure 3 in Appendix): the more "linearly" related to `mpg` are the weight `wt`, the displacement `disp` and the horse power `hp`; the most useful discrete variable seems to be the number of cylinders `cyl`: different numbers corresponds to almost disjoint sets of values of `mpg`.

Modelling

The percentage of variance explained (R^2) by simple linear models with a single predictor chosen among `wt`, `disp`, `hp`, `cyl` and `am` is the following:

```
##   mpg ~ wt mpg ~ disp   mpg ~ hp  mpg ~ cyl   mpg ~ am
##  0.7445939  0.7089548  0.5891853  0.7170527  0.3384589
```

The model that uses am alone can obviously only predict the mean values of mpg for the two types of transmission with a slope equal to the difference of the means (Figure 4 in Appendix), hence its poor performance. The results show that the best single predictor is the weight wt, but in order to quantify the effects of transmission on mileage per gallon a model with am as predictor is needed. The models with two predictors, one of them being am, and the other being chosen among the previously listed variables, show almost always higher values of R^2:

```
##                      mpg ~ wt+am mpg ~ disp+am mpg ~ hp+am mpg ~ cyl+am
## Adjusted R-squared     0.7357889     0.7149405   0.7670025    0.7423938
```

but the standard error of the am coefficient is sufficiently low with respect to its estimate only in the models with hp and cyl, of which the first allows a narrower estimate of the am coefficient:

```
##            mpg ~ wt+am mpg ~ disp+am mpg ~ hp+am mpg ~ cyl+am
## Estimate   -0.02361522      1.833458    5.277085     2.567035
## Std. Error  1.54564533      1.436100    1.079541     1.291428
```

The final choice is between two models based on hp and am: one with and the other without the hp^2 predictor, both showing residuals with no apparent pattern (Figure 5 in Appendix). The best model in terms of both R^2 and residual standard error is the one with the quadratic term in hp that shows an R^2 quite identical to the linear model using all the 11 variables as predictors (0.8066423):

```
## $coefficients
##                  Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) 33.7758515492 3.1474764229 10.731090 1.982239e-11
## hp          -0.1484779371 0.0363793627 -4.081378 3.377049e-04
## I(hp^2)      0.0002520292 0.0001003027  2.512685 1.801788e-02
## am           3.7512142899 1.1634918539  3.224100 3.203424e-03
## 
## $adj.r.squared
## [1] 0.8030831
```

and whose 95% confidence interval on the am coefficient estimate is:

```
## [1] 1.367909 6.134519
```

Results

The mean mileage per gallon is higher for manual transmission than for automatic transmission with 95% confidence, so manual transmission is on average better than automatic transmission in terms of miles per gallon.

The main variable influencing the miles per gallon figure, besides transmission type, is the horse power.

For fixed horse power, manual transmission brings an increment between 1.3679093 and 6.1345193 miles per gallon with respect to automatic transmission, with 95% confidence.

Appendix

Figure 1 − Miles per gallon and Transmission

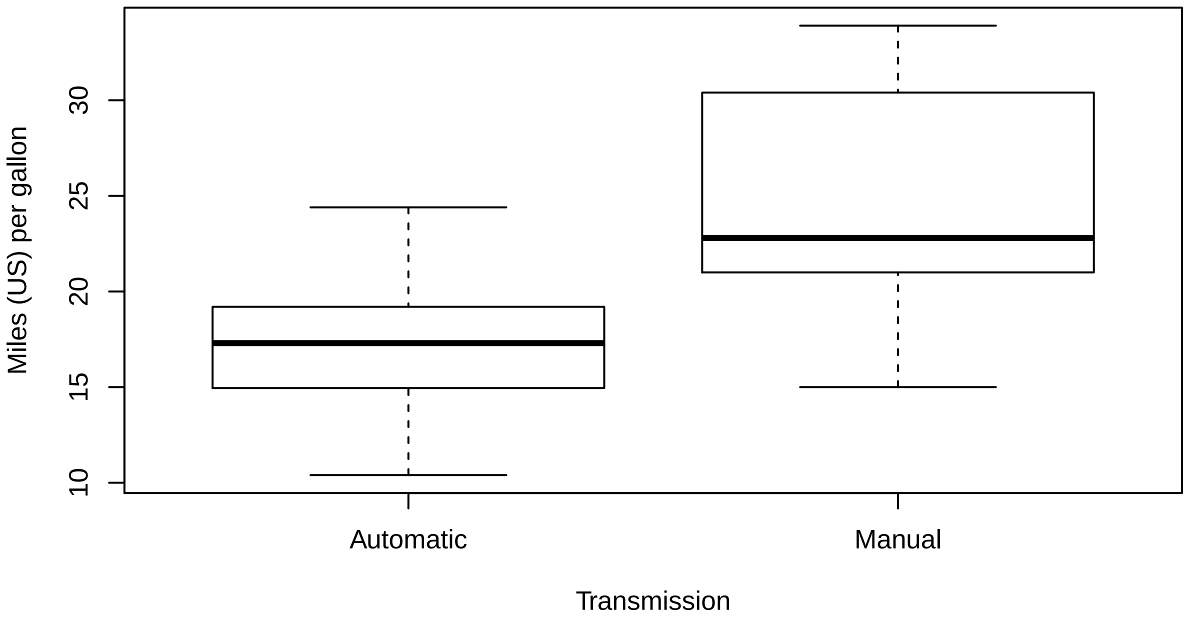


Figure 2 − Miles (US) per gallon densities

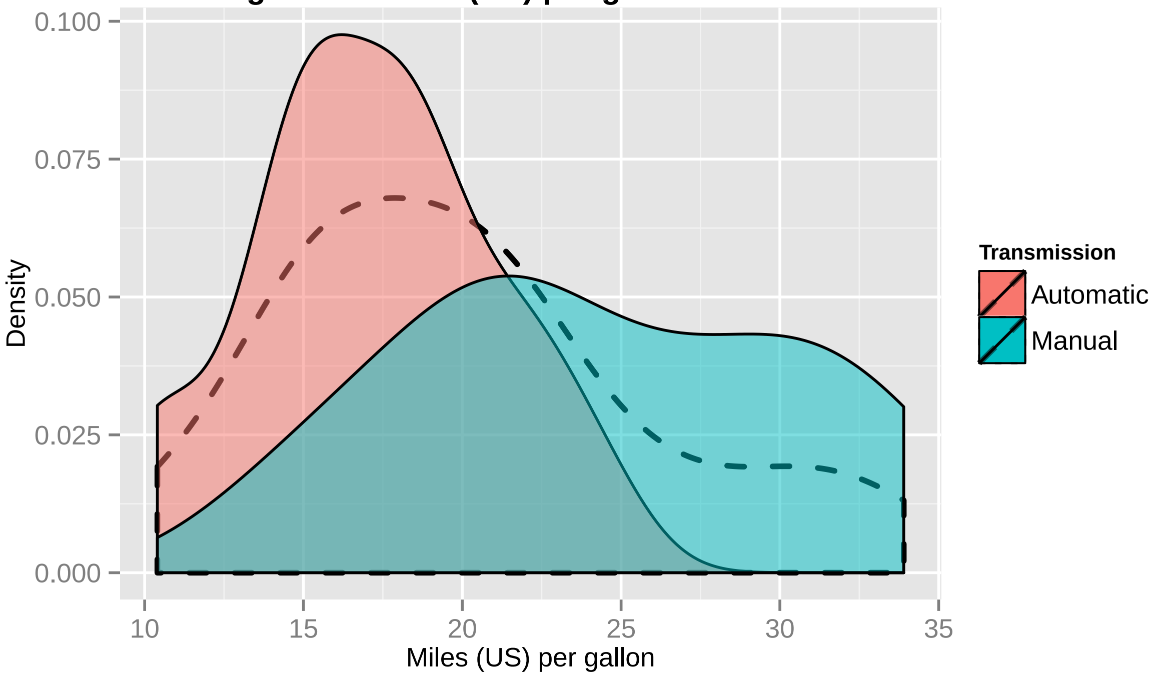


Figure 3 − Relationship between couples of variables

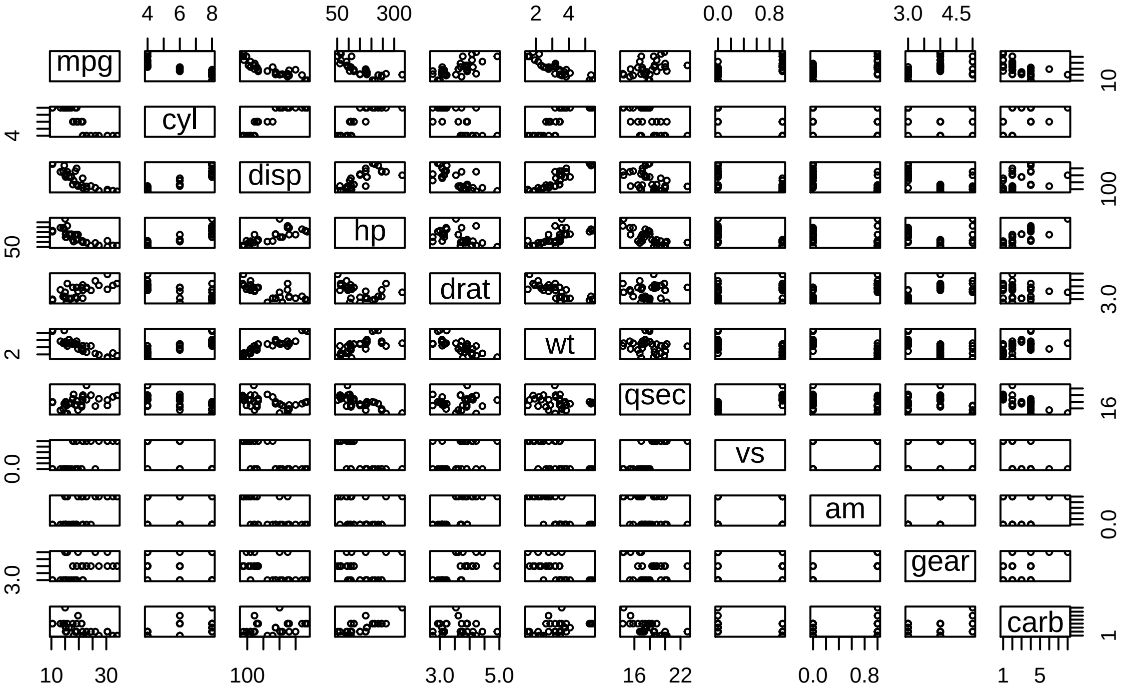


Figure 4 − The simple linear model mpg ~ am

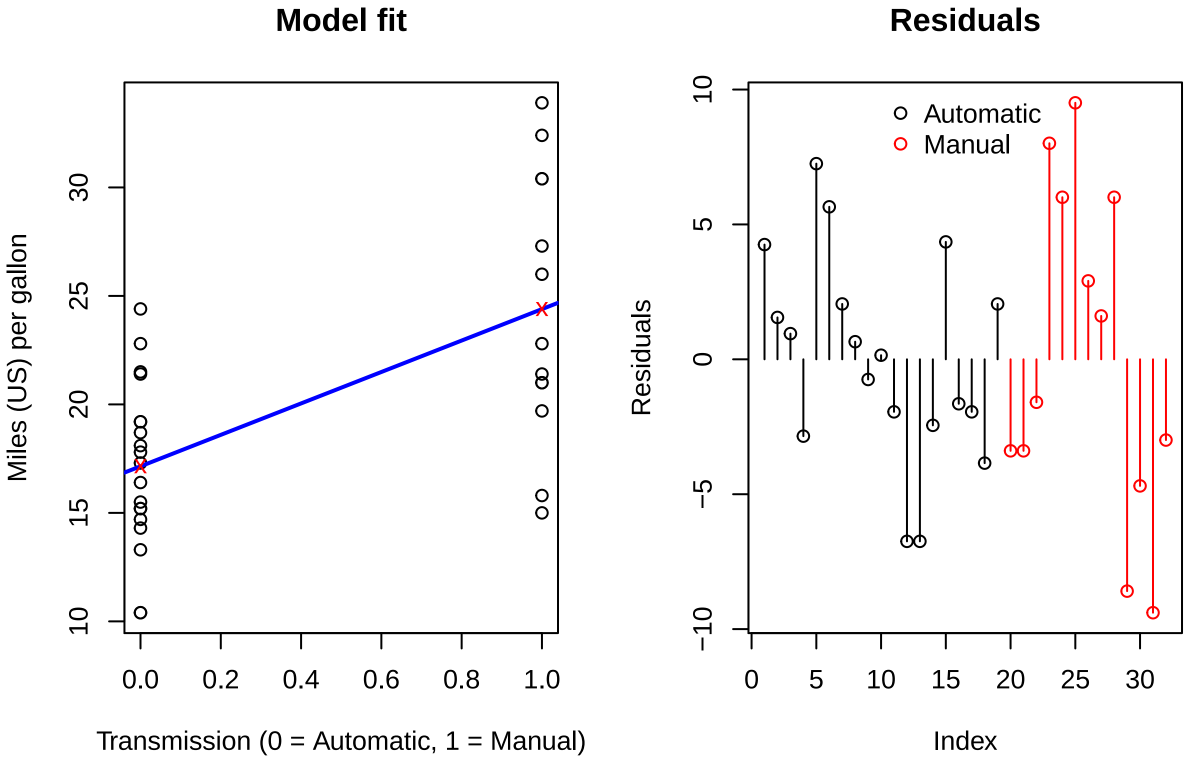
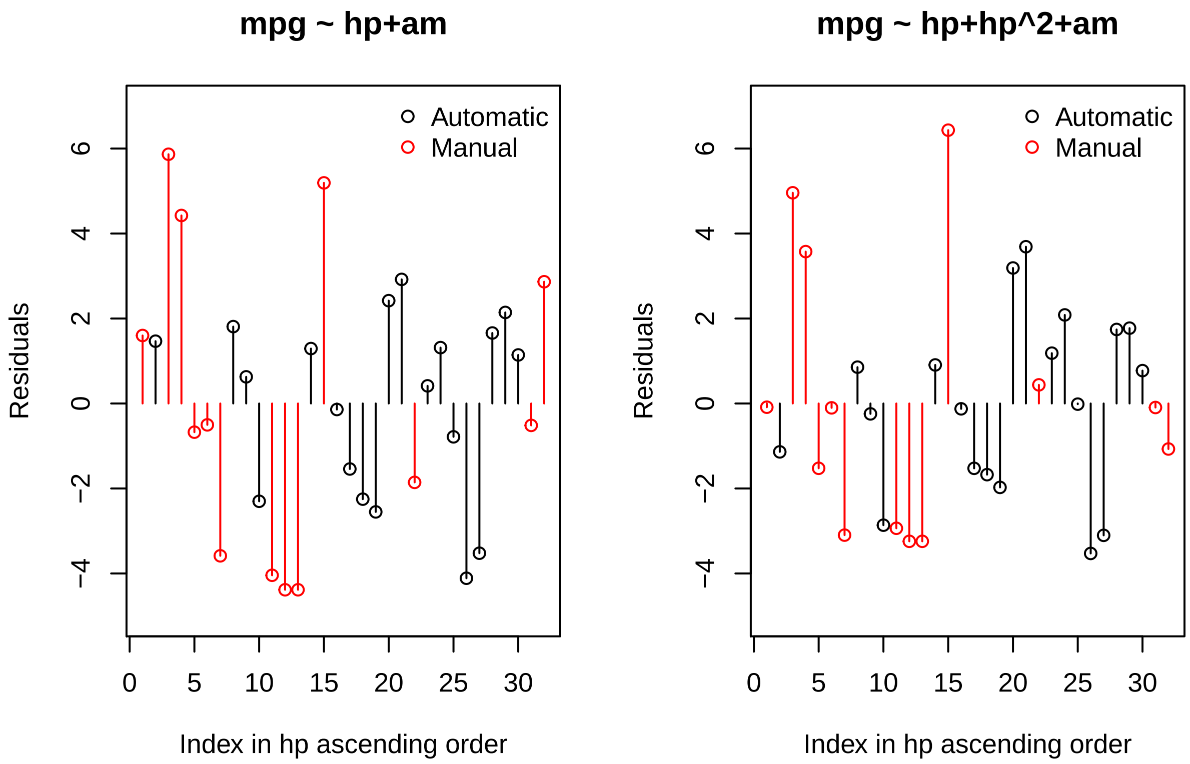


Figure 4 − Residual comparison of linear models



mpg ~ hp+am

mpg ~ hp+hp^2+am