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, but also that other variables are in a much stronger relationship with the outcome. A T-test has been used to check the hypothesis of a higher mean mileage per gallon for manual transmission and a few linear models have been analysed and compared in terms of R^2 and residuals. Finally an approximate quantitative impact of transmission type on `mpg` has been estimated, based on linear model coefficients associated to the variable `am`.

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:

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

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg[mtcars$am == 0] and mtcars$mpg[mtcars$am == 1]
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

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 wight `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 captured by simple linear models with a single predictor chosen among `wt`, `disp`, `hp`, `cyl` and `am` is the following:

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

The poor performance of the model that uses `am` alone is due to the fact that it can obviously only predict the mean values of `mpg` for the two types of transmission with a slope, the coefficient associated to `am` itself, equal to the difference of the means (Figure 4 in Appendix). The results show that the best single predictor is the weight `wt`.

A linear model using all the 11 variables as predictors reaches a R^2 of 0.8066423.

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:

```
##     wt.am   disp.am     hp.am    cyl.am
## 0.7357889 0.7149405 0.7670025 0.7423938
```

but the most interesting result in terms of R^2 is obtained using all the 5 variables as predictors:

```
summary(lm(mpg ~ wt+hp+disp+cyl+am, data=mtcars))
```

```
##
## Call:
## lm(formula = mpg ~ wt + hp + disp + cyl + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5952 -1.5864 -0.7157  1.2821  5.5725
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.20280    3.66910  10.412 9.08e-11 ***
## wt          -3.30262    1.13364  -2.913  0.00726 **
## hp          -0.02796    0.01392  -2.008  0.05510 .
## disp         0.01226    0.01171   1.047  0.30472
## cyl         -1.10638    0.67636  -1.636  0.11393
## am           1.55649    1.44054   1.080  0.28984
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.505 on 26 degrees of freedom
## Multiple R-squared:  0.8551, Adjusted R-squared:  0.8273
## F-statistic:  30.7 on 5 and 26 DF,  p-value: 4.029e-10
```

Results

Appendix

Figure 1 − Miles per gallon and Transmission

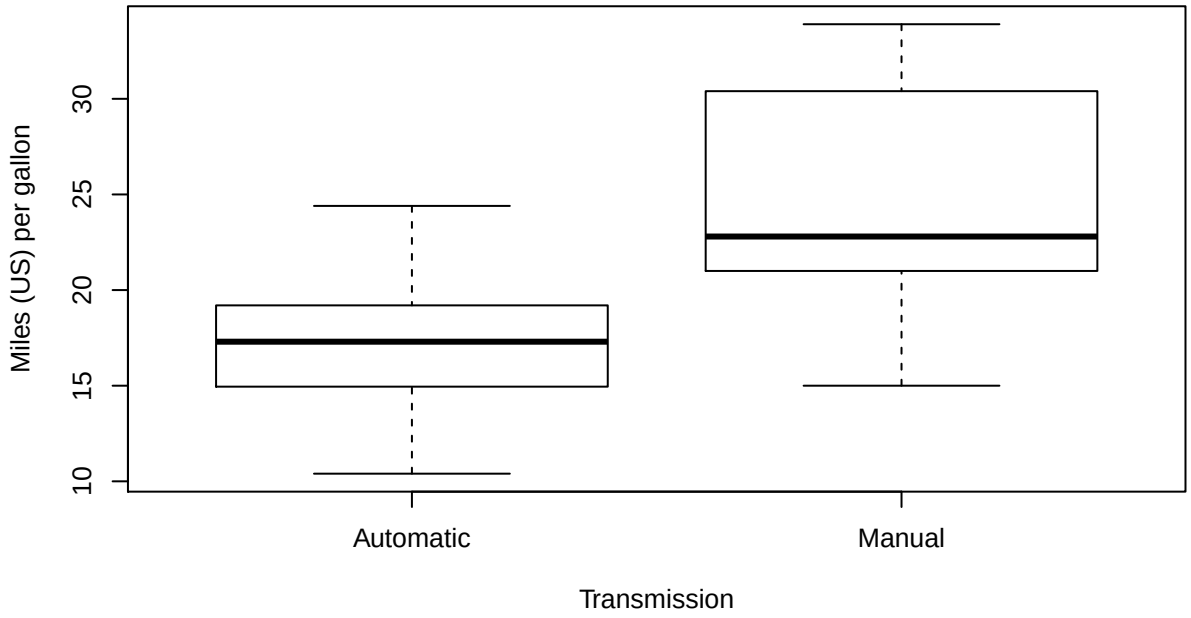

Figure 2 − Miles (US) per gallon densities

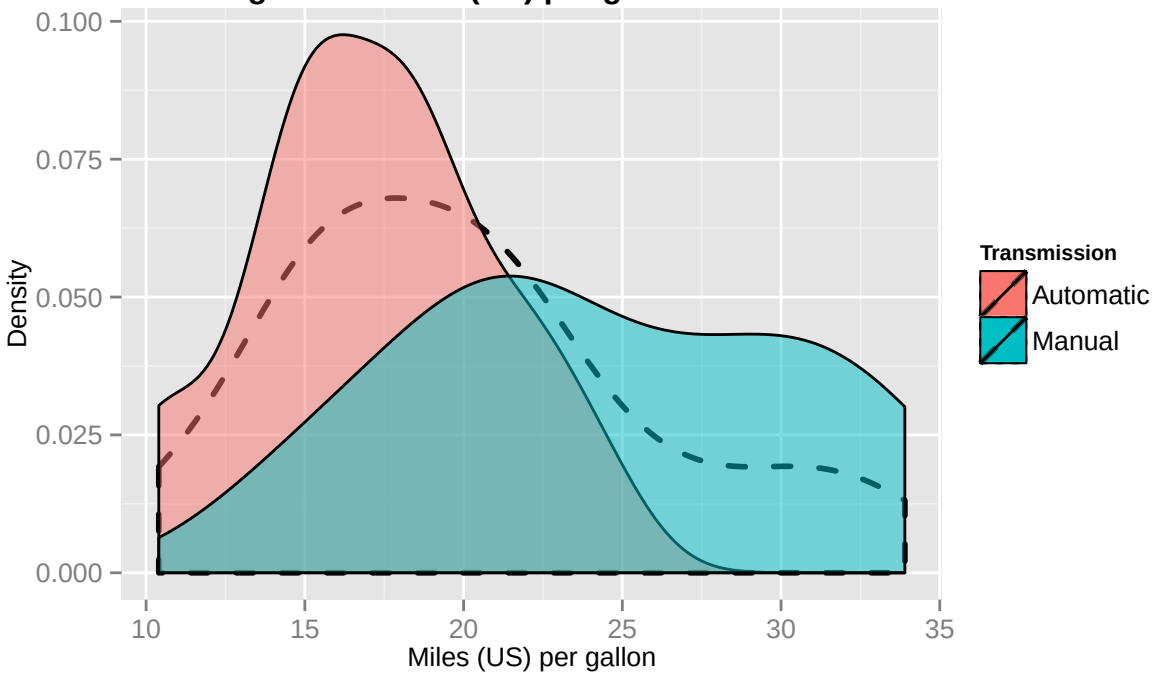

Figure 3 − Relationship between couples of variables

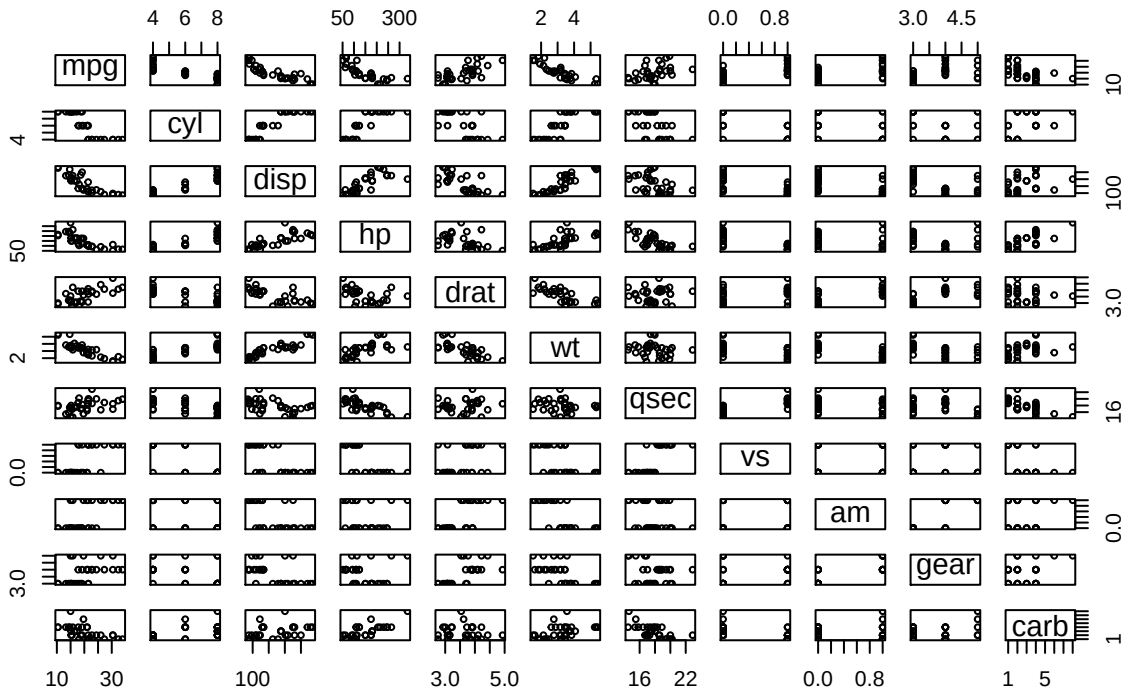

Figure 4 − The simple linear model mpg ~ am