Transmission type versus MPG in 1973-74 classic cars

by R. Rodríguez

NOTE: I understand this course is not about R, but about regression models, therefore, the R code and outputs are mostly hidden or embedded in text. If you are interested in the code, or just to see how I did X, the source code for this document is available at https://github.com/ricrogz/regmods

Summary

In this issue of MotorTrend we are going to have a statistical look at our favourite models from 1973-74,and we will try to determine which is more fuel efficient: automatic transmission or manual. First we will try a "naive" approximation on which kind of transmission is more fuel efficient in average, and after that, we will take a deeper look into the features of our cars to find out the reason of this difference.

Analysis

To start our study, we will do a very naive plot to compare the mpg averages of manual and automatic tramsission cars. This plot is shown in Figure 1 in the Appendix. The graph clearly shows that the average mpg values are quite different, with the average mpg for automatic cars being clearly higher than the one for manual transmission cars (24.39 versus 17.15).

A paired t-test shows that this difference is signifficant (p.value = 0.001). When trying to fit to a simple linear regression model, we get a regression coefficient of R.squared = 0.36, which means the model we get is only able to explain 36 % of the variance in the data.

But, this is only a very simple comparison; it might be that it is not directly the manual transmission who is responsible for the higher fuel consumption, but other characteristics which are more frequent in manual transmission cars than in automatic ones. Now we will try to evaluate such factors.

For first, we will do another plot. We will build a grid of pair plots to explore the relation between mpg and all the other variables in our data. This paired plot is shown in Figure 2 in the appendix. The plots above the diagonal show the data plus a LOWESS smoother, and the values in the boxes under the diagonals are the correlation values between the variables indicated by the row and column.

Since we are most interested in the mpg, we observe the first row of plots and the first column of correlation factors in the grid, and see that variables "cyl" (number of cylinders), "disp" (displacement in inches), "hp" (gross horsepower), "wt" (weight in lb/1000) have the strongest correlation with mpg, but the correlations between each pair of these variables are also strong, which makes it difficult to predict which of these variables would be the best candidates as regressors in a multiple regression model.

To further investigate this relations, we are going to build several multiple variable regression models, and evaluate them stepwise to find the one which the most significante refressors. For this, we will use R's step function, using the AIC algorithm, using both backwards elimination and forward selection. The result of this process is the following model:

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.939 -1.256 -0.401  1.125  5.051
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.7083     2.6049   12.94  7.7e-13 ***
## cyl6         -3.0313     1.4073   -2.15   0.0407 *
## cyl8         -2.1637     2.2843   -0.95   0.3523
## hp           -0.0321     0.0137   -2.35   0.0269 *
## wt           -2.4968     0.8856   -2.82   0.0091 **
## am1           1.8092     1.3963    1.30   0.2065
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.866,  Adjusted R-squared:  0.84
## F-statistic: 33.6 on 5 and 26 DF,  p-value: 1.51e-10
```

As it was expected, the model uses three of the regressors we detected in the pair plots, but ignores the "disp" one. At the same time, the transmission type is also included in the model, meaning that it indeed has some influence on mpg.

As we can see, the selected model has an adjusted R squared value of 0.84 (it explains 84 % of the variance in the data. Further than that, we can identify the impacts of the changes in each of the regressors: 4-cylinder motors have the highest mpg values, and going up to 6 or 8 cylinders produces a decrease mpg (-3.03 mpg and -5.19 mpg, respectively); at the same time, an increase of 1 hp in motor power translates into a loss of -0.03 mpg, while weight also has an important impact on mpg, falling almost 2.5 mpg per 1000 lb of car weight. The only feature that has a positive impact on mpg is the transmission type, which improves in 1.81 mpg when changing from manual transmission to automatic, but is, at the same time, the regressor with the smallest influence (p.value = 0.20646 for 'am == 1', in respect to the only alternative of 'am == 0').

Finally, to make sure we did not overlook any effect, we will analyze the residuals plots for our chosen model. The plots are shown in Figures 3 to 6 in the Appendix. In the plots, we can see that there are no recognizable patterns in the Residuals versus fitted plot (Fig. 3), so we did not miss any important relation in our model. Also, most of the residuals lay on or very close to the diagonal in the Normal Q - Q plot (Fig. 4), meaning the residuals are distributed almost normally. In the Scale - Location plot (Fig. 5), the points randomly scattered and uniformly distributed, so that we can conclude that the variance is constant.

To conclude this analisis, we observe an outlier in the Residuals versus Leverage plot (Fig. 3). This outlier corresponds to the Maserati Bora, which is the car with the highest hp number in our data (~25% more hp than the second most powerful car). Despite the high hp value, and having 8 cylinders, the car is suprisingly fuel efficient, with a mpg value of 15, versus the expected value of 13.6835. Since there is no evidence of wrong data or error in this case, the `names(outlier)` should not be excluded from the fitted data, and therefore, our analysis is correct.

Conclusions

Despite in a simple comparison automatic transmission cars have a signigicantly better mpg average than manual ones, a detailed study reveals the tramsission type has only a small effect on the mpg: we found out that a different transmission type only accounts for a change of approximately 1.81 mpg (keeping every other parameter constant), while the difference in the average between manual and automatic cars is around -7.2449 mpg.

More than that, our study revealed that the factos that are most influential in mpg are the number of cylinders, weight, horsepower (all of them with a negative influence: the higher this value, the lower the mpg), as well as the transmission type, as we already mentioned.

Appendix

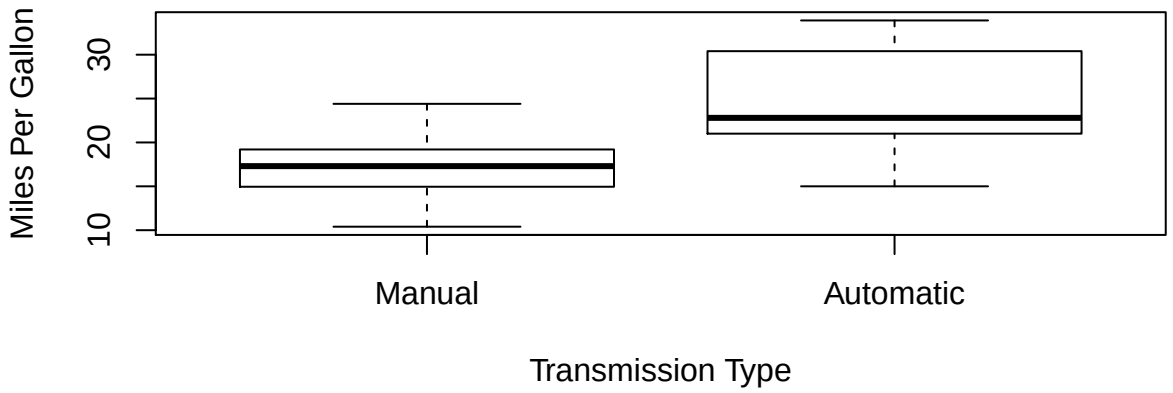

Figure 1: Figure 1 - Naive Fuel efficiency by transmission type.

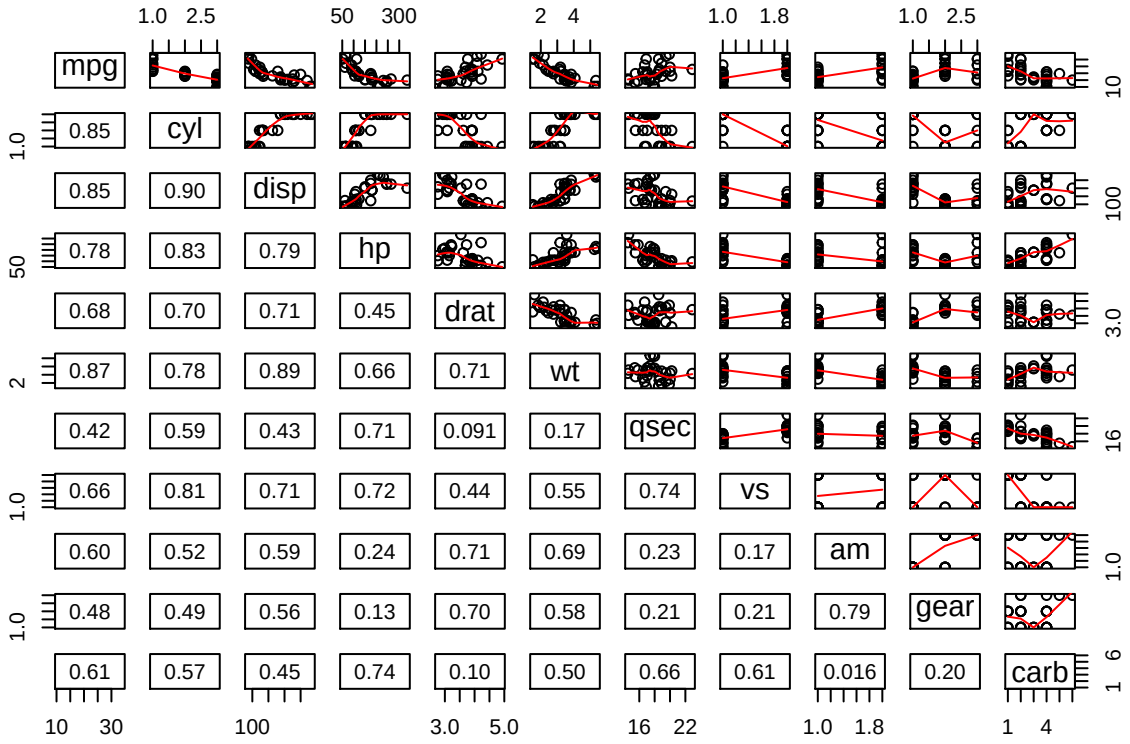

Figure 2: Figure 2 - Paired variable plots.

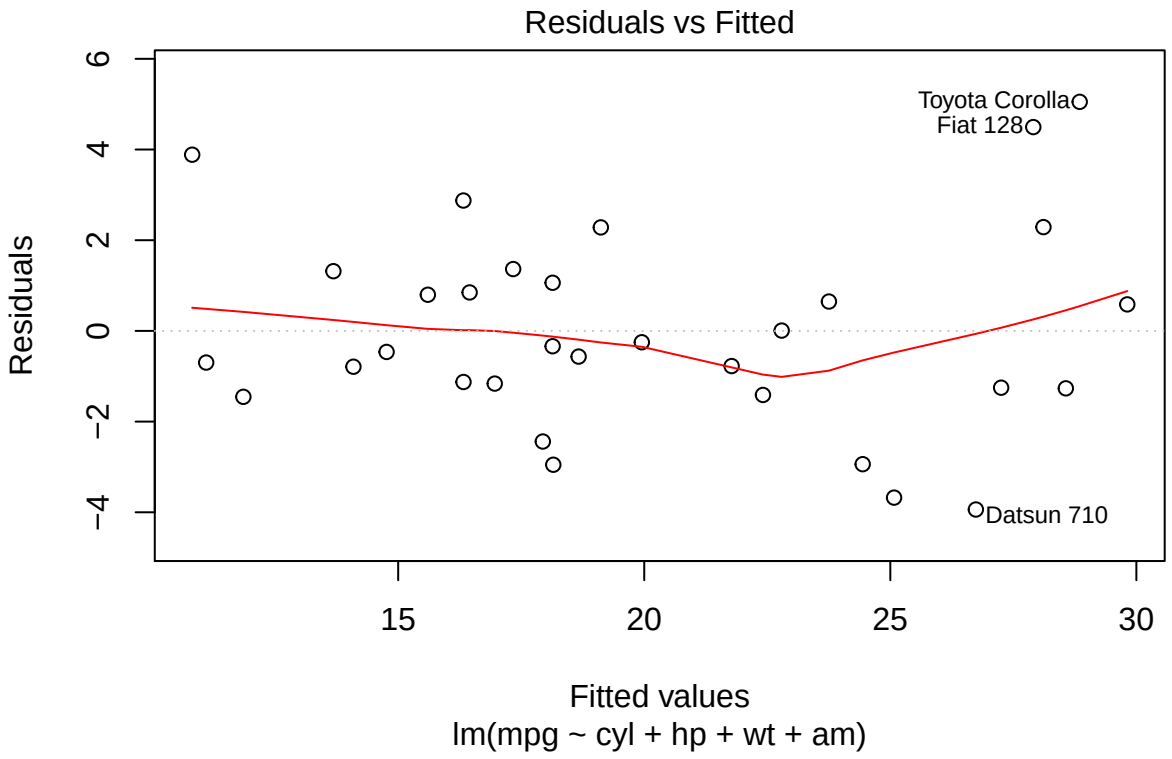

Figure 3: Figure 3 - Residual plots: residuals versus fitted

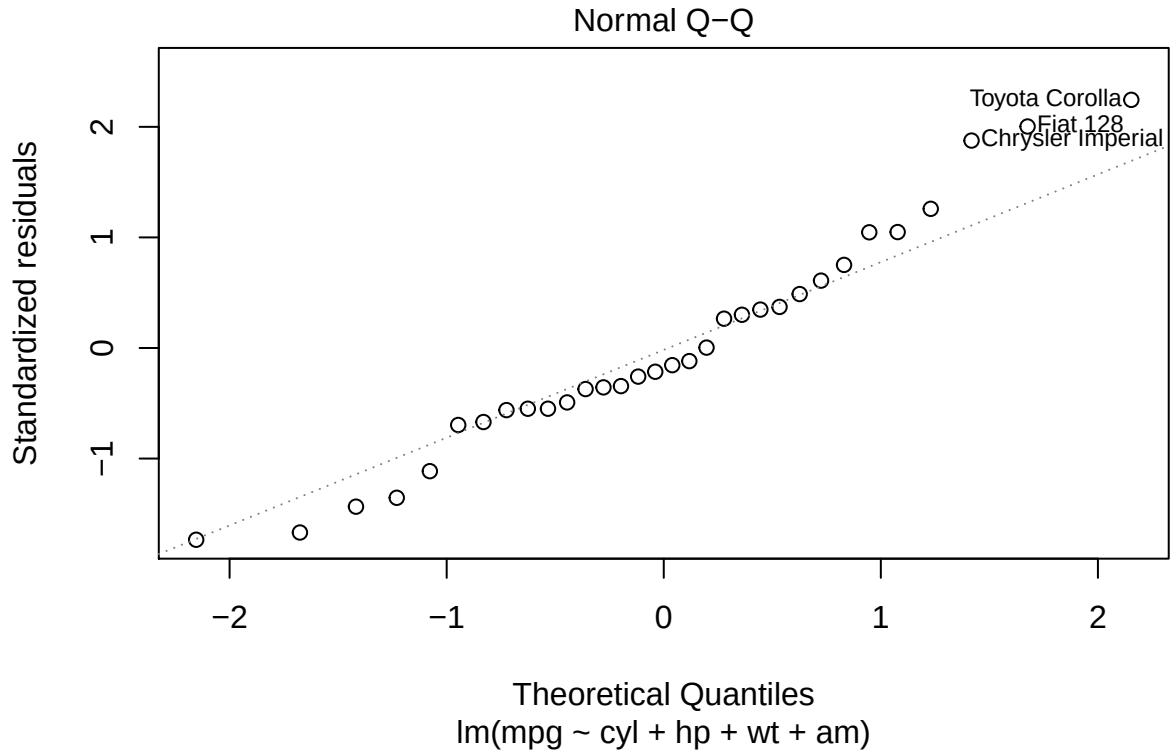

Figure 4: Figure 4 - Residual plots: Normal Q - Q.

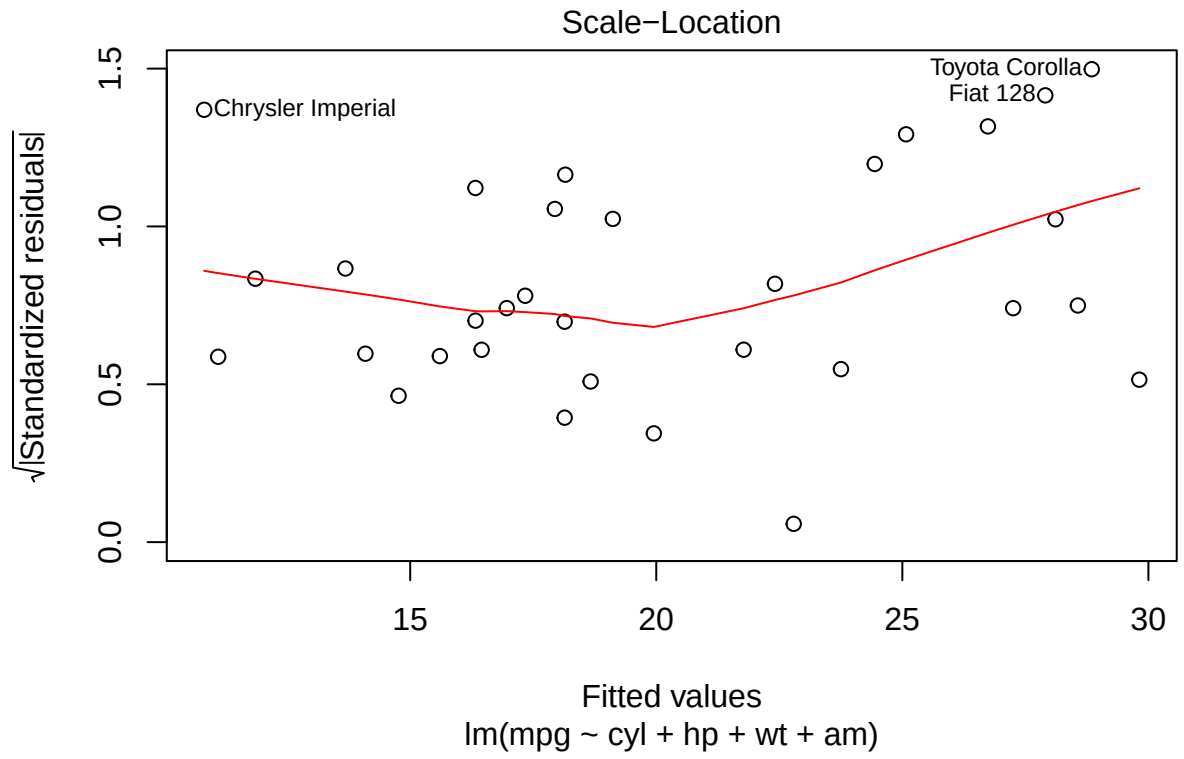

Figure 5: Figure 5 - Residual plots: Scale - Location.

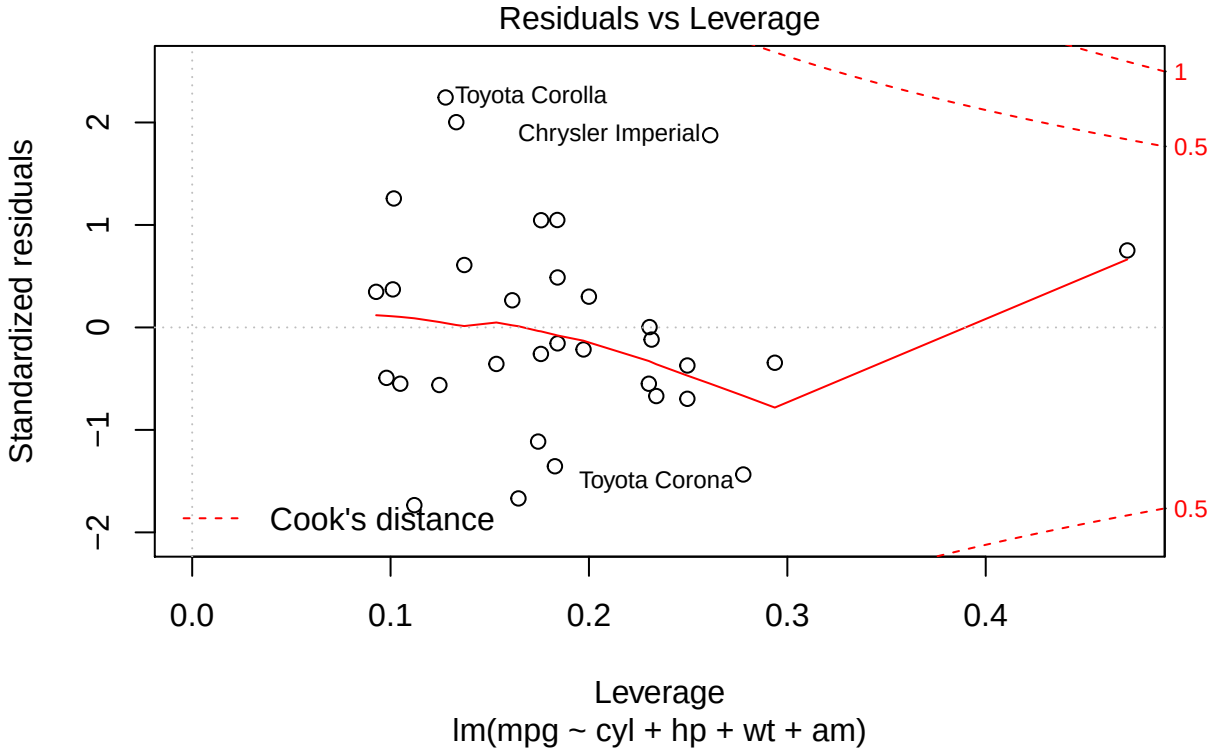

Figure 6: Figure 6 - Residual plots: Residuals versus Leverage.