Predicting the Carbon Footprint of Automobiles Through Regression Models

J Faleiro

April 14, 2015

Executive Summary

This report is an analysis of data extracted from the 1974 issue of *Motor Trend*, an US automobile magazine. The data comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

This analysis concentrates in answering two questions:

- *Is an automatic or manual transmission better for MPG*
- *Quantify the MPG difference between automatic and manual transmissions*

Our analysis will be conducted in R and will consist of tables, figures and several summaries. We will take special care to document each of the steps, making sure they follow a fully scripted flow and therefore allow for the entire analysis to be entirely reproducible.

Data Analysis

Exploratory Data Analysis

For our models we need a factor of transmission type, according to the value 0 or 1 in `am` will be respectively associated to `automatic` and `manual`.

With that we can get the first glimpse on how mpg might be correlated to transmission type (**Figure 1**).

We can see that there is an indication of which is better, on this case manual transmissions. But we know enough about cars and there might be uncorrelated regressors that are affecting the outcome. Our best guess at this point are `wt`, `hp` and `disp` **(Figure 2)**.

As expected, `mpg` is negativelly correlated to all regressors, and on the other hand all regressors are positivelly correlated amongst themselves. We need to select a model that relies on the relevant regressors to infer the lowest error and highest influence of that on the final outcome variable `mpg`.

Model Selection Strategy

Our model selection strategy will consider two different types of fitting: linear and logistic poisson regression.

For each type, we will quantify the uncertainty of that fitting through *nested models* that will consider, incrementally, a set of regressors:

- `disp`: Displacement (cu.in.)
- `hp`: Gross horsepower
- `wt`: Weight (lb/1000)

For each nested model, we will use `anova()`, analysis of variances, to quantify the uncertainty of adding that additional regressor to infer the model with lowest error and highest influence of that final outcome variable `mpg`.

The `F statistic` in the analysis of variances tests the predictive capability of the model as a whole - the larger the F value the better our model is at predicting the dependent variable `mpg`. On the other hand, lower F values indicate the model is not as good at predicting the dependent variable.

Three asterisks (***) at the lower right of the printed table indicate that the null hypothesis is rejected at the 0.001 level, so at least one of the two additional regressors is significant. Rejection is based on a right-tailed F test, `Pr(>F)`, applied to an F value.

We will select the model with highest break on `F` value and lowest `Pr(>F)` on the analysis of variances.

Next, we will check patterns of residuals, according to a few guidelines for *well-behaved* Residuals vs. Fitted plot and Residuals vs Leverage plot, in what they suggest about the appropriateness of the simple linear regression model:

- Residuals vs Fitted plot
  - The residuals "bounce randomly" around the 0 line. This suggests that the assumption that the relationship is linear is reasonable.
  - The residuals roughly form a "horizontal band" around the 0 line. This suggests that the variances of the error terms are equal.
  - No one residual "stands out" from the basic random pattern of residuals. This suggests that there are no outliers.
- Residuals vs Leverage plot
  - There should be no points with (Cook's distance greated than 0.5.

Linear Fitting

Following our strategy, the first type of fitting is a linear regression, defined through `lm()`. The strategy also describes a selection of an optimal model through nesting:

```
## Analysis of Variance Table
##
## Model 1: mpg ~ transmission
## Model 2: mpg ~ transmission + hp
## Model 3: mpg ~ transmission + hp + wt
## Model 4: mpg ~ transmission + hp + wt + disp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 245.44  1    475.46 71.3552 4.646e-09 ***
## 3     28 180.29  1     65.15  9.7773    0.0042 **
## 4     27 179.91  1      0.38  0.0576    0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see a clear break on the F statistic value from model 2 to model 3, when F got to **71.3552** before dropping to **9.7773**. The `Pr(>F)` at that level is also very low at $\frac{4.646}{10^9}$ indicating that with formula `mpg ~ transmission + hp` our null-hyphotesis H_0 can be safely rejected.

Let's investigate the residuals of the selected model `mpg ~ transmission + hp`, removing the intersect with `-1` (**Figure 3**).

According to the guidelines outlined in our strategy to verify residuals this is an appropriate linear regression model. We can move ahead and use coefficients to quantify the mpg difference between automatic and manual

```
##                         Estimate  Std. Error   t value     Pr(>|t|)
## transmissionautomatic 26.5849137 1.425094292 18.654845 1.073954e-17
## transmissionmanual    31.8619991 1.282279151 24.847943 4.252195e-21
## hp                    -0.0588878 0.007856745 -7.495191 2.920375e-08
```

```r
automatic <- coef(fit)[1]
manual <- coef(fit)[2]
(manual - automatic) / automatic
```

```
## transmissionmanual
##          0.1984992
```

According to our selected linear model, a manual transmission gives a **19.85%** better mpg ratio, plus or minus a standard error of **1.42**.

Logistic Fitting

The second type of fitting is a generalized linear regression, GLM, defined through `glm()`. Given the presence of a factor regressor we will be using a Poison family of regressions.

The strategy also describes a selection of an optimal model through nesting:

```
## Analysis of Deviance Table
##
## Model 1: as.integer(mpg) ~ transmission
## Model 2: as.integer(mpg) ~ transmission + hp
## Model 3: as.integer(mpg) ~ transmission + hp + wt
## Model 4: as.integer(mpg) ~ transmission + hp + wt + disp
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        30     34.858
## 2        29     10.686  1  24.1728 8.807e-07 ***
## 3        28      6.554  1   4.1314   0.04209 *
## 4        27      6.550  1   0.0039   0.95012
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see a clear break on the `Deviance` value from model 2 to model 3, when deviance got to 24.1728 before dropping to 4.1314 The Pr(>Chi) at that level is also very low at $\frac{8.807}{10^7}$ indicating that with formula mpg ~ transmission + hp our null-hyphotesis H_0 can be safely rejected.

Let's investigate the residuals of the selected model `mpg ~ transmission + hp`, removing the intersect with `-1` **(Figure 4)**:

According to the guidelines outlined in our strategy to verify residuals this is an appropriate logistic poisson regression model. We can move ahead and use coefficients to quantify the mpg difference between automatic and manual

```
##                           Estimate  Std. Error   z value      Pr(>|z|)
## transmissionautomatic  3.310020391 0.116488160 28.415080 1.316830e-177
## transmissionmanual     3.550104532 0.093215300 38.084998  0.000000e+00
## hp                    -0.003158038 0.000673509 -4.688932  2.746352e-06
```

```
## transmissionautomatic    transmissionmanual                    hp
##            27.3856839            34.8169568             0.9968469
```

```
automatic <- exp(coef(fitl)[1])
manual <- exp(coef(fitl)[2])
(manual - automatic) / automatic
```

```
## transmissionmanual
##          0.2713561
```

According to our selected logistic model, a manual transmission gives a **27.13%** better mpg ratio, plus or minus a standard error of **1.09**.

Conclusions

In conclusion, the answers to our initial questions:

- *Is an automatic or manual transmission better for MPG*
  - In general **manual transmission vehicles performed better than automatic transmission vehicles**, yielding a higher ratio of miles per gallons
- *Quantify the MPG difference between automatic and manual transmissions*
  - Our regression models provided slightly different answers:
    * Linear regression model: **19.85% +/- 1.42**
    * Logistic poisson model: **27.13% +/- 1.09**

Appendix of Figures

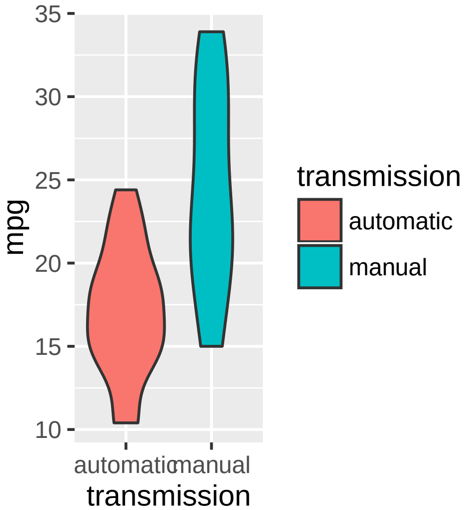

Figure 1: Violin Plot of MPG Consumption per Transmission Type

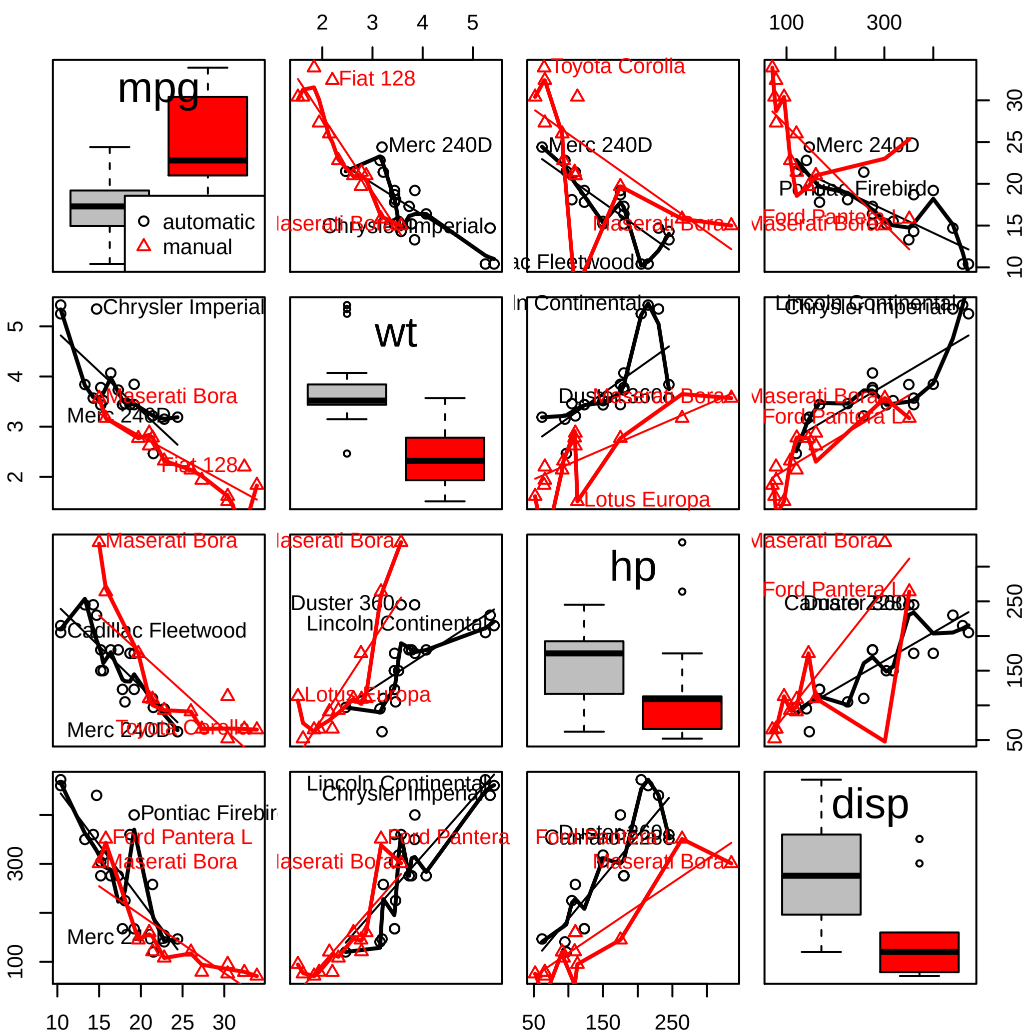

Figure 2: Scatter Plot Matrix of Regressors Candidates

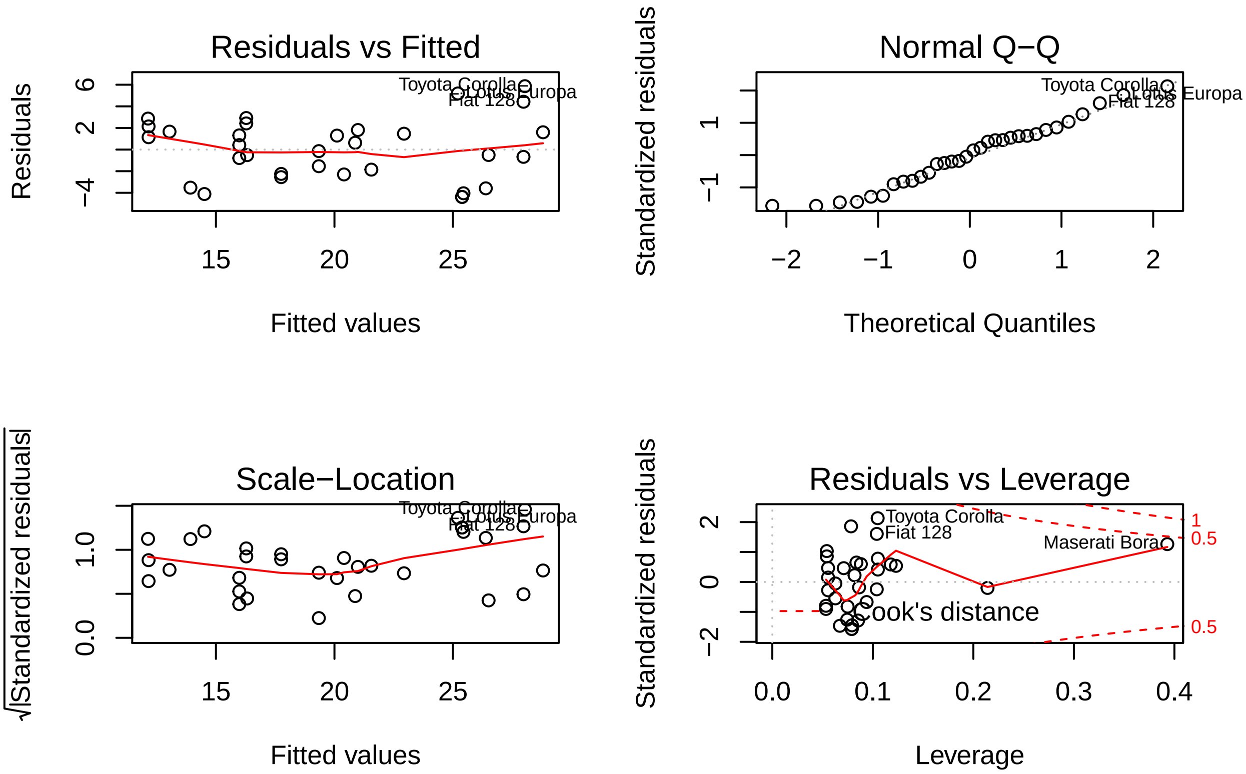

Figure 3: Diagnostic Plots of Linear Model

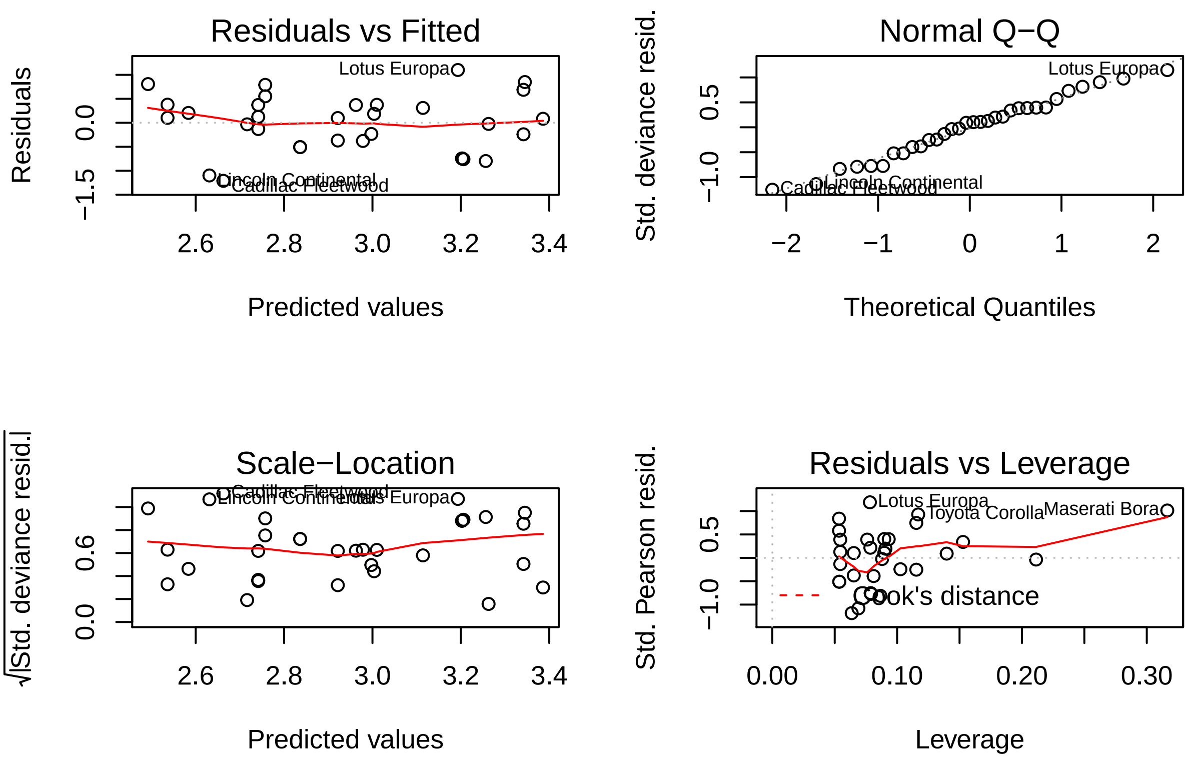

Figure 4: Diagnostic Plots of Poisson Logistic General Model