Regression Models Course Project

A study of the relationship between transmission type and fuel economy in vehicles in the mtcars dataset

I. Beauser

2015-10-25

Executive Summary

Motor Trend, a magazine about the automobile industry, is interested in how fuel economy, measured in miles per gallon (MPG), differs between those automobiles with automatic transmissions, and those with manual transmissions. The data for their analysis is sourced from the `mtcars` dataset (Henderson and Velleman 1981). The magazine's analysis targeted two main requirements:

1. Answer the question, "Is an automatic or manual transmission better for MPG?", and
2. Quantify the MPG difference between automatic and manual transmissions.

In summary, the analysis shows that an automobile with manual transmission has better fuel economy than that of a car with automatic transmission. After other factors affecting fuel economy are held constant, on average manual transmission vehicles are expected to exceed the fuel economy of those with automatic transmissions by ~42%.

The Data

The description (R Core Team and contributors worldwide) from the `mtcars` dataset says the data was "extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models)." The documentation provides the variable names and descriptions.

Examing the correlation between `mpg` (miles-per-gallon) and the remaining variables suggests which variables might be included in the model. The results below have eliminated those variables with a weaker (< 0.5) correlation to `mpg` than the others.

```
##    Var1 Var2      value
## 5   mpg   wt -0.8676594
## 1   mpg  cyl -0.8521620
## 2   mpg disp -0.8475514
## 3   mpg   hp -0.7761684
## 4   mpg drat  0.6811719
## 7   mpg   vs  0.6640389
## 8   mpg   am  0.5998324
## 10  mpg carb -0.5509251
```

The variable `am` (Automatic vs Manual transmission), though having the second to lowest correlation to `mpg` of the remaining variables, is kept as the predictor of principle interest. Those variables that have a strong (> 0.7) correlation to `mpg` are kept too as probable confounders. The variable `am` is converted to a factor and labelled according to the dataset description, for further analysis.

```
## 'data.frame':    32 obs. of  6 variables:
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ am  : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
```

Model Selection

A linear model *lm(mpg ~ factor(am), mtcars)* has a statistically significant p-value (0.00029 < 0.05) suggesting that changes in `mpg` are strongly associated with a change in transmission type.

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## amManual     7.244939   1.764422  4.106127 2.850207e-04
```

However, the influences of the other strongly correlated variables `wt` (weight in thousands of pounds), `cyl` (number of cylinders), `hp` (horsepower), and `disp` (cubic inches of displacement) need to be examined.

Intuition suggests and testing shows a very strong correlation (0.9020329) and statistically significant (p-value = 1.802838e-12) relationship between displacement and number of cylinders, so that `disp` can be a proxy for `cyl` in further testing. Similarly, `disp` will serve as a proxy for `hp` due to the strong correlation (0.7909486) and statistically significant (p-value = 7.142679e-08) relationship between the two. With these substitutions, a nested likelihood ratio test is run using the `anova` function on three models: the first, the simple one with only `am` as a predictor; the other two to test if the additional parameters `wt` and `disp` are needed.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + disp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 50.2610 1.032e-07 ***
## 3     28 246.56  1     31.76  3.6072   0.06788 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The test including the second model is highly significant (p-value = 1.031522e-07), suggesting the additional parameters should be included. Adding displacement is not statistically significant.

Additionally, the influence of each parameter on the coefficient can be tested.

```
##                                  Estimate     Pr(>|t|)
## Model 1: mpg ~ am              7.24493927 0.0002850207
## Model 2: mpg ~ am + wt        -0.02361522 0.9879145855
## Model 3: mpg ~ am + wt + disp  0.17772414 0.9055482918
## Model 4: mpg ~ am + disp       1.83345825 0.2118396121
```

In this test, the p-value of Model 2 is insignificant, as is that of both the combined parameters and the `disp` parameter alone. Therefore, Model 1 is selected and it estimates, by the intercept coefficient, that a manual transmission car gets an expected 7.24 miles per gallon more than that obtained by an automatic, all other parameters held constant.

Residual Analysis

The Normal Q-Q plot does not raise any concerns with the fit of Model 1 (Fig 1), as the residual distrubtion meets the "pencil test" (link to B. Caffo video).

Using the `predict` function, the confidence interval of the prediction can be calculated.

```
##        fit      lwr     upr
## 1 24.39231 14.00311 34.7815
```

The calulation implies a 95% confidence that the mean MPG for a manual transmission car falls within the range 14 to 35 mpg. A similar calulation provided 95% confidence that the mean MPG for an automatic transmission car falls in the range 7 to 27 mpg. Thus, on average, the manual car in the sample set is approximately 42% more fuel efficient than the automatic.

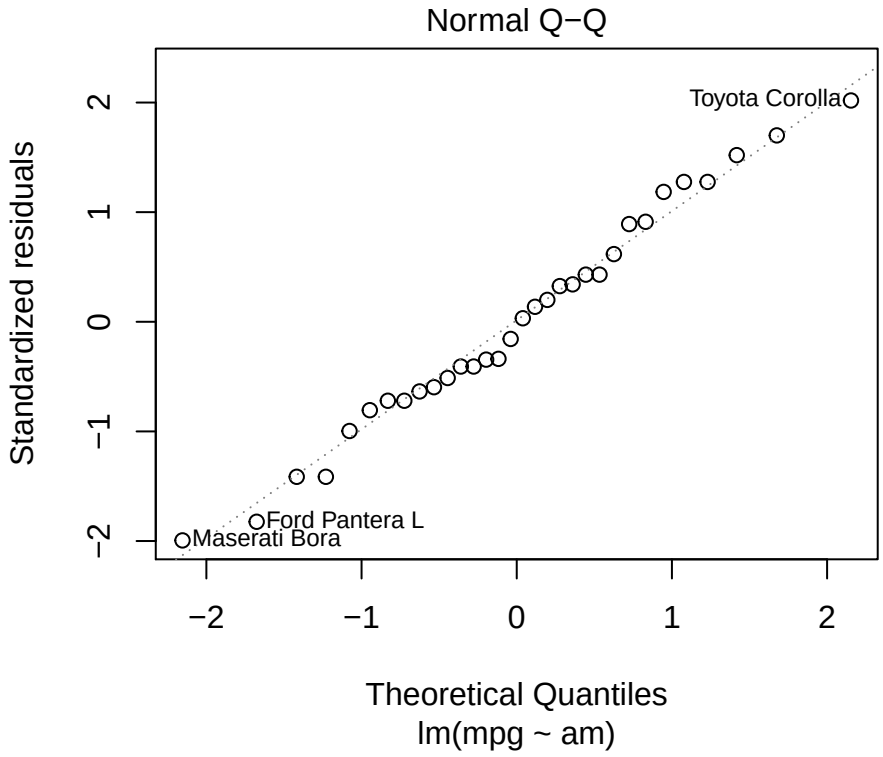

Figure 1: Model 1 Residual Analysis

References

Henderson and Velleman. 1981. "Building Multiple Regression Models Interactively." *Biometrics*.

R Core Team and contributors worldwide. "The R Datasets Package." R Core Team R-core@r-project.org.