Regression Models - Analysis of Transmission against MPG for the mtcars dataset

Bill Gourley

November 21, 2015

Context

Explore the relationship between a set of variables and miles per gallon (MPG) (outcome). The following two questions need to be answered:

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

For this analysis, the primary aim is to describe the regressor relationships from a modelling perspective, and not necessarily from a best prediction estimate perspective.

Executive Summary

The analysis shows that, for the data available in the `mtcars` dataset, the following statements can be made :

- When no other variables are involved, manual transmission is better for MPG. Cars with automatic transmission achieve, on average, 17.15 MPG with a 95% confidence interval of [14.85,19.44] whereas cars with manual transmission achieve, on average, 24.39 MPG with a 95% confidence interval of [18.49,30.29]. However, this model only explains 36% of the total variation. Analysis is shown in the Appendix.
- When all of the mtcars variables are included in the model and regression diagnostics and model selection are carried out, a model which regresses mpg against transmission (am), horsepower (hp) and weight(wt) explaining 89% of the total variation is achieved. Results indicate that the choice of transmission is statistically insignificant, makes little difference to MPG, and could be left out of the model. The model indicates that, when other variables are accounted for, cars with automatic transmission achieve, on average, 35.5 MPG with a 95% confidence interval of [31.2,39.74] whereas cars with manual transmission achieve, on average, 35.98 MPG with a 95% confidence interval of [29.56,42.39].

Details of the analysis follows. The supporting code can be found in the `mtcarsAnalysis.rmd` file.

Analysis

Variables cyl, vs, am, gear and carb have been converted into factors for the analysis.

Exploratory Data Analysis and Model Selection

Nested models were created and two ANOVA analyses were carried out to identify significant variables to include in the final model. The output from the second analysis can be found in the Appendix as Figure 2.

The p-value for disp indicates that this variable is not significant, see Figure 3a in the Appendix, therefore we can improve the model by removing it. Variables cyl and hp are related, there seems to be 3 distinct

groupings as can be seen in Figure 3b in the Appendix. We can therefore simplify the model by replacing cyl + hp with hp. (Note that there is a clear non-linear relationship between mpg and hp, and the model could be improved by using a log transformation of hp. This has not been done in this case in order to simplify the interpretation of the model).

From the model residual and Q-Q plots shown in the Appendix as Figure 4a, 3 observations are highlighted as influential, `Chrysler Imperial`, `Toyota Corolla`, and `Fiat 128`. The model fit can be improved by removing these observations.

The resultant model coefficients and R squared values can be found in the Appendix as Figure 4b. The 95% confidence intervals can be found in the Appendix as Figure 4c.

The coefficients of the model variables can be interpreted as follows:

- As it doesn't make sense to have a car with 0 hp and 0 wt, the Intercept, in this case, merely anchors the regression estimates.
- The coefficient for hp indicates that, for every 100 increase in horsepower, there is a decrease of 2.8 MPG.
- the coefficient for wt indicates that, for every 1000lb increase in weight, there is a decrease of 3.75 MPG

In the final model, all variables are significant except am. The variable am has been left in the model as this was the primary focus of the questions raised.

The summary details of the final model are shown in the Appendix as Figure 5.

Appendix

Comparison of mpg against am in Isolation

A boxplot of the comparison can be found in Figure 1 below. The coefficients, 95% confidence intervals and R Squared values from `lm(formula = mpg ~ am, data = mtcars)` are as follows :

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am1          7.244939   1.764422  4.106127 2.850207e-04
```

```
## [1] 0.3597989
```

```
##                2.5 %   97.5 %
## (Intercept) 14.85062 19.44411
## am1          3.64151 10.84837
```

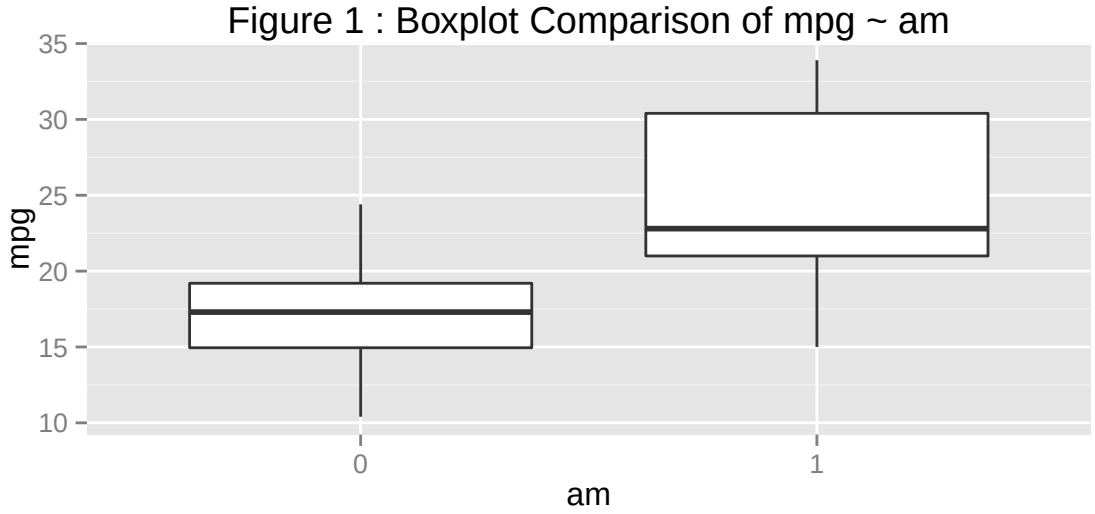

Figure 2 : ANOVA Analysis 2

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + disp
## Model 4: mpg ~ am + cyl + disp + hp
## Model 5: mpg ~ am + cyl + disp + hp + wt
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 264.50  2    456.40 37.9300 2.678e-08 ***
## 3     27 230.46  1     34.04  5.6572  0.025339 *
## 4     26 183.04  1     47.42  7.8820  0.009541 **
## 5     25 150.41  1     32.63  5.4236  0.028246 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3a : Model Coefficients and R Squared Values for mpg ~ am + cyl + disp + hp + wt

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 33.864276061 2.69541569 12.5636562 2.668321e-12
## am1          1.806099494 1.42107933  1.2709350 2.154510e-01
## cyl6        -3.136066556 1.46909031 -2.1346996 4.277253e-02
## cyl8        -2.717781289 2.89814941 -0.9377644 3.573375e-01
## disp         0.004087893 0.01276729  0.3201848 7.514890e-01
## hp          -0.032480178 0.01398322 -2.3227963 2.862128e-02
## wt          -2.738694608 1.17597755 -2.3288664 2.824553e-02
```

```
## [1] 0.8664276
```

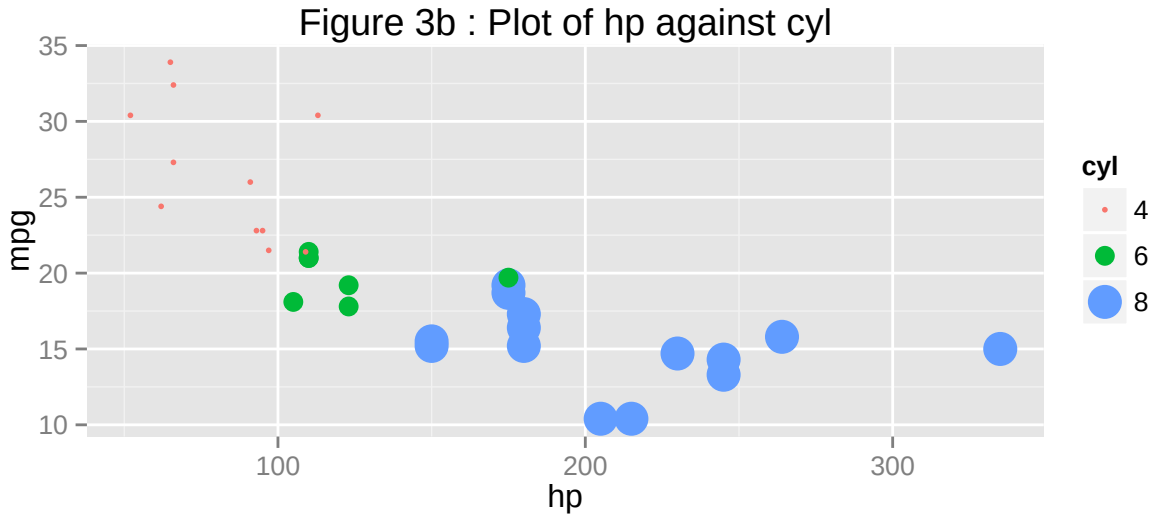

Figure 4a : Model Diagnostic Plots for mpg ~ am + hp + wt

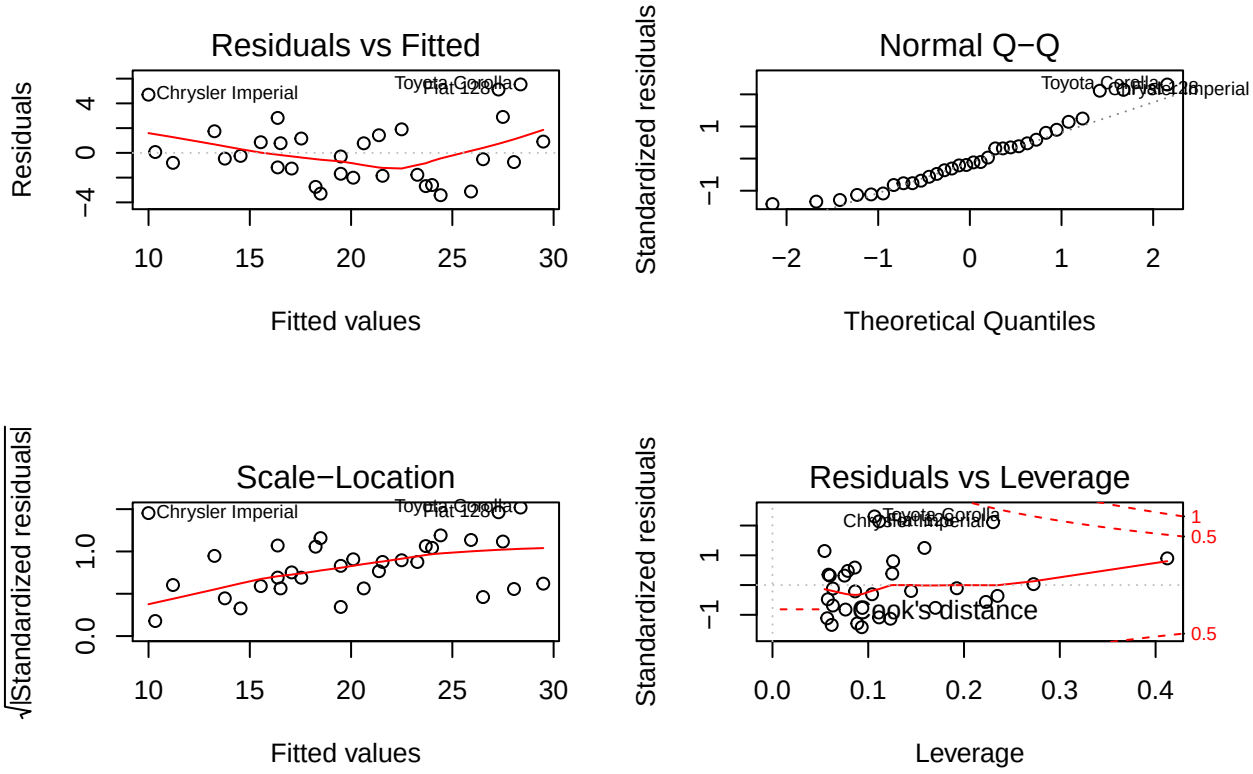

Figure 4b : Model Coefficients and R Squared Values for mpg ~ am + hp + wt

```
##                Estimate  Std. Error    t value     Pr(>|t|)
```

```
## (Intercept) 35.47152353 2.071991008 17.1195355 2.556304e-15
## am1          0.50558410 1.039672494  0.4862917 6.309960e-01
## hp          -0.02802033 0.007147566 -3.9202619 6.078746e-04
## wt          -3.75030468 0.712336605 -5.2647929 1.881295e-05
```

```
## [1] 0.8873634
```

Figure 4c : 95% Confidence Intervals for mpg ~ am + hp + wt

```
##                   2.5 %      97.5 %
## (Intercept) 31.20417817 39.73886889
## am1         -1.63566148  2.64682968
## hp          -0.04274102 -0.01329964
## wt          -5.21738938 -2.28321998
```

Figure 5 : Model Summary for mpg ~ am + hp + wt

```
##
## Call:
## lm(formula = mpg ~ am + hp + wt, data = mtcars.filtered)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.186 -1.113  0.076  1.236  3.263
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.471524   2.071991  17.120 2.56e-15 ***
## am1          0.505584   1.039672   0.486 0.630996
## hp          -0.028020   0.007148  -3.920 0.000608 ***
## wt          -3.750305   0.712337  -5.265 1.88e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.826 on 25 degrees of freedom
## Multiple R-squared:  0.8874, Adjusted R-squared:  0.8738
## F-statistic: 65.65 on 3 and 25 DF,  p-value: 5.441e-12
```