Exploring the Effects of Transmission Type on MPG

Juan Agustin Melendez

Executive Summary

The goal of this analysis is to determine if automatic or manual transmission is better for gas mileage and to quantify the differences using linear regression and the *mtcars* dataset in R. It was found that considering only transmission type, manual transmission vehicles have 7.24 more *mpg* than vehicles with automatic transmission. Further evaluation of the data showed that confounding variables *wt, qsec, and am* had the most influence on the response variable. When considering these confounders, a vehicle with manual transmission had 2.93 more *mpg* than automatic transmission vehicles holding *wt & qsec* constant.

Exploratory Data Analysis and Single Variable Linear Regression

A density plot reveals that the distribution of *mpg* is relativel normal therefore satisfying the assumption of normality in linear regression. A box plot of *mpg* versus *am* shows that the average *mpg* in manual transmission is higher than the average for automatic transmission. Further more, other variables seem to be correlated to *mpg* as can be observed from the pairs plot (see figures 1, 2 and 3 in the appendix).

To start with, a model was fitted using solely *am* as the independant variable. The data was preloaded and factor variables were assigned as such (code hidden for space purposes).

```
base_model <- lm(mpg ~ am, mtcars)
base_model$coef
```

```
FALSE (Intercept)    amManual
FALSE   17.147368    7.244939
```

The mean mpg for automatic transmission in the base model is the intercept **17.15** and the mean for manual transmission is the addition of the coefficients, **24.39**. While the estimated coefficient were significant with p-values close to zero, the adjusted R-squared value is very low at approximately 0.34. This means that the model explains only about 34% of the variability of the data and suggests additional confounders should be considered.

Multivariate Regression

The approach taken was to fit a model with all the variables as confounders and systematically reduce the number of confounders while evaluating the adjusted R-squared values and coefficient significance. Variables were removed by evaluating the correlation with *mpg* and with other variables, excluding first those that did not have much correlation with mpg and eventually those that were correlated with each other (e.g. *hp & cyl*). ANOVA was performed every time variables were removed from the model to evaluate the impact on the model. This process was repeated until all significant confounding variables were tested. The fitted models can be seen in the ANOVA result output (Figure 5) in the appendix.

Two models were given considerable attention: *mpg ~ am + wt + qsec* and *mpg ~ am + wt + hp*. The adjusted R-squared values for these models differed only by about 1% but the coefficient for the manual transmission on the model having *hp* as a confounder had p-values of 0.141 and thus the estimate was not considered to be significant. In the model containing *qsec* as a confounder, all the coefficients had significant levels with p-values below 0.05. Also, when analyzing the residual plots for this models, the residuals vs fitted plot appeared to be more randomly distributed than the plot with *hp* as confounder. From the ANOVA test it can be seen that removing confounders off the model containing all variables as confounders had little significance until the qsec variable was eliminated. The F statistic grew drastically to 13.58 and had a p-value of 0.002 rejecting the null hypothesis that removing the qsec variable from the model had no effect on the model.

```
FALSE              Estimate Std. Error   t value     Pr(>|t|)
FALSE (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
FALSE amManual     2.935837  1.4109045  2.080819 4.671551e-02
FALSE wt          -3.916504  0.7112016 -5.506882 6.952711e-06
FALSE qsec         1.225886  0.2886696  4.246676 2.161737e-04
```

For these reasons *mpg ~ am + wt + qsec* was selected as the final model. This model improved the adjusted R-squared values from 0.779 to 0.839 and all coefficients were estimated to statistically significant levels. This model explains 83.9% of the variability in the data and suggests that on average, manual transmission vehicles get 2.93 more *mpg* than automatic transmission vehicles when holding *wt & qsec* constant.

Residual Analysis and Diagnostics

Residual plots were analyzed for this model and can be seen in Figure 4 on the appendix. The fitted values plotted againts the residual are normally distributed showing no noticeable pattern. The Normal Q-Q plot confirms the residual are normally distributted points laying closely to the line. The points on the Scale-Location plot were close to the line confirming constant variance.

```r
dfb <- dfbetas(model6)
dfb[which(abs(dfb)>1)]
```

```
FALSE [1] 1.093842
```

When analysing influence of the points using the dfbetas function, the weight of the Chrysler Imperial was found to have some influence in the model. The beta coefficient on the *wt* variable was 1.093 meaning that the weight of the vehicle was influential if it would have been removed from the model. This could be due to its relative higher values of *mpg* while having other parameters close in value when compared to other vehicles in its category as can be seen on the summary of cars with *wt* values above 5 listed below.

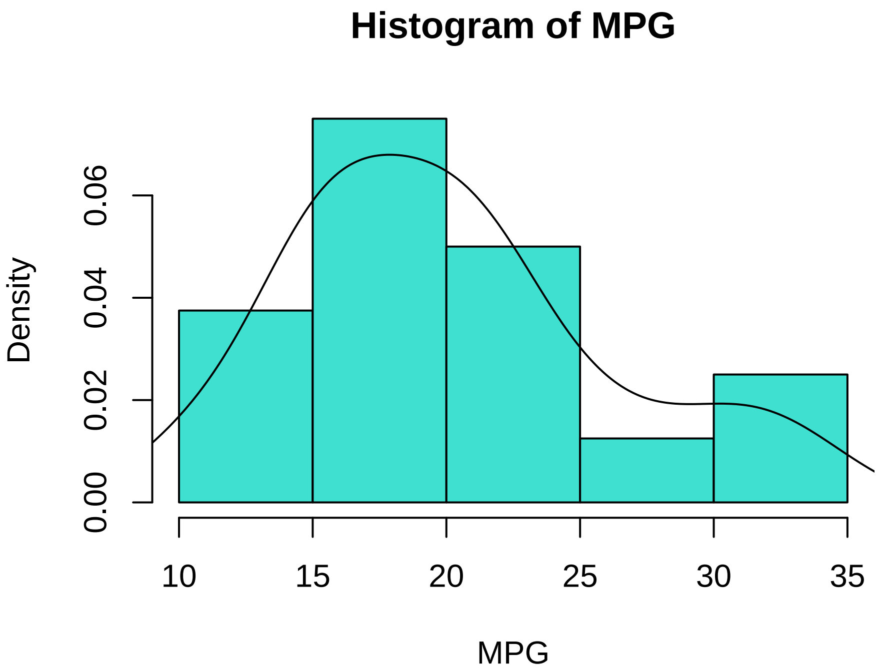
```
FALSE                      mpg    wt  qsec        am
FALSE Cadillac Fleetwood  10.4 5.250 17.98 Automatic
FALSE Lincoln Continental 10.4 5.424 17.82 Automatic
FALSE Chrysler Imperial   14.7 5.345 17.42 Automatic
```

Conclusion

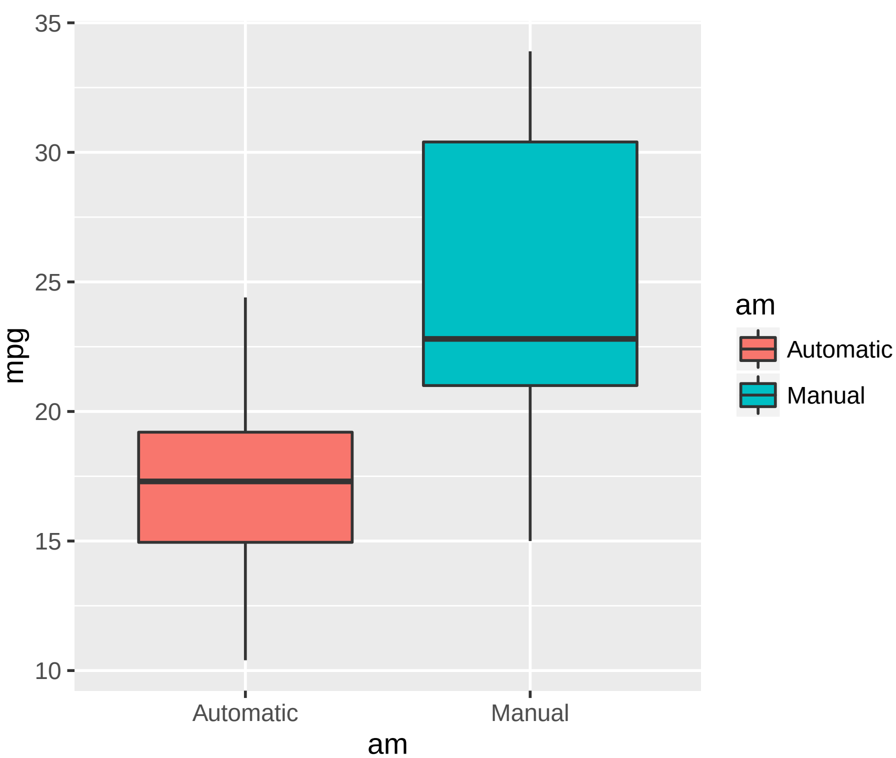
This analysis shows that vehicles with manual transmission have higher mpg in general but other confounders such as *wt & qsec* had influence in the response. The difference in average between manual and automatic transmission was 7.24 when not considering other confounders. When considering *qsec and wt* as additional confounders, manual transmission vehicles had 2.93 more *mpg* than vehicles with automatic transmission.

Appendix

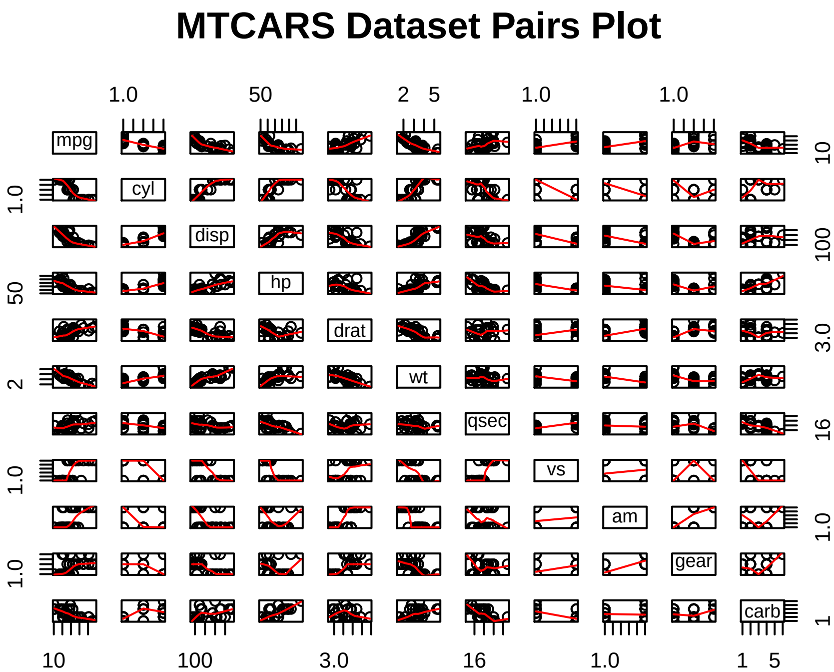
1. Histogram of MPG



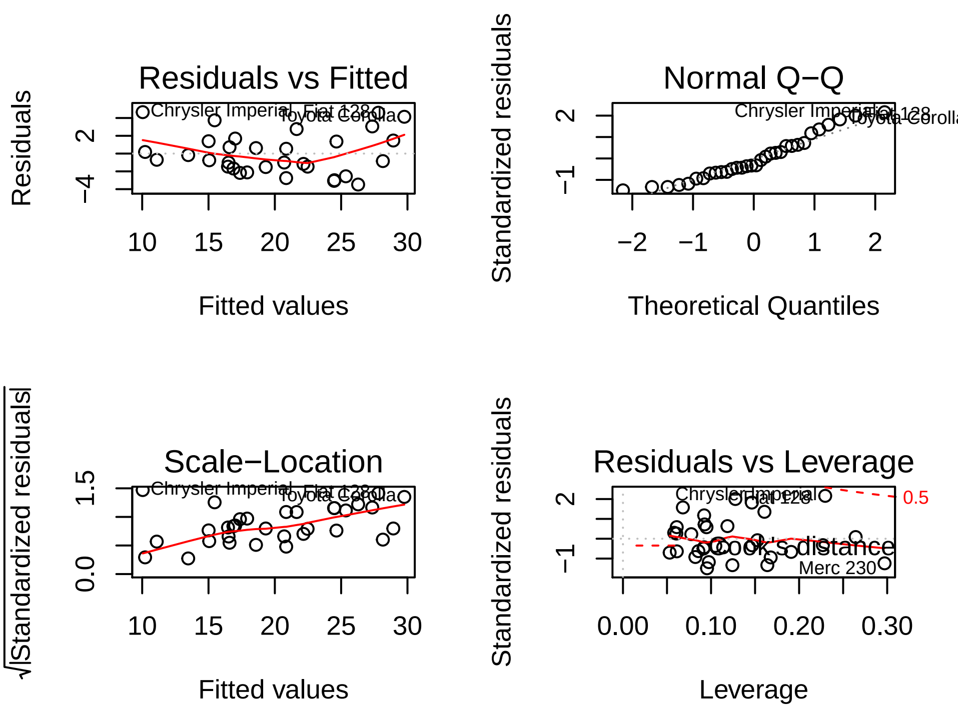
2. Boxplot MPG vs AM



3. Pairs Plot MTCARS Data



4. Residuals Plot of *mpg* ~ *am* + *wt* + *qsec* model.



5. ANOVA Test Results.

```
FALSE Analysis of Variance Table
FALSE
FALSE Model 1: mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear + carb
FALSE Model 2: mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs
```

```
FALSE Model 3: mpg ~ am + cyl + hp + wt + qsec
FALSE Model 4: mpg ~ am + hp + wt + qsec
FALSE Model 5: mpg ~ am + wt + qsec
FALSE Model 6: mpg ~ am + wt
FALSE   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
FALSE 1     15 120.40
FALSE 2     22 139.02 -7   -18.620  0.3314 0.927349
FALSE 3     25 143.98 -3    -4.959  0.2059 0.890698
FALSE 4     27 160.07 -2   -16.085  1.0019 0.390460
FALSE 5     28 169.29 -1    -9.219  1.1486 0.300788
FALSE 6     29 278.32 -1  -109.034 13.5836 0.002203 **
FALSE ---
FALSE Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```