Analysis of MPG for Automatic and Manual Transmission Vehicles

Johannes Rebane

November 23, 2014

Executive Summary

In this report, we first explore the data provided in the `mtcars` set and find that, in general, manual transmission has a higher mean mpg than automatic transmission; however, we quickly note that there are other correlative attributes that could be more explanatory than transmission type than transmission type.

By performing **three separate regression analyses** (a single variable regression, a regression with all variables included, and an optimized model derived using a stepwise algorithm), and by **measuring and comparing R-Squared Transmission P-Values** for each model, and by **quantifying and plotting model residual traits** we are able to conclude that **transmission type is not a statistically significant indicator of mpg**.

Exploratory Data Analysis

Before we dive into model selection, we perform some basic exploratory data analysis. In this we want to get a visual sense of automatic vs. manual mpg performance and also get a high level view of the correlation coefficients of each of the variables to start hypothesizing what confounding variables may exist.

Graphical Analysis: As we can see in **Figure 1**, by comparing the distribution of mpg for automatic and manual transmission cars, manual transmission vehicles seem to generally be associated with a higher mpg value. However, there also seems to be quite a bit of overlap, and a great deal of variance for mpg values for each transmission mode.

Basic Correlation Assessment: By running a basic assessment of correlations of the different variables with `mpg`, we can see that there are a number of other variables that have a higher absolute correlation with `mpg`. As you can see in **figure 2**, only `qsec`, `gear`, and `carb` are less correlated with `mpg`. To generate this chart, we plot the absolute values generated from running the following code: `abs(cor(mtcarsorig$mpg,mtcarsorig[,-1]))`.

Regression Analysis

First Model: Regression using a single variable

For our first regression analysis, we set transmission type as the sole predictor of the outcome, mpg, and run a regression accordingly.

```
single_fit <- lm(mpg ~ am, mtcars)
summary(single_fit)$coefficients
```

```
##              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
```

The coefficent, 7.2 suggests that, if we switch from automatic to manual transmission, we should expect an improvement (increase) of 7.2 miles per gallon. If we look at the **residual plots (Figure 3)** and see the distribution of residuals, we can see that the residual variation suggests that the model is a poor fit. Further supporting this, our `Adjusted R-Squared value` is ~0.36, implying that only ~36% of the changes in mpg are explained by this model. Therefore we will try different models to bring us closer to a good fit.

Second Model: Regression using all variables

Next, we plot a regression using all variables to see if this provides a better fit.

```
fit_all <- lm(mpg ~ am + cyl + gear + disp + hp + drat + wt, data = mtcars)
print(fit_all$coefficients)
```

```
##  (Intercept)     amManual         cyl6         cyl8        gear4
## 33.928387971  1.212894339 -3.043784271 -2.160826866  1.039781004
##        gear5         disp           hp         drat           wt
##  1.326883050  0.006425349 -0.036367740 -0.059724725 -2.874979253
```

The coefficent of ~1.35 suggests that, if we switch from automatic to manual transmission, we should expect an improvement (increase) of 1.35 mpg, much lower than our previous model. Now our `Adjusted R-Squared value` is ~0.868, implying that ~86% of the changes in mpg are explained by this model. Let's try one more model to see if we can increase the R-Squared even further.

Third Model: Regression using an optimized model

Here we leverage a stepwise algorithm to find a model that optimizes upon our previous model stored in `fit_all`.

```
step_model <- step(fit_all, trace = 0)
print(step_model)
```

We extract the ANOVA table to get a sense of how impactful the transmission factor is in this optimized model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.71	2.60	12.94	0.00
amManual	1.81	1.40	1.30	0.21
cyl6	-3.03	1.41	-2.15	0.04
cyl8	-2.16	2.28	-0.95	0.35
hp	-0.03	0.01	-2.35	0.03
wt	-2.50	0.89	-2.82	0.01

Based on this table, we can see that, while the coefficient indicates an increase in mpg of ~1.81 with manual transmission, the P-Value of `amManual` is much greater than some of the other factors in the table and, in fact, would not be statistically significant if we were assessing the importance with a 95% confidence interval. If we calculate our R-Squared as `cor(mtcars$mpg,predict(step_model))^2`, we see that the predictive accuracy of this model is higher than both previous models at 0.8658799.

Finally, the residual deviance of this model is higher than our previous model, and therefore has better fit. Therefore, we can conclude that transmission is not statistically significant as a predictor of mpg.

Appendix I: Figures

The following pages include figures referenced in the report.

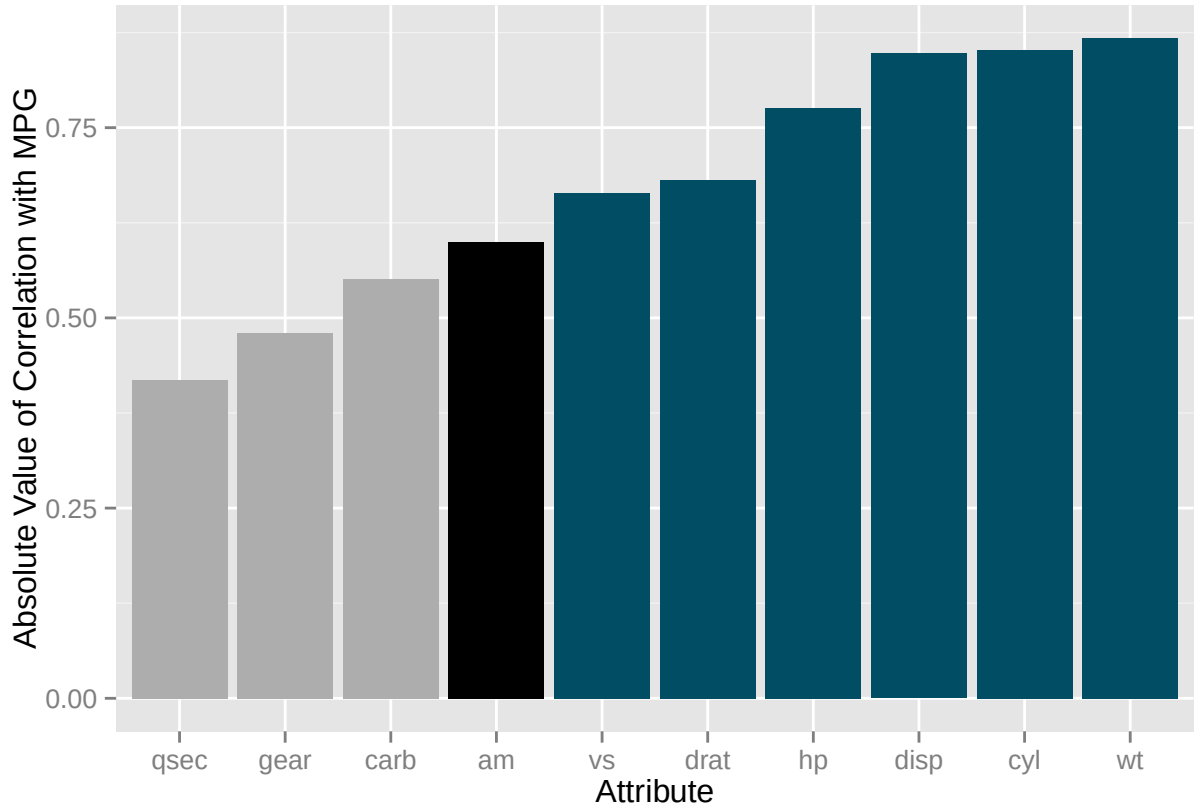

Figure 1: Absolute Value of correlations of vehicle attributes to mpg performance

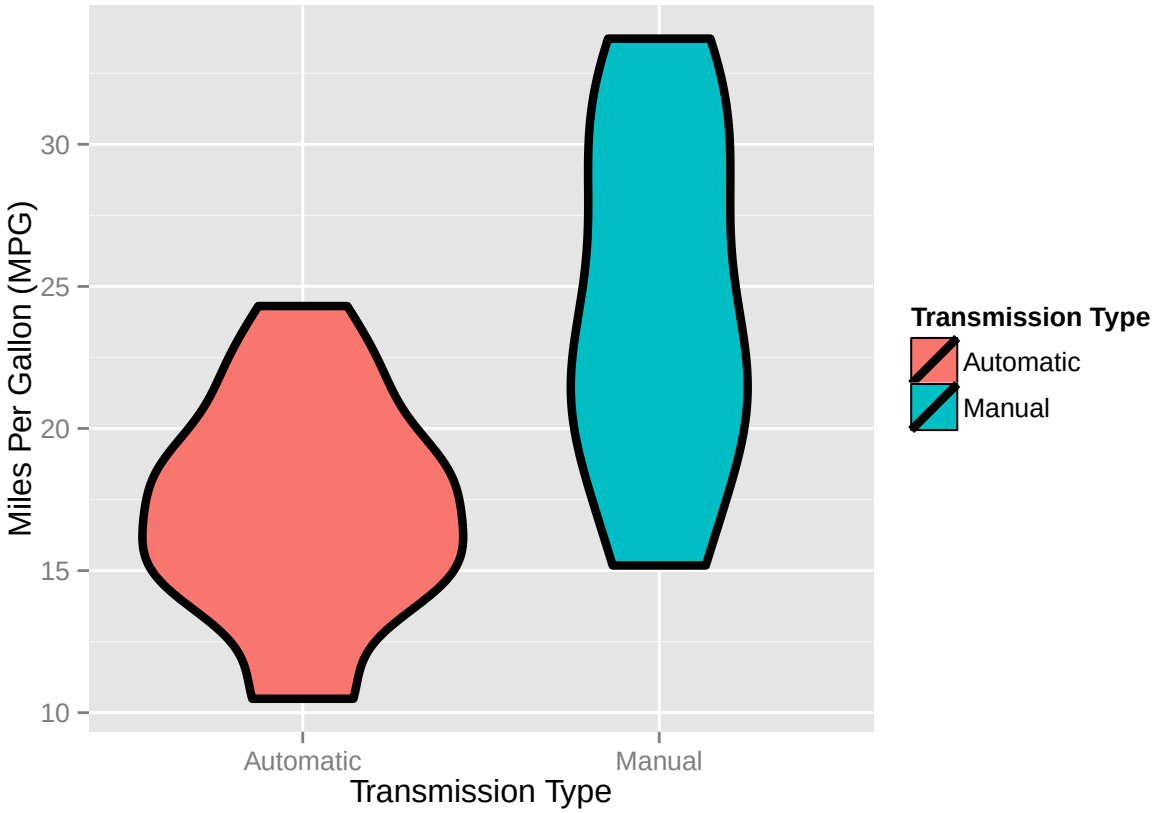

Figure 2: Distribution of MPG Performance by Transmission Type

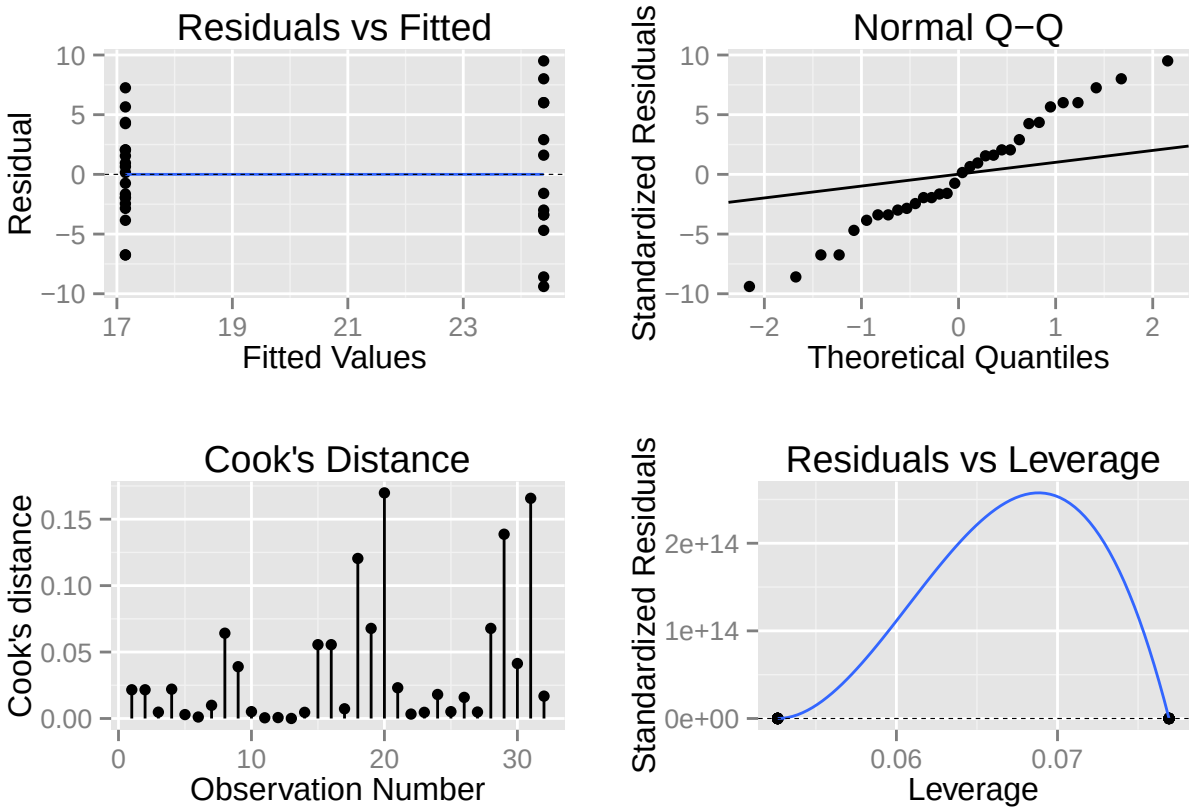

Figure 3: Residual Analysis of Single Variable Linear Regression