MPG Analysis on Motor Trends Data

Ash Chakraborty

July 24, 2015

EXECUTIVE STATEMENT

This report looks for the impact of transmission type (automatic or manual) on vehicle Miles Per Gallon (MPG). The dataset used is *Motor Trend* magazine's *mtcars* dataset of 1973-74 models. Multivariate regression analysis is the strategy used to build competing models in an attempt to incorporate any predictors that confound transmission's relationship with MPG.

The report concludes, *with 95% confidence*, that vehicles with manual transmissions have a statistically significant advantage of **between 0.05 to 5.83 MPG** (holding all other predictors constant) over vehicles that have an automatic transmission. The mean advantage for manuals observed in this dataset is **2.94 MPG**. Note: The code and complete analysis may be found at my github repo.

EXPLORATORY ANALYSIS

The *mtcars* dataset contains 32 observations across 11 variables. We take an initial look at the relationship between our focus of inquiry: MPG and Transmission Type. See the *violin plot* in *Figure 1.1* (appendix). We immediately note that there is non-constant variance between the two groups.

COMPETING MODELS

Model 0: Base Model

From the Exploratory Analysis above, we have our base model:

MODEL 0, **mpg = 17.1 + 7.24 * amManual**

Although this model suggests a significant difference of 7.24 in MPG for manual transmissions over automatics, the poor *adjusted* R^2 value of *0.33* is cause for concern. Moreover, the non-constant variance shown in *Figure 1.1* pretty much renders this model as unsuitable. *We keep it around for baseline comparisons only.*

Model 1: Step-wise Addition/Elimination

In order to consider the entire spectrum of possible predictors in the dataset (all are potentials), we perform a step-wise regression on all potential predictors (*mpg ~ .*):

```
model1 <- step(lm(mpg~., data=mtcars), direction="both", trace=FALSE)
```

MODEL 1: **mpg = 9.62 -3.92 * wt + 1.23 * qsec + 2.94 * amManual** ; *adjusted* R^2 is *0.83*.

Note: There's a concern here that the mean MPG for automatic transmissions (the intercept), holding other predictors constant, is *not* significant.

Model 2: Linear Correlation + Step + VIF

We turn to linear relationship with the response to help us pick likely predictors (*Figure 1.2*, as shown by *adjusted* R^2, and a significant coefficient). The following candidates emerge (in decreasing strength of R^2): **wt, cyl, disp, hp, and drat**. We then use the *step-wise* addition/elimination process on these candidates, combined with a *VIF* analysis to eliminate *multicollinearity*; we get the equation: $mpg\ am + wt + hp$.

MODEL 2, **mpg = 34 -3.92 * wt - 1.23 * hp + 2.94 * amManual** ; *adjusted* R^2 is *0.82*.

Note: There's a concern here that the mean MPG difference for manual transmissions, holding other predictors constant, is *not* significant.

VERIFYING REGRESSION ASSUMPTIONS

Variance of Means Test

We want to be certain that the predictors added to each model cause a significant difference in the sum of squares and overall variation. We verify this with an ANOVA test for Models 1 and 2 (**ONLY 1 is SHOWN**):

Table 1: Model 1 ANOVA Test

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
30	720.8966	NA	NA	NA	NA
29	278.3197	1	442.5769	73.20250	0.0000000
28	169.2859	1	109.0338	18.03425	0.0002162

We see that in each model (entire code available at my github repo), adding the predictors causes significant change in group variance.

Residual Diagnostics

Finally, we take a loot at some Residual plots (*Figure 2.1*). The residual analysis shows us that Models 1 and 2 are roughly homoskedastic, while their residuals approximate a normal distribution. Model 2, however, does have some outliers far away from the quantile line. These need to be investigated. Model 1 also has some outliers that need a closer look. *Can we eliminate these?*

Evaluating Influence

Figure 2.1 also shows us some leverage points for Models 1 and 2. We're concerned about leverage points that *also* exert more influence on the model; this might cause the regression line to bend unfairly towards such values. In order to quantify this combination of influence and leverage, we see the major *Cook's Distance* measures for each such leverage point (**ONLY model 2 is SHOWN**):

	Model 2: Cook's Distances of Concern
Chrysler Imperial	17
Fiat 128	18
Toyota Corolla	20

In reviewing these outlier records, however, there is *no* indication of erroneous data points. These are merely extreme specimens of a combination of predictor values. We therefore err on the side of caution with our regression assumptions by choosing *not* to remove these data points from our models.

BEST FIT MODEL and CONCLUSIONS

We see that Model 1 has a slight advantage over Model 2 in terms of the *adjusted* R^2. Moreover, the overall outlier exertion on the model (judged in terms of their mean Cook's Distance) seems to be slightly better for Model 1. We therefore choose model 1 to represent our best fit model:

BEST FIT MODEL: **mpg = 9.62 -3.92 * wt + 1.23 * qsec + 2.94 * amManual**

The average MPG for vehicles with manual transmissions - while holding weight, quarter mile time, and the automatic transmission coefficients constant - sees an advantage of **2.94 Miles Per Gallon** over the automatics in this dataset. Furthermore, we state with *95% confidence*, that manual transmissions enjoy a positive advantage in the range of **0.05 to 5.83 MPG** over their automatic counterparts, while holding weight and qsec values constant at the coefficients shown by the equation above.

Finally, *Figure 3* summarizes the relationship between the automatic and manual transmission groups in the 3d scatter. We see that the best fit plane of both groups have different slopes, as suggested by our model. It's interesting to note that heavier cars seem to have automatic transmissions.

APPENDIX

Entire markdown, including my code is available at my github repo.

Figure 1.1: Violin Plot Exploring MPG Vs. Transmission

This scatter superimposed on a violin plot shows us that there is non-constant variance between two transmission types. This pretty much *eliminates* the viability of Base Model 0 (mpg ~ am) as a suitable candidate for our analysis.

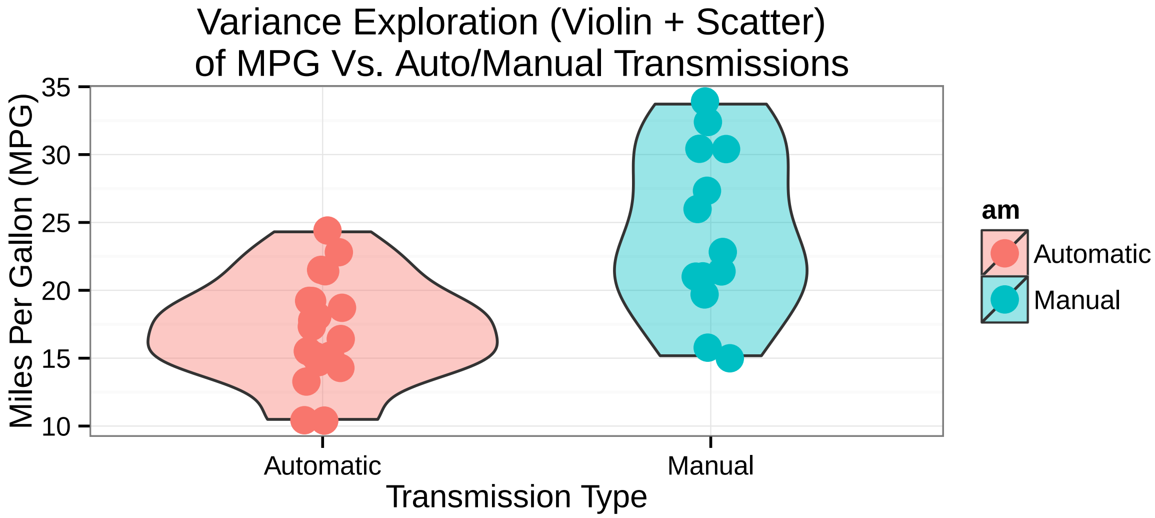


Figure 1.2: Linear Correlations (Response-Predictor Pairs)

This scatterplot produces a pairs plot with the lower panel showing the *adjusted* R^2 value between response and predictor, as well as the coefficient's p-value. The font size increases by strength of the correlation.

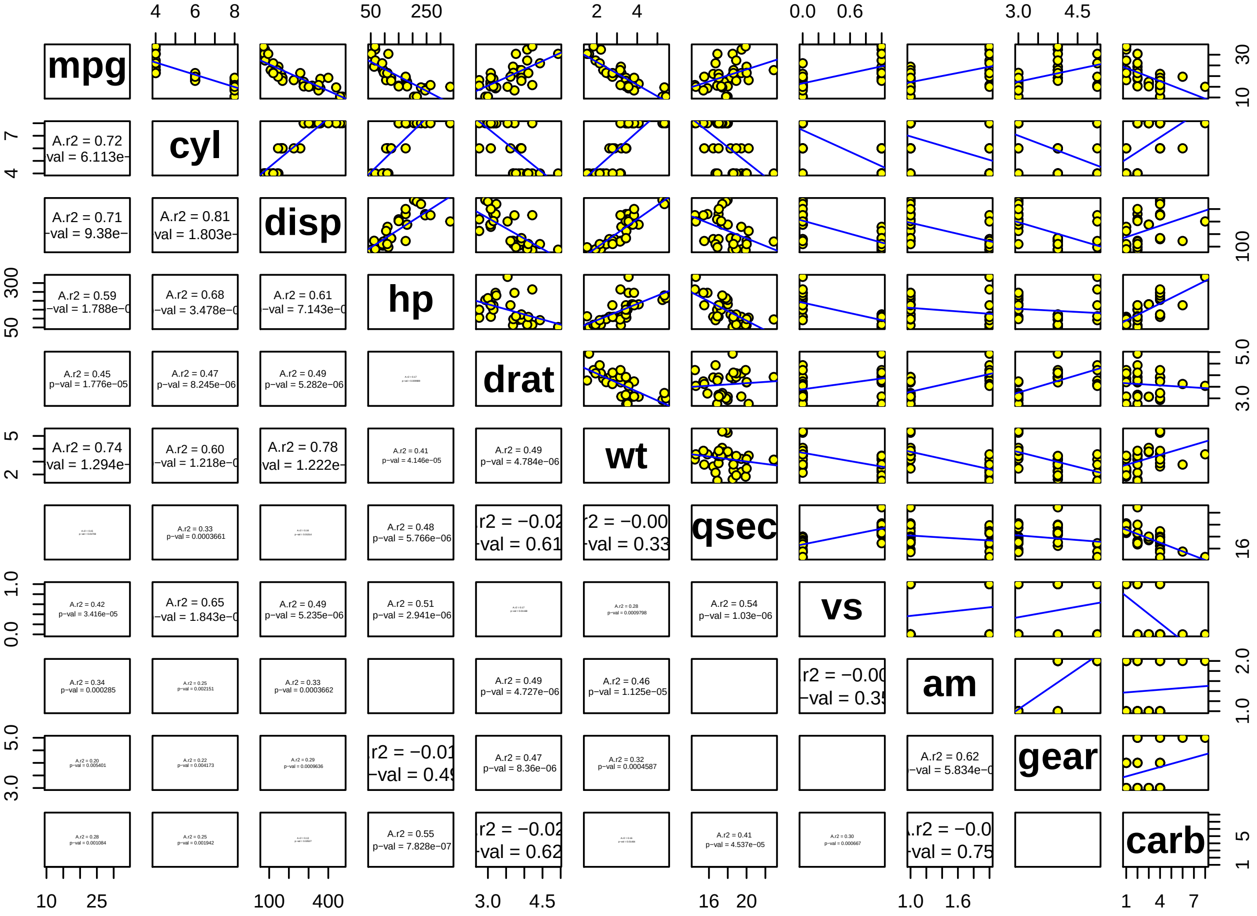
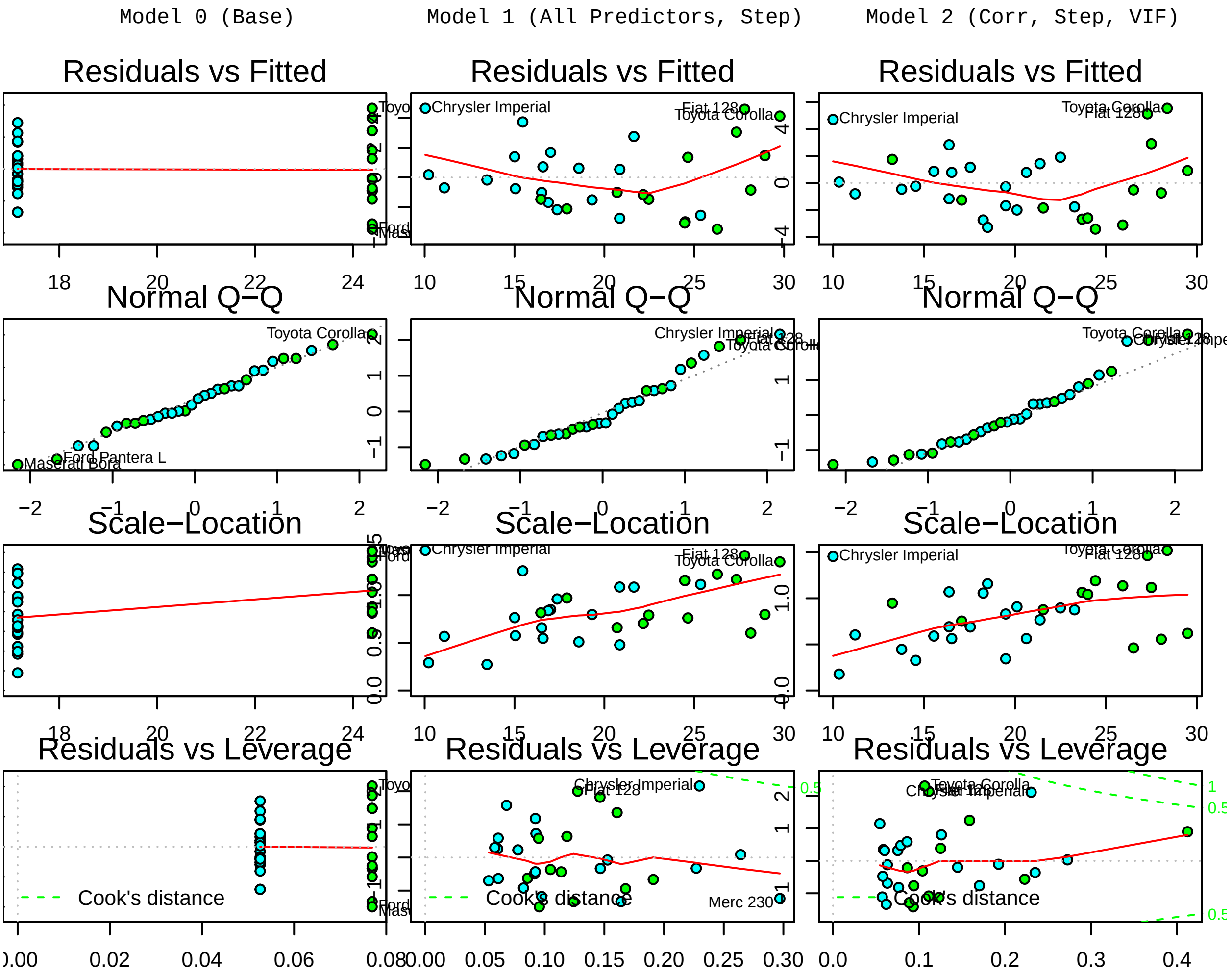


Figure 2.1: Residual Diagnostic Plots

We plot 3 Residual diagnostics for each model in each column of the grid below:

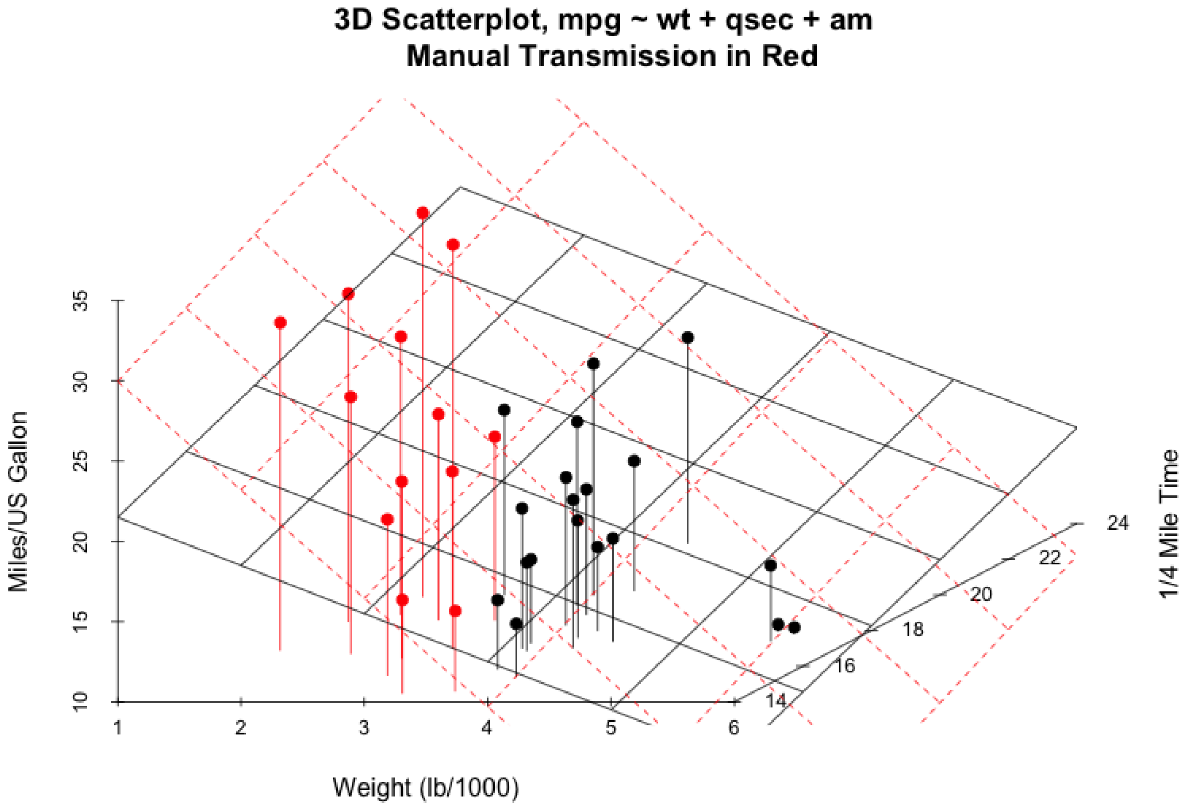


Influence Plots (not shown) may be seen here.

Figure 3: Best Fit Model (Scatter)

Finally, we visualize a 3D scatter of our best fit model. The red plane shows the best fit plane for manual transmissions. As expected, it highlights a difference in slope from the best fit automatic plane in black:

mpg = 9.62 -3.92 * wt + 1.23 * qsec + 2.94 * amManual



Note: *Knitr* has a bug rendering my second regression plane on the 3d plot, so I inserted an image instead. The code for the plot may be seen at my github repo.