# MPG Analysis on Motor Trends Data

*Ash Chakraborty*

*July 24, 2015*

## Executive Statement

This report looks for the impact of transmission type (automatic Vs. 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**.

## 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 "1Figure 1.1 (appendix). We immediately note that there is non-constant variance between the two groups. This renders this basic relationship as incomplete.

## Competing Models

### Model 0: Base Model

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

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 17.147368 | 1.124602 | 15.247492 | 0.000000 |
| amManual | 7.244939 | 1.764422 | 4.106127 | 0.000285 |

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 ($mpg \sim .$):

| | estimate |
|---|---|
| (Intercept) | 9.617781 |
| wt | -3.916504 |
| qsec | 1.225886 |
| amManual | 2.935837 |

The most optimal model coming out of Stepwise Addition/Elimination is:

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

The concern above has led us to look for yet other means of explaining confounding predictors. We turn to linear relationship with the response to help us pick likely candidates. We thus take a look at scatter plots of MPG against every potential predictor in Figure 1.2. We then select predictors based on the following criteria:

- High *Adjusted* $R^2$ (showing a high linear relationship with response)
- Slope coefficient that's significant.
- Approximate linear relationship with fitted line, few outliers acceptable.

The following candidates emerge (in decreasing strength of $\mathrm{R}^2$): **wt, cyl, disp, hp, and drat**. We now use the *step-wise* addition/elimination process on the potential equation, $mpg\ am + wt + cyl + disp + hp + drat$. We get:

Table 3: Coefficients for Model 2

| | estimate |
|---|---|
| (Intercept) | 38.7517874 |
| wt | -3.1669731 |
| cyl | -0.9416168 |
| hp | -0.0180381 |

Next, we force "am" into the model, and thus extract the result from the second-last step the procedure above. Finally, we guard against *multicollinearity* by performing a *Variance Inflation Test* between the predictors and *eliminating* those that show an abnormal bump in standard deviation:

Table 4: SD Inflation for Model 2

| | SD Inf |
|---|---|
| am | 1.595669 |
| wt | 1.997074 |
| cyl | 2.309477 |
| hp | 2.076061 |

After removing the co-dependent predictor "cyl" we finally arrive at a fairly orthogonal predictor set:

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 Models

## 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:

Table 5: 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 |

Table 6: Model 2 ANOVA Test

| Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|
| 30 | 720.8966 | NA | NA | NA | NA |
| 29 | 278.3197 | 1 | 442.57690 | 68.73415 | 0.0000000 |
| 28 | 180.2911 | 1 | 98.02863 | 15.22428 | 0.0005464 |

We see that in each model, adding the predictors causes significant change in group variance.

## Residual Diagnostics

Finally, we take a loot at some Residual plots (Figure 2.1) to verify that the following assumptions **hold true** for each model:

1. Homoskedasticity of residuals around fitted line: This holds approximately for all models.
2. Error terms are approximately normally distributed: This holds approximately for all models, with some outliers for Models 1 and 2.
3. Non-systematic residuals (no apparent patterns): This is mostly pattern free.

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 2.2 shows us the concerning leverage points in each model. We're concerned about leverage points that *also* exert more influence as shown by the deviation from the standardized residuals. This might cause the regression line to bend unfairly towards such values. In order to quantify this combination of influence and leverage, we list the highest *Cook's Distance* measures for each such leverage point:

| | Model 1: Concerning Cook's Distances |
|---|---|
| Merc 230 | 9 |
| Chrysler Imperial | 17 |
| Fiat 128 | 18 |

| | Model 2: Concerning Cook's Distances |
|---|---|
| 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

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**

## Conclusions

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. We can see these summarized in the confidence intervals generated below:

Table 9: Confidence Intervals for Coefficients of Model 1

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | -4.6382995 | 23.873860 |
| wt | -5.3733342 | -2.459673 |
| qsec | 0.6345732 | 1.817199 |
| amManual | 0.0457303 | 5.825944 |

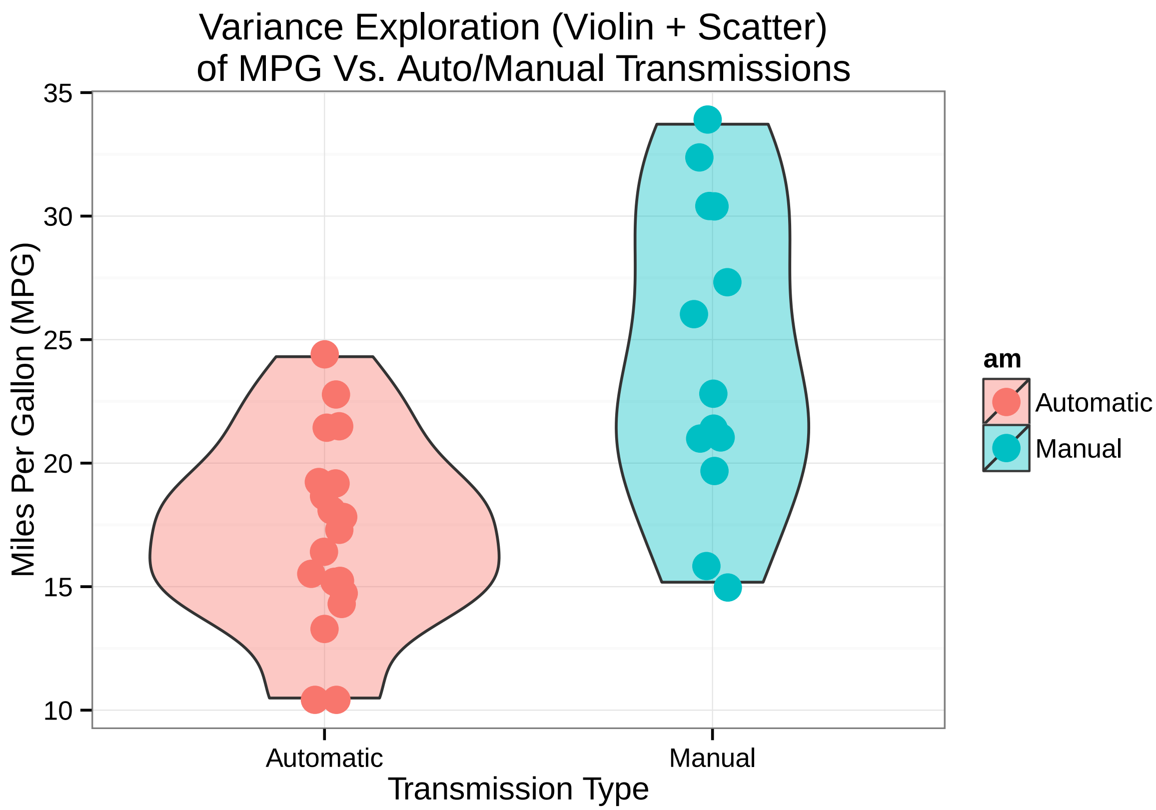
Finally, *Figure 3* summarizes the relationship between the automatic and manual transmission groups in the 3d scatter. We see that the best fit plane of both groups have different slopes, as suggested by our model. It's interesting to note that heavier cars seem to have automatic transmissions.

**END OF REPORT**

## APPENDIX

### 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.

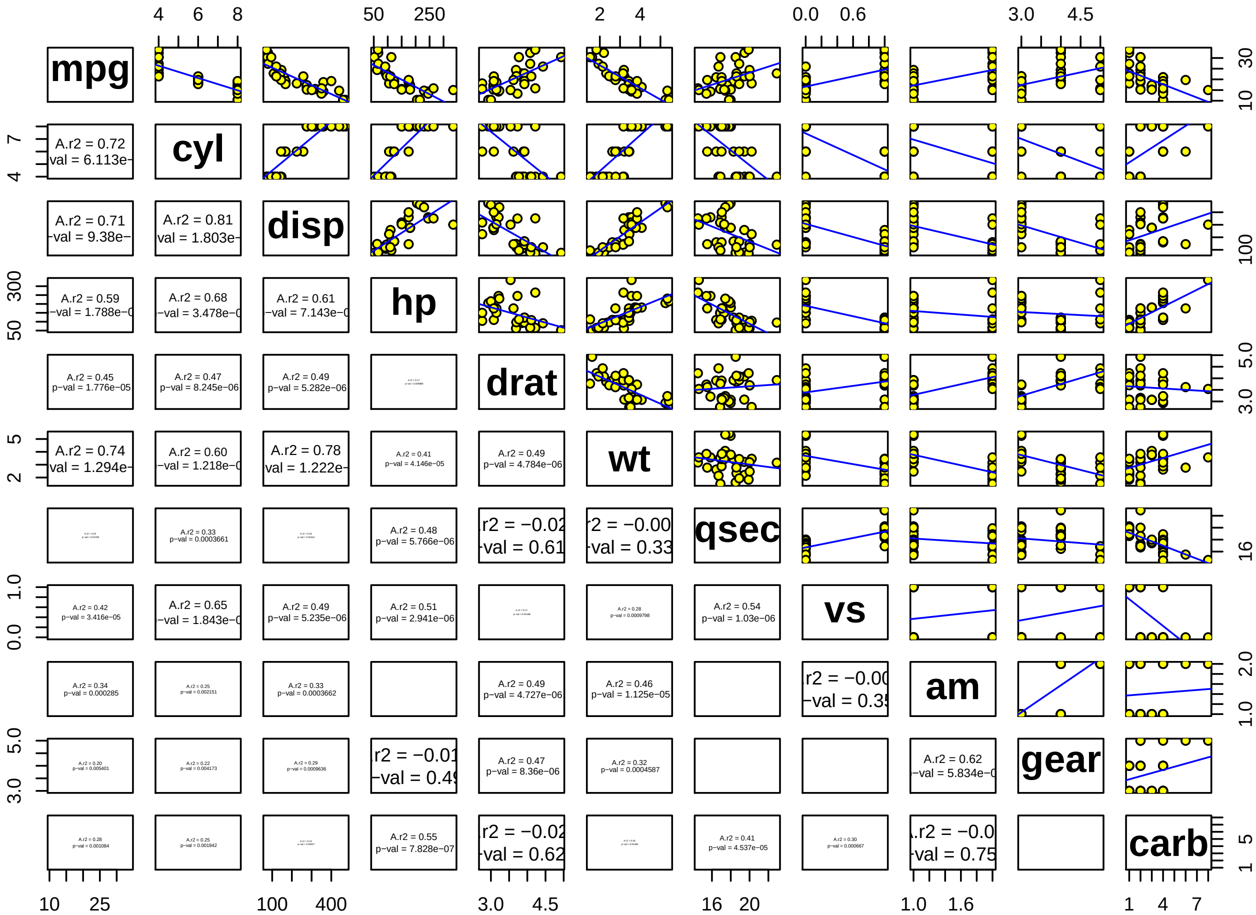



**Figure 2.1: Residual Diagnostic Plots**

We plot 4 Residual diagnostics for each model in each column of the grid below:

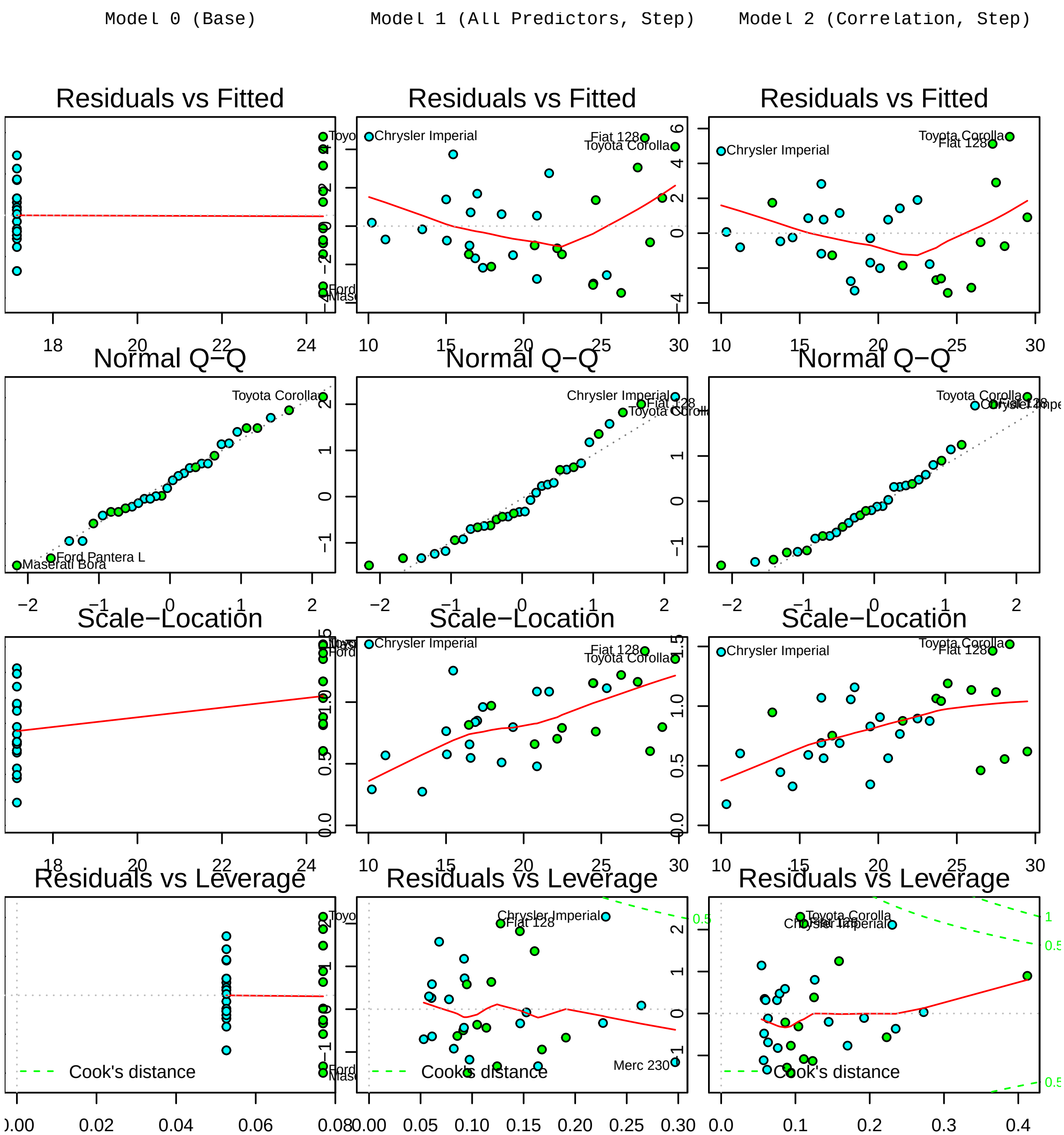
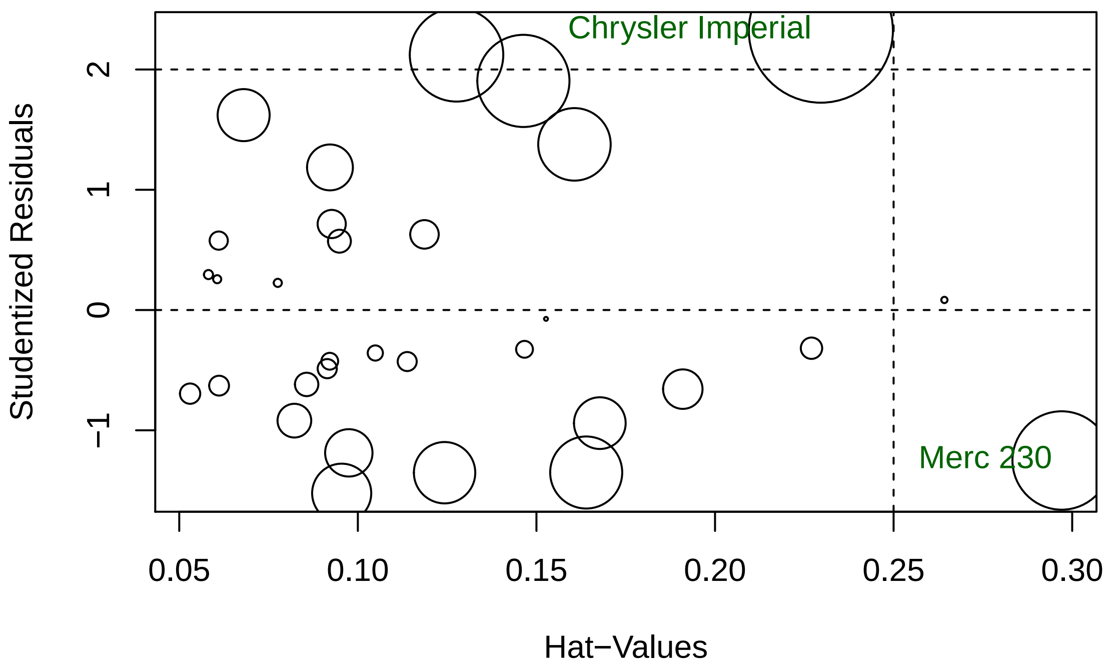
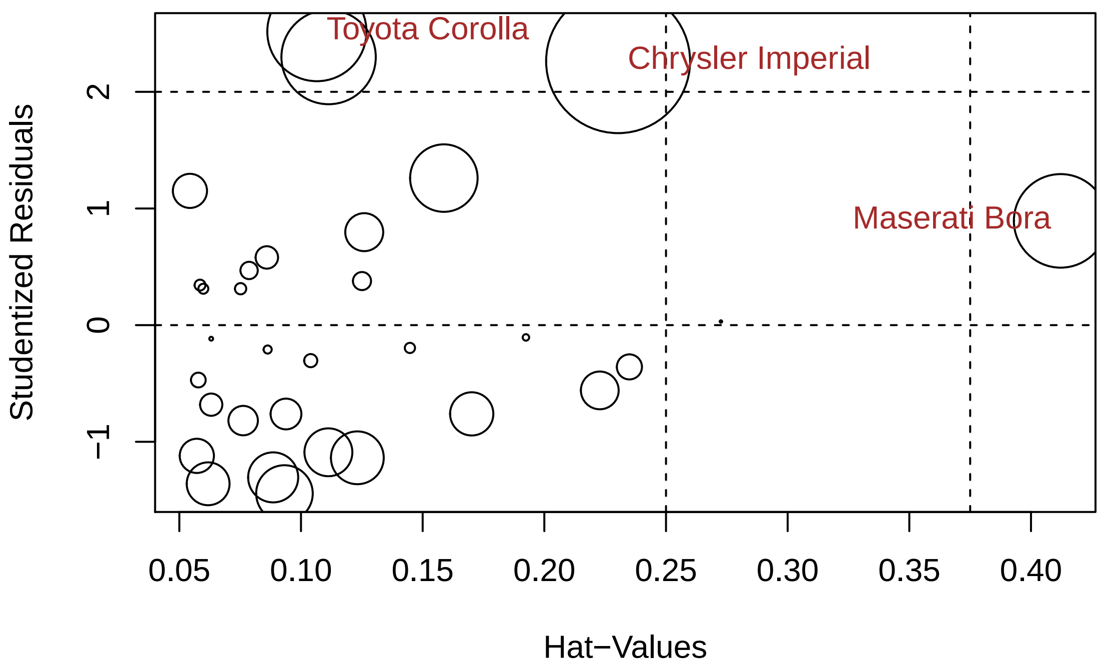



**Figure 2.2: Influence Plots**

Closer look at Leverage points (hatvalues) and their deviation from standard residuals, for each model:

```
##                     StudRes       Hat     CookD
## Merc 230          -1.251106 0.2970422 0.4025949
## Chrysler Imperial  2.323119 0.2296338 0.5895739
```




```
##                     StudRes       Hat     CookD
## Chrysler Imperial 2.2627843 0.2303244 0.5778554
## Toyota Corolla    2.5163024 0.1065328 0.3981850
## Maserati Bora     0.8938865 0.4121968 0.3756236
```

### 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**