Effect of Car Design Aspects on Fuel Economy

Jonathan Owen

July 26th 2015

Executive Summary

The fuel economy of a car should be related to some of its design aspects. The following study uses reported values of fuel economy and these design aspects to establish relationships in the form of linear regression models. Horsepower, weight, and time to travel $\frac{1}{4}$ mile are found to be the most useful predictors of fuel consumption.

Data

The `mtcars` dataset contains observations of fuel economy and 10 design aspects reported in *Motor Trend* magazine for 32 car models available in the United States between 1973 and 1974. The variables are described in the table below.

`mpg`	fuel economy, miles/US gallon	`drat`	rear axle ratio	`am`	transmission
`cyl`	number of cylinders	`wt`	weight, 1000 lb	`gear`	number of forward gears
`disp`	displacement, cubic inches	`qsec`	1/4 mile time, seconds	`carb`	number of carburettors
`hp`	gross horsepower	`vs`	V or straight cylinder		

Exploratory data analysis identified several possible predictors as well as issues prior to constructing parsimonious models for fuel economy. A panel of scatterplots and correlations for all variable pairs is included as Figure 1 in the appendix. For reasons explained in the **Initial Model** section, it is easier to understand the model in terms of fuel consumption `gpm` in US gallons per mile rather than fuel economy `mpg`. The two quantities are related by each being the reciprocal of the other. Plots of fuel consumption against the 10 design aspect variables show strong negative correlations (absolute values 0.78 - 0.87) with displacement, horsepower, and weight. It is also clear that these variables are positively correlated with each other (0.66 - 0.89), which may cause problems if they are all are used in the model. The number of cylinders falls into this same group, but is not an initial choice for the model because it is not continuous and is correlated with displacement (0.90) as would be expected.

Initial Model

The exploratory data analysis can be supplemented with some theory to create an initial model. The work done by a moving car during a time period Δt, such as `qsec` in the dataset, is $P(t)\Delta t$ where $P(t)$ is the power at time t. Although the exact form of $P(t)$ is not reported for these cars, it should be proportional to the gross horsepower, which *is* available. The work done should be proportional to both `hp` and `qsec`.
The energy required to do this work comes from the fuel and is proportional to the volume of fuel used. Not all of this energy goes into driving the car. There are losses that can be grouped into an efficiency factor η, which should depend on some of the design aspects. The following relationship should be true for a car travelling $\frac{1}{4}$ mile

$$\frac{\eta}{4 \cdot \texttt{mpg}} \propto \texttt{hp} \cdot \texttt{qsec}$$

or

$$\frac{\eta \cdot \texttt{gpm}}{4} \propto \texttt{hp} \cdot \texttt{qsec}$$

The inital model is a fit of `gpm` against the interaction `hp:qsec`.

$$\texttt{gpm}_i = \beta_0 + \beta_1 \texttt{hp}_i \cdot \texttt{qsec}_i + \epsilon_i$$

The intercept β_0 is not significant ($p = 0.32$). R^2 increases from 0.69 to 0.96 when the intercept is removed from the model.

Model Improvement by Residuals Adjustment

Residuals adjustment were used to find any additional predictors—ones that are rolled into the efficiency factor η. Residuals from fitting `gpm` and the remaining car design aspects using `hp:qsec` are plotted against each other. These are included in Figure 2. High correlation in the residuals plots identifies predictors that will best fit variation that remains from the initial model. Of the unused predictors, `wt` has the highest correlation (0.81) with `gpm` after removing the effects of `hp` and `qsec`. A second model that included `wt` as a predictor

$$\texttt{gpm}_i = \beta_1 \texttt{hp}_i \cdot \texttt{qsec}_i + \beta_2 \texttt{wt}_i + \epsilon_i$$

produces a model with significant coefficients $\beta_1 = 5.92 \times 10^{-6}$ and $\beta_2 = 1.20 \times 10^{-2}$ and R^2 of 0.99.

ANOVA and Residuals Analysis

Analysis of variance ANOVA on the nested models indicates that reduction of the sum of the residual squared from 0.0042 to 0.0041 by including `wt` is significant ($p = 1.3 \times 10^{-8}$).

Lastly, four plots of the residuals can be used to assess the fit. These are included in Figure 3. In the plot of residuals vs. fitted values points should be randomly distributed around a horizontal line corresponding to zero residual. This is approximately the case in Figure 2 although there is some deviation at higher fitted values. Similar behavior is expected for standardized residuals vs. fitted values and this is also approximately the case for the model. Of note is that all the standardized residuals appear to be within 2 standard deviations of zero.

If the residuals are normal they should lie close to a diagonal line in the Q-Q plot. Again, this is mainly true but there is some deviation at lower values.

In the plot of residuals vs. leverage, there are some concerns about high leverage points, such as the Maserati Bora; however, the standardized residuals for this point does show it has a large effect despite the potential leverage. Of greater concern is the location of Chrysler Imperial point in relation to the Cook's distance 0.5 contour. Cook's distance measure the effect of deleting a point on the model. Higher Cook's distances correspond to outliers (high residuals) or high leverage. Ideally, the Chrysler Imperial requires further examination.

Overall, the residuals analyses suggest that, although the model has a high R^2 value, there may be missing predictors because the distribution of residuals does not appear entirely random. However, testing with other variables in the mtcars dataset did nothing to improve the appearance of residuals plots. Additional predictors that could improve the fit do not seem to have been included in the dataset.

Intepretation of the coefficients and uncertainty in the model

Coefficient β_1 implies that for an increase of 1 hp the expected fuel consumption increases by 5.92×10^{-6} US gallons per mile if all other variables are unchanged. The same change occurs for each 1 second extra on the $\frac{1}{4}$ mile time. Coefficient β_2 implies that for each 1000lb increase in the weight of the car, the expected fuel consumption increases by 1.20×10^{-2} US gallons per mile. Although these are the expected values, there is uncertainty in these estimates. The 95% confidence intervals are 1.99×10^{-6} to 9.86×10^{-6} for β_1 and 0.89×10^{-2} to 1.52×10^{-2} for β_2. Separate prediction intervals would need to be calculated for predictions using the model with new data.

Effect of transmission type

Unfortunately, the model does not directly address the effect of transmission type. As noted in the exploratory data analysis, `am` is correlated with `wt` (0.69) as well as other variables of interest. In the dataset the weight range for manual transmission cars is 1,500 - 3,500 lbs and for automatics is 2,500 - 4,500 lbs. The overlap of the two ranges doesn't provide enough data points to try to separate the effect of transmission from weight. Also, outside of the data, the effect of automatic transmissions is consider to be primarily due to the increased weight. So even though the transmission isn't include the difference between a manual and automatic transmission versions of otherwise identical cars could be predicted using their weights.

Appendix

The color of the points indicates transmission type in all plots. **Red points** are automatic. **Blue points** are manual.

Figure 1 Scatterplot and correlation of variable pairs from the `mtcars` dataset

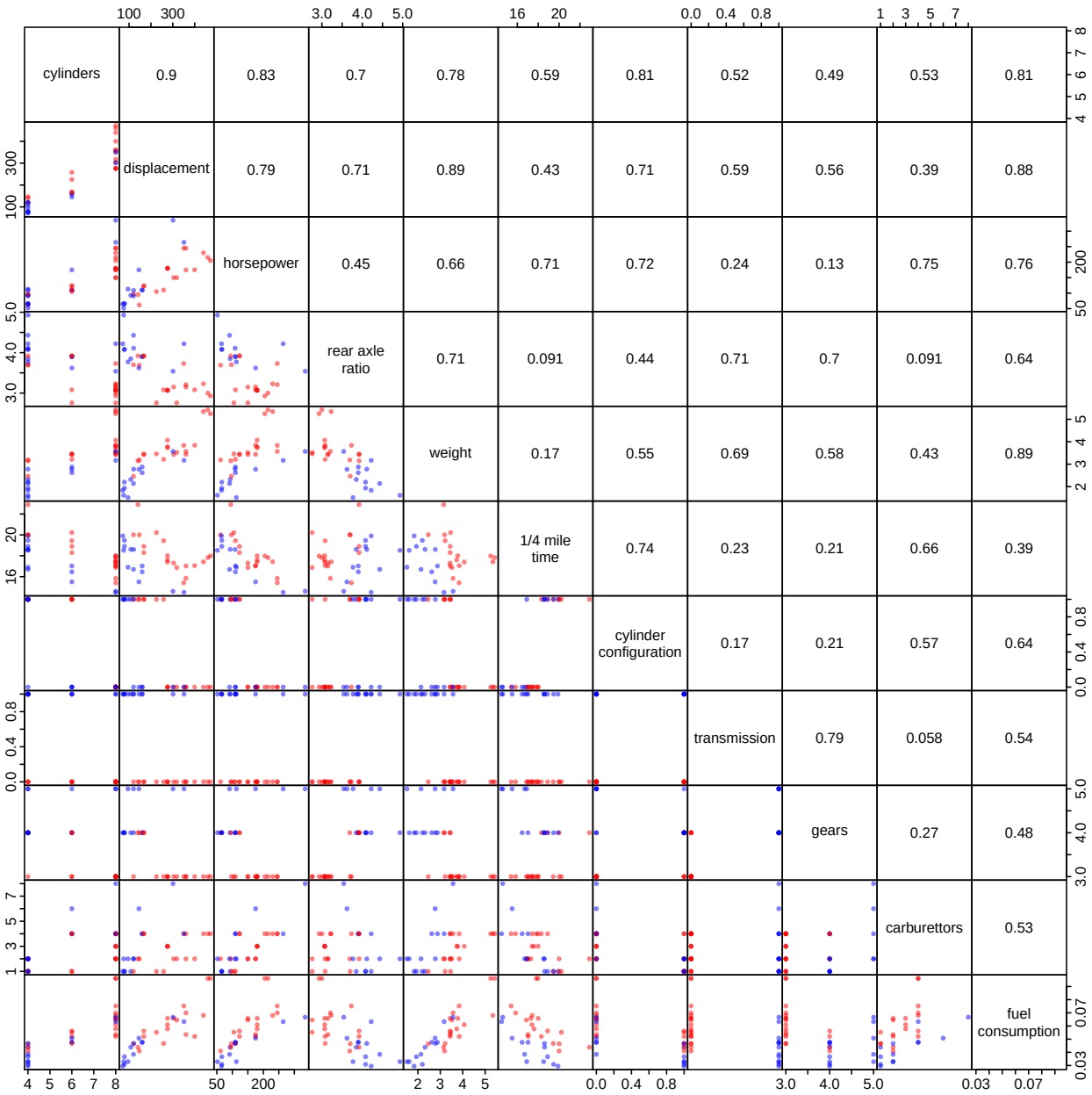

Figure 2 Residuals of `gpm` vs. other variables after fitting with `hp:qsec`

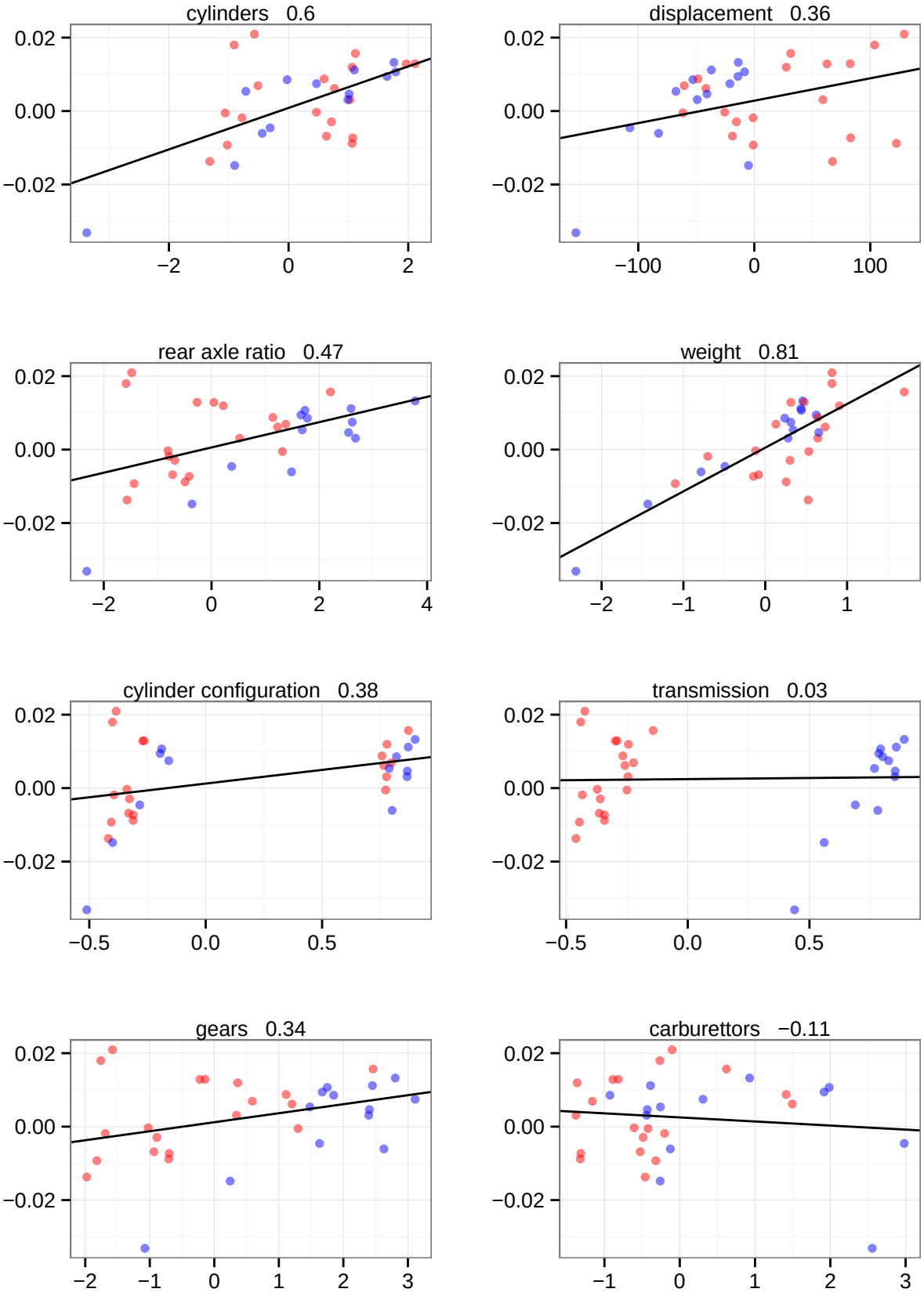

Figure 3 Residuals analysis of the final model