Motor Trend Analysis

atops

July 14, 2015

Context

This analysis is performed on behalf of Motor Trend, a magazine about the automobile industry. We are interested in the relationship between a set of variables and miles per gallon (MPG) (outcome). In particular, we are interested in the following two questions:

1. “Is an automatic or manual transmission better for MPG”
2. "Quantify the MPG difference between automatic and manual transmissions"

Executive Summary

Multiple linear regression models were created interactively answer these questions. Our focus was on seeking to understand the data rather than on purely optimizing the statistical values. The best fit model with the simplest form was to predict mpg based on transmission (am) and horsepower (hp). Manual transmissions are better for mpg. One can expect a manual transmission car get 5.2 mpg better miles per gallon than an automatic with the same horsepower.

Approach

In order to fit a model to answer the questions above, we will follow a process as follows:

1. Create some scatter plots showing relationships between predictor (and response) variables
2. In addition to transmission type (the predictor in question), make judgments about which variables may be most relevant to mpg, based on the scatter plots and what we know about cars.
3. Start with the simplest and most naive model of mpg as a function of transmission type. Call this the base model.
4. Seek to improve the base model through the addition (or subtraction) of variables based on F-statistics from ANOVA analysis of adjusted and unadjusted models (the impact of adding the variable).
5. Validate no outliers are having undue influence on the model based on hat values (measures of leverage) and df betas (change in individual coefficients when the ith point is deleted in fitting the model).
6. Settle on the best predicting yet most parsimonious model for mpg based on the available predictor variables. Draw conclusions on impact of transmission type by the coefficient, which shows the impact on mpg with different transmission types, holding all other model variables constant. The affect of all other unmodeled variables is captured in the error term, which we will try to keep as manageable as possible.

Exploratory Data Analysis

The Motor Trend data set has 32 observations on 11 variables:

Table 1. Motor Trend Vehicle Data Set Variables

var	Variable Description
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (lb/1000) *
qsec	1/4 mile time *
vs	V/S (V engine or straight engine)
am	Transmission (0 = automatic, 1 = manual) *
gear	Number of forward gears
carb	Number of carburetors

A pairs plot is included in the appendix as Figure 1. Based on inspection, the following variables appear significant:

1. am: The predictor in question, transmission type (manual or automatic) should have an impact on gas mileage. See also the box plot in Figure 2. Cars with manual transmissions should get better gas mileage than automatic.
2. cyl: Cars with smaller engines (fewer cylinders) should get better gas mileage. And there appears to be a relationship in Figure 1 to support this. mpg goes down with an increase from 4 to 6 to 8 cylinders.
3. wt: Weight logically should impact gas mileage as it takes more energy to move a heavier vehicle. The data appear to support this with a strong downward trend in mpg with an increase in weight.

Several other variables appear to have a relationship with mpg, but perhaps not as strong. The only variable I would exclude off the bat is number of forward gears (gear), which seems to have little or no relationship with mpg. We will have to consider which ones truly add value to the model.

Model Selection: Base Model

Using only transmission type (am) as a predictor for mph, the base model coefficients are as follows:

	Estimate	Std. Error	t value	Pr(>|t|)
am	7.245	1.764	4.106	0.0003
(Intercept)	17.147	1.125	15.248	0.0000

Table 2. Base Model--mpg as predicted by transmission (am)

Observations	Residual Std. Error	R^2	Adjusted R^2
32	4.902	0.3598	0.3385

Based on a 95% confidence level (p-values below 0.05), the intercept and predictor variable are significant so we reject the hypothesis that their coefficients are 0. The intercept term represents the expected mpg for automatic transmissions (am=0) and the intercept plus the am coefficient represents the mpg for manual transmissions (am=1). These values are 17.1 and 24.4, respectively.

R^2 is low, however, suggesting the model can be improved. Observation of the data set reveals several exotic sports cars (e.g., Lotus, Ferrari, Maserati) with manual transmissions. These are not necessarily representative of the population of vehicles and likely reduce the expected gas mileage of cars with manual transmissions. This suggests we should look to variables such as horsepower or displacement to control for this effect.

Regarding outliers, based on the plots of Figures 3, 4, and 5, there are no points that exert undue leverage on the regression results. The hat values are measures of leverage and df betas are the change in individual coefficients when the ith point is deleted in fitting the model.

Model B: Adding Horsepower

The addition of Horespower to the model gives the results in Table 2. Adding horsepower is a major improvement to the model. R^2 goes from 0.3598 to 0.78203 compared with the base model. The differentiating effect of hp can be seen in Figure 6. In Table 3, the Analysis of Variance Table, the high F value and low Pr(>F) enable us to conclude the hp variable is a significant predictor of mpg.

Conclusion

In the Appendix, there are several other models that were created but discarded as they did not improve upon the model with transmission (am) and horsepower (hp) as predictors for mpg. One of these models included an interaction term between the two predictor variables. This did not improve the model, somewhat surprisingly. In answer to the questions, (1) Manual transmissions are better for mpg. (2) One can expect a manual transmission car get 5.2 mpg better miles per gallon than an automatic with the same horsepower within a 95% confidence interval of (3.4428087, 7.1113619).

Appendix

Exploratory Analysis: Pairs Plot

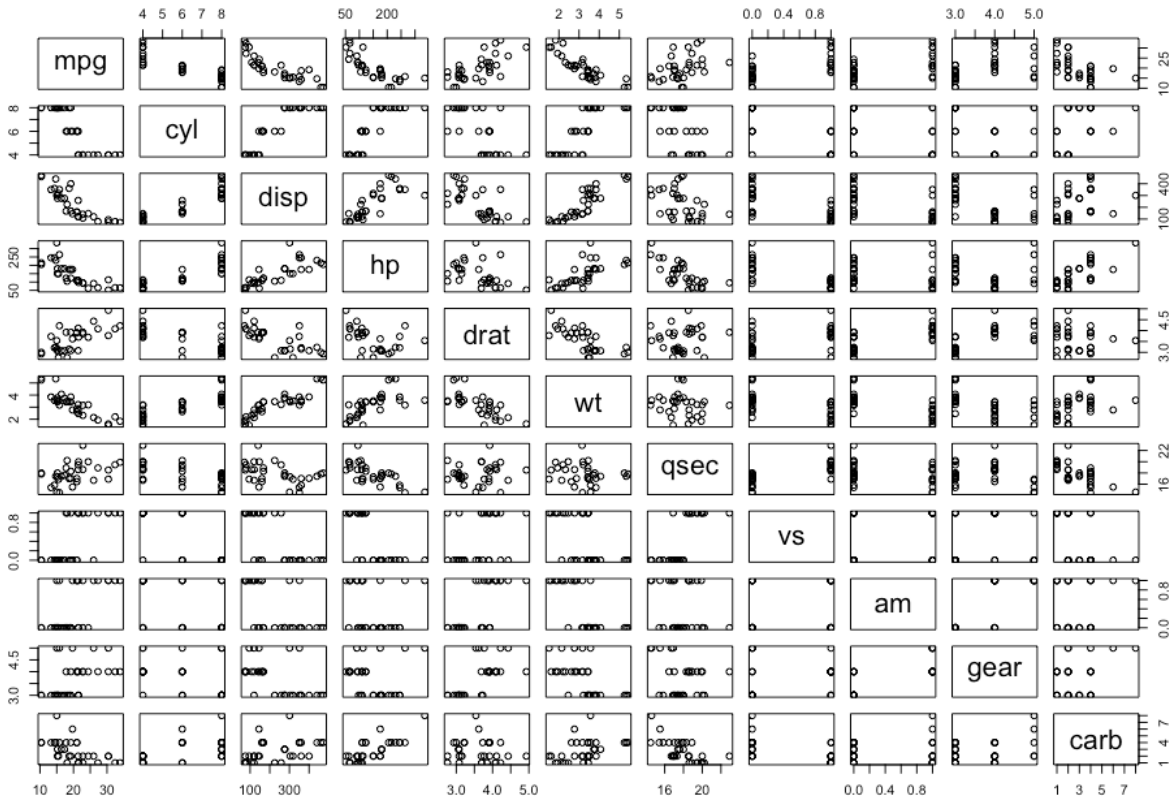

Figure 1. Pairwise relationships between all variables

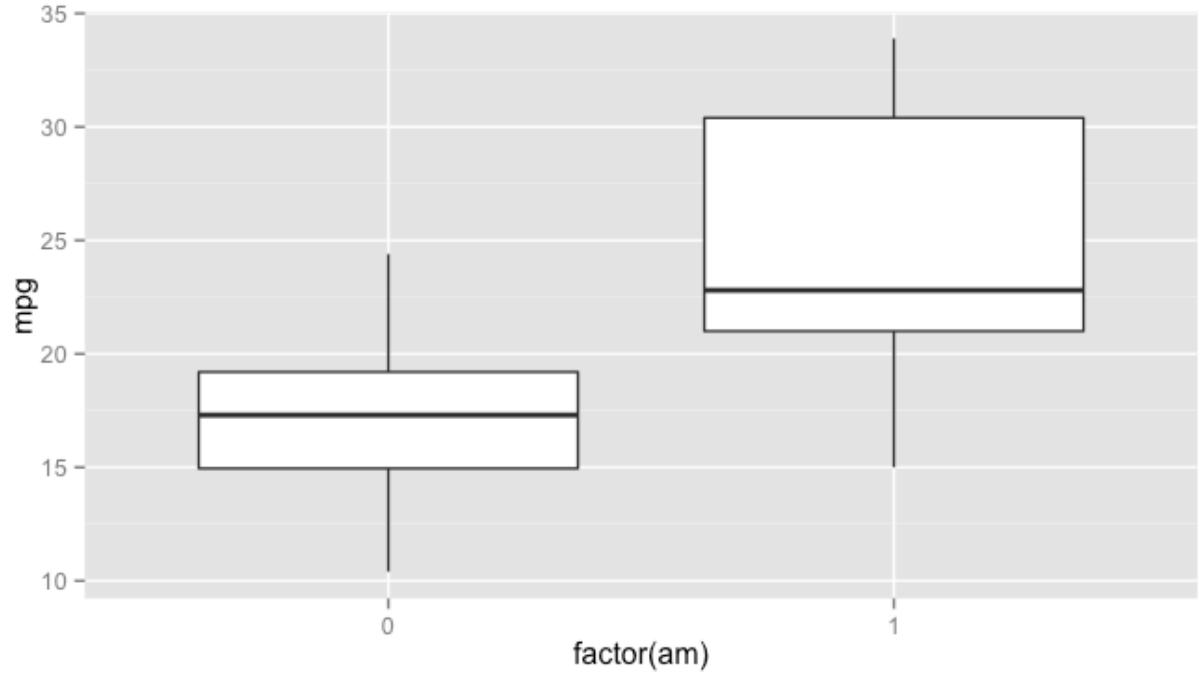

Figure 2. Relationship between transmission type and mpg

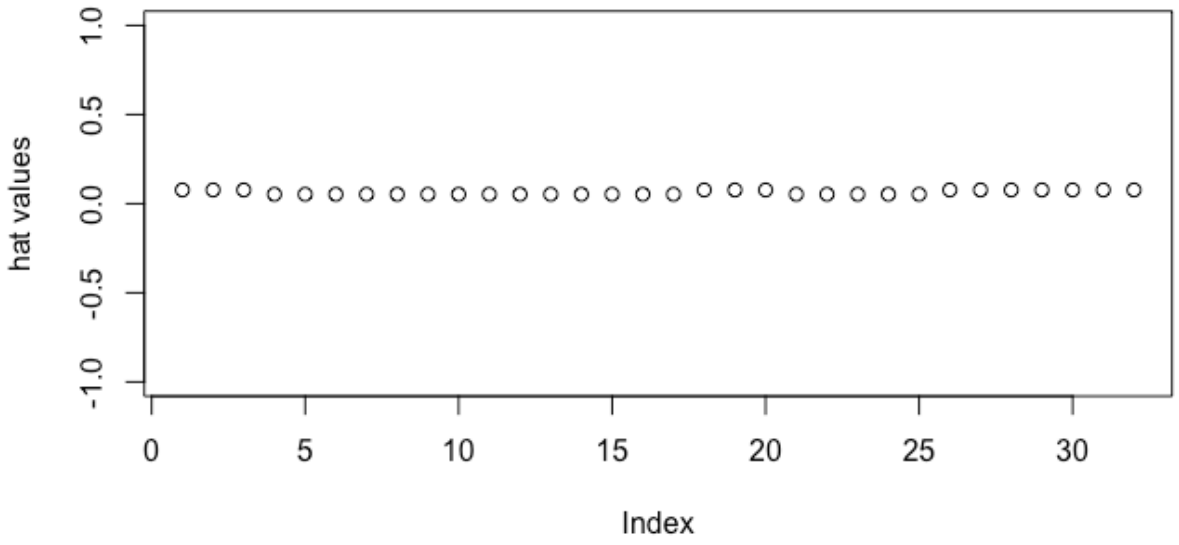

Figure 3: Effects of outliers on base regression model

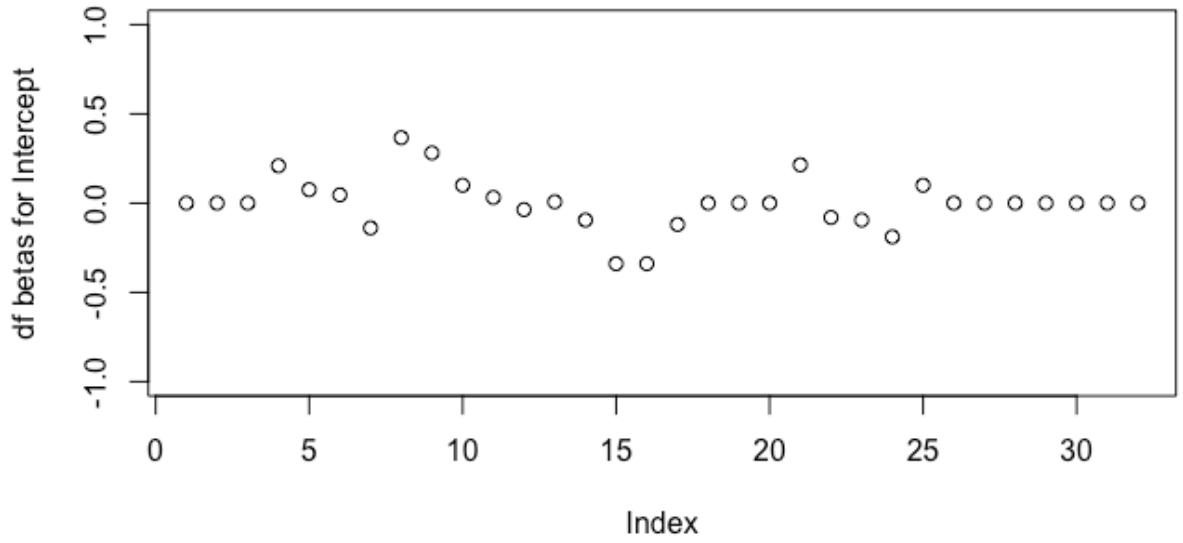

Figure 4: Effects of outliers on base regression model

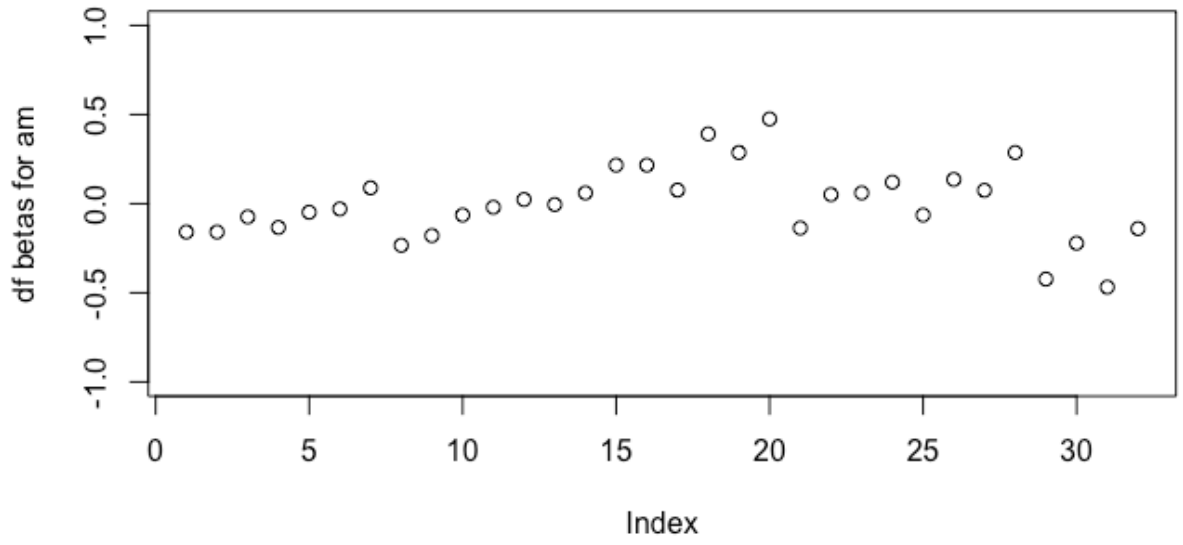

Figure 5: Effects of outliers on base regression model

Model B: Adding horsepower

	Estimate	Std. Error	t value	Pr(>|t|)
am	5.2771	1.0795	4.888	0
hp	-0.0589	0.0079	-7.495	0
(Intercept)	26.5849	1.4251	18.655	0

Table 2. Model B--mpg as predicted by transmission (am) and horsepower (hp)

Observations	Residual Std. Error	R^2	Adjusted R^2
32	2.909	0.782	0.767

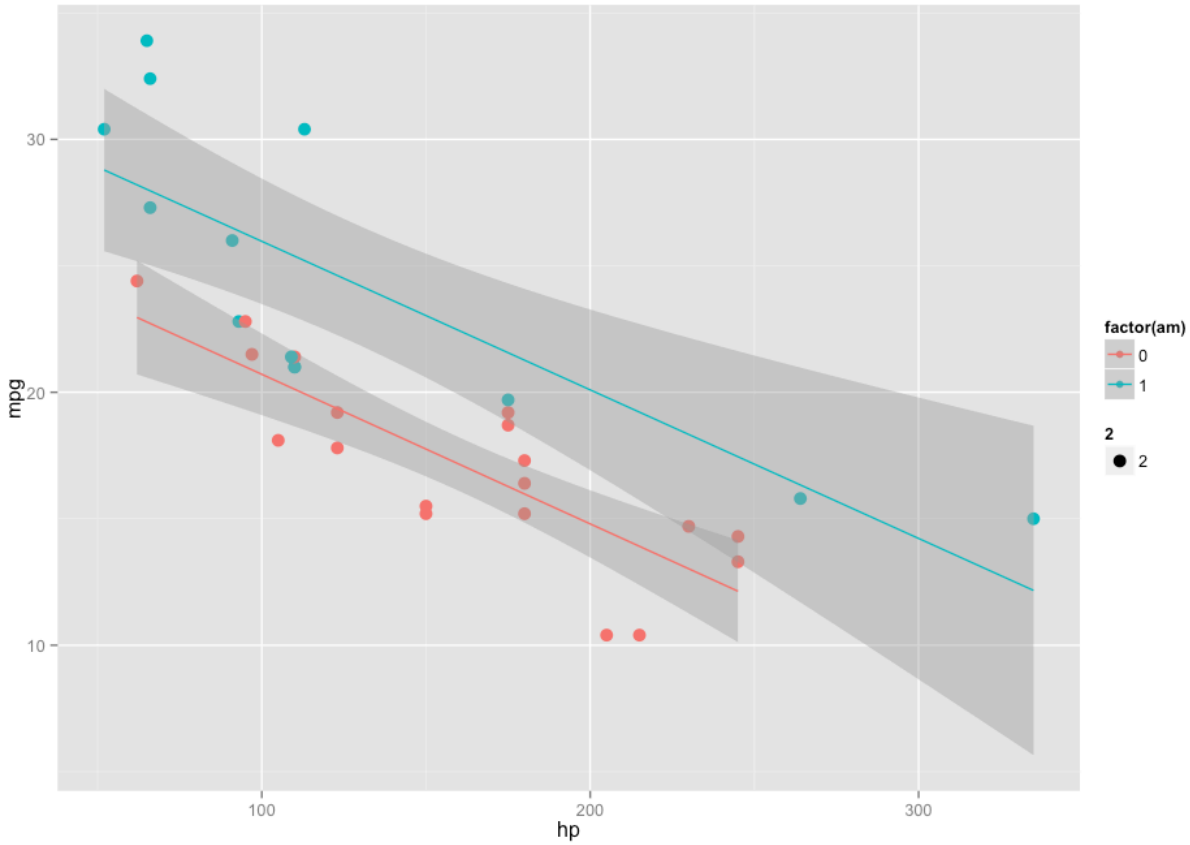

Figure 6. Scatter plot of am and hp on mpg

Table 3. Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
30	720.9	NA	NA	NA	NA
29	245.4	1	475.5	56.18	0

Model B: Adding weight

	Estimate	Std. Error	t value	Pr(>|t|)
am	-0.0236	1.5456	-0.0153	0.9879
wt	-5.3528	0.7882	-6.7908	0.0000
(Intercept)	37.3216	3.0546	12.2180	0.0000

Fitting linear model: mpg ~ am + wt

Observations	Residual Std. Error	R^2	Adjusted R^2
32	3.098	0.7528	0.7358

Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
30	720.9	NA	NA	NA	NA
29	278.3	1	442.6	46.12	0

Model C: Adding cylinder

	Estimate	Std. Error	t value	Pr(>|t|)
am	2.560	1.298	1.973	0.0585
factor(cyl)6	-6.156	1.536	-4.009	0.0004
factor(cyl)8	-10.068	1.452	-6.933	0.0000
(Intercept)	24.802	1.323	18.752	0.0000

Fitting linear model: mpg ~ am + factor(cyl)

Observations	Residual Std. Error	R^2	Adjusted R^2
32	3.074	0.7651	0.7399

Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
30	720.9	NA	NA	NA	NA
28	264.5	2	456.4	24.16	0

Renders transmission type not significant. Not helpful.

Model E: Adding transmission and horsepower interaction term

	Estimate	Std. Error	t value	Pr(>|t|)
am	5.2177	2.6651	1.9578	0.0603
hp	-0.0591	0.0129	-4.5684	0.0001
am:hp	0.0004	0.0165	0.0245	0.9806
(Intercept)	26.6248	2.1829	12.1968	0.0000

*Fitting linear model: mpg ~ am * hp*

Observations	Residual Std. Error	R^2	Adjusted R^2
32	2.961	0.782	0.7587

Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
30	720.9	NA	NA	NA	NA
28	245.4	2	475.5	27.12	0

Not an improvement. R^2 goes from 0.76511 to 0.78204, which is basically unchanged. The interaction term has p-value of 0.98 (not significant), and the p-value for transmission type is 0.06 (>0.05), which is not significant at the 95% level we would like to see. Revert to model without interaction term.