Motor Trend Fuel Consumption Regression Analysis

Connor Gaspar – March 23, 2017

Executive Summary

The following analysis utilizes data collected by US Motor Trend magazine published in 1974. The dataset compares different cars (rows) across different objective measures (columns). In particular, this analysis is concerned with regressing fuel economy/efficiency as defined by miles per gallon. Specifically, the following questions (and more) will be answered through statistical inference and regression analysis:

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions

To this end, exploratory data analysis, hypothesis testing, regression analysis, model fit analysis were performed to make reasoned inferences from this data. The results were that manual transmissions offered an increased fuel efficiency of 2.94mpg relative to automatic transmissions.

Closer examination the regression equation led to a more effective model including acceleration and weight as regressors. Using this methodology the coefficient of determination increased from `0.36` to `0.85`.

Data Overview

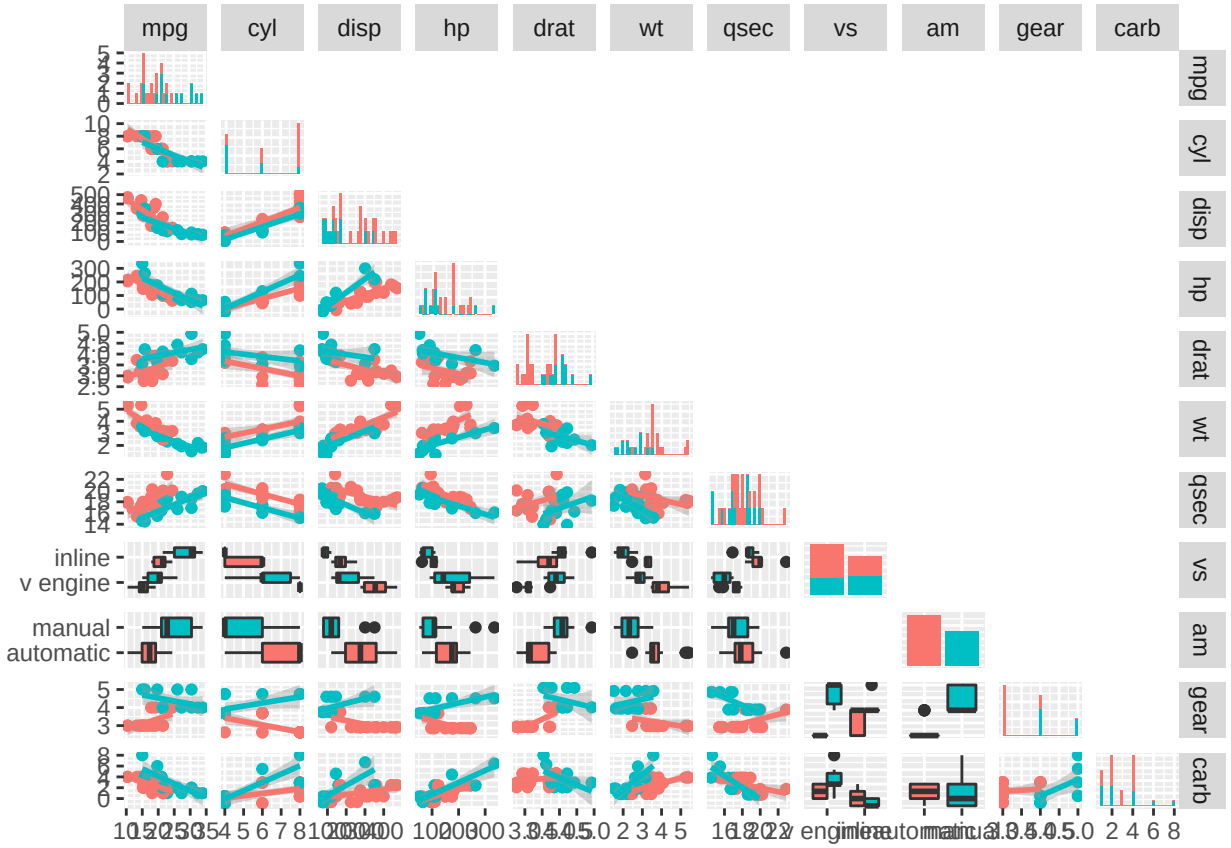
Below is the first five rows of the `mtcars` data set. The response variable is `mpg` and the critical regressor is `am` (transmission type)

N.B. The variables `am` and `vs` have been modified from continuous (numeric) to categorical (factor) to reflect their binary nature. For `am`, labels have been applied to enhance interpretability of the factor levels.

Table 1: Overview of Data Set

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	v engine	manual	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	v engine	manual	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	inline	manual	4	1

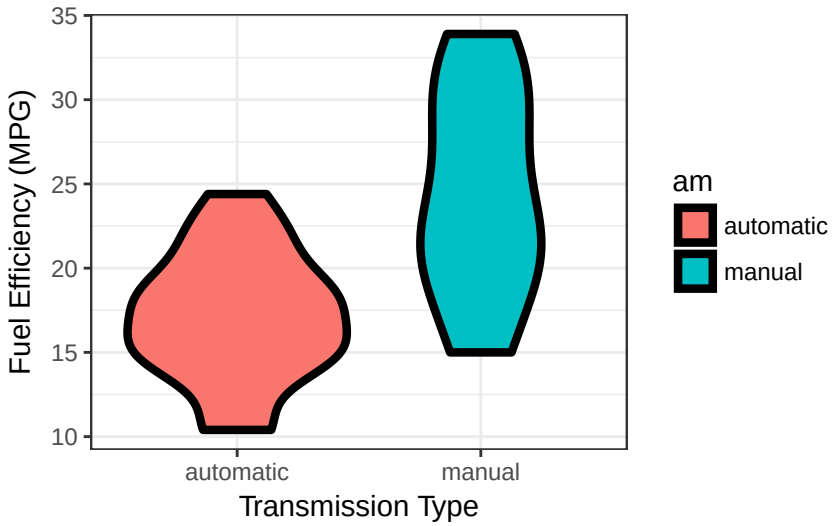
Data Exploration



Looking at the first column of the pairwise plot, clear differences between fuel consumption by automatic and manual transmission can be observed. Scatterplot points have been coloured by the level of `am` to detect any heterogeneity. Briefly examing `mpg` as a function of colour demonstrates clear heterogeneity, thereby showing preliminary evidence for an effect of `am`.

To elucidate this effect, a violin plot of fuel economy by transmission type was constructed.

Visualization of Hypothesized Effect



From this violin plot a clear visual pattern arises that cars with manual transmissions have a higher fuel efficiency than cars with automatic transmissions. To test this hypothesis, In order to statistically test this hypothesis, a test will be utilized to detect whether differences are likely to arise given that no true differences exist between the groups. Given that population parameters are unknown, the t-distribution will be used.

Given the observed differences, a directional hypothesis is formulated that automatic cars are less fuel efficient than manual cars. This can be expressed as:

H0 = HD = 0;

H1 = HD < 0

With the directional hypothesis, a one-tailed t-test was conducted to increase the post-hoc power of the test from `0.91` to `0.96`.

Hypothesis Testing

Table 2: Welch's t-test Output

df	test.statistic	p.value	CI.lower	CI.upper	AT.mean	MT.mean	power
18.332	-3.767	0.001	-Inf	-3.913	17.147	24.392	0.955

The Welch's t-test demonstrates that there's a statistically significant difference in fuel economy between transmission types, $t(18.3) = 3.77$, $p < .001$, such that cars with manual transmissions are more fuel efficient than cars with automatic transmissions.

With a statistically significant difference established between transmission types, a simple linear regression model is generated with `mpg` as the outcome and `am` as the regressor.

Simple Linear Regression

Table 3: Simple Linear Regression Table

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.147	1.125	15.247	0
ammanual	7.245	1.764	4.106	0

With a categorical regressor, the estimate of the intercept represents the average fuel consumption for cars with automatic transmissions. Therefore, the variable below the intercept, `ammanual`, represents the difference in fuel economy from the intercept term.

Since the estimate for `ammanual` is positive, this means that fuel economy *increases* by `7.245` miles per gallon when considering a manual transmission car rather than an automatic transmission car.

The extent of significance observed in the regression summary indicates that this magnitude of difference between transmissions would be observed *less than* 1 in 1000 times (< .1%) should no difference truly exist in fuel economy between the groups. Given the level of significance attained, we can conclude with high certainty that cars with manual transmission are more fuel efficient than automatic transmissions.

Multivariate Linear Regression

Fitting a model with a singular regressor is certainly informative. However, with the availability of many other pertinent variables, it is best to determine whether models with other and more than one regressors exist that explain more variation in fuel economy. Given that the simple linear regression model accounted for `36`% of variance in `mpg`, it's likely that more variance can be determined through additional variables.

Table 4: All Variable Regression Table

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.303	18.718	0.657	0.518
cyl	-0.111	1.045	-0.107	0.916
disp	0.013	0.018	0.747	0.463
hp	-0.021	0.022	-0.987	0.335
drat	0.787	1.635	0.481	0.635
wt	-3.715	1.894	-1.961	0.063
qsec	0.821	0.731	1.123	0.274
vsinline	0.318	2.105	0.151	0.881
ammanual	2.520	2.057	1.225	0.234
gear	0.655	1.493	0.439	0.665
carb	-0.199	0.829	-0.241	0.812

By including all possible regessors in the equation, the standard error of the equation increases significantly. Therefore, in producing a model, careful thought needs to be employed to determine which variables to include in the model.

Another important consideration is *collinearity*, which is defined as the extent to which two variables in a model are correlated. Although multicollinearity doesn't dimish overall predictive power of a model, it does severely diminishes variable specific predictive power. Thus, in the pursuit of the model most effective in describing differences between transmission type in fuel economy, mutlicollinearity should be avoided.

Using variance inflation factors (VIFs), it is possible to analyze the extent to which variables are collinear in a model.

Table 5: Variance Inflation Factors (All Regessors)

regressor	VIF
cyl	15.374
disp	21.620
hp	9.832
drat	3.375
wt	15.165
qsec	7.528
vs	4.966
am	4.648
gear	5.357
carb	7.909

As a rule of thumb, when a VIF is greater than `10`, mutlicollinearity is high. Additionally, VIFs ranging `5 - 10` are of concern, albeit less detrimental to a model. Therefore careful consideration will be taken in inclusion of these variables in a final regresson model

Stepwise Selection

Bidirectional stepwise regression using all measured variables was conducted to fit the regression model of fuel economy.

The stepwise regression demonstrated that the variables `wt` (weight), `qsec` (acceleration), and `am` (transmission type) were statistically significant contributors to explaining variance in fuel economy. Given these three variables are methodologically sound regressors for fuel economy, they were included within the final model fit.

Implications of the Multivariate Model Fit

Table 6: Summary of the Final Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.618	6.960	1.382	0.178
wt	-3.917	0.711	-5.507	0.000
qsec	1.226	0.289	4.247	0.000
ammanual	2.936	1.411	2.081	0.047

Through generating a new model, the coffefficient of determination, r2, increased to `0.85` .

With the updated model, 85% of the observed variation in fuel economy is explained. This differs from the simple linear regression where only 36% of the variance was explained.

Conclusions

Along with with the change in regression model, the interpretation of differences in fuel economy differences changes too. The updated conclusion to draw from the model is that manual cars are `2.94` MPG more fuel efficient than automatic cars for the observed data set.

This statement is made with greater confidence than the simple linear model estimates, as more variance in fuel economy has been explained through inclusion of other variables. Thus, the interpretation of the influence of transmission types has been adjusted accordingly.

Appendix