Regression Analysis of Transmission Type and MPG

Bill Rowe

May 26, 2016

Executive Summary

The project goal was to determine if there is a quantifiable effect of transmission type (automatic or manual) on fuel efficiency. Data from a Motor Trend dataset was explored, highly correlated predictors removed, data tidied and different models investigated.

There is a relationship between fuel efficiency and transmissiong type, however, that relationship is weak, accounting for only 34% of the variability in fuel efficiency (mpg). Two other models are discussed that have higher adjusted R-squared values. One model finds the number of cylinders and transmission type account for . The third model involves more variables (number of cylinders, rear axle ratio, time to cover 1/4 mile, engine type, transmission type and number of forward gears.). That model accounted for

All models were evaluated against a record set with highly correlated variables removed.

Data Exploration

The dataset, mtcars (Motor Trend Cars), is from a 1974 Motor Trend US issue. Fuel efficiency and 10 other vehicle attributes were collected for 32 cars in 1973-1974. Each line of data in the mtcars dataframe represents one model of a car with each column representing one attribute of that car as shown in Table A-1 in the Appendix.

19 of the 32 records are foreign cars (1 Ferrari, 2 Fiats, 1 Honda, 1 Lotus, 1 Maserati, 2 Mazdas, 7 Mercedes, 1 Porsche, 2 Toyotas and 1 Volvo). Figure 1 shows scatter plots of MPG data versus each of the other variables. There is a fair amount of curvative in the possible response of MPG to different car attributes. The most fuel efficient car is the Toyota Corolla at 33.9 mpg. The least fuel efficient cars are the Cadillac Fleetwood and Lincoln Continental at 10.4 mpg.

Tidying Data

Since the question is about the effect of transmission type on fuel efficiency, the transmission type variable (AM) was recoded as 0 = Automatic and 1 = Manual. The variables Transmission Type (am), Number of Cylinders (cyl), Engine Type(vs) and Number of Forward Gears (gear), while numeric, are categorical and were converted to factors. All other variables were not altered. Variable names, while not memorable, were left as abbreviations.

Highly correlated predictors lead to regression coefficients that depend on other included predictors, standard errors increase which leads to wider confidence interval and less precise slope estimates and marginal model improvement due to an added variable depends on other variables in the model. The findCorrelation function in the caret package was used to evaluate a correlation matrix (Table A-2) of the numerical variables. Gross HorsePower (hp), Displacement (disp), Weight (wt) and Number of Carburetors (carb). The reduced data frame includes the following data elements: model, Miles Per Gallon (mpg), Number of Cylinders (cyl), Rear Axle Ratio (drat), Quarter Mile Time (qsec), Engine Type (vs), Transmission Type (am) and Number of Gears gear).

Analysis

Two preliminary linear models Transmission Type only (MPG ~ AM) and all remaining variables (MPG ~ .) were constructed. The simplest model, MPG ~ AM, has each term as significant but an adjusted R-squared of 0.3385. The more complicated model (MPG ~ .), that includes all remaining variables has an adjuste R-squared of 0.7026 but no term is very significant.

Step wise regresssion (stepAIC from the MASS package) was used to find combinations of more significant terms from the MPG ~ . model. Akaike information criterion (AIC) was used as the criteria for forward, backward and both direction executions of the stepAIC function. Bayesian information criterion (BIC) with log(32) degrees of freedom was used as the criteria for the last step wise regression.

The resulting models were then diagnostically evaluated for multicollinearity (VIF), autocorrelation (Durbin-Watson) ,Shapiro-Wilk normality of residuals, heteroscedasticity, Non-constant Variance Score. For each of the following models, there were no concerns resulting from the evaluations. Eliminating the variables with high correlations before creating models improved the results of the diagnostic tests.

The models tested are shown in the following table.

Table 1: Models evaluated.

Model	Adj. R-Squared	Formula
stepBack	0.7399	mpg ~ cyl + am
stepBoth	0.7399	mpg ~ cyl + am
stepFor	0.7026	mpg ~ cyl + drat + qsec + vs + am + gear
stepBIC	0.7399	mpg ~ cyl + am
fit.am.r	0.3385	mpg ~ am
fit.all.r	0.7026	mpg ~ cyl + drat + qsec + vs + am + gear

The most common model found, and the one with a marginally higher adjust R-squared of 0.7399, was the fuel efficiency as function of the number of cylinders and transmission type (mpg ~ cyl + am).

Figures 11-13 show plots of studentized residuals against a normal distribution. Figure 14-16 are Q-Q plots of studentized residuals for the three models discussed below. Figures 17-19, are spread level plots of the absolute studentized residuals against the fitted Values for the three Models. All three collections indicate the studentized residuals are nearly normal in their distribution.

Results

The original question was whether transmission type by itself was a good predictor of Miles Per Gallon. The simplest model, MPG ~ AM had an adjusted R-squared of 0.3385. That simple model only explains 34% of the variaton in the data. A still simple model of mpg ~ cyl + am had an adjusted R-Squared of 0.7399. That is, it explains 74% of the variation. A third model of mpg ~ cyl + drat + qsec + vs + am + gear explained 70% of the variability.

The following sections compare the three models. The LL and UL columns indicate the upper and lower confidence interval limits for the 95% level. The last column corresponds to the R significance codes:’
0‘***’ =0.001‘**’ =0.01‘*’ =0.05‘.’ =0.1‘’ =1 =
’

Simple Model (mpg ~ am)

The following table shows the regression results for fuel efficiency effects due to transmission type.

Table 2: Simple Model of mpg vs am

	Estimate	Std. Error	LL	UL	t value	Pr(>|t|)	
Intercept)	17.15	1.13	14.85	19.44	15.25	1.13E-15	***
am1	7.25	1.76	3.64	10.85	4.11	0.000285	***

The coefficient for manual transmission (am1), shows that fuel efficiency is 7.24 miles per gallon higher than cars with automatic transmissions (am0). However, this simple model has a low adjusted R-squared value of 0.3385. That is, transmission type differences account for only 34% of the variation in fuel efficiency. For comparison, the most fuel efficient car in, the Toyota Corolla, is a manual transmission and the least efficient cars, the Cadillac Fleetwood and Lincoln Continental, are automatic transmission.

Highest Adjusted R-Squared Model (mpg ~ cyl + am)

The following table shows the regression results for fuel efficiency effects due to the number of cylinders and transmission type.

Table 3: Model with two terms (mpg ~ cyl + am)

	Estimate	Std. Error	LL	UL	t value	Pr(>|t|)	
(Intercept)	24.80	1.32	22.09	27.51	18.75	< 2.00E-16	***
cyl6	-6.16	1.54	-9.30	-3.01	-4.01	0.000411	***
cyl8	-10.07	1.45	-13.04	-7.09	-6.93	1.55E-07	***
am1	2.56	1.30	-0.10	5.22	1.97	0.058457	.

Three variations of step wise regression found this model with an adjusted R-Squared of 0.7399. What does not show in the model is the cyl4 (Number of cylinders = 4) variable. Moving from 4 to 6 cylinders drops fuel efficiency by 6.16 miles per gallon and moving to 8 cylinders drops fuel efficiency by 10.07 miles per gallon while using a manual transmission (am1) increases fule efficiency by 2.56 miles per gallon. Transmission type as a variable has low significance. The Toyota Corolla has 4 cylinders and a manual transmission and the Cadillac Fleetwood and Lincoln Continental both have 8 cylinders and automatic transmissions.

Second Highest Adjusted R-Squared Model (mpg ~ cyl + drat + qsec + vs + am + gear)

This model is included in the results since it was the starting model for step wise regression and had an adjusted R-squared of 0.7026. Unlike the previous model, no variable was considered significant. Increasing the number of cylinders and forward gears descreases fuel efficiency (), drat increases MPG by 1.4201, qsec just a little bit 0.3208 while engine type (straight vs VB) increases mileage by 1.58 MPG and manual transmission (am1) by 4.65 MPG.

Table 4: Complex Model (mpg ~ cyl + drat + qsec + vs + am + gear)

	Estimate	Std. Error	LL	UL	t value	Pr(>|t|)	
(Intercept)	11.99	21.35	-32.17	56.15	0.56	0.5797	
cyl6	-4.38	2.59	-9.74	0.98	-1.69	0.1044	
cyl8	-7.18	4.08	-15.63	1.27	-1.76	0.0921	.
drat	1.42	2.41	-3.56	6.40	0.59	0.5609	
qsec	0.32	0.84	-1.42	2.06	0.38	0.7059	
vs1	1.58	2.52	-3.62	6.79	0.63	0.5351	
amManual	4.65	2.62	-0.77	10.07	1.77	0.0893	.
gear4	-2.25	2.87	-8.19	3.70	-0.78	0.4424	
gear5	-2.41	3.03	-8.67	3.85	-0.80	0.4336	

While cyl8 and amManual have the lowest significance, there are some interesting points here in this model. Rear axle ratios (drat) are associated with pickup (qsec) and engine revolutions per minute (RPMs). A lower drat means quicker pickup but higher RPMs which means lower fuel efficiency. Conversely, a higher drat means slower pickup, but lower engine RPMs and better fuel efficiency. This shows in the data. The Cadillac Fleetwood and Lincoln Continental have drats of 2.93 and 3.00 with qsec values of 17.98 and 17.82. The more fuel efficient Toyota Corolla has a drat of 4.22 and a qsec of 19.9. Considering, the relationship between drat and qsec, it is surprising the variables were not consider correlated.

Four cylinder engines are often straight or inline (with the exception of the 4 cylinder Porsche 914-2 which has a Vee engine) while all the 8 cylinder engines are Vee engines. Automatic transmission 6 cylinder engines are Vee engines while the manual transmission 6 cylinder have straignt or inline. While this model is not as simple as the MPG ~ AM or MPG ~ CYL + AM models, it includes many components directly related to fuel efficiency.

Conclusions

In this data set vehicles achieve a fuel effiency of 7.2 miles per gallon more than automatic vehicles. Transmission type, by itself, is a a poor predictor of fuel efficiency. The best predictors were a combination of the number of cylinders and transmission type. A third model with more terms was not as explanatory as the cylinders and transmission type model but has other terms that do explain changes in fuel efficiency.

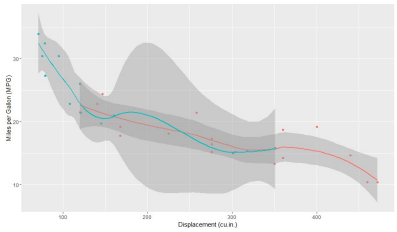
Appendix

Table A-1: MTCars Data Elements

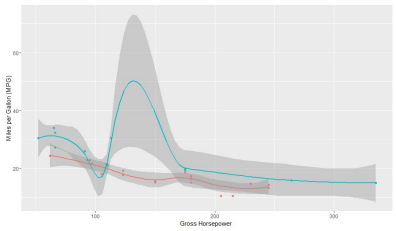
Variable	Data Type	Description	Levels
Model	character	Brand and model	%nbsp;
MPG	number	Miles/(US) gallon	Continuous
cyl	number	Number of cylinders	4, 6, 8
disp	number	Displacement (cu.in.)	Continuous
hp	number	Gross horsepower	52,62,65,66,91,93,95,97,105,109,110,113,123,150,175,180,205,215,230,245,264,335
drat	number	Rear axle ratio	Continuous
wt	number	Weight (lb/1000)	Continuous
qsec	number	1/4 mile time	Continuous
vs	number	Engine type	0 = vee engine, 1 = straight or inline engine
am	number	Transmission type	0 = automatic, 1 = manual
gear	number	Number of forward gears	3, 4, 5
carb	number	Number of carburetors	1, 2, 3, 4, 6, 8

Figures A1-A10: MPG vs Attributes. Automatic Transmission data are colored pink, while manual transmission data are colored teal.

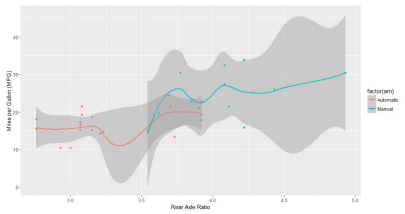
MPG vs Displacement



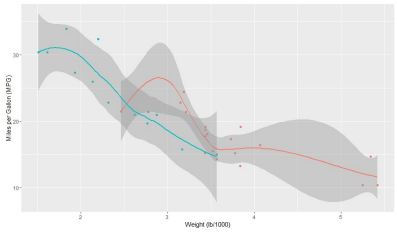
MPG vs Gross Horsepower



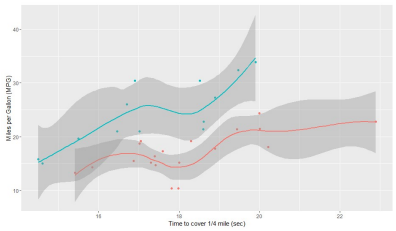
MPG vs Rear Axle Ratio



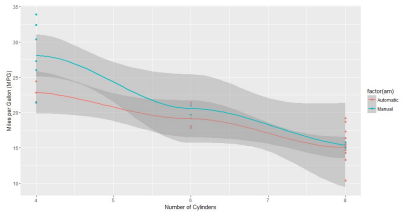
MPG vs Weight



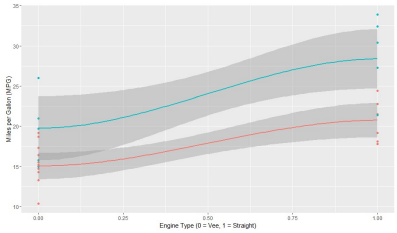
MPG vs Time to 1/4 Mile



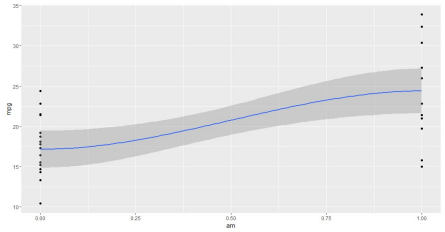
MPG vs Number of Cylinders



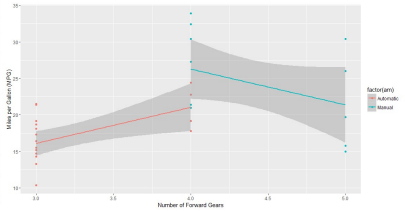
MPG vs Engine Type



MPG vs Transmission Type



MPG vs Number of Forward Gears



MPG vs Number of Carburetors

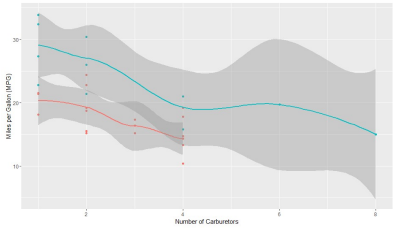


Table A-2: Correlation Matrix for MTCars Variables. Variables with values > 0.6 were removed.

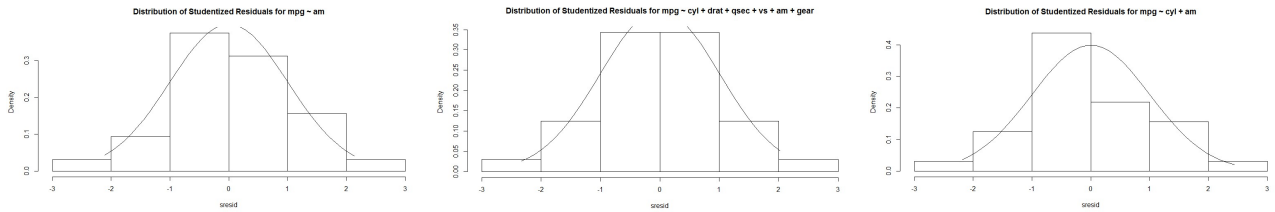
	disp	**hp**	**drat**	**wt**	**qsec**	**carb**
disp	1.00					
hp	0.79	1.00				
drat	-0.71	-0.45	1.00			
wt	0.89	0.66	-0.71	1.00		
qsec	-0.43	-0.71	0.09	-0.17	1.00	
carb	0.39	0.75	-0.09	0.43	-0.66	1.00

Figures A11-A13: Distribution of Studentized Residuals for Three Models.

mpg ~ am

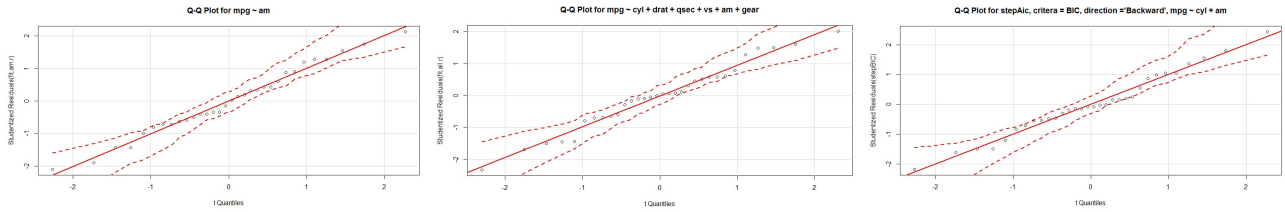
mpg ~ cyl + drat + qsec + vs + am + gear

mpg ~ cyl + am



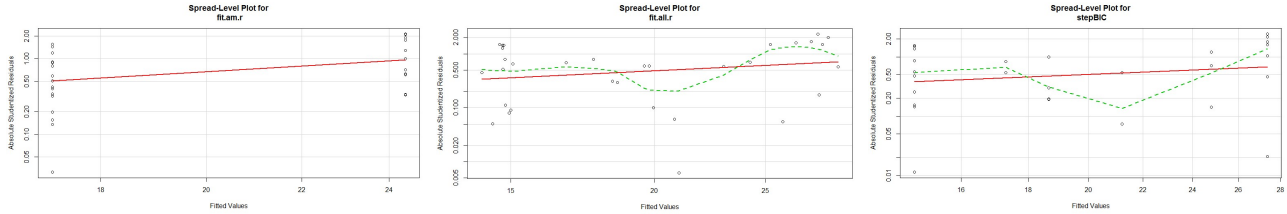
Figures A14-A16: Q-Q Plots of Studentized Residuals for Three Models.

mpg ~ am	mpg ~ cyl + drat + qsec + vs + am + gear	mpg ~ cyl + am



Figures A17-A19: Spread Level Plots of Absolute Studentized Residuals vs Fitted Values for Three Models.

mpg ~ am	mpg ~ cyl + drat + qsec + vs + am + gear	mpg ~ cyl + am



System Information

The analysis was run with the following software and hardware.

```
sessionInfo()
```

```
## R version 3.2.5 (2016-04-14)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 10586)
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] magrittr_1.5    formatR_1.3     tools_3.2.5     htmltools_0.3.5
##  [5] yaml_2.1.13     Rcpp_0.12.4     stringi_1.0-1   rmarkdown_0.9.5
##  [9] knitr_1.12.3    stringr_1.0.0   digest_0.6.9    evaluate_0.8.3
```

The source code for this document and the analysis is stored in GitHub at https://github.com/wer61537/Regression_Models_Course_Project (https://github.com/wer61537/Regression_Models_Course_Project). The R code is at https://github.com/wer61537/Regression_Models_Course_Project/blob/master/regr_transmission_mpg.R (https://github.com/wer61537/Regression_Models_Course_Project/blob/master/regr_transmission_mpg.R). The Markdown document is at https://github.com/wer61537/Regression_Models_Course_Project/blob/master/regr_transmission_mpg.Rmd (https://github.com/wer61537/Regression_Models_Course_Project/blob/master/regr_transmission_mpg.Rmd).