A Study of Gas Mileage Based on Drive Train: Choosing Regression Predictors *(by Jay Gendron)*

Executive Summary

This analysis is presented in the context of someone working for Motor Trend Magazine assigned to a data analysis project. Of interest to the analyst are the variables that help explain the relationship between drive train (automatic or manual) and miles per gallon (MPG). Specifically, two research questions were presented to frame the analysis:

- Is an automatic or manual transmission better for MPG?
- How different is the MPG between automatic and manual transmissions?

The research questions are addressed first by exploratory data analysis to become familiar with the variables and then statistical modeling to develop a regression model using step-wise regression. The Results section interprets the coefficients and provides two findings: a) manual transmissions provide better MPG; and b) with 95% confidence, a manual transmission is estimated to result in a 0.05 to 5.82 MPG increase (statistically significant difference at 0.05%). Further investigation determined the model was well constructed by using residuals and diagnostics.

Data Processing

Data Collection: The data is the `mtcars` set included with the base R package. It was originally found in Motor Trend Magazine (1974) and has 11 variables representing 10 auto design features plus one variable indicating the MPG. There are 32 observations - each representing a different model of automobile from 1973 and 1974.

Exploratory Data Analysis: All aspects of this analysis employ packages from the base R package. A summary() command showed there were no missing data points. One transformation was made to explicitly label the transmission variable from (0,1) to ((0 = automatic, 1 = manual). Another transform converted five variables from numeric to factors: *cyl, vs, am, gear, and carb*.

A pairs() plot showed many interesting correlations (see Appendix). The variable *mpg* was highly correlated with *wt (0.87)*, engine size *cyl (0.85), disp (0.85)*, and *hp (0.78)*. These may be valuable in a model. Additionally, low correlations with the designated predictor variable *am* included *vs (0.17)*, *qsec (0.23)*, and *hp (0.24)*. Low correlation is interesting with regard to predictors because highly correlated variables result in variance inflation. It should be noted that *hp, cyl, and disp* are highly correlated with one another; however *am* has a low correlation with *hp and qsec*. This makes is favorable that these latter variable could appear in model selection.

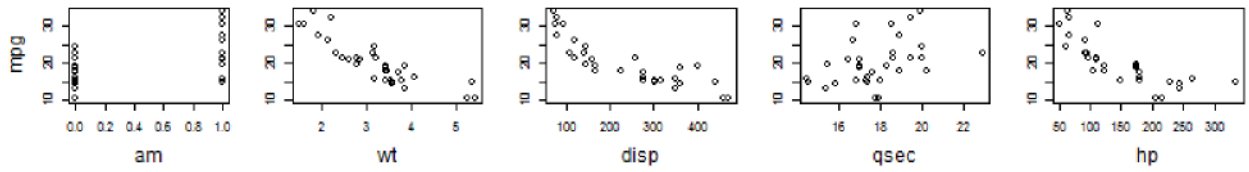


Figure 1: Looking for Model Elements. Consideration of variables based on bi-variate correlations (sample of some *pairs()* plots in R.

Statistical Modeling: Model selection was accomplished using the backwards elimination method based on adjusted R-squared as described in Diez, Barr, and Rundel (2012, p. 361) http://www.openintro.org/download.php?file=os2_08&referrer=/stat/textbook/textbook_os2_chapters.php. The method begins with a full model composed of all variables and eliminates one variable at a time then repeating the reduction with the revised model until there is no further increase in adjusted R-squared. This required 41 total model evaluations. Table 1 shows the intermediate results of the backwards elimination method. Notice the continued enhancement of the model correlation and notice how much more the final model explained error (83.75%) versus the base model as defined by the research question (33.85%)

Table 1: Model Selection. Intermediate steps showing elimination of variables based on increasing adjusted R-squared.

Model Content	/ Adjusted R-squared	> Resulting Model
Full Model	/ 0.8066	> lm(mpg~cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb)
remove (cyl)	/ 0.8153	> lm(mpg~disp+hp+drat+wt+qsec+vs+am+gear+carb)
remove (vs)	/ 0.8230	> lm(mpg~disp+hp+drat+wt+qsec+am+gear+carb)
remove (carb)	/ 0.8296	> lm(mpg~disp+hp+drat+wt+qsec+am+gear)
remove (gear)	/ 0.8347	> lm(mpg~disp+hp+drat+wt+qsec+am)
remove (drat)	/ 0.8375	> lm(mpg~disp+hp+wt+qsec+am)
Base Model	/ 0.3385	> lm(mpg~am)

It is interesting to note how well the pairs plot built during exploratory data analysis pointed out the variables likely to emerge in the step-wise, backwards elimination method. The proposed model of five variables was run to investigate p-values. That showed two variables(*disp and hp*) were not significant with p-values of 0.300 and 0.156, respectively. They were removed and interestingly, the adjusted R-squared decreased by only a third of a percent to 0.8336. Having established a final model of three predictor variables, this analysis moves into studying results and interpreting them to address the research questions.

Table 2: Final Model. Refinements to backward elimination method based on resulting p-Values.

Model Content / Adjusted R-squared	> Resulting Model

Proposed Model	/ 0.8375	> lm(mpg~disp+hp+wt+qsec+am)
remove (disp)	/ 0.8368	> lm(mpg~hp+wt+qsec+am)
Final Model	/ 0.8336	> lm(mpg~wt+qsec+am)

Results

Here is the underlying information of the final model – most importantly the coefficients (slope) of each of the predictor variables.

```
##                   Estimate Std. Error t value  Pr(>|t|)
## (Intercept)          9.618     6.9596   1.382 1.779e-01
## wt                  -3.917     0.7112  -5.507 6.953e-06
## qsec                 1.226     0.2887   4.247 2.162e-04
## factor(am)manual     2.936     1.4109   2.081 4.672e-02
```

The estimated expected change in mpg is discerned by interpreting the coefficients:

- Estimate a **3.9 decrease in MPG** per 1,000 pound *increase* in weight
- Estimate a **1.2 increase in MPG** per 1 second *increase* in quarter mile run time
- Estimate a **2.9 increase in MPG** if the vehicle has a *manual* transmission (versus the reference of an automatic transmission)

Answering the research questions

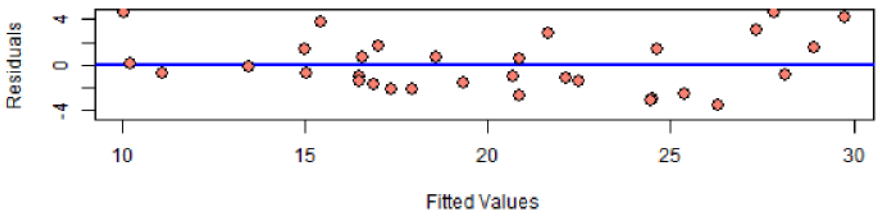
Is an automatic or manual transmission better for MPG? Model coefficients answer this first question. With a significance level of under 0.05, the coefficient of slope for the factor *am* is 2.936. The label indicates this is when the variable equals “manual”. The positive value indicates that as compared to the reference level (“automatic”), a manual transmission was nearly three times better for MPG.

Howdifferent is the MPG between automatic manual transmission? One can determine a 95% confidence interval using inferential statistics and the coefficient and standard error taken from the linear model for the variable *am*. It takes the form {*am* coeff} +/- t(alpha=0.025)*(Std error). Interpretation yields that with 95% confidence (while holding all other variables constant) we estimate that a manual transmission results in a 0.05 to 5.82 increase in MPG. This confidence interval does not include zero; therefore, one may reject a null hypothesis that the mean difference on MPG between manual and automatic transmissions is zero. Here is the formulation of the confidence interval coded in R.

```
summary(final)$coefficients[4,1] + c(-1, 1) * qt(.975, df = final$df) *
summary(final)$coefficients[4,2]
```

Assessing the model quality

The final model has good properties as seen in the residual plot below. It shows the model has good fit and displays constant variance (no heteroskedasticity), no missing terms, and no patterns that would occur from non-linear fit or from time-based effects.



Additionally, the influence measures of the model were investigated for leverage (hat values) and to determine the stability of the model relative to removal of observations (dfbetas). The R commands *hatvalues()* and *dfdbetas()* show “hatvalues” within a reasonable range and the “dfbetas” also display low values, indicating low likelihood of any observation exerting possible influence.

```
summary(hatvalues(final)) #assess model for observations with high leverage
summary(dfbetas(final)[,2]);summary(dfbetas(final)[,3]); summary(dfbetas(final)[,4])
```

Influence Measure	/ Range	Median
hatvalues	/ [0.05-0.3]	0.10
dfbetas(wt)	/ [-0.43-1.09]	-0.02
dfbetas(qsec)	/ [-0.70-0.50]	-0.00
dfbetas(am)	/ [-0.42-0.56]	-0.04

The results presented in this analysis were validated through various diagnostics to ensure the confidence intervals are reliable. Overall, the degree of uncertainty in the conclusions is low. This stems from the holistic combination of a) a well-fit model using multiple strategies; b) p-values significant to 0.05 or better; and c) good results from diagnostic assessments. The variability explained by the model is the ratio of variability among the variables (MSG) to the variability within the variables (MSE) = MSG/MSE = 956.8/6.05 = 158.4 (F value) resulting in an overall p-value of 0.0001. Additionally, the variability explained by the model (SSG) as compared to total error (SST) = SSG/SST = 956.8/1126.1 = 85%. All this stems from the building of a good model having an adjusted R-squared value of more than 83%.

Appendix

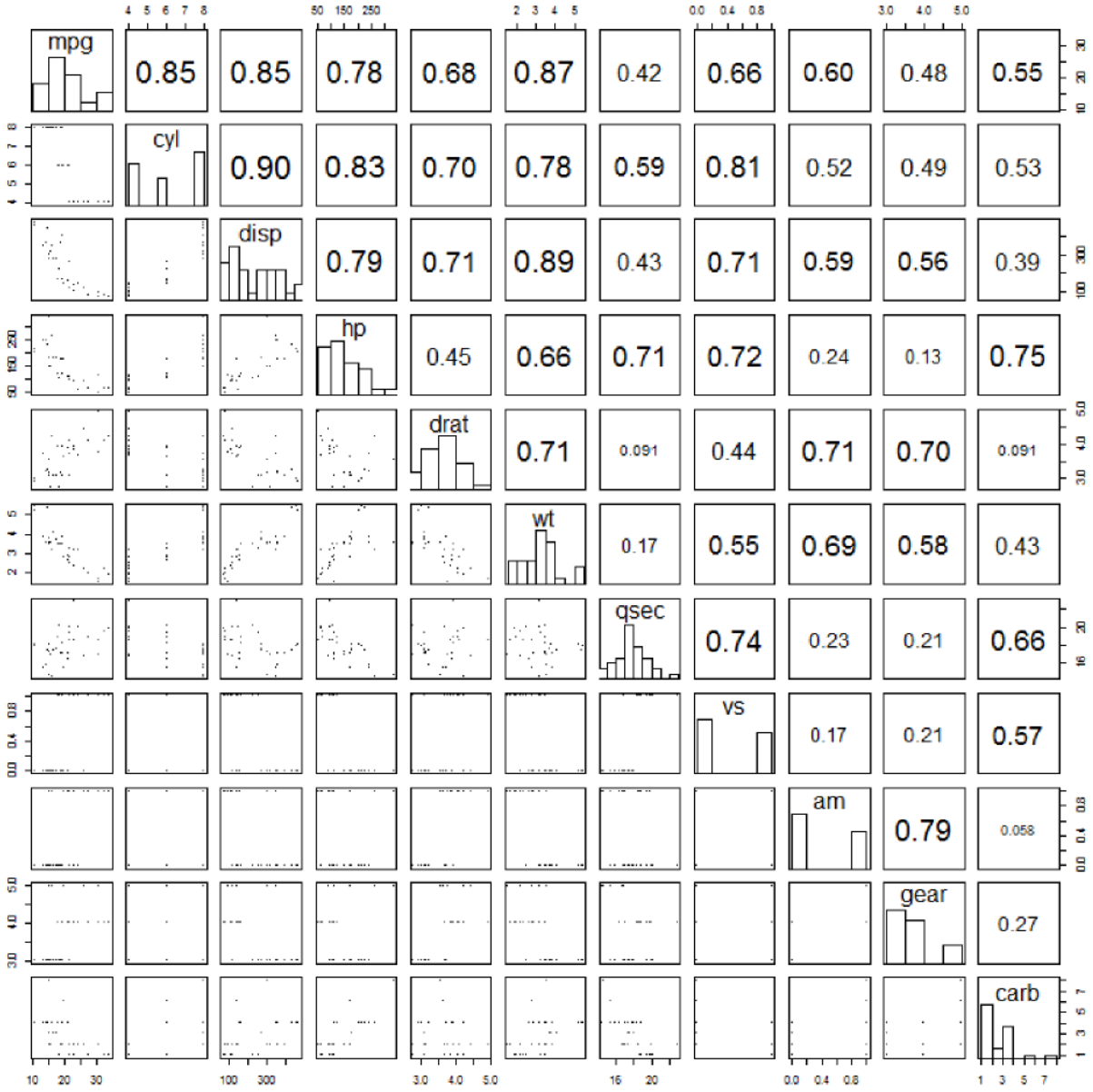


Figure A-1: Pairs Plot for *mtcars* Dataset. The purpose of the pairs plot is to show the correlation of all variables in a pairwise manner. The purpose of conducting this during exploratory data analysis is to identify those variables which are highly correlated with one another. Highly correlated variables can be reduced into other combination – thereby reducing variance inflation. This pairs plot shows the variable *mpg* was highly correlated with *wt (0.87)*, engine size *cyl (0.85); disp (0.85)*, and *hp (0.78)*. Additionally, low correlations are useful to identify high correlation variables that are not correlated well with other predictors. For example, *hp, cyl, and disp* are highly correlated with one another; however *am* has a low correlation with *hp and qsec*. This makes is favorable that these latter variable could appear in model selection.