Regression Model Course Project

Jim White

December 23, 2015

Executive Summary

In 1974 Motor Trend, a magazine about the automobile industry was interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They were particularly 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

An analysis was completed using regression models and exploratory data analysis. Based on this analysis, the affect of predictor variables other tham “am” on the response “mpg” make it difficult to answer the questions directly. During exploratory analysis (see Figure 1 in appendix), apparent differences do appear to occur. Further analysis, using all of the predictor variables in the data set and linear modeling techniques, proved isolating the “am” effect on “mpg” difficult. A multivariate model was developed using the predictor variables am, cyl, wt, and hp.

Exploratory Analysis

The data set contains 32 observations with 11 variables. A brief statistical description of the data set variables can be found in Table 1 Descriptive Statistics in the appendix. Also, from this brief examination, some variables appear to be **categorical** (cyl, vs, am, gear, carb) and some are **continuous** (mpg, disp, hp, drat, wt, qsec). For further exploratory analysis, the categorical variables werer changed to factor variables.

Perhaps the first step should be to compare the “mpg”" responses against the “am”" predictor for each of the types of transmission. In Figure 1 Transmission Type and MPG (located in the appendix), a relationship appears to be straightforward between transmission types and mpg results; hypothesis testing was used to determine if the means are different.

$$H_0 : \mu_{auto} = \mu_{manual} \quad H_a : \mu_{auto} < \mu_{manual}$$

The t-test results are:
With a 95% confidence interval, the p-value is 6.910^{-4}, consequently, the null hypothesis can be rejected. The average value for mpg for auto trans = 17.15 and the average value for mpg for manual trans = 24.39. *A difference of* 7.24 *in favor of the manual transmission*.

While a clear difference exists between the means μ_{manual} and μ_{auto} other variables may be further influencing the “mpg” response. The next step is to examine any relationship that may exist among all of the variables in the dataset.

Variable Relationships Correlation is used to determine if possible relationships appear to exist between any of the variables and “mpg”. Both positive and negative significant results appear for most of the predictor variables and “mpg.”"

```
##   mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##  1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
```

When examining the relationship of “am” to other variables in the data set, the results (below) indicate that “am” has a significant positive or negative correlation with gear, drat, wt, disp, and cyl. This is possible evidence that some of the predictor variables other than “am” influence the “mpg” results - creating concerns about quantifying the differences in auto or manual transmission influences.

```
##   mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##  0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
```

Note: At this point the “qsec” variable (1/4 mile time) is removed from the dataset. An explanation is provided in the appendix in the *Description of Variables* section.

Regression and Residual Analysis

One way to measure to relationship between “am” on “mpg” is to use a simple linear model and calculate R^2 (coefficient of determination). That results follows:

The R-squared result from the single variable regression model = 0.3598. The R^2 value result appears to indicate the transmission type “am”" has only a minor impact on “mpg”" and other variables apparently contribute to the “mpg” results. Thus providing straightforward answers to the two Motor Trend questions will be difficult.

Further evidenced is provided by comparing the coefficients for “am”, 1) first when only “am” is used in the linear model to predict “mpg”, and 2) second the coefficient for “am” when all of the variables are included in the linear model. The results follow:

[1] am coefficient (linear model) with only am and mpg: 7.245

[2] am coefficient (linear model) including all variables [except qsec]: 1.887.

Accounting for other variables should not change the relationship between the predictor (am) and the response (mpg), unless covariance exists between predictor variables. The above results indicate that predicting “mpg”" from transmission type (am) alone may be difficult due to the influence of other predictor variables on the mpg response.

Residuals, Diagnostics, and Variation

Before determining which predictor variables should be included in a model to predict “mpg”, an analysis of the residuals may be helpful. The first step may be to determine model fit or lack thereof. This can be done by examining the charts in Figure 2 Residual Analysis (located in the appendix). The model used in this analysis includes all of the variables except “qsec.”

Description of Figure 2 Residual Analysis (appendix): The *Residuals vs Fitted* plot may indicate some lack of linearity influenced by values on the right end of the chart. The *Scale Location* chart tends to indicate some level of heteroscedasticity in which some parts of the data appear to have different variabilities than others. The *QQ Plot* appears to indicate a normal distributions for the residuals (as should be expected). In the *Residuals vs Leverage* plot, there does not appear to be any points of significant influence or leverage.

Also, when running the **hatvalues** function (from the stats package), there are no points in the data identified as having enough leverage to influence model fit. (Table 2 Results of hatvalues Function are in the appendix.)

Model Fitting

When evaluating which variables to include in building a model, one step is to include an analysis of the variance inflation factor (vif). In this case the standard deviation inflation factor is examined (the sqrt of the vif). If the predictor variables are not related (not correlated), then the values reported should be close to 1.

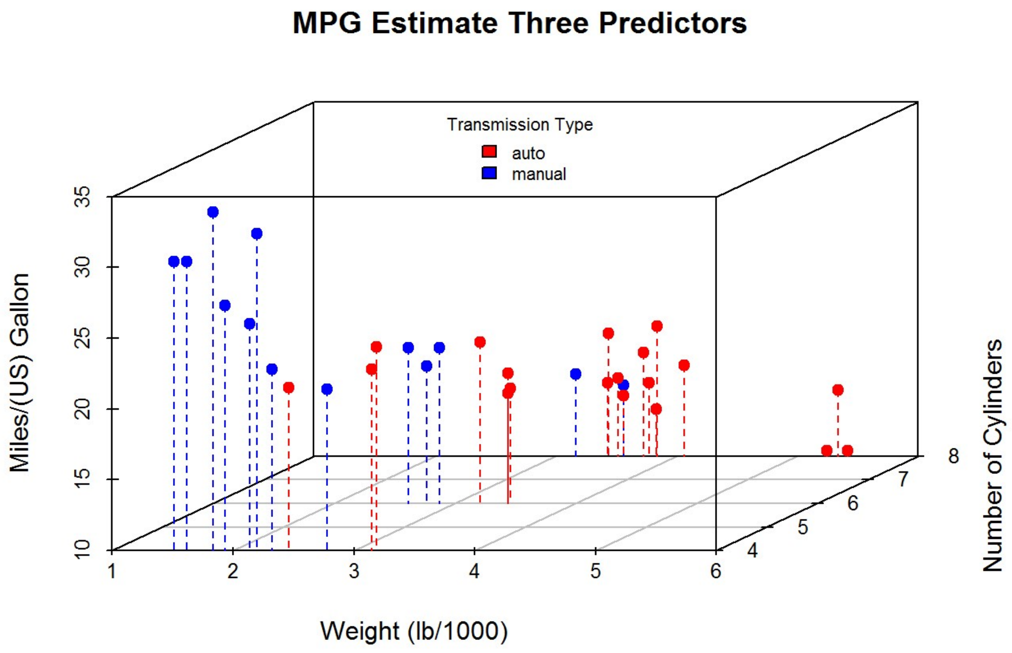
```
##   cyl disp   hp drat   wt   vs   am gear carb
## 3.77 4.46 3.12 1.84 3.36 2.07 2.07 2.31 2.71
```

The results can be interpreted as follows: e.g., the “disp”" variable standard error is four times what it would be if it were not correlated with other variables. Most of the predictor variables results are high indicating significant correlations between predictor variables. This makes model selection difficult.

Model selection for this analysis uses the nested model approach, where one’s interest is in how one predicator variable affects the response variable, and how adding other predictor variables affects the model. This process along with an anova calculation was conducted and the results follow. (Table 3 Results for All Models is located in the appendix.) The results recorded here are based on the selection of the predictor variables that appear to create the best model. Models 2 or 3 (below) may be the best choices.

```
##     Model         Predictors  F_Values     p_values
## 1 Model 1                 am        NA           NA
## 2 Model 2      am + cyl + hp 39.733688 9.036232e-09
## 3 Model 3 am + cyl + wt + hp  8.029469 8.603218e-03
```

Finally, a 3D Model representation of Model 2, just for fun. Apparently, heavier cars tend to have automatic transmissions and get less miles per gallon.



Appendix

Description of Variables: [1] mpg = Miles/(US) gallon, [2] cyl = Number of cylinders, [3] disp = Displacement (cu.in.), [4] hp = Gross horsepower, [5] drat = Rear axle ratio, [6] wt = Weight (1000 lbs), [7] qsec = 1/4 mile time, [8] vs V/S (piston position, v shaped or straight; V=0, S=1), [9] am = Transmission (0 = automatic, 1 = manual), [10] gear = Number of forward gears, [11] carb = Number of carburators.

The “qsec” variable was removed from the data set because the time for running the quarter mile is more likely related to hp, which in turn is likely related to the other predictor variables. While “qsec” and “mpg” may be correlated, the likelihood is that they are both dependent upon the same predictors.

Table 1 Descriptive Statistics

```
## 
## ============================================
## Statistic N   Mean   St. Dev.  Min     Max  
## --------------------------------------------
## mpg       32 20.091   6.027   10.400 33.900 
## cyl       32  6.188   1.786     4       8   
## disp      32 230.722 123.939  71.100 472.000
## hp        32 146.688  68.563    52     335  
## drat      32  3.597   0.535   2.760   4.930 
## wt        32  3.217   0.978   1.513   5.424 
## qsec      32 17.849   1.787   14.500 22.900 
## vs        32  0.438   0.504     0       1   
## am        32  0.406   0.499     0       1   
## gear      32  3.688   0.738     3       5   
## carb      32  2.812   1.615     1       8   
## --------------------------------------------
```

Figure 1 Transmission Type and MPG

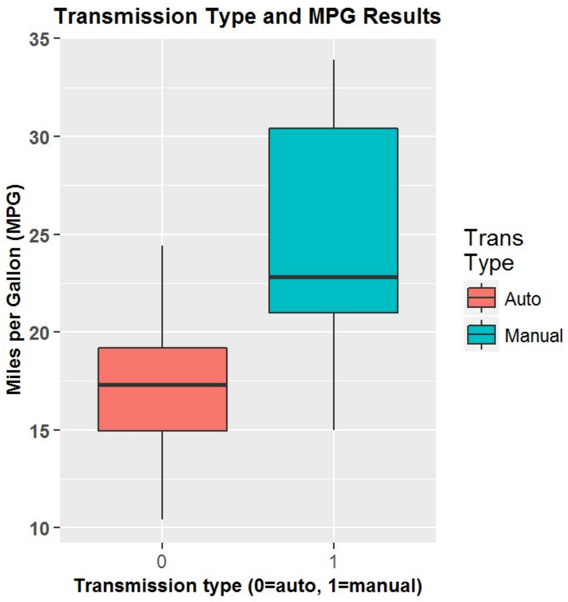


Figure 2 Residual Analysis

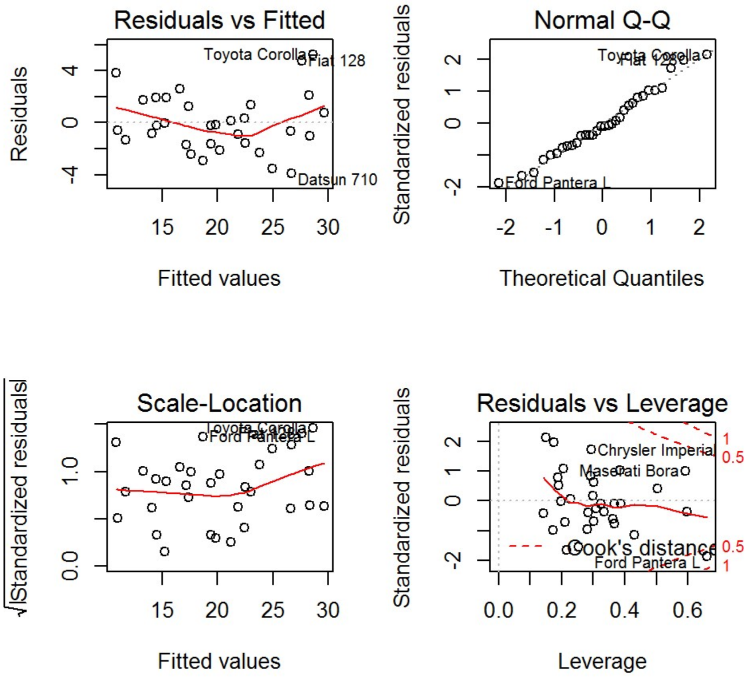


Table 2 Results of hatvalues Function:

```
##           Mazda RX4       Mazda RX4 Wag          Datsun 710      Hornet 4 Drive   Hornet Sportabout 
##           0.3021679           0.2837555           0.2148455           0.2277084           0.1910947 
##             Valiant          Duster 360           Merc 240D            Merc 230            Merc 280 
##           0.2802018           0.3235086           0.3005933           0.2997062           0.3681555 
##           Merc 280C          Merc 450SE          Merc 450SL         Merc 450SLC  Cadillac Fleetwood 
##           0.3681555           0.2918091           0.1870342           0.1969380           0.3633075 
## Lincoln Continental   Chrysler Imperial            Fiat 128         Honda Civic      Toyota Corolla 
##           0.3088452           0.2933027           0.1763924           0.5046273           0.1493008 
##       Toyota Corona    Dodge Challenger         AMC Javelin          Camaro Z28    Pontiac Firebird 
##           0.4302743           0.2098348           0.1727814           0.3334880           0.2044873 
##           Fiat X1-9       Porsche 914-2        Lotus Europa      Ford Pantera L        Ferrari Dino 
##           0.1420648           0.5970469           0.3856969           0.6605028           0.3868809 
##       Maserati Bora          Volvo 142E 
##           0.5930931           0.2523986
```

Astericks would identify data points with influence (none are apparent)

Table 3 Results of anova test on various models created by incrementally adding in predictor variables:

```
## Analysis of Variance Table
## 
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + disp
## Model 4: mpg ~ am + cyl + disp + hp
## Model 5: mpg ~ am + cyl + disp + hp + drat
## Model 6: mpg ~ am + cyl + disp + hp + drat + wt
## Model 7: mpg ~ am + cyl + disp + hp + drat + wt + vs
## Model 8: mpg ~ am + cyl + disp + hp + drat + wt + vs + gear
## Model 9: mpg ~ am + cyl + disp + hp + drat + wt + vs + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)    
## 1     30 720.90                                   
## 2     28 264.50  2    456.40 30.0155 3.846e-06 ***
## 3     27 230.46  1     34.04  4.4768   0.05039 .  
## 4     26 183.04  1     47.42  6.2373   0.02380 *  
## 5     25 182.38  1      0.66  0.0866   0.77239    
## 6     24 150.10  1     32.28  4.2459   0.05599 .  
## 7     23 142.66  1      7.45  0.9793   0.33711    
## 8     21 139.26  2      3.39  0.2230   0.80257    
## 9     16 121.64  5     17.62  0.4635   0.79766    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```