Regression Models Course Project

Tiffany Stoeke

September 26, 2015

Executive Summary:

In this project, we are asked to imagine that we work for Motor Trend magazine. Looking at a data set of a collection of cars, our magazine is interested in exploring the relationship between a set of variables and miles per gallon (MPG). They are particularly interested in determining if an automatic or manual transmission is better for MPG, and to quantify the MPG difference between automatic and manual transmissions.

Our data set comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles, all of which are 1973-74 models. (The data was extracted from the 1974 Motor Trend US magazine.)

After performing multivariable regression, our findings were that while manual transmission is better for MPG than automatic, there are other variables that affect MPG more significantly that transmission type, such as horsepower and weight of the vehicle.

Exploratory Data Analysis

To begin our exploration of the data, we created a boxplot to compare automatic and manual transmissions and their effect on MPG. Per our boxplot (see Appendix, Plot 1), the mean MPG for automatic transmissions is 17.14 and the mean for manual transmissions is 24.39, an improvement in 7.24 mpg on average. Our confidence intervals (see appendix) verifies these values and shows that the p-value is low, indicating our results are significant and the means of these two categories are not equivalent (null hypothesis).

This simplistic boxplot does answer the two questions posed by the project rubric. However, with so many variables in our data set and many that appear to be specific to the engine itself, let's check if there are confounding variables that skew our results.

Exploration of Correlations Between Variables

Per our review of the correlation between variables, mpg is highly correlated with weight (wt), number of cylinders (cyl), engine displacement (disp), and horsepower (hp).

```
##        mpg        cyl       disp         hp       drat         wt
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
##       qsec         vs         am       gear       carb
##  0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

Because of this correlation, we will evaluate these variables for confounders. The data, as set up in our table, includes a number of variables that should be coerced as factors rather than as numeric values - these are V/S (vs), automatic/manual (am), number of gears (gear), number of cylinders (cyl), and number of carburetors (carb). Cylinders and carburetors are considered factor variables in this report per forum discussions and the conclusion that number of cylinders/carburetors does not affect fuel consumption as would be mathematically suggested if the values in these columns were treated as continuous integers. Making these changes creates the baseline reference model as one with 4 cylinders, automatic transmission, 3 gears, and 1 carburetor.

Modeling Results

"Remember that all models are wrong, but some are useful." – George Box

In our first model, we concern ourselves with only MPG and transmission type, as requested in the project rubric. Per the model summary (see Appendix), our p values show significance, but our adjusted R squared value shows this model only accounds for 34% of variance. If we include all variables in our model, the adjusted R squared value jumps to 78%; however, our summary lm function shows that each individual variable has a p value well over our 5% threshold and are therefore not significant while the model as a whole is significant. Hence we know we have confounders in our midst!

The R function "step" leads us to a new potential model - per the step results (not shown due to page constraints), R indicates that the best model fit is cyl+hp+wt+am. We create a new lm fit model (called "fit1") of this set of variables and see that in this set, our p values for the number of cylinders is split, while the transmission type is shown as not significant. Standard error is relatively low, however, and our R squared value is a whopping 84%. Are we guilty of overfitting? Can we do better?

Removing cylinders from consideration and running an lm summary on the new fit ("fit2") gives us similar values. This is actually good news - our F statistic has jumped, our standard error is still low and our R squared is still a high 82%. Our p values are all significant except for transmission type, so even though our project asks for review of mpg based on transmission type, what happens if we remove transmission as a confounder itself?

"Fit3" accomplishes this step. Our p values now indicate that both remaining variables (wt and hp) are significant, as is the p value for the model as a whole. Our F statistic has more than doubled from our initial "fit1" model, and our R squared is still a very high 81%. Therefore removing transmission type did little to affect our model outcome and therefore is not necessary when running regression analysis on this data set.

Diagnostics/Residuals

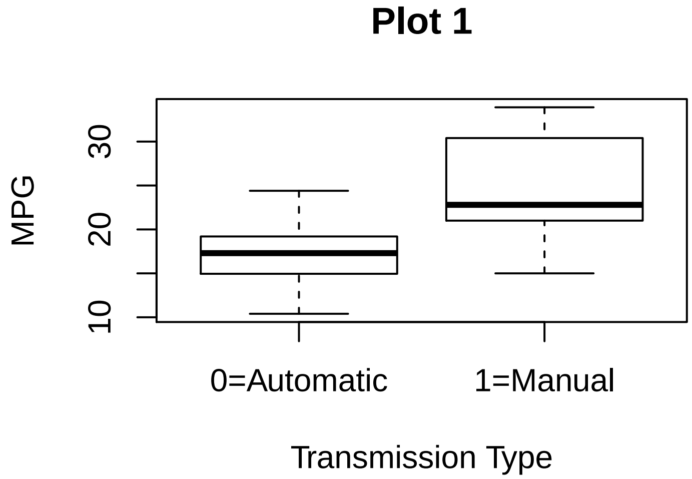
We next ran an analysis of variation (ANOVA) on our selected models to further compare/verify our model selection. Two sets of anova results were generated to compare like variables in nested order. After full anova comparisons were completed, fit3 (mpg~wt+hp) has come out on top as the best of the models reviewed in this project.

An exploration of the residuals of our chosen fit (see Appendix, Plot 3) shows that our selected model works quite well. The "Residuals vs Fitted" plot does not indicate strong patterns that would require further analysis, therefore heteroskedasticity does not appear to be an issue. Our "Normal Q-Q"" plot shows a linear relationship, therefore normality of errors. The "Scale-Location" plot shows that all points are less than 1.5 standard deviations away with no distinct pattern that requires further evaluation. The final plot ("Residuals vs Leverage") shows that while there are a few individual points that indicate potential outliers, no individual plot shows to have both high leverage and high influence at the same time.

Quantification of Results - Conclusion

Per our final results, an interpretation of our coefficients show that an increase in 1,000lb of weight leads to a decrease in an average of 4 mpg while an increase in 1 hp leads to a decrease in approximately 0.03 mpg. The horsepower result appears minuscule, but our hp range in the data set is 52-335 hp, therefore this is not an insignificant affect when total hp is considered. Our 95% confidence intervals are also included in the appendix and reinforce these averages. Note that if we included automatic vs. manual transmission in the data, we would see that the switch to manual transmission adds roughly 2.1 mpg, however due to confounding variables this data is not reliable.

Appendix



```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am1            7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt + hp + am, data = mtcars)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.9387 -1.2560 -0.4013  1.1253  5.0513 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *  
## cyl8        -2.16368    2.28425  -0.947  0.35225    
## wt          -2.49683    0.88559  -2.819  0.00908 ** 
## hp          -0.03211    0.01369  -2.345  0.02693 *  
## am1          1.80921    1.39630   1.296  0.20646    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401 
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

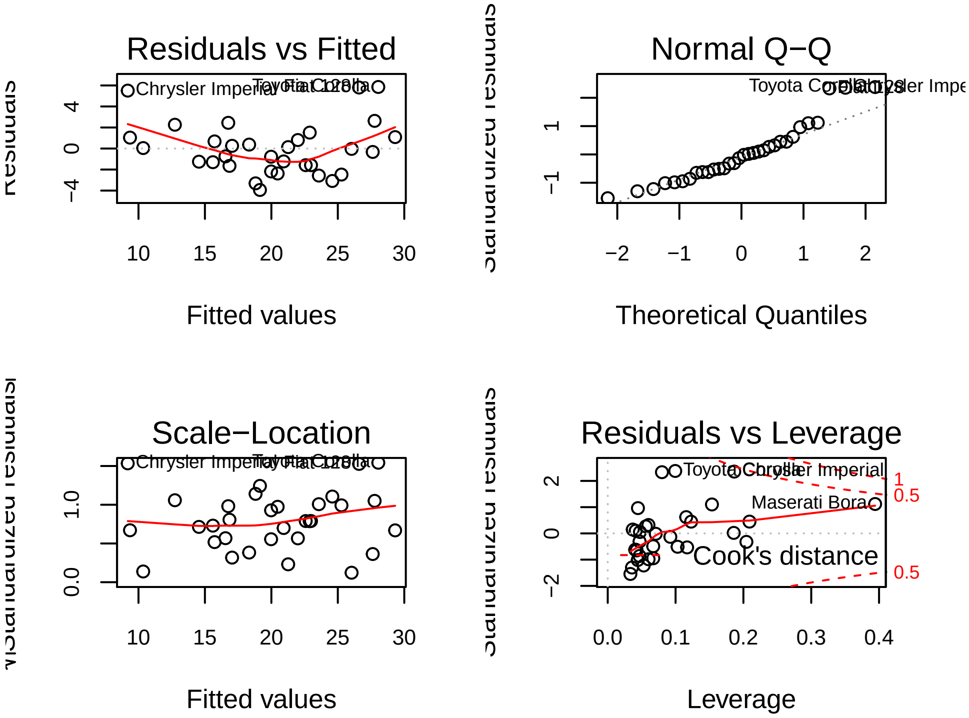
```
## 
## Call:
## lm(formula = mpg ~ wt + hp + am, data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.4221 -1.7924 -0.3788  1.2249  5.5317 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## wt          -2.878575   0.904971  -3.181 0.003574 ** 
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## am1          2.083710   1.376420   1.514 0.141268    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227 
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

```
## 
## Call:
## lm(formula = mpg ~ wt + hp, data = mtcars)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -3.941 -1.600 -0.182  1.050  5.854 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
## wt          -3.87783    0.63273  -6.129 1.12e-06 ***
## hp          -0.03177    0.00903  -3.519  0.00145 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.593 on 29 degrees of freedom
## Multiple R-squared:  0.8268, Adjusted R-squared:  0.8148
## F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12

## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + hp + am
## Model 3: mpg ~ cyl + wt + hp + am
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 180.29  2    540.61 46.5343 2.566e-09 ***
## 3     26 151.03  2     29.27  2.5191       0.1 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Variance Table
##
## Model 1: mpg ~ wt + hp
## Model 2: mpg ~ wt + hp + am
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     29 195.05
## 2     28 180.29  1    14.757 2.2918 0.1413

##                    2.5 %      97.5 %
## (Intercept) 33.95738245 40.49715778
## wt          -5.17191604 -2.58374544
## hp          -0.05024078 -0.01330512
```


Residuals vs Fitted
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals
Fitted values
Normal Q−Q
Toyota Corolla
Fiat 128
Chrysler Imperial
Standardized residuals
Theoretical Quantiles
Scale−Location
Chrysler Imperial
Toyota Corolla
Fiat 128
Standardized residuals
Fitted values
Residuals vs Leverage
Toyota Corolla
Chrysler Imperial
Maserati Bora
Cook's distance
Standardized residuals
Leverage