### Miles per Gallon (MPG) on Manual vs Automatic Transmission Vehicles

#### Frank D. Evans - Johns Hopkins Data Science, Regression Modeling

### Executive Summary

The key questions of interest concern the car mileage (mpg) and type of transmission (0:automatic, 1:manual). Considering the relationship only between those factors shows a strong relationship--though one with a low probability of being statistically signicant given the size of the data. When the other key attributes of the car are held constant, the relationship drops strongly to the transmission type--and other factors about the car show a high probability of being more strongly related to the mileage.

### Exploratory Analysis

Load the data set, and required libraries.

library(datasets); library(MASS); data(mtcars)

Based on data documentation, it can safely be assumed that transmission type is a discrete factor dimension, while mpg is measured on a bounded, yet continuous scale that will consist exclusively of positive values. To verify assumptions and get a starting point of scale of the data, we will call a summary of the mpg attribute.

summary(mtcars$mpg)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10.4 15.4 19.2 20.1 22.8 33.9

To get a sense of scale, we'll examine the number of records by factor for the transmission variable.

table(mtcars$am)

##   
## 0 1   
## 19 13

To get a prelimiary sense of data difference, we can run a summary separated by the transmission factor.

by(mtcars$mpg, mtcars$am, summary)

## mtcars$am: 0  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10.4 15.0 17.3 17.1 19.2 24.4   
## --------------------------------------------------------   
## mtcars$am: 1  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 15.0 21.0 22.8 24.4 30.4 33.9

### Regression Analysis

#### Single Variable

We first want to test a relationship directly between the variables of interest. A single variable regression seems to imply a considerable relationship.

fit\_1var <- lm(mtcars$mpg ~ mtcars$am)  
summary(fit\_1var)$coefficients; summary(fit\_1var$residuals)

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 17.147 1.125 15.247 1.134e-15  
## mtcars$am 7.245 1.764 4.106 2.850e-04

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -9.390 -3.090 -0.297 0.000 3.240 9.510

The coefficient of 7.2 would imply that in the abscence of other factors, it should be expected that a manual tranmission car (encoded in the am variable as 1) attributes to an increase of 7.2 mpg. However, with a t value of 4.1 and a considerable band of variation around the residuals--this model appears to be of poor fit. Further investigation of the plot of the residuals (Appendix: Figure 1) confirms the probable poor fit of the single variable model.

#### Multi-Variable Analysis

To find a better fitting model that best captures the variability given the variables given, a generalized linear model is fit against the principle variables. (Note: factor dimensional variables are fed into the model with a factor object conversion to automatically generate the necessary dummy variables.) Then, we leverage a stepwise algorithm to consider alternate multi-variate regression models for better candidates.

step\_fit <- glm(mpg ~ as.factor(am) + as.factor(cyl) + as.factor(gear) + disp + hp + drat + wt, data = mtcars)  
model\_fit <- stepAIC(step\_fit, direction = 'both')

model\_fit$anova

## Stepwise Model Path   
## Analysis of Deviance Table  
##   
## Initial Model:  
## mpg ~ as.factor(am) + as.factor(cyl) + as.factor(gear) + disp +   
## hp + drat + wt  
##   
## Final Model:  
## mpg ~ as.factor(am) + as.factor(cyl) + hp + wt  
##   
##   
## Step Df Deviance Resid. Df Resid. Dev AIC  
## 1 22 148.4 161.9  
## 2 - as.factor(gear) 2 1.6987 24 150.1 158.3  
## 3 - drat 1 0.3083 25 150.4 156.3  
## 4 - disp 1 0.6168 26 151.0 154.5

The results of an ANOVA test from the stepwise fit algorithm suggest a final model based on the transmission type compunded with the number of cylinders in the car's engine (cyl), the horsepower of the car (hp) and the weight of the car (wt). From an intuition perspective, this makes probable sense.

A final model is built from these factors.

final\_model <- glm(mpg ~ as.factor(am) + as.factor(cyl) + hp + wt, data = mtcars)  
summary(final\_model)$coefficients; summary(final\_model$residuals)

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 33.70832 2.60489 12.9404 7.733e-13  
## as.factor(am)1 1.80921 1.39630 1.2957 2.065e-01  
## as.factor(cyl)6 -3.03134 1.40728 -2.1540 4.068e-02  
## as.factor(cyl)8 -2.16368 2.28425 -0.9472 3.523e-01  
## hp -0.03211 0.01369 -2.3450 2.693e-02  
## wt -2.49683 0.88559 -2.8194 9.081e-03

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -3.940 -1.260 -0.401 0.000 1.130 5.050

The results of the sugested model suggest a much lower relationship, with a manual transmission attributable to a 1.8 mile per gallon lift when controlled for key corroborating factors. However, the variability of this measure is still present as the high p-value of 0.21 for the transmission factor means that there is a relatively low probability that the relationship as observed in this data is significant.

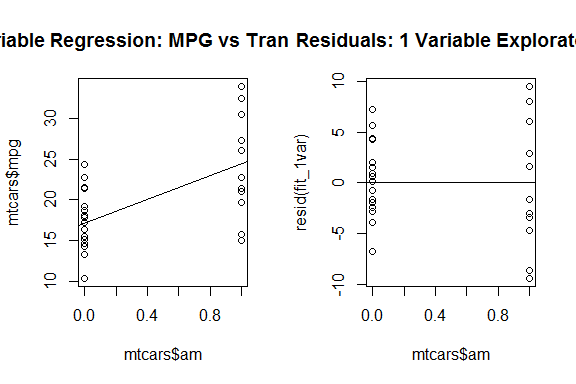
More significant factors are found among the engine size, the horsepower of the engine, and the weight of the car. The weight of the car being the most sginificant factor statistically, and interpreted as a probable 2.5 mile per gallon decrease with every 1,000 pounds of additional weight in the car itself. Behind this factor, is a likely significant factor that the presence of a 6 cylendar engine as compared to a 4 cylandar baseline attributes a reduction of mpg by 3.0 miles per gallon.

A major corroborating factor surrounding statistical signifigance in this case is the high ratio of the total number of available variables (11) to the number fo observations (32). The resultant model is likely to improve considerably with the addition of more data. However, an examination of the model fit and residual fit (Appendix: Figure 4) show that given the data and variables for analysis, incorporating the additional variable results in a significantly better explantory and predictive model.

### Appendix

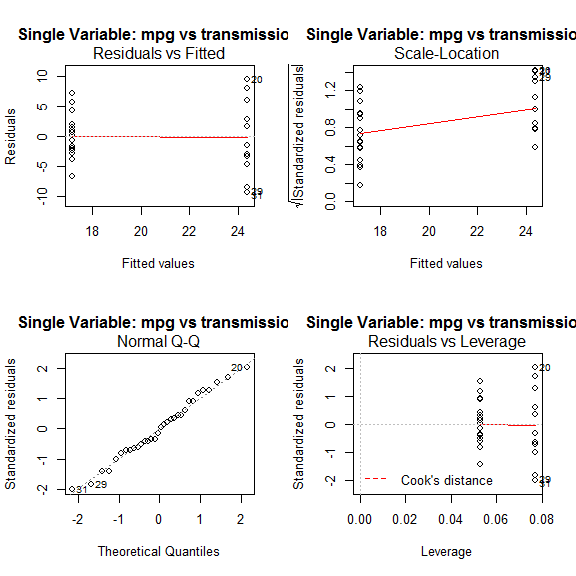
#### Figure 1: Single Variable Regression Plot and Residuals

par(mfrow = c(1,2))  
plot(mtcars$am, mtcars$mpg, main = '1 Variable Regression: MPG vs Transmission')  
abline(fit\_1var)  
plot(mtcars$am, resid(fit\_1var), main = 'Residuals: 1 Variable Exploratory')  
abline(h = 0)



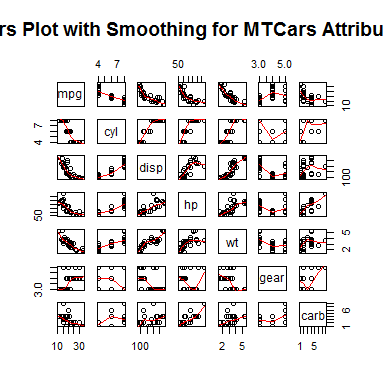
#### Figure 2: Single Variable Regression - Full Model Fit Plots

layout(matrix(c(1,2,3,4),2,2))  
plot(fit\_1var, main = 'Single Variable: mpg vs transmission')



#### Figure 3: Pairs Plot for MTCars Attributes

test\_vars <- c('mpg','cyl','disp','hp','wt','gear','carb')  
pairs(x = mtcars[,test\_vars],   
 panel = panel.smooth,  
 main = 'Pairs Plot with Smoothing for MTCars Attributes')



#### Figure 4: Final Model Regression - Full Model Fit Plots

layout(matrix(c(1,2,3,4),2,2))  
plot(final\_model, main = 'Final Model: mpg vs transmission and other factors')

