Regression Models Course Project

Executive Summary:

This report was commissioned to examine the relationship between type of transmission and fuel efficiency of a car (miles per gallon). The objective is to analyze predictors of high fuel efficiency and then quantify the relationship between type of transmission and miles per gallon (mpg).
There are two types of analyses performed in this report. The first one is a simple exploratory effort to draw conclusions based on simple visualizations. The second type consists of applying ordinary least squares and multivariate linear regressions to the data in an attempt to quantify the effect of transmission type on mpg.

The following are the results and recommendations based on the analyses:

- Regarding fuel efficiency, a manual transmission is slightly better than an automatic transmission.
  - This, of course, depends on the style of driving of any particular individual since they have control of rpm.
- Standard transmission yields, on average, a 2.08 mpg increase in fuel efficiency compared to automatic transmissions with a 1.37 standard error. It must be noted, however, that the P-value of this parameter is relatively high at .14, which suggests the need for additional study.

Analysis:

In this exercise, the data is given in a tidy, tabular format that is easily analyzed. The variables 'cyl', 'vs', 'am', 'gear', and 'carb' will be changed to factor variables as necessary.

1. Exploratory Analysis:

One method used in this exercise was Exploratory Analysis. The purpose of this type of analysis is to find relationships in the data that were not previously known, but it cannot be used alone to predict or generalize (correlation is not causation!).

The first step in exploratory analysis was to plot the outcome and regressor of interest .in this case "mpg" vs. "am" [Appendix: Fig. 1]. Based on the plot, we can hypothesize that manual transmission cars have, on average, higher mpg than automatic cars.

2. Ordinary Least Squares (OLS)

In order to accurately determine the response that type of transmission causes on mpg, its effect needs to be isolated from other regressors. Simple OLS fit [Appendix: Fit 0 - OLS] indicates that standard transmission yields, on average, a 7.24 mpg increase in fuel efficiency compared to automatic transmission. It must be noted, however, that the standard error of the residuals is 4.9 and the R^2 value is only 36%.

3. Multivariate Linear Regression & Nested Model Testing

A relatively high error and low Rˆ2 value in the OLS indicate a poor fit, and therefore a multivariate regression fit will be performed. It is assumed that some regressors have more influence on mpg than others, and therefore a Nested Model test approach will be implemented. Based on basic automobile knowledge, it can be assumed that weight of a car and engine size have a considerable effect on fuel efficiency. The effect that all other variables have on mpg will also be tested, but their addition to the model will depend on their statistical significance.

Upon inspection of the data columns, it is apparent that some of the variables are strongly correlated. For example, there could be multiple descriptors of 'engine size' such as number of cylinders, displacement and horsepower. Before a model is tested, the correlations between all variables must be tested. A visualization of all the correlations was performed using the corrplot package [Appendix: Fig. 2]. It is apparent from the plot that wt, hp, cyl, and disp have the most effect on mpg. At the same time, however, wt-disp, hp-cyl, cyl-disp, hp-disp all have strong correlations. Based on these observations,the regression model will use engine size, weight, and transmission type as regressors. Hp will be used as the 'engine size' variable (instead of cyl or disp) due to its low correlation with transmission type.

The model summary below [Details in Appendix: Fit 1 - Multivariate] shows that most of the variation in mpg from the first OLS model can be 'explained away' by other predictors namely hp and weight. The low P values of wt and hp in our new model indicate that they are majorly significant and therefore relevant to the model. This model indicates that that standard transmission yields, on average, a 2.08 mpg increase in fuel efficiency compared to automatic transmission with a 1.37 standard error. It must be noted, however, that the P-value is .14 for the factor variable(am)1, which is typically not considered statistically significant in a hypothesis test.

Fit 1 coefficients:

```r
fit1 <- lm(mpg ~ wt + hp + factor(am), data=df)
summary(fit1)$coef
```

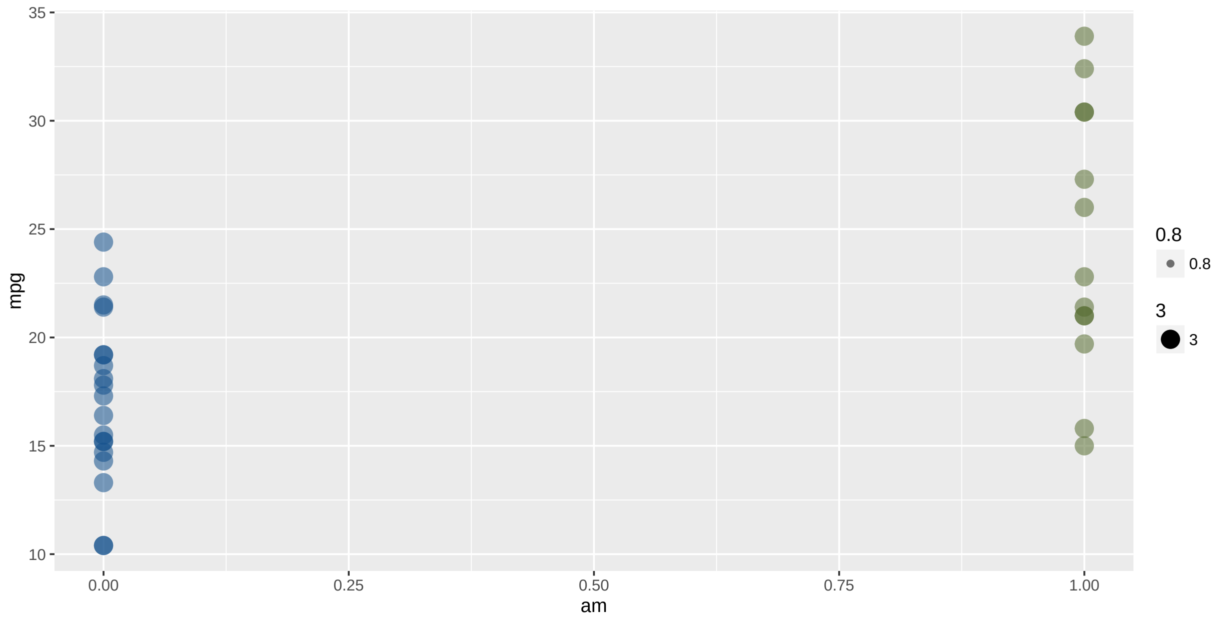
```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 34.00287512 2.642659337 12.866916 2.824030e-13
## wt          -2.87857541 0.904970538 -3.180850 3.574031e-03
## hp          -0.03747873 0.009605422 -3.901830 5.464023e-04
## factor(am)1  2.08371013 1.376420152  1.513862 1.412682e-01
```

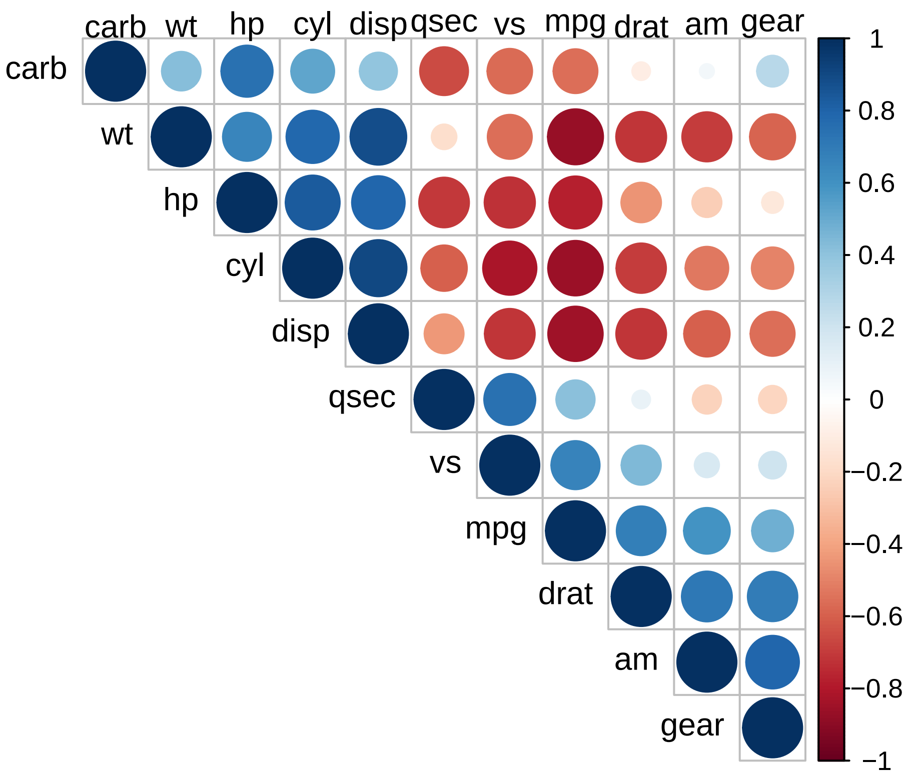
For completeness, a model summary including the effect of all available predictors is included for comparison [Appendix: Fit ALL - Multivariate]. This model indicates a .96 increase in mpg for standard transmissions. Note that fitting all the variables makes all the P-values statistically insignificant. In addition, the error is larger and the adjusted Rˆ2 value is lower, indicating a poor model fit. Based on these results, the final model for this project will remain Fit 1.

The residual plots for the chosen model (fit 1) is located in the appendix. The residual plots indicate that the errors are random with no particular pattern. As a result, we can be confident that there is no obvious correlation left unexplained by the model variables.

Appendix:

Fig. 1 - mpg vs. transmission type (0=auto, 1=std)





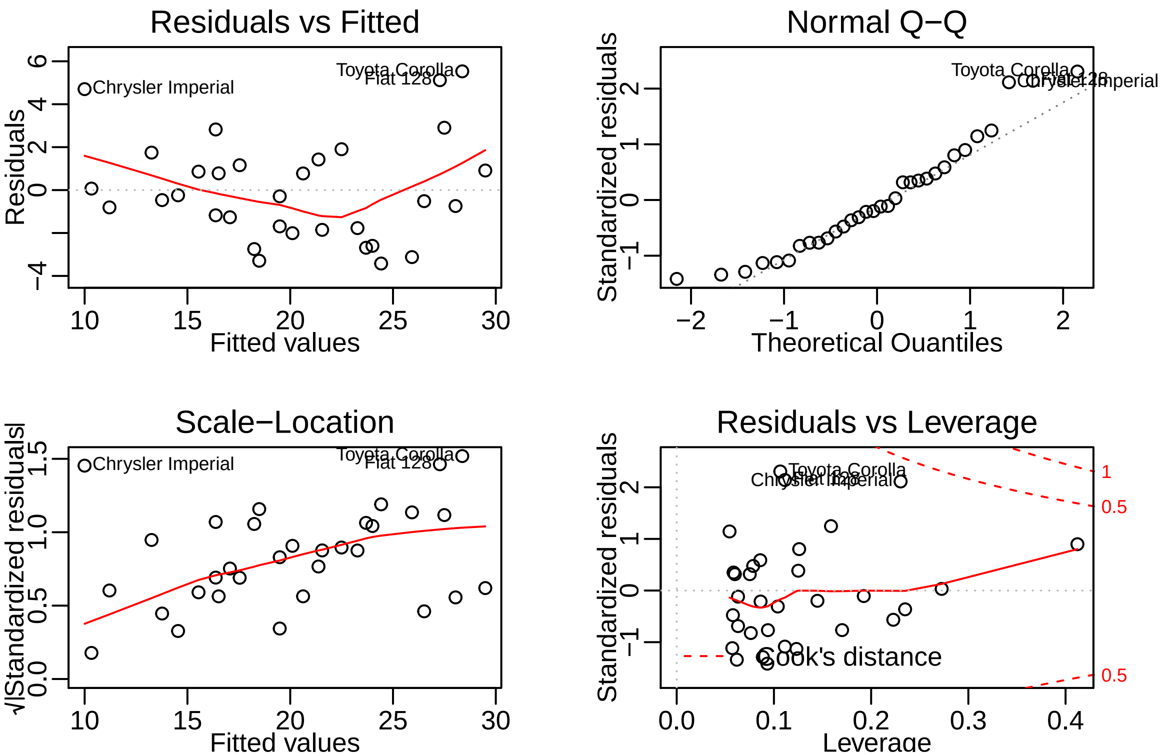
#### Fig. 2 - Correlation Plot:
#### Fit 0 - OLS

```
##
## Call:
## lm(formula = mpg ~ factor(am), data = df)
##
```

```
## 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 ***
## factor(am)1    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
```

Fit 1 - Multivariate

```
## 
## Call:
## lm(formula = mpg ~ wt + hp + factor(am), data = df)
## 
## 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 ***
## factor(am)1  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
```



#### Fit ALL - Mutivariate

```
##                    Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept)     25.31994337 23.88164477  1.0602261 0.30478503
## wt              -4.63539945  2.52736612 -1.8340831 0.08530813
## hp              -0.04881225  0.03189192 -1.5305523 0.14541042
## cyl             -1.02343435  1.48131027 -0.6908980 0.49953134
## disp             0.04376554  0.03057568  1.4313841 0.17156359
## drat             1.82084238  2.38100971  0.7647354 0.45556110
## qsec             0.26966987  0.92631150  0.2911222 0.77469794
## factor(vs)1      1.04907556  2.70494812  0.3878357 0.70324874
## factor(gear)4    1.75359631  3.72533672  0.4707216 0.64419293
## factor(gear)5    1.87898502  3.65935137  0.5134749 0.61463655
## factor(carb)2   -0.93427482  2.30934499 -0.4045627 0.69115583
## factor(carb)3    3.42168886  4.25512809  0.8041330 0.43310616
## factor(carb)4   -0.99363962  3.84682616 -0.2583011 0.79946771
## factor(carb)6    1.94388997  5.76982873  0.3369060 0.74056649
## factor(carb)8    4.36998439  7.75434447  0.5635530 0.58087155
## factor(am)1      0.96265239  3.19137777  0.3016416 0.76681049
```