Regression Models Course Project

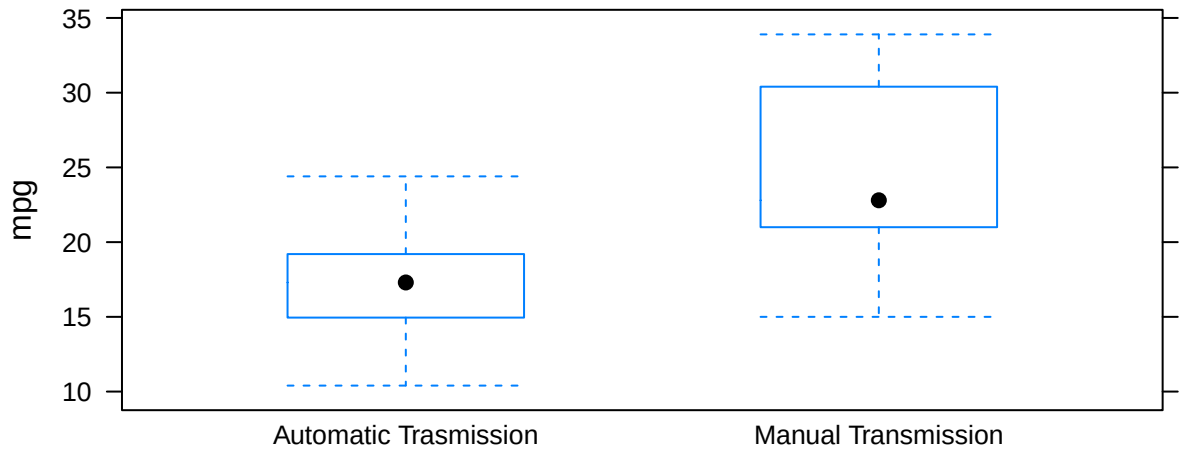
Executive Summary

This analysis uses the mtcars dataset available in R. According to the help file for this data set, "this data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models)." The goal of this analysis is to answer the question "Is an automatic or manual transmission better for MPG?" and to "Quantify the MPG difference between automatic and manual transmissions." The analysis will show that a manual transmission provides better fuel economy and that in a car with average horse power, and a 4 cylinder, straight engine, a manual transmission that expected difference is about 5.2 miles per gallon.

Exploratory Data Analysis

In order to decide which variables are appropriate for inclusion attempts to determine the effect of transmission type on fule efficiency, some exploratory data analysis was done. Note that prior to this analysis, some pre-processing was performed on the data. This can be found in the appendix. A summary of the data can be found in the appendix as well.

The oucome of interest - mpg by transmission type - is graphed below.



From this plot it is clear to see that manual transmissions have better fule economy. However, it is unknown if this will hold true when other variables are controlled for.

Linear Models

Several models will be presemnted and analyzed below. Three main creteria will be used to evaluate the appropriateness of each model. 1. The adjusted R-squared, 2. The interpretablility of the model (e.g. how much sense does it make to include the included variables), and 3. The signifigance of the term of interest - the trasmission type variable.

The first model is the model which contains only the fuel efficiency and transmission type variables. The coefficients and adjusted R-squared of which are below.

```
## $coefficients
##                          Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)           17.147368   1.124603 15.247492 1.133983e-15
## amManual Transmission  7.244939   1.764422  4.106127 2.850207e-04
##
## $adj.r.squared
## [1] 0.3384589
```

The benefits of this model are that it is very interpretable and that the Transmission term is significant. However, no ohter factors are accounted for and the adjusted R-squared is low.

The second model is the model which contains all of the variables in the mtcars dataset. The adjusted R-squared of which is presented below. For space considerations, the coefficients of this model appear in the appendix.

```
## $adj.r.squared
## [1] 0.7790215
```

The benefit of this model is that all variables are accounted for and that the adjusted R-squared is much higher than the first model. However, it is not clear that it is appropriate to include all of the variables. Many may be measuring almost the same thing or may be simply unrelated. The transmission term in the model is not significant, likely due to a high number of other terms causing overfitting.

The third and final model that will be included here is the model that predicts fuel economy using the transmission type, horse power, number of cylinders, and engine shape variables. For the sake of later interpretablity, the horse power variable has been replaced by the difference between the horse power and the mean horse power. The coefficients and adjusted R-squared of which are below.

```
## $coefficients
##                             Estimate Std. Error     t value     Pr(>|t|)
## (Intercept)           20.13974477 1.80440355 11.16144159 2.051244e-11
## amManual Transmission  5.16287130 1.45386237  3.55114170 1.489216e-03
## hpavg                  0.04687855 0.01451486  3.22969437 3.346750e-03
## cyl6                  -2.65245486 1.79590505 -1.47694604 1.517011e-01
## cyl8                  -0.27710473 3.48664077 -0.07947613 9.372625e-01
## vsV Engine            -2.56902830 1.94243080 -1.32258421 1.974901e-01
##
## $adj.r.squared
## [1] 0.8043595
```

The benefits of this model are that the included variables make sense because they each have an obvious effect on fuel economy (see the appendix for plots demostrating each of these effects.) This model also has a higher adjusted R-squared than either previous model. Finally, the term regarding the transmission type is significant in this model. That makes interpreting that coefficient more meaningful. Several diagnostic plots of this model appear in the appendix. No major issues appear in them.

Conclusions

The final model (and, in fact, all of the models included) show that manual transmissions had better fuel efficiency in 1974. The model has a positive coefficient for the "Manual Transmission" term. This matches what we expected from the initial exploratory analysis.

This model tells us that if a car changes from the base case - automatic transmission, average horse power, 4 cylinders, and a straight engine - and switches to a manual transmission, it can expect to gain 5.16 miles per gallon of fuel efficiency. The 95% confidence interval for this value is [2.1744144, 8.1513282]:

Appendix

Pre-processing

```
     data(mtcars)
# Replace automatic/manual factor with descriptive terms
    mtcars$am[which(mtcars$am==0)]<-"Automatic Trasmission"
    mtcars$am[which(mtcars$am==1)]<-"Manual Transmission"
    mtcars$am<-as.factor(mtcars$am)

# Replace "vs" factor with descriptive terms
    mtcars$vs[which(mtcars$vs==0)]<-"V Engine"
    mtcars$vs[which(mtcars$vs==1)]<-"Straight Engine"
    mtcars$vs<-as.factor(mtcars$vs)

# Convert cylinders, # of gears, and # of carburators to factors
    mtcars$cyl<-as.factor(mtcars$cyl)
    mtcars$gear<-as.factor(mtcars$gear)
    mtcars$carb<-as.factor(mtcars$carb)
```

Summary of mtcars data

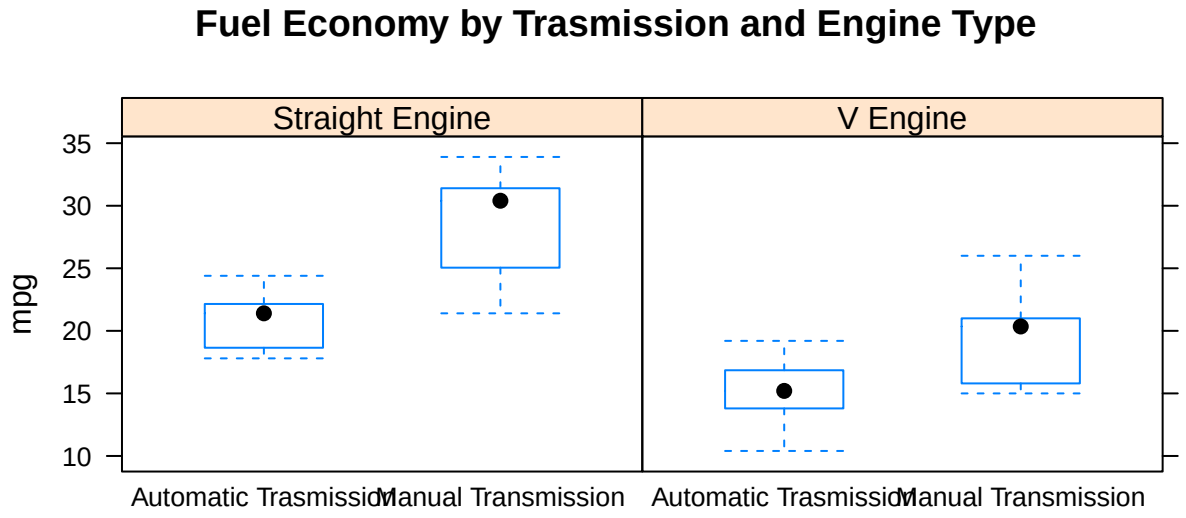
```
##       mpg        cyl         disp             hp             drat
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930
##        wt             qsec                     vs
##  Min.   :1.513   Min.   :14.50   Straight Engine:14
##  1st Qu.:2.581   1st Qu.:16.89   V Engine       :18
##  Median :3.325   Median :17.71
##  Mean   :3.217   Mean   :17.85
##  3rd Qu.:3.610   3rd Qu.:18.90
##  Max.   :5.424   Max.   :22.90
##                      am     gear   carb
##  Automatic Trasmission:19   3:15   1: 7
##  Manual Transmission  :13   4:12   2:10
##                             5: 5   3: 3
##                                    4:10
##                                    6: 1
##                                    8: 1
```

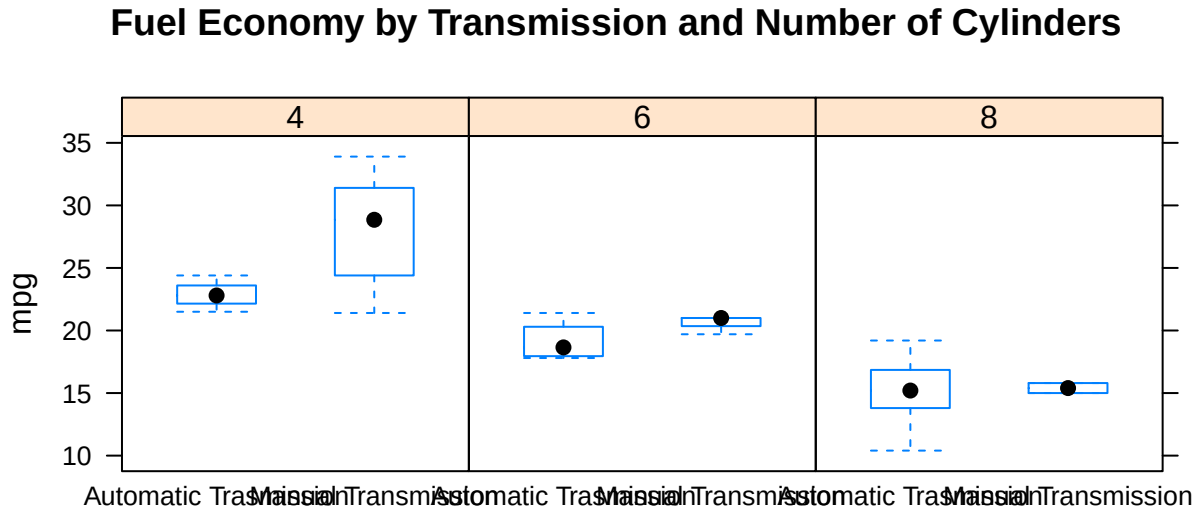
Coefficients of linear model with all variables

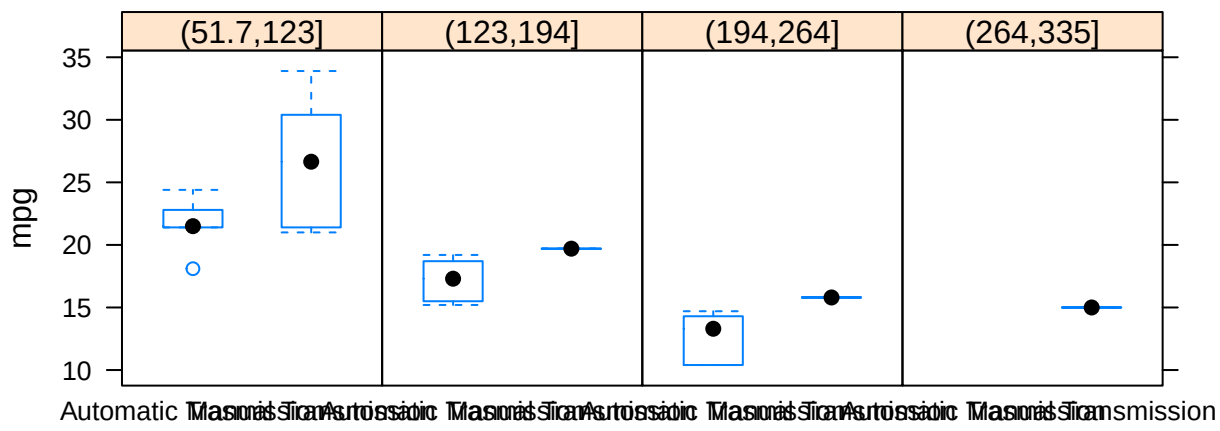
```
## $coefficients
##                          Estimate  Std. Error     t value   Pr(>|t|)
## (Intercept)          25.80998298 20.26412882  1.27367839 0.22216055
## cyl6                 -2.64869528  3.04089041 -0.87102622 0.39746642
## cyl8                 -0.33616298  7.15953951 -0.04695316 0.96317000
```

```
## disp                   0.03554632  0.03189920  1.11433290 0.28267339
## hp                    -0.07050683  0.03942556 -1.78835344 0.09393155
## drat                   1.18283018  2.48348458  0.47627845 0.64073922
## wt                    -4.52977584  2.53874584 -1.78425732 0.09461859
## qsec                   0.36784482  0.93539569  0.39325050 0.69966720
## vsV Engine            -1.93085054  2.87125777 -0.67247551 0.51150791
## amManual Transmission  1.21211570  3.21354514  0.37718957 0.71131573
## gear4                  1.11435494  3.79951726  0.29328856 0.77332027
## gear5                  2.52839599  3.73635801  0.67670068 0.50889747
## carb2                 -0.97935432  2.31797446 -0.42250436 0.67865093
## carb3                  2.99963875  4.29354611  0.69863900 0.49546781
## carb4                  1.09142288  4.44961992  0.24528452 0.80956031
## carb6                  4.47756921  6.38406242  0.70136677 0.49381268
## carb8                  7.25041126  8.36056638  0.86721532 0.39948495
```

Effects of included variables on fuel economy







Diagnostic Plots on Final Model