The Effect of Transmission on Fuel Economy

Executive Summary

The purpose of this analysis is to quantify the relationship between vehicle transmission and fuel economy, measured in miles per gallon (mpg). The data we use is taken from the `mtcars` dataset included in the R programming language . This data is taken from the 1974 Motor Trend magazine, and details 11 aspects of motor vehicle design and performance. We find that on average, cars with manual transmissions are more fuel efficient than those with automatic transmissions by 2.08 miles per gallon. This effect, however, is not statistically significant after adjusting for variations in weight and horsepower.

Analysis

This analysis requires the `ggplot2` and `GGally` graphics packages, and the `mtcars` dataset. In Figure 1, we create a matrix plot of the columns of mtcars to see what relationships there are between the variables. Figure 2 is a boxplot comparing fuel economy between automatic and manual transmission cars.

We find many relationships between various columms. This can be problematic for linear regression; omitting interaction terms can make models less accurate, but including these terms makes models much harder to interpret. We will fit a linear model in a stepwise fashion without interaction terms, starting with transmission type. To limit the effect of the corelation between variables, we will limit our model to 3 predictors.

First, we create the one-predictor model based on transmission type.

```
mpg_am <- lm(mpg ~ am, data=mtcars)
summary(mpg_am)[c("coefficients","r.squared")]
```

```
## $coefficients
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.134e-15
## am             7.245      1.764   4.106 2.850e-04
##
## $r.squared
## [1] 0.3598
```

As expected, our model does not do a great job of explaining the variance in the data set; the R squared is only 0.338.

The next predictor we add is the one which improves the R squared statistic the most; to find this predictor, we caclulate the linear regression for each remaining column and compare the R squared.

```
variablesLeft <- setdiff(names(mtcars), c("mpg", "am"))
models1 <- lapply(variablesLeft, FUN=function(x) {
    form <- paste0("mpg ~ am + ", x)
    lm(form, data=mtcars)
})
results <- cbind(column=variablesLeft,
                 adj.r.squared=sapply(models1,
                                      FUN=function(x) summary(x)$adj.r.squared))
results[order(results[,2], decreasing=TRUE)[1],]
```

```
##              column       adj.r.squared
##                "hp" "0.767002539013904"
```

```
summary(models1[[3]])[c("coefficients","r.squared")]
```

```
## $coefficients
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 26.58491   1.425094  18.655 1.074e-17
## am           5.27709   1.079541   4.888 3.460e-05
## hp          -0.05889   0.007857  -7.495 2.920e-08
##
## $r.squared
## [1] 0.782
```

We find that the model evaluating horsepower has the highest R squared statistic. We repeat the process to find our last predictor to add to the model.

```
variablesLeft <- setdiff(variablesLeft, "hp")
models2 <- lapply(variablesLeft, FUN=function(x) {
    form <- paste0("mpg ~ am + hp + ", x)
    lm(form, data=mtcars)
})

results <- cbind(column=variablesLeft,
                 adj.r.squared=sapply(models2,
                                      FUN=function(x) summary(x)$adj.r.squared))
results[order(results[,2], decreasing=TRUE)[1],]
```

```
##              column       adj.r.squared
##                "wt" "0.822735694896529"
```

```
summary(models2[[4]])[c("coefficients","r.squared")]
```

```
## $coefficients
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 34.00288   2.642659  12.867 2.824e-13
## am           2.08371   1.376420   1.514 1.413e-01
## hp          -0.03748   0.009605  -3.902 5.464e-04
## wt          -2.87858   0.904971  -3.181 3.574e-03
##
## $r.squared
## [1] 0.8399
```

We find that weight explains the most variance in the rest of the data. Holding these two covariates constant, we find that the difference between manual and automatic transmission fuel economy is 2.08, but is not significant. Looking at the Residuals plot in Figure 4, we see that there is a distinct U shape in the plot; this is not unexpected, as we intentionally disregarded the effect of interactions between the variables and limited the number of predictors in the model. Incorporating interaction terms would likely improve the fit of the model and reduce the pattern in this residuals plot.

The major weakness of this analysis is that by ignoring interaction terms, we sacrifice power in order to make the interpretation of the model easier. As such, our conclusion of a lack of relation between mpg and transmission type must be taken cautiously. A study with a larger sample size that considers the interaction between the variables may yield better results.

Appendix: Figures

Figure 1: Matrix plot of mtcars

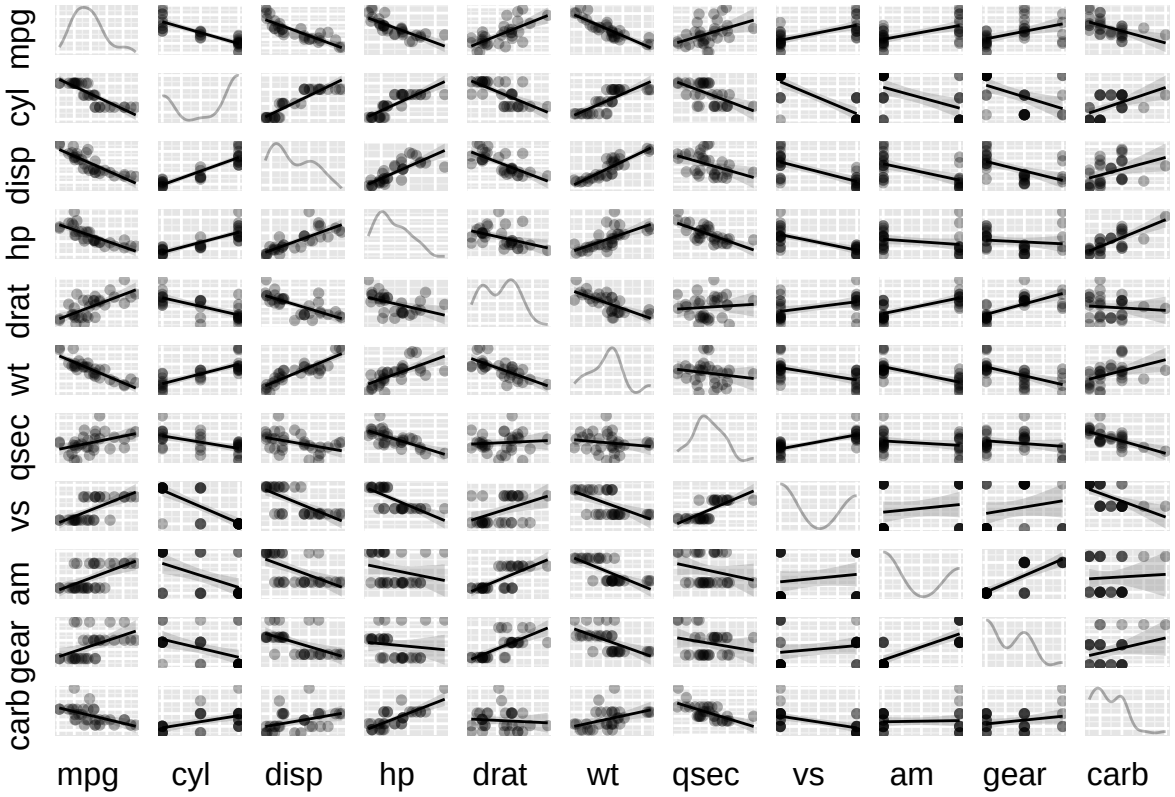

Figure 2: Fuel Economy vs. Transmission

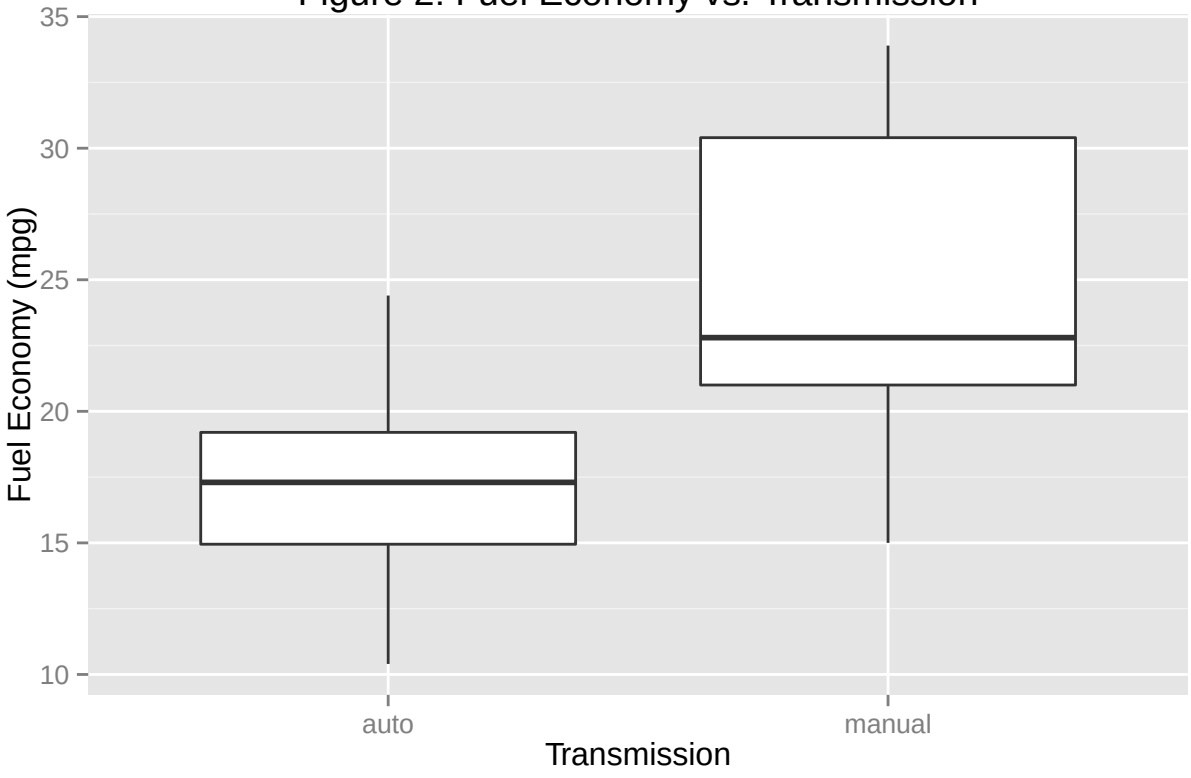

Figure 3: Fuel Economy vs. Horsepower

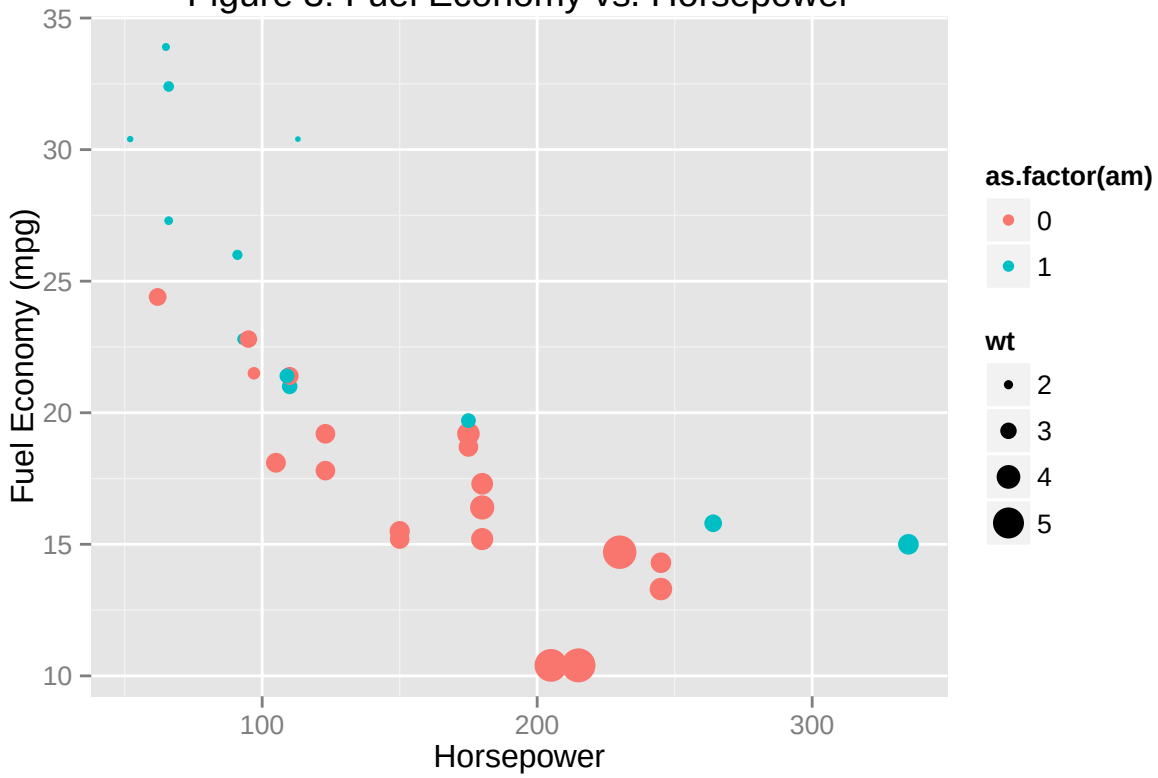

Figure 4: Residuals versus Fitted Values