Motor Trend Analysis

Anthony Cato

April 24, 2017

Summary

The goal of this analysis is to answer the following questions:

1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions

The data source for this project comes from the **mtcars** dataset which is made available by default in R.

Data Prep

Import data.table

```
library(data.table, quietly = TRUE)
library(ggplot2, quietly = TRUE)
```

```
data(mtcars)
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("Automatic", "Manual")
mtcars <- as.data.table(mtcars)
```

Hypothesis Testing

Average MPG for each transmission type

```
avg.trans <- mtcars[, .(avg_mpg = mean(mpg)), by = am]
names(avg.trans)[1] <- c("trans_typ")
avg.trans
```

```
##    trans_typ  avg_mpg
## 1:    Manual 24.39231
## 2: Automatic 17.14737
```

In comparing the miles per gallon for automatic and manual transmission cars, we observe that the average manual car in this dataset have better fuel economy than the automatic cars. (see Fig. 1)

T Test

```
test <- t.test(mpg ~ am, data = mtcars)
```

Using the `t.test` function, we get a p-value of 0.0013736, indicating that there is strong evidence to reject the null hypothesis. In other words, a p-value this low suggests that there is a relationship between mpg and transmission type that is worth examining. (see Fig. 2)

Simple Linear Regression

```
fit <- lm(mpg ~ am, data = mtcars)
```

A simple linear regression model (see Fig. 3) shows that while transmission types have significant p-values, they are poor predictors of mpg. This poorness-of-fit is summarized by the adjusted r-squared of `0.3385`. This means that roughly `33.85`% of the variance in our data can be explained by this model. To get a better look at the relationship, we must include other variables. To this end, I've tested a few multivariate regression models to find the features that when combined serve as a better predictor of a car's fuel economy.

Multivariate Regression

```
# Factoring in Transmission and Number of Cylinders
fit.cyl <- update(fit, mpg ~ am + factor(cyl))
# Factoring in Transmission, Number of Cylinders, and Weight
fit.cyl.wt <- update(fit.cyl, mpg ~ am + factor(cyl) + wt)
# Factoring in Transmission, Number of Cylinders, Weight, and Horsepower
fit.cyl.wt.hp <- update(fit.cyl.wt, mpg ~ am + factor(cyl) + wt + hp)
```

Anova Comparison

The `anova` comparison (see Fig. 4) shows that model 2 shows that much of the fluctuations in mpg can be explained by the transmission and type and number of cylinders as indicated by a r-squared of `0.7399`. The p-value, `1.3875174\times 10^{-8}` suggests that these two factors are a good starting point for the analysis.

Adjusted R Square for each model

```
# Formula - mpg ~ am + factor(cyl)
summary(fit.cyl)$adj.r.squared
```

```
## [1] 0.7399447
```

```
# Formula - mpg ~ am + factor(cyl) + wt
summary(fit.cyl.wt)$adj.r.squared
```

```
## [1] 0.8134405
```

```
# Formula - mpg ~ am + factor(cyl) + wt + hp
summary(fit.cyl.wt.hp)$adj.r.squared
```

```
## [1] 0.8400875
```

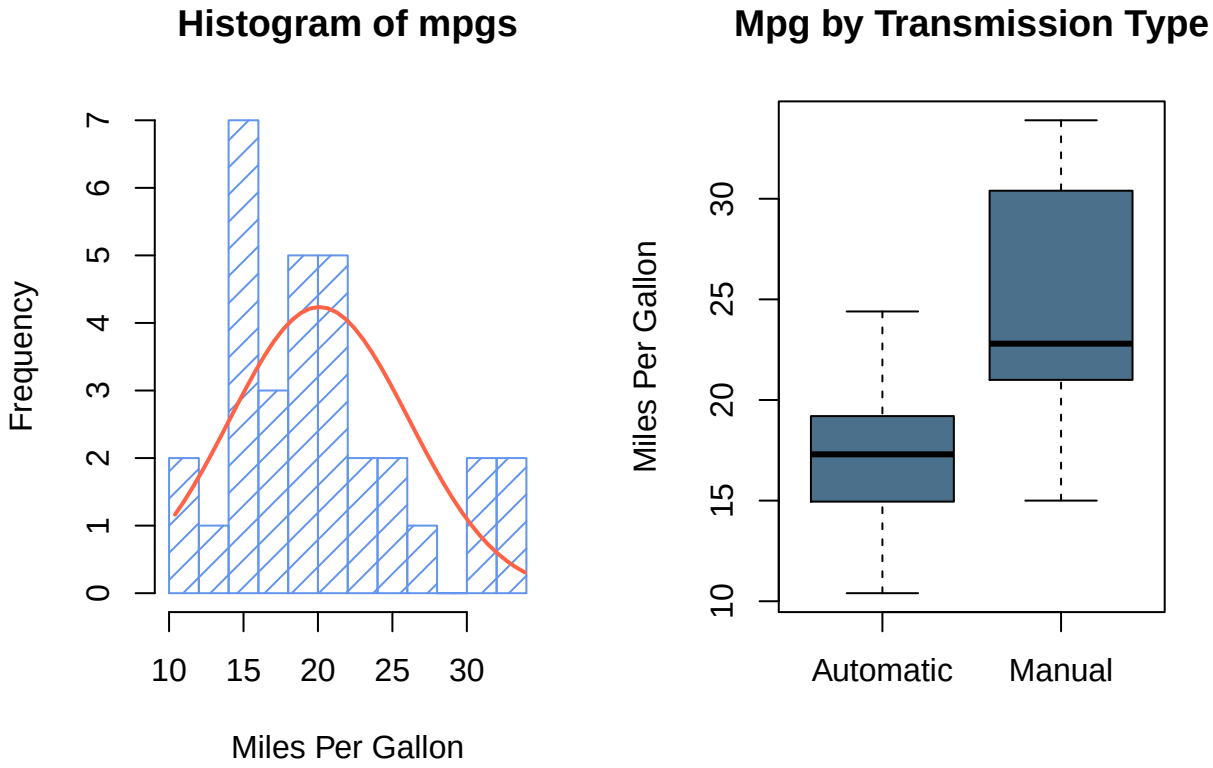
```
result <- round(summary(fit.cyl.wt.hp)$adj.r.squared, 4)
result.pct <- result * 100
```

Conclusion

The third model, which explains how mpg is influenced by transmission type, number of cylinders, weight, and horsepower, yields the highest adjusted r_squared at **0.8401**. This r squared means that more than **84.01%** of the variance in our data can be explained by this model. Because of this high predictive accuracy, I would choose this model. For supporting plot, see Figure 5.

Appendix

Histogram and Boxplot of Mpgs (Fig. 1)



T Test (Fig. 2)

```
test
```

```
##
##  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 Automatic    mean in group Manual
##                17.14737                24.39231
```

Simple Linear Regression (Fig. 3)

```
summary(fit)
```

```
##
## 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 ***
## amManual       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
```

Anova Comparison (Fig. 4)

```
result.models
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + factor(cyl)
## Model 3: mpg ~ am + factor(cyl) + wt
## Model 4: mpg ~ am + factor(cyl) + wt + hp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 264.50  2    456.40 39.2861 1.388e-08 ***
## 3     27 182.97  1     81.53 14.0354 0.0009026 ***
## 4     26 151.03  1     31.94  5.4991 0.0269346 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual Plotting for Best Model (Fig. 5)