Regressive Analysis on Fuel Efficiency

Jiachang (Ernest) Xu

6/16/2017

Executive Summary

This project will explore the relationship between miles per gallon (MPG) and other variables from the Motor Trends dataset. We are particularly interested in the following two questions:

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

Our first guess is manual transmission is better for MPG than automatic transmission. The exploratory t.test verifies this guess. We try different regressors to find the optimal model based on the highest adjust R-squared with a significant p-value.

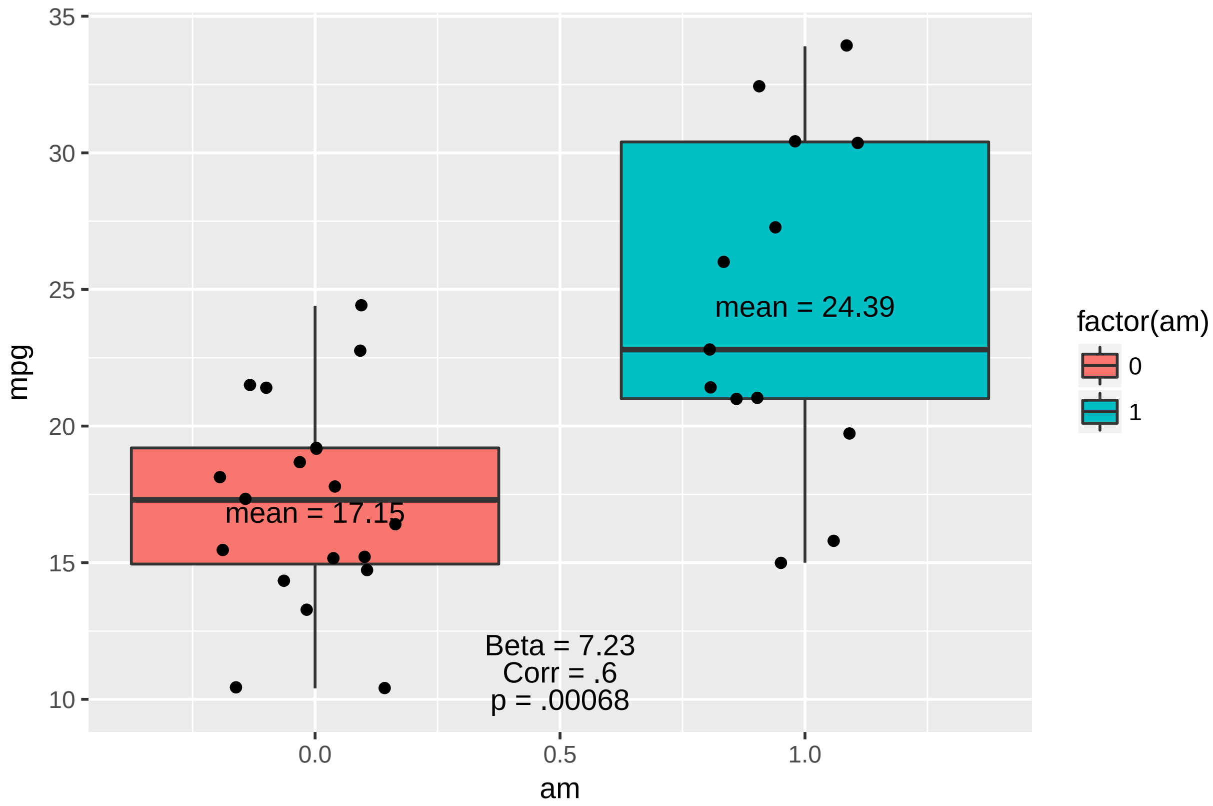
Exploratory Analysis

The following t.test and boxplots (Figure 1) show that the average MPG of manual transmission is significantly higher that that of automatic transmission. Therefore, based on the siginificant level of 5%, we can conclude that **manual transmission is better for MPG than automatic transmission**. Please refer to Appendix I about loading and splitting data.

```
t.test(manual$mpg, auto$mpg, alternative = "less")$p.value
```

```
## [1] 0.0006868192
```

Figure 1: Sample Distribution of MPG based on Transmission Types



Model Selection

Model 1 includes all variables, but has an insignificant p-value.

```
model_1 <- lm(mpg~., mtcars)
model_1
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Coefficients:
## (Intercept)          cyl         disp           hp         drat
##    12.30337     -0.11144      0.01334     -0.02148      0.78711
##          wt         qsec           vs           am         gear
##    -3.71530      0.82104      0.31776      2.52023      0.65541
##        carb
##    -0.19942
```

```
summary_1 <- append(summary(model_1)$coef[2,], c(summary(model_1)$adj.r.squared, summary(model_1)$r.squa
names(summary_1)[5:6] <- c("Adj R^2", "R^2")
summary_1
```

```
##   Estimate Std. Error    t value   Pr(>|t|)    Adj R^2        R^2
## -0.1114405  1.0450234 -0.1066392  0.9160874  0.8066423  0.8690158
```

Model 2 includes am only, but only explains 34% of the variability.

```r
model_2 <- lm(mpg~am, mtcars)
model_2
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Coefficients:
## (Intercept)           am
##      17.147        7.245
```

```r
summary_2 <- append(summary(model_2)$coef[2,], c(summary(model_2)$adj.r.squared, summary(model_2)$r.squa
names(summary_2)[5:6] <- c("Adj R^2", "R^2")
summary_2
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)      Adj R^2
## 7.2449392713 1.7644216316 4.1061269831 0.0002850207 0.3384589082
##          R^2
## 0.3597989434
```

Model 3 uses algorithm to select optimal regressors. It has both higher variabiliry coverage and significant p-value.

```r
model_3 <- df[[memdex]]
model_3
```

```
##
## Call:
## FUN(formula = X[[i]], data = ..1)
##
## Coefficients:
## (Intercept)           am         disp           hp           wt
##    14.36190      3.47045      0.01124     -0.02117     -4.08433
##        qsec
##     1.00690
```

```r
summary_3 <- append(summary(model_3)$coef[2,], c(summary(model_3)$adj.r.squared, summary(model_3)$r.squa
names(summary_2)[5:6] <- c("Adj R^2", "R^2")
summary_3
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
## 3.47045340 1.48578009 2.33577865 0.02748781 0.83753338 0.86373768
```

Conclusion

Our analysis shows that **manual transmission is more fuel eifficient than automatic transmission, by an average of 3.407 MPG higher holding all other variables constant**.

Appendix I: Loading and Splitting Data

- Loading Data

```r
require(datasets)
require(plyr)
require(ggplot2)
```

```r
require(GGally)
require(car)
data("mtcars")
attach(mtcars)
```

- Splitting Data

```r
auto <- mtcars[which(am == 1), ]
manual <- mtcars[which(am == 0), ]
```

Appendix II: Figure 1 Plotting Code

```r
figure_1 <- ggplot(mtcars, aes(y=mpg,x=am)) + geom_boxplot(aes(fill=factor(am))) + labs(title = "Figure

figure_1
```

Appendix III: Model Selection

```r
model_1 <-lm(mpg~., mtcars)
model_1
```

```r
summary_1 <- append(summary(model_1)$coef[2,], c(summary(model_1)$adj.r.squared, summary(model_1)$r.squa
names(summary_1)[5:6] <- c("Adj R^2", "R^2")
summary_1
```

```r
model_2 <-lm(mpg~am, mtcars)
model_2
```

```r
summary_2 <- append(summary(model_2)$coef[2,], c(summary(model_2)$adj.r.squared, summary(model_2)$r.squa
names(summary_2)[5:6] <- c("Adj R^2", "R^2")
summary_2
```

```r
model_3 <- df[[memdex]]
model_3
```

```r
summary_3 <- append(summary(model_3)$coef[2,], c(summary(model_3)$adj.r.squared, summary(model_3)$r.squa
names(summary_2)[5:6] <- c("Adj R^2", "R^2")
summary_3
```