Regression Models Course Project

Charles Yoo

June 21, 2016

Executive Summary

Using the mtcars dataset, the relationship between miles per gallon (MPG) and a set of variables. Exploratory data analysis and regression analysis will be used to explore how automatic and manual tramissions affect MPG. The difference between automatic and manual transmissions quantifying MPG.

Cars with manual transmission hold a higher MPG than those with automatic transmissions.

Refer to Appendex for extended output and plots.

Data Processing

Load the dataset

```
data(mtcars)
```

Summary of Data

```
head(mtcars)
```

Exploratory Data Analysis

Looking at MPG

```
summary(mtcars$mpg)
```

Boxplot of MPG vs Transmission Type

```
boxplot(mpg ~ am, data = mtcars,
        xlab = "Transmission Type (0 = Automatic, 1 = Manual",
        ylab = "MPG",
        main = "MPG vs Transmission Type")
```

We see that the manual transmission has a higher MPG than automatic.

Statistical Inference

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

P-value is basically 0, so significant. The differences between the transmissions are significant.

Regression Analysis

Correlation

```r
cor(mtcars)[1,]
```

Based on correlations, we see very strong correlations with variables

`cyl, disp, hp, wt, drat, vs, am`

Simple Linear Regression Model

Base Model

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

```r
summary(baseModel)
```

A decent p-value, but have a 33.9% Adjusted R-Squared, indicating not sufficient to explain the variance around MPG.

Multivariable Regression Model

Model 1
Looking at Everything

```r
initialModel <- lm(mpg ~ ., data = mtcars)
# step to find a better model
betterModel <- step(initialModel, direction = "both")
```

Compare Models

```r
anova(baseModel, betterModel)
summary(initialModel)
summary(betterModel)
```

Including all variables, our Adjusted R-squared rises to 80.7%, and a superior p-value. Can we do better?

Model 2 Since maybe a relationship between am and wt

```r
amwtModel <- lm(mpg ~  wt + qsec + am + wt:am, data = mtcars)
```

Look at new model

```r
summary(amwtModel)
```

Our best model so far, with the best p-value and Adjusted R-Squared of 88%

Plot New Model

```r
par(mfrow = c(2, 2))
plot(amwtModel)
```

Check all models

```
anova(baseModel, betterModel, amwtModel)
```

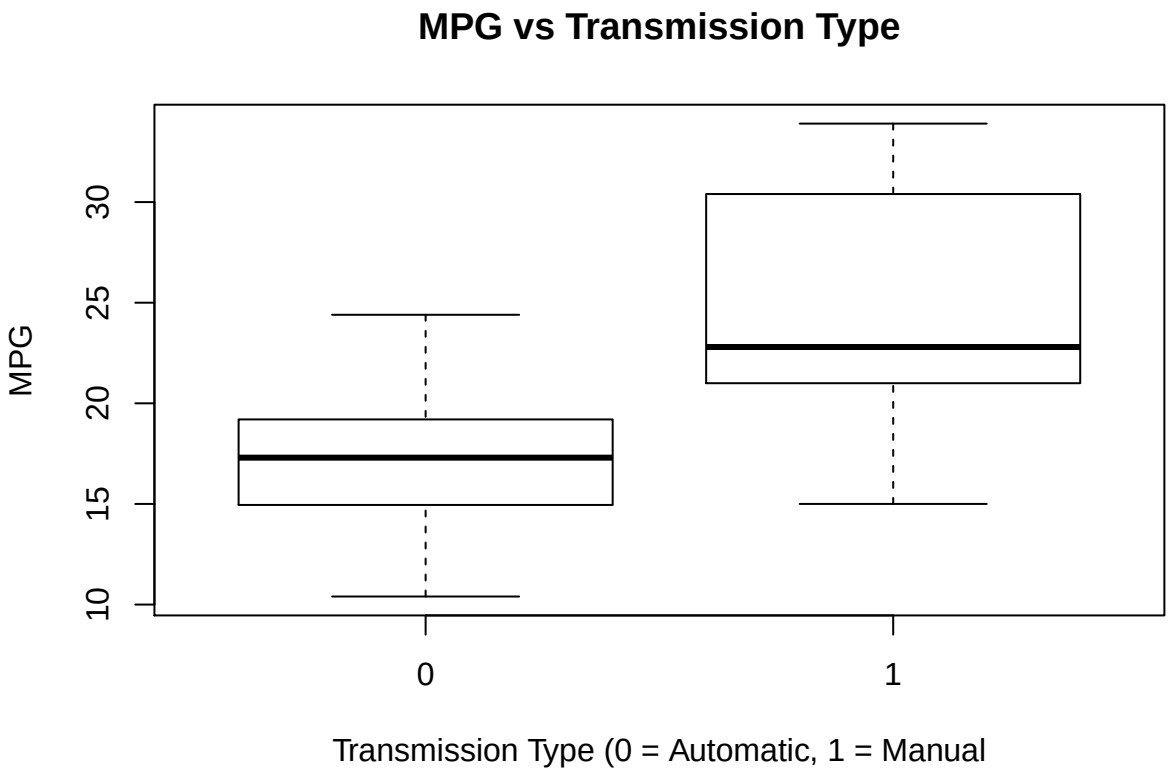
Conclusion

The results show that manual transmission has a greater positive influence on MPG than automatic.

Appendix

Boxplot of MPG vs Transmission Type

```
boxplot(mpg ~ am, data = mtcars,
        xlab = "Transmission Type (0 = Automatic, 1 = Manual",
        ylab = "MPG",
        main = "MPG vs Transmission Type")
```



Statistical Inference

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

```
##
##  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 0 mean in group 1
##        17.14737        24.39231
```

Compare BaseModel to BetterModel

```
anova(baseModel, betterModel)
summary(baseModel)
summary(initialModel)
summary(betterModel)
```

Other Summary Tools

```
# description of variables
 str(mtcars)

# summary of dataset
 summary(mtcars)

# count number of rows
 nrow(mtcars)
 dim(mtcars)

# exploration of data
 unique(mtcars)
```

Step Model of Everything

```
bestModel <- step(initialModel, direction = "both")
```

Pair Graph of Motor Trend Car Road Tests

```
pairs(mtcars, panel=panel.smooth, main="Pair Graph of Motor Trend Car Road Tests")
```

Plots and Other Unsuccessful Models

Based on initial model and coefficients

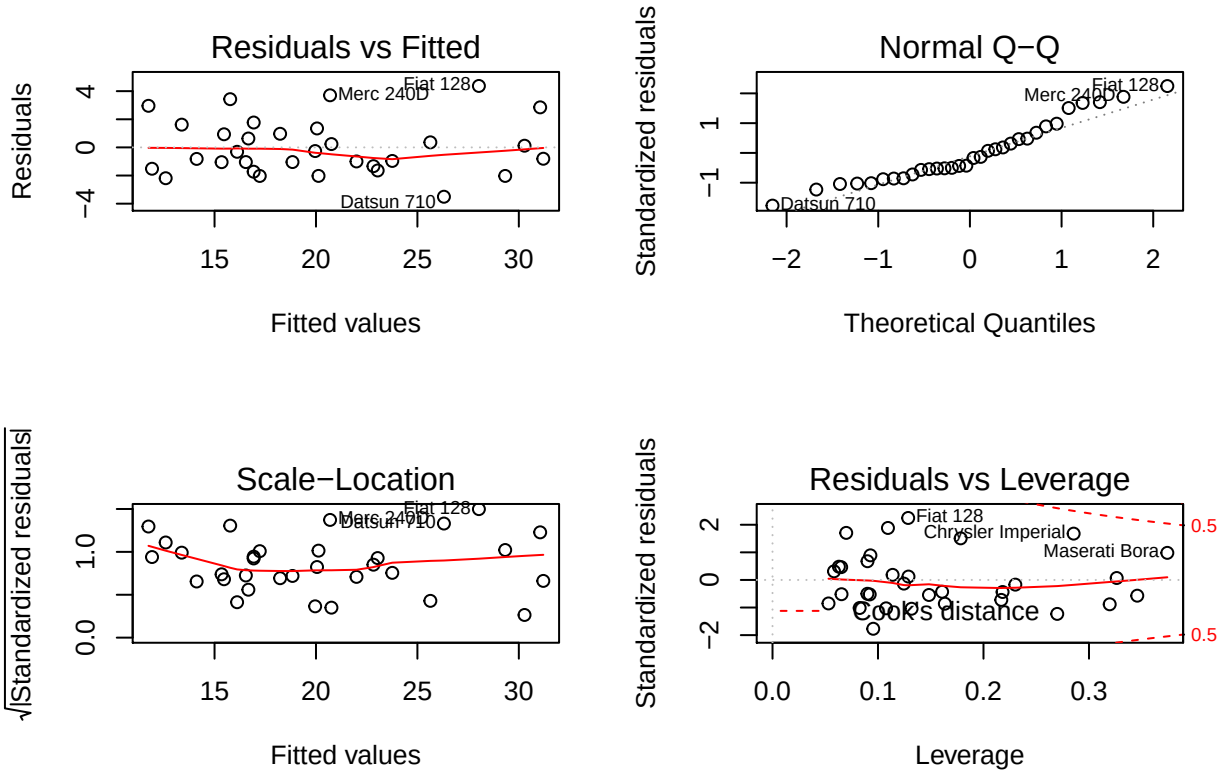
Model 2 Plot New Model

```
amwtModel <- lm(mpg ~  wt + qsec + am + wt:am, data = mtcars)
summary(amwtModel)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am + wt:am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.723      5.899   1.648 0.110893
## wt            -2.937      0.666  -4.409 0.000149 ***
## qsec           1.017      0.252   4.035 0.000403 ***
## am            14.079      3.435   4.099 0.000341 ***
## wt:am         -4.141      1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

```
par(mfrow = c(2, 2))
plot(amwtModel)
```



Model 3

```
newModel <- lm(mpg ~ cyl + wt + vs + am, data = mtcars)
#summary(newModel)
#par(mfrow = c(2, 2))
#plot(newModel)
```

Model 4

```
new2Model <- lm(mpg ~  wt + qsec + am, data = mtcars)
#summary(new2Model)
```

```
#par(mfrow = c(2, 2))
#plot(new2Model)
```

Model 5

```
improvedModel <- lm(mpg ~ cyl + disp + wt + hp + am, data = mtcars)
#summary(improvedModel)
#par(mfrow = c(2, 2))
#plot(improvedModel)
```

Factor variables

```
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am)
```

Scatter Plot of MPG vs Weight by Transmission

```
#ggplot(mtcars, aes(x = wt, y = mpg, group = am, color = am)) +
#        geom_point() +
#        scale_colour_discrete(labels = c("Automatic", "Manual")) +
#        xlab("weight") +
#        ggtitle("Scatter Plot of MPG vs Weight by Transmission")
```

References

https://rstudio-pubs-static.s3.amazonaws.com/29778_09382a14e5294dce9c8758b5be04968c.html
http://rodrigodealexandre.github.io/DS-Courses/Regression%20Models/Data/Course%20Project.html
https://github.com/Xiaodan/Coursera-Regression-Models/blob/master/motor_trend_project/report.Rmd