Final Project

Ernest Jum

11/24/2016

Executive Summary

In this project, we shall use the mtcars data set to address the following two questions:

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

Our analysis focuses on inference with a simple linear regression model and a multiple regression model. We shall observe that both models will support the conclusion that the cars in this study with manual transmissions have on average significantly higher MPG's than cars with automatic transmissions. In the simple model (**starting_model**), the mean MPG difference is 7.245 MPG; the average MPG for cars with automatic transmissions is 17.147 MPG, and the average MPG for cars with manual transmissions is 24.392 MPG. In the multiple regression model, the MPG difference is 2.9358 MPG at the mean weight and qsec. This conclusion holds whether we consider the relationship between MPG and transmission type alone or transmission type together with 2 other predictors: wt / weight; and qsec / 1/4 mile time.Thus, there is a statistically significant difference between the mean and median MPG for automatic and manual transmission cars.

Data Processing

Here we load the mtcars data set and then transform categorical variables to factors using the factor function.

```r
mydata<-mtcars
mydata$cyl <- factor(mydata$cyl)
mydata$vs <- factor(mydata$vs)
mydata$gear <- factor(mydata$gear)
mydata$carb <- factor(mydata$carb)
mydata$am <- factor(mydata$am,labels=c('Automatic','Manual'))
```

Exploratory Data Analysis

Here, we examine the relationships between different variables in the mtcars data set using the plots presented in Appendix 1. One observes that the seven variables (cyl, disp, hp, drat, wt, vs, and am) are correlated with MPG.This relationship cannot be fully quantified unless we use linear models or we examine each one at a time.

We also observed from the box and density plots in Appendix 2 that MPG is higher when the car transmission is manual. The box and density plots gives an indicaton that the effects of transmission on MPG but we need to use the regression analysis to quantify and verify this conclusion.

```r
str(mydata)
summary(mydata)
```

```
cdat <- ddply(mydata, "am", summarise, mpg.mean=mean(mpg))
cdat
```

```
##          am mpg.mean
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

Inference

Here, we test wheter the two types of transmissions (automatic and manual) have equal means.This is accomplished via the t-test. The test shows that the distribution of MPG is significantly different for manual and authomatic transmission cars with a p-value of 0.001374.

```
t.test(mydata[mydata$am == "Automatic",]$mpg, mydata[mydata$am == "Manual",]$mpg)
```

Model Building and Selection

In this section, we build different regression models using with different variables in the model and then find the best model fit. Next, I perform analysis of residuals. In order to build our model, we commence by building a model with all the variables as the predictors of MPG. The final model (final_model) with significant predictors is obtained using AIC with a stepwise model selection. Thus, only variables that are significant in predicting MPG are included in the final model.

```
starting_model <- lm(mpg ~ ., data = mtcars)
final_model <- step(starting_model, direction = "both")
```

Here is a summary of the simple linear regression invloving *MPG* and *am*. We shall note here that the mean MPG difference is 7.245 MPG; the average MPG for cars with automatic transmissions is 17.147 MPG, and the average MPG for cars with manual transmissions is 24.392 MPG.

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

```
##
## 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 ***
## am             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
```

Below we examine the summary of the selected final model with the most signinficant paramters.

```
summary(final_model)
```

Here, we compare this model with the model with only *am* as the independent variable.

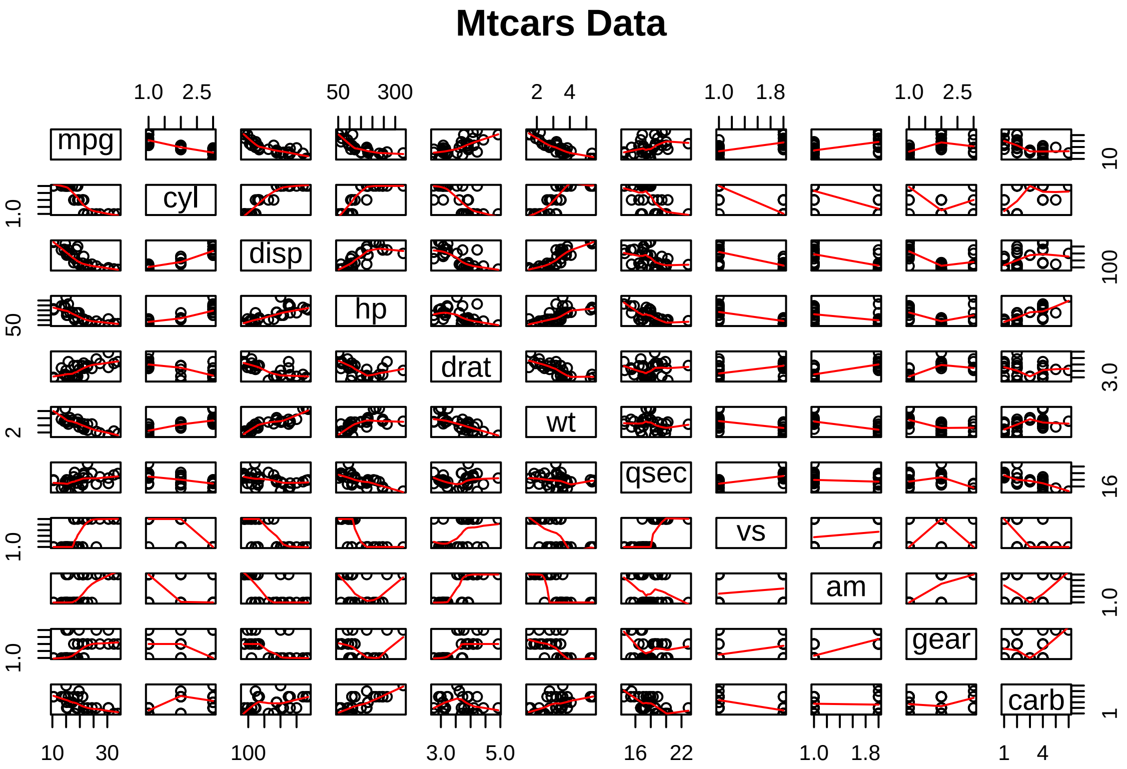
```
anova(base_model, final_model)
final_model$anova
```

Since the p-value of the above anova test is significant, we shall reject the null hypothesis and conclude that the three variables cyl, wt, and hp contribute significantly to the linear model

Appendices

Appendix 1: Correlations

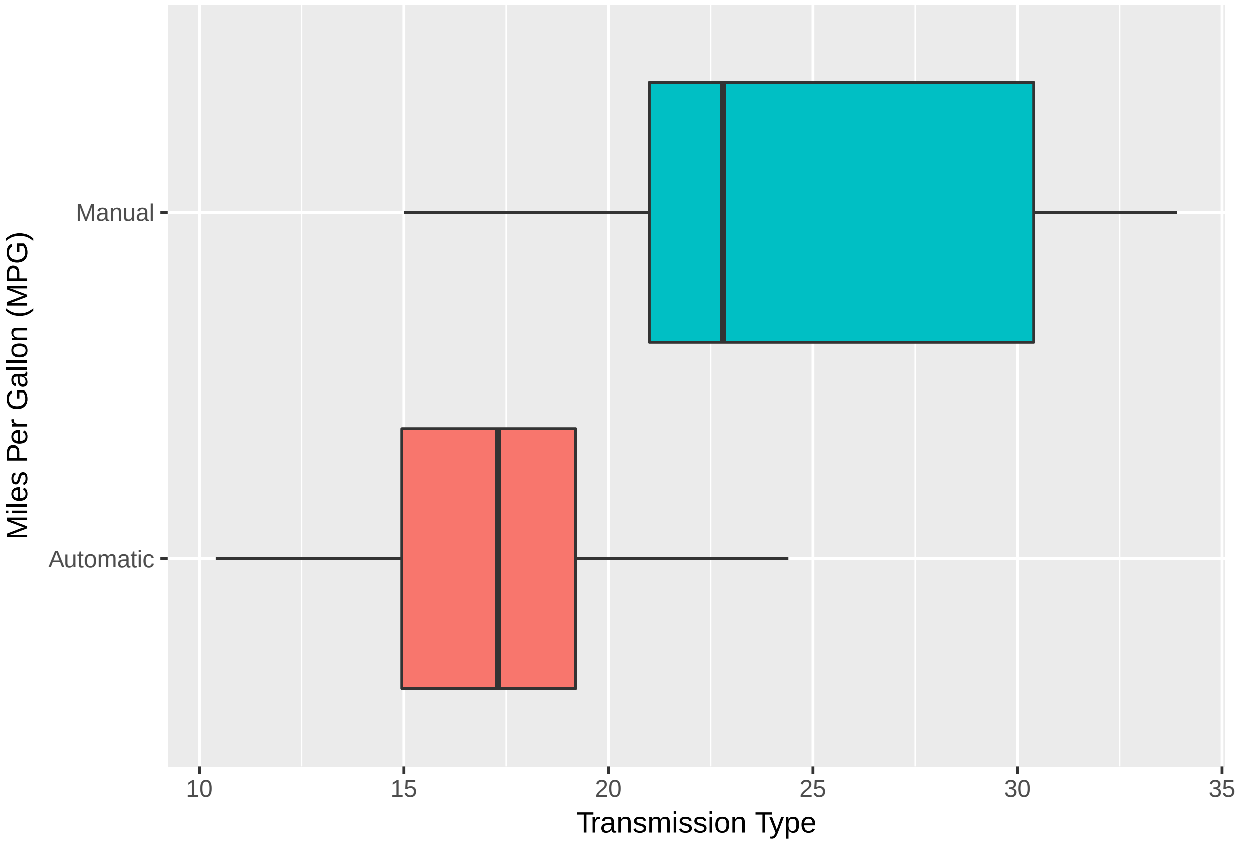
```
pairs(mydata, panel = panel.smooth, main="Mtcars Data")
```



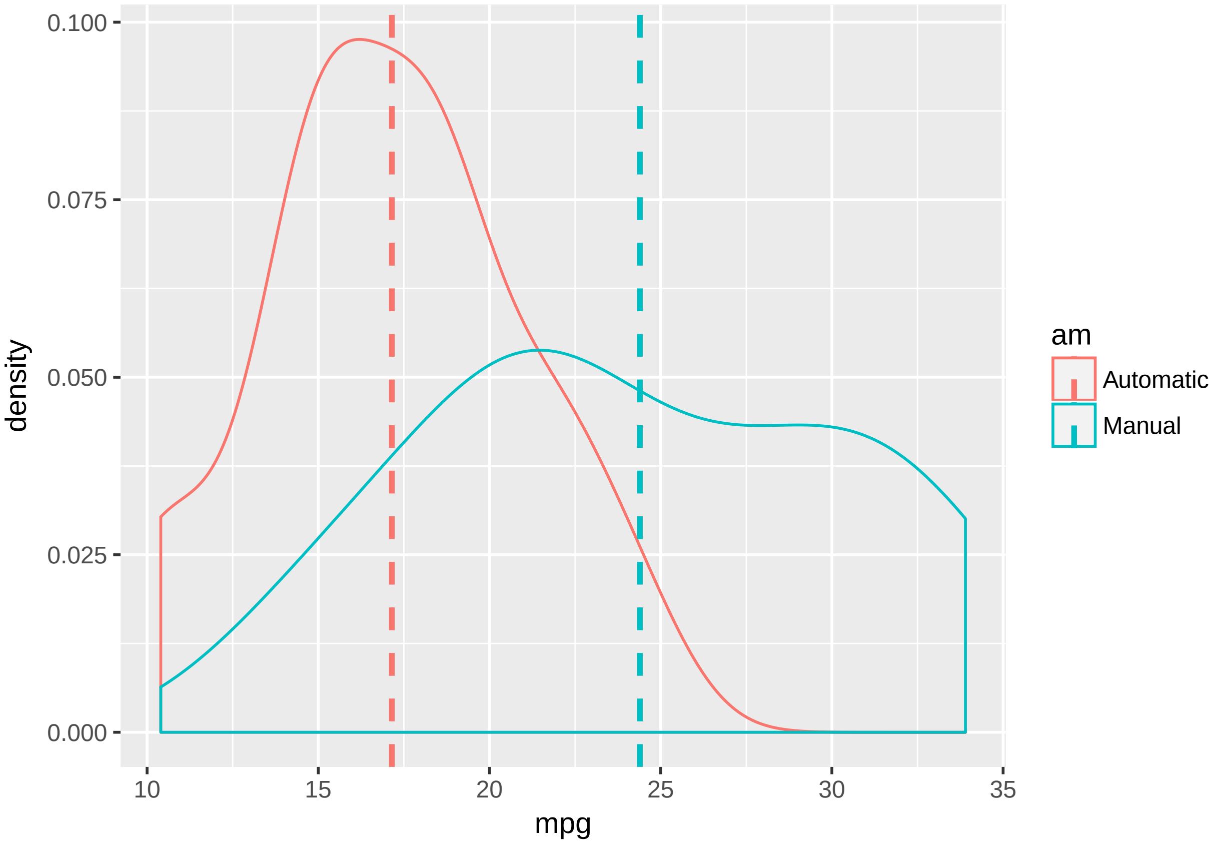
Appendix 2: Box and Density Plot

```
g<-ggplot(data=mydata, aes(x=am, y=mpg, fill=am))
g<-g+ geom_boxplot()
g<-g+ guides(fill=FALSE) + coord_flip()
```

```
g<-g+xlab("Miles Per Gallon (MPG)")
g<-g+ylab("Transmission Type")
g
```

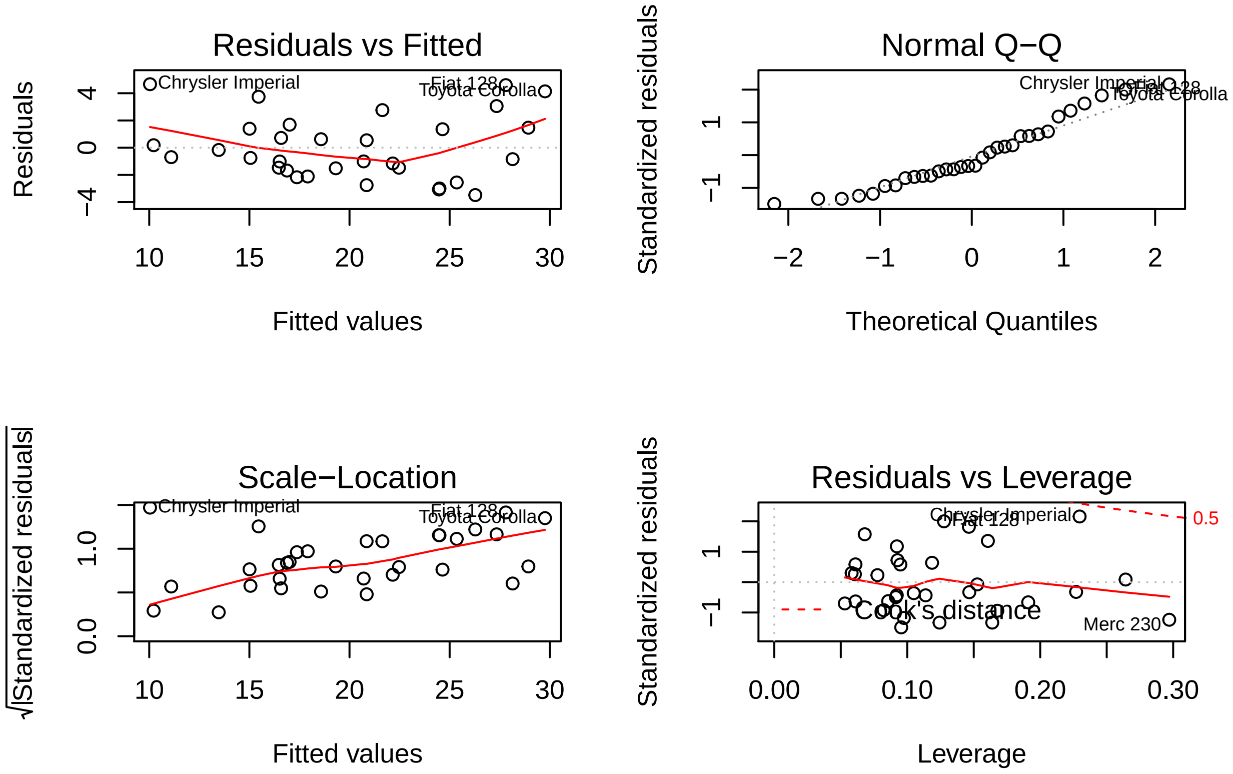


```
# Density plots with means
g<-ggplot(data=mydata, aes(x=mpg, colour=am)) +geom_density()
g<-g+geom_vline(data=cdat, aes(xintercept=mpg.mean,  colour=am),
             linetype="dashed", size=1)
g
```



Appendix 3: Model Dignostics

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



```
outlier <- hatvalues(final_model)
tail(sort(outlier),3)
```

```
##   Chrysler Imperial Lincoln Continental            Merc 230
##           0.2296338           0.2642151           0.2970422
```

```
influential <- dfbetas(final_model)
tail(sort(influential[,3]),3)
```

```
## Chrysler Imperial    Toyota Corolla          Fiat 128
##         0.3366781         0.4514928         0.4968861
```