Regression Models

Aaditya Bhat

July 21, 2015

## Executive Summary

The goal of this analysis is to explore the relationship between a set of variables such as number of cylinders, displacement, gross horsepower, etc and miles per gallon (MPG). For the analysis 'Motor Trend Car Road Tests' dataset in R is used. This data was extracted from 1974 Motor Trend US magazine. The analysis answers the following two questions:

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

## Data processing

Loading the required packages

library(ggplot2)

Loading the 'mtcars' data set. Coverting appropriate variables into factors.

data("mtcars")  
mtcars$cyl <- factor(mtcars$cyl)  
mtcars$vs <- factor(mtcars$vs, labels = c('V-Engine', 'Straight Engine'))  
mtcars$gear <- factor(mtcars$gear)  
mtcars$carb <- factor(mtcars$carb)  
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))

## Exploratory Analysis

Refer apendix, to see the exploratory plots.

## Hypothesis Testing.

Checking the variance of both the samples.

var(mtcars$mpg[mtcars$am == 'Automatic'])

## [1] 14.6993

var(mtcars$mpg[mtcars$am == 'Manual'])

## [1] 38.02577

As the variance isn't equal, performing Welch's t test.

t.test(mtcars$mpg~mtcars$am,conf.level=0.95)

The summary of the Welch's t test is in the appendix. As p-value < 0.05, we reject the Null Hypothesis that mean MPG is same for both the transmission types.

## Fitting Models

Fitting a linear model with 'mpg' as the response and 'am' as the regressor.

model1 <- lm(mpg~am, data = mtcars)  
summary(model1)

The summary for this model can be found in the appendix. As R-Squarred value is 0.3598, this model only accounts for 36% variablity in mpg. Hence this models isn't a good fit. A linear model with 'mpg' as the response and all the remaining variables as the regressors will result in overfitting. Hence we obtain the best model by backward selection, using 'step' function in r.

model2 <- lm(mpg~., data = mtcars)  
best\_model <- step(model2, direction = "backward")

summary(best\_model)

##   
## Call:  
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.9387 -1.2560 -0.4013 1.1253 5.0513   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 33.70832 2.60489 12.940 7.73e-13 \*\*\*  
## cyl6 -3.03134 1.40728 -2.154 0.04068 \*   
## cyl8 -2.16368 2.28425 -0.947 0.35225   
## hp -0.03211 0.01369 -2.345 0.02693 \*   
## wt -2.49683 0.88559 -2.819 0.00908 \*\*   
## amManual 1.80921 1.39630 1.296 0.20646   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.41 on 26 degrees of freedom  
## Multiple R-squared: 0.8659, Adjusted R-squared: 0.8401   
## F-statistic: 33.57 on 5 and 26 DF, p-value: 1.506e-10

As the R-squarred value is 0.8659, the model accounts for 86.6% variablity in the mpg. This model is a good fit for the data. Various diagnostic plots are included in appendix.

## Conclusion

According to the Welch's test and regression model, Manual Transmittion is better for mpg. Manual Transmission results in an increase of 1.8092 in mpg, keeping other variables constant. However this relation isn't very significant.

## Apendix

### Welch's t test summary

##   
## Welch Two Sample t-test  
##   
## data: mtcars$mpg by mtcars$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

### Model 1 summary

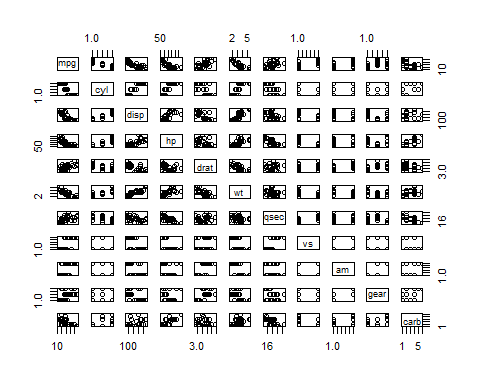
model1 <- lm(mpg~am, data = mtcars)  
summary(model1)

##   
## 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

### Exploratory Analysis

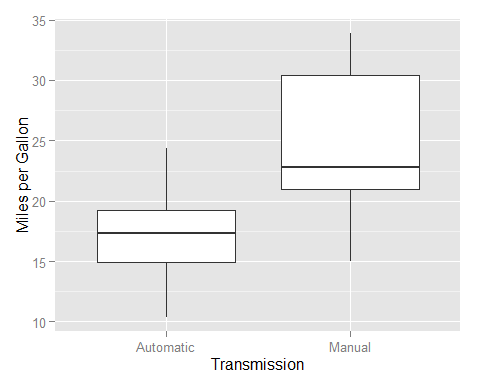
Mtcars pair plot

pairs(mtcars)



box plot of mpg ~ am

g <- ggplot(data = mtcars, aes(am,mpg))  
g+geom\_boxplot()+labs(x = 'Transmission', y = 'Miles per Gallon')



### Residual and Diagnostic Plots

par(mfrow=c(2, 2))  
plot(best\_model)

