Regression Models Course Project

J. Thatcher

Tuesday, January 19, 2016

# 1.0 Introduction

In this project we consider the mtcars dataset. We will explore the relationship between transmission type (automatic or manual) and miles per gallon (MPG).

# 2.0 Data Analysis

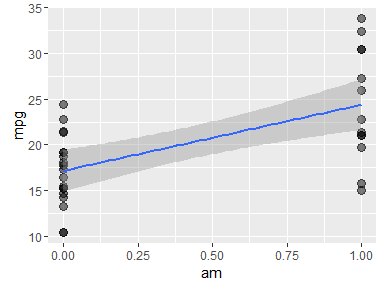
### 2.1 Exploratory Data Analysis

In data(mtcars) we find that the first column variable is mpg (Miles/US Gallon) and the nineth column variable am is transmission type (0 = automatic, 1 = manual)

library(ggplot2)  
DF <- data.frame(mtcars) #grab data as data.frame

Fist, an exploratoty graph of the response and predictor using ggplot2.

#plot the data and a linear fit  
ggplot(DF, aes(x = am, y = mpg)) +   
 geom\_point(alpha = 0.5, cex = 3) +   
 geom\_smooth(method = "lm")



The graph shows that there may be a significant correlation between mpg and transmission type.

### 2.2 Relationship Between MPG and Transmission Type

Initially we will ask: is MPG better for manual or automatic transmission vehicles?

let's explore this using a linear model. First, we will calculate the coefficients of a linear model using the lm function.

#fit a linear model  
fit <- lm(mpg ~ am, data = DF)  
summary(fit)

##   
## Call:  
## lm(formula = mpg ~ am, data = DF)  
##   
## 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

### 2.3 Interpretation of Correlation Coefficient, R-Squared

Reviewing the calculations performed by the lm funciton, we can find many imporant values including the correlation coefficient. The corelation coefficeint is a measure of the variance in vehicle MPG that is explained by the variable transmission type. The R-squared value for the model is 0.36. Which means that transmission type explains 36% of the variance in vehicle MPG.

### 2.4 Interpretation of the Intercept, Beta0

Following these calculations, the y-intercept Beta0 = 17.15. This values is the mean mpg for the automatic transmission vehicles.

### 2.5 Interpretation of the Slope, Beta1

Reviewing the slope, Beta1, in the model will answer our initial quesiton, is MPG better for manual or automatic transmission vehicles? Beta1 is 7.25, and is interpreted as the increase in the average vehicle's MPG for one unit increase in the predictor variable, transmission. In this case transmission is coded as 0 = automatic and 1 = manual. Therefore, those vehicles that have an automatic transmission have an change in mpg equal to beta1 \* 1 or simply Beta1. Beta1 is positive, which means that there is an increase in the response variable mpg for an increase in the predictor variable, transmission type.

The quantified the difference in MPG between automatic and manual transmission vehicles is 7.25 mpg. In other words, there is a 7.25 mpg increase in the average vehicle's MPG for those vehicles that have a manual transmission.

### 2.6 Is the slope significant?

We cannot assume this 7.25 mpg increase in mpg for manual transmission vehicles is not different than 0. We can perform a hypothesis test on the Beta1 coefficient to determine that the slope is actually different than zero. The hypotheses are:

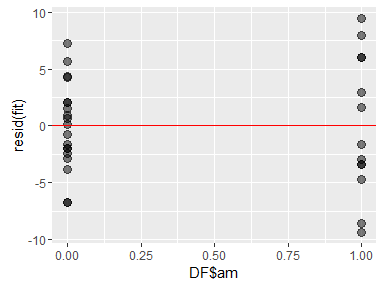
H0: Beta1 = 0  
H1: Beta1 != 0

To determine whether we can reject the null hypothesis, H0, we will go back to our results from the lm function. In the second row and fourth column we find the probability that the the slope is equal to zero equals 0.000285. This is much lower than our level of statistical significance, alpha = 0.05. Therefore, we can reject H0 and determine that Beta1 is significantly different that zero.

### 2.7 Residuals

An important way to determine the effectiveness of the linear model is to plot the residuals. In a residual plot, we are looking for the residual points to be spread in a random fashion on either side of the horizantle line. This will help confirm that the linear model was an appropriate model for this data.

#plot the residuals  
ggplot(fit, aes(x = DF$am, y = resid(fit))) +   
 geom\_point(alpha = 0.5, cex = 3) +   
 geom\_hline(yintercept=0, color="red")



The plot does show that the residuals are spread evenly above and below the line at 0. However, the graph shows that the data slightly resembles a property called Heteroscedasticity. This is when the spread of the data differs at different levels of the predictor variable. In this case the variance of mpg appears to be slightly higher in cars with manual transmissions compared to automatic transmissions.

### 2.8 Alternative Models

Upon reviewing the residual plot, we will consider a few other models that may show better performance compared to the linear model.

# 3.0 Executive Summary

The results of this analysis answer the two main questions asked. One, "Is an automatic or manual transmission better for MPG?" Answer, the manual transmission is on average better for MPG. Two, "Quantify the MPG difference between automatic and manual transmissions." Answer: using a linear regression model we find that there is an average increase of 7.25 mpg for cars with a manual transmission.