Regression Models Course Project

jacethedatascientist

August 13, 2019

Regression Models Course Project

Jace Galleon

I. Project Overview

In this project, we will work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome).

We will examine the mtcars data set and explore how miles per gallon (MPG) is affected by different variables. In particularly, we will answer the following two questions: (1) Is an automatic or manual transmission better for MPG, and (2) Quantify the MPG difference between automatic and manual transmissions.

II. Project Objective

Motor Trend is particularly interested in the following two questions: 1. "Is an automatic or manual transmission better for MPG" 2. "Quantify the MPG difference between automatic and manual transmissions"

III. Project Analysis

A. Data Preparation

We will use the **mtcars** dataset installed in the R Package and will take a sample.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
data(mtcars)
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

There are other variables that are numerical in nature but are actually entered as *"types"* so we will convert them to factors.

```
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,labels=c("Automatic","Manual"))
```

B. Data Exploration

*Based on **Figure F-1**, there is an obvious difference between the impact of each Transmission Types to MPG. It can be seen that Automatic Transmission has a better impact compared to Manual.*

```
aggregate(mpg~am, data = mtcars, mean)
```

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

On hypothesis, there is an average of 7.245 difference in the MPG between Manual and Automatic Transmission. But, to determine if there is a significant difference, we'll be using T-Test.

Null : *There is no significant difference between the mean of MPG for both Transmission Types.*
Alternative : *There is a significant difference between the mean of MPG for both Transmission Types.*

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg[mtcars$am == "Automatic"] and mtcars$mpg[mtcars$am == "Manual"]
## 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 of x mean of y
##  17.14737  24.39231
```

Based on the P-value (0.001374), there is a significant difference between the mean of MPG for both Transmission Types.

C. Data Analysis

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

```
##
## 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
```

This summary shows us that the average MPG for cars with AT has an **average of 17.147** while for Manual is **7.245** higher.

If you look at the value of the R^2, the value is **0.36**, which means that this accounts to **36% of the variation** in the MPG. To get a more accurate result, we'll be doing a multi-variate regression that will use the rest of the variables.

```
multi_var <- lm(mpg~am + cyl + disp + hp + wt, data = mtcars)
anova(linmod,multi_var)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + disp + hp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     25 150.41  5    570.49 18.965 8.637e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on this, we can infer that **cyl**, **disp**, **hp**, **wt** have a stronger correlation with mpg than **am**. We, then, build a new model using these variables and compare them to the initial model with the anova function to determine if there's a significant difference.

The P-value is **8.637e-08** and is almost 0, thus, we can infer that this new model is **better** than our initial (*linmod*).

Double-check the residuals for *non-normality* (see Figure F-2) and we can see that they are all *normally distributed*.

D. Assumptions and Conclusions

```
summary(multi_var)
```

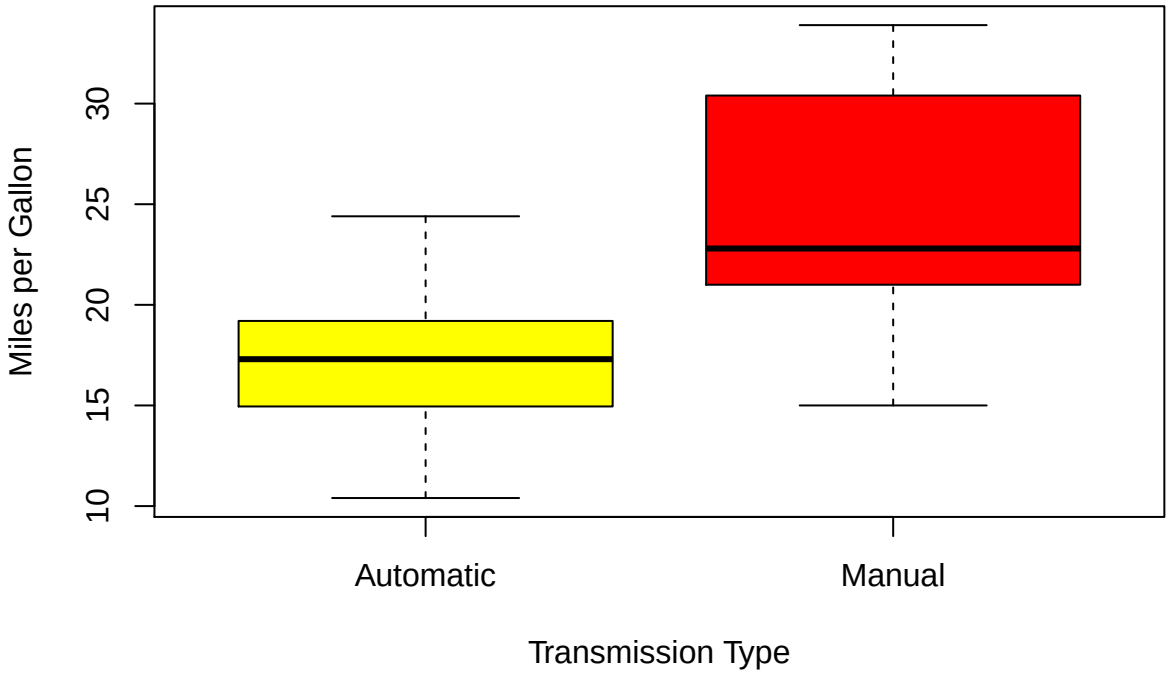
```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.864276   2.695416  12.564 2.67e-12 ***
## amManual     1.806099   1.421079   1.271   0.2155
## cyl6        -3.136067   1.469090  -2.135   0.0428 *
## cyl8        -2.717781   2.898149  -0.938   0.3573
## disp         0.004088   0.012767   0.320   0.7515
## hp          -0.032480   0.013983  -2.323   0.0286 *
## wt          -2.738695   1.175978  -2.329   0.0282 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

The model accounts to **86.64%** of the variance and as a result, **cyl**, **disp**, **hp**, and **wt** did affect the correlation between mpg and am by **roughly 51%**. Thus, we can say that the difference between automatic and manual transmissions is 1.81 MPG with the manual being higher.

This formally ends the Course Project. Thank You!

E. Appendices

1. Boxplot for the impact of the Transmission Type (am) to the MPG.



2. Plot of the latest model (multi_var) to show the distribution.