Linear Regression course project

jbassard

27 july 2017

Executive Summary

This is the course project for the statistical inference class on coursera. In this report, we will examine the "mtcars" dataset and explore how miles per gallon (MPG) is affected by automatic or manual transmission. In particularly, we have to answer the two questions:

+ Is an automatic or manual transmission better for MPG?

+ Quantify the MPG difference between automatic and manual transmissions.

Loading necessary packages and dataset

```
if (!require("knitr")) {
    install.packages("knitr")}

## Loading required package: knitr

data(mtcars) #Loading the dataset
library(knitr)
```

Setting the default of echo and cache to be True throughout the whole report

```
knitr::opts_chunk$set(echo = TRUE) #Make the code always visible
knitr::opts_chunk$set(cache=TRUE)
```

Data Cleaning and Exploration

```
dim(mtcars)

## [1] 32 11

str(mtcars)

## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

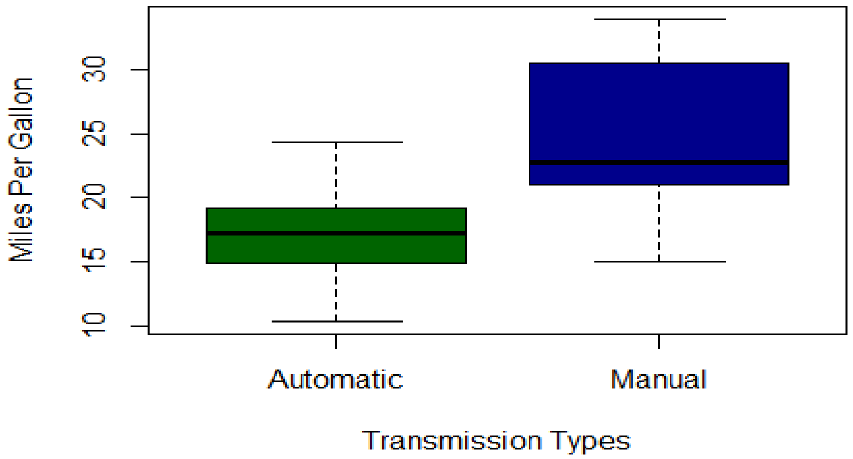
```
head(mtcars, 5)
```

```
##                    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
```

```
mtcars$am <- factor(mtcars$am, labels=c("Automatic","Manual")) # for better
readibility, O and 1 are renamed: 0=automatic, 1=manual
```

Boxplot of MPG vs. Transmission.

```
boxplot(mpg ~ am, data = mtcars, col = (c("dark green","dark blue")), ylab =
"Miles Per Gallon", xlab = "Transmission Types") #Display a quick and dirty
graph for the data
```



From the boxplot, it seems there is an **impact of the transmission type on MPG** with Automatic transmission having a lower MPG.

Inference

We set the null hypothesis as the MPG of the Automatic and Manual transmissions has no difference. We use the two sample T-test to test the hypothesis.

```
inference <- t.test(mpg ~ am, data=mtcars)
inference$p.value
```

```
## [1] 0.001373638
```

```
inference$estimate
```

```
## mean in group Automatic    mean in group Manual 
##                17.14737                24.39231
```

Since the p-value is 0.0014 which is less than 0.05, we reject our null hypothesis. We can say that there is a **significant difference in MPG between the 2 transmission types**. The mean for MPG of manual transmitted cars and automatic cars are 24.39 and 17.18, respectively (difference of 7.245 miles per gallon).

Linear regression models

Simple linear regression model

Now to quantify the difference confirmed by the T.test, we will do a linear regression. We will use MPG as the dependent variable and AM as the independent variable to fit a linear regression.

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

```
## 
## 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
```

Since the p-value = 0.000285 is less than 0.05 so we rejected null hypothesis. We have a significant difference. Looking at the coefficien, we see again **the difference of 7.245 miles per gallon at the advantage of the manual cars**. But the adjusted R squared value is 0.3385 which means our model only explains 33.85% of the variance. Thus, we need to include other predictors in the linear regression, we need to build a multivariate linear regression.

Multivariate linear regression model

The new model will use other variables to make it more accurate. We run an analysis of variance to find the variables to add.

```
AOV <- aov(mpg ~ ., data = mtcars)
summary(AOV)

##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          1  817.7   817.7 116.425 5.03e-10 ***
## disp         1   37.6    37.6   5.353  0.03091 *
## hp           1    9.4     9.4   1.334  0.26103
## drat         1   16.5    16.5   2.345  0.14064
## wt           1   77.5    77.5  11.031  0.00324 **
## qsec         1    3.9     3.9   0.562  0.46166
## vs           1    0.1     0.1   0.018  0.89317
## am           1   14.5    14.5   2.061  0.16586
## gear         1    1.0     1.0   0.138  0.71365
## carb         1    0.4     0.4   0.058  0.81218
## Residuals   21  147.5     7.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above Analysis of Variance, looking for p-values of less than .5, gives us cyl, disp, and wt to consider in our model in addition to transmission type (am).

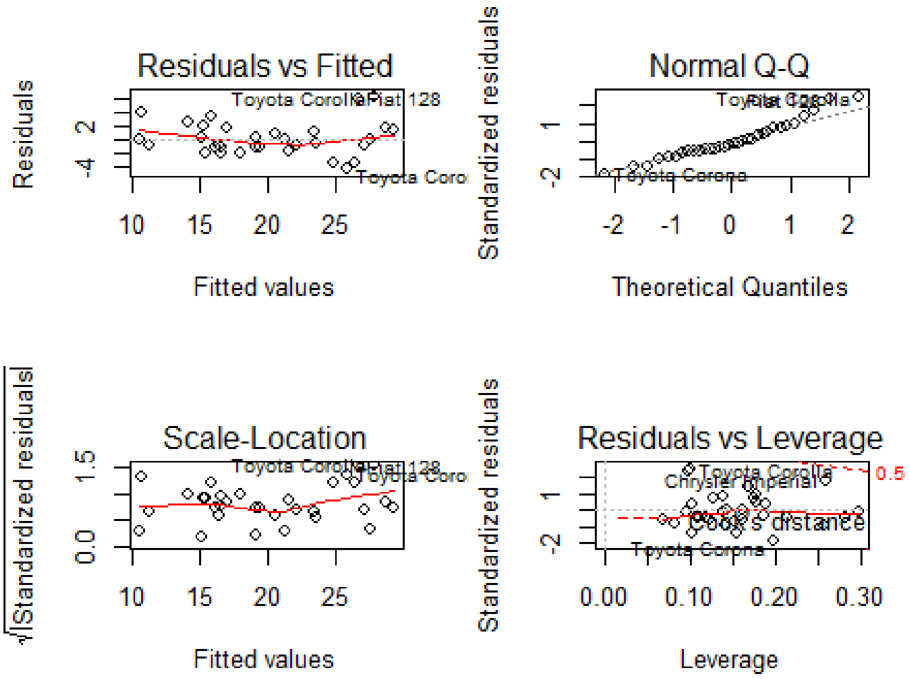
```
multivar <- lm(mpg ~ am + cyl + disp + wt, data = mtcars)
summary(multivar)

##
## Call:
## lm(formula = mpg ~ am + cyl + disp + wt, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.318 -1.362 -0.479  1.354  6.059
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.898313   3.601540  11.356 8.68e-12 ***
## amManual     0.129066   1.321512   0.098  0.92292
## cyl         -1.784173   0.618192  -2.886  0.00758 **
## disp         0.007404   0.012081   0.613  0.54509
## wt          -3.583425   1.186504  -3.020  0.00547 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.642 on 27 degrees of freedom
## Multiple R-squared:  0.8327, Adjusted R-squared:  0.8079
## F-statistic: 33.59 on 4 and 27 DF,  p-value: 4.038e-10
```

This Multivariate Regression Model now gives us an R-squared value of 0.8079, suggesting that 81% of variance can be explained by this multivariate model. P-values for cyl and wt are below 0.5, clearly showing that these are confounding variables in the relation between car Transmission Type and Miles Per Gallon. Looking at estimates, we can say the **difference between automatic and manual transmissions is 0.13 MPG**.

Residual Analysis and Diagnostics

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



The "Residuals vs Fitted" plot here shows us that there is no patterns in the residuals from the Multivariate model. We can also see that they are normally distributed (normal Q-Q plot), with the exception of a few outliers.

Conclusion

Is an automatic or manual transmission better for MPG? It appears that manual transmission cars are better for MPG compared to automatic cars. However when modeled with confounding variables like cyl, disp and wt, the difference is not as significant as it seems in the beginning. In fact, a big part of the difference is explained by other variables.

Quantify the MPG difference between automatic and manual transmissions Analysis shows that when only transmission was used in the model, manual cars have an MPG increased of 7.245. However, when other variables are included, the automatic car advantage drops to 0.13.