Motor Trend Analysis

Sam

03/08/2020

I 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). They are 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.

We use the mtcars dataset from the dataset library for analysis.

Our analysis demonstrates the following

1. Manual transmission will yield better miles per gallon, when compared with Automatic. On average, a manual car will achieve 24 mpg, versus 17 mpg for automatics.
2. Further analysis shows a correlation between MPG and the following confounding variables :

   wt (Weight). The greater the weight of the car, the less MPG

   cyl (number of engine cylinders).

Exploratory Data Analysis

At the onset of our analysis we load the required libraries and data . Then we preprocess the data for the analysis.

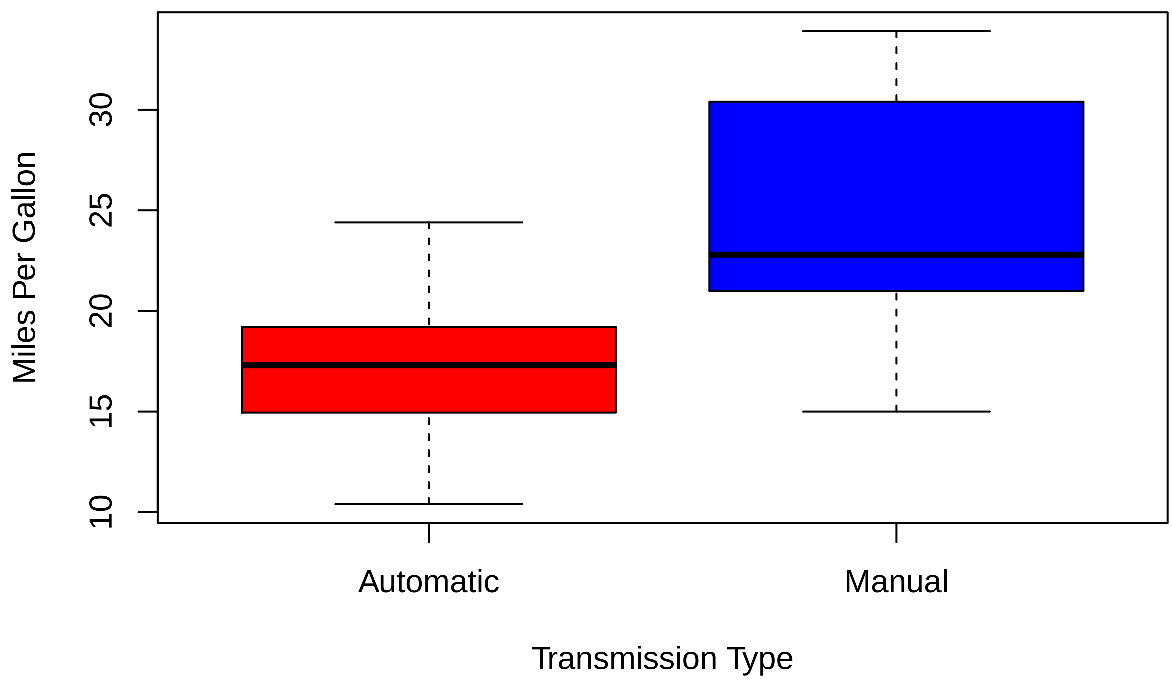
Load the mtcars dataset.

Convert some variables for easy interpretation.

```
library(ggplot2)
data("mtcars")
mtcars$vs <- factor(mtcars$vs)
mtcars$am.label <- factor(mtcars$am, labels=c("Automatic","Manual")) # 0=automatic, 1=manual
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
head(mtcars)
```

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

```r
boxplot(mpg ~ am.label, data = mtcars, col = (c("red","blue")),
        ylab = "Miles Per Gallon", xlab = "Transmission Type")
```



From the above Box plot we see that Manual Transmission provides more MPG than the Automatic Transmission. So let's make sure in the upcoming analysis.

Regression Analysis

Let's calculate the mean MPG for Automatic and Manual Transmissions.

```r
aggregate(mtcars$mpg , by=list(mtcars$am.label), FUN=mean)
```

```
##     Group.1        x
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

Now again we see that Manual transmission yields more MPG than the Automatic Transmission by 7 MPG. Now let's test this hypothesis with the simple Linear Regression test.

```r
T_simple <- lm(mpg ~ factor(am), data=mtcars)
summary(T_simple)
```

```
##
```

```
## Call:
## lm(formula = mpg ~ factor(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 ***
## factor(am)1    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
```

The p-value is less than 0.0003, so we will not reject the hypothesis. However, the R-squared value for this test is only ~= .35, suggesting that only a third or so of variance in MPG can be attributed to transmission type alone. Let's perform an Analysis of Variance for the data:

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

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## cyl          1  817.7   817.7 102.591 2.3e-08 ***
## disp         1   37.6    37.6   4.717 0.04525 *
## hp           1    9.4     9.4   1.176 0.29430
## drat         1   16.5    16.5   2.066 0.16988
## wt           1   77.5    77.5   9.720 0.00663 **
## qsec         1    3.9     3.9   0.495 0.49161
## vs           1    0.1     0.1   0.016 0.90006
## am           1   14.5    14.5   1.816 0.19657
## gear         2    2.3     1.2   0.145 0.86578
## carb         5   19.0     3.8   0.477 0.78789
## Residuals   16  127.5     8.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

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

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + wt + am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

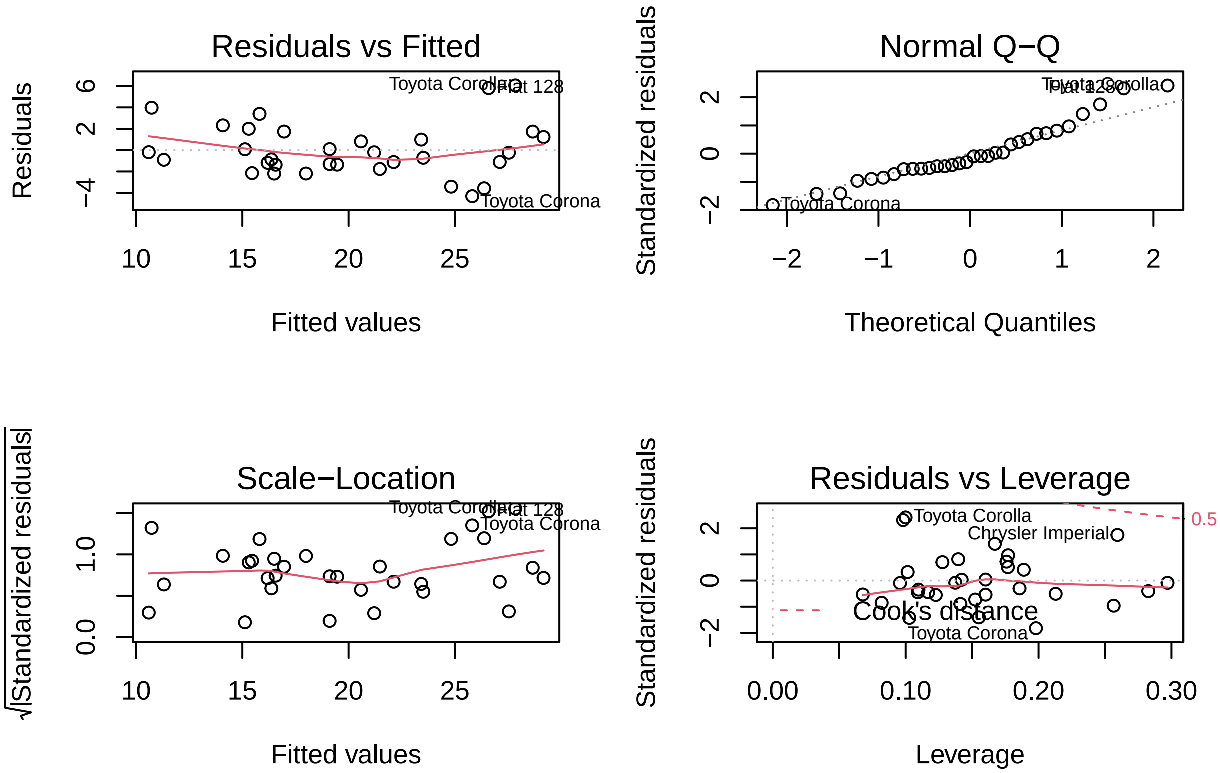
```
## -4.318 -1.362 -0.479  1.354  6.059
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.898313   3.601540  11.356 8.68e-12 ***
## cyl         -1.784173   0.618192  -2.886  0.00758 **
## disp         0.007404   0.012081   0.613  0.54509
## wt          -3.583425   1.186504  -3.020  0.00547 **
## am           0.129066   1.321512   0.098  0.92292
## ---
## 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 Multivariable Regression test now gives us an R-squared value of over .83, suggesting that 83% or more of variance can be explained by the multivariable model.

P-values for cyl (number of cylinders) and weight are below 0.5, suggesting that these are confounding variables in the relation between car Transmission Type and Miles per Gallon.

Residual plot and Analysis

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



The “Residuals vs Fitted” plot here shows us that the residuals are homoscedastic. We can also see that they are normally distributed, with the exception of a few outliers.