Motor Trend Analysis

Philong Ho

3/4/2020

Motor Trend Analysis

Overview

In this analysis, we will be looking at the mtcars dataset in R. Let's pretend we work for Motor Trend, a magazine about the automobile industry. Looking at the collection of cars, we are interested in the relationship between a set of variables and miles per gallon (mpg), our outcome. We are interested in the two questions:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

Our findings show that manual transmission has a larger mean output of mpg than that of automatic transmission. We also find that our linear model comparing mpg to transmission type is not as optimized as a model that compares mpg to transmission type plus other variables such as cylinders, horsepower, and weight included as regressors. This shows that our outcome is correlated with the other variables and can not be left out of our linear models.

Data Summary

Let's load in the data and take a look at it.

```r
library(datasets)
data(mtcars)

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 ...
```

```r
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
```

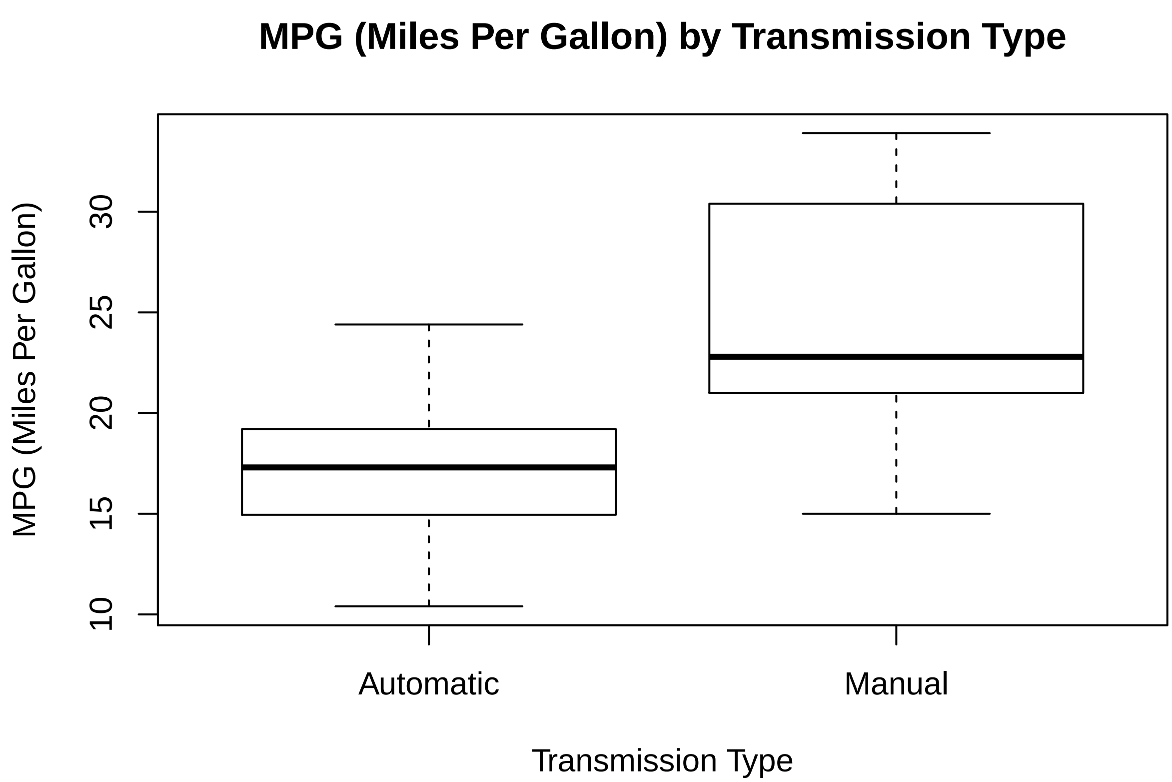
We can see that this is a dataset of 32 observations and 11 variables. We can answer the first question above by first looking at the variables mpg and am (automatic). The variable am is a numeric class which assigns a 1 to automatic and a 0 to manual. We will change the class to a factor and assign 0 to automatic and 1 to manual for simplicity.

```r
#Converting the 'am' variable to a factor with 2 lvls: 'Manual' as 1, and 'Automatic' as 0
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
```

Now, we want to just do a quick analysis and see the mean mpg's of both transmission types.

```r
mean.mpg <- aggregate(mpg ~ am, mtcars, mean)

boxplot(mpg ~ am, data = mtcars, xlab = "Transmission Type", ylab = "MPG (Miles Per Gallon)", main = "MI
```



We can see that manual transmission has a mean mpg of 24.39 and automatic transmission has a mean mpg of 17.15. This is also represented in the boxplot.

Now, we can do a quick hypothesis testing using a T Test.

```
t.test(mtcars$mpg ~ mtcars$am)
```

```
##
##  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
```

Looking at the p-value, we can see that it is less than 0.05 so we reject the null hypothesis that there is no difference in mpg by transmission type. Let's make a linear model to see if there is a good linear relationship between mpg and am.

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

```
##
## 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
```

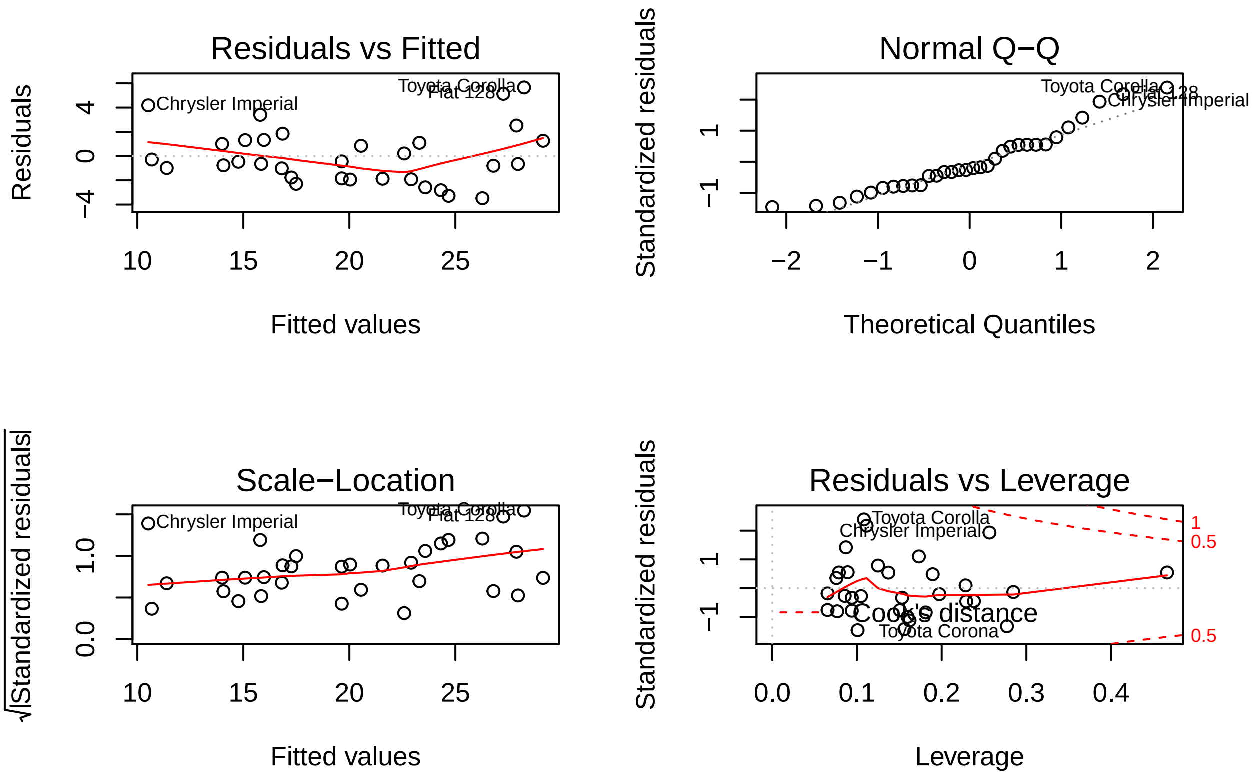
We can see that the coefficients for automatic and manual are the means we calculated before. We can look at R^2 and see that its value is 0.3598 which means this model only explains for about 36% of the variance. Let's add in some more regressors to check for a better model. I chose to use cylinders, horsepower, and weight because they are easily understandable.

```
fit2 <- lm(mpg ~ am + cyl + hp + wt, data = mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4765 -1.8471 -0.5544  1.2758  5.6608
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.14654    3.10478  11.642 4.94e-12 ***
## amManual     1.47805    1.44115   1.026   0.3142
## cyl         -0.74516    0.58279  -1.279   0.2119
## hp          -0.02495    0.01365  -1.828   0.0786 .
## wt          -2.60648    0.91984  -2.834   0.0086 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.509 on 27 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.8267
## F-statistic: 37.96 on 4 and 27 DF,  p-value: 1.025e-10
```

The intercept, which is our mpg for automatic transmission, has a coefficient of 36.17 and our manual transmission is 1.48 above that of automatic. The other variables are negatively correlated. This time, the R^2 value is 0.849 which explains 85% of the variance. This is a much better linear model.

```r
par(mfrow = c(2,2))
plot(fit2)
```



Our qq plot shows that our model is fairly normal but tails off at both ends. Our residuals vs fitted plot is not very linear as we can see a V pattern. Our scale-location plot is a positive sloped line so it's not homoscedastic. Our residuals vs leverage plot doesn't show any influential outliers.

Conclusion

Our analysis shows that manual transmission is better than automatic transmission for mpg outcome. Our multivariate linear model was a fairly good model to show this conclusion even though it didn't have much homoscedasticity. We can not leave out the other variables because they are all correlated with mpg.