Regression Models Course Project

Antonio Vitor Villas Boas

07/02/2022

Executive Summary

In this article it will be presented an analysis comparing automatic and manual transmission impacts on Miles Per Gallon (MPG) and quantify the MPG difference difference between the two transmissions.

The data set to be studied is the `mtcars` . The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

```
library(ggplot2)
library(dplyr)
library(regclass)

data(mtcars)

mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
```

The exploratory analysis is shown in the appendix 1 and 2

Considering only the transmission as a predictor we can conclude that manual transmission cars have a better MPG performance over automatic transmission ones. Although there are other variables that may impact this conclusion such as weight, number of cylinders, gross horsepower etc. Appendix 3 has the `summary` of the model with all predictors.

It is important to have in mind that excluding correlated variables can lead to biased conclusions, including uncorrelated predictors can inflate the variance of the model. To measure this inflation, the `VIF` function calculates the Variation Inflation Factors of all predictors in a regressions model.

First let’s use a different fit to the model as an comparison to `full_fit` . One with and extra predictors: `cyl` .

```
fit2 <- lm(mpg ~ am + cyl, mtcars)
VIF(fit2)
```

```
##       am      cyl
## 1.375739 1.375739
```

```
full_fit <- lm(mpg ~ ., mtcars)
VIF(full_fit)
```

```
##       cyl      disp        hp      drat        wt      qsec        vs        am
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873  4.648487
##      gear      carb
##  5.357452  7.908747
```

comparing both am Variation Inflation Factors it shows that `am` is much more inflated when using all predictors in the model.

The function `step` in useful to select which predictors are better correlated to the model with the lowest Akaike Information Criterion (AIC).

```
final_fit <- step(full_fit, direction = "backward", trace = F)
final_fit
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Coefficients:
## (Intercept)           wt         qsec     amManual
##       9.618       -3.917        1.226        2.936
```

```
VIF(final_fit)
```

```
##       wt     qsec       am
## 2.482952 1.364339 2.541437
```

This shows that the lowest IAC = 61.31 and the VIF = 2.5414372 using three predictors: `wt` , `qsec` and `am` .

With these coefficients, we can conclude that Manual transmission cars get 2.936 more MPG on average than automatic ones.

the confidence interval for this statement is:

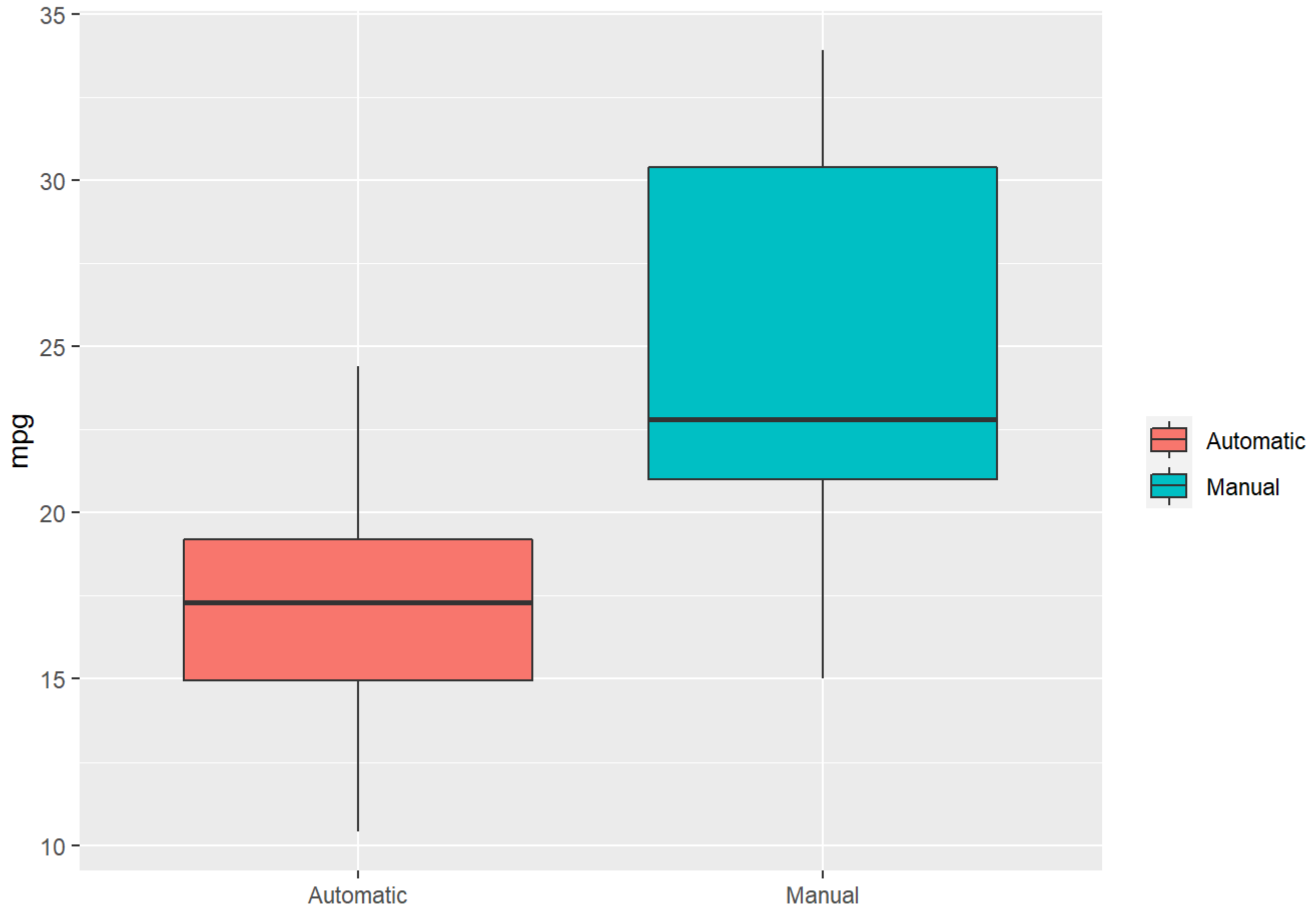
```
confint(final_fit)['amManual',]
```

```
##      2.5 %     97.5 %
## 0.04573031 5.82594408
```

Appendix

1. Simple model fit with `mpg` as an outcome and `am` as the only predictor

```
ggplot(mtcars,
       aes(x = am,
           y = mpg,
           fill = am)) +
    geom_boxplot() +
    labs(x = "", fill = "")
```



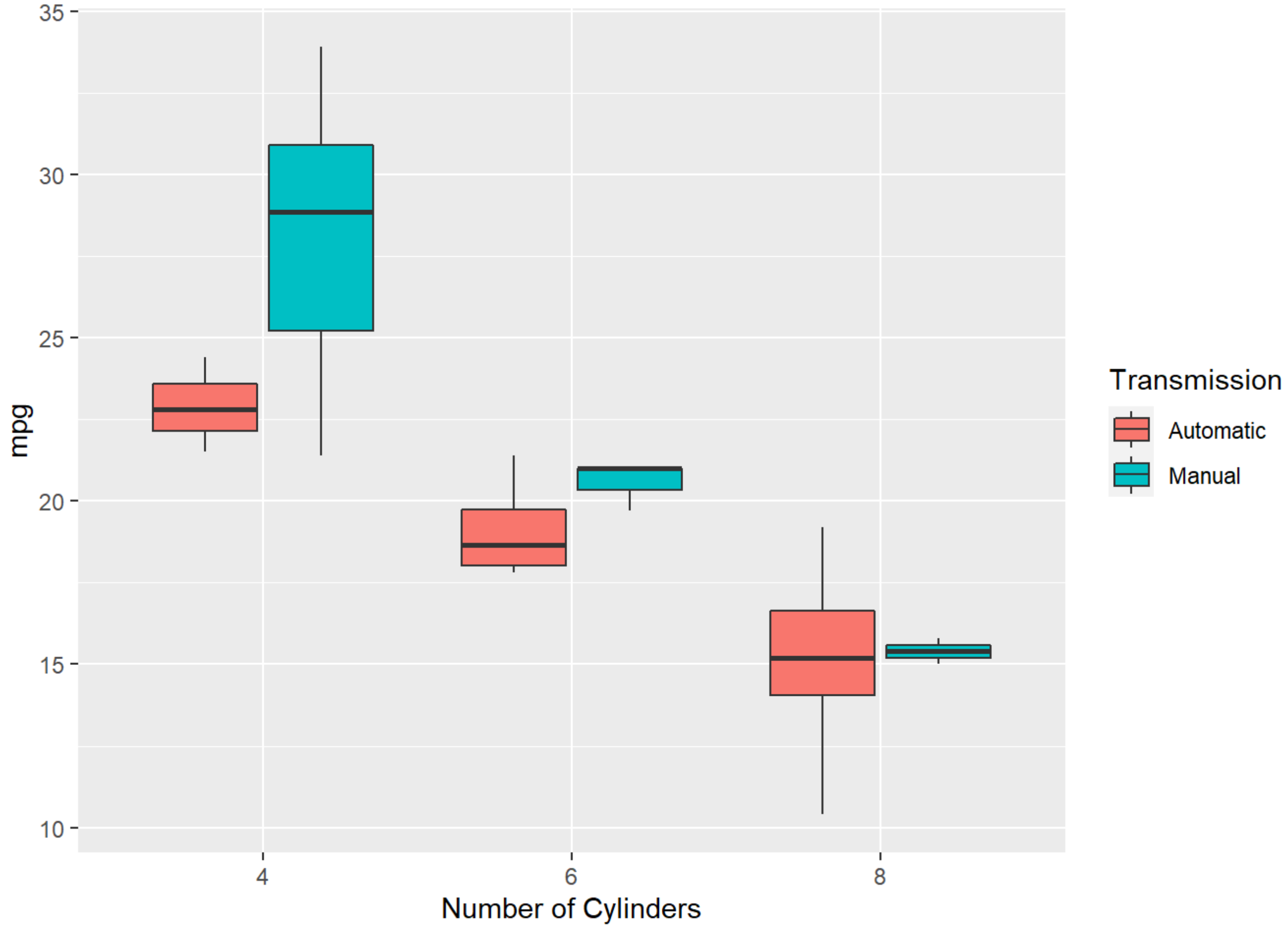

```
fit <- lm(mpg ~ am, mtcars)
summary(fit)
```

```
##
## 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
```

2. mpg ~ am separated by number of cylinders considered.

```
mtcars$cyl <- factor(mtcars$cyl)

ggplot(mtcars,
       aes(x = cyl,
           y = mpg,
           fill = am)) +
     geom_boxplot() +
     labs(fill = "Transmission",
          x = "Number of Cylinders")
```



3. Summary of the model fit with all predictors

```
summary(full_fit)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## amManual     2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

4. Null Hypothesis test

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

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means between group Automatic and group Manual
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
```