Regression Models Course Project

Endri Raco

4/12/2020

Executive Summary

This paper presents some insights regarding the relationship between a set of variables and miles per gallon (MPG) (outcome).

For this research we will analyse **mtcars** data set. 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).

We will write up an analysis to answer their question using regression models and exploratory data analyses.

The research aims to address the following questions:

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

Methods and Analysis

Importing data

We start our work by loading our dataset:

```
# Load data
data(mtcars)
```

Explore our data

Let's have a look in our dataset

```
# First dataset look
mtcars %>% explore_tbl()
```

The results of running **explore_tbl()** are shown in **Appendix 1**

We can see that our dataset **mtcars** has 11 variables and 32 observations.

Now that we have a general look in our data let's focus in separate variables:

```
mtcars %>% explore_all(ncol = 3)
```

The results are shown in Appendix 2. And now we will see difference between automatic and manual transmission

```
# Difference between automatic and manual transmission
mtcars %>% explore(am)
```

The results of running **explore()** are shown in Appendix 3.

We see that we have 59.4% of cars correnspondingto automatic transmission and 40.6% of cars correnspondingto manual transmission.

Regression Analysis

We start our regression analysis by creating a general idea regarding relationship between a set of variables and miles per gallon (MPG) (outcome):

```
# Relationship between several variables and mpg
mtcars %>% select(gear, mpg, hp, cyl, am) %>% explore_all(target = mpg)
```

The results are shown in Appendix 4. From the fourth boxplot we can see that cars with automatic transmission have more miles per gallon (mpg) compared to cars with manual transmission.

Saying that our H_0 hypothesis will be:

- MPG is the same for both transmission methods.

And our alternative hypothesis H_a:

- MPG is not the same for both transmission methods.

Let's use T test to see if it supports our H_0 hypothesis:

```
# T test
t_test <- t.test(mtcars$mpg ~ mtcars$am, conf.level = 0.95)
print(t_test)
```

The results of test are shown in Appendix 5.

Our **p-vaue** is **0.001374** so we reject null hypothesis, and accept our alternative hypothesis *MPG is not the same for both transmission methods.* Anyway the only way to accept this assumption is to keep all other characteristics constant for both types of transmission.

Let's have a deeper look on the question by using linear regression analysis.

We will build a multiple linear regression model, and use function **step** in R for a stepwise regression.

```
# Multiple linear regression model
reg_model1 = step(lm(data = mtcars, mpg ~ .), trace = 0, steps = 10000)
summary(reg_model1)
```

Model summary is shown in Appendix 6.

By looking at coefficients seems that **wt** variable, together with **am** and **qsec**

This model explains 0.85 of total variance. Let's make a final examination by checking how $mpg \sim wt + qsec$ is correlated to **am**.

```r
# Examine the correlation
reg_model2 <- lm(mpg ~ factor(am):wt + factor(am):qsec, data = mtcars)
summary(reg_model2)
```

The results of examination are shown in Appendix 7

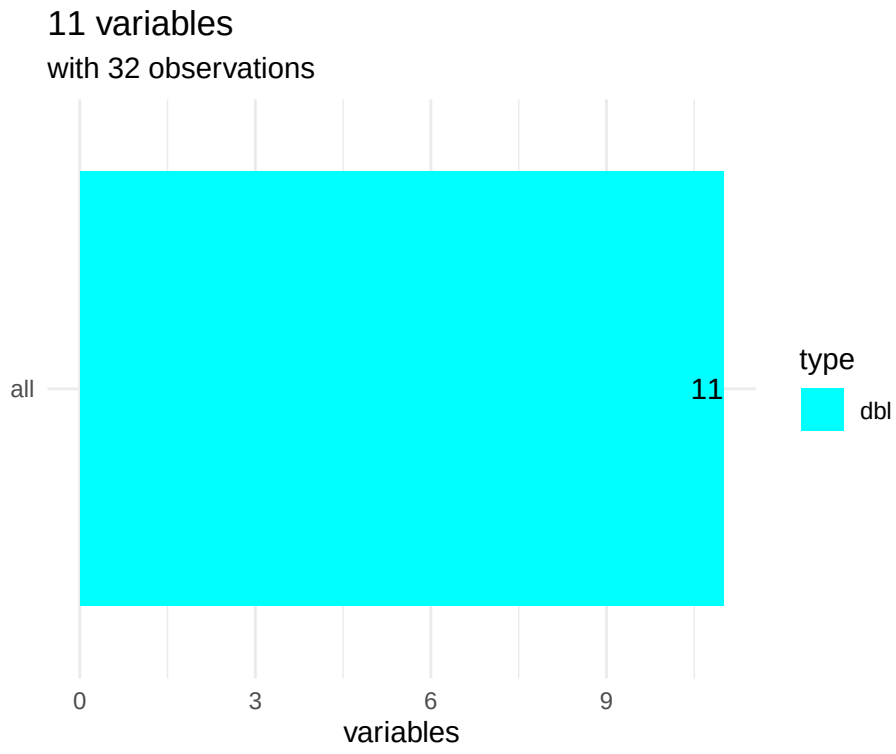
Results

Our model has:

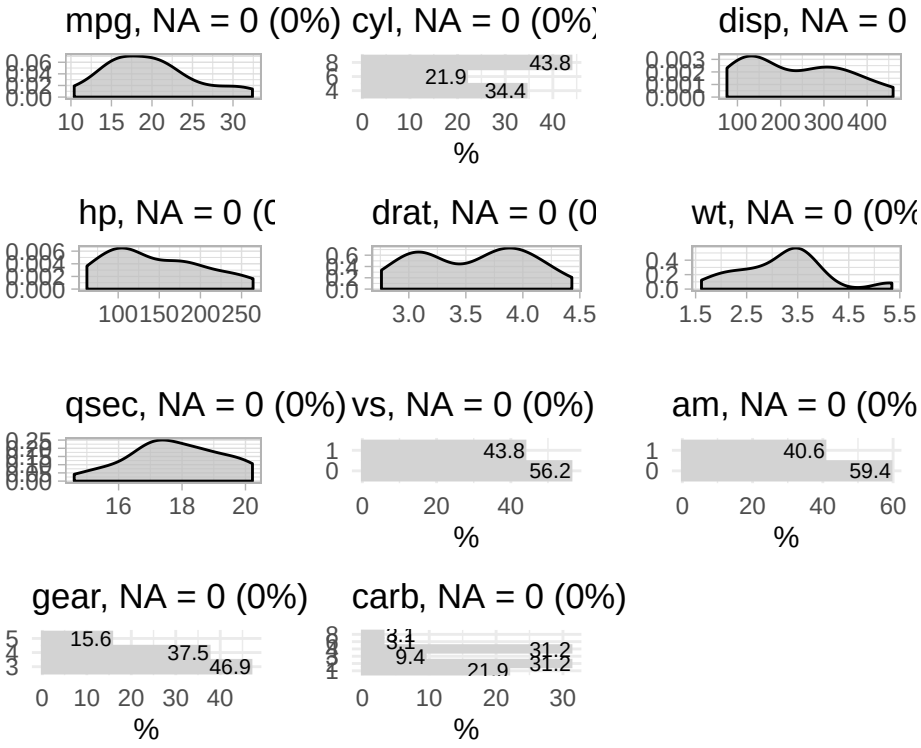
- 89.46% total variance
- adjusted R^2 0.879.

By dropping acceleration speed, and **qsec** increased 1 second, **mpg** increases by 0.834 miles when automatic transmission, and 1.446 miles when manual. So it seems that in given conditions of weight and accelaration manual transmission performs better regarding mpg.

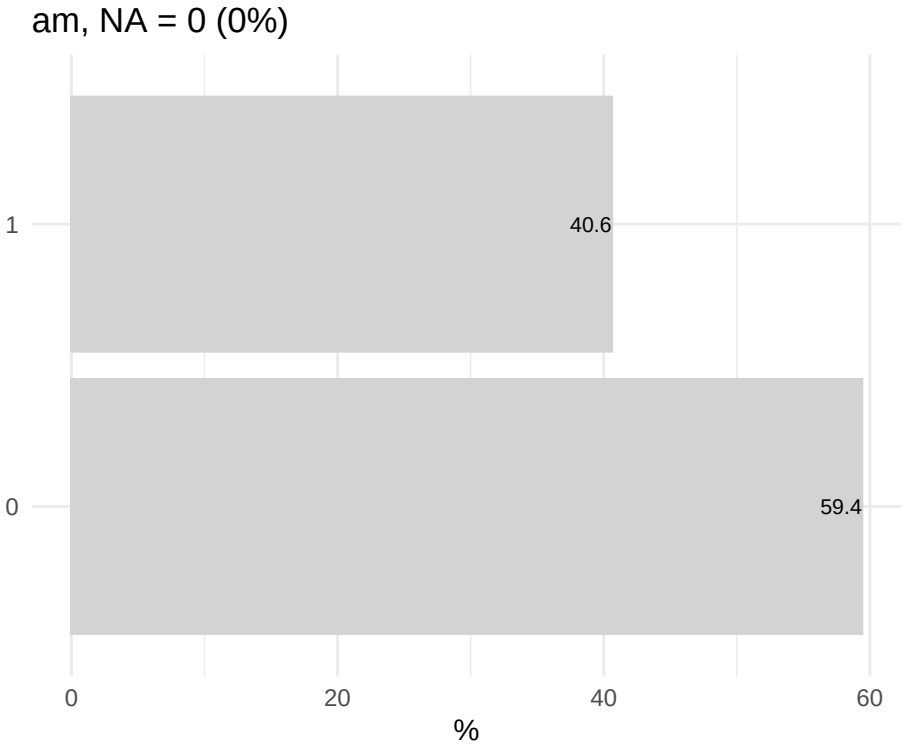
Appendix 1



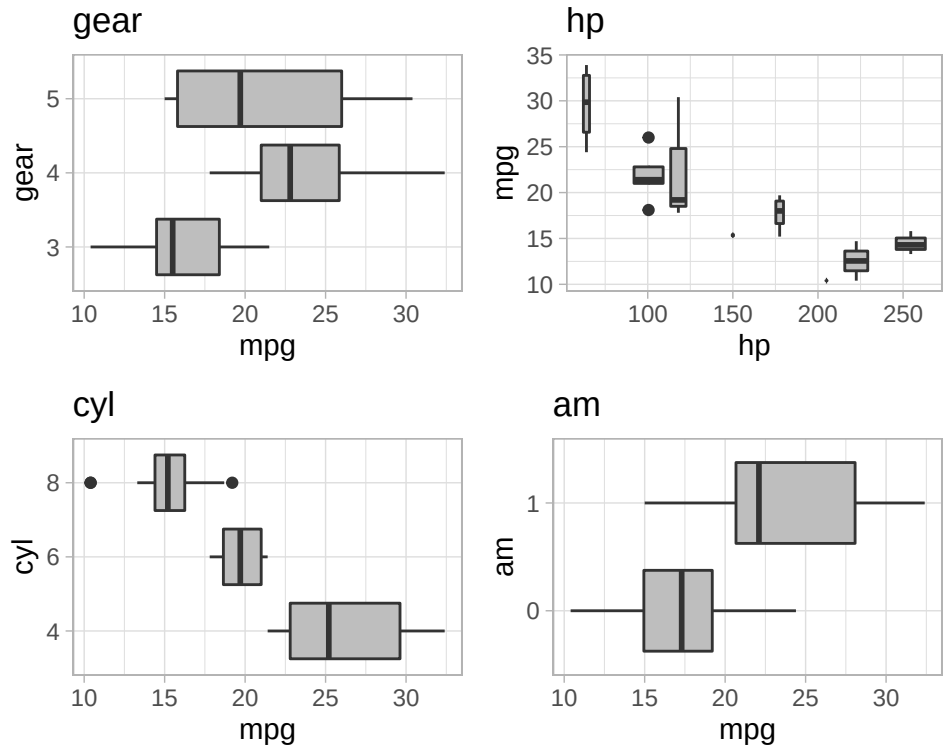
Appendix 2



Appendix 3



Appendix 4



Appendix 5

```
##
##  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 0 mean in group 1
##        17.14737        24.39231
```

Appendix 6

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Appendix 7

```
##
## Call:
## lm(formula = mpg ~ factor(am):wt + factor(am):qsec, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9361 -1.4017 -0.1551  1.2695  3.8862
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       13.9692     5.7756   2.419  0.02259 *
## factor(am)0:wt    -3.1759     0.6362  -4.992 3.11e-05 ***
## factor(am)1:wt    -6.0992     0.9685  -6.297 9.70e-07 ***
## factor(am)0:qsec   0.8338     0.2602   3.205  0.00346 **
## factor(am)1:qsec   1.4464     0.2692   5.373 1.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.097 on 27 degrees of freedom
## Multiple R-squared:  0.8946, Adjusted R-squared:  0.879
## F-statistic: 57.28 on 4 and 27 DF,  p-value: 8.424e-13
```