Analysis of relationship between car transmission and fuel consumption

Executive summary

We will use the mtcars dataset to determine the kind of transmission (automatic/manual) that has the most Miles per Gallon, and in which proportion. After a short description of the dataset, we will prepare the data. Then t-test group comparison will confirm the difference between the two groups. After, we will create a model to try to predict mpg according to transmission. Finally, we will try to improve the model and see that there are other variables that are better predictor than only the transmission.

Data description

We will use the mtcars dataset included in the R core. The dataset contains 32 observations, and 11 variables. You will find an excerpt of the dataset in appendix. Interesting variables in this dataset are:

- mpg: Miles/(US) gallon, it will be our outcome variable that we will analyze against the following predictors.
- cyl: Number of cylinders
- wt: Weight (lb/1000)
- am: Transmission(0 = automatic, 1 = manual)

After loading the data we prepare the data to have an additional column named *transmission* that match the label *Automatic* to the level 0 and *Manual* to the level 1.

Exploratory analysis

We would like to determine if there is a relationship between transmission an mpg. So the first idea is to draw a boxplot for the two groups. It shows there is a relation between transmission and mpg: cars use more fuel when transmission is manual than automatic. See Appendix for the plot.

A means of group comparison using Student t-test has been performed and confirms the relation. See appendix for more detail.

Regression model

Now, we will try to quantify the difference between the two types of transmission on mpg:

```
model.1 <- lm(mpg~transmission, data=mtcars)
```

This model shows that:

- the coefficient is statistically significant (with a p-value of 2.8502 × 10-4)
- transmission explains 35.9799% of the variance
- when the transmision change from automatic (0) to manual(1), there is an increase in mpg of 7.2449
- the 95% confidence interval is 3.6415, 10.8484

The residuals plot (see Appendix) shows that the points are randomly placed around the blue line, so there is no relation between the residual and the outcome.

Regression model improvement

Next we will try to improve our model by adding other variables, first let's look at the correlation matrix in appendix, it shows that wt and cyl have impact on mpg. So, we first create a model with transmission and cylinder as predictor of mpg:

```
model.2 <- lm(mpg ~ transmission + factor(cyl), mtcars)
```

Summary of different models are in the appendix. Here we can see that transmission coefficient is no more significant (with a p-value of 0.0585). This model also shows that:

- the coefficients for cylinder are statistically significant (with a p-value of 4.1061 $\times$ 10-4 for 6-cyl and 1.5466 $\times$ 10-7)
- transmission and cylinder explains 76.5111% of the variance
- when the transmision change from automatic (0) to manual(1), the coefficient is now 2.56

Another model with transmission and weight as predictor shows that transmission coefficient is no more significant (with a p-value of 0.9879):

```
model.3 <- lm(mpg ~ transmission + wt, mtcars)
```

Let's compare those 2 new models with the first one with ANOVA test:

```
pvalue.1vs2 <- anova(model.1, model.2)[[6]][2]
pvalue.1vs3 <- anova(model.1, model.3)[[6]][2]
```

We see that the 2 new models significantly explain more variance than the transmission variable alone:

- mpg ~ transmission vs mpg ~ transmission + cyl: 8.0101 $\times$ 10-7
- mpg ~ transmission vs mpg ~ transmission + wt: 1.8674 $\times$ 10-7

Conclusion

This study shows several points. First, there is a relationship between mpg and transmission: automatic transmission tends to have less mpg.

But if we go further by adding other variables like cylinder or weight, we see that those parameters give more accurate information on mpg variance.

So we can conclude that there is a difference between automatic and manual transmission but it is not the main variable that can explain the difference of consumption between the vehicles.

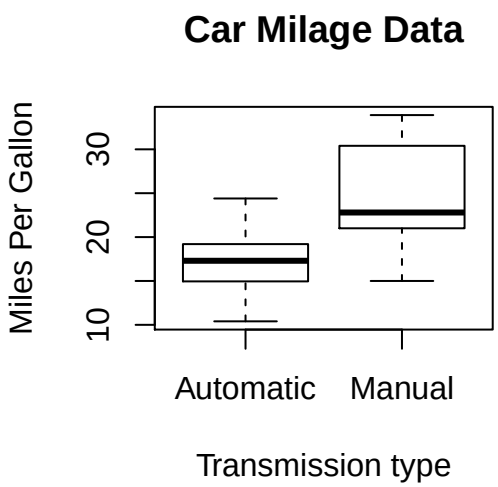
Appendix

Data excerpt

```
##                   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
##                   transmission
## Mazda RX4               Manual
## Mazda RX4 Wag           Manual
## Datsun 710              Manual
## Hornet 4 Drive       Automatic
## Hornet Sportabout    Automatic
## Valiant              Automatic
```

Boxplot of both transmission type



Means of group comparison

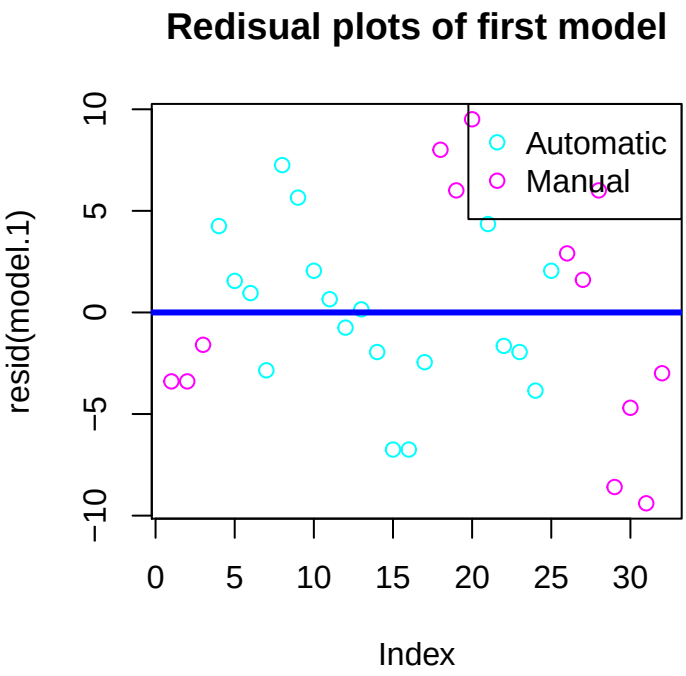
We will perform a Student's t-test to analyze the difference of mean mpg between automatic and manual transmissions. Let's suppose that:

- H0 (null hypothesis) is "there are no difference in means between the two groups"
- Ha (alternative hypothesis) is "the means are different between the two groups"

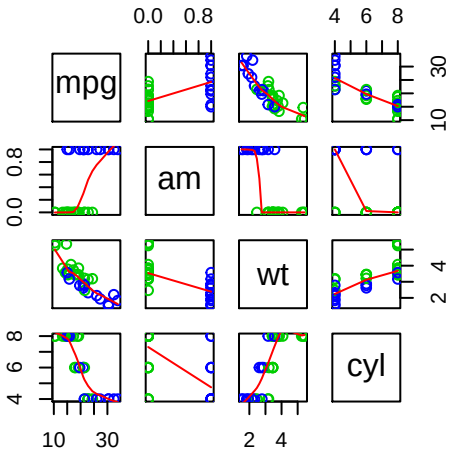
The test gives a p-value of 0.0014 less than 0.05, so we reject the null hypothesis and assume there is a difference between the two groups.

The confidence interval being negative (-11.2802, -3.2097), it tends to prove that manual transmission have greater mpg.

Residuals of model predicting mpg according to transmission



Correlation matrix



Model summaries

```
##
## Call:
## lm(formula = mpg ~ transmission, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.392 -3.092 -0.297  3.244  9.508
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)           17.15       1.12   15.25  1.1e-15 ***
## transmissionManual     7.24       1.76    4.11  0.00029 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.9 on 30 degrees of freedom
## Multiple R-squared:  0.36,   Adjusted R-squared:  0.338
## F-statistic: 16.9 on 1 and 30 DF,  p-value: 0.000285
```

```
##
## Call:
## lm(formula = mpg ~ transmission + factor(cyl), data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.962 -1.497 -0.206  1.891  6.538
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)           24.80       1.32   18.75  < 2e-16 ***
## transmissionManual     2.56       1.30    1.97  0.05846 .
## factor(cyl)6          -6.16       1.54   -4.01  0.00041 ***
## factor(cyl)8         -10.07       1.45   -6.93  1.5e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.07 on 28 degrees of freedom
## Multiple R-squared:  0.765,  Adjusted R-squared:  0.74
## F-statistic: 30.4 on 3 and 28 DF,  p-value: 5.96e-09
```

```
##
## Call:
## lm(formula = mpg ~ transmission + wt, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.530 -2.362 -0.132  1.403  6.878
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         37.3216     3.0546   12.22  5.8e-13 ***
## transmissionManual  -0.0236     1.5456   -0.02     0.99
## wt                  -5.3528     0.7882   -6.79  1.9e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.1 on 29 degrees of freedom
## Multiple R-squared:  0.753,  Adjusted R-squared:  0.736
## F-statistic: 44.2 on 2 and 29 DF,  p-value: 1.58e-09
```