# Course project Regression models

*kwatanwa17*

*12/9/2018*

## Executive summmary

In this paper, we analyse "mtcar" data set to compare cars with different transmission type based on males per gallon (mpg). We conclude that in general manual cars are more efficinet than automatic ones, but this relationship is also explained by other variables.

## Data description

The "mtcars" data set 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). It is a data frame with 32 observations on 11 (numeric) variables.

## Prepearing for the analysis

Before we start the analysis, we need to change the classes of some variables.

```
#change variable clasees
mtcars$cyl <- factor(mtcars$cyl)
mtcars$am <- factor(mtcars$am, levels = c(0,1), labels = c("automatic", "manual"))
mtcars$vs <- factor(mtcars$vs, levels = c(0,1), labels = c("V-shaped", "straight"))
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

## First analysis (mpg vs am)

Firstly, we check only mpg and am variables. See Figure 1 in the appendix.

```
#ttest
t.test(mtcars$mpg ~ mtcars$am, alternative = "two.sided", paired = FALSE, var.equal = FALSE)
```

```
    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
```

It is quite obvious that the manual cars have heigher mean than the automatic ones, but we suspect other variables that afects this relationship.

## Explanation

Now we check the all possibles relationships in the variables. To do that, we use pairs function to investigate possible correlation between variables. As a result of the plot, cyl, disp, hp, drat, wt, vs, and am seem to have strong relationship with the mpg variable. See the Figure 2.

## Modeling

### Linear model

```
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
```

Ajusted R-squred value tells us that 33% of the variation of mpg will be explained by am.

## Modeling test

As we see in Figure 2, there are some strong relationships between mpg and some variables of the data set. We have to check if these variables are statistically significant in the regression. In order to obtain the optimal regression, we conduct AIC algorithms. First, we take all dependent variables and then elminate on by one until AIC score reaches a limit.

```
initial_fit <- lm(mpg ~ ., data = mtcars)
final_fit <- step(initial_fit, direction = "both")
```

And this regression seems to be better than the regression with only am variable.

```
anova(fit,final_fit)
```

```
Analysis of Variance Table

Model 1: mpg ~ am
Model 2: mpg ~ cyl + hp + wt + am
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     30 720.90
2     26 151.03  4    569.87 24.527 1.688e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Analysis of final fit model

## Residuals

The QQ-plot shows that residuals are normally distributed and Residual vs Fitted plot indicates some outliers. See Figure 3.

## final fit

```
summary(final_fit)
```

```
Call:
lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9387 -1.2560 -0.4013  1.1253  5.0513

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
cyl6        -3.03134    1.40728  -2.154  0.04068 *
cyl8        -2.16368    2.28425  -0.947  0.35225
hp          -0.03211    0.01369  -2.345  0.02693 *
wt          -2.49683    0.88559  -2.819  0.00908 **
ammanual     1.80921    1.39630   1.296  0.20646
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.41 on 26 degrees of freedom
Multiple R-squared:  0.8659,    Adjusted R-squared:  0.8401
F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

# Conclusion

To sum up, we conclude the following:

- First, munual transmission cars have more miles per galon (mpg) than automatic transmission cars. In concreatly?, the difference on the mean is 1.8 ajusted by cyl, hp and wt variables.
- secondly, with respect to the number of cylinders, the cylinder 4 seems to have the highest mean in both two transmission type. See Figure 4.
- thirdly, both hp and wt have negative relationship with mpg. If we increment hp or wt, mpg will decrease.

# Appendix

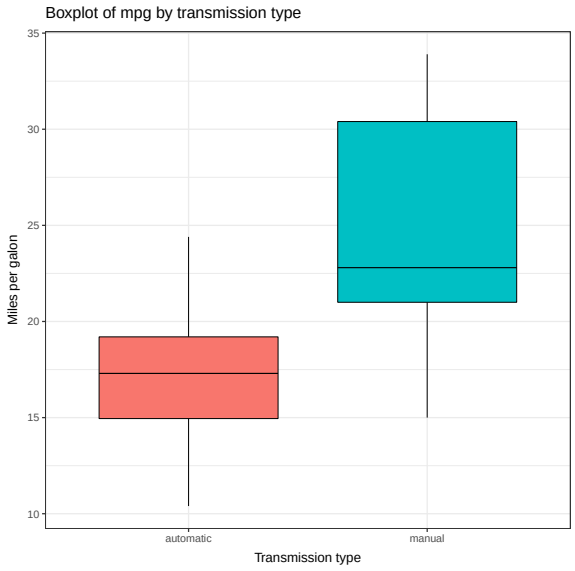


Figure 1: Boxplot of mpg by transmission type

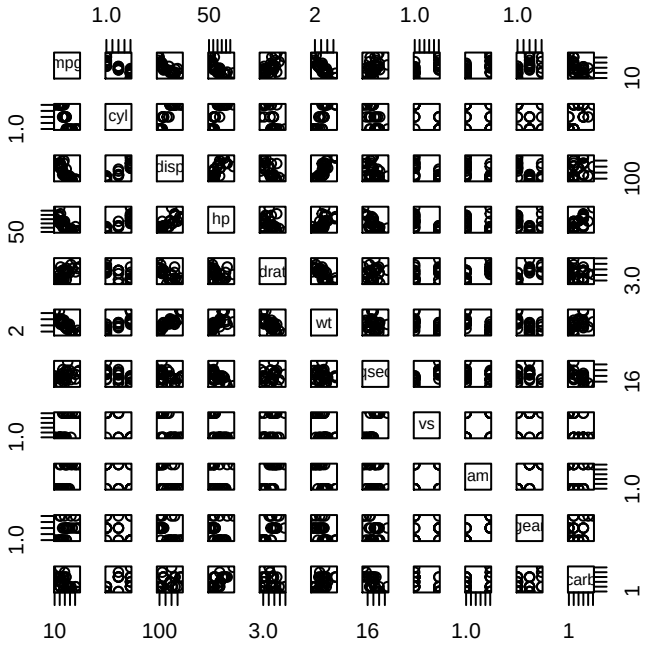


Figure 2: Pairs plots

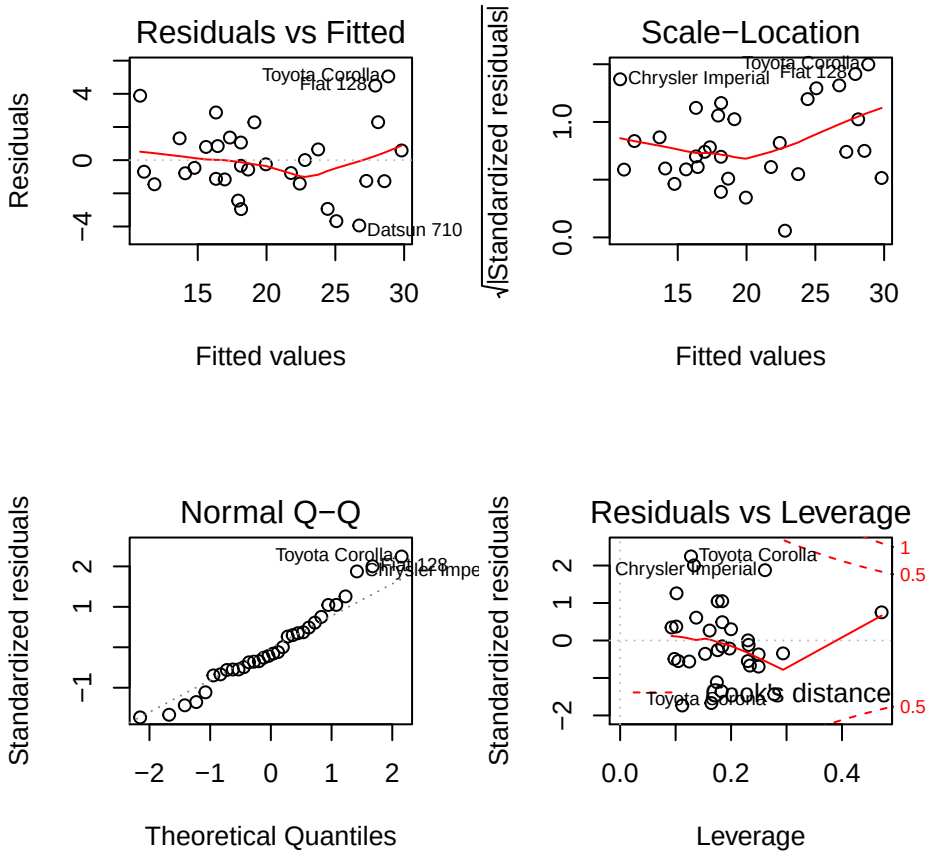


Figure 3: Residual analysis of final fit model

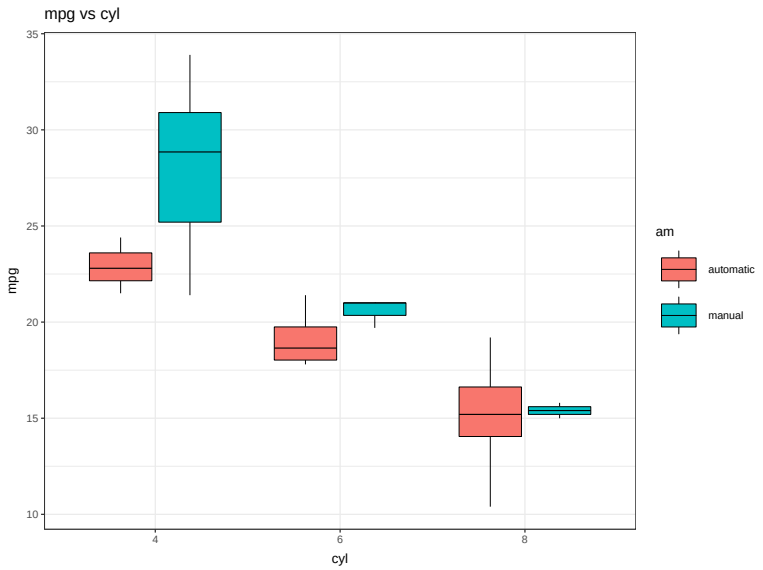


Figure 4: mpg vs cyl