# Regression models course project

## Executive summary

This project uses the **mtcars** data set in R to answer the following couple of questions. We are interested in exploring the relationship between a set of variables and the miles per gallon (MPG). In particular, we will answer these questions:

*Is an automatic or manual transmission better for MPG*

*Quantify the MPG difference between automatic and manual transmissions*

Out final preferred model for miles per galollon as outcome (mpg) includes as explanatory variables the type of transmission (am), the weight of the car (wt), an interaction between the two (since the weight differs across transmission type), and factors for the number of cylinders (cyl). Our overall conclusion is that **manual tranmission has higher mpg**.

## Exploratory analysis of the data

We start with loading the data and doing some basic summary.

```r
library(datasets)
head(mtcars, 3)
str(mtcars) #output not shown
```

## Do automatic and manual transmission cars differ in mpg use?

We are interested in whether the mpg differes by type of transmission (manual vs automatic). In the Appendix, we draw a boxplot showing the mean mpg by type of transmission. The average mpg for automatic transmission is about 17, and for manual transmission about 25.

A t-test for the difference in means has been performed (see Appendix).The p-value of the null hypothesis (of equal means) is 0.001374, so we reject the null and infer that the mean mpg significantly differs between manual and automatic cars.

## Which variables could explain mpg?

The most difficult part is to pick the appropriate variable (other than the transmission type) that we will include in our model. We draw pairwise graphs between mpg and a number of variables that we believe could affect it. We also present a function (taken from the pairs Help page), which calculates to correlation between the variables. See the Appendix for the syntax of the function and all the graphs. The variables with highest correlation with mpg are ***weight(wt)*** and ***number of cylinders(cyl)***. Since we want our covariates to have high explanatory power for the outcome, we will consider using these variables in our model.

## Regression analysis

We already have our candidates for variables to include in the regression model. To see which is the most appropriate one, we will fit a few models and compare how they perform.

We start with a simple model, fitting mpg only with the type of transmission. Then, we add the weight and cylinders and compare how that one is performing. However, after performing other exploratory analysis, we realised that the weight variable differes across transmission types. See the Appendix for very illustrative ggplot and for a t-test for difference in means. The t-test indeed showed that the weight across the two types is signifantly different. Therefore, a final model we will consider is one interacting the weight with the transmission type.

In the first model (fit1), the coefficient of the am variable is 7.245, which is the difference between manual cars (am=1) and automatic cars (am=0). It explains about 34% of the variation in mpg (adjusted R-squared is 0.3385). Interestingly, in the second model the adjusted R-squared increases to 0.8134 but the factor for transmission is no longer statistically significant. However, as we expect a strong relationship between **wt** and **am**, this insignificance could be due to omitted variables bias. Our final model (fit3) has R-sqared of 0.8775, and the coefficient for the type of transmission is **11.569**. Overall, if we hold wt and cyl constant, a manual car (am=1) has 11.569 -2.399*wt more mpg than automatic (am=0) cars. For example, a manual car that weight 1000lbs, has 9.17 more mpg than an automatic car with the same weight and number of cyclinders. Our preferred model has 26 df and residual standard error of about 2.304.

```r
fit1<-lm(mpg~factor(am), data=mtcars)
summary(fit1)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

```r
#second model
fit2<-lm(mpg~wt+factor(am)+factor(cyl), data=mtcars)
summary(fit2)$coef
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  33.7535920  2.8134831 11.9970836 2.495549e-12
## wt           -3.1495978  0.9080495 -3.4685309 1.770987e-03
## factor(am)1   0.1501031  1.3002231  0.1154441 9.089474e-01
## factor(cyl)6 -4.2573185  1.4112394 -3.0167231 5.514697e-03
## factor(cyl)8 -6.0791189  1.6837131 -3.6105432 1.227964e-03
```

```r
#third model
fit3<-lm(mpg~wt+factor(am)+wt*factor(am)+factor(cyl), data=mtcars)
summary(fit3)$coef
```

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    29.774836  2.8403415 10.482836 7.870715e-11
## wt             -2.398713  0.8439884 -2.842116 8.603904e-03
## factor(am)1    11.568790  4.0877912  2.830083 8.853842e-03
## factor(cyl)6   -2.709777  1.3573517 -1.996370 5.646509e-02
## factor(cyl)8   -4.776110  1.5558306 -3.069814 4.964603e-03
## wt:factor(am)1 -4.067981  1.3974151 -2.911075 7.295503e-03
```

To test whether adding additional terms is necessary, we used ANOVA likelihood ratio test for nested models. The output is given below. The conclusion from the ANova test is that adding the variables is appropriate.

```r
anova(fit1, fit2, fit3)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ wt + factor(am) + factor(cyl)
## Model 3: mpg ~ wt + factor(am) + wt * factor(am) + factor(cyl)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)    
## 1     30 720.90                                   
## 2     27 182.97  3    537.93 33.7850 4.031e-09 ***
## 3     26 137.99  1     44.98  8.4744  0.007296 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### Some robustness tests

Finally, we performed some robustness analysis on the model we chose. We did the following (see the Appendix for code and graphs):

Plotting the fitted valued and the residuals failed to display some pattern.

Normall Q-Q plot indicates the residuals are normally distributed, as we see no large deviations from the line.

The Scale-Location does not appear problematic.

Finally, residuals versus Leverage does not show the presence of outliers.

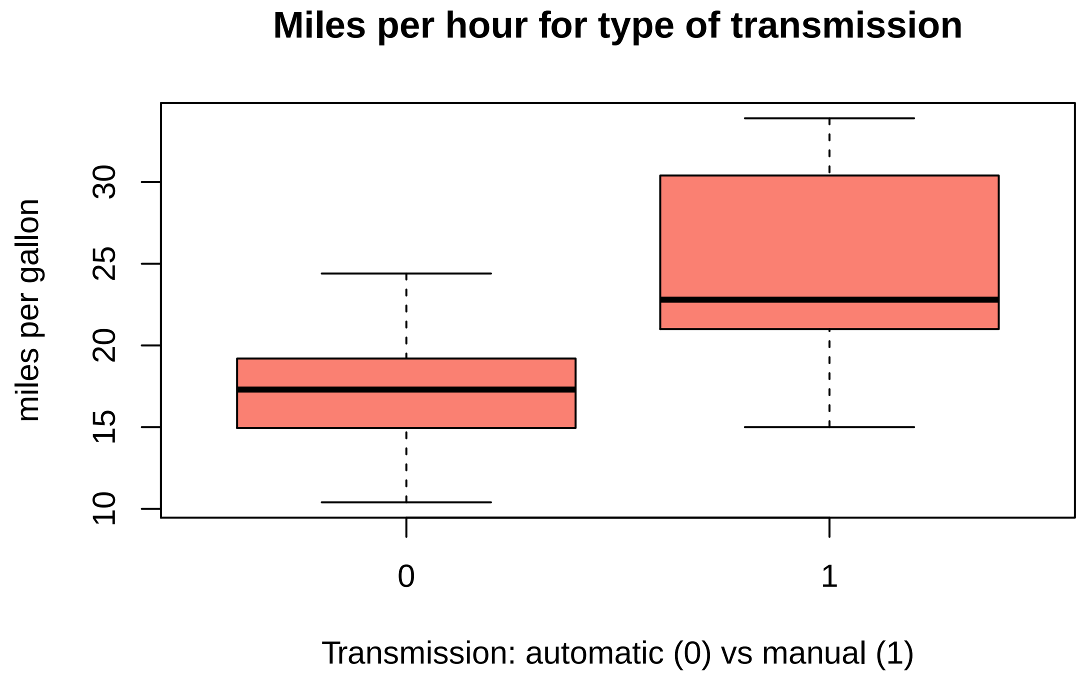
### Concluding

In conclusion, we fitted a model that explains the miles per gallon (mpg) with the type of transmission of a car (manual/automatic), the weight, the cylinders, and and interaction between the weight and the type of a car. We find that manual cars are overall associated with higher mpg than automatic ones.

### Appendix

Boxplot for mean mpg across manual and automatic cars.

```
boxplot(mpg~am, data=mtcars, col="salmon", xlab="Transmission: automatic (0) vs manual (1)", ylab="miles
```



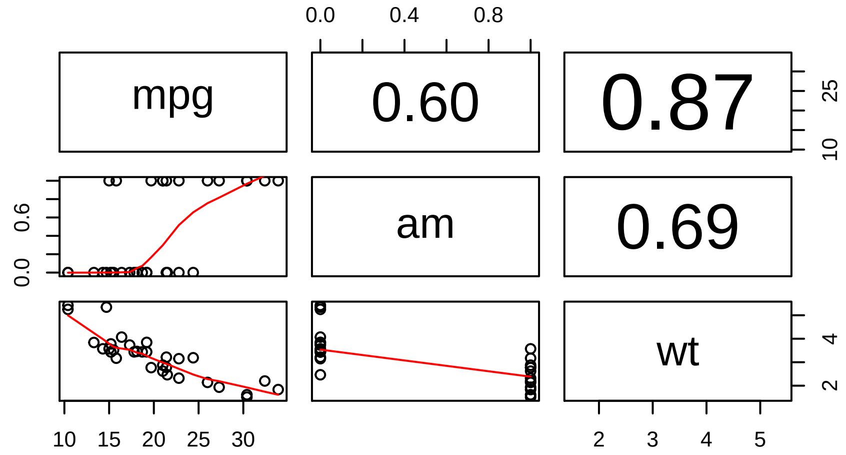
T-test for difference of means between manual and automatic

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

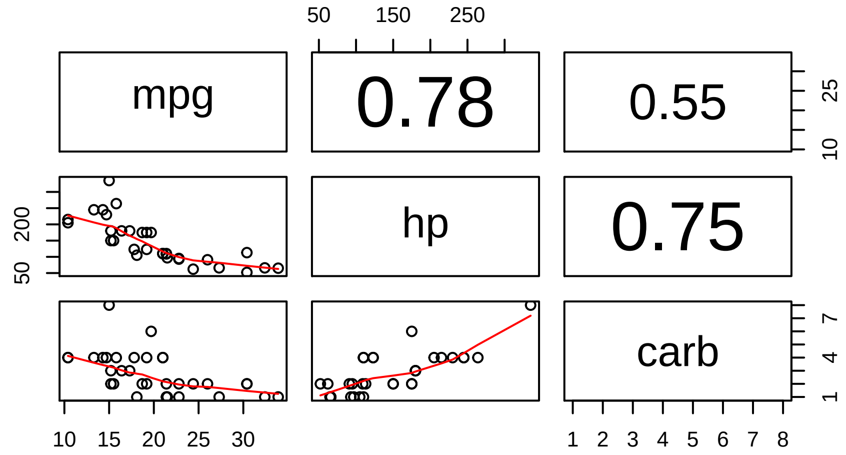
```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg by mtcars$am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

A code for function that in the upper panel of pairwise plots, writes the correlations between the respective variables (note, this code was in the *pairs Help page* but was useful to use in our context).The code is followed by some paired plots between mpg and other variables. For better visibility, we plot mpg with only 2 other variables at a time, though it is, of course, possible to draw the relationships between all simultaneously.

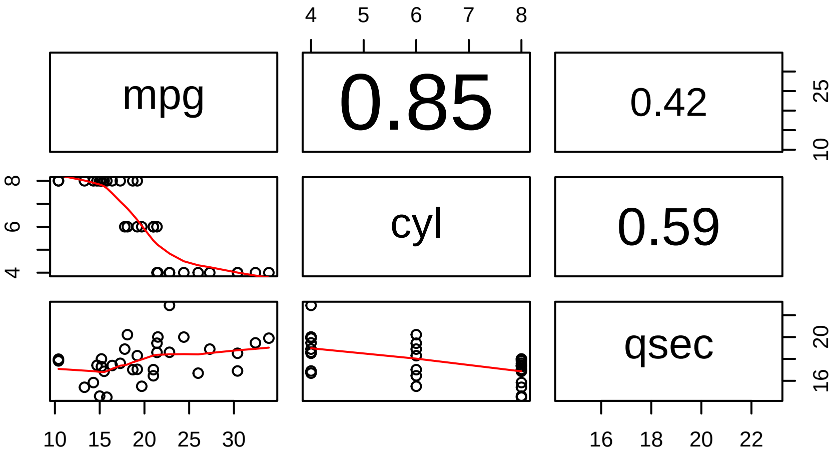
```
panel.cor <- function(x, y, digits=2, prefix="", cex.cor)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits=digits)[1]
    txt <- paste(prefix, txt, sep="")
    if(missing(cex.cor)) cex <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex * r)
}
pairs(mpg~am+wt, data=mtcars, lower.panel=panel.smooth, upper.panel=panel.cor)
```



```
pairs(mpg~hp+carb, data=mtcars, lower.panel=panel.smooth, upper.panel=panel.cor)
```



```
pairs(mpg~cyl+qsec, data=mtcars, lower.panel=panel.smooth, upper.panel=panel.cor)
```



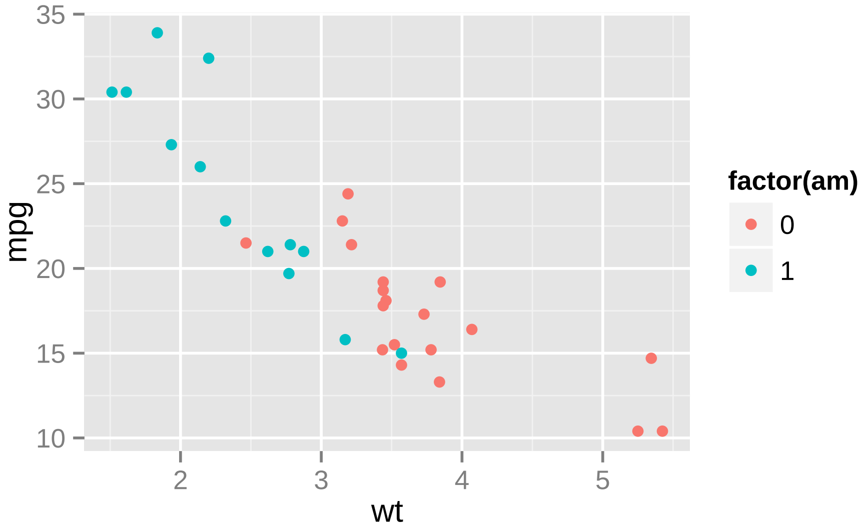
### Weight differences in automatic and manual transmission cars

Do automatic versus manual transmission cars have difference in weight and number of cylinders? We can do a t-test for difference in means.

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

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$wt by mtcars$am
## t = 5.4939, df = 29.234, p-value = 6.272e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8525632 1.8632262
## sample estimates:
## mean in group 0 mean in group 1
##        3.768895        2.411000
```

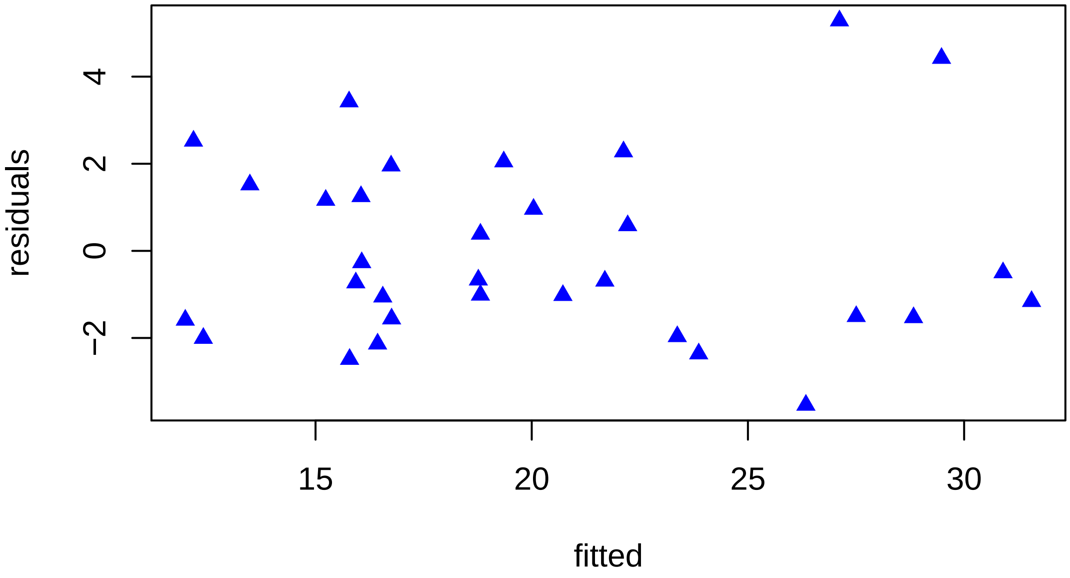
```
library(ggplot2)
ggplot(mtcars, aes(wt, mpg, color=factor(am)))+geom_point()
```



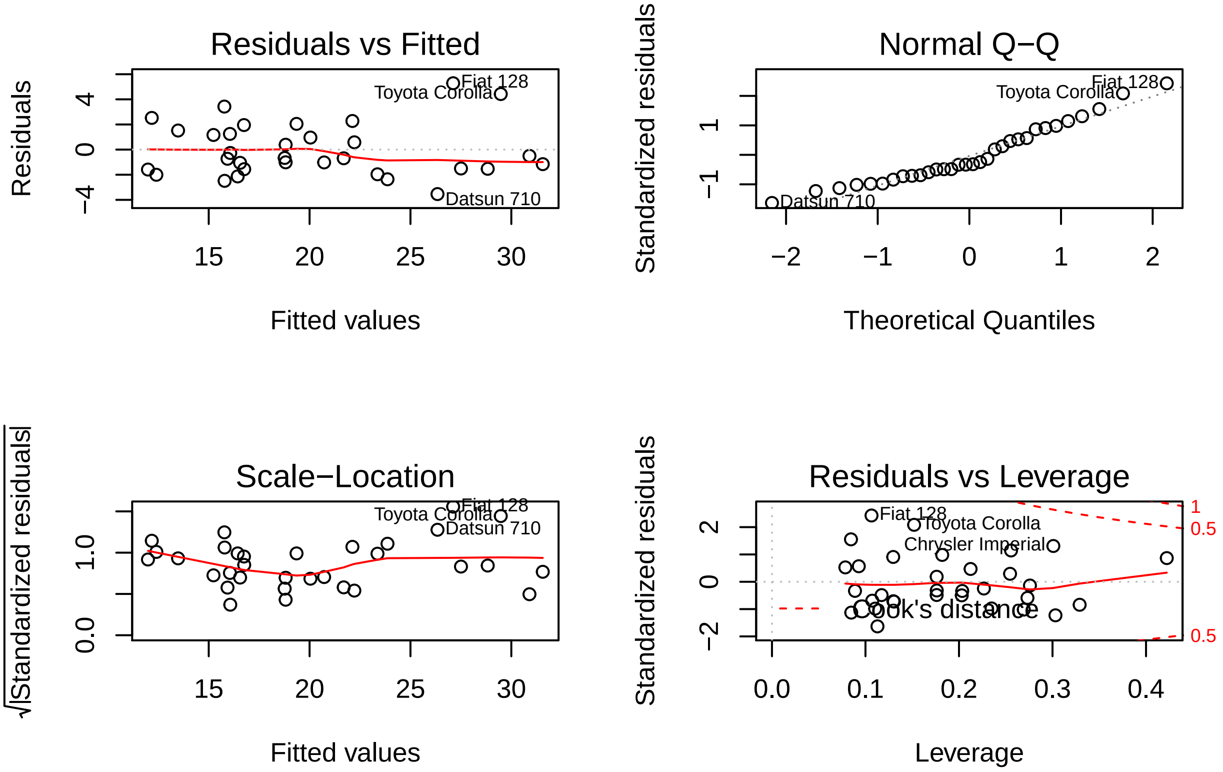
### Residuals

Residual plots

```
plot(fit3$fitted, fit3$res, col="blue", pch=17, ylab="residuals", xlab="fitted")
```



```
#Overall residual plots
par(mfrow=c(2,2))
plot(fit3)
```


Residuals vs Fitted
Residuals
−4
0
4
15
20
25
30
Fitted values
Fiat 128
Toyota Corolla
Datsun 710
Normal Q−Q
Standardized residuals
−1
1
−2
−1
0
1
2
Theoretical Quantiles
Fiat 128
Toyota Corolla
Datsun 710
Scale−Location
√|Standardized residuals|
0.0
1.0
15
20
25
30
Fitted values
Fiat 128
Toyota Corolla
Datsun 710
Residuals vs Leverage
Standardized residuals
−2
0
2
0.0
0.1
0.2
0.3
0.4
Leverage
Fiat 128
Toyota Corolla
Chrysler Imperial
Cook's distance
1
0.5
0.5