R workshop #2: multivariate regression analysis and factor interactions

Nicola Romanò

Introduction

Last year we have talked about linear models and their use to perform regression and analysis of variance (ANOVA). We only considered simple situations where we have one independent variable influencing our measured variable (one-way ANOVA) or two factors (two-ways ANOVA) that do not interact with each other. In the lectures we have now talked about interactions and how they change our interpretation of linear models. In this workshop we will have a look at how to deal with interactions in R.

Learning objectives

After completing this workshop you will be able to:

- Use linear models to perform multiple regression
- Use linear models to perform analysis of variance with multiple factors
- Correctly interpret the results of your analysis in the presence of interactions

Section 1 - A refresher on linear models

We start this workshop with a little refresher of linear models. A linear model is a statistical model that relates the changes in a dependent variable (Y) with the changes in one or more independent variables ($X_1, X_2, ..., X_n$).

The general equation for such model is:
$Y = X_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$
Where:

- Y is our measured variables
- $X_1, ..., X_n$ are the factors (or predictors) that influence Y. They are generally the other variables in your dataset, or transformations/combination of them [1].
- $\beta_1, ...\beta_n$ are the regression coefficients, scaling factor that indicate the importance of each predictor and its effect on the outcome variable.

[1] For instance, we may have collected the weight of the subjects in our study, but use log(weight) as a predictor for our model. Or we may have collected two different values and use their ratio as a model parameter

- ϵ is the error, or residual. It represents the difference between what is explained by the model prediction, and what we have observed. It includes the effect of all the factors that we did not measure in our experimental setup, as well as measurement errors. We assume that it is normally distributed.

When we use R (or any other software!) to generate the model, what it does is estimating the coefficients β in such a way to minimise the error[2].

In this formula each predictor acts independently from the others. In other words, if we have two predictors, X_1 and X_2, the effect of X_1 on Y will always be the same, independently of the value of X_2. As we have seen in the lecture this is not always the case.

Simple regression

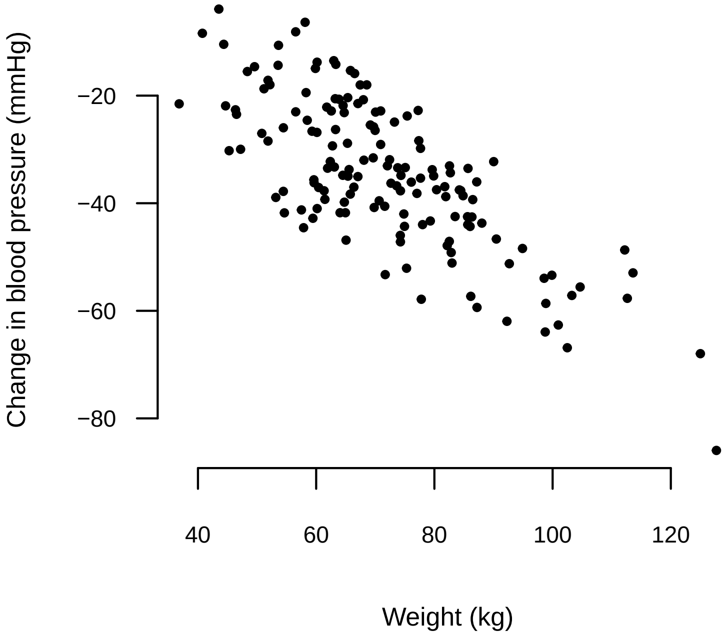
As a first example let's consider the dataset *pressure-workshop2.csv*. In this study the effect of a drug on reducing blood pressure (measured in mmHg) has been investigated on 150 patients of different age, weight (measured in kg), and sex.

Start by familiarising with the data. How many men and women are there? What range of age and weight? Try and plot the various variables against each other and see if any particular patterns emerge [3].

Let's forget for a moment about the other variables and concentrate on the relation between Weight and response; it looks like the largest effect is seen in heavier patients.

[2] In the case of `lm`, this is called a least-square estimation. In statistics books and publications you may see the estimated parameters indicated as $\hat{\beta}$ (read as "beta hat"). This is to indicate that this is the result of an estimation, that is an approximation of the true value of β, which remains unknown

[3] If you do not remember how to do that, see Workshop 1.



We can use a linear model to test whether such a relation exists.

As always, we start by stating our null hypothesis ________________________

____

Do you remember how to perform a linear regression in R? Try it, if you don’t remember see the following page!

```
model <- lm(Response ~ Weight, data = pressure)
```

This generates the model

$Response \ = \beta_0 + \beta_1 * Weight + \epsilon$

What are the assumption of this model? Do you remember how to verify that they are satisfied? [4]

This is one of the simplest linear models we can generate, where the value of the outcome depends on a single parameter. This is called *simple regression*.

Let's look at the output of the model

```
summary(model)
```

```
##
## Call:
## lm(formula = Response ~ Weight, data = pressure)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.190  -6.354   0.874   6.568  19.554
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.78143    3.32703   3.541 0.000533 ***
## Weight      -0.64852    0.04521 -14.346  < 2e-16 ***
## ---
## Signif. codes:
## 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 9.133 on 148 degrees of freedom
## Multiple R-squared:  0.5817, Adjusted R-squared:  0.5789
## F-statistic: 205.8 on 1 and 148 DF,  p-value: < 2.2e-16
```

The summary gives us a lot of information.

First of all, it tells us the parameters β (coefficients) that have been estimated by the model.

$\hat{\beta}_0 = 11.78$ and $\hat{\beta}_1 = -0.65$

Therefore

$Response \ = 11.78 - 0.65 * Weight + \epsilon$

This means that for any increase of 1 Kg in weight there is a decrease of 0.65 mmHg in blood pressure following the intake of the drug. The effect of weight on the response to the drug is statistically significant ($F_{1,148} = 205.8, p = 2 * 10^{-16}$)[5].

[4] Let's discuss this in the forum! I would say that for this case the assumptions are generally satisfied, what do you think?

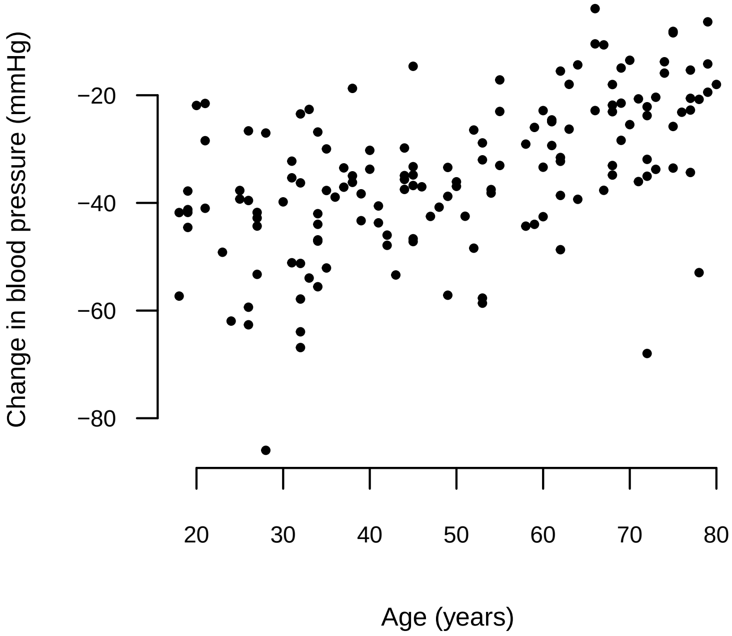
[5] R also reports a p-value for the intercept; this is the result of a one sample t-test comparing the intercept to 0. In other words, in this case the intercept is statistically different from 0. The intercept is the value corresponding to a change in blood pressure where all of the factors (in this case weight) are equal to zero. Since a weight of 0 is not biologically meaningful we can ignore this value in this instance

Another important value is the coefficient of determination (R^2, sometimes called deviance). This is a measure of how good the model is, or how much of the variation in the data it explains. R^2 is not a great way to compare two different models, since it depends on the number of parameters; that is, if we add an extra descriptor to our model R^2 will always increase. For this reason, R reports also an "adjusted" version of it. In this case $adj.R^2 = 0.5789$; this means that our model describes/explains ~57.9% of the variability in our data, which is OK but not great. It means that there are other factors that we have not considered accounting for >40% of the variability! [6] So, what are these other factors?

Multiple regression

Our dataset contains two other descriptors: Age and Sex. It is very biologically plausible that these would affect blood pressure, so we should add them to our model[7]. To keep things simple, we will start with Age, and consider gender later.

It is useful, at this point, to also plot the change in blood pressure against age.



We see a possible in the response to the drug depending on age. Let's incorporate age in our model.

```
model.2 <- lm(Response ~ Weight + Age, data = pressure)
```

This will generate a model that considers the effect of weight and

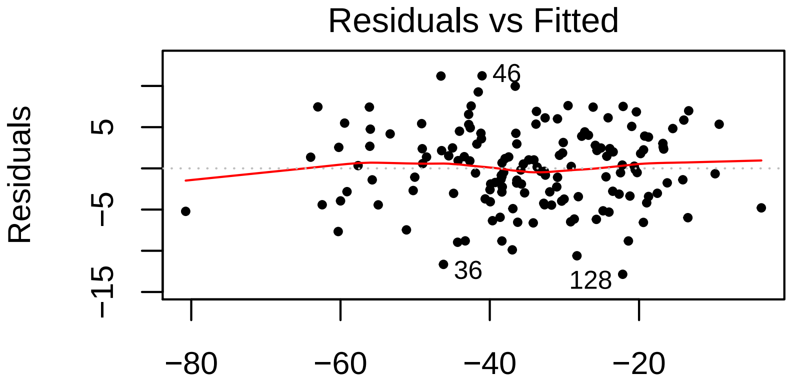
[6] What do you think is the maximum value of R^2? Why?

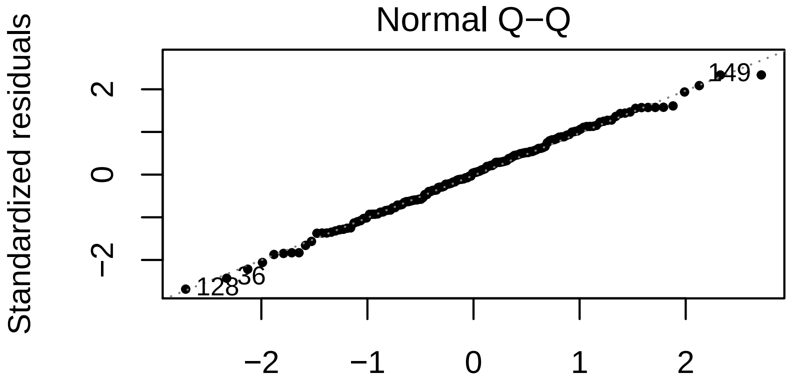
[7] Note that, although for the sake of simplicity we are adding these descriptors one at a time, in practice we would start from a complete model, including all of the descriptors!

the effect of age, independently of each other [8]

What are the null hypotheses[9] that this model is testing?

Again, we want to check the assumptions of the model by using diagnostic plots.





Since the diagnostic plots look good (equal variance throughout and normally distributed residuals), we can continue with this model.

```
summary(model.2)
```

```
##
## Call:
## lm(formula = Response ~ Weight + Age, data = pressure)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8545  -3.3021   0.1972   3.1070  11.2320
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.60214    2.04592  -4.205 4.53e-05 ***
## Weight      -0.65779    0.02392 -27.501  < 2e-16 ***
## Age          0.42390    0.02169  19.541  < 2e-16 ***
## ---
## Signif. codes:
```

[8] This means that the model will look at the effect of the weight of the individual on his/her response to the drug, independently of his/her age, and vice versa for age.

[9] There is more than one!!!

```
## 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.832 on 147 degrees of freedom
## Multiple R-squared:  0.8837, Adjusted R-squared:  0.8821
## F-statistic: 558.6 on 2 and 147 DF,  p-value: < 2.2e-16
```

You can interpret the result of this model just like you did for the previous one. It tells us that there is a statistically significant effect of weight ($p < 2 * 10^{-16}$) and of age ($p < 2 * 10^{-16}$) on the response to the drug ($F_{2,147} = 558.6$ [10]) Note that now the model explains 88% of the variability!

Let’s now consider gender and add it to our model. Plot the data, do you think gender affects the response to the drug? We now are dealing with a discrete qualitative variable, with two levels, F and M. All that we said so far still applies, and `lm` is able to deal with this type of variables with no issue. However, the way we deal with these type of variables is slightly different.

```
model.3 <- lm(Response ~ Weight + Age + Sex, data = pressure)
```

This is modelling the following:

$Response = \beta_0 + \beta_1 * Weight + \beta_2 * Height$

I will leave to you the exercise of checking the assumptions and discussing the output of this third model.

[10] Note that the F statistic reported by `summary` refers to the whole model. If you wanted to know the F statistic for specific components of the model you could run `anova(model.2`, which would give you F and DF for the various descriptors of your model.)