Session 3: Help! my data don't fit! Transformations to meet assumptions

In a fix? What to do!

In many cases, biological data do not even approximately conform to the assumptions of parametric tests. For a one-way ANOVA then you can simply use the non-parametric equivalent (Kruskal-Wallis test) but for more complex analyses such as two-way ANOVA, you are a bit stuck.

There are however, some things that you can do to try and deal with this, through what is known as **data transformation**.

This involves rescaling the data in some way to conform to the assumptions of the analysis (like normality of the residuals). It may seem like cheating to some extent but what it does it change the size of the 'gaps' between each data value, but not the actual order of values (from high to low).

Here we are going to look at the use of transformation in the context of a two-way ANOVA.

Download the `algae.csv` file from Moodle and import the data into a dataframe called `algae`. There are three columns (`particle_size`, `concentration` and `absorbance`). As for the previous example, as concentration is a number (check out `str(algae)` or inspect the data through the Environment tab), you need to convert it to a factor (f_conc) before proceeding.

These data come from a study of the effects of silver, as either nanoparticles or bulk sized particles on the growth of an algal species, *Scenedesmus*, as measured by the absorbance of light at 665nm, which reflects chlorophyll a concentrations.

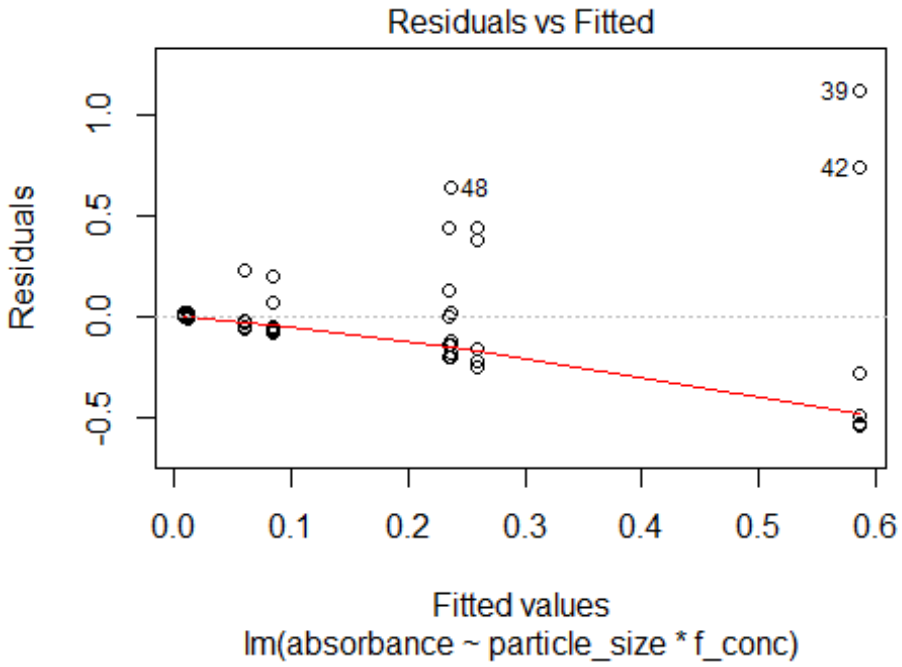
Four different mass doses of particles were used, and the absorbance of six replicates of each combination of particle size and concentration measured after one week of exposure.

The experimental design is similar to the previous session, with two factors (particle size and concentration) to consider, along with the interaction between them.

Following the pattern of the analysis in the last session, use the `lm` command to store the results of the analysis in an object called `a1`.

Now we need to check the assumptions. First the plots.

```
plot(a1, which=1)
```



Houston, we have a problem!

Even without going any further, there is clear evidence of a 'wedge'-like pattern in the residuals, with the spread of values on the y axis increasing from left to right. We could confirm this by doing the Levene test, but it is pretty obvious!

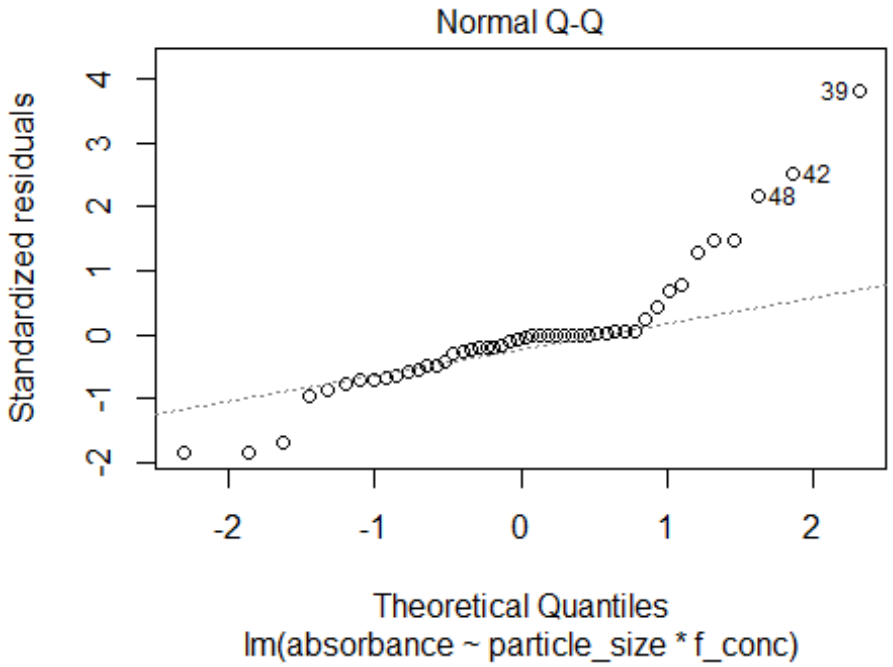
```
library(car)

leveneTest(absorbance~particle_size*f_conc, data=algae)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  7  2.2909 0.04627 *
##       40
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So there is a significant difference in the variance of the absorbance data in each treatment group. At this point we can't carry on with the ANOVA.

Just for completeness, we will check the normality of the residuals as well.

```
plot(a1, which=2)
```



The residuals do not look anything like normal (you can check formally by doing a Shapiro-Wilk test on the residuals if you want).

```
shapiro.test(a1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  a1$residuals
## W = 0.84932, p-value = 2.086e-05
```

A transformation

If data deviate substantially from the assumptions it can lead to false conclusions being drawn in relation to the hypotheses being posed. Time for a transformation!

A pattern in the residuals versus fits plot like the one shown would tend to suggest that a **logarithmic transformation** of the DV (absorbance) is appropriate. This is probably the most common transformation in biology. We shall try this to see if it helps meet the assumptions.

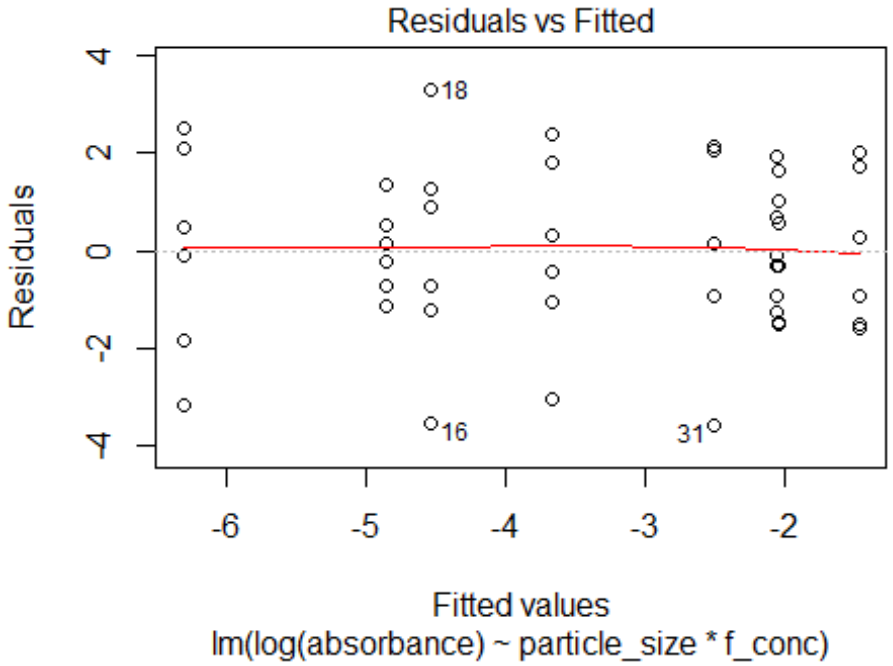
It is easy to add a transformation into the mix. We just specify the transformation within the code we use. The code for taking a log of a number is either `log` for natural or Naperian log or `log10` for log to the base 10. As you are studying at the institution closely associated with John Napier, who invented logarithms, we will wave the Napier flag and use the `log` option!

To keep things separate, we will store this in `a2`.

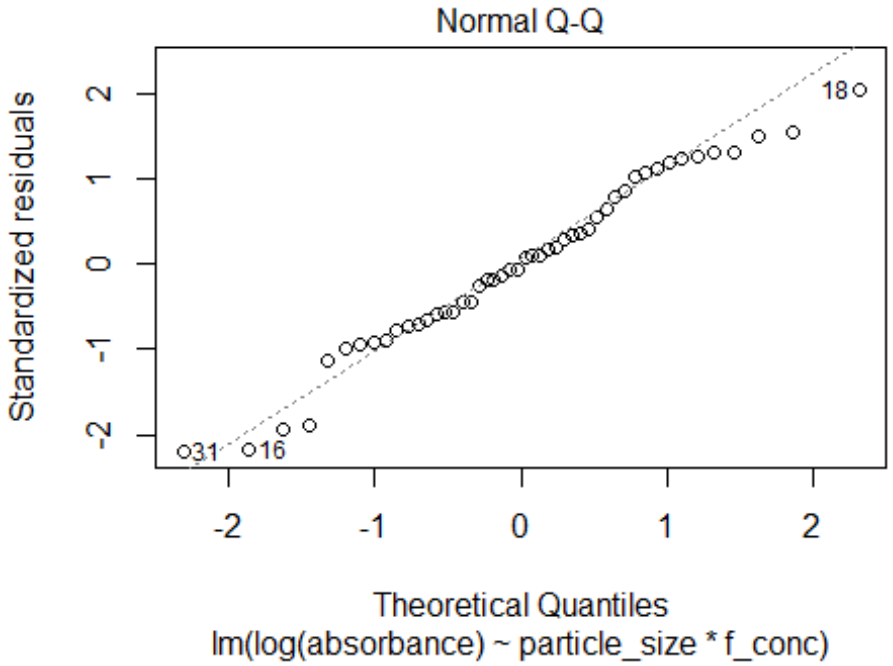
```
# the absorbance data is logged by using log() around the variable
# the rest is the same as before.
a2 <- lm(log(absorbance)~particle_size*f_conc, data=algae)
```

Repeat the plots and tests

```
plot(a2, which=1)
```




```
# need to specify the transformation in the code here too
leveneTest(log(absorbance)~particle_size*f_conc, data=algae)

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  7   0.989  0.453
##       40

plot(a2, which = 2)
```




```
shapiro.test(a2$residuals)

##
##  Shapiro-Wilk normality test
##
```

```
## data:  a2$residuals
## W = 0.97558, p-value = 0.4105
```

Everything is looking a bit more rosy now! As the assumptions have now been met, we can go ahead and look at the results of the analysis:

```
Anova(a2, type = 3)

## Anova Table (Type III tests)
##
## Response: log(absorbance)
##                       Sum Sq Df F value   Pr(>F)
## (Intercept)           12.916  1  4.0995  0.04961 *
## particle_size          1.025  1  0.3254  0.57155
## f_conc                90.929  3  9.6203 6.57e-05 ***
## particle_size:f_conc   8.977  3  0.9498  0.42577
## Residuals            126.023 40
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So there is no significant interaction, a significant effect of concentration and no significant effect of particle size (**make sure you can see this from the results**)

Perform Tukey post-hoc testing of the significant effect to determine where the difference are.

A plus

In this example, you performed a log transformation of your response or DV data for the ANOVA. If your data contained zero values, then you would not be able to transform them just using the log transformation as you cannot calculate the log of zero. In this case, you could use what is known as the log + 1 transformation. As the name suggests, this simply involves adding 1 to all values prior to taking logs. You would write this as shown

```
# put the +1 after the response variable
results <- lm(log(response+1)~factor1*factor2, data=the_data)
```

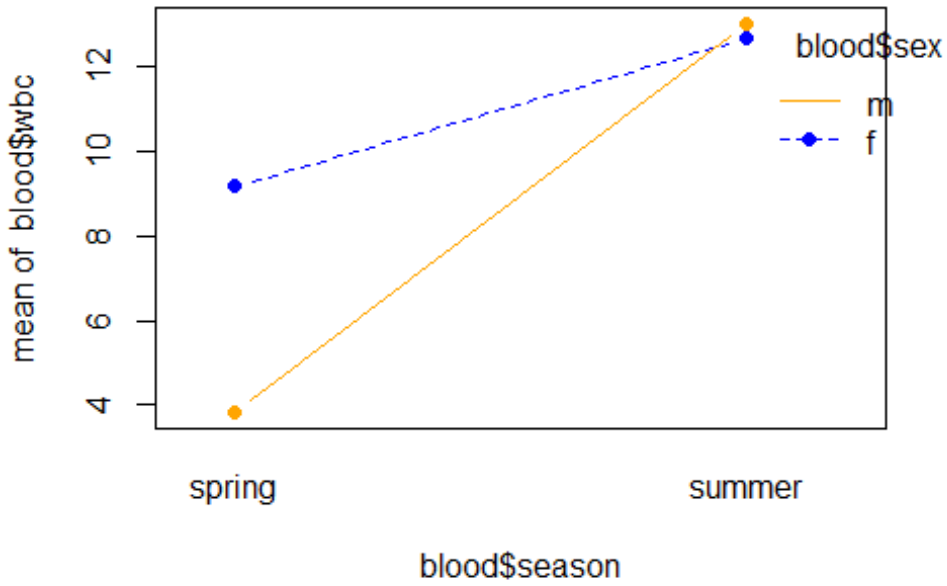
Getting interactive

Download the WBC.csv file from Moodle and import the data into `blood`. These data come from a study of the effect of season and sex on white blood cell (WBC) counts (cells per view in haemocytometer) in common sandpipers (*Actitis hypoleucos*).

Write code to analyse them using a two-way ANOVA and check the assumptions.

Here we have a significant interaction. Although the main effects (season and sex individually) are also significant, we can't interpret them directly as the significant interaction indicates that they do not work independently. In order to visualise what this means, we can use an interaction plot.

```
interaction.plot(blood$season, blood$sex, blood$wbc, type="b", pch=19,
col=c("Blue", "Orange"))
```



This gives you a visual assessment of the differences shown, but to be certain which means are significantly different we need to perform the Tukey post-hoc test. Here the interaction is significant, so you need to write it slightly different to before. Previously we specified which factor we were interested in comparing. This time we have the interaction to look at.

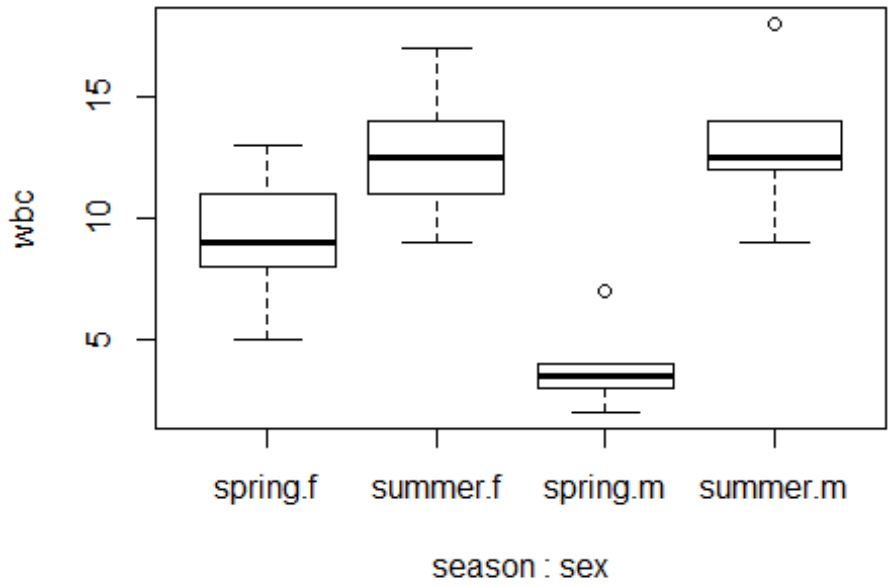
This is specified as follows:

```
# assuming that your results are stored in 'b1' - change this to
# whatever you have stored them in
TukeyHSD(aov(b1), which=c("season:sex"))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = b1)
##
## $`season:sex`
##                           diff         lwr       upr     p adj
## summer:f-spring:f  3.5000000  -0.7293892  7.729389 0.1277403
## spring:m-spring:f -5.3333333  -9.5627225 -1.103944 0.0104263
## summer:m-spring:f  3.8333333  -0.3960558  8.062723 0.0842904
## spring:m-summer:f -8.8333333 -13.0627225 -4.603944 0.0000561
## summer:m-summer:f  0.3333333  -3.8960558  4.562723 0.9960847
## summer:m-spring:m  9.1666667   4.9372775 13.396056 0.0000347
```

Rather a lot to go through in this case

```
# first do a boxplot to see what the data look like
# note that we can put both factors in here
boxplot(wbc~season*sex, data=blood)
```



Use the boxplot in conjunction with the outcome of the Tukey test to interpret the outcome and which means are significantly different.

If you are not clear on this, review the 'Two-way ANOVA interpretation' document on Moodle.

Further transformations

If you have whole number (integer) data, such as counts of species, cells etc. that are not normally distributed, then a square root transformation may be the appropriate choice. This can be done as follows.

```
# code to convert the data using square root transformation
data$sqrt_count <- sqrt(data$count)

# or do it when fitting the 'lm'
s1 <- lm(sqrt(count)~factor1*factor2, data=the_data)
```

If you have proportion or percentage data that are not normally distributed, then the appropriate transformation would be the arcsin (arcsine square root, or angular) transformation. Note that if you have percentages then you need to divide by 100 first to make them proportions.

This is the most complex of the common transformations to calculate as you have to do two steps (square root first and then taking the arcsine). An example is shown below.

```
# code to convert the data using arcsin
data$arcsine <- asin(sqrt(data$proportion))
```

```
# or do it when fitting the 'lm'
a1 <- lm(asin(sqrt(count))~factor1*factor2, data=the_data)
```

When to test for an interaction?

This is important. The potential for an interaction between the two factors should be something that should be considered at the point of experiment or study design, rather than something that is decided upon afterwards. Basically, there should be a **good biological reason** why you would expect the two variables being considered as factors to potentially work interactively on the values of your dependent variable. You should always remember that statistics are a tool to help you understand the biology, so the analysis should be underpinned by the biological reasoning which links to your hypothesis/es.

The second consideration is whether you have sufficient data to be able to reliably detect an interaction. It is not necessary to have exactly the same number of replicates in each combination of treatments, but if there is strong variation, and low numbers in certain combinations of your two factors, then you may end up with a non-significant interaction simply because you are limited in your ability to detect it, rather than it not existing.