# BIO782P ANOVA and Regression

*Rob Knell / Joe Parker*

*7 November 2018*

## ANOVA

Understanding how animals respond in the face of changing temperature is of obvious importance nowadays. One of the important questions that we need to understand is whether unpredictable changes in temperature have different effects from predictable, constant ones. This question was examined by Manenti ***et al.*** (2014) who described an experiment in which 25 isofemale lines of ***Drosophila simulans*** were reared in one of three treatments: a constant temperature (C) of 23°, a predictably fluctuating temperature (F) which rose to 28° during the day and declined to 13° at night, and an unpredictably fluctuating temperature (U) which rose to a randomly determined value between 23° and 28° during the day and fell to a randomly determined value between 23° and 13° at night. A variety of measurements were taken of flies from each treatment.

The file `Manenti_heat_shock.txt` contains data for one of these measurements: time to heat knock-down, which is the period for which the flies could withstand a temperature of 37.5° before becoming incapacitated. Each measurement is a mean value from about 10 measurements, each from a separate isofemale line.

1) Set up an object in R called "Heatshock" and read the data from the manenti_heatshock.txt text file into it using the `read.table()` function. You might find the chapter on importing data in Introductory R useful here.
2) Check that the data have been imported properly using the `str()` function. You should have a data frame with two variables: treatment, which indicates the temperature treatment that each group of insects was exposed to, and heatshock.time which is the mean time that insects from a particular line were able to withstand a high temperature for. Treatment should be a factor with three levels, heatshock.time should be a numeric variable.
3) To get an idea of what your data look like, draw a boxplot with the factor levels on the x-axis and longevity on the y-axis. NB to refer to the variables within the data frame you'll either need to attach the data frame or refer to them using the name of the data frame and then the name of the variable with a dollar sign between them, e.g. `Heatshock$treatment`.
4) What do you see? Is there anything that might make an ANOVA unsuitable for analysing these data as they are?
5) We're going to carry out an ANOVA on these data in three ways: by calculating it, by using the `aov()` function and by using the `lm()` function. Let's start by doing it the hard way, by hand.

Firstly you need to calculate the total sums of squares for heatshock.time. Remember that this is calculated as each number in the vector, minus the overall mean value for the vector and squared. Make sure you square the numbers before you sum them: the sum of the squared differences is not the same as the sum of the differences, squared. Set up an object called `SStotal` to store this number in.

Now calculate it a different way using the `var()` function and multiplying the variance by the degrees of freedom. The two numbers should be the same.

Now to calculate the error sums of squares. This is the residual variance left after the effect of the factor (in this case the temperature treatment) has been removed. You need to start by calculating the mean value for heatshock.time for each of the three factor levels: the easy way to do this is by using the `tapply()` function. Look up the help file and try to get it to work, and once you're there store the output as another object.

Once you have the means for each factor level then you can calculate the error sum of squares by working out the sum of the squared deviations from the factor mean for each factor level, and adding the sums of squares for each factor level together. Store this value as an object called `SSerror`.

What about the treatment sum of squares? Well, the total sums of squares is equal to the error sum of squares plus the treatment sums of squares, so we can work this last value out by subtracting `SSerror` from `SStotal`.

Having calculated these values we can fill in an ANOVA table.

| Source of Variation | df | Sum of squares | Mean squares | F | p |
|---|---|---|---|---|---|
| Treatment | | | | | |
| Error | | | | | |
| Total | | | | | |

Now that we know the calculated test statistic, F, we want to know if there is a significant effect. In other words, we want to know what the probability is of getting the observed value of F given that our degrees of freedom are Treatment df and Error df. R has all sorts of statistical distributions built in and we can calculate out probability using the `pf()` function. This gives us the probability density of F, in other words the probability of observing that value or a smaller one given the degrees of freedom. We want the probability of observing our F value or bigger, so we subtract the value `pf()` from one to give us our p-value:

```
# e.g.
#   F-value = 4.01
#   df1 (treatments) = 2
#   df2 (residuals)  = 17
F_value=4.01
df1=2
df2=17
# get probability density for this F-ratio with these d.f.
1-pf(F_value, df1,df2)
```

```
## [1] 0.03744327
```

6) Now use R to carry out an ANOVA on the same data using the `lm()` function. You'll need to set up a new object and allocate the output of the `lm()` function to it, and then you can get an ANOVA table using the `anova()` function. It should be the same as the one you've constructed above.

- What is your null hypothesis?
- What is the alternative hypothesis?
- What do you conclude from the results of your ANOVA?

7) Use `summary()` to produce a summary table for your fitted ANOVA model which you produced using `lm()`. Try to work out what the coefficients mean: you might find the section on sexual signalling in yeast in Introductory R (the latest version) helpful.

8) Use `plot()` to produce some diagnostic plots for your fitted ANOVA. Have a look: does your model have an acceptable fit?

9) What do you conclude?

Manenti, T., Sørensen, J.G., Moghadam, N.N., and Loeschcke, V. (2014). Predictability rather than amplitude of temperature fluctuations determines stress resistance in a natural population of *Drosophila simulans*. *J. Evol. Biol.* **27**:2113–2122.

## Regression

Cope's rule refers to the tendency of animal body sizes to increase in time with increasing age of the lineage. De Souza and Santucci (2014) tested this using a set of data from the dinosaur clade called the Titanosaurs. This is a group of sauropods that included some of the largest animals known in the history of the Earth (e.g. ***Argentinosaurus***) and which were widespread and diverse during the Cretaceous period. Most Titanosaurs are known from only a few bones, and the massive limb bones are particularly common. De Souza and Santucci assembled measurements of limb bones from 46 species of Titanosaur. Measurements of both humerus and femur were available for 20 of these, with 12 being represented only by the humerus and 14 by the femur only. In order to carry out an analysis of how the body size of these animals varied with time, it is necessary to estimate the femur sizes for those animals that are only represented by the humerus, and to do this we can use a linear regression to work out the expected values for each of these missing data points.

1. Load the dataset called Titanosaurs.txt into R – save it as an object called (for example) Titanosaurs. Use `str()` to check the structure of the new data frame – there should be a variable called "Taxa" which is the species name, the one called "Mean.time.MA" which is the age of the fossil in millions of years, then "Femur" and "Humerus" which are the lengths of the respective bones in metres.
2. Plot a scatterplot to show how the length of the two limb bones relate to each other. Remember we are going to try to predict femur length from humerus length so decide which measurement should go on the y- and x- axes on this basis.

Have you got a plot? Great - let's try out some lines for size. The R fuction for this is `abline(a,b)` - e.g lines of the form y = a + bx, where ***a*** is the y-intercept, and ***b*** is the coefficient (slope):

```
abline(0,1,col="grey")
# Not quite - try translating it up a bit
abline(0.25,1,col="grey")
# Maybe a bit shallower?
abline(0.25,0.8,col="light blue")
# Hmm..
```

3. Have a look at the data. Can you see anything that might mean that you should be cautious in using a linear regression on these data? If so, can you fix it in any way?
4. Fit a linear regression to the data using the `lm()` function and save the fitted model to an object in the R workspace.

You can bring up a summary of the fitted model using the `summary()` function as before. What information does this give you that you didn't have before?

5. An alternative way to check the statistical significance of a regression is to calculate how much of the variance in the data is explained by the regression line and compare that to the error variance in exactly the same way as we do for an ANOVA. You can do this by using the `anova()` function.
6. Check the diagnostic plots for the regression using `plot()` as before – in particular pay attention to the plot of residual versus predicted values. Do you see anything that causes concern?
7. Instead of fitting a straight line to your data, try fitting a curve. I've given you the code to fit a second order polynomial:

```
model3 <- lm(Femur~Humerus + I(Humerus^2))
```

8. Check your diagnostic plots and compare them with the plots for the simple linear model. Also have a look at the table produced by `summary()` – is the addition of the quadratic term justified?
9. Plot your data with the fitted curve by using the `curve()` function to add a curve to a scatterplot.
10. Now that you have calculated your slope and intercept you can use the equation for the curve that relates femur length to humerus length to workout what the predicted values are for the femurs of the titanosaur species that are only represented in this dataset by their humerus measurements.

De Souza, L.M., and Santucci, R.M. (2014). Body size evolution in Titanosauriformes (Sauropoda, Macronaria). ***J. Evol. Biol.*** **27**:2001–2012.