# Exercises for "When assumptions are not met"

Timothée Bonnet

May 2, 2019

Find the content for today and previous workshops at https://github.com/timotheenivalis/RSB-R-Stats-Biology.

## Contents

**1 "Error" with error distribution** **2**
Exercise 1 *Is that normality normal?* . . . . . 2
Exercise 2 *What is missing?* . . . . . 2

**2 Non-independence** **2**
Exercise 3 *Simpson's paradox* . . . . . 2
Exercise 4 *Genetic isolation distance or adaptation?* . . . . . 2

**3 DIY for complex issues** **2**
Exercise 5 *Measurement error* . . . . . 2
Exercise 6 *Non-linear model* . . . . . 3
Exercise 7 *When something is missing* . . . . . 3
Exercise 8 *Markovian process* . . . . . 4

# 1 "Error" with error distribution

### * Exercise 1 Is that normality normal?

Load the dataset `norm.csv`. It contains three response variables and one predictor. Fit a simple linear model for each reponse and check the properties of the residuals. What is wrong (or right)? What can you do about it?

### * Exercise 2 What is missing?

Load the dataset `boldness.csv`. We want to know how boldness relates to size and we fit a simple linear model of boldness as a function of size. Check the assumptions of such a model. What is wrong, how to fix it?

# 2 Non-independence

### * Exercise 3 Simpson's paradox

Load the dataset `thorndata.txt`. According to simple linear models, does thorns have an effect on herbivory? Is the effect consistent across sites (fit models on subsets)?

### ** Exercise 4 Genetic isolation distance or adaptation?

Load the dataset `genotype.csv`. Let's imagine that in a tree species we have measured allele frequencies at a gene that we suspect is related to thermal adaptation. Fit a simple linear model of AllFreq as a function of temperature. Check model assumptions. Are you confident this gene controls local adaptation to temperature?

# 3 DIY for complex issues

Knowing how to relax assumptions is a huge topic, pretty much synonymous to knowing how to do statistical modelling. We cannot cover all aspects in 2 hours, but we can train to understand what you may want, and learn the name of a few models, so that you can look them up later. So, without worrying about software and code, let's try and imagine what the models would look like (verbally or in equations).

### ** Exercise 5 Measurement error

You have measured body mass in small wild animals, in the field, on windy days. You are trying to quantify how much variation in mass explains survival to the next year. You model survival probability as $p_i \sim logit(\mu + M_i\beta)$ (and $Survial_i \sim Bernouilli(p_i)$), but you know $M$ are not the true masses, but only measurements with a lot or error... that violates a fundamental assumtion of your model. How to write a model that relaxes that assumption? (What new assumptions would you need then?)

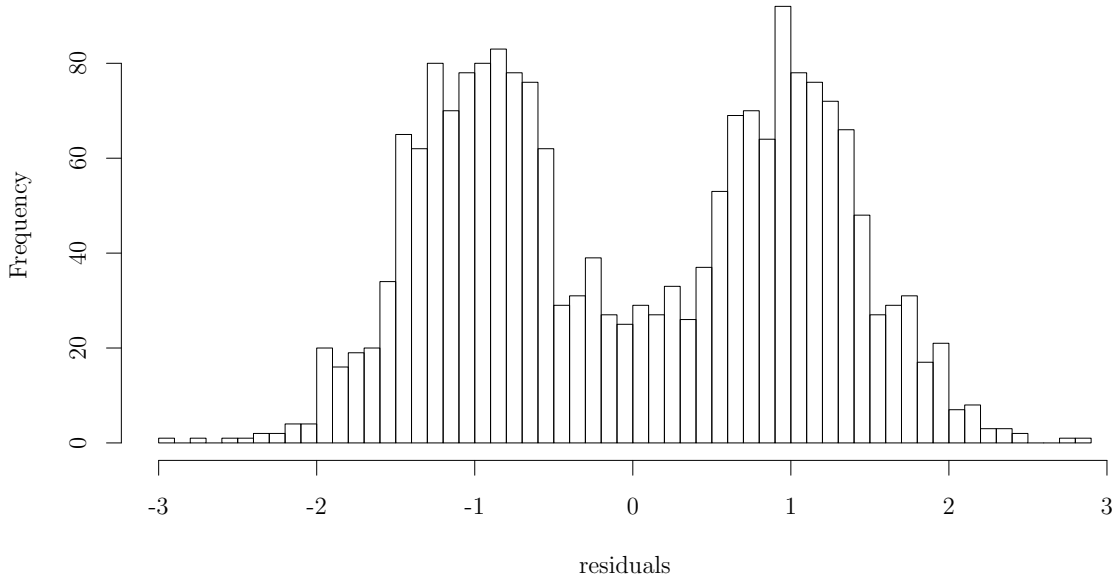
## * Exercise 6 Non-linear model

You study the effect of temperature on population growth rate in a species of yeast. You start from single cells in bottles and grow populations at 10 different temperatures (3 bottles per temperature) and measure population size at five different times. You know from litterature that population size ($P$) can often be modelled as $P(t) = \frac{K}{1+A\exp(-rt)}$ where $K$ is the carrying capacity, $A$ is a constant we do not care about, $t$ is time, and $r$ is the growth rate. At first you try to fit the equation above for every bottle, extract the estimate of $r$, and then correlate $r$ with temperature, but you quickly realize this is not ideal, because the estimate of $r$ and all other parameters are very imprecise (you have only 5 points per bottle, and population count data are noisy!) and the error in the estimates of $r$ are not accounted for.

How would you write a model to estimate the effect of temperature on population growth? What assumptions will need?

## ** Exercise 7 When something is missing

You study diet in a population of sea elephants on their colony by looking at the isotopic composition of Nitrogen in their blood (different preys are differently enriched in some isotops, so you can tell whether an animal rather eats a lot of invertebrates or a lot of fish). You suspect animals migrate to two different areas which are known to differ in food resources (one is fish-rich, the other krill-rich), but have no way to observe them there and there are no data on who goes where. You fitted a linear model of concentration in nitrogen-15 with age and sex as covariates. You obtained this distribution for your residuals:




What model could you fit to get better residuals, and learn something new? What

new assumptions do you need?

## ** Exercise 8 Markovian process

What if the current state of your response variable depends heavily on the previous state. For instance, you study kangaroo movements and have GPS locations every 2 minutes. You try to model habitat selection, but realize residuals are far from independent in your model... How to relax assumptions, what new assumptions would you make?