Models in MORSE package

December 21, 2015

This document describes the statistical models used in MORSE to analyze survival and reproduction data, and as such serves as a mathematical specification of the package. For a more practical introduction, please consult the "Tutorial" vignette; for information on the structure and contents of the library, please consult the reference manual.

Model parameters are estimated using Bayesian inference, where posterior distributions are computed from the likelihood of observed data and prior distributions on the parameters. These priors are specified after each model description.

1 Survival bioassays

In a survival bioassay, subjects are exposed to a controlled dose of a contaminant over a given period of time and the number of surviving individuals is measured at certain time points during exposition. In most standard bioassays, the dose is held constant throughout the whole experiment, which we will assume here. An experiment is generally replicated several times and also repeated for various levels of the contaminant.

In so-called *final time* bioassays, the mortality is measured only at the end of the experiment. The chosen time point is called *target time*. Let us see how this particular case is handled in MORSE.

1.1 Analysis of final time survival bioassays

A dataset from a final time survival bioassay is a collection $D = \{(c_i, n_i^{init}, n_i)\}_i$ of experiments, where c_i is the tested concentration, n_i^{init} the initial number of individuals and n_i the number of individuals at the target time. Triplets such that $c_i = 0$ correspond to control experiments.

Modelling In the particular case of endpoint assays, the model used in MORSE is defined as follows. Let t be the target time in days. We suppose the *mean survival rate after t days* is given by a function f of the contaminant level c. We also suppose that the death of two individuals are two independent events. Hence, given an initial number n_i^{init} of individuals in the bioassay, we obtain that the number N_i of surviving individuals at time t follows a binomial distribution:

$$N_i \sim \mathcal{B}(n_i^{init}, f(c_i))$$

Note that this model neglects inter-replicate variations, as a given concentration of pollutant implies a fixed value of the survival rate. There may be various possibilities for f. In MORSE we assume:

$$f(c) = \frac{d}{1 + (\frac{c}{e})^b}$$

where b, e and d are (positive) parameters. In particular d corresponds to the survival rate in absence of pollutant and e corresponds to the LC_{50}. The parameter b is related to the effect intensity of the contaminant.

Inference Posterior distributions for parameters b, d and e are estimated using JAGS with the following priors:

- we assume the range of tested concentrations in an experiment is chosen to contain the LC_{50} with high probability. More formally, we choose:

$$\log_{10} e \sim \mathcal{N}(\frac{\log_{10}(\min_i c_i) + \log_{10}(\max_i c_i)}{2}, \frac{\log_{10}(\max_i c_i) - \log_{10}(\min_i c_i)}{4})$$

  which implies e has a probability slightly higher than 0.95 to lie between the minimum and the maximum tested concentration.

- we choose a quasi non-informative prior distribution for the shape parameter b:

$$\log_{10} b \sim \mathcal{U}(-2, 2)$$

The prior on d is chosen as follows: if we observe no mortality in control experiments then we set $d = 1$, otherwise we assume a uniform prior for d between 0 and 1.

2 Reproduction bioassays

In a reproduction bioassay, we observe the number of offspring produced by a population of adult individuals subjected to a certain dose of a contaminant over a given period of time. The offspring (young individuals, clutches or eggs) are regularly counted and removed from the medium at each observation, so that the reproducing population cannot increase. It can decrease however, if some individuals die during the experiment. The same procedure is usually repeated with various concentrations of contaminant, in order to establish a quantitative relationship between the reproduction rate and the concentration of contaminant in the medium.

As mentionned already, it is often the case that part of the individuals of an bioassay die during the observation period. In previous approaches, it was proposed to consider the cumulated number of reproduction outputs without accounting for mortality [2, 3], or to exclude replicates where mortality occurred [4]. However, individuals may have reproduced before dying and thus contributed to the observed response. In addition, individuals dying the first are the most sensitive, so the information on reproduction of these prematurely dead individuals is valuable and ignoring it is likely to bias the results in a non-conservative way. This is particularly critical at high concentrations, when mortality may be very high.

In a bioassay, mortality is usually regularly recorded, *i.e.* at each timepoint when reproduction outputs are counted. Using these data, we can approximately estimate for each individual the period it has stayed alive (which we assume coincides with the period it may reproduce). As commonly done in epidemiology for incidence rate calculations, we can then calculate, for one replicate, the total sum of the periods of observation of each individual before its death (see next paragraph). This sum can be expressed as a number of individual-days. Hence, reproduction can be evaluated through the number of outputs per individual-day.

In the following, we denote M_{ijk} the observed number of surviving individuals for concentration c_i, replicate j and time t_k.

2.1 Estimation of the effective observation time

We define the effective observation time as the sum for all individuals of the time they spent alive in the experiment. It is counted in individual-days and will be denoted NID_{ij} for concentration c_i and replicate j. As mentionned earlier, mortality is observed at particular time points only, so the real life time of an

individual is unknown and in practice we use the following simple estimation: if an individual is alive at t_k but dead at t_{k+1}, its real life time is approximated as $\frac{t_{k+1}+t_k}{2}$.

With this assumption, the effective observation time for concentration c_i and replicate j is then given by:

$$NID_{ij} = \sum_k M_{ij(k+1)}(t_{k+1} - t_k) + (M_{ijk} - M_{ij(k+1)})(\frac{t_{k+1} + t_k}{2} - t_k)$$

2.2 Target time analysis

In this paragraph, we describe our so-called "target time analysis", where we model the cumulated number of offspring up to a target time as a function of pollutant concentration and effective observation time in this period (cumulated life times of all individuals in the experiment, as described above). A more detailed presentation can be found in [1].

We keep the convention that the index i is used for concentration levels and j for replicates. The data will therefore correspond to a set $\{(nid_{ij}, n_{ij})\}_i$ of pairs, where nid_{ij} denotes the effective observation time and n_{ij} the number of reproduction output. These observations are supposed to be drawn independently from a distribution that is a function of the level of contaminant c_i.

Modelling We assume here that the effect of the considered contaminant on the reproduction rate[1] does not depend on the exposure time, but only on the concentration of the contaminant. More precisely, the reproduction rate in an experiment with a concentration c_i of contaminant is modelled by a three-parameter log-logistic model, that writes as follows:

$$f(c;\theta) = \frac{d}{1 + (\frac{c}{e})^b} \quad \text{with} \quad \theta = (e, b, d)$$

Here d corresponds to the reproduction rate in absence of contaminant (control condition), and e to the value of the EC_{50}, that is the concentration dividing the average number of offspring by two with respect to the control condition. Now the number of reproduction outputs N_{ij} for concentration c_i in replicate j can be modelled using a Poisson distribution:

$$N_{ij} \sim Poisson(f(c_i;\theta) \times NID_{ij})$$

This model is later referred to as "Poisson model". If there happens to be a non-negligible variability of the reproduction rate between replicates for a some fixed concentration, we propose a second model, named "gamma-Poisson model", stating that:

$$N_{ij} \sim Poisson(F_{ij} \times NID_{ij})$$

where the reproduction rate F_{ij} for at c_i in replicate j is a random variable following a gamma distribution. Introducing a dispersion parameter ω, we assume that:

$$F_{ij} \sim gamma(\frac{f(c_i;\theta)}{\omega}, \frac{1}{\omega})$$

Note that a gamma distribution of parameters α and β has mean $\frac{\alpha}{\beta}$ and variance $\frac{\alpha}{\beta^2}$, that is here $f(c_i;\theta)$ and $\omega f(c_i;\theta)$ respectively. Hence ω can be considered as an overdispersion parameter (the greater its value, the greater the inter-replicate variability)

[1] that is, the number of reproduction outputs during the experiment per individual-day

Inference Posterior distributions for parameters b, d and e are estimated using JAGS with the following priors:

- we assume the range of tested concentrations in an experiment is chosen to contain the EC_{50} with high probability. More formally, we choose:

$$\log_{10} e \sim \mathcal{N}(\frac{\log_{10}(\min_i c_i) + \log_{10}(\max_i c_i)}{2}, \frac{\log_{10}(\max_i c_i) - \log_{10}(\min_i c_i)}{4})$$

  which implies e has a probability slightly higher than 0.95 to lie between the minimum and the maximum tested concentration.

- we choose a quasi non-informative prior distribution for the shape parameter b:

$$\log_{10} b \sim \mathcal{U}(-2, 2)$$

- as d corresponds to the reproduction rate without contaminant, we set a normal prior $\mathcal{N}(\mu_d, \sigma_d)$ using the data:

$$\mu_d = \frac{1}{r_0} \sum_j \frac{n_{0j}}{nid_{0j}}$$

$$\sigma_d = \sqrt{\frac{\sum_j (\frac{n_{0j}}{nid_{0j}} - \mu_d)^2}{r_0(r_0 - 1)}}$$

  where r_0 is the number of replicates in the control condition. Note that since they are used to estimate the prior distribution, the data from the control condition are not used in the fitting phase.

- we choose a quasi non-informative prior distribution for the ω parameter of the gamma-Poisson model:

$$log_{10}(\omega) \sim \mathcal{U}(-4, 4)$$

For a given dataset, the procedure implemented in MORSE will fit both models (Poisson and gamma-Poisson), and use an information criterion known as Deviance Information Criterion (DIC) to choose the most appropriate. In situations where overdispersion (that is inter-replicate variability) is negligible, using the Poisson model will provide more reliable estimates. That is why a Poisson model is preferred unless the gamma-Poisson model has a sufficiently lower DIC (in practice we require a difference of 10).

References

[1] Marie Laure Delignette-Muller, Christelle Lopes, Philippe Veber, and Sandrine Charles. Statistical handling of reproduction data for exposure-response modeling. *Environmental science & technology*, 48(13):7544–7551, 2014.

[2] OECD. Guidelines for testing of chemicals n.220. *Enchytraeid* reproduction test. Technical report, Organisation for Economic Cooperation and Development, 2004.

[3] OECD. Guidelines for testing of chemicals n.226. Predatory mite (*Hypoapsis (Geolaelaps) aculeifer*) reproduction test in soil. Technical report, Organisation for Economic Cooperation and Development, 2008.

[4] OECD. Guidelines for testing of chemicals n.211. *Daphnia magna* reproduction test. Technical report, Organisation for Economic Cooperation and Development, 2012.