Additional exercise Ch. 6

Bob Douma

22 November 2017

Fitting models to data

In this exercise you will learn how to fit models to data through means of maximum likelihood and compare the likelihood of different models (hypotheses). Fitting a model to data through likelihood requires that you take four steps:

1. Specify how the dependent variable depends on the independent variable, i.e. specify a function how the mean of y depends on the value of x.
2. Specify a probability distribution to describe the deviations of the observations from the mean
3. Choose the parameters of the deterministic model and the probability model such that the negative log likelihood is lowest.
4. Compare the likelihood of alternative models (change the deterministic function or the stochastic function) and compare with AIC(c) or BIC which model is most parsimonious.

To fit a model through means of maximum likelihood you need to specify a function that calculate the negative log likelihood (NLL) based on the data and the parameter values. For example to calculate the NLL of a linear model and a normal distribution the following function works:

```
nll = function(par,y,x){
  a = par[1]
  b = par[2]
  sd = par[3]
  # this calculates the mean y for a given value of x: the deterministic function
  mu = a+b*x
  # this calculates the likelihood of the function given the probability
  # distribution, the data and mu and sd
  nll = -sum(dnorm(y,mean=mu,sd=sd,log=T))
  return(nll)
}
```

Note that `-sum(log(dnorm(y,mean=mu,sd=sd)))` should not be used as it may lead to underflow (the computer cannot store very very small probabilities) and therefore to optimisation problems.

Next we specify a function to find the maximum likelihood estimate

```
par=c(a=1,b=1,c=1) # initial parameters
opt1 = optim(par=par,nll,x=x,y=y) # y represents the data, x the independent variable
```

It can also be done through `mle2`

```
nll.mle = function(a,b,sd){
  # this calculates the mean y for a given value of x: the deterministic function
  mu = a+b*x
  # this calculates the likelihood of the function given the probability
  # distribution, the data and mu and sd
  nll = -sum(dnorm(y,mean=mu,sd=sd,log=T))
  return(nll)
}
```

```r
# the data should be supplied through data and the parameters through list().
mle2.1 = mle2(nll.mle,start=list(a=1,b=1,sd=1),data=data.frame(x,y))
summary(mle2.1)
```

The following steps will lead you through the model fitting procedure.

1. Take the first dataset and tweak the above functions such that it matches with the deterministic and stochastic model that you have chosen. In case you got stuck in the previous exercises where you had to choose a deterministic function and a stochastic function see next page for suggestions.

   hint: In a previous exercise you have eyeballed the parameter values of the functions, you can use these as starting values.

   hint: In case you get convergence problems, further adapt your starting values, or choose a different optimizer. For example Nelder-Mead is a robust one, e.g. `method = "Nelder-Mead"`.

2. Change the determinstic function for a possible alternative determinstic function
3. Compare the likelihoods of the data given both models
4. Apply model selection criteria and conclude which model fits that data best.
5. Does the model makes sense from a biological perspective?

Optional and time permitting:
6. Repeat the above procedure for the other 5 datasets

Hints for choosing deterministic functions and stochastic functions

1. Deterministic functions

   dataset 1 light response curve. There are a number of options of functions to choose from, depending on the level of sophistication: $\frac{ax}{(b+x)}$, $a(1-e^{(-bx)})$, $\frac{1}{2\theta}(\alpha I + p_{max} - \sqrt{(}\alpha I + p_{max})^2 - 4\theta I p_{max})$ see page 98. A parameter `d` can be added in all cases to shift the curve up or down.

   dataset 2 The dataset describes a functional responses. Bolker mentions four of those $min(ax, s)$ $\frac{ax}{(b+x)}$, $\frac{ax^2}{(b^2+x^2)}$, $\frac{ax^2}{(b+cx+x^2)}$

   dataset 3 Allometric relationships generally have the form ax^b

   dataset 4 This could be logistic growth $n(t) = \frac{K}{1+(\frac{K}{n_0})e^{-rt}}$ or the gompertz function $f(x) = e^{-ae^{-bx}}$

   dataset 5 What about a negative exponential? $ae{-}bx$ or a power function ax^b

   dataset 6 Species reponse curves are curves that describe the probability of presence as a function of some factor. A good candidate good be a unimodel response curve. You could take the equation of the normal distribution without the scaling constant: e.g. $ae^{\frac{-(x-\mu)^2}{2\sigma^2}}$

2. Stochastic functions/Probability distributions

   dataset 1 y represents real numbers and both positive and negative numbers occur. This implies that we should choose a continuous probability distribution. In addition, the numbers seems unbound. Within the family of continuous probability distributions, the normal seems a good candidate distribution because this one runs from -inf to +inf. In contrast the Gamma and the Lognormal only can take positive numbers, so these distributions cannot handle the negative numbers. In addition, the beta distribution is not a good candidate because it runs from 0-1.

   dataset 2 y represents real numbers and only positive numbers occur. The data represents a functional response (intake rate of the predator), and it is likely that you can only measure positive numbers (number of prey items per unit of time). This implies that we should choose a continuous probability distribution. Within the family of continuous probability distributions, the Gamma and the Lognormal could be taken as candidate distributions because they can only take positive numbers (beware that the Gamma cannot take 0). However, you could try to use a normal as well.

   dataset 3 y seems represents counts (this is the cone dataset that is introduced in ch. 6.). Given that it contains counts we can pick a distribution from the family of discrete distributions. The Poisson and the Negative Binomial could be good candidates to describe this type of data.

   dataset 4 y represents population size over time. From looking at the data, they seems to represent counts. Given that it contains counts we can pick a distribution from the family of discrete distributions. The Poisson and the Negative Binomial could be good candidates to describe this type of data.

   dataset 5 No information is given on y. The data clearly seems to represent counts. Thus the same reasoning applies here as to the two previous datasets.

   dataset 6 The data (y) represents species occurences (presence/absence). The binomial model would be a good model to predict the probability of presence.