Maximum likelihood

Data

The dataset thermal_range.csv is the result of an experiment to determine the effect of temperature (*temp*) on the number of eggs (*num_eggs*) produced by a species of mosquito. Three replicates were measured for temperature values between 10 and 32 degrees Celsius.

```
therm <- read.csv("../donnees/thermal_range.csv")
head(therm)
```

```
##   temp num_eggs
## 1   10        1
## 2   10        1
## 3   10        2
## 4   12        4
## 5   12        4
## 6   12        6
```

1. Poisson model

We assume that the mean number of eggs produced follows a Gaussian function centered on an optimal temperature. (This function has the same shape as a normal distribution, but it is not a probability.)

$$N = N_o \exp\left(-\frac{(T - T_o)^2}{\sigma_T^2}\right)$$

In this equation, N is the mean number of eggs produced at T, T_o is the optimum temperature, N_o is the number of eggs produced at T_o and σ_T represents the tolerance (the larger σ_T is, the slower the production decreases around the optimum).

a) Produce a graph of *num_eggs* vs. *temp* for this dataset. Using this graph, answer the following questions.

- Does the number of eggs appear to reach a maximum with a symmetrical decrease on both sides of the maximum as predicted by the equation above?
- Does the variance between replicates appear to be homogeneous?

b) To model these data, we will use a Poisson distribution, where the mean λ corresponds to the N calculated according to the above equation.

Why is it preferable here to use a Poisson distribution rather than a normal distribution to represent the random variation in the number of eggs around the predicted mean value?

c) Create an R function to calculate the negative log-likelihood of the observed number of eggs as a function of the model parameters N_o, T_o and σ_T. *Note*: The function `dpois(y, lambda, log = TRUE)` is used to compute the log of the probability of a vector of data `y` following a Poisson distribution with a vector of means `lambda`.

d) Use the `mle2` function to estimate the three parameters of the model according to maximum likelihood.

For this problem, it is necessary to specify bounds for each parameter, to prevent the optimizer from moving too far away from plausible values. In the function `mle2`, the lower and upper bounds are given by the arguments `lower` and `upper`, e.g.: `mle2(..., start = list(...), lower = c(no = 1, to = 5, s_t = 1), upper = c(...)`. Note that these arguments are specified by a vector `c(...)` whereas `start` (the initial values) are specified by a list.

You can try different values for the bounds, however the lower bounds of N_o and σ_T should be at least 1, the upper bound of σ_T should not exceed the range (max-min) of temperatures tested; likewise, the bounds for T_o should be realistic values of temperature.

Note: You can ignore the warning *Warning: bounds can only be used with method L-BFGS-B (or Brent)*. However, if you get an error, try again by adjusting the parameter bounds.

e) Visualize the profile likelihood for each parameter and calculate the 95% confidence intervals. Do you think the quadratic approximation would be good for these parameters?

f) Let us now compare the model predictions with the data.

- Add a column to the dataset for the model mean predictions (the λ of the Poisson model for each observation), obtained by replacing the maximum likelihood estimates in the N equation above.
- Simulate 1000 data sets from the Poisson distribution with the estimated λ values. To generate a dataset, use `rpois(n, lambda)` where n is the number of observations (the number of rows in the original dataset) and λ is the column of mean predictions. To generate 1000 datasets, use `replicate`. The result of `replicate` should be a matrix of n rows and 1000 columns (1 column per simulation).
- To obtain a 95% prediction interval for each observation, calculate the appropriate quantiles for each row of the matrix of simulations with `apply`. For example, `apply(sim_mat, 1, quantile, prob = 0.025)` applies the `quantile` function to each row of `sim_mat`, with the `prob` argument of `quantile` set to 0.025. Do the same for the quantile at $p = 0.975$ and you will get two vectors for the lower and upper bounds of the interval, which you can add to the dataset.

Note: These prediction intervals assume that the parameter estimates are exact and therefore ignore their uncertainty.

- At this point, in addition to `temp` and `num_eggs`, your dataset contains three columns respectively representing the mean predictions, and the lower and upper bounds of the 95% prediction interval for each observation. Add the mean prediction and the interval to the graph of *num_eggs* vs. *temp*, e.g. with `ggplot`, you can add `geom_line(aes(y = mean_pred))` to the graph to add a line representing the `mean_pred` column of mean predictions, same for the lower and upper bounds of the interval.

From the results, can you tell whether the model represents the general trend of the data and the random variation around that trend?

2. Negative binomial model

a) Repeat parts (c) to (f) of the previous number using a negative binomial distribution instead of the Poisson distribution for the random portion of the model.

Reminder: In the Poisson distribution, the mean and variance are equal to λ. In the negative binomial distribution, the mean is equal to μ and the variance is equal to $\mu + \mu^2/\theta$. For this problem, we will use $k = 1/\theta$ as parameter. If $\theta > 0$, k must take a value greater or equal to 0. Since the variance as a function of k is $\mu + k\mu^2$, the Poisson distribution corresponds to the case $k = 0$. Here are the main changes to be made to replace the Poisson model with the negative binomial model:

- Add the parameter k to the log-likelihood function. Replace the call to `dpois` with `dnbinom(y, mu, size = 1/k, log = TRUE)` where `mu` is the mean prediction, so it is equivalent to the `lambda` of `dpois`.
- Use a lower bound of 0 for the `k` parameter in `mle2`; the upper bound should be less than 100.

- To simulate the data, replace `rpois` with `rnbinom` and specify the arguments `mu` (mean prediction) and `size = 1/k`.

b) Would it be correct to use the likelihood-ratio test to compare the Poisson model from the previous section to the negative binomial model?

c) Whether or not you answer yes or no to (b), do the results clearly show whether or not a negative binomial model is justified (relative to the simpler Poisson model) for these data?