M0019 Alternative assessed Practical 2020

Only do this project if you have discussed this with Arne Kovac in advance. It is worth 20% of the marks for this unit.

To compensate for the effects of strike action, this practical has been simplified and made more prescriptive than usual.

The data loaded with the command

```
bugs <- read.table("https://people.maths.bris.ac.uk/~sw15190/TOI/bugs.dat",header=TRUE)
```

give data on counts of a pest aphid on a farm over the course of 41 days. The data are from standard traps set up in the fields, and monitored daily.

Suppose that you are working for a company specialising in agricultural data science analysing these data for the farm in order to offer advice. Early in an agricultural pest outbreak a simple model is that the pest population will grow exponentially, so that, for daily data we have an iteration for the pest population N, of

$$N(t) = e^r N(t-1)$$

where r is a parameter relating to the pest reproductive rate. *The farm would like to know if there is any indication that the rate of growth is slowing over time, indicating that the measures taken to combat the aphids are working.*

A simple model that allows for departure from unbounded exponential growth is

$$N(t) = e^{\sum_{i=1}^{K} \beta_i t^{i-1}} N(t-1)$$

where K has to be selected and the β_i are parameters to be estimated (along with the initial N). A reasonable model for the observed data is that the number of aphids at time t is $y_t \sim \text{Poi}(N(t))$, with the y_t being independent.

You should do the following, obviously checking results as you go to make sure the answers are sensible.

1. Write R code to evaluate the negative log likelihood of the model given above suitable for optimization using `optim`. Make your code sufficiently general that you can control K simply by changing the dimension of the parameter vector that you supply to your function. Notice that the model is invariant to the units that time is measured in within the exponential: all linear rescalings of time are equivalent there. Within the exponential term it is best to have time run from -1 to 1 in equal steps (see R function `seq`, for example): this will ensure that your estimated β_i are 'nicely scaled' giving good numerical performance from `optim`.

2. The simplest way to select K is to use AIC covered in section 8.9 of the notes. Fit models with values of K from 1 to 5 and select the best one by AIC. Check that the selected model could not be further simplified using a GLRT.

3. Given the non-linear dependence of the model $N(t)$ on the model parameters, it is not completely straightforward to obtain confidence intervals for $N(t)$ (at all the observed times, of course). One approach is non-parametric bootstrapping. Use this to generate a set of 100 bootstrap replicate $N(t)$ trajectories and overlay these (see R command `lines`) on a plot of the raw data. A simple way of implementing non-parametric bootstrapping is to sample the indices of the data, counting how many times each index is sampled and using these counts to weight the sum of log-likelihood contributions - other approaches are also possible.

4. The results from the previous part seem very variable. Why is this and is the variability a sensible reflection of the uncertainty here?

5. An alternative bootstrapping approach is *parametric bootstrapping*. Take the estimated $N(t)$ from your best fit model and generate a complete set of new data from $y_t^* \sim \text{Poi}(\hat{N}(t)) \; \forall \; t$. Refit the model to each set of bootstrap data generated in this way. Again overlay 100 bootstrap replicate $N(t)$ trajectories on a plot of the original data, and comment.

What to hand in.

1. A 4 page report (A4, normal margins, at least 10pt font, PDF file) consisting of 2 sections.
   (a) A one page summary for farmers, addressing the key question of whether there is evidence for the rate of spread slowing, or not and how reliable the evidence is. Make sure to be careful about what it is important for the farmer to know, and what is 'too much detail'.
   (b) A three page (maximum) report explaining what you did statistically for other statisticians. This should include no, or minimal, R code. The idea is that you explain what you did and why and what you concluded so that a competent statistician could replicate it for themselves, and could judge how well founded your conclusions are based on the results presented.

   Your report should not assume that the reader has seen this assignment sheet.
2. A text file containing the R code you used, carefully structured and commented. Again a statistician should be able to take your code and use it to replicate your analysis, while understanding what it is doing.

One report (as a pdf file) and one R code file (plain text is best) per group should be emailed to `simon.wood@bristol.ac.uk` with the subject `M0019 APHIDS` followed by your surnames, by the deadline on the course web page (unless you have been given an alternative date).

Mark scheme guidance

First class marks will be awarded for work that could be passed on to the farm essentially without modification. That is to say the statistics is appropriate and clearly explained, the conclusions appropriately drawn and any limitations are discussed fairly.

Upper second class marks will be awarded for work that could be passed on to the farm, after a round of revision correcting some errors of presentation, interpretation or statistics that are relatively minor.

Lower second class marks will be awarded to work that has some more substantial flaws of presentation, interpretation or statistical reasoning which would require some more work to correct.

Third class marks will be awarded for work that contains some indication of substantive understanding and engagement, but contains more serious errors and misunderstandings.