Data Analysis Prelim Exam

Department of Statistics, University of Chicago

Assigned Wednesday Sept 16 2020 at 4pm; Due Friday Sept 18 2020 at 4pm (Chicago time)

1 Data

The data set for this exam is taken from the paper

> *Shifts in timing and duration of breeding for 73 boreal bird species over four decades*, Hällfors et al, *PNAS* vol. 117, no. 31, pages 18557–18565 (2020)

The data set consists of *ringing records* for birds across 73 species in Finland from year 1975–2017. Ringing is when a human volunteer places a metal ring on a the leg of a bird living in the wild; these rings are marked and can be used to track the bird population (see Figure 1 for how ringing is performed). Ringing can only be performed on baby birds during a narrow window of age and is also dependent on conditions such as weather. This data set records the date, location, and bird species for all ringing events during the range of years studied. The dates of the ringing records can then be used to study the timing of the breeding season of the birds.

The data files are downloaded from `https://datadryad.org/stash/dataset/doi:10.5061/dryad.wstqjq2ht`.

- `73_species.csv` contains the ringing records.
- `Traits_73_species.csv` contains some information about each of these 73 species.
- `README_RingingDataBorealBirds.txt` contains some additional details (how data were gathered, etc).

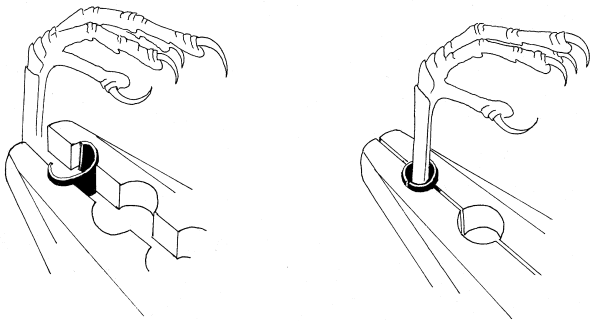

Figure 1: How ringing is performed (source: `http://safring.birdmap.africa/downloads/ring-manual-05.pdf`)

2 Questions

Your report should include code for all plots and results that you present. The report should be *concise, well-organized, and easy to read.* For example, if you examine plots for each individual bird species, you may choose to show only a small selection of representative plots.

1. Please read the original PNAS paper (including the supplementary information appendix) and briefly summarize the design and goals of this study. Include a high-level description of the data collection, the modelling approach used, and the main conclusions drawn by the authors. Highlight some of the main statistical issues, as you see them, that arose during the analysis.

2. In Figure 2 of Hällfors et al, left-hand panel, the authors show eight estimated densities (4 species in each of 2 different years) of ringing events across the seasons. Several of these estimated densities are multi-modal.

   Attempt to produce versions of these eight plots from the same data, improving them where possible. Comment on whether the published density estimates accurately represent the raw observed data. Do you think the multi-modal features in the published density estimates likely represent clustering in the actual nesting behavior of the birds? Are there other possible explanations?

3. In Supplementary Text S2, entitled "Assessing change in ringer behavior", the authors discuss whether ringer behavior and effort may change over time. In particular they examined changes in wing length (Fig S7) [the data for which are not included in the dataset we used here] and changes in total numbers of ringing events (Figures S9-S10) over the years of the study.

   (a) Explain, briefly, why the authors examined changes in wing length over time (Figure S7).

   (b) Explain how changes in ringing effort over time, if they existed, could impact the main results and conclusions of the study. Summarize the arguments that the authors make, based on Figures S9-S10, that changes in ringing effort are unlikely to impact their study. Discuss the strengths and limitations of the argument. What assumptions would make it possible to distinguish changes in ringing effort from changes in bird behavior?

4. For this question, we will work with the *duration* of the breeding season as the response variable, which (as in the paper) is defined as the number of days between the 5th and 95th percentile of ringing events. For the species CORRAX, use a linear regression to assess whether the duration of the breeding season is showing a decrease in length over time. (For this problem, you should treat duration as the observed response, and ignore the fact that it was computed based on individual ringing events. However, you may choose to include other information such as boreal zone in your analysis.) Discuss your findings.

5. For this question, we will also work with the *duration* of the breeding season as the response variable, and consider the linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{1}$$

   where y_i is the estimated duration of the breeding season for year i and x_i is corresponding time variable for the i-th year (e.g., $x_i = 1990$). We focus on fitting this model for the species HIRRUS in the south boreal zone.

   The following histogram (Figure 2) reveals that, for this species-zone combination, the number of observed ringing events in each year has a dramatic increase after year 1997. Explain why this fact

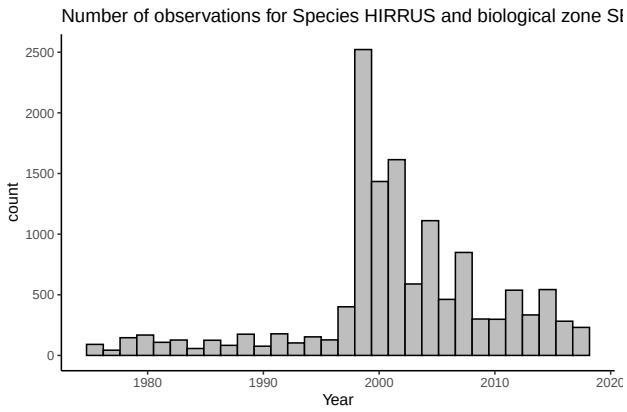

Figure 2: Number of ringing records per year for the species HIRRUS.

   might call into question the use of ordinary least squares (OLS) for fitting the model (1). Propose and

implement an alternative approach, and compare the results with those from OLS. In what scenarios might this kind of issue be a serious concern in practice?

6. In this question we consider an alternative way of measuring shifts in the breeding season than the quantile-based approaches used in the paper. Specifically, for each year i, let p_{ijs} denote the *proportion of ringing events that occur before Day* 170 *of the year* for species j in bioclimatic zone s. Changes in p_{ijs} with year would suggest shifts in the breeding season (for species j in zone s).

   Devise analyses to estimate how p_{ijs} changes with year i, and whether there are differences in this year effect among species. Further, assess whether year effects differ among species that have different numbers of broods or migration types.[1] Clearly explain the models you use, the results of the analyses, and your conclusions.[2]

[1] As in the paper, you may combine the "R" and "S" levels of migration into a single level.

[2] You do not need to run a full Bayesian analysis for this as done in the paper; in particular, while the paper uses phylogenetic information (recording how the species are related to each other), you are not expected to incorporate any such information.