Department of Statistics
STATS 762: Regression for Data Science
Assignment 2
Semester 1, 2023

Total: 100 marks Due: 23:59, 11 April 2023

Notes:

(i) Write your assignment using R Markdown. Knit your report to either a PDF or HTML document. Submit both a (1) .Rmd file, and (2) either a .html or .pdf file on Canvas, one after the other.

(ii) Create a section for each question, and a subsection for each sub-question. Include all relevant code and output in the final document.

(iii) Please keep your code tidy and your plots neat and professional. For example, it's very useful for the reader if you use informative, readable axis labels rather than allowing the default behaviour of printing the R object name.

(iv) These assignments do require you to write R code, but we appreciate this course is not specifically about programming. If you are struggling with the programming aspects of this assignment, please ask for help. If you can describe specifically what you want your code to do, then we can point you in the right direction.

Introduction

The Cape peninsula moss frog *Arthroleptella lightfooti* is endemic to a mountain range surrounding the city of Cape Town, South Africa. They are listed as a near-threatened species by the International Union for the Conservation of Nature, largely because they are severely range restricted: the entire population lives within a single national park, and the mossy seepages they inhabit are vulnerable to invasive vegetation. We do not know if their numbers are in decline.

Monitoring population size of these frogs is a real challenge: individuals are incredibly difficult to see or capture because they are tiny (about the size of your fingernail; females reach a length of 22 mm, while males are shorter) and they live in dense vegetation known as fynbos (the literal translation from Afrikaans is "fine bush") that can reach 2 m in height. However, males produce loud and distinctive advertisement calls, which can be easily heard. Consequently, the quickest and cheapest way to monitor the species is using passive acoustic surveys: researchers place an array of microphones in *A. lightfooti* habitat and record the chirping of male frogs. Each individual frog chirps like clockwork at the same rate, about once every three seconds.

We now have a statistical challenge: how do we convert recordings of frogs chirping into estimates of frog population density? The prevailing solution (at least for this species) is a statistical model known as spatial capture-recapture (SCR). If you are interested in reading a little more about the statistical details of SCR models for frog populations, then take a look at **?** and **?**. If you are interested in reading about an application of these models to *A. lightfooti* data, then take a look at **?**. Reading these papers is certainly not necessary to successfully complete this assignment.

A fundamental component of SCR models is the detection function, which relates the distance between a microphone and a frog to the probability that the microphone detects a call produced by the frog. In other words, it roughly describes how far a microphone can 'hear' (e.g., see Figure 1 in **?**). We can use SCR models to estimate the detection function, along with population density.

? developed an SCR model that is appropriate when individuals remain stationary while calling, and illustrated its use with an application to *A. lightfooti* acoustic survey data. In this assignment, your task is to use generalised linear models (GLMs) to estimate detection functions from a different data set collected at the same location.

You might find it useful to know that frogs all produce calls at about the same volume. We wouldn't expect some frogs to be more detectable than others based on the characteristics of its call. However, some microphones might be better at detecting calls than others: one placed inside a bush might not be as effective as on placed out in the open.

The data

Seven microphones were deployed in a mossy seepage and recorded the chirping of *A. lightfooti* males. The researchers applied a sound-processing algorithm to the recordings from each microphone to determine which sounds were frog chirps rather than background noise or sounds from other animals. They carried out call identification (i.e., determining when different microphones had detected the same call), individual identification (i.e., determining when different detected calls were produced by the same individual), and localised individual frogs (i.e., determined where each detected frog was calling from) using sophisticated procedures that I won't describe here.

Although the researchers left the microphone array to record for a full hour, here we only include data from a single minute. In total, 26 individual frogs were detected.

The data set `frog-detections.csv` is available on Canvas, and contains seven rows for every detected animal, one for each microphone, so it has a total of $26 \times 7 = 182$ rows. The following variables are included for each row, with all coordinates measured in metres:

`animal.id`: An identification number for the individual. If two rows have the same value for `animal.id`, then they include data related to calls produced by the same individual.

`mic.id`: An identification number for the microphone. If two rows have the same value for `mic.id`, then they include data related to the same microphone.

`mic.x`: An x-coordinate for the location of the microphone related to this row.

`mic.y`: A y-coordinate for the location of the microphone related to this row. If two rows have the same `mic.id`, then they will also have the same `mic.x` and `mic.y`, because microphones do not move.

`animal.x`: An x-coordinate for the location of the individual related to this row.

`animal.y`: A y-coordinate for the location of the individual related to this row. If two rows have the same `animal.id`, then they will also have the same `animal.x` and `animal.y`, because frogs do not move during such a short survey.

`detected`: The number of calls produced by the animal that were detected by the microphone.

`n.calls`: The total number of calls produced by the animal during the survey.

Here are rows 1–7 of the data set, relating to the first detected animal, which was produced by the 11th frog, located at the coordinates $(-6.7, 4.0)$:

Figure 1 *An* A. lightfooti *male sitting on a microphone.*

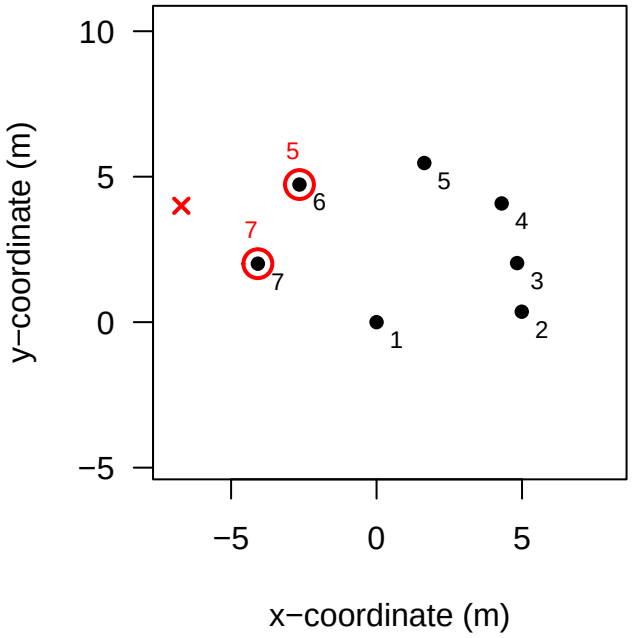

Figure 2 *The layout of the microphones, and a summary of the data related to the first detected frog. The solid black dots are the microphone locations, and appear alongside their respective microphone identification numbers. The red cross is the location of the frog. Red circles indicate which microphones detected the animal at least once, and the red number indicates how many of its calls were detected.*

```
#   animal.id mic.id mic.x mic.y animal.x animal.y detected n.calls
# 1         1      1  0.00  0.00    -6.71        4        0       8
# 2         1      2  4.99  0.36    -6.71        4        0       8
# 3         1      3  4.83  2.03    -6.71        4        0       8
# 4         1      4  4.30  4.08    -6.71        4        0       8
# 5         1      5  1.64  5.47    -6.71        4        0       8
# 6         1      6 -2.65  4.73    -6.71        4        5       8
# 7         1      7 -4.08  2.01    -6.71        4        7       8
```

This frog produced 8 calls in total, and was only detected by Microphones #6 and #7. Four

of the calls were detected by Microphone #6, and seven of them were detected by Microphone #7. This makes sense when we look at Figure 2, because these are the two closest microphones to the frog.

Question 1

(a) Fit a model that estimates the effect of distance between a frog and a microphone on the probability that a call produced by the frog is detected by the microphone. For now, your model should include distance as the only explanatory variable.

(b) The biologist who collected the data reports that there was a bit of a technical mishap with the recording equipment on the survey, and she doesn't trust all of the data that were collected. Unfortunately, she can't quite remember what went wrong.

In Lecture Set 3, we learned how to identify suspicious observations. By using these techniques on your model from Question 1(a), can you figure out what went wrong on the survey? Which observations are untrustworthy?

Provide code, output, and plots related to your investigation, along with a brief explanation of what you are doing at each step. In 2–3 sentences, communicate what went wrong with the recording equipment. Please keep your answer succinct, and your code and output neat and tidy.

Question 2

Fix or remove the untrustworthy observations from the data set, based on your findings from Question 1.

(a) Find the model you think is best for your fixed-up data set. You can now use other explanatory terms, in addition to distance. Use diagnostic techniques to show that your model is appropriate.

(b) Once you've found a model that you're happy with, briefly present some conclusions. Remember that the main goal here is to estimate the effect of distance on detection probabilities.