Department of Statistics
STATS 762: Statistical Modelling
Assignment 1
Semester 1, 2023

Total: 100 marks, plus 5 bonus marks Due: 23:59, 21 March 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

Body size of animals can provide insights into both individual- and population-level health for a wildlife population. For example, body size can be an indicator of an individual's maturity stage and its reproductive status, while also informing the population's demographics and overall health, which are possibly related to environmental conditions. A population that becomes smaller over time may be suffering from a negative change in their environment.

Traditionally, obtaining accurate measurements of body size is acheived by capturing live individuals; however, this approach can be difficult, dangerous for the researchers, and disruptive to the animals. The population's habitat might be inaccessible, the animals might have sharp teeth and claws, and they may react negatively to live capture. Recently, researchers have been using drones equipped with video cameras to overcome these issues: we can fly drones to places that human's can't access, and we can measure animals without having to physically capture them.

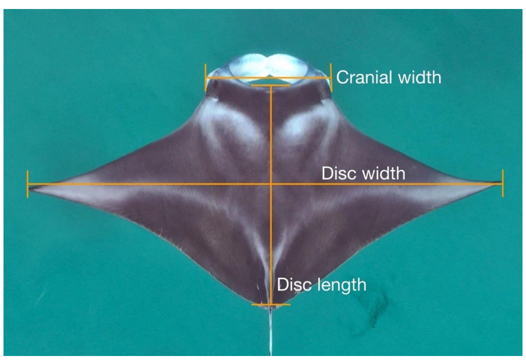

Figure 1 *An image of a manta ray captured by a drone, with the three measurements indicated in yellow lines.*

Measurements of body size from a population of reef manta rays *Mobula alfredi* have recently been collected using drones in Raja Ampat, Indonesia (Setyawan et al., 2022). Traditionally, manta ray body size is measured using the "disc width", or the distance from one wingtip to the other when the individual is perfectly flat (Figure 1). When a manta ray is physically captured, this measurement can simply be taken using a tape measure. However, the disc width is more difficult to measure with a drone, because you need to perfectly capture the precise moment at which the animal's wings are fully extended. Click **here** for a video that includes swimming reef manta rays.

Therefore, in some cases, a direct drone measurement of the disc width may not be available. However, it is much easier to take measurements of a manta ray's cranial width and the disc length from a drone, because these remain stable from one frame to the next in a video recording. In this assignment, we will consider a linear model that can predict an individual's disc width from measurements of its disc length and cranial width, using data collected by drones.

The data

A drone was used to measure the disc length, disc width, and cranial width of 24 individuals. The sex of mature individuals can reliably be determined from images, with the presence of claspers indicating a male, and the presence of either a cloaca or mating scars indicating a female. The data set `manta-measurements.csv` includes a row for each individual manta ray, and the following columns:

- `DL`: disc length, in metres.
- `DW`: disc width, in metres.
- `CW`: cranial width, in metres.
- `sex`: either `M` for male, or `F` for female.

Question 1

Consider the following model:

$$\begin{aligned} Y_i &\sim \text{Normal}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 d_i + \beta_2 c_i + \beta_3 s_i, \end{aligned}$$

where

- Y_i is a random variable denoting the ith individual's disc width,
- d_i is the ith individuals disc length,
- c_i is the ith individual's cranial width, and
- s_i indicates the sex of the ith individual, where $s_i = 0$ if it is female and $s_i = 1$ if it is male.

(a) Why is a linear model more appropriate than other types of GLMs we've met in this course so far?

[5 marks]

(b) Create plots to investigate relationships between the variables in this data set. Briefly comment on the plots (2–3 sentences).

[10 marks]

For the rest of Question 1, you need to calculate various quantities directly in R, **without** using `lm()`, to demonstrate that you understand what is going on 'under the hood' in this function. For example, direct calculation of $\widehat{\boldsymbol{\beta}}$ is carried out for the catheter example on Slide 35 in Lecture Set 1.

Include your code and output in your final document. For each sub-question, your output should just comprise printing whatever you have been asked to calculate. Please separate each calculation into a different code chunk, and include the question label (i.e., (b), (c), and so on) as a section heading ahead of each chunk in your final document.

We recommend that you **do** use `lm()` to check your answers, but you shouldn't include the code and output in your submission.

(c) Create the model's design matrix. Only print out the first six rows.

[5 marks]

(d) Calculate $\widehat{\boldsymbol{\beta}}$, the vector of estimated coefficients.

[5 marks]

(e) Calculate $\widehat{\boldsymbol{\mu}}$, the vector of estimated expected values (or "fitted values"). Only print out the first six values.

[5 marks]

(f) Calculate $\widehat{\sigma}^2$, the estimated error variance.

[10 marks]

(g) Calculate standard errors for the estimated coefficients.

[5 marks]

(h) Calculate t-test statistics and p-values to test individual null hypotheses that the coefficients are equal to zero.

[10 marks]

(i) Calculate confidence intervals for the coefficients.

[5 marks]

A researcher was interested in predicting the disc width of two hypothetical manta rays:

- Morag: a female manta ray with a disc length of 1.3 m and a cranial width of 0.75 m.
- Evelyn: a female manta ray with a disc length of 1.625 m and a cranial width of 0.9375 m.

In reality, we wouldn't be able to measure Evelyn's dimensions to quite this accuracy, but it's important for a later question that Evelyn's measurements are both exactly 25% larger than Morag's.

The researcher didn't manage to measure Morag's or Evelyn's disc widths. Use the model to predict the missing measurements.

(j) Calculate point predictions (equivalent to estimated expectations) for Morag's and Evelyn's disc widths.

[5 marks]

(k) Calculate confidence intervals for Morag's and Evelyn's expected disc widths.

[5 marks]

(l) Calculate prediction intervals for Morag's and Evelyn's observed disc widths.

[5 marks]

(m) **Bonus question!** Write your own R function that accepts the arguments

- `y`, a vector of observed responses, and
- `X`, a design matrix,

and returns output containing estimated coefficients, standard errors, t-test statistics, and p-values, similar to the `summary()` function. Demonstrate that it works by using it to fit the model in this question.

This shouldn't be too difficult if you already know how to write your own R function, and you have successfully completed questions (c)–(h).

[5 bonus marks]

Question 2

For this question, you can use `lm()` or `glm()` if you like, or you can continue with direct calculations. Either way is acceptable.

The researcher is interested in determining whether or not the reef manta ray data are consistent with a hypothesis of isometric growth. If growth is isometric, then the ratios between measurements are constant as the individual increases in size. For example, humans do **not** demonstrate isometric growth from infancy to adulthood, because babies have short legs relative to their height in comparison to adults.

Evelyn is exactly 25% larger than Morag in terms of both disc length and cranial width. If manta ray growth is isometric, then Evelyn's expected disc width will be 25% larger than Morag's. If the ratio is less than 25%, then we have negative allometric growth (manta rays tend to get less wide as they grow, relative to the other dimensions), otherwise we have positive allometric growth (manta rays tend to get wider as they grow, relative to the other dimensions).

(a) Under the model from Question 1, what is the estimated ratio of Evelyn's expected disc width to Morag's expected disc width?

[5 marks]

(b) Compute a confidence interval for this ratio.

[15 marks]

(c) Is it plausible that manta ray growth is isometric? If not, do we have evidence for negative allometric growth, or positive allometric growth?

[5 marks]

In case you're interested...

If you read the paper, you'll notice Setyawan et al. (2022) didn't use linear models to analyse the data. The actual situation is slightly more complicated that what was presented in this assignment:

- The linear model considered here can only predict disc width, and it requires measurements of both disc length and cranial width. However, the authors wanted a single model that could predict any of the dimensions based on any subset of the others (e.g., we might want to predict disc length with a measurement of disc width, without observing the cranial width).
- Measurements from drones have non-negligible measurement error, so if you have multiple images of the same individual then you get different measurements from each one. In this assignment, each observation actually comprises average values

taken across multiple images of the manta ray, which can violate assumptions of a linear model (e.g., averages that were calculated from lots of images will have less variance than averages taken from a smaller number of images).

Setyawan et al. (2022) developed a multivariate hierarchical model that could predict any dimension from drone measurements of any subset of the others, and also accounted for drone measurement error.

References

Setyawan, E., Stevenson, B. C., Izuan, M., Constantine, R., and Erdmann, M. V. (2022). How big is that manta ray? A novel and non-invasive method for measuring reef manta rays using small drones. *Drones*, 6:1–19.