ZZSC5855 Project

General Instructions

This is a small data analysis project. You are provided with a substantive setting, a dataset, and a series of substantive questions to answer using what you learned in Multivariate Analysis. You are to document your reasoning and code you ran in an R Notebook generated from an R Markdown document, and write a short report explaining what you did and your findings using plain language.

For detailed instructions, rubric, and submission information, see https://moodle.telt.unsw.edu.au/mod/turnitintooltwo/view.php?id=5066093 .

Problem

Disclaimer

This is a hypothetical problem. I have endeavoured to make it reasonably realistic, but my own knowledge in this domain is somewhat limited. (Though, I did once spend a summer figuring out how to use a sonar to better count fish in the lakes of Upstate New York.) If you possess actual expertise in this area, I apologise in advance for the likely mangling.

Background

You have been hired by an Australian company that manufactures equipment used by divers harvesting abalone, a kind of marine snail used primarily for its meat but also its eggs. For their next generation product, they want to develop an integrated set of calipers and diving goggles that will allow the diver to quickly measure the dimensions (length, diameter, and height) of an abalone and immediately view some statistically predicted information about it in a heads up display, aiding in the decision whether to harvest it or to leave it alone for the time being.

In particular, to assess the profitability of harvesting a particular abalone, they are interested in predicting the shucked weight of the abalone (the amount of meat for human consumption), its viscera weight (biomass useful for other purposes), and the relationship between the two, as well as the abalone's sex. (This is because female abalone can also be sold for their eggs, which are more valuable.) To assess the sustainability impact of harvesting the abalone, they are also interested in predicting the age and the sex of an abalone. (In particular, harvesting female abalone has a greater sustainability impact, in that it has a stronger effect on the size of the next generation of abalone.)

It is also helpful to predict how profitable a particular abalone is likely to be given the market prices.

Furthermore, the requirement that the microcircuitry built into the diving equipment be rugged, compact, energy-efficient, and inexpensive means that its computing power is limited, and so your client would prefer prediction methods that (once fitted) are computationally cheap for predicting new observations, so “deep learning” techniques are out, but, say, something that uses some (potentially transformed) linear or quadratic prediction or a support vector machine that doesn’t use too many support vectors would be suitable.

Data

Data were collected about a large sample of abalone and can be found in `abalone.csv`.

Researchers collected the following information about each specimen:

`Sex`
: **M**ale, **F**emale, or **I**nfant

`Length`
: (mm) longest shell measurement

`Diameter`
: (mm) perpendicular to length

`Height`
: (mm) with meat in shell

`Whole weight`
: (grams) whole abalone

`Shucked weight`
: (grams) weight of meat

`Viscera weight`
: (grams) gut weight (after bleeding)

`Shell weight`
: (grams) after being dried

`Rings`
: number of rings (can be used to estimate the mollusc's age: adding 1.5 gives the age in years)

Questions

Question 1: Sustainability

Propose, justify, and assess (compare) methods for predicting the sex of the abalone (with “Infant” being considered its own sex for the purposes of this) based on its exterior measurements (length, diameter, and height). Of interest are:

1. predicting the sex of the abalone in general;
2. predicting specifically Infants as opposed to others (to avoid harvesting them);
3. predicting specifically Females as opposed to others (when profitability is prioritised); and
4. predicting specifically Males as opposed to others (when sustainability is prioritised).

Question 2: Profitability

Propose, justify, and assess methods for predicting its shucked and visceral weights (transformed, if necessary) of an abalone based on its exterior measurements (length, diameter, and height).

In particular, because prices fluctuate, the relative profitability of meat and viscera can vary over time. This means that it would be helpful to let the user get a profitability index for an abalone given that day's prices without having to reprogram it from scratch. Develop an algorithm (i.e., a series of steps or a function) that takes as its inputs:

- length, diameter, and height of an abalone;
- v_{shucked} , the dollar value of 1 gram of shucked weight; and
- v_{viscera} , the dollar value of 1 gram of viscera weight;

and produces:

1. an estimate of the value of that abalone: $S = v_{\text{shucked}} \times X_{\text{shucked}} + v_{\text{viscera}} \times X_{\text{viscera}}$, where X_{shucked} is the abalone's shucked weight in grams and X_{viscera} is the abalone's viscera weight in grams;
2. a prediction interval that contains the true value of the abalone some specified (e.g., 90) percent of the time.

Note that due to the computational constraints, this algorithm must rely in precomputed summaries of the data (e.g., means and covariance matrices), and one that requires refitting the prediction model for every new v_{shucked} and v_{viscera} is not a valid solution. On the other hand, since the sample size is quite large, you may assume that any parameters you estimate are not meaningfully different from their true population parameters.

Other Rules and Guidelines

Assumptions and requirements

You are responsible for ensuring that the assumptions and requirements of the techniques you use are met, as well as for cleaning data of outliers (if appropriate) and making appropriate transformations; be sure to document them and provide the appropriate graphical and other diagnostics in the R notebook.

Techniques

Obviously, there are some techniques that you have learned outside of this course, such as logistic regression, that are also suitable for some of these tasks. Feel free to attempt them and mention them in your report if they turn out to be superior, but you must also provide

evidence that you have attempted to apply the appropriate techniques covered in this course and mention them in your report.

Collaboration

This is an individual assessment. The work must be your own, and you may not discuss the assessment or approaches to it with anyone except for the course instructor. You may use "static" external resources, including journal articles and the Wikipedia. You may also use Q&A sites such resources as StackOverflow, but you may not ask questions on such sites. All resources you utilise must be referenced, either in the R notebook or in a separate appendix (which does not count towards any word limit).

Questions and clarifications

If you have any questions, please contact me by e-mail or private forum post. Note that my response is likely to depend on the question:

- For substantive questions about the data and the analysis (e.g., "Am I using this technique correctly?") I will emulate a client with limited statistical knowledge but who is happy to clarify their requirements.
- Logistical, formatting, and other non-substantive questions, I will answer in my capacity as the course coordinator. If I believe that a clarification is valuable to others, I may post it publicly.
- I will, of course, be happy to answer any questions about the course material, as I would during any other week of the course.

Sources

The data set is based (with modifications) on

> Warwick J. Nash, Tracy L. Sellers, Simon R. Talbot, Andrew J. Cawthorn and Wes B. Ford (1994) "The Population Biology of Abalone (*Haliotis* species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of Bass Strait", Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288). Retrieved from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/abalone).

Some of the questions are motivated by this article: https://www.theatlantic.com/health/archive/2010/03/how-to-sex-an-abalone-a-sea-snails-story/37198/ .