Department of Statistics
STATS 330: Statistical Modelling
Assignment 2
Summer School, 2022

Total: 100 marks Due: 14:00, 20 January 2022

Notes:

(i) Write your assignment using R Markdown. Knit your report to either a Word or PDF document.

(ii) Create a section for each 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 at our labs or office hours. If you can describe specifically what you want your code to do, then we can point you in the right direction.

(v) I have done my best to write a clear description of the data and the details of the studies in this assignment, but it is your responsibility to ensure you understand. Asking any questions to clarify your understanding is part of the work required for this assignment.

Introduction

Figure 1 *Rita out for a nice walk.*

Rita is a chicken: one of the finest chickens there is (Figure 1). Rita likes dust baths, adventures with her friends, bok choy, and bullying her flockmate Barbara. Rita dislikes mealworms, rainy days, foot baths, and Barbara. Rita thinks Barbara is arrogant and self-important. Barbara thinks Rita is just jealous of her beautiful double-laced feathering (Figure 2). Rita's owner loves her very much, even though she can be a bit of a bully. He would never admit this to the other chickens, but Rita is one of his favourites.

One day Rita's owner noticed that she wasn't quite herself: she was hunched up on her own in the corner of the field and wasn't participating in all the fun and games with the other chickens. She didn't even take the opportunity to chase Barbara away from a mid-afternoon treat of thinly sliced cabbage. Rita's owner took her to the local chicken vet, who diagnosed her with an internal infection. The vet prescribed a course of antibiotics, and Rita eventually made a full recovery (much to Barbara's disappointment).

When a hen takes a course of antibiotics, her eggs are not suitable for human consumption: there is a seven-day egg withholding period following the final dose. This is a bit of a problem for Rita's owner, because all of the hens lay their eggs in the same set of egg boxes, and it's impossible to know who laid which egg unless he keeps a very close eye on things. Without being able to determine which eggs were laid by Rita, her owner is forced to throw them all away—which is a big waste of all the hard work the chickens put into egg production.

Figure 2 *Barbara showing off her magnificent feathering.*

Luckily, Rita's owner is not only a big fan of chickens, but also enjoys statistics. Over the last few months he has been putting together a data set relating to the eggs laid by his chickens: whenever he can reliably identify the hen responsible for a particular egg (e.g., by virtue of a hen laying an egg before his eyes while he cleans the coop), then he adds data relating to the egg to the data set.

Your task in this assignment is to fit models to the data set to determine which eggs were not laid by Rita, and are therefore safe to eat. You have two data sets at your disposal: one collected prior to Rita's egg withholding period, and another during her egg withholding period for which the identity of the layer is not known.

The data set `eggs.csv` is available on Canvas, and contains a row for each egg measured prior to Rita's withholding period. The following variables are included:

`name`: The name of the hen who laid the egg.

`width`: The width of the egg in cm.

`length`: The length of the egg in cm.

`weight`: The weight of the egg in g.

`colour`: The colour of the egg; either light brown, medium brown, or dark brown.

`calcium.spots`: Whether or not calcium spots are present on the shell; either yes or no. A calcium spot is an small deposit of calcium on the eggshell, usually caused by the egg being inside the chicken for a little longer than usual.

Question 1

First, create a new variable called `rita`, which is equal to `1` if an egg was laid by Rita and is equal to `0` otherwise. (Hint: `y <- ifelse(x == 3, "a", "b")` will create `y` so that `y` is `"a"` when `x` is 3, and `y` is `"b"` otherwise.)

In this question we'll only consider the effect of the variable weight on the probability that an egg was laid by Rita.

(a) Create side-by-side boxplots of the weights of the eggs laid by the different hens. Briefly comment on how the weights of Rita's eggs compare to the weights of the eggs laid by other hens.

[5 marks]

(b) Based on this plot alone and without fitting any models, briefly and informally describe how you expect the weight of a randomly selected egg is related to the probability that the egg was laid by Rita.

[10 marks]

(c) Fit a model that estimates the effect of `weight` on the probability that an egg was laid by Rita by replacing `...` in the code below with something sensible. Consider which distribution is most appropriate for the response variable.

```
glm(rita ~ weight, ...)
```

Print the `summary()` output. Do you think this is an appropriate model? Describe why or why not, using techniques covered in class to justify your answer, including any relevant code and output. Use more than one technique to demonstrate that you have met multiple learning objectives. (Hint: If you don't think the model is particularly good, then some ways to demonstrate that the model is not appropriate involve fitting a slightly improved model that you think *is* appropriate, and then demonstrating the new one is better.)

[15 marks]

Question 2

(a) Find and select a model that you think does the best job at modelling the probability that an egg was laid by Rita. Describe how you found your model in a few simple, easy-to-understand bullet points, explaining each step you took. Neatly present the R code and output you used to select this model. Show your final model's `summary()` output.

There should be a little code for each of your bullet points. If any functions you used resulted in many pages of output, then you do not need to include it all. If you tried a few things that didn't work out, then you can mention what you tried very briefly, but you do not have to include the relevant code here.

Overall, the markers should be able to easily understand how you found your final model by reading your bullet points, and quickly skimming through your code. Adding brief comments to your code may help. If they are wading through pages and pages of material and they're struggling to understand what you've done, they will deduct marks.

There are various ways to complete this question. There is not one single correct answer. It is possible for two students to both get full marks, even if they have very different final models, as long as they have sensibly justified each of their decisions.

[40 marks]

(b) For each egg in the data set, use your model to calculate the fitted probability that the egg was laid by Rita. Calculating fitted values can be achieved using the `predict()` function. See Tutorial 2 for examples. Note that if you leave out the `newdata` argument, then by default the function will calculate fitted values for the observations you fitted the model to, which is what you want here.

Make side-by-side boxplots of these fitted probabilities separated into the different hens. Whose eggs does your model do a good job at disinguishing from Rita's? Whose eggs does your model not distinguish from Rita's so well?

[10 marks]

Question 3

For this question you will be producing fitted values from your model to new data, which can be achieved using the `predict()` function and by providing the new data set as the argument `newdata`. Take a look at Tutorial 2 for an example.

The data set `eggs-new.csv` is available on Canvas, and contains a row for each egg collected during Rita's withholding period. However, this data set does not include the names of the hens who laid the eggs.

Rita's owner is only prepared to discard 16 eggs. Use your model from Question 2 to decide which eggs should be kept, and which should be discarded. Create a vector called `decision` with an element for each egg in `eggs-new.csv`. If you wish to keep the egg, then the element should be `NA`. If you wish to discard the egg, then the element should be a number between `1` and `16`, indicating the priority in which you are discarding the eggs. For example, the egg associated with an element of `1` should be the egg you think is most likely to be Rita's, the egg with an element of `2` is the egg you think is second-most likely to be Rita's, and the egg with an element of `16` is the egg you think is sixteenth-most likely to be Rita's.

Record your 16 selected eggs in a `.csv` file using the following code:

```r
## Loading in the new data set.
new.eggs.df <- read.csv("eggs-new.csv")
## Write code to make your predictions here.
## Your final predictions should be in a vector called "decision".
decision.df <- data.frame(decision, new.eggs.df)
write.csv(decision.df, file = "YOUR_UPI.csv", row.names = FALSE)
```

This code will create a `.csv` file in your working directory. Make sure you replace the file name in the code above with your University of Auckland UPI. Upload this `.csv` file to Canvas when you submit your assignment.

The identities of the chickens who laid these eggs are actually known, but are being kept a secret while students work on the assignment. This question carries a total of 20 marks, 10 of which are awarded for uploading a correctly formatted `.csv` file to Canvas. The remaining 10 marks are awarded proprotional to the number of Rita's eggs that were correctly discarded. If the 16 eggs you chose to discard included all of Rita's eggs, then you will receive full marks.[1]

This question will be automatically marked, so please make sure you have followed the instructions. It's important that

1. you upload your `.csv` file to Canvas in addition to the file with your assignment answers,
2. the filename prefix is your UPI (mine would be called `bste085.csv`),
3. the column containing your decision is called `"decision"` (all lowercase),
4. the entries in this column are either a value between 1 and 16 (for eggs you are discarding) or `NA` (for eggs you are keeping), and
5. you don't discard more than 16 eggs.

The student who successfully discards all of Rita's eggs and has them listed at a higher priority than anyone else wins a prize!

[1] I think almost all sensible models will score at least 18/20 using this scheme. Beyond that, obtaining full marks will involve a small element of luck!