STAD29 / STA 1007 assignment 7

Due Tuesday Mar 17 at 11:59pm on Quercus

You will need to load `tidyverse`, `MASS`, and `car`. If you want to avoid problems, load `MASS` first, or, load `conflicted` as well and deal with any conflicts as they occur (the important one being to prefer `dplyr::select`, the `tidyverse` one, over `MASS::select`).

1. One of the ways to measure the effectiveness of a drug is to measure its concentration in the bloodstream at different times after it is taken. A small study was designed to compare the effectiveness of two different forms of the same drug: a tablet and a capsule. Ten subjects were used and were randomly assigned to the `form` of drug that they would receive (each subject received only one form, either the tablet or the capsule, not both). Each subject was measured at five times after receiving the drug in their assigned form, at 0.5, 1, 2, 3, and 4 hours. At each time, a blood sample was taken, and the concentration of the drug in the subject's bloodstream was measured. The data are in `http://ritsokiguess.site/STAD29/bloodstream.csv`.

   (a) (2 marks) Read in and display (some of) the data. Is the data frame in long or wide format? Explain briefly. (If you prefer, talk about whether the data frame is "tidy" or "untidy".)

   (b) (4 marks) Run a repeated-measures ANOVA to see whether concentration (as measured in the columns `t0.5` through `t4`) depends on `form`, time or the combination of both. Remember the steps: create a response variable, run `lm`, create the within-subjects structure, run `Manova` from `car`.

   (c) (2 marks) What does your MANOVA tell you about the data? Explain briefly, in the context of the data.

   (d) (4 marks) Make a spaghetti plot. That is, plot blood concentration of the drug over time, with the points for each subject being joined by lines, and the lines coloured by the form of the drug that subject received. Do you need wide or long format for your plot?

   (e) (2 marks) What does your spaghetti plot tell you about the reason for the significance or non-significance of the interaction term? Explain briefly.

2. Work through Chapter 24 of PASIAS.

3. On a previous assignment, we learned about researchers who are comparing different ways to give technical information about diet. 33 subjects were randomly assigned to one of three groups: technical dietary information from a website; same information from a nurse practitioner; same information from a video. Each subject then made three ratings: difficulty, usefulness, and importance of the information in the presentation.

   The data are in `http://ritsokiguess.site/STAD29/dietary.csv`.

   (a) (1 mark) Again read in and display (some of) the data.

   (b) (2 marks) Previously we ran a MANOVA on these data and found a significant result. In this assignment, we aim to find out what the significant result means. To begin, run a suitable discriminant analysis, saving the result. Display the saved result.

   (c) (2 marks) Would you prefer to look at one, two or more linear discriminants? Explain briefly.

   (d) (3 marks) For each of your proposed linear discriminants, say whether each of the original variables have a positive, negative or zero effect on it, and what kind of values of those variables would make that discriminant score large and positive.

(e) (3 marks) Obtain predicted group memberships and posterior probabilities. Display at least some of them, side by side with the values they are predictions for. Save the results.

(f) (3 marks) Make a suitable plot of the discriminant scores against group for your chosen number of discriminants.

(g) (2 marks) Comment briefly on how the discriminant scores distinguish the groups, if you think they do.

(h) (2 marks) What do your conclusions above tell you about how the groups are distinguished by their values on the *original* measured variables, if at all? Explain briefly.

(i) (2 marks) Make a cross-tabulation of the people actually in each group with the groups they were predicted to be in. To do this, use `table` or `count` as you prefer.

(j) (2 marks) How many people were misclassified: that is, how many people had a different predicted group from their actual group?

(k) (3 marks) Find a person that was wrongly classified, and display that person's true `group`, predicted `class` and the posterior probabilities for all three groups. It doesn't matter which person you choose. Comment briefly on how close that person was to being predicted correctly.

Notes

[1] This is because using `cbind` to glue a data frame `dietary` to a `matrix p` gets you an old-fashioned `data.frame`. In an R Notebook you'll get the first ten rows as usual, but for me I get the whole thing. If it were a `tibble` I'd get the same as you.