University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 (K. Butler), Final Exam
April 24, 2023

Aids allowed (on paper, no computers):

- My lecture overheads (slides)
- Any notes that you have taken in this course
- Your marked assignments
- My assignment solutions
- Non-programmable, non-communicating calculator

This exam has 10 numbered pages of questions plus this cover page.
In addition, you have an additional booklet of Figures to refer to during the exam.
The maximum marks available for each part of each question are shown next to the question part.
If you need more space, use the last page of the exam. Anything written on the back of the page will not be graded.
You may assume throughout this exam that the code shown in Figure 1 of the booklet of Figures has already been run.

1. Two different treatment drugs are available for the treatment of cocaine addiction. These are called Desipramine and Lithium. In a study, 72 subjects (cocaine addicts seeking treatment) were randomly assigned to one of these treatments or to a placebo. The subjects were followed for a certain period of time, and a record was kept of whether they "relapsed", which means that they went back to using cocaine during that time period. This was recorded as `yes` or `no`. Some of the dataframe is shown in Figure 2. The aim of the study was to see whether there was a difference between the treatments, and if so, which one of the treatments made relapses less likely.

   (a) [3] What feature of the dataset makes logistic regression, as run with `glm`, a reasonable technique to use in studying these data?

   (b) [2] In Figure 3, a logistic regression model is fitted. How do you know that this model is predicting the probability that a subject *does* relapse, rather than the probability that the subject *does not* relapse? Explain briefly.

   (c) [2] In Figure 3, is there evidence of any difference among the treatments on the probability of relapse? In your answer, explain which part of the Figure you are basing your answer upon.

   (d) [3] According to the evidence in Figure 3, which of the three treatments is the most effective? Explain briefly.

2. Samples of worsted yarn (wool) were stretched until they broke, and the time until they broke was recorded. Specifically, there were two factors that might affect how long it took for a sample of worsted to break: the length of the piece of worsted (250, 300, or 350 mm), and the "amplitude of the loading cycle". The worsted is stretched for a certain period of time before being released and then stretched again. Each sample was repeatedly stretched for one of three periods of time (8, 9, and 10 minutes), and the response variable `cycles` is the number of stretch-release cycles that the sample of worsted underwent before it broke. A higher value of `cycles` thus indicates that the yarn is stronger.

   The dataframe is shown in Figure 4. The explanatory variables mentioned above are in columns named `len` and `amp` (respectively); the response variable is called `cycles`. There is an additional column called `load` in the Figure that we ignore in this question.

   This was a designed experiment, with the values of `len` and `amp` chosen by the experimenter. We therefore want to treat `len` and `amp` as categorical (even though their values are actually numbers); we don't want to assume a linear relationship with their numerical values.

   (a) [2] A grouped boxplot is shown in Figure 5. From this plot, why would you expect to see a significant interaction between length and amplitude?

   (b) [2] An analysis of variance is shown in Figure 6. What do you conclude from this Figure?

   (c) [4] Figure 7 shows two parts of a simple-effects analysis. What specifically do you conclude from this Figure? (There are two parts in the Figure, labelled SE1 and SE2. You need a conclusion from each.)

   (d) [2] Why was it not surprising, given what we have seen so far, that your two conclusions in the previous part were different from each other? Explain briefly.

(e) [2] Why does the boxplot in Figure 5 cause you to doubt the analysis we have just done, and why in particular does it suggest that a transformation of `cycles` like square root or log might be useful?

3. Salmon are born in fresh water (rivers) before travelling to salt water (the sea) before returning to fresh water to have their young. Some of the 100 salmon in our dataset were born in Alaskan rivers, and some in Canadian rivers; some were male, and some were female. Do these salmon grow differently? One way to measure this is to look at the diameter of growth rings for the first year of life in the river, and also for the first year of life in the sea. Some of the data are shown in Figure 8. The columns are respectively the `Gender` of each salmon (1 is female, 2 is male), the growth ring diameter in the river (`Freshwater`), the growth ring diameter in the sea (`Marine`), where the salmon was born (`Origin`), and the combination of origin and gender, for use later.

   The growth ring diameters are measured in hundredths of an inch.

   (a) [2] Why would MANOVA be suitable for analyzing these data?

   (b) [2] What code would create a suitable response variable for a MANOVA?

   (c) [4] A MANOVA analysis is shown in Figure 9. There are three parts to the analysis. Explain briefly why I did the analysis the way I did, and say what your final conclusion from the Figure is in the context of the data. (Note that `response` in the Figure is the response variable you created earlier.)

(d) [2] Why is a discriminant analysis worth considering at this point?

(e) [2] A discriminant analysis is shown in Figure 10. Why are there *two* linear discriminants, and why is only one of them apparently important in distinguishing the origin-gender combinations?

(f) [2] What values of the original variables would make LD1 *large*? (You may assume that neither coefficient is close to zero.)

(g) [4] Some further discriminant analysis is shown in Figure 11, and a plot of the scores of each salmon on the first two linear discriminants is shown in Figure 12. From the plot, would you say that either the origins or the genders of the salmon are distinguishable? Explain briefly.

Note: there are four colours on the plot: black, orange, blue, and green. If you cannot tell all four colours apart, ask an invigilator to help you identify some points. You might like to think first about which points would be helpful to identify. I tried to find a colour-blind-friendly collection of colours (see Note above the Figure).

(h) [2] Figure 13 shows a table of actual and predicted genders and origins for the salmon. How does this table support your answer to the previous part? Explain briefly.

(i) [3] Taking the MANOVA and the discriminant analysis together, what do you conclude overall about how `Freshwater` and `Marine` differ according to the origin and gender of the salmon? (You may wish to consider what you would write in the Conclusion of a report about these data.)

4. Thirty-four individuals took part in a study of self-esteem. Each subject was randomly assigned to one of three treatments: diet, diet plus exercise, and a control group with no special instructions for diet or exercise. Each subject had their self-esteem assessed 1, 2, and 3 months after the start of their treatment. The data are shown in Figure 14. The columns `se1` through `se3` are the measurements of self-esteem at the three times.

(a) [2] One of the study authors tells you that this is a two-way ANOVA with two factors treatment and time. Why do you think this is *not* an appropriate way to analyze the data?

Note: I am not asking about the arrangement of the data; your answer should explain why the data, *even if suitably rearranged*, should not be analyzed as described by the study author.

(b) [2] An interaction plot is drawn, using the code shown in Figures 15 and 16. Why is the code in Figure 15 necessary?

(c) [2] Does the plot in Figure 16 suggest that there will be an interaction between treatment and time? Explain briefly why or why not.

(d) [4] A suitable analysis is shown in Figure 17. Using the output shown, what do you conclude from this Figure? Be specific about where you are drawing your conclusions from. (There are usually four parts to this output, with the Multivariate Analysis shown first. Draw an appropriate conclusion based on the results shown here, which are displayed in the same order as on the complete output.)

(e) [2] How might you gain some insight about whether the diet plus exercise group consistently ends up with higher self-esteem than the diet-only group?

5. In 1969, cars were much less reliable than they are now. A car magazine kept track of the service records of 33 car models that were popular in 1969, and rated each model on the 13 items shown in Figure 18. The data are shown in Figure 19. A `+` in the dataframe means that this model of car needed repair for the item shown more often than average, and a `-` means that repair was needed less often than average. Our aim is to cluster the car models into groups of similar ones.

   Some of the car names end in a number (which is often the number of cylinders in the engine), and some of them end in Full, which is short for "Full Size". These last were big cars that consumed a lot of gas (which in those days was not considered to matter very much).

   (Question continues on next page.)

(a) [2] A dissimilarity measure that is often used for data like this is the "simple matching coefficient". This is defined as the fraction of items on which two car models disagree, out of the 13 items that they might disagree on. The simple matching coefficient for the AMC Ambassador 8 and the Buick Special 6 is 0.385, to three decimals. Verify that this value is correct, showing your process.

(b) [1] I computed the simple matching coefficient for each pair of car models (not shown), and ran a hierarchical cluster analysis using Ward's method, shown in Figure 20. What does the code `rect.hclust` shown in the Figure do?

(c) [2] Why do you think I decided to use two clusters? Explain briefly.

(d) [3] I counted up, for each car model, the number of the 13 items that are above average. This calculation is shown in Figure 21, based on the original dataframe in Figure 19, with the column `n` in the dataframe `problems` indicating the number of items for each model that were `+` rather than `-`.

Some additional computation and a plot are shown in Figures 22 and 23. I have shown you the dataframe `clusters`. Describe in words what the three lines of code in Figure 23 are doing, in such a way that somebody who knows about cluster analysis and some basic statistics but who does *not* know R can understand your description. The line beginning `##` above the plot is output, not code.

You need to be familiar with all the code in Figures 22 and 23; you are *not allowed* to ask an invigilator about it.

(e) [2] In a few words, explain what you conclude from the graph at the bottom of Figure 23, in the context of the data.

6. The city of Boston, Massachusetts, had 506 distinct census tracts in 1970. For each one, 14 variables were measured, as shown in Figure 24. Some of the dataframe is shown in Figure 25. We will be doing a principal components analysis.

   (a) [3] Figure 26 shows a scree plot. What do you conclude from this plot? Explain briefly.

   (b) [2] Figure 27 shows the standard deviations explained by each principal component. How does this Figure support your preferred number of principal components that you obtained from the scree plot?

   (c) [3] Figure 28 shows the loadings of all the principal components on all of the original variables. Which *three* variables are the most important in component 1? Make sure to translate the variable names into something your reader can understand.

   (d) [2] In Figure 28, which *two* variables are the most important in component 2?

(e) [4] Figure 29 is a plot of the first two component scores. Figure 30 is a five-number summary (plus the mean) of all the variables in the original dataframe. Fig 31 shows the data values for observations 163 and 164. Given this information and what you have found out so far, does it make sense that these observations appear on Figure 29 where they do? Explain briefly.

(f) [3] A biplot is shown in Figure 32. What kind of loadings should the variable `chas` (which indicates whether the census tract borders onto the Charles River) have on component 1 and component 2? Verify that it does have loadings like that.

Use the rest of this page if you need more space. Be sure to label any answers here with the question and part they belong to.