University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 / STA 1007 (K. Butler), Midterm Exam
March 11, 2022

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 8 numbered pages of questions. Check to see that you have all the pages.
In addition, you should have an additional booklet of output to refer to during the exam. Contact an invigilator if you do not have this.
Answer each question in the space provided (under the question).
The maximum marks available for each part of each question are shown next to the question part.
You may assume throughout this exam that the code shown in Figure 1 of the booklet of code and output has already been run.
The University of Toronto's Code of Behaviour on Academic Matters applies to all University of Toronto Scarborough students. The Code prohibits all forms of academic dishonesty including, but not limited to, cheating, plagiarism, and the use of unauthorized aids. Students violating the Code may be subject to penalties up to and including suspension or expulsion from the University.

Question 1 (10 marks)

Leukemia is, according to the Mayo Clinic, "cancer of the body's blood-forming tissues, including the bone marrow and the lymphatic system". Like any cancer, a sign of a successful treatment is "remission", meaning that the symptoms of leukemia have reduced. Part of our data set is shown in Figure 2. For each patient, the variables recorded are:

- `remiss`: whether or not the patient shows remission (1 = yes, 0 = no)
- `cell`: cellularity of the marrow clot section
- `smear`: smear differential percentage of blasts
- `infil`: percentage of absolute marrow leukemia cell infiltrate
- `li`: percentage labeling index of the bone marrow leukemia cells
- `blast`: absolute number of blasts in the peripheral blood
- `temp`: highest temperature prior to start of treatment

I don't know what any of these mean, apart from the information given here. We want to see whether any of the other variables have an effect on remission.

(a) (2 marks) Explain briefly why logistic regression would be a suitable method to use to analyze these data.

(b) (2 marks) Two logistic regression models are shown in Figures 3 and 4. What precisely are these models predicting? Explain briefly.

(c) (3 marks) Why was it necessary to do the test in Figure 5? What do you conclude from this test?

(d) (3 marks) Some predictions are shown in Figure 6. How are these predictions consistent with the output shown in the appropriate one of Figure 3 and Figure 4? Explain briefly.

Question 2 (12 marks)

A random sample of adult residents of Alachua County, Florida, had their mental health assessed by a professional. Three variables were recorded for each person:

- `impairment`: the professional's overall assessment of mental health for each person, on a scale from Well (good), Mild, Moderate, Impaired (bad).
- `ses`: socio-economic status: high or low
- `life_events`: a scale reflecting the number and severity of important life events such as birth of child, new job, divorce, or death in family that occurred to the subject within the past three years. (A higher number means that the person has experienced more of these events, or the events they have experienced have been more severe.)

The entire data set is shown in Figure 7. (The data values are separated by single spaces.)

(a) (3 marks) The data was read into a dataframe `mh`. Some possible models were fitted in Figure 8. Explain briefly why the model labelled `mh.2` is more appropriate than *each* of the models labelled `mh.1` and `mh.3`.

(b) (2 marks) In Figure 8, why did it make sense to use `fct_inorder` in defining the model `mh.2`? Explain briefly.

(c) (2 marks) Some more output is shown in Figure 9. What do you learn from this output? Explain briefly, in the context of the data.

(d) (2 marks) Another model was fitted, as shown in Figure 10. Why was it necessary to run `drop1` again, even though the remaining explanatory variable was significant in Figure 9?

(e) (3 marks) Some predictions are shown in Figure 11. Would you say that a person with a higher score on the life events scale is likely to have better or worse mental health overall than someone with a lower score? Explain briefly. Based on what you know about mental health, and what you have learned about the life events scale used in this data set, do you find this surprising? Explain (very) briefly.

Question 3 (11 marks)

49 patients took part in a trial of a new treatment, called linoleic acid, for a particular form of colorectal cancer. Think of these patients as a random sample of all patients with this particular cancer. 25 of these patients were randomized to the new treatment, and the 24 received a control treatment (the current best treatment for this form of cancer). Some of the data are shown in Figure 12. For each patient, the experimenters recorded the treatment received, whether or not the patient died, and the length of time that the patient was observed, in months.

(a) (3 marks) Figure 13 shows some code to create a new column in the dataframe `cancer` as was read in from the spreadsheet. If you were to look at the first four values of the new column `y`, what would you see? Explain briefly why you would see that.

(b) (2 marks) A Cox proportional hazards model is fitted, as shown in Figure 14. In this output, why does `treatment` display as it does? Explain briefly. You do not need to discuss any numeric values here (that comes later).

(c) (2 marks) Estimated survival curves are shown in Figure 22 (at the end of the booklet of Figures), along with the calculations that led up to them. Interpret the plot. In particular, which treatment appears to be more successful? How do you know? Explain briefly.

(d) (2 marks) In Figure 14, which number supports the conclusion that you drew from the previous part? Explain briefly.

(e) (2 marks) Based on what you see here, do you think that this conclusion would generalize to *all* patients with this type of colorectal cancer? Explain briefly.

Question 4 (17 marks)

Three different treatments, labelled A, B, and C, are being investigated to see whether they have any effect on the growth of plants. The experimenters choose to assess plant growth in three different ways: the height, width, and weight of the plant. Fifteen plants were grown, five for each treatment. The data, in dataframe `plants`, are shown in Figure 15.

(a) (2 marks) What feature of this data set would make multivariate analysis of variance (MANOVA) an appropriate method to use? Explain briefly.

(b) (2 marks) In Figure 16, the MANOVA analysis is shown. It uses a variable `y` that I had to define. How did I define it (in code or in words)?

(c) (2 marks) What do you conclude from the MANOVA output in Figure 16?

(d) (2 marks) What was the purpose of running the discriminant analysis in Figure 17? Explain briefly, in the context of the data.

(e) (2 marks) Why is it that there are two linear discriminants, and why is it that I only need to consider the first one? Explain briefly.

(f) (2 marks) Which of the response variables contribute the most to distinguishing the treatments? Explain briefly.

(g) (3 marks) A plant has small weight, small height and average width. Using the graph in Figure 18, which treatment do you think it received? Describe your thought process clearly enough so that your reader is convinced by your logic.

(h) (2 marks) How does Figure 18 confirm what you said earlier about the relative importance of LD1 and LD2 for these data? Explain briefly.

Question 5 (10 marks)

Investigators at the University of North Carolina Dental School were interested in the growth of children's skulls. They measured 27 children, 11 female and 16 male (as the children identified themselves). Each child was measured at ages 8, 10, 12, and 14 years. The quantity measured was the distance (in millimetres) between the centre of the pituitary to the pterygo-maxillary fissure. This distance usually increases with age, but because both of the two points can move, the distance occasionally decreases with age. The quantity measured is known as "the distance" for the rest of the question; you do not need to know any more about what it is.

Some of the dataset is shown in Figure 19. There are six columns: the number code of each child, the gender of the child (labelled `sex`), and the distance as measured at each age, labelled `d` followed by the age (as two digits).

(a) (2 marks) Figure 20 shows the mean distance for each gender and age. In the code for the graph, why was the `pivot_longer` necessary before drawing the graph? Explain briefly.

(b) (2 marks) What about this dataset makes a repeated measures ANOVA a suitable method of analysis? Explain briefly.

(c) (2 marks) The analysis is shown in Figure 21. What do you conclude about the interaction between gender and time? Is this consistent with the graph in Figure 20? Explain briefly.

(d) (2 marks) From Figure 21, what do you conclude about the effect of time (age)? Is this consistent with the graph in Figure 20? Explain briefly.

(e) (2 marks) From Figure 21, what do you conclude about the effect of gender? Is this consistent with the graph in Figure 20? Explain briefly.

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