University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 / STA 1007 (K. Butler), Midterm Exam
March 2, 2019

Aids allowed:

- 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 9 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.

In my Crowdmark grading, I use the shaded (blue) rectangle to denote something wrong (in your answer) or something missing (in the question).

Question 1 (7 marks)

An office equipment repair company services photocopiers in large businesses. Data are collected from a number of service calls, as shown in Figure 2. The column `machines` shows the number of photocopiers needing service on a service call, and the column `minutes` shows the total number of minutes the technician spends on site to service all the photocopiers that need service on that visit.

(a) (2 marks) A plot of the dataset is shown in Figure 3. Describe briefly what you conclude from this plot.

(b) (2 marks) A regression model is fitted, predicting the number of minutes spent on site by the technician from the number of machines needing service. The output is shown in Figure 4. Quote *one* number from this output that is consistent with what you saw in the plot, and explain briefly how it is consistent.

(c) (3 marks) The repair company receives a call from a business with 6 photocopiers needing service. Use the appropriate part of Figure 5 to obtain an interval estimate of how long the technician will be on site. Justify your choice briefly.

Question 2 (16 marks)

In some places, you pay a deposit when you buy a soft-drink bottle, say 10 cents, and then you can return the empty bottle to the store where you bought it and get the deposit back. This is environmentally friendly, because the soft-drink manufacturer can get the bottles back from the store, wash them and re-use them. (This is usually done with glass bottles rather than plastic ones.)

A convenience store chain ran an experiment to see whether soft drink bottles are more likely to be returned if the deposit is higher. 3,000 one-litre bottles of a certain soft drink were individually bar-coded and randomly assigned to one of six different deposits ranging from 2 cents to 30 cents, labelled so that a customer could see what the deposit was after they bought the bottle. Records were kept of whether or not each bottle was returned to the store where it was bought, and what the deposit was on that bottle. The dataset is shown in Figure 6, with each row containing a summary of how many bottles were sold at each deposit level, and how many were returned.

(a) (2 marks) Why is logistic regression a sensible idea here?

(b) (3 marks) A logistic regression was run in Figure 7, to predict the probability of a bottle being returned, as it depends on the deposit cost. The output is shown. A variable `y` was used. Give the R code that was used to create `y`.

(c) (2 marks) Does the size of deposit have a significant effect on whether or not a bottle is returned?

(d) (2 marks) Why does it make sense that the Estimate for `deposit` in Figure 7 is positive? Explain briefly.

(e) (1 mark) In Figure 8, I obtain some predictions. What does the `type="response"` do? Explain (very) briefly.

(f) (4 marks) Give R code to make a graph showing the predicted probabilities of a bottle being returned, joined by lines, along with the proportions in the data of bottles that were returned at each deposit level, shown as points. Your graph should look something like Figure 9. Hint: you have a calculation to do before you draw the graph. You may use any of the variables that I obtained in Figures 6 through 8 to make your graph.

(g) (2 marks) The plot that you gave code for in the previous part is shown in Figure 9. What does this plot tell you about how well the logistic regression fits? Explain briefly.

Question 3 (15 marks)

Arthritis refers to pain or stiffness in the joints of the human body. A study investigates a new treatment for arthritis. In the study, patients are randomly assigned to the new treatment (labelled `Treated`) or to a placebo (`Placebo`), and the age and sex of each patient is recorded. Some of the data is shown in Figure 10. The column `impr` is a numerical code: 0 is no improvement, 1 is moderate improvement, and 2 is complete recovery from the arthritis. Figure 11 shows the fitting of a model, and some output from that model.

(a) (2 marks) Why did I need to use `factor(impr)` rather than just `impr` in the first line of Figure 11? Explain briefly.

(b) (2 marks) Why did I use `polr` rather than `multinom`? Explain briefly.

(c) (2 marks) Should I remove any explanatory variables from the model? If so, which ones? Explain briefly.

(d) (3 marks) In Figure 12 I make some predictions for the two sexes, the two treatments and two representative ages. I first created a data frame containing all the combinations of these, which I called `new`. What code would get predicted probabilities of each improvement category for each of these combinations, using the model that I fitted in Figure 11?

(e) (3 marks) From Figure 12, is being older rather than younger associated with a better or worse recovery overall? Explain briefly.

(f) (3 marks) Based on the output in Figures 11 and 12, would you say that the treatment is *helpful*? Explain briefly.

Question 4 (16 marks)

A number of patients suffering from some disease are given one of two treatments, labelled A and B. Each of these patients is observed for a time; either the patient dies, or something else happens to them (they are definitely known to have survived, or their status is unknown, but they were never observed to die). Each patient's age is also recorded. The data are shown in Figure 13, with survival times in months. The event of interest is "death".

(a) (3 marks) Give R code to create a suitable response variable `y` for a Cox proportional hazards model.

(b) (2 marks) My response variable is as shown:

```
##  [1]  1   1   4   5   6+  8   9+  9  12  15+ 22  25+ 37  55  72+
```

Why is the fifth value (and some of the others) marked with a `+`? Explain briefly. (The `[1]` is not one of the values.)

(c) (2 marks) A Cox proportional-hazards model is fitted, with the results shown in Figure 14. Is there a significant effect of age? Using this output, describe the effect of age on survival, justifying your conclusion briefly.

(d) (3 marks) The quartiles of age are 67 and 75. Give code to create a data frame called `new` that contains all combinations of these ages and the two treatments, that is to be used in a moment for obtaining predicted survival probabilities.

(e) (2 marks) Give R code to obtain predicted survival probabilities, in such a way that we will be able to plot the survival curves using `ggsurvplot`.

(f) (2 marks) Predicted survival curves are shown in Figure 23. Which of the two treatments is more effective at prolonging life? Explain briefly.

(g) (2 marks) How does something in Figure 14 support your conclusion about treatments in the previous part? Explain briefly.

Question 5 (16 marks)

Three treatments for headache, labelled A, B, and C, are being tested. Thirty people, fifteen men and fifteen women, are each randomly assigned to one of the treatments. Each person is instructed to take their assigned treatment when they next get a headache, and to note how many minutes it takes "until the pain subsides".

(a) (2 marks) The data as I originally received it is shown in Figure 15. The blank treatments are repeats of the one above, thus for example five men and five women were assigned to treatment A. Describe briefly how the data are not suitable for analysis in their current form, even after I have fixed up the blank treatments.

(b) (3 marks) I put the data into a suitable format, and calculated mean pain relief times for each `Treatment-Gender` combination. I saved the group means in a data frame `d`, with the column of means called `mean_time`. In Figure 24 I use `d` to create an interaction plot. Give the R code that was used to create this plot, assuming that you already have the data frame `d`.

(c) (2 marks) What do you conclude from Figure 24? Explain briefly.

(d) (3 marks) An analysis of variance is shown in Figure 16. What do you conclude from it, and what does that mean in the context of the data?

(e) (2 marks) Some further analysis is shown in Figures 17, 18, and 19. Which of this analysis is appropriate here and which is not? Explain briefly.

(f) (4 marks) What do you conclude from the appropriate analyses out of Figures 17, 18, and 19, in the context of the data? Explain briefly.

Question 6 (11 marks)

How moral do people think it is to eat meat? Ninety university professors were asked to rate their answers to this question on a nine-point scale, with 1 being "very morally bad', 5 being neutral and 9 being "very morally good". The professors studied ethics, philosophy but not ethics, and other disciplines. Some of the data are shown in Figure 20. The professors' areas of study are in `discipline`, denoted `Eth` (ethics), `Phil` (philosophy but not ethics), `Other`. The column `discipline` is an R `factor`.

The researchers who collected the data had two research questions in mind:

- Are the Ethics professors and the Philosophy professors different in terms of their attitudes towards eating meat?
- How does the average of the Ethics professors and the Philosophy professors compare with the Other professors?

(a) (2 marks) Explain briefly how contrasts would provide an efficient way of answering the researchers' questions.

(b) (2 marks) Write two contrasts that will represent the researchers' questions of interest, in the same order as above. Call your contrasts `c1` and `c2`. That is, give R code that will create two vectors, one for each contrast, that can be used in an `lm` later to make tests. Note that the `disciplines` are to be taken in alphabetical order.

(c) (2 marks) Verify that your two contrasts are orthogonal.

(d) (2 marks) Give R code to set up your two contrasts as contrasts for the model we are about to fit using `lm`. (Two steps. The data frame is called `meat`.)

(e) (3 marks) What do you conclude from Figure 22, in the context of the data? You may wish to use Figure 21 to give a complete conclusion.