University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 / STA 1007 (K. Butler), Midterm Exam
February 29, 2020

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 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 (13 marks)

According to Wikipedia, Intensive Care Units in hospitals "...cater to patients with severe or life-threatening illnesses and injuries, which require constant care, close supervision from life support equipment and medication in order to ensure normal bodily functions". A number of patients do not live long enough to be discharged from the Intensive Care Unit (ICU).

A study was carried out to investigate factors associated with a patient living long enough to be discharged from the ICU (or of dying before then). The study collected a large amount of information; we will look only at a few explanatory variables. Our data set is a sample of 200 patients from a much larger number in the original study.

- `id`: identification code for patient
- `sta`: patient's vital status: Lived (until being discharged from the ICU), Died.
- `age`: in years
- `can`: Cancer is part of the patient's present problem: No, Yes
- `cpr`: patient had CPR (cardio-pulmonary resuscitation) before admission to the ICU: No, Yes
- `inf`: probable infection at admission to ICU: No, Yes
- `race`: patient's race: White, Black, Other

The variables above that are actually categorical have been turned into R `factor`s so that they will properly be treated as categorical. The categories are ordered as shown above. Some of the data set is shown in Figure 2.

(a) (2 marks) The response variable is `sta`. How do we know that the logistic regression models that we fit will predict the probability of dying rather than the probability of living? Explain briefly.

(b) (2 marks) A logistic regression is shown in Figure 3. Based on the output shown there, which explanatory variable would you consider removing first, if any? Explain briefly.

(c) (3 marks) I used `step` to remove unimportant explanatory variables from my model, obtaining the output shown in Figure 4. I decided to keep `inf` even though its P-value is slightly greater than 0.05. The remaining categorical variables all have two levels. Explain briefly, based on what you know or can guess about patients entering the ICU, why it makes sense that each of the last three Estimates in the Figure (that is, except for the Intercept) is positive.

Question 1 continues... This page: 7 marks.

(d) (4 marks) We want to obtain predicted probabilities of dying for all combinations of representative values of the explanatory variables in the model shown in Figure 4. Using Figure 5 to help you, give code that will do this. You have to decide what "representative values" means for you. In your code, end by displaying the predictions with the values they are predictions for.

(e) (2 marks) The predictions I made are shown in Figure 6. Describe the effect of an increase in age on probability of dying, all else equal. Explain briefly.

Question 2 (16 marks)

Two surveys were taken, one in 1960 and one in 1970, of different randomly selected groups of 1000 people. One of the questions in the survey asked respondents to say in which of a number of categories their annual income fell. The incomes are in thousands of dollars; in 1960, \$15,000 was a good income! Our question of interest is to say what happened to income over that time period. The data are shown in Figure 7. Note the types of the variables. For example, `ord` means "ordered factor".

(a) (2 marks) Describe briefly an appropriate graph for these data. (I want a description, not code.)

(b) (3 marks) Give code for fitting an appropriate model, named `income.1`, for predicting income category from year. (You may assume any necessary packages are loaded with `library()`.)

(c) (2 marks) In your code of the previous part, did you have a `weight`? Explain briefly why you need it (if you had it) or why you don't need it (if you didn't have it).

(d) (2 marks) Some output from your fitted model is shown in Figure 8. The `tidy` output at the top is basically the same as the `summary` output for the model. Is the distribution of incomes the same or different for the two years? Explain briefly.

(e) (3 marks) Give code to obtain predicted probabilities of a person falling in each income category, for each year in the data set, and to display your predictions next to the values they are predictions for.

(f) (2 marks) The predictions from your code are shown in Figure 9. Describe briefly how the probabilities are changing between 1960 and 1970.

(g) (2 marks) Would you say, according to the fitted model, that incomes between 1960 and 1970 are typically going down, staying the same or going up? Explain briefly.

Question 3 (14 marks)

100 patients with a certain disease are randomly allocated to one of four treatments, labelled A, B, C, and D. Each patient is monitored periodically until the disease comes back (labelled "recurrence" in the data set), or until the study ended and no recurrence had been seen until that point. Some of the data are shown in Figure 10. Our intention is to do a survival analysis with these data, to see which treatment is most effective at delaying recurrence, but with the data as they were given to us, we have some work to do first. For each patient, we have: the treatment they were on, whether or not recurrence occurred (the disease came back), the month, day and year that they were enrolled into the study, and the month, day and year of the last followup (the last time they were seen by a doctor). The data frame as read in from the file is called `disease`.

(a) (1 mark) When did the study end? Explain (very) briefly.

(b) (3 marks) Give code to create the date of enrollment as an R date.

(c) (3 marks) After running your code of the previous part, repeating the process on the followup dates, and removing columns no longer needed, the data is as shown in Figure 11. A survival analysis needs "survival times": that is, the number of days between enrolment into the study and last follow up. Give code to obtain these from the data shown in Figure 11.

(d) (2 marks) The data after we have finished tidying it is shown in Figure 12. What code will create a suitable response variable `y` for a survival analysis of these data?

(e) (2 marks) In Figure 13, a Cox model is fitted and some output shown. Is there evidence of any difference among the treatments? Explain briefly.

(f) (3 marks) Figure 22 shows estimated survival curves. Which treatment is best at delaying recurrence? Cite *two* pieces of evidence, using this Figure and Figure 13, to support your assertion.

Question 4 (16 marks)

Arthritis sufferers often feel pain in their joints (knees, elbows, wrists) which interferes with their daily lives. Three competing treatments for joint pain are compared in terms of their mean time to pain relief in patients with osteoarthritis. Because the investigators hypothesize that there may be a difference in time to pain relief in men versus women, they randomly assign each of 15 participating men to one of the three competing treatments and randomly assign each of 15 participating women to one of the three competing treatments. Thus there are 5 men and 5 women assigned to each treatment. Participating men and women do not know to which treatment they are assigned. They are instructed to take the assigned medication when they experience joint pain and to record the time, in minutes, until the pain subsides. A shorter time to pain relief is better. The data are shown in Figure 14 as data frame `joint`.

(a) (3 marks) There are two categorical explanatory variables. One way to understand the nature of their effect on time to pain relief is to draw an interaction plot. Give code to do this. Hint: your code could do a calculation first.

(b) (2 marks) My interaction plot is shown in Figure 23. What do you conclude from it? Explain briefly.

(c) (3 marks) What other plot could you draw with this data set? Describe and justify briefly one advantage *and* one disadvantage of your proposed plot, compared with the interaction plot.

(d) (3 marks) Two analyses of variance are shown in Figure 15. What do you conclude from `joint.1`? Is it a good idea to do the analysis `joint.2`? Explain briefly.

Question 4 continues... This page: 11 marks.

(e) (3 marks) Figures 16 and 17 show two possible further analyses. Which one of these is more appropriate, and what do you conclude? Explain briefly. If you think neither of these analyses is appropriate, explain briefly why.

(f) (2 marks) Which treatment or treatments would you recommend? Is that different for males and females? Explain briefly.

Question 5 (14 marks)

To assess the relative merits of three methods of instruction (labelled `a, b, c`) for elementary computer programming, a curriculum researcher randomly selected 12 fifth graders from each of three elementary schools in a certain school district. Each group, within the setting of its home school, then received a six-week course of instruction. At the end of the six weeks, each student did a test to see how well they had learned the prescribed elements of the subject matter. However, two of the schools (the ones at which methods `a` and `b` were taught) were in more affluent neighbourhoods, and so, before the course began, each student in each of the schools was given a test of basic computer familiarity as well.

The data are shown in Figure 18.

(a) (3 marks) A graph is shown in Figure 24. Give the code that was used to draw the graph.

(b) (2 marks) An analysis of covariance is shown in Figure 19. What is the one principal thing that this analysis tells you **about the data**?

(c) (2 marks) A second analysis of covariance is shown in Figure 20. Is this an appropriate analysis to conduct? Explain briefly. If it is appropriate, what do you conclude from it?

(d) (2 marks) What does Figure 24 suggest about the likely reason for the significance or non-significance of `method` that you observed earlier? Explain briefly.

(e) (3 marks) Another test is shown in Figure 21. What does this one say? Why do you think this conclusion is different from what you have seen elsewhere in this question? What made it come out different in the way it did? Explain briefly.

(f) (2 marks) What do you think would be a statistically better way to design this study? Why do you think it was done the way it was? 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.