University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 (K. Butler), Midterm Exam
March 4, 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 7 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. A study was done to investigated the effectiveness of a new teaching method, called PSI, in economics. This was a comparative study with a control group. Four variables were measured:

   - `grade_improved`: whether the student's exam grades were improved during the study period (yes or no), as compared to before the study period, response
   - `psi` whether the student had been exposed to PSI during the study period (yes or no). The `yes` group is the treatment group, and the `no` group is the control group.
   - `tuce` measure of student's academic ability at the start of the study period
   - `gpa` student's grade point average at the start of the study period.

   Some of the data are shown in Figure 2. One of the authors on this study was named Spector, hence my use of the name `spector` for the dataframe and models.

   (a) [2] Why would logistic regression be suitable here? Explain briefly.

   (b) [2] Some analysis is shown in Figures 3 and 4. Why do you think I used `update` to fit the model `spector.2`? Explain briefly.

   (c) [2] The variable `tuce` that was removed from the logistic regression was a measure of the student's academic ability. You might expect such an explanatory variable to have an impact on whether or not the student's exam grades improved during the study period. Why do you think it would have been removed from this logistic regression?

   (d) [3] From the model `spector.2` in Figure 4, interpret the numbers in the Estimate column (not including the intercept). Hint: unless you are very careful, you will have to make sure not to over-interpret these numbers.

(e) [3] Some predictions are shown in Figure 5. Explain briefly how these predictions are consistent with your interpretations of each of the two Estimates in the model `spector.2`.

(f) [3] Comment briefly on the width of the confidence intervals in Figure 5. Given what you know about the data, does your comment make sense? Explain briefly.

2. Boxes each containing approximately 100 trout eggs were buried at five different locations in a river, and retrieved (removed from the river) at four different times. (There were thus $5 \times 4 = 20$ boxes of eggs altogether, and each box was retrieved only once.) After each box was retrieved, the number of surviving eggs was recorded. The data are shown in Figure 6. The columns are, respectively, the number of eggs in the box that survived, the total number of eggs in the box to begin with, the location in the river (coded A through E), and the number of days the box was left in the river.

   (a) [3] Some code is shown in Figure 7. Why is it necessary to run this code? Explain briefly.

   (b) [2] A logistic regression is shown in Figure 8. Does this logistic regression predict the probability that an egg survives, or the probability that the egg does not survive? How do you know? Explain briefly.

(c) [2] From Figure 8, at which location is the probability of an egg surviving predicted to be highest, all else equal? (That is, you may assume that the comparison of locations is done for the same value of `period`.) Explain briefly.

3. In 1993, a survey was carried out in western Germany about attitudes towards science. We focus on one of the survey questions, here labelled `D`. The responses to question `D` are labelled 1, "strongly disagree", through 5, "strongly agree". A person who answered 5 to this question has a strong (positive) belief in the value of science. For each survey respondent, the `sex` they identified as, and their level `edu` of education, were also recorded. The column `edu` was actually recorded in categories, but we will treat this as quantitative, with a higher value of `edu` corresponding with more education. Some of the data is shown in Figure 9.

(a) [3] A model was fitted, as shown in Figure 10. Even though the responses to item D are given as numbers, I used `polr` instead of an ordinary linear regression. Why was that? Explain briefly.

(b) [2] Some predictions are shown in Figure 11. Why was the `pivot_wider` a good idea? Explain briefly.

(c) [3] Is it true based on Figure 11 that someone with more education has a stronger belief in the value of science overall, all else equal? Explain briefly.

(d) [3] In Figure 11, would you describe the effect of `sex` as large or small? Explain briefly. From the output in Figure 10, why would you have expected to see the size of effect you did? Explain briefly again.

4. An important part of any business is making sure that the business gets paid for the work it does. Most businesses issue an invoice stating how much money the customer owes, and giving the customer a certain amount of time to pay the invoice. When the customer pays an invoice, the business marks the invoice as "settled", and no further action is needed. (If the customer does not pay an invoice by the due date, further action will need to be taken to make sure the invoice gets paid.)

   A certain company kept track of over two thousand invoices it issued between January 2012 and November 2014. The information we will use is:

   - `invoice_date`: when the invoice was issued
   - `due_date`: when payment is due
   - `invoice_amount`: how much money ($) the invoice is for
   - `settled_date`: when the invoice was settled (paid)

   The dataset also contains information about the customer who received each invoice, which we will not use in this question. Some of the (relevant) data, in dataframe `receivables`, is shown in Figure 12. You will be writing some code in this question.

   (a) [3] All the dates in the data file were recorded as text. What code would re-define the `invoice_date` column to be an R date? (For the rest of the question, assume that the other two dates have also been re-defined as R dates.)

   (b) [2] Someone tells you that this business always gives all its customers the same amount of time to pay their invoices. What code would tell you how many days that was?

(c) [2] We are asked to find out how many of the invoices were paid on or before the due date, and how many were paid after the due date. What code would find this out?

(d) [3] The business management wants to know whether invoice amounts are changing over time. Bear in mind that there are over two thousand invoices of varying sizes, so the management may need some help in seeing any trend. What code will draw a suitable graph to help the management find out what they want to know? Use whichever dates you feel are appropriate.

(e) [4] The management also wants to know what the mean invoice amount is for each month over the time that the data were collected (that is, treating February of 2012 separately from February of 2013). What code would calculate these means?

5. Twenty-six psychiatric inpatients, admitted to the University of Iowa hospitals during the years 1935-1948, were observed over a period of several years.

   Data for each patient consists of:

   - `age` at first admission to the hospital
   - `sex` (1 = male, 2 = female)
   - `time`: number of years of follow-up (years from admission to death or censoring)
   - `death`: patient status at the follow-up time (1 = dead, 0 = alive when last observed).

   Our aim is to understand how age and sex influence survival time among psychiatric patients. The dataset is shown in Figure 13.

   (a) [3] Some code is shown in Figure 14. What is `Surv` doing, what are its inputs, and why are some of the values in the last column displayed with a plus sign? Explain briefly.

   (b) [2] A Cox proportional-hazards model and its output is shown in Figure 15. Describe the effect of `age` as far you can deduce it from this Figure.

   (c) [3] A dataframe of values to predict for is set up in Figure 16, and Figure 17 shows estimated survival curves for these values. Describe how this plot supports your conclusions about the effects of `age` in the previous part (or does not support them, if that's the case).

   (d) [2] Figure 15 shows that there is no significant effect of `sex`. How does this show up on the graph in Figure 17? Explain briefly.

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.