00D7×

University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 / STA 1007 (K. Butler), Midterm Exam
February 18, 2017

Aids allowed:

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

Before you begin, complete the signature sheet, but sign it only when the invigilator collects it. The signature sheet shows that you were present at the exam.

This exam has 11 numbered pages of questions. Please 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). If you need more space, use the backs of the pages, but be sure to draw the marker's attention to where the rest of the answer may be found.

The maximum marks available for each part of each question are shown next to the question part. In addition, the total marks available for each page are shown at the bottom of the page, and in the table on the next page.

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.

Last name: ______________________________

First name: ______________________________

Student number: ______________________________

For marker's use only:

Page	Points	Score
1	8	
2	11	
3	6	
4	8	
5	9	
6	6	
7	6	
8	7	
9	7	
10	8	
11	8	
Total:	84	

You may assume throughout this exam that the code shown in Figure 1 of the booklet of code and output has already been run.

1. Spatio-temporal reasoning is the ability to understand a pattern in time and space, and to understand how objects fit into it. For example, suppose you are meeting a friend at a coffee shop across town later today. You use spatio-temporal reasoning to figure out how to get there, and when you will have to leave to get there in time. This kind of reasoning is important for children to develop.

   How can we help children to develop spatio-temporal reasoning? We might give them lessons of various different kinds. In one study, 40 children were randomly allocated to groups who received piano, singing, computer or no lessons. Each child only did one kind of lesson; the "no lessons" was a control group. Each child's ability at spatio-temporal reasoning was measured twice, once before they started the lessons and again after they finished. The score that was recorded for each child was the difference between the measurement after and the measurement before. That is, a child with a positive score had better spatio-temporal reasoning after their lessons than before, and a child with a negative score had worse spatio-temporal reasoning afterwards. (Note that the apparent matched-pairs nature of the data has been removed by looking at the differences; we now have independent measurements.)

   The data are shown in Figure 2 of the booklet of code and output.

   (a) (3 marks) Some code is given in Figure 3. Describe what this code does (without using the word "gather"). Your answer should contain a description of what the output from `gather` looks like, and should explain what role `reasoning2` has in the process.

   (b) (3 marks) What output will the code in Figure 4 produce? Describe it precisely.

   (c) (2 marks) What output will the code in Figure 5 produce? Describe it precisely.

2. Fish can be healthy to eat, but fish that is contaminated with mercury is not so healthy. What influences the amount of mercury contamination? Water samples were taken from 38 lakes and analyzed. From those same 38 lakes, samples of fish were taken, and the mercury contamination in their muscle tissue was recorded. Fish absorb mercury over time (older fish have higher concentrations on average, with a relationship that is known), so standardized mercury levels for 3-year-old fish were calculated for each lake (based on the fish that were caught there).

   The data are summarized in Figure 6. The variables are:

   - `mercury`: standardized mercury level of fish in parts per million (response)
   - `alkalinity` of water in mg/l
   - `calcium` concentration of water, in mg/l
   - `pH` of water (7 is neutral, lower values are acid, and higher values are alkaline).

   (a) (3 marks) Scatter plots of `mercury` against each of the explanatory variables are shown in Figure 7. A regression model was fit in Figure 8. Why specifically do you think that fitting this regression model was a bad idea?

   (b) (2 marks) A second regression is fit in Figure 9, with the output shown. A third regression is fit in Figure 10, but the output is not shown. Which explanatory variables does the third regression contain? You may assume that this third regression model is satisfactory for the rest of the question.

   (c) (4 marks) What precisely is being predicted in Figure 11? In particular, what do the last two numbers in the second row of the output of the `cbind` tell you?

   (d) (2 marks) In what way are Figures 9 and 11 telling the same story? (You may ignore the fact that models `mercury.2` and `mercury.3` are not the same.)

3. 33 leukemia patients were studied. For each patient, their white blood cell count was recorded (`wbc` in the data file), and the presence or absence of a certain morphological characteristic in the white blood cells was also recorded (`ag` in the data file, denoted `+` for present and `-` for absent). It was noted whether each patient lived for at least a year (`live=1`) or not (`live=0`). The actual lifetime of each patient was not recorded, only whether it was at least a year or not. The researchers were interested in whether either of the two explanatory variables `wbc` or `ag` helped to predict survival.

   (a) (2 marks) Look at the plot in Figure 13. Why do you think the researchers decided to use the log of white blood cell count in their analysis?

   (b) (2 marks) In the code for Figure 13, why did I have to say `x=factor(live)` instead of `x=live`?

   (c) (2 marks) In the logistic regression of Figure 14, what probability is being predicted? Explain briefly.

(d) (2 marks) Which, if any, of the explanatory variables would you consider removing from the regression? Explain briefly.

(e) (2 marks) Looking at Figure 14, what would be the effect of a higher white blood cell count?

(f) (4 marks) I did some predictions based on the model in Figure 14. These are shown in Figure 15. Based on the latter Figure, what can you say about the effect of `ag`? Is that consistent with Figure 14? Explain briefly.

4. A clinical trial was designed to compare the effectiveness of three pain-relief drugs to be taken after an operation. The drugs were called `c15`, `c60` and `z100`. Each patient in the trial was given one randomly-chosen drug after their operation. Each patient was then asked to rate the pain relief offered by their drug on a scale "poor, fair, good, very good". After the trial was complete, the number of patients giving each drug each rating was tabulated, as shown in Figure 16.

   (a) (3 marks) What is the response variable here, and for what *two* reasons would you choose `polr` from package `MASS` to do your analysis of these data?

   (b) (2 marks) What is the *purpose* of the `mutate` in the creation of the new data frame `painrelief` in Figure 16? Explain briefly. (That is, *what* the code is doing is part of the answer, but I am mainly interested in *why* I have to do it.)

   (c) (2 marks) A plot is shown in Figure 33 (at the end of the booklet of code and output). Based on this plot, would you expect to see a significant difference in pain relief among the drugs? Explain briefly.

   (d) (2 marks) Why did I need `weight=` in my modelling statement in Figure 17?

(e) (2 marks) In constructing my data frame `new` in Figure 17, why did I *not* need to use `expand.grid`? Explain briefly.

(f) (2 marks) Should I remove `drug` from the model? Why or why not?

(g) (2 marks) Are the predictions in Figure 17 consistent with the bar charts in Figure 33? Discuss briefly.

5. *Clematis ligusticifolia* or western white clematis is a climbing vine with showy flowers. It is also known as “old man’s beard”, since that is what its flowers look like. An ecologist is studying this vine, in particular whether its male or its female flowers are visited more frequently by insects. The ecologist observed waiting times during the blooming period. She watched a particular flower and waited until an insect landed on it, recording (i) how many minutes this was, (ii) whether it was a male or female flower. In some cases, she watched a flower for a long time and never saw an insect land on it. She therefore also recorded whether an insect was observed at the recorded time (`observed=yes`) or whether the time was when she stopped watching (`observed=no`). The structure of the data is shown in Figure 18, along with some randomly-chosen rows of the dataset.

   (a) (2 marks) Figure 19 shows side-by-side boxplots of waiting times for insects to arrive at the two genders of flowers. Why could these boxplots give a misleading comparison for these data? Explain briefly.

   (b) (3 marks) Figure 20 shows the construction of a variable `y`. What is the purpose of the `observed=="yes"` in the `Surv` line, and what do the `+` signs mean next to some of the values in the display of `y`?

   (c) (1 mark) In Figures 21 and 22, two Cox proportional-hazards models are fitted. Which explanatory variables are included in the model `clematis.0` of Figure 22?

(d) (2 marks) Using Figures 21 and 22, is there a significant effect of gender (of the flower)? How can you tell? What does that mean in the context of these data? Explain briefly.

(e) (3 marks) The ecologist's research hypothesis was that male flowers would be more attractive to insects than female flowers. What would that imply for waiting times for insects on male and female flowers? Does the graph in Figure 34 (at the end of the booklet of code and output) support that hypothesis? Explain briefly.

(f) (2 marks) One of the numbers on Figure 21 supports your conclusion of the previous part. Which one, and why?

This page: ______________ of possible 7 points.

6. An evil Statistics lecturer has his students write a final exam in essay form. A randomly chosen half of the students write the final essay exam using a regular exam book (`bluebook`), and the other half use a laptop computer (`computer`). In addition, he assesses how much typing experience each student has, classified as "none", "some" or "lots". The handwritten exams are transcribed, and all exams are printed out onto the same paper, so that the lecturer grades each exam without knowing whether it was originally handwritten or typed. The response variable is the `score`. The data are shown in Figure 23.

   (a) (2 marks) In Figure 24, how do the factors `ability` and `ability.ord` differ? Explain briefly.

   (b) (2 marks) In Figure 24, I calculate the mean `score` for each combination of `ability` level and `method`, and in Figure 35 (and the end of the booklet of code and output) I draw a graph with the values I calculated. By looking at Figure 35, what is likely to happen in the analysis, and why? Explain briefly.

   (c) (3 marks) In Figures 25 and 26, two analyses are shown. Which one is the more reasonable to base our conclusions on? What, therefore, do you conclude? Explain briefly.

(d) (2 marks) Look at Figure 27. Which students are included in this analysis, and what is being tested here?

(e) (3 marks) What do you conclude from Figure 27? Give a complete answer, using the results from the whole of the Figure, as appropriate. **Use α of 0.10**.

(f) (3 marks) What do you conclude from Figure 28? Again, give a complete answer, using the results from the whole of the Figure, as appropriate. Again, **use α of 0.10**.

7. 26 samples of Romano-British pottery were found at four different sites. The samples were analyzed by atomic absorption photometry to measure the percentages of oxides of various different metals contained in the samples. The aim of the study was to see whether pottery collected from different locations had a different profile of metal oxides. The data are shown in Figure 29. In this question, we focus on iron oxide, labelled `Fe` in the data. The four sites are Llanederyn and Caldicot, both in the Gwent area of south Wales, and Island Thorns and Ashley Rails, both in the New Forest area of England. Boxplots of the iron oxide values for the four sites are shown in Figure 30.

   The researchers were most interested in whether (i) the two sites in south Wales were different from each other, (ii) the two sites in England were different from each other, and (iii) the two sites in England were on average different from the two sites in Wales.

   (a) (2 marks) Why is it better to address the researcher's interests using contrasts rather than a regular ANOVA followed by Tukey? Explain briefly.

   (b) (3 marks) Some computations are shown in Figure 31. Explain briefly but specifically how the computations are related to the researchers' questions.

   (c) (3 marks) What do you conclude from Figure 32? There are three things to say.