University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 / STA 1007 (K. Butler), Midterm Exam
February 24, 2018

Aids allowed:

- My lecture overheads (slides)
- Any notes that you have taken in this course
- Your marked assignments
- My assignment solutions
- Notes from STAC32
- 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 8 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	9	
3	13	
4	5	
5	11	
6	6	
7	7	
8	11	
Total:	70	

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. A common problem for hospitals is that their patients can get infections while in hospital (which means that the hospital has to treat the infection *and* whatever it is that the patient is in hospital for). There is a standard procedure for assessing the infection risk to a patient, resulting in a numeric (quantitative) score. 113 patients at different hospitals in different regions of the US were assessed; we will assess how the infection risk (`InfctRsk`) depends on the number of days the patient stayed in the hospital (`Stay`), how often they were given X-rays (`Xray`) and the `Region` of the US where the hospital is located (a numerical code: 1 is north-east, 2 is north-central, 3 is south and 4 is west).

   Some of the data set is shown in Figure 2.

   I don't know the units of `InfctRsk` (other than that a higher score is a greater infection risk) or of `Xray`.

   (a) (2 marks) There is an extra line of code at the top of Figure 3 (beyond what would normally be seen in fitting a multiple regression). What does it do, and why does it need to be there? Explain briefly.

   (b) (3 marks) A regression and its `summary` output is shown in Figure 3. Interpret the slope for `Stay`, whose value is 0.349.

   (c) (3 marks) In Figure 3, interpret the slope for `Region4`, whose value is 0.833.

(d) (2 marks) For assessing the (statistical) significance of `Region`, explain briefly why it is better to look at Figure 4 than Figure 3.

(e) (2 marks) What do you conclude about regions from Figure 4? Explain briefly.

(f) (2 marks) Look at the first line of Figure 5. In `inf.2`, what is being predicted from what? Explain briefly.

(g) (3 marks) In Figure 5, what does the prediction tell you? Explain briefly.

2. A local health clinic sent flyers to all of its clients to encourage everyone to get a flu shot (vaccination) before winter. Later, the clinic randomly sampled 50 of its clients and asked each one whether they got a flu shot or not. The clinic also collected data on each client's age and health awareness (via a survey). The results of the health awareness survey were summarized into a health awareness score (called `awareness` in the data set), where a higher value means greater health awareness. A client who received the flu shot is denoted 1 in the `shot` column, and one who did not is denoted 0. Some of the data set is shown in Figure 6.

   (a) (2 marks) Why is it that I do *not* need a two-column response variable to fit a logistic regression to these data?

   (b) (2 marks) A logistic regression is shown in Figure 7. Would you consider removing either of the explanatory variables from the model? Explain briefly.

   (c) (2 marks) What probability is being modelled in Figure 7? Explain briefly how you know.

   (d) (3 marks) Look at the slope coefficients in Figure 7. What do they tell you about what makes a person more or less likely to get a flu shot? Explain briefly.

   (e) (4 marks) Using the information in Figure 8, give R code to obtain predicted probabilities of getting a flu shot for all combinations of first and third quartiles of `age` and of `awareness`, and to display the predictions side by side with the values they are predictions for.

3. The LSYPE is a "longitudinal study of young people in England". (Longitudinal means that individuals are followed over a period of time.) Over 15,000 people, born in 1989 and 1990, were followed starting in 2004. Each person was interviewed 8 times between 2004 and 2016. There were 57 variables collected for each person. We will focus on just a few, related to educational achievement.

   A few rows of the data are shown in Figure 9 and the values are summarized in Figure 10. The variables are described below.

   English young people take important national exams at age 14, called "Key Stage 3". We will focus on the Key Stage 3 English exam. Each young person receives a whole-number grade between 3 and 7 (inclusive), where higher is better. The actual exam mark is not known, nor indeed is the precise process by which exam marks are turned into these grades. This is the value in column `k3en` in Figure 9. At age 11, these young people have previously been tested in English and math at what is called "Key Stage 2".

   The other variables are:

   - `gender`: of the young person, 0 is male, 1 is female.
   - `sec`: socio-economic status of the young person's family, on this scale:
     - 0: higher managerial/professional
     - 1: lower managerial/professional
     - 2: Intermediate occupations
     - 3: Small employer/self-employed
     - 4: Lower supervisory/technical
     - 5: Semi-routine (semi-skilled)
     - 6: routine (unskilled)
     - 7: never worked/long term unemployed
   - `ks2stand`: the overall Key Stage 2 score, a decimal number (unlike the Key Stage 3 English grade). Higher is better. The scale is set so that 0 is "average", so that a `ks2stand` value can be negative.

   All of these explanatory variables will be treated as quantitative. (It is questionable whether `sec` should be treated as such, but assume that this is reasonable.)

   Our aim is to predict Key Stage 3 English grades from the other variables.

   (a) (3 marks) Describe briefly (in words) what the code in Figure 11 should be doing, and how you know it has succeeded in doing that. (You are supposed to know what `is.na` and `!` mean in this context.)

   (b) (2 marks) Even though `k3en` grade is apparently quantitative, it would be a mistake to treat it as such. Explain briefly why this is.

(c) (3 marks) Give R code to create an ordered factor called `en3` *within the data frame* ***`kids`***, using the values in `k3en` in that data frame in a sensible order.

(d) (2 marks) An ordered logistic regression is fit in Figure 12. Using the information in this Figure, do you need to keep all the explanatory variables? Explain (very) briefly.

(e) (3 marks) Some predictions are shown in Figure 13. Describe the effect of Key Stage 2 score on predicted Key Stage 3 English grade, explaining how you got your answer. (Note that the code that was used to obtain these predictions is not shown.)

(f) (3 marks) Again using Figure 13, describe the effect of socio-economic status on the predicted Key Stage 3 score, all else equal. Is it what you would expect, given your understanding about socioeconomic status and education? Explain briefly. Hint: remind yourself of the scale on which the socio-economic status is measured for these data.

4. What are the factors that determine how long someone remains unemployed, and what effect do they have? To find out, about 1,900 unemployed people were followed from the time they became unemployed until the time they found a new job (or until the study ended). A large amount of information was collected on these people. We will look at the variables shown in Figure 14, which are:

   - `spell`: the number of months spent unemployed
   - `event`: whether or not the person found a job by the end of the study (1 is yes, 0 is no)
   - `ui`: whether or not the person was claiming unemployment insurance (1 is yes, 0 is no)
   - `logwage`: the logarithm of the person's last salary before unemployment
   - `work_area`: the person's previous area of work, categorized as:
     - `constr`: Construction, for example building of houses
     - `fire`: Emergency services such as fire, police, paramedic
     - `mining`: for example coal mining
     - `pubadmin`: public service or administration, for example working in government
     - `services`: work that does not produce a physical object
     - `trade`: for example electrician, plumber
     - `transp`: Transportation, for example truck driver.

   (a) (3 marks) Figure 15 shows the construction of a variable `y` and the display of its first 20 values. What does the value 1 at the start of the first line of output mean, and why does the next value 3 have a plus sign next to it? If your answer uses the word "censored" at any point, you should explain what that means in the context of these data.

   (b) (3 marks) Figure 16 shows a Cox model for these data, together with the output from `summary` and `drop1` for this model.

   In Figure 16, does `ui` have a significant effect on the time taken to find a new job? Does a person who receives unemployment insurance typically take a longer or a shorter time to find a new job than someone who does not? How can you tell? Explain briefly.

(c) (2 marks) Figure 17 shows some code to obtain predictions of "survival" for the median `logwage`, `ui` of 1 (receiving unemployment insurance), and the various different work areas. These predictions are shown on a plot in Figure 22. How do you know which coloured prediction is which? In particular, what prediction is the purple "survival curve" for? (If you have trouble telling the colours apart, ask an invigilator.)

(d) (3 marks) According to Figure 22, for people with median log-wage and receiving unemployment insurance, people previously in which work area are most likely to find a new job the quickest? Explain briefly.

(e) (2 marks) Go back to Figure 16. For the work area you chose in the previous part, how does it appear as the one in which people are most likely to find a new job the quickest? Explain briefly.

5. In a manufacturing process, a plastic rod is made by melting a plastic and then extruding it through a nozzle. It is better if the rod can be made more quickly, that is, if the extrusion rate is higher. Two factors can be controlled in the extrusion process: temperature (200 or 300 degrees Fahrenheit), and pressure (40 or 60 pounds per square inch). What effect do these factors have on extrusion rate? An experiment was run in which three replicates of each combination of temperature and pressure was used, and the extrusion rate measured, for a total of 12 observations.

   The data set is shown in Figure 18.

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

   (b) (2 marks) Look at Figure 23. What additional information does this plot give that is not shown in Figure 24? Does that strengthen or weaken your conclusion from the previous part? Explain briefly.

   (c) (2 marks) An analysis of variance is shown in Figure 19. What do you conclude from it, in the context of the data?

   (d) (2 marks) Figure 20 shows some more analysis. The company who ran the experiment is looking for the highest extrusion rate. Based on this Figure, what recommendation would you make for the combination or combinations of pressure and temperature to use? Explain briefly. (If you recommend only one combination of pressure and temperature, justify that recommendation.)

   (e) (3 marks) Figure 21 shows a final piece of analysis. Explain briefly what you conclude from this analysis, in the context of the data. (Explaining what the code does will not help you much. I want to know what this Figure tells you *about the data*.)