University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 / STA 1007 (K. Butler), Midterm Exam
February 13, 2016

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 9 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	8	
3	7	
4	6	
5	6	
6	11	
7	12	
8	5	
9	9	
Total:	72	

1. The North American professional basketball league, called the NBA, had 30 teams in the 2008–2009 season. What factors affect the average number of fans that come to each team's games? (This is called the team's "attendance".) Three possible factors are:

   - the recent success of the team, measured as number of wins in the 82-game 2008–2009 season.
   - the size of the city. This was measured as the size of the team's TV market (as a percentage of all households in the US). This variable is called "market share".
   - Whether or not the team has played in the NBA finals in the last five years. (This is another measure of recent success).

   The data are shown in Figure 1 in the booklet of code and output.

   (a) (2 marks) A regression model is shown in Figure 2. Would you say that the factors named above together give a good, moderate or poor prediction of team attendances? Explain briefly.

   (b) (2 marks) Interpret the slope of `wins`: that is, express in words in the context of the data what it is telling you.

   (c) (3 marks) Why is the slope of `finals` shown in the output as `finalsyes`? Interpret this value.

   (d) (1 mark) Figure 3 shows the residual plots from the regression in Figure 2. What was the purpose of the `par(mfrow=c(2,2))` above the plots?

(e) (2 marks) What do you conclude from the top left plot in Figure 3? Does that indicate any problems with the regression? Explain briefly.

(f) (2 marks) What do you conclude from the top right plot in Figure 3? Does that indicate any problems with the regression?

(g) (2 marks) What do you conclude from the bottom left plot in Figure 3?

(h) (2 marks) Make a recommendation for the next thing to do in this analysis, if anything. Support your recommendation briefly.

This page: ______________ of possible 8 points.

2. Some scientists were studying the ozone layer over the Antarctic. These scientists developed a measure of the degree to which oceanic phytoplankton production (labelled `inhibit` in the data) is inhibited by exposure to ultraviolet radiation (`uvb` in the data set). The statistical question is how, if at all, `inhibit` is related to `uvb`. To complicate matters, measurements were taken at two different depths, labelled `surface` and `deep`, and the scientists wanted to see whether depth had any impact as well. The first task the scientists wanted to do is to make a plot of `inhibit` against `uvb`, with the points labelled according to depth.

   Unfortunately, the data come to us in the form shown in Figure 4. There are two columns of `inhibit`, one for depth `deep` and one for depth `surface`. Likewise, there are two columns of `uvb` measurements.

   Underneath the listing of the data, I load `dplyr` in such a way as not to show us those messages it would otherwise show us.

   (a) (3 marks) There are four lines of `dplyr` code in Figure 5. (There are actually only two lines of R code altogether, but each one is split into two to fit on the paper). Describe in words what the first two printed lines, from `ozone %>%` to `-> deep` inclusive, are doing. Do this in such a way that someone who does not understand `dplyr` can see what is being done.

   (b) (2 marks) Figure 6 contains some code and output that uses the data frames `deep` and `surface` created in Figure 5. I do not expect you to have seen `bind_rows` before, but try to explain in words what it is doing. There are two specific things that I would like you to observe.

   (c) (2 marks) Figure 7 shows some code to make a `ggplot` plot, and the output that it produces. (The warnings below the code are about the missing values in the data frame `v`. You do not need to worry about these.) Explain briefly what role the `aes` plays here.

(d) (2 marks) In Figure 7, what role does the `geom_smooth` play here?

(e) (1 mark) Suppose I had wanted to make a scatterplot of all the points, with *one* regression line for all the data. How would I modify the code in Figure 7 to achieve that?

(f) (3 marks) What do you conclude about the relationship, if any, between `uvb` and `inhibit`? Is the nature of the relationship the same or different for the two different depths?

3. Back in the 1800s, people would try to bring their favourite bird species from their homeland to where they were living. They would release the birds, and hope that the birds would become successfully established in the new country. This was done a great deal in New Zealand. In 1996, an exploratory study was done, as an attempt to determine what factors might lead to a bird species becoming successfully established. They looked at 79 bird species that had been introduced to New Zealand prior to 1907, and measured a number of variables related to the species and introduction(s), as follows:

   - `Species`: abbreviated name of species
   - `Status`: whether species still present (1) or absent (0) in New Zealand (response variable)
   - `Length`: female body length (mm)
   - `Mass`: female body mass (g)
   - `Range`: geographic range (% of area of Australia. Why not New Zealand I have no idea.)
   - `Migr`: migratory (3), somewhat migratory (2) or not migratory (1). ("Migratory" means the bird flies to a warmer climate in winter or a cooler one in summer.)
   - `Insect`: number of months of year with insects in diet
   - `Diet`: eats plants (1), everything (2), meat (3)
   - `Clutch`: number of eggs per breeding episode
   - `Broods`: number of breeding episodes per year
   - `Wood`: uses woodland areas frequently (1), infrequently (0)
   - `Upland`: uses upland areas (hills or mountains) frequently (1), infrequently (0)
   - `Water`: uses water frequently (1) or infrequently (0)
   - `Release`: number of times birds of the species released
   - `Indiv`: total number of individual birds released.

   I have treated the variables `Migr` and `Diet` as numeric. Although they are really categorical, we can think of an implied underlying continuous scale in each case. (For example, a bird species might have a greater or lesser tendency to eat meat.)

   The structure of the data is shown in Figure 8. The original data had some missing values. I deleted all the observations with any missing values.

   (a) (2 marks) Why is logistic regression a plausible technique to use here?

   (b) (2 marks) In a logistic regression, what probability will R model? Explain briefly.

   (c) (2 marks) A logistic regression is fit in Figure 9. Why do you think I chose to begin with this model? Explain briefly.

(d) (3 marks) A second analysis is shown in Figure 10. Why do you think I chose to do this analysis next? Bearing in mind the exploratory nature of this problem (I am trying to find some potentially interesting factors in determining whether an introduced bird species becomes successfully established or not), do you think I have now found a good model, or not? Explain briefly.

(e) (2 marks) Look at the test shown in Figure 11. What do you conclude from it, in the context of the data?

(f) (2 marks) In Figure 10, I used `update`. Explain briefly, in words, what that is doing here.

(g) (2 marks) Look at the output in Figure 10. Does this indicate that a species for which more individuals were released is more, less or equally likely to be still present in New Zealand, other things being equal? How can you tell? Explain briefly.

(h) (2 marks) In Figure 10, the slope coefficient for `Upland` is significantly negative. What does that mean, in the context of the data?

This page: ______________ of possible 11 points.

4. On January 28, 1986, the space shuttle Challenger took off and exploded, killing all the astronauts aboard. Might it have been possible to anticipate the problems and stop the launch on that day? Data were collected on all the space shuttle flights, as follows:

   - `flight`: flight number, 1 through 25 (two flights had missing data and were omitted from the data set)
   - `distress` (response): The number of "thermal distress incidents" in which hot gas damaged the joint seals of a flight's booster rockets. Damage to the joint seals helped lead to the Challenger disaster. Categorized as 0, 1–2 or 3+ (3 or more): the higher the distress, the riskier the flight.
   - `temp`: Calculated joint temperature at launch time (degrees F).
   - `date`: the date (number of days since Jan 1, 1960). There may have been changes in the shuttle program or the hardware over time that may have had an impact on `distress`.
   - `z.computed.`: ignore.

   A summary of the data is shown in Figure 12.

   (a) (2 marks) What is it about the data that makes the function `polr` more suitable for carrying out the analysis than `multinom`?

   (b) (2 marks) In Figure 13, some models are fitted. What do you conclude, in the context of the data, from Figure 14?

   (c) (2 marks) What do you conclude, in the context of the data, from Figure 15?

   (d) (4 marks) Figure 16 shows some predictions for representative values of `temp` and `date`. (These values are approximately the first and third quartiles of each variable.) Describe the effects of (i) increased temperature and (ii) increased date on the response variable.

   (e) (2 marks) On the date of the Challenger launch, which was day 9524 on this scale, the temperature was 31 F. A prediction is shown in Figure 17. (Note that the format is different for some reason: the predictions go down the page instead of across.) What would you say to the people in charge of the launch, and why?

This page: ______________ of possible 12 points.

5. Primary biliary cirrhosis is a disease in which the bile ducts in the liver are slowly destroyed. If this happens, harmful substances can build up in the liver and lead to scarring of the liver tissue (which is what "cirrhosis" is). A study was carried out at the Mayo Clinic between 1974 and 1984. This was a placebo-controlled trial of the then-new drug D-penicillamine (with some patients receiving a placebo). In the study, a number of other variables were measured that might have some impact on a subject's survival. (I know nothing about what these other variables are.) The aim of the study was to see whether patients who took the new drug tended to survive longer than patients on the placebo, after adjusting for the effects of the other variables. The structure of the data is shown in Figure 18. The variables are these:

   - `id`: subject ID
   - `days`: survival time (days) after diagnosis
   - `status`: what happened to the patient: alive at last doctor visit (0), removed from study to have liver transplant (1), dead (2).
   - `drug`: D-penacillamine (1), placebo (2).
   - `age`: patient's age at diagnosis (days).
   - `edema`: yes (1), no (0). There is also a value 0.5, "edema resolved by diuretics". We treat this variable as a continuous numeric variable.
   - `bilirubi`: serum bilirubin (mg/dl).
   - `albumin`: serum albumin (mg/dl).
   - `prothom`: prothrombin time (seconds).

   (a) (3 marks) Look at Figure 19. What is happening there? Your answer should address three things: (i) why I am going to need this variable `y`, (ii) the purpose of the `status==2`, (iii) precisely why some of the values of `y` have `+` next to them.

   (b) (2 marks) Are the researchers looking for a coefficient of `drug` that is positive or negative? Explain briefly.

This page: ______________ of possible 5 points.

(c) (2 marks) A Cox proportional hazards model is fitted in Figure 20. What does this output tell you about any effect of `drug`? Explain briefly.

(d) (2 marks) *Not* including `drug`, do any or all of the other variables have a significant effect on survival time? Explain briefly.

(e) (3 marks) For *each* of the significant variables in the model, would a high value or a low value be associated with surviving for a long time?

(f) (2 marks) Figure 21 shows the set-up for some predictions of survival curves, using representative values for the variables, and Figure 22 shows the plot of those predicted survival curves. Explain how the plot is consistent with *two* things that you have previously seen in this question.