University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 / STA 1007 (K. Butler), Final Exam
April 16, 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 15 numbered pages of questions. Please check to see that you have all the pages.

In addition, you should have an additional booklet of code and output to refer to during the exam. Contact an invigilator if you do not have this. References to numbered Figures in this exam refer to Figures in that booklet. The captions with Figure numbers are *underneath* each Figure.

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	4	
2	7	
3	4	
4	7	
5	5	
6	8	
7	8	
8	9	
9	11	
10	7	
11	12	
12	10	
13	6	
14	10	
15	8	
Total:	116	

1. "Inattentional blindness" is a phenomenon where a person who is concentrating on one task fails to see something that is right in front of them. The classic example of this is a video showing students passing a basketball around, and your task is to count how many passes the players in white shirts made. See, for example, `http://www.theinvisiblegorilla.com/videos.html` (check it out after the exam). During the video, a person wearing a gorilla suit walks across the screen, and the question is, "did you notice anything else while you were counting the passes?". If you didn't notice the gorilla, you have inattentional blindness.

   There is a psychological test called the Stroop Colour Word test. This comes in three parts, each to be done as fast as possible:

   - `W`: word alone, or reading a list of colour names.
   - `C`: colour alone, or naming the colours in which a series of X's is printed.
   - `CW`: colour and word, or naming the colour in which a *different* colour name is printed.

   Each individual gets a score on each of the three parts. Is any of this related to whether or not a person saw the gorilla on the video? Some data were collected from 49 subjects, and these are shown in Figure 1 of the booklet of code and output. The variable `seen` is 1 or 0 according to whether the subject saw, or did not see, the gorilla when they watched the video. We aim to predict whether or not a subject saw the gorilla from the test scores.

   (a) (2 marks) Why is logistic regression a suitable procedure for what we are trying to do?

   (b) (2 marks) What probability is the logistic regression in Figure 2 modelling? Explain briefly.

(c) (2 marks) What do you conclude from the `anova` at the bottom of Figure 2? Note that a model with a `1` on the right side of the squiggle has no explanatory variables, just an intercept. Explain briefly, in the context of the data.

(d) (2 marks) How is the output from `summary(gorilla.1)` consistent with what you've just said?

(e) (2 marks) Figure 3 shows two boxplots side by side. What do these boxplots tell you? Is the message from the boxplots similar to what you already found? Explain briefly.

(f) (1 mark) What would you recommend that the researcher do next?

2. I have a collection of variables measured on a collection of 1978-model cars. The variables I have are these:

   - `make`: the name (model) of the car
   - `price`: the manufacturer's recommended sale price for the car
   - `mpg`: gas mileage (in miles per US gallon, so larger is better)
   - `rep78` (response variable): repair record in 1978 (categorized; takes the values 1–5 with a higher value being better). Five car models had a missing value for this variable; these models were excluded from the analysis.
   - `headroom`: distance from a car's roof to the bottom of the seat (measured in feet).
   - `trunk`: total volume in cubic feet of the car's trunk space (or other cargo-carrying space).
   - `weight`: total weight, in pounds.
   - `length`: front to back, in inches.
   - `turn`: the diameter of the smallest U-turn the car can make, in feet (known as the "turning circle radius").
   - `displacement`: total volume of the cylinders in the car's engine, in (I think) cubic centimetres. The larger this is, the more powerful the car.
   - `gear_ratio`: the larger this is, the faster the engine spins (in RPM) for a given size of tires.

   We want to use (some or all of) these variables to predict the repair record. Some of the data set is shown in Figure 4.

   (a) (2 marks) For my modelling, I used `polr` from package `MASS`. Explain briefly why that was a good choice.

   (b) (2 marks) I fitted the two models shown in Figure 5. What do you conclude from the `anova` statement at the bottom of that Figure? Which model should we use to do predictions with? Explain briefly.

(c) (1 mark) In Figure 6, I do some predictions, for representative values of the explanatory variables, based on model `autos.2`, which we assume is an appropriate model to do predictions from (this may or may not agree with your answer to the previous part). What did I achieve by using `expand.grid` in Figure 6?

(d) (2 marks) Describe the effect of car price on the repair record. As price increases, does the repair record improve? Explain briefly.

(e) (2 marks) Describe the effect of turning circle radius on the repair record. Is an increase in the turning circle radius associated with an improvement in the repair record, or not? Explain briefly.

(f) (2 marks) Does the output of Figure 6 say anything about the effect of gas mileage? Explain briefly why or why not.

3. A study of breast cancer was done in Germany. A total of 720 patients with "primary node positive breast cancer" were included in the study. Patients were recruited into the study between July 1984 and December 1989. A number of variables were recorded for each patient:

   - `id`: patient's ID in study
   - `diagdate`: date of diagnosis of breast cancer
   - `recdate`: date of recurrence (the cancer came back) or of "recurrence-free survival" (the cancer went away) after treatment.
   - `deathdate`: date of death
   - `age`: age at diagnosis
   - `menopause`: patient has reached menopause, 1=yes, 2=no
   - `hormone`: patient had hormone therapy, 1=yes, 2=no
   - `size`: tumour size (mm)
   - `grade`: tumour grade (1–3)
   - `nodes`: number of nodes involved
   - `prog_recp`: number of progesterone receptors
   - `extrg_recp`: number of estrogen receptors
   - `rectime`: time to recurrence or survival (days)
   - `censrec`: whether cancer came back (1) or did not (0)
   - `survtime`: time to death (days)
   - `censdead`: whether died (1) or survived (0)

   Our aim in this question is to predict the time to recurrence of the breast cancer, and to see how that depends on any of the other variables. The data are summarized in Figure 7.

   (a) (2 marks) Figure 8 shows some analysis of these data. Given our aims here, explain briefly how the `Surv` statement uses the right variables in the right way.

   (b) (3 marks) I did some preliminary analysis (not shown) to determine which explanatory variables to keep. My final model is the one shown in Figure 8. Consider each of the three explanatory variables in turn. Based on the information in Figure 8, what effect does each variable have on the breast cancer coming back? Be specific about the nature of the effect in each case. Assume that $\alpha = 0.10$ for this question.

(c) (2 marks) What is the code in Figure 9 doing? There are two things to comment on: first, the purpose of the first four lines of code (these together achieve *one* thing, that you need to describe) and second, specifically what `pp` contains (that you would see if you looked at a `summary` of `pp`).

(d) (2 marks) On the plot in Figure 11, which of the values of the explanatory variables are associated with the "best" survival, that is, having the best chance of the breast cancer taking the longest to return? Explain briefly. The variables in `combo` are listed in the same order as the variables in `new` in Figure 9. In this part and the next, if you cannot distinguish the colours on the plot, ask.

(e) (4 marks) For each of the three explanatory variables in the proportional-hazards model, explain briefly how their effect on the survival plot is the same as the effect you described back in part (b). Or explain how it is inconsistent, if that is what you see.

4. Three new textbooks are being tried out for (university) students learning intermediate French. The textbooks are labelled `a`, `b` and `c`. 15 students take part in a study, with five being randomly assigned to each of the new textbooks. These students have all previously passed an introductory French course. Also recorded is each student's (cumulative) grade point average, `gpa` in the data set. At the end of the study, each student takes a French test and the score is recorded. The test covers what the students are supposed to have learned from the textbook they used.

   The data set is shown in Figure 12.

   (a) (2 marks) An analysis of variance is shown in Figure 13. I checked that the assumptions for this analysis of variance were satisfied. What do you conclude from Figure 13, in the context of the data?

   (b) (2 marks) Figure 14 shows a scatterplot of test score against GPA, with the observations labelled by which textbook the student used. Do you think there is an effect of GPA on test score? Explain briefly.

   (c) (2 marks) In Figure 14, do you think there is a difference in test scores among the textbooks? Explain briefly.

   (d) (2 marks) What do you conclude from the analysis `french.2` in Figure 15? What does that mean for the data? Explain briefly.

(e) (2 marks) Explain what the `update` statement in Figure 15 is doing, and why it is a sensible thing to do.

(f) (2 marks) What do you learn from the first part of the output of `french.3` (the `anova`)? There are two things to note.

(g) (3 marks) Compare your conclusions in parts (a) and (f). Are they consistent or inconsistent? Why did they come out that way? Explain briefly.

(h) (2 marks) Look at the coefficients of `bookb` and `bookc` in the last part of Figure 15 (the `summary` of model `french.3`). What do those numbers mean, in the context of the data?

5. Ten people with low self-esteem took part in a study that compared a standard therapy method (labelled `baseline` in the data) with a new therapy (labelled `new`). Five people were randomly allocated to each therapy. Each person underwent the appropriate therapy for four weeks, and a self-esteem score was assessed for each patient at the end of each week. The data are shown in Figure 16.

   (a) (2 marks) What does it mean to say that these data are "repeated measures"? Why would a two-way analysis of variance (based on therapy and weeks) not be appropriate? (That is, what would a two-way analysis assume, that is actually not true?)

   (b) (3 marks) What do you conclude from the analysis in Figure 17? (There are three things to conclude.)

   (c) (2 marks) Figure 18 shows the code to do some processing of the data. Describe what the output data frame `selfesteem.long` looks like. (I'm looking for a description in words.)

   (d) (1 mark) Figure 19 shows a spaghetti plot of the observations for each person against time. Which part of the code (above the plot) produces a *separate* line for each person?

   (e) (3 marks) Describe how the spaghetti plot supports *each of the three* of your conclusions from part (b).

6. A market research company commissioned a survey of families in order to understand what makes a family choose to subscribe to one magazine rather than another. We will investigate some data from families that subscribe to exactly one of four magazines (listed below). In the survey, a lot of demographic information was collected, as described below:

   - `id` of family (ignore in analysis)
   - `magazine` subscribed to, a number indicating which magazine:
     1. Better Homes and Gardens
     2. Reader's Digest
     3. TV Guide
     4. Newsweek
   - `i1` through `i4`: ignore
   - `famsize`: number of people in family, with 6 indicating "6 or more".
   - `income`: family income, coded with 1 meaning "less than $10,000" up to 11 meaning "$75,000 and above".
   - `race`: 1 is "white", 0 is "other".
   - `tv`: number of televisions (3 is "3 or more").
   - `newspaper`: whether family subscribes to a newspaper (1=yes, 0=no)
   - `nomale`: whether family has a father living in the house (1=no father, 0=father)
   - `nofemale`: whether family has a mother living in the house (1=no mother, 0=mother)
   - `child18`: whether there are any children over 18 living in the house (1=yes, 0=no)
   - `headage`: age group of head of household, from 1 (18–29) to 6 (65 and above)
   - `headeduc`: education level reached by head of household, from 1, "some grade school" to 7, "went to graduate school".

   The structure of the data is shown in Figure 20. Some of these variables can take only the values 1 and 0, but that is not a problem for the analysis that is done here.

   (a) (2 marks) Why could discriminant analysis help the market research company with their aims here?

   (b) (3 marks) Look at Figure 21. Which two variables does the first linear discriminant mainly depend upon? When will the first discriminant score be large? Explain briefly.

   (c) (2 marks) Figure 22 shows a plot of the first two discriminant scores. What does this plot suggest about the ability of the demographic variables to distinguish the magazines that a family might subscribe to? Explain briefly.

(d) (3 marks) Explain briefly what the table in Figure 23 is telling you. In particular, what does the number 16 in the table tell you?

(e) (2 marks) If the discriminant analysis were doing a better job of explaining magazine readership, how would the table in Figure 23 be different? Explain briefly.

(f) (2 marks) Are the table in Figure 23 and the plot in Figure 22 telling a consistent or an inconsistent story? Explain briefly.

(g) (3 marks) Figure 24 shows the predicted group membership and posterior probabilities for a sample of families. Note that the output is too wide to fit on the paper, so continues below as a second table with the same row names. For the family with `id` 132, what is the name of the magazine they actually subscribed to? What is the name of the magazine they are predicted to subscribe to?

(h) (2 marks) Would you say the prediction for family 132 in Figure 24 was clear-cut or not? Explain briefly.

7. How do people perceive the similarity of different kinds of car? In one study, several people were asked to assess the similarity of the following cars, all (at the time) priced between \$30,000 and \$35,000: BMW 328i, Ford Explorer, Infiniti, Jeep Grand Cherokee, Lexus, Chrysler Town and Country, Mercedes, Saab 900, Porsche Boxster, Volvo. Each person was given a list of all 45 possible pairs of these cars, and was asked to rank the pairs from most similar (1) to least similar (45). The results for one person are shown in Figure 25. Some of the car names have been abbreviated; if it is not clear which car they are, ask.

   (a) (2 marks) Why are all those missing values in the data in Figure 25 not going to be a problem?

   (b) (2 marks) A multidimensional scaling analysis is carried out in Figure 26. Why is `isoMDS` a better idea for these data than `cmdscale`?

   (c) (2 marks) Will the non-metric multidimensional scaling produce a map that does a good job of reproducing the dissimilarities? Explain briefly, giving a number from one of the Figures that helps you decide.

   (d) (1 mark) Why did I need to use `xlim` in the code above Figure 27? Explain (very) briefly.

   (e) (3 marks) Look at the map in Figure 27. The cars overwriting each other are Ford and Jeep top left, and Infiniti and Lexus at the bottom. Find the Chrysler Town and Country and Porsche Boxster on the map. Are they close together or far apart on the map? Give *two* reasons why that is what you would have expected to see, given the other Figures.

8. I collected some statistics on the major-league baseball teams in 2015. Each team played 162 games. The data are shown in Figure 28. The variables have abbreviated names, which stand for these:

   - `TEAM`: team's name
   - `RS`: total runs scored by the team
   - `H`: total hits made by the team
   - `HR`: total home runs hit by the team
   - `AVG`: team's overall batting average (number of hits divided by number of total number of "at-bats"). Higher is better.
   - `W`: wins by the team (number of losses not included since it is 162 minus wins)
   - `ERA`: "earned run average": a measure of how good the team's pitchers were, a lower number being better
   - `RA`: total runs allowed by the team
   - `SO`: strikeouts made by the team's pitchers (more is better)
   - `BB`: "bases on balls" or "walks" allowed by the team's pitchers (lower is better)
   - `E`: errors made by the team's fielders (lower is evidently better).

   (a) (2 marks) I ran a principal components analysis in Figure 29. In the `princomp`, why did I use `baseball[,-1]` instead of just `baseball`?

   (b) (2 marks) Why was it a sensible idea, considering the nature of the data, to use `cor=T` in Figure 29? Explain briefly.

   (c) (2 marks) Use the scree plot in Figure 30 and the output in Figure 29 to identify an appropriate number of components to use. Justify your choice briefly.

(d) (2 marks) I decided to try a factor analysis. I ran the factor analysis with 4 factors (this may or may not have been indicated by your choice above), with the results shown in Figure 31. Are four factors enough, or should I have used more? Explain briefly.

(e) (4 marks) Baseball fans might divide the variables up into hitting-related ones (`RS, H, HR, AVG`), pitching-related ones (`ERA, SO, BB, RA`), fielding-related ones (`E`), and one that relates to overall team quality (`W`). Discuss how well these correspond to the high loadings on our four factors.

(f) (2 marks) Are there any variables that are not included on one of my four factors? What is the best way to tell? Explain briefly.

(g) (2 marks) Figure 32 shows a biplot of the factor analysis. The Toronto Blue Jays are team #1 on the plot. Which variables would you expect to be especially high or low for the 2015 Blue Jays? Explain briefly, basing your assessment on what you see on the biplot (not on what you remember from last summer!)

9. A study is investigating (categorical) variables associated with the birth of a child. These variables are assessed for each of 738 childbirths:

   - the gender of the child (male or female)
   - premature rupturing of membranes, called `rupture` (yes or no). (Rupturing of the membranes normally happens when the mother goes into labour, when it is known as "breaking the water". Here, it is noted whether rupture happens earlier than that.)
   - whether labour was induced (yes or no)
   - whether the child was born by cesarean section (yes or no)

   These variables are listed in the chronological order that makes sense. The data for this study are listed in Figure 33. The ultimate aim is to discover which variables, if any, are associated with the need for a mother to have a Cesarean section.

   (a) (2 marks) In Figure 34, only the start and finish of my analysis is shown. Why did I stop at model `births.10`?

   (b) (2 marks) Why did I choose to obtain the particular results that I obtained in Figure 35, and not others? Explain briefly.

   (c) (4 marks) Suppose we are looking for factors that are related to the mother having a Cesarean section (`cesarean` in the data). Use Figures 34 and 35 as necessary to describe which other factors are related to the mother having a Cesarean section, and to describe how those factors are related.