University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 / STA 1007 (K. Butler), Final Exam
April 6, 2017, 7:00–10:00pm (3 hours)

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 12 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	6	
2	8	
3	10	
4	11	
5	8	
6	11	
7	8	
8	13	
9	3	
10	6	
11	8	
12	9	
Total:	101	

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. American football is a tough physical game, and players wear helmets to protect their heads. There is concern, however, about a possible association between helmet design and neck injuries. As a preliminary to a study of this association, data were collected on three groups of individuals, labelled `group` in the data: high school football players (labelled 1 in the data set), college football players (labelled 2) and non-football players (3). For each individual, six head measurements were made, as follows:

   - `wdim`: head width (at its widest)
   - `circum`: head circumference
   - `fbeye`: front-to-back head size at eye level
   - `eyehd`: eye to top of head measurement
   - `earhd`: ear to top of head measurement
   - `jaw`: jaw width.

   All of the measurements are in inches.

   Some randomly-chosen rows of the data are shown in Figure 2. Our aim is to predict which one of the three groups an individual is in, based on their head measurements.

   (a) (2 marks) What two things about the data suggest that I should use `multinom` for the modelling?

   (b) (2 marks) Some modelling is shown in Figure 3. Explain briefly in words what the `update` line is doing.

   (c) (2 marks) What do you conclude, in the context of the data, from the test done in Figure 3? Explain briefly.

(d) (3 marks) I did some predictions for the model `football.2`. (You may assume that this is a reasonable model to be predicting from.) These are shown in Figure 4. Using the information in this Figure, what can you say about an individual who has head width (at its widest) 15, eye-to-top-of-head 10, ear-to-top-of-head 13.5, and jaw width 12.5 inches?

(e) (1 mark) The data for this problem had one individual per row of the data file. Suppose, instead, each row represented several individuals, with a column called `count` saying how many individuals were summarized in that row. How would I have to change my `multinom` statement to accommodate this?

2. The "windshield" method of distributing questionnaires consists of randomly placing the questionnaires under the windshield wipers of cars in parking lots. What method leads to the best response rate (that is, the fraction of questionnaires filled out and returned)? One idea is to use brightly-coloured paper to print the questionnaires on. In one study, blue, green and orange paper was used. 15 supermarket parking lots in a metropolitan area were chosen. One colour was randomly chosen for each parking lot, and questionnaires of that colour were distributed. Does the colour of the paper make a difference to the response rate? One of the researchers thought the size of the parking lot (number of parking spaces contained in the entire lot) might also matter, so this was recorded as well. The data are shown in Figure 5. The response rate is in percent.

   A scatterplot of the data is shown in Figure 35 (at the end of the booklet of code and output, this plot being in colour), with response rate plotted against parking lot size, with the points coloured according to the colour of the questionnaire paper and regression lines added for each colour. (The purpose of the last line of code above the plot is to make the colours on the plot correspond to the colours of the paper. Otherwise the line labelled "blue" would actually have been coloured red!)

   (a) (2 marks) Looking at Figure 35, how does it appear that parking lot size affects the response rate, if at all? Explain briefly.

   (b) (2 marks) Again looking at Figure 35, does it appear that the colour of the paper the questionnaire is printed on has any effect on the response rate, or not? Explain briefly.

(c) (2 marks) In Figure 6, two models are fit and compared. Explain briefly why I can remove the interaction between size and colour, based on the information in this Figure.

(d) (2 marks) What does the model without interaction mean for these data?

(e) (3 marks) Interpret each of the `size`, `colourgreen` and `colourorange` estimates in Figure 7. That is, explain what each of those three numbers mean, in the context of the data.

(f) (3 marks) Figure 8 shows another analysis. What does this one show, and how, despite appearances, is it consistent with what you concluded earlier? Explain briefly.

3. Two treatments, labelled A and B, for post-traumatic stress disorder (PTSD) are being compared. Nine patients are available; each one is randomly allocated to one of the two treatments, labelled `trt` in the data. The two treatments include different amounts of supportive counselling and other therapies. The response variable is the number of symptoms of PTSD observed by the therapist (so a smaller number is better).

   The research questions are:

   1. Are the treatments doing something (changing the number of symptoms over time, ideally reducing the number)?
   2. Is one treatment more effective than the other?
   3. Is the pattern of change in symptoms over time different for the two treatments?

   Each patient was observed on three occasions: before therapy (`pre`), after therapy (`post`), and three months after therapy ended (`followup`). The data are shown in Figure 9.

   (a) (2 marks) Why is a repeated-measures analysis necessary here? Explain briefly.

   (b) (3 marks) A repeated-measures analysis of variance is shown in Figure 10. From this analysis, what do you conclude about each of the three research questions given above, in the context of the data?

   (c) (2 marks) A spaghetti plot is shown in Figure 36 (at the end of the booklet of code and output). In the code above the plot, why did I need the `mutate` line? (Think about what would have happened if I had left it out.)

   (d) (4 marks) How does the spaghetti plot support each of your conclusions of part (b)? In addition, does the spaghetti plot suggest that the treatments have a beneficial effect? (“Beneficial” means “does good”.) Explain briefly.

4. In each of the cases below, describe specifically how the *data* appropriate for the two techniques given would be different.

   (a) (2 marks) Logistic regression using code as in Figure 11, and using `polr` from package `MASS`.

   (b) (2 marks) ANOVA and MANOVA. (You need to know what these two abbreviations mean.)

   (c) (2 marks) Analysis of variance and analysis of covariance.

   (d) (2 marks) Hierarchical clustering, as with `hclust`, and K-means clustering.

5. In this question, I will show you some examples of R code and ask you to explain *in words* what each code example does.

   (a) (3 marks) Figure 12 shows a data frame `x` and some code, on the last line. The results of the pipe on the last line are saved in a variable `y`. What will `y` look like?

   (b) (2 marks) Using the same data frame `x` as above, look at Figure 13. What does `w` contain?

   (c) (2 marks) Look at Figure 14. Starting from the vector `xx` as shown, what does the vector `yy` contain? (We have used the function `sqrt` in this course, so you need to know what it does.)

   (d) (4 marks) Look at Figure 15. There, a function `f` is defined, a data frame `data` is shown, and a last line of code saves something into a variable `res`. What, in words, does the function `f` do, and therefore what does `res` contain? For full credit, explain how `res` is laid out.

6. The data set we use for this question consists of 88 observations of a number of variables taken along a 0.8km stretch of Summit Creek in eastern Oregon. The members of the research group that collected these data were interested in the effect of cattle grazing on the nature of the creek. They divided the creek into three zones, labelled `Reach` in the data set: an upstream area A in which cattle were permitted to graze, a middle area B from which cattle were excluded, and a downstream area C in which cattle were again permitted to graze. In each zone, the creek was divided into "hydrologic units" such as pools, shallow sections or straight sections, such that the variables measured (the widths and depths described below) would be expected to be similar within a hydrologic unit but possibly different for different hydrologic units.

   The variables measured for each hydrologic unit were these:

   - `DepthWS`: depth from water surface to bottom (metres)
   - `WidthWS`: width at water surface (metres)
   - `WidthBF`: width at "bankfull stage" (metres). This is a way of measuring the width of a creek that does not depend on how much water it contains. (That's all I know. I have no more information.)
   - `HUAreaWS`: area of hydrologic unit at water surface (square metres)
   - `HUAreaBF`: area of hydrologic unit at bankfull stage (square metres)
   - `wsgrad`: water-surface gradient (dimensionless).

   Some of the data is shown in Figure 16. There are other variables that will not be used in this question.

   (a) (2 marks) What do you conclude from the analysis in Figure 17 about the zones of the creek?

   (b) (2 marks) What would a discriminant analysis tell us about the creek that would be worth knowing? Explain briefly.

   (c) (2 marks) The variables are measured on very different scales, so I decided to standardize them before proceeding. The calculation is shown in Figure 18. The variables in the data frame `summit.s` have the same names as variables in data frame `summit`, but all the numeric variables in `summit.s` have been standardized. The standardized variables are used in the rest of the question. (This is for your information; there is no question here. The question for this part is in the short paragraph below.)

   Figure 19 shows the results of a discriminant analysis on the (standardized) data. Why are there two linear discriminants?

   (d) (2 marks) Do you think both linear discriminants are worth paying attention to, or not? Explain briefly.

(e) (3 marks) Which of the variables in `summit.s` does the first linear discriminant mainly depend on? What would make a hydrological unit have a *large* score on `LD1`? Explain briefly.

(f) (3 marks) Which three variables make the largest contribution to `LD2`, and what, therefore, would make a hydrological unit's `LD2` score *small*? Explain briefly.

(g) (3 marks) Zone B is the only one in which cattle are not allowed to graze. Looking at either or both of Figure 37 and Figure 38 (in colour at the end of the booklet of code and output), how is zone B different from the other zones? What does that imply about the values on the measured variables for hydrological units in zone B?

(h) (2 marks) Look again at Figure 37 and Figure 38. Is there an indication that `LD2` does anything to distinguish the zones? Explain briefly.

(i) (2 marks) Would you say the zones are easy to classify on the basis of the measured variables, or not? Explain briefly.

This page: ______________ of possible 13 points.

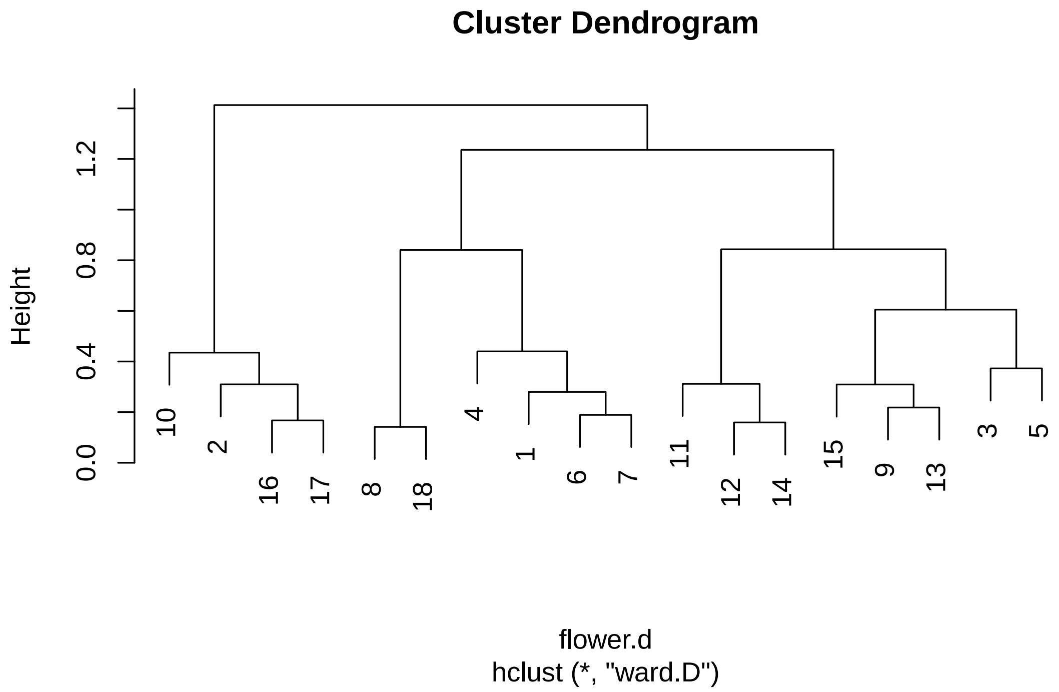
7. Data were collected on 18 popular kinds of flowers, numbered 1–18. For each flower, the following was recorded:

   - `winters`, 1 if the flower may be left in the ground when it freezes, 0 if not.
   - `shadow`, 1 if the flower grows in the shade, 0 if it need to grow in sunshine.
   - `tubers`, 1 if the flower has tubers, 0 if not
   - `colour` (unordered):
     1. white
     2. yellow
     3. pink
     4. red
     5. blue
   - `soil`: grows in dry (1), normal (2) or wet (3) soil. This should be treated as ordered.
   - `preference`: a reviewer's preference ranking, from 1 (best) to 18 (worst). Treat as ordered.
   - `height` in centimetres (a number)
   - `distance` how much space, in centimetres, that should be left between the flowers (a number).

   The data are shown in Figure 20. A `dist` object called `flower.d` is created (code not shown) and used as input into a cluster analysis. The code for the cluster analysis is shown in Figure 21.

   (a) (3 marks) The dendrogram for the cluster analysis in Figure 21 is shown below. How many clusters do you think the flowers should be divided into? Explain briefly. (There is a little space for writing below the dendrogram.) Draw your clusters on the dendrogram, in the same kind of way that `rect.hclust` does, so that it is clear which of the flowers belong in which cluster.

```
## Warning: package 'cluster' was built under R version 3.5.2
```



(b) (3 marks) Pick two flowers that are in the same one of your clusters (any such pair of flowers will do). By looking back at the original data in Figure 20, does it make sense that the two flowers you chose would be in the same cluster? Explain briefly. (The numbers of the flowers on the dendrogram are the same as the row numbers in the original data frame. You might like to copy the appropriate rows of the data frame `flower0` here to refer to them.)

(c) (3 marks) The same data underwent a multidimensional scaling, with code shown in Figure 22 and map shown in Figure 23. In part (b) you chose a pair of flowers. Find those same two flowers on Figure 23. (Do not hand in Figure 23.) Are your flowers close together or far apart? Why would you not be so surprised if they were not close together on the map? Explain briefly.

8. Some socioeconomic variables were measured for twelve census districts in Los Angeles (in 1976). The variables were:

   - `district`: a single letter identifying the district
   - `population`: total population
   - `school`: median school years of district residents
   - `employment`: total number of employed people living in district
   - `services`: professional services, whatever that is, in suitable units
   - `housevalue`: median house value in the district (in 1976 dollars).

   We are interested in what makes the districts different, or, equivalently, seeing whether we can find fewer variables to describe them with.

   (a) (2 marks) A principal components analysis is carried in Figure 25, and a scree plot is shown in Figure 26. What is an appropriate number of principal components to use? Explain briefly, using *both* Figures to help you decide.

   (b) (2 marks) Component loadings are shown in Figure 27. A blank loading is close to zero and can be ignored. What about a district would make it have a low (negative) score on *component 2*? Explain briefly.

   (c) (4 marks) A plot of the first two principal component scores, with the districts labelled, is shown in Figure 28. Find district D on this plot. Where is it? What should be unusual about this district, given where it is on the plot? Is that the case, looking back at the original data in Figure 24?

9. A study was made of the classroom behaviour of school students, and factors relating to this. 97 students took part in the study. Over the period of the study, each student's behaviour was classified by the teacher as "deviant" or "non-deviant" (that is, unacceptable or acceptable). In addition, each student was classified as "at risk" or "not at risk" based on their home background, and the adversity level of the student's school was classified as low, medium or high. The primary aim of the study was to investigate the effect of home background and school on behaviour.

   (a) (2 marks) Data of this kind are usually presented as a contingency table, as shown in Figure 30. (The function `ftable` displays the contingency table in the format shown.) Explain briefly why the data frame shown in Figure 29 is more suitable for our analysis.

   (b) (2 marks) The first stage of my analysis is shown in Figure 31. My next step is to remove the `behaviour:risk:adversity` term. Why? Explain briefly.

   (c) (2 marks) The rest of my analysis is shown in Figures 32 and 33. Was my analysis appropriate? Explain briefly why or why not, and if my analysis was not appropriate, what I should have done instead.

   (d) (3 marks) Figure 34 shows some contingency tables for subsets of the data. Based on my analysis, or the correct version of it if my analysis was incorrect, and using the appropriate tables from Figure 34, what do you conclude about the data, especially about any impact of risk and adversity on behaviour? (If you need to see a different subtable to draw your conclusion, explain what subtable you need to see and how you would use it to draw your conclusion.)