University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 / STA 1007 (K. Butler), Final Exam
April 10, 2019

Aids allowed:

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

This exam has 11 numbered pages of questions. 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).
The maximum marks available for each part of each question are shown next to the question part.
You may assume throughout this exam that the code shown in Figure 1 of the booklet of code and output has already been run.
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.

Question 1 (9 marks)

An experiment is run to test what effect the dose of a drug (measured in mg) has on how lethargic a rat is. This is measured by the amount of time a rat spends sleeping or resting in a four-hour period. It is also suspected that the age of a rat (measured in months) will have some impact on how much time the rat spends sleeping or resting. The data are shown in Figure 2. The column `dose` is a number, but is treated here as a factor.

(a) (2 marks) What feature or features of this data set mean that analysis of covariance is a suitable method to analyze it? Explain briefly.

(b) (3 marks) A scatterplot is shown in Figure 23. What does this plot suggest about effects of age, dose and interaction between them? Explain briefly. (Note: the graphs in colour are at the end of the Booklet of Code and Output. I apologize in advance for any flipping back and forth you need to do.)

(c) (2 marks) An analysis of covariance for these data is shown in Figure 3. What do you conclude from it, in the context of the data?

(d) (2 marks) In class, we didn't talk about simple effects in this kind of model, but describe briefly how you might use a simple-effects idea to understand your conclusion for this data set.

Question 2 (19 marks)

An experiment is conducted, with the results shown in Figure 4. The columns are: treatment (labelled A, B, or C), time point (labelled T1, T2, and T3), subject ID, and a response variable value, called `y`. There are 27 observations in total.

(a) (1 mark) How many different treatments does each subject undergo?

(b) (2 marks) What feature of the data indicates that a repeated-measures analysis of variance will be necessary? Explain briefly.

(c) (3 marks) Give code to create a spaghetti plot for these data (that is, a plot showing how the response variable changes over time for each treatment, with the values for each subject joined by lines). Assume that the data frame shown in Figure 4 is called `rm`.

(d) (2 marks) What advantage does a spaghetti plot have over an interaction plot for repeated-measures data? Explain briefly.

(e) (3 marks) The spaghetti plot produced by your code is shown in Figure 24. What does this plot suggest about likely (i) treatment effect, (ii) time effect, (iii) treatment-time interaction? Explain briefly.

(f) (4 marks) What code would run a suitable repeated-measures ANOVA, using `Manova`? Pay close attention to the layout of the data in Figure 4, which is the data frame `rm`.

(g) (2 marks) The analysis for which you gave code in the previous part is shown in Figure 5. What do you conclude from this, in the context of the data?

(h) (2 marks) Why do you think that taking out the first time point *would not* change the significance of the interaction as you found it in the previous part, in contrast to the dogs example in class? (Taking out the first time point may change the P-value, but would not change whether or not the interaction is significant at, say, $\alpha = 0.05$.)

Question 3 (23 marks)

Crude oil samples were taken from sandstone of different types, known as "zones". The three zones are Wilhelm, Sub-Mulinia, and Upper (Mulinia). The zone names are abbreviated in the data set. The aim is to see whether the following measurements are associated with the zone of sandstone from which the oil sample was taken:

- vanadium
- iron
- beryllium
- saturated hydrocarbons
- aromatic hydrocarbons.

A random sample of the data set is shown in Figure 6.

(a) (2 marks) Why might MANOVA be a sensible method of analysis for these data? Explain briefly.

(b) (2 marks) A MANOVA analysis is shown in Figure 7. What do you conclude from it, in the context of the data?

(c) (2 marks) Given the results of the MANOVA, why might we want to do a discriminant analysis? Explain briefly.

(d) (2 marks) Based on Figure 8, how many linear discriminants should we use? Explain briefly.

(e) (3 marks) Which of the original variables contribute to LD1, and how do they do so? (For example, what kind of values would make LD1 large?)

(f) (2 marks) I created a plot of the LD scores, shown in Figure 25. I did this by making predictions, then creating a data frame `d` containing both the original data and the predictions. The code I used to make the plot is as shown. (The numbers beside the points are the numbers of the oil samples in the data set.)

Would you say that the zones are relatively distinct, or not? Explain briefly.

(g) (2 marks) For oil samples from the Upper zone, what in terms of the original variables distinguishes them from the other zones? Explain briefly.

(h) (3 marks) My data frame `d` contains all the original data plus a predicted zone for each observation. What R code would use `d` to calculate the proportion of all the observations that were misclassified, that is, for which the predicted zone is different from the actual zone? (If you also wish to calculate the proportion that were *correctly* classified, that is fine too.)

(i) (2 marks) What *change* to your code of the previous part would obtain the misclassification proportions for *each* zone?

(j) (3 marks) Find an observation in Figure 25 that could be misclassified, given where it is on the plot, and justify your choice briefly. The observations are numbered according to which oil sample they are. By assessing the appropriate row of Figure 9, describe briefly whether there is doubt about the zone of the oil sample you chose, given its values on the quantitative variables. (The column `r` contains the numbers of the oil samples.)

Question 4 (14 marks)

For people who enjoy listening to music, the loudspeaker is an important part of the listening experience, because the quality of the loudspeaker has a big impact on the quality of the sound that is produced. People who like listening to high-quality sound are called "audiophiles". A magazine for audiophiles tested 19 brands of mid-sized loudspeakers for several characteristics:

- Price: manufacturer's suggested list price, in dollars
- Accuracy: how accurately the loudspeaker can reproduce every frequency in the musical spectrum (scale of 0 to 100, higher better).
- Bass: how well the loudspeaker handles very loud bass notes (scale of 1 to 5, higher better).
- Power: the number of watts per channel needed to reproduce moderately loud music.

Our aim is to group the loudspeakers, labelled A through S, into clusters of similar ones. The data are shown in Figure 10.

(a) (3 marks) The numerical values in Figure 10 are on very different scales. Give code to create a data frame called `speakers.s` that contains only the four numerical columns, and replaces the values shown in Figure 10 with their standardized values. For full credit, do this *without naming any of the numerical columns.*

(b) (2 marks) What code would produce a K-means clustering of the standardized data, obtaining four clusters, and allowing suitably for the fact that K-means is a randomized algorithm so may not produce the same result every time?

(c) (4 marks) Give code to create a scree plot, which will enable us to choose a sensible number of clusters to divide the loudspeakers into. Go up to 15 clusters.

(d) (2 marks) My scree plot is shown in Figure 11. What do you think is a sensible number of clusters? Explain briefly.

(e) (3 marks) Describe a procedure by which you could obtain a graph of the results with your chosen number of clusters. I am looking for a description in words.

Question 5 (17 marks)

How do crabs of the species *Leptograpsus variegatus* differ from one other? Body measurements were taken on 200 of these crabs, collected at Fremantle, Western Australia. For each crab the following was recorded:

- `sp`: colour, blue (B) or orange (O).
- `sex`: male (M) or female (F).
- `index`: ID number, 1 through 200.
- `FL`: frontal lobe size.
- `RW`: rear width.
- `CL`: carapace length.
- `CW`: carapace width.
- `BD`: body depth.

All the body measurements were in millimetres. A sample of the data is shown in Figure 12.

(a) (2 marks) A principal components analysis is shown in Figure 13. Why did I use `select_if(is.double)` in my code? Explain briefly.

(b) (2 marks) A scree plot is shown in Figure 14. What do you conclude from this? Is your conclusion consistent with Figure 13? Explain briefly.

(c) (2 marks) The component loadings are shown in Figure 15. The first principal component is often a measure of "size". Do you think that has happened here? Explain briefly.

(d) (2 marks) Which of the original variables is most associated with component 2? Explain briefly. (Note: your previous answers may have said not to look at component 2. If that's what they said, humour me and look at component 2 here anyway.)

(e) (3 marks) A plot of the principal component scores on the first two components is shown in Figure 16. The crabs are labelled on the plot by the value of `index`. Find two crabs that are shown in Figure 12 and that differ substantially on component 1. By looking at Figure 12, explain how those crabs differ. (Note: `geom_text` plots text *at* the points, as opposed to `geom_text_repel`, which adds text *next to* the plotted points.)

(f) (3 marks) Find two crabs on Figure 16 that are also on Figure 12 and differ substantially on component *2*. By looking at Figure 12, explain how those crabs differ.

(g) (3 marks) Another plot of component scores is shown in Figure 26. What do you conclude from this plot, and how does that relate to the original variables that were measured? Explain briefly.

Question 6 (9 marks)

Basketball fans often believe in the "hot hand": a player who has just successfully made a shot is more likely to make the next one as well. A study was made of this in the early 1980s. The study used free throws (also known as "foul shots"), because a free throw is always taken from the same place, and opposing players are not allowed to interfere with the shot.

A player that is fouled while shooting (and also at certain other times during the game) is awarded two free throws. Is a player who makes their first free throw, when they are awarded two, more likely to make the second one? The data we use comes from the Boston Celtics, 1980–1982; this is the data that was used for the original "hot hand" study, and is shown in Figure 17. The data is shown in "long" format, with one column of frequencies: the number of times the player shown "hit" (made) or missed the first free throw, and hit or missed the second one. For example, Larry Bird missed his first free throw and made the second one 48 times during the period the data were collected.

(Basketball fans among you will note that I have simplified things a little: a player who is fouled while shooting and *makes the shot anyway* only gets one free throw. Such single shots are not counted here.)

(a) (2 marks) The first analysis totalled up over all players. This is shown in Figure 18. The first part of the output shows that when the first shot was hit, the second shot was hit 79% of the time; when the first shot was missed, the second shot was hit 74% of the time. The bottom part of the output shows that this small difference is significant; there is a (small) association between hitting or missing the first shot and hitting or missing the second one. (The test is an ordinary chi-squared test for independence.)

Figure 19 shows the proportion of *second* shots hit by each player according to whether they hit or missed the first one. Using Figure 19, criticize the analysis in Figure 18.

(b) (2 marks) A log-linear analysis is shown in Figures 20 through 22. Why is Figure 22 a good place to stop? Explain briefly.

Question 6 continues…

(c) (2 marks) What do the effects remaining in Figure 22 tell you, in the context of the data?

(d) (3 marks) What does the log-linear analysis in Figures 20 through 22 tell you about the evidence for a hot-hand effect? How is that consistent with Figure 18? Explain briefly.