University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 (K. Butler), Final Exam
April 18, 2024

Aids allowed (on paper, no computers):

- 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 plus this cover page.
In addition, you have an additional booklet of Figures to refer to during the exam.
The maximum marks available for each part of each question are shown next to the question part.
If you need more space, use the last page of the exam. Anything written on the back of the page will not be graded.
You may assume throughout this exam that the code shown in Figure 1 of the booklet of Figures 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.

1. When babies are born with low birth weight, defined as birth weight less than 2500 grams, physicians are concerned, because infant mortality rates and birth defect rates are higher for low birth weight babies. A woman's behaviour during pregnancy (including diet, smoking habits, and receiving prenatal care) can greatly alter the chances of carrying the baby to term, and, consequently, of delivering a baby of normal birth weight. We will focus on predicting whether or not a baby has low birth weight, in column `low`, from the mother's weight at her last menstrual period (`lwt`) and whether or not the mother smoked during pregnancy (`smoke`). The two categorical variables have levels `yes` and `no`; the weight is measured in pounds. Some of the data, in dataframe `birthwt`, is shown in Figure 2.

   (a) [2] Why is it sensible to analyze these data using logistic regression?

   (b) [2] Figure 3 shows the results of running a logistic regression. How do you know that it is predicting the probability that a baby *is* of low birth weight (rather than the probability that a baby *is not*)?

   (c) [2] In Figure 3, is there any reason to remove any explanatory variables? Explain briefly.

   (d) [2] According to Figure 3, precisely what is predicted to happen when a woman's weight at last menstrual period increases by 1 pound, all else equal?

(e) [2] Some predictions are shown in Figure 4. Describe the effects of mother's weight at last menstrual period and of smoking on the probability of the baby being of low birth weight.

(f) [2] Explain briefly how the predictions in Figure 4 are consistent with the output in Figure 3.

2. An air traffic controller must be able to respond quickly to an emergency condition indicated on her display panel. Three (numbered) types of display panel were compared. Each panel was tested under four (numbered) simulated emergency situations. Two well-trained controllers were assigned to each of the 12 combinations of emergency condition and display panel type, for a total of 24 controllers in all, and the time taken to respond to the emergency situation was recorded. A lower time is better.

   The data, in dataframe `display`, are shown in Figure 5. The numbered display panel types, in `panel`, and emergency situations, in `emergenc`, are already `factor` variables, and so will be treated as categorical in the analyses that follow.

   (a) [2] Two analyses are shown in Figure 6. What do you conclude from the first analysis in `display.1`?

   (b) [3] Why is it necessary to do the analysis in `display.2` in Figure 6? What do you conclude from it?

(c) [2] What do you conclude from the analysis in Figure 7, in the context of the data? (If you think that it was not appropriate to run this analysis, explain briefly why.)

(d) [2] The main interest in this study was the display panels. On the basis of the information given here, which panel or panels would you recommend? Explain briefly.

(e) [3] Why would it be a mistake to run a simple effects analysis here? What do you expect would happen if you did run a simple effects analysis of display panels for each emergency type? Explain briefly.

3. The observations in dataframe `vocab` are respondents to U.S. General Social Surveys, 1972-2016. I selected a random sample of 100 respondents (in total) from the three years 1974, 1994, 2014 from a much larger number of respondents in several different years. The three variables of interest to us are

   - `year`: the year of the survey (1974, 1994, or 2014)
   - `education`: the number of years of education (recorded as a whole number)
   - `vocabulary`: the number of items correct on a 10-word test (also recorded as a whole number).

   Some of the data are shown in Figure 8.

   (a) [2] A scatterplot is shown in Figure 9. Why did I use `geom_jitter` rather than `geom_point`?

(b) [2] Based on what you know or can guess about education and vocabulary, does the general trend of the lines on the plot of Figure 9 make sense? Explain briefly.

(c) [2] An analysis of covariance is shown in Figure 10. Why is it correct to use the `drop1` table to assess the significance of the interaction, and incorrect to use the `summary` output to do this?

(d) [3] Is there a significant interaction between education and year? What does the significance (or non-significance) of that interaction term tell you about what you see on the scatterplot in Figure 9? Explain briefly.

(e) [3] A second analysis of covariance is shown in Figure 11. Interpret the effects of education and year shown in the Estimate column of the `summary` table.

4. Grazing cattle can ingest larvae, which deprives the host animal of nutrients and weakens the immune system, affecting the growth of the animal. Each of 60 animals were randomly allocated to one of two treatments A and B, with 30 animals receiving each treatment. Each animal was weighed 11 times at (mostly) two-week intervals, and the weight was recorded. Some of the data, in dataframe `cattle`, are shown in Figure 12. The animals are identified by the letter of their treatment followed by the number of that animal within its treatment group.

   (a) [2] What feature of these data makes a repeated measures analysis appropriate? Explain briefly.

   (b) [2] How do you know that the dataframe shown in Figure 12 is laid out appropriately for drawing a graph such as a spaghetti plot or an interaction plot?

   (c) [2] Figure 13 shows an interaction plot. Assuming that the amount of variation about these means is small, what would you expect a repeated measures analysis to demonstrate? Explain briefly.

   (d) [1] A mixed model analysis is shown in Figure 14. Why did I use `factor(day)` rather than just `day`?

(e) [2] What do you conclude from Figure 14? Explain briefly.

(f) [2] The code needed to run a repeated measures analysis via MANOVA is shown in Figure 15. Why specifically was it necessary to begin with the `pivot_wider`?

(g) [4] The output from a repeated measures MANOVA is long. Some selected parts of the output, in the same order that they would be in the full output, are shown in Figure 16. What do you conclude from this output? Describe your process clearly (there are several steps to get right), and give the (rounded-off) numerical value of any P-values you use.

5. A number of measurements were taken on the skulls of kangaroos. These kangaroos were known to be of three different species. Some of the data is shown in Figure 17. The species have long names; in this question, you may refer to them by the first letter of their names, F, G, or M as appropriate.

   (a) [2] Some analysis is shown in Figure 18. What do you conclude from this analysis?

   (b) [2] Why is it reasonable to run a discriminant analysis here? Explain briefly.

   (c) [2] A discriminant analysis was run, with the results shown in Figure 19. Why are there two linear discriminants, rather than some other number?

   (d) [2] In Figure 19, would you expect the second linear discriminant to distinguish any of the species? Explain briefly.

(e) [2] Which two of the measurements are the *least* important in LD1? Explain briefly.

(f) [2] What combination of values of *two* of the variables would make the score on `LD2` *small* (that is, negative)? Explain briefly.

(g) [2] Some further analysis and a graph is shown in Figure 20. In the code above the graph, the `shape = species` draws points of different shapes for the different species (rather than drawing them all as circles). What does the graph tell you about how easy the species are to distinguish? I am looking for two distinct comments.

(h) [2] A table is shown in Figure 21. How does this table support your two conclusions of the previous part?

6. 15 sports fans were asked to rank seven sports according to which one they would most like to play (rated 7) down to which one they would least like to play (rated 1). The data are shown in Figure 22, in a dataframe called `ranks`.

   (a) [2] In preparation for a cluster analysis, a function to work out the dissimilarity between rows (individuals) `i` and `j` is shown in Figure 23. Why does it make sense for the dissimilarity to be based on the sum of squares of the differences between the ratings awarded to each sport by the two individuals? (Hint: what does it mean for individuals to be similar or dissimilar?)

   (b) [3] A cluster analysis is run as shown in Figure 24. The individuals are numbered 1 through 15, in the order shown in Figure 22. In the display of `ranks.1$merge` at the bottom of the output, what is the meaning of the two numbers in row 6 (displayed with `[6,]` on the left), and how does this appear on the dendrogram?

   (c) [2] Figure 25 shows the individuals (in column `id`), the cluster membership, and the rating given by each individual to each sport. Cluster 3 is shown as the middle cluster on the dendrogram. Why do you think these individuals ended up in the same cluster?

7. A consumer's organization collected information about 111 models of car. Eleven variables were measured, specifically: `Length` of entire car, `Wheelbase` (distance between front and rear wheels), `Width` of car, `Height` of car, `FrontHd` (the distance between a front-seat passenger's head and the car roof), `RearHd` (same, but for a rear-seat passenger), `FrtLegRoom` (leg room in the front), `RearSeating` (distance from back of front seats to back of rear seats), `FrtShld` (shoulder room in front), `RearShld` (shoulder room in rear), `Luggage` (luggage area). The variables are scaled to have a median and IQR that will make them resemble z-scores. Some of the data are shown in Figure 26.

   We will aim to distinguish the cars by clearly fewer variables than these eleven. The full dataframe is called `cars`; I also created a dataframe `cars_numeric` that has only the quantitative columns.

   (a) [2] Using the information in Figure 27, how would you justify looking at three principal components, rather than more or fewer? Explain briefly. (I am looking for two distinct points.)

(b) [2] The loadings on the first three principal components are shown in Figure 28. Does component 1 have a clear interpretation, or not? Explain briefly.

(c) [2] I ran a factor analysis and obtained the factor loadings shown in Figure 29. Which are the most important four variables in factor 1? Explain (very) briefly.

(d) [2] What do cars that score high on factor 1 have in common, in a few words? Explain briefly.

(e) [3] What do cars that score high on factor 3 have in common? Explain briefly.

(f) [2] Figure 30 shows the uniquenesses. Why does `FrtLegRoom` have a high uniqueness?

Use this page if you need more space. Be sure to label any answers here with the question and part they belong to.