University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 (K. Butler), Midterm Exam
March 7, 2025

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 10 numbered pages of questions including 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.

Houses in Canton, New York

Canton is a small town in upstate New York, USA (meaning, not near to New York City). 53 houses were sold there in one year. Some of the data are shown in Figure 2. The variables recorded were the selling `Price` in thousands of dollars, the number of bedrooms (`Beds`), the number of bathrooms (`Baths`), the floor area of the house in square feet (`Size`), and the `Lot` size in acres. We will be interested in predicting selling price from the floor area of the house and the number of bedrooms it has.

(1) (2 points) Scatterplots of `Price` against `Beds` and `Size` are shown in Figure 3. From this Figure, do you think that `Price` depends on either or both of `Beds` and `Size`? Explain briefly.

(2) (2 points) The realtors analyzing these data decided to predict selling price itself, and not, for example, the log of selling price, from the two explanatory variables. On the basis of Figure 4, why do you think they chose to do that?

(3) (2 points) Two regressions are shown in Figure 5 and Figure 6. Why is it that `Beds` is significant in the second one but not the first one?

(4) (2 points) Some predictions are shown in Figure 7, along with confidence limits. What precisely are `conf.low` and `conf.high` limits for? Explain briefly.

(5) (3 points) The second interval in Figure 7 is longer than the first one. Explain briefly why that makes sense. You may find the information in Figure 8 useful.

Turtles

The temperature at which turtle eggs are kept can, it is hypothesized, affect the chance that a turtle that hatches from those eggs turns out to be male or female. To assess this hypothesis, an experiment was run, in which the temperature was controlled at various different values. The data from the experiment are shown in Figure 9. The columns are, in the order shown, the temperature in degrees Celsius, the number of male turtles that hatched from the eggs at that temperature, and the number of female turtles that hatched.

(The experiment was in fact replicated at the same temperatures on three different days. This does not affect our analysis in any way.)

(6) (2 points) Why is logistic regression a sensible technique to use to assess the hypothesis of interest?

(7) (3 points) Is there one or more than one individual per row of the data in Figure 9? How can you tell? How does this show up in the analysis of Figure 10?

(8) (2 points) In Figure 10, how can you tell that the model is predicting the probability that a hatched turtle is *female*?

(9) (2 points) Interpret the *sign* (positive or negative) of the number in the `temp` row of the `Estimate` column in Figure 10.

(10) (2 points) Interpret the *value*, including its sign, of the number in the `temp` row of the `Estimate` column in Figure 10.

(11) (3 points) A plot is shown in Figure 11, along with the code that produced the plot. The researchers were interested in estimating the temperature at which 50% of the turtles would be female. What do you think that temperature is? Do you think that temperature has been estimated accurately or inaccurately? Explain briefly in both cases.

Marijuana use

The National Youth Survey collected data on marijuana use among young people aged 11 to 17. This survey was carried out every year from 1976 to 1980, and different young people were sampled each year. Each young person sampled was asked their sex (as they identified), and whether and how often they used marijuana (classified as "never", "once a month or less", "more than once a month"). The responses on marijuana use were abbreviated as `never`, `<1m`, and `>1m` respectively. The data for each individual were summarized into counts of how many individuals fell into each combination of sex, year, and use category (the column `n` contains the counts).

The data are shown in Figure 12, in dataframe `potuse`. There are 30 rows altogether, of which 15 randomly chosen rows are shown in the Figure. A summary is shown in Figure 13.

The survey organization was interested in whether there was a trend over time (`year` is treated as quantitative), and whether males and females used marijuana at a different level within any time trend.

(12) (2 points) A model is fit using the code in Figure 14. Why was it necessary to use `polr` (rather than, say, `glm`)?

(13) (2 points) How can you tell that the response categories are in a sensible order, based on anything you have seen about these data so far?

(14) (2 points) Some output is shown in Figure 15. What do you conclude from it?

(15) (4 points) Some predictions are shown in Figure 16. Describe the effects of both `year` and `sex`.

(16) (3 points) A plot is shown in Figure 17. In the `condition =` part of the code, what was the effect of entering those three variables in that order, and why was the order sensible?

Treatments for lung cancer

14 patients with lung cancer were randomly allocated to either a new treatment (`newdrug`) or the standard treatment (`control`). The researchers were interested in whether the patients receiving the new treatment lived for longer than the patients who received the standard one.

The data, in dataframe `lungcancer`, are shown in Figure 18. The columns are:

- `time`: time from diagnosis until last observation in days
- `cens`: whether the patient was alive (0) or dead (1) when last observed
- `group`: the treatment received.

(17) (2 points) In Figure 19, some of the output values have plus signs. Why is this?

(18) (2 points) What would have been another way to write the `Surv` code in Figure 19? Explain briefly why your alternative way would have worked.

(19) (3 points) A Cox proportional-hazards model is fitted, with output shown in Figure 20. According to this Figure, does the new treatment have (i) a significant effect, (ii) a *positive* effect on survival, compared to the standard one? Explain briefly.

(20) (2 points) A plot is shown in Figure 21. Explain briefly how this plot is consistent with your answer to (ii) of the previous question (or is not consistent, if that's what you think).

(21) (2 points) Another plot is shown in Figure 22. What do you conclude from this plot?

Shock

A psychologist designed a experiment to test the effect of electric shock on the number of attempts it took to successfully complete a (difficult) task. They compared three treatments: no shock, medium shock, severe shock. These are labelled respectively as `Group1` through `Group3` in column `group` in the dataset. 27 subjects were randomly assigned to one of the three treatments.

The psychologist wanted to know two things: (i) if there is an effect of any shock vs. no shock, and (ii) how medium shock compared to severe shock. For the response variable, `attempts`, a smaller value is better. The data, in dataframe `Shock`, are shown in Figure 23.

(22) (2 points) Why is it better to use contrasts to analyze these data than the standard one-way ANOVA followed by Tukey?

(23) (2 points) What R code will create contrasts `c_any` and `c_med_sev` that we will be able to use to test the comparisons of interest? Note that `Group1` through `Group3` are in that order.

(24) (2 points) Verify that your two contrasts are orthogonal. Show your calculation (that is, *not* R code that will do your calculation).

(25) (2 points) What R code will set it up so that running the ANOVA as a regression will test the two contrasts of interest? You may assume that `group` is a `factor`. This question is *not* asking about how to run the ANOVA as a regression; it is asking what you do *before* that, in order to make it work.

(26) (2 points) What do you conclude from Figure 24? Assume that all tests are two-sided. If you use a P-value to draw a conclusion, say which P-value you are using to draw that conclusion from.

Growth of pigs

In the data shown in Figure 25, fifty pigs were randomly allocated to one of five feed treatments, labelled `T1` through `T5` in column `treatment`. Each pig's weight was measured before the study (in `weight1`) and again after the study (in `weight2`); the column `gain` is the difference `weight2` minus `weight1` and reflects how much weight the pig gained over the course of the study. The column `feed` shows how much of their allocated feed the pig consumed during the study. We ignore the column `rep`.

Interest is in whether a pig's weight gain depends on the feed treatment they were on and the amount of feed that they consumed. The dataframe is called `crampton.pig`.

(27) (2 points) Figure 26 shows a graph of `gain`, `feed`, and `treatment`. In predicting `gain`, do you think there is a significant interaction between `feed` and `treatment`? Explain briefly.

(28) (2 points) An analysis is shown in Figure 27. What do you conclude from this Figure?

(29) (2 points) Why is it better to use Figure 27 to answer the previous question, rather than the output in Figure 28?

(30) (2 points) On the graph in Figure 26, what was the most important piece of evidence that you used in answering Question 27? How does this evidence show up in the output from `summary` for this model shown in Figure 28?

If you need any more space, use this page, labelling each answer with the question number it belongs to.