University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 / STA 1007 (K. Butler), Midterm Exam
March 11, 2022

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 19 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 (10 marks)

Leukemia is, according to the Mayo Clinic, "cancer of the body's blood-forming tissues, including the bone marrow and the lymphatic system". Like any cancer, a sign of a successful treatment is "remission", meaning that the symptoms of leukemia have reduced. Part of our data set is shown in Figure 2. For each patient, the variables recorded are:

- `remiss`: whether or not the patient shows remission (1 = yes, 0 = no)
- `cell`: cellularity of the marrow clot section
- `smear`: smear differential percentage of blasts
- `infil`: percentage of absolute marrow leukemia cell infiltrate
- `li`: percentage labeling index of the bone marrow leukemia cells
- `blast`: absolute number of blasts in the peripheral blood
- `temp`: highest temperature prior to start of treatment

I don't know what any of these mean, apart from the information given here. We want to see whether any of the other variables have an effect on remission.

(a) (2 marks) Explain briefly why logistic regression would be a suitable method to use to analyze these data.

My answer:
The response variable `remiss` is categorical with two categories, yes (1) and no (0).

(b) (2 marks) Two logistic regression models are shown in Figures 3 and 4. What precisely are these models predicting? Explain briefly.

My answer:
They are predicting something to do with a probability of remission, given the values on the other variables for a leukemia patient.

Specifically, in a logistic regression, we are predicting the probability of the second category (or, to be precise, the log-odds of the second category), with the first category being the baseline. In this case, the first category is 0 (the baseline), and the second category is 1. Thus we are predicting the probability (or log-odds) of a remission *occurring*, given the values of the other variables.

There is no value in discussing explanatory variables here, since both models are predicting the same thing. Say something about probability of remission (or not) for more than 1 point, and say how you know whether it's the probability of remission (as opposed to the probability of no remission) for both points.

The point of this part was to show that you know what feature of remission is being predicted (the probability that it happens). Saying just that we are predicting remission without further detail is only 1 point.

(c) (3 marks) Why was it necessary to do the test in Figure 5? What do you conclude from this test?

Question 1 continues...

My answer:

It was necessary to do the test because more than one (actually five) explanatory variables were removed from the model `leuk.1` to make `leuk.2`. It is thus necessary to test whether removing all of those at once was appropriate. (If we had only removed one, we could have used one of the tests in the `summary` of `leuk.1`). The key word was "necessary": why was there no way to do it other than this way?

The null hypothesis that the larger and smaller models fit equally well is not rejected (P-value 0.4827), and therefore the smaller simpler model with just `li` is preferred.

(d) (3 marks) Some predictions are shown in Figure 6. How are these predictions consistent with the output shown in the appropriate one of Figure 3 and Figure 4? Explain briefly.

My answer:

According to the predictions in Figure 6, as `li` increases, the probability of remission increases: that is to say, a higher percentage labelling index of the bone marrow leukemia cells is associated with remission being more likely. Specifically, the predicted probability of remission goes up from 0.09 when `li` is 0.5 to 0.64 when it is 1.5.

The predictions come from the second model, the one shown in Figure 4 (an additional clue being that values for the other explanatory variables were not supplied for the prediction, which would have failed if it had been based on model `leuk.1`). In Figure 4, the Estimate for `li` is 2.897, positive, meaning that as `li` increases, the probability of remission also increases, which is the same conclusion that we draw from the predictions.

A less good answer is to see that the P-value of 0.015 in Figure 4 is small, so that there is a significant effect of `li` on the probability of remission. This shows up in the predictions as a substantial *change* in the probability of remission. However, looking at the P-value only tells you that you would expect to see the predictions *change* as `li` increases; it doesn't tell you whether the change is upward or downward, and for that you need to look at the Estimate as well.

In summary, full marks for looking at the Estimate, or the Estimate and the P-value, and making the case. Only two marks if you only look at the P-value but otherwise correctly explain what is happening. (I am not usually keen on you mentioning two things when I only want one, but if you talk about the small P-value and the positive Estimate, I am OK with that; leading off with the small P-value indicating a "real" relationship, so that the positive Estimate is not just chance, actually is a strong answer.)

If you try to interpret the size of the Estimate, you will need to be careful: as `li` goes up by 1, the *log-odds of remission* goes up by 2.897 (we are not predicting remission itself, which is categorical, but the *probability* that it happens). Expect to lose something if you convey the impression that this is an ordinary linear regression. The easiest (satisfactory) way to approach this is to look only at whether the Estimate is positive or negative.

Question 2 (12 marks)

A random sample of adult residents of Alachua County, Florida, had their mental health assessed by a professional. Three variables were recorded for each person:

- `impairment`: the professional's overall assessment of mental health for each person, on a scale from Well (good), Mild, Moderate, Impaired (bad).
- `ses`: socio-economic status: high or low
- `life_events`: a scale reflecting the number and severity of important life events such as birth of child, new job, divorce, or death in family that occurred to the subject within the past three years. (A higher number means that the person has experienced more of these events, or the events they have experienced have been more severe.)

The entire data set is shown in Figure 7. (The data values are separated by single spaces.)

(a) (3 marks) The data was read into a dataframe `mh`. Some possible models were fitted in Figure 8. Explain briefly why the model labelled `mh.2` is more appropriate than *each* of the models labelled `mh.1` and `mh.3`.

My answer:

Model `mh.2` is more appropriate than model `mh.1` because the categories of `impairment` are *ordered* (from Well at the good end to Impaired at the bad end), and model `mh.1` is only suitable for a response with *unordered* categories. Two points.

Model `mh.2` is more appropriate than model `mh.3` because the response for a `polr` model must be a `factor`, which `fct_inorder` will create. Having the response variable be text will not even run. (It is enough to say that the response variable has to be a factor rather than the text it is here; the `fct_inorder` thing is coming up.) One point. Another way to approach this is that the levels are ordered, but they won't necessarily come out in the `right` order. This overlaps with the next part, and so wasn't what I was originally thinking of here, but it seems to be a reasonable answer to the question as posed, so that's also one point.

(b) (2 marks) In Figure 8, why did it make sense to use `fct_inorder` in defining the model `mh.2`? Explain briefly.

My answer:

The response variable in a `polr` model has to be a categorical variable (actually a `factor`) with the categories in a sensible order. Looking at the data in Figure 7, the categories in impairment are listed with Well first, then Mild, then Moderate, then Impaired: that is to say, best to worst. This is a sensible order. (Say why the ordering in the data is a sensible order.) Thus, using `fct_inorder` to create the response will take the categories in a sensible order, because it uses the order in the data.

Say, somehow, that `fct_inorder` uses the ordering in the data, and that this ordering makes sense (and how you know).

The reason for showing you all the data (instead of just some of it) in Figure 7 was precisely that I wanted to ask you this.

(c) (2 marks) Some more output is shown in Figure 9. What do you learn from this output? Explain briefly, in the context of the data.

Question 2 continues...

My answer:

This output shows that in the model `mh.2`, the socioeconomic status can be removed because it is not significant (P-value 0.064, greater than the usual α of 0.05). That is to say, whether a person's socioeconomic status is high or low does not affect their level of mental health impairment. Life events, though, is significant and needs to stay in the model.

Some sensible discussion of whether each explanatory variable is significant or not, or can be kept in the model or not, is called for here. Mentioning just one of them, particularly if that one is `ses`, is probably OK as well.

If you want to take the AIC angle, you can do that. This says that the AIC of dropping nothing is the best (lowest), so from that point of view you should keep both explanatory variables. (This often happens when you have an explanatory variable like `ses` here with a small P-value, but not quite small enough to be significant: AIC says to keep it, the P-value says to drop it.)

(d) (2 marks) Another model was fitted, as shown in Figure 10. Why was it necessary to run `drop1` again, even though the remaining explanatory variable was significant in Figure 9?

My answer:

Figure 9 only shows that `life_events` was significant in a model containing both it and `ses`. If we take `ses` out, as we did in Figure 10, `life_events` might have become non-significant (even though this seems unlikely), and so we should check to make sure that it is still significant and needs to be kept in our model (it does).

There are different ways to say this; for example you could say that the two explanatory variables might be correlated, so that the P-value of `life_events` in a model alone could be different from the P-value of the same variable in a model with `ses` as well. Anything that gets at these ideas is good. (The key word in the question was "necessary"; why did I *have to* run `drop1` again?)

(e) (3 marks) Some predictions are shown in Figure 11. Would you say that a person with a higher score on the life events scale is likely to have better or worse mental health overall than someone with a lower score? Explain briefly. Based on what you know about mental health, and what you have learned about the life events scale used in this data set, do you find this surprising? Explain (very) briefly.

My answer:

As the score on the life events scale increases, the probability of a person being Well on mental health decreases, and their probability of them having moderate impairment or being impaired increases. Thus, an increase on the life events scale goes with worse mental health overall. Two points.

A person who has more or more severe of these (bad or stressful) life events happen to them, as the life scale is defined in the question preamble, would be expected to be in a worse place in mental health terms (just from practical knowledge), so our finding is not at all surprising. One point. (Pretty much any sensible discussion here will work.)

For the first two points, I'm looking for an overall assessment of whether a person's mental health is likely to get better or worse as the life events score increases. So combine your assessment of the probability of being at the good end decreasing, and the probability of being at the bad end increasing, to come up with an overall "likely to be worse".

Question 3 (11 marks)

49 patients took part in a trial of a new treatment, called linoleic acid, for a particular form of colorectal cancer. Think of these patients as a random sample of all patients with this particular cancer. 25 of these patients were randomized to the new treatment, and the 24 received a control treatment (the current best treatment for this form of cancer). Some of the data are shown in Figure 12. For each patient, the experimenters recorded the treatment received, whether or not the patient died, and the length of time that the patient was observed, in months.

(a) (3 marks) Figure 13 shows some code to create a new column in the dataframe `cancer` as was read in from the spreadsheet. If you were to look at the first four values of the new column `y`, what would you see? Explain briefly why you would see that.

My answer:

Look again at Figure 12 to get the values. You would see the values 1+, 5+, 6, and 6 in the column `y`.

The event of interest here is death, so the third and fourth observations are "complete" or "uncensored" in that the patients in question did die and the number in the time column is a time until death. For the first and second observations, however, the patients did not die; these patients were "lost to followup" (as the expression is); they are in survival analysis jargon "censored", and will show in `y` as the time observed with a plus sign.

Make sure you demonstrate your knowledge of: the event of interest here; the difference between a censored and an uncensored observation; how you tell whether a particular observation is censored or uncensored; how a censored observation is displayed.

(If you have been reading PASIAS, you'll have seen that I often create the response variable *outside* of the dataframe. The column `y` here is actually a list-column, because it contains within itself both a lifetime and a censoring status, and didn't used to display very nicely, but sometime within the last couple of years someone made it display properly within a dataframe. Hence, I am now happy to leave it within the dataframe, so as to keep things all in one place. This was discussed in a tutorial.)

Some people seemed to think that this was *only* selecting the patients who died. It is not a `filter`. Indeed, the patients that didn't die contain some information, because we know they lived at least as long as they were observed for. A censored patient who was observed to live for a long time suggests that the treatment they were on was effective, because when they eventually do die (whenever that is) they will end up with a long lifetime after diagnosis.

(b) (2 marks) A Cox proportional hazards model is fitted, as shown in Figure 14. In this output, why does `treatment` display as it does? Explain briefly. You do not need to discuss any numeric values here (that comes later).

My answer:

Treatment is a categorical explanatory variable, and, as in any regression, there is an Estimate for every level except for the first "baseline" level, whose Estimate is zero. In this case, the treatments are `control` (first alphabetically) and `linoleic_acid`. Thus `control` is the baseline, and we see an Estimate only for the other treatment, `linoleic_acid`, relative to the baseline.

The facts that the Estimate for the linoleic acid treatment is negative and that its P-value is large are important, but not relevant here. (If you mention them here, I am not likely to penalize you, but if you read ahead, you can avoid wasting your time.)

(c) (2 marks) Estimated survival curves are shown in Figure 22 (at the end of the booklet of Figures), along with the calculations that led up to them. Interpret the plot. In particular, which treatment appears to be more successful? How do you know? Explain briefly.

My answer:

Looking at the dataframe `new` (of treatments to predict for), we see that Stratum 1 on the plot is Control and Stratum 2 is linoleic acid. On the plot, stratum 2 is above and to the right of stratum 1, which means that the (undesirable) event of death is less likely to happen sooner for that treatment. Thus, the linoleic acid treatment is more successful because it delays death.

The best answer recognizes that the event, death, is undesirable here and says something about how you know which treatment postpones death for longer.

Survival curves always drop over time (remember the saying "in the long run we are all dead") but the concern here is which one is higher at any given time than the other (the linoleic acid one).

(d) (2 marks) In Figure 14, which number supports the conclusion that you drew from the previous part? Explain briefly.

My answer:

The Estimate for the linoleic acid treatment, -0.249, is negative. This means that the hazard of event (death) is less for the linoleic acid treatment than for the control treatment: that is, a patient on the linoleic acid treatment is less likely to die sooner. This is the same conclusion that we drew from the survival curves.

The "hazard of death" thing is saying that if this is less, which it will be at any time for a patient on linoleic acid compared to one in the control group, the chance of that patient dying in the next small time interval is smaller. (If this sounds to you like a derivative in the sense of calculus, you would be right.) Everyone is going to die in the end, but the effectiveness of a treatment like this one is that it is more likely to make you live longer (you are less likely to die sooner). Some people like to look at the `exp(coef)` below, but the key point if you do that is to compare that with *1* rather than zero.

As in one of the earlier questions, this is better than looking at the P-value, because that actually says here that there is *no* significant difference between the treatments. We take that up in the next part.

(e) (2 marks) Based on what you see here, do you think that this conclusion would generalize to *all* patients with this type of colorectal cancer? Explain briefly.

My answer:

Asking whether you can generalize from the data you are looking at to a wider population requires the use of a hypothesis test, in this case one that compares the two treatments. This means asking whether the linoleic acid treatment is significantly different from (better than) the control. The P-value for this test comes from Figure 14, either the one on the end of the linoleic acid treatment line, or from the `drop1` table below. The P-value of 0.563 (0.5623) is not small, so there is no evidence of a significant difference between the two treatments, one that would generalize to the population of which these patients are a sample.

I added the second sentence to the question preamble so that you wouldn't be able to say that these patients are not a random sample of anything. Maybe they aren't, in fact, but the nature of this kind of study is that researchers get the patients they get, and have to act as if the patients they have are typical of all patients. (If it was clear that they were in fact *not* anything like a random sample, this would come up in the peer review of the paper based on this analysis.)

In other words, the patients on the new treatment did (a little) better in this study, but this was very much something that could have happened by chance if in fact there were no difference between the treatments. So it is far from being convincing evidence that the linoleic acid treatment is something to recommend for all patients with this form of cancer.

Some people said no, because the sample size was too small. But *this is exactly what the P-value is designed to assess*: was the kind of difference I observed likely to have happened by chance if there was no treatment effect? And *in the calculation of the P-value* the sample size plays a role. You can get a small P-value out of a sample size this small if the effect is large, or if the effect is small but very consistent (to exaggerate, if all the people in control lived for exactly 35 months after diagnosis, and all the people on linoleic acid lived for exactly 36 months after diagnosis). In the latter case, you then get into issues of statistically significant (yes) vs. practically important (maybe not).

Another direction to take is that there could be other variables (not measured here) that would make a difference, and that would affect the generalizability. But that would lead to a conclusion of "maybe not", and by looking at the P-value, you can make the stronger conclusion "no". This, or another sensible comment, can certainly net you one point.

Extra: the output from `summary(s)` looks like this:

```
Call: survfit(formula = cancer.1, newdata = new, data = cancer)

 time n.risk n.event survival1 survival2
    6     46       6     0.856     0.886
    8     40       2     0.808     0.847
   10     37       2     0.760     0.807
   12     34       6     0.613     0.683
   20     19       1     0.578     0.652
   24     16       2     0.500     0.583
   30      9       1     0.441     0.528
   32      7       1     0.372     0.462
   42      2       1     0.212     0.298
```

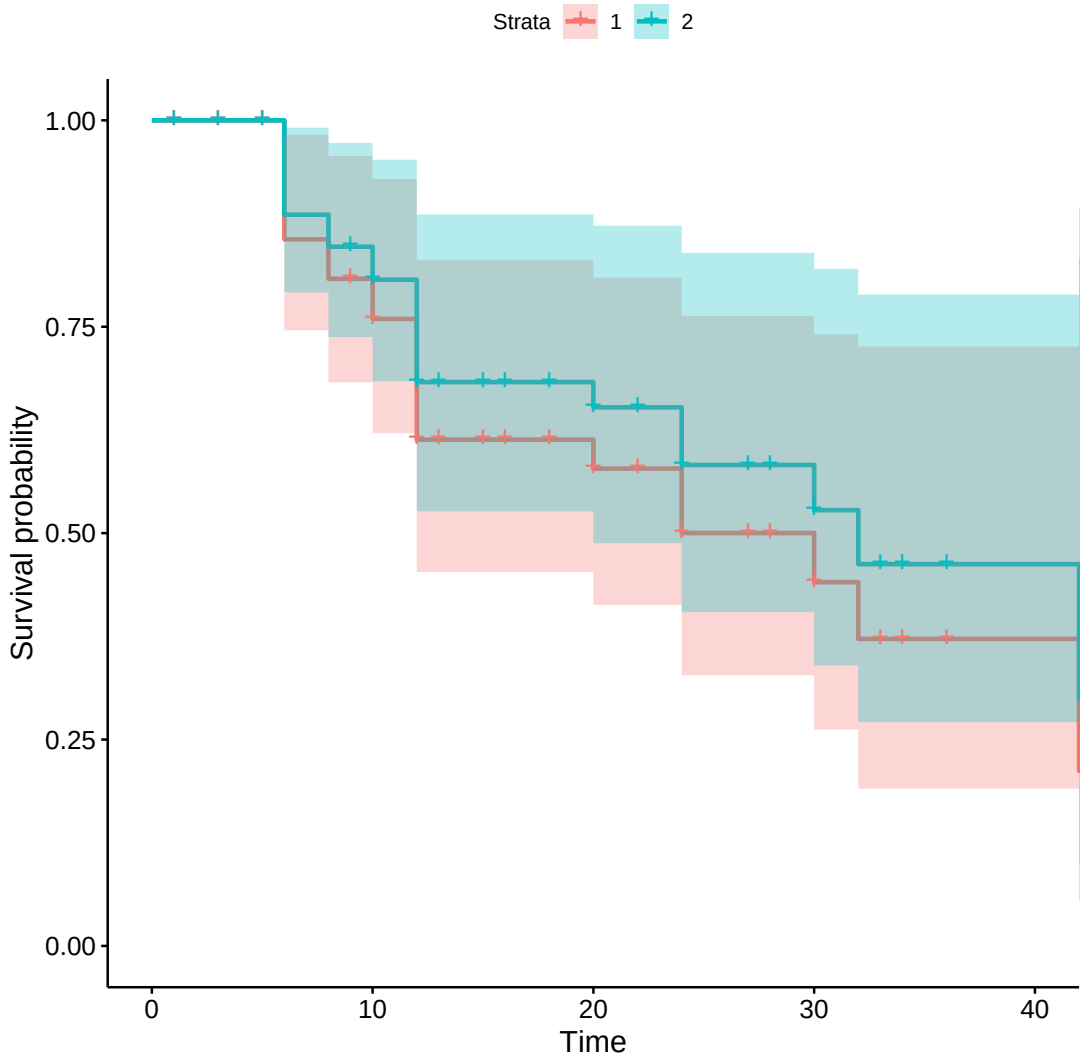
The numbers in the two `survival` columns are the estimated probabilities of surviving the number of months in the `time` column (or longer), for the two treatments control and linoleic acid respectively. You see that the probabilities in the linoleic acid column are near 0.08 bigger most of the way down (for example, the estimated probabilities of surviving 24 months are 0.500 and 0.583). This may strike you as a decently large effect size (wouldn't *you* like to have an extra 8 percentage points chance of living for a certain time?), but the fact remains that with these kinds of sample sizes, you would need to observe a much bigger effect to have any chance of it being significant. (If you felt that the survival curves were "close together", this is how that plays out.)

Perhaps now is a good time to look at those survival curves *with* the confidence intervals attached (remove the `conf.int = FALSE`):

```
cancer %>% count(treatment) -> new
new
```

```
## # A tibble: 2 x 2
##   treatment         n
##   <chr>         <int>
## 1 control          24
## 2 linoleic_acid    25
```

```
s <- survfit(cancer.1, newdata = new, data = cancer)
ggsurvplot(s)
```



This shows that the confidence intervals overlap by quite a bit, so that it was not surprising that the two treatments were not significantly different.

Question 4 (17 marks)

Three different treatments, labelled A, B, and C, are being investigated to see whether they have any effect on the growth of plants. The experimenters choose to assess plant growth in three different ways: the height, width, and weight of the plant. Fifteen plants were grown, five for each treatment. The data, in dataframe `plants`, are shown in Figure 15.

(a) (2 marks) What feature of this data set would make multivariate analysis of variance (MANOVA)

an appropriate method to use? Explain briefly.

My answer:
MANOVA is the method of choice when there is more than one quantitative response variable. Here, there are three: height, width, and weight.
(One point for knowing when MANOVA applies, and one for convincing me that you know which variables are responses in *this* data set.)

(b) (2 marks) In Figure 16, the MANOVA analysis is shown. It uses a variable `y` that I had to define. How did I define it (in code or in words)?

My answer:
One of these:

1. `y` needs to be all the response variables collected together into an R `matrix`
2. `y` is defined by this code: `y <- with(plants, cbind(Height, Width, Weight))` (or the equivalent with dollar signs).

You need only one of these.

(c) (2 marks) What do you conclude from the MANOVA output in Figure 16?

My answer:
The P-value is smaller than 0.05, so the treatment definitely has an effect on the combined response. More specifically, one or more of the treatments has a different mean for one or more of height, width, or weight.

More than that you cannot say.

"It is significant" or "reject H_0" is, as I think you know by now, not any kind of complete answer. If you say that without even trying to say what it means in the context of the data we have, expect no more than 0.5.

I saw some people trying to use the word "diagonally". This is what I said in the lecture notes because we had two response variables that we could draw on a graph and the difference actually was diagonally between the treatments (fertilizer levels in that case). It is not clear how that applies to (i) three response variables, (ii) whether the difference actually *is* diagonal or not, whatever that means, because we haven't done the discriminant analysis yet.

(d) (2 marks) What was the purpose of running the discriminant analysis in Figure 17? Explain briefly, in the context of the data.

My answer:
I wanted to understand which treatments had an effect on which of my response variables. That is to say, I want more information than the MANOVA gave me, and the discriminant analysis is a way to get it.

(Discriminant analysis plays a similar role in MANOVA to that played by Tukey in ANOVA.)

There are lots of different ways to say something relevant, and I was pretty relaxed about how you said it as long as it got at the point. For example, you could take the angle of predicting treatment based on height, width, and weight.

(e) (2 marks) Why is it that there are two linear discriminants, and why is it that I only need to consider the first one? Explain briefly.

My answer:
One mark for each:

There are two linear discriminants because there are 3 response variables and 3 treatments, and the smaller of (the first) 3 and (the second) $3-1$ is 2. (Get both of these, or you may lose a half point.)

The second linear discriminant can be ignored because its "proportion of trace" is almost zero, so that it contributes almost nothing to separating the treatments. LD1 always does more to separate the groups than LD2, but the point here is that it does *much* more, to the extent that LD2 is not worth considering at all. (If the proportions of trace had been even 0.7 and 0.3, LD2 would have had *something* to say, if maybe not very much.)

(f) (2 marks) Which of the response variables contribute the most to distinguishing the treatments? Explain briefly.

My answer:

Look at the Coefficients of Linear Discriminants in Figure 17, and in particular the LD1 column. The coefficients for Weight and Height are clearly positive, and the one for Width is close to zero. Hence Weight and Height are what distinguish the treatments.

The idea here is to name *all* the response variables that are far from zero, not just the one that is biggest in size. If you want to make the call that `Height`'s coefficient is close to zero, go ahead and say so, but you need to say so. It should be clear whether you consider each one of the response variables to be part of LD1 or not. (If I wanted you to name only one of the response variables, I would have asked "which one of the response variables"; the question as asked leaves open the possibility that there could be more than one.)

(g) (3 marks) A plant has small weight, small height and average width. Using the graph in Figure 18, which treatment do you think it received? Describe your thought process clearly enough so that your reader is convinced by your logic.

My answer:

From Figure 17, the coefficients of LD1 are positive for both weight and height (and close to zero for width). This means that such a plant will have a *negative* LD1 score (smaller than average makes a negative contribution on the variables that count), or at least a small one. (Scores average out to zero, so small is negative.) LD1 score is on the x-axis of Figure 18, so this plant will be on the left of the graph, in with the circles, so it received treatment A.

For yourself, you can confirm this answer from Figure 17 simply by looking at the Group Means: a small Weight and small Height points towards treatment A, and an average Width towards B, so the overall picture is that it's probably A. But I wanted to make sure you understood what the graph was saying, which means working via LD1 scores.

If I had given you a biplot, it would of course have made it much easier: weight would have pointed right and down, height would have pointed right and up, and width would have had a short arrow. To be low on weight and height implies being on the left somewhere, in with the treatment A plants.

(h) (2 marks) How does Figure 18 confirm what you said earlier about the relative importance of LD1 and LD2 for these data? Explain briefly.

My answer:

Earlier, we said that LD1 is much more important than LD2, meaning that LD1 does a much better job of distinguishing the treatments than LD2 does. On the graph, LD1 does indeed distinguish the treatments left-to-right, with A on the left, B in the middle and C on the right. LD2, on the other hand, does nothing to distinguish any treatments; having a high or low score on LD2 does not distinguish any of the treatments from others.

Extra: the graph uses shapes rather than colours because of our print shop. They always give me grief if I try to put anything coloured on an exam, so I found another way to distinguish the treatments, using `shape` on the plot. I also found that the default size for the points came out rather small (and made them harder to distinguish), so I made them bigger with `size = 2`. That reminds me, the survival curve graph has colours. I think they'll make me put it at the end.

(Edit to add: this time, the print shop seemed happy to print the entire Booklet of Figures in colour, so I don't think I needed to worry, but I have definitely had grief from them in the past.)

Question 5 (10 marks)

Investigators at the University of North Carolina Dental School were interested in the growth of children's skulls. They measured 27 children, 11 female and 16 male (as the children identified themselves). Each child was measured at ages 8, 10, 12, and 14 years. The quantity measured was the distance (in millimetres) between the centre of the pituitary to the pterygo-maxillary fissure. This distance usually increases with age, but because both of the two points can move, the distance occasionally decreases with age. The quantity measured is known as "the distance" for the rest of the question; you do not

need to know any more about what it is.

Some of the dataset is shown in Figure 19. There are six columns: the number code of each child, the gender of the child (labelled `sex`), and the distance as measured at each age, labelled `d` followed by the age (as two digits).

(a) (2 marks) Figure 20 shows the mean distance for each gender and age. In the code for the graph, why was the `pivot_longer` necessary before drawing the graph? Explain briefly.

My answer:

The original data in Figure 20 is "wide" and the graph needs the data to be "long". That's about a minimum for one point.

For the second point, you need to show that you know what this means for the data we have here. Specifically, the graph needs all the distances in one column, and a second column saying which age each distance is for. The original data has the distances in four different columns, one for each age, not what is needed for a graph. The `pivot_longer` converts from the original format to what we need.

(b) (2 marks) What about this dataset makes a repeated measures ANOVA a suitable method of analysis? Explain briefly.

My answer:

Each individual child has not just one but four measurements, one for each of the four ages. Said differently, the same children are measured at four different ages rather than just once. That's enough for the points, but you could also say that some children are just bigger than others, so the four measurements are likely to be correlated (some children are high all the way through, and some always low, compared to the others).

The key thing here is to be specific about what's happening in this dataset. If you want to copy something from my notes to guide yourself, go ahead, but you need to say something about distances measured on children at four different ages to get two points.

Extra: for the analyses we have seen earlier to be appropriate, each measurement would have to be on a *different* child, requiring $4 \times 27 = 108$ children altogether.

(c) (2 marks) The analysis is shown in Figure 21. What do you conclude about the interaction between gender and time? Is this consistent with the graph in Figure 20? Explain briefly.

My answer: Remember the steps: check for sphericity first, then look at either the univariate or Huynh-Feldt adjusted P-values as appropriate (and then, if you like, compare with the multivariate one).

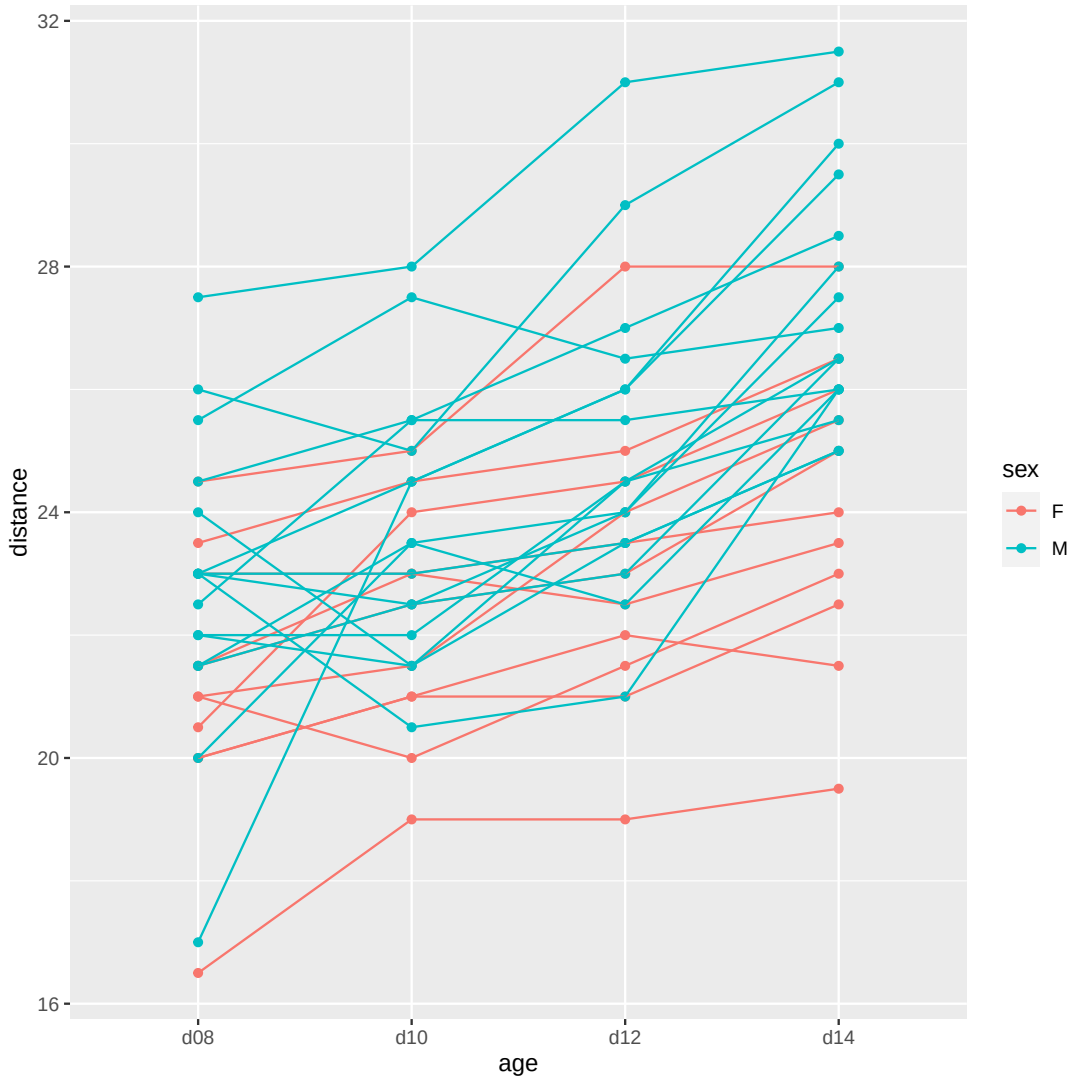
The sphericity test is significant, P-value 0.02. The right P-value is the Huynh-Feldt adjusted one at the bottom; the scientific notation in the P-value translates to 0.053, so it is not quite significant. Note that this P-value is only a little bigger than the one in the univariate table (0.041) and, indeed, close to the one in the multivariate table, 0.046, but has flipped just the other side of significance.

This means that the effect of time on the distance is not different for males and females. This shows up in the graph by the two traces being close to parallel: for the age 14 children, the males are bigger than the females by more than at other ages in this data set, but this difference is not big enough to be significant.

If you want to say that the lines are not parallel and therefore that the test and interaction plot are inconsistent, be my guest. I care more about the quality of discussion than the precise conclusions you draw. Bear in mind, though, that the traces can be a bit non-parallel just by chance (when there is actually no interaction).

Extra: this is how a spaghetti plot looks:

```
growth %>%
  pivot_longer(starts_with("d"), names_to = "age", values_to = "distance") %>%
  ggplot(aes(x = age, y = distance, colour = sex, group = sub)) +
    geom_point() + geom_line()
```



Maybe some of the blue traces (males) go up more steeply between ages 12 and 14 than the red traces (females) do between the same ages. There is a tiny bit of evidence here, but not very much, and evidently not quite convincing enough to make the interaction significant. (At other ages, the trends are mostly parallel.)

Extra 2: I have discovered that the complete `summary` output is typically too long to fit on one page of the booklet of Figures, so my usual procedure for an exam is to abbreviate the (very long) multivariate output, and then call for the univariate tests, sphericity tests and adjusted P-values in that order. The whole `summary` output is much longer, because of the length of the multivariate part, but, after the long multivariate part, it contains the same things in the same

order as you see here.

(d) (2 marks) From Figure 21, what do you conclude about the effect of time (age)? Is this consistent with the graph in Figure 20? Explain briefly.

My answer:

The sphericity test for age is rejected as well, so again go to the Huynh-Feldt table to get the right P-value, 8.3×10^{-15}. This is still very small, so there is clearly an effect of age (time). This shows up on the graph in that both traces are going clearly uphill: the overall trend of mean distances is increasing with age, for both males and females.

Extra: on the spaghetti plot above, pretty much all of the trends are going uphill too.

(e) (2 marks) From Figure 21, what do you conclude about the effect of gender? Is this consistent with the graph in Figure 20? Explain briefly.

My answer:

The P-value for gender is 0.010, also small and significant. Because gender is the same for all times — it plays the role of a treatment here — you don't test this for sphericity first, and the P-value is the same whether you look at the univariate or the multivariate table.

This is shown on the graph by the trace for males being consistently above the trace for females: that is, males, at any age, have a larger mean distance than females do.

If you are concerned about looking at main effects when the interaction is (almost) significant in these last parts, you can rationalize doing so in a couple of ways. One is that any interaction effect must be small, and that the main effects show up pretty clearly on the interaction plot (the males are consistently above the females, both male and female trends go up with age). So you could think of the main effects as being over and above the (small) interaction effect, noting that this is a repeated measures, so that you cannot (this way) remove terms with time in them. A second way to think about this is to imagine what would happen with simple effects: there would be a strong effect of age for both genders separately, and there would be an effect of gender for each of the four time points separately. It is, however, fair to say that we should not be looking at main effects when the interaction is significant, and I have tried to make sure that you have not been short-changed marks-wise if that's what you concluded.

Extra: on the spaghetti plot above, most of the male (blue) traces are above most of the female (red) ones. Though it's not terribly clear, the overall trend seems to be that males have larger distances at any age than females do, at least on average. (This mixed-upness of individuals is probably why the P-value is small but not very small.)

The overall pattern is clearer on the interaction plot, which is why I gave you that to answer these questions with rather than the spaghetti plot, but if you were doing a real analysis, you would probably want to look at both. My goal with this question is to see whether you knew what it was about the data that was driving the significant effects that you saw, and I wanted to make that part of it as straightforward for you as I could.

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