University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 / STA 1007 (K. Butler), Final Exam
April 16, 2016

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 22 numbered pages of questions. Please check to see that you have all the pages.

In addition, you should have an additional booklet of code and output to refer to during the exam. Contact an invigilator if you do not have this. References to numbered Figures in this exam refer to Figures in that booklet. The captions with Figure numbers are *underneath* each Figure.

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	4	
2	7	
3	4	
4	7	
5	5	
7	8	
9	8	
10	9	
11	11	
13	7	
15	12	
17	10	
18	6	
19	6	
20	4	
21	8	
Total:	116	

1. "Inattentional blindness" is a phenomenon where a person who is concentrating on one task fails to see something that is right in front of them. The classic example of this is a video showing students passing a basketball around, and your task is to count how many passes the players in white shirts made. See, for example, `http://www.theinvisiblegorilla.com/videos.html` (check it out after the exam). During the video, a person wearing a gorilla suit walks across the screen, and the question is, "did you notice anything else while you were counting the passes?". If you didn't notice the gorilla, you have inattentional blindness.

   There is a psychological test called the Stroop Colour Word test. This comes in three parts, each to be done as fast as possible:

   - `W`: word alone, or reading a list of colour names.
   - `C`: colour alone, or naming the colours in which a series of X's is printed.
   - `CW`: colour and word, or naming the colour in which a *different* colour name is printed.

   Each individual gets a score on each of the three parts. Is any of this related to whether or not a person saw the gorilla on the video? Some data were collected from 49 subjects, and these are shown in Figure 1 of the booklet of code and output. The variable `seen` is 1 or 0 according to whether the subject saw, or did not see, the gorilla when they watched the video. We aim to predict whether or not a subject saw the gorilla from the test scores.

   (a) (2 marks) Why is logistic regression a suitable procedure for what we are trying to do?

   My answer: The response variable, "saw the gorilla", is a categorical yes/no variable. (And therefore we will be predicting the probability that it will be "yes" from the values on the Stroop test scores. See below.)

   (b) (2 marks) What probability is the logistic regression in Figure 2 modelling? Explain briefly.

   My answer: The two levels of `seen` are 0 and 1. The first is used as a baseline, and the probability modelled is of the second one, ie. the probability that the subject *saw* the gorilla.

(c) (2 marks) What do you conclude from the `anova` at the bottom of Figure 2? Note that a model with a `1` on the right side of the squiggle has no explanatory variables, just an intercept. Explain briefly, in the context of the data.

> **My answer:** There is no significant difference between the model with no explanatory variables and the one with all three, and we should prefer the one with none: that is, we should get rid of all three!
>
> Another way to say this is that *none* of the test scores have any impact on whether the subject will see the gorilla.

(d) (2 marks) How is the output from `summary(gorilla.1)` consistent with what you've just said?

> **My answer:** This says that none of the three explanatory variables are significant, and that any one of them could be removed. This is the kind of thing we'd expect to see if none of the explanatory variables had any predictive value (which we just concluded).

(e) (2 marks) Figure 3 shows two boxplots side by side. What do these boxplots tell you? Is the message from the boxplots similar to what you already found? Explain briefly.

> **My answer:** The boxplots say that the median scores on the CW test (which was the one where you had to see a colour name and then say the colour in which the word was printed, without saying the colour name that was printed) are almost exactly the same for people who saw the gorilla and people who did not.
>
> This is exactly the same thing as we already saw: the test score is the same whether the subject saw the gorilla or not. If the CW test had any impact, the mean score would be higher, or at least different, for gorilla-seers than for people who missed the gorilla. I observe that the CW test is the most demanding of the three, and requires you to think about what colour the word is printed in, while ignoring what the word actually says. (I know from personal experience that this is very hard to do, and requires considerable concentration.) But, it has nothing to do with inattentional blindness.

(f) (1 mark) What would you recommend that the researcher do next?

> **My answer:** Give up on finding any association between these tests and inattentional blindness! Or find another test to test for association with inattentional blindness. Or something like that. There seems to be nothing to be gained from any further analysis of these data.

2. I have a collection of variables measured on a collection of 1978-model cars. The variables I have are these:

   - `make`: the name (model) of the car
   - `price`: the manufacturer's recommended sale price for the car
   - `mpg`: gas mileage (in miles per US gallon, so larger is better)
   - `rep78` (response variable): repair record in 1978 (categorized; takes the values 1–5 with a higher value being better). Five car models had a missing value for this variable; these models were excluded from the analysis.
   - `headroom`: distance from a car's roof to the bottom of the seat (measured in feet).
   - `trunk`: total volume in cubic feet of the car's trunk space (or other cargo-carrying space).
   - `weight`: total weight, in pounds.
   - `length`: front to back, in inches.
   - `turn`: the diameter of the smallest U-turn the car can make, in feet (known as the "turning circle radius").
   - `displacement`: total volume of the cylinders in the car's engine, in (I think) cubic centimetres. The larger this is, the more powerful the car.
   - `gear_ratio`: the larger this is, the faster the engine spins (in RPM) for a given size of tires.

   We want to use (some or all of) these variables to predict the repair record. Some of the data set is shown in Figure 4.

   (a) (2 marks) For my modelling, I used `polr` from package `MASS`. Explain briefly why that was a good choice.

   My answer: The response variable `rep78` is categorical *and ordered.* The numbers don't mean anything as numbers, but 5 is better than 4 and so on. (You should justify in some way your assertion that the response variable is ordered.)

   (b) (2 marks) I fitted the two models shown in Figure 5. What do you conclude from the `anova` statement at the bottom of that Figure? Which model should we use to do predictions with? Explain briefly.

   My answer: The P-value is not small, so the two models fit equally well, and we should prefer the smaller one, `autos.2` (since `autos.1` must contain some explanatory variables that are not important).

(c) (1 mark) In Figure 6, I do some predictions, for representative values of the explanatory variables, based on model `autos.2`, which we assume is an appropriate model to do predictions from (this may or may not agree with your answer to the previous part). What did I achieve by using `expand.grid` in Figure 6?

> **My answer:** To get all possible combinations of the variable values I used.
>
> There were only two values for two of my variables and only one for the rest, so there were in the end only four combinations. I did this to stop things getting too out of hand for you below.

(d) (2 marks) Describe the effect of car price on the repair record. As price increases, does the repair record improve? Explain briefly.

> **My answer:** We have to let car price increase while holding everything else fixed, for example by looking at the first two rows of predictions. In this case, the probability of being in the first four categories decreases (slightly) while the probability of being in the last category increases (substantially). Having a higher probability of being in a higher-numbered repair record category is better, so increasing the car price improves the repair record.
>
> Looking at the third and fourth rows gives a similar (but different) picture. This time the probability of being in category 2 goes down and category 4 goes up. But it is similarly an improvement in the repair record.

(e) (2 marks) Describe the effect of turning circle radius on the repair record. Is an increase in the turning circle radius associated with an improvement in the repair record, or not? Explain briefly.

> **My answer:** Same idea: let turning circle increase, while holding everything else fixed. So compare something like the first and third rows of predictions. The effect is actually rather more dramatic than for price (rather to my surprise): as `turn` increases, the probability of being in the first three categories increases sharply, and for the last two categories decreases sharply. So a vehicle with a larger turning circle has a worse repair record.
>
> Comparison of the second and fourth predictions tells exactly the same story.

(f) (2 marks) Does the output of Figure 6 say anything about the effect of gas mileage? Explain briefly why or why not.

> **My answer:** No, because there are no rows with different values for `mpg`. All the rows in the predictions have `mpg` of 20. Also, when I'm constructing `new`, I only have that one value of `mpg` in `mpgs`: to assess the effect of `mpg`, I'd have to have two (or more) values here.

3. A study of breast cancer was done in Germany. A total of 720 patients with "primary node positive breast cancer" were included in the study. Patients were recruited into the study between July 1984 and December 1989. A number of variables were recorded for each patient:

   - `id`: patient's ID in study
   - `diagdate`: date of diagnosis of breast cancer
   - `recdate`: date of recurrence (the cancer came back) or of "recurrence-free survival" (the cancer went away) after treatment.
   - `deathdate`: date of death
   - `age`: age at diagnosis
   - `menopause`: patient has reached menopause, 1=yes, 2=no
   - `hormone`: patient had hormone therapy, 1=yes, 2=no
   - `size`: tumour size (mm)
   - `grade`: tumour grade (1–3)
   - `nodes`: number of nodes involved
   - `prog_recp`: number of progesterone receptors
   - `extrg_recp`: number of estrogen receptors
   - `rectime`: time to recurrence or survival (days)
   - `censrec`: whether cancer came back (1) or did not (0)
   - `survtime`: time to death (days)
   - `censdead`: whether died (1) or survived (0)

   Our aim in this question is to predict the time to recurrence of the breast cancer, and to see how that depends on any of the other variables. The data are summarized in Figure 7.

   (a) (2 marks) Figure 8 shows some analysis of these data. Given our aims here, explain briefly how the `Surv` statement uses the right variables in the right way.

   > **My answer:** Our aim is to predict recurrence time, so we need to use `rectime` rather than (say) `survtime`, since we are not interested in predicting time to death. The second part of `Surv` is the censoring variable: this is correctly `censrec`, since that is the variable that indicates this; the reason for the `==1` is that `censrec` is 1 if the cancer came back (the event happened).

   (b) (3 marks) I did some preliminary analysis (not shown) to determine which explanatory variables to keep. My final model is the one shown in Figure 8. Consider each of the three explanatory variables in turn. Based on the information in Figure 8, what effect does each variable have on the breast cancer coming back? Be specific about the nature of the effect in each case. Assume that $\alpha = 0.10$ for this question.

   > **My answer:** The slope describes the effect of increasing the value of that variable on the *hazard of the event occurring.* Thus a positive slope means that increasing that variable makes the event happen *quicker.* In this case, the event is an undesirable thing (recurrence of breast cancer).
   > Specifically:
   >
   > - The slope of `hormone` is negative. Increasing the value of `hormone` means *not* having the hormone therapy, so not having the hormone therapy decreases the hazard of recurrence. Or, in plainer English, having the hormone therapy is associated with the breast cancer coming back *sooner.*

- The `size` of the tumour has a positive slope, so if the tumour is larger, the hazard of recurrence increases: that is, if the tumour is larger, the breast cancer is likelier to come back *sooner*. (This is exactly what you would guess. But are you guessing it for the right reason?)
- The slope of `nodes` is positive, so this is just like the last part: if there are more nodes, the breast cancer is more likely to come back sooner. This too is what you would guess.

(c) (2 marks) What is the code in Figure 9 doing? There are two things to comment on: first, the purpose of the first four lines of code (these together achieve *one* thing, that you need to describe) and second, specifically what `pp` contains (that you would see if you looked at a `summary` of `pp`).

> **My answer:** The overall picture is that we are obtaining predicted survival curves (survival probabilities at a number of times).
>
> The first *three* lines are some values of the three variables in our survival model; the `expand.grid` line makes all possible combinations of these variables (there are $2^3 = 8$ of them).
>
> The last line actually obtains the predictions. So `pp` contains predicted survival probabilities (in rows, if you look at `summary(pp)`) under the various combinations of explanatory variables (in columns, so there are 8 columns of predictions).

(d) (2 marks) On the plot in Figure 11, which of the values of the explanatory variables are associated with the "best" survival, that is, having the best chance of the breast cancer taking the longest to return? Explain briefly. The variables in `combo` are listed in the same order as the variables in `new` in Figure 9. In this part and the next, if you cannot distinguish the colours on the plot, ask.

> **My answer:** The best survival curve (the one for which the chance of the event is lowest for longest) is as ever the top-right one. This is the pale blue one, which goes with the combination `2-20-1`. This is: no hormone therapy, tumour size 20, nodes 1. (I would like you to translate what `hormone=2` means.)

(e) (4 marks) For each of the three explanatory variables in the proportional-hazards model, explain briefly how their effect on the survival plot is the same as the effect you described back in part (b). Or explain how it is inconsistent, if that is what you see.

> **My answer:** The cleanest way of demonstrating the effect of an explanatory variable is to change it, while leaving the other variables fixed. To demonstrate a "positive" effect on recurrence (that is, that the breast cancer is likely to take longer to come back), show that the survival curve moves up and to the right:
>
> - For `hormone`, changing 1 to 2 (eg. comparing the pale blue curve to the red one) improves the chances of recurrence taking longer. This means that things are better with *no* hormone therapy, as we said before.
> - For `size`, changing 20 to 35 (eg. comparing the red curve to the olive green one) makes things *worse*, so having a tumour of larger size is likely to make recurrence happen *sooner*. This is also as we said before.
> - Last, for `nodes`, changing 1 to 7 also moves the survival curve down (compare eg. the red and orange curves), so having a tumour with a larger number of nodes is likely to make the breast cancer come back sooner. This, too, is what we said back in part (b).
>
> I don't mind which pairs of survival curves you compare, as long as they demonstrate what you need to demonstrate. For example, for `nodes`, you could also compare the purple survival curve with the pink one, or you could compare the two green ones.
>
> What is not so good, though it is worth something, is to do something like comparing *all* the curves with `hormone=1` with all the curves that have `hormone=2`. It is true that the hormone-2 curves are higher up the page, but this is confounding the effects of `hormone` with those of the other variables (for example, if `size` had a very big effect, it would be hard to discern the effect of `hormone` this way). The all-else-equal idea is a much cleaner way to determine the kind of effect an explanatory variable has.

I am looking for two things here: a reasonable comparison strategy (that you explain clearly), and the results of that comparison on the three variables.

4. Three new textbooks are being tried out for (university) students learning intermediate French. The textbooks are labelled `a`, `b` and `c`. 15 students take part in a study, with five being randomly assigned to each of the new textbooks. These students have all previously passed an introductory French course. Also recorded is each student's (cumulative) grade point average, `gpa` in the data set. At the end of the study, each student takes a French test and the score is recorded. The test covers what the students are supposed to have learned from the textbook they used.

   The data set is shown in Figure 12.

   (a) (2 marks) An analysis of variance is shown in Figure 13. I checked that the assumptions for this analysis of variance were satisfied. What do you conclude from Figure 13, in the context of the data?

   My answer: The P-value of 0.489 is by no means smaller than any reasonable α like 0.05, so we cannot reject the null hypothesis that the mean test scores are equal for all three textbooks. That is, the three textbooks are equally good in terms of (average) test scores.

   (b) (2 marks) Figure 14 shows a scatterplot of test score against GPA, with the observations labelled by which textbook the student used. Do you think there is an effect of GPA on test score? Explain briefly.

   My answer: There is an upward trend of test score with GPA, so I would expect to see a (strong) effect of GPA.

   (c) (2 marks) In Figure 14, do you think there is a difference in test scores among the textbooks? Explain briefly.

   My answer: On the plot, *for any given GPA*, the blue dots (book C) are at the top and the red ones (book A) are at the bottom. This means that textbook C is consistently the best (once you allow for GPA) and textbook A is consistently the worst (ditto). Because this picture is so consistent, I *would* expect to see an effect of textbook.

   This is inconsistent with the analysis of variance in part (a), which I will ask you about in a moment.

   (d) (2 marks) What do you conclude from the analysis `french.2` in Figure 15? What does that mean for the data? Explain briefly.

   My answer: There is no (significant) interaction between `gpa` and `book`.

   That is, one of these:

   - The way in which test score depends on `gpa` is the same for each book
   - The way in which test score depends on `book` is the same for all `gpa` value.
   - The lines relating `test` and `gpa` for each book are parallel.

   Any one of those. But if you start talking about main effects here, you risk losing a point, because that discussion starts only *after* you've gotten rid of the non-significant interaction.

(e) (2 marks) Explain what the `update` statement in Figure 15 is doing, and why it is a sensible thing to do.

> **My answer:** It is fitting a new model that is taking the `gpa` by `book` interaction out of the model `french.2`. It is a sensible thing to do because the interaction was not significant, and therefore it should be removed. (The second mark was an easy one if you got the first mark.)

(f) (2 marks) What do you learn from the first part of the output of `french.3` (the `anova`)? There are two things to note.

> **My answer:** Both P-values are extremely small, so there is an effect of `gpa` on test score, and there is also an effect of `book` on test score.
>
> I didn't need any further explanation here.

(g) (3 marks) Compare your conclusions in parts (a) and (f). Are they consistent or inconsistent? Why did they come out that way? Explain briefly.

> **My answer:** Mine are inconsistent: earlier, I said that `book` had no effect, but now I am saying that `book` has a very strong effect.
>
> As to why this happened: we also found (just now) that `gpa` had a strong effect, and the analysis of covariance that we just did actually says something slightly different: *after you allow for a student's GPA*, their scores on the test are different for the different books. You can also see this same point from the scatterplot in Figure 14.
>
> Another way of saying this is that `gpa` and `book` are both strongly significant, so that both of them should be in your model, and so having only one of them (the `aov` is predicting test score from `book` only) is a mistake.
>
> Or you can look at the graph in Figure 14, and see that when you don't consider `gpa`, the scores for each `book` are all over the place, and so there is no hope of finding any significant differences among the textbooks this way.
>
> There are several ways to come to a conclusion, but what you should end up saying is that the analysis of covariance is trustworthy, or the analysis of variance is unreliable (not making the best use of the data), with something to justify your point of view.

(h) (2 marks) Look at the coefficients of `bookb` and `bookc` in the last part of Figure 15 (the `summary` of model `french.3`). What do those numbers mean, in the context of the data?

> **My answer:** There is no longer any interaction, so the slopes of the lines for the regressions of `test` on `gpa` are the same for each `book`. The coefficients of `bookb` and `bookc` express the changes in *intercept* for these two books compared to book A, or to put it another way, for any `gpa`, book B adds about 6 points to the average test score compared to book A, and book C adds about 12.5 points to the average test score, compared to book A.
>
> This is another way of saying that book A is worst and book C is best, after you allow for `gpa`, the same story that we get from the scatterplot in Figure 14.
>
> I tried adding those three regression lines to the plot in Figure 14, and they really are very parallel. This is partly because the trends for each group are so unambiguous.
>
> Interpretation of the coefficients of the model in an analysis of covariance is really a lot easier if the interaction is removed. If the interaction has to stay, the interaction's coefficients reflect the changes in *slope* for the books, compared to the slope for book `a`. But that's harder to get your head around, and I didn't want to put you through that.

5. Ten people with low self-esteem took part in a study that compared a standard therapy method (labelled `baseline` in the data) with a new therapy (labelled `new`). Five people were randomly allocated to each therapy. Each person underwent the appropriate therapy for four weeks, and a self-esteem score was assessed for each patient at the end of each week. The data are shown in Figure 16.

   (a) (2 marks) What does it mean to say that these data are "repeated measures"? Why would a two-way analysis of variance (based on therapy and weeks) not be appropriate? (That is, what would a two-way analysis assume, that is actually not true?)

   > **My answer:** There is more than one measurement on each person (one per week for the four weeks). A two-way analysis of variance would assume that we had *one* measurement for each person, taken at one of those four weeks (and thus that there were $4 \times 10 = 40$ people altogether), which is not what we have.

   (b) (3 marks) What do you conclude from the analysis in Figure 17? (There are three things to conclude.)

   > **My answer:**
   >
   > - The interaction is not significant (P-value 0.59), so the pattern of treatment effects is the same over all times (or, the effect of time is the same for both treatments; either way around is good.)
   > - There is *no* difference between the two therapies (P-value 0.35); they are equally effective as far as self-esteem goes.
   > - There *is* an effect of time (P-value 0.00002): self-esteem levels during therapy vary over time.
   >
   > These can be in any order, as long as you get all three of them somehow. The intercept should not be tested, since we are not interested in whether that's zero or not.
   >
   > In a repeated measures, there's no way to take out the non-significant interaction, so we make all of our conclusions at once rather than taking out the interaction and then saying something about the main effects, as we would in a regular ANOVA.

   (c) (2 marks) Figure 18 shows the code to do some processing of the data. Describe what the output data frame `selfesteem.long` looks like. (I'm looking for a description in words.)

   > **My answer:** The name is a clue: it's turning the data from wide format, as used in the `Manova`, to long format, as required for the spaghetti plot. Specifically, it's turning the four columns of self-esteem scores, one per week, into *one* column of self-esteem scores, with an additional column `week` labelling the week that each score came from.
   >
   > If you want to give a few rows of what the long data frame looks like *as part of your explanation*, that's OK, but simply listing the whole thing (assuming you have time to write it down) doesn't show that you understand what `gather` does *in general*.

   (d) (1 mark) Figure 19 shows a spaghetti plot of the observations for each person against time. Which part of the code (above the plot) produces a *separate* line for each person?

   > **My answer:** The piece `group=subject` inside the `aes`.

   (e) (3 marks) Describe how the spaghetti plot supports *each of the three* of your conclusions from part (b).

My answer: I'll do it in the same order as I did in part (b):

- The pattern over time (usually up and then down again) is the same for the red and blue lines). This supports the lack of interaction between therapy and time.
- The red and blue lines are all mixed up (there is no tendency, for example, for the red lines to be at the top and the blue ones at the bottom). This supports the lack of therapy effect.
- The lines for each subject go up from week 1 to week 3, and then down again in week 4. This seems to be consistent over subjects (though the *size* of the changes varies). This consistency is more than enough to obtain a significant time effect.

6. A market research company commissioned a survey of families in order to understand what makes a family choose to subscribe to one magazine rather than another. We will investigate some data from families that subscribe to exactly one of four magazines (listed below). In the survey, a lot of demographic information was collected, as described below:

   - `id` of family (ignore in analysis)
   - `magazine` subscribed to, a number indicating which magazine:
     1. Better Homes and Gardens
     2. Reader's Digest
     3. TV Guide
     4. Newsweek
   - `i1` through `i4`: ignore
   - `famsize`: number of people in family, with 6 indicating "6 or more".
   - `income`: family income, coded with 1 meaning "less than \$10,000" up to 11 meaning "\$75,000 and above".
   - `race`: 1 is "white", 0 is "other".
   - `tv`: number of televisions (3 is "3 or more").
   - `newspaper`: whether family subscribes to a newspaper (1=yes, 0=no)
   - `nomale`: whether family has a father living in the house (1=no father, 0=father)
   - `nofemale`: whether family has a mother living in the house (1=no mother, 0=mother)
   - `child18`: whether there are any children over 18 living in the house (1=yes, 0=no)
   - `headage`: age group of head of household, from 1 (18–29) to 6 (65 and above)
   - `headeduc`: education level reached by head of household, from 1, "some grade school" to 7, "went to graduate school".

   The structure of the data is shown in Figure 20. Some of these variables can take only the values 1 and 0, but that is not a problem for the analysis that is done here.

   (a) (2 marks) Why could discriminant analysis help the market research company with their aims here?

   My answer: The idea of discriminant analysis is to understand how groups (here, magazines) differ according to some other measured variables, of which we have a large number here. This will help the market research company determine what kinds of household tend to purchase which kind of magazines.

   (b) (3 marks) Look at Figure 21. Which two variables does the first linear discriminant mainly depend upon? When will the first discriminant score be large? Explain briefly.

   My answer: Look at the "coefficients of linear discriminants" and look down the LD1 column for values far from zero. The two that stand out for me are `child18`, −0.75, and `nofemale`, −0.58.

   That is to say, `LD1` will be large principally when there are *no* children over 18 in the house (`child18` is small), and/or when there *is* a mother at home (when `nofemale` is small). Small because of the negative signs. This takes a bit of decoding, and I expect you to show that you have done it.

   (c) (2 marks) Figure 22 shows a plot of the first two discriminant scores. What does this plot suggest about the ability of the demographic variables to distinguish the magazines that a family might subscribe to? Explain briefly.

My answer: The numbers (representing the magazines) are all mixed up, basically all over the plot, so the demographic variables are doing a poor job of separating the magazines. If you look carefully, there is a little something happening: the 1s are mostly at the bottom, the 2s are mostly on the right, the 3s really are all over the place and the 4s are mostly on the left.

So you can say either that there is no separation of the groups (that you justify), or a little separation (that you describe). Either is good.

(d) (3 marks) Explain briefly what the table in Figure 23 is telling you. In particular, what does the number 16 in the table tell you?

> **My answer:** This table is saying how many families who actually subscribe to the magazine indicated are predicted (by their values on the demographic variables) to subscribe to each magazine. The 16 says that 16 of the families that actually subscribed to magazine 3 (TV Guide) were predicted to subscribe to magazine 2 (Reader's Digest).

(e) (2 marks) If the discriminant analysis were doing a better job of explaining magazine readership, how would the table in Figure 23 be different? Explain briefly.

> **My answer:** More of the observations (families) would be on the top-left to bottom-right diagonal and fewer of them would be elsewhere in the table. This is because observations on the diagonal were correctly predicted and observations elsewhere were wrongly predicted.

(f) (2 marks) Are the table in Figure 23 and the plot in Figure 22 telling a consistent or an inconsistent story? Explain briefly.

> **My answer:** The plot and the table are both saying that magazine readership is unpredictable, and thus they are telling a consistent story. The plot says this by, for example, readers of magazine 1 being all over the place, and the table says this by there being a lot of misclassifications (values off the top-left, bottom-right diagonal). (There is a certain amount of overlap here with the last part, which is intended to give you a clue for this part.)
>
> Having said that, the table is actually not so bad, given how awful the plot looks: a decent fraction of subscribers to each magazine is predicted correctly, the best being magazine 2, Reader's Digest, where the proportion correctly predicted is
>
> ```
> 32/(8+32+7+2)
> ## [1] 0.6530612
> ```
>
> This is a considerable improvement over the 0.25 that you would get by guessing. It is possible that the third LD (which is not shown in the plot) is indeed adding something to the prediction of which magazine a family is subscribed to.
>
> So you can alternatively say something like "the plot is awful, but the table is not completely awful, and therefore they are inconsistent". As long as the assessment of plot and table are reasonable, and the inference of consistency or not is sound, I'm happy.

(g) (3 marks) Figure 24 shows the predicted group membership and posterior probabilities for a sample of families. Note that the output is too wide to fit on the paper, so continues below as a second table with the same row names. For the family with `id` 132, what is the name of the magazine they actually subscribed to? What is the name of the magazine they are predicted to subscribe to?

> **My answer:** Magazine 1 (in the `magazine` column), which is Better Homes and Gardens.
>
> Family 132 is in the third line of each part of the output, the one with row name 32 (it was the 32nd line of the original data frame). The column `guess` has the predicted magazine for them, which is magazine 2 or Reader's Digest.

(h) (2 marks) Would you say the prediction for family 132 in Figure 24 was clear-cut or not? Explain briefly.

> **My answer:** To see whether the prediction was clear-cut or not, look at the posterior probabilities. These are in the last four columns (numbered 1 through 4), as you can tell from the

`cbind` above the table. For the family with `id=132`, the posterior probabilities for magazines 1 and 2 are 0.287 and 0.293 respectively, which are very close; the prediction was of magazine 2 because that posterior probability was slightly higher: it was a very close call, not clear-cut at all.

In fact, none of the posterior probabilities are small, so on the evidence of the demographic variables, this family really could have subscribed to *any one* of the magazines and it wouldn't have been at all surprising.

7. How do people perceive the similarity of different kinds of car? In one study, several people were asked to assess the similarity of the following cars, all (at the time) priced between $30,000 and $35,000: BMW 328i, Ford Explorer, Infiniti, Jeep Grand Cherokee, Lexus, Chrysler Town and Country, Mercedes, Saab 900, Porsche Boxster, Volvo. Each person was given a list of all 45 possible pairs of these cars, and was asked to rank the pairs from most similar (1) to least similar (45). The results for one person are shown in Figure 25. Some of the car names have been abbreviated; if it is not clear which car they are, ask.

   (a) (2 marks) Why are all those missing values in the data in Figure 25 not going to be a problem?

   My answer: The dissimilarity between cars A and B is the same as the dissimilarity between cars B and A, so we only need to know one of the pair (the other one will be the same). For example, Ford-BMW being 34 means that BMW-Ford must also be 34.

   (b) (2 marks) A multidimensional scaling analysis is carried out in Figure 26. Why is `isoMDS` a better idea for these data than `cmdscale`?

   My answer: These data are only ordinal (they are ranks) and don't represent anything real. So it is enough to produce a map that gets the dissimilarities in (approximately) the right order. Non-metric scaling does this, and that is accomplished by `isoMDS`. `cmdscale` would have done a *metric* scaling, which we just said is not justified.

   (c) (2 marks) Will the non-metric multidimensional scaling produce a map that does a good job of reproducing the dissimilarities? Explain briefly, giving a number from one of the Figures that helps you decide.

   My answer: The relevant number is the stress, at the bottom of Figure 26. Here, that is 3.99 (percent). On the scale given in the notes, that is an "excellent" fit, being smaller than 5%, so the map does not just a good job, but an excellent one.

   (d) (1 mark) Why did I need to use `xlim` in the code above Figure 27? Explain (very) briefly.

   My answer: To get the name "Porsche" on the map. Otherwise it would have disappeared off the right-hand side. (`Porsche` has the largest x-coordinate, about 22, which would be about as far right as the map would otherwise go.)

   (e) (3 marks) Look at the map in Figure 27. The cars overwriting each other are Ford and Jeep top left, and Infiniti and Lexus at the bottom. Find the Chrysler Town and Country and Porsche Boxster on the map. Are they close together or far apart on the map? Give *two* reasons why that is what you would have expected to see, given the other Figures.

   My answer: They are on the left, and at the top right, respectively, and so are far apart on the map.

   If you go back to Figure 25, you'll see that the actual dissimilarity between these two cars is 45, which is the largest of all. So we'd expect to see these cars a long way apart on the map. That's the first reason.

   The second reason is that the stress is very small, so that actual dissimilarities and map distances should be very consistent. Cars far apart on the map should be consistently far apart in actuality, including the two cars we looked at.

8. I collected some statistics on the major-league baseball teams in 2015. Each team played 162 games. The data are shown in Figure 28. The variables have abbreviated names, which stand for these:

   - `TEAM`: team's name
   - `RS`: total runs scored by the team
   - `H`: total hits made by the team
   - `HR`: total home runs hit by the team
   - `AVG`: team's overall batting average (number of hits divided by number of total number of "at-bats"). Higher is better.
   - `W`: wins by the team (number of losses not included since it is 162 minus wins)
   - `ERA`: "earned run average": a measure of how good the team's pitchers were, a lower number being better
   - `RA`: total runs allowed by the team
   - `SO`: strikeouts made by the team's pitchers (more is better)
   - `BB`: "bases on balls" or "walks" allowed by the team's pitchers (lower is better)
   - `E`: errors made by the team's fielders (lower is evidently better).

   (a) (2 marks) I ran a principal components analysis in Figure 29. In the `princomp`, why did I use `baseball[,-1]` instead of just `baseball`?

   > **My answer:** The first column has team names in it (and not numbers like all the others), so I wanted to omit this column when I ran the analysis. The stuff inside the square brackets means "use all the rows but omit column 1".

   (b) (2 marks) Why was it a sensible idea, considering the nature of the data, to use `cor=T` in Figure 29? Explain briefly.

   > **My answer:** This will use the correlation matrix (or, implicitly, standardize all the variables). This makes sense because they are all measured on different scales (or are typically different sizes), and using the correlation matrix gives each variable equal importance.

   (c) (2 marks) Use the scree plot in Figure 30 and the output in Figure 29 to identify an appropriate number of components to use. Justify your choice briefly.

   > **My answer:** There are elbows on the scree plot at 3, 5 and possibly 7. That would mean 2, 4 or 6 components. There are 10 variables, so I think 6 is too many. Two components explain 66% of the variability, but four explain 89%, a big improvement, so I think 4 components is best.
   >
   > A judgement call; if you can justify two components, for example by saying that 4 is too many for 10 variables, that's good too.

(d) (2 marks) I decided to try a factor analysis. I ran the factor analysis with 4 factors (this may or may not have been indicated by your choice above), with the results shown in Figure 31. Are four factors enough, or should I have used more? Explain briefly.

My answer: This is the P-value in PVAL. It's not at all small, so 4 factors are not rejected (that is to say, they're enough).

(e) (4 marks) Baseball fans might divide the variables up into hitting-related ones (RS, H, HR, AVG), pitching-related ones (ERA, SO, BB, RA), fielding-related ones (E), and one that relates to overall team quality (W). Discuss how well these correspond to the high loadings on our four factors.

My answer: There are different ways to handle this: you could start with what loads heavily on the four factors, and see how they correspond to the "departments" of the game, or you could start with hitting, pitching, fielding and overall quality, and see which factors they seem to belong to. I'll tackle it both ways, but picking one way is good.

Starting from the factors:

- Factor 1 is wins and strikeouts (negatively), earned run average, runs against and bases on balls (positively). Note that a good team will come out very *negative* on factor 1. These are all pitching variables, except for the number of wins. (This says that wins are more associated with pitching than with hitting.)
- Factor 2 is principally hits and batting average. You can include runs scored if you like, but then you ought to include strikeouts as well. The last of those doesn't really belong, since the others are all hitting variables, but a good hitting team will be high on factor 2 (otherwise). According to my source, the variable SO really is the number of strikeouts earned by the team's pitchers, but its presence here would make a lot of sense if it were the number of strikeouts suffered by the team's hitters (so that low would be good). However, this is not how it is.
- Factor 3 is runs scored and home runs. This is also hitting, but you could say that this one is "big hitting" rather than factor 2, which is more about "getting the ball in play": the variables there are big for *any* hit, even one that only gets you to first base. (Or, if you didn't say that, you could say that hitting is split into two parts, one part in factor 2 and the other in factor 3.)
- Factor 4 is fielding errors. This is the only fielding variable, so it's clear that this factor is fielding.

Or you can take it the other way around and see where the variables of the different types ended up factor-wise:

- Of the hitting variables RS, H, HR, AVG, two of them, H and AVG, ended up in factor 2 and the others were in factor 3. So we have two hitting factors. As discussed above, the distinction seems to be getting-the-ball-in-play hitting and big hitting.
- The pitching variables were ERA, SO, BB, RA. These all appear in factor 1.
- The one fielding variable is E. This is on its own in factor 4.
- The overall team quality variable W is in with the pitching variables in factor 1. That suggests, as above, that winning is associated principally with pitching.

When you've gone through this, I'd like to see an overall assessment of how well the parts of the game correspond to the factors. I would say that the correspondence is good, but I would note that hitting has gotten split over two factors and wins have gone in with the pitching variables. Also, if you included strikeouts `SO`, note that, as it is defined, this doesn't belong with hitting variables. I think these are the most important oddities to note.

(f) (2 marks) Are there any variables that are not included on one of my four factors? What is the best way to tell? Explain briefly.

My answer: The uniquenesses at the top of Figure 31. These are all acceptably small, nothing big like 0.9 (not by any means).

You can check this by reading across the loadings table and seeing that each variable (row) has a big loading on at least one factor (column). For example, `AVG` appears strongly in factor 2 and nowhere else, while `SO` is kind of split between factor 1 and factor 2, but has non-small loadings on both. Most of the variables have a large loading on exactly one of the factors (and the ones that have a *very* large loading have also a *very* small uniqueness).

(g) (2 marks) Figure 32 shows a biplot of the factor analysis. The Toronto Blue Jays are team #1 on the plot. Which variables would you expect to be especially high or low for the 2015 Blue Jays? Explain briefly, basing your assessment on what you see on the biplot (not on what you remember from last summer!)

My answer: The Blue Jays are near the top of the plot, a little to the left. The `RS` arrow points towards them, so their runs scored ought to be high. Also, the `AVG` and `H` arrows point up, so we'd expect them to have a high team batting average and a high number of hits. That's what I expect to see.

If you look back at the data in Figure 28, you'll see that the Jays scored easily the most runs of any team (`RS`). Their batting average `AVG` was almost the highest, very close with Kansas City (the eventual World Series winners) and Detroit. The team's hits `H` was the fifth highest in the league, so definitely nearer the top than the bottom.

None of the pitching variables were unusual for the Blue Jays, so their success was built on their hitting. Compare with team 24, the St Louis Cardinals: they had a lot of wins, and *low* values on the pitching variables `BB, ERA, RA` (which is good), so their success was built on their pitching. This also checks out from the data.

9. A study is investigating (categorical) variables associated with the birth of a child. These variables are assessed for each of 738 childbirths:

   - the gender of the child (male or female)
   - premature rupturing of membranes, called `rupture` (yes or no). (Rupturing of the membranes normally happens when the mother goes into labour, when it is known as "breaking the water". Here, it is noted whether rupture happens earlier than that.)
   - whether labour was induced (yes or no)
   - whether the child was born by cesarean section (yes or no)

   These variables are listed in the chronological order that makes sense. The data for this study are listed in Figure 33. The ultimate aim is to discover which variables, if any, are associated with the need for a mother to have a Cesarean section.

   (a) (2 marks) In Figure 34, only the start and finish of my analysis is shown. Why did I stop at model `births.10`?

   My answer: Everything that could be dropped is significant (at 0.05 level), so there is nothing that I can reasonably remove.

   (b) (2 marks) Why did I choose to obtain the particular results that I obtained in Figure 35, and not others? Explain briefly.

   My answer: These were the three interactions that were significant in the log-linear model, and I want to understand *why* they are significant: that is, what kind of associations there are between the variables in question.

   (c) (4 marks) Suppose we are looking for factors that are related to the mother having a Cesarean section (`cesarean` in the data). Use Figures 34 and 35 as necessary to describe which other factors are related to the mother having a Cesarean section, and to describe how those factors are related.

   My answer: In the context of a log-linear model, relationships with the "outcome" factor show up as *interactions* with that factor. The model `births.10` at the end of Figure 34 shows that `cesarean` is associated with `rupture` and `induced` but not with `gender`. As to how? Well, that comes from the tables in Figure 35.

   I showed you the tables of proportions from `prop.table` rather than the original subtables of frequencies (because the proportions are easier to interpret, done the right way).

   Since we are only interested in relationships with `cesarean`, the second table (the one from `xt2`) is not relevant to us, and any comment on it here is a mistake.

   The first table shows the relationship between `rupture` and `cesarean`. It says that if a woman has suffered a premature rupture of the membranes, having a cesarean is slightly *less* common (11% of the time vs. 18% of the time). This is not a big difference, but is evidently big enough, given the number of births that was observed, to be significant. Note that the "if-then" was deliberately set up this way around, because it's the rows in this table that add up to 1.

   The third table (from `xt3`) shows the relationship between having a Cesarian section and whether or not labour was induced. (The latter means that the mother was given a drug to help labour start.) Here, noting that the conditioning is now on the *columns*, if labour was induced, then it is *more* likely that the mother will have a Cesarean section, 23% vs. 14%. (You might argue that if labour is induced, it is more likely to be a problematic birth for other reasons.) Again, the if-then has to be this way around, because it is the *columns* that add up to 1.

Neither effect is huge, but there is enough data to declare it significant (and also, we would imagine that there would be enough power to declare any `gender` associations significant if they actually were).