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

Aids allowed (on paper, no computers):

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

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

The University of Toronto's Code of Behaviour on Academic Matters applies to all University of Toronto Scarborough students. The Code prohibits all forms of academic dishonesty including, but not limited to, cheating, plagiarism, and the use of unauthorized aids. Students violating the Code may be subject to penalties up to and including suspension or expulsion from the University.

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

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

   My answer:

   The response variable `low` is categorical, and to predict a categorical response we need to run a logistic regression.

   Strictly, it matters that `low` has only two categories (so that the use of `glm` later is correct), but my major concern here is that you know when to run a logistic regression of any kind.

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

   My answer:

   The categorical response `low` has two levels, `yes` and `no`. The first alphabetically, `no`, is the baseline, and the model predicts the probability of the second level alphabetically, `yes`.

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

   My answer:

   No, because they are both significant at $\alpha = 0.05$, with P-values 0.029 and 0.037 respectively.

   An easy two points.

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

   My answer:

   The estimate for `lwt` is -0.0133, so the log-odds that the woman's baby is of low birth weight decreases by 0.0133. (In this model, it is actually *better* for the mother to be heavier.)

   One point for mentioning log-odds, one for saying what happens to it. If you think that something else is decreasing by that much, your maximum is 1 out of 2.

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

My answer:

The first set of predictions (from the second code chunk in the Figure) show the effect of mother's weight when smoking is held fixed. As the mother's weight increases, the probability of a low birthweight baby decreases.

The second set of predictions (from the code chunk at the bottom of the Figure) show that if the mother smokes, the probability of a low birth weight baby is (substantially) higher, holding weight constant.

(This is why mothers-to-be are advised to quit smoking before getting pregnant.)

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

My answer:

- The predictions say that a larger mother's weight goes with a smaller probability of low birth weight, which is consistent with the negative Estimate for `lwt` (a larger weight goes with a smaller log-odds and hence a smaller probability)
- According to the predictions, a mother that smokes is more likely to have a low birth weight baby. This is consistent with the positive Estimate for `smokeyes`, which says that the probability of low birth weight is higher for a woman that smokes compared with a woman who does not (baseline).

Extra: I keep reading `smokeyes` as "smokey eyes", but it's actually `smoke` with `yes` on the end.

This was meant to be a straightforward warmup question.

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

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

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

   My answer:

   Test the interaction between display panel and emergency type. With a P-value of 0.567, this is nowhere near significant, and thus the effect of display panel is the same for each emergency type (and vice versa).

We should remove the non-significant interaction and fit a model without it (which is the next part, but if you say that here, it's also good).

It is an error to discuss the main effects at this point, because we need to re-fit the model without the interaction to assess their significance.

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

My answer:

The interaction in `display.1` was not significant, so we need to remove it and re-fit the model without it. (One point.)

The two remaining main effects are both strongly significant, with P-values of 1.8×10^{-5} for `panel` and 8.3×10^{-10} for `emergenc`. There are therefore some differences in time among display panels and among emergency situations. This is as far as we can go for now.

Extra: note that by taking the interaction out, the P-values for the two main effects have decreased, so that they are even more significant than they were before. This is not so much because the F-statistics have changed greatly, but more because we have gained six df for error, which is making similar F-statistics a lot more significant.

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

My answer:

It is appropriate to run the Tukey because we have two significant main effects that we want to understand better:

- all the emergency situations took a different mean time to respond to (all six P-values are less than 0.05, if only just in one case)
- display 3 is significantly different (worse) than the other two displays, which do not differ significantly in mean time.

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

My answer:

Display panel 3 was associated with the largest mean time (from the Tukey), since the time was longest on average (longer than either of the other two panels). There was no significant difference between panels 1 and 2, so we can recommend either of these panels.

Make sure your answer (i) eliminates panel 3 and (ii) says that we would recommend either of the other two because there is no significant difference between them.

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

My answer:

It would be a mistake to run a simple effects analysis here because the interaction between display panel and emergency situation is not significant. We only run a simple effects analysis when the interaction is significant. One point.

If we had run the analysis described, we would expect the effect of display panel to be the same for each emergency situation, because there is, according to the analysis, *an* effect of display panel that applies for all emergency situations. That is, display panel 3 would come out worst every time, and panels 1 and 2 would be not significantly different every time.

Extra: that actually isn't quite how it comes out:

Emergency situation 1:

```
display %>% filter(emergenc == 1) %>%
  aov(time ~ panel, data = .) -> mod
summary(mod)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
panel        2  89.33   44.67   9.926 0.0476 *
Residuals    3  13.50    4.50
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(mod)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = time ~ panel, data = .)

$panel
    diff         lwr       upr     p adj
2-1   -2 -10.8644949  6.864495 0.6549468
3-1    7  -1.8644949 15.864495 0.0901656
3-2    9   0.1355051 17.864495 0.0480695
```

Panels 1 and 3 are not quite significantly different.

Emergency situation 2:

```
display %>% filter(emergenc == 2) %>%
  aov(time ~ panel, data = .) -> mod
summary(mod)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
panel        2   64.0   32.00   17.45 0.0223 *
Residuals    3    5.5    1.83
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(mod)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = time ~ panel, data = .)

$panel
    diff       lwr       upr     p adj
2-1   -4 -9.658072  1.658072 0.1167813
3-1    4 -1.658072  9.658072 0.1167813
3-2    8  2.341928 13.658072 0.0196715
```

3 vs 1 is some way from being significant.

Emergency situation 3:

```
display %>% filter(emergenc == 3) %>%
  aov(time ~ panel, data = .) -> mod
summary(mod)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
panel        2   52.0   26.00    1.88  0.296
Residuals    3   41.5   13.83
```

```
TukeyHSD(mod)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = time ~ panel, data = .)

$panel
    diff        lwr      upr     p adj
2-1    2 -13.542152 17.54215 0.8595269
3-1    7  -8.542152 22.54215 0.2875938
3-2    5 -10.542152 20.54215 0.4668071
```

No significant differences anywhere.

Emergency situation 4:

```
display %>% filter(emergenc == 4) %>%
  aov(time ~ panel, data = .) -> mod
summary(mod)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
panel        2  56.33   28.17   9.389 0.0511 .
Residuals    3   9.00    3.00
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(mod)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = time ~ panel, data = .)
```

```
$panel
    diff         lwr      upr     p adj
2-1 -4.0 -11.2378298  3.23783 0.1975218
3-1  3.5  -3.7378298 10.73783 0.2541129
3-2  7.5   0.2621702 14.73783 0.0455779
```

Depending on how you read this: no significant differences, or 3 vs 2 is *just* significant.

What I think has happened here is that we started off with 24 observations, which gave us a decent amount of power to find differences. But when we did the simple effects, there were 4 of them, so each one was based on only 6 observations, 2 per display panel. This does not give us a lot of evidence to demonstrate that display panels are different, which is why we saw less significance in the simple effects than we did in `display.2`.

This is just in case you were curious.

The answer I wanted was the one I gave above, showing an understanding that the effect of panel applies across all emergency situations, so that the panels in the simple effects would be expected to show the same significance as in `display.2`, and with more data, they would. We could predict the results of the simple effects here (not altogether accurately, as it turned out) because the interaction was not significant; if the interaction had been significant, the results of the simple effects analyses would have been different from each other, because there was no longer "an" effect of panel. In that case, certain panels would have been better for only some of the emergency situations, and other panels better for others, in the same way that on the midterm question, it was better to have wrinkled fingers if the object was wet, but if the object was dry, it didn't make any difference.

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

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

   Some of the data are shown in Figure 8.

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

My answer:

Both the years of education and vocabulary test scores are recorded as whole numbers, so there is a limited number of places on the graph to plot points. For example, a lot of respondents had 12 years of education (ie. graduated high school), and a lot of those will have the same vocabulary score. Hence, a number of points on the scatterplot will need to be plotted in one place. (One point for getting to that, or something like it.)

To spread the points out so that we can see them all, we use `geom_jitter` to move each point

slightly away from where it would be plotted by `geom_point`, and then all 100 points will be visible. (The other point, showing that you know the difference between `geom_jitter` and `geom_point`).

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

My answer:

All three lines show an upward trend: the more education, the greater the vocabulary score. This makes sense because someone who has more education has probably done more reading and is therefore likely to know more words. (I'm willing to be flexible here, as long as you can come up with some kind of plausible link between education and how many words a person knows, or their chances of happening to know the words that are on the vocabulary test.)

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

My answer:

Because `year` is being treated as categorical (the `factor(year)`), we need to use `drop1` to test anything with `year` in it. This is obligatory here because there are three different years; we want to know whether we can remove the interaction *as a whole*, and the two tests on the last two lines of the Coefficients only compare each year (precisely, the slopes of the lines for each year) with the baseline year 1974.

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

My answer:

In the `drop1` at the bottom of Figure 10, the P-value is 0.319, which is not even close to being significant. Hence there is no significant interaction. (One point.)

This tells us that the *slopes* of the lines in Figure 9, even though they look a bit different, are actually not significantly different from each other. (The other two points).

My suspicion is that this is because there is a lot of scatter about the lines (there are a lot of points not very close to their lines), and so the slopes are not estimated very accurately (as you see from the wide envelopes around them, which I left in as a clue).

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

My answer:

From the `summary` output, the slope for education is 0.294. This is an ordinary regression slope, so an additional year of education is associated with an increase of 0.294 points on the vocabulary test. This holds for any year. One point.

The two year Estimates say that vocabulary scores in 1994 were predicted to be about 0.14 less than in 1974, and in 2014 were predicted to be about 0.15 less than in 1974, holding education constant. The other two points. These two effects are not significant (as well as being very small), but I wanted you to show me that you knew what they meant.

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

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

   My answer:

   Each animal gives us more than one measurement: specifically, 11 measurements, because that same animal was measured 11 times.

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

   My answer:

   The data are laid out in "longer" format, with all the weights in one column, labelled in column `day` with the time at which they were taken, and multiple rows for each animal. (The contrast is with the standard MANOVA layout, where each animal is in *one* row and there is one column for the weight on each day.)

   Say something about why longer format makes it easier to draw a graph, and what precisely we have here that will make it work.

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

   My answer:

   The traces over time are not quite parallel, so we would expect to see an interaction between weight and time (day). That's all. Remember that if there is an interaction, you stop there.

   I added the piece about "small amount of variation" because otherwise you could say that the two traces are more or less parallel and that you would expect to see no interaction (and, in that case, a main effect of time but probably not any effect of treatment). If you say that there is likely no interaction and no treatment effect, but a time effect, you haven't used the "small amount of variation" part, but your interpretation is otherwise reasonable, so that is 1.5. Note: If you say there is no interaction, you are entitled to say something about likely treatment and time effects: (i) there is a time effect (upward trend) for both treatments, and (ii) the two treatments are similar for all times.

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

My answer:

Day was recorded as a number (the number of days since the start of the study), but I wanted to treat it as categorical. This is the usual reason for needing to use `factor`. (Question 2, in the second paragraph of the description of the data, gives you a hint for this one.)

`lmer` is like `lm` in that if you have an explanatory variable that is recorded as a number, it will be treated as quantitative, so that without the `factor`, these models will fit a linear trend in `day` for each treatment (like an ANCOVA model, but one with a random effect of animal). As you saw from the interaction plot, the trend is not linear: the animals grow more slowly at the beginning and end, and you might guess that by the last observation, at 133 days, they are fully grown and will not gain any more weight.

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

My answer:

The top analysis, `cattle.1`, says that there is a significant interaction (P-value 0.00012) between treatment and day: that is, the pattern of weight over time is different for the two treatments. Give me a P-value so that I know you're looking at the correct one.

That's where you stop: you cannot (yet) say anything about *why* the interaction is significant, and because the interaction is significant, you should not look at the analysis of main effects in `cattle.1a` at all. (A student looked at `cattle.1a` first and said there was no effect of treatment *by itself* but then looked at the interaction and said that the effect of time *does* depend on treatment. This seemed like a fair way to say it, getting at the idea that the interaction is the most important thing. I wouldn't have done it this way, but I can't really fault doing so.)

Extra: I admit that I put `cattle.1a` in as a bit of misdirection, to make sure that you were confident enough to know not to look at it. The place where you *would* look at it is if the interaction had *not* been significant, and then you would have been able to conclude that there was an effect of time but not of treatment. Such a conclusion, though, is inconsistent with what we actually found: if there is actually no effect of treatment, it cannot be that the effect of time is different for the two treatments (that we just said are actually the same).

If you go back to Figure 13, you see that the reason for that significant interaction must be that the blue trend pops above the red one after about day 100. We have seen much more spectacular interaction effects than this one, and it seems like a very small P-value for such a small interaction effect. My guess is that once you allow for differences among animals (using the animal random effect) and the clear time effect, there is not a lot of random variation left, and so even a small effect can be strongly significant.

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

My answer:

The MANOVA analysis needs the data wider (one point) — specifically, it needs all the measurements for one animal in one row, with a separate column for each time point (the other point). If you get the second point, even without specifically saying anything about "wider", you get the first one as well. If you can say clearly what kind of data layout you need, that's the key thing.

Saying that we need to make a response variable is another way to get the first point, but you still need to say "one column per day" (or "one row per animal") to get the second one.

Extra: When you were doing (b), you might have been thinking about how normally `pivot_longer` is done before drawing a graph or doing a mixed-model analysis. That's because we usually get the data in wider format first and we have to make it longer to do those things. But this one is the other way around: the data came to us long, so it was easier to do

a graph and the mixed-model analysis first, and *then* make it *wider* to do the MANOVA-style analysis.

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

My answer:

We need to test the interaction.

- Begin by testing for sphericity, using the output from the second line of code. The P-value, from the second line of the table, is 8.9×10^{-85}, astonishingly small, so sphericity fails.
- That means we need to use one of the adjusted P-values below the third line of code. My recommendation is Huynh-Feldt (on the right); the P-value for the interaction is 0.0235. (I have no objection if you use Greenhouse-Geisser (on the left) instead, and get a P-value of 0.0254.)
- The interaction is significant still, and so the trend of weight over time is different for the two treatments.
- As before, going further is a mistake.

One point for doing (or appropriately not doing) each of those. Crowdmark lets you add point values to comments, either as a deduction (which is what I usually use), or as an addition, which is what I used here. Thus, a correct answer has four one-point comments on it, and if you are missing any of them, it means you did not get far enough to get the relevant point.

The P-value from the first line of code, 0.00018, is very like the one for the mixed model, but that is *wrong* here because sphericity fails (and therefore you cannot use the output from the first line of code at all). Note that the adjustment to the P-value is a substantial one: it goes from very strongly significant to barely significant at all, perhaps because sphericity fails so badly. (The `eps` values in the bottom output are a lot less than 1, indicating that sphericity has gone badly wrong. On one of the assignments, you might have run into a warning about `eps` values being bigger than 1, which they logically shouldn't be, but such a value meant that there was nothing wrong with sphericity at all.)

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

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

   My answer:

   This is a MANOVA, predicting the six quantitative variables from species. The P-value is very small, so species has some effect on some or all of the response variables. More than that we can not say, but there are some differences on something among species to find.

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

   My answer:

   We want to find an association between the categorical variable `species` and the various measurements, to determine what makes the species different or how we can most easily tell them apart.

   Or, take the angle that this is a followup to the MANOVA: we learned that there are some differences on some of the responses between the species, and the discriminant analysis will tell us which ones and how.

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

   My answer:

   There are three groups (species) and six variables, and the smaller of $3-1$ and 6 is 2.

   This was meant to be an easy one.

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

   My answer:

   No (or, at least, "probably not"), because it has a (very) small proportion of trace. We would expect almost all of the distinguishing of species to happen with the first linear discriminant. (But see later.)

   Make sure you mention the words "proportion of trace", or else it is not clear where the number you are citing comes from.

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

My answer:

The *most* important ones are those whose coefficient is farthest from zero, so the two least important ones have coefficient closest to zero: `nasal.width` and `mandible.depth`.

That means that the four variables `nasal.length`, `ramus.height`, `zygomatic.width`, and `lacrymal.width` are important in LD1, but it was easier to ask you (and easier to grade!) if I asked you for only two variables.

Strictly speaking, "smallest" is not quite right by itself, because what you want is "smallest in size" or "smallest in absolute value" or "closest to zero". What about a very negative coefficient, like the one for `zygomatic.width`? Does that count as "smallest"?

Extra: These coefficients are all pretty small. This is because the data values are on the large side (values in the hundreds: see Figure 17) so that a one-unit change in any of them would go with a tiny change in LD1 score.

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

My answer:

The two most important variables in `LD2` are `nasal.width` (negative) and `lacrymal.width` (positive). That means that in order to get a small (negative) LD2 score, you need a *large* value of `nasal.width` and a *small* value of `lacrymal.width`.

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

My answer:

This is a plot of `LD1` score vs `LD2` score for all the kangaroos.

- The fuliginosus (F) kangaroos are mostly on the left of the plot (low `LD1` score), and so seem to be easy to distinguish from the others.
- The other two species (green and blue points) are mainly mixed together on the right of the plot (high `LD1` score). If you prefer, say that melanops (M) is mainly at the top and giganteus (G) is mostly at the bottom (high and low LD2 scores respectively), but I think you ought to mention that they are at least somewhat mixed up.

In the light of what we said about `LD2` before, I actually think it is surprising that it distinguishes these two species as much as it does.

Extra: I found the green and blue circles I drew first a little difficult to distinguish (and, as

far as I know, my colour vision is all right), so I added the `shape` to the graph to make green triangles and blue squares. I hope that makes it a bit easier to distinguish the two sets of points on the right. I will have this printed in colour for the exam, but I never quite know how well the printing is going to come out.

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

My answer:

This is a "confusion matrix", or table of how often kangaroos of each species are misclassified:

- all 36 of the fuliginosus (F) kangaroos were correctly classified, which is consistent with them being on the left in the graph.
- several of the kangaroos of the other two species were wrongly classified, which is consistent with their being intermingled on the graph. (You might imagine that a kangaroo with an `LD2` score near zero could have come from either species G or M.)

You might add that the majority of kangaroos of the other two species were correctly classified, particularly giganteus (G), which is consistent with them being mostly, though not exclusively, at the bottom of the graph.

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

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

   My answer:

   Two individuals should be similar if their ratings for each of the seven sports are about the same (example: individuals 2 and 3 in Figure 22, which are actually identical). In that case, the differences will all be small, and the sum of squares of them will be small.

   Two individuals should be dissimilar if at least some of their ratings are very different (example: individuals 3 and 4 in the Figure). In that case, the big differences will become even bigger when squared, and the sum of squares of them will be large.

   I was pretty relaxed about how close to this you needed to get, and if you said something sort of relevant without getting to what I thought was worth 2, you got 1.

   Extra: I'm not expecting you to be able to figure out exactly what the function does, especially not under exam conditions, but the important line is the one that defines `diss` (the second-last line of the function), which is indeed based on the sum of the squared differences between the individuals' ratings.

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

   My answer:

   Negative numbers denote individuals, and positive numbers denote clusters. Thus, in row 6, individual 10 is being joined to the cluster formed in step 2 (one point), which consists of individuals 7 and 11 (the second point).

   This appears on the dendrogram in the middle cluster, where individuals 7 and 11 are combined (on the right), and later individual 10 is combined with that cluster to form a cluster with individuals 7, 10, and 11 in it. The third point. Describe as clearly as you can where to find this happening. (Individuals 6 and 9 are combined with that cluster, but not until later.)

   I might have deducted only a half point from you for missing the last thing if you otherwise clearly understood what was going on, on the grounds that the most important thing here was recognizing that the `merge` output meant that individual 10 was joined to the cluster containing individuals 7 and 11.

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

My answer:

They ended up in the same cluster because they are somehow similar. Look at Figure 25 to see what they have in common. The most obvious thing is that they all gave a high score to Swimming (7 with one 6), and so cluster 3 consists of individuals who want to participate in swimming.

I think this is the best answer, but it is also reasonable to look for *low* scores. For example, cluster 3 mostly consists of people who do *not* want to play baseball or football.

Or you can take the angle that all their ranks are close together (that is to say, they like the same things and don't like the same things).

Somebody even went a step further and suggested that these people like individual sports and don't like team sports.

Extra: you can eyeball the other clusters in the same way. People in cluster 1 want to play tennis, and people in cluster 2 want to play baseball.

These data came from a much larger dataset (130 people), but I restricted it to these 15 people so that you would have an easier dendrogram to look at. (A dendrogram with 130 individuals is rather a hard to read plot!)

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

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

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

My answer:

- In the scree plot, there is a clear elbow at 4, suggesting that we should use $4 - 1 = 3$ components.
- In the `summary` of the principal components, 3 components explain 81% of the variability, which is acceptably high. (Or something like that.)

As an alternative to the second point, note that the elbow at 4 is "a long way down the mountain", or "in the scree", or something similar, which gets at the same point. There is

an elbow at 7 also, suggesting six components (or one at 9 suggesting eight), but this is too many given that there are 11 variables and we are trying to summarize them by a much smaller number.

As an alternative alternative, another way of assessing whether the components you are thinking about using are "meaningful" is to see whether their SD is greater than 1 (this is called Kaiser's criterion, which I didn't talk about in lecture this time but is probably in the course materials somewhere). Here, that also points to 3 components. One of the nice things about this dataset is that you can get to 3 components several different ways, and as long as you can find elbow plus one of the other ways, you are good.

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

My answer:

The loadings of the first component on all the variables are about equal, so the first component is really "a bit of (almost) everything". Or, take the angle that it is hard to tell whether a variable belongs to the first component or not, because none of the loadings are clearly large or (mostly) small.

So I would say it does not have a clear interpretation.

If you want to say that it *does* have a clear interpretation because it's basically an average of everything, and that interpretation is "size", then be my guest. (The idea of component 1 being "size" is quite common, and this is a good example of it happening. A high scorer on component 1 here is a generally "big" car. We see later that a high scorer on factor 1 is also a "big" car, but in a different sense.) I'll read your answer and grade it not according to how well it agrees with mine, but on how consistent a job it does of answering the question, which this sort of answer 100% does.

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

My answer:

The ones with the largest loadings (in size, but they are all positive): length, wheelbase, width, and front shoulder room.

I think most people got these.

Note how the interpretation is much clearer for this factor analysis than it was for the principal components; I asked you for four variables because it pretty clear that those were *in* factor 1 and the others were *out*. This is the value of the rotation that goes with a factor analysis.

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

My answer:

These cars must be large on the four variables you mentioned in the previous part. Large length and wheelbase mean that the car is long; large width and shoulder room mean that it is wide. So these cars must be just generally big (large in size).

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

My answer:

Do the same two things you just did for factor 1: find out which variables need to be high, and

then say what those have in common.

Here that is rear seating, rear shoulder room, luggage space.

These cars seem to have a lot of space for passengers or cargo.

Extra: I would have expected these to be vans, but the vans are actually top of factor *2*, which is mainly height.

I figured you might like to see which actual cars came out on top of each factor.

```
scores <- data.frame(carname = cars$Carname, score = cars.2$scores)
```

There is a technicality here, because the scores are actually in a `matrix`, but if you add them to an old-fashioned `data.frame`, the columns of the matrix will be used as columns of the dataframe. (`tibble` is a lot more picky about what it will let you glue together; you could use it here, but I suspect there would be some messing about with `unnest`.)

Anyhow, the largest ones on Factor 1 are:

```
scores %>% slice_max(score.Factor1, n = 10)
```

carname	score.Factor1	score.Factor2	score.Factor3
LincolnTownCar	2.820438	-0.0688464	0.7860368
FordLTDCrownVictoria	2.727730	-0.3949726	0.5242082
CadillacBrougham	2.648630	-0.1074826	0.7697818
ChevroletCaprice	2.132903	-0.0245915	0.6672689
ChevroletAstro	1.727722	2.9742072	-1.7096809
FordThunderbird	1.578431	-0.6781103	0.2977158
CadillacDeVille	1.552789	0.1486800	0.6931368
BuickRiviera	1.374394	-0.5402926	0.1834544
ChevroletLuminaAPV	1.349803	1.8891140	-0.6582082
ChevroletCamaro	1.325792	-0.9693265	-0.2338936

You might recognize these as "boats", giant American cars from back in the day when the people who drove them didn't have to worry about the price of gas.

Factor 2 was height:

```
scores %>% slice_max(score.Factor2, n = 10)
```

carname	score.Factor1	score.Factor2	score.Factor3
VolkswagenVanagon	0.5213784	4.254193	-1.2488753
ChevroletAstro	1.7277220	2.974207	-1.7096809
NissanVan	-0.6878471	2.893194	-1.7345945
MitsubishiWagon	-0.8400708	2.881734	-1.8116823
FordAerostar	0.7921322	2.826875	-0.8355978
MazdaMPV	0.5462066	2.266413	-1.2458215
DodgeGrandCaravan	0.8244277	2.024106	-0.5077064
NissanAxxess	-0.6688422	1.967635	-0.5597348
DodgeCaravan	0.7579424	1.941608	-1.3700894
ChevroletLuminaAPV	1.3498029	1.889114	-0.6582082

and this is where the vans show up.

Factor 3 was a bit odd: there was nothing unusually large, but there were some unusually *small* ones, so I display these in *ascending* order:

```
scores %>% slice_min(score.Factor3, n = 5)
```

carname	score.Factor1	score.Factor2	score.Factor3
Nissan300ZX	1.0573148	-1.718415	-3.876875
ChevroletCorvette	0.9084246	-2.069313	-3.755734
HondaCivicCRX	-0.4674789	-1.536427	-3.693839
MazdaRX7	-0.0314903	-1.617555	-3.559693
MazdaMX5Miata	-0.7060277	-1.669412	-3.522470

These are cars that don't have much room for either passengers or cargo. You see that they are mostly Japanese cars; when the price of gas first went up (in the 1970s), the American carmakers didn't realize fast enough that there was a market for smaller fuel-efficient cars, and the Japanese carmakers took advantage.

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

My answer:

This must be because `FrtLegRoom` does not feature in any of our factors. If you look back at Figure 29, you'll see that `FrtLegRoom` is not an important part of any of the three factors (its biggest loading in size is -0.361 on factor 2, not nearly big enough to be considered an important part of factor 2).

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

Numbered Figures begin here, in with caption and label

```
library(MASS)
library(ggbiplot)
library(tidyverse)
library(marginaleffects)
library(lme4)
library(car)
```

Figure 1: Packages loaded

low birth weight (logistic)

low	lwt	smoke
no	120	yes
yes	148	no
yes	105	no
no	155	no
no	95	yes
yes	109	no
yes	102	yes
no	140	yes
yes	101	yes
no	109	yes
no	182	no
no	131	no
no	250	yes
no	130	no
no	133	yes

Figure 2: Low birth weight data (randomly chosen rows)

```r
birthwt.1 <- glm(factor(low) ~ lwt + smoke, data = birthwt, family = "binomial")
summary(birthwt.1)
```

```
Call:
glm(formula = factor(low) ~ lwt + smoke, family = "binomial",
    data = birthwt)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.62200    0.79592   0.781   0.4345
lwt         -0.01332    0.00609  -2.188   0.0287 *
smokeyes     0.67667    0.32470   2.084   0.0372 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 224.34  on 186  degrees of freedom
AIC: 230.34

Number of Fisher Scoring iterations: 4
```

Figure 3: Low birth weight logistic regression

```
new <- datagrid(model = birthwt.1, lwt = c(110, 125, 140))
new
```

low	smoke	lwt
no	no	110
no	no	125
no	no	140

```
cbind(predictions(birthwt.1, newdata = new)) %>%
  select(smoke, lwt, estimate)
```

smoke	lwt	estimate
no	110	0.3007605
no	125	0.2604667
no	140	0.2238432

```
new <- datagrid(model = birthwt.1, smoke = c("no", "yes"))
new
```

low	lwt	smoke
no	130	no
no	130	yes

```
cbind(predictions(birthwt.1, newdata = new)) %>%
  select(smoke, lwt, estimate)
```

smoke	lwt	estimate
no	130	0.2478400
yes	130	0.3932927

Figure 4: Low birth weight predictions

display (two-way anova)

time	panel	emergenc
17	1	1
14	1	1
15	2	1
12	2	1
21	3	1
24	3	1
25	1	2
24	1	2
22	2	2
19	2	2
29	3	2
28	3	2
31	1	3
24	1	3
28	2	3
31	2	3
32	3	3
37	3	3
14	1	4
13	1	4
9	2	4
10	2	4
15	3	4
19	3	4

Figure 5: Display panels data

```r
display.1 <- aov(time ~ panel * emergenc, data = display)
summary(display.1)
```

```
              Df Sum Sq Mean Sq F value   Pr(>F)
panel          2  232.7   116.4  20.094 0.000148 ***
emergenc       3 1052.5   350.8  60.573 1.61e-07 ***
panel:emergenc 6   28.9     4.8   0.832 0.567501
Residuals     12   69.5     5.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
display.2 <- aov(time ~ panel + emergenc, data = display)
summary(display.2)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
panel        2  232.7   116.4   21.29 1.81e-05 ***
emergenc     3 1052.5   350.8   64.16 8.28e-10 ***
Residuals   18   98.4     5.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 6: Display panels: two analyses

```r
TukeyHSD(display.2)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = time ~ panel + emergenc, data = display)

$panel
      diff       lwr        upr     p adj
2-1 -2.000 -4.983847  0.9838467 0.2284740
3-1  5.375  2.391153  8.3588467 0.0006236
3-2  7.375  4.391153 10.3588467 0.0000173

$emergenc
          diff        lwr          upr     p adj
2-1   7.333333   3.517810  11.14885716 0.0001980
3-1  13.333333   9.517810  17.14885716 0.0000001
4-1  -3.833333  -7.648857  -0.01780951 0.0487057
3-2   6.000000   2.184476   9.81552383 0.0016255
4-2 -11.166667 -14.982190  -7.35114284 0.0000009
4-3 -17.166667 -20.982190 -13.35114284 0.0000000
```

Figure 7: Display panels: Tukey analysis

vocab (ancova)

year	education	vocabulary
1974	8	3
1994	12	5
1974	12	7
1974	15	5
1994	12	3
1974	13	10
1994	20	10
2014	14	6
2014	12	6
1994	12	5
1994	12	7
2014	19	6
2014	16	7
2014	20	9
1994	12	5
1994	12	9
1994	16	7
1974	14	9
1974	12	1
1974	18	9

Figure 8: Education and vocabulary data (some randomly chosen rows)

```
ggplot(vocab, aes(x = education, y = vocabulary, colour = factor(year))) +
  geom_jitter() + geom_smooth(method = "lm")
```

```
`geom_smooth()` using formula = 'y ~ x'
```

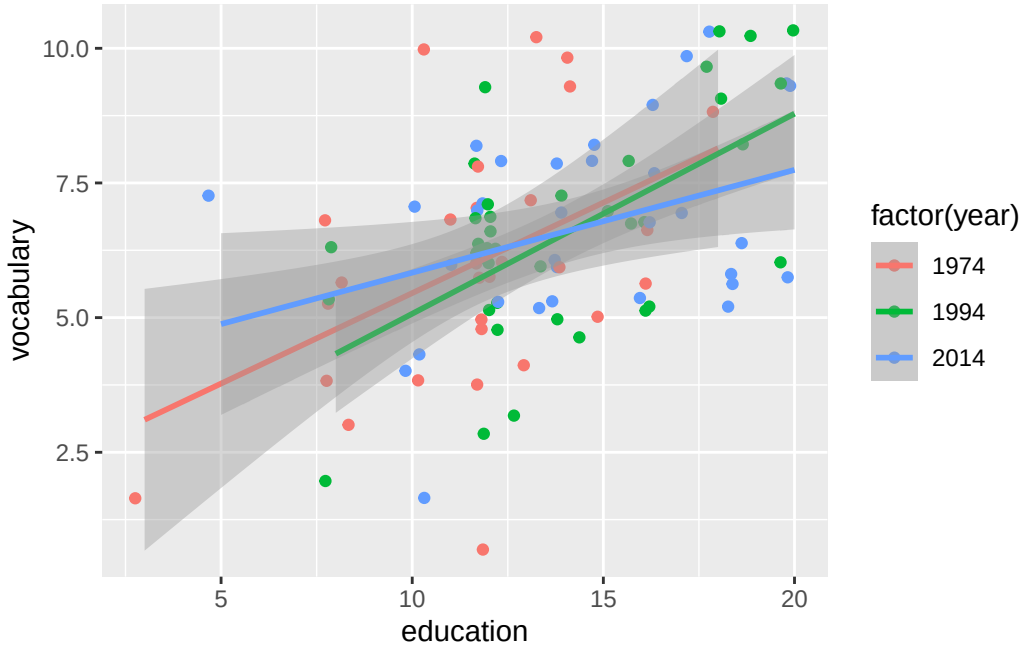


Figure 9: Scatterplot of education and vocabulary by year

```r
vocab.1 <- lm(vocabulary ~ education * factor(year), data = vocab)
summary(vocab.1)
```

```
Call:
lm(formula = vocabulary ~ education * factor(year), data = vocab)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1272 -1.3613  0.0457  1.2094  4.5447

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  2.09583    1.32000   1.588  0.11570
education                    0.33594    0.10909   3.080  0.00272 **
factor(year)1994            -0.73778    1.83257  -0.403  0.68816
factor(year)2014             1.83017    1.84771   0.991  0.32447
education:factor(year)1994   0.03545    0.14032   0.253  0.80111
education:factor(year)2014  -0.14510    0.13968  -1.039  0.30158
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.758 on 94 degrees of freedom
Multiple R-squared:  0.2675,    Adjusted R-squared:  0.2285
F-statistic: 6.866 on 5 and 94 DF,  p-value: 1.679e-05
```

```r
drop1(vocab.1, test = "F")
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
	NA	NA	290.6165	118.6834	NA	NA
education:factor(year)	2	7.157061	297.7735	117.1163	1.157477	0.3187163

Figure 10: Education and vocabulary analysis of covariance

```r
vocab.2 <- lm(vocabulary ~ education + factor(year), data = vocab)
summary(vocab.2)
```

```
Call:
lm(formula = vocabulary ~ education + factor(year), data = vocab)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1155 -1.3130  0.0278  1.1293  4.4720

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.59073    0.71278   3.635  0.00045 ***
education         0.29373    0.05402   5.438  4.1e-07 ***
factor(year)1994 -0.14335    0.45417  -0.316  0.75297
factor(year)2014 -0.14745    0.46826  -0.315  0.75352
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.761 on 96 degrees of freedom
Multiple R-squared:  0.2495,    Adjusted R-squared:  0.226
F-statistic: 10.64 on 3 and 96 DF,  p-value: 4.221e-06
```

Figure 11: Education and vocabulary analysis of covariance 2

kenward cattle (repeated measures)

animal	trt	day	weight
A1	A	0	233
A1	A	14	224
A1	A	28	245
A1	A	42	258
A1	A	56	271
A1	A	70	287
A1	A	84	287
A1	A	98	287
A1	A	112	290
A1	A	126	293
A1	A	133	297
A10	A	0	232
A10	A	14	240
A10	A	28	247
A10	A	42	263
A10	A	56	275
A10	A	70	286
A10	A	84	294
A10	A	98	302
A10	A	112	308
A10	A	126	319
A10	A	133	326
A11	A	0	234
A11	A	14	237
A11	A	28	259

Figure 12: Cattle data (some)

```
cattle %>%
  group_by(trt, day) %>%
  summarize(mean_weight = mean(weight)) %>%
  ggplot(aes(x = day, y = mean_weight, colour = trt, group = trt)) +
    geom_point() + geom_line()
```

```
`summarise()` has grouped output by 'trt'. You can override using the `.groups`
argument.
```

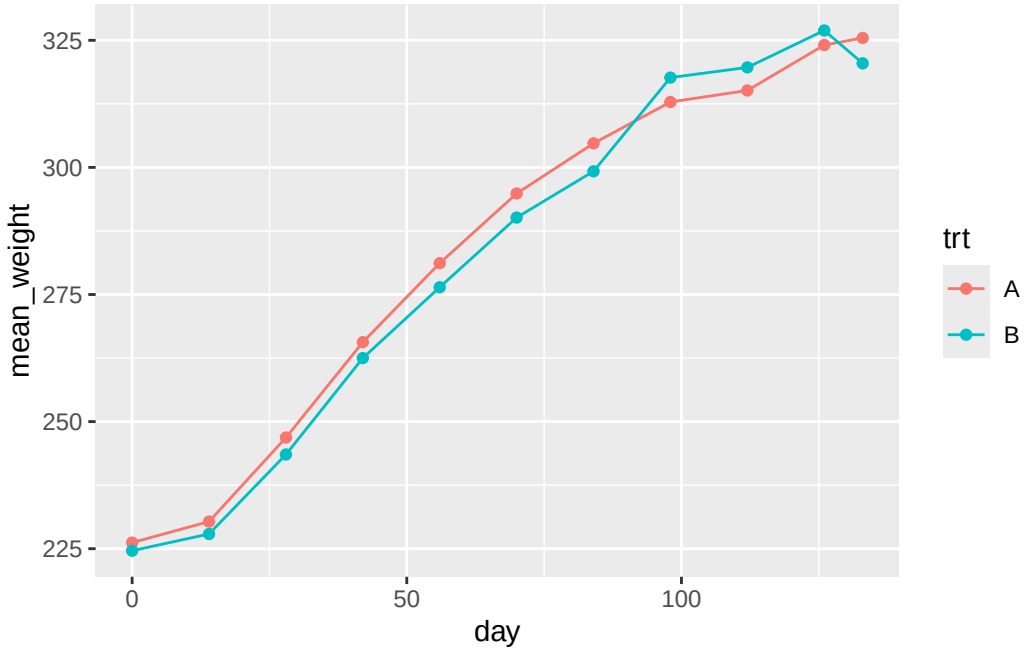


Figure 13: Cattle data interaction plot

```r
cattle.1 <- lmer(weight ~ trt * factor(day) + (1 | animal), data = cattle)
drop1(cattle.1, test = "Chisq")
```

	npar	AIC	LRT	Pr(Chi)
	NA	4866.584	NA	NA
trt:factor(day)	10	4881.634	35.04987	0.0001224

```r
cattle.1a <- lmer(weight ~ trt + factor(day) + (1 | animal), data = cattle)
drop1(cattle.1a, test = "Chisq")
```

	npar	AIC	LRT	Pr(Chi)
	NA	4881.634	NA	NA
trt	1	4879.839	0.2047196	0.650938
factor(day)	10	6721.837	1860.2032319	0.000000

Figure 14: Cattle data mixed model analysis

```r
# pivot wider
cattle %>%
  pivot_wider(names_from = day, values_from = weight) -> cattle_wider
# set up for manova
cattle_wider %>% select(-animal, -trt) %>%
  as.matrix() -> response
cattle.2 <- lm(response ~ trt, data = cattle_wider)
times <- colnames(response)
times.df <- data.frame(times = factor(times))

cattle.3 <- Manova(cattle.2, idata = times.df, idesign = ~times)
```

Figure 15: Cattle repeated measures MANOVA code

```
  summary(cattle.3)$univariate.tests
```

```
             Sum Sq num Df Error SS den Df    F value    Pr(>F)
(Intercept) 53035479      1   133128     58 23106.1031 < 2.2e-16 ***
trt             455      1   133128     58     0.1982 0.6578077
times         846142     10    37638    580  1303.9048 < 2.2e-16 ***
trt:times       2264     10    37638    580     3.4891 0.0001767 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
  summary(cattle.3)$sphericity
```

```
         Test statistic    p-value
times          3.399e-05 8.9223e-85
trt:times      3.399e-05 8.9223e-85
```

```
  summary(cattle.3)$pval.adjustments
```

```
            GG eps   Pr(>F[GG])    HF eps    Pr(>F[HF])
times     0.2415572 2.496758e-96 0.2528023 1.100120e-100
trt:times 0.2415572 2.535322e-02 0.2528023  2.346705e-02
attr(,"na.action")
(Intercept)         trt
          1           2
attr(,"class")
[1] "omit"
```

Figure 16: Cattle repeated measures MANOVA output

kanga (discrim)

species	nasal.length	nasal.width	ramus.height	zygomatic.width	mandible.depth	lacrymal.width
fuliginosus	552	205	751	919	194	454
giganteus	755	268	754	902	206	467
fuliginosus	574	212	641	822	191	405
giganteus	756	249	731	903	198	467
melanops	565	204	556	764	156	385
giganteus	687	223	688	873	205	432
giganteus	682	253	706	875	194	455
fuliginosus	522	190	629	799	179	374
fuliginosus	719	253	765	946	215	473
fuliginosus	554	195	657	837	188	392
melanops	893	260	824	994	216	499
fuliginosus	625	250	739	934	211	470
fuliginosus	497	167	648	807	178	390
giganteus	629	222	643	824	181	416
giganteus	616	220	652	805	180	412
melanops	800	245	813	939	240	492
fuliginosus	737	278	880	1090	271	535
melanops	690	242	708	855	210	451
giganteus	626	226	651	839	173	441
giganteus	734	245	724	920	193	462

Figure 17: Kangaroo skull data, variables of interest, randomly chosen rows

```
kanga %>%
  select(nasal.length:lacrymal.width) %>%
  as.matrix() -> response
kanga.0 <- manova(response ~ species, data = kanga)
summary(kanga.0)
```

```
          Df Pillai approx F num Df den Df    Pr(>F)
species    2 1.0065   15.873     12    188 < 2.2e-16 ***
Residuals 98
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

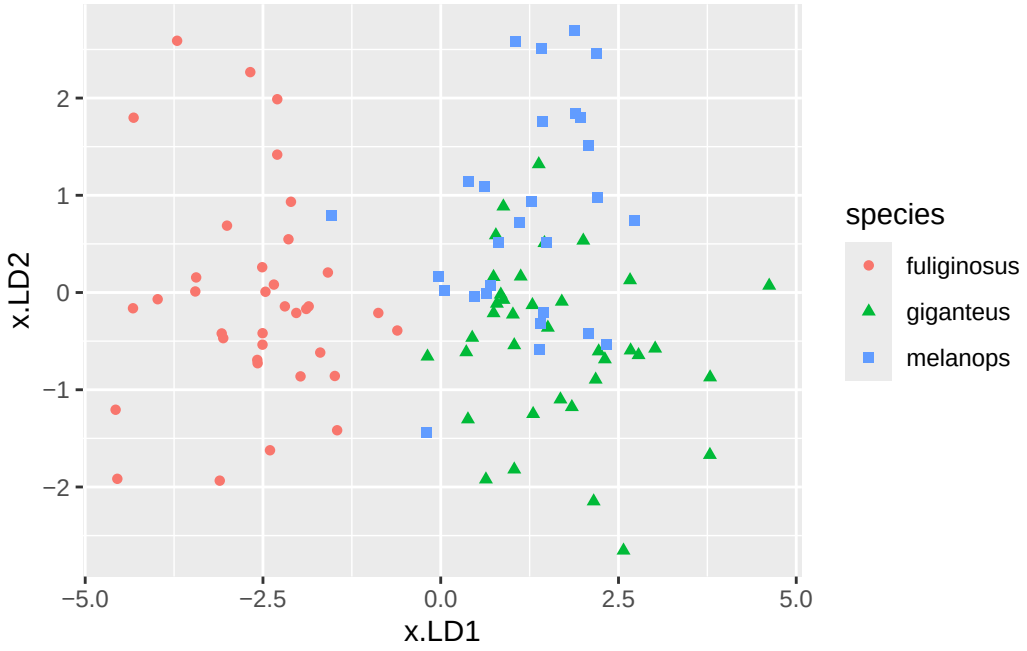
Figure 18: Kangaroo skull analysis 1

```r
kanga.1 <- lda(species ~ nasal.length + nasal.width + ramus.height +
                 zygomatic.width + mandible.depth + lacrymal.width, data = kanga)
kanga.1
```

```
Call:
lda(species ~ nasal.length + nasal.width + ramus.height + zygomatic.width +
    mandible.depth + lacrymal.width, data = kanga)

Prior probabilities of groups:
fuliginosus   giganteus    melanops
  0.3564356   0.3663366   0.2772277

Group means:
            nasal.length nasal.width ramus.height zygomatic.width
fuliginosus     614.1944    222.4722     728.5000        912.6944
giganteus       707.3243    246.8919     686.4595        869.8649
melanops        684.1071    233.2857     695.2143        860.5000
            mandible.depth lacrymal.width
fuliginosus       205.7500       442.3056
giganteus         194.3784       444.2973
melanops          194.7500       444.3214

Coefficients of linear discriminants:
                        LD1          LD2
nasal.length     0.027729200  0.003536769
nasal.width     -0.006006286 -0.067051346
ramus.height    -0.013524008  0.016312892
zygomatic.width -0.025345101 -0.032317300
mandible.depth  -0.008867705 -0.003186253
lacrymal.width   0.022430867  0.060543739

Proportion of trace:
   LD1    LD2
0.9359 0.0641
```

Figure 19: Kangaroo skulls, discriminant analysis

```
p <- predict(kanga.1)
d <- cbind(kanga, p)
ggplot(d, aes(x = x.LD1, y = x.LD2, colour = species, shape = species)) + geom_point()
```



Note that the points on the plot are distinguished by both colour and shape (plotting symbol).

Figure 20: Kangaroo skulls, further analysis

```
with(d, table(obs = species, pred = class))
```

```
            pred
obs          fuliginosus giganteus melanops
  fuliginosus          36         0        0
  giganteus             0        32        5
  melanops              1        10       17
```

Figure 21: Kangaroo skulls, a table

cluster

Baseball	Football	Basketball	Tennis	Cycling	Swimming	Jogging
7	5	1	6	4	3	2
7	5	6	3	4	1	2
7	5	6	3	4	1	2
4	1	5	7	3	2	6
6	5	7	1	2	3	4
2	5	4	1	6	7	3
3	1	5	4	6	7	2
3	4	1	7	6	5	2
2	1	3	5	4	7	6
1	3	5	7	4	6	2
4	2	3	5	6	7	1
7	4	3	2	1	5	6
3	1	4	7	6	2	5
4	7	5	6	2	3	1
6	7	5	3	2	1	4

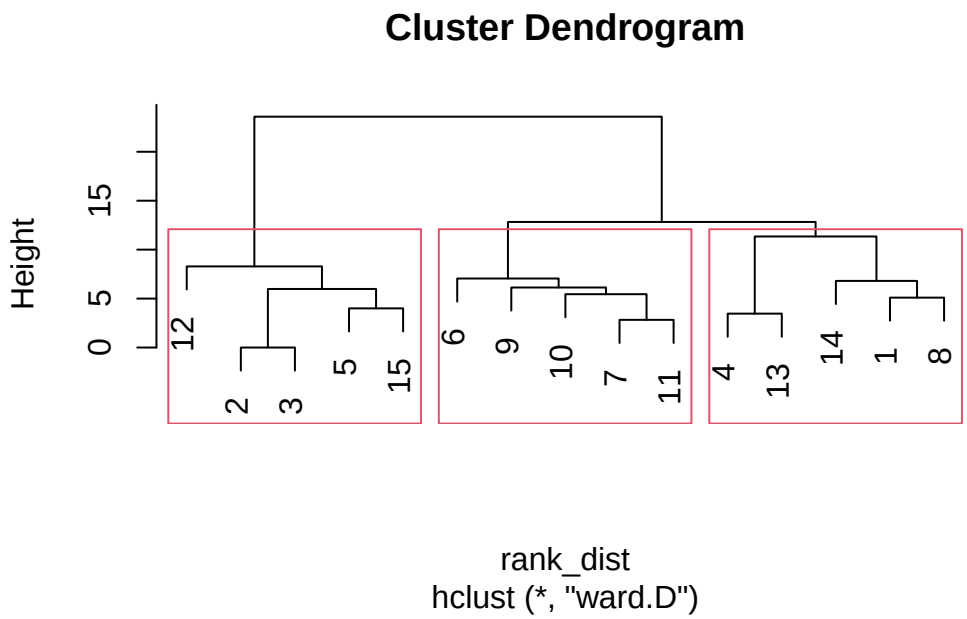
Figure 22: Sports preference data

```
my_dist <- function(d, i, j) {
  d %>% slice(i, j) %>%
    mutate(indiv = c("row1", "row2")) %>%
    pivot_longer(-indiv, names_to = "sport", values_to = "rating") %>%
    pivot_wider(names_from = indiv, values_from = rating) %>%
    summarize(diss = sqrt(sum((row1 - row2)^2))) %>%
    pull(diss)
}
```

Figure 23: Function to compute dissimilarity for sports ranking data

The data in `ranks` were first turned into a `dist` object, called `rank_dist`, and then the following code was run:

```
ranks.1 <- hclust(rank_dist, method =  "ward.D")
plot(ranks.1)
rect.hclust(ranks.1, 3)
```




```
ranks.1$merge
```

```
      [,1] [,2]
 [1,]   -2   -3
 [2,]   -7  -11
 [3,]   -4  -13
 [4,]   -5  -15
 [5,]   -1   -8
 [6,]  -10    2
 [7,]    1    4
 [8,]   -9    6
 [9,]  -14    5
[10,]   -6    8
[11,]  -12    7
[12,]    3    9
[13,]   10   12
[14,]   11   13
```

Figure 24: Cluster analysis for sports data

id	cluster	Baseball	Football	Basketball	Tennis	Cycling	Swimming	Jogging
1	1	7	5	1	6	4	3	2
4	1	4	1	5	7	3	2	6
8	1	3	4	1	7	6	5	2
13	1	3	1	4	7	6	2	5
14	1	4	7	5	6	2	3	1
2	2	7	5	6	3	4	1	2
3	2	7	5	6	3	4	1	2
5	2	6	5	7	1	2	3	4
12	2	7	4	3	2	1	5	6
15	2	6	7	5	3	2	1	4
6	3	2	5	4	1	6	7	3
7	3	3	1	5	4	6	7	2
9	3	2	1	3	5	4	7	6
10	3	1	3	5	7	4	6	2
11	3	4	2	3	5	6	7	1

Figure 25: Individuals (in column `id`) and cluster memberships for sports preference data

cars (pca/factor)

Carname	Length	Wheelbase	Width	Height	FrontHd	RearHd	FrtLegRoom
NissanPulsarNX	-1.3333333	-1.2	-1.0	-4.0	1	-2.0	-1
InfinitiQ45	2.3333333	2.2	2.0	0.5	-2	-0.5	0
HondaPrelude	-0.1111111	-0.2	-0.5	-3.0	-4	-2.0	2
HondaAccord	0.6666667	1.0	0.0	-0.5	2	0.0	1
FordAerostar	-0.4444444	3.4	2.0	16.5	3	1.5	1
ChevroletCavalier	0.0000000	-0.2	-1.0	0.0	1	1.0	-1
Porsche944	-1.1111111	-1.4	0.0	-4.0	0	-2.0	4
SubaruLoyale	-0.4444444	-1.0	-1.5	-1.0	-2	0.0	-1
GEOMetro	-3.2222222	-1.8	-2.5	-1.5	-2	-0.5	-1
Mazda626	0.0000000	-0.2	-0.5	-0.5	0	0.0	-2
BuickRiviera	2.1111111	1.2	2.5	-1.5	0	0.0	0
PontiacLeMans	-0.7777778	-0.6	-1.0	0.0	3	0.5	0
ChevroletLumina	2.1111111	1.2	1.5	1.0	2	1.5	1
ToyotaCelica	-0.5555556	-0.6	0.5	-3.0	-3	-2.0	1
PontiacBonneville	2.2222222	1.8	2.0	0.5	1	2.0	1

Figure 26: Cars data (some rows and columns)

```r
cars.1 <- princomp(cars_numeric, cor = TRUE)
summary(cars.1)
```

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
Standard deviation     2.1859859 1.4764164 1.3821796 0.82691759 0.73804436
Proportion of Variance 0.4344122 0.1981641 0.1736746 0.06216297 0.04951904
Cumulative Proportion  0.4344122 0.6325763 0.8062509 0.86841391 0.91793295
                           Comp.6     Comp.7     Comp.8      Comp.9     Comp.10
Standard deviation     0.55143677 0.44712743 0.41450063 0.308496259 0.269616682
Proportion of Variance 0.02764386 0.01817481 0.01561916 0.008651813 0.006608469
Cumulative Proportion  0.94557681 0.96375163 0.97937079 0.988022601 0.994631070
                          Comp.11
Standard deviation     0.24301900
Proportion of Variance 0.00536893
Cumulative Proportion  1.00000000
```

```r
ggscreeplot(cars.1)
```

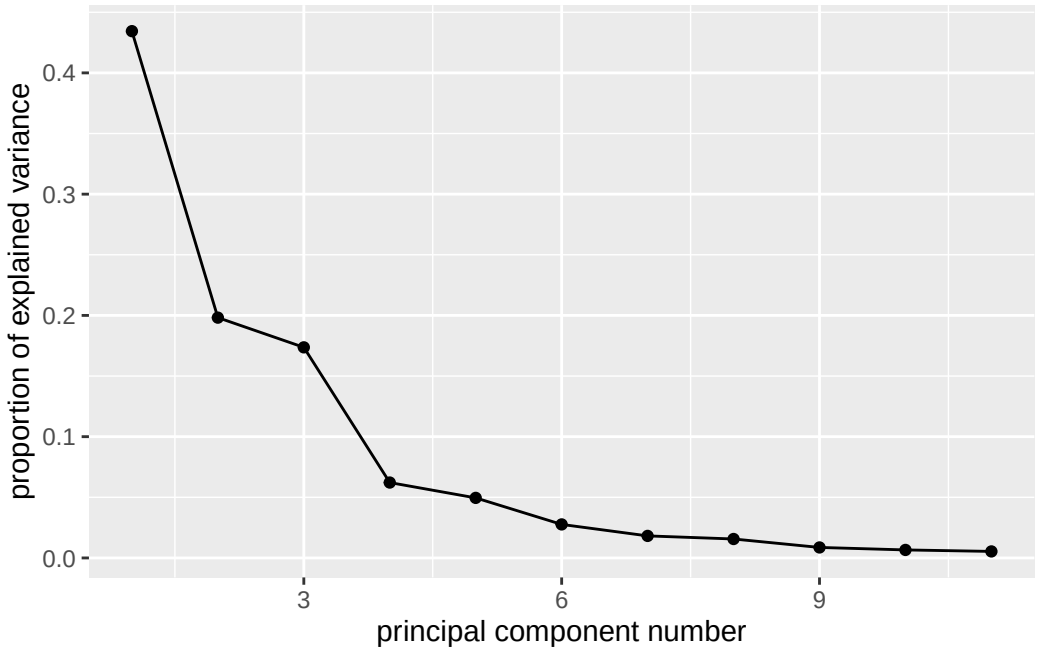


Figure 27: Cars principal components analysis and scree plot

```
               Comp.1      Comp.2      Comp.3
Length      0.36800447  0.26491429  0.15382216
Wheelbase   0.39303399  0.19208728  0.16487812
Width       0.36271313  0.08827474  0.37020532
Height      0.25231975 -0.48930844  0.04282917
FrontHd     0.27563231 -0.27414127 -0.05442509
RearHd      0.32033083 -0.37823667 -0.19691964
FrtLegRoom  0.03600793  0.41677436  0.32261159
RearSeating 0.28877687  0.13354004 -0.45373369
FrtShld     0.39776393 -0.04124462  0.27343479
RearShld    0.28585066  0.09344998 -0.46883122
Luggage     0.10803017  0.47489313 -0.40497703
```

Figure 28: Cars data: loadings of first three principal components

```
cars.2 <- factanal(cars_numeric, 3, scores = "r")
cars.2$loadings
```

```
Loadings:
            Factor1 Factor2 Factor3
Length       0.837           0.346
Wheelbase    0.810   0.172   0.309
Width        0.959   0.157
Height       0.193   0.935
FrontHd      0.318   0.481   0.150
RearHd       0.200   0.831   0.291
FrtLegRoom   0.341  -0.361
RearSeating  0.180   0.251   0.803
FrtShld      0.867   0.388
RearShld     0.123   0.343   0.822
Luggage      0.106  -0.400   0.799

               Factor1 Factor2 Factor3
SS loadings      3.382   2.472   2.295
Proportion Var   0.307   0.225   0.209
Cumulative Var   0.307   0.532   0.741
```

Figure 29: Cars data: factor analysis and factor loadings

```
cars.2$uniquenesses
```

```
    Length   Wheelbase       Width      Height     FrontHd      RearHd
0.18046602  0.21837389  0.05494901  0.08052403  0.64476164  0.18494410
FrtLegRoom RearSeating     FrtShld    RearShld     Luggage
0.75282443  0.25953381  0.09399943  0.19096693  0.18980421
```

Figure 30: Cars data: uniquenesses