University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 (K. Butler), Final Exam
April 9, 2026

Aids allowed (on paper, no computers or devices):

- 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 12 numbered pages of questions including this cover page.

In addition, you have an additional booklet of Figures to refer to during the exam.

The maximum marks available for each question are shown next to the question number.

For more space, use space on the last page. 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.

Budworms

Budworms are insects that attack tobacco plants. A study was made of the effectiveness of a particular insecticide at killing budworms. It was expected that the effectiveness of the insecticide would depend on the dose and would be different for male and female budworms. Twenty randomly chosen budworms of each sex were exposed to each of six different doses (different budworms each time), and the number that lived and died were recorded. The data from the study are shown in Figure 2. The columns are: `ldose`, the dose of the insecticide, the `sex` of the budworms at that dose (abbreviated M and F), the number of budworms that were alive at the end of the study, and the number that were dead.

(1) (2 points) In Figure 3, a logistic regression is fitted. In the first line of code above the output, why was the `cbind` necessary?

This makes a two-column response. This is necessary because each row of Figure 2 represents not one but twenty budworms (more than one), and each row of the dataframe summarizes how *many* budworms of that sex at that dose were alive or dead at the end of the study.

Contrast this with the logistic regression problem on the midterm, where each row of the dataframe there, in column `x`, indicated whether the seed germinated or not, so it was only talking about *one* seed. This question is therefore the opposite of question 5 on the midterm.

Points:

- 2: the `cbind` makes a two-column response because each row of the dataframe represents more than one budworm (actually twenty)
- 1.5: there is more than one individual per row (without being specific about what the individuals are or how many of them there are)
- 1: `cbind` makes a two-column response without correct or complete explanation of why it is needed here
- 0.5: other relevant comment not enough for 1

(2) (2 points) In Figure 3, how do you know that the model is predicting the probability that a budworm is *alive* and not that a budworm is dead?

With a two-column response, the probability being predicted is that of the *first column* in the two-column response. In the `cbind` at the top of Figure 3, the first column is `numalive`, so we are predicting the probability that a budworm is alive (at the end of the study).

The fact that `numalive` is alphabetically before `numdead` is true but *irrelevant* here. Compare question 9 on the midterm: with a two-column response (and multiple observations per

row of the data), the first column of the response is what matters, but with one observation per row, the first level alphabetically of the response is what matters.

Points:

- 2: the model predicts the probability of whatever is in the *first column* of the response
- 1: some progress towards the 2-pt answer but not clear or complete enough
- 0: `numalive` is alphabetically before `numdead`
- 0: no points

(3) (2 points) In Figure 3, both explanatory variables are significant. Interpret the numerical value of the Estimate for `ldose`.

The Estimate for `ldose` is -1.06 (rounded). This says that if dose increases by 1, the *log-odds* of a budworm being alive *decreases* by 1.06, all else equal (or, for either sex).

- 2: if dose increases by 1, the *log-odds* of a budworm being alive *decreases* by 1.06, all else equal (or, for either sex)
- 1.5: as 2, but "increases by -1.06" (hard to for reader to understand)
- 1: missing "all else equal" (or equivalent such as "for both sexes")
- 0.5: as 2, but "probability" instead of log-odds
- 0.5: reasonable comment but not enough for more points

Extra: We are testing an insecticide, so this is as you would guess. If the dose increases, you would expect fewer budworms to survive, and this is also the impression you get from Figure 2: within a sex, as the dose increases, the number alive out of 20 decreases.

(4) (2 points) In Figure 3, interpret the numerical value of the Estimate for `sexM`.

The Estimate for `sexM` is -1.10 (rounded). The explanatory variable `sex` is categorical, so this says that, compared to the baseline sex (`F`, female), the log-odds of a male budworm being alive is less by 1.10 at any dose (or, holding dose constant, or all else equal). "Log-odds is greater for females" by this much is equivalent and thus also good.

"Increasing `sex` by 1" is nonsense because `sex` is not quantitative.

That is to say, male survival is worse than female, at any dose. You can guess this by looking at Figure 2: for any value of `ldose`, there were more females alive than males.

My major focus here was on the comparison of males *with females*, so if you interpreted the actual number in the same way as in the previous question, you were likely to be good (as long as you interpreted it somehow).

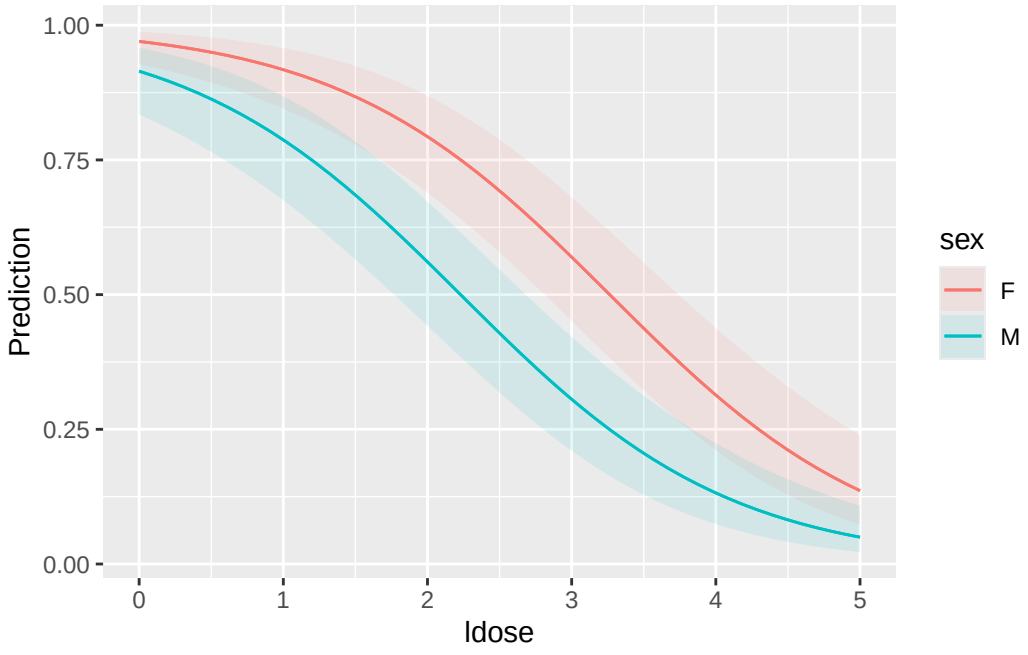
Points:

- 2: compared to the baseline sex (`F`, female), the log-odds of a male budworm being alive is less by 1.10 at any dose (or, holding dose constant, or all else equal).
- 1: not explicitly comparing males *with females*
- 1: not assessing numerical value
- 1: backwards (eg concluding prob. of survival is *higher* for males)
- 0.5: something correct but too many errors/omissions for more points

(5) (3 points) After running the code in Figure 3, I ran the code `plot_predictions(budworm.1, condition = c("ldose", "sex"))`. Sketch the graph that would be the result of running that code. Hint: if your graph needs different colours, instead identify the things that would be different colours by labelling them. (The exam will be scanned in black and white, so even if you can draw with different colours, it is still better to label things.)

I can once again cheat and show you exactly what the result would be:

```
plot_predictions(budworm.1, condition = c("ldose", "sex"))
```



From a model like this one:

- the y-axis is the predicted probability of being alive, so `P(alive)` is enough as a label for it, with values going from 0 to 1. "P(alive)" is in any case arguably a better label

than the one `plot_predictions` gives. (A lot of people thought this was the *number* alive, but it's actually the probability. Giving the y-axis some values shows that you have this right.)

- the x-axis is the first thing in `condition`, namely `ldose`, and you know from Figure 2 that the values go from 0 to 5. Just the lowest and highest values are sufficient.
- the colours are the second thing in `condition`, so you will get one curve for each sex, one above the other all the way along. Label the upper one "female" or "F" in some way, for example by an arrow pointing to it saying "female", and label the lower one "male" or "M" in the same way. (You know that the probability of survival is higher for females from your interpretation of the Estimate for `sex`, or from looking at the original data.)
- you will also get a legend showing which colour is which, but as long as you identify which curve goes with which sex somehow, you are good. (I guess you can label the top curve "red" and the bottom one "blue", and then somehow explain which colour is which sex.)

The actual graph has coloured "envelopes" showing the accuracy of the predictions. I do not need you to draw these on your sketch.

It is best if the predictions curve a bit as they approach 0 or 1, to make it clear that they cannot pass those limits.

Points:

- 3: graph with axes appropriately labelled and with at least two values (each); two curves with the female one above the male one, appropriately labelled; something indicating that the probabilities don't go beyond 0 or 1
- 2.5: as 3, but with eg. straight-line trends that might go beyond 0 or 1
- 2: as 3, but with an axis label / values missing
- 2: as 3, with two curves, but labelled the wrong way around
- 1.5: not labelling the male and female curves
- 1.5: as 3, but putting `sex` on x-axis and `ldose` as `fill` somehow
- 1: only one curve instead of two
- 1: as 3 but more than one thing missing / in error
- 0.5: something correct but not enough for more points
- 0: no points

Birth defects of the central nervous system

Babies can be born with various defects of the central nervous system. There are several such defects; our data contain counts of Anencephalus (`An`) and Spina Bifida (`Sp`), with

other birth defects of the central nervous system collected in `Other`. The data also contain counts of the number of babies born with no central nervous system birth defect (in `NoCNS`), along with two demographic variables: the hardness of the `Water` where the parents live (measured in parts per million), and the kind of `Work` done by the parents (classified as `Manual` and `NonManual`). The two demographic variables are believed to have some kind of effect on the kind of birth defect (if any). The data were collected in South Wales (UK) in 1971; at that time, some people were employed in coal-mining and related industries (`Manual` work), but others were employed in offices (`NonManual` work). The data, in dataframe `cns0`, are shown in Figure 4, in the form that I received them.

(6) (2 points) Why will it be appropriate to use `multinom` (from the `nnet` package) to model the relationship of interest?

The response variable is the kind of birth defect if any. This is categorical with four levels (more than two): `NoCNS`, `An`, `Sp`, `Other`. The categories do not have a natural order; even if you put `NoCNS` first, there is no order to the other three categories. Hence, `multinom` rather than `polr` (which would be appropriate if the response categories were ordered) or `glm` with `family = "binomial"` (which would be appropriate if the response had two categories only).

This might be confusing since the birth defect types are not (yet) all in one column in dataframe `cns0`. You might like to look at the next question first if you are unsure (this is another good reason to at least scan the whole exam before you write anything). The thing to be guided by is the *description* of the data, rather than its current layout, or to look at the dataframe `cns` in Figure 5 which is the one actually used for model-fitting.

Points:

- 2: birth defect type has *more than two unordered* levels (naming the levels will do for "more than two")
- 1: missing one of "more than two levels" or "unordered"

Extra: about water hardness: hard water has a high concentration of calcium or magnesium. When the water is hard, this can have a bad effect on household plumbing or appliances (such as kettles), and also on health, which is the issue here. This is a good read on the subject. The units are parts per million *of calcium and magnesium ions* dissolved in the water, which can be high if the water has come from an area where there is limestone.

(7) (2 points) Look at Figure 5. Why is it necessary to run this code before doing any modelling?

We need all the birth defect categories in one column (along with a column of counts of each one) so that we can fit a model that predicts birth defect category from our explanatory variables.

Points:

- 2: need all the birth defect categories in one column
- 1: something relevant but not clear or complete enough

(8) (2 points) A model is fitted in Figure 6. What would have happened if I had not included the `weights = count` in the code?

This is how the model knows *how many* incidences there were of birth defects of each type (for each water hardness and parental work category). If I had left it out, the model would have assumed *one* case (baby) in each row of the dataframe. (This is actually the same issue as in the logistic regression question, but with a different model it is handled a different way.)

The model would have fitted perfectly well, but the results, such as prediction later, would have been nonsense. This is the same issue as in the Extra to question 4 on Assignment 3; the predicted probabilities for each birth defect category would have been $1/4 = 0.25$ regardless of the values of anything else, which cannot be correct because the vast majority of the babies born had no central nervous system birth defects at all, and so the probability of `NoCNS` should be predicted as close to 1.

Points:

- 2: each row would have represented *one* baby, or we need to account for the number of babies per row, which is what the `weights` does.
- 1: some relevant comment, but not precise enough

(9) (2 points) The `step` function was applied to the model of the previous question. The output is shown in Figure 7. What is the most important conclusion from this output? Explain briefly.

The function `step` tries to build a better model by seeing which explanatory variables can be removed. There are several ways to see what we should conclude from the `step` output:

- the `Call:` near the bottom summarizes where `step` finished. The best model predicts `birth_type` from both `Water` and `Work`.
- just above the `Call`, the table of AIC values shows that the best (smallest) one goes with removing nothing, and therefore that removing nothing is better than removing either explanatory variable.

- the table of Coefficients at the bottom shows that the final model contains a coefficient for both `Water` and the non-baseline category of the categorical `Work`. These are in columns here because that is the default `summary` of this kind of model (that we didn't really look at) with the non-baseline response categories in rows.

All of these are saying that we need to keep both explanatory variables `Water` and `Work` in the model; removing either of them would be a mistake. Any one of them is good for the "explain briefly" part.

Points:

- 2: we should not remove either of the explanatory variables, along with something that says how you know
- 1: as 2, but incomplete or absent explanation of why

Extra 1: for a `step` output, this one is pretty short. Here's what was done:

- start from a model with both explanatory variables
- see how well a model without `Water` fits (that is, including `Work`)
- see how well a model without `Work` fits (including `Water`)
- summarize the AIC values for removing each explanatory variable and for removing nothing
- conclude that the best thing is to remove nothing
- stop there.

Normally, there would be something worth removing, and `step` would remove that, and then do another round to see what else can be removed, but here there is nothing more to do.

Extra 2: there are no P-values attached to any of this, because `step` uses AIC to decide what to do. If you want some, you have to explicitly fit models without the things you want to test, and then use `anova` to compare the fits (exactly what we did in lecture with the brand preference example to see whether `age` or `sex` had an effect). For these data, I found that the P-values are actually 0.004 for `Water` and 0.002 for `Work`, so both variables are strongly significant. The logic is that if `step` says not to keep an explanatory variable, it's not significant, but if it says to keep it, the P-value might be a bit bigger than 0.05. Or, to flip it around, if an explanatory variable is significant, `step` will say to keep it, so that when you use `step`, there is no danger of "losing" an important explanatory variable.

(10) (2 points) Some predictions are shown in Figure 8. What do you learn from these predictions? Explain briefly. (Hint: here, all the probabilities are close to 0 or 1, so a significant effect may appear to be small.)

This changes `Work` while holding `Water` constant. It says that manual workers have a higher probability of having a baby with any of the birth defects, and a lower probability of having none, all else equal (holding water hardness constant). You can be more specific about the birth defects if you wish, but the changes are similar in size for all of them.

This is a better conclusion than "the probabilities are unchanged", because of the hint, and also that `Work` is significant, so there must be *some* differences of note.

We don't have an interaction in this model, so there is *an* effect of work type that holds for any water hardness, which is what we are investigating here. The value 94 for water hardness is the mean value, the result of not specifying another value for `Water` in the `datagrid` line above the predictions in the Figure.

Points:

- 2: manual workers have a higher probability (than non-manual) of having a baby with any of the birth defects, and a lower probability of having a baby with none, for the same water hardness
- 1: the differences in probability are all small (true but not insightful enough)
- 1: some other relevant but not complete comment

Extra: the changes in probability in these predictions are very small, but because the probabilities are all close to 0 or 1, the *log-odds* of them change substantially, which is where the significance of the explanatory variables comes from. Another way to look at this is to look at *relative* changes in probability for the small ones. For the predictions here, being parents who are non-manual workers (vs. being manual workers) reduces the probability of anencephalitis by about a third, and reduces the probability of other birth defects by almost a half. (I didn't talk about this in class this time, but a ratio of (small) probabilities is what public health people call "relative risk".)

(11) (3 points) We are going to fit another model. The setup for the model-fitting is shown in Figure 9, and the model-fitting itself, along with the model output, is shown in Figure 10. What is being predicted here, and what does the output tell you? (All the code used to produce the output from the original data is shown in these two Figures.)

This is predicting the probability of *any* central nervous system birth defect (vs. none). The column `CNS` adds up the counts of all the actual birth defects (`An`, `Sp`, `Other`). Each row of `cns_yn` represents more than one baby, so we use a two-column response in fitting `cns.2`. The first column is `CNS`, so we are predicting the probability of *any* central nervous system birth defect. The output shows that both explanatory variables are significant, and therefore that the hardness of the water and the parents' work both predict whether or not there will be a birth defect of the central nervous system of some kind.

Additionally, we can see that the Estimates for the two explanatory variables are significantly negative, so:

- as water hardness increases, the probability of any birth defect decreases
- parents that are non-manual workers have a lower probability of having a baby with any birth defect, compared to parents that are manual workers. (This is very similar to the conclusion from the previous question, but does not differentiate between the types of birth defect.)

Points:

- 3: all of:
  - the model is fitting probability of any birth defect vs. none
  - as water hardness increases, prob of any birth defect decreases
  - non-manual workers have lower prob of a baby with any birth defect (or manual, higher)
- 2: what the model is fitting, plus saying that the two explanatory variables are significant (without saying what kind of effect they have)
- 2: as 3, saying what the model is fitting, but with only one of the two other comments
- 1.5: as 3, but without either of the comments about explanatory variables
- 1: incorrect description of what the model is fitting, plus a logical consequence of the description given that seems to follow from the output.

This is a three-point question, so the implication is that you need to say three things.

(12) (2 points) A third model is fitted in Figure 11, and `anova` output from this model, in comparison to another model as shown, is in Figure 12. What do you conclude from these two Figures? Explain briefly.

In Figure 11, we are eliminating all the babies without birth defects, so we are modelling the type of birth defect, *given that there is one*, from water hardness and parents' work. The `anova` output in Figure 12 says that in this case the model with both explanatory variables is not significantly better than the model with neither, and we should prefer the smaller model with neither (that is, just an intercept).

This looks a lot like the first model we fitted to these data, but you should be careful to see (and say) how it is different.

The conclusion from these two Figures is that *if there is a birth defect*, knowing the water hardness and the parents' work *does not* tell you what kind of birth defect it is.

Points:

- 2: both of: (i) modelling type of birth defect *given that there is one* from `Water` and `Work`; (ii) concluding that the model with neither explanatory variable should be preferred over the one with both.
- 1.5: as 2, but for the second point saying that there is no significant difference between the two models (without saying which one should be preferred)
- 1: concluding that the model with neither explanatory variable should be preferred over the one with both, without saying correctly what is being modelled
- 0.5: as 1, but saying only that there is no significant difference between the two models (without saying which one should be preferred)

(13) (2 points) Summarize your findings about the central nervous system birth defect data, based on what you have seen in the questions about these data, in *one* sentence.

The water hardness and parents' work help to predict whether there will be a central nervous system birth defect, but not what kind of birth defect it will be.

I am expecting you to find it challenging to put all this together. There is nothing new here in terms of model-fitting, but as a graduating applied statistician you ought to be able to see how the models are different and to summarize the conclusions briefly (which is why I said "in one sentence").

If you were unable to make sense of the second or third models, you will have trouble getting full marks here. Normally, I would try to account for getting earlier questions wrong, but if you don't get these questions, you are making it easier here and I cannot award full points for that.

Points:

- 2: The water hardness and parents' work help to predict whether there will be a central nervous system birth defect, but not what kind of birth defect it will be. Or something equivalent.
- 1.5: a reasonable summary, but more than one sentence
- 1: summary of some of the key ideas
- 0.5: some otherwise relevant comment.

Extra: this conclusion comes from models `cns.2` and `cns.3` (the significance in the first one and the non-significance in the second one of those), but the conclusion is consistent with the predictions from model `cns.1` as well. There, we found that the predicted probabilities of all the actual birth defects (`An`, `Sp`, and `Other`) changed in about the same way as the explanatory variable changed, so that the explanatory variable had no distinguishing effect on the type of birth defect. (I originally had another question asking about the effect of water hardness, and the picture there was the same as well, but that question did not survive the editing process.)

This dataset comes from Faraway's book Extending the Linear Model with R, starting on page 113. He investigates the models I called `cns.2` and `cns.3` and comes to the same conclusion I did. (He only fits my first model at the end, but I wanted to do a multinomial model where it was clear what was going on, so that you didn't lose your way for all the questions on these data.)

Glass fragments

A forensic scientist collected over 200 fragments of glass while investigating crimes. Each glass fragment was classified by where it came from (in column `type`), and the percentage of calcium oxide by weight was recorded (in column `Ca`). Some of the data are shown in Figure 13, and the abbreviations in `type` are described in Figure 14.

The forensic scientist is specifically interested in the following comparisons of percentage of calcium by weight:

1. window float glass vs. window non-float glass
2. the average of window float and non-float glass vs. vehicle window glass
3. the average of all three types of window glass vs. vehicle headlamp glass
4. the average of all types of window or headlamp glass vs. the average of containers and tableware
5. containers vs. tableware.

These comparisons will be numbered `c1` through `c5` below.

(Added afterwards) When you are doing a contrast, you are always comparing one mean (or group of means) against some other mean (or group of means). So the fourth contrast says that on one side you have "all types of window and/or headlamp glass" and on the other you have "containers and tableware". Maybe it would have been clearer with "and" instead of "or", but what is meant is "all types of glass that are either window glass or headlamp glass" for the one side of the comparison.

(14) (2 points) Why would an analysis based on contrasts be better than an analysis based on `aov` followed by Tukey?

The forensic scientist is interested in the five specific comparisons named above, not in comparisons between all possible pairs of glass types. Interest is in specific comparisons (without looking at the data), which suggests contrasts as the preferred analysis.

If you say that the scientist is interested in specific comparisons without saying which ones, you could have copied this from your notes without understanding anything.

Points:

- 2: interested in *specific* comparisons (namely, the ones listed in the question) and not in all possible pairs of glass types
- 1: interested in specific comparisons without saying which ones we are interested in *here*

(15) (2 points) What code would produce a vector encoding contrast #5 listed above?

Contrast #5 is comparing containers against tableware, which are the 4th and 5th types of glass listed in Figure 14. Hence the contrast should have 1 and -1 in the 4th and 5th places, with zeros elsewhere, such as

```
c5 <- c(0, 0, 0, 1, -1, 0)
```

Also acceptable: this vector multiplied through by any nonzero number (thus, for example, the 1 and -1 interchanged is also acceptable), but the zero values must be as shown. For the marker: look for the same non-zero number but with opposite signs in positions 4 and 5, and zeros elsewhere.

Technically, according to the instructions, this contrast should be named `c5`, but the important thing is to get the numerical values in the vector correct.

Points:

- 2: as in solutions, or equivalent
- 1: progress towards a solution without being correct

(16) (2 points) What code would produce a vector encoding contrast #4 listed above?

Contrast #4 is comparing the average of all types of window (float or non-float) or headlamp glass to the average of containers and tableware. That is, glass types 1, 2, 3, and 6 (four of them) vs. glass types 4 and 5 (two of them). That means you need a contrast like this:

```
c4 <- c(1/4, 1/4, 1/4, -1/2, -1/2, 1/4)
```

Again, you can multiply through by anything nonzero, so that as long as glass types 4 and 5 have double the weight and the opposite sign to glass types 1, 2, 3, and 6, you will be good.

Points:

- 2: as in solutions, or equivalent
- 1: progress towards a solution without being correct.

(17) (2 points) Determine whether the contrasts you constructed in the previous two questions are orthogonal.

Multiply the corresponding numbers together and add up the results (using your calculator if necessary):

$(0)(1/4)+(0)(1/4)+(0)(1/4)+(1)(-1/2)+(-1)(-1/2)+(0)(1/4) = 0+0+0+(-1/2)+(1/2)+0 = 0$.

One point for doing this calculation correctly (with your numbers). The second point for correctly using the result of your calculation to make a statement about orthogonality. My calculation gave zero, so my two contrasts are orthogonal.

Half a point for the calculation part if you give correct code to work out whether the two contrasts are orthogonal. Your aim here is to show whether the contrasts you wrote down are *actually* orthogonal, and you have everything you need to do that.

If you make a mistake in your calculation or in the definition of your contrasts, you can still get the second point if you use the result you got to say whether you think your contrasts are orthogonal or not. If you were unable to do one or both of the previous questions, write down what you would do to determine whether your contrasts were orthogonal. The next question uses these contrasts (and the others that I didn't ask you to write down) to do some tests, so the implication is that they *should* be orthogonal (we haven't tested non-orthogonal contrasts), so that is what you are aiming for, and if yours are not, you are invited to consider why not.

Points:

- 2: both of: (i) correct calculation to assess orthogonality (based on answers to two previous questions) (ii) statement about orthogonality based on calculation result
- 1.5: as 2, but giving correct code to do calculation instead of actually doing calculation
- 1: calculation incorrect, but correct conclusion about orthogonality based on result obtained
- 0.5: correct code to do calculation, but incorrect or incomplete statement about orthogonality

(18) (2 points) I constructed contrasts `c1` through `c5` to represent the five comparisons of interest (respectively), and ran the code shown in Figure 15, with the results shown. Interpret any relevant significant results in the Figure.

The test for the intercept is not relevant, so we ignore that. (This is actually saying that the overall mean percentage of calcium by weight is not zero, which is something the forensic scientist would expect to see and would not be especially interested in.)

The only other significant result is for contrast `c4`, shown as `typec4` because it is a contrast of types of glass. This is the second one you encoded above: there is a significant difference in percent of calcium by weight between window and headlamp glass on the one hand and container and tableware glass on the other.

None of the other differences are significant, and I haven't given you a plot or a table of means, so you have no evidence about which way the significant difference goes, only that there is one. So this is as much as you can say. (The sign of the Estimate depends on which numbers in my contrast were negative and which ones were positive, which you don't know, so it is a mistake to say which glass types had more calcium: the window-and-headlamp ones or the containers-and-tableware.)

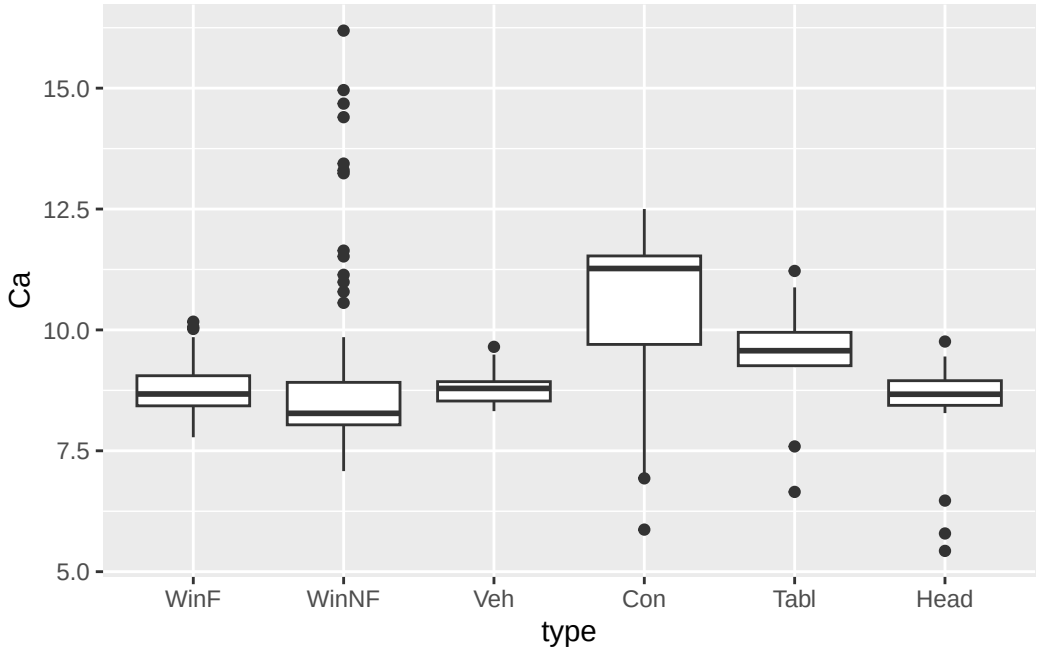
Points:

- 2: interpretation of significant `c4` in context of data
- 1.5: as 2 but attempted interpretation of intercept
- 1: assertion that `c4` is significant, but incomplete or incorrect interpretation of what that means for data
- 0.5: any other potentially relevant comment

Extra:

I can make a boxplot to see what is actually going on:

```
ggplot(glass, aes(x = type, y = Ca)) + geom_boxplot()
```



It looks as if we have a lot of problems with outliers. I didn't want to get into that, so I didn't show you this graph before (and besides, we have no way to do contrasts with medians).

Leaving the outliers aside, it looks as if the window, vehicle and headlamp glass all have about the same median percentage of calcium by weight, while the glass in containers and tableware has a higher percent of calcium by weight.

So, all the contrasts that are *within* all window and vehicle headlamp glass should not be significant (contrasts `c1`, `c2`, `c3`) and the contrast between the window-and-headlamp glass and the other types of glass *should* be significant (contrast `c4`). That leaves contrast `c5`, between containers and tableware. These look different on the boxplot, but there were actually only 13 and 9 observations of these types respectively:

```
glass %>% count(type)
```

type	n
WinF	70
WinNF	76
Veh	17
Con	13

type	n
Tabl	9
Head	29

with the other types of glass having larger sample sizes. This is, after all, observational data, so you get the sample sizes you get. Along with the large amount of variability that you see on the boxplot, this is probably why contrast `c5` did not come out significant.

Cadmium

Cadmium is a silvery-white metal that is chemically similar to zinc and mercury. It is toxic, and so workers exposed to cadmium over time may suffer health effects. One suspected health effect is a decrease in lung volume. Data were collected from 84 workers that might have been exposed to cadmium in their work. For each worker, their exposure was rated as "none", "low", or "high"; their age was recorded, and they did a test of lung volume resulting in the number recorded in `vital.capacity` (higher number means greater lung volume). One observation was collected for each worker. Some of the data are shown in Figure 16.

(19) (2 points) A graph is shown in Figure 17. Using the information in this graph, how can you argue that there may be a significant interaction between age and exposure level?

In the Figure, the red line, for high exposure, appears to descend more quickly than the other two lines. That is to say, the slopes appear to not all be the same, and that would result in a significant interaction between age and exposure.

It is not enough to say "the slopes are not all equal"; talk about how you know.

Points:

- 2: the slopes are not all equal because the high-exposure (red) one goes down faster (is more negative)
- 1: the slopes are not all equal without a clear enough explanation of how you know.

Extra: in this question and the next one, you might have a strong belief one way or the other about the significance of the interaction. That you have to set aside in order to answer both questions. The issue is "can you make an argument that is persuasive enough to someone else?" It is like debate club in that regard: it has to be a convincing argument, not necessarily one that you actually believe.

(20) (2 points) Again using Figure 17, how can you argue that there *may not* be a significant interaction between age and exposure? Hint: this may seem to be in direct contradiction to the previous question, but note that these questions are *not* asking what you believe, but instead to come up with an argument one way or the other.

There appears to be a lot of variability around the lines. For example, the blue points (no exposure) can be a long way above or below the blue line. Thus, the slopes are not accurately estimated, and so the differences observed in the slopes might just be chance.

Points:

- 2: there is a lot of variability (points far from the lines), so none of the slopes might be significantly different.
- 1: as 2, but not clear enough

(21) (3 points) A model is fitted in Figure 18. Some output from the model is shown in Figure 19, and some more output is shown in Figure 20. Can we remove the interaction between age and exposure from the model? Explain briefly. Your answer should clearly state which Figure(s) you are basing your answer on, and how you are using those Figure(s).

To decide whether we can remove the interaction term from the model, we need to test it for significance (obtaining one overall P-value). The categorical variable `exposure` has three levels, so we *must* use the `drop1` output in Figure 19 to test the overall significance of the interaction. The P-value of 0.034 is significant (less than 0.05, if only just), so the interaction must be kept and we cannot drop it.

It is an error to use Figure 20 here, because this compares the slope of the line for each exposure shown (in the bottom two lines of the table of Coefficients) to the slope of the line for the baseline exposure `high`, rather than assessing an overall effect of the interaction. (Note that in Figure 19, the interaction term has 2 degrees of freedom, corresponding to the three levels all being compared together.) You might use Figure 20 to help understand *why* the interaction is significant (that is, which slopes are different from which), but that's not the issue in this question: the interaction as a whole is significant, so it needs to be kept in the model as a whole (you cannot remove "part" of an interaction), and only Figure 19 will give you an answer for that.

Points:

- 3: all of: (i) use `drop1` in Figure 19 to test the whole interaction term; (ii) P-value is 0.034, smaller than 0.05, so conclude that interaction term is significant; (iii) need to keep it in the model (cannot remove it).
- 2: as 3, but without citing P-value

- 2: as 3, but without stating a conclusion about dropping the term from the model
- 2: as 3, but using Figure 20 as well
- 1: trying to draw a reasonable conclusion using only Figure 20

Extra: the fact that the P-value for the interaction is only just significant is consistent with the two earlier questions about Figure 17: on those, you could make plausible arguments either for the significance or the non-significance of the interaction (based on the slopes of the lines). Here, you see that the "significant" side wins, but not overwhelmingly. If you were using $\alpha = 0.01$, the "non-significant" side would win.

(22) (2 points) Look at the slopes of the lines in Figure 17. How are the relevant Estimate values in Figure 20 consistent with that feature of the graph? (If you think Figure 20 is *inconsistent* with the graph, say how that is.)

In Figure 20, the Estimates that relate to slopes are the two at the bottom, `age:exposurelow` and `age:exposurenone`. These say how the slopes of the lines for each exposure group compare to the slope of the line for the baseline group `high` (as you can infer (i) because there is no `age:exposurehigh` in Figure 20, it being zero, and (ii) because `high` is alphabetically before `low` and `none`).

Both of these numbers in the Estimate column are positive. This means that the slopes of the lines for `low` and `none` are bigger than for `high`, by respectively 0.039 and 0.055. Since all the slopes are negative, "bigger" actually means "less negative" (which you need to say): that is to say, all the trends with age go downhill, but the ones for exposure `low` and `none` go downhill less fast than the one for exposure `high`.

You will need to be careful in arguing from the positive Estimates to the *less* negative slopes.

This is a better answer than "the two Estimates are close to zero" (and therefore that all three slopes are about the same, which is inconsistent with the graph) because the Estimates, though they are all small, are not all the same (in the sense that the null that they are all equal was rejected), as Figure 19 said.

Points:

- 2: both of: (i) the two slopes shown are bigger than the one for `high`; (ii) this means that they are *less negative* than the one for `high`
- 1: the slopes for `low` and `none` are bigger than the one for `high` without properly connecting that with the steeper (but negative) slope for high
- 1: the two Estimates shown are close to zero, and therefore all three slopes are about the same.

Extra: As it happens, the slopes for `high` and `none` are significantly different (P-value 0.0116) but the slopes for `high` and `low` are not (P-value 0.1014). My guess is that there are a lot of observations with no exposure to cadmium (many blue points on the graph), so it was easier to demonstrate a significant difference in slope between `high` and `none`, but there are relatively fewer with low exposure, so it was relatively more difficult to demonstrate a significant difference between `high` and `low`.

Had this been a designed experiment, we would have made sure to have the same number of people in each exposure group, but of course you cannot ethically randomly assign people to experience a chosen level of exposure to something toxic!

Extra extra: something else I could have asked you about is that all of the lines go downhill, meaning that as a worker gets older, regardless of the level of exposure, their lung capacity gets smaller. (This is not really a surprise.) The suspicion is, from the graph, that this decrease happens more quickly, or at younger ages, if the worker has had a high level of exposure to cadmium.

The other thing to say is that the only way to get to a high level of exposure is to work for a long time in an industry that uses cadmium: that is, to be older in the first place. Nobody in our data set with a high level of exposure was younger than about 39 years old, but there were plenty of workers with no exposure, and some with low exposure, who were younger than that.

Heat and cognition

Forty-six college students were asked to solve cognitive problems first thing in the morning during a heat wave in their Northeastern (US) city. Twenty of the students had air-conditioning in their rooms and twenty-six did not. For each student, it was recorded whether or not the student had air-conditioning or not (in column AC), and also their times taken to solve a number of math problems (in column `MathZRT`) and a number of colour dissonance problems (in column `ColorsZRT`). The times have been expressed as z-scores, so that a very negative value is a fast time and a very positive value is a slow time. Some of the data, in dataframe `heat`, are shown in Figure 21.

(23) (2 points) Some analysis is shown in Figure 22. Why is this an appropriate analysis to use for these data? Explain briefly, showing that you know what kind of analysis it is.

It's a multivariate analysis of variance (which is rather a giveaway given that the code uses `manova`). This is appropriate when we have more than one quantitative response variable, and at least one categorical explanatory variable. In this case, whether or not a student has air-conditioning in their room is explanatory (categorical), and we might

expect that variable to affect how fast they can solve the two types of problem (`MathZRT` and `ColorsZRT` are response variables), so we do indeed have more than one quantitative response variable.

Points:

- 2: manova; two quantitative responses that you name
- 1: manova; two quantitative responses but not named or otherwise indicated
- 0.5: manova but no or incomplete indication of why

(24) (2 points) What do you conclude from Figure 22?

The P-value shown is 0.033, which is less than 0.05, so we conclude that whether or not a student's room has air-conditioning has an effect on the time taken to solve math problems or colour dissonance problems or both.

That's all we can say: we have no idea what kind of effect, or whether it's one response variable or both. All we can say is that there is some kind of effect, and it is wrong at this point to say more.

Points:

- 2: give P-value; effect of AC on one or both of time taken to solve different types of problems
- 1.5: as 2, but without giving P-value
- 1: as 2, but trying to be more specific than warranted
- 0.5: some progress towards answer but not enough for 1

Extra: the P-value is small, but not *very* small. This suggests that whatever effect there is may not be very convincing.

(25) (2 points) A discriminant analysis is shown in Figure 23. According to this Figure, what values on the original variables would result in a student having a large (positive) score on LD1? Explain briefly.

Look at the Coefficients of Linear Discriminants. These are both negative, so the way to get a large (positive) score on LD1 is to have small (negative) values on both `MathZRT` and `ColorsZRT`.

- 2: small (negative) values on both because (both) coefficients negative
- 1: as 2 but with not clear enough reasoning
- 1: right reasoning but error in logic

Extra: That is to say, a student who solves these problems quickly will have a large positive score on LD1. This doesn't say yet what the relationship with air-conditioning is, but if you look at Group Means in the Figure, you can see that students with air-conditioning are quicker than average compared to students who don't (who are slower than average), for both kinds of problems, but the difference is especially noticeable for the colour dissonance problems. An advantage to our data being z-scores is that we know both groups have the same spread (variance), so that a bigger difference in group means for the colour dissonance test actually *is* because the effect is bigger there, not just because times on the colour dissonance test are more variable.

(26) (2 points) Figure 24 shows some further analysis, and Figure 25 shows some of the output from that analysis. Which is the first student in Figure 25 to be misclassified? Explain briefly. Use the number at the left of the row to identify the student.

Look for the first row where `AC` (actual AC status) and `class` (predicted AC status) are different. This is the second row, student number 38, who actually did not have AC in their room, but was predicted to have it.

There are other misclassified students, but I asked you to find the first one, to make the question easier to mark correctly.

Points:

- 2: student number 38 (or one on second row) because actual AC status (`AC`) and predicted (`class`) are different
- 1: student number 38 (or one on second row) with incomplete or absent explanation

Extra: this student was slightly slower than average at solving the math problems, but noticeably quicker than average at solving the colour problems (the two `ZRT` columns being z-scores makes it easier to deduce this). They also have a clearly positive LD1 score. This pattern is more characteristic of a student that *does* have AC in their room. See the graphs in an Extra below.

(27) (2 points) For the student you found in the previous question, would you describe the misclassification as a serious error or a close call? Explain briefly.

To assess this, look at the posterior probabilities for this student. These are 0.30 (for No) and 0.70 (for Yes). These are not very close to 50–50, so this was a serious error in my opinion. (If you can make a convincing case that it is still close to 50–50, go ahead, but you are likely to find that challenging.)

Points:

- 2: look at posterior probabilities and make a well-reasoned call about whether they are close to 50–50 (close call) or not (serious error).
- 1: apparently reasonable decision made on some other basis, or unclear explanation

Extra: these are a randomly-chosen collection of students, set with a random number seed so that when I rebuild this exam, I get the same list of students in the same order.

(28) (2 points) Starting from the results obtained in Figure 24, what code would tell you, for students who actually had air-conditioning in their rooms and for those who did not, how many of them were predicted to have air-conditioning in their rooms or not?

Count up the numbers of students in each combination of `AC` and `class`, using what I called `heat_p` to get them from. There are two ways, the base R `table` way, which is the easiest way to get a nice table (you won't have the actual results, of course):

```
with(heat_p, table(AC, class))
```

```
     class
AC    No Yes
  No  21   5
  Yes  8  12
```

I like having the actual AC status in the rows, because then you can say "out of the students who actually had AC, how many of them were predicted to have AC?" and so on.

Also good is the `tidyverse` way, which is a straightforward `count`:

```
heat_p %>% count(AC, class)
```

AC	class	n
No	No	21
No	Yes	5
Yes	No	8
Yes	Yes	12

I would accept this on an exam, but to make a nice table like the output from `table` I would go a step further:

```
heat_p %>%
  count(AC, class) %>%
  pivot_wider(names_from = class, values_from = n)
```

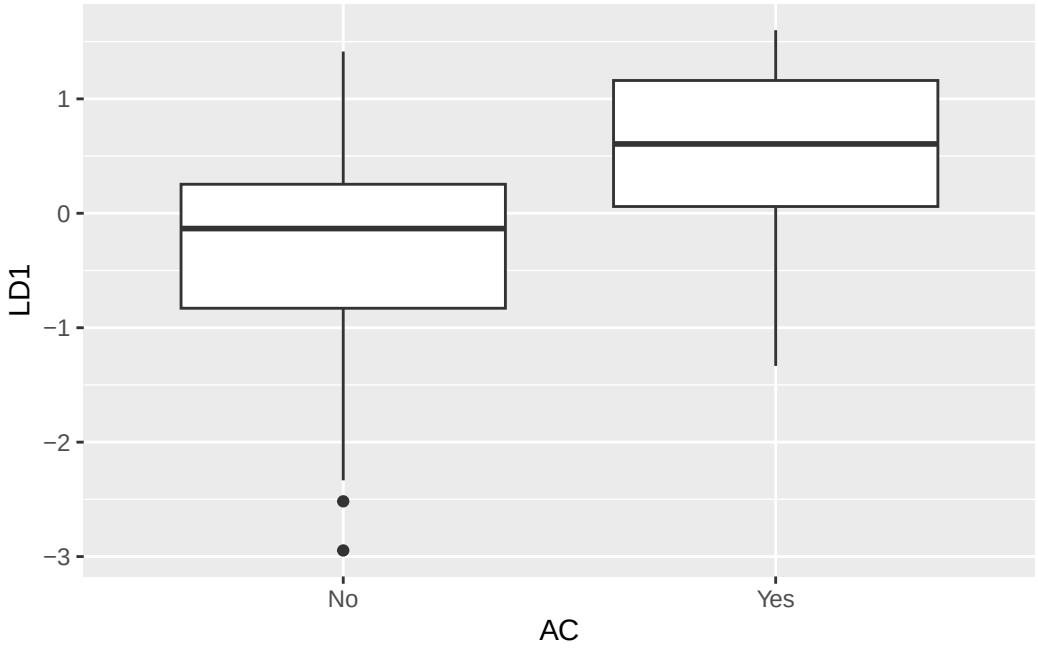
AC	No	Yes
No	21	5
Yes	8	12

Points:

- 2: table correctly made using `table` or `count` approaches
- 1: a reasonable approach but with error(s)

Extra: you see that $21 + 12 = 33$ out of the 46 students were correctly predicted, and only 13 were gotten wrong. This looks not too bad, if perhaps not great. With only one LD, it's hard to make a good graph like the ones we have seen. One option is a boxplot of LD scores vs. actual AC status:

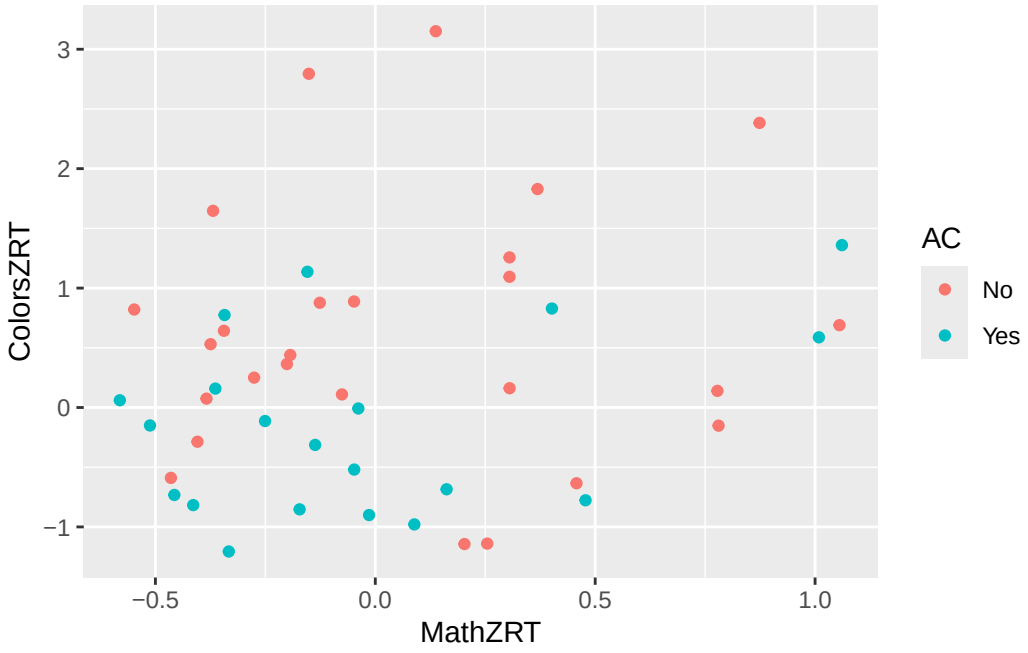
```
ggplot(heat_p, aes(x = AC, y = LD1)) + geom_boxplot()
```



This shows that students who actually had air-conditioning tended to have a higher score on LD1, but there is a lot of overlap. This is the sort of thing that goes with the kind of misclassification rates we saw. (Based on our earlier work, it also says that being quicker to solve the problems was positively associated (somewhat) with having AC in the room.)

Another graph is to realize that we have two original variables (as for the seed yield and seed weight example in class), so we can plot them against each other coloured by AC status:

```
ggplot(heat_p, aes(x = MathZRT, y = ColorsZRT, colour = AC)) + geom_point()
```

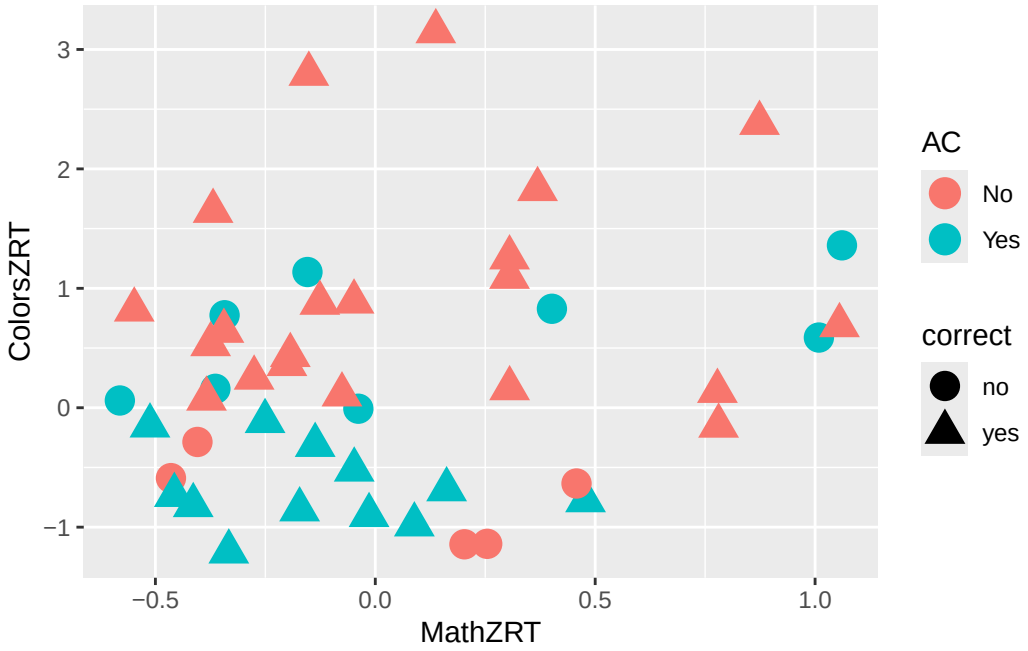


This is rather a long way from being clear-cut. A *slight* majority of the blue points are at the bottom left of that plot, with most of the points at the top (especially) being red ones. This suggests that students with AC in their rooms tended to solve both kinds of problems more quickly. Students without AC were sometimes just as quick, but sometimes, especially on the colour dissonance problems, they were a lot slower.

Here's an embellishment of that plot:

```
heat_p %>%
  mutate(correct = ifelse(AC == class, "yes", "no")) %>%
```

```
ggplot(aes(x = MathZRT, y = ColorsZRT, colour = AC, shape = correct)) +
geom_point(size = 5)
```



The triangles are correctly classified, and the circles are misclassified. (I made the plot points bigger[1] so that you could see the shapes of the points more clearly.) The pattern, such as it is, is that the correctly-classified students with AC are at the bottom left (fast), and the correctly-classified students without are not (slow on at least one, although not usually slow on both).

In other words, the effect is sort-of what you would have guessed: that not having AC is associated, at least in part, with students being slower to solve the problems.

Beat the Blues

52 people with depression ("subjects") took part in a study of a new treatment called Beat the Blues. Each subject was randomized to either the Beat the Blues treatment (`BtheB` in

[1]There is a trick here: I wanted *all* the points to be the same amount bigger, so I put the `size` inside the `geom_point` and *not* inside an `aes`. If I had wanted the size to be different according to one of my variables, I would have put it inside the `aes`, but since the size is not connected to a variable, I put it outside the `aes`.

the data) or to a standard treatment called `TAU`. After each of 2, 4, 6, and 8 months on their treatment, each subject's depression was assessed using the Beck Depression Inventory, with a lower score indicating that the depression was having less of an effect on the subject's daily life (that is, a lower score is better). Some of the data, in dataframe `blues`, are shown in Figure 26.

(29) (2 points) Why would it *not* be appropriate to analyze these data by rearranging the dataframe as shown in Figure 27, and then using `aov(beck ~ treatment * time, data = blues_long)`?

This code assumes that each observation is independent, and they are not because there are four observations from the same subject (which will be correlated). This is evident because in Figure 27, the same `subject` appears on several lines (one for each time point), which ought to be a clue that something is amiss.

Another way to say this is that for an `aov` analysis we need *one* observation per subject, and Figure 27 indicates that we actually have *four* observations per subject, one at each of the four time points.

Points:

- 2: requires that obs are independent, and why they are not. Or, requires one obs per subject and we have four (one each at the four times).
- 1: right conclusion but incomplete or incorrect reasoning.

Extras:

- The treatment `TAU` actually stands for "treatment as usual"! That is, it really *is* a control in this context.
- There was originally also a measurement at time zero (before the treatment had a chance to start working), but including time zero makes the interaction between treatment and time (later) almost significant, so I omitted time zero completely, as for the second go at the dogs data in lecture.
- Compare the kind of data we needed in C32 for a two-sample test and for matched pairs. For a two-sample test we needed one observation per subject/individual ("two independent samples"), but for matched pairs we had two observations per individual, such as a before and after measurement. Matched pairs is really also repeated measures, but we avoided that in C32 by taking differences to give us only *one* difference per individual, and then we could do an ordinary one-sample test.

(30) (4 points) Code for an analysis is shown in Figure 28, with output from `summary(blues.2)` shown in Figure 29. Give a complete interpretation of the results from this analysis, as far as Figure 29 permits. If you use a P-value, make it

clear where you got that P-value from. If you need any output that is not shown in Figure 29, explain what output you need and why.

First, check sphericity. This is not rejected (P-value 0.81, from Sphericity Tests, both for the interaction and the time effect).

This means that we use the univariate type II tests (top table in the Figure) the rest of the way, and do not use the Huynh-Feldt adjusted P-values at all. (These are for when sphericity fails.)

Second, test the interaction between treatment and time (from the top table). This, with P-value 0.14, is not significant. Thus, the effect of treatment is the same over all times (or, the effect of time is the same for both treatments) and it makes sense to look at main effects.

Third, *in this kind of analysis only*, you are allowed to ignore the non-significant interaction and immediately look at the main effects (also in the top table). This is because we have no way in this kind of model to remove anything involving time:

- the two treatments differ significantly (P-value 0.0057)
- there is a (strongly) significant effect of time (P-value 0.00027).

We are not able to say what *kind* of treatment or time effects there are on the basis of this Figure, so it is a mistake to go further at this point.

Points:

- 4: correctly check sphericity and make it clear that type II table will be used; test interaction (not rejected); test main effects and conclude sig. effect of time and of treatment.
- 3: as 4 but not making it clear whether type II or Huynh-Feldt P-values are being used.
- 2: as 4 but not checking sphericity first
- 2: as 4 but not assessing main effects
- 2: as 4 but not assessing interaction
- 1: reasonable ideas but not enough for 2

Extra: you will see that the Huynh-Feldt P-values, for effects involving time, seem to be exactly the same as the ones in the univariate type II tests. This is because, when there is a problem with sphericity, the `HF eps` in the second-last column of the P-value adjustments table will be less than 1 and the P-values will be adjusted upwards. When `HF eps` is bigger than 1, as here, there is definitely no problem with sphericity, and there are no adjustments. So if you use the bottom table, you will in this case get the right P-values *for the wrong reason*, namely that you should not even be looking at these P-values. This is why it is

important to say that you are using the univariate type II tests for everything, and not using the Huynh-Feldt adjusted P-values at all. On the other hand, for the Greenhouse-Geisser tests (that we are never using), the `GG eps` is a little less than 1, and so the GG P-values are a little bigger than the ones in the Univariate Type II tests, even though the difference is very small.

(31) (3 points) A spaghetti plot and an interaction plot are shown in Figure 30 and Figure 31 (respectively). Use either or both of these plots to explain the significance or non-significance of any effects you examined in the previous question, and to interpret any significant effects. Be clear about which plot(s) you are using and how.

There are three effects to assess: the non-significant interaction, and the significant time and treatment effects.

- The interaction term was not significant. We would expect the two lines on the interaction plot to be parallel, which they are not quite, but the spaghetti plot indicates that there is a lot of variability, so the two lines are not significantly different from being parallel.
- Both lines on the interaction plot go downhill, so the significant time effect is that depression scores are decreasing over time (that is, depression is on average improving over time) regardless of treatment. (This is also suggested by the spaghetti plot, but not very clearly.)
- The red line on the interaction plot is above the blue one, so the significant treatment effect is that depression scores on the new treatment are lower (better) than on the standard treatment, at all times. That is to say, the new treatment really is better. (On the spaghetti plot, most of the red traces are above the blue ones, and the pattern is a bit clearer than for the time effect.)

If you *only* assessed the interaction in the previous question, you get 2 marks here for correctly using both plots to explain its non-significance (not full marks because you have made this question easier by only giving yourself one thing to assess).

Points:

- 3: assess interaction using both plots, main effects using interaction plot
- 2.5: as 3, but not using spaghetti plot to assess whether lines are really parallel
- 2: correctly assessing two of the three effects but not the third
- 2: correctly assessing interaction *if* that was the only effect discussed in previous question
- 1: correctly assessing one of the three effects
- 0.5: some reasonable comment but not enough for 1

Farmland in Ohio

For 15 counties in Ohio (US), the percentage of farmland devoted to various crops was recorded, as shown in Figure 32. The first column shows the name of the county, and the remaining columns are the percentage of farmland devoted to the crop named in the column, with the column `mixed` denoting "mixed small grains".

(32) (2 points) A K-means cluster analysis is run as shown in Figure 33, and the counties in each cluster are shown in Figure 34. Why does the cluster membership of Adams and Warren counties make sense, based on what you have seen so far about these data?

In Figure 34, Adams and Warren counties are both in cluster 2: that is to say, they are both in the same cluster. They should therefore be similar in terms of what crops are grown there. Look back at Figure 32 and scan along the rows for these two counties:

- in Adams county, the highest percentages are for corn, wheat, and hay (and the lowest are for mixed grains and barley)
- in Warren county, the highest percentages are also for corn, wheat, and hay (and the lowest are also for mixed grains and barley).

These patterns are very similar, and so it is no surprise that they ended up in the same cluster. The actual numbers are not so close, but they are sufficiently similar to make it plausible that the two counties would end up in the same cluster.

Another way to go about this is to see from Figure 33 that cluster 2 has a high percentage of corn and wheat (and a low percentage for oats), and both Adams and Warren counties have these (and so it is no surprise that they ended up in cluster 2 rather than one of the other clusters).

Points:

- 2: Adams and Warren in same cluster and finding a convincing way to say how they are similar
- 1: Adams and Warren in same cluster without correctly going further

Extra: If you want to calibrate "similar", try two counties that are in different clusters, such as Washington and Putnam. Putnam is much higher on oats and soy, and Washington is much higher on hay. These seem much less similar than Adams and Warren counties.

(33) (2 points) A plot is shown in Figure 35, with the code used to make the plot above the plot. What do you conclude from this plot? Explain briefly.

This is a scree plot, to help determine a good number of clusters to use. A good number of clusters is where there is a downward-facing elbow reasonably far down the mountain. The best elbow I see is at 6 clusters, so that is a good number of clusters to use. You are looking to make a call about where the "scree" starts and the "mountain" ends, which is admittedly subjective.

There are only 15 counties, so even 6 is a largish number of clusters, but fewer clusters than that is too far "up the mountain" and you will end up with dissimilar counties in the same small number of clusters. You might also be able to make a successful case for 4 or 5 clusters (these are both elbows), but in my opinion 4 clusters (especially) is on the mountain rather than the scree, meaning that the counties within those four clusters may not be very similar. (In the earlier question about Adams and Warren counties, I deliberately picked two counties that seemed to be similar so that the logic of the question would make sense.)

Points:

- 2: finding an elbow far enough down the mountain, eg. 6 clusters
- 1: finding an elbow too far up (or down) the mountain

(34) (3 points) Some further analysis is shown in Figure 36. This uses a number of clusters that may or may not be suggested by your previous work. What, if anything, does the output tell you about the cluster membership of Adams and Warren counties based on the analysis in `ohio.6`? Explain briefly. Your answer should make it clear that you know what the code in Figure 36 is doing.

The code is doing this (say enough of this to make it clear that you know what is going on):

- run a K-means cluster analysis with 6 clusters
- make a dataframe with the county names, and the cluster membership for 3 and 6 clusters
- make a cross-tabulation of how many counties were in each cluster in the three-cluster solution and the six-cluster solution.

Adams and Warren counties were in cluster 2 in the three-cluster solution. The table tells us that all 4 of the counties in cluster 2 in the three-cluster solution ended up in cluster 4 in the six-cluster solution. Hence, Adams and Warren counties must *still* be in the same cluster, that is, cluster 4, now.

Points:

- 3: a clear enough description of what code is doing, plus reasoned argument for which clusters the two counties must now be in

- 2: a clear enough description of what code is doing without successfully going further
- 1: some relevant comment but not enough for 2 points

Diagnosing multiple sclerosis

Two random samples of patients, one group from Winnipeg and one from New Orleans, were each tested by two neurologists, who we will call A and B. Each neurologist was asked to assess how likely each patient was to have multiple sclerosis, on a scale "Certain", "Probable", "Possible", "Doubtful", in order from most to least likely. Each patient was assessed by both neurologists. Some of the data, in dataframe `ms_patients`, are shown in Figure 37. The columns in the dataframe are respectively: the diagnosis by neurologist A, the diagnosis by neurologist B, where the patient is from, and the number of times that combination of categories was observed. We are interested in any associations between the categorical variables in these data.

(35) (2 points) A log-linear model is fitted in Figure 38. What specifically would you do next? Explain briefly.

The three-way interaction is not significant, so remove it and fit a model without it.

The place to get to is "fit a model without the non-significant three-way interaction"; "remove the non-significant interaction" is not enough of an answer by itself, because it doesn't say what to do next (fit the model without it).

Points:

- 2: remove three-way interaction and fit model without it
- 1: remove three-way interaction without saying what to do next
- 0.5: other relevant comment but not enough for 1 pt

(36) (2 points) After some further model-building, we arrive at the output shown in Figure 39. The output shows what is still in the model. What does this output tell us, in the context of the data?

The two interaction terms shown are (strongly) significant. That means that there are associations between:

- the diagnoses of the two neurologists A and B
- the diagnosis of neurologist B and where the patient was from.

What *kind* of association, we don't know yet (that's the purpose of the graphs coming up).

These are "interactions" in the model, but what that actually *means* for anyone using the model is that the categorical variables in question are *associated*, so that's what the output is actually telling us.

Points:

- 2: associations between: (i) diagnoses of the two neurologists A and B; (ii) diagnosis of neurologist B and where the patient was from.
- 1.5: as 2, but using the word "interaction" instead of "association"
- 1: correctly noting one of the two associations
- 0.5: other relevant comment

(37) (2 points) Based on what you know or can guess about this scenario, explain briefly why there is something surprising shown in Figure 39, and something that is not surprising at all.

The surprising thing is that the diagnosis for neurologist B is associated with where the patient was from (and, by inference, the diagnosis for neurologist A was not, because that association is not in the final model). It seems strange that the diagnosis should have anything to do with where the patient is from, or that it should for one neurologist and not for the other. (I carefully said in the data description that these were "samples" of patients, so they ought to be similar unless there is really a population difference.) You can also reasonably suggest that neurologist B is "biased" in that their diagnosis seems to depend on information they would not normally be using.

The not-surprising thing is that there is an association between the diagnoses for the two neurologists A and B. They are both neurologists and they are both assessing the same patients, so they should agree at least somewhat on the diagnoses. (A significant association means that if you know the diagnosis from one neurologist, you know something about the diagnosis from the other. They don't have to agree to have an association, but here it seems likely that they would agree, something that will be supported by Figure 41 when you get to it.)

I was prepared to allow some latitude here, as long as what you said here was consistent with what you said above and at least somewhat relevant.

Points:

- 2: convincingly identifying something surprising and something not
- 1: convincingly identifying only one of the two
- 0.5: other relevant comment

(38) (2 points) Interpret Figure 40.

When a patient is from Winnipeg, neurologist B is more likely to diagnose them as a Certain case (and less likely to diagnose them as a Probable case) than if they are from New Orleans.

This is the significant association between neurologist B's diagnosis and where the patient is from, that we said was surprising above. (If you couldn't identify something surprising above, you might try this question first and see whether that helps you with the earlier question.)

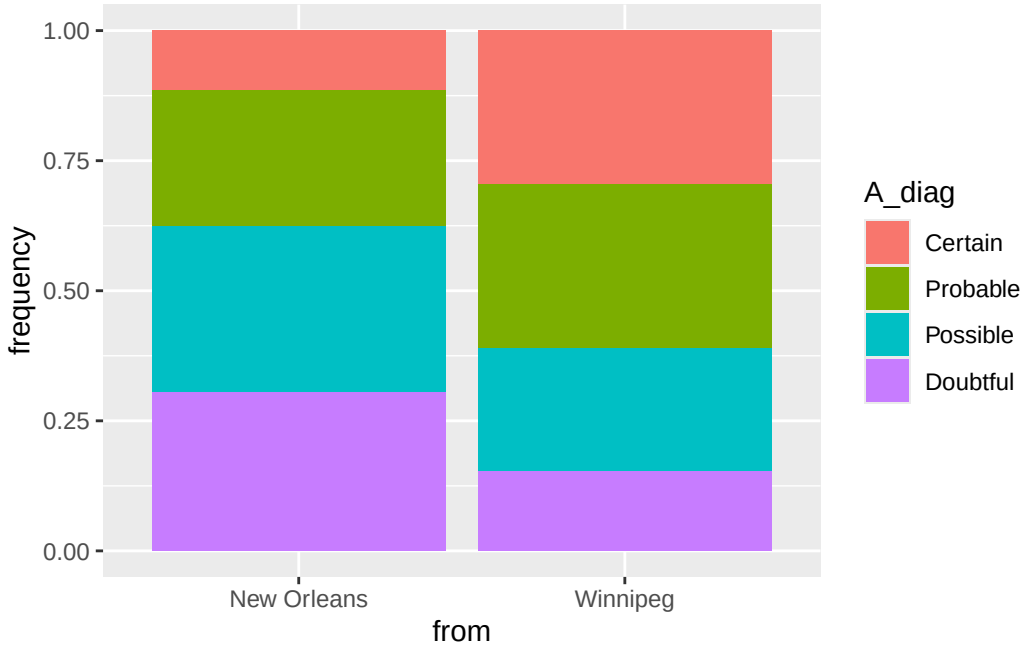
You need to mention that this is neurologist B (that is, you need to decode `B_diag` on the output).

Points:

- 2: proper comparison of diagnoses between patients from the two different places
- 1: as 2, but explanation unclear or confused
- 0.5: other relevant comments

Extra: I said that there is no such association for neurologist A, inferred from the `diag_A:from` interaction no longer being in the model (it was actually nowhere near significant). This is how the corresponding graph looks:

```
ggplot(ms_patients, aes(x = from, y = frequency, fill = A_diag)) +
  geom_col(position = "fill")
```



There is a bit of a difference here, but (evidently) not enough to be significant. It seems that the patients from Winnipeg really are more likely to be Certain cases than the ones from New Orleans (as judged by both neurologists), but the association is only significant for neurologist B.

(39) (2 points) Interpret Figure 41.

This says that the neurologists sort of agree:

- when A says "Certain", B tends to do so as well
- when A says "Probable", B says "Certain" or "Probable"
- when A says "Possible", B usually says "Probable"
- when A says "Doubtful", so does B usually.

For you, pick a couple of these, and come to an overall conclusion that the two neurologists tend to agree, or that they tend to pick similar points in the scale, or that B is calibrated further up the scale (more likely to say Certain or Probable than A is), or some similar relevant comment. You are allowed to say that the two neurologists disagree sometimes if you can support that from the Figure, for example when A says Probable or Possible, B is further up the scale than that. But I think you need to say that the neurologists agree at least sometimes (eg on Certain or, usually, on Doubtful).

Points:

- 2: successfully explaining how the neurologists sort-of agree (or how they sort-of don't), using at least two categories to make the point
- 1: as 2 but explanation unclear or incomplete (eg. compares only one category)
- 0.5: other relevant comment

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

Figures

```
library(tidyverse)
library(marginaleffects)
library(nnet)
library(MASS, exclude = "select")
library(car)
```

Figure 1: Packages loaded

```
budworm
```

ldose	sex	numalive	numdead
0	M	19	1
1	M	16	4
2	M	11	9
3	M	7	13
4	M	2	18
5	M	0	20
0	F	20	0
1	F	18	2
2	F	14	6
3	F	10	10
4	F	8	12
5	F	4	16

Figure 2: Budworm data (all)

```r
budworm.1 <- glm(cbind(numalive, numdead) ~ ldose + sex,
                 family = "binomial", data = budworm)
summary(budworm.1)
```

```
Call:
glm(formula = cbind(numalive, numdead) ~ ldose + sex, family = "binomial",
    data = budworm)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.4732     0.4685   7.413 1.23e-13 ***
ldose        -1.0642     0.1311  -8.119 4.70e-16 ***
sexM         -1.1007     0.3558  -3.093  0.00198 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124.8756  on 11  degrees of freedom
Residual deviance:   6.7571  on  9  degrees of freedom
AIC: 42.867

Number of Fisher Scoring iterations: 4
```

Figure 3: Budworm data logistic regression

```
cns0
```

Area	NoCNS	An	Sp	Other	Water	Work
Cardiff	4091	5	9	5	110	NonManual
Newport	1515	1	7	0	100	NonManual
Swansea	2394	9	5	0	95	NonManual
GlamorganE	3163	9	14	3	42	NonManual
GlamorganW	1979	5	10	1	39	NonManual
GlamorganC	4838	11	12	2	161	NonManual
MonmouthV	2362	6	8	4	83	NonManual
MonmouthOther	1604	3	6	0	122	NonManual
Cardiff	9424	31	33	14	110	Manual
Newport	4610	3	15	6	100	Manual
Swansea	5526	19	30	4	95	Manual
GlamorganE	13217	55	71	19	42	Manual
GlamorganW	8195	30	44	10	39	Manual
GlamorganC	7803	25	28	12	161	Manual
MonmouthV	9962	36	37	13	83	Manual
MonmouthOther	3172	8	13	3	122	Manual

Figure 4: Birth defect data

```
cns0 %>%
  pivot_longer(NoCNS:Other, names_to = "birth_type", values_to = "count") -> cns
cns %>% slice_sample(n = 15)
```

Area	Water	Work	birth_type	count
GlamorganC	161	NonManual	NoCNS	4838
GlamorganE	42	Manual	An	55
GlamorganW	39	NonManual	Sp	10
Cardiff	110	NonManual	Sp	9
MonmouthV	83	Manual	An	36
MonmouthOther	122	Manual	Other	3
MonmouthV	83	Manual	Sp	37
Swansea	95	NonManual	An	9
Newport	100	NonManual	Other	0
MonmouthOther	122	NonManual	Other	0
Swansea	95	Manual	Sp	30
GlamorganE	42	NonManual	Sp	14
Swansea	95	NonManual	NoCNS	2394
Cardiff	110	NonManual	Other	5
Newport	100	NonManual	An	1

Figure 5: Creating dataframe `cns` from `cns0` (randomly chosen rows of result shown)

```
cns.1 <- multinom(birth_type ~ Water + Work, data = cns, weights = count)
```

```
# weights:  16 (9 variable)
initial  value 117209.801938
iter  10 value 21036.698341
iter  20 value 15617.111557
iter  30 value 10896.654673
iter  40 value 5835.171062
iter  50 value 4705.925019
final  value 4695.482603
converged
```

Figure 6: Model for birth defect data

```
step(cns.1)
```

```
Start:  AIC=9408.97
birth_type ~ Water + Work

trying - Water
# weights:  12 (6 variable)
initial  value 117209.801938
iter  10 value 12579.677501
iter  20 value 4766.837068
iter  30 value 4702.153552
final  value 4702.151109
converged
trying - Work
# weights:  12 (6 variable)
initial  value 117209.801938
iter  10 value 20958.882465
iter  20 value 6020.214535
iter  30 value 4702.872861
iter  40 value 4702.854773
final  value 4702.853815
converged
        Df      AIC
<none>   9 9408.965
- Water  6 9416.302
- Work   6 9417.708

Call:
multinom(formula = birth_type ~ Water + Work, data = cns, weights = count)

Coefficients:
      (Intercept)        Water WorkNonManual
NoCNS    5.455468  0.002904707     0.3637960
Other   -1.139496  0.002395556    -0.2785110
Sp       0.384321 -0.001418227     0.1203003

Residual Deviance: 9390.965
AIC: 9408.965
```

Figure 7: Output from `step` for model `cns.1`

```
new2 <- datagrid(model = cns.1, Work = c("Manual", "NonManual"))
cbind(predictions(cns.1, new2)) %>%
  select(group, estimate, Water, Work) %>%
  pivot_wider(names_from = group, values_from = estimate)
```

Water	Work	An	NoCNS	Other	Sp
94	Manual	0.0032238	0.9913406	0.0012921	0.0041436
94	NonManual	0.0022462	0.9938162	0.0006814	0.0032562

Figure 8: Predictions from model `cns.1`

```
cns0 %>% mutate(CNS = An + Sp + Other) -> cns_yn
cns_yn
```

Area	NoCNS	An	Sp	Other	Water	Work	CNS
Cardiff	4091	5	9	5	110	NonManual	19
Newport	1515	1	7	0	100	NonManual	8
Swansea	2394	9	5	0	95	NonManual	14
GlamorganE	3163	9	14	3	42	NonManual	26
GlamorganW	1979	5	10	1	39	NonManual	16
GlamorganC	4838	11	12	2	161	NonManual	25
MonmouthV	2362	6	8	4	83	NonManual	18
MonmouthOther	1604	3	6	0	122	NonManual	9
Cardiff	9424	31	33	14	110	Manual	78
Newport	4610	3	15	6	100	Manual	24
Swansea	5526	19	30	4	95	Manual	53
GlamorganE	13217	55	71	19	42	Manual	145
GlamorganW	8195	30	44	10	39	Manual	84
GlamorganC	7803	25	28	12	161	Manual	65
MonmouthV	9962	36	37	13	83	Manual	86
MonmouthOther	3172	8	13	3	122	Manual	24

Figure 9: Setup for another model for birth defects

```
cns.2 <- glm(cbind(CNS, NoCNS) ~ Water + Work, data = cns_yn, family = "binomial")
summary(cns.2)
```

```
Call:
glm(formula = cbind(CNS, NoCNS) ~ Water + Work, family = "binomial",
    data = cns_yn)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.4325803  0.0897889 -49.367  < 2e-16 ***
Water         -0.0032644  0.0009684  -3.371 0.000749 ***
WorkNonManual -0.3390577  0.0970943  -3.492 0.000479 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.047  on 15  degrees of freedom
Residual deviance: 12.363  on 13  degrees of freedom
AIC: 102.49

Number of Fisher Scoring iterations: 4
```

Figure 10: Another model for birth defects of central nervous system

```
cns %>% filter(birth_type != "NoCNS") -> cns_type
cns.3 <- multinom(birth_type ~ Water + Work, data = cns_type, weights = count)
```

```
# weights:  12 (6 variable)
initial  value 762.436928
iter  10 value 685.762336
final  value 685.762238
converged
```

Figure 11: A third model for birth defects of central nervous system. `!=` means "not equal to".

```
cns.4 <- multinom(birth_type ~ 1, data = cns_type, weights = count)
```

```
# weights:  6 (2 variable)
initial  value 762.436928
final  value 687.227416
converged
```

```
anova(cns.4, cns.3, test = "Chisq")
```

Model	Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
1	94	1374.455		NA	NA	NA
Water + Work	90	1371.524	1 vs 2	4	2.930357	0.5695467

Figure 12: Output from `anova` for third model. The model `cns.4` contains only an intercept, no explanatory variables.

```
glass %>% slice_sample(n = 20)
```

Ca	type
8.95	Head
5.87	Con
9.42	WinNF
8.05	WinNF
9.07	Head
9.13	WinNF
8.22	WinNF
8.68	WinF
7.59	Tabl
8.55	WinF
8.27	WinNF
9.57	WinF
8.17	WinF
8.79	Veh
7.83	WinNF
8.12	WinNF
8.76	Head
13.30	WinNF
8.21	WinNF
11.14	WinNF

Figure 13: Glass fragment data (20 randomly selected rows)

type	description
WinF	window float
WinNF	window non-float
Veh	vehicle window
Con	container
Tabl	tableware
Head	vehicle headlamp

Figure 14: Descriptions of glass types. “Float glass” is made by a process good for large windows.

```r
m <- cbind(c1, c2, c3, c4, c5)
contrasts(glass$type) <- m
glass.1 <- lm(Ca ~ type, data = glass)
summary(glass.1)
```

```
Call:
lm(formula = Ca ~ type, data = glass)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2538 -0.7273 -0.1655  0.3454  7.1163

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.1043     0.1289  70.655  < 2e-16 ***
typec1       -0.1382     0.1152  -1.199  0.23182
typec2        0.1017     0.2377   0.428  0.66924
typec3        0.2949     0.2191   1.346  0.17963
typec4       -1.2719     0.4333  -2.935  0.00371 **
typec5        0.3836     0.3017   1.272  0.20494
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.391 on 208 degrees of freedom
Multiple R-squared:  0.06667,   Adjusted R-squared:  0.04423
F-statistic: 2.971 on 5 and 208 DF,  p-value: 0.01297
```

Figure 15: Analysis of glass fragment data (note that the column `type` is already a `factor`)

```
cadmium %>% slice_sample(n = 20)
```

age	vital.capacity	exposure
41	3.88	none
21	5.22	low
35	4.24	none
43	4.61	low
45	4.02	high
32	4.55	none
36	4.36	none
38	5.09	low
43	4.62	low
29	5.17	low
29	5.21	low
65	4.83	none
29	4.51	low
42	4.89	none
42	5.12	low
58	2.88	low
36	4.02	none
48	5.00	none
48	4.06	low
42	4.04	none

Figure 16: Cadmium exposure data (random sample of observations)

```
ggplot(cadmium, aes(x = age, y = vital.capacity,
                    colour = exposure)) +
  geom_point() + geom_smooth(method = "lm")
```

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

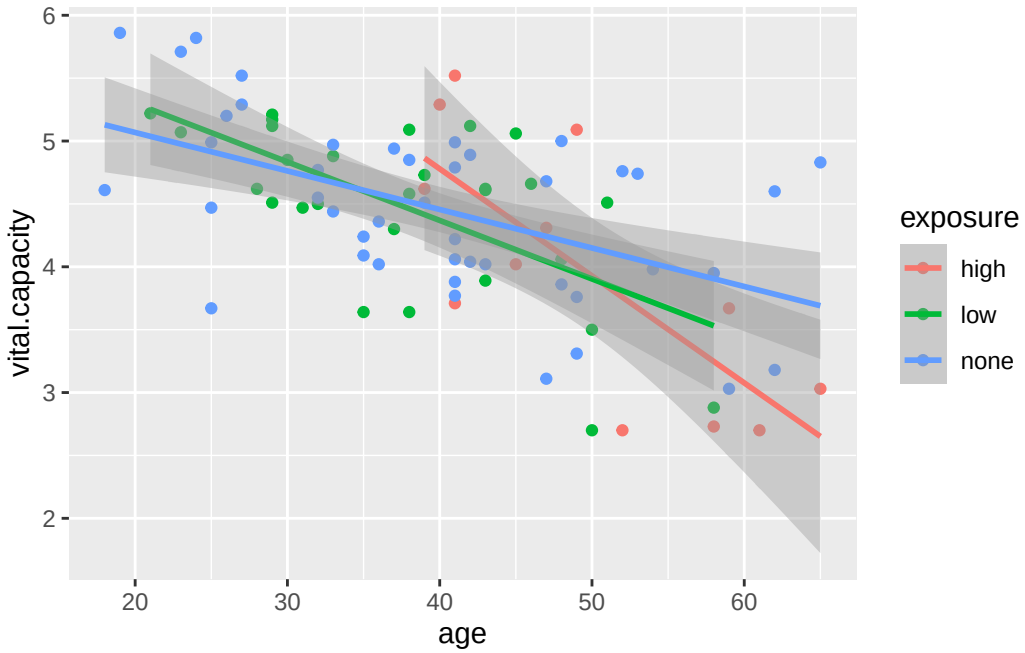


Figure 17: Graph of cadmium data

```
cadmium.1 <- lm(vital.capacity ~ age * exposure, data = cadmium)
```

Figure 18: Model fit to cadmium data

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

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
	NA	NA	27.53521	-81.68950	NA	NA
age:exposure	2	2.499458	30.03467	-78.39101	3.540153	0.0337565

Figure 19: Output 1 from model for cadmium data

```r
summary(cadmium.1)
```

```
Call:
lm(formula = vital.capacity ~ age * exposure, data = cadmium)

Residuals:
     Min       1Q   Median       3Q      Max
-1.24497 -0.36929  0.01977  0.43681  1.13953

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.18344    0.99358   8.236 3.28e-12 ***
age              -0.08511    0.01967  -4.327 4.44e-05 ***
exposurelow      -1.95341    1.10481  -1.768   0.0810 .
exposurenone     -2.50315    1.04184  -2.403   0.0187 *
age:exposurelow   0.03858    0.02327   1.658   0.1014
age:exposurenone  0.05450    0.02107   2.587   0.0116 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5942 on 78 degrees of freedom
Multiple R-squared:  0.422, Adjusted R-squared:  0.385
F-statistic: 11.39 on 5 and 78 DF,  p-value: 2.871e-08
```

Figure 20: Output 2 from model for cadmium data

```
heat %>% slice_sample(n = 20)
```

AC	MathZRT	ColorsZRT
No	0.3689486	1.8291605
Yes	-0.1542739	1.1364296
Yes	-0.0385349	-0.0081918
No	-0.3838167	0.0744950
Yes	0.4016167	0.8286923
Yes	-0.3427270	0.7745810
Yes	0.4780500	-0.7766711
Yes	1.0086719	0.5876084
No	0.1375749	3.1508342
No	0.7779180	0.1392707
No	-0.3744340	0.5302846
Yes	-0.4569153	-0.7322503
No	-0.4646578	-0.5893349
No	-0.4043569	-0.2865224
Yes	0.0888260	-0.9789665
Yes	-0.3330683	-1.2057150
No	0.3052599	1.2569187
No	-0.1932635	0.4394465
Yes	-0.1720559	-0.8528630
No	0.3052090	1.0944743

Figure 21: Heat and cognition data, randomly selected rows

```
heat %>% select(-AC) %>%
  as.matrix() -> response
heat.1 <- manova(response ~ AC, data = heat)
summary(heat.1)
```

```
          Df  Pillai approx F num Df den Df  Pr(>F)
AC         1 0.14688   3.7015      2     43 0.03287 *
Residuals 44
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 22: Heat and cognition data, analysis 1

```
heat.3 <- lda(AC ~ MathZRT + ColorsZRT, data = heat)
heat.3
```

```
Call:
lda(AC ~ MathZRT + ColorsZRT, data = heat)

Prior probabilities of groups:
       No       Yes
0.5652174 0.4347826

Group means:
        MathZRT  ColorsZRT
No   0.07174467  0.6229526
Yes -0.03086761 -0.1573737

Coefficients of linear discriminants:
                LD1
MathZRT   -0.229593
ColorsZRT -1.018859
```

Figure 23: Heat and cognition data, discriminant analysis

```
p <- predict(heat.3)
heat_p <- cbind(heat, p)
```

Figure 24: Heat and cognition data, further analysis

```
heat_p %>% slice_sample(n = 15)
```

	AC	MathZRT	ColorsZRT	class	posterior.No	posterior.Yes	LD1
21	Yes	-0.1720559	-0.8528630	Yes	0.3364193	0.6635807	1.2037097
38	No	0.2027137	-1.1434487	Yes	0.2991778	0.7008222	1.4137312
13	Yes	-0.3427270	0.7745810	No	0.6561051	0.3438949	-0.4152421
7	Yes	1.0610924	1.3597289	No	0.8018446	0.1981554	-1.3337326
29	No	0.3689486	1.8291605	No	0.8401444	0.1598556	-1.6531061
42	No	0.2545139	-1.1395287	Yes	0.3019117	0.6980883	1.3978444
19	Yes	-0.4139825	-0.8172580	Yes	0.3329073	0.6670927	1.2229779
30	No	-0.0483233	0.8874223	No	0.6889963	0.3110037	-0.5978046
6	Yes	-0.1542739	1.1364296	No	0.7277427	0.2722573	-0.8271825
39	No	-0.1511125	2.7944868	No	0.9142510	0.0857490	-2.5172356
34	No	0.1375749	3.1508342	No	0.9380890	0.0619110	-2.9465842
31	No	-0.4646578	-0.5893349	Yes	0.3741411	0.6258589	1.0023909
11	Yes	-0.0141754	-0.9004402	Yes	0.3341887	0.6658113	1.2159359
26	No	-0.2009270	0.3640760	No	0.5818221	0.4181779	-0.0295515
45	Yes	-0.1367077	-0.3131837	Yes	0.4445949	0.5554051	0.6457366

Figure 25: Heat and cognition data, results of further analysis (some). The numbers to the left are the row numbers of the original data: that is, IDs for the students who took part in the study.

```
blues %>% slice_sample(n = 20)
```

subject	treatment	bdi.2m	bdi.4m	bdi.6m	bdi.8m
84	TAU	9	6	7	1
56	BtheB	7	5	4	0
7	TAU	7	7	3	7
31	TAU	7	15	16	0
95	BtheB	11	4	2	3
9	BtheB	13	14	20	11
94	BtheB	4	3	3	3
71	BtheB	6	10	1	0
88	TAU	0	6	0	1
62	TAU	23	15	25	17
20	BtheB	12	10	8	10
37	TAU	14	20	1	8
78	BtheB	10	5	5	12
35	BtheB	24	20	29	14
18	BtheB	8	8	7	6
43	TAU	29	2	0	0
19	TAU	30	33	31	22
47	BtheB	27	16	30	15
99	TAU	5	5	0	6
4	BtheB	17	16	10	9

Figure 26: Depression data (some)

```
blues %>%
  select(subject, treatment, ends_with("m")) %>%
  pivot_longer(ends_with("m"),
               names_to = "time",
               values_to = "beck") -> blues_long
blues_long %>% slice(1:15)
```

subject	treatment	time	beck
2	BtheB	bdi.2m	16
2	BtheB	bdi.4m	24
2	BtheB	bdi.6m	17
2	BtheB	bdi.8m	20
4	BtheB	bdi.2m	17
4	BtheB	bdi.4m	16
4	BtheB	bdi.6m	10
4	BtheB	bdi.8m	9
6	BtheB	bdi.2m	0
6	BtheB	bdi.4m	0
6	BtheB	bdi.6m	0
6	BtheB	bdi.8m	0
7	TAU	bdi.2m	7
7	TAU	bdi.4m	7
7	TAU	bdi.6m	3

Figure 27: Depression data, long format (some)

```
blues %>%
  select(ends_with("m")) %>%
  as.matrix() -> response
times <- colnames(response)
times.df <- data.frame(times=factor(times))
blues.1 <- lm(response ~ treatment, data = blues)
blues.2 <- Manova(blues.1,   idata = times.df,
  idesign = ~times)
```

Figure 28: Depression data, repeated measures analysis setup

Univariate type II tests, below:

```
                Sum Sq num Df Error SS den Df  F value    Pr(>F)
(Intercept)      36438      1  15133.7     50 120.3852 6.468e-15 ***
treatment         2530      1  15133.7     50   8.3588 0.0056665 **
times              497      3   3684.9    150   6.7466 0.0002672 ***
treatment:times    138      3   3684.9    150   1.8681 0.1374278
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sphericity tests (Mauchly's test) below:

```
                Test statistic p-value
times                  0.95419 0.80852
treatment:times        0.95419 0.80852
```

P-value adjustments, Greenhouse-Geisser and Huynh-Feldt, below. You may ignore the `na.action` stuff below that.

```
                   GG eps   Pr(>F[GG])   HF eps   Pr(>F[HF])
times           0.9723176 0.0003124699 1.039048 0.0002671774
treatment:times 0.9723176 0.1392150057 1.039048 0.1374278130
attr(,"na.action")
(Intercept)   treatment
          1           2
attr(,"class")
[1] "omit"
```

Figure 29: Depression data, repeated measures output

```
ggplot(blues_long, aes(x = time, y = beck, colour = treatment, group = subject)) +
         geom_point() + geom_line()
```

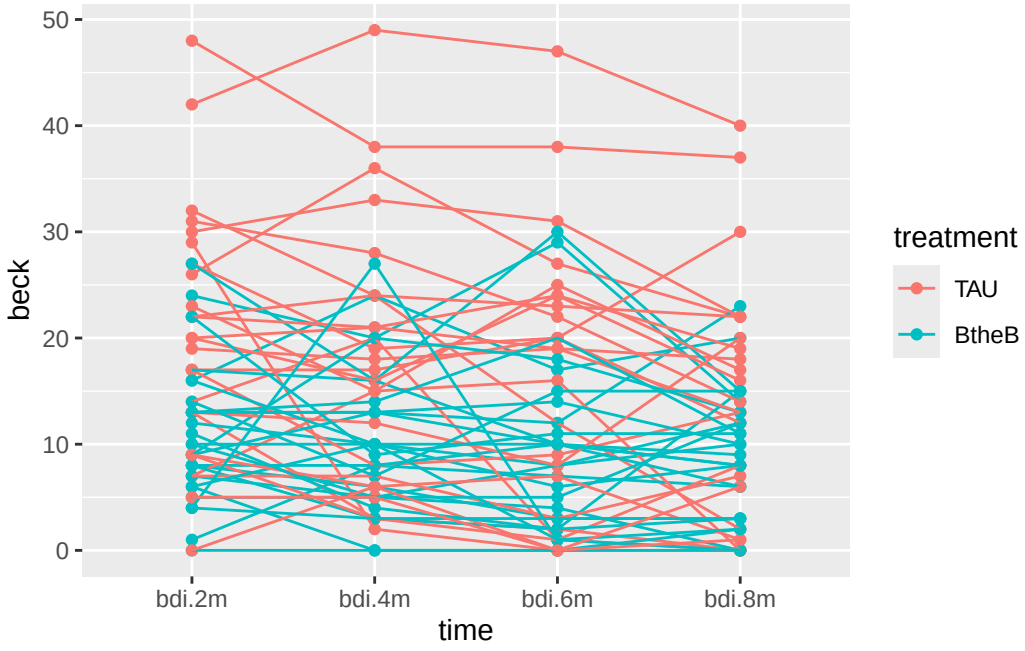


Figure 30: Depression data, spaghetti plot

```
blues_long %>%
  group_by(treatment, time) %>%
  summarize(mean_beck = mean(beck)) %>%
  ggplot(aes(x = time, y = mean_beck, colour = treatment, group = treatment)) +
  geom_point() + geom_line()
```

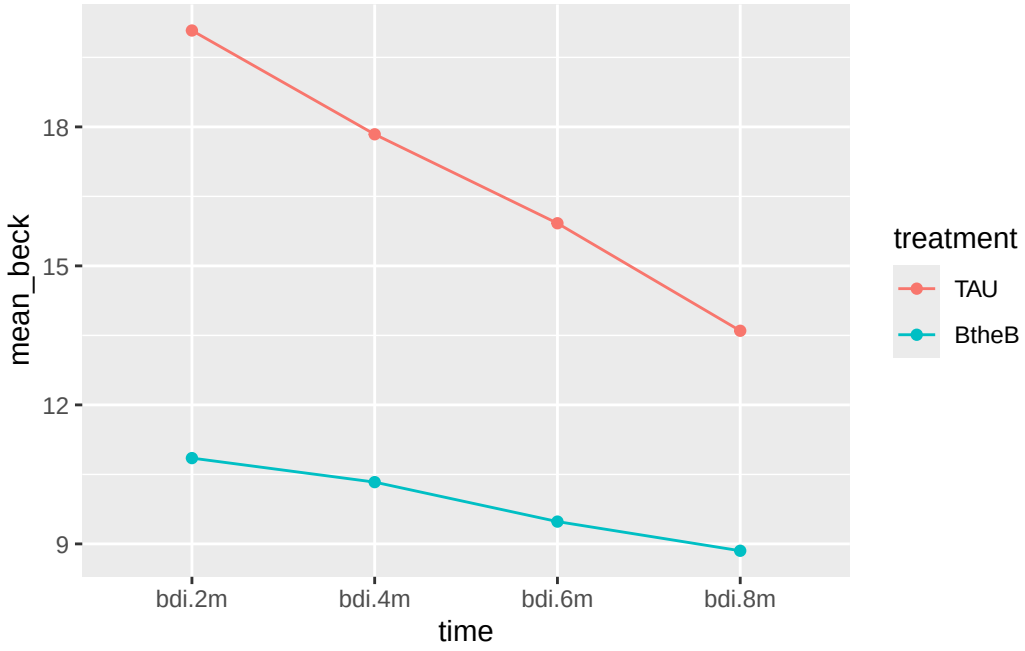


Figure 31: Depression data, interaction plot

```
ohio
```

county	corn	mixed	wheat	oats	barley	soy	hay
Adams	42.41	0.21	22.47	1.07	0.37	0.62	27.80
Allen	34.43	0.13	23.76	18.35	0.11	12.18	15.31
Ashtabula	22.88	0.24	13.52	15.67	0.02	1.30	38.89
Athens	26.61	0.18	8.89	3.42	0.05	0.71	53.91
Delaware	33.52	0.13	17.60	11.33	0.16	11.82	22.69
Clinton	48.45	0.24	29.50	3.10	0.25	2.72	9.85
Gallia	31.38	0.83	13.07	2.03	0.60	0.71	44.07
Geauga	23.04	0.21	12.68	17.44	0.11	0.41	37.80
Hancock	36.13	0.12	24.64	16.56	0.13	13.91	16.46
Highland	47.83	0.11	31.57	1.59	0.05	1.46	16.10
Meigs	28.20	0.28	14.08	3.06	0.18	0.67	46.71
Portage	26.67	0.11	19.13	18.67	0.03	0.69	27.33
Putnam	30.97	0.13	24.16	15.28	0.13	14.10	12.61
Warren	43.23	0.09	24.97	3.20	0.24	4.68	18.72
Washington	25.08	0.08	13.43	1.96	0.66	1.06	50.27

Figure 32: Ohio farmland data

```r
ohio %>%
  select(-county) %>%
  mutate(across(everything(), \(x) scale(x))) -> ohio_scaled
```

```r
ohio.3 <- kmeans(ohio_scaled, 3)
ohio.3
```

```
K-means clustering with 3 clusters of sizes 6, 4, 5

Cluster means:
        corn      mixed      wheat       oats     barley        soy        hay
1 -0.8355741  0.5310194 -1.0151217 -0.2167225  0.3263680 -0.6701208  1.0806934
2  1.4051093 -0.2373220  1.1041559 -0.9034358  0.1096393 -0.3844435 -0.7490024
3 -0.1213986 -0.4473657  0.3348213  0.9828157 -0.4793530  1.1116998 -0.6976301

Clustering vector:
 [1] 2 3 1 1 3 2 1 1 3 2 1 3 3 2 1

Within cluster sum of squares by cluster:
[1] 28.055703  4.474602  7.409624
 (between_SS / total_SS =  59.2 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

Figure 33: Ohio farmland 3-cluster analysis

```
tibble(county = ohio$county, cluster = ohio.3$cluster) %>%
  arrange(cluster)
```

county	cluster
Ashtabula	1
Athens	1
Gallia	1
Geauga	1
Meigs	1
Washington	1
Adams	2
Clinton	2
Highland	2
Warren	2
Allen	3
Delaware	3
Hancock	3
Portage	3
Putnam	3

Figure 34: Ohio farmland 3-cluster membership

```
ss <- function(i, d) {
  d %>%
    select(where(is.numeric)) %>%
    kmeans(i, nstart = 20) -> km
  km$tot.withinss
}
tibble(n_cluster = 2:13) %>%
  rowwise() %>%
  mutate(ssq = ss(n_cluster, ohio_scaled)) %>%
  ggplot(aes(x = n_cluster, y = ssq)) +
  geom_point() + geom_line()
```

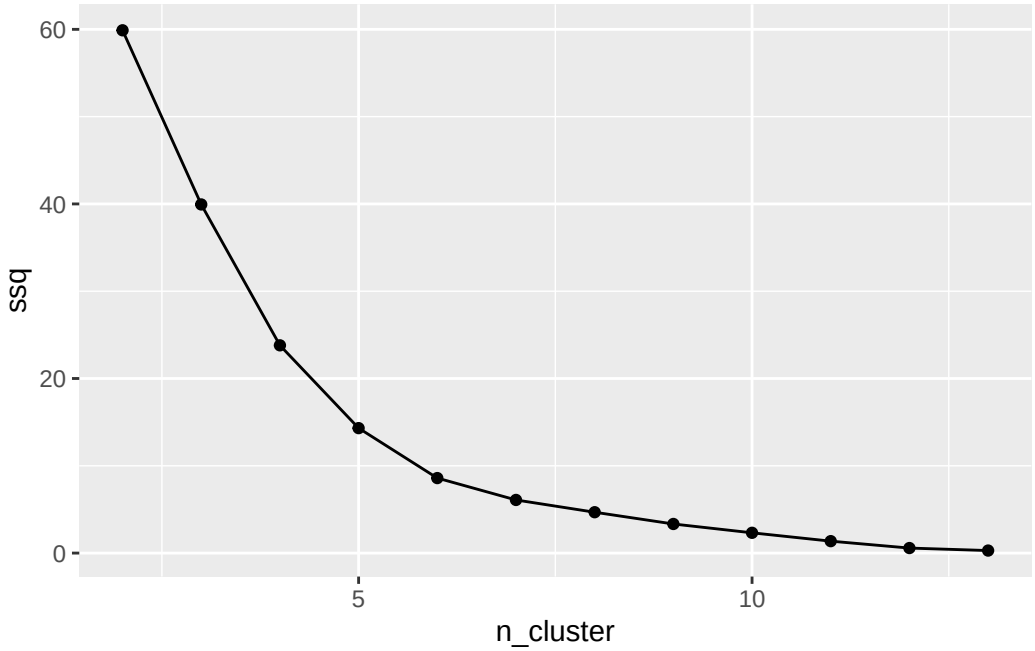


Figure 35: Ohio farmland plot

```r
ohio.6 <- kmeans(ohio_scaled, 6)
tibble(county = ohio$county,
       cluster3 = ohio.3$cluster,
       cluster6 = ohio.6$cluster) -> cluster_membership
with(cluster_membership, table(cluster3, cluster6))
```

```
        cluster6
cluster3 1 2 3 4 5 6
       1 3 0 0 0 2 1
       2 0 0 0 4 0 0
       3 0 3 1 0 1 0
```

Figure 36: Ohio farmland further analysis

```
ms_patients %>% slice_sample(n = 20)
```

A_diag	B_diag	from	frequency
Doubtful	Certain	Winnipeg	3
Doubtful	Doubtful	Winnipeg	10
Possible	Probable	New Orleans	13
Probable	Doubtful	New Orleans	0
Doubtful	Possible	Winnipeg	3
Possible	Probable	Winnipeg	14
Possible	Possible	New Orleans	3
Doubtful	Probable	New Orleans	2
Probable	Doubtful	Winnipeg	0
Certain	Certain	New Orleans	5
Possible	Possible	Winnipeg	5
Certain	Probable	New Orleans	3
Certain	Doubtful	Winnipeg	1
Doubtful	Doubtful	New Orleans	14
Possible	Doubtful	New Orleans	4
Doubtful	Certain	New Orleans	1
Probable	Probable	Winnipeg	11
Possible	Certain	New Orleans	2
Probable	Probable	New Orleans	11
Certain	Possible	New Orleans	0

Figure 37: MS patients data (some)

```
ms.1 <- glm(frequency ~ A_diag * B_diag * from,
            data = ms_patients, family = "poisson")
drop1(ms.1, test = "Chisq")
```

	Df	Deviance	AIC	LRT	Pr(>Chi)
	NA	0.00000	159.9112	NA	NA
A_diag:B_diag:from	9	6.64285	148.5541	6.64285	0.6742479

Figure 38: MS patients log-linear model 1

```
drop1(ms.2, test = "Chisq")
```

	Df	Deviance	AIC	LRT	Pr(>Chi)
	NA	6.963563	142.8748	NA	NA
A_diag:B_diag	9	129.297607	247.2088	122.33404	0e+00
B_diag:from	3	41.501477	171.4127	34.53791	2e-07

Figure 39: MS patients log-linear model 2

```
ggplot(ms_patients, aes(x = from, y = frequency, fill = B_diag)) +
  geom_col(position = "fill")
```

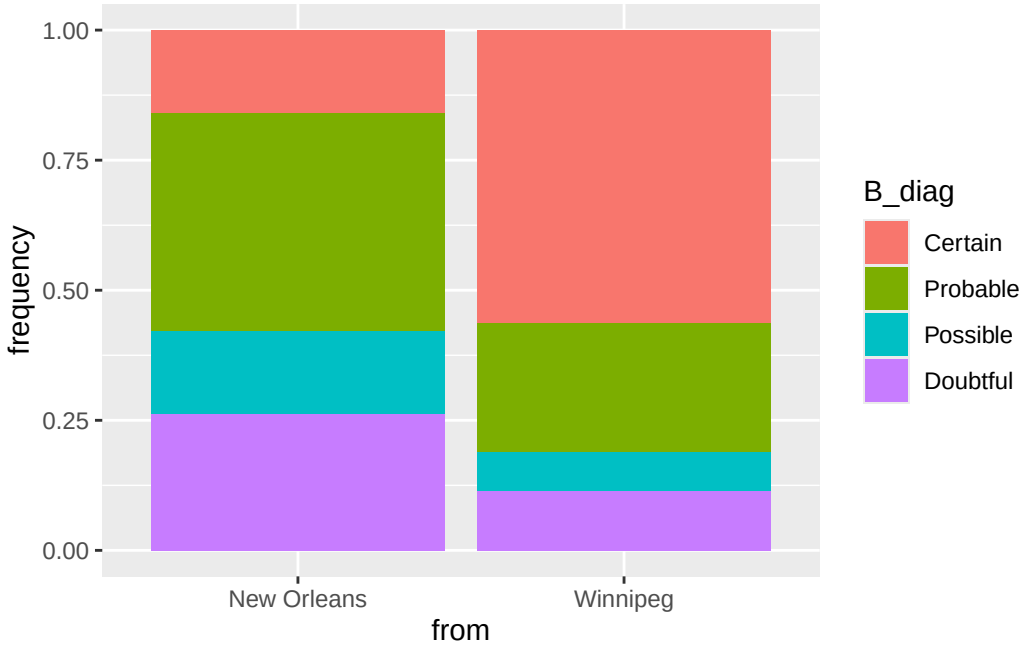


Figure 40: Bar chart 1 of MS patients

```
ggplot(ms_patients, aes(x = A_diag, y = frequency, fill = B_diag)) +
  geom_col(position = "fill")
```

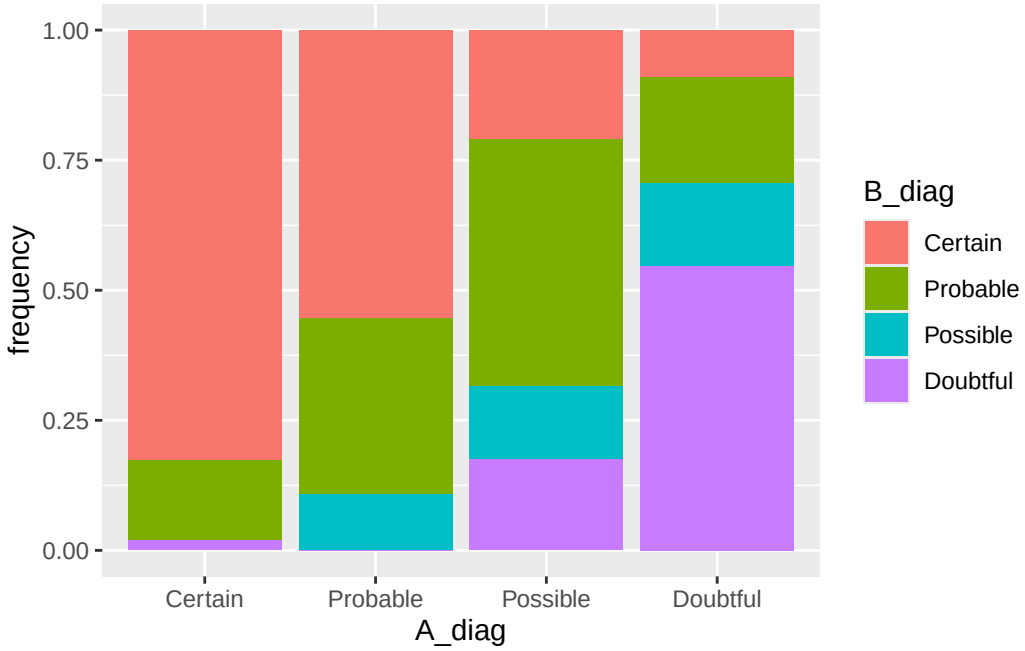


Figure 41: Bar chart 2 of MS patients