University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 (K. Butler), Midterm Exam
February 28, 2026

Aids allowed (on paper, no computers or other 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 10 numbered pages of questions including this cover page.

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

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

If you need more space, use the last page of the exam. Anything written on the back of the page will not be graded.

You may assume throughout this exam that the code shown in Figure 1 of the booklet of Figures has already been run.

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

Stack loss

A chemical plant is making nitric acid by oxidizing ammonia. Data were obtained from 21 days of operation on the following variables:

- `Air.Flow`: the rate of air flow through the plant during the process
- `Water.Temp`: temperature of cooling water in the tower where the nitric acid is collected
- `Acid.Conc.`: concentration of the acid
- `stack.loss`: how much of the input ammonia escapes from the tower without being converted to nitric acid (lower better).

We are interested in how `stack.loss` depends on the other variables. The data, in dataframe `stackloss`, are shown in Figure 2.

(1) (2 points) Some plots are shown in Figure 3. A regression analysis is shown in Figure 4. What is the most important conclusion to be drawn from both the plots and the regression output (one conclusion based on *both* figures)? Explain briefly.

The plots are scatterplots of stack loss against each of the explanatory variables (as you can work out from the code above the plots). The relationship with `Acid.Conc.` is much weaker than the others; in the regression, `Acid.Conc.` is not significant. Both Figures are indicating that `Acid.Conc.` should be removed from the regression.

This is a stronger conclusion than saying that one of the other variables has a strong relationship with `stack.loss` and therefore should stay in the regression. I asked for one conclusion (singular), so mentioning that the two significant explanatory variables both have reasonably strong relationships with `stack.loss` is not answering the question either, though you'll get partial credit for an answer like that.

Points:

- 2: Both the scatterplot and regression say to remove `Acid.Conc`, and why (weak relationship, not significant)
- 1: Right conclusion but missing or incomplete explanation of why
- 1: More than one relevant conclusion, eg. to keep the two significant explanatory variables (with stronger relationships)
- 1: Saying to keep one explanatory variable and why
- 0.5: relevant comments but not enough for 1 point.

Extra: this is one of the "classic" datasets included with R, alongside `mtcars` which you met at the beginning of C32. A hint that this dataset goes back to the early days of R is the use of dots to separate words in variable names; since the advent of `tidyverse`, the fashion is to use underscores rather than dots, hence `Air_Flow` or even `air_flow` instead of `Air.Flow`.

(2) (2 points) A second regression is shown in Figure 5. According to this Figure, what can you say about the most desirable values of the explanatory variables? Explain briefly.

The most desirable value of the response `stack.loss` is a small one (less of the ammonia is not converted into nitric acid, or, more clearly, more of it *is* converted into nitric acid). The slopes of `Air.Flow` and `Water.Temp` are both positive, so the way to make `stack.loss` small is to also have both of these explanatory variables small.

This has nothing to do with P-values or with R-squared.

Points:

- 2: Best response value is small one, positive slopes, so best values of explanatory variables are both small.
- 1: as 2 points, but slip in logic (eg claiming that something needs to be large)
- 1: reasonable thinking, but not reaching answer (such as investigating the effects of changing the explanatory variables)
- 0.5: right answer, but missing or incomplete explanation
- 0.5: some progress

(3) (2 points) Some predictions are shown in Figure 6, using a function from the `marginaleffects` package. What precisely is being estimated here? Explain briefly.

These use `predictions` from the `marginaleffects` package, so they are estimating the mean stack loss for *all possible* observations that have the values shown for `Air.Flow` and `Water.Temp`. The issue is actually not the predictions themselves (they are the same whether you are predicting mean response for all observations or a single response), but about what those upper and lower limits are actually an interval *for*.

Points:

- 2: mean response for all possible observations with those values for the two explanatory variables.
- 1: apparently the right answer but not clearly enough explained.
- 0.5: claiming that they are prediction intervals
- 0.5: something relevant but not addressing the upper and lower limits

Extra: you can tell that these come from `marginaleffects` because of the columns called `estimate`, `conf.low`, and `conf.high`. They are not prediction intervals: that is to say, they are not estimating the uncertainty in how much stack loss there would be in an individual observation of stack loss for those values of the two explanatory variables. If you wanted to do that, you would need to do something like this:

```r
new <- tibble(Air.Flow = c(50, 60), Water.Temp = c(18, 21))
p <- predict(stackloss.2, new, interval = "p")
cbind(new, p)
```

Air.Flow	Water.Temp	fit	lwr	upr
50	18	6.515207	-0.6641464	13.69456
60	21	17.112805	10.1482280	24.07738

These intervals are much wider, because there is a lot more uncertainty attached to an individual observation than there is to the average of all possible observations. (Looking ahead to the next question, the first interval here is also longer than the second one, but it is much more difficult to tell by eye that this is the case.)

(4) (2 points) In Figure 6, the first interval is longer than the second one. Why is that?

The first interval is for predicting stack loss when `Air.Flow` is 50 and `Water.Temp` is 18. The second interval is for values of 60 and 21 respectively.

Go back and look at the data in Figure 2. An `Air.Flow` of 50 is the smallest value; similarly, a `Water.Temp` of 18 is close to the smallest value. The values of 60 and 21 seem to be more typical of the data, and you can reasonably infer that these values are close to the means of the two variables. The predictions will be more accurate (the confidence intervals will be shorter) if the explanatory variables are close to their means, and less accurate (longer confidence intervals) if the explanatory variables are far from their means, as the values 50 and 18 are.

Only one point for "because the explanatory variables are further from their means"; the second point is for a convincing explanation of how you know.

Points:

- 2: "the explanatory variable values are further from the means for the first interval", along with a reasonable explanation of how you know
- 1: "the explanatory variable values are further from the means for the first interval" with missing or incomplete explanation

The grader can also award 1.5 if they think you got close but not close enough.

Seed germination

Seeds of a certain species were planted in an experiment. Each seed was treated with a certain amount of fertilizer (in suitable units), and it was recorded whether the seed germinated (started to grow into a plant) or not. The data are shown in Figure 7. There are two columns: `fert`, the amount of fertilizer, and `x`, where 1 indicates germination and 0 indicates non-germination.

(5) (2 points) How do you know that each row of Figure 7 represents only one seed, as opposed to more than one seed?

Look at the column `x`: the only possible values are 1 (germinated) and 0 (did not), so it must be talking about a single seed each time. Alternatively, if each row had been referring to more than one seed, we would have observed *counts* of the number of seeds that did or did not germinate (like, 3 germinated and 2 did not) rather than the values 1 and 0.

Points:

- 2: correct explanation of why each row refers to one seed or how it cannot refer to more than one seed.
- 1: partial but incomplete explanation

(6) (1 point) Why did I use `group = x` in the code for my boxplot in Figure 8?

The problem is that `x`, though it is really categorical, is recorded as a number. We need to tell `ggplot` that it is actually categorical. One way would have been to use `factor(x)` inside `aes`, but this way also works: the variable `x` divides the data up into groups.

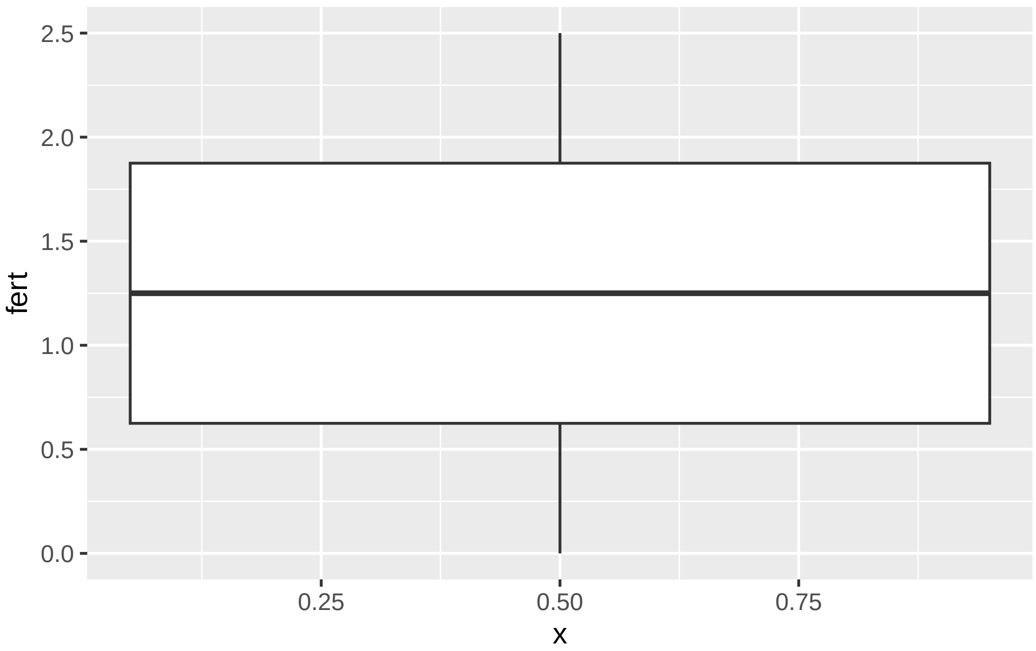
Points:

- 1: `x` is actually categorical but recorded as a number.

Extra: You'll notice that the boxplot has a continuous x-scale, as if all values of `x` are possible, not just 0 and 1, but there are correctly two boxes. If you leave out the `group`, this happens:

```
ggplot(germination, aes(x = x, y = fert)) + geom_boxplot()
```

```
Warning: Continuous x aesthetic
i did you forget `aes(group = ...)`?
```



You only get *one* box, along with two warnings:

- the first one says that both variables appear to be quantitative. A boxplot has one quantitative variable (on the y) and one categorical (on the x). With two apparently-quantitative variables, it's saying that it's putting them on the `x` and `y` you specified, but is warning you that the result may not be what you were hoping for.
- the second warning says that you might have meant to use a `group`. This has probably happened to you while you have been doing your assignments, and the solution is either to use a `group` (as here), or put the actually-categorical variable inside a `factor`.

(7) (2 points) Interpret the boxplot in Figure 8.

This requires a bit of care, because the roles of explanatory and response have been reversed here. The interpretation of a boxplot is "for each group on the x-axis, what happens to the variable on the y-axis?":

The seeds that germinated tended to be treated with more fertilizer than the seeds that did not germinate.

In your answer, make sure you express "if germinated, then more fertilizer" rather than the other way around, because that's the way a boxplot works in terms of cause and effect, even though you would *like* to be able to say "a seed that gets more fertilizer is more likely to germinate", which the logistic regression (coming up) *will* enable you to say.

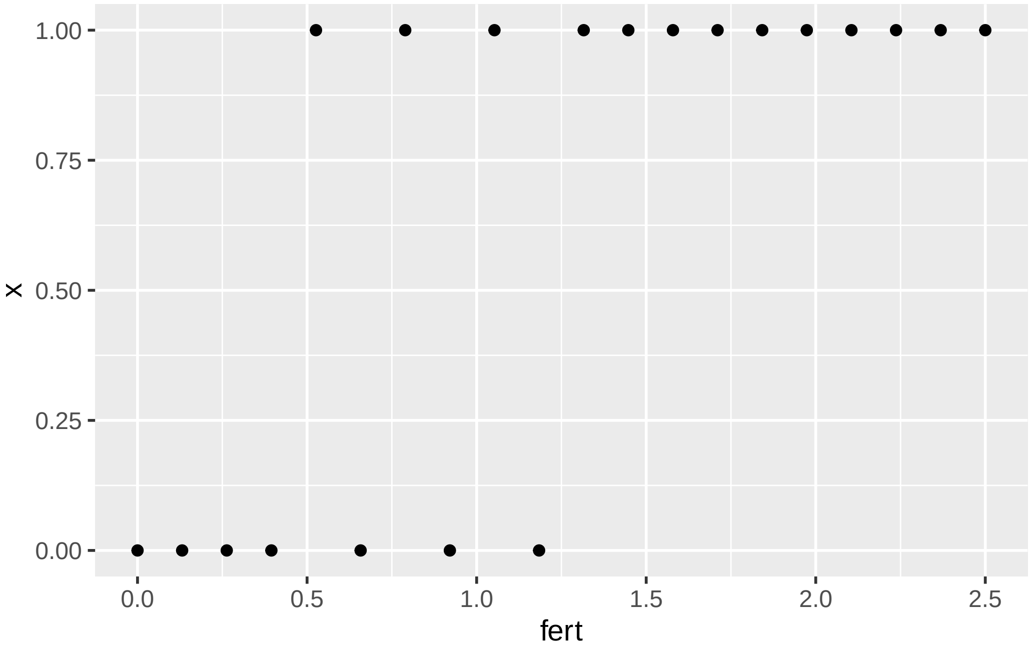
Points:

- 2: if a seed germinated, then it tended to be treated with more fertilizer than if it did not.
- 1: if a seed was treated with more fertilizer, then it was more likely to germinate (backwards logic from boxplot)
- 0.5: something else possibly relevant

Extra 1: you usually use a boxplot in an ANOVA-type situation, where the categorical variable is explanatory and the quantitative one is the response, so that the interpretation is then as you would expect. But here, the quantitative variable is *explanatory* and the categorical one is the *response*, hence the need for care in interpreting the boxplot. We run into this issue again later in interpreting a MANOVA (where the categorical variable is explanatory) by using a discriminant analysis (where the categorical variable acts like a response).

Extra 2: if you wanted a graph from which you *could* conclude the right thing, you could put fertilizer on the x-axis and make a kind of scatterplot with the actual values of x:

```
ggplot(germination, aes(x = fert, y = x)) + geom_point()
```



(note the weirdness inside `aes`), and then you can say that if `fert` is bigger, then `x` tends to be bigger as well: that is, with more fertilizer a seed is more likely to germinate. (Note that for the four lowest values of `fert` the seed did not germinate, and for about the ten highest values of `fert`, the seed *did* germinate.)

(8) (2 points) A logistic regression is shown in Figure 9. Assuming that these are a random sample of seeds of this type, can you conclude that there is a relationship between fertilizer amount and germination probability for all seeds of this species? Explain briefly.

This is another way of saying "is there a significant relationship, one that generalizes to the population?". To test this, look at the P-value for `fert`, which is 0.0239. This is less than 0.05, so reject the null hypothesis (that there is no relationship) in favour of the alternative that there is one.

Make sure to say what P-value you are using, since there are two.

Points:

- 2: "yes", for a correct reason, including quoting the correct P-value
- 1: "yes" without citing the correct P-value
- 0.5: "no" with the P-value 0.504 (from the intercept), or some other confusion

(9) (2 points) How do you know that the logistic regression in Figure 9 is predicting the probability that a seed *does* germinate, as opposed to the probability that it does not?

The actual answer is that if the response variable in a logistic regression takes the values 0 and 1, `glm` will fit the probability of 1.

An equally acceptable answer (and the one I think you'll give) is that the first response category "alphabetically" (actually numerically) is 0, and the model will predict the probability of the other one, 1.

The above is the best answer, but there may be something for saying that the boxplot showed seeds that germinated tended to have more fertilizer, and the slope in Figure 9 is positive (looking ahead to the next question), hence the logistic regression must be predicting the probability that a seed germinates. This is an inference, however, whereas the above is an absolute certainty because of the way the modelling works.

Points:

- 2: "probability of 1 category" or "non-baseline category, ie. 1"
- 1: comparison of boxplot and positive slope with logical inference
- 1: as 2, but explanation not clear enough

(10) (2 points) Interpret the sign (positive or negative) of the Estimate for `fert` in Figure 9.

3.811 is positive, so as the amount of fertilizer goes up, the probability of the seed germinating also goes up. (Or, a higher amount of fertilizer is associated with a larger probability of germination.)

Points:

- 2: positive, so increase in fertilizer goes with increase in probability of germination (or logical equivalent such as increase in log-odds of germination)
- 1: as 2, but explanation not clear enough (such as not being explicit about what increases when fertilizer increases)
- 0.5: potentially relevant comment but not enough for 1

(11) (2 points) Interpret the numerical value of the Estimate for `fert`.

The value is 3.811, so if the amount of fertilizer increases by 1 unit, the predicted *log-odds* of the seed germinating increases by 3.811.

I hope you can manage to stop yourself from saying "the probability increases by 3.811", since a probability must be between 0 and 1. Looking back at Figure 7, an increase of 1 in `fert` *is* something that was observed in the data, so there is no way the 3.811 can be a change in probability.

Points:

- 2: as amount of fertilizer increases by 1 unit, log-odds of germination increases by 3.811
- 1: as 2, but not clear enough
- 0.5: trying to interpret the wrong number
- 0: a claim that the *probability* increases by 3.811

Extra: This is actually a very large change on the log-odds scale, but it makes sense because very few of the seeds with close to 0 fertilizer germinate, but by the time the amount of fertilizer gets up to 1 and beyond, almost all of the seeds are germinating.

(12) (3 points) Using the `marginaleffects` package and starting from the output in Figure 9, what code would display a table of predicted probabilities of germination for amounts of fertilizer from 0 to 2.5 in steps of 0.5, together with lower and upper 95% confidence limits for those predictions, as done in this class?

This is `predictions`. There are three steps:

- make a dataframe of values to predict for (the thing I usually call `new`)

- get the predictions
- from the predictions, display the columns you need.

First:

```
new <- tibble(fert = seq(0, 2.5, 0.5))
new
```

fert
0.0
0.5
1.0
1.5
2.0
2.5

The column must be called exactly `fert`, but the dataframe can have any name (as long as you use that name below).

Second and third (the two lines of code below, respectively):

```
cbind(predictions(germination.1, new)) %>%
  select(fert, estimate, conf.low, conf.high)
```

fert	estimate	conf.low	conf.high
0.0	0.0380068	0.0015504	0.5013062
0.5	0.2099031	0.0385550	0.6376856
1.0	0.6411190	0.2924235	0.8853492
1.5	0.9231516	0.5001208	0.9931145
2.0	0.9877716	0.6063462	0.9997640
2.5	0.9981623	0.6816470	0.9999927

The inputs to `predictions` have to be in that order, but as long as you `select` those four columns in some not-insane order, that part is good. Naming the `model` (first) and `newdata` (second) inputs to `predictions` is optional, but if you do name them (as `model` and `newdata`), they can be in either order. The columns you `select` have to be exactly those, in some order: `fert` from the original data, and the other three from the output of `predictions`.

Points:

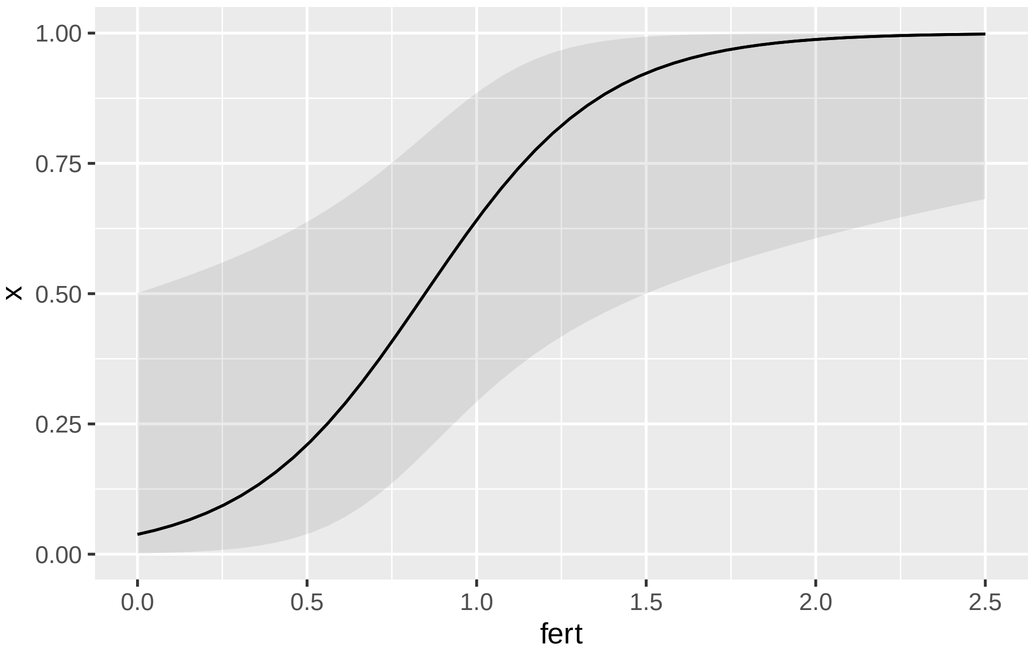
- 3: correct code to create `new` (via `tibble` or `datagrid` with model specified, or other equivalent), correct run of `cbind(predictions())`, correct selection of columns
- 2.5: as 3 but with small error
- 2: as 3 but with important error, or one step missing or incomplete
- 2: as 3 but with two small errors (by grader's judgement)
- 1.5: as 3 but with more than two errors (or one large and one small) and at least two steps attempted
- 1: one step correct but no more
- 0.5: something relevant but not enough for 1

Extra: you won't have the output, of course, but now that you see it, you see again that the probability of germination is very small for low amounts of fertilizer and close to 1 for high amounts. This strong trend is why the fertilizer effect was significant.

Having said that, though, note that the confidence intervals are very wide, so we really have not estimated the probability of germination very accurately at all. This is because a sample size of only 20 seeds is very small, when all you observe about the seeds is whether or not they germinated. This small sample size is why the P-value for `fert` was small but not *very* small (0.0239), even though the predictions seemed definitely to be increasing sharply.

Another way to see that is a plot of the predictions:

```
plot_predictions(germination.1, condition = "fert")
```



Here, you see both the strong upward trend in the predictions, but also the fact that we are very uncertain.

Married women participating in the labour force

In a 1977 survey, 263 Canadian married women (in heterosexual marriages) were asked about whether they worked outside the home, as well as about some variables that might influence this: their husband's income, whether or not there were children in the household, the region of the country in which they live. Some of the data are shown in Figure 10, and details about the variables recorded are shown in Figure 11. The aim of the survey was to see whether and how the other variables influenced participation in the labour force outside the home. The dataframe is called `womenlf`.

(13) (2 points) Why would it be appropriate to use `polr` from the `MASS` package to analyze these data?

The `polr` function is for ordinal logistic regression. This is appropriate here because the response variable `partic` is categorical with:

- more than two categories (here, three)
- the categories have an order that makes sense, eg. from least to most work outside the home.

If there were only two categories, it wouldn't matter whether they were ordered or not, because whatever order you put them in, they would actually be in a sensible order (either high to low or low to high) already.

Points:

- 2: more than 2 categories *and* there is an order that makes sense (that you describe)
- 1: one of the two items for 2 points
- 1: as 2 points, but without describing an order that makes sense (or otherwise explaining how you know there is one)

(14) (2 points) Figure 12 shows the number of women in each employment category. Something in this Figure indicates that `polr` will not run appropriately. What is that?

The categories are in *alphabetical* order, rather than a logical one, such as from least to most work (outside the home).

It is enough to say that the order is not logical, and why (or to give an example of a logical order).

Points:

- 2: these categories in alphabetical or non-logical order, and how you know or example of an order that the categories need to be in.
- 1: categories in inappropriate order without clear enough explanation of why that is

(15) (2 points) Some analysis is shown in Figure 13. It was necessary to use `drop1` here to obtain P-values. What are *two* distinct reasons why that was necessary?

These:

- we have categorical explanatory variables (`children` and `region`). This is the same reason as for any regression.
- in this kind of model, the `summary` output does not give helpful P-values, or indeed any P-values at all.

Points:

- 2: both of: categorical explanatory variables; `summary` output does not give P-values.
- 1: one of the two reasons for 2 points.

Extra: to illustrate the second point:

```
summary(womenlf.1)
```

```
Re-fitting to get Hessian


Call:
polr(formula = partic ~ hincome + children + region, data = womenlf)

Coefficients:
                   Value Std. Error t value
hincome         -0.05691    0.02017 -2.8215
childrenpresent -2.00989    0.29431 -6.8292
regionBC         0.15299    0.56438  0.2711
regionOntario    0.26867    0.46220  0.5813
regionPrairie    0.47966    0.54127  0.8862
regionQuebec    -0.07109    0.49377 -0.1440

Intercepts:
                  Value   Std. Error t value
not.work|parttime -1.7492  0.5436    -3.2177
parttime|fulltime -0.8315  0.5325    -1.5617

Residual Deviance: 439.766
AIC: 455.766
```

Even if you interpret a t-value over 2 in size as significant, the comparison here is with the baseline category, so that you don't get a P-value for the categorical variable as a whole. Here, that matters especially for `region`, since there are five regions: nothing appears to be significantly different from the `Atlantic` region, but that does not rule out two of the other regions being significantly different from each other.

To interpret the actual estimates here (in the column labelled `Value`) requires you to know which way around the ordering goes. I chose to go in increasing order of work outside the home, from none through part-time to full-time, so the values in the Coefficients table reflect that:

- `hincome` is (apparently) significantly negative, so a higher husband's income goes with being lower on the employment scale (more likely to work less outside the home)

- `childrenpresent` is (apparently) strongly significantly negative, which means that having children in the household as opposed to not doing so also goes with being lower on the employment scale (more likely to work less outside the home)
- none of the `region`s seem to be significantly different from the Atlantic region, but if anything, married women living in the Prairie region are most likely to work more (most positive t-value) and those in Quebec are least likely (most negative t-value). (You could say something about availability of child care in the different provinces, but bear in mind that the `drop1` table says that there are no differences among the regions overall.)

This is, as you see, harder work than looking at predictions, which is why we don't spend much time with the `summary` output for this kind of model.

(16) (2 points) Some further analysis is shown in Figure 14. Why is model `womenlf.2` better than model `womenlf.1`? Explain briefly.

In model `womenlf.1`, `region` is not significant, so should be removed from the model, as has been done in model `womenlf.2`.

This is meant to be an easy one.

Points:

- 2: (i) `region` is not significant and (ii) should be removed from the model.
- 1: `region` is not significant (only)
- 1: `region` should be removed from the model (only).

(17) (2 points) Some predictions are shown in Figure 15. The code that produced these predictions (not shown) ended with a pivot-wider. Why was that pivot-wider necessary?

The purpose of the pivot-wider was to display all three probabilities for each combination of `children` and `hincome` side by side to allow easier comparison.

Make sure you say what the importance of the pivot-wider is, or what happens if you don't do it.

(I didn't show the code because some of it is similar to what I asked you to give earlier, in the logistic regression question.)

Points:

- 2: to display predictions beside each other to allow easier comparison (or some other good reason for having them next to each other or not having them in one long column)

- 1: to display predictions beside each other without clear reason for doing so

Extra: this, in fact, is what happens without the pivot-wider:

```r
new <- datagrid(model = womenlf.2, hincome = seq(10, 50, 20),
                children = c("absent", "present"))
cbind(predictions(womenlf.2, newdata = new)) %>%
  select(group, estimate, children, hincome)
```

```
Re-fitting to get Hessian
```

group	estimate	children	hincome
not.work	0.2119801	absent	10
not.work	0.6590193	present	10
not.work	0.4415146	absent	30
not.work	0.8502980	present	30
not.work	0.6990959	absent	50
not.work	0.9434786	present	50
parttime	0.1888712	absent	10
parttime	0.1687698	present	10
parttime	0.2213518	absent	30
parttime	0.0835928	present	30
parttime	0.1533743	absent	50
parttime	0.0330006	present	50
fulltime	0.5991487	absent	10
fulltime	0.1722109	present	10
fulltime	0.3371336	absent	30
fulltime	0.0661092	present	30
fulltime	0.1475298	absent	50
fulltime	0.0235209	present	50

This way, you get "long" predictions, with a column called `group` that says which response category it is a prediction for. This is, however, not easy to read, because for your desired combination of `children` and `hincome`, you have to go hunting for the three probabilities of the three response categories, `not.work`, `parttime`, and `fulltime`. It is even more of a challenge to assess, say, the effects of the husband's income (next question) with the predictions laid out this way. Try it.

(18) (2 points) According to Figure 15, do married women with children in the household have a greater or lesser tendency to work outside the home, compared to women without children in the household, all else equal? Explain briefly.

Look at Figure 15, and compare two rows where `children` is different but `hincome` is the same, such as the first and second rows (or the fifth and sixth). Whichever pair of rows you pick, the probability of a married woman not working outside the home is much *higher* if there are children (`present`) compared to if there are not (`absent`). Also, the probability of the woman working full time is much *lower* if there are children present vs. if there are not. Hence, women with children have a much *lesser* tendency to work outside the home. (For full credit, compare *both* the probability of not working *and* of working full time.)

Be sure to have a statement that actually answers the question (such as my last sentence).

The three probabilities in each row add up to 1 (a woman with the values of `children` and `hincome` must fall into one of the three `partic` categories), so if one of the three probabilities is bigger, one of the others must be smaller. This points you towards an overall conclusion about where on the scale from less employment outside the home to more employment a woman sits.

Points:

- 2: compare suitable pair of rows (eg 1st and 2nd); by comparing corresponding predictions in those two rows, argue that woman has lesser tendency to work outside the home.
- 1.5: comparing only one pair of corresponding predictions (argument is stronger if you compare `not.work` and also (say) `fulltime`)
- 1: compare suitable pair of rows without making overall conclusion about tendency to work outside the home.
- 1: get partway towards an answer, or come to the right conclusion without making a clear enough argument. (For example, compare two outcome categories such as `not.work` becomes more likely and `fulltime` becomes less likely as husband's income increases, to make sure your argument is consistent.)

(19) (2 points) According to Figure 15, what is the effect of the husband's income on the tendency of a married woman to work outside the home, all else equal? Explain briefly.

Look again at Figure 15, and compare rows where `hincome` is different but `children` is the same, such as the first, third, and fifth rows. As the husband's income increases, the predicted probability of a woman not working outside the home increases, and the probability of the woman working full time decreases. (It is not clear what is happening to the probability of working part time, but if you are looking at the cases where children are present, that goes down with husband's income as well.)

Hence, if the husband's income is larger, there is a lesser tendency for the woman to work outside the home.

This is the same direction of effect as for the presence of children, but apparently not so big, which is consistent with the P-value for `children` in Figure 14 being the smallest.[1]

Points:

- 2: compare relevant rows of predictions (eg., 1st, 3rd, 5th); by comparing corresponding predictions in those rows, say that as husband's income increases, woman has lesser tendency to work outside the home.
- 1: compare suitable pair of rows without making overall conclusion about tendency to work outside the home.
- 1: get partway towards an answer, or come to the right conclusion without making a clear enough argument.

Extra: I didn't ask about `plot_predictions`, because showing you the output from that would take away from the prediction questions here, but now that we have answered those, we can think about `plot_predictions`. The tricky part of these for this kind of model is what goes in `condition` and in what order. There are more things to consider than you might think:

- `hincome`, quantitative (so this would be a good candidate to go on the x-axis)
- `children`, categorical
- but also `group` (see the question about `pivot_wider`, in particular my Extra), categorical.

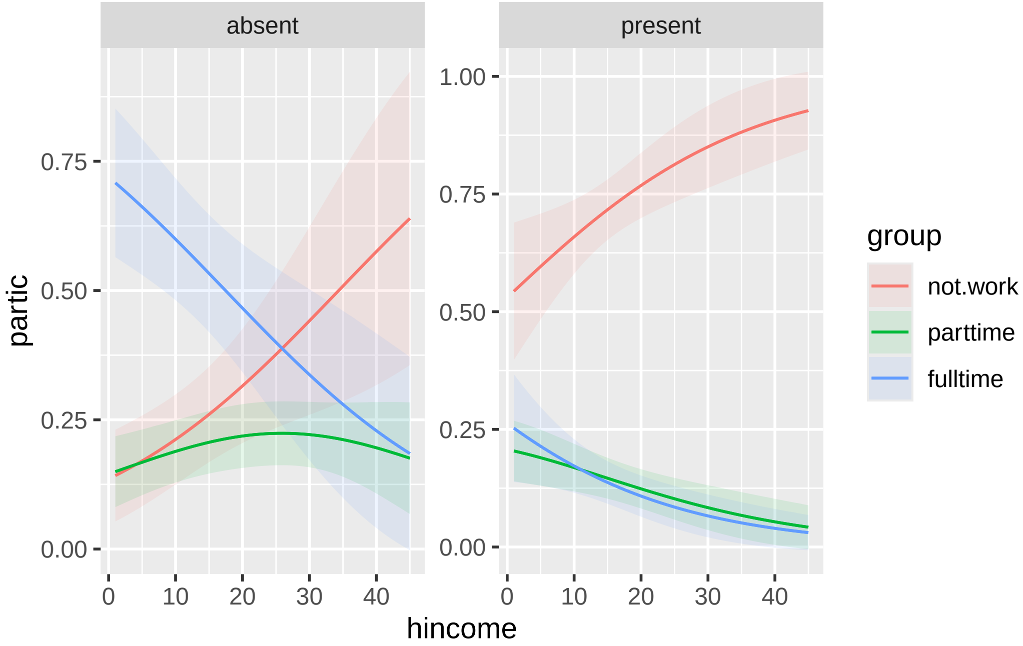
The `group` represents an outcome category, so it generally makes sense to have this as colour with this kind of model (that is, the second thing in `condition`), and that leaves the other categorical variable `children` to be facets (the third thing in `condition`). There is no `pivot_wider` needed any more (the `colour` takes care of that):

```r
plot_predictions(womenlf.2, condition = c("hincome", "group", "children"))
```

```
Re-fitting to get Hessian
```

[1] I don't usually like to compare P-values with each other, but the idea here is that if the P-value is smaller, the effect is larger.

```
Ignoring unknown labels:
* linetype : "group"
```



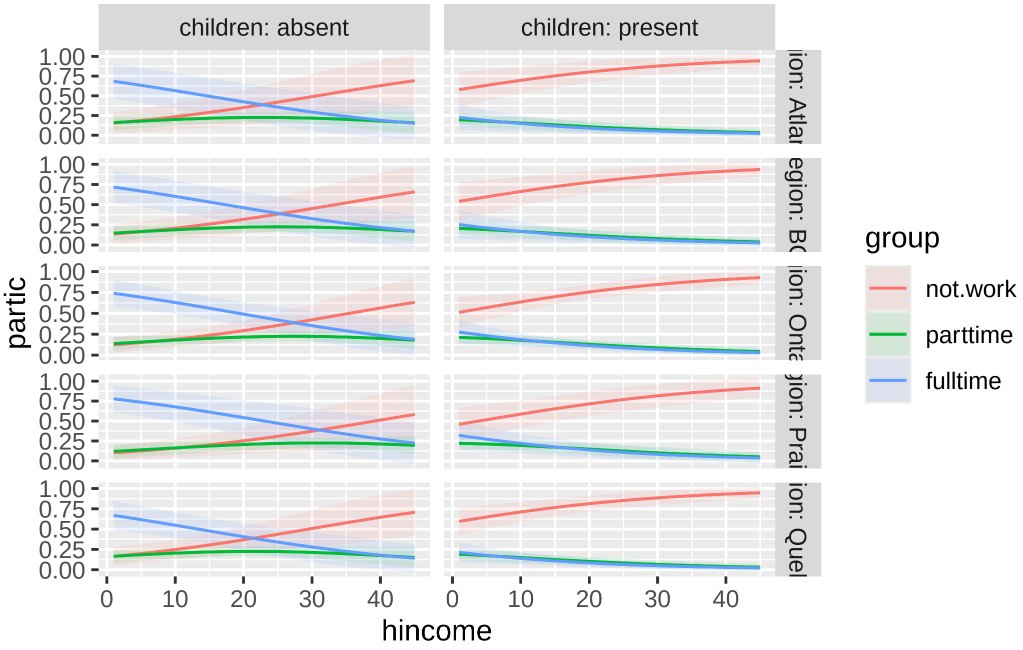
The message here is the same one that we got from the predictions in Figure 15: in both facets, as husband's income increases, a woman becomes more likely to not work outside the home, and less likely to work full-time. Comparing the left and right facets, if there are children in the household, a woman is more likely to not work outside the home, and is a lot less likely to work full-time.

If we had needed to keep `region` in the model, this would have had to go in `condition` as well. What `plot_predictions` does in this case is to have facets both up and down and across the page, in the way that `facet_grid` does (in fact, it uses `facet_grid` to make the graph):

```
plot_predictions(womenlf.1, condition = c("hincome", "group", "children", "region"))
```

```
Re-fitting to get Hessian
```

```
Ignoring unknown labels:
* linetype : "group"
```



With so many facets, it becomes harder to see what is going on. But the effects of `hincome` and `children` are the same as above for any region, and in fact there seems to be very little difference among the regions (the five plots in one column of facets look almost the same as each other). This is telling us that any effect of region is very small, consistent with the fact that `region` was not significant in model `womenlf.1`.

Plasma cell neoplasm

Plasma cell neoplasm is a condition in which the body produces too many white blood cells. People can be diagnosed with a "monoclonal gammopathy of undetermined significance" or MGUS, which is not a problem in itself. People with MGUS are examined from time to time by a doctor to see whether anything problematic has developed. One problem that can occur (in people with MGUS) is a "plasma cell malignancy" (PCM), which is a cancer of the white blood cells. If PCM is observed, it needs to be treated, but people with MGUS may never develop PCM.

1338 observations on patients with MGUS were collected (over many years). The data are shown in Figure 16, and descriptions of the variables are given in Figure 17. The dataframe is called `mgus2`. We are interested in the time it takes for PCM to develop (if it does) after diagnosis with MGUS, and how this time depends on other variables.

(20) (2 points) What code will create a suitable response variable for a Cox proportional hazards model?

Typically, in a survival model, we are modelling time until death, so you might be tempted to write `Surv(futime, death == "died")`. But that is not what we care about here; we are interested in the time until PCM, which is in the column `ptime`, and whether or not PCM was observed (`pstat`):

```
mgus2 %>% mutate(y = Surv(ptime, pstat == 1)) %>%
  select(y) %>%
  slice(1:20)
```

y
30+
25+
46+
92+
8+
4+
151+
2+
57+
136+
2+
108+
10+
14+
18+
43+
34+
67+
16+
62+

All you need is the `Surv` bit with the right two variables (and the right value for the second one) in it; I just did it this way to demonstrate that my code works without showing you all 1384 values. (You see that all of these observations were “censored”.)

This also works:

```
mgus2 %>% mutate(y = Surv(ptime, pstat)) %>%
  select(y) %>%
  slice(1:20)
```

y
30+
25+
46+
92+
8+
4+
151+
2+
57+
136+
2+
108+
10+
14+
18+
43+
34+
67+
16+
62+

because 1 is logically equivalent to TRUE (as in the dancing example in lecture). But if you go this way, you need to say something about why it will work (because otherwise it looks as if you are guessing).

Points:

- 2: `Surv(ptime, pstat == 1))`
- 2: `Surv(ptime, pstat)` along with explanation of why it will work
- 1.5: `Surv(ptime, pstat)` without explanation
- 1: error, such as wrong value for `pstat`
- 0.5: `Surv(futime, death == "died")`
- 0.5: non-trivial progress to solution without being worth more points

My guideline was: if you messed up one of the inputs to Surv, 1 point; if you messed up both, 0.5.

Extra 1: if you go back and look at the data in Figure 16, you'll see that ptime and futime are often the same. In some of these cases, this is because the patient died without having developed PCM; that is to say, they died *of something else*. It is usually only for patients who developed PCM (with pstat equal to 1) that ptime and futime are different. In those cases, patients were diagnosed with PCM, were followed up for a while, and then either died or the PCM became under control and they lived.

Extra 2: these are the mgus2 data from the survival package. One of the hazards of this kind of data is all the technical terms. I had to learn about what MGUS and PCM are: I learned about MGUS at the Mayo Clinic website and about plasma cell neoplasms at the US National Cancer Institute website. It seems to me that these are reliable websites, at least at the time of writing.[2] Then I had to simplify it enough to go into an exam question. On the other hand, as John Tukey says, the best thing about being a statistician is that you get to play in everyone's backyard, and thus get to learn a little about a lot of things. (You might recognize the picture on that website.)

(21) (2 points) Imagine that you were able to look at the output from your code of the previous question. Some of the values of the response variable would be displayed as a number followed by a plus sign. What does that mean, in the context of the data?

These are censored observations. That means that these patients were never observed to develop PCM (the event of interest).

Points:

- 2: patients were never observed to develop PCM
- 1: patients were never observed to die
- 1: "censored observations" without explaining what that means here

Extra: this is actually most of the patients, which suggests that it may take a long time to reach a appreciable probability of developing PCM:

```
mgus2 %>% count(pstat)
```

[2] What you should *not* do for this kind of work is to use an LLM; in that case, you have *no idea* whether the descriptions it gives you are correct, because the LLM will do no more than produce an answer that looks plausible on first reading. An LLM has *no mechanism* to give you an answer that will actually *be* correct.

pstat	n
0	1226
1	112

Only 112 of the 1338 patients developed PCM at all.[3] Quite a few of the patients were old to begin with:

```
ggplot(mgus2, aes(x = age)) + geom_histogram(bins = 10)
```

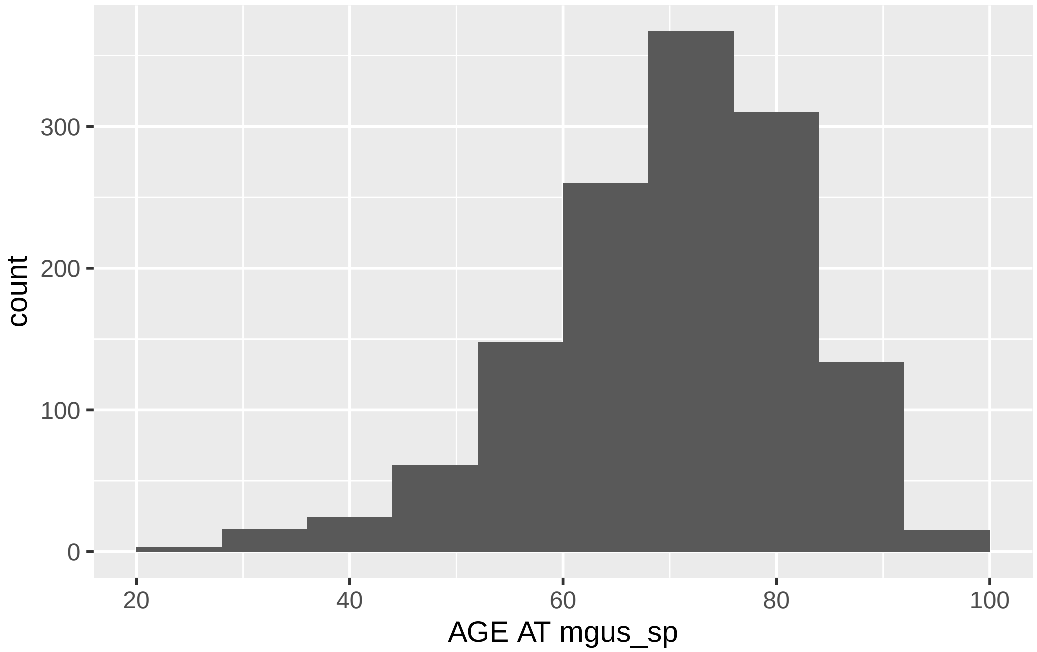


Figure 16 is deceiving in that it really displays too many patients who got PCM; the story is better summarized here:

```
mgus2 %>% count(pstat, death) %>%
  pivot_wider(names_from = death, values_from = n)
```

[3] For the data display in Figure 16, I deliberately displayed a larger fraction of the patients who developed PCM so that you could see what was happening.

pstat	died	lived
0	838	388
1	100	12

If a patient died (`died` column), it was mostly likely not from PCM, but if a patient got PCM (`1` row), they were very likely to die during the study period (presumably from the PCM).

(22) (3 points) Output from a Cox model is shown in Figure 18. What kind of effect, if any, do each of these two explanatory variables have? Explain briefly.

The first thing to check is that both explanatory variables are significant. They both are, with P-values 0.0088 and 3.4×10^{-8} respectively. So it makes sense to talk about the effects of both of them. (If one of them had *not* been significant, that explanatory variable would have been inferred to have no effect.) This was the reason for the words "if any" in the question. Read carefully.

`hgb` has a negative coefficient, so a larger value of `hgb` (hemoglobin) means that a patient has less hazard of developing PCM: that is, such a patient has a high probability of taking a long time to develop PCM.

`mspike` has a positive coefficient, so a larger value of `mspike` (the size of the monoclonal serum spike) means that a patient has greater hazard of developing PCM: that is, they are more likely to develop it more quickly.

The "event" here is "develops PCM", not "death", so in this case "survival" has the specific meaning of "has not developed PCM (yet)".

Points:

- 3: all of: check both explanatory variables for significance; patient with high `hgb` will take longer to develop PCM; patient with larger value of `mspike` likely to get PCM quicker.
- 2: proper discussion of effects of `hgb` and `mspike` but not of their significance
- 2: two other of the items for 3 points
- 2: proper discussion of effect of coefficients on hazard rates but not linking that to chance of developing PCM
- 1: proper discussion of only one of the items for 3 points (eg. the significance).
- 1: proper discussion of hazard rates for both explanatory variables but not linking to chance of developing PCM, and no discussion of significance

Extra: I got to this model by first fitting a model with more explanatory variables (all the ones in the original data, including some that I didn't tell you about):

```
mgus2.1 <- coxph(Surv(ptime, pstat == 1) ~ age + sex + hgb + creat + mspike,
                 data = mgus2)
summary(mgus2.1)
```

```
Call:
coxph(formula = Surv(ptime, pstat == 1) ~ age + sex + hgb + creat + 
    mspike, data = mgus2)

  n= 1338, number of events= 112 

            coef exp(coef)  se(coef)      z Pr(>|z|)    
age     0.011171  1.011234  0.008533  1.309   0.1905    
sexM    0.098700  1.103736  0.205748  0.480   0.6314    
hgb    -0.134653  0.874019  0.055565 -2.423   0.0154 *  
creat  -0.145029  0.864997  0.181878 -0.797   0.4252    
mspike  0.912379  2.490240  0.165093  5.526 3.27e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

       exp(coef) exp(-coef) lower .95 upper .95
age        1.011     0.9889    0.9945    1.0283
sexM       1.104     0.9060    0.7374    1.6520
hgb        0.874     1.1441    0.7838    0.9746
creat      0.865     1.1561    0.6056    1.2355
mspike     2.490     0.4016    1.8018    3.4417

Concordance= 0.67  (se = 0.031 )
Likelihood ratio test= 38.98  on 5 df,   p=2e-07
Wald test            = 41.23  on 5 df,   p=8e-08
Score (logrank) test = 41.82  on 5 df,   p=6e-08
```

There are several non-significant variables there, so I could either remove them one by one or use `step`. I went the latter route:

```
mgus2.2 <- step(mgus2.1)
```

```
Start:  AIC=1364.46
Surv(ptime, pstat == 1) ~ age + sex + hgb + creat + mspike

         Df    AIC
- sex     1 1362.7
- creat   1 1363.6
- age     1 1364.2
<none>      1364.5
- hgb     1 1368.0
- mspike  1 1391.8

Step:  AIC=1362.69
Surv(ptime, pstat == 1) ~ age + hgb + creat + mspike

         Df    AIC
- creat   1 1361.7
- age     1 1362.4
<none>      1362.7
- hgb     1 1366.2
- mspike  1 1390.1

Step:  AIC=1361.67
Surv(ptime, pstat == 1) ~ age + hgb + mspike

         Df    AIC
- age     1 1361.4
<none>      1361.7
- hgb     1 1364.5
- mspike  1 1389.5

Step:  AIC=1361.43
Surv(ptime, pstat == 1) ~ hgb + mspike

         Df    AIC
<none>      1361.4
- hgb     1 1366.0
- mspike  1 1388.7
```

It turns out, perhaps surprisingly, that neither age nor sex have any impact on the time to develop PCM; the only things that do are the ones that appear in Figure 18, though the decision about whether to keep or remove `age` was a close one.

(23) (2 points) A plot is shown in Figure 19. Explain briefly how this plot supports one of your conclusions from the previous question.

These are predicted survival curves for various values of hemoglobin. Patients with the highest value of hemoglobin (the purple[4] curve) have the highest chance of not developing PCM for the longest time (best "survival"), while patients with the lowest value of hemoglobin (the red curve) are most likely to develop PCM soonest. This is the same conclusion that we drew about hemoglobin in the previous question.

I was prepared to be more relaxed here if you thought earlier that "survival" meant "not dying" rather than "not getting PCM".

Points:

- 2: best "survival" (longest time to develop PCM) goes with highest value of hemoglobin (purple/pink curve) (or worst goes with lowest (red curve)); this is same conclusion as about `hgb` in previous question.
- 1: discussion of survival curves without linking to previous question
- 0.5: some relevant comment but not enough for 1

Extra 1: I can't very well stop you doing this question first and the previous one second, if it's easier for you that way around. You might look at Figure 19 and see that a higher value of `hgb` goes with better "survival" (which in this case is not developing PCM), because the purple curve is at the top right, and *then* go back and ask yourself how that is consistent with the negative coefficient. This also shows the value of reading the whole exam before you do any questions: by doing the questions "out of order", you might make it easier for yourself.

Extra 2: I realized that the default y-axis label gives away the answer to the earlier question about the response variable for the model, so I used `labs` to change the text displayed on the y-axis.

Battery

A battery has been designed to operate under extreme temperature conditions. Three types of plate material are being considered. An experiment is designed to compare the lifetime of the battery for each of the types of plate material (denoted `M1` through `M3`) and for each of three different temperatures: an extremely low one, `Low`, a moderate one, `Med`, and an extremely high one, `Xtr`. For four replications of each combination of temperature and material, the battery life is measured, in hours of operation. The data, in dataframe `battery`, are shown in Figure 20. A larger value of `Life` is better.

[4] Or maybe you think this is pink.

(24) (2 points) A grouped boxplot is shown in Figure 21. What does this plot tell you about the likely presence or absence of an interaction between temperature and material? Explain briefly.

Make a well-reasoned point about whether or not there is a consistent effect of `Material` at each `Temperature`. For example, at a `Low` temperature, the three materials are about the same in terms of `Life`. But at a `Med` temperature, the materials are very different, with `M1` being a lot worse than the other two. One point for this kind of discussion.

The second point is for saying that because there appears *not* to be a consistent effect of `Material` (it's different for each `Temperature`), we would expect to see an interaction.

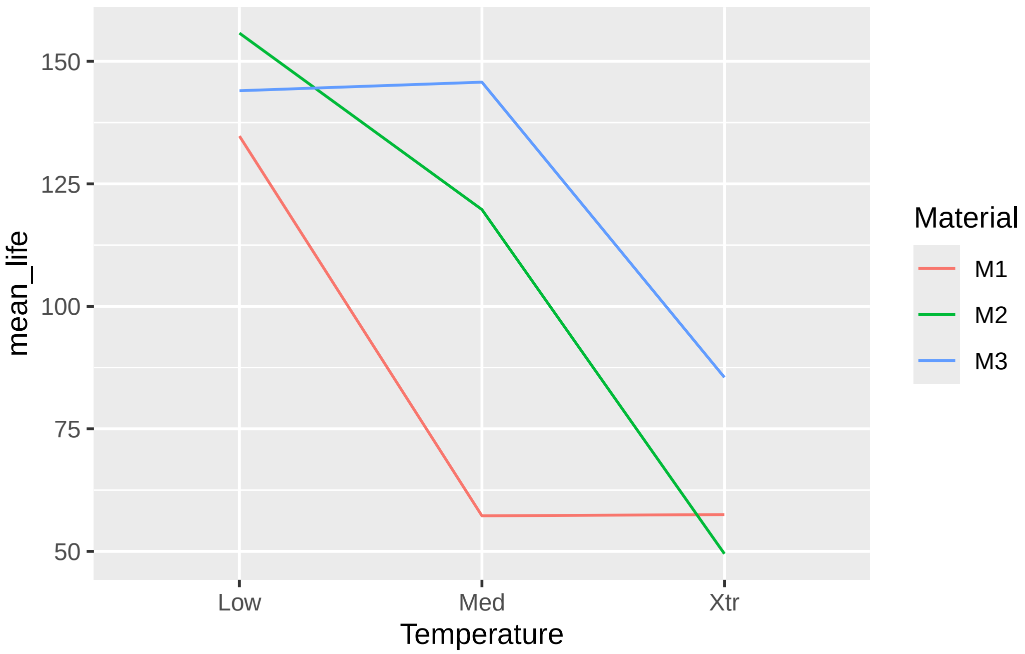
Points:

- 2: there is not a consistent effect of material at each temperature; therefore we expect to see an interaction
- 1: one of the two things for 2 points, but not connecting them

Extra: this is not an interaction plot, so you cannot immediately talk about traces being parallel. The way to convert this plot to an interaction plot is to imagine lines drawn through the median of each box, joining up the boxes of the same colour with a line of that colour. The red line would go down and across, the green line would keep going down, and the blue line would go across and down:

```r
battery %>%
  group_by(Temperature, Material) %>%
  summarize(mean_life = mean(Life)) %>%
  ggplot(aes(x = Temperature, y = mean_life,
             colour = Material, group = Material)) + geom_line()
```

```
`summarise()` has regrouped the output.
i Summaries were computed grouped by Temperature and Material.
i Output is grouped by Temperature.
i Use `summarise(.groups = "drop_last")` to silence this message.
i Use `summarise(.by = c(Temperature, Material))` for per-operation grouping
  (`?dplyr::dplyr_by`) instead.
```



These lines are definitely not parallel, so we would on this basis expect to see an interaction effect as well.[5]

Extra extra: I labelled the highest temperature `Xtr` so that the temperatures would display in a logical order on the boxplot without my having to do anything else: in alphabetical order, the temperatures go from low to high.

(25) (2 points) Some analysis is shown in Figure 22. What do you conclude from this Figure?

Look at the interaction term first. It is significant (P-value 0.019), so that is the finding: there is a significant interaction between temperature and material (or, the effect of temperature on battery lifetime is different for each material, or the other way around).

This ought to be the same thing that you concluded from the boxplot (if it is not, you are invited to figure out why not). The idea is that even allowing for the variability, those lines on my interaction plot are definitely not parallel.

It is an error to talk about main effects in the presence of a significant interaction. The significant interaction shows that the effect of material is different for each temperature, so it makes no sense to talk about "an" effect of material that holds for all temperatures.

[5] There is the technical detail of the boxplot showing medians and the interaction plot showing means, but the boxplots suggest that the means and medians are unlikely to be too far different.

Points:

- 2: significant interaction citing P-value; "there is a significant interaction between temperature and material" or "the effect of material is different for each temperature"
- 1.5: as 2 points, but without giving P-value (or showing that it is small)
- 1: as 2 points, but discussing main effects as well
- 0.5: as 1.5 points, but discussing main effects as well
- 0.5: discussing main effects *without* discussing interaction
- 0.5: other relevant commentary, but not enough for 1

(26) (2 points) What feature of the data made it possible to test for an interaction effect in Figure 22?

The fact that there was replication: there were four observations made for each combination of `Temperature` and `Material`. (This is something to remember from your second stats course.) "Replication" is the key word, but you need to show that you know what that means *for these data*.

Points:

- 2: repeated observations for each combination of temperature and material, or "there are replicates" plus something that shows that you know what "replicates" means
- 1: "there are replicates" with missing or incomplete discussion of what that means here
- 1: something relevant but comments unclear
- 0.5: something relevant but not enough for 1 point

The fact that R gave us a test for interaction here does not explain *why* it was able to do that. That's why I said "feature of the data"; go back and look at Figure 20 and try to figure it out from there. (That's actually why I showed you the whole dataset: so that it was clear that each combination of temperature and material was tested more than once.)

Additional discussion: in order to be able to assess whether the effect of material is different for each temperature, you need to have repeated observations for each combination of the two explanatory variables. Another way to say this is, if you don't have that, you run out of degrees of freedom for error. For data like these, one observation for each combination of material and temperature would give us $n = 9$ so 8 total degrees of freedom; temperature and material each have two degrees of freedom and so the interaction has $2 \times 2 = 4$. This leaves no degrees of freedom for error.

What you have to do in a randomized block design with no replication is to *assume* that there is no interaction, and then you get tests for the main effects, with the degrees of freedom that would have gone into the interaction being used for error.

(27) (2 points) Some further analysis is shown in Figure 23. Why is it appropriate to run this kind of analysis here?

These are (two of the three) simple effects of material at different temperatures. This is appropriate to run because the interaction is significant: the effect of material is going to be different at the different temperatures.

A minimal answer is "these are simple effects because the interaction is significant". Show that you know what you are looking at and why this analysis is being run.

Points:

- 2: these are simple effects because interaction is significant
- 1: these are simple effects but not saying why we are running them
- 1: this kind of analysis is run when the interaction is significant, but without naming that it is simple effects
- 0.5: other possibly relevant comments

(28) (3 points) What do you conclude from Figure 23? Explain briefly.

There are three bits of output, one point for interpretation of each:

- when temperature is low, there are no differences among the materials (P-value 0.686)
- when temperature is medium, the materials do not all have the same mean lifetime (P-value 0.00047)
- at medium temperature, material `M1` has a significantly different (lower) mean lifetime than the other two materials, which are not significantly different from each other (the Tukey analysis).

The three marks are a hint that you need to say three things.

Points:

- 3: all of: temperature low: no differences among materials; temperature medium: there is an effect of materials (means are not all the same); at medium temperature, material M1 has lower mean lifetime than the other two materials
- 2.5: as 3, but saying only that material M1 has "different" mean lifetime than the others
- 2: analysis of the two F-tests but not of the Tukey
- 2: missing analysis of F-test for medium temperature (you only look at the Tukey if the F-test is significant, so need to establish that first)
- 1: something relevant, but analysis confused (for example, not saying that it matters which temperature you are looking at)
- 0.5: other relevant comment

Extra: There is no Tukey analysis for low temperatures because there are no differences among the materials to find. Also, I didn't show you the simple effects for temperature `Xtr` because (i) the question was long enough already and (ii) I found the result surprising given Figure 21:

```
battery %>%
  filter(Temperature == "Xtr") -> d
d.3 <- aov(Life ~ Material, data = d)
summary(d.3)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
Material     2   2859  1429.3    2.93  0.105
Residuals    9   4391   487.9
```

There are actually no significant differences among materials at the extremely high temperature, even though the boxplot suggests that at least `M2` and `M3` will be different. This is possibly because each boxplot is only based on four observations (the four replicates at that temperature and material), so we are not dealing with large sample sizes.

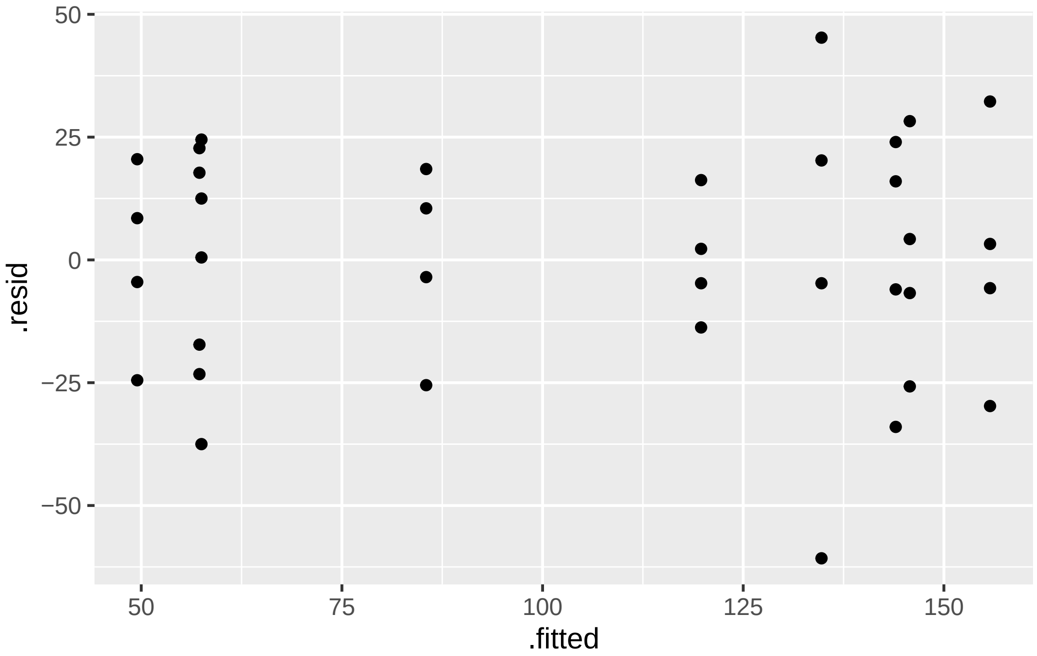
Extra 2: the other thing I might have asked you about is residuals. The boxplots, being based on only four observations each, don't give a very clear picture about normality, equal spreads and the like:

```
battery.2 <- lm(Life ~ Temperature * Material, data = battery)
anova(battery.2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Temperature	2	39118.722	19559.361	28.967692	0.0000002
Material	2	10683.722	5341.861	7.911372	0.0019761
Temperature:Material	4	9613.778	2403.444	3.559535	0.0186112
Residuals	27	18230.750	675.213	NA	NA

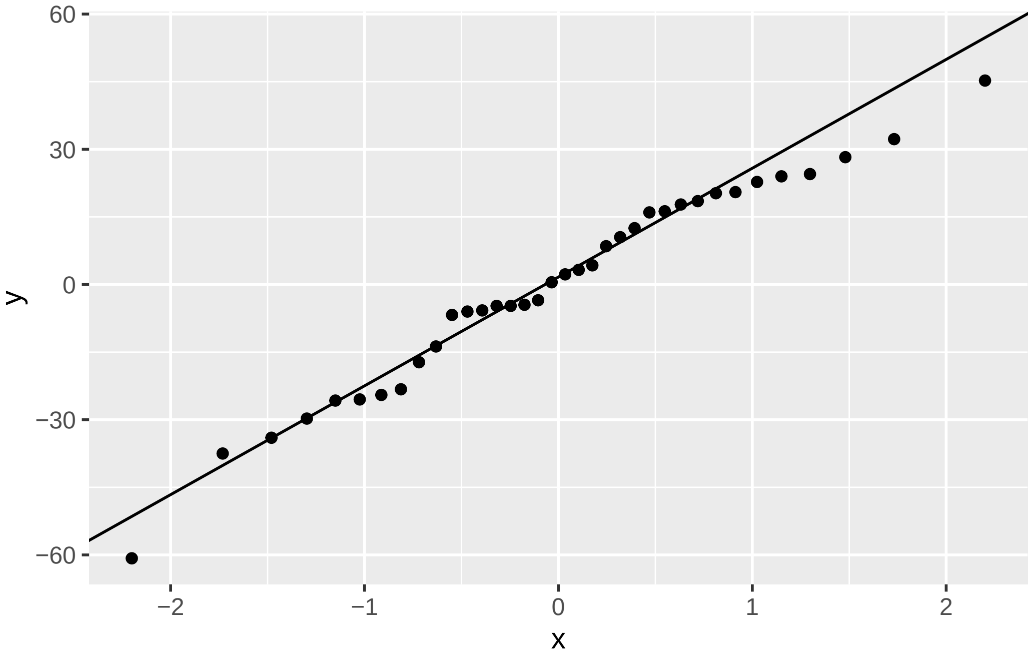
(the same result as before) and

```
library(broom)
battery.2 %>% augment(battery) -> battery.2a
ggplot(battery.2a, aes(x = .fitted, y = .resid)) + geom_point()
```



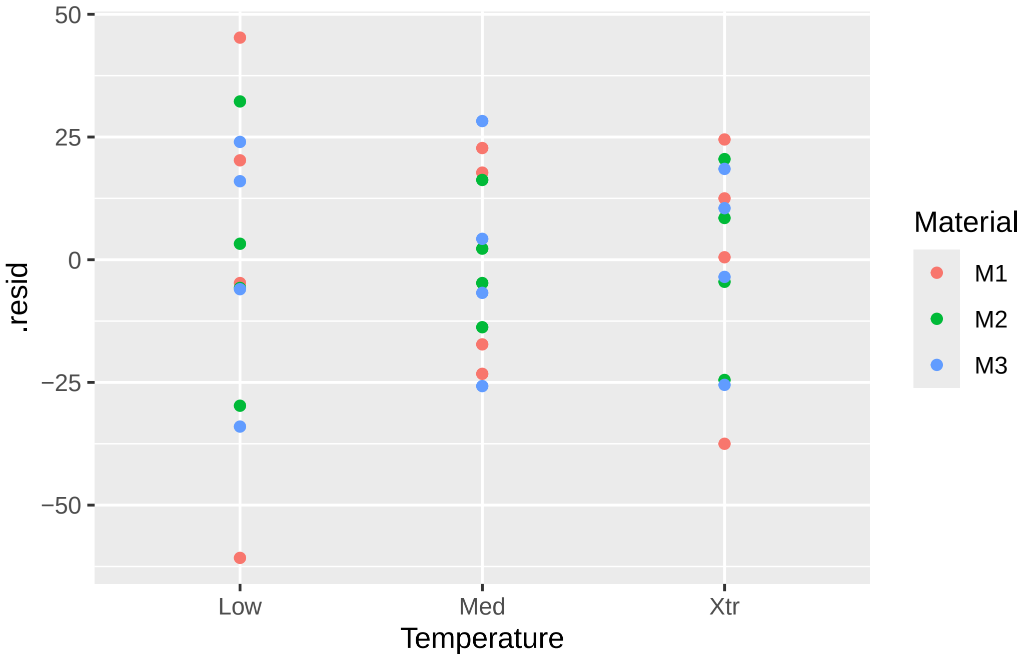
This to me is near enough a random pattern, not enough to call "fanning out".

```
ggplot(battery.2a, aes(sample = .resid)) + stat_qq() + stat_qq_line()
```



This is close to normal, with only the smallest hint of a left skew.

```
ggplot(battery.2a, aes(x = Temperature, y = .resid, colour = Material)) +
  geom_point()
```



For this one, I was trying to get a hint of whether the residuals for temperature or material or their combination had any pattern. I think the residuals for low temperature and/or for material M1 are more spread out than the others. The red dots seem to be nearer the top and bottom, and the dots for low temperature may seem to be more spread out because of the M1 residuals there.

Overall I'd say the residual picture is pretty good.

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

Figures

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

Figure 1: Packages loaded

```r
stackloss
```

Air.Flow	Water.Temp	Acid.Conc.	stack.loss
80	27	89	42
80	27	88	37
75	25	90	37
62	24	87	28
62	22	87	18
62	23	87	18
62	24	93	19
62	24	93	20
58	23	87	15
58	18	80	14
58	18	89	14
58	17	88	13
58	18	82	11
58	19	93	12
50	18	89	8
50	18	86	7
50	19	72	8
50	19	79	8
50	20	80	9
56	20	82	15
70	20	91	15

Figure 2: Stack loss data

```
stackloss %>%
  pivot_longer(-stack.loss, names_to = "xnames", values_to = "xvals") %>%
  ggplot(aes(x = xvals, y = stack.loss)) + geom_point() +
  facet_wrap(~ xnames, scales = "free", ncol = 2)
```

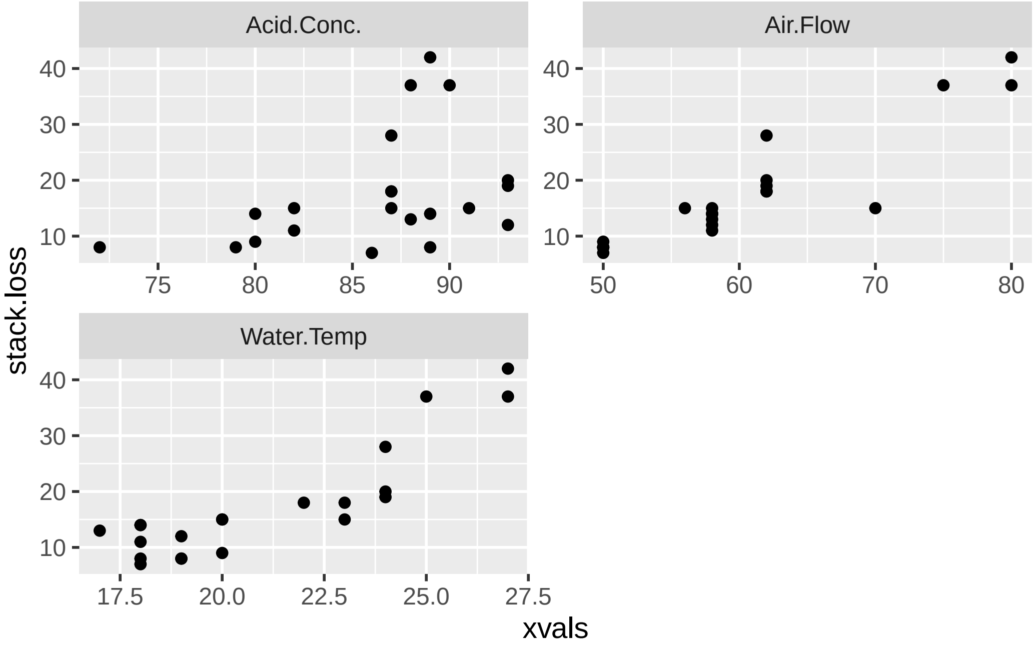


Figure 3: Graphs of stack loss data

```r
stackloss.1 <- lm(stack.loss ~ Air.Flow + Water.Temp + Acid.Conc., data = stackloss)
summary(stackloss.1)
```

```

Call:
lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.,
    data = stackloss)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.2377 -1.7117 -0.4551  2.3614  5.6978 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -39.9197    11.8960  -3.356  0.00375 ** 
Air.Flow      0.7156     0.1349   5.307  5.8e-05 ***
Water.Temp    1.2953     0.3680   3.520  0.00263 ** 
Acid.Conc.   -0.1521     0.1563  -0.973  0.34405    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.243 on 17 degrees of freedom
Multiple R-squared:  0.9136,	Adjusted R-squared:  0.8983 
F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09
```

Figure 4: Regression 1 of stack loss data

```r
stackloss.2 <- lm(stack.loss ~ Air.Flow + Water.Temp, data = stackloss)
summary(stackloss.2)
```

```

Call:
lm(formula = stack.loss ~ Air.Flow + Water.Temp, data = stackloss)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.5290 -1.7505  0.1894  2.1156  5.6588 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -50.3588     5.1383  -9.801 1.22e-08 ***
Air.Flow      0.6712     0.1267   5.298 4.90e-05 ***
Water.Temp    1.2954     0.3675   3.525  0.00242 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.239 on 18 degrees of freedom
Multiple R-squared:  0.9088,	Adjusted R-squared:  0.8986 
F-statistic: 89.64 on 2 and 18 DF,  p-value: 4.382e-10
```

Figure 5: Regression 2 of stack loss data

estimate	conf.low	conf.high	Air.Flow	Water.Temp
6.515207	4.378133	8.652281	50	18
17.112805	15.725987	18.499623	60	21

Figure 6: Stack loss data predictions.

fert	x
0.0000000	0
0.1315789	0
0.2631579	0
0.3947368	0
0.5263158	1
0.6578947	0
0.7894737	1
0.9210526	0
1.0526316	1
1.1842105	0
1.3157895	1
1.4473684	1
1.5789474	1
1.7105263	1
1.8421053	1
1.9736842	1
2.1052632	1
2.2368421	1
2.3684211	1
2.5000000	1

Figure 7: Seed germination data

```
ggplot(germination, aes(x = x, y = fert, group = x)) + geom_boxplot()
```

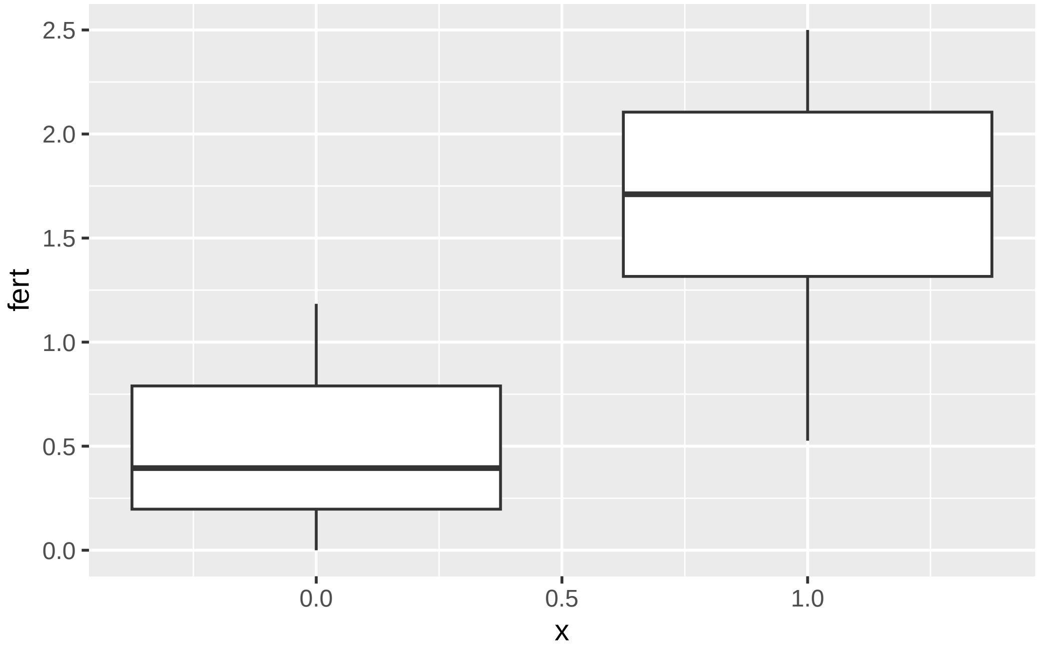


Figure 8: Seed germination boxplot

```
germination.1 <- glm(x ~ fert, data = germination, family = "binomial")
summary(germination.1)
```

```

Call:
glm(formula = x ~ fert, family = "binomial", data = germination)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.231      1.651  -1.957   0.0504 .
fert           3.811      1.687   2.259   0.0239 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25.898  on 19  degrees of freedom
Residual deviance: 12.412  on 18  degrees of freedom
AIC: 16.412

Number of Fisher Scoring iterations: 6
```

Figure 9: Seed germination logistic regression

	partic	hincome	children	region
139	not.work	28	present	Ontario
148	not.work	19	present	BC
29	not.work	17	present	Prairie
244	fulltime	6	present	Quebec
186	parttime	15	present	BC
190	not.work	23	present	BC
112	fulltime	17	present	Atlantic
36	not.work	19	absent	Ontario
76	parttime	38	present	Ontario
85	parttime	23	present	Ontario
197	fulltime	5	present	Prairie
37	not.work	15	present	Ontario
243	not.work	17	present	Quebec
203	not.work	15	present	Quebec
246	parttime	19	present	Quebec
87	fulltime	16	present	BC
39	not.work	9	present	Atlantic
8	fulltime	7	present	Ontario
127	not.work	17	present	Atlantic
202	not.work	13	present	Quebec

Figure 10: Women in labour force data (20 randomly chosen rows). The number in the left-hand column (without a column heading) is an ID for the woman interviewed.

- `partic`: participation in work outside the home: `fulltime` or `parttime` work, or `not.work` (not working outside the home).
- `hincome`: husband's income in thousands of (1977) dollars
- `children`: whether the household had one or more children (`present`) or no children (`absent`). (The actual number of children was not recorded.)
- `region` of the country in which the woman lived, classified as `BC` (British Columbia), `Prairie` (Alberta, Saskatchewan or Manitoba), `Ontario`, `Quebec`, `Atlantic` (New Brunswick, Nova Scotia, Prince Edward Island, or Newfoundland).

Figure 11: Variables in labour force data as recorded

```
womenlf %>% count(partic)
```

partic	n
fulltime	66
not.work	155
parttime	42

Figure 12: Counts of women in each `partic` category

```
womenlf.1 <- polr(partic ~ hincome + children + region, data = womenlf)
drop1(womenlf.1, test = "Chisq")
```

	Df	AIC	LRT	Pr(>Chi)
	NA	455.7660	NA	NA
hincome	1	462.2948	8.528759	0.0034958
children	1	504.0503	50.284316	0.0000000
region	4	449.6630	1.896942	0.7547066

Figure 13: Some analysis of women in labour force data. `partic` has been modified so that `polr` will run appropriately.

```
womenlf.2 <- polr(partic ~ hincome + children, data = womenlf)
drop1(womenlf.2, test = "Chisq")
```

	Df	AIC	LRT	Pr(>Chi)
	NA	449.6630	NA	NA
hincome	1	455.8696	8.206679	0.0041736
children	1	498.3117	50.648770	0.0000000

Figure 14: Further analysis of women in labour force data.

children	hincome	not.work	parttime	fulltime
absent	10	0.2119801	0.1888712	0.5991487
present	10	0.6590193	0.1687698	0.1722109
absent	30	0.4415146	0.2213518	0.3371336
present	30	0.8502980	0.0835928	0.0661092
absent	50	0.6990959	0.1533743	0.1475298
present	50	0.9434786	0.0330006	0.0235209

Figure 15: Predictions for women in labour force

age	sex	hgb	mspike	ptime	pstat	futime	death
82	F	5.7	0.6	3	0	3	died
55	M	15.8	0.9	46	0	46	died
82	F	11.0	2.9	94	0	94	died
80	M	14.8	0.6	208	0	208	died
66	F	14.4	1.3	58	0	58	lived
86	F	14.5	2.4	57	0	57	lived
83	M	14.4	0.4	90	0	90	died
69	M	14.3	0.5	57	0	57	lived
79	F	14.8	0.5	50	0	50	died
53	M	17.9	1.6	31	0	31	died
76	F	15.4	1.5	128	1	133	died
78	M	12.1	2.1	12	1	44	died
66	M	14.5	1.9	80	1	111	died
74	M	15.2	1.7	35	1	91	died
60	F	13.4	2.1	84	1	122	died
76	F	14.0	0.7	9	1	49	died
68	M	14.2	1.4	34	1	35	died
62	F	13.5	0.5	91	1	164	lived
83	F	13.0	1.0	23	1	52	died
79	F	11.8	1.0	83	1	88	died

Figure 16: MGUS data (selected observations)

Variables shown in Figure 16:

- `age` in years
- `sex`: M (male), F (female)
- `hgb`: hemoglobin (suitable units)
- `mspike`: size of the monoclonal serum spike (suitable units)
- `ptime`: time from diagnosis of MGUS until PCM observed or last contact, in months
- `pstat`: whether PCM occurred (1) or not (0)
- `futime`: time from diagnosis until death or last contact, in months
- `death`: status of patient at last contact: `lived` or `died`

Figure 17: Variables in MGUS data

```
Call:
coxph(formula = Surv(ptime, pstat == 1) ~ hgb + mspike, data = mgus2)

  n= 1338, number of events= 112 

           coef exp(coef) se(coef)      z Pr(>|z|)    
hgb    -0.13111   0.87712  0.05002 -2.621  0.00876 ** 
mspike  0.91413   2.49461  0.16570  5.517 3.45e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

       exp(coef) exp(-coef) lower .95 upper .95
hgb       0.8771     1.1401    0.7952    0.9675
mspike    2.4946     0.4009    1.8029    3.4518

Concordance= 0.66  (se = 0.031 )
Likelihood ratio test= 36.01  on 2 df,   p=2e-08
Wald test            = 37.74  on 2 df,   p=6e-09
Score (logrank) test = 38.42  on 2 df,   p=5e-09
```

Figure 18: Cox model `mgus2.2` for MGUS data

```
plot_predictions(mgus2.2, condition = c("ptime", "hgb"), type = "survival") +
  labs(y = "Survival probability")
```

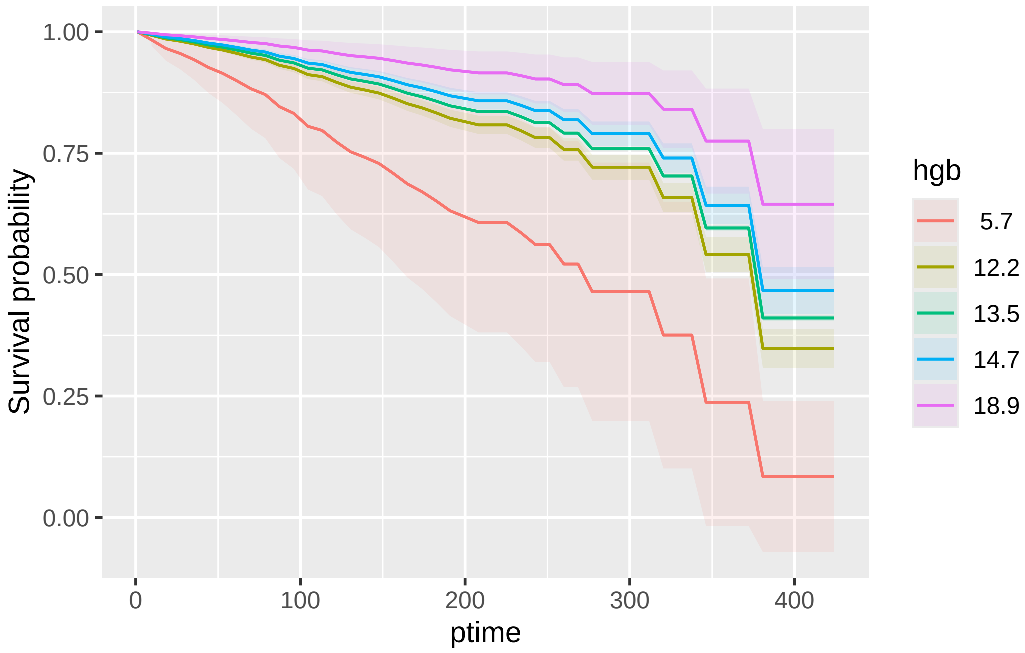


Figure 19: MGUS data plot (the `labs` is the text to go on the y axis)

```
battery
```

Life	Material	Temperature
130	M1	Low
74	M1	Low
155	M1	Low
180	M1	Low
150	M2	Low
159	M2	Low
188	M2	Low
126	M2	Low
138	M3	Low
168	M3	Low
110	M3	Low
160	M3	Low
34	M1	Med
80	M1	Med
40	M1	Med
75	M1	Med
136	M2	Med
106	M2	Med
122	M2	Med
115	M2	Med
174	M3	Med
150	M3	Med
120	M3	Med
139	M3	Med
20	M1	Xtr
82	M1	Xtr
70	M1	Xtr
58	M1	Xtr
25	M2	Xtr
58	M2	Xtr
70	M2	Xtr
45	M2	Xtr
96	M3	Xtr
82	M3	Xtr
104	M3	Xtr
60	M3	Xtr

Figure 20: Battery data

```
ggplot(battery, aes(x = Temperature, y = Life, fill = Material)) + geom_boxplot()
```

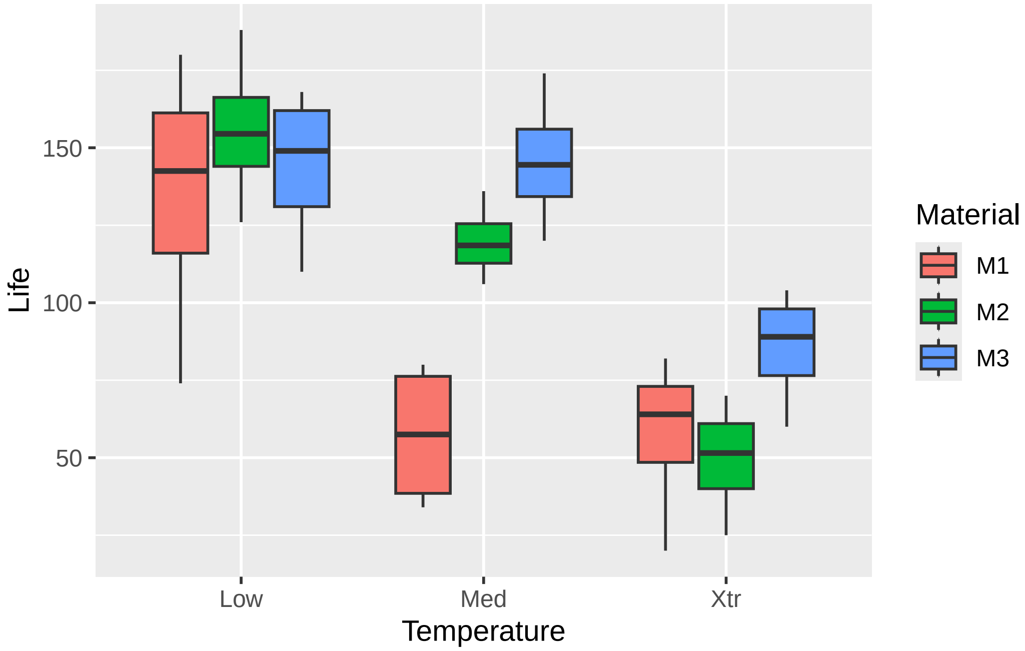


Figure 21: Battery data boxplot

```
battery.1 <- aov(Life ~ Temperature * Material, data = battery)
summary(battery.1)
```

```
                     Df Sum Sq Mean Sq F value   Pr(>F)    
Temperature           2  39119   19559  28.968 1.91e-07 ***
Material              2  10684    5342   7.911  0.00198 ** 
Temperature:Material  4   9614    2403   3.560  0.01861 *  
Residuals            27  18231     675                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 22: Battery data analysis 1

```
battery %>%
  filter(Temperature == "Low") -> d
d.1 <- aov(Life ~ Material, data = d)
summary(d.1)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
Material     2    886   443.1   0.392  0.686
Residuals    9  10164  1129.3
```

```
battery %>%
  filter(Temperature == "Med") -> d
d.2 <- aov(Life ~ Material, data = d)
summary(d.2)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
Material     2  16553    8276   20.26 0.000465 ***
Residuals    9   3676     408
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(d.2)
```

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

Fit: aov(formula = Life ~ Material, data = d)

$Material
      diff       lwr       upr     p adj
M2-M1 62.5  22.59911 102.40089 0.0045670
M3-M1 88.5  48.59911 128.40089 0.0004209
M3-M2 26.0 -13.90089  65.90089 0.2177840
```

Figure 23: Battery data analysis 2