University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 / STA 1007 (K. Butler), Midterm Exam
February 24, 2018

Aids allowed:

- My lecture overheads (slides)
- Any notes that you have taken in this course
- Your marked assignments
- My assignment solutions
- Notes from STAC32
- Non-programmable, non-communicating calculator

Before you begin, complete the signature sheet, but sign it only when the invigilator collects it. The signature sheet shows that you were present at the exam.

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

In addition, you should have an additional booklet of output to refer to during the exam. Contact an invigilator if you do not have this.

Answer each question in the space provided (under the question). If you need more space, use the backs of the pages, but be sure to draw the marker's attention to where the rest of the answer may be found.

The maximum marks available for each part of each question are shown next to the question part. In addition, the total marks available for each page are shown at the bottom of the page, and in the table on the next page.

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

Last name: _____

First name: _____

Student number: _____

For marker's use only:

Page	Points	Score
1	8	
3	9	
5	13	
7	5	
9	11	
11	6	
13	7	
16	4	
17	2	
21	2	
22	3	
Total:	70	

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

1. A common problem for hospitals is that their patients can get infections while in hospital (which means that the hospital has to treat the infection *and* whatever it is that the patient is in hospital for). There is a standard procedure for assessing the infection risk to a patient, resulting in a numeric (quantitative) score. 113 patients at different hospitals in different regions of the US were assessed; we will assess how the infection risk (`InfctRsk`) depends on the number of days the patient stayed in the hospital (`Stay`), how often they were given X-rays (`Xray`) and the `Region` of the US where the hospital is located (a numerical code: 1 is north-east, 2 is north-central, 3 is south and 4 is west).

   Some of the data set is shown in Figure 2.

   I don't know the units of `InfctRsk` (other than that a higher score is a greater infection risk) or of `Xray`.

   (a) (2 marks) There is an extra line of code at the top of Figure 3 (beyond what would normally be seen in fitting a multiple regression). What does it do, and why does it need to be there? Explain briefly.

   > **My answer:** This is turning the numerical variable `Region` into a factor (categorical variable), so that it will be handled as a grouping variable in the regression. This is necessary because the numbers 1–4 for the regions have no meaning as numbers; they are just labels to distinguish the regions. (`Region` is nominal rather than interval or ratio, if you like to use that jargon.)
   >
   > Somehow you need to get at the idea that the number values for `Region` are not really numbers because Region is really categorical and needs to be treated as such.
   >
   > Expect one point for saying what the code does, and another for saying something convincing about why Region needs to be a factor (treated as categorical) and currently isn't one. (In regression, explanatory variables can be quantitative or categorical, but if they are actually categorical and yet look like numbers, they will get treated as numbers unless we do this `factor` thing.)

   (b) (3 marks) A regression and its `summary` output is shown in Figure 3. Interpret the slope for `Stay`, whose value is 0.349.

   > **My answer:** A patient who stays one day longer in hospital is predicted to have an infection score 0.349 greater, all else being equal. (This is an ordinary slope, but note the "all else equal" because it's a multiple regression in which all else might *not* be equal.)
   >
   > How you say it is up to you, but you have to get at (i) what changes by 1, (ii) what increases by 0.349, (iii) "all else equal" (or "the other variables stay the same" or something similar). One point for each of those.
   >
   > I observe that this slope is positive, meaning that a patient who stays longer in hospital is at greater risk of infection. This (unfortunately) makes sense, since hospitals are good places to pick up infections. (An observation like this might be worth a point if you missed out on saying something else. I didn't quite feel it was fair to ask you whether the positiveness of the slope made practical sense, so I didn't.)

   (c) (3 marks) In Figure 3, interpret the slope for `Region4`, whose value is 0.833.

   > **My answer:** `Region` is a categorical variable, so this is saying that the predicted infection risk for a patient who is in Region 4 (west) is 0.833 higher than for a patient (with the same length of stay and the same value of `Xray`) who is in Region 1 (north-east).
   >
   > You need to get at the "all else equal" again somehow, and you need to say that this is a comparison with the baseline region 1 (you can tell it's the baseline because it is missing, or

because it is first "alphabetically"). If you missed out the "all else equal" on the previous part, I won't punish you again if you miss it out here, though. There should be a maximum total deduction of one point (for the question) if you missed it out.

I want you to *name* the regions being compared, and just talking about region 4 (West) doesn't tell the whole story, because region 4 is being *compared with* something.

I would also go with something like the West region being "most dangerous" in terms of infection, because its slope is the highest of all the regions (including the zero for the baseline north-east). Precisely, if you were to think about patients who'd stayed in hospital the same number of days and had the same number of X-rays, the infection risk would be highest in the west (and lowest in the north-east).

To be more precise on the ordering: text is sorted according to this table: `http://www.asciitable.com/` (you'll have to copy and paste that), in which numbers come before uppercase letters which come before lowercase letters, but if you're comparing just numbers, or text like `region1` and `region4`, it sorts the way you'd expect.

To be even more precise, this is if you have only regular letters and numbers. If not, you might have accented and other special characters to work with, and you have to delve into Unicode to find exactly how it works. R uses a thing called UTF-8 to handle characters other than regular letters and numbers, and things called `locale`s to handle the fact that different languages sort characters different ways (for example, Scandinavian languages put the accented letters at the *end* of the alphabet: see `https://en.wikipedia.org/wiki/Danish_and_Norwegian_alphabet`).

(d) (2 marks) For assessing the (statistical) significance of `Region`, explain briefly why it is better to look at Figure 4 than Figure 3.

My answer: The `summary` output only compares the significance of each region vs. region 1. This is not the whole story, since there are other regions to be compared with each other (eg. 2 vs 3). The `drop1` output is testing whether there is evidence of *any* differences in infection rates among regions, which is what we want to test if we are trying to decide whether to keep `Region` in the model. (If there are any differences, we want to keep `Region`, if not, not. We don't take out "some" of the regions, even if we can see how to do that. Think about what, if anything, that would even *mean*.)

You ought to say something about why Figure 3 is bad, as well as why Figure 4 is useful. You can say something along the lines of the former being good for comparing regions, but only the latter for deciding whether region as a whole makes a difference. You don't need to say much, but say something.

You might be concerned that the region 1 – region 4 difference is significant in Figure 3 but not in Figure 4. Bear in mind, though, that the former figure is doing three tests all for `region` at one time, so we really ought to do some kind of adjustment (like Bonferroni or Holm: see the ANOVA section) to allow for this. This is another reason why the latter Figure is better: one test for `region` instead of more than one. Both Bonferroni and Holm would multiply the 0.0313 P-value by 3, making it non-significant.

(e) (2 marks) What do you conclude about regions from Figure 4? Explain briefly.

My answer: At $\alpha = 0.05$, `Region` is not significant. There are no significant differences among regions (in terms of infection rates, all else being equal, blah blah blah), and so `Region` ought to be removed from the model.

If you got to that last point, you have a leg up on the next part.

You can choose your own α here. As long as your conclusion matches up with your α and the P-value, I'm good with it. The AIC actually says "don't remove `region`", because the AIC for `region` is actually a bit bigger than for `none` (which is also a conclusion I would accept). This is another one of those cases where AIC is a bit more generous about keeping things than the P-value would be.

A few people talked about "region 3" here, and I wondered what this was about, until I noticed the number 3 (the degrees of freedom for `region`) next to the word "region". It's a test of *all four* regions at once, so it has $4 - 1 = 3$ degrees of freedom. That's what that is.

(f) (2 marks) Look at the first line of Figure 5. In `inf.2`, what is being predicted from what? Explain briefly.

My answer: The model `inf.2` is the same as `inf.1`, except that `Region` has been removed. That is to say, it is predicting infection risk from `Stay` and `Xray` only.

Don't confuse this with the next part. I'm only asking you to look at and talk about the `update`. One point for saying that `Region` has been removed, two for saying precisely what is left. Here, you can use the variable names.

(g) (3 marks) In Figure 5, what does the prediction tell you? Explain briefly.

My answer: This is a prediction interval, so it's talking about the predicted infection risk for *an individual* who has stayed in the hospital for 15 days, who has an `Xray` of 70 (and is from the northeast, although that is actually irrelevant since the model `inf.2` does not contain `Region`).

Such an individual is predicted to have an infection risk between 3.44 and 7.97.

The word "individual" needs to be in your answer somewhere, or the implication that the person to whom this interval applies is someone not in the original data set. The word "you" didn't quite capture this, and the word "patients" suggests that you don't understand whether this is a confidence interval or prediction interval.

Extra: in the spirit of assignment 1 (where I asked you to say something about the effect of knowing the explanatory variable on the confidence interval for the mean response), the corresponding thing here would be to compare the prediction interval with infection risks that 95% of the individuals as a whole lie between. This means finding the 2.5th and 97.5th percentiles:

```
infection %>% summarize(q025=quantile(InfctRsk,0.025),
                        q975=quantile(InfctRsk,0.975))
```

```
## # A tibble: 1 x 2
##    q025  q975
##   <dbl> <dbl>
## 1  1.56   6.8
```

Compared to this, the prediction interval, though not short, is shorter (4.53 long as against 5.24), and also shifted upwards compared to these sample percentiles. This is because (i) the `Days` and `Xray` values are both above average (and both slopes are positive), so we'd expect a greater infection risk for this individual compared to the population as a whole, and (ii) the prediction interval is shorter because knowing how long a person has been in the hospital and how many X-rays they have had says *something* about the accuracy with which you can estimate their infection risk.

2. A local health clinic sent flyers to all of its clients to encourage everyone to get a flu shot (vaccination) before winter. Later, the clinic randomly sampled 50 of its clients and asked each one whether they got a flu shot or not. The clinic also collected data on each client's age and health awareness (via a survey). The results of the health awareness survey were summarized into a health awareness score (called `awareness` in the data set), where a higher value means greater health awareness. A client who received the flu shot is denoted 1 in the `shot` column, and one who did not is denoted 0. Some of the data set is shown in Figure 6.

   (a) (2 marks) Why is it that I do *not* need a two-column response variable to fit a logistic regression to these data?

   My answer: Each row of the data frame is only one person. The way you can tell is that `shot` is either 1 (had the flu shot, the "success") or 0 (did not), rather than columns containing numbers of people who had the flu shot and didn't. In that case, each row of the data frame would have represented more than one person and you would have had to construct a two-column response out of the numbers of "successes" and "failures".

   I would also take a description of the kinds of response columns that tend to go with one-column or two-column responses in the logistic regression. Something that talks about the response column without quite getting to the point is likely to receive 1.

   (b) (2 marks) A logistic regression is shown in Figure 7. Would you consider removing either of the explanatory variables from the model? Explain briefly.

   My answer: No, because they are both significant (small P-values) and removing them would be a mistake. A gimme.

   (c) (2 marks) What probability is being modelled in Figure 7? Explain briefly how you know.

   My answer: The response variable `shot` takes the values 0 and 1. The first value 0 is the baseline, and the modelled probability is of the second one, thus the probability that a person *does* get a flu shot.

   The point of this part is that we could be modelling either the probability that a person *does* get a flu shot, or that they *do not*, and it's not obvious which one it is until you stop and work it out. (If it's the probability that a person *does not* get a flu shot, that changes around all the conclusions below.)

   (d) (3 marks) Look at the slope coefficients in Figure 7. What do they tell you about what makes a person more or less likely to get a flu shot? Explain briefly.

   My answer: The slopes for both `age` and for `awareness` are both positive. This means that a person who is older is more likely to get a flu shot, and also a person who has greater health awareness is more likely to get a flu shot. (Neither of these is terribly surprising, at least not to me.)

   (e) (4 marks) Using the information in Figure 8, give R code to obtain predicted probabilities of getting a flu shot for all combinations of first and third quartiles of `age` and of `awareness`, and to display the predictions side by side with the values they are predictions for.

   My answer:

   The quartiles of age are 40.2 and 53; the quartiles of awareness score are 43.2 and 59, so:

```
ages=c(40.2,53)
```

```
awarenesses=c(43.2,59)
new=crossing(age=ages, awareness=awarenesses)
p=predict(shot.1,new,type="response")
cbind(new,p)
##    age awareness          p
## 1 40.2      43.2 0.02027946
## 2 40.2      59.0 0.34023424
## 3 53.0      43.2 0.26135917
## 4 53.0      59.0 0.89811856
```

That's the code I would like to see, or something equivalent to it.

The order of these lines, and their spelling, matters. You don't have to define `ages` and `awarenesses` as I did; you can put them in the `crossing`, but the result of `crossing` needs to be a data frame containing columns named *exactly* `age` and `awareness`, since otherwise `predict` will not work. (What I called `ages` and `awarenesses`, if you defined the quartiles into columns first, can be called whatever you like, but the contents of what I called `new` *must* be what I said.) Thus, if you can't spell "awareness", expect to lose a mark. My scale was one point for each of these correctly done:

- defining vectors containing the values to make all combinations of
- using `crossing` (or `expand.grid`, if you must) to make the combinations, where the columns have the right names
- running `predict` with the model, data frame of values to predict for, and the right `type`
- using `cbind` or equivalent to put things side by side

If you didn't do the first of those, then you get two for coming up with the right `crossing`, with the right values (of things to make combinations of) inside.

Extra: the results agree with intuition and our previous observations: an older person is more likely than a younger person to get a flu shot, regardless of the `awareness` value, while a person with greater health awareness is more likely to get a flu shot, regardless of age. Both effects look pretty big, so it is not surprising that they are both significant even with only 50 observations.

3. The LSYPE is a "longitudinal study of young people in England". (Longitudinal means that individuals are followed over a period of time.) Over 15,000 people, born in 1989 and 1990, were followed starting in 2004. Each person was interviewed 8 times between 2004 and 2016. There were 57 variables collected for each person. We will focus on just a few, related to educational achievement.

   A few rows of the data are shown in Figure 9 and the values are summarized in Figure 10. The variables are described below.

   English young people take important national exams at age 14, called "Key Stage 3". We will focus on the Key Stage 3 English exam. Each young person receives a whole-number grade between 3 and 7 (inclusive), where higher is better. The actual exam mark is not known, nor indeed is the precise process by which exam marks are turned into these grades. This is the value in column `k3en` in Figure 9. At age 11, these young people have previously been tested in English and math at what is called "Key Stage 2".

   The other variables are:

   - `gender`: of the young person, 0 is male, 1 is female.
   - `sec`: socio-economic status of the young person's family, on this scale:
     - 0: higher managerial/professional
     - 1: lower managerial/professional
     - 2: Intermediate occupations
     - 3: Small employer/self-employed
     - 4: Lower supervisory/technical
     - 5: Semi-routine (semi-skilled)
     - 6: routine (unskilled)
     - 7: never worked/long term unemployed
   - `ks2stand`: the overall Key Stage 2 score, a decimal number (unlike the Key Stage 3 English grade). Higher is better. The scale is set so that 0 is "average", so that a `ks2stand` value can be negative.

   All of these explanatory variables will be treated as quantitative. (It is questionable whether `sec` should be treated as such, but assume that this is reasonable.)

   Our aim is to predict Key Stage 3 English grades from the other variables.

   (a) (3 marks) Describe briefly (in words) what the code in Figure 11 should be doing, and how you know it has succeeded in doing that. (You are supposed to know what `is.na` and `!` mean in this context.)

   > **My answer:** It is removing all the missing values from our four variables of interest. (Two marks.) I know it has succeeded because the `summary` at the bottom of Figure 11 is not showing any missing values any more, whereas Figure 10 shows rather a lot of missing values. (The other one mark).
   >
   > The clue is the two `summary` outputs: one of them has a lot of missing values, while the other has none.
   >
   > I am expecting that your version of "removing the missing values" will be longer than mine, so I allotted two marks to be able to give you partial credit if I felt your answer deserved it.
   >
   > I could have used `complete.cases`, but that would have removed any observations with missing values on *other* variables as well, and left us with (even) fewer observations. Perhaps I should have done a `select` first and *then* used `complete.cases`. There are many different ways to do things.

   (b) (2 marks) Even though `k3en` grade is apparently quantitative, it would be a mistake to treat it as such. Explain briefly why this is.

My answer: The numbers are really labels for grades (in the same kind of way that A, B, C, D, F are). All we know about them is their order; we don't know that the gap between 3 and 4 is the same as the gap between 4 and 5, for example.

The key point is that `k3en` is only ordinal: the only information contained in it is the order of the categories. The actual numbers have not much meaning as numbers.

"Because we need it to be a factor for the logistic regression" isn't really an explanation because it doesn't get at *why* I set up this question the way I did.

(c) (3 marks) Give R code to create an ordered factor called `en3` *within the data frame* ***kids***, using the values in `k3en` in that data frame in a sensible order.

My answer: This is what I used:

```
kids = kids %>%
  mutate(en3=ordered(k3en,3:7))
```

This produces a column with label `ord`, meaning that it is an ordered factor. The second thing in `ordered` has to be the possible values of `k3en`, *in the order that you want them.*

Feel free to get the values out of `k3en`, for example by `unique(kids$k3en)`. (This works on things that are factors or any other kind of variable, so you don't need to convert first). I suspect that `distinct` doesn't work so well, since that takes the Key Stage 3 English scores *in the order they appear in the data*, which might not be the order you want. I didn't want to deduct anything for that, though, since it shows the thing I was asking about here: getting the values to make the ordered factor from.

Also, you should use `en3` for the name of the new column (the thing before the `=` inside the `mutate`), and save the new data frame *back into* ***kids***. There appeared to be some confusion about this.

(d) (2 marks) An ordered logistic regression is fit in Figure 12. Using the information in this Figure, do you need to keep all the explanatory variables? Explain (very) briefly.

My answer: Yes. All of them are significant (P-values less than 0.05). Just that. (What a gimme!)

Alternatively, the AIC of the model with nothing removed, at 22208, is less than for the models with anything removed, so removing nothing is the best thing to do.

We still have a ton of data (even after removing missing values), so the effects could be this significant and yet not very big. This we investigate in a moment.

(e) (3 marks) Some predictions are shown in Figure 13. Describe the effect of Key Stage 2 score on predicted Key Stage 3 English grade, explaining how you got your answer. (Note that the code that was used to obtain these predictions is not shown.)

My answer: Look at any set of three neighbouring rows, since they differ only in `ks2stand`, not in anything else. The effect of having a higher Key Stage 2 score is to dramatically reduce the probability of getting a low grade (3 or 4) and dramatically increase the probability of a high grade (6 or 7).

That is to say, a young person who did well at age 11 is probably also going to do well at age 14, all else equal.

For full marks, I expect an answer that includes some understanding of what you see in the Figure: that the Key Stage 3 English grade is typically higher if the Key Stage 2 score is higher, or there is a "positive relationship" or words like that. (I was fairly relaxed about what I would take, but I wanted to see *something* about the overall picture. I was happy to see that in a lot of cases.)

(f) (3 marks) Again using Figure 13, describe the effect of socio-economic status on the predicted Key Stage 3 score, all else equal. Is it what you would expect, given your understanding about socioeconomic status and education? Explain briefly. Hint: remind yourself of the scale on which the socio-economic status is measured for these data.

My answer: A low value on `sec` means *high* socio-exonomic status: 1 is lower managerial, and 6 is unskilled. Compare rows where this changes and nothing else does, for example rows 1 and 4, or rows 9 and 12. An `sec` of 1 goes with a lower probability of a low grade (3 or 4) and a higher probability of a high grade (6 or 7).

To summarize: young people with `sec` 1 are predicted to do better overall than young people with `sec` 6. This makes sense because young people from a family with at least one professional parent would be expected to do better in school. I would be happy for you to say that, but if you do, I would like you to back it up with something (like my discussion in the previous paragraph).

The effect on the probability of a middling grade (5) depends on who you're looking at. Rows 1 and 4 are males (young men) who did badly at Key Stage 2; a 5 is a good grade for these people, and more of them are predicted to get a 5 if their `sec` is 1. On the other hand, rows 9 and 12 are females (young women) who did well at Key Stage 2. 5 is a bad grade for them, and *fewer* of them are predicted to get a 5 if their `sec` is 1.

I wanted you to get as far as saying that the `sec` 6's (whichever ones you were looking at) had a higher probability of getting a lower score and a lower probability of getting a higher score, compared to the `sec` 1's. Don't forget to compare your finding to what you would expect to see (without any data)!

There were lots of ways to tackle this question; anything sensible is fine by me. Even if you got something mixed up, it was possible to get 2 out of 3.

Extra: I didn't ask you to compare males and females. That is much as you would guess. Compare rows 1 and 7 (or 6 and 12, say): females also have a lower probability of a low grade and a higher probability of a high one than males: that is, they are predicted to do better, all else equal.

4. What are the factors that determine how long someone remains unemployed, and what effect do they have? To find out, about 1,900 unemployed people were followed from the time they became unemployed until the time they found a new job (or until the study ended). A large amount of information was collected on these people. We will look at the variables shown in Figure 14, which are:

   - `spell`: the number of months spent unemployed
   - `event`: whether or not the person found a job by the end of the study (1 is yes, 0 is no)
   - `ui`: whether or not the person was claiming unemployment insurance (1 is yes, 0 is no)
   - `logwage`: the logarithm of the person's last salary before unemployment
   - `work_area`: the person's previous area of work, categorized as:
     - `constr`: Construction, for example building of houses
     - `fire`: Emergency services such as fire, police, paramedic
     - `mining`: for example coal mining
     - `pubadmin`: public service or administration, for example working in government
     - `services`: work that does not produce a physical object
     - `trade`: for example electrician, plumber
     - `transp`: Transportation, for example truck driver.

   (a) (3 marks) Figure 15 shows the construction of a variable `y` and the display of its first 20 values. What does the value 1 at the start of the first line of output mean, and why does the next value 3 have a plus sign next to it? If your answer uses the word "censored" at any point, you should explain what that means in the context of these data.

   My answer: The first person found a new job in 1 month (one point), and the second person was observed for 3 months but did not find a job during that time (two points).

   The second person was "censored", meaning that the event "finding a job" was never observed for them. This, if you mention it, is part of the two points for talking about the `3+` observation.

   I was not referring to the `[1]` at the start of the line; I was referring to the observation `1` next to the `3+` in the data. (Having it be a different value would have make things clearer, but the words "next to it" in the question ought to have been clear enough.)

   Strictly speaking, when talking about the "1" observation, you need to say that this person found a job at the end of the 1 month. I was willing to accept you not saying this, if I felt that you had "strongly implied" this by whatever else you said (by making a contrast with the `3+` observation, for example). But the safe way is to say it.

   I realized, the way I wrote the question, that I didn't ask for an explanation of the number 3, just of what the `+` part of `3+` meant. I think I originally intended an explanation of the 3 part as well (and it was certainly not wrong to provide one), but the way the question came out, the explanation required was for the "censored" part only.

   (b) (3 marks) Figure 16 shows a Cox model for these data, together with the output from `summary` and `drop1` for this model.

   In Figure 16, does `ui` have a significant effect on the time taken to find a new job? Does a person who receives unemployment insurance typically take a longer or a shorter time to find a new job than someone who does not? How can you tell? Explain briefly.

   My answer: Two things:

   1. Yes, `ui` is significant, since its P-value is extremely small, less than 2×10^{-16}. (One mark.)

2. Look in the `coef` column: for `ui`, the slope coefficient is -0.99, negative. This means the "hazard of event" is less for a person receiving unemployment insurance. The "event" here is finding a new job, so this is saying that a person receiving UI is *less* likely to find a new job soon (or will typically take *longer* to find a new job). (Two marks.)

The right answer makes logical sense because a person receiving unemployment insurance does have *some* money coming in, so is not under quite the same financial pressure as someone who needs "a job, now" to pay the bills.

As to whether a person receiving UI takes a longer or shorter time to find a new job, I was looking for enough clarity of thinking to convince me that you did understand what was going on. "A negative relationship" is not clear enough (for me), because there is another step in the logic. If you talk about the "hazard of finding a new job" being less for the UI people, or if you use the word "probability" in a sensible way, I probably found you full marks. Strictly, the probability approach requires you to focus on a fixed amount of time (let's say 6 months, though any fixed time will do), and then you say that the probability of having found a new job within that time is lower for people receiving UI, all else equal, blah blah. The reason the actual time doesn't matter is that our models don't have interactions, and so the survival curves on the plot don't cross over each other. That means that one of the survival curves is "best" (whether best is a short time or a long time: see later) for all times. (If we had an interaction term, the survival curves could cross over each other, and the "best" one depends then on the time you're looking at.)

Marking: everybody got the one mark for saying that `ui` is significant (another gimme). Altogether, if you said that a person receiving UI would take longer to find a job for what I thought was a good, complete reason, 3 points; if you said "longer" for a reasonable but incomplete reason, 2 points.

This problem is backwards because here the "event" is a desirable thing (finding a new job), while often the event is an undesirable thing like dying. This plays out again later on when you interpret the plot of survival curves.

(c) (2 marks) Figure 17 shows some code to obtain predictions of "survival" for the median `logwage`, `ui` of 1 (receiving unemployment insurance), and the various different work areas. These predictions are shown on a plot in Figure 22. How do you know which coloured prediction is which? In particular, what prediction is the purple "survival curve" for? (If you have trouble telling the colours apart, ask an invigilator.)

My answer: The numbers on the "strata" at the top of the plot are the same as the numbers on the rows in `unemp_new`. In particular, the purple survival curve is stratum 6, which is row 6 of `unemp_new`, which is `trade`.

Expect one point if you made some reasonable attempt to identify which curve was which work area without explicitly identifying the purple one, or if you misidentified which work area was which and identified stratum 6 of what you thought the work areas were. The strata are in `unemp_new`, the table at the *bottom* of Figure 17. We know this because the strata on the plot of survival curves are in the order predicted in `survfit`, that is, in the order in `unemp_new`, since that was the input to `survfit`. The `work_areas` higher up in the Figure come out of `distinct`, so are in the order they appear in the data (not alphabetical order), but `crossing` sorts them.

(d) (3 marks) According to Figure 22, for people with median log-wage and receiving unemployment insurance, people previously in which work area are most likely to find a new job the quickest? Explain briefly.

My answer: This one is backwards from the usual "death from disease"-type examples, in that we want the event to happen *sooner*, so we are looking for the survival curve that is furthest *down and to the left*. This is the yellow-brown one, stratum number 2, which corresponds to `fire`, police and paramedics (emergency services).

If you thought up and to the right was the thing, stratum 4 is the one (match up the two shades of green and get the right one). This is `pubadmin`, public service and administration. These people are actually expected to take *longest* to find a new job. Expect to get one mark if you picked this one, because "top right is best" or similar reasoning. Admittedly, on the exam paper, the two shades of green were hard to tell apart, so I gave you one point if you picked the other one as well.

The message here is that I want you to *think*, not just react. (I was pleased with how many people worked it out.)

If you are colour-blind, ask an invigilator for any needed help. I briefed them on what kind of help to offer. You can ask, for example, "which stratum is the bottom one", or do it in two goes, "what colour is the bottom one", and, if you can't identify which stratum is the yellow one, "which one of the strata is that?". What I am looking for is that you know *what* to do even if your colour-blindness prevents you from doing it. (In that case, think for yourself about what kind of graph would enable you to identify the strata. Maybe a colour-blind-aware colour scheme would help, or the use of line types rather than colours. I invite you to find out how to produce those kinds of graphs yourself; if your future involves data analysis of any kind, this is something you will need to figure out or get help with figuring out.)

(e) (2 marks) Go back to Figure 16. For the work area you chose in the previous part, how does it appear as the one in which people are most likely to find a new job the quickest? Explain briefly.

My answer: `fire` has the most positive slope coefficient out of all the `work_area` ones, 0.54. This supports the idea that out of all the work areas, this one has the greatest "hazard of event", that is, that the event, finding a new job, is most likely to happen soonest.

If you picked `pubadmin`, you'll find that its slope coefficient is the most *negative*, which is

normally the best (when the event is something like death), but here it means that the hazard of the event is lowest: that is, it is most likely to take *longest* for people in this field to find a new job. This is consistent logic if you picked `pubadmin`, so you get 2 marks here for this as well.

Talking about P-values is not relevant here. When you're looking at a categorical variable like `work_area`, the only relevant P-value is the one for the factor as a whole, the one in the `drop1` table at the bottom of Figure 16. This tells you that `work_area` as a whole has some effect on time spent unemployed (that is, unemployment time is not the same for all work areas, even if the other variables are the same), so `work_area` has to stay in the model (even if some of the work areas have the same average unemployment time as the others). If you mentioned the slope as being the most extreme *and* you talked about P-values, I really ought to have deducted one mark, but I didn't. Colour me generous (on this, at any rate). In a regression (or a survival analysis like this one), we don't do anything like Tukey to determine which categories differ from which; we note that the categorical variable has *some* effect, and then look at predictions to see what kind of effect it has (like the graph of survival curves here).

Of course, this is assuming all else equal, and all else is *not* likely to be equal; people who worked in certain fields, such as construction, are likely to have had a lower-than-average salary before, and so *that* will tend to make them take longer to find a new job.

Extra: I had to do a bit of reorganizing to get the data in this form. I won't show you the whole thing, but I will illustrate the kind of thing I mean. You know those surveys, where it says "check one of these boxes"? This is how those survey results are encoded. Let's simplify things a bit and suppose that the alternatives you have to check one of are called A, B, C and D. The data entry person then creates columns A, B, C, D in their spreadsheet and puts a 1 in for the one a person checked, and a 0 in each of the others, like this:

```
id A B C D
p1 1 0 0 0
p2 0 0 1 0
p3 0 0 0 0
p4 0 1 0 0
p5 0 0 1 0
```

There are five people, labelled `p1` through `p5`, who were each supposed to give a response to this question Q, with one of the responses to A-D being a 1 (checked) and the rest 0 (unchecked). Except that person 3 didn't check anything.

These data are untidy, since all of A through D are responses to one question Q. This makes us think of `gather`, except not quite:

```
survey=read_delim("zero-one.txt"," ")
```

```
## Parsed with column specification:
## cols(
##   id = col_character(),
##   A = col_double(),
##   B = col_double(),
##   C = col_double(),
##   D = col_double()
## )
```

```
survey
```

```
## # A tibble: 5 x 5
##   id        A     B     C     D
##   <chr> <dbl> <dbl> <dbl> <dbl>
## 1 p1        1     0     0     0
## 2 p2        0     0     1     0
```

```
## 3 p3        0     0     0     0
## 4 p4        0     1     0     0
## 5 p5        0     0     1     0
```

If we `gather` as expected, we get:

```
survey %>% gather(response,Q,A:D)
```

```
## # A tibble: 20 x 3
##    id    response     Q
##    <chr> <chr>    <dbl>
##  1 p1    A            1
##  2 p2    A            0
##  3 p3    A            0
##  4 p4    A            0
##  5 p5    A            0
##  6 p1    B            0
##  7 p2    B            0
##  8 p3    B            0
##  9 p4    B            1
## 10 p5    B            0
## 11 p1    C            0
## 12 p2    C            1
## 13 p3    C            0
## 14 p4    C            0
## 15 p5    C            1
## 16 p1    D            0
## 17 p2    D            0
## 18 p3    D            0
## 19 p4    D            0
## 20 p5    D            0
```

so that each person appears four times, one for each of their four responses to A through D. But we only want to keep the "1" responses, so:

```
survey %>% gather(response,Q,A:D) %>%
  filter(Q==1)
```

```
## # A tibble: 4 x 3
##   id    response     Q
##   <chr> <chr>    <dbl>
## 1 p1    A            1
## 2 p4    B            1
## 3 p2    C            1
## 4 p5    C            1
```

and the Q column has now served its purpose, so can be removed. Now, by `count`ing the `response` column, we can see how many people gave each response, and by looking at the `id` column, we can see that person `p3` has disappeared, because they didn't give a response to any of the options A through D.

The `work_area` in the original data set I had was like this; there were seven columns, labelled with the names of the seven work areas, and six of them were always zero. I realized that this was not tidy, and so had the idea of gathering the columns and picking out the ones. This meant that I briefly created a data frame with seven times as many rows, but the data-wrangling functions in the `tidyverse` are very efficient, so I had no need to worry.

5. In a manufacturing process, a plastic rod is made by melting a plastic and then extruding it through a nozzle. It is better if the rod can be made more quickly, that is, if the extrusion rate is higher. Two factors can be controlled in the extrusion process: temperature (200 or 300 degrees Fahrenheit), and pressure (40 or 60 pounds per square inch). What effect do these factors have on extrusion rate? An experiment was run in which three replicates of each combination of temperature and pressure was used, and the extrusion rate measured, for a total of 12 observations.

   The data set is shown in Figure 18.

   (a) (2 marks) What do you conclude from Figure 24? Explain briefly.

   My answer: This is an interaction plot. We are looking to see whether the two traces are approximately parallel. I think they are not, so I expect to see an interaction between temperature and pressure: the effect of one of the explanatory variables on extrusion rate depends on the value of the other one.

   Or, take the view that the lines are approximately parallel, and that therefore there is no interaction. That could also reasonably be supported by this picture, so I am good with it.

   Make a call about whether there's an interaction or not (one point), and tell me why you say that (the lines are or are not approximately parallel). I don't need "in the context of the data" yet (that's coming up). The first point is actually a freebie, since you can say what you like about whether you think there's an interaction or not, as long as you say *something* about that.

   If you suggest interaction without using that word, that's fine (eg. "the effect of pressure is greater at the larger temperature"). If you talk sensibly about main effects, that's one point; likewise, if you say there is or is not an interaction without saying why, or if you say that the lines are or are not parallel without saying what it tells us. (The reason it's only one point for talking about the main effects is that in a two-way ANOVA such as this one, the *first* order of business is to think about whether there's an interaction or not, since what we do next depends on what we conclude about that.)

   (b) (2 marks) Look at Figure 23. What additional information does this plot give that is not shown in Figure 24? Does that strengthen or weaken your conclusion from the previous part? Explain briefly.

   My answer: The grouped boxplot also shows how much variability there is in the extrusion rates within the groups. There is very little. That supports the idea that the lines on the interaction plot really are not parallel, and that therefore the interaction effect really does exist. (The conclusion about interaction is strengthened.)

   If you said in the previous part that the lines are "approximately parallel": approximately parallel plus *large* variability points to no interaction, but approximately parallel plus small variability might point to not actually parallel after all. (The conclusion about no-interaction is weakened.)

   If you like, you can also use Figure 23 to assess the assumptions for ANOVA: that is, normality with equal spread within each group. I'd say the normality is OK, especially considering that each box is based on only four observations, but the spread tends to be bigger when the mean (median) extrusion rate is larger, which would suggest that a transformation of extrusion rate is called for. If you somehow have doubts about the appropriateness of the analysis, that would weaken your conclusion about interaction. What I am guessing is that an appropriate transformation would reduce the difference between the two highest means, and reduce the evidence for an interaction.

   Of course, I will explore this later.

   In any case, mention something *different* that comes from the grouped boxplots, and try to say something about how that affects your conclusions from the previous part, whatever they

were. When I was marking this part, I made a point of referring back to what you said in (a) to check for this. This means that an appropriate answer for (b) will be *different* depending on what you said in (a) (which I thought would make this very slow to mark, but I seem to have improved over the years in deciding whether the answer in front of me is a 2 or a 1).

(c) (2 marks) An analysis of variance is shown in Figure 19. What do you conclude from it, in the context of the data?

My answer: This two-way analysis includes an interaction, so look at the interaction first. This is significant, so the effect of temperature on extrusion rate depends on the pressure (or, if you prefer, the effect of pressure on extrusion rate depends on the temperature. Either way around is good.) I was fairly relaxed about how much interpretation I insisted on, since the most important thing is just below.

The interaction is significant, so **stop here** with the interpretation. That's the most important thing. If you try to assert, for example, a significant effect of temperature, this is *wrong* until you have untangled what the interaction effect looks like. (If you look back at Figure 24, you'll see that there clearly *is* an effect of pressure, in that the extrusion rate is higher when the pressure is higher, but the *size* of the effect is bigger at temperature 300 than at temperature 200. That would be *understanding* the interaction, and if you say that, which is beyond what I was asking here, since I only asked about the ANOVA, you are entitled to talk about a pressure effect, but only then.)

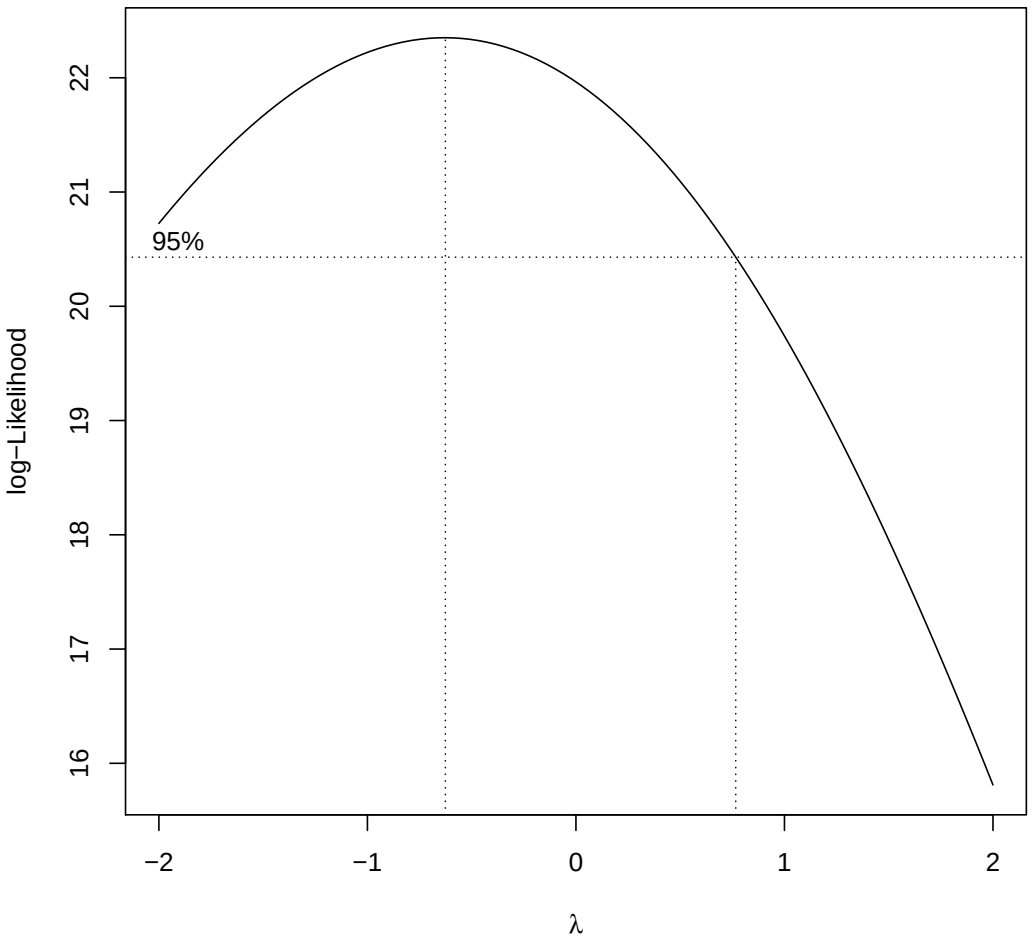
Memo to self: cut down on the parenthetical remarks. You, meanwhile, might like to take a deep breath after reading that sentence with all the commas in it.

Extra: I mentioned earlier about the possibility of a transformation. I'm guessing that square root or log would be the thing, but we can fire up Box-Cox to help us decide:

```
rods=read_csv("rodmold.csv")
```

```
## Parsed with column specification:
## cols(
##   temperature = col_double(),
##   pressure = col_double(),
##   batch = col_double(),
##   extrusion_rate = col_double()
## )
```

```
rods = rods %>% mutate(pressure=factor(pressure),
                       temperature=factor(temperature))
boxcox(extrusion_rate~temperature*pressure,data=rods)
```



This is not the clearest (we don't have much data), but log is as good as anything, and doing nothing is not *quite* defensible in fact (the lower end of the CI for λ is off the left side of the picture). So:

```
extr.2=aov(log(extrusion_rate)~temperature*pressure,data=rods)
summary(extr.2)
```

```
##                      Df Sum Sq Mean Sq F value   Pr(>F)
## temperature           1 1.0631  1.0631 330.714 8.59e-08 ***
## pressure              1 0.3172  0.3172  98.680 8.92e-06 ***
## temperature:pressure  1 0.0127  0.0127   3.936   0.0825 .
## Residuals             8 0.0257  0.0032
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interaction is indeed no longer significant (that was a guess before). So in that case, the right thing to do is to take it out:

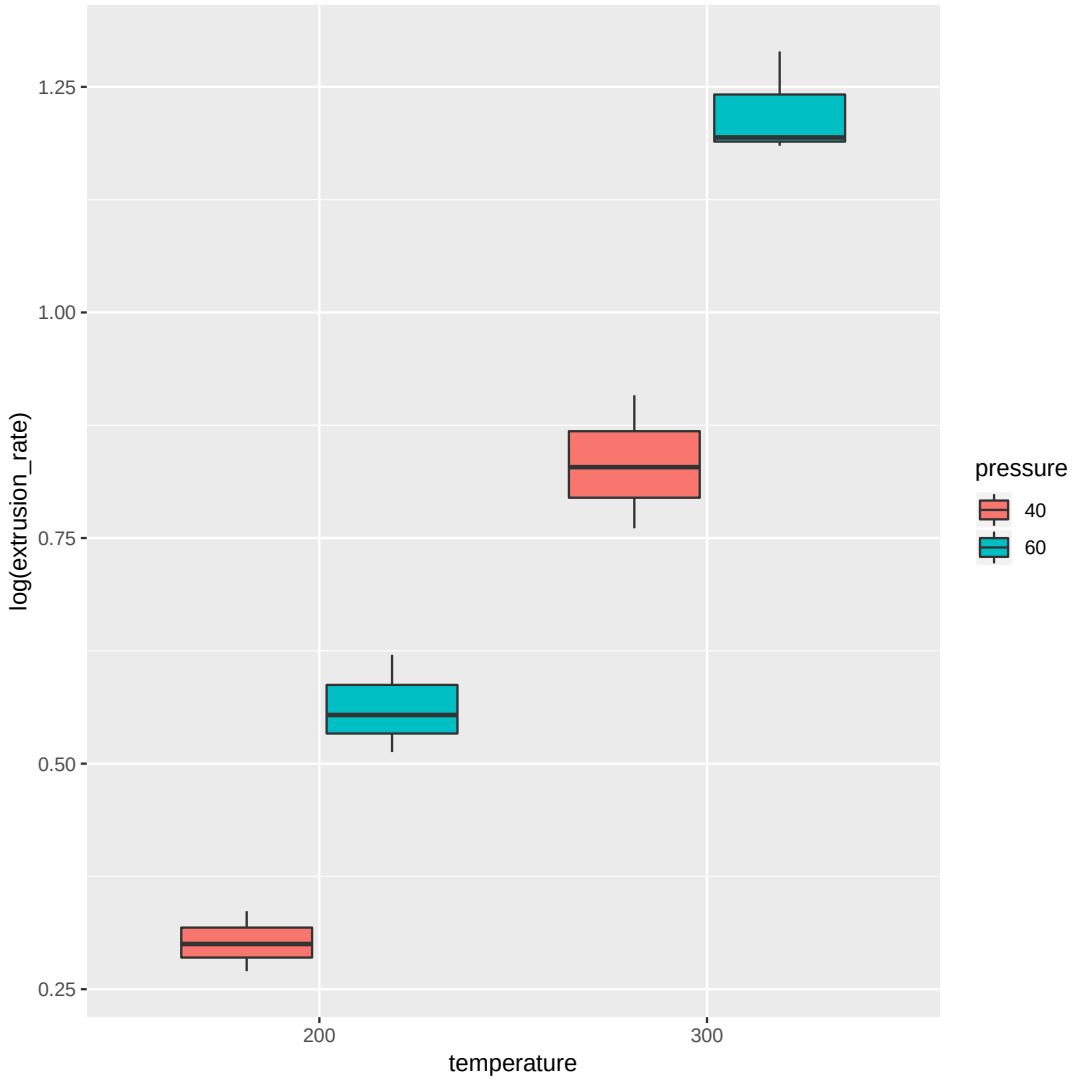
```
extr.3=update(extr.2,.~.-temperature:pressure)
summary(extr.3)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## temperature  1 1.0631  1.0631  249.37 7.21e-08 ***
## pressure     1 0.3172  0.3172   74.41 1.21e-05 ***
## Residuals    9 0.0384  0.0043
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now there are significant main effects of temperature and pressure, which means that, on the log scale, the effect of temperature is the same at both pressures, and vice versa. That would have made a simpler analysis for you, but I wanted you to handle one where the interaction was significant, so I didn't do the transformation in the original question. (If you saw that a transformation might have been a good idea looking at the grouped boxplots, you ought to have received some credit for it.)

The reason I expected this to happen is that transformations like log tend to bring larger values down and make them less spread-out, but often they also simplify the model. This is what happened here. We can re-draw the boxplots and see that they are more equally spread, and that the difference between low and high pressure is similar for the two temperatures:

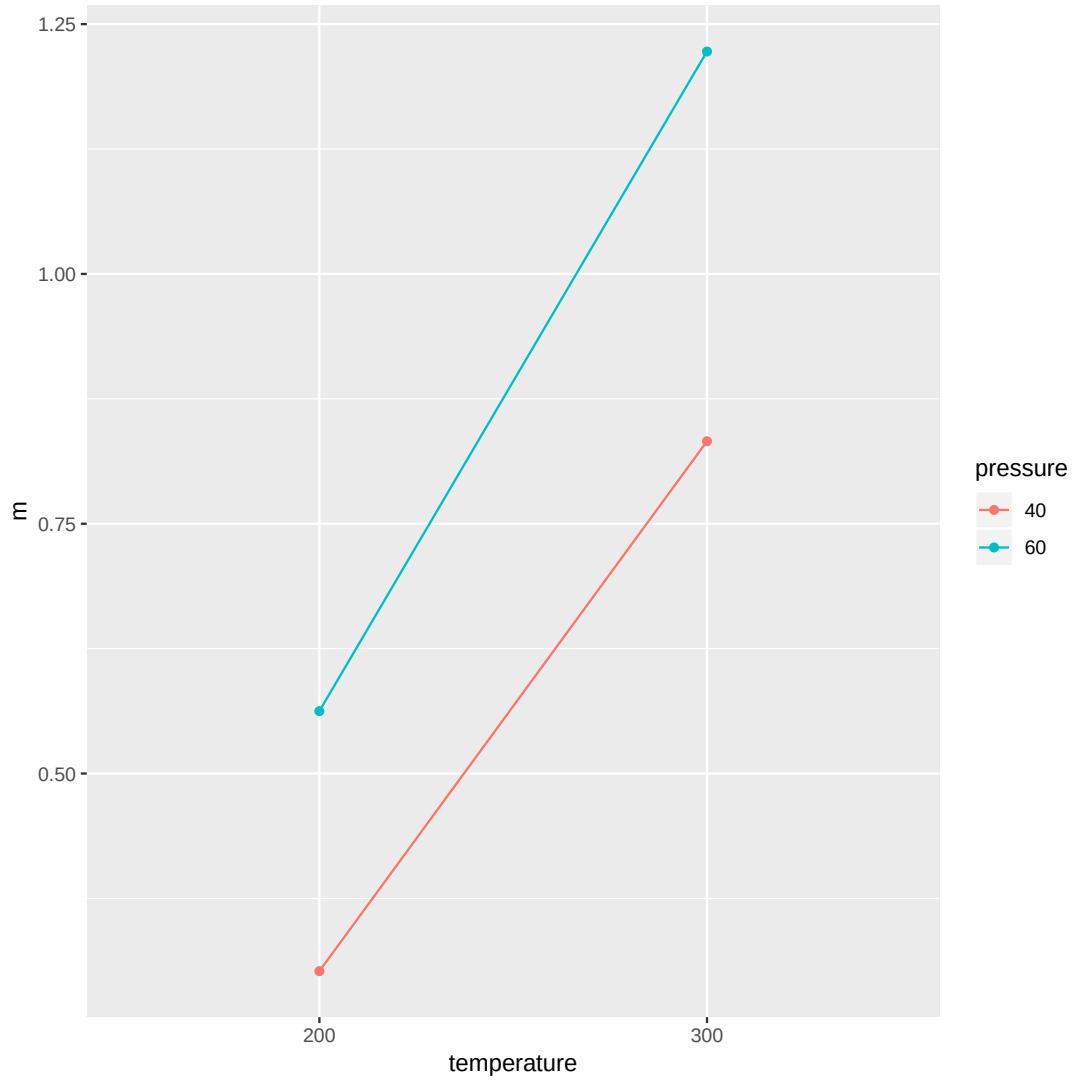
```
ggplot(rods,aes(y=log(extrusion_rate),x=temperature,fill=pressure))+
    geom_boxplot()
```



Those boxplots are more nearly the same height, and the blue ones are above the corresponding red ones by more nearly the same amount. Not perfect, but a lot better.

And the interaction plot:

```
rods %>% group_by(temperature,pressure) %>%
    summarize(m=mean(log(extrusion_rate))) %>%
ggplot(aes(y=m,x=temperature,colour=pressure,group=pressure))+
    geom_point()+geom_line()
```



Not perfectly parallel, but very close.

(d) (2 marks) Figure 20 shows some more analysis. The company who ran the experiment is looking for the highest extrusion rate. Based on this Figure, what recommendation would you make for the combination or combinations of pressure and temperature to use? Explain briefly. (If you recommend only one combination of pressure and temperature, justify that recommendation.)

My answer: The highest mean extrusion rate is 3.40, at temperature 300 and pressure 60. But this might have come out highest by chance. Look at the Tukey table (the temperature/pressure combination one at the bottom) to see that the `300:60` combination is in fact *significantly* higher in terms of extrusion rate than all the other combinations, so the inference is that this combination really is better than all the others (rather than just happening to come out better this time).

If the best one had not been significantly higher than, say, one of the others, the right answer would have been to recommend *either of the top two* combinations. That is, for the two points, you need to identify the combination with the highest extrusion rate, *and* state that its mean

extrusion rate is significantly higher than that of the other combinations.

(e) (3 marks) Figure 21 shows a final piece of analysis. Explain briefly what you conclude from this analysis, in the context of the data. (Explaining what the code does will not help you much. I want to know what this Figure tells you *about the data*.)

My answer: This is doing an analysis of the simple effects of pressure for each value of temperature. The two P-values (in the bottom table) are both very small, so there is a definite effect of pressure at each of the two values of temperature. Going back to the graphs, or the table of means, tells you that the extrusion rate at each temperature is significantly *higher* when the pressure is higher. (I have to decide whether you need to say that last bit to get the third mark.)

I showed you the code so that you could compare it with your notes from class (or my solutions to Assignment 4.5) and see that this is how we do simple effects the all-at-once (split-apply-combine) way (rather than doing two separate `filter`s, one for each temperature).

Note that we are *not* assessing an effect of temperature here. We are, in effect, holding temperature *fixed* (at 200 first and then at 300) and asking what is the effect of *pressure* at our chosen temperature. Because there is an interaction, the effect of pressure is different at each temperature: a different size (both significant), but the same direction (higher pressure is higher extrusion rate).

As to what the code is actually doing (for your edification afterwards): first we write a function to do the one-way ANOVA of extrusion rate as it depends on pressure, from any input data frame that contains those two variables, and to pull out the P-value. The easiest way to do *that* is to use `glance` from `broom` to get the one-line summary of the `aov`, and then pull out the `p.value` from that. (The overall P-value is *always* called `p.value` in a `glance` output, which saves having to remember what it's called, for example if we were trying to get it from the `summary` output, which is nice to look at but a pain to compute with.)

Having set that up, we:

- "split" the data frame into two sub-data-frames, one for each temperature, using `group_by` and `nest` (this is the data-frame-within-data-frame thing)
- "apply" our function to each one using `map`, in this case `map_dbl` because `pval` returns a decimal number (a P-value)
- and "combine" the results into what you see at the bottom of the Figure, with the two small P-values, one for each temperature. (This is done by including the `map_dbl` within a `mutate`.)

I thought it was too much to ask you for code to do that, but I think it's perfectly reasonable (if challenging) to ask you to interpret what that code and output means.

I might find a point for a sufficiently clear explanation of what the code does, if you can't tell me what you learn from the results.