00D7×

University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 / STA 1007 (K. Butler), Midterm Exam
February 18, 2017

Aids allowed:

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

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

This exam has 19 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	3	
2	5	
4	9	
5	2	
6	6	
7	8	
8	9	
9	4	
10	2	
12	2	
13	3	
14	1	
15	7	
16	7	
17	8	
18	8	
Total:	84	

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. Spatio-temporal reasoning is the ability to understand a pattern in time and space, and to understand how objects fit into it. For example, suppose you are meeting a friend at a coffee shop across town later today. You use spatio-temporal reasoning to figure out how to get there, and when you will have to leave to get there in time. This kind of reasoning is important for children to develop.

   How can we help children to develop spatio-temporal reasoning? We might give them lessons of various different kinds. In one study, 40 children were randomly allocated to groups who received piano, singing, computer or no lessons. Each child only did one kind of lesson; the "no lessons" was a control group. Each child's ability at spatio-temporal reasoning was measured twice, once before they started the lessons and again after they finished. The score that was recorded for each child was the difference between the measurement after and the measurement before. That is, a child with a positive score had better spatio-temporal reasoning after their lessons than before, and a child with a negative score had worse spatio-temporal reasoning afterwards. (Note that the apparent matched-pairs nature of the data has been removed by looking at the differences; we now have independent measurements.)

   The data are shown in Figure 2 of the booklet of code and output.

   (a) (3 marks) Some code is given in Figure 3. Describe what this code does (without using the word "gather"). Your answer should contain a description of what the output from `gather` looks like, and should explain what role `reasoning2` has in the process.

My answer: The effect of the `gather` is to turn the "wide" data input (four columns of scores with 10 rows) into "long" format (with 40 rows). Specifically, there are two new columns: `changescore` which has all the scores regardless of the type of lessons, and `lesson` which has the type of lesson that the score next to it was obtained for. The column `row` was not mentioned, and so it exists in the new data frame `reasoning2` as well.

The `-> reasoning2` means "take the output from `gather` and save it in a data frame called `reasoning2`". Thus, the long 40-row data frame is saved in `reasoning2` to use later.

To prove all this:

```
reasoning=read.csv("reasoning.csv",header=T)
reasoning %>%
  gather(lesson,changescore,piano:none) -> reasoning2
reasoning2
```

```
##    row   lesson changescore
## 1    1    piano           2
## 2    2    piano           5
## 3    3    piano           7
## 4    4    piano          -2
## 5    5    piano           2
## 6    6    piano           7
## 7    7    piano           4
## 8    8    piano           1
## 9    9    piano           0
## 10  10    piano           7
## 11   1  singing           1
## 12   2  singing          -1
## 13   3  singing           0
## 14   4  singing           1
## 15   5  singing          -4
## 16   6  singing           0
## 17   7  singing           0
## 18   8  singing          1
## 19   9  singing          0
## 20  10  singing         -1
## 21   1 computer          0
## 22   2 computer          1
## 23   3 computer          1
## 24   4 computer         -3
## 25   5 computer         -2
## 26   6 computer          4
## 27   7 computer         -1
## 28   8 computer          2
## 29   9 computer          4
## 30  10 computer          2
## 31   1     none          5
## 32   2     none         -1
## 33   3     none          7
## 34   4     none          0
## 35   5     none          4
## 36   6     none          0
## 37   7     none          2
## 38   8     none          1
## 39   9     none         -6
## 40  10     none          0
```

(b) (3 marks) What output will the code in Figure 4 produce? Describe it precisely.

My answer: The number of observations and the mean change score for the children who had each different kind of lesson. (Three things: number of observations, mean, for each lesson. Expect to lose a point for each one you missed):

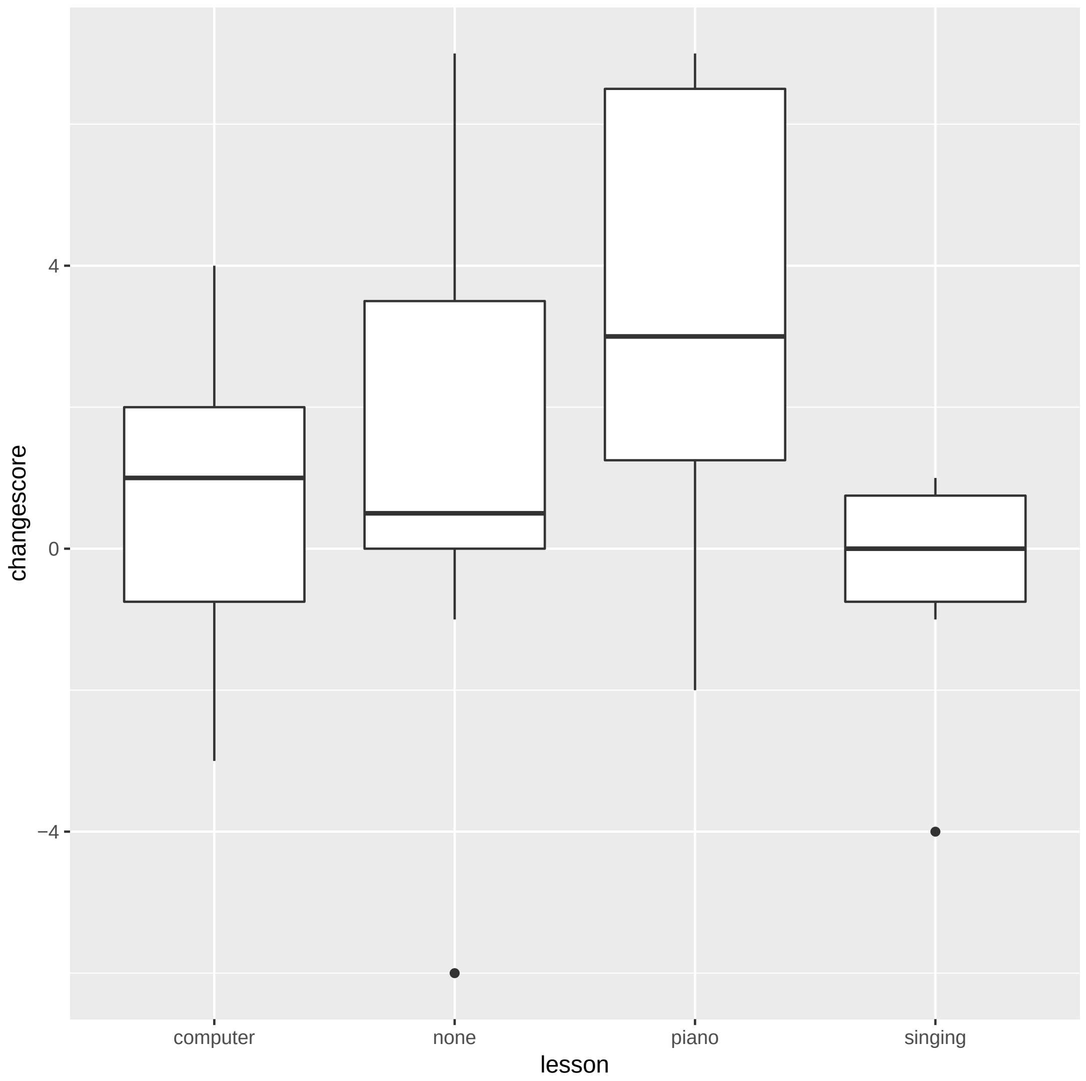
```
reasoning2 %>%
  group_by(lesson) %>%
  summarize(count=n(),m=mean(changescore))
```

```
## # A tibble: 4 x 3
##   lesson   count     m
##   <chr>    <int> <dbl>
## 1 computer    10   0.8
## 2 none        10   1.2
## 3 piano       10   3.3
## 4 singing     10  -0.3
```

(c) (2 marks) What output will the code in Figure 5 produce? Describe it precisely.

My answer: Side-by-side boxplots of change score for each lesson type:

```
reasoning2 %>%
  ggplot(aes(x=lesson,y=changescore))+geom_boxplot()
```



I thought "a boxplot with `lesson` on the x-axis and `changescore` on the y-axis" didn't really display enough insight, since you could guess that from looking at the code even if you had no idea about `ggplot`. I wanted to see something that told me how many or which boxplots I should see: "for each lesson type", or "four side-by-side boxplots of change score", or draw a picture, or anything like that. Show me that you know exactly what kind of thing you'd see. (That was why I put "precisely" in the question. I could have asked you "what kind of plot does the Figure produce", to which a good answer would be "a boxplot with `lesson` on the x-axis and `changescore` on the y-axis", but that would have been only a one-point question.)

2. Fish can be healthy to eat, but fish that is contaminated with mercury is not so healthy. What influences the amount of mercury contamination? Water samples were taken from 38 lakes and analyzed. From those same 38 lakes, samples of fish were taken, and the mercury contamination in their muscle tissue was recorded. Fish absorb mercury over time (older fish have higher concentrations on average, with a relationship that is known), so standardized mercury levels for 3-year-old fish were calculated for each lake (based on the fish that were caught there).

The data are summarized in Figure 6. The variables are:

- `mercury`: standardized mercury level of fish in parts per million (response)
- `alkalinity` of water in mg/l
- `calcium` concentration of water, in mg/l
- `pH` of water (7 is neutral, lower values are acid, and higher values are alkaline).

(a) (3 marks) Scatter plots of `mercury` against each of the explanatory variables are shown in Figure 7. A regression model was fit in Figure 8. Why specifically do you think that fitting this regression model was a bad idea?

> **My answer:** The regression is assuming a linear relationship between `mercury` and the other variables. From the scatterplots, the relationships with `alkalinity` and with `calcium` are definitely curved. The relationship with `pH` appears to be acceptably straight, but the fact that there is at least one non-linear relationship should be enough to make us think again. (I want you to name the two variables whose relationships appear to be curved.)
>
> The clue is that I gave you the scatterplots, which I presumably did for some reason, and so you need to go beyond for example saying that the R-squared is low. It is, but it is for a reason, namely the relationship are not linear.
>
> Also, those are not residual plots in Figure 7; they are plots of the actual data. So we *want* to see relationships on these plots, just not the two curved ones that we see.

(b) (2 marks) A second regression is fit in Figure 9, with the output shown. A third regression is fit in Figure 10, but the output is not shown. Which explanatory variables does the third regression contain? You may assume that this third regression model is satisfactory for the rest of the question.

> **My answer:** `log(calcium)` and `pH` have been removed, so only log(alkalinity) is left. There is a clue here: in the regression of Figure 9, the two variables removed are not close to significance, and the one remaining is the only significant one. (I checked by `anova` that the smaller regression was satisfactory.)

(c) (4 marks) What precisely is being predicted in Figure 11? In particular, what do the last two numbers in the second row of the output of the `cbind` tell you?

> **My answer:** The prediction is of `log(mercury)` (careful!) for alkalinities 6 and 67. (The log of the alkalinity is calculated inside the `predict`.) A clue that it is not `mercury` itself is in the output of `summary`: the median `mercury` is 445, with quartiles 250 and 490 or so, while the predictions are a lot smaller than that.
>
> Almost everybody failed to notice that log of mercury was being predicted.
>
> The last two numbers on the second row are a confidence interval for the mean of all `log(mercury)` values when `alkalinity` is 67. That is to say, this is *not* a prediction interval (which would be `interval="p"`) and your answer should make that clear. With 95% confidence, the mean log-mercury in fishes from all lakes where the alkalinity is 67 lies between 5.35 and 5.73.
>
> You needed to be precise about what kind of interval I had here: not a prediction interval for the mean log-mercury of a single fish, but a confidence interval for the mean log-mercury of *all*

fish in lakes where the alkalinity is 67. (If you thought we were predicting mercury itself, you can say that again here and I will count that as good.)

We are *not* predicting alkalinity here; we are predicting log-mercury *from* the alkalinity.

A number of people noted that as alkalinity increased, the predicted (log)-mercury decreased. This doesn't really belong here, but I'll count it as if you said it in part (d), since it is part of the answer there.

Brahms' first piano concerto playing here. A gorgeous piece of music.

(d) (2 marks) In what way are Figures 9 and 11 telling the same story? (You may ignore the fact that models `mercury.2` and `mercury.3` are not the same.)

My answer: Figure 9 has the output from a regression, with slopes, R-squared and so on. Figure 11 has (at the end) predictions. What I noticed is that the predicted log-mercury goes down as `alkalinity` (and therefore log-alkalinity) goes up: both the fitted values and the intervals. Looking back at the regression, the slope for `log(alkalinity)` is *negative* (and significantly so), which also points to the same thing: as alkalinity goes up, mercury and log-mercury go down.

I thought this was the most insightful answer, but I decided that if you noticed that log-alkalinity was significant in Figures 9, and you made some comment about the two predicted (log-) mercury values being different for the different alkalinity values, I was good with that too.

I left this one open (on purpose) to see what you would find. If you came up with something else sensible that the two Figures have in common, I'm good with that.

3. 33 leukemia patients were studied. For each patient, their white blood cell count was recorded (`wbc` in the data file), and the presence or absence of a certain morphological characteristic in the white blood cells was also recorded (`ag` in the data file, denoted `+` for present and `-` for absent). It was noted whether each patient lived for at least a year (`live=1`) or not (`live=0`). The actual lifetime of each patient was not recorded, only whether it was at least a year or not. The researchers were interested in whether either of the two explanatory variables `wbc` or `ag` helped to predict survival.

   (a) (2 marks) Look at the plot in Figure 13. Why do you think the researchers decided to use the log of white blood cell count in their analysis?

   My answer: Because the distribution of `wbc` appears (very) skewed right. Or that there are some very large values that could have undue influence over the analysis. Or the two groups don't have similar spread. (`wbc` does not need to be normally distributed as such, though having a symmetric distribution helps.)

   I was generally relaxed about marking this: pretty much anything that suggested "get rid of the few very large values" was good. (But if you don't talk about `wbc`, don't expect much.)

   (b) (2 marks) In the code for Figure 13, why did I have to say `x=factor(live)` instead of `x=live`?

   My answer: Because `live` is a number, and to make side-by-side boxplots, the grouping variable has to be a factor (categorical variable), so I had to make it one.

   If you use a numeric variable for the `x=` in side-by-side boxplots, you get just one boxplot (for all the data combined), along with a warning.

   I am obviously in a good mood this morning, since anything close enough to my answer seems to be getting the points: saying that `lived` needs to be a factor and it isn't one (because the values 0 and 1 will be treated as numbers) appears to be enough.

   We are actually *not* using `factor(live)` in the logistic regression, since `glm` will treat the 0-1 response as a factor (that's in the next part) anyway. But for the boxplots (in particular for `ggplot`'s interpretation of boxplots), the grouping variable needs to be an actual factor, and that's the important thing here. (It can be text, which will be treated as a factor, but if it looks like a number it won't work.)

   (c) (2 marks) In the logistic regression of Figure 14, what probability is being predicted? Explain briefly.

   My answer: The response variable `live` is 0 or 1, which is being treated as a factor, so the first value 0 is the baseline and we are predicting the probability of the second, 1: that the patient lives for at least a year.

   I need to see in your answer somewhere why it is *not* probability of *dying* within a year. There is one point for saying that it is the probability of living, and one point for saying why it is that and not the probability of dying.

(d) (2 marks) Which, if any, of the explanatory variables would you consider removing from the regression? Explain briefly.

> **My answer:** In Figure 14, both variables have P-values less than 0.05 (0.0162 and 0.0209), so neither of them should be removed.

(e) (2 marks) Looking at Figure 14, what would be the effect of a higher white blood cell count?

> **My answer:** If `wbc` is larger, so is `log(wbc)`. The slope of `wbc` is negative, so if `wbc` is larger, the probability of living is *smaller*.
>
> I didn't even ask for an explanation (maybe I should have done), so "a higher white blood cell count decreases the probability of survival for a year" is a complete answer. But you do well to supply some kind of explanation, in case I don't think your answer is good but there is something sound in the explanation (in which case you might get some part marks).
>
> I was only looking for "the slope is negative"; I didn't need you to interpret the number -1.10. It's actually saying this: "if `log(wbc)` increases by 1, the *log-odds* of survival decreases by 1.10." This is hard to interpret, which is why I didn't ask you about it. It is *not* saying that the *probability* decreases by 1.10 (which wouldn't make much sense). But, the way I phrased the question, I couldn't really deduct marks if you said that; as long as you said something about the probability of living going down, I was good. (Or the probability of whatever you thought was being modelled, earlier.)
>
> There is a clue to all this here with the boxplot in Figure 13. There, most of the high `wbc` values went with `live=0` (that is, with dying within a year). This is telling the same story, only backwards. (When you come to look at discriminant analysis, you see some of the same backwards thinking.)

(f) (4 marks) I did some predictions based on the model in Figure 14. These are shown in Figure 15. Based on the latter Figure, what can you say about the effect of `ag`? Is that consistent with Figure 14? Explain briefly.

> **My answer:** Compare `ag` being `+` or `-` with the *same* value of `wbc`, that is, rows 1 and 4 of the table of predictions (or rows 2 and 5, or rows 3 and 6). In all those cases, the probability of living is a lot higher for a patient with AG present, as compared to AG being absent.
>
> Going back to Figure 14: `ag` is a categorical variable with levels `-` and `+`. The first of these is the baseline (not shown, slope of 0), while the slope shown, 2.5196, is for the second one. That means that, all else equal, the probability of living is a lot higher for `ag` being `+` than for it being `-`. This is, as it should be, entirely consistent with the results of the predictions.
>
> I was going to give 2 points each for appropriate consideration of Figure 14 and Figure 15. Then I decided that this part was really about the predictions, so if you did a complete assessment of the predictions without going back to Figure 14, I would give you 3 out of 4.

4. A clinical trial was designed to compare the effectiveness of three pain-relief drugs to be taken after an operation. The drugs were called `c15`, `c60` and `z100`. Each patient in the trial was given one randomly-chosen drug after their operation. Each patient was then asked to rate the pain relief offered by their drug on a scale "poor, fair, good, very good". After the trial was complete, the number of patients giving each drug each rating was tabulated, as shown in Figure 16.

   (a) (3 marks) What is the response variable here, and for what *two* reasons would you choose `polr` from package `MASS` to do your analysis of these data?

   > **My answer:** The response variable is the amount of pain relief. This is categorical, with more than two categories, and the categories are in order from poor to very good. So an ordinal logistic regression is appropriate to assess how pain relief category depends on drug, which is what `polr` does.
   >
   > A minimal answer: `rating`; categorical; ordered.
   >
   > I didn't ask for a reason, but you help yourself by listing the categories in order.

   (b) (2 marks) What is the *purpose* of the `mutate` in the creation of the new data frame `painrelief` in Figure 16? Explain briefly. (That is, *what* the code is doing is part of the answer, but I am mainly interested in *why* I have to do it.)

   > **My answer:** The code is creating an ordered factor (and storing it back in `rating`) with the levels in the order "poor, fair, good, very good" (that is, the logical, correct order). I need to do this because, as the output from `lev` shows, the levels read in from the file are in alphabetical order, which is not the order we want to use (either in the ordinal logistic regression or in the bar chart).
   >
   > I actually made the bar chart first, and saw that the levels of `rating` came out in the wrong order, so I inserted some code to put them in the right order, and then I figured I should ask you about it.

   (c) (2 marks) A plot is shown in Figure 33 (at the end of the booklet of code and output). Based on this plot, would you expect to see a significant difference in pain relief among the drugs? Explain briefly.

   > **My answer:** We are looking for a difference in the pattern of the ratings for each drug. The drugs `c15` and `c60` are similar in that most patients rate them "good" or less (about the same number in each category) and very few patients rate them "very good". The drug `z100`, however, is different: a majority of patients rate it "very good", with very few rating it "poor" or "fair".
   >
   > That is to say, drug `z100` looks much better than the other two (which are about the same), so I would expect a significant difference in rating due to drugs because `z100` is different from (better than) the others.
   >
   > I'm looking for a call about whether any of the drugs are different from the others, and also some kind of discussion of how the different one(s) are different (or, if you think they're all the same, why you think that). I think it makes most sense to say "given the drugs, here's how the responses stack up for them" rather than conditioning on the responses ("most of the very good ratings were for Z100"), but if the discussion makes sense, I'm good with it.

   (d) (2 marks) Why did I need `weight=` in my modelling statement in Figure 17?

   > **My answer:** Because each line of the data file is a summary of several patients (the ones who took the same drug and gave it the same rating), rather than just one patient, and the modelling statement needs to know this.

(e) (2 marks) In constructing my data frame `new` in Figure 17, why did I *not* need to use `expand.grid`? Explain briefly.

My answer: There is only one explanatory variable, `drug`, so there is no need to make “all possible combinations” of one thing. As long as `new` contains all the different drugs (which it does), I am good.

(f) (2 marks) Should I remove `drug` from the model? Why or why not?

My answer: The output from `drop1` in Figure 17 says that `drug` should not be dropped from the model: either because `<none>` (ie. remove nothing) has a smaller AIC, or because the P-value attached to `drug` is very small, and will make the model fit worse if it is removed.

What happens if I remove `drug` anyway? Well, the predictions are going to be awful, but we can still get some. First, though, I have to repeat most of what is on the Code and Output:

```
library(MASS)
##
## Attaching package: 'MASS'
## The following object is masked from 'package:dplyr':
##
##     select
painrelief0=read.table("drugcomp.txt",header=T)
lev=levels(painrelief0$rating)
painrelief0 %>%
  mutate(rating=ordered(rating,lev[c(3,1,2,4)])) -> painrelief
painrelief.1=polr(rating~drug, weight=frequency, data=painrelief)
new=data.frame(drug=levels(painrelief$drug))
p=predict(painrelief.1,new,type="probs")
cbind(new,p)
##   drug       poor       fair      good  verygood
## 1  c15 0.32446172 0.25034061 0.2972508 0.1279469
## 2  c60 0.34244375 0.25200600 0.2863636 0.1191867
## 3 z100 0.06486819 0.09848111 0.3327126 0.5039381
```

Now, I can fit a model without `drug` by taking `drug` out of `painrelief.1`, thus:

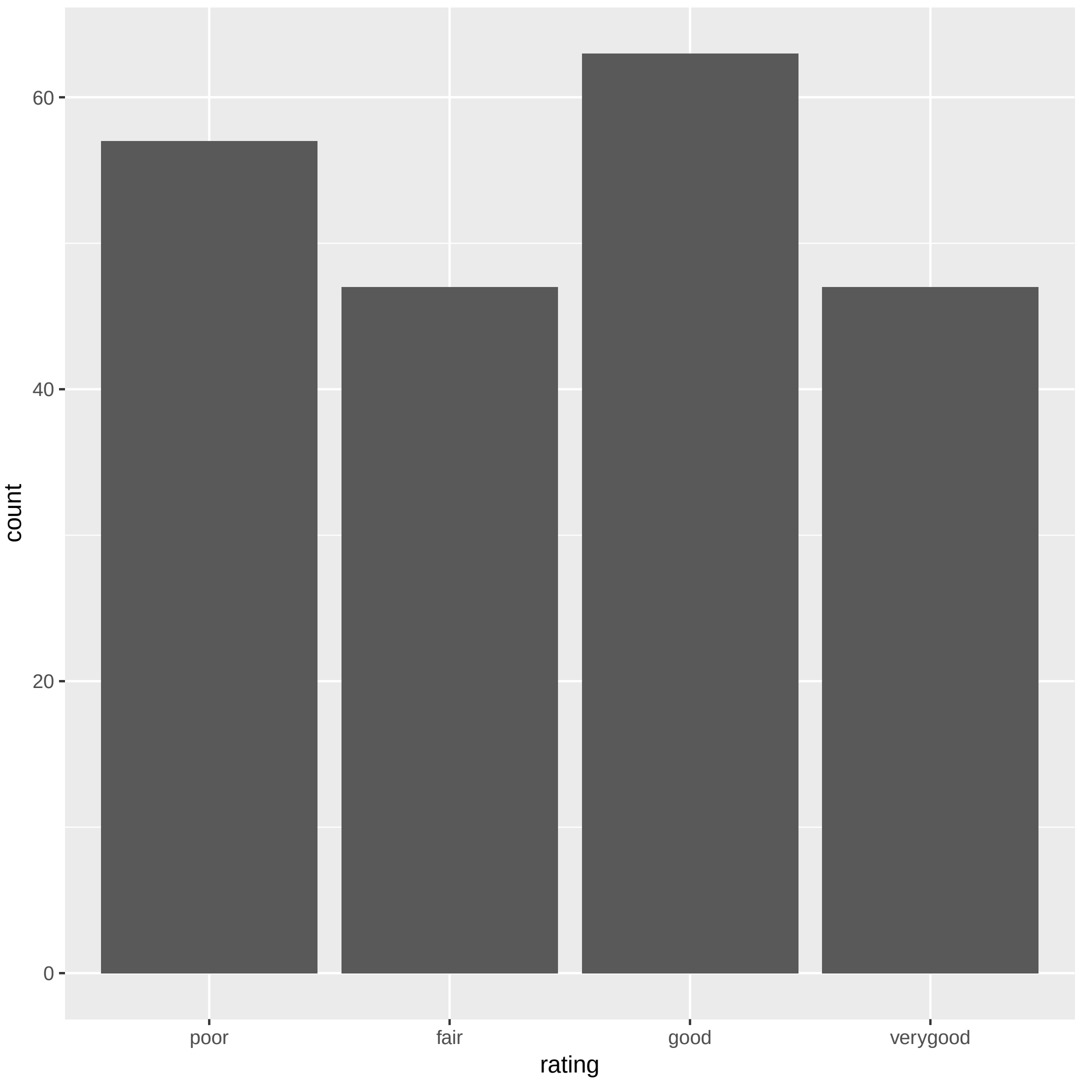
```
painrelief.0=update(painrelief.1,.~.-drug)
```

How can I remove the last explanatory variable, you ask? Well, let’s see what kind of predictions it gives:

```
p=predict(painrelief.0,new,type="probs")
cbind(new,p)
##   drug      poor      fair      good verygood
## 1  c15 0.2663551 0.2196262 0.2943927 0.219626
## 2  c60 0.2663551 0.2196262 0.2943927 0.219626
## 3 z100 0.2663551 0.2196262 0.2943927 0.219626
```

The predictions are *the same for each response category regardless of the drug (that is, reading down the columns)*. This is a bad fit because we already know that drug `z100` was much better than the others, but it still works. What it does is to pool all the ratings together regardless of drug, and estimate the percentage of all the ratings that were very good, good, etc. In other words, instead of working from the “grouped” bar chart in Figure 33, it works from this one, where we *don’t* group by drug:

```
ggplot(painrelief,aes(x=rating,weight=frequency))+
  geom_bar()
```



The frequency of each rating is pretty uniform all the way across, but what this hides, and is important to know, is that most of those "very good" ratings came from drug `z100` and most of the "fair" and "poor" came from the other two drugs. This shows in another way that it's worth keeping `drug`, not just because it's significant, but because it helps us explain what is going on. (We don't have an R-squared-like thing here, but if we did, it would be much higher for the model including `drug` than for the model without, since we are explaining so much more of what is going on when we include `drug`.)

In the `ggplot`, there were 12 rows in our data set; for this bar chart, the 3 rows for each rating have their frequencies added together.

(g) (2 marks) Are the predictions in Figure 17 consistent with the bar charts in Figure 33? Discuss briefly.

My answer: The predictions say that a patient who took drug `c15` or `c60` is about equally likely to rate the drug "poor", "fair" or "good", and unlikely to rate it "very good". A patient taking drug `z100`, however, is very likely to rate it "good" or "very good". This is exactly what the bar charts said, so the model is very much consistent with the data.

I'm happy with any sensible discussion here, though I want you to go a bit further than saying "drug has a significant effect and the bars have different heights for the different drugs" (which is worth one point by itself): I'd like you to compare, somehow, the predictions with the relative heights of the bars on the bar chart.

The model has done a bit of "smoothing": for example, the prediction for "poor" is bigger than for "fair" for both `c15` and `c60`, even though for drug `c15` more people rated it "fair" than "poor". Feel free to point out inconsistencies like this in your answer, though I think the "big picture" is that the bar charts are, overall, consistent with the predictions.

5. *Clematis ligusticifolia* or western white clematis is a climbing vine with showy flowers. It is also known as "old man's beard", since that is what its flowers look like. An ecologist is studying this vine, in particular whether its male or its female flowers are visited more frequently by insects. The ecologist observed waiting times during the blooming period. She watched a particular flower and waited until an insect landed on it, recording (i) how many minutes this was, (ii) whether it was a male or female flower. In some cases, she watched a flower for a long time and never saw an insect land on it. She therefore also recorded whether an insect was observed at the recorded time (`observed=yes`) or whether the time was when she stopped watching (`observed=no`). The structure of the data is shown in Figure 18, along with some randomly-chosen rows of the dataset.

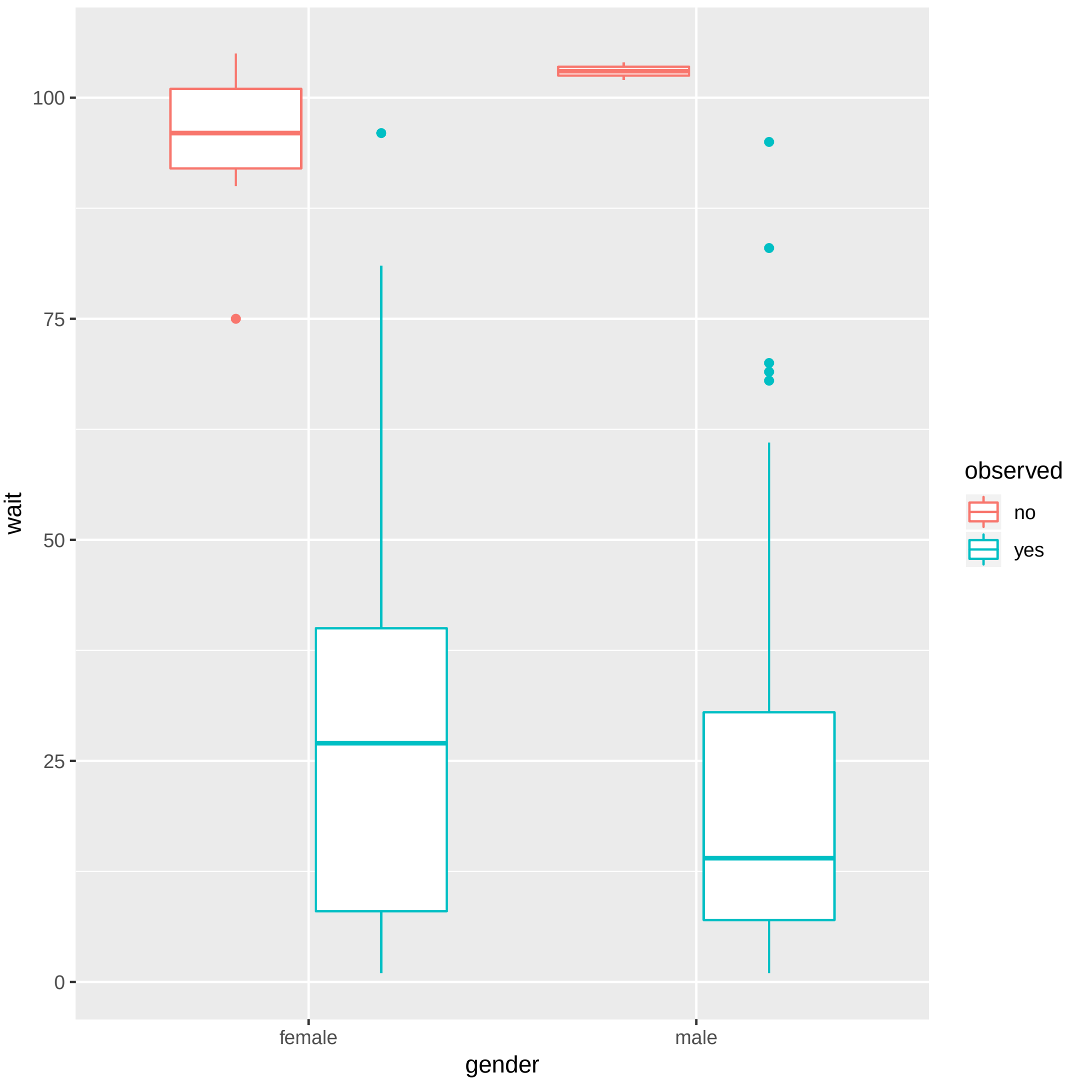
   (a) (2 marks) Figure 19 shows side-by-side boxplots of waiting times for insects to arrive at the two genders of flowers. Why could these boxplots give a misleading comparison for these data? Explain briefly.

My answer: The waiting times were of two types: ones where an insect was observed, and one where the ecologist gave up waiting. These are *not distinguished* on the boxplots. The actual waiting time for an insect could be a lot greater than the "censored" times, and treating the censored times as if they were observed could give a false impression.

Some other discussion of the boxplots themselves is worth one point.

Here's what I think *should* be done:

```
clematis=read.csv("muenchow.csv",header=T)
ggplot(clematis,aes(x=gender,y=wait,colour=observed))+geom_boxplot()
```



Now we are comparing like with like, and we can say "out of those times when insects were actually observed, the typical (median) waiting time for an insect to land on a female flower is longer than on a male flower".

The shape of the distribution (or the presence of outliers) doesn't matter very much, since the Cox model can handle that; indeed, distributions of waiting times are *typically* skewed to the right, since you can't see something sooner than "now", but you might wait a very long time. (Think about waiting for a number 38 bus on a snowy day.)

(b) (3 marks) Figure 20 shows the construction of a variable `y`. What is the purpose of the `observed=="yes"` in the `Surv` line, and what do the `+` signs mean next to some of the values in the display of `y`?

My answer: `Surv` requires two things: the variable containing the times-to-event, which is `wait` here, and also an indicator of whether the event happened or not. In this case, `observed` is `yes` if an insect was observed to land on a flower, which is the "event" we are dealing with here.

The values of `y` with a `+` next to them are "censored": that is, they correspond to a waiting time where no insect arrived (the ecologist stopped looking).

I want to see that you know what the variables represent *here*, so that copying something out of your notes will only help you a little. Note also that we are using *all* the data, including the censored values, in the analysis; what we want to know is *which observations are which*, observed or censored. If your answer suggested that we *only* used the observations where an insect was observed to land, I took a point off, since that's not what is happening.

Some people were getting explanatory and response confused here. Waiting time and `observed` are both part of the response, and `gender` is the only explanatory variable.

(c) (1 mark) In Figures 21 and 22, two Cox proportional-hazards models are fitted. Which explanatory variables are included in the model `clematis.0` of Figure 22?

My answer: None. It is a model that predicts time-to-event from all the flowers combined, without distinguishing gender of flower. "Just an intercept" or similar is also good.

(d) (2 marks) Using Figures 21 and 22, is there a significant effect of gender (of the flower)? How can you tell? What does that mean in the context of these data? Explain briefly.

> **My answer:** In the `anova` to compare `clematis.0` and `clematis.1`, the P-value is small, so that the bigger model (the one containing `gender`) is better. That is to say, `gender` of flower makes a difference in time until an insect lands: one of the genders attracts insects faster than the other.
>
> I wanted two things: a consideration of whether we should keep `gender` (based on either or both of the two Figures), and then a statement about time until an insect landed on flowers of different genders: that the waiting times until an insect lands are significantly different for male and female flowers. (Which flower gets insects quicker is for later, and I didn't judge whether you had it the right way around here, so that anything suggesting what "an effect of gender" means is good. I wanted something a bit more than "there is an effect of gender".
>
> I realized that I messed up the question: I only meant you to look at Figure 22 (the `anova`) here. Inviting you to look at the other Figure as well probably confused the issue, or at least said the same thing a different way. I have tried to mark allowing for the possibility that I confused you.

(e) (3 marks) The ecologist's research hypothesis was that male flowers would be more attractive to insects than female flowers. What would that imply for waiting times for insects on male and female flowers? Does the graph in Figure 34 (at the end of the booklet of code and output) support that hypothesis? Explain briefly.

> **My answer:** If male flowers are more attractive to insects, then the waiting time for an insect on a male flower should be *less* than on a female flower (one point for this, stated or clearly implied). On the graph in Figure 34, the male survival curve (the blue one) is lower than the female (red) one, which implies that for the males, the event (seeing an insect) happens *more quickly*. (The other two points for this. If you mess up along the way, but you make a logically consistent statement somewhere, you'll get a point.)
>
> This is the opposite way around to the usual survival-analysis situation, where the event is something like death that we don't want to see. Here, the event is something we *do* want to see, so we want it to happen more quickly. You can reason this out from the graph: at time 50 (minutes), say, the probability that an insect has not been seen yet on a female plant is about 0.4, but the probability that an insect has not been seen yet on a male plant is only about 0.2. (You can tell it's "not seen yet", since the probability as time passes goes from 1 down to 0.)
>
> I don't want you to religiously stick to "upper-right survival curve is better", since here it isn't: I want you to *think*.

(f) (2 marks) One of the numbers on Figure 21 supports your conclusion of the previous part. Which one, and why?

> **My answer:** The slope coefficient for `gendermale`, 0.5069. This is positive (and significantly nonzero) compared to `genderfemale`, which, as the baseline, is fixed at zero. This means that for male flowers, the event (an insect landing) is more likely to happen faster. This is the same conclusion as we obtained from the survival curves.
>
> The small P-value (0.0195) says that there is an effect of gender *but it doesn't say which way it goes*: that is, the P-value shows that one of the genders attracts insects faster than the other, but not that males are more attractive than females. For that, you need the positive coefficient and a decent explanation of why that means male plants are more attractive.

6. An evil Statistics lecturer has his students write a final exam in essay form. A randomly chosen half of the students write the final essay exam using a regular exam book (`bluebook`), and the other half use a laptop computer (`computer`). In addition, he assesses how much typing experience each student has, classified as "none", "some" or "lots". The handwritten exams are transcribed, and all exams are printed out onto the same paper, so that the lecturer grades each exam without knowing whether it was originally handwritten or typed. The response variable is the `score`. The data are shown in Figure 23.

   (a) (2 marks) In Figure 24, how do the factors `ability` and `ability.ord` differ? Explain briefly.

   My answer: `ability.ord` was created as an ordered factor, with the levels in the order shown (the logical order). The factor `ability` has its levels in alphabetical order, since it was read in from the data file.

   I figured that the logical order would be easier for you to make sense of in the graph later, so I created it and then figured I could ask you about it.

   (b) (2 marks) In Figure 24, I calculate the mean `score` for each combination of `ability` level and `method`, and in Figure 35 (and the end of the booklet of code and output) I draw a graph with the values I calculated. By looking at Figure 35, what is likely to happen in the analysis, and why? Explain briefly.

   My answer: This is an interaction plot, so we need to assess whether the red and blue "traces" are approximately parallel or not. I think they are not even close to parallel, which means that we would expect to see a significant interaction between `ability` and `method`.

   I'll take some sensible discussion that gets at something like this, for example that the highest (average) exam score for people writing on a computer is for those with some typing experience, whereas for those writing in a blue book, the highest score is for those with lots of typing experience. This gets at the interaction and shows a sensible reading of the graph.

   (c) (3 marks) In Figures 25 and 26, two analyses are shown. Which one is the more reasonable to base our conclusions on? What, therefore, do you conclude? Explain briefly.

   My answer: The procedure is to fit the model with interaction first (this is Figure 26) and look at the results. In this case, the interaction is significant, and so we stop here and say that a student's score depends on the combination of their typing ability and their method of writing the final exam. Or, the effect of a student's typing ability on their score depends on their method of writing the final exam (which sounds obvious when you say it that way).

   That is to say, we do not even look at Figure 25, or at the main effects in Figure 26.

   The way in which the score depends on the `method-ability` combination is something we explore in a moment, but for now we can conclude that there is some dependence there.

   I was fairly relaxed in terms of what I would accept, but I did want to see the word "interaction" somewhere in your answer, and not just something like "`ability:method` is significant" without any indication that you know what that meant.

(d) (2 marks) Look at Figure 27. Which students are included in this analysis, and what is being tested here?

My answer: The students who wrote their essay on a computer; we are testing to see whether, for these students, their `score` depends on their typing ability. This is the so-called *simple effect* of ability on score for the students who wrote their essay on a computer.

Using the words "simple effect" was *almost* enough to get both marks all by itself! Saying "testing the simple effect of typing ability on score for those students who wrote the essay on a laptop" *was* enough for both marks.

(e) (3 marks) What do you conclude from Figure 27? Give a complete answer, using the results from the whole of the Figure, as appropriate. **Use** α **of 0.10**.

My answer: From the ANOVA, for the students who wrote the essay using a computer, there is a significant effect of typing ability on score. So we look at the Tukey. At $\alpha = 0.10$, having `some` typing ability results in a significantly different score than having `none` or `lots`. If you look at the confidence intervals, which I made 90% intervals to match our $\alpha = 0.10$, you see that having `some` ability at typing results in a *higher* score than either having ability `none` or (surprisingly) `lots`.

You can hypothesize about why that might be.

This is the same picture as from the interaction plot, Figure 35: on the blue trace, the mean is much higher for `some` than for either `none` or `lots`.

I figured that a complete answer was to get to "`some` significantly different from `none` or `lots`", since this was the end-point and I didn't specify any detail in the question. I might have asked "for the students who wrote their essay on a computer, which level of typing ability is associated with the highest scores?", expecting the answer "those with some typing ability", but I didn't do that, so I couldn't demand that you said that. It was nice if you did, but I couldn't insist on it.

(f) (3 marks) What do you conclude from Figure 28? Again, give a complete answer, using the results from the whole of the Figure, as appropriate. Again, **use** α **of 0.10**.

My answer: This time, there is no evidence of an effect of typing ability on score, for those students who wrote their essay by hand. (This is hardly a surprise: that typing ability has nothing to do with how well you can hand-write an essay.)

Because the ANOVA F-test is not significant, you don't look at the Tukey, though if you did, it would tell you the same story: that there are no significant differences between any of the groups.

If you look at the red trace on the interaction plot, it looks as if score is increasing with typing ability, but these differences are not large enough to be significant: even the difference between `none` and `lots` for these students only gets down to a P-value of 0.12.

7. 26 samples of Romano-British pottery were found at four different sites. The samples were analyzed by atomic absorption photometry to measure the percentages of oxides of various different metals contained in the samples. The aim of the study was to see whether pottery collected from different locations had a different profile of metal oxides. The data are shown in Figure 29. In this question, we focus on iron oxide, labelled `Fe` in the data. The four sites are Llanederyn and Caldicot, both in the Gwent area of south Wales, and Island Thorns and Ashley Rails, both in the New Forest area of England. Boxplots of the iron oxide values for the four sites are shown in Figure 30.

   The researchers were most interested in whether (i) the two sites in south Wales were different from each other, (ii) the two sites in England were different from each other, and (iii) the two sites in England were on average different from the two sites in Wales.

   (a) (2 marks) Why is it better to address the researcher's interests using contrasts rather than a regular ANOVA followed by Tukey? Explain briefly.

   My answer: The researchers have specific research hypotheses to test, which (i) they can enumerate before looking at the data and (ii) only involve comparing some means, and not all possible means (so that Tukey would be a waste).

   It was possible to answer this question by copying something from my notes, but you should be careful about this because it won't often work: you'll usually have to adapt something.

   (b) (3 marks) Some computations are shown in Figure 31. Explain briefly but specifically how the computations are related to the researchers' questions.

   My answer: These are constructing contrasts that will enable us to test the three research hypotheses. Specifically, `c.a` compares the two sites in England (hypothesis (ii)), `c.b` compares the average of the two sites in Wales with the average of the two sites in England (hypothesis (iii)), and `c.c` compares the two sites in Wales (hypothesis (i)).

   The last two lines are constructing a matrix of contrasts, and setting things up so that `lm` will test those contrasts in the `lm` to follow. (That isn't really related to addressing the researchers' questions, but may be worth something if you say it.)

   I wanted, mainly, for you to link the contrasts I defined to the researcher's questions.

   (c) (3 marks) What do you conclude from Figure 32? There are three things to say.

   My answer:
   In the order that the contrasts are listed:

   - there is no difference in iron content between the two English sites (P-value 0.6584)
   - the Welsh sites are very different from the English sites in iron content (P-value 2.42×10^{-11})
   - there is no difference in iron content between the two Welsh sites (P-value 0.0865), or, if you prefer, there is marginal evidence of a difference between these two sites. If you want to call it significant, say what α you are using, as in "significant at $\alpha = 0.10$", or use words like "marginally significant", or else I'll assume you meant $\alpha = 0.05$ and got confused.

   I put the boxplots in Figure 30 so that you could sanity-check these results. The two sites in Wales (the second and fourth ones) have way more iron oxide than the two sites in England; given the variability, the two sites in England are very similar; Llanederyn has more iron oxide than Caldicot, but the latter only has two pottery samples, so it's not clear whether the difference will be significant.

I wanted the comparisons that reflected the research hypotheses. Admittedly, if you couldn't do (b) this was a bit challenging, so if that was the case, I've tried to be generous with your mark for (c).

I went on Google Maps to find out where Llanederyn is: it's an eastern suburb of Cardiff, but nobody seems to know whether it's spelled Llanederyn or Llanedeyrn. The former spelling turns up mostly this pottery data set (in the R world, a very well-known data set), while the latter seems to hit the place. I knew where Caldicot was: it's further north on the Welsh side of the River Severn. There's a famous Roman fortress called Caerleon near there; I remember going on a school trip there from my elementary school, when we were doing a module on Roman history. We all rode on a train to Newport, and then a graffiti-filled local bus. (I don't remember much about the actual Roman fortress.)