STAD29 / STA 1007 assignment 3

Due Tuesday Jan 28 at 11:59pm on Blackboard

Hand in the indicated questions. In preparation for the questions you hand in, it is worth your while to work through (or at least read through) the other questions as well.

Hand in your work on Quercus. If you did STAC32 last fall, it's the same procedure. A reminder is here: `https://www.utsc.utoronto.ca/~butler/c32/quercus1.nb.html`

You are reminded that work handed in with your name on it must be *entirely your own work*. It is as if you have signed your name under it. If it was done wholly or partly by someone else, *you have committed an academic offence*, and you can expect to be asked to explain yourself. The same applies if you allow someone else to copy your work. The grader will be watching out for assignments that look suspiciously similar to each other (or to my solutions). Besides which, if you do not do your own assignments, you *will* do badly on the exams, because the struggle to figure things out for yourself is an important part of the learning process.

Some packages that you will probably need too:

```
library(MASS)
library(tidyverse)
library(nnet)
```

I am also using `conflicted` which requires (for this assignment) this code:

```
library(conflicted)
conflict_prefer("select", "dplyr")

## [conflicted] Will prefer dplyr::select over any other package

conflict_prefer("filter", "dplyr")

## [conflicted] Will prefer dplyr::filter over any other package
```

I used `conflicted` as well because both `tidyverse` and `MASS` have a `select` and I want to make sure I get the right one. There is also a `filter` in base R that I make sure I don't get by mistake.

1. Review the questions in Chapter 19 of PASIAS that relate to multiple logistic regression (where you have several explanatory variables but a yes/no response): 19.4, 19.5, 19.6, 19.8, 19.10.

   Hand the next one in.

2. A study was made on the effects of analgesics (painkillers) on neuralgia patients. Two different treatments were used, labelled A and B, along with a placebo, labelled P. The researchers also recorded the age and sex of each patient, and the duration of pain before the treatment began. The response variable was whether or not the patient reported pain after the treatment.

   (a) (2 marks) What feature of this study makes logistic regression a plausible method of analysis?

   (b) (2 marks) The data are in `http://ritsokiguess.site/STAD29/neuralgia.txt`. Read the data into R, and check that you have all the promised variables. (Hint: check the layout of the data.)

   (c) (2 marks) Fit a logistic regression predicting `pain` from everything else. Display the output. Hint: what kind of thing does your response variable need to be?

   (d) (3 marks) When you have a mixture of quantitative and categorical explanatory variables, it's hard to tell from the `summary` output which ones should be kept. Pass your model into `drop1` with `test="Chisq"`. Which explanatory variable(s) are candidates to be removed? Explain briefly.

(e) (2 marks) Remove any non-significant explanatory variables and fit again, and display the results.

(f) (1 mark) What is the thing that we are predicting the probability of? Explain (very) briefly.

(g) (2 marks) By looking at the numbers in the Estimate column of the output from your most recent model, describe how the treatments differ, and thus which one(s) are best or worst.

(h) (3 marks) Use your most recent model to do some predictions, as below. We will use all possible combinations of the three treatments, the two sexes, and the two ages 65 and 75 (which are about the 1st and 3rd quartiles of the ages, so they are more or less "typical" of the ages we have). We're going to use `predict`. Use the following line to get the "new" data to predict from:

```
new <- crossing(treatment=c("A","B","P"),sex=c("F","M"),age=c(65,75))
```

and check that this did indeed give you all possible combinations. (Whenever you want "all possible combinations", `crossing` is probably what you want.) Obtain the predicted probabilities of pain for each of these, using `predict` with a suitable `type` to make sure that you get probabilities. Put the predicted probabilities side by side with the things they're predictions for.

(i) (2 marks) By looking at the predicted probabilities from the last part, and by comparing predictions for the same sex and age, what is the effect of treatment on pain? Specifically, are treatments A and B better than the placebo? Do they differ much from each other?

(j) (3 marks) Again looking at the predicted probabilities, how do you describe the effect of age? Of sex?

3. Work through, or look at the problems in Chapter 20 of PASIAS that relate to ordinal responses: 20.1, 20.2, 20.8, 20.10. You don't need to worry about nominal responses yet.

   Hand the next one in.

4. Breast cancer affects a lot of women, but it can be treated if caught early enough. One way this can be done is to ask women over 40 to have a procedure called a "mammogram" every year. This is an x-ray that uses a low dose of radiation to help find any lumps or unusual areas of breast tissue, which can then be investigated further to rule out cancer or anything else malignant. There is a lot more information at `https://www.cancer.ca/en/cancer-information/diagnosis-and-treatment/tests-and-procedures/mammography/?region=on`. The kind of procedure we are concerned with here is called "screening mammography" there.

   Encouraging otherwise healthy women over 40 to get a mammogram is a public health benefit because treating breast cancer is easier and cheaper if it is caught earlier. It is, of course, also better for the women who undergo a mammogram, at least in the long run.

   A study was carried out on attitudes towards mammography in women over 40. Several things were recorded, mostly categorical:

   - `obs`: a code identifying each respondent (ignore)
   - `me`: "mammograph experience", values `less1` (last mammogram was less than 1 year ago), `more1` (last one was more than 1 year ago), `never`. This is the response variable.
   - `sympt`: response to the statement "you do not need a mammogram unless you develop symptoms", on a 4-point scale from Strongly Disagree to Strongly Agree. (Later, you will need to think carefully about what a Strongly Disagree response to this statement actually *means*.)
   - `pb`: Perceived Benefit of mammography. This is the sum of five responses on a 1–4 scale (with, therefore, a minimum score of 5 and a maximum of 20). A *low* score indicates strong agreement with the benefits of mammography.
   - `hist`: family history of breast cancer (meaning, mother or a sister), yes or no.
   - `bse`: "has anyone taught you how to examine your own breasts" (such as a doctor or a nurse), yes or no.
   - `detc`: "how likely do you think it is that a mammogram could find a new case of breast cancer?", not likely, somewhat likely, very likely.

I had to do a bit of data reorganization (which I talk about in an Extra in my solutions), so the data is in a (to you) unusual format, which I will lead you through reading in.

(a) (2 marks) Read in and display (some of) the data. The data frame is at `http://ritsokiguess.site/STAD29/meexp.rds`. (You won't see anything if you try to look at this.) Use `readRDS` to read it in. First store the URL in a variable called (for example) `my_url`, and then do something like this to put it in a data frame called `meexp`:

```
meexp <- readRDS(url(my_url))
```

The extra `url()` seems to be necessary.[6]

(b) (2 marks) What order are the categories of the response variable `me` in? Does this order make sense? Explain briefly. (Hint: investigate `levels` or `distinct` or `summary` of a data frame.)

(c) (2 marks) Why is it that a logistic regression with `polr` would make more sense than one with `multinom`? Explain briefly.

(d) (2 marks) Fit a suitable model with `polr`. You don't need to display the output.

(e) (3 marks) Use `drop1` with `test="Chisq"` to demonstrate that none of the explanatory variables should be removed.

(f) (3 marks) The usual way of understanding a model of this kind is to do some predictions. This is made harder here by the large number of significant explanatory variables. One place to start is to find some "typical" values to work from. To set this up, find the median of each of the quantitative variables, and find the combination of categories of all the categorical variables that has the most observations. (Hint: for the categorical variables, count them all, and put the frequencies in order.)

(g) (4 marks) Investigate the effect of family history of breast cancer on the recency of a mammogram. To do this:

- create a new data frame with two rows: the first row contains what you got in the previous part, and the second row is the same as the first, except that the family history variable takes the other value. The names of the variables in your new data frame must be the same as the ones in the original data frame.
- Then obtain predicted probabilities for each response category for each of those two rows, displaying the predictions next to the values they are predictions for.
- Finally, explain whether the predictions are what you'd expect, in language that a hospital *administrator* or other public health official would be able to understand.

(h) (4 marks) Repeat the previous part, but this time assessing the effect of answer to the statement "you do not need a mammogram unless you develop symptoms" on the recency of mammogram. Again, do the results make practical sense?

Notes

[1] Like ANOVA.

[2] The slight difference between A and B was just chance.

[3] As the model dictates.

[4] If it were not, you would have an ANOVA-style *interaction* between the variables concerned.

[5] This is why age was significant.

[6] This is because the input to `readRDS` has to be what R knows of as a "connection". If you just give it the URL, it'll think it's some kind of file which it then won't be able to find, but if you wrap it in `url()` it knows that it's really a URL and should be treated as such.

[7] I wasn't sure whether they were this or tabs, but I tried this and it worked.

[8] Think about what will happen if I leave the `never` as 0: will the three categories then be in a sensible order?

[9] You can't get less recent than "never"!