STAD29 / STA 1007 assignment 4

Due Tuesday Feb 4 at 11:59pm on Quercus

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.

You will need this first:

```
library(tidyverse)
```

Hand in questions 2 and 4.

1. Work through, or look at the problems in Chapter 20 of PASIAS that relate to nominal responses: 20.4, 20.5, 20.9, 20.11. (The other questions in Chapter 20 are about the organization of some of the data sets used in the chapter; feel free to go through those too.)

2. A survey called High School and Beyond was given to a large number of American high school seniors (grade 12) by the National Center of Education Statistics. The data set at `http://ritsokiguess.site/STAD29/hsb.txt` is a random sample of 200 of those students.

   The variables collected are:

   - `gender`: student's gender, female or male.
   - `ses`: Socio-economic status of student's family (low, middle, or high)
   - `schtyp`: School type, public or private.
   - `prog`: Student's program, general, academic, or vocational.
   - `read`: Score on standardized reading test.
   - `write`: Score on standardized writing test.
   - `math`: Score on standardized math test.
   - `science`: Score on standardized science test.
   - `socst`: Score on standardized social studies test.

   Some of these variables are quantitative and some are categorical.

   You will recognize this as one of the data sets from PASIAS. This time, we take a different angle: we try to predict which program a student went into based on the values of the other variables.

   (a) (2 marks) Read in and display (some of) the data.

   (b) (2 marks) Use `multinom` to fit a model predicting program from everything else that makes sense as an explanatory variable. Exclude everything after `socst`. No need to display the output.

(c) (3 marks) There are a lot of explanatory variables. To see whether any of them can be removed, we can use `step` on one of these models.[1] Run `step` with `direction="backward"` on your model from the previous part, saving the result and then displaying the `summary` of it. Which explanatory variables seem to be still in the model? Hint: `step` produces a lot of output; the input `trace` controls how much output there is. See if you can minimize the amount of output.

(d) (3 marks) The `summary` output is very hard to understand. Let's set up to do some predictions. First, for each of the explanatory variables in your final model from `step`, find out all its levels if it is categorical, and find its (first and third) quartiles if it is quantitative. Do this how you like. You don't need to be clever.

(e) (3 marks) Make a data frame that includes all combinations of values you obtained in the previous part. (I have five variables, two of which are categorical, and one of those has three levels, so I have 48 rows.)

(f) (3 marks) Obtain predictions, using your model that came from `step` and the data frame of values to predict for that you just created. Be sure to obtain predicted *probabilities* for each response category. Display (some of) your predictions side by side with the values they are predictions for. Save your final data frame.

(g) (3 marks) Assess the effect of `ses` on the probabilities of a student being in each of the three programs by displaying appropriate rows of your saved data frame of predictions (using `slice` or `filter`). Which programs do students of each socioeconomic status mostly end up in?

(h) (3 marks) By looking at an appropriate choice of rows from your data frame of predictions, assess the effect of an increase in `math` score on a student's choice of program.

3. Work through, or at least read through, chapter 21 of PASIAS.

4. A small clinical trial is run to compare two combination treatments in patients with advanced gastric cancer. Twenty participants with stage IV gastric cancer who consent to participate in the trial are randomly assigned to receive chemotherapy before surgery or chemotherapy after surgery. The primary outcome is death and participants are followed for up to 48 months (4 years) following enrollment into the trial. (The foregoing is directly taken from the website where I got the data; it's not how I would have written it, but I think it's good for you to see how this kind of thing is written by others.)

   The data are shown in `http://ritsokiguess.site/STAD29/chemo.csv`. There are three columns: whether each patient had chemotherapy before or after surgery, how many months they were observed for, and whether or not they were observed to have died.

   (a) (2 marks) Read in and display (some of) the data.

   (b) (2 marks) In the context of *this* data set, what would a censored observation look like? Give an example of one from the data set.

   (c) (3 marks) Create a response variable suitable for a Cox proportional-hazards regression. Do this outside the data frame, and display at least some of its values. How are the censored observations distinguished?

   (d) (3 marks) Fit a Cox proportional-hazards model to predict survival time from the treatment `chemo`. Display the results.

   (e) (2 marks) Is there any evidence that one of the treatments is better than the other? If so, which treatment is better? Explain briefly. (Hint: you will get this backwards unless you are careful.)

   (f) (3 marks) Make a suitable plot that shows which treatment is better, and explain briefly why it shows that.

Notes

[1] Oddly, `step` works but `drop1` doesn't. I have yet to figure out why that is.

[2] The principle with `filter` is that you omit the explanatory variable whose effect you want to see, and supply values, any ones, for the others.