Department of Statistics
STATS 330 Statistical Modelling
Assignment 3 (2020; Semester 1)

Total: 50 marks **Due: 23:59 NZDT, May 24, 2020**

Notes:

(i) Write your assignment using R Markdown. Knit your report to either a Word or PDF document.
(ii) Create a section for each question. Include all relevant code and output in the final document.
(iii) 7 presentation marks will be used; these will be deducted if offences occur. Please think of your markers—keep your code and plots neat and remember to check your spelling.
(iv) Please remember to upload your Word or PDF document to Canvas by the due date.
(v) Please remember to upload your R Markdown file to Canvas before the deadline, too. If the markers identify an error in your work, being able to run the code you have written can help determine what you did wrong
(vi) **Remember to comment on all your output!**

1. [10 marks] **Heart attacks** Consider Question 5 of Assignment 1 concerning the `hearthealth` data. The model answers suggested that an object called `ha.lin` was the most appropriate. Let's do a quick double check.

   (a) Using VGAM, fit a GAM to `heartattack` with the variable `age` smoothed and cholesterol in the model too. Apply `summary()` and comment. [5 marks]

   Notes:

   - Don't unpack the variables from the data frame. Keep the variables in there and use the `data` argument.
   - If you are using R 4.0.0 then it might be good to create

   ```
   > Heart.df$fchol <- factor(Heart.df$chol)  # Good idea for R 4.0.0
   ```

   and use `fchol` instead of `chol`.

   (b) Plot the estimated component function and comment. Write down the mathematical formula for the model. [5 marks]

2. [20 marks] **Womens' BMI** Consider the BMI data placed on Canvas.

   (a) Read in the data into a data frame. It may be a good idea to sort by `age`. Find the proportion of "obese" people and the proportion of "overweight" people who are not "obese", according to the World Health Organization. Give details of any website that you use. Comment. [4 marks]

   (b) Produce a high quality scatterplot Y = body mass index (BMI) versus X = `age` for European-type women. As usual, remember to comment. [5 marks]

   (c) Use `smooth.spline()` to fit a smoother to the data. Print out its EDF and add your smooth to the scatter plot. Comment, and in particular, give an explanation for any trend. [5 marks]

   (d) What BMI value does your smooth suggest for a 50 year old European-type women? Comment. [2 marks]

   (e) Try fitting a quadratic to these data and add it to your plot in (c). Comment. [4 marks]

3. [20 marks] **COVID-19** Consider the NZ COVID-19 data placed on Canvas and/or the class webpage. *As with all questions, please comment on all your output, especially plots—demonstrate that you understand the results and what the data is saying.*

   (a) Read in the data.

   - Some counts are -1, so replace them by 0.
   - Create a numerical variable called `Day` that runs from 1 to the number of rows of the data frame—you will be smoothing with respect to `Day` below.
   - Hint: `as.Date()` may be useful.

   Then plot the data, either in the form of a histogram (but don't use `hist()`) or an ordinary scatterplot. The latter is preferred because you will be adding fitted values to it below. Comment. [4 marks]

   (b) Using `mgcv`, fit a simple GAM to the data. Plot the estimated component function with standard error bands. And add the fitted values to your plot from (a). Comment. [5 marks]

   (c) Fit a quadratic model to the data. Add the fitted values to your plot in another colour. Comment on how the parametric and nonparametric models compare, e.g., using AIC. [4 marks]

   (d) Using your answers to (b) and (c), estimate the date where the number of new cases peaked. Is there much difference between the two models? Hint: `which.max()` may be useful. [4 marks]

   (e) A number of 0s were added to the data prior to the first case. This was deliberate. Why? Show some evidence for your answer. [3 marks]