Assignment 1
STAT8178: Statistical Computing

Session 1, 2020 Due Friday 20 March, via *iLearn*, by 9:00 p.m.

General instructions

For each of the questions below, you are expected to document your answer fully. Try to work within scripts or functions, rather than issuing commands at the command line, or manually manipulating spreadsheets, so as to leave a repeatable record of what you've done. If you *do* decide to manipulate data manually, please fully describe what you've done in your assignment submission, and submit any files (such as manipulated spreadsheets) that may clarify your work.

Where you are required to write code, please state and interpret the output, as well as submitting the code itself.

Also, some of your answers should be mathematical expressions. The best way to write these is LaTeX, but it has a long learning curve, and is probably not worth learning if you don't expect to do much technical writing after this unit. *Word*'s mathematical typesetting functionality has improved a great deal in the last decade; a quick intro to get you started appears at `https://en.wikibooks.org/wiki/Typing_Mathematics_in_Microsoft_Word`. If even that seems too hard and you want to just handwrite your equations, photograph them, and paste them into your answer submission, you may; I won't actually deduct marks, but you should feel vaguely ashamed, as if you've come to class with your shirt inside-out or something.

For convenience, the questions describe working with either *Matlab* or *R*. As always, if you believe you can do equivalent computations on a different platform to the one described, you are free to do so, and I am happy to discuss it if you're not sure. Regardless of which programming language you're using, you will probably need to look things up. You should freely use the built-in documentation and internet searches to learn the details of how to use each function or command that you need.

Question 1: MLE with Newton-Raphson (21 marks)

In this question, we will fit a seasonal model via maximum likelihood estimation.

For health resources planning, the UK's National Health Service keeps records of the use of its services, and much of this data is publicly available. The provided file

`MAR_Comm-Timeseries-Dec-19-REVISED-Apr-18-to-Nov-19-9tun8.xls`[1]

shows the number of admissions of various types throughout England, for each of the $N = 141$ months from April 2008 to December 2019. We would like to model the "total non-elective general and acute admissions" data (column `K`), allowing for the possibility of a seasonal trend.

If there *is* a seasonal trend, we'll assume it's a simple annual sinusoid, and therefore can be described as a linear combination of the cosine and sine of an angle ϕ_i that completes a full circle each year. Therefore, we will assume that

- $\phi_1 = 30° = \pi/6$ (corresponding to the first month in the sample, April 2008),
- $\phi_2 = 60° = \pi/3$ (corresponding to May 2008),
- $\phi_3 = 90° = \pi/2$ (corresponding to June 2008),

et cetera. Since there is clearly an upward trend as well, which looks as though it could be exponential, we will model the monthly admission counts y_i as

$$y_i \sim \text{Poisson}(\mu_i),$$

where

$$\mu_i = \exp\left(\beta_1 + \beta_2 i + \beta_3 \cos(\phi_i) + \beta_4 \sin(\phi_i)\right). \tag{1}$$

[1] downloaded on 4 March 2020 from
`https://www.england.nhs.uk/statistics/statistical-work-areas/hospital-activity/monthly-hospital-activity/`

(a) Describe a matrix X (with elements x_{ij}) such that the model (1) can be written in the neater and more convenient form

$$\mu_i = \exp\left(\sum_{j=1}^{J} x_{ij}\beta_j\right), \tag{2}$$

where $J = 4$. **(1 mark)**

(b) Find an expression for the log-likelihood ℓ. **(2 marks)**

(c) Differentiate Equation (2) with respect to β_j, and simplify the result by expressing it as a multiple of μ_i. Show the steps of your calculation. **(2 marks)**

(d) Find an expression for the jth component of the gradient of the log-likelihood, that is, $\partial\ell/\partial\beta_j$. **(2 marks)**

(e) Find an expression for the (j,k)th component of the Hessian of the log-likelihood, that is, $\partial^2\ell/\partial\beta_j\partial\beta_k$. **(1 mark)**

Let β, y and μ be the column vectors with entries β_j, y_i and μ_i, respectively, and let M denote $\mathrm{diag}(\mu)$, the diagonal matrix with the values μ_1 to μ_N along the diagonal. That is,

$$\beta = \begin{pmatrix}\beta_1\\ \vdots \\ \beta_J\end{pmatrix}, \qquad y = \begin{pmatrix}y_1\\ \vdots \\ y_N\end{pmatrix}, \qquad \mu = \begin{pmatrix}\mu_1\\ \vdots \\ \mu_N\end{pmatrix}, \qquad M = \begin{pmatrix}\mu_1 & 0 & \cdots & & 0\\ 0 & \mu_2 & & & \\ \vdots & & \ddots & & \vdots \\ & & & \mu_{N-1} & 0 \\ 0 & & \cdots & 0 & \mu_N\end{pmatrix}.$$

(f) Express the gradient and Hessian of the log-likelihood using matrix algebra, in terms of X, y, μ and M. **(2 marks)**

(g) Choose a reasonable starting value for the parameter vector β, and explain your choice. (Nothing very sophisticated is required; we just need to be in the general vicinity of the best fit. Probably only a single non-zero entry is necessary.) **(2 marks)**

(h) Write a script in *Matlab* to

- read the admissions data from the range `K15:K155` of the provided *Excel* file (you'll probably want to use the *Matlab* function `xlsread`);
- construct the matrix X;
- initialise the vector β;
- perform iterations of the Newton-Raphson method until the estimated values μ are no longer changing very much;
- output the resulting estimates for β.

Comment your code thoroughly to explain its logic. Run this script and state and interpret the output. **(9 marks)**

Question 2: Posterior distribution after a single coin toss (9 marks)

In this question we will produce a histogram of the posterior distribution of a coin's tails probability after a single coin toss.

In class we discussed prior distributions in the context of a coin toss. Suppose someone you don't know very well has a coin. You can't see the coin very clearly; it's too far away to see what's on either side. In particular, it may be a genuine coin, but you don't know that for certain. The coin is then tossed once and shown to you, and you can see that it shows tails.

(a) Describe and justify a prior distribution (i.e., prior to the single toss that you witnessed) for the parameter π, the coin's probability of showing tails on a single toss. This is a matter of opinion, so the right answer is not unique, but the prior distribution should accurately describe a reasonable opinion. **(3 marks)**

(b) Use an appropriate Bayesian inference method to produce a histogram of your posterior distribution, after the single coin toss that you witnessed. (Hints: depending on your prior, this may or may not be straightforward in *JAGS*. Another option is rejection sampling. Also, you'll probably need a large sample and narrow histogram bins to clearly display the posterior distribution.) **(6 marks)**