MAS6003(2)/MAS474 Extended Linear Models

Exercises 3

1. In the lecture notes we considered exponential random variables which were right-censored (i.e., values bigger than some threshold c were missing). Here, we will consider left-censored exponential random variables.

   Suppose the sequence of random variables $X_1, \dots, X_n$ are iid with an exponential distribution with mean $\frac{1}{\lambda}$, i.e., they have pdf

$$f(x;\lambda) = \begin{cases} \lambda \exp(-\lambda x) & \text{for } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

   Suppose further that X_i is missing if $X_i < c_i$ for some set of thresholds $c_1, \dots, c_n$. Finally, suppose that we are told that $X_1, \dots, X_r$ are observed, but $X_{r+1}, \dots, X_n$ are missing. The data can thus be summarized by the thresholds, the observed values of x, and the pattern of missingness.

   (i) Show that the likelihood function for λ is

$$L(\lambda) = \lambda^r \exp(-\lambda \sum_{i=1}^{r} x_i) \prod_{i=r+1}^{n} (1 - \exp(-c_i \lambda))$$

   and explain why this would be difficult to maximize.

   (ii) We will now consider using the EM algorithm to estimate λ. The complete data will consist of the observed x_i and the missing values $X_{r+1}, \dots, X_n$.

   a) Explain why

$$\mathbb{E}(X_i | X_i > c_i) = \frac{1}{\lambda} + c_i.$$

   b) The law of total expectation says that

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X | X < c_i))$$

   so that

$$\mathbb{E}(X) = \mathbb{E}(X | X < c_i)\mathbb{P}(X < c_i) + \mathbb{E}(X | X \geq c_i)\mathbb{P}(X \geq c_i)$$

   as $\{X < c_i\}' = \{X \geq c_i\}$.
   Hence, or otherwise, show that

$$L_i := \mathbb{E}(X_i | X_i \leq c_i) = \frac{1 - (1 + \lambda c_i)e^{-\lambda c_i}}{\lambda(1 - e^{-\lambda c_i})}.$$

   c) Hence derive an EM-algorithm for finding the maximum likelihood estimator of λ. You may leave your answer in terms of L_i.

   d) Generate some toy data in R (i.e., pick a values for the parameters, $n = 1000, \lambda = 1, c_i = 1$ say, simulate some exponential random variables, and then discard all the x values with $x < 1$), and implement your EM algorithm for estimating λ.

2. The `mice` package contains the dataset `mammalsleep`. The data are from Allison and Cicchetti (1976), and records information on the interrelationship between sleep, ecological, and constitutional variables for 62 mammal species. The dataset contains missing values on five variables. The covariates are

   - species - Species of animal
   - bw - Body weight (kg)
   - brw - Brain weight (g)
   - sws - Slow wave (”nondreaming”) sleep (hrs/day)
   - ps - Paradoxical (”dreaming”) sleep (hrs/day)
   - ts - Total sleep (hrs/day) (sum of slow wave and paradoxical sleep)
   - mls - Maximum life span (years)
   - gt - Gestation time (days)
   - pi - Predation index (1-5), 1 = least likely to be preyed upon
   - sei - Sleep exposure index (1-5), 1 = least exposed (e.g. animal sleeps in a well-protected den), 5 = most exposed
   - odi - Overall danger index (1-5) based on the above two indices and other information, 1 = least danger (from other animals), 5 = most danger (from other animals)

   i) Use the commands `md.pattern` and `md.pairs` to describe the pattern of missingness in the data.

   ii) Use the `mice` package to create 10 imputed datasets.

   iii) For each of your imputed datasets, fit a linear model to predict `ts` from `brw` and `bw`.

   iv) Using the `pool` command, combine these estimates to give an expected value and a standard error for the coefficient of `bw`.

   v) Repeat this calculation by coding up the pooling estimates by yourself (i.e. using Rubin’s formulas).