UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS
INCOMPLETE DATA ANALYSIS

Workshop 3

1. Suppose $Y_1, \ldots, Y_n$ are independent and exponentially distributed with density $f(y;\theta) = \theta \exp\{-\theta y\}$.

   (a) The survival function $S(y)$ is defined as

   $$S(y) = \Pr(Y > y).$$

   Show that $S(y;\theta) = \exp\{-\theta y\}$.

   (b) Suppose that observations are right censored if $Y_i > C$ for some known C and let

   $$X_i = \begin{cases} Y_i & \text{if } Y_i \leq C, \\ C & \text{if } Y_i > C, \end{cases} \qquad R_i = \begin{cases} 1 & \text{if } Y_i \leq C, \\ 0 & \text{if } Y_i > C. \end{cases}$$

   That is, we only observe the realisation of Y_i if it is less or equal than the censoring value C. If the realised value is greater than C, we only observe C.
   When Y_i is interpreted as time until some event (e.g., death), right censored data occur when the observation period ends before the event has taken place. It can also occur, for instance, when measuring a person's weight, if the weight exceeds the scale maximum.
   Show that the likelihood function of the observed data $\{(x_i, r_i)\}_{i=1}^n$ is given by

   $$\log L(\theta) = \log \theta \sum_{i=1}^{n} r_i - \theta \sum_{i=1}^{n} x_i.$$

   From it, derive the maximum likelihood estimator for θ.

   (c) Show that the expected Fisher information for the observed data likelihood is

   $$I(\theta) = \frac{n(1 - e^{-\theta C})}{\theta^2}.$$

   Comment the two limiting cases: $C \to \infty$ and $C \to 0$.

   (d) From the missing data mechanism perspective, are right censored data MCAR, MAR, or MNAR? Justify.

   (e) You will now conduct a small simulation study to investigate the impact of the naive approaches of

      (i) dropping all censored observations from the analysis, and

(ii) including the censored observations in the analysis but ignoring that C is not the true realized value,

on the estimation of the mean of the distribution. To this end, suppose that $\theta = 1/5$. Then, the mean of the Y_i's is

$$\frac{1}{\theta} = 5.$$

Generate 1000 samples, each with $n = 100$ observations. Then, for each sample, generate censored observations by assuming that observations are censored if they are larger than 10 (that is, $C = 10$). Also, for each sample, compute the estimate of the mean based on

$$\widehat{\theta}_{\text{naive}_1} = \frac{\sum_{i=1}^n r_i}{\sum_{i=1}^n r_i x_i}, \qquad \text{(case (i) above)},$$
$$\widehat{\theta}_{\text{naive}_2} = \frac{n}{\sum_{i=1}^n x_i}, \qquad \text{(case (ii) above)}.$$

Compute also the estimate of the mean based on $\widehat{\theta}$ found in (b). For each of the three cases, display the 1000 estimates in a histogram along with the true value. Comment.

2. The following question, although not involving incomplete/missing data, provides you with further practice about likelihood approaches and about numerical methods for finding maximum likelihood estimates, something that will be useful in the next weeks.

In linear regression the response is usually assumed to be on a continuous scale. Many other types of responses do however exist, making the need for different regression methods. Assume Y_i is a binary response variable ($Y_i \in \{0, 1\}$), while $\mathbf{x}_i$ is a vector of explanatory variables. In linear regression, the expected response is modelled as a linear function of the explanatory variable:

$$E[Y_i] = \mathbf{x}_i'\boldsymbol{\beta}.$$

Note however that in the case of a binary response, the expectation is equal to $\Pr(Y_i = 1)$. The linear regression model is surely inappropriate since the expected value may vary from $-\infty$ to ∞.
In a (binary) logistic regression, the response is modelled as

$$Y_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\{p_i(\boldsymbol{\beta})\},$$
$$E[Y_i] = \Pr(Y_i = 1) = p_i(\boldsymbol{\beta})$$

where

$$p_i(\boldsymbol{\beta}) = \frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}.$$

The parameter of interest to be estimated from the data is $\boldsymbol{\beta}$.

(a) For the case where there is only one covariate, so that

$$p_i(\boldsymbol{\beta}) = \frac{\exp\{\beta_0 + x_i\beta_1\}}{1 + \exp\{\beta_0 + x_i\beta_1\}}, \quad \boldsymbol{\beta} = (\beta_0, \beta_1)',$$

write down the likelihood, the log likelihood, and the score function.

(b) Using both the functions `maxLik` and `optim` (as in the lectures), find the maximum likelihood estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ for the (simulated) data `dataw3` available on Learn. Provide also the standard error associated to the estimates.

3. Suppose that as in 2., we have that

$$Y_i \overset{\text{ind}}{\sim} \text{Bernoulli}\{p_i(\boldsymbol{\beta})\}$$
$$p_i(\boldsymbol{\beta}) = \frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}$$

Assume that the covariates $\mathbf{x}$ are always observed, but that Y can be missing and, as usually, let R be the missing data indicator, taking the value one if Y is observed and zero otherwise. We further assume that

$$R_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\{\pi_i(\boldsymbol{\phi}_0, \phi_1)\},$$
$$\pi_i(\boldsymbol{\phi}_0, \phi_1) = \frac{\exp\{\mathbf{x}_i'\boldsymbol{\phi}_0 + y_i\phi_1\}}{1 + \exp\{\mathbf{x}_i'\boldsymbol{\phi}_0 + y_i\phi_1\}}.$$

State conditions on $\boldsymbol{\phi}_0$ and ϕ_1 under which (i) the missing data are MAR, and (ii) the missing data mechanism is ignorable for likelihood inference.