UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS
INCOMPLETE DATA ANALYSIS

Assignment 2

- To be uploaded to Learn by 23:59, November 22, 2020.
- This assignment is worth 40% of your final grade for the course.
- Assignments should be typed (LaTeX, word, etc.).
- Answers to questions should be in full sentences and should provide all necessary details.
- Any output (e.g., graphs, tables) from `R` that you use to answer questions must be included with the assignment. Also, please include your `R` code in the assignment (screenshots of the `R` console are not allowed) or make it available in a public repository (e.g., GitHub).
- The assignment is out of 100 marks.

1. Suppose $Y_1, \ldots, Y_n$ are independent and identically distributed with cumulative distribution function given by

$$F(y;\theta) = 1 - e^{-y^2/(2\theta)}, \quad y \geq 0, \quad \theta > 0.$$

Further suppose that observations are (right) censored if $Y_i > C$, for some known $C > 0$, and let

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

(a) (**7 marks**) Show that the maximum likelihood estimator based on the observed data $\{(x_i, r_i)\}_{i=1}^n$ is given by

$$\widehat{\theta} = \frac{\sum_{i=1}^n X_i^2}{2\sum_{i=1}^n R_i}.$$

(b) (**13 marks**) Show that the expected Fisher information for the observed data likelihood is

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

Note: $\int_0^C y^2 f(y;\theta)\mathrm{d}y = -C^2 e^{-C^2/(2\theta)} + 2\theta(1 - e^{-C^2/(2\theta)})$, where $f(y;\theta)$ is the density function corresponding to the distribution function $F(y;\theta)$ above.

(c) (**3 marks**) Appealing to the asymptotic normality of the maximum likelihood estimator, provide a 95% confidence interval for θ.

2. Suppose that $Y_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, for $i = 1, \ldots, n$. Further suppose that now observations are (left) censored if $Y_i < D$, for some known D and let

$$X_i = \begin{cases} Y_i & \text{if } Y_i \geq D, \\ D & \text{if } Y_i < D, \end{cases} \qquad R_i = \begin{cases} 1 & \text{if } Y_i \geq D, \\ 0 & \text{if } Y_i < D. \end{cases}$$

Left censored data commonly arise when measurement instruments are inaccurate below a lower limit of detection and, as such, this limit is then reported.

(a) (**6 marks**) Show that the log likelihood of the observed data $\{(x_i, r_i)\}_{i=1}^n$ is given by

$$\log L(\mu, \sigma^2 \mid \mathbf{x}, \mathbf{r}) = \sum_{i=1}^{n} \left\{ r_i \log \phi(x_i; \mu, \sigma^2) + (1 - r_i) \log \Phi(x_i; \mu, \sigma^2) \right\},$$

where $\phi(\cdot; \mu, \sigma^2)$ and $\Phi(\cdot; \mu, \sigma^2)$ stands, respectively, for the density function and cumulative distribution function of the normal distribution with mean μ and variance σ^2.

(b) (**6 marks**) Determine the maximum likelihood estimate of μ based on the data available in the file `dataex2.Rdata`. Consider σ^2 known and equal to 1.5^2. **Note**: You can use a built in function such as `optim` or the `maxLik` package in your implementation.

3. Consider a bivariate normal sample (Y_1, Y_2) with parameters $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_{12}, \sigma_2^2)$. The variable Y_1 is fully observed, while some values of Y_2 are missing. Let R be the missingness indicator, taking the value 1 for observed values and 0 for missing values. For the following missing data mechanisms state, justifying, whether they are ignorable for likelihood-based estimation.

(a) (**5 marks**) $\text{logit}\{\Pr(R = 0 \mid y_1, y_2, \theta, \psi)\} = \psi_0 + \psi_1 y_1$, $\psi = (\psi_0, \psi_1)$ distinct from θ.

(b) (**5 marks**) $\text{logit}\{\Pr(R = 0 \mid y_1, y_2, \theta, \psi)\} = \psi_0 + \psi_1 y_2$, $\psi = (\psi_0, \psi_1)$ distinct from θ.

(c) (**5 marks**) $\text{logit}\{\Pr(R = 0 \mid y_1, y_2, \theta, \psi)\} = 0.5(\mu_1 + \psi y_1)$, scalar ψ distinct from θ.

4. (**25 marks**) Suppose that

$$Y_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\{p_i(\boldsymbol{\beta})\},$$
$$p_i(\boldsymbol{\beta}) = \frac{\exp(\beta_0 + x_i \beta_1)}{1 + \exp(\beta_0 + x_i \beta_1)},$$

for $i = 1, \ldots, n$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Although the covariate x is fully observed, the response variable Y has missing values. Assuming ignorability, derive and implement an EM algorithm to compute the maximum likelihood estimate of $\boldsymbol{\beta}$ based on the data available in the file `dataex4.Rdata`. **Note**: 1) For simplicity, and without loss of generality because we have a univariate pattern of missingness, when writing down your expressions, you can assume that the first m values of Y are observed and the remaining $n - m$ are missing. 2) You can use a built in function such as `optim` or the `maxLik` package for the M-step.

5. Consider a random sample $Y_1, \ldots, Y_n$ from the mixture distribution with density

$$f(y) = p f_{\text{LogNormal}}(y; \mu, \sigma^2) + (1-p) f_{\text{Exp}}(y; \lambda),$$

with

$$f_{\text{LogNormal}}(y; \mu, \sigma^2) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(\log y - \mu)^2\right\}, \qquad y > 0, \quad \mu \in \mathbb{R}, \quad \sigma > 0,$$
$$f_{\text{Exp}}(y; \lambda) = \lambda \exp\{-\lambda y\}, \qquad y \geq 0, \quad \lambda > 0,$$

and $\theta = (p, \mu, \sigma^2, \lambda)$.

(a) (**13 marks**) Derive the EM algorithm to find the updating equations for $\theta^{(t+1)} = (p^{(t+1)}, \mu^{(t+1)}, (\sigma^{(t+1)})^2, \lambda^{(t+1)})$.

(b) (**12 marks**) Using the dataset `dataex5.Rdata` implement the algorithm and find the maximum likelihood estimates for each component of θ. As starting values, you might want to consider $\theta^{(0)} = (p^{(0)}, \mu^{(0)}, (\sigma^{(0)})^2, \lambda^{(0)}) = (0.1, 1, 0.5^2, 2)$. Draw the histogram of the data with the estimated density superimposed.