UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS
INCOMPLETE DATA ANALYSIS

Assignment 2

- To be uploaded to Learn by 4pm, March 23, 2023.
- Location for submission: Gradescope over Learn. **Important**: When uploading your report to Gradescope please tag separately each subquestion (e.g. 1a), 1b), 1c), etc).
- 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 X and Y are independent, Pareto-distributed, with cumulative distributions given by

$$F_X(x;\lambda) = 1 - \frac{1}{x^\lambda}, \quad F_Y(y;\lambda) = 1 - \frac{1}{y^\mu},$$

with $x, y \geq 1$ and $\lambda, \mu > 0$. Let $Z = \min\{X, Y\}$ and define the (non)censoring indicator

$$\delta = \begin{cases} 1 & \text{if } X < Y, \\ 0 & \text{otherwise.} \end{cases}$$

(This type of censoring is often known as "type I censoring.")

(a) (**10 marks**) Obtain the density function of Z (f_Z) and the frequency function of δ (f_δ). What are the distributions of Z and δ?

(b) (**5 marks**) Let $Z_1, \ldots, Z_n$ be a random sample from $f_Z(z;\theta)$, with $\theta = \lambda + \mu$, and let $\delta_1, \ldots, \delta_n$ be a random sample from $f_\delta(d;p)$, with $p = \lambda/(\lambda+\mu)$. Derive the maximum liklihood estimators of θ and p.

(c) (**8 marks**) Appealing to the asymptotic normality of the maximum likelihood estimator, provide a 95% confidence interval for θ and for p.

2. Suppose that $Y_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, for $i = 1, \dots, n$. 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 $i = 1, \dots, n$ 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 cumulative distribution function

$$F(y) = pF_X(y;\lambda) + (1-p)F_Y(y;\mu),$$

where $F_X(x;\lambda) = 1 - x^{-\lambda}$, $F_Y(y;\mu) = 1 - y^{-\mu}$, with $x, y \geq 1$ and $\lambda, \mu > 0$.

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

(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, consider $\theta^{(0)} = (p^{(0)}, \lambda^{(0)}, \mu^{(0)},) = (0.3, 0.3, 0.4)$ and as stopping criterion use

$$\left|\beta_0^{(t+1)} - \beta_0^{(t)}\right| + \left|\beta_1^{(t+1)} - \beta_1^{(t)}\right| < 0.0001.$$

Draw the histogram of the data with the estimated density superimposed. Hint: Use the Freedman–Diaconis rule for selecting the number of breaks in the histogram.