UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS
INCOMPLETE DATA ANALYSIS

Assignment

- To be uploaded to Learn by 4pm, March 21, 2025.
- 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 20% 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 $X_1, \ldots, X_n$ are independent and distributed according to a Pareto distribution with cumulative distribution function given by

$$F(x;\theta) = 1 - \frac{1}{x^\theta}, \qquad x > 1, \qquad \theta > 0.$$

The probability density and survival functions are

$$f(x;\theta) = \frac{\theta}{x^{\theta+1}}, \quad \text{and } S(x;\theta) = \frac{1}{x^\theta}$$

where $S(x) = Pr(X > x)$.

(a) **(11 marks)** Suppose that observations are censored if $X_i > C$, for some fixed and known $C > 1$, and let

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

be the observed variables. Write down the observed data likelihood and derive from it the maximum likelihood estimator.

(b) **(6 marks)** Suppose you have an approximate solution $\theta^{(0)}$. Give an expression for the Newton-Raphson iteration to obtain an improved estimate $\theta^{(1)}$.

(c) **(6 marks)** From the missing data mechanism perspective, are data censored this way MCAR, MAR, or MNAR? Justify.

2. (**25 marks**) Each of the 100 datasets contained in the file `dataex2.Rdata` was generated in the following way

$$y_i \mid x_i \overset{\text{ind.}}{\sim} \text{N}(\beta_0 + \beta_1 x_i, 1), \quad x_i \overset{\text{iid}}{\sim} \text{Unif}(-1, 1), \quad \beta_0 = 1, \quad \beta_1 = 3,$$

for $i = 1, \ldots, 100$. Additionally, some of the responses were set to be missing using a MAR mechanism. The goal of this exercise is to study the effect that acknowledging/not acknowledging parameter uncertainty when performing step 1 of multiple imputation might have on the coverage of the corresponding confidence intervals. Further suppose that the analysis of interest in step 2 is to fit the regression model that was used to generate the data, i.e., a normal linear regression model where the response is y and the covariate is x. With the aid of the `mice` package, calculate the empirical coverage probability of the 95% confidence intervals for β_1 under the following two approaches: stochastic regression imputation and the corresponding bootstrap based version. Comment. For both approaches, please consider $M = 20$ and `seed=1`. **NOTE 1**: In order to calculate the empirical coverage probability, you only need to compute the proportion of the time (over the 100 intervals) that the interval contains the true value of the parameter. **NOTE 2**: The data are stored in an array structure such that, for instance, `dataex2[, , 1]`, corresponds to the first dataset (which has 100 rows and 2 columns, with the first column containing the values of x and the second the values of y).

3. Suppose that $Y_i \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$, 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 `dataex3.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.

4. 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.

5. (**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 `dataex5.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.