IDA Assignment 2

Johnny Lee, s1687781

Q1.

Suppose $Y_1, \cdots, Y_n$ 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

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

a)

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

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

Answer :

We first define the Survival function (from **Workshop 3**)as

$$S(C;\theta) = \mathbb{P}(Y_i > C;\theta) = 1 - F(y_i;\theta)$$

which also represents the censored observations. For the uncensored observation, we have

$$f(y_i;\theta) = \frac{d}{dy_i}F(y_i;\theta) = \frac{ye^{-y^2/2\theta}}{\theta}$$

Given that $Y_1, \ldots, Y_n$ are independent and identically distributed, we have the likelihood function as,

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n \left([f(y_i;\theta)]^{r_i} [S(C;\theta)]^{1-r_i} \right) \\ &= \prod_{i=1}^n \left([\frac{ye^{-y^2/2\theta}}{\theta}]^{r_i} [e^{-C^2/\theta}]^{1-r_i} \right) \\ &= \left(\frac{y_i}{2\theta} \right)^{\sum_{i=1}^n r_i} \exp\left(\frac{\sum_{i=1}^n (r_i y_i^2 + (1-r_i)C^2)}{2\theta} \right) \end{aligned} \tag{1}$$

Now we can rewrite the term in the exponential as X_i can we expressed as

$$x_i = r_i y_i + C(1 - r_i)$$

Then by taking square on both sides we have,

$$x_i^2 = r_i^2 y_i^2 + (1-r_i)^2 C^2 + 2r_i y_i C(1 - r_i)$$

Noting that R_i is binary, we can then conclude with the expression as

$$x_i^2 = r_i y_i^2 + (1 - r_i)C^2 \tag{2}$$

Now we substitute (2) into (1) to have,

$$\begin{aligned} L(\theta) &= \left(\frac{y_i}{2\theta}\right)^{\sum_{i=1}^n r_i} \exp\left(\frac{\sum_{i=1}^n x_i}{2\theta}\right) \\ \implies \log(L(\theta)) &= -\log\theta \sum_{i=1}^n r_i - \frac{\sum_{i=1}^n x_i^2}{2\theta} \\ \implies \frac{d}{d\theta}\log L(\theta) &= \frac{1}{\theta}\sum_{i=1}^n r_i + \frac{1}{2\theta^2}\sum_{i=1}^n x_i^2 \end{aligned} \tag{3}$$

By equating the derivative to 0, we can obtain the maximum likelihood estimate of θ as below.

$$\hat{\theta}_{MLE} = \frac{\sum_{i=1}^n x_i^2}{2\sum_{i=1}^n r_i}$$

b)

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

Answer :

From (3), we take another derivative of it and thus obtain as below

$$\frac{d^2}{d\theta^2}\log L(\theta) = \frac{1}{\theta^2}\sum_{i=1}^n r_i - \frac{x_i^2}{\theta^3}$$

Then, the Fisher Information for the observed data likelihood is,

$$\begin{aligned} I(\theta) &= -\mathbb{E}\left(\frac{\sum_{i=1}^n r_i}{\theta^2} - \frac{x_i^2}{\theta^3}\right) \\ &= \frac{n\mathbb{E}(R)}{\theta^2} + \frac{n\mathbb{E}(X^2)}{\theta^3} \\ &= \frac{n\mathbb{E}(R)}{\theta^2} + \frac{1}{\theta^3}\left(n\mathbb{E}(RY^2) + nC^2\mathbb{E}(1-R)\right) \end{aligned} \tag{4}$$

$$$$

c)

Appealing to the asymptotic normality of the maximum likelihood estimator, provide a 95% confidence interval for θ.

Answer :

Q2.

Suppose that a dataset consists of 100 subjects and 10 variables. Each variable contains 10% of missing values. What is the largest possible subsample under a complete case analysis? What is the smallest? Justify.

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

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

a)

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

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

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.

Answer :

b)

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.

Answer :

Q3.

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)

$\text{logit}\{\mathbb{P}(R = 0|y_1, y_2, \theta, \psi)\} = \psi_0 + \psi_1 y_1, \psi = (\psi_0, \psi_1)$ distinct from θ.

Answer :

b)

$\text{logit}\{\mathbb{P}(R = 0|y_1, y_2, \theta, \psi)\} = \psi_0 + \psi_1 y_2, \psi = (\psi_0, \psi_1)$ distinct from θ.

Answer :

c)

$\text{logit}\{\mathbb{P}(R = 0|y_1, y_2, \theta, \psi)\} = 0.5(\mu_1 + \psi y_1)$, scalar ψ distinct from θ.

Answer :

Q4.

$$Y_i \overset{\text{ind.}}{\sim} \text{Bernoulli}(p_i(\boldsymbol{\beta})$$

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

for $i = 1, \cdots, n$ 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 the EM algorithm to compute the MLE of $\boldsymbol{\beta}$ based on the data available in `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

Answer :

Q5

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

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

with

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

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

Answer :

b)

Using the dataset `datasetex5.Rdata` implement the EM algorithm and find the MLEs 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. ### **Answer** :