# Incomplete Data Analysis Hand-in

Ken Deng, S2617343

March 21, 2025

The R code is available at `https://github.com/kdeng-gzcn/incomplete_data_analysis`.

# 1 Q1

## 1.1 (a)

The likelihood function for the observed data with censoring can be written as follows,

$$L(\theta) = \prod_{i=1}^{n} \left[ f(Y_i \mid \theta)^{\Delta_i} S(C \mid \theta)^{1-\Delta_i} \right],$$

where $f(Y_i \mid \theta)$ is the pdf of the distribution, $S(C \mid \theta)$ is the survival function at the $C$, and $\Delta_i$ is an indicator variable that equals 1 if the observation is uncensored.

Since the density function is given as

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

the survival function is

$$S(x; \theta) = P(X > x) = \frac{1}{x^{\theta}}.$$

Given the censoring at $C$, the observed data $Y_i$ are either equal to $X_i$ if $X_i \leq C$ or equal to $C$ if $X_i > C$. The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \left[ \left( \frac{\theta}{Y_i^{\theta+1}} \right)^{\Delta_i} \left( \frac{1}{C^{\theta}} \right)^{1-\Delta_i} \right].$$

Hence, we can compute the log-likelihood function, which is

$$\ell(\theta) = \sum_{i=1}^{n} \left[ \Delta_i \log \theta - (\theta + 1)\Delta_i \log Y_i - (1 - \Delta_i)\theta \log C \right].$$

To find the MLE, we take the derivative of the log-likelihood function with respect to $\theta$, which is

$$\frac{d}{d\theta}\ell(\theta) = \sum_{i=1}^{n} \left[ \frac{\Delta_i}{\theta} - \Delta_i \log Y_i - (1 - \Delta_i) \log C \right].$$

Setting the derivative equal to 0 and solving for $\theta$, we get the MLE, which is

$$\hat{\theta} = \frac{\sum_{i=1}^{n} \Delta_i}{\sum_{i=1}^{n} (\Delta_i \log Y_i + (1 - \Delta_i) \log C)}.$$

## 1.2 (b)

Suppose there is a initial solution $\theta^{(0)}$, with the Newton-Raphson iteration, the expression for the improved $\theta^{(1)}$ is given by

$$\theta^{(1)} = \theta^{(0)} - \frac{\ell'(\theta^{(0)})}{\ell''(\theta^{(0)})},$$

where we want the improved solution to be more suitable for the MLE, i.e., $\theta$ s.t. $\ell'(\theta) = 0$.

The first derivative of the log-likelihood function is

$$\ell'(\theta) = \sum_{i=1}^{n} \left[ \frac{\Delta_i}{\theta} - \Delta_i \log Y_i - (1 - \Delta_i) \log C \right],$$

and the second derivative is

$$\ell''(\theta) = -\sum_{i=1}^{n} \frac{\Delta_i}{\theta^2}.$$

### 1.3 (c)

The censoring way is **MNAR** (Missing Not At Random). This is because the data are censored depending on the values of $X_i$. If $X_i > C$, the data are censored and the observed value is reported as $C$, which depends on the value of the unobserved $X_i$. Therefore, the missingness is related to the unobserved data, making the mechanism MNAR.

# 2 Q2

The problem is calculating the empirical coverage probability of the 95% confidence intervals for the parameter $\beta_1$ under two imputation approaches: *stochastic regression imputation* and *bootstrap-based imputation*.

The data is generated based on the model:

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

With `mice`, for each of the 100 datasets, a normal linear regression model is fitted with the observed and imputed values of $x_i$ and $y_i$, the formula is

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where $\epsilon$ is the error term. The objective is to estimate $\beta_1$ and its 95% confidence interval.

For each imputed dataset, we compute the 95% confidence interval for $\beta_1$. The formula for empirical coverage probability is:

$$\text{Empirical Coverage Probability} = \frac{\sum_{i=1}^{100} \mathbf{1}_{[\beta_1 \in CI(\hat{\beta}_1^{(i)})]}}{100}$$

where $\beta_1$ is the true value, and $CI(\hat{\beta}_1^{(i)})$ is the confidence interval for $\beta_1$ for the $i$-th imputed dataset.

The results are as Table 1.

Table 1: Empirical Coverage Probability of 95% Confidence Intervals for $\beta_1$

| Imputation Method | Prob. |
|---|---|
| Stochastic Regression | 0.94 |
| Bootstrap-based | 0.95 |

# 3 Q3

The observations are represented by $Y_i$, where $i = 1, \ldots, n$, and the corresponding censored data is denoted as $X_i$, where:

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

and $R_i$ is an indicator variable defined as:

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

## 3.1 (a)

If the observation $X_i = Y_i \geq D$, the likelihood contribution for this case is the density function of the normal distribution evaluated at $x_i$, which is

$$f(x_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right),$$

where the log-likelihood is $\log f(x_i \mid \mu, \sigma^2) \equiv \log \phi(x_i \mid \mu, \sigma^2)$.

If the observation $X_i = D$ is censored, so we cannot directly observe $Y_i$ but know that it is less than $D$. The likelihood is the probability that $Y_i$ is less than $D$, which is given by the cdf of the normal distribution, which is

$$\Pr(Y_i < D) = \Phi(x_i \mid \mu, \sigma^2),$$

where the log-likelihood is $\log \Phi(x_i \mid \mu, \sigma^2)$.

The log-likelihood is the sum of the log-likelihood for each observation, with $r_i$ indicating whether the data is observed or censored, which is

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

This is the desired log-likelihood function for the observed data, accounting for both censored and uncensored observations.

### 3.2 (b)

Given that $\sigma^2 = 1.5^2$ is known, we can use the R code to compute the MLE of $\mu$ with `optimize`, where the result is $\hat{\mu} = 5.533$.

# 4 Question 4

## 4.1 (a)

We are given:

$$\operatorname{logit}\left(\Pr(R=0 \mid y_1, y_2, \theta, \psi)\right) = \psi_0 + \psi_1 y_1, \quad \psi = (\psi_0, \psi_1) \text{ distinct from } \theta.$$

The logistic model for the missingness mechanism is:

$$\log\left(\frac{\Pr(R=0)}{\Pr(R=1)}\right) = \psi_0 + \psi_1 y_1.$$

The missingness probability depends on $Y_1$ and not directly on $Y_2$, which is MAR. Also, we note that $\psi$ is distinct from $\theta$. Hence, this mechanism is **ignorable**.

## 4.2 (b)

We are given:

$$\operatorname{logit}\left(\Pr(R = 0 \mid y_1, y_2, \theta, \psi)\right) = \psi_0 + \psi_1 y_2, \quad \psi = (\psi_0, \psi_1) \text{ distinct from } \theta.$$

The logistic model for the missingness mechanism is:

$$\log\left(\frac{\Pr(R = 0)}{\Pr(R = 1)}\right) = \psi_0 + \psi_1 y_2.$$

In this case, the missingness depends on the unobserved value of $Y_2$, an MNAR case, which makes this mechanism **non-ignorable**.

### 4.3 (c)

We are given:

$$\operatorname{logit}\left(\Pr(R = 0 \mid y_1, y_2, \theta, \psi)\right) = 0.5(\mu_1 + \psi y_1), \quad \psi \text{ distinct from } \theta.$$

The logistic model for the missingness mechanism is:

$$\log\left(\frac{\Pr(R = 0)}{\Pr(R = 1)}\right) = 0.5(\mu_1 + \psi y_1).$$

Although the missingness mechanism depends on $Y_1$ and not directly on $Y_2$, it involves $\mu_1$. Hence, this mechanism is **non-ignorable**.

# 5 Q5

We are given a logistic regression model

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

where $Y_i$ is the binary response variable and $x_i$ is the covariate. The goal is to estimate the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1)$ with missing data.

**E-step** In the E-step, we estimate the missing values of $Y$ based on the current estimates of $\boldsymbol{\beta}^{(t)}$, the parameter values at the $t$-th iteration. The expectation of $Y_i$ given $x_i$ and the current parameter estimates is:

$$\mathbb{E}[Y_i \mid x_i, \boldsymbol{\beta}^{(t)}] = p_i(\boldsymbol{\beta}^{(t)}) = \frac{\exp(\beta_0^{(t)} + x_i\beta_1^{(t)})}{1 + \exp(\beta_0^{(t)} + x_i\beta_1^{(t)})}.$$

Thus, in the E-step, we replace the missing values of $Y_i$ with their expected values, $p_i(\boldsymbol{\beta}^{(t)})$.

The objective function for the EM algorithm is the log-likelihood function, which is

$$\ell(\boldsymbol{\beta} \mid \mathbf{x}, \mathbf{y}, \boldsymbol{\beta}^{(t)}) = \left( \sum_{i \in \text{observed}} + \sum_{i \in \text{missing}} \right) \left[ y_i \log p_i(\boldsymbol{\beta}) + (1 - y_i) \log(1 - p_i(\boldsymbol{\beta})) \right],$$

where $y_i = y_i(\boldsymbol{\beta}^{(t)})$ when $i$ is missing.

**M-step** The M-step involves maximizing this log-likelihood with respect to $\boldsymbol{\beta}$, where we have

$$\boldsymbol{\beta}^{(t+1)} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta} \mid \mathbf{x}, \mathbf{y}, \boldsymbol{\beta}^{(t)}).$$

This maximization is done using optimization methods `optim` function in R. The results are as Table 2.

Table 2: Estimated Parameters

| Parameter | Value |
|---|---|
| $\beta_0$ | 0.976 |
| $\beta_1$ | -2.480 |