Incomplete Data Analysis: Assignment 2

s2445245

March 20, 2023

Question 1

(1a)

The cumulative distributions of two independent variables X and Y are each given as;

$$F_X(x;\lambda) = 1 - \frac{1}{x^\lambda}$$

and

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

with $x, y \geq 1$ and $\lambda, \mu > 0$.

Let $Z = min\{X, Y\}$, the density function of Z (f_z) can be expressed as follows;

First, obtain the cumulative distribution function (CDF) of Z,

$$Z = min\{1 - \frac{1}{x^\lambda}, 1 - \frac{1}{y^\mu}\}$$

$$P(Z) = P(min\{x, y\} \leq Z)$$

$$P(Z) = P(x \leq Z) + P(y \leq Z) - P(x \leq Z) \cdot P(y \leq Z)$$

It can be done so as X and Y are assumed to be independent.

Substituting the definition of $F_X(x;\lambda)$ and $F_Y(y;\lambda)$, this CDF can be rewritten and simplified as;

$$\begin{aligned} P(Z) &= (1 - \frac{1}{z^\lambda}) + (1 - \frac{1}{z^\mu}) - ((1 - \frac{1}{z^\lambda}) \cdot (1 - \frac{1}{z^\mu})) \\ &= (1 - \frac{1}{z^\lambda}) + (1 - \frac{1}{z^\mu}) - (1 - \frac{1}{z^\lambda} - \frac{1}{z^\mu} + \frac{1}{z^{\mu+\lambda}}) \\ &= 1 - \frac{1}{z^{\mu+\lambda}} \end{aligned}$$

Then, to obtain the density function (PDF), take the first derivative of the above CDF;

$$f_Z(z) \equiv \frac{\mathrm{d}}{\mathrm{d}z} P(Z) = (\mu + \lambda) \cdot z^{-(\mu+\lambda+1)}$$

From the above expression, it can be seen that Z follows a Pareto distribution. The scale parameter is 1 and the shape parameter is $\lambda + \mu$.

The censoring indicator, δ, is defined as;

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

The frequency function of δ (f_δ) can be expressed as follows; First, define the PDF of X and Y as;

$$f_X(x) = \lambda \cdot x^{-(\lambda+1)}$$

$$f_Y(y) = \mu \cdot y^{-(\mu+1)}$$

They are both obtained by taking the first derivatives of $F_X(x)$ and $F_Y(y)$, their CDFs.

Then, from the definition of δ, it can be written that $P(\delta = 1) \equiv P(X < Y)$, which can be obtained by taking the integral of the products of the two PDFs defined above.

$$P(X < Y) = \int_1^\infty \int_1^y f_X(x) f_Y(y) \mathrm{d}x \mathrm{d}y$$

$$\begin{aligned} P(X < Y) &= \int_1^\infty \int_1^y \lambda \cdot x^{-(\lambda+1)} \cdot \mu \cdot y^{-(\mu+1)} \mathrm{d}x \mathrm{d}y \\ &= \int_1^\infty \mu \cdot y^{-(\mu+1)} \cdot (-y^{-\lambda} + 1) \mathrm{d}y \\ &= \frac{\mu}{-\lambda - \mu} + 1 \\ &= \frac{\lambda}{\lambda + \mu} \end{aligned}$$

Hence,

$$P(\delta = 1) = \frac{\lambda}{\lambda + \mu}$$

and,

$$\begin{aligned} P(\delta = 0) &= 1 - P(\delta = 1) \\ &= \frac{\mu}{\lambda + \mu} \end{aligned}$$

From the above expression, it can be seen that δ follows a Bernoulli distribution.

(1b)

Let $Z_1, ..., Z_n$ be random samples from $f_Z(z;\theta)$, with $\theta = \lambda + \mu$. Let $\delta_1, ..., \delta_n$ be random samples from $f_\delta(d;p)$, with $p = \frac{\lambda}{\lambda+\mu}$

The maximum likelihood estimator (MLE) of θ can be expressed as follows;

Firstly, given $\theta = \lambda + \mu$, $L(\theta)$ can be written as;

$$L(\theta) = \prod_{i=1}^{n} \theta z_i^{-(\theta+1)}$$

Taking its log,

$$log(L(\theta)) = \sum_{i=1}^{n} log(\theta) - \sum_{i=1}^{n} (\theta + 1) \cdot log(z_i)$$

Then, take the first derivative and equate to zero to obtain the MLE,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\theta} log(L(\theta)) &= \sum_{i=1}^{n} \frac{1}{\theta} - \sum_{i=1}^{n} log(z_i) \\ &= \frac{n}{\theta} - \sum_{i=1}^{n} log(z_i) \end{aligned}$$

Solving for 0,

$$\begin{aligned} 0 &= \frac{n}{\theta} - \sum_{i=1}^{n} log(z_i) \\ \frac{n}{\theta} &= \sum_{i=1}^{n} log(z_i) \end{aligned}$$

Hence, the MLE of θ, $\hat{\theta}$,is given as;

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} log(z_i)}$$

Next, the MLE of p can be expressed as follows;

Firstly, given $p = \frac{\lambda}{\lambda+\mu}$, $L(p)$ can be expressed as;

$$L(p) = \prod_{i=1}^{n} p^{\delta_i} \cdot (1-p)^{1-\delta_i}$$

Then, taking its log,

$$log(L(p)) = \sum_{i=1}^{n} \delta_i log(p) + \sum_{i=1}^{n} (1 - \delta_i) \cdot log(1-p)$$

Taking the first derivative and equate it to zero to obtain the MLE of p, $\hat{p}$,

$$\frac{\mathrm{d}}{\mathrm{d}p} log(L(p)) = \frac{\sum_{i=1}^{n} \delta_i}{p} - \frac{\sum_{i=1}^{n} (1-\delta_i)}{1-p} = 0$$

Solving the equation,

$$\frac{\sum_{i=1}^{n} \delta_i}{p} = \frac{\sum_{i=1}^{n} (1-\delta_i)}{1-p}$$

$$\sum_{i=1}^{n} \delta_i (1-p) = \sum_{i=1}^{n} (1-\delta_i) p$$

$$\sum_{i=1}^{n} \delta_i - \sum_{i-1}^{n} \delta_i p = \sum_{i=1}^{n} p - \sum_{i-1}^{n} \delta_i p$$

$$\sum_{i=1}^{n} \delta_i = np$$

Therefore,

$$\hat{p} = \frac{\sum_{i=1}^{n} \delta_i}{n}$$

(1c)

Appealing to the asymptotic normality of the maximum likelihood estimator, the 95% confidence interval for θ and for p can be obtained as follows;

To obtain an expression for the variance of the MLE, solve for the Fisher Information and take its inverse.

The second derivative of θ can be written as,

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} log(L(\theta)) = -n\theta^{-2}$$

given that,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} log(L(\theta)) = \frac{n}{\theta} - \sum_{i=1}^{n} log(z_i)$$

Fisher information is obtained by taking the negative expectation of the second derivative,

$$I(\theta) = -E[-n\theta^{-2}] = \frac{n}{\theta^2}$$

Taking its inverse, the variance of $\hat{\theta}_{MLE}$ is,

$$I(\theta)^{-1} \equiv V(\hat{\theta}_{MLE}) = \frac{\theta^2}{n}$$

Therefore, the distribution of $\hat{\theta}_{MLE}$ can be written as,

$$\hat{\theta}_{MLE} \sim N(\theta, \frac{\theta^2}{n})$$

and the the 95% confidence interval for $\hat{\theta}_{MLE}$ as,

$$\theta \pm 1.96\sqrt{\frac{\theta^2}{n}}$$

Following the same steps, the 95% confidence interval for p can be obtained by;
Taking the second derivative of p for Fisher information,

$$\frac{\mathrm{d}^2}{\mathrm{d}p^2} log(L(p)) = \frac{-\sum_{i=1}^{n} \delta_i}{p^2} - \frac{\sum_{i=1}^{n}(1-\delta_i)}{(1-p)^2}$$

given that,

$$\frac{\mathrm{d}}{\mathrm{d}p} log(L(p)) = \frac{\sum_{i=1}^{n} \delta_i}{p} - \frac{\sum_{i=1}^{n}(1-\delta_i)}{1-p}$$

Fisher information is obtained by taking the negative expectation of the second derivative,

$$\begin{aligned} I(p) &= -E[\frac{-\sum_{i=1}^{n} \delta_i}{p^2} - \frac{\sum_{i=1}^{n}(1-\delta_i)}{(1-p)^2}] \\ &= -E[\frac{-\sum_{i=1}^{n} \delta_i}{p^2} - \frac{n-\sum_{i=1}^{n} \delta_i}{(1-p)^2}] \end{aligned}$$

Substituting an expression of

$$\sum_{i=1}^{n} \delta_i = np$$

obtained from Q1b, the Fisher information can be rewritten and simplified as,

$$\begin{aligned} I(p) &= -E[\frac{-np}{p^2} - \frac{n-np}{(1-p)^2}] \\ &= \frac{np}{p^2} + \frac{n-np}{(1-p)^2} \\ &= \frac{n}{p} + \frac{n}{1-p} \\ &= \frac{n}{p(1-p)} \end{aligned}$$

Taking its inverse, the variance of $\hat{p}_{MLE}$ is,

$$I(p)^{-1} \equiv V(\hat{p}_{MLE}) = \frac{p(1-p)}{n}$$

Therefore, the distribution of $\hat{p}_{MLE}$ can be written as,

$$\hat{p}_{MLE} \sim N(p, \frac{p(1-p)}{n})$$

and the the 95% confidence interval for $\hat{p}_{MLE}$ as,

$$p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

Question 2

(2a)

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

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

$$R_i = \begin{cases} 1, & \text{if} \quad Y_i \geq D, \\ D, & \text{otherwise} \end{cases}$$

The log-likelihood of the observed data $\{(x_i, r_i)\}_{i=1}^n$ can be derived as follows;

X_i can take values either Y_i or D, $X_i = Y_i$ given $Y_i \geq D$, and $X_i = D$ if $Y_i < D$. As such, the likelihood of a single observation x_i can be represented in a binomial form of;

$$p^{r_i} \cdot (1-p)^{1-r_i}$$

in which p represent the probability that the variable has $Y_i \geq D$, and $(1-p)$ other wise.

Assuming $Y_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, the '$p$' component of the above likelihood follows the PDF of a normal distribution,

$$\phi(x_i; \mu, \sigma^2)$$

as $x_i = y_i$ if $y_i \geq D$.

On the other hand, the '$(1-p)$' component, which represents the cases where $y_i < D$, can be written as the CDF of a normal distribution $\Phi(y_i; \mu, \sigma^2)$. It is because it gives the probability that the variable is less than a certain threshold, in this case, D. Replacing the 'y_i' in $\Phi(\cdot)$ as 'x_i' gives a CDF

$$\Phi(x_i; \mu, \sigma^2)$$

Combining the two together , the likelihood of a single observation x_i can be written as;

$$L(\mu, \sigma^2 | x_i, r_i) = \phi(x_i; \mu, \sigma^2)^{r_i} \cdot \Phi(x_i; \mu, \sigma^2)^{1-r_i}$$

For the likelihood of all n observed x_i, using the product rule, it can be written as;

$$L(\mu, \sigma^2 | \mathbf{x}, \mathbf{r}) = \prod_{i=1}^{n} \phi(x_i; \mu, \sigma^2)^{r_i} \cdot \Phi(x_i; \mu, \sigma^2)^{1-r_i}$$

The binomial component is omitted as the probability of being censored is independent for each x_i, and dependent on the values of y_i but not x_i itself.

Then, taking the log of the above likelihood, it can be written as;

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

(2b)

```r
# loading data for Q2,Q4, and Q5
load("dataex2.rdata")
load("dataex4.rdata")
load("dataex5.rdata")

# Determine the maximum likelihood estimate of mu based on the data available
# in the file dataex2.Rdata. Consider sigma2 known and equal to 1.5^2
mean.all <- mean(dataex2$X)
sigma2 <- 1.5

# Data extraction
X <- dataex2$X
R <- dataex2$R

# Define the negative log-likelihood function for the model
n.log.ll <- function(mu, sigma2, x, r) {
  log.ll <- 0
  for (i in 1:length(x)){
    # Update the log-likelihood value by summing over all data points
    log.ll <- log.ll - (r[i] * dnorm(x[i], mu, sqrt(sigma2)) +
                          (1 - r[i]) * pnorm(x[i], mu, sqrt(sigma2)))
  }
  log.ll
}

# Optimize the negative log-likelihood function to find the maximum likelihood
# estimate of mu
mle <- suppressWarnings(optim(mean.all, n.log.ll,
                              sigma2 = sigma2, x = X, r = R))

print(paste('The maximum likelihood estimate of mu based on the data is',
            mle$par))
```

```
## [1] "The maximum likelihood estimate of mu based on the data is 5.30569817835934"
```

Question 3

(3a)

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 missing indicator, taking the value 1 for observed values and 0 for missing values.

In the case for;

$$logit\{Pr(R = 0|y_1, y_2, \theta, \psi)\} = \psi_0 + \psi_1 y_1$$

$\psi = (\psi_0, \psi_1)$ distinct from θ, the missing data mechanism is ignorable for likelihood-based estimation.

It is because the missing data mechanism is said to follow the Missing at Random (MAR) mechanism. In the above equation, the missingness only depends on observed value y_1 and a set of distinct missingness parameters ψ, but not on missing values y_2 or model parameters θ.

Writing in mathematical form,

$$logit\{Pr(R = 0|y_1, y_2, \theta, \psi)\} = logit\{Pr(R = 0|y_1, \psi)\} = \psi_0 + \psi_1 y_1$$

which follows the mathematical definition of MAR,

$$f(r|\mathbf{y}, \varphi) = f(r|y_{obs}, \varphi), \quad \forall y_{miss}, \varphi$$

The missingness in the question satisfies the ignorability condition that the mechanism is MAR and the missingness mechanism parameter ψ is distinct from the data model parameters θ.

Therefore, the missing data mechanism in this case is MAR and ignorable for likelihood-based estimation.

(3b)

In the case for;

$$logit\{Pr(R = 0|y_1, y_2, \theta, \psi)\} = \psi_0 + \psi_1 y_2$$

$\psi = (\psi_0, \psi_1)$distinct from θ, the missing data mechanism is not ignorable for likelihood-based estimation.

It is because the missing data mechanism is said to follow the Missing Not at Random (MNAR) mechanism.

Contrary to Q3a, in the above equation, the missingness now depends on unobserved value y_2 and a set of distinct missingness parameters ψ. Writing in mathematical form,

$$logit\{Pr(R = 0|y_1, y_2, \theta, \psi)\} = logit\{Pr(R = 0|y_2, \psi)\} = \psi_0 + \psi_1 y_2$$

which follows the mathematical definition of MNAR,

$$f(r|\mathbf{y}, \varphi) = f(r|y_{obs}, y_{miss,}\varphi), \quad \forall \varphi$$

Although the missingness in the question satisfies one of the ignorability condition that the missingness mechanism parameter ψ is distinct from the data model parameters θ, it violates the other criterion that the missing mechanism is MNAR, but not MCAR or MAR.

Therefore, the missing data mechanism in this case is MNAR and not ignorable for likelihood-based estimation.

(3c)

In the case for;

$$logit\{Pr(R = 0|y_1, y_2, \theta, \psi)\} = 0.5 \cdot (\mu_1 + \psi y_1)$$

scalar ψdistinct from θ, the missing data mechanism is ignorable for likelihood-based estimation under certain assumptions.

It is because the missing data mechanism is said to follow the Missing at Random (MAR) mechanism.

In the above equation, the missingness depends on observed value y_1, data model parameter μ_1, and a set of distinct missingness parameters ψ, but not on missing values y_2.

Writing in mathematical form,

$$logit\{Pr(R = 0|y_1, y_2, \theta, \psi)\} = logit\{Pr(R = 0|y_1, \theta, \psi)\} = 0.5 \cdot (\mu_1 + \psi y_1)$$

which follows the mathematical definition of MAR,

$$f(r|\mathbf{y}, \varphi) = f(r|y_{obs}, \varphi), \quad \forall y_{miss}, \varphi$$

The missingness in the question satisfies one of the ignorability condition that the missingness mechanism parameter ψ is distinct from the data model parameters θ.

All of the derivation above to justify the mechanism to be MAR is conditioned on an assumption that μ_1 independent from unobserved y_2. If this assumption is violated, the missingness will become dependent on unobserved and violates the ignorability condition that the mechanism is not either one of MAR or MCAR. It is important to check whether μ_1 is independent from y_2, and if this is invalid, the missing mechanism follows MNAR and not ignorable for likelihood-based estimation. In this case, it is a valid assumption as μ_1 is held to be constant.

Therefore, under the assumption made here that μ_1 is independent from y_2, the missing data mechanism in this case is MAR and ignorable for likelihood-based estimation.

Question 4

Suppose that

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

for $i = 1, ..., n$ and $\beta = (\beta_0, \beta_1)'$.

The covariate x are fully observed, but the response variable Y has some missing values. Assuming ignorability, the EM algorithm to compute the maximum likelihood estimate of β based on the data available can be derived and implemented as follows;

Firstly, EM algorithm in this case is a valid approach as the missing data mechanism does not depends on the unobserved Y_i, hence, it is a MAR mechanism.

Given that Y_i follows a Bernoulli distribution (with iid), the likelihood of $p_i(\beta)$ can be expressed as,

$$
\begin{aligned}
L(p_i(\beta)) &= \prod_{i=1}^{n} p_i(\beta)^{y_i} \cdot (1-p_i(\beta))^{1-y_i} \\
&= \prod_{i=1}^{n} \left(\frac{exp(\beta_0+\beta_1 x_i)}{1+exp(\beta_0+\beta_1 x_i)} \right)^{y_i} \cdot \left(1 - \frac{exp(\beta_0+\beta_1 x_i)}{1+exp(\beta_0+\beta_1 x_i)} \right)^{1-y_i} \\
&= \prod_{i=1}^{n} \left(\frac{exp(\beta_0+\beta_1 x_i)}{1+exp(\beta_0+\beta_1 x_i)} \right)^{y_i} \cdot \left(\frac{1}{1+exp(\beta_0+\beta_1 x_i)} \right)^{1-y_i}
\end{aligned}
$$

Taking its log, the likelihood function of complete data can be rewritten as,

$$
\begin{aligned}
log(L(p_i(\beta))) &= \sum_{i=1}^{n} y_i \cdot log(p_i(\beta)) + (1-y_i) \cdot log(1-p_i(\beta)) \\
&= \sum_{i=1}^{n} y_i \cdot log\left(\frac{exp(\beta_0+\beta_1 x_i)}{1+exp(\beta_0+\beta_1 x_i)} \right) + (1-y_i) \cdot log\left(\frac{1}{1+exp(\beta_0+\beta_1 x_i)} \right) \\
&= \sum_{i=1}^{n} y_i \cdot (log(exp(\beta_0+\beta_1 x_i)) - log(1+exp(\beta_0+\beta_1 x_i))) + (1-y_i) \cdot (log(1) - log(1+exp(\beta_0+\beta_1 x_i))) \\
&= \sum_{i=1}^{n} y_i \cdot (\beta_0+\beta_1 x_i) - y_i \cdot log(1+exp(\beta_0+\beta_1 x_i)) - log(1+exp(\beta_0+\beta_1 x_i)) + y_i \cdot log(1+exp(\beta_0+\beta_1 x_i)) \\
&= \sum_{i=1}^{n} y_i \cdot (\beta_0+\beta_1 x_i) - log(1+exp(\beta_0+\beta_1 x_i))
\end{aligned}
$$

Assume that the first $i = 1, ..., m$ values of Y are observed and the remaining $i = m+1, ...n$ are missing.

Then, the log likelihood can be expressed as,

$$
log(L(p_i(\beta))) = \sum_{i=1}^{m} y_i \cdot (\beta_0+\beta_1 x_i) - log(1+exp(\beta_0+\beta_1 x_i)) + \sum_{i=m+1}^{n} y_i \cdot (\beta_0+\beta_1 x_i) - log(1+exp(\beta_0+\beta_1 x_i))
$$

At iteration $(t+1)$ of the algorithm, the E-step calculates the conditional expectation, with respect to the missing observations, of the complete data log-likelihood given the observed data and the estimate of β from iteration t, β^t,

$$
Q(\beta|\beta^t) = E_{Y_{miss}}[log(L(p_i(\beta)|\mathbf{y}_{obs}, \mathbf{y}_{miss}))|\mathbf{y}_{obs}, \beta^t]
$$

In this case, conditional expectation becomes,

$$
\begin{aligned}
Q(\beta|\beta^t) &= E_{Y_{miss}}[\sum_{i=1}^{m} y_i \cdot (\beta_0+\beta_1 x_i) - log(1+exp(\beta_0+\beta_1 x_i)) + \sum_{i=m+1}^{n} y_i \cdot (\beta_0+\beta_1 x_i) - log(1+exp(\beta_0+\beta_1 x_i))] \\
&= \sum_{i=1}^{m} y_i \cdot (\beta_0+\beta_1 x_i) - log(1+exp(\beta_0+\beta_1 x_i)) + \sum_{i=m+1}^{n} E_{Y_{miss}}[y_i \cdot (\beta_0+\beta_1 x_i)] - log(1+exp(\beta_0+\beta_1 x_i)) \\
&= \sum_{i=1}^{m} y_i \cdot (\beta_0+\beta_1 x_i) + \sum_{i=m+1}^{n} E_{Y_{miss}}[y_i \cdot (\beta_0+\beta_1 x_i)] - \sum_{i=1}^{n} log(1+exp(\beta_0+\beta_1 x_i))
\end{aligned}
$$

Given that for a random variable X which follows a Bernoulli distribution,

$$P(X = x) = p^x \cdot (1-p)^{1-x}$$

The expectation is,

$$E[X] = p$$

The above conditional expectation can be completed as,

$$\begin{aligned} Q(\beta|\beta^t) &= \sum_{i=1}^{m} y_i \cdot (\beta_0 + \beta_1 x_i) + \sum_{i=m+1}^{n} E_{Y_{miss}}[y_i] \cdot (\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} log(1 + exp(\beta_0 + \beta_1 x_i)) \\ &= \sum_{i=1}^{m} y_i \cdot (\beta_0 + \beta_1 x_i) + \sum_{i=m+1}^{n} \left(\frac{exp(\beta_0^t + \beta_1^t x_i)}{1 + exp(\beta_0^t + \beta_1^t x_i)} \right) \cdot (\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} log(1 + exp(\beta_0 + \beta_1 x_i)) \end{aligned}$$

Following the derivation of the E-step as above, the M-step can be computed as,

$$\beta^{t+1} = argmax_\beta Q(\beta|\beta^t)$$

The maximisation of the likelihood function is run with respect to β, and continued until some convergence is reached.

The following R codes computes the maximum likelihood estimate of β based on the data available

```r
x <- dataex4$X
y <- dataex4$Y

# Extract observed and missing data
y.obs <- y[!is.na(y)]
x.obs <- x[!is.na(y)]
x.miss <- x[is.na(y)]

# Define the negative log-likelihood function
nll.q4 <- function(beta, y, x) {
  p <- exp(beta[1] + x * beta[2]) / (1 + exp(beta[1] + x * beta[2]))
  -sum(y * log(p) + (1 - y) * log(1 - p))
}

# EM algorithm
# Initialize the parameter estimates and convergence settings
beta <- c(0, 0)
tolerance <- 1e-4
converged <- FALSE

while (converged==FALSE) {
  # E-step: estimate the missing response values based on the current beta
  # values
  miss <- exp(beta[1] + x.miss * beta[2]) /
    (1 + exp(beta[1] + x.miss * beta[2]))

  # M-step: update the beta values using the complete data
  # (observed + estimated missing values)
  y.comp <- c(y.obs, miss)
```

```
  x.comp <- c(x.obs, x.miss)

  # Optimize the negative log-likelihood function
  mle.beta <- optim(beta, function(beta) nll.q4(beta, y.comp, x.comp))
  beta.new <- mle.beta$par

  # Check convergence: if the change in beta values is less than
  # the tolerance, break
  if (sum(abs(beta.new - beta)) < tolerance) {
    converged <- TRUE
    beta <- beta.new
  } else {
    beta <- beta.new
  }
}

print(paste('The maximum likelihood estimate of beta0 based on the data is',
            beta[1]))
```

```
## [1] "The maximum likelihood estimate of beta0 based on the data is 0.975707884191108"
```

```
print(paste('The maximum likelihood estimate of beta1 based on the data is',
            beta[2]))
```

```
## [1] "The maximum likelihood estimate of beta1 based on the data is -2.47998737558292"
```

Question 5

(5a)

Consider a random sample $Y_1, ..., Y_n$ 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 $\lambda, \mu > 0$.

Let $\theta = (p, \lambda, \mu)$. The EM algorithm to find the updating equation for $\theta^{(t+1)} = (p^{(t+1)}, \lambda^{(t+1)}, \mu^{(t+1)})$ can be derived as follows;

Firstly, the densities of $F_X(x;\lambda)$ and $F_Y(y;\mu)$ can be written as,

$$f_X = \lambda y^{-(\lambda+1)}$$
$$f_Y = \mu y^{-(\mu+1)}$$

Then, the likelihood of $f(y)$ can be written as,

$$f(y) = p \cdot (\lambda y^{-(\lambda+1)}) + (1-p) \cdot (\mu y^{-(\mu+1)})$$

Let z_i be binary latent variables such that,

$$z_i = \begin{cases} 1, & \text{if} y_i \text{belongs to the first component,} \\ 0, & \text{if} y_i \text{belongs to the second component} \end{cases}$$

The observed data in this context is $\mathbf{y} = (y_1, ..., y_n)$ and the missing data is $\mathbf{z} = (z_1, ..., z_n)$.

The likelihood of the complete data (**y,z**) is,

$$L(\theta; \mathbf{y,z}) = \prod_{i=1}^{n} p \cdot (\lambda y^{-(\lambda+1)})^{z_i} + (1-p) \cdot (\mu y^{-(\mu+1)})^{1-z_i}$$

And taking its log, the log likelihood becomes,

$$log(L(\theta; \mathbf{y,z})) = \sum_{i=1}^{n} z_i \cdot (log(p) + log(\lambda y^{-(\lambda+1)})) + (1-z_i) \cdot (log(1-p) + log(\mu y^{-(\mu+1)}))$$

For the E-step, the expectation is formulated as,

$$\begin{aligned} Q(\theta|\theta^t) &= E_z[log(L(\theta; \mathbf{y,z}))|\mathbf{y}, \theta^t] \\ &= \sum_{i=1}^{n} E[Z_i|y_i, \theta^t] \cdot (log(p) + log(\lambda y^{-(\lambda+1)})) + (1 - E[Z_i|y_i, \theta^t]) \cdot (log(1-p) + log(\mu y^{-(\mu+1)})) \end{aligned}$$

Given that $E[Z_i|y_i, \theta^t] = Pr(Z_i = 1|y_i, \theta^t)$, applying Bayesian theorem and the law of total probability,

$$\begin{aligned} E[Z_i|y_i, \theta^t] &= Pr(Z_i = 1|y_i, \theta^t) \\ &= \frac{p^t \cdot (\lambda^t y^{-(\lambda^t+1)})}{p^t \cdot (\lambda^t y^{-(\lambda^t+1)}) + (1-p^t) \cdot (\mu^t y^{-(\mu^t+1)})} \\ &= \tilde{p}_i^t, \qquad\qquad i = 1, ..., n \end{aligned}$$

As such,

$$Q(\theta|\theta^t) = \sum_{i=1}^{n} \tilde{p}_i^t \cdot (log(p) + log(\lambda y^{-(\lambda+1)})) + (1-\tilde{p}_i^t) \cdot (log(1-p) + log(\mu y^{-(\mu+1)}))$$

For the M-step, compute the partial derivatives of $Q(\theta|\theta^t)$ with respect to all three components of $\theta = (p, \lambda, \mu)$.

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}p} Q(\theta|\theta^t) = 0 \Longrightarrow p^{t+1} = \frac{1}{n} \sum_{i=1}^{n} \tilde{p}_i^t, \\ &\frac{\mathrm{d}}{\mathrm{d}\lambda} Q(\theta|\theta^t) = 0 \Longrightarrow \lambda^{t+1} = argmax_\lambda \sum_{i=1}^{n} \tilde{p}_i^t \cdot log(\lambda y^{-(\lambda+1)}) \\ &\frac{\mathrm{d}}{\mathrm{d}\mu} Q(\theta|\theta^t) = 0 \Longrightarrow \mu^{t+1} = argmax_\mu \sum_{i=1}^{n} (1-\tilde{p}_i^t) \cdot (log(\mu y^{-(\mu+1)}) \end{aligned}$$

The latter two becomes,

$$\begin{aligned}
\lambda^{t+1} &= argmax_\lambda \sum_{i=1}^{n} \tilde{p}_i^t \cdot log(\lambda y^{-(\lambda+1)}) \\
&= argmax_\lambda \sum_{i=1}^{n} \tilde{p}_i^t \cdot (log(\lambda) - (\lambda+1) + log(y)) \\
\mu^{t+1} &= argmax_\mu \sum_{i=1}^{n} (1-\tilde{p}_i^t) \cdot log(\mu y^{-(\mu+1)}) \\
&= argmax_\mu \sum_{i=1}^{n} (1-\tilde{p}_i^t) \cdot (log(\mu) - (\mu+1) + log(y))
\end{aligned}$$

and continued until some convergence is reached.

(5b)

```
y <- dataex5

# Define the density functions for the mixture model components
fx <- function(y, lambda) {
  lambda * y^(-lambda - 1)
}

fy <- function(y, mu) {
  mu * y^(-mu - 1)
}

# EM algorithm
# Initialize the parameter estimates and convergence settings
theta <- c(0.3, 0.3, 0.4)
tolerance <- 1e-4
converged <- FALSE

while (converged == FALSE) {
  # E-step: Compute the posterior probability of each observation
  # belonging to component 1
  p <- (theta[1] * fx(y, theta[2])) /
    (theta[1] * fx(y, theta[2]) + (1 - theta[1]) * fy(y, theta[3]))

  # M-step: Update the parameter estimates
  theta.new <- theta
  # Update the proportion of component 1
  theta.new[1] <- mean(p)
  mle.lambda <- suppressWarnings(
    optim(theta[2],
          function(lambda) -sum(p * (log(lambda) - (lambda + 1) * log(y)))))
  # Update lambda using the MLE
  theta.new[2] <- mle.lambda$par
  mle.mu <- suppressWarnings(
    optim(theta[3],
```

```
          function(mu) -sum((1 - p) * (log(mu) - (mu + 1) * log(y)))))
  # Update mu using the MLE
  theta.new[3] <- mle.mu$par

  # Check convergence: If the change in parameter estimates is less
  # than the tolerance, break
  if (sum(abs(theta.new - theta)) < tolerance) {
    converged <- TRUE
    theta <- theta.new
  } else {
    theta <- theta.new
  }
}

print(paste('The maximum likelihood estimate of p based on the data is',
            theta[1]))
```

```
## [1] "The maximum likelihood estimate of p based on the data is 0.79285150529194"
```

```
print(paste('The maximum likelihood estimate of lambda based on the data is',
            theta[2]))
```

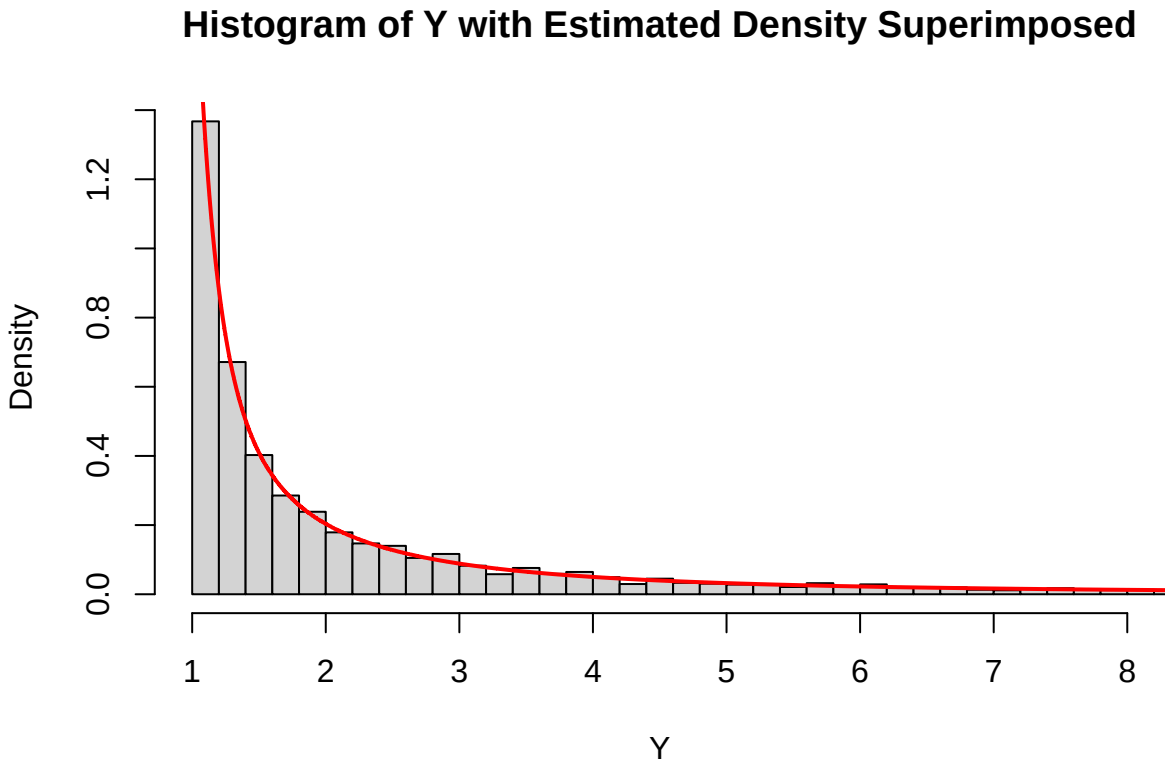
```
## [1] "The maximum likelihood estimate of lambda based on the data is 0.975180177109972"
```

```
print(paste('The maximum likelihood estimate of mu based on the data is',
            theta[3]))
```

```
## [1] "The maximum likelihood estimate of mu based on the data is 6.62975796848295"
```

```
# Calculate density
density <- theta[1] * fx(sort(y), theta[2]) +
  (1 - theta[1]) * fy(sort(y), theta[3])

# Plot histogram and density
hist(y, freq = FALSE, breaks = 'FD', border = "black",
     xlab = "Y", ylab = "Density", xlim = c(1,8),
     main = "Histogram of Y with Estimated Density Superimposed")
lines(sort(y), density, col = "red", lwd = 2)
```


Histogram of Y with Estimated Density Superimposed
Density
0.0
0.4
0.8
1.2
1
2
3
4
5
6
7
8
Y