Incomplete Data Analysis
Assignment 2

Callum Abbott

Question 1

Suppose $Y_1, ..., Y_n$ are independent and identically distributed with cumulative distribution function given by

$$F(y;\theta) = 1 - e^{-y^2/(2\theta)}, \quad y \geq 0,\ \theta > 0$$

Further suppose that observations are (right) censored if $Y_i > C$, for some known $C > 0$, and let

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

Question 1a

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

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

Solution:

- To derive the MLE we must maximize the log-likelihood of the observed data $\{x_i, r_i\}_{i=1}^n$. In this context, there are two contributions to the likelihood function:
    1. $f(y_i;\theta) = dF(y_i;\theta)/dy_i$ from *non-censored* observations.
    2. $Pr(Y_i > C;\theta) = S(C;\theta) = 1 - F(y_i;\theta)$ from *censored* observations.
- All observations $Y_i, ..., Y_n$ are iid, hence,

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

- Note that in order to understand how one goes from line 3 to line 4 in the equation defined above, we recall that we can write the variable X_i as $X_i = Y_i R_i + C(1 - R_i)$
and due to the binary nature of R_i:

$$\implies X_i^2 = Y_i^2 R_i^2 + C^2(1-R_i)^2 + 2Y_i R_i C(1-R_i)$$
$$\implies X_i^2 = Y_i^2 R_i + C^2(1-R_i)$$

- We can now define the log-likelihood to be

$$\log L(\theta) := l(\theta) = \sum_{i=1}^{n} r_i \log\left(\frac{y_i}{2\theta}\right) - \frac{1}{2\theta}\sum_{i=1}^{n} x_i^2$$

- Maximising this quantity through taking its derivative

$$\frac{\mathrm{d}}{\mathrm{d}\theta} l(\theta) = -\frac{\sum_{i=1}^{n} r_i}{\theta} + \frac{\sum_{i=1}^{n} x_i^2}{2\theta^2}$$

- leading to

$$\widehat{\theta}_{\text{MLE}} = \frac{\sum_{i=1}^{n} X_i^2}{2\sum_{i=1}^{n} R_i}.$$

- Note that we have assumed here that $\widehat{\theta}_{\text{MLE}}$ is indeed a maximum and have not computed the second derivative since our result matches the one given in the question.

Question 1b

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.

Solution:

- We first recall the general definition of the expected Fisher information to be

$$I(\theta) = -\mathbb{E}\left[\frac{d^2 l(\theta)}{d\theta^2}\right]$$

- We now compute the second derivative of the log-likelihood and re-introduce the variables r_i and y_i for x_i which will allow us to take expectations more clearly. This yields,

$$I(\theta) = -\frac{\sum_i \mathbb{E}[R_i]}{\theta^2} + \frac{\sum_i \mathbb{E}[R_i Y_i^2]}{\theta^3} + \frac{\sum_i C^2\mathbb{E}[(1-R_i)]}{\theta^3}$$

- Note that R is a binary random variable and so

$$\begin{aligned}\mathbb{E}[R] &= 1 \times \Pr(R=1) + 0 \times \Pr(R=0)\\ &= \Pr(R=1)\\ &= \Pr(Y \leq C)\\ &= F(C;\theta)\\ &= 1 - e^{-C^2/2\theta}.\end{aligned}$$

- And hence

$$\begin{aligned}I(\theta) &= -\frac{\sum_i \mathbb{E}[R_i]}{\theta^2} + \frac{\sum_i \mathbb{E}[R_i Y_i^2]}{\theta^3} + \frac{\sum_i C^2 \mathbb{E}[(1-R_i)]}{\theta^3}\\ &= -\frac{n}{\theta^2}(1 - e^{C^2/2\theta}) + \frac{n}{\theta^3}\left\{-C^2 e^{-C^2/2\theta} + 2\theta(1 - e^{-C^2/2\theta})\right\} + \frac{n}{\theta^3}e^{-C^2/2\theta}\\ &= \frac{n}{\theta^2}(1 - e^{-C^2/2\theta})\end{aligned}$$

Question 1c

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

Solution:

- We recall the asymptotic normality of the MLE as

$$\widehat{\theta}_{\text{MLE}} \sim N(\theta, I(\theta)^{-1})$$

- Therefore

$$\frac{\widehat{\theta}_{\text{MLE}} - \theta}{\sqrt{I(\theta)^{-1}}} \sim N(0,1)$$

- Using the properties of the standard Gaussian distribution ($\alpha = 0.05$)

$$Pr\left(z_{-\alpha/2} \leq \frac{\widehat{\theta}_{\text{MLE}} - \theta}{\sqrt{I(\theta)^{-1}}} \leq z_{\alpha/2}\right) = 1 - \alpha = 0.95$$

- The 95% CI for $\widehat{\theta}_{\text{MLE}}$ is hence $\left[\sqrt{I(\theta)^{-1}} z_{-\alpha/2} + \theta, \sqrt{I(\theta)^{-1}} z_{\alpha/2} + \theta\right]$ where $z_{\alpha/2} = 1.959964$, $z_{-\alpha/2} = -1.959964$, and $\sqrt{I(\theta)^{-1}} = \theta/\sqrt{n(1 - e^{-C^2/2\theta})}$

```r
alpha = 0.05

z = qnorm(1-alpha/2)
```

Question 2

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

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

Question 2a

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

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

where $\phi(x_i; \mu, \sigma^2)$ and $\Phi(x_i; \mu, \sigma^2)$ stands, respectively, for the density function and cumulative distribution function of the normal distribution with mean μ and variance σ^2.

Solution:

- Similar to 1a, our likelihood function has two contributions:
  1. $\phi(x_i; \mu, \sigma^2)$ from *non-censored* observations.
  2. $Pr(X_i < D; \mu, \sigma^2) = S(D; \mu, \sigma^2) = 1 - \Phi(x_i; \mu, \sigma^2)$ from *censored* observations.
- All observations $X_i, ..., X_n$ are iid, hence,

$$\begin{aligned} l(\mu, \sigma^2|\boldsymbol{x}, \boldsymbol{r}) &= \log \prod_{i=1}^{n} \left\{\phi(x_i; \mu, \sigma^2)\right]^{r_i} [1 - \Phi(x_i; \mu, \sigma^2)]^{1-r_i}\right\} \\ &= \log \left\{\phi(x_i; \mu, \sigma^2)\right]^{\sum_i r_i} [1 - \Phi(x_i; \mu, \sigma^2)]^{\sum_i (1-r_i)}\right\} \\ &= \sum_{i=1}^{n} \left\{r_i \log \phi(x_i; \mu, \sigma^2) + (1 - r_i) \log \Phi(x_i; \mu, \sigma^2)\right\} \end{aligned}$$

- Note that we have made use of the fact that $\log(1) = 0$.

Question 2b

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.

Solution:

- $\widehat{\mu}_{\text{MLE}} = 5.5328$ to 4 d.p.

```
library(maxLik)

# Loading in data

load('dataex2.Rdata')


# Log likelihood function set to maximized

get_log_likelihood = function(param, data) {

  mu = param

  x = data[,1]; r = data[,2]

  return(sum(r*dnorm(x, mean=mu, sd=1.5, log=TRUE) +

               (1 - r)*pnorm(x, mean=mu, sd=1.5, log.p=TRUE)))
```

```r
}
# Get MLE
mle = maxLik(logLik = get_log_likelihood, data = dataex2, start = c(mu=1))
# Present results
summary(mle)
```

```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 3 iterations
## Return code 2: successive function values within tolerance limit
## Log-Likelihood: -336.3821
## 1  free parameters
## Estimates:
##    Estimate Std. error t value Pr(> t)
## mu   5.5328     0.1075   51.48  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

Question 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.

Solution:

- A missing data mechanism (MDM) is said to be ignorable for likelihood based inference if and only if the following two criteria are met:

  1. The missing data are missing at random (MAR) or missing completely at random (MCAR).
  2. The parameter ψ (missingness mechanism) and θ (data model) are distinct in the sense that the joint parameter space of (ψ, θ) is the product of the parameter spaces Ψ and Θ (separability condition).

- The three missing data mechanisms presented below all meet criterion 2 hence we simply need to justify whether the data caused to be missing by each mechanism meets criterion 1.

(a) $logit\,\{Pr(R = 0|y_1, y_2, \theta, \psi)\} = \psi_0 + \psi_1 y_1; \quad \psi = (\psi_1, \psi_2)$ distinct from θ.

- We observe that the MDM is dependent on the fully observed variable, y_1, only. The missing data resulting from this mechanism will hence be MAR indicating MDM (a) is ignorable for likelihood-based estimation.

(b) $logit\,\{Pr(R = 0|y_1, y_2, \theta, \psi)\} = \psi_0 + \psi_1 y_2; \quad \psi = (\psi_1, \psi_2)$ distinct from θ.

- We observe that the MDM is dependent on the missing variable, y_2, only. The missing data resulting from this mechanism will hence be MNAR indicating MDM (b) is **NOT** ignorable for likelihood-based estimation.

(c) $logit\,\{Pr(R = 0|y_1, y_2, \theta, \psi)\} = 0.5(\mu_1 + \psi_1 y_1); \quad \psi$ (scalar) distinct from θ.

- We observe a similar MDM to (a) with an added dependency on μ_1. Whether the data are MAR or MNAR now depends on whether σ_{12} is equal to 0. In the case where σ_{12} is equal to 0, Y_1 and Y_2 would be independent variables and hence the missing data from the MDM would be MAR rendering the MDM ignorable for likelihood-based estimation. In the case where σ_{12} is **NOT** equal to 0, Y_1 and Y_2 would be dependent variables meaning μ_1 would have some Y_2 dependency rendering the data from the MDM MNAR. We would then be unable to rule out the MDM for likelihood-based estimation.

Question 4

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 $i = 1, ..., n$ 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`.

Solution:

- We begin by first obtaining the likelihood for $\boldsymbol{\beta}$ which is given by

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

The corresponding log likelihood is hence

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{e^{\beta_0 + x_i\beta_1}}{1 + e^{\beta_0 + x_i\beta_1}} \right) + (1 - y_i) \log\left(\frac{1}{1 + e^{\beta_0 + x_i\beta_1}} \right) \right\} \\ &= \sum_{i=1}^{n} \{y_i(\beta_0 + x_i\beta_1) - \log(1 + e^{\beta_0 + x_i\beta_1})\}. \end{aligned}$$

- Now that we are in possession of the log-likelihood, we can use this to conduct the expectation step of the EM algorithm and define our Q function.
- Note that the expectation is taken under the distribution of the **missing data**. We hence make use of our univariate pattern of missingness and assume that the first m values of Y are reserved and the remaining $n - m$ are missing i.e. $\boldsymbol{y_{obs}} = y_1, ..., y_m$ and $\boldsymbol{y_{mis}} = y_m + 1, ..., y_n$.

$$\begin{aligned} Q(\boldsymbol{\beta}|\boldsymbol{\beta^{(t)}}) &= \mathbb{E}_{\boldsymbol{y_{mis}}}\left[l(\boldsymbol{\beta})|\boldsymbol{y_{obs}}, \boldsymbol{x}, \boldsymbol{\beta^{(t)}} \right] \\ &= \sum_{i=1}^{m} \{y_i(\beta_0 + \beta_1 x_i)\} - \sum_{i=1}^{n} \log(1 + e^{\beta_0 + \beta_1 x_i}) + \sum_{i=m+1}^{n} (\beta_0 + \beta_1 x_i)\mathbb{E}_{\boldsymbol{y_{mis}}}[y_i|\boldsymbol{y_{obs}}, \boldsymbol{x}, \boldsymbol{\beta^{(t)}}] \\ &= \sum_{i=1}^{m} \{y_i(\beta_0 + \beta_1 x_i)\} - \sum_{i=1}^{n} \log(1 + e^{\beta_0 + \beta_1 x_i}) + \sum_{i=m+1}^{n} (\beta_0 + \beta_1 x_i)p_i(\boldsymbol{\beta}) \end{aligned}$$

- Where we have used the result that $\mathbb{E}[Y_i] = p_i(\boldsymbol{\beta})$ since $Y_i \overset{\text{ind}}{\sim} \text{Bernoulli}\{(p_i(\boldsymbol{\beta})\}$. We remind the reader of the definition of $p_i(\boldsymbol{\beta})$ is $p_i(\boldsymbol{\beta}) = \frac{exp(\beta_0 + x_i\beta_1)}{1 + exp(\beta_0 + x_i\beta_1)}$ as defined in the question.
- Following our definition of the Q function, we conduct the maximization step of the EM algorithm and maximize this function with respect to the parameters, $\boldsymbol{\beta} = (\beta_0, \beta_1)'$.
- The R code below presents our results where we yield values of $\beta_0 = 0.7636$ to 4 d.p and $\beta_1 = -4.1510$ to 4 d.p as our MLEs.

```r
# Loading relevant packages
library(maxLik)
library(dplyr)
library(tidyr)
library(magrittr)

# Loading data
load('dataex4.Rdata')

dataex4 = dataex4 %>%
  # Sorting data in ascending order in column Y (0 -> 1 -> NA)
  arrange(Y) %>%
  # Creating indicator variable column (if Y == NA -> R = 0, else -> R = 1)
  mutate(R = (Y == 0 | Y == 1)*1) %>%
  # Replacing NAs with 0s in R column
  tidyr::replace_na(list(R = 0)) %>%
  # Replacing NAs with 2s in Y column to prevent coercion problems i.e. 0*NA=NA
  tidyr::replace_na((list(Y = 2)))

# Sigmoid probability function
prob = function(beta0, beta1, x) {
  return(exp(beta0 + x*beta1) / (1 + exp(beta0 + x*beta1)))
}

# Defining Q function for EM algorithm
q_function = function(params, data){
  beta0 = params[1]; beta1 = params[2]
  xx = data$X
  yy = data$Y
  rr = data$R
  sum(yy*rr*(beta0 + beta1*xx) - log(1 + exp(beta0 + beta1*xx)) +
        (1 - rr)*(beta0 + beta1*xx)*prob(beta0, beta1, xx))
```

```r
}


# Get MLE
mle_q = maxLik(q_function, data = dataex4, start = c(beta0=1, beta1=1))
# Present results
summary(mle_q)
```

```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 7 iterations
## Return code 1: gradient close to zero
## Log-Likelihood: -222.825
## 2  free parameters
## Estimates:
##       Estimate Std. error t value  Pr(> t)
## beta0   0.7636     0.1400   5.453 4.94e-08 ***
## beta1  -4.1510     0.3336 -12.442  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

Question 5

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

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

and $\theta = (p, \mu, \sigma^2, \lambda)$

Question 5a

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

Question 5b

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.