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

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

- Note that we have assumed here that $\widehat{\theta}_{\mathrm{MLE}}$ 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

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

- 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,

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

- Note that R is a binary random variable and so

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

- And hence

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

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 $\sqrt{I(\theta)^{-1}} = \theta/\sqrt{n(1 - e^{-C^2/2\theta})}$

```
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,

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

- 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.

```r
library(maxLik)

# Loading in data

load('dataex2.Rdata')
```

```r
# 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)))
}
# 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
## --------------------------------------------
```