Assignment 2

Arturo Esquivel

```r
require(maxLik)
```

1

a) All observations are assumed independent. And given the way in which the observations occur, to construct the likelihood function only $f(y;\theta)$ (for non-censored observations) and $\Pr(Y > C;\theta)$ (for censored observations) are needed.

$$f(y_i;\theta) = \frac{dF(y_i;\theta)}{dy_i} = \frac{y_i}{\theta}e^{-\frac{y_i^2}{2\theta}}$$
$$\Pr(Y > C;\theta) = 1 - F(C;\theta) = e^{-\frac{C^2}{2\theta}}$$

And the likelihood function is given by:

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

Note that, given its definition, $x_i^2 = y_i^2 r_i + C^2(1-r_i)$ so:

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

Taking the *log*:

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

And so

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

And making it equal to 0 yields:

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

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

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

$$\begin{aligned} I(\theta) &= -E\left[\frac{\sum_{i=1}^{n} r_i}{\theta^2} - \frac{\sum_{i=1}^{n} x_i^2}{\theta^3}\right] \\ &= -\frac{nE[R]}{\theta^2} + \frac{nE[X^2]}{\theta^3} \end{aligned}$$

R is a *Bernoulli* variable with $p = F_y(C;\theta)$ and therefore:

$$-\frac{nE[R]}{\theta^2} = -\frac{n}{\theta^2}\left(1 - e^{-\frac{C^2}{2\theta}}\right)$$

Now, $X^2 = Y^2$ when $Y \le C$ and C^2 otherwise, thus:

$$\begin{aligned} \frac{nE[X^2]}{\theta^3} &= \frac{n}{\theta^3}\left(\int_0^C y^2 f(y;\theta)\, dy + C^2\left(e^{-\frac{C^2}{2\theta}}\right)\right) \\ &= \frac{n}{\theta^3}\left(-C^2\left(e^{-\frac{C^2}{2\theta}}\right) + 2\theta\left(1 - e^{-\frac{C^2}{2\theta}}\right) + C^2\left(e^{-\frac{C^2}{2\theta}}\right)\right) \\ &= \frac{2n}{\theta^2}\left(1 - e^{-\frac{C^2}{2\theta}}\right) \end{aligned}$$

Then:

$$\begin{aligned} I(\theta) &= -\frac{nE[R]}{\theta^2} + \frac{nE[X^2]}{\theta^3} \\ &= -\frac{n}{\theta^2}\left(1 - e^{-\frac{C^2}{2\theta}}\right) + \frac{2n}{\theta^2}\left(1 - e^{-\frac{C^2}{2\theta}}\right) \\ &= \frac{n}{\theta^2}\left(1 - e^{-\frac{C^2}{2\theta}}\right) \end{aligned}$$

c) Given the asymptotic normality of the MLE, $\theta_{0.95} = \widehat{\theta}_{\text{MLE}} \pm 1.96 I(\theta)^{-1}$, therefore:

$$CI(\theta)_{0.95} = \left(\widehat{\theta}_{\text{MLE}} - \frac{1.96(\theta^2)}{n\left(1 - e^{-\frac{C^2}{2\theta}}\right)}\ ,\ \widehat{\theta}_{\text{MLE}} + \frac{1.96(\theta^2)}{n\left(1 - e^{-\frac{C^2}{2\theta}}\right)}\right)$$

2

a) A non-censored observation contributes $f(y;\mu,\sigma^2) = \phi(y;\mu,\sigma^2)$ to the likelihood function. A censored one contributes $\Pr(Y < D;\mu,\sigma^2) = \Phi(D;\mu,\sigma^2)$. Given that $X_i = Y_i$ when $Y_i \ge D$ (*and* $R_i = 1$) and D otherwise ($R_i = 0$), the likelihood function is defined as follows:

$$L(\mu, \sigma^2; x, r) = \prod_{i=1}^{n} \left\{ [\phi(x_i; \mu, \sigma^2)]^{r_i} [\Phi(x_i; \mu, \sigma^2)]^{1-r_i} \right\}$$

and taking the *log* leads to

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

b) The function `maxLik` is used to obtain the MLE for μ by maximizing the log likelihood function previously defined. 4, the value when censure occurs, is used as the starting value.

```
load("/cloud/project/dataex2.Rdata")
x <- dataex2$X
r <-dataex2$R
log_like_normal <- function(x, mu){
  sum(r*dnorm(x, mean = mu, sd=1.5, log = TRUE ) +
  (1-r)*pnorm(x, mean = mu, sd =1.5, log = TRUE))
}
mle <- maxLik(logLik = log_like_normal, x = x, start = c(4))
summary(mle)
```

```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 3 iterations
## Return code 1: gradient close to zero
## Log-Likelihood: -336.3821
## 1  free parameters
## Estimates:
##      Estimate Std. error t value Pr(> t)
## [1,]   5.5328     0.1075   51.45  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

The MLE is 5.53 with an associated standard error of 0.11.

3

a)

$$Pr(R = 0 \mid y_1, y_2, \theta, \psi) = \frac{e^{\psi_0 + \psi_1 y_1}}{1 + e^{\psi_0 + \psi_1 y_1}}$$

So it is ignorable because missingness depends only on the fully observed data (Y_1), Y_2 is MAR. Also the vector ψ is distinct from θ.

b)

$$Pr(R = 0 \; | y_1, y_2, \theta, \psi) = \frac{e^{\psi_0 + \psi_1 y_2}}{1 + e^{\psi_0 + \psi_1 y_2}}$$

So it is not ignorable because missingness depends on the missing data (Y_2), Y_2 is MNAR. The vector ψ is distinct from θ, but the first condition is more important for ignorability to be possible.

c)

$$Pr(R = 0 \; | y_1, y_2, \theta, \psi) = \frac{e^{0.5(\mu_1 + \psi y_1)}}{1 + e^{0.5(\mu_1 + \psi y_1)}}$$

So it is ignorable because missingness depends only on the fully observed data (Y_1), Y_2 is MAR. Also the scalar ψ is distinct from θ.

4

First, vectors are created to store the $Y's$ observed and their associated $X's$. Another vector is created to store the $X's$ associated to the missing $Y's$. There are 500 observations from which 95 are missing.

```
load("/cloud/project/dataex4.Rdata")
x_obs <- dataex4$X[is.na(dataex4$Y) == FALSE]
y_obs <- dataex4$Y[is.na(dataex4$Y) == FALSE]
x_mis <- dataex4$X[is.na(dataex4$Y) == TRUE]
```

The likelihood for $\boldsymbol{\beta}$ is given by

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

And assuming that the first 405 for Y are observed:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{405} \left\{ \left(\frac{e^{\beta_0 + x_i \beta_1}}{1 + e^{\beta_0 + x_i \beta_1}} \right)^{y_i} \left(\frac{1}{1 + e^{\beta_0 + x_i \beta_1}} \right)^{1-y_i} \right\} \prod_{i=406}^{500} \left\{ \left(\frac{e^{\beta_0 + x_i \beta_1}}{1 + e^{\beta_0 + x_i \beta_1}} \right)^{y_i} \left(\frac{1}{1 + e^{\beta_0 + x_i \beta_1}} \right)^{1-y_i} \right\}$$

The corresponding log likelihood is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{405} \left\{ y_i(\beta_0 + x_i \beta_1) - \log(1 + e^{\beta_0 + x_i \beta_1}) \right\} + \sum_{i=406}^{500} \left\{ y_i(\beta_0 + x_i \beta_1) - \log(1 + e^{\beta_0 + x_i \beta_1}) \right\}$$

Now we proceed to the E-step

$$
\begin{aligned}
Q(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) &= E_{Y_{mis}}\left[l(\boldsymbol{\beta}) \mid y_{obs}, \boldsymbol{\beta}^{(t)}\right] \\
&= \sum_{i=1}^{405}\left\{y_i(\beta_0 + x_i\beta_1) - \log(1 + e^{\beta_0 + x_i\beta_1})\right\} + \sum_{i=406}^{500}\left\{E\left[y_i \mid y_{obs}, \boldsymbol{\beta}^{(t)}\right](\beta_0 + x_i\beta_1) - \log(1 + e^{\beta_0 + x_i\beta_1})\right\} \\
&= \sum_{i=1}^{405}\left\{y_i(\beta_0 + x_i\beta_1) - \log(1 + e^{\beta_0 + x_i\beta_1})\right\} + \sum_{i=406}^{500}\left\{p_i(\boldsymbol{\beta}^{(t)})(\beta_0 + x_i\beta_1) - \log(1 + e^{\beta_0 + x_i\beta_1})\right\} \\
&= \sum_{i=1}^{405}\left\{y_i(\beta_0 + x_i\beta_1) - \log(1 + e^{\beta_0 + x_i\beta_1})\right\} + \sum_{i=406}^{500}\left\{\left(\frac{e^{\beta_0^{(t)} + x_i\beta_1^{(t)}}}{1 + e^{\beta_0^{(t)} + x_i\beta_1^{(t)}}}\right)(\beta_0 + x_i\beta_1) - \log(1 + e^{\beta_0 + x_i\beta_1})\right\}
\end{aligned}
$$

For the M-step `maxLik` is used to obtain the $\boldsymbol{\beta}$ that maximizes the Q function. An $\epsilon = 0.00001$ is used to determine convergence and $\boldsymbol{\beta}^{(0)} = (0, 0)$.

```
log_like <- function(beta, x_obs, y_obs, x_mis, b_0_t, b_1_t){
  b_0 <- beta[1]
  b_1 <- beta[2]
  sum(y_obs*(b_0 + b_1*x_obs) - log(1 + exp(b_0 + b_1*x_obs))) +
  sum((exp(b_0_t + b_1_t*x_mis)/(1 + exp(b_0_t + b_1_t*x_mis)))*(b_0 + b_1*x_mis)
  - log(1 + exp(b_0 + b_1*x_mis)))
}
beta <- c(0,0)
diff <- 1
eps <- 0.00001
while(diff > eps){
  b_0_t <- beta[1]
  b_1_t <- beta[2]
  Mstep <- maxLik(logLik = log_like, x_obs = x_obs, y_obs = y_obs, x_mis = x_mis,
            b_0_t = b_0_t, b_1_t = b_1_t, start = c(b_0 = 0, b_1 = 0))
  beta <- Mstep$estimate
  diff <- max(abs(beta[1] - b_0_t), abs(beta[2] - b_1_t))
  }
summary(Mstep)
```

```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 5 iterations
## Return code 1: gradient close to zero
## Log-Likelihood: -239.8308
## 2  free parameters
## Estimates:
##     Estimate Std. error t value  Pr(> t)
## b_0   0.9755     0.1225   7.966 1.63e-15 ***
## b_1  -2.4804     0.2376 -10.440  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

The estimate computed is $\widehat{\boldsymbol{\beta}}_{\text{MLE}}$ = (0.9755, -2.4804).

5

a) First, lets define the following variable:

$$z_i = \begin{cases} 1 & \text{if } y_i \text{ belongs to the LogNormal distribution,} \\ 0 & \text{if } y_i \text{ belongs to the Exponential distribution.} \end{cases}$$

With this definition, the likelihood function for θ is defined as

$$L(\theta) = \prod_{i=1}^{n} \left\{ \left[\frac{p}{y_i\sqrt{2\pi\sigma^2}} \left(e^{-\frac{(log(y_i)-\mu)^2}{2\sigma^2}} \right) \right]^{z_i} \left[(1-p)\lambda e^{-\lambda y_i} \right]^{1-z_i} \right\}$$

with the corresponding log likelihood being

$$l(\theta) = \sum_{i=1}^{n} z_i \left\{ \log(p) - \log(y_i) - \log(\sqrt{2\pi\sigma^2}) - \frac{(log(y_i)-\mu)^2}{2\sigma^2} \right\} + \sum_{i=1}^{n} (1-z_i)\{\log(1-p) + \log(\lambda) - \lambda y_i\}$$

z is the missing data since we observe all values for Y but we don't know which distribution each of them comes from. To proceed to the E-step we need to calculate

$$\begin{aligned} Q(\theta \mid \theta^{(t)}) &= E_Z \left[\log L(\theta;) \mid y, \theta^{(t)} \right] \\ &= \sum_{i=1}^{n} E\left[Z_i \mid y_i, \theta^{(t)} \right] \left\{ \log(p) - \log(y_i) - \log(\sqrt{2\pi\sigma^2}) - \frac{(log(y_i)-\mu)^2}{2\sigma^2} \right\} \\ &+ \sum_{i=1}^{n} \left(1 - E\left[Z_i \mid y_i, \theta^{(t)} \right] \right) \{\log(1-p) + \log(\lambda) - \lambda y_i\} \end{aligned}$$

Since $Z_i \mid Y_i, \theta$ is Bernoulli, $E[Z_i \mid y_i, \theta^{(t)}] = \Pr(Z_i = 1 \mid y_i, \theta^{(t)})$. Using the Bayes theorem and the law of total probability we get that:

$$E\left[Z_i \mid y_i, \theta^{(t)} \right] = \Pr(Z_i = 1 \mid y_i, \theta^{(t)}) = \frac{p^{(t)} \frac{1}{y_i\sqrt{2\pi(\sigma^{(t)})^2}} \left(e^{-\frac{(log(y_i)-\mu^{(t)})^2}{2(\sigma^{(t)})^2}} \right)}{p^{(t)} \frac{1}{y_i\sqrt{2\pi(\sigma^{(t)})^2}} \left(e^{-\frac{(log(y_i)-\mu^{(t)})^2}{2(\sigma^{(t)})^2}} \right) + (1-p^{(t)})\lambda^{(t)} e^{-\lambda^{(t)} y_i}} = \tilde{p}_i^{(t)}$$

Thus

$$Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^{n} \tilde{p}_i^{(t)} \left\{ \log(p) - \log(y_i) - \log(\sqrt{2\pi\sigma^2}) - \frac{(log(y_i)-\mu)^2}{2\sigma^2} \right\} + \sum_{i=1}^{n} (1-\tilde{p}_i^{(t)})\{\log(1-p) + \log(\lambda) - \lambda y_i\}$$

For the M-step each parameter should maximize the Q function for which we compute the partial derivatives. We obtain the following updating equations:

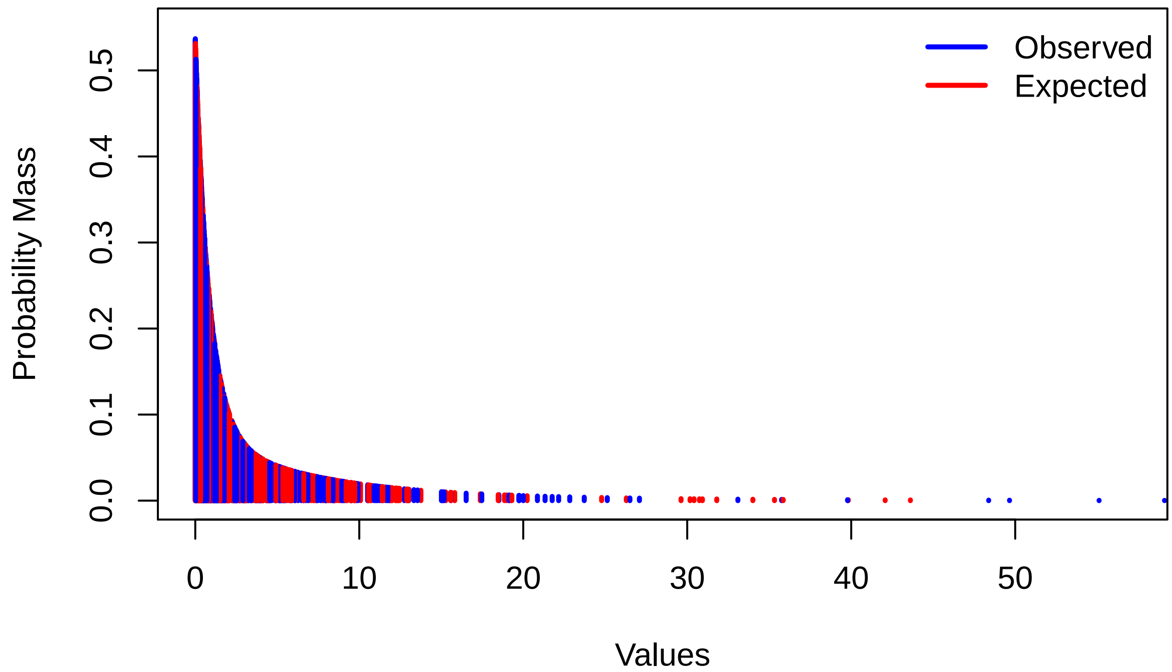
$$
\begin{aligned}
&\frac{\partial}{\partial p} Q(\theta \mid \theta^{(t)}) = 0 \Rightarrow \sum_{i=1}^{n} \frac{1-\tilde{p}_i^{(t)}}{1-p^{(t+1)}} = \sum_{i=1}^{n} \frac{\tilde{p}_i^{(t)}}{p^{(t+1)}} \Rightarrow p^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \tilde{p}_i^{(t)}, \\
&\frac{\partial}{\partial \mu} Q(\theta \mid \theta^{(t)}) = 0 \Rightarrow \mu^{(t+1)} = \frac{\sum_{i=1}^{n} \tilde{p}_i^{(t)} \log(y_i)}{\sum_{i=1}^{n} \tilde{p}_i^{(t)}}, \\
&\frac{\partial}{\partial \sigma^2} Q(\theta \mid \theta^{(t)}) = 0 \Rightarrow (\sigma^{(t+1)})^2 = \frac{\sum_{i=1}^{n} \tilde{p}_i^{(t)} (\log(y_i) - \mu^{(t)})^2}{\sum_{i=1}^{n} \tilde{p}_i^{(t)}}, \\
&\frac{\partial}{\partial \lambda} Q(\theta \mid \theta^{(t)}) = 0 \Rightarrow \lambda^{(t+1)} = \frac{\sum_{i=1}^{n} (1-\tilde{p}_i^{(t)})}{\sum_{i=1}^{n} y_i (1-\tilde{p}_i^{(t)})}
\end{aligned}
$$

b) Now that we have the updating equations we conduct the estimation. An $\epsilon = 0.00001$ is used to determine convergence. $\theta^{(0)} = (0.1, 1, 0.25, 2)$ is used as initial value.

```r
load("/cloud/project/dataex5.Rdata")
EM <- function(y, theta0, eps){
  theta <- theta0
  p <- theta[1]
  mu <- theta[2]
  sigma2 <- theta[3]
  lambda <- theta[4]
  diff <- 1
  while(diff > eps){
    theta_old <- theta

    # E-step
    plogN <- p*dlnorm(y, meanlog = mu, sdlog = sigma2^(0.5))
    pexp <- (1-p)*dexp(y, rate = lambda)
    p_t <- plogN/(plogN + pexp)

    # Update Values
    p <- mean(p_t)
    mu <- sum(p_t*log(y))/sum(p_t)
    sigma2 <- sum(p_t*(log(y)-mu)^2)/sum(p_t)
    lambda <- sum(1-p_t)/sum(y*(1-p_t))
    theta <- c(p, mu, sigma2,lambda)
    diff <- sum(abs(theta-theta_old))
  }
  return(theta)
}

Est <- EM(y=dataex5, theta0 = c(.1,1,.25,2), eps = .00001)
```

The estimate obtained is $\widehat{\theta}_{\text{MLE}} = (0.4796, 2.0132, 0.8638, 1.0331)$. And if we graph the observed and the expected distribution we can see that the estimated model is quite good. The distributions only differ on a couple of outlier observations that are not shown in the plot.

```r
mixture <- Est[1]*dlnorm(dataex5, meanlog = Est[2], sdlog = Est[3]^(0.5)) +
  (1-Est[1])*dexp(dataex5, rate = Est[4])

plot(dataex5, mixture, type = "h", lwd = 2.5, lty = 1,
     col = c("blue", "red"), xlim = c(0,57), ylim = c(0, 0.55),
     xlab = "Values",ylab = "Probability Mass")
legend("topright", legend = c("Observed", "Expected"),
lty = c(1, 1), lwd = c(2.5, 2.5), col = c("blue", "red"), bty = "n")
```



The entire code can be found in the following GitHub repository: https://github.com/Arturo-Esquivel/Incomplete-Data-Analysis