Zero inflated Poisson distribution

Sophie Donnet. For Master 2 Math SV

10/02/2022

1. The data

We study the abundance of fish species at $n = 89$ sites in the Barents Sea (Fossheim, Nilssen, and Aschan (2006)). The data are available in the file BarentsFish.csv where the first 4 columns correspond to four environmental covariates covariates (latitude, longitude, depth, temperature) and the next 30 columns are the abundances of 30 species.

```
abundance <- read.csv("BarentsFish.csv", sep=";")
View(abundance)
```

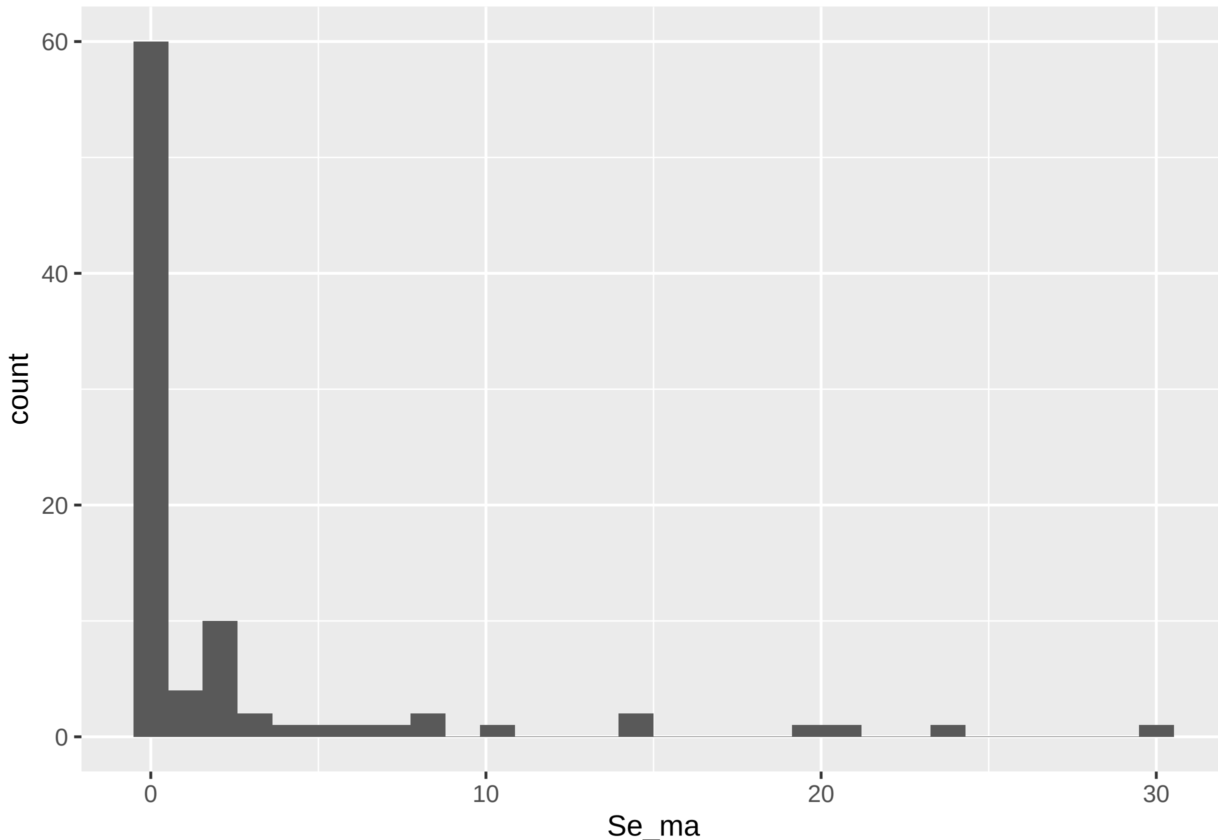
In the following, we will consider only one fish species, for example the 20th ('Se_ma = Sebastes marinus = Golden redfish) and we will note $1 \leq i \leq n$.

$$Y_i = \text{ abundance of golden redfish in station i.}$$

1. Explore the data with standard tools (means, histograms…)

```
library(ggplot2)
ggplot(abundance,aes(Se_ma))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Observe an over-representation of null values. We propose to modelize this over inflation of 0.

2. Zero-inflated Poisson model

We propose to consider the following Zero Inflation Poisson distribution (ZIP) Let Z_i be a latent variable such that

$$Z_i \sim_{i.i.d} \mathcal{B}ern(\pi)$$

Then

$$Y_i|Z_i \sim (1 - Z_i)\delta_{\{0\}} + Z_i\mathcal{P}(\mu_i) \tag{1}$$

Z_i represents the presence of the species. where $\mathcal{P}$ is the Poisson distribution.

2. Write the marginal distribution of Y_i

$$f_Y(y) = (1 - \pi)\delta_{\{0\}}(y) + \pi e^{-\mu}\frac{\mu^y}{y!}$$

3. Derive $\mathbb{E}[Y_i]$ and $P(Y_i = 0)$

$$\mathbb{E}[Y_i] = \pi\mu$$

$$P(Y_i = 0) = (1 - \pi) + \pi e^{-\mu}$$

4. Write the complete log likelihood $\log p_\theta(\mathbf{Y}, \mathbf{Z})$ of the model where $\theta = (\pi, \mu)$.

$$
\begin{aligned}
\log p_\theta(\mathbf{Y},\mathbf{Z}) &= \sum_{i=1}^n \log p_\theta(Y_i|Z_i) + \log p_\theta(Z_i)\\
&= \sum_{i=1}^n \mathbf{1}_{Z_i=0}\log p(Y_i|Z_i=0;\theta) + \mathbf{1}_{Z_i=1}\log p(Y_i|Z_i=1;\theta) + [Z_i\log(\pi) + (1-Z_i)\log(1-\pi)]\\
&= \sum_{i=1}^n \mathbf{1}_{Z_i=1}\log\left(e^{-\mu}\frac{\mu^{Y_i}}{Y_i\boldsymbol{!}}\right) + \mathbf{1}_{Z_i=0}\log(\delta_{\{0\}}(Y_i)) + [Z_i\log(\pi) + (1-Z_i)\log(1-\pi)]\\
&= \sum_{i=1}^n Z_i\left(-\mu + Y_i\log\mu + Cste\right) + [Z_i\log(\pi) + (1-Z_i)\log(1-\pi)]
\end{aligned}
$$

We propose to maximize likelihood with respect to the parameters using the EM algorithm

5. Write the corresponding E-step.

We need $p_{\theta^{(t-1)}}(\mathbf{Z}|\mathbf{Y}) = \prod_{i=1}^n p_{\theta^{(t-1)}}(Z_i|Y_i)$

$$
\begin{aligned}
\tau_i^{(t)} = p_{\theta^{(t-1)}}(Z_i=1|Y_i) &= \frac{p_{\theta^{(t-1)}}(Y_i|Z_i=1)p_{\theta^{(t-1)}}(Z_i=1)}{p_{\theta^{(t-1)}}(Y_i)}\\
&= \frac{\pi^{(t-1)}e^{-\mu^{(t-1)}}\frac{(\mu^{(t-1)})^{Y_i}}{Y_i!})}{(1-\pi^{(t-1)})\delta_{\{0\}}(Y_i) + \pi^{(t-1)}e^{-\mu^{(t-1)}}\frac{(\mu^{(t-1)})^{Y_i}}{Y_i!}}
\end{aligned}
$$

6. Write the corresponding M-step.

$$
\begin{aligned}
\mathbb{E}_{\tau^{t-1}}[\log p_\theta(\mathbf{Y},\mathbf{Z})|\mathbf{Z}] &= \sum_{i=1}^n \mathbb{E}_{\tau^{t-1}}[Z_i]\left(-\mu + Y_i\log\mu + Cste\right) + \mathbb{E}_{\tau^{t-1}}[Z_i\log(\pi) + (1-Z_i)\log(1-\pi)]\\
&= \sum_{i=1}^n \tau_i^{t-1}\left(-\mu + Y_i\log\mu + Cste\right) + \tau_i^{t-1}\log(\pi) + (1-\tau_i^{t-1})\log(1-\pi)
\end{aligned}
$$

$\mu^{(t)}$ such that

$$
\begin{aligned}
\sum_{i=1}^n \tau_i^{t-1}\left(-1 + \frac{Y_i}{\mu^{(t)}}\right) &= 0\\
\mu^{(t)} &= \frac{\sum_{i=1}^n \tau_i^{t-1}Y_i}{\sum_{i=1}^n \tau_i^{t-1}}
\end{aligned}
$$

$\pi^{(t)}$ such that

$$
\begin{aligned}
\sum_{i=1}^n \tau_i^{t-1}\frac{1}{\pi^{(t)}} - (1-\tau_i^{t-1})\frac{1}{1-\pi^{(t)}} &= 0\\
\pi^{(t)} &= \frac{\sum_{i=1}^n \tau_i^{t-1}}{n}
\end{aligned}
$$

7. Suggest an initial value for the parameter θ.

$$\mu^{(0)} = \frac{1}{n}\sum_{i=1}^{n} Y_i \mathbf{1}_{Y_i \neq 0}, \qquad \pi^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Y_i > 0}$$

8. Code the EM algorithm.

```
n <- nrow(abundance)
y <- abundance$Ra_ra
#y <- abundance$Me_ae
```

```
#n=100
#prob.pi = 0.7
#mu = 10
#z = rbinom(n,1,prob.pi)
#y <- z*rpois(n,mu)
```

```
lik = function(y,mu,prob.pi){(1-prob.pi)*(y==0) + prob.pi*dpois(y,mu)}
tol <- 1e-9;

diff <- 2*tol;
prob.pi <- mean(y>0)
mu <-  2*mean(y[y>0])
log.lik <-c(sum(log(lik(y,mu,prob.pi))))

#
print((c(mu,prob.pi)))

while (diff > tol) {

  ### E step
  tau1 <- prob.pi*dpois(y,mu)/lik(y,mu,prob.pi)

  #### M step
  mu.new <- sum(y*tau1)/sum(tau1)
  prob.pi.new <- mean(tau1)

  #### diff mu, prob.pi
  diff = max(abs(mu.new-mu), abs(prob.pi.new-prob.pi))

  mu <- mu.new
  prob.pi <- prob.pi.new

  ll<- sum(log(lik(y,mu,prob.pi)))
  print((c(mu,prob.pi,ll)))

  #####  log.Lik
  log.lik = c(log.lik,ll)

}
plot(log.lik,type='b',pch=20)
lines(log.lik)
```

3. ZIP with covariates

We now consider a model similar to ZIP but taking into account the environmental covariates. We note x_i the vector comprising these covariates for the site i:

$$x_i = [1, \text{latitude}_i, \text{longitude}_i, \text{depth}_i, \text{temperature}_i].$$

We therefore pose : $(Z_i)_{1,\leq i\leq n}$ independent,$(Y_i|Z_i)_{1,\leq i\leq n}$ independent and

$$\begin{array}{rclcrcl} Z_i & \sim & \mathcal{B}\text{ern}(\pi_i) & \text{with} & \log\left(\frac{\pi_i}{1-\pi_i}\right) & = & x_i^T\alpha \\ Y_i|Z_i & \sim & (1-Z_i)\delta_{\{0\}} + Z_i\mathcal{P}(\mu_i) & \text{with} & \log\mu_i & = & x_i^T\beta \end{array} \tag{2}$$

$$\pi_i = \frac{1}{1+e^{-x_i^T\alpha}}, \qquad 1-\pi_i = \frac{e^{-x_i^T\alpha}}{1+e^{-x_i^T\alpha}}$$

The vectors α and β contain the regression coefficients to predict absence and abon- dance conditional on the presence of the species at each site.

9. Write the full log likelihood $p_\theta(\mathbf{Y},\mathbf{Z})$ of this new model as a function of the parameter $\theta = (\alpha,\beta)$.

$$\begin{aligned} \log p_\theta(\mathbf{Y},\mathbf{Z}) &= \sum_{i=1}^n \log p_\theta(Y_i|Z_i) + \log p_\theta(Z_i) \\ &= \sum_{i=1}^n \mathbf{1}_{Z_i=1}\log\left(e^{-\mu_i}\frac{\mu_i^{Y_i}}{Y_i!}\right) + \mathbf{1}_{Z_i=0}\log(\delta_{\{0\}}(Y_i)) + [Z_i\log(\pi_i) + (1-Z_i)\log(1-\pi_i)] \\ &= \sum_{i=1}^n Z_i\left(-e^{x_i^T\beta} + Y_i x_i^T\beta + Cste\right) + [Z_i\log\left(\frac{\pi_i}{1-\pi}\right) + \log(1-\pi)] \\ &= \sum_{i=1}^n Z_i\left(-e^{x_i^T\beta} + Y_i x_i^T\beta + Cste\right) + \left[Z_i x_i^T\alpha + \log\left(\frac{e^{-x_i^T\alpha}}{1+e^{-x_i^T\alpha}}\right)\right] \\ &= \sum_{i=1}^n Z_i\left(-e^{x_i^T\beta} + Y_i x_i^T\beta + Cste\right) + \left[(Z_i-1)x_i^T\alpha - \log\left(1+e^{-x_i^T\alpha}\right)\right] \end{aligned}$$

10. Write the E-step.

$$\begin{aligned} \mu_i^{(t-1)} &= x_i^T\beta^{(t-1)} \\ \pi_i^{(t-1)} &= \frac{1}{1+e^{-x_i^T\alpha^{(t-1)}}} \\ \tau_i^{(t)} &= \frac{\pi^{(t-1)}e^{-\mu_i^{(t-1)}}\frac{(\mu_i^{(t-1)})^{Y_i}}{Y_i!})}{(1-\pi_i^{(t-1)})\delta_{\{0\}}(Y_i) + \pi_i^{(t-1)}e^{-\mu^{(t-1)}}\frac{(\mu_i^{(t-1)})^{Y_i}}{Y_i!}} \end{aligned}$$

11. Write the M-step.

$$\begin{aligned} \beta^{(t)} &= \arg\max_\beta \sum_{i=1}^n \tau_i^{(t)}\left(-e^{x_i^T\beta} + Y_i x_i^T\beta\right) \\ \alpha^{(t)} &= \arg\max_\alpha \sum_{i=1}^n (\tau_i^{(t)}-1)x_i^T\alpha - \log\left(1+e^{-x_i^T\alpha}\right) \end{aligned}$$

It is a maximum likelihood estimator.

12. Propose an initial value for the parameter θ.

- α estimated by Logit regression on $\mathbf{1}_{Y_i>0}$
- β estimated by Poisson regression on $(Y_i)_{i|Y_i>0}$

```
x <- as.matrix(cbind(rep(1, n), abundance[, 1:4]));
alpha0 <- as.vector(glm(1*(y>0) ~ -1 + x, family='binomial')$coef)
beta0<- as.vector(glm(y[y>0] ~ -1 + x[y >0, ], family='poisson')$coef)
```

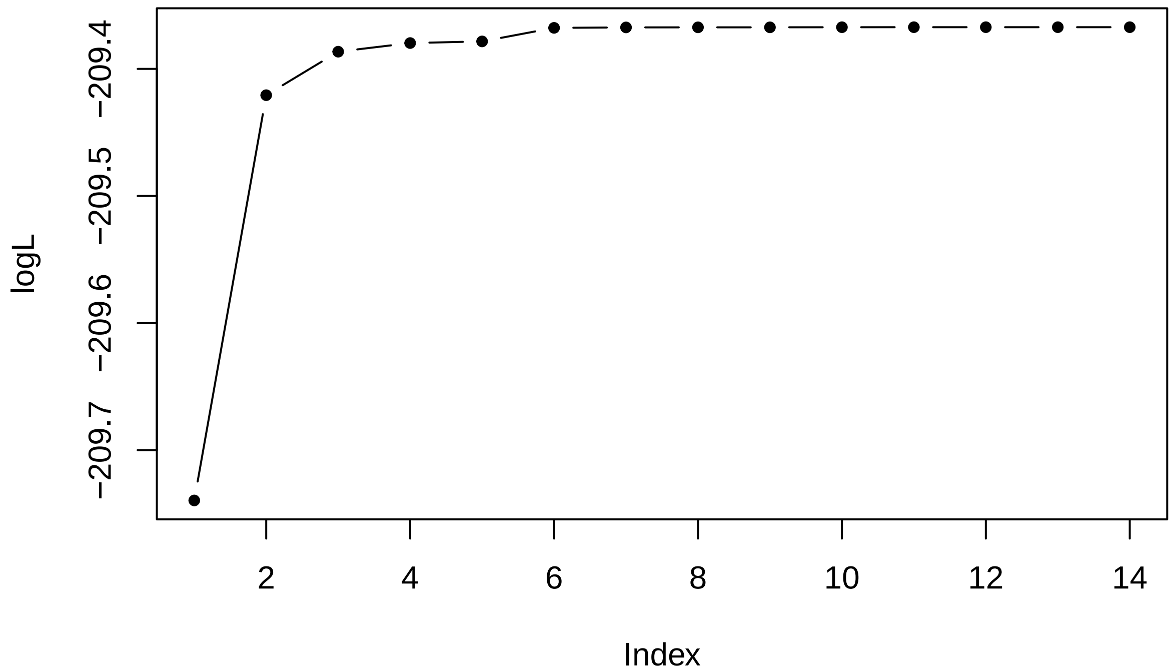
13. Code the EM algorithm

Util functions

```
LogLikBernoulli <- function(alpha, x, tau){sum((tau-1) *  x%*%alpha) - sum(log(1+exp(-x%*%alpha)))}
LogLikPoisson <- function(beta, x, y, tau){
  u <- x%*%beta
  sum(tau*(y*u-exp(u)))
  }
LogLik <- function(alpha, beta, x, y){
  prob.pi <- plogis(x%*%alpha); mu <- exp(x%*%beta)
  return(sum(log((1-prob.pi)*(y==0) + prob.pi*dpois(y, mu))))
}
```

EM

```
tol <- 1e-6; diff <- 2*tol;
iterMax <- 1e3; iter <- 1
logL <- rep(NA, iterMax)
alpha <- alpha0; beta <- beta0
logL[iter] <- LogLik(alpha, beta, x, y)
while((diff > tol) & (iter < iterMax)){
  iter <- iter+1;
  # E step
  prob.pi <- plogis(x%*%alpha); mu <- exp(x%*%beta)
  tau <- 1- (y==0)*(1-prob.pi)/((1-prob.pi)*(y==0) + prob.pi*dpois(y, mu))
  alphaNew <- optim(par=alpha, f=LogLikBernoulli,  x=x, tau=tau,
                    control=list(fnscale=-1))$par
  betaNew <- optim(par=beta, f=LogLikPoisson, y=y, x=x, tau=tau,
                   control=list(fnscale=-1))$par
  # Test & update
  diff <- max(abs(c(alphaNew, betaNew)-c(alpha, beta)))
  alpha <- alphaNew; beta <- betaNew
  logL[iter] <- LogLik(alpha, beta, x=x, y=y)
  #cat(alpha, beta, diff, logL[iter], '\n')
}
logL <- logL[1:iter]
plot(logL, type='b', pch=20)
```




```
### References
```

Fossheim, Maria, Einar M. Nilssen, and Michaela Aschan. 2006. "Fish Assemblages in the Barents Sea." *Marine Biology Research* 2 (4): 260–69. https://doi.org/10.1080/17451000600815698.