# 1 Modèles de Poisson avec excès de zéros

## Données

On s'intéresse à l'abondance du Sébaste doré (*Sebastes marinus* = *Golden redfish*) dans $n = 89$ stations réparties dans la mer de Barents. Les données sont disponibles dans le fichier `GoldenRedfish.csv` dont les 4 premières colonnes correspondent à quatre covariables environementales (latitude, longitude, profondeur, température) et la colonne suivante à l'abondance (comptage) de sébastes dorés.

Dans la suite, on notera

$$Y_i = \text{ abondance dans la station } i \qquad (1 \leq i \leq n).$$

## 1.1 Modèle sans covariable

On considère d'abord un modèle sans covariable prévoyant l'absence ou la présence de l'espèce dans la station $i$ et, conditionnellement à sa présence, une abondance poissonnienne. On pose :

$$\begin{aligned} &\{Z_i\}_{1\leq i\leq n} \text{ iid}, && Z_i \sim \mathcal{B}(\pi), \\ &\{Y_i\}_{1\leq i\leq n} \text{ indépendants} \mid \{Z_i\}, && Y_i \mid Z_i \sim Z_i\delta_0 + (1-Z_i)\mathcal{P}(\lambda). \end{aligned} \tag{1}$$

La variable latente $Z_i$ est donc l'indicatrice d'*absence* de l'espèce dans la station $i$. L'objectif est d'implémenter un algorithme EM afin d'obtenir l'estimateur du maximum de vraisemblance de $\theta = (\pi, \lambda)$.

1. Écrire la vraisemblance complète $\log p_\theta(Y, Z)$ du modèle (1) en fonction de $\theta$.
2. Écrire l'étape E.
3. Écrire l'étape M.
4. Proposer une valeur initiale pour le paramètre $\theta$.
5. Coder l'algorithme EM.
6. Comparer ce modèle au modèle de Poisson simple

$$\{Y_i\}_{1\leq i\leq n} \text{ iid}, \qquad Y_i \sim \mathcal{P}(\lambda). \tag{2}$$

## 1.2 Modèle avec covariables

On considère maintenant un modèle analogue au modèle (1) mais prenant en compte les covariables environnementales. On note $x_i$ le vecteur comprenant ces covariables pour la station $i$, ainsi qu'une terme constant :

$$x_i = [1 \text{ latitude}_i \text{ longitude}_i \text{ profondeur}_i \text{ température}_i]^\intercal.$$

On pose :

$$\begin{aligned} &\{Z_i\}_{1\leq i\leq n} \text{ indépendants}, && Z_i \sim \mathcal{B}(\pi_i), && \log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^\intercal \alpha, \\ &\{Y_i\}_{1\leq i\leq n} \text{ indépendants} \mid \{Z_i\}, && Y_i \mid Z_i \sim Z_i\delta_0 + (1-Z_i)\mathcal{P}(\lambda_i), && \log \lambda_i = x_i^\intercal \beta. \end{aligned} \tag{3}$$

Les vecteurs $\alpha$ et $\beta$ contiennent les coefficients de régression permettant de prédire respectivement l'absence et l'abondance conditionnelle à la présence de l'espèce en chaque site.

1. Écrire la vraisemblance complète $\log p_\theta(Y, Z)$ du modèle (1) en fonction du paramètre $\theta = (\alpha, \beta)$.
2. Écrire l'étape E.
3. Écrire l'étape M.
4. Proposer une valeur initiale pour le paramètre $\theta$.
5. Coder l'algorithme EM.
6. Comparer les trois modèles (1), (2) et (3) ainsi que le modèle de régression poissonnienne

$$\{Y_i\}_{1\leq i\leq n} \text{ indépendants}, \qquad Y_i \sim \mathcal{P}(\lambda_i), \qquad \log \lambda_i = x_i^\intercal \beta. \tag{4}$$

## Rendu attendu

Vous enverrez à l'adresse `stephane.robin@sorbonne-universite.fr` un fichier '.R' contenant l'implémentation en R de l'algorithme EM pour le **modèle avec covariables** (section 1.2). Ce programme devra :

— prendre en entrée un fichier de la même forme que `GoldenRedfish.csv`,
— tracer l'évolution de la log-vraisemblance $\log p_{\theta^{(h)}}(Y)$ au cours des itérations de l'algorithme EM,
— afficher les estimations des vecteurs de coefficients de régression $\alpha$ et $\beta$ du modèle (3),
— afficher les vraisemblances associées aux modèles (1), (2), (3) et (4).