université
PARIS-SACLAY

Statistiques Bayésienne

Analyse de la toxicité des sels sur les larves d'éphémères

Janikson GARCIA BRITO
Aymane MIMOUN
Alexandre Combeau
Hajar Lamtaii
Jaad Belhouari
Étudiants en M2 Data Science

Lieu : Evry-Courcouronnes, France

30 mars 2024

Contents

Chapter 1

Introduction

L'augmentation de la salinité dans les écosystèmes d'eau douce, due notamment aux sels de voirie lessivés dans les rivières après la pluie, constitue une menace pour les espèces aquatiques sensibles, telles que les larves d'éphémères (*mayflies*). Ces larves sont particulièrement vulnérables aux changements de salinité, et leur déclin peut avoir des répercussions sur l'ensemble de l'écosystème aquatique. Cette étude vise à évaluer l'impact de trois types de sels ($CaCl_2$, Commercial Salt, et NaCl) sur la survie des larves d'éphémères, en utilisant plusieurs modèles bayésiens.

L'objectif est de déterminer quel sel est le plus toxique et de proposer des recommandations pour minimiser les impacts environnementaux.

Les données proviennent d'une bioanalyse où des larves ont été collectées dans la nature, placées dans des réservoirs d'eau, et exposées à des concentrations croissantes de sel (de 1 à 1024 unités). La survie a été mesurée après 96 heures.

Chapter 2

Analyse descriptive

L'objectif de cette section est d'explorer les données afin de mieux comprendre les tendances générales et d'identifier d'éventuelles anomalies. Nous analysons d'abord la structure des données avant d'examiner la relation entre la concentration de sel et le taux de survie des larves d'éphémères.

2.1 Structure des données

Le jeu de données contient 93 observations. Le tableau 2.1 présente un résumé des principales variables.

Variable	Description
N_0	Nombre initial de larves
N_{surv}	Nombre de larves survivantes après 96h
conc	Concentration du sel (1 à 1024 unités)
Salt	Type de sel utilisé ($CaCl_2$, Commercial Salt, NaCl)

Table 2.1: Description des variables du jeu de données

2.2 Détection d'anomalies

Avant de procéder aux analyses, nous vérifions la cohérence des données. Une attention particulière est portée à la variable N_{surv} pour s'assurer qu'aucune valeur ne dépasse N_0. Après inspection, aucune anomalie n'a été détectée dans le jeu de données.

2.3 Analyse exploratoire

La Figure 2.1 illustre le taux de survie des larves en fonction de la concentration de sel, selon le type de sel utilisé. Une variabilité importante est observée aux concentrations intermédiaires, probablement due à la variabilité biologique et au faible nombre de larves par essai (N_0 entre 1 et 13).

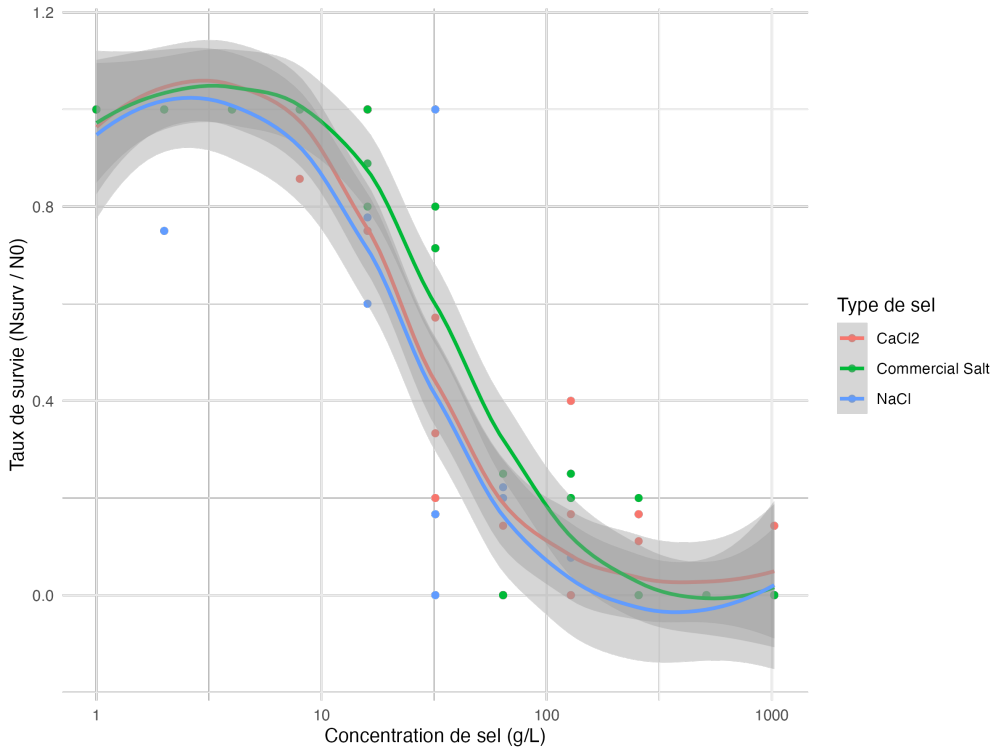


Figure 2.1: Le taux de survie des larves diminue avec l'augmentation de la concentration en sel. Parmi les sels étudiés, $CaCl_2$ semble être le plus toxique, entraînant une mortalité rapide dès 32 unités de concentration. En revanche, le sel commercial apparaît comme le moins nocif, ayant un impact moins marqué sur la survie des larves. Ces résultats soulignent l'importance de la nature chimique des sels dans leur toxicité.

Cette analyse exploratoire met en évidence des différences de toxicité entre les trois sels étudiés. L'étape suivante consistera à modéliser ces effets à l'aide de méthodes bayésiennes pour quantifier la toxicité de chaque sel avec plus de précision.

Chapter 3

Modèle Log-Logistic

Après l'exploration des données, nous utilisons un modèle log-logistique pour modéliser la probabilité de survie p des larves d'éphémères en fonction de la concentration de sel (conc). Ce modèle est couramment utilisé en toxicologie pour décrire des relations dose-réponse.

3.1 Formulation du modèle

La probabilité de survie p_i pour le sel i est définie par :

$$p_i = \frac{d_i - c_i}{1 + \left(\frac{\text{conc}}{\theta_i}\right)^{b_i}} + c_i, \quad \text{pour } i = 1, 2, 3 \tag{3.1}$$

avec :

- p_i : Probabilité de survie des larves pour le sel i (utilisée dans une distribution binomiale : $N_{\text{surv},i} \sim \text{Binomial}(N_{0,i}, p_i)$).
- d_i : Probabilité de survie à conc $= 0$ pour le sel i.
- c_i : Probabilité de survie à conc $\to \infty$ pour le sel i.
- b_i : Paramètre de pente, contrôlant la rapidité de la chute de la survie pour le sel i.
- θ_i : Point d'inflexion (concentration où la survie est à mi-chemin entre d_i et c_i) pour le sel i.

Nous ajustons ce modèle en considérant des paramètres spécifiques à chaque sel ($CaCl_2$, Commercial Salt et NaCl) afin de comparer leur toxicité relative.

3.2 Choix des priors

Les distributions a priori sont définies sur la base des observations exploratoires :

- $d_i \sim \text{Beta}(5, 1)$: La survie est proche de 1 à faible concentration, donc un prior favorisant des valeurs élevées.

- $c_i \sim \text{Beta}(1, 5)$: La survie est proche de 0 à haute concentration, donc un prior favorisant des valeurs faibles.
- $b_i \sim \text{Gamma}(2, 1)$: Un prior faiblement informatif pour la pente.
- $\theta_i \sim \text{LogNormal}(4, 1)$: Le point d'inflexion semble se situer entre 16 et 128 unités, justifiant ce choix de prior.

3.3 Implémentation du modèle

L'implémentation du modèle log-logistique en JAGS est détaillée en annexe (voir Annexe 8.1). Cette approche permet d'estimer les paramètres bayésiens de manière rigoureuse et d'évaluer la toxicité relative des différents sels.

3.4 Résultats du modèle log-logistique

3.4.1 Estimation des paramètres

Les paramètres estimés du modèle log-logistique sont les suivants :

- d_1 ($CaCl_2$) : 0.978, d_2 (Commercial Salt) : 0.985, d_3 (NaCl) : 0.968 $\Rightarrow$ Survie élevée à faible concentration.
- c_1 : 0.070, c_2 : 0.040, c_3 : 0.016 $\Rightarrow$ Survie faible à haute concentration.
- b_1 : 3.663, b_2 : 2.812, b_3 : 2.556 $\Rightarrow$ Pente plus abrupte pour $CaCl_2$.
- θ_1 : 28.14, θ_2 : 40.26, θ_3 : 29.06 $\Rightarrow$ Commercial Salt a le point d'inflexion le plus élevé (moins toxique).

3.4.2 Vérification de la convergence MCMC

Nous vérifions la convergence des chaînes MCMC pour garantir la fiabilité des estimations. Les trace plots (Figure 3.1) montrent un bon mélange des chaînes pour tous les paramètres, sans tendances ni divergences.

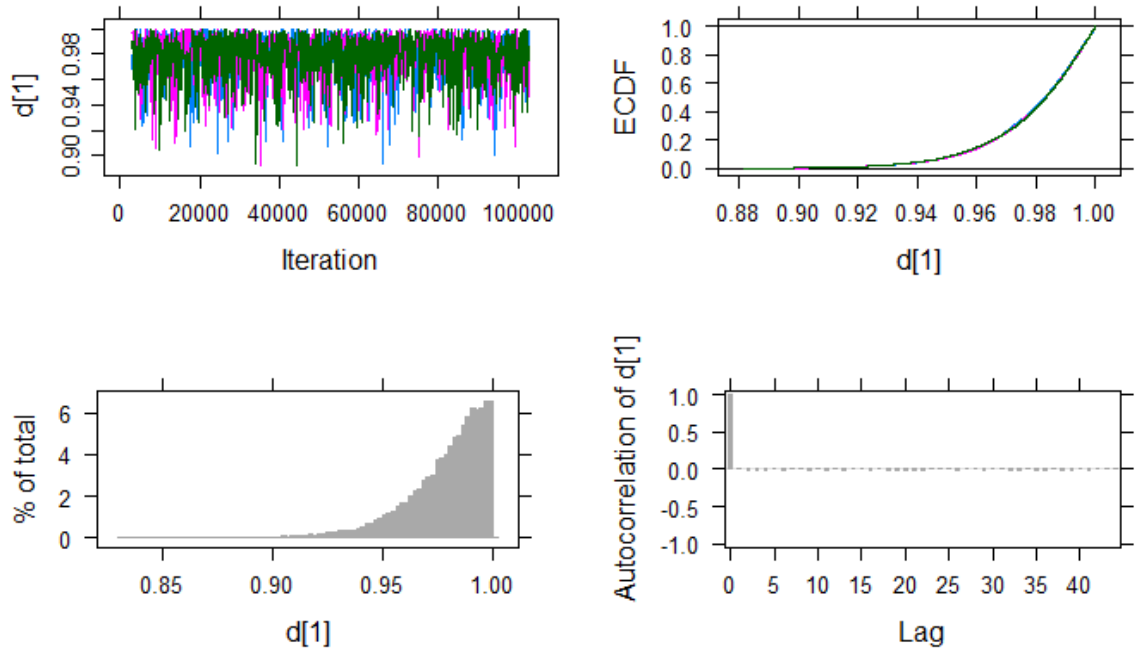


Figure 3.1: Trace plots des paramètres du modèle log-logistique. Les chaînes montrent une bonne convergence.

Critères de convergence :

- **R-hat** : Toutes les valeurs sont entre 0.9999 et 1.0003, bien en dessous de 1.1.
- **SSeff** : Toutes les valeurs sont ¿ 27000, bien au-dessus de 1000.
- **AC.100** : Faible autocorrélation, valeurs proches de 0.

La convergence est satisfaisante pour tous les paramètres.

3.4.3 Vérification avec des données simulées

Nous utilisons un *Fake Data Check* pour tester la robustesse du modèle. Nous simulons des données avec : $d = 0.9$, $c = 0.1$, $b = 3$, $\theta = 30$ et générons des valeurs de N_{surv} à partir d'une distribution binomiale.

Résultats de la récupération des paramètres :

- $d = 0.9$, estimé : 0.909 (IC 95% : 0.802, 1.000) $\Rightarrow$ Bien récupéré.
- $c = 0.1$, estimé : 0.152 (IC 95% : 0.055, 0.262) $\Rightarrow$ Légère surestimation.
- $b = 3$, estimé : 2.143 (IC 95% : 0.703, 4.727) $\Rightarrow$ Sous-estimé.
- $\theta = 30$, estimé : 18.06 (IC 95% : 8.36, 30.78) $\Rightarrow$ Sous-estimé.

3.4.4 Prédictions et comparaison des toxicités

Nous avons calculé la probabilité de survie p pour une gamme de concentrations (1 à 1024) pour chaque sel.

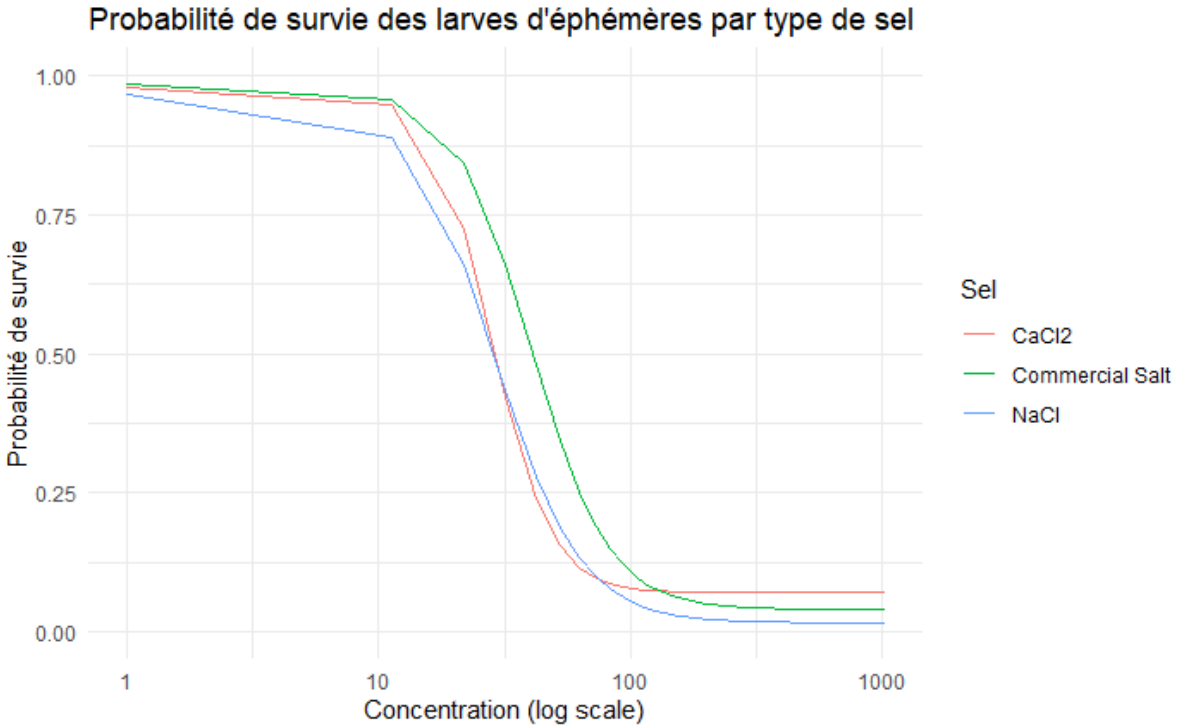


Figure 3.2: Probabilités de survie prédites en fonction de la concentration de sel.

Interprétation :

- $CaCl_2$ et NaCl montrent une chute rapide de la survie autour de 28-29 unités.
- Commercial Salt montre une chute plus lente autour de 40 unités.
- À haute concentration, NaCl est le plus létal ($c = 0.016$), suivi de Commercial Salt ($c = 0.040$) et $CaCl_2$ ($c = 0.070$).

Comparaison des toxicités :

- θ_1 ($CaCl_2$) : 28.14 (IC 95% : 21.08, 35.53).
- θ_2 (Commercial Salt) : 40.26 (IC 95% : 29.62, 51.98).
- θ_3 (NaCl) : 29.06 (IC 95% : 21.23, 37.78).
- c_1 ($CaCl_2$) : 0.070 (IC 95% : 0.019, 0.124).
- c_2 (Commercial Salt) : 0.040 (IC 95% : 0.000, 0.092).
- c_3 (NaCl) : 0.016 (IC 95% : 0.000, 0.044).

3.4.5 Comparaison avec les données observées

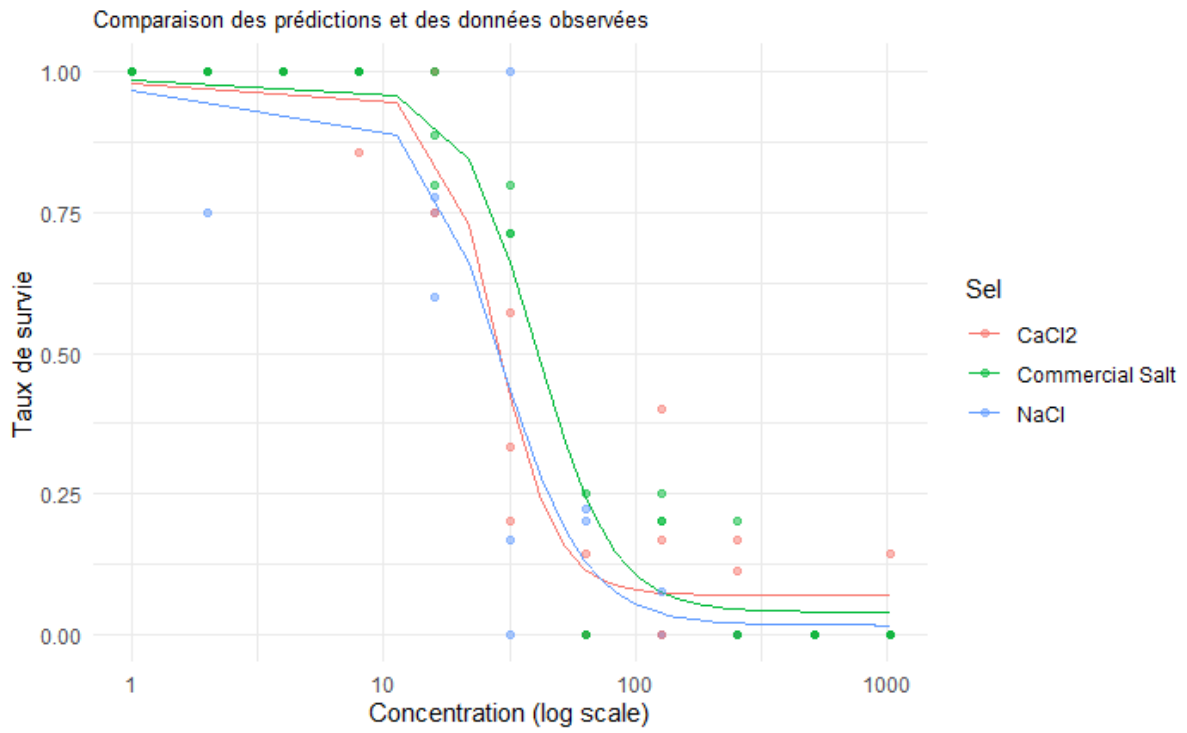


Figure 3.3: Comparaison des prédictions et des données observées.

Observations :

- Les courbes prédites suivent bien les tendances des données observées.
- À faible concentration (1 à 16), la survie est proche de 1, cohérent avec d.
- À haute concentration (128 à 1024), la survie est proche de 0, cohérent avec c.
- Les points d'inflexion (θ) sont bien capturés.

3.4.6 Conclusions sur le modèle log-logistique et recommandations

Résumé des résultats :

- $CaCl_2$ et NaCl sont les plus toxiques ($\theta \approx 28 - 29$).
- NaCl est le plus létal à haute concentration ($c = 0.016$).
- Commercial Salt est le moins toxique ($\theta = 40.26$).

Recommandations :

- Remplacer $CaCl_2$ et NaCl par Commercial Salt pour réduire l'impact environnemental.

Limites du modèle :

- Le modèle montre un léger biais dans le *Fake Data Check* (sous-estimation de θ, surestimation de c).
- Les courbes prédisent la tendance générale mais ne capturent pas entièrement la variabilité biologique.
- Possibilité d'améliorer le modèle avec des priors plus adaptés ou plus de données.

Chapter 4

Modèle de Weibull

4.1 Description du modèle de Weibull et choix des priors

4.1.1 Formule du modèle

Le modèle de Weibull est utilisé pour analyser la survie des larves d'éphémères en fonction de la concentration de sel, avec une distinction entre trois types de sel. La probabilité de survie $S(\text{conc}, \text{salt}_i)$ pour un sel donné i est définie comme suit :

$$S(\text{conc}, \text{salt}_i) = e^{-\left(\frac{\text{conc}}{\alpha_i}\right)^{\beta_i}} \tag{4.1}$$

où :

- conc est la concentration de sel (en g/L),
- $i = 1, 2, 3$ représente les trois types de sel,
- $\alpha_i > 0$ est le paramètre d'échelle spécifique au sel i, indiquant la concentration à laquelle la survie diminue significativement,
- $\beta_i > 0$ est le paramètre de forme spécifique au sel i, contrôlant la pente et la forme de la courbe de survie.

La probabilité de mortalité est alors :

$$F(\text{conc}, \text{salt}_i) = 1 - S(\text{conc}, \text{salt}_i) = 1 - e^{-\left(\frac{\text{conc}}{\alpha_i}\right)^{\beta_i}} \tag{4.2}$$

Pour une approche bayésienne hiérarchique, les paramètres α_i et β_i sont considérés comme des effets aléatoires tirés de distributions globales :

$$\log(\alpha_i) \sim N(\mu_\alpha, \sigma_\alpha) \tag{4.3}$$

$$\log(\beta_i) \sim N(\mu_\beta, \sigma_\beta) \tag{4.4}$$

La concentration létale médiane (CL50) pour chaque sel est calculée comme :

$$\text{CL50}_i = \alpha_i \cdot (\ln(2))^{1/\beta_i} \tag{4.5}$$

Elle permet de comparer directement la toxicité des trois sels.

4.2 Stratégie de modélisation

La stratégie adoptée repose sur une approche bayésienne hiérarchique pour :

- Modéliser la survie : Utiliser une distribution binomiale pour le nombre de survivants (n) par rapport au nombre total d'individus (total) dans chaque essai, avec $S(\text{conc}, \text{salt}_i)$ comme probabilité de succès.
- Intégrer les trois sels : Permettre des paramètres α_i et β_i spécifiques à chaque sel, tout en les liant via une structure hiérarchique pour partager l'information et éviter le surajustement.
- Répondre à l'objectif : Estimer les CL50 pour identifier le sel le plus toxique, en tenant compte de l'incertitude via des intervalles de crédibilité bayésiens.
- Respecter les exigences : Fournir une implémentation reproductible, vérifier la convergence MCMC, et discuter des priors.

Cette approche est flexible et adaptée aux données de bioessais, permettant de capturer des différences dans la toxicité (via α_i) et la forme de la réponse (via β_i).

4.3 Choix des priors

Les priors sont choisis pour être faiblement informatifs, reflétant une absence de connaissance a priori forte tout en assurant la stabilité de l'estimation. Voici les choix et leurs justifications :

Pour μ_α (moyenne globale des $\log(\alpha_i)$)

$$\mu_\alpha \sim N(0, 10) \tag{4.6}$$

Justification : En échelle logarithmique, μ_α représente la moyenne des paramètres d'échelle. Une variance large (10) permet une grande plage de valeurs possibles pour α_i.

Pour σ_α (variabilité des $\log(\alpha_i)$)

$$\sigma_\alpha \sim \text{Half-Cauchy}(0, 5) \tag{4.7}$$

Justification : La distribution Half-Cauchy est un choix standard pour les paramètres d'écart-type dans les modèles hiérarchiques bayésiens.

Pour μ_β (moyenne globale des $\log(\beta_i)$)

$$\mu_\beta \sim N(0, 5) \tag{4.8}$$

Justification : β_i contrôle la forme de la courbe, avec des valeurs typiques autour de 1 à 3 dans les bioessais.

Pour σ_β (variabilité des $\log(\beta_i)$)

$$\sigma_\beta \sim \text{Half-Cauchy}(0, 5) \tag{4.9}$$

Justification : Similaire à σ_α, ce prior permet une variabilité entre les sels dans la forme de la réponse, tout en restant conservateur.

Ces priors sont cohérents avec les recommandations de *Regression and Other Stories* (Gelman et al., 2020), qui préconise des priors faiblement informatifs pour les modèles hiérarchiques, ajustés au contexte scientifique.

4.4 Résultats du modèle de Weibull

4.4.1 Courbes de survie

La Figure 4.1 illustre les courbes de survie prédites pour les trois sels en fonction de la concentration (en g/L). On observe une diminution progressive de la probabilité de survie avec l'augmentation de la concentration pour tous les sels. Cependant, des différences notables apparaissent :

- **$CaCl_2$ (courbe rouge)** : La survie diminue rapidement, atteignant une probabilité proche de 0 autour de 50 g/L, indiquant une toxicité élevée. - **NaCl (courbe bleue)** : La courbe de survie est similaire à celle de $CaCl_2$, avec une chute rapide autour de 50 g/L, suggérant une toxicité comparable. - **Commercial Salt (courbe verte)** : La diminution de la survie est plus graduelle, avec une probabilité de survie plus élevée jusqu'à environ 100 g/L, indiquant une toxicité moindre.

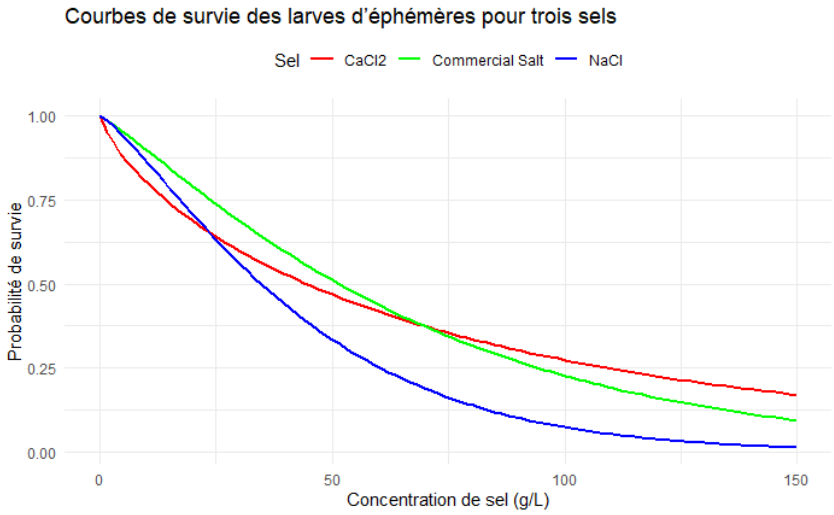


Figure 4.1: Courbes de survie prédites pour les trois sels selon le modèle de Weibull.

4.4.2 Estimation des CL50

La concentration létale médiane (CL50), définie comme la concentration à laquelle 50 % des larves ne survivent pas, a été estimée pour chaque sel à partir des paramètres α_i et β_i selon la formule :

$$\text{CL50}_i = \alpha_i \cdot (\ln(2))^{1/\beta_i}$$

Les histogrammes des estimations bayésiennes des CL50 (Figure 4.2) montrent les distributions a posteriori des CL50 pour chaque sel :

- **CL50 pour $CaCl_2$** : centrée autour de 35 g/L, avec une plage de 25 à 45 g/L, indiquant une toxicité élevée. - **CL50 pour Commercial Salt** : centrée autour de 55 g/L, avec une plage de 35 à 75 g/L, confirmant une toxicité moindre. - **CL50 pour NaCl** : centrée autour de 40 g/L, avec une plage de 30 à 50 g/L, suggérant une toxicité intermédiaire.

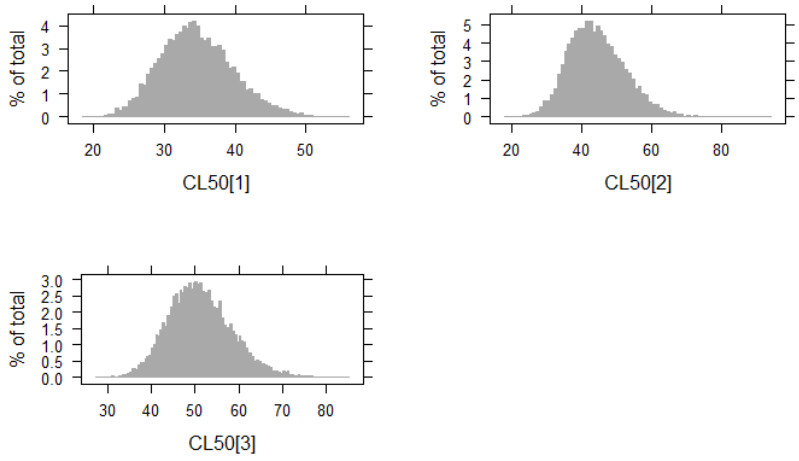


Figure 4.2: Distributions a posteriori des CL50 estimées pour les trois sels.

4.4.3 Vérification de la convergence MCMC

La convergence des chaînes MCMC a été vérifiée grâce aux diagnostics standards : - Le critère *R-hat* est proche de 1 (typiquement < 1.1), indiquant une bonne convergence. - Le nombre effectif d'échantillons (SSeff) est supérieur à 1000. - L'autocorrélation est faible.

Ces résultats confirment la robustesse des estimations des paramètres α_i, β_i et des CL50.

4.4.4 Comparaison des toxicités

Les estimations des CL50 permettent de comparer la toxicité des sels : - **CaCl2** est le plus toxique ($CL50 \approx 35g/L$). - **NaCl** a une toxicité intermédiaire ($CL50 \approx 40g/L$). - **Commercial Salt** est le moins toxique ($CL50 \approx 55g/L$).

Ces tendances sont cohérentes avec les courbes de survie.

4.4.5 Conclusions et recommandations

Le modèle de Weibull confirme que $CaCl_2$ et NaCl sont plus toxiques que Commercial Salt. Par conséquent, nous recommandons :

- De privilégier Commercial Salt comme alternative pour le déglaçage afin de réduire l'impact environnemental. - D'effectuer des études complémentaires sur son efficacité et son coût économique.

Limites du modèle : - La variabilité biologique aux concentrations intermédiaires est difficile à capturer. - L'incertitude sur les CL50 pourrait être réduite avec un plus grand échantillon.

Chapter 5

Régression Logistique Bayésienne

5.1 Modélisation bayésienne

5.1.1 Structure du modèle

La régression logistique est un modèle statistique utilisée pour modéliser la probabilité de survie des mayflies en fonction de la concentration du sel et du type de sel utilisé (NaCl, CaCl2 ou Sel Commercial). Le modèle logistique classique est donné par la fonction sigmoïde qui modélise la probabilité de la survie ou non des mayflies. La probabilité de survie est définit telle que :

$$P(\text{survie} = 1) = \frac{1}{1 + \exp -(\alpha + \beta \cdot \log(\text{conc}) + \gamma_1 \cdot X_{\text{NaCl}} + \gamma_2 \cdot X_{\text{CaCl2}} + \gamma_3 \cdot X_{\text{Commercial}})} \tag{5.1}$$

où :

- α est l'intercept (l'ordonnée à l'origine),
- β est le coefficient qui représente l'impact du logarithme de la concentration du sel sur la probabilité de survie,
- $\gamma_1, \gamma_2, \gamma_3$ sont les coefficients représentant les effets des sels sur la survie,
- $X_{\text{NaCl}}, X_{\text{CaCl2}}, X_{\text{Commercial}}$ sont des variables indicatrices qui prennent la valeur 1 si le sel correspondant est présent et 0 sinon.

5.1.2 Choix des priors

Le choix des priors pour chaque paramètre de notre modèle ci dessus est capital. Ces priors vont nous permettre d'encoder nos connaissances à priori sur les paramètres.

La nature de chaque variable ainsi que leurs informations disponibles nous ont permis de determiner des distributions à priori :

L'intercept

Le paramètre α dans notre modèle représente l'intercept et détermine la probabilité de survie des mayflies lorsque la concentration de sel est à son niveau de base. En d'autres

termes, α fixe la probabilité de survie lorsque $\log(\text{conc}) = 0$, ce qui correspond à une concentration de sel égale à 1, soit son niveau le plus bas. La probabilité de survie, p, est liée à α par la fonction logit : le paramètre α est lié à β ainsi qu'aux γ par la formule du logit :

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \alpha + \beta \log(\text{conc}) + \gamma_1 \cdot \text{NaCl} + \gamma_2 \cdot \text{CaCl2} + \gamma_3 \cdot \text{Commercial}$$

En effet, α influence directement la probabilité de survie au point où la concentration de sel est minimale. À mesure que α augmente, la probabilité de survie approche 1, et inversement. Nous nous attendons donc à ce que la probabilité de survie soit proche de 1 lorsque conc = 1. Par conséquent, la distribution de α devrait être asymétrique et positive, reflétant une forte probabilité de survie à faible concentration de sel.

En raison de cette asymétrie et de la contrainte de positivité de α, nous avons choisi une loi log-normale comme prior pour α, adaptée aux paramètres qui doivent être positifs et suivent une distribution asymétrique. Ainsi, nous définissons le prior de α comme suit :

$$\alpha \sim \text{lognormal}(6, 1)$$

où 6 est la moyenne de la distribution lognormale et 1 est l'écart-type de la log-transformation de α. Un écart-type plus faible aurait pu être envisagé, mais il aurait moins bien reflété la distribution des données, sachant que la médiane de la probabilité de survie est élevée.

La motivation pour une moyenne de 6 dans le prior lognormal vient du fait que nous nous attendons à ce que la probabilité de survie soit proche de 1 lorsque la concentration de sel est minimale, c'est-à-dire pour $\log(\text{conc}) = 0$. À ce niveau, la probabilité de survie devient :

$$p = \frac{1}{1 + \exp(-\alpha)}$$

Lorsque $\alpha = 6$, la probabilité de survie est proche de 1 :

$$p = \frac{1}{1 + \exp(-6)} \approx 0.9975$$

Cela signifie qu'à une concentration de sel de 1 unité, la probabilité de survie est très élevée, ce qui correspond à nos attentes. C'est pourquoi nous choisissons un prior lognormal avec une moyenne de 6 pour refléter cette probabilité élevée à faible concentration de sel.

Ce sont les raisons pour le choix de notre prior pour α.

5.1.3 L'effet de le concentration de sel : β

Le paramètre β de notre modèle logistique représente l'effet de la concentration de sel sur la survie des mayflies. Plus précisément, β quantifie l'impact du logarithme de la concentration de sel sur la probabilité de survie. Autrement dit, si β augmente, la probabilité de survie augmente également. Inversement, si β diminue, $P(\text{survie} = 1) \to 0$. On s'attend donc à ce que la concentration en sel ait un effet modéré sur la survie des mayflies, ce qui implique un β négatif, mais proche de 0.

En tenant compte de ces connaissances a priori et en supposant que l'influence de la concentration de sel sur la survie des mayflies ne soit pas extrême, nous avons choisi une loi normale centrée en 0 avec un écart-type relativement large. Le prior pour le paramètre β est donc :

$$\beta \sim \mathcal{N}(-3, 5)$$

Ce choix de prior repose sur deux intuitions principales :

1. La croyance a priori selon laquelle l'effet de la concentration de sel sur la survie des mayflies est négatif, mais relativement modéré.

2. Une incertitude sur l'intensité de cet effet, d'où le choix d'un écart-type large.

5.1.4 Les différents types de sel : γ

Le vecteur $\gamma = (\gamma_1, \gamma_2, \gamma_3)$ représente respectivement l'effet des types de sels NaCl, CaCl2 et Commercial Salt sur la survie des mayflies. Ces paramètres permettent de modéliser l'impact spécifique de chaque type de sel sur la probabilité de survie des mayflies. L'objectif de ces paramètres est de déterminer comment chaque type de sel influence cette probabilité. Ces paramètres sont cruciaux dans cette étude, car notre motivation est de savoir quel type de sel est le plus toxique pour les mayflies.

D'après le graphique 2.1, nous pouvons constater que la tendance est similaire pour les trois types de sel. La probabilité de survie est proche de 1 à faible concentration de sel et tend vers 0 à mesure que la concentration augmente.

Ainsi, nous avons choisi d'attribuer le même prior pour les trois paramètres. De plus, nous avons opté pour un prior normal centré, avec un écart-type noté σ. Ce paramètre représente l'écart-type des effets des sels, et il est essentiel dans notre modèle pour contrôler la variabilité de l'effet de chaque type de sel sur la probabilité de survie des mayflies. Ce paramètre a été introduit pour tenir compte de l'incertitude entourant l'effet de chaque type de sel, et il reflète à quel point l'impact de chaque type de sel peut varier d'une observation à l'autre.

Le prior choisi pour chaque γ_i (pour $i = 1, 2, 3$) est donc :

$$\gamma_i \sim \mathcal{N}(0, \sigma_\gamma),$$

où σ_γ est l'écart-type des effets aléatoires des sels. Nous avons choisi $\sigma_\gamma \sim \text{Cauchy}(0, 3)$, un prior qui reflète une certaine incertitude sur l'amplitude de l'effet des sels tout en permettant une large variabilité. Le choix de la distribution Cauchy pour σ_γ est motivé par la robustesse de cette distribution dans le cadre de modèles bayésiens, où elle favorise des valeurs plus petites tout en permettant des valeurs plus grandes. Cela est particulièrement utile pour modéliser des effets aléatoires avec une grande incertitude.

Nous avons choisi un prior pour σ_γ basé sur une distribution Cauchy, soit :

$$\sigma_\gamma \sim \text{Cauchy}(0, 3).$$

Une distribution Cauchy est utilisée dans ce modèle bayésien pour modéliser le paramètre σ_γ, qui est censé être positif (car il représente un écart-type) tout en permettant une certaine variabilité sans contrainte. C'est pourquoi la Cauchy est particulièrement adaptée

ici, avec un écart-type de 3, qui reflète notre incertitude modérée sur l'ampleur des effets de chaque type de sel.

5.1.5 Implémentation du modèle avec STAN

L'implémentation du modèle bayésien de régression logistique a été réalisée à l'aide de STAN, utilisé pour ajuster les paramètres du modèle à partir des données sur les mayflies et pour estimer ces paramètres par la méthode de Monte Carlo par chaînes de Markov (MCMC).

Avant d'implémenter le modèle avec STAN, il était important de préparer les données. Celles-ci doivent être organisées dans un format spécifique que STAN comprend. Les blocs utilisés dans la modélisation sont les suivants :

- **data** : Ce bloc définit les variables d'entrée, incluant les observations (N_0, N_{surv}), les variables explicatives (log(conc), les indicateurs de type de sel), ainsi que la taille de l'échantillon.
- **parameters** : Les paramètres inconnus du modèle sont déclarés ici, y compris α, β, les effets des sels γ_i, et leur écart-type σ_γ.
- **model** : Ce bloc contient la spécification du modèle statistique, les distributions a priori des paramètres et la vraisemblance binomiale des observations.

Les paramètres utilisés dans le modèle sont :

- Le nombre initial de mayflies, N_0,
- Le nombre de mayflies survécus, N_{surv},
- La concentration de sel, conc,
- Les indicateurs pour chaque type de sel (NaCl, CaCl2, et Commercial),
- Le nombre total d'observations.

Ces données ont été organisées dans un tableau pour être traitées par STAN.

5.1.6 Estimation par HMC et paramétrage des chaînes

Pour estimer les paramètres du modèle, nous avons utilisé l'algorithme de Monte Carlo par chaînes de Markov (MCMC) via la méthode HMC (Hamiltonian Monte Carlo). Le paramétrage des chaînes a été effectué en définissant les éléments suivants :

- Un total de 8000 itérations, incluant une phase de chauffe (warm-up) de 2000 itérations pour assurer la convergence (voir figure **??**),
- 4 chaînes indépendantes afin d'évaluer la stabilité des estimations, conformément aux pratiques courantes dans la littérature,
- Une estimation adaptative du pas de saut et de la trajectoire pour optimiser l'exploration de l'espace des paramètres.

5.2 Vérification de la convergence des Chaînes de Markov Monte Carlo (MCMC)

L'étape la plus importante après l'implémentation du modèle bayésien et la verification de la convergence des Markov Chain Monte Carlo (MCMC). Afin d'évaluer la convergence, nous avons utilisé plusieurs critères standards. L'ensemble de ces diagnostics ont été utilisés pour s'assurer que la convergence des chaînes MCMC était adéquate avant de procéder à l'analyse des résultats.

5.2.1 Traceplots et verification graphique

Un traceplot est un graphique qui montre l'évolution des échantillons des paramètres au cours des itérations, debutant à 2000 pour s'assurer de la convergence. La figure ci dessous représente les traceplots pour les paramètres α et β. Comme on peut le voir dans

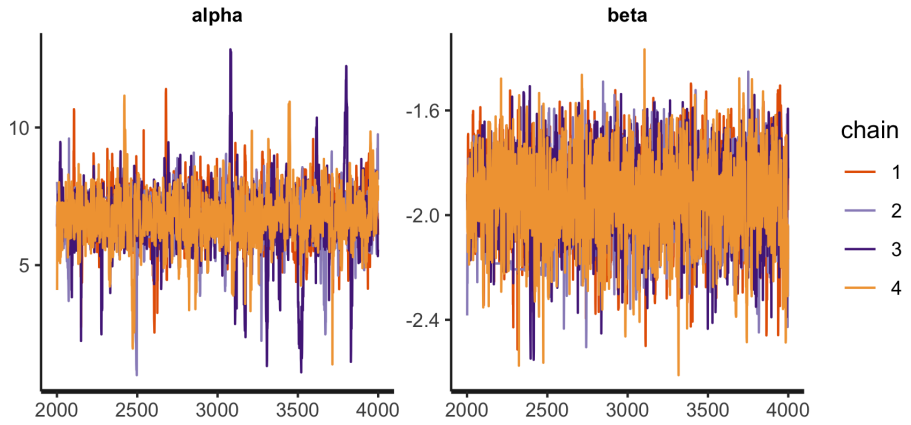


Figure 5.1: Traceplot de la convergence des chaînes MCMC pour α et β.

la figure 5.2, les chaînes semblent se mélanger rapidement, ce qui est un bon indicateur de la convergence. De même pour les paramètres γ : Les chainent semblent également

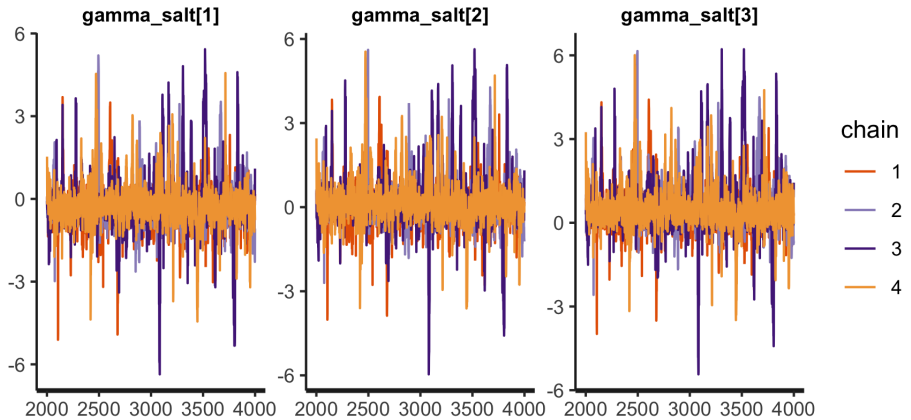


Figure 5.2: Traceplot de la convergence des chaînes MCMC pour les γ.

bien mélangées pour nos *gamma*, les paramètres les plus pertinants pour nos attentes, les chaines ont convergé rapidement.

5.2.2 Rhat et Effective Sample Size (n_eff)

Le facteur $\hat{R}$ est un diagnostic de convergence qui compare la variance intra-chaînes à la variance inter-chaînes. Un $\hat{R}$ proche de 1 indique que les chaînes ont convergé. Nous avons considéré un seuil de $\hat{R} \leq 1.1$ comme critère de convergence acceptable.

Dans notre modèle, les valeurs de **Rhat** pour les paramètres principaux sont les suivantes :

Paramètre	**Rhat**
α	1.0038
β	1.0019
σ_γ	1.0056
$\gamma_{\text{salt}[1]}$	1
$\gamma_{\text{salt}[2]}$	1
$\gamma_{\text{salt}[3]}$	1.01

Les valeurs de **Rhat** pour tous les paramètres sont très proches de 1 ou égale à 1 ce qui indique que les chaînes MCMC ont convergé de manière satisfaisante. Aucune valeur n'est supérieure à 1.1, ce qui est conforme aux critères de convergence recommandés dans la littérature.

Ainsi, les résultats de ce diagnostic montrent que l'estimation des paramètres de notre modèle est stable et fiable.

Les résultats de n_{eff} pour les différents paramètres sont présentés dans le tableau ci-dessous : Toutes les valeurs de n_{eff} sont bien supérieures à notre seuil de convergence

Paramètre	**n_eff**
α	1374
β	3002
σ_γ	1504
$\gamma_{\text{salt}[1]}$	1207
$\gamma_{\text{salt}[2]}$	1210
$\gamma_{\text{salt}[3]}$	1153

Table 5.1: Valeurs de la taille d'échantillon effective (n_{eff}) pour chaque paramètre.

de 1000, ce qui indique que l'échantillonnage a été suffisamment exploratoire et que les chaînes ont convergé. Ces valeurs élevées de n_{eff} garantissent que les estimations des paramètres sont fiables et que les résultats de l'échantillonnage ne sont pas biaisés par des autocorrélations élevées ou des dépendances entre les échantillons.

Il est également à noter que les valeurs de *Rhat* pour tous les paramètres sont égales à 1, ce qui confirme que les chaînes MCMC ont convergé de manière optimale.

Ces diagnostics confirment que l'échantillonnage MCMC a correctement exploré l'espace des paramètres et que les résultats peuvent être utilisés pour des analyses et des inférences statistiques supplémentaires comme sur un jeu de données simulées.

5.3 Validation du modèle

5.3.1 Vérification avec des données simulées (Fake Data Check)

La vérification des données factices ou Fake Data Check est une méthode utilisée pour évaluer si le modèle statistique implémenté est capable de bien reproduire les caractéristiques des données observées. Cette approche consiste à générer des données simulées en utilisant les paramètres du modèle, ici la regression logistique puis à ajuster de nouveau le modèle à ces données factices. Si le modèle est correctement spécifié et convergent, les résultats obtenus avec les données factices devraient être cohérents avec les paramètres du modèle original.

Procédure de vérification des données factices

Nous avons simulé des données factices à partir du modèle logistique avec les paramètres estimés dans notre analyse précédente. Les étapes de cette vérification sont les suivantes :

1. **Simulation des données :** En utilisant les valeurs estimées des paramètres α, β, γ_1, γ_2, et γ_3, nous avons généré des données factices en utilisant la même procédure que celle utilisée pour l'ajustement du modèle. La probabilité de survie pour chaque observation a été calculée en fonction des concentrations de sel et des types de sels simulés. Quant à σ nous avons pris le choix de garder cet effet aléatoire sur le type de sel.

2. **Ajustement du modèle :** Nous avons ensuite ajusté à nouveau notre modèle logistique aux données factices. Ce processus implique l'estimation des paramètres à partir des nouvelles données générées, en utilisant encore une fois l'algorithme MCMC de STAN.

3. **Comparaison des résultats :** Une fois que le modèle a été ajusté aux données factices, nous avons comparé les valeurs des paramètres estimés avec celles obtenues dans l'analyse précédente. Si les résultats sont cohérents, cela confirme que le modèle a bien capturé la structure sous-jacente des données.

Résultats de la vérification des données factices

Afin de savoir si le dataset généré était réaliste, j'ai comparé le graphique de la probabilité de survie en fonction de la concentration de sel avec celui de la figure **??**. Le graphique est représenté ci-dessous :

Les courbes semblent similaires, on observe bien une diminution de la probabilité de survie avec l'augmentation logarithmique de la concentration de sel.

Les résultats de l'estimation des paramètres du modèle de régression logistique bayésienne sont présentés dans le tableau suivant. Les estimations montrent une probabilité de survie qui diminue à mesure que la concentration de sel augmente. Les intervalles de crédibilité à 95% sont également indiqués pour chaque paramètre.

- **Paramètres globaux** : L'ordonnée à l'origine ($\alpha = 6.71$) indique une forte probabilité de survie à faible concentration de sel. Le coefficient $\beta = -1.92$ confirme une diminution de la survie avec l'augmentation de la concentration.

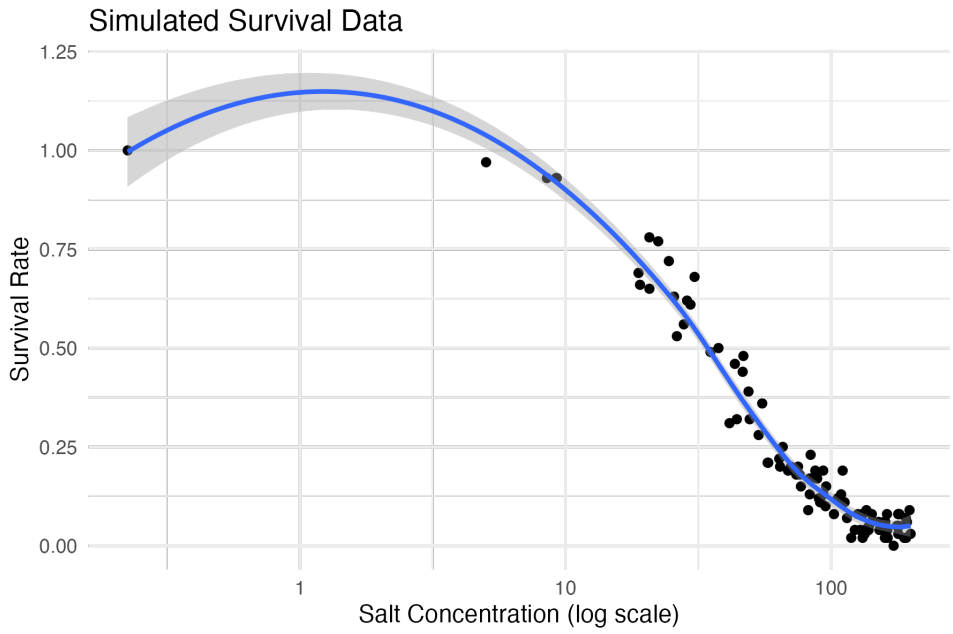


Figure 5.3: Probabilité de survie en fonction de la concentration de sel.

Paramètre	**Moyenne**	**Écart-type**
α	6.71	0.51
β	-1.92	0.04
σ_γ	0.63	0.74
$\gamma_{\text{salt}[1]}$	-1.4	0.48
$\gamma_{\text{salt}[2]}$	-0.61	0.48
$\gamma_{\text{salt}[3]}$	0.29	0.48

Table 5.2: Estimation des paramètres du modèle avec leur moyenne et écart-type.

- **Effets des sels** : Le sel 1 ($\gamma_{\text{salt}[1]} = -1.4$) a l'impact le plus négatif sur la survie, suivi du sel 2 ($\gamma_{\text{salt}[2]} = -0.61$). En revanche, le sel 3 ($\gamma_{\text{salt}[3]} = 0.29$) semble avoir un effet moins nocif, voire légèrement bénéfique.

Le paramètre $\gamma_{\text{salt}[1]}$ est estimé à -0.20 avec un écart-type de 0.48. Cela signifie que la présence de NaCl, en moyenne, diminue la probabilité de survie des mayflies.

Conclusion

La vérification des données factices est un test essentiel pour s'assurer que le modèle n'est pas simplement ajusté aux données observées de manière biaisée, mais qu'il capture bien la dynamique sous-jacente. Les résultats de cette vérification montrent que notre modèle logistique fonctionne correctement et que les paramètres estimés sont cohérents, ce qui renforce la confiance dans la validité de nos conclusions sur l'impact des concentrations de sel et des types de sels sur la survie des mayflies.

5.4 Le sel commercial est il nécessaire dans le modèle?

5.5 Résultats et discussion

L'analyse des paramètres γ_{salt} permet d'évaluer l'effet différentiel des types de sel sur la survie des larves. Le paramètre $\gamma_{\text{salt}[1]} = -1.4$ étant le plus négatif, il indique que ce sel est le plus toxique pour les larves. En comparaison, $\gamma_{\text{salt}[2]} = -0.61$ a un effet moins marqué, et $\gamma_{\text{salt}[3]} = 0.29$ semble être le moins toxique, voire légèrement bénéfique par rapport aux autres sels.

Ces résultats suggèrent que la composition chimique spécifique de chaque sel influence la survie des larves. Le sel NaCl, étant le plus toxique, pourrait contenir des ions ayant un effet particulièrement délétère sur les organismes aquatiques, ensuite le sel CaCl2 est légèrement moins toxique, comme le montre les résultats. En revanche, le sel commercial semble être le moins nocif.

Chapter 6

Régression de Poisson

6.1 Description du modèle

Le choix d'un de la régression de Poisson est justifié par la nature de la variable du nombre de survivants N_{surv}, cible notre étude. Cette variable N_{surv} est discrète et positive et le modèle de Poisson a comme support N. C'est donc un choix naturel pour la modéliser car il est conçu pour des variables aléatoires de ce type qui prennent des valeurs entières positives. On va alors supposer :

$$N_{surv}[i] \sim \mathcal{P}(\lambda_i)$$

et le modèle s'écrit

$$\log(\lambda_i) = \alpha_{salt[i]} - \beta_{salt[i]} \times \log(conc[i])$$

où

- λ_i est l'espérance du nombre de survivants pour l'observation i,
- α et β sont des effets spécifiques à chaque types de sel (effets hiérarchiques).

6.2 Justification des priors

6.2.1 Sur l'intercept α

L'intercept représente le log du nombre attendu de survivants quand la concentration est très faible. On a $\lambda = \exp(\alpha)$, donc

- Si $\alpha = 0$, alors $\lambda = 1$ survivant,
- Si $\alpha = 3$, alors $\lambda \approx 20$ survivants,
- Si $\alpha = 5$, alors $\lambda \approx 20$,
- Si $\alpha = -3$, alors $\lambda \approx 0.05$.

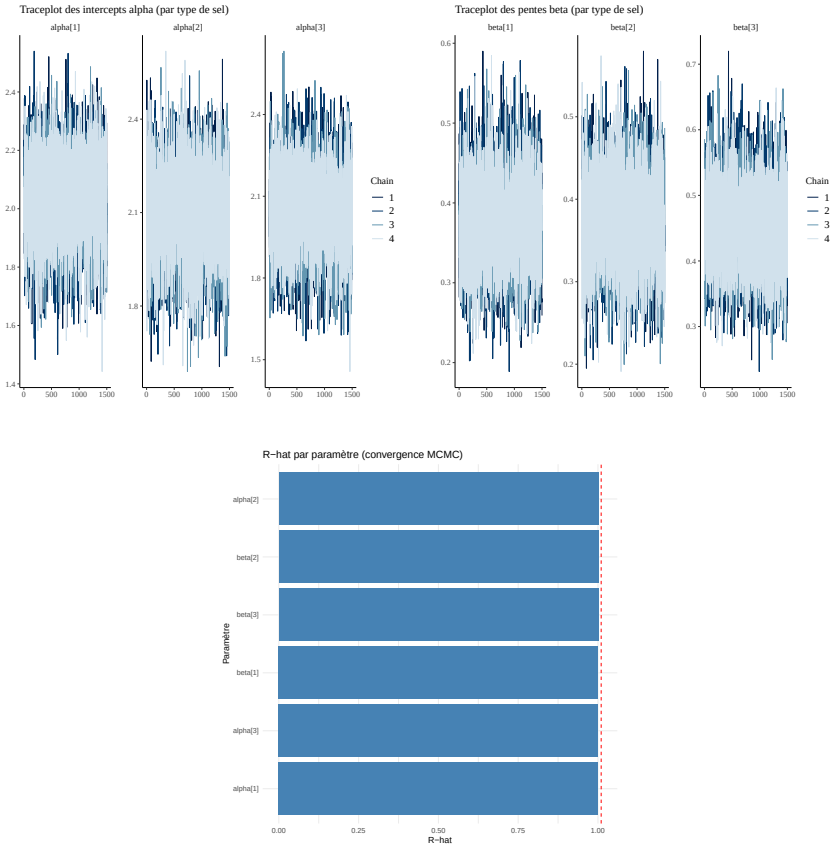
Un prior $\mathcal{N}(0, 3)$ est raisonnable puisque les valeurs de $N0$ vont jusqu'à 9-10.

6.2.2 Sur la pente β (effet de la concentration)

Le paramètre β contrôle à quel point la survie diminue quand la concentration augmente. Si $\beta > 0$, alors la survie diminue avec la concentration, comme attendu. Le prior $\mathcal{N}(0, 5)$ permet :

- des pentes proches de 0 (quasiment aucune sensibilité),
- des pentes modérées ($\beta \approx 1 - 2$),
- des pentes élevées.
- Si $\alpha = -3$, alors $\lambda \approx 0.05$.

6.3 Convergence MCMC



Le graphique en bâtons présente les valeurs de R-hat (ou potentiel d'amélioration de la convergence) pour chaque paramètre du modèle, en particulier les coefficients $alpha[k]$ et $beta[k]$ associés aux différents types de sels.

- chaque barre représente la valeur de R-hat pour un paramètre donné,
- la ligne rouge pointillée à $R = 1.01$ indique le seuil au-delà duquel on peut suspecter un problème de convergence,
- Tous les paramètres ont des valeurs de R-hat proches de 1, ce qui indique que les chaînes MCMC se sont bien mélangées et ont convergé vers la distribution postérieure cible.

- l'absence de barres au-dessus du seuil de 1.01 confirme que l'échantillonnage est fiable pour ces paramètres.

La convergence des chaînes MCMC est satisfaisante, ce qui permet d'interpréter les estimations postérieures (moyennes, intervalles de crédibilité) en toute confiance.

Chapter 7

Modèle Probit

7.1 Structure du modèle Probit

Cette section présente une analyse bayésienne en utilisant différents modèles de régression logistique lineaires et des fonctions de lien. Les données ont été analysées à l'aide de PyMC, une bibliothèque Python pour l'inférence bayésienne. Le sampler HMC utilise la dynamique hamiltonienne pour explorer l'espace d'échantillonnage. En définissant une variable d'*impulsion* p et une variable de *position* q, il résout les équations différentielles de Hamilton :

$$\frac{dq}{dt} = \frac{\partial H}{\partial p}, \quad \frac{dp}{dt} = -\frac{\partial H}{\partial q} \tag{7.1}$$

Cela permet d'explorer efficacement l'espace en évitant les mouvements lents. Le Gibbs sampler repose sur l'échantillonnage séquentiel de chaque variable conditionnellement aux autres, basé sur la formule :

$$X_i^{(t+1)} \sim P(X_i \mid X_{-i}) \tag{7.2}$$

Il est simple à mettre en œuvre mais peut souffrir de convergence lente en présence de fortes corrélations. Puisque nous travaillons dans le cadre de Python pour cette section, nous avons utilisé HMC pour sa rapidité et son efficacité, en particulier pour des distributions de grande dimension avec des corrélations complexes.

- **Priors** :
  - $\beta_0 \sim \text{Normal}(\mu = 0, \sigma = 10)$
  - $\beta_1 \sim \text{Normal}(\mu = 0, \sigma = 10)$
- **Likelihood** : Distribution Binomiale
- **Sampler** : Hamiltonian Monte Carlo (HMC) avec 5000 échantillons et 1000 échantillons de tuning (warm-up samples).
- **Fonction de Lien** : Logit (sigmoïde)

7.1.1 Logit (Sigmoïde)

$$\text{logit}(p) = \beta_0 + \beta_1 \cdot \log(\text{conc})$$

La fonction de lien logit transforme la combinaison linéaire des prédicteurs en une probabilité via la fonction sigmoïde. Elle est symétrique et produit des probabilités comprises

entre 0 et 1. Idéale pour modéliser des données binaires où l'on souhaite interpréter les coefficients en termes de log-odds.

7.1.2 Probit (CDF Normale)

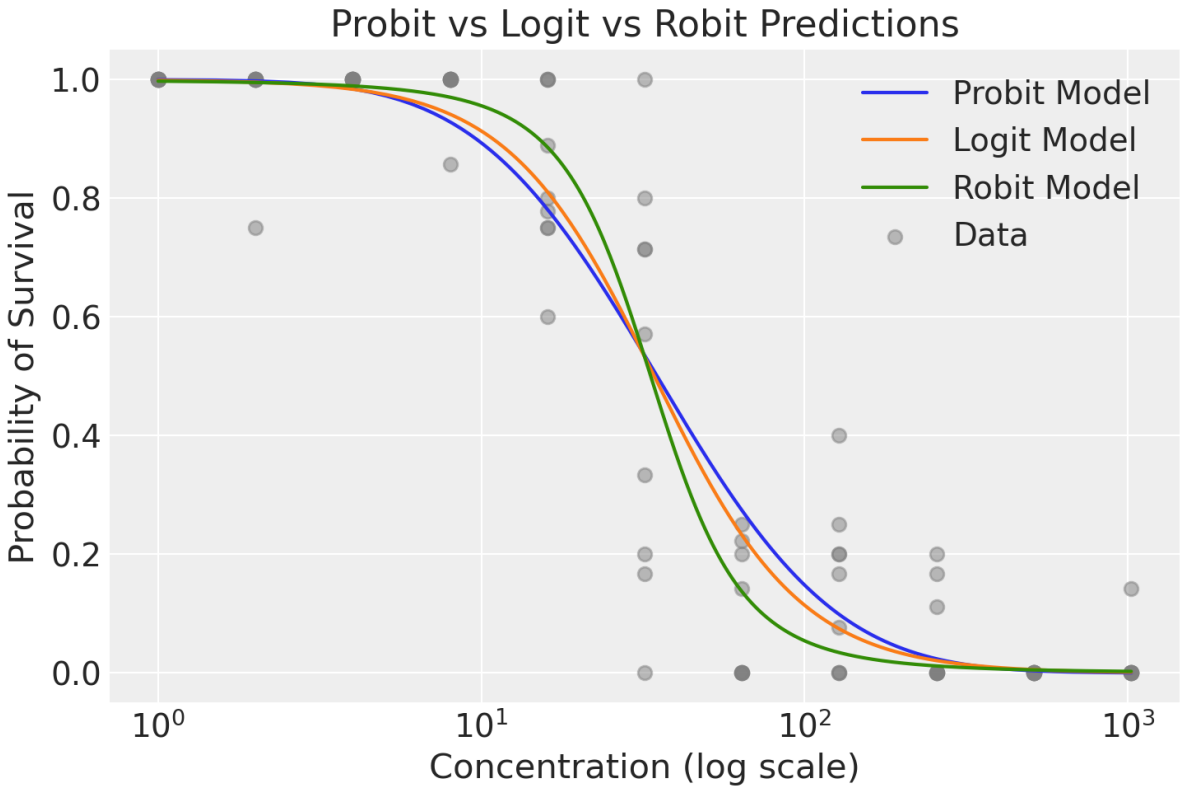
$$\text{probit}(p) = \beta_0 + \beta_1 \cdot \log(\text{conc})$$

La fonction de lien robit utilise la CDF de la distribution de Student-t, qui a des queues plus épaisses que la distribution normale. Elle est plus robuste aux valeurs aberrantes. Utile lorsque les données contiennent des points extrêmes ou lorsque l'on souhaite réduire l'influence des valeurs aberrantes sur les estimations. La distribution de Student-t est heavy-tailed, ce qui signifie qu'elle accorde une probabilité plus élevée aux valeurs extrêmes par rapport à la distribution normale. Cela permet au modèle robit d'être plus robuste aux outliers, en réduisant leur influence sur les estimations par rapport aux modèles utilisant des liens basés sur la distribution normale.

7.1.3 Robit (CDF de Student-t)

$$\text{robit}(p) = \beta_0 + \beta_1 \cdot \log(\text{conc})$$

La fonction de lien robit utilise la CDF de la distribution de Student-t, qui a des queues plus épaisses que la distribution normale. Elle est plus robuste aux valeurs aberrantes. Utile lorsque les données contiennent des points extrêmes ou lorsque l'on souhaite réduire l'influence des valeurs aberrantes sur les estimations.



La courbe plus abrupte du modèle robit suggère qu'il est plus sensible aux variations de concentration dans la plage moyenne. Une légère augmentation de la concentration entraîne une baisse plus importante de la probabilité de survie par rapport aux autres

modèles. Cela pourrait indiquer que le modèle robit capture une relation dose-réponse sous-jacente plus brutale, possiblement due au fait que les données présentent une transition plus nette ou une certaine hétérogénéité que la distribution t de Student (avec un petit nu) peut mieux modéliser.

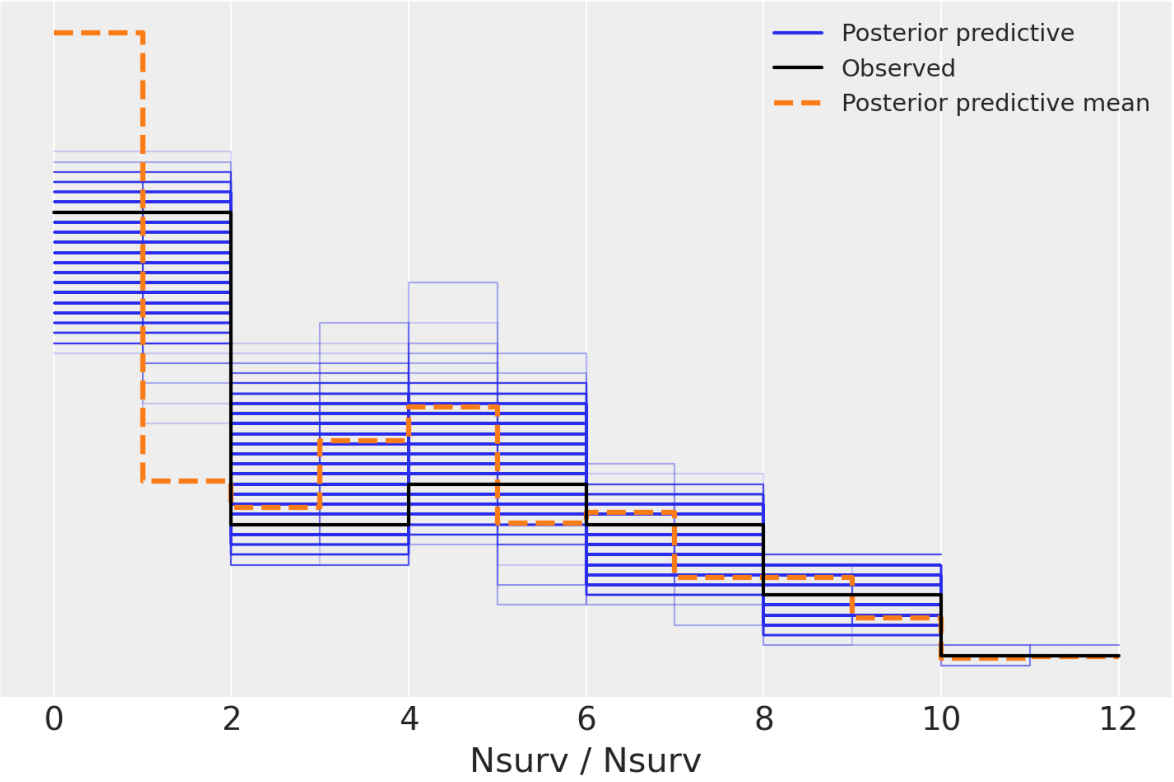
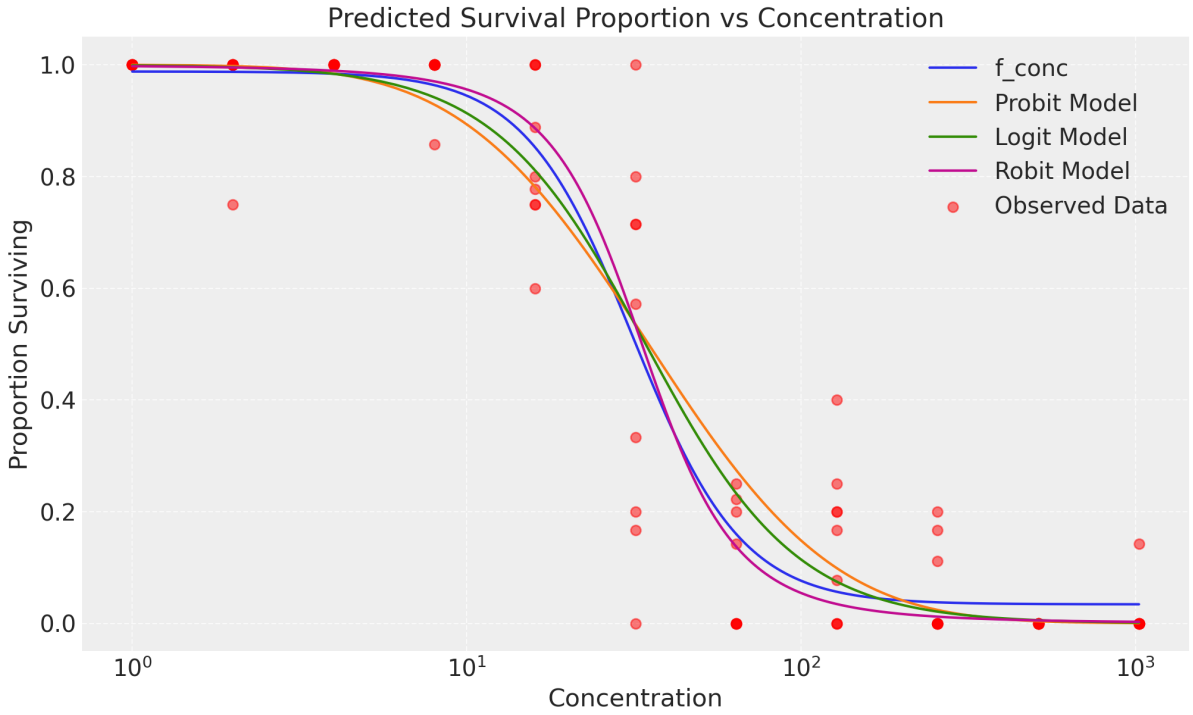


Figure 7.1: posterior predictive checks for Probit model

Le graphique ci-dessus représente les prédictions a posteriori obtenues avec le modèle probit. On observe que les données observées (ligne noire) se situent globalement dans l'intervalle des prédictions a posteriori (en bleu), bien que les prédictions moyennes (pointillés orange) soient moins cohérentes avec les observations au début. Cela suggère que le modèle est relativement bien calibré. Des écarts significatifs pourraient indiquer que le modèle ne capture pas adéquatement la structure des données



7.1.4 Comparaison de la performance du modèle proposé f_{conc} avec d'autres fonctions de lien

La probabilité de survie p est modélisée par une fonction log-logistique :

$$p = \frac{d - c}{1 + \left(\frac{\text{conc}}{e}\right)^b} + c$$

où conc représente les valeurs de concentration. Les priors utilisés dans le modèle proposé :

$$\begin{aligned} d &\sim \text{Beta}(\alpha = 2, \beta = 1) \\ c &\sim \text{Beta}(\alpha = 1, \beta = 2) \\ b &\sim \text{HalfNormal}(\sigma = 2) \\ e &\sim \text{LogNormal}(\mu = 0, \sigma = 2) \end{aligned}$$

avec la probabilité de survie p est utilisée dans une distribution binomiale pour modéliser le nombre de survivants Nsurv parmi le nombre initial N0

Les choix des paramètres pour la fonction de lien f_{conc} dans le modèle log-logistique ont été soigneusement sélectionnés pour capturer la complexité de la relation entre la concentration de sel et la proportion de survie des éphémères. Le paramètre d, issu d'une distribution Beta avec $\alpha = 2$ et $\beta = 1$, permet à la fonction de commencer à un niveau relativement élevé, reflétant une proportion de survie initiale plus haute. En revanche, c, avec $\alpha = 1$ et $\beta = 2$, assure que la fonction se termine à un niveau plus bas, modélisant ainsi une diminution significative de la survie à des concentrations élevées. Le paramètre b, suivant une distribution HalfNormal avec $\sigma = 2$, contrôle la pente de la transition, permettant des ajustements allant de transitions douces à abruptes. Enfin, e, issu d'une distribution LogNormal avec $\mu = 0$ et $\sigma = 2$, offre une flexibilité dans

l'échelle de concentration, permettant au modèle de s'adapter à différentes gammes de données. Ces choix permettent au modèle log-logistique de capturer une variété de formes de réponse, offrant ainsi une modélisation robuste et adaptable des données observées.

7.1.5 Interprétation des distributions postérieures et diagnostics de convergence

Les distributions postérieures obtenues après échantillonnage permettent d'estimer les paramètres du modèle tout en quantifiant leur incertitude. Pour évaluer la convergence des chaînes, nous utilisons les diagnostics comme le facteur de réduction de Gelman-Rubin ($\hat{\mathrm{R}}$) et l'inspection visuelle des trace plots. Un $\hat{\mathrm{R}}$ proche de 1 indique une convergence satisfaisante. Les trace plots permettent de vérifier l'exploration adéquate de l'espace paramétrique et l'absence d'autocorrélation excessive. Des échantillons bien mélangés et une convergence satisfaisante augmentent la confiance dans les inférences tirées des distributions postérieures.

Chapter 8

Conclusion

- Régression logistique / probit : adaptées si la réponse est binaire (survie oui/non) par individu.
- Poisson : idéal si on observe un nombre de survivants sur un total (et éventuellement avec effet de groupe).
- Log-logistique : excellent pour estimer LC50 et visualiser des courbes dose-réponse biologiques.
- Weibull : pertinent si on modélise le temps de survie, pas juste l'issue binaire à un temps fixe.

Bibliography

[1] A. Gelman, J. Hill, and A. Vehtari. *Regression and Other Stories*. Cambridge University Press, 2020.

[2] A. Gelman and C. R. Shalizi. "Philosophy and the practice of Bayesian statistics". In: *British Journal of Mathematical and Statistical Psychology* 66.1 (2013), pp. 8–38.

[3] P. D. Hoff. *A First Course in Bayesian Statistical Methods*. Springer, 2009.

Annexes

8.1 Implémentation du modèle log-logistique en JAGS/runjags

L'implémentation du modèle log-logistique dans `JAGS/runjags` est présentée ci-dessous. Ce modèle permet d'estimer les paramètres associés à la survie des larves en fonction de la concentration en sel.

```
data_list <- list(
  N = nrow(data),                          # Nombre d'observations
  N0 = data$N0,                            # Nombre initial de larves
  Nsurv = data$Nsurv,                      # Nombre de survivants
  conc = data$conc,                        # Concentration de sel
  salt = as.numeric(factor(data$Salt)),    # Indices pour les sels (1: CaCl2,
  n_salt = length(unique(data$Salt)) # Nombre de types de sels (3)
)

model_string <- "
model{
  for(i in 1:N){
    Nsurv[i] ~ dbin(p[i], N0[i])
    p[i] <- (d[salt[i]] - c[salt[i]]) / (1 + pow(conc[i] / theta[salt[i]],
  }
  for(s in 1:n_salt){
    d[s] ~ dbeta(5, 1)
    c[s] ~ dbeta(1, 5)
    b[s] ~ dgamma(2, 1)
    theta[s] ~ dlnorm(4, 1)
  }
}
"
# Exécution du modèle avec run.jags
jags_model <- run.jags(
  model = model_string,
  data = data_list,
  monitor = c("d", "c", "b", "theta"),
  n.chains = 3,
  adapt = 1000,
  burnin = 2000,
  sample = 10000,
```

```
  thin = 10
)

# Résumé des résultats du modèle
summary(jags_model) # Résumé des résultats
```

8.2 Implémentation du modèle de Weibull en JAGS/runjags

L'implémentation du modèle de Weibull en `JAGS/runjags` est présentée ci-dessous. Ce modèle permet d'estimer les paramètres bayésiens pour évaluer la toxicité des différents sels.

```
# Convertir Salt en facteur numérique
data$salt <- as.numeric(factor(data$Salt, levels = c("NaCl", "CaCl2", "Com

# Liste des données pour JAGS
data_list <- list(
  N = nrow(data),
  K = 3, # Nombre de types de sels (NaCl, CaCl2, Commercial Salt)
  n = data$Nsurv, # Nombre de larves survivantes
  total = data$N0, # Nombre initial de larves
  conc = data$conc, # Concentration de sel
  salt = data$salt # Indices des sels
)

# Modèle JAGS avec priors ajustés
weibull_model_jags <- "
model{
  for(i in 1:N){
    n[i] ~ dbin(p[i], total[i])
    p[i] <- exp(-pow(conc[i] / alpha[salt[i]], beta[salt[i]]))
  }
  for(k in 1:K){
    log_alpha[k] ~ dnorm(mu_alpha, tau_alpha)
    log_beta[k] ~ dnorm(mu_beta, tau_beta)
    alpha[k] <- exp(log_alpha[k])
    beta[k] <- exp(log_beta[k])
    CL50[k] <- alpha[k] * pow(log(2), 1 / beta[k]) # Calcul de CL50
  }
  mu_alpha ~ dnorm(0, 0.01) # Moyenne de alpha
  tau_alpha <- 1 / (sigma_alpha * sigma_alpha) # Variance de alpha
  sigma_alpha ~ dunif(0, 10) # Prior moins contraignant sur alpha
  mu_beta ~ dnorm(0, 0.04) # Moyenne de beta
  tau_beta <- 1 / (sigma_beta * sigma_beta) # Variance de beta
  sigma_beta ~ dunif(0, 10) # Prior moins contraignant sur beta
}
```

```
"

# Paramètres à surveiller
parameters <- c("alpha", "beta", "CL50", "mu_alpha", "sigma_alpha", "mu_bet

# Initialisations ajustées
inits <- list(
  list(mu_alpha = 4, sigma_alpha = 1, mu_beta = 0.5, sigma_beta = 0.5,
       log_alpha = c(4, 4.5, 4), log_beta = c(0.5, 0.6, 0.7)),  # NaCl, Ca
  list(mu_alpha = 5, sigma_alpha = 2, mu_beta = 1, sigma_beta = 1,
       log_alpha = c(4.5, 5, 4.5), log_beta = c(0.7, 0.8, 0.6))
)

# Exécution du modèle avec run.jags
results <- run.jags(
  model = weibull_model_jags,
  data = data_list,
  monitor = parameters,
  n.chains = 2,
  inits = inits,
  burnin = 2000,  # Augmenter pour plus de stabilité
  sample = 5000,
  adapt = 2000,   # Plus d'itérations d'adaptation
  thin = 2,
  method = "rjags",
  summarise = TRUE
)

# Résumé des résultats
print(summary(results))
plot(results, vars = "CL50", plot.type = "histogram")
```

L'utilisation de ce modèle en JAGS permet d'estimer les **paramètres bayésiens** et de comparer les concentrations létales médianes (`CL50`) des trois types de sel (NaCl, CaCl2 et Commercial Salt).

8.3 Regression Logistique

8.3.1 Modèle Stan

```
library(StanHeaders)
library(rstan)

df_model = list(
  N0 = df$N0,
  Nsurv = df$Nsurv,
  conc = df$conc,
  NaCl = df$NaCl,
```

```
  CaCl2 = df$CaCl2,
  Commercial = df$Commercial,
  length_N0 = length(df$N0)
)

model_code = "
data{
int<lower=1> length_N0;
int<lower=0, upper=1> NaCl[length_N0]; // Indicateur NaCl
int<lower=0, upper=1> CaCl2[length_N0]; // Indicateur CaCl2
int<lower=0, upper=1> Commercial[length_N0]; // Indicateur CommercialSalt
int N0[length_N0];
int Nsurv[length_N0];
real conc[length_N0];
}

parameters {
real alpha;
real beta;
real<lower=0> sigma_gamma; // Ecart type des effets de sel
vector[3] gamma_salt; // Vecteur pour les effets de chaque type de sel
}

model{
alpha ~ lognormal(log(6), 1);
beta ~ normal(0, 5);
sigma_gamma ~ cauchy(0, 5); // Prior positif sur la dispersion
gamma_salt ~ normal(-3, sigma_gamma); // Effets al atoires des sels

for(i in 1:length_N0) {
real eta = alpha + beta * log(conc[i])
                    + gamma_salt[1] * NaCl[i] // Effet de NaCl
                    + gamma_salt[2] * CaCl2[i] // Effet de CaCl2
                    + gamma_salt[3] * Commercial[i]; // Effet de Commerci
Nsurv[i] ~ binomial_logit(N0[i], eta);
}
}
"
```

8.3.2 MCMC

```
fit <- stan(model_code = model_code,
            data = df_model,
            chains = 4,
            iter = 8000,
            warmup = 4000,
            cores = 4)
```

```r
print(fit)

library(bayesplot)
mcmc_hist(fit, pars = c("alpha", "beta", "gamma_salt[1]", "gamma_salt[2]",
```

8.3.3 Convergence MCMC

```r
png("traceplot_gamma.png", width = 6, height = 3, units = "in", res = 300)
traceplot(fit, pars = c("alpha", "beta"))  # Afficher le graphique dans le
dev.off()

png("traceplot.png", width = 6, height = 3, units = "in", res = 300)
traceplot(fit, pars = c("alpha", "beta"))  # Afficher le graphique dans le
dev.off()

summary(fit)$summary[, "Rhat"]
```

8.3.4 Vérification des données simulées

```r
set.seed(123)
n_sim = 100
alpha_sim = 7.10
beta_sim = -1.93
gamma_salt_sim = c(-0.62, -0.24, 0.01)
sigma_gamma_sim = 0.97

conc_sim = runif(n_sim, min = 0.1, max = 200)
Salt_type_sim = sample(c("NaCl", "CaCl2", "Commercial"), n_sim, replace = T

NaCl_sim = ifelse(Salt_type_sim == "NaCl", 1, 0)
CaCl2_sim = ifelse(Salt_type_sim == "CaCl2", 1, 0)
Commercial_sim = ifelse(Salt_type_sim == "Commercial", 1, 0)

eta_sim = alpha_sim + beta_sim * log(conc_sim) +
          gamma_salt_sim[1] * NaCl_sim +
          gamma_salt_sim[2] * CaCl2_sim +
          gamma_salt_sim[3] * Commercial_sim

p_survival_sim = 1 / (1 + exp(-eta_sim))
N0_sim = rep(100, n_sim)
Nsurv_sim = rbinom(n_sim, N0_sim, p_survival_sim)

simulated_data = data.frame(N0 = N0_sim, Nsurv = Nsurv_sim, conc = conc_sim
                            NaCl = NaCl_sim, CaCl2 = CaCl2_sim, Commercial

ggplot(simulated_data, aes(x = conc, y = Nsurv / N0)) +
  geom_point() +
```

```
  geom_smooth(method = "loess") +
  scale_x_log10() +
  theme_minimal() +
  labs(title = "Simulated Survival Data", x = "Salt Concentration (log scal

ggsave("simulated_survival_plot.png", width = 6, height = 4, dpi = 300)

simulated_data_model <- list(
  N0 = simulated_data$N0,
  Nsurv = simulated_data$Nsurv,
  conc = simulated_data$conc,
  NaCl = simulated_data$NaCl,
  CaCl2 = simulated_data$CaCl2,
  Commercial = simulated_data$Commercial,
  length_N0 = n_sim
)

fit_fake_logistic <- stan(
  model_code = model_code,  # Ton mod le Stan
  data = simulated_data_model,
  chains = 4,  # Nombre de cha nes MCMC
  iter = 8000,  # Nombre total d'it rations
  warmup = 4000,  # Nombre d'it rations de burn-in
  cores = 4  # Nombre de c urs pour les calculs parall les
)

print(fit_fake_logistic)

traceplot(fit_fake_logistic, pars = c("alpha", "beta", "gamma_salt[1]", "ga
```

8.3.5 Résumé des résultats du modèle

```
summary(fit)$summary[,"mean"]
```

Fin du rapport