Projet HAB904B

Basile Pajot (DARWIN), Marion Themeze–Leroy(ECOSYSTEMES),

2023-12-20

Contents

1 Lecture et exploration des données **1**

1.1 La variable à expliquer 1

1.2 Les variables explicatives 1

1.3 Exploration des données 2

2 Ajustement d'un modèle simple **5**

3 Comparaison de modèles **10**

4 Inférence et interprétation des résultats **11**

5 Discussion **11**

1 Lecture et exploration des données

1.1 La variable à expliquer

La variable d'intérêt à expliquer, est `Shells`, soit le nombre de carapaces de tortues récentes trouvées lors des relevés sur le terrain. Cette variable est un proxy pour estimer le nombre de tortues mortes d'une année sur l'autre.

1.2 Les variables explicatives

- `Prev` est une variable explicative qualitative qui correspond à la prévalence pour *Mycoplasma agassizii*, soit le rapport entre le nombre de tortues séropositives sur l'effectif total de tortues par année pour chaque site.
- `Site` est une variable qualitative qui correspond au site d'échantillonnage. Elle a 10 modalités : le parc national *Big Shoals (BS)*, l'aire de gestion de la faune sauvage *Camp Blanding (CB)*, l'aire de gestion de la faune sauvage et de l'environnement *Cecil Field/Branan Field (CF)*, une *propriété privé en Floride centrale (CE)*, le parc national *Fort Cooper (FC)*, l'aire de gestion de la faune sauvage *Flying Eagle (FE)*, le parc national *Gold Head Branch (GH)*, l'aire de gestion de la faune sauvage et de l'environnement *Perry Oldenburg (OL)*,la station biologique *Ordway-Swisher (OR)*, l'aire de gestion de la pêche *Tenoroc Fish (TE)*.

- `Area` est une variable quantitative qui correspond à l'aire couverte par site lors des relevés.
- `Year` est une variable qualitative qui correspond à l'année pour laquelle les relevés ont été faits. Elle a 3 modalités : 2004, 2005, 2006.

1.3 Exploration des données

Nous regardons un résumé statistique des variables de notre jeu de donnée.

```
##       Site        year              shells        type         Area
##  BS     : 3   Length:30          Min.   :0.00   Fresh:30   Min.   : 5.30
##  CB     : 3   Class :character   1st Qu.:0.00              1st Qu.:15.20
##  Cent   : 3   Mode  :character   Median :1.00              Median :27.30
##  CF     : 3                      Mean   :1.80              Mean   :29.02
##  FC     : 3                      3rd Qu.:2.75              3rd Qu.:43.20
##  FE     : 3                      Max.   :9.00              Max.   :61.00
##  (Other):12
##     density           prev        total_turtle      standprev
##  Min.   : 1.80   Min.   : 1.00   Min.   : 44.80   Min.   :-0.9163
##  1st Qu.: 2.50   1st Qu.: 1.20   1st Qu.: 74.75   1st Qu.:-0.9088
##  Median : 3.50   Median :15.95   Median :104.22   Median :-0.3559
##  Mean   : 8.76   Mean   :25.45   Mean   :119.72   Mean   : 0.0000
##  3rd Qu.: 4.80   3rd Qu.:42.05   3rd Qu.:174.90   3rd Qu.: 0.6223
##  Max.   :33.00   Max.   :80.70   Max.   :200.79   Max.   : 2.0709
##
##        H             Cov.y.1          Cov.y.2
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.4333   Mean   :0.3333   Mean   :0.3333
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
```

Nous avons un plan d'expérience équilibré avec un même nombre d'observations par site et par année.

Pour la prevalence `prev` la moyenne est supérieure à la médiane, c'est-à-dire que plus de 50% des valeurs sont inférieures à la moyenne. Il en est de même pour le nombre de carapces `shells`. De plus, pour la prévalence, la différence entre le troisème quartile et le minimum est d'environ 40, tout comme la différence entre le maximum et le 3ème quartile. Ainsi, la gamme de valeurs prise par 25% des données est égales à celle prise par 75% des données. Pour le nombre de carapaces, la gamme de valeurs prise par 25% des données est plus de trois fois supérieur à celle prise par 75% des données. Ceci est illustré par les *figures 1 et 2*.

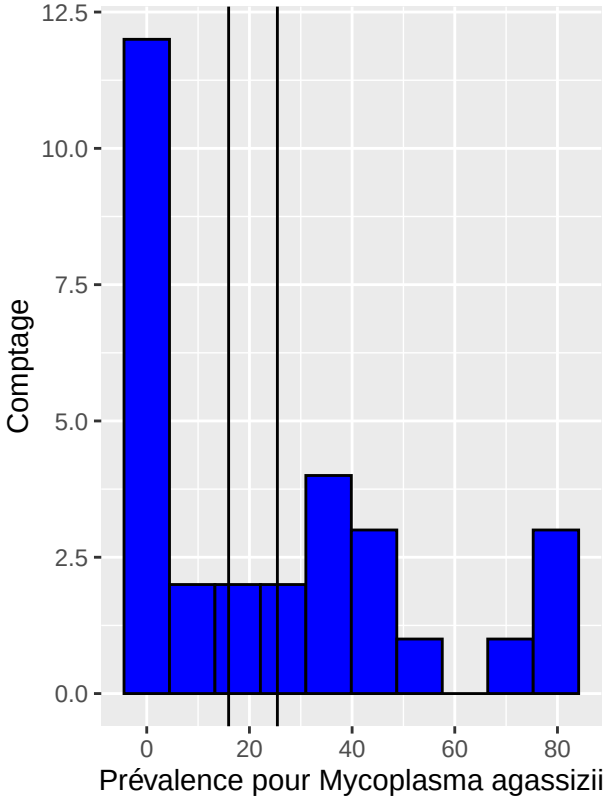


Figure 1 : Distribution de la prévalen

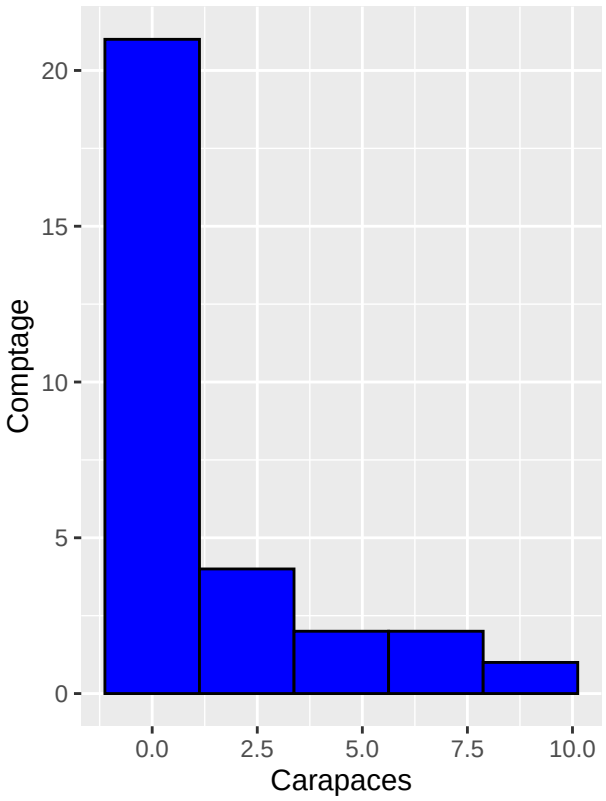


Figure 2 : Distribution du nombre de c

La distribution du nombre de carapaces ressemble à une distribution de Poisson.

La figure 3 donne plusieurs informations sur notre jeu de données.

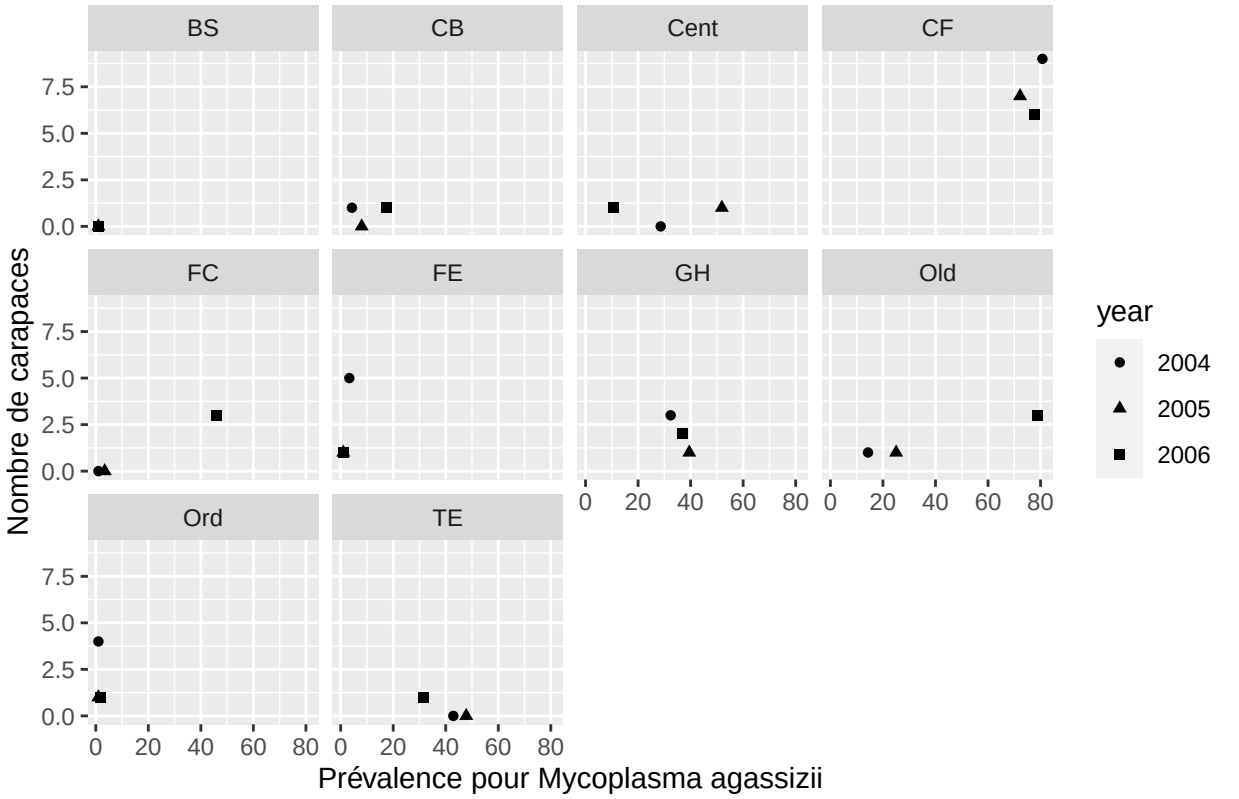


Figure 3 : Le nombre de carapaces en fonction de la prevalence par site par anné

Tout d'abord, le nombre de carapaces récentes trouvées varie ou non en fonction des années et cette variation n'est pas la même en fonction des sites. Le nombre de carapaces récentes par rapport à l'année précédente reste constante augmente ou diminue. On observe des évolutions différentes pour les sites : on observe une diminution du nombre de carapaces pour le site CF sur les trois ans ou un changement de tendance se traduisant par une diminution puis une augmentation pour le site CB.
Ensuite, la prévalence en fonction des sites peut également varier en fonction des années. Comme précédemment, cette variation n'est pas la même en fonction des sites. La prévalence reste constante, augmente ou diminue. La variation peut être globale sur les trois année d'étude (augmentation de la prévalence pour le site Old) ou changer (augmentation puis diminution pour le site Cent). Lorsque le prevalence augmente d'une année à l'autre `prev[n] < prev[n+1]`, le nombre de carapaces récentes trouvées l'année suivante augmente `shells[n+2]>shells[n+1]` (sites CB, Old), et inversement (sites CF).
Ainsi, les variations du nombre de carapaces récentes trouvées pourrait être expliquée par les variations de la prévalence.

Il apparaît également que certains sites ont de faibles prévalences (BS, Ord, FE) quelque soit l'année et d'autres des prévalences élevées (CF, GH, TE). Ceci concourt avec les observations faites précedemments avec le résumé de la variable `prev` et la distribution de la prévalence. Ainsi, nous pourrons séparer les sites en deux catégories, ceux à faible ou forte prévalence. Cette variables sera donc traitée de deux manières : de manière continue et de manière discontinue avec deux catégories : - faible prévalence (0) : $\texttt{Prev} < 0.25$ - haute prévalence (1) : $\texttt{Prev} > 0.25$

Nous allons donc essayer de déterminer si la prévalence et l'année permettent d'expliquer les variations du nombre de carapaces. Nous avons vu que la prévalence et le nombre de carapaces trouvées diffère entre les sites et entre les années. Afin de pouvoir nous concentrer sur l'effet de la prévalence, nous allons mettre un effet aléatoire sur la variable `Site`. Nous pourrions faire de même pour la variable `Année` mais par souci de simplification, nous allons garder cette variable en effet fixe.

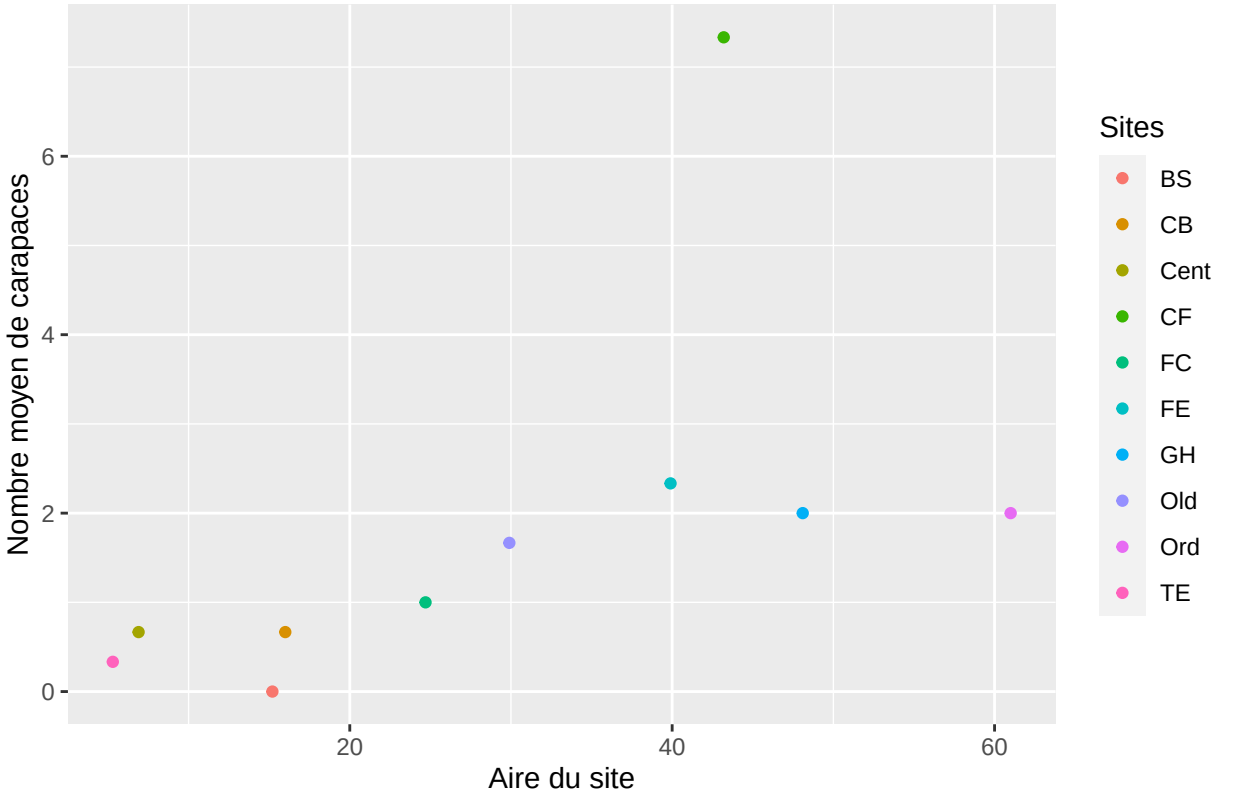


Figure 4 : Le nombre moyen de carapaces en fonction de l'aire du site d'étude

D'après la *figure 4* les sites n'ont pas tous la même aire et il semble qu'un plus grand nombre de carapces sont trouvés sur les sites avec une plus grande aire. Afin de pourvoir comparer les sites entre eux, nous allons prendre le rapport entre le nombre de carapaces trouvées par site et l'aire du site.

D'après les observations faites précedemment, nous souhaitons donc déterminer si : - le nombre de carapaces récentes trouvées est correlée avec la prevalence pour *Mycoplasma agassizii* pour une année donnée. - le nombre de carapces récentes trouvées est plus grand dans les sites à haute prévalence par rapport aux sites à basse prévalence.

2 Ajustement d'un modèle simple

Nous commençons par un modèle simple `M1` en considérant uniquement l'effet de la prévalence sur le nombre de carapces récentes trouvées.

L'équation (approche fréquentiste) du modèle linéaire simple est la manière suivante :

$$\frac{shells}{aire_{site}} = \mu_0 + \beta * prev$$

Ceci se traduit en approche bayesienne par un modèle considérant les hypothèses suivantes :

- `shells` suit un loi de poisson de paramètre λ (*cf figure 2*), c'est-à-dire que c'est une variable discrète de comptage dans un intervale de temps et un espace donnés ; avec une variance égale à la moyenne $E(shells) = V(shells) = \lambda$
- toutes les observations de `shells` sont **indépendantes**
- le logarithme de la moyenne de `shells` peut être exprimée comme la combinaaison linéaire des variables explicatives sélectionnées.

- que les paramètres à estimer (ordonnée à l'origine et coefficients de regression) suivent des lois connues, explicités ci-après.

Nous avons donc :

$$Shells_i \overset{i.i.d}{\sim} Pois(\lambda_i) \, avec \, i = 1, ..30 \, le \, nombre \, d'observations$$

$$log(\lambda_i) = log(aire_i) + \mu_0 + \beta * prev_i$$

Nous utilisons comme priors les distribution suivantes :

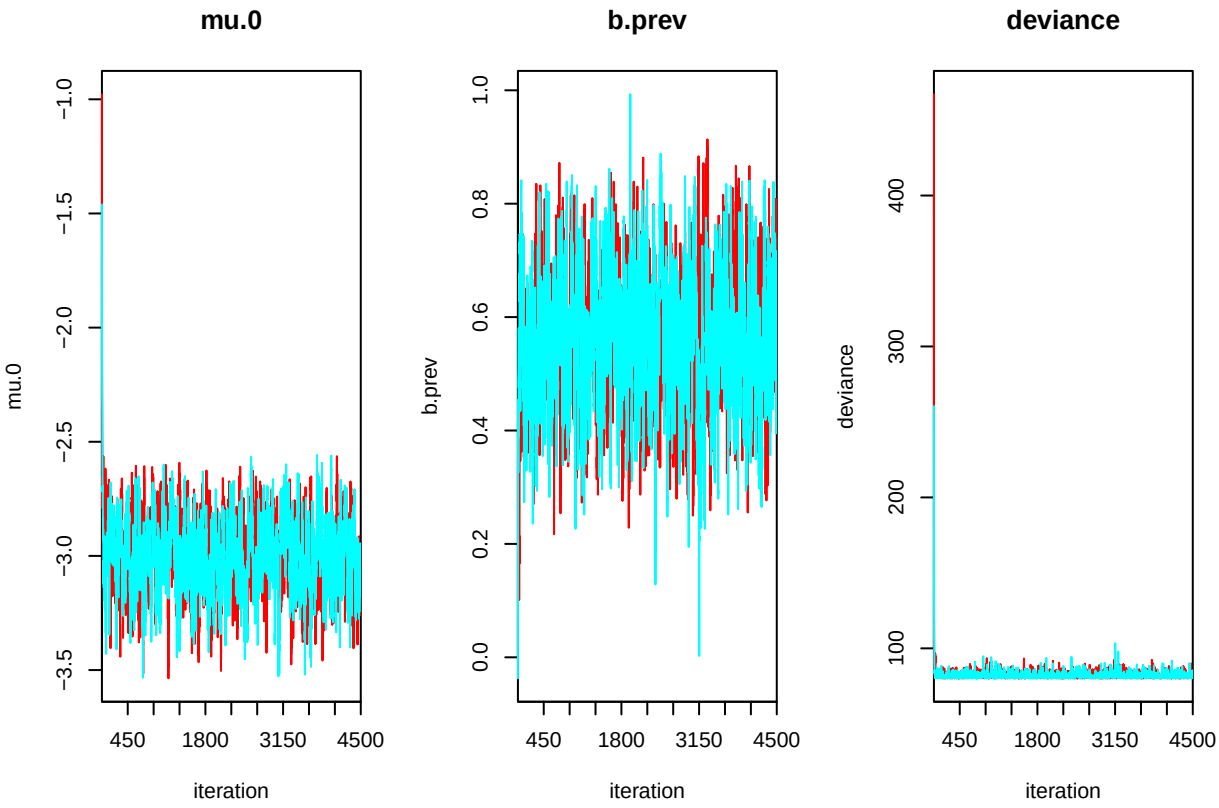
$$\mu_0 \sim \mathcal{N}(0, 100)$$
$$\beta \sim \mathcal{N}(0, 100)$$

```
## module glm loaded
```

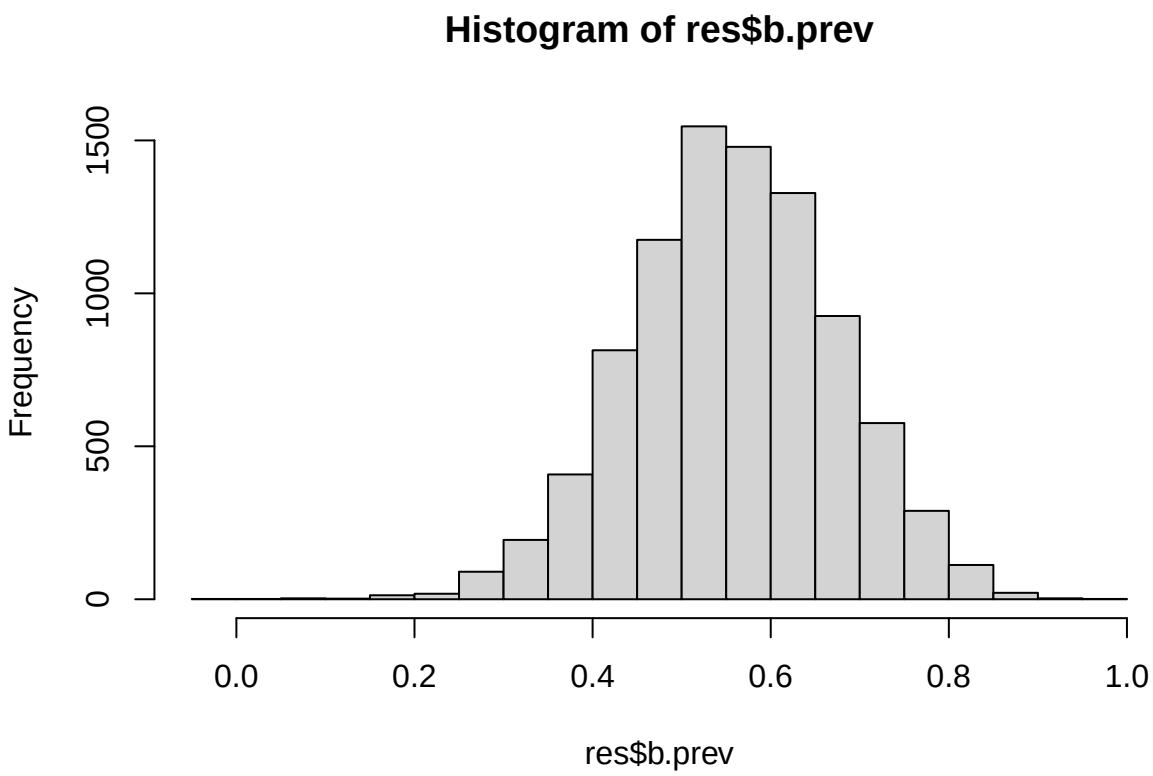
```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 30
##    Unobserved stochastic nodes: 2
##    Total graph size: 182
##
## Initializing model
```

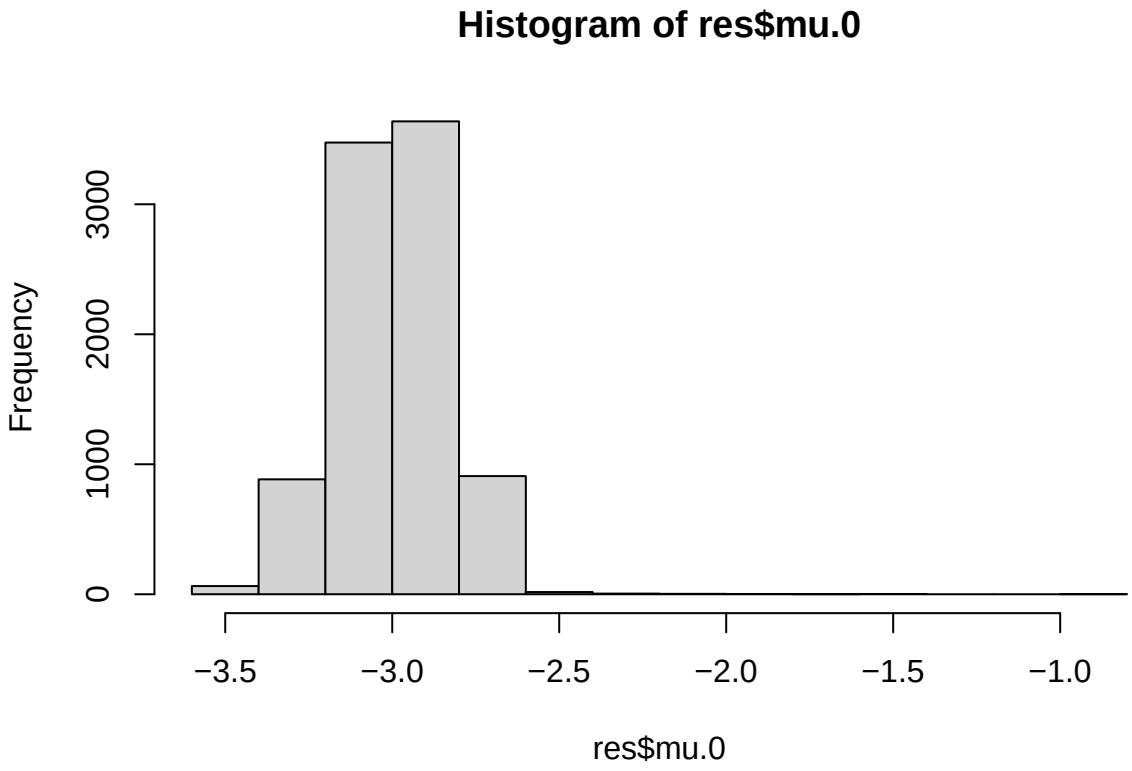
```
## Inference for Bugs model at "/var/folders/l9/dzz054qd4_q8g63vvgsts36w0000gn/T//RtmpkHp8De/model1792f
##  2 chains, each with 9000 iterations (first 4500 discarded)
##  n.sims = 9000 iterations saved
##          mu.vect sd.vect   2.5%    25%    50%    75%  97.5%  Rhat n.eff
## b.prev     0.558   0.115  0.331  0.482  0.558  0.636  0.782 1.001  8300
## mu.0      -2.997   0.164 -3.320 -3.104 -2.996 -2.888 -2.698 1.003   670
## deviance  82.225   5.408 80.160 80.691 81.500 82.905 87.282 1.017  4000
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
## DIC info (using the rule, pD = var(deviance)/2)
## pD = 14.6 and DIC = 96.8
## DIC is an estimate of expected predictive error (lower deviance is better).
```

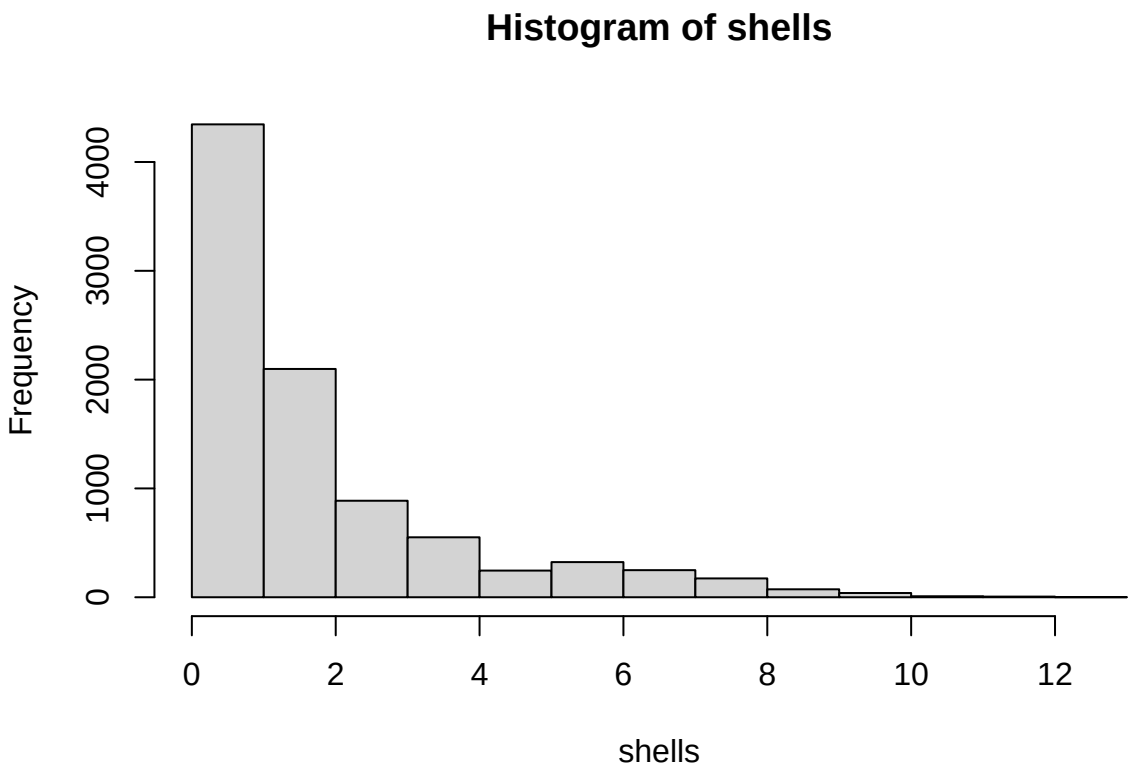
Nous vérifions que le modèle a bien convergé.



Les 2 chaînes se mélangent bien et convergent toutes deux. Ceci est aussi confirmé par la statistique de Gelman-Rubin $\hat{R}$ qui est inférieure à 1.1 pour chaque paramètre estimé. Nous notons aussi que `n.eff` est supérieur à 100. Nous avons donc des estimations de nos paramètres qui sont stables et des chaines peu autocorrélées.


Histogram of res$b.prev
Frequency
0
500
1000
1500
0.0
0.2
0.4
0.6
0.8
1.0
res$b.prev


Histogram of res$mu.0
Frequency
0
1000
2000
3000
−3.5
−3.0
−2.5
−2.0
−1.5
−1.0
res$mu.0




```
## [1] 1.81121
```

3 Comparaison de modèles

Une analyse statistique passe toujours par l'ajustement d'un premier modèle simple. Commencez par décrire le modèle qui vous semble pertinent pour répondre à une des questions traitées dans l'article, en vous aidant d'équations ou pas, et en justifiant son utilisation. Rappelez les hypothèses sur lesquelles repose le modèle. Notez qu'il ne vous est pas demandé de vérifier la validité de ces hypothèses (l'analyse des résidus). Vous pouvez ignorer les effets aléatoires sur les pentes (slopes) et ne considérer que des modèles avec un effet aléatoire sur l'ordonnée à l'origine (intercept).

En vous aidant de l'article, formez quelques hypothèses et construirez les modèles correspondant. Ajustez et comparez ces modèles pour déterminer l'hypothèse la mieux supportée par les données. N'oubliez de standardiser les variables explicatives continues avec la fonction scale() de R par exemple. Dans le cas d'une variable année, la standardisation est un peu différente, n'utilisez pas la fonction scale(). Si year <- c(2006, 2007, ... 2021) par exemple, utilisez simplement la variable year - 2005 qui vaut (1, 2, ... 16) pour avoir toujours une variable entière, mais qui commence à 1 et ne prend plus de grandes valeurs.

Le modèle null : pas de covariable, pas d'effet aléatoire Modèle où on ajoute les effets aléatoires Modèle avec les covariables explicatives : une, l'autre, les deux Eventuellement tester une intéraction si elle est pertinente Partir directement avec les effets aléatoires (prendre en compte la structuration des données)

4 Inférence et interprétation des résultats

Sur la base du meilleur modèle, donnez les estimations des paramètres ainsi qu'une mesure de l'incertitude associée. Interprétez vos résultats.

5 Discussion

Comparez vos résultats à ceux du papier. Sont-ils semblables ou différents? Pourquoi selon vous? Si cela vous semble pertinent, proposez des pistes d'amélioration de l'analyse.