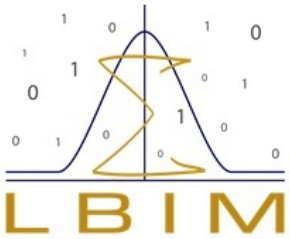
Programmation de modèles hiérarchiques et mixtes sous WinBUGS

Sophie Hammann



Programmation de modèles hiérarchiques et mixtes sous WinBUGS

HAMMANN Sophie
sophie.hammann@etu.unistra.fr
Université de Strasbourg
UFR de Mathématiques et d'Informatiques
Master 1 Statistique

27 août 2012

Remerciements

Tout d'abord, je souhaiterais remercier le Professeur Nicolas Meyer, mon tuteur professionnel, pour son encadrement et le temps qu'il m'a consacré. Son savoir et ses expériences m'ont permis d'enrichir grandement mes connaissances. Je le remercie par ailleurs d'avoir partagé ses anecdotes, cela m'a permis de constater la difficulté de l'utilisation des statistiques.

Je remercie également les Docteurs Erik-André Sauleau et François Lefebvre d'avoir partagé leur connaissance ainsi que leur point de vue durant ces 2 mois.

Je tiens à présent à remercier Monsieur Schaeffer Mickael, un camarade de Master 2 Statistiques, pour son aide, son soutien et sa joviale compagnie tout au long de ce stage.

Table des matières

Chapitre 1

Laboratoire de Biostatistique

1.1 Présentation et Activité

Situé au coeur du campus de la faculé de Médecine, le Laboratoire de Biostatistique et d'Informatique Médicale est composé du Dr Erik-André Sauleau, Dr François Lefebvre et Pr Nicolas Meyer, responsable du laboratoire. Le Laboratoire de Biostatistique travaille sur plusieurs axes de recherche tels que les méthodes PLS (Partial Least Squares), les modèles spatiaux pour l'épidémiologie et les applications médicales des méthodes bayésiennes.

1.2 Présentation du Sujet

Avertissement : les données traitées dans ce rapport sont soit purement fictives, soit extraites d'exemples connus.

1.2.1 Croissance chez les rats

Le jeu de données utilisé ici est issu d'un exemple de Winbugs provenant de la section 6 de Gelfand et al (1990). Le poids de 30 jeunes rats a été mesuré de façon hebdomadaire pendant 5 semaines. L'ajout d'un facteur inter-sujet *groupe* permettra de prendre en compte le sexe de chaque rat. On étudiera ainsi si l'effet *sexe* a eu un impact significatif sur la prise de poids.
L'objectif principal de cette étude est de comparer les résultats obtenus par des méthodes bayésiennes implémentées dans WinBUGS avec les résultats obtenus par des méthodes fréquentistes implémentées dans les logiciels R et SAS.

1.2.2 Marquage radioactif

Sur un ensemble de 18 patients, l'intensité de fixation d'un marqueur radioactif a été mesuré sur différentes zones d'un cliché du corps obtenu chez des patients présentant une pathologie infectieuse. Le nombre de points de mesures est variable d'un patient à l'autre et chaque mesure est faite à deux temps différents. On dispose donc au total de 104 mesures. Deux écritures différentes de modèle seront étudiées, l'objectif étant ici d'ajuster au mieux le modèle aux données.

1.2.3 Croissance infantile

L'évaluation de la croissance chez les enfants se fait principalement par la taille. Il existe de nombreuses autres manières pour juger de cette croissance et notamment, en mesurant la distance en millimètres entre la glande pituitaire (hypophyse) et la fissure ptérygo-maxillaire. Pendant 8 ans, ces distances ont été mesurées biennalement (tous les 2 ans) chez 27 enfants, dont 16 garçons et 11 filles. Les enfants étaient tous âgés de 8 ans lors de la première mesure. Le sexe de l'enfant est ici pris en compte car chez les garçons, la glande pituitaire et la fissure ptérygo-maxillaire sont plus éloignées que chez les filles.

Chapitre 2

L'Analyse Bayésienne

2.1 Définition

Un modèle statistique bayésien est constitué d'un modèle statistique paramètrique, f(x|θ), et d'une distribution a priori pour les paramètres, $\pi(\theta)$.

Une loi a priori sur θ résume nos connaissances a priori sur ce paramètre. Cette inférence est basée sur le Théorème de Bayes, dont elle tire son nom : Analyse Bayésienne.

Théorème de Bayes :

Si A et B sont des évènements tels que P(A) $\neq 0$, P(A|B) et P(B|A) sont reliés par la relation :

$$\mathrm{P(B|A)} = \frac{\mathrm{P(A|B)P(B)}}{\mathrm{P(A)}}.$$

En appliquant ce théorème au cas d'une densité de fonction, on peut obtenir la loi conditionnelle de θ sachant $\underline{x}$, les observations. Cette loi est appelée *loi a posteriori*, notée $\pi(\theta|\underline{x})$ et donnée par :

$$\pi(\theta|\underline{x}) = \frac{f(x|\theta)\pi(\theta)}{\int_{\Theta'} f(x|\theta')\mathrm{d}\theta'}.$$

Le dénominateur ne dépendant pas du paramètre θ, nous avons donc :

$$\pi(\theta|\underline{x}) \propto f(x|\theta)\pi(\theta).$$

La loi a posteriori combine les informations a priori et les informations apportées par les observations. Elle permettra d'inférer sur θ : par exemple, lorsque l'on cherche une estimation de θ, un estimateur bayésien peut alors être $\pi(\theta|\underline{x})$. Mais parfois, la loi a posteriori peut être difficile à étudier analytiquement, notamment lorsque θ est multivarié ou que la loi a priori n'est pas conjuguée (cf Annexe B) .

Ce problème peut être évité grâce aux méthodes MCMC : Monte Carlo par Chaîne de Markov. Avant d'étudier cette méthode, il nous faut comprendre ce qu'est une Chaîne de Markov.

2.1.1 Chaîne de Markov

On considère une suite de variables aléatoires $\{X_t\}_{t\geq 0}$ avec $t = 0, 1, 2...$ telles que les variables aléatoires X_i sont à valeurs dans E un espace continu. $\{X_t\}$ est une Chaîne de Markov si $\forall\ A \in T(E)$, la tribu engendrée par E,

on a :

$$\mathbb{P}(X_{t+1}\in A|X_0 = x_0, X_1 = x_1, ..., X_t = x_t) = \mathbb{P}(X_{t+1}\in A|X_t = x_t).$$

2.2 Méthode MCMC

La méthode de Monte Carlo utilise des tirages aléatoires pour réaliser le calcul d'une quantité déterministe. La Méthode de Monte Carlo par Chaînes de Markov tire son nom du fait que si nous voulons produire des approximations acceptables d'intégrales et de fonctions dépendants d'une certaine loi d'intérêt, il nous suffit de simuler une chaîne de Markov $\{\theta^m\}_m$ qui a pour loi stationnaire (cf Annexe A) la loi d'intérêt. De nos jours, les algorithmes MCMC sont fréquemment utilisés ; en effet, ces algorithmes sont itératifs et étant donné l'avancée technologique actuelle, les simulations des Chaînes de Markov sont obtenues facilement.

2.2.1 Principe

Le principe résulte en la simulation de variables aléatoires θ_1, θ_2 ... θ_T de loi $\pi(\theta|\underline{x})$. A l'aide de ces variables aléatoires, on peut par exemple approximer $\mathbb{E}(\theta|\underline{x})$ par

$$\frac{1}{T}\sum_{i=1}^{T}\theta_i.$$

Selon la Loi Forte des Grands Nombres par le Théorème Ergodique (cf Annexe A), cet estimateur converge presque sûrement vers $\mathbb{E}(\theta|\underline{x})$ lorsque T tends vers $+\infty$.
L'algorithme de Gibbs fait parti des algorithmes MCMC classiques et fréquemment utilisés. En particulier, cet algorithme est utilisé dans le logiciel WinBUGS pour simuler des chaînes de Markov.

2.3 L'algorithme de Gibbs

Supposons que l'on ait $\theta = (\theta^1, \theta^2, ..., \theta^d)\in\mathbb{R}^d$, $(d\in\mathbb{N})$ et que l'on sache simuler selon les lois conditionnelles $\pi(\theta^1|\theta^2, \theta^3, ..., \theta^d, \underline{x}), \pi(\theta^2|\theta^1, \theta^3, ..., \theta^d, \underline{x}), ..., \pi(\theta^d|\theta^1, \theta^2, \theta^3, ..., \theta^{d-1}, \underline{x})$. Le principe de l'algorithme de Gibbs consiste à partir de l'état de la chaîne à l'instant t, $\theta_t = (\theta_t^1, \theta_t^2, ..., \theta_t^d)$, à simuler l'état de la chaîne à l'instant $t + 1$ par :

$$\theta_{t+1}^1 \sim \pi(\theta^1|\theta_t^2, \theta_t^3, ..., \theta_t^d, \underline{x})$$
$$\theta_{t+1}^2 \sim \pi(\theta^2|\theta_{t+1}^1, \theta_t^3, ..., \theta_t^d, \underline{x})$$
$$...$$
$$\theta_{t+1}^d \sim \pi(\theta^d|\theta_{t+1}^1, \theta_{t+1}^2, ..., \theta_{t+1}^{d-1}, \underline{x})$$

La chaîne de Markov $\{(\theta_t^1, \theta_t^2, ..., \theta_t^d)\}_{t\geq 0}$ produite par cet algorithme admet $\pi(\theta^1, ..., \theta^d|\underline{x})$ pour loi stationnaire.

2.4 WinBUGS

WinBUGS est un logiciel statistique utilisable sur le système d'exploitation Windows et adapté à l'analyse bayésienne. Il utilise le programme BUGS : Bayesian Analysis using the Gibbs Sampler, qui prend en entrée un modèle composé de loi(s) a priori et de valeurs initiales, et retourne les sorties de l'algorithme de Gibbs pour cette loi a posteriori. Le logiciel calcule également des statistiques relatives à cette loi, et propose des diagnostics de validité, de type convergence. Une hypothèse fondamentale est la convergence de l'algorithme MCMC, c'est-à-dire la convergence des éléments de la chaîne de Markov obtenus, $\{(\theta_t^1, \theta_t^2, ..., \theta_t^d)\}_{t\geq 0}$, vers la loi stationnaire $\pi(\theta^1, ..., \theta^d|\underline{x})$. Pour juger de cette convergence, nous exploiterons les diagnostics d'autocorrélation, de Gelman et Rubin et d'Heidelberg et Welch (cf Annexe C).

Il existe d'autre diagnostics importants en plus de ceux que nous utiliserons, tel que le Diagnostic de Geweke et le Diagnostic de Raftery et Lewis. Réciproquement, le premier test l'égalité de moyenne de deux sous-suites de la chaîne de Markov $(Z_k)_{k\in\mathbb{N}}$ (car si la chaîne a bien convergé, ce test devrait être accepté) et le deuxième exprime le nombre d'itérations nécessaire pour pouvoir estimer un quantile de la loi a posteriori à partir de la chaîne de Markov simulée $(Z_k)_{k\in\mathbb{N}}$. On utilisera ici que les diagnostics les plus courants.

Notons qu'une des particularités lorsque l'on code dans un programme de type WinBUGS, est que les lois normales ne prennent pas en paramètre moyenne et variance, mais moyenne et précision, où la précision est définie comme l'inverse de la variance.

Chapitre 3

Différents Modèles

3.1 Les modèles Mixtes

3.1.1 Définition

Les modèles mixtes sont des modèles composés d'une partie à effets fixes et d'une partie à effets aléatoires. Mathématiquement, cela s'écrit :

$$y_i = \mu + \alpha_1 x_{i1} + \alpha_2 x_{i2} + .. + \alpha_p x_{ip} + \beta_{i1} z_{i1} + \beta_{i2} z_{i2} + ... + \beta_{iq} z_{iq} + \mathcal{E}_i$$

avec i allant de 1 à n, représentant l'individu i, p paramètres à effets fixes et q paramètres à effets aléatoires. Matriciellement, ce modèle s'écrit :

$$y = X\alpha + Z\beta + e$$

où $\beta = (\beta_1, \beta_2, ..., \beta_q)'$ est le vecteur des effets aléatoires,
$\alpha = (\mu, \alpha_1, \alpha_2, ..., \alpha_p)'$, vecteur des effets fixes,
$e = (e_1, e_2, ..., e_n)'$, représente les résidus et $y = (y_1, y_2, ..., y_n)$, représente les observations.

$$X = \begin{pmatrix} 1 & x_{11} & x_{21} & ... & x_{n1} \\ 1 & x_{12} & x_{22} & ... & x_{n2} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{1p} & x_{2p} & ... & x_{np} \end{pmatrix}$$

matrice des effets fixes de dimension p×(n+1), et

$$Z = \begin{pmatrix} z_{11} & z_{21} & ... & z_{n1} \\ z_{12} & z_{22} & ... & z_{n2} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ z_{1q} & z_{2q} & ... & z_{nq} \end{pmatrix}$$

matrice des effets aléatoires de dimension q×n.

La variance du modèle est définie par la matrice de variance-covariance, V :

$$V = var(y) = var(X\alpha + Z\beta + e)$$

On suppose que les effets aléatoires et les résidus sont indépendants.

$$V = var(X\alpha) + var(Z\beta) + var(e)$$

Comme α décrit les effets fixes, on a $var(X\alpha) = 0$.

$$V = var(Z\beta) + var(e) = Zvar(\beta)Z' + var(e) = ZGZ' + R$$

avec $G = var(\beta)$ et $R = var(e)$ qui représente la matrice des résidus.

La structure des matrices G et R peut être définie de façons différentes. En fait, il existe 3 grands types de modèle mixte ; les modèles à effets aléatoires, les modèles à coefficients aléatoires et les modèles avec motifs de covariance. Les modèles que l'on étudiera seront les modèles à coefficients aléatoires et les modèles avec motifs de covariance.

Les données longitudinales sont des exemples de modèles à coefficients aléatoires. On verra des exemples d'utilisation dans les paragraphes 4.2 et 5.2.
Le modèle avec motif de covariance sera quant à lui exploité dans l'étude de la croissance infantile. Il faut cependant introduire ce modèle avant de l'exploiter.

Le modèle avec motifs de covariance (covariance pattern model en anglais) est un modèle de type mixte dont la matrice de variance-covariance est spécifiée. La structure de cette matrice n'est en fait pas définie ici par les effets ou les coefficients aléatoires, mais par une structure propre au contexte de l'étude reflétée dans la matrice des résidus R.

En Statistique Bayésienne, il est d'usage d'utiliser comme loi a priori pour cette matrice de variance-covariance, la loi inverse de Wishart. Une loi Wishart est une version multivariée de la loi du $\mathcal{X}^2$. Si nous avons $\{X_i\}^v$ une suite de variables aléatoires gaussiennes indépendantes et identiquement distribuées, $\{X_i\}^v \sim \mathcal{N}(0, P)$, avec P une matrice symétrique définie positive, alors par définition
$Y = \sum_{i=1}^{v} X_i X_i'$ est distribué selon une loi de Wishart.

3.2 Les modèles Hiérarchiques

3.2.1 Définition

Un modèle à struture hiérarchique est aussi appelé modèle à multi-niveaux "emboîtés". Par exemple, lorsque l'on a des mesures intra patients, intra cliniques et intra régions.

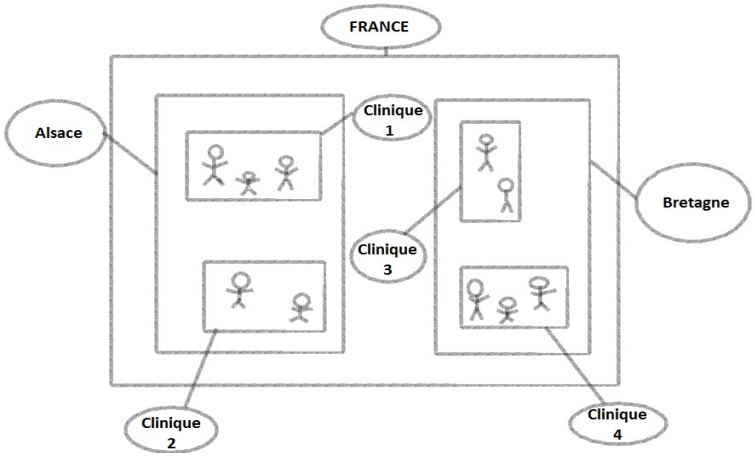


FIGURE 3.1: Exemple de Structure hiérarchique

Les modèles hiérarchiques décrivent notamment des ensembles de données complexes incluant de la corrélation ou comprenant d'autres propriétés, c'est pour cela qu'ils se combinent avec les modèles mixtes.

Il ne faut pas confondre cette définition du mot "hiérarchique" avec celle de modèle hiérarchique bayésien. En effet, en analyse bayésienne, un modèle hiérarchique correspond à un modèle bayésien où la loi a priori $\pi(\theta)$ est décomposée en une suite hiérarchique de lois de probabilité conditionnelles. Par exemple, pour un entier m la loi $\pi(\theta)$ se décompose de la façon suivante :

$$\pi(\theta) = \int_{\Theta_1} ... \int_{\Theta_m} \pi(\theta|\theta_1)\pi(\theta_1|\theta_2)...\pi(\theta_{m-1}|\theta_m)\pi(\theta_m)\mathrm{d}\theta\mathrm{d}\theta_1..\mathrm{d}\theta_m.$$

Concrètement, cela signifie que le modèle statistique bayésien met en jeu plusieurs niveaux de distributions a priori conditionnelles. On appelle les paramètres θ_i des *hyperparamètres* de niveau i, les $\pi(\theta_i|\theta_j)$ sont appelés les lois a priori conditionnelles ou les *hyperpriors*.

Chapitre 4

Croissance chez les rats

Mathématiquement, le modèle de type analyse de la variance à mesures répétées ajusté aux données s'écrit :

$$Y_{isj} = \mu + \alpha_i + \tau_{s(i)} + \beta_j + (\alpha\beta)_{ij} + \mathcal{E}_{isj}$$

avec le *groupe* i ,allant de 1 à 2, le *rat* s, allant de 1 à 30, et le *temps* j, allant de 1 à 5. Les facteurs α et β étant à effet fixes, on fixe des contraintes supplémentaires pour éviter toute surparamétrisation :

$$\sum_{i=1}^{2} \alpha_i = 0,\ \sum_{j=1}^{5} \beta_j = 0,\ \sum_{i=1}^{2} (\alpha\beta)_{ij} = 0\ \forall j,\ \sum_{j=1}^{5} (\alpha\beta)_{ij} = 0\ \forall i.$$

Y_{isj} représente le poids du rat s, appartenant au groupe i, au temps j.

Les $\tau_{s(i)}$ représentent un échantillon de population de taille 30 prélevé dans une population plus importante. On admet que les $\tau_{s(i)}$ sont distribués suivant une loi normale centrée de variance $\sigma^2_{\tau|\alpha}$. Les facteurs τ permettent de prendre en compte la variabilité intra-rats. L'ajustement d'un modèle d'analyse de la variance classique n'est pas possible vu que la condition d'indépendance des mesures n'est pas respectée. En effet, pour chaque rat on a mesuré à 5 temps différents leur poids, cela signifie que l'on a fait plusieurs mesures sur un même rat. Les mesures ne peuvent donc pas être indépendantes.

Ce modèle peut être utilisé si certaines conditions sont vérifiées, on détaillera ces hypothèses lorsque l'on utilisera les logiciels SAS et R.

Cependant, lorsque l'on modélise ce modèle sur le logiciel WinBUGS, aucune condition ne sera vérifiée. En fait, le logiciel WinBUGS permet de simuler des chaînes de Markov, aucune statistique de test ne sera calculée et donc aucune p-valeur n'apparaîtra. Les résultats obtenus par simulation nous donneront à titre indicatif une idée de l'effet significatif ou non du facteur *sexe*. Pour réellement interpréter cet effet, l'hypothèse de normalité des résidus est nécessaire.

On souhaite également étudier une autre écriture du modèle et ainsi comparer les résultats obtenus avec ceux qui précédent. Le terme τ ne sera alors plus présent dans le modèle, c'est-à-dire que la corrélation due aux données répétées ne sera pas prise en compte par le facteur *rat*. On comparera également les

résultats obtenus dans WinBUGS avec les résultats obtenus par des méthodes fréquentistes implémentées dans les logiciels R et SAS.

On étudie sous le logiciel WinBUGS le modèle suivant :

$$Y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \mathcal{E}_{ij}$$

4.1 Première écriture

On écrit tout d'abord le modèle sous la forme suivante :

$$\boxed{\begin{array}{l} Y[i,j] \sim dnorm(mu[i,j], tau.c) \\ mu[i,j] \leftarrow alpha + bb[groupe[i]] + beta[x[j]] + iac[groupe[i], x[j]] \end{array}}$$

Le paramètre *alpha* représente la constante commune c'est-à-dire le μ de l'écriture précédente. $bb[groupe[i]]$ représente en fait le groupe auquel appartient l'individu i, $beta[x[j]]$ fait référence au temps j et le paramètre $iac[groupe[i], x[j]]$ représente l'interaction entre l'individu i et le temps j.

Une loi a priori doit être spécifiée pour chaque paramètre. Si l'on veut faire en sorte que seulement les données nous apportent de l'information, c'est-à-dire éviter toute idée préconçue, on doit choisir des lois a priori peu informatives. C'est pour cela que l'on choisit des lois gaussiennes avec de grandes variances pour tous les paramètres, exceptés pour la précision, τ égal à $\frac{1}{\sigma^2}$, qui a pour distribution une loi *Gamma*.

Le paramètre *groupe* est de nature qualitatif. Lorsqu'un rat appartient au *groupe* 1,cela signifie que c'est une femelle, et lorsqu'il appartient au *groupe* 2, que c'est un mâle. x est également un facteur qualitatif, il représente les différents temps auquels les poids ont été mesurés. $x[1]$représente le *temps* 1 c'est-à-dire la première semaine, $x[2]$ le *temps* 2, la deuxième semaine etc.. jusqu'à $x[5]$ le *temps* 5, la cinquième semaine.

Le modèle où les contraintes sur les sommes ne sont pas prises en compte sera tout d'abord étudié, c'est-à-dire :

$$\sum_{i=1}^{2} \alpha_i \neq 0,\ \sum_{j=1}^{5} \beta_j \neq 0,\ \sum_{i=1}^{2} (\alpha\beta)_{ij} \neq 0\ \forall j,\ \sum_{j=1}^{5} (\alpha\beta)_{ij} \neq 0\ \forall i$$

puis, l'ajout de ces contraintes permettra l'étude d'un autre modèle. En effet, il faut voir quel modèle s'ajuste le mieux aux données.

4.1.1 Modèle sans contrainte

On utilise le modèle écrit précédemment sous le logiciel WinBUGS.

```
model
{
for( i in 1 : N ) {
for( j in 1 : T ) {
Y[i , j] ~ dnorm(mu[i , j],tau.c)
mu[i , j] ← alpha+ bb[groupe[i]]+beta[x[ j ]]+iac[groupe[i],x[j]]
predicte[i,j]← alpha+ bb[groupe[i]]+beta[x[ j ]]+iac[groupe[i],x[j]]
}
}
alpha ~ dnorm(243,0.001)
for (k in 1 : 2) { bb[k]~dnorm(0,0.001) }

for(l in 1 : 5){beta[l] ~ dnorm(0,0.001)}

for(k in 1 : 2){
for(l in 1 : 5){
iac[k,l]~dnorm(0,0.0001)
}
}

tau.c ~ dgamma(0.001,0.001)
sigma ← 1 / sqrt(tau.c)
}
```

Aucune contrainte n'a ainsi été prise en compte. Tous nos paramètres ont pour distribution une loi normale de moyenne 0, excepté le paramètre *alpha*. En effet, l'ordre de grandeur de la variable réponse étant de 243, l'a priori sur la constante est donc fixée à cette valeur.

Après avoir monitoré 100 000 itérations qui vont constituer notre Chaîne de Markov, on obtient les estimations suivantes :

	mean	sd	MC error	2.5%	median	97.5%
alpha	242.7	25.05	0.07501	193.7	242.7	291.6
bb[1]	-3.82	27.22	0.08227	-57.3	-3.815	49.4
bb[2]	3.139	27.28	0.08338	-50.33	3.119	56.66
beta[1]	-15.1	29.27	0.09517	-72.41	-15.12	42.2
beta[2]	-6.727	29.25	0.08837	-64.04	-6.793	50.64
beta[3]	0.2448	29.07	0.09361	-56.95	0.2534	57.12
beta[4]	7.79	29.24	0.09557	-49.25	7.703	64.89
beta[5]	13.67	29.05	0.08793	-43.11	13.42	70.87
iac[1,1]	-78.57	40.5	0.1279	-158.1	-78.54	0.7523
iac[1,2]	-38.47	40.65	0.131	-118.0	-38.51	41.59
iac[1,3]	-5.705	40.49	0.1249	-84.78	-5.688	73.72
iac[1,4]	29.07	40.59	0.1246	-50.83	29.08	108.4
iac[1,5]	57.98	40.38	0.1302	-21.37	57.78	136.9
iac[2,1]	-72.37	40.47	0.1282	-151.7	-72.4	6.829
iac[2,2]	-30.24	40.46	0.1286	-110.1	-30.27	49.05
iac[2,3]	9.077	40.5	0.1275	-70.2	9.143	88.42
iac[2,4]	47.83	40.57	0.1225	-32.1	47.8	127.3
iac[2,5]	77.65	40.39	0.1265	-1.62	77.59	156.8
sigma	12.0	0.7202	0.002419	10.69	11.97	13.51

WinBUGS a ainsi estimé les différents paramètres "alpha", "bb[1]".... 100 000 fois. Les résultats obtenus représentent les moyennes des 100 000 estimations ainsi que leurs écarts-type. Avec ces estimations, sont également présents des intervalles de confiance ; si l'on désigne I un de ces intervalles, θ le paramètre estimé, $\underline{x}$ les observations et le risque $\alpha \in]0,1[$, on a

$$P(\theta \in I|\underline{x}) = \int_I \pi(\theta|\underline{x})\,\mathrm{d}\theta = 1-\alpha.$$

Ces intervalles permettent de juger la significativité d'un facteur ou non. En fait, si les bornes supérieures et inférieures de notre intervalle sont du même signe, le facteur dû au paramètre a un effet significatif. Cet effet sera positif ou négatif selon le signe de l'estimation du paramètre, c'est-à-dire que la variable réponse sera augmentée ou diminuée par rapport à l'effet en question. Ainsi, si la valeur 0 est comprise dans cet intervalle, le facteur n'aura pas d'effet significatif.

Ici, aucun des facteurs n'a un effet significatif, *alpha* est le seul paramètre dont la valeur 0 n'appartient pas à son intervalle, cependant ce paramètre ne représente pas un facteur mais une constante commune à tous les rats.

On modèlise à présent les densités des lois a posteriori à l'aide du logiciel WinBUGS :

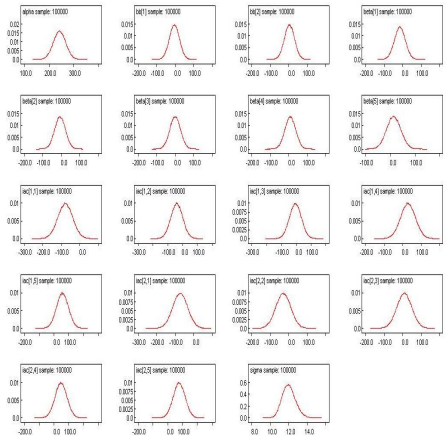


FIGURE 4.1: Densité des lois a posteriori

La qualité des graphes précédents est satisfaisante, la loi normale étant conjuguée avec elle-même, la loi a posteriori est ainsi également de densité normale. Les autocorrélations pour chaque paramètre sont par ailleurs faibles, le diagnostic d'Heidelberg et de Welch (cf Annexe C) indique que la chaîne est stationnaire, la convergence de la chaîne est appuyée par la statistique de Gelman et Rubin (cf Annexe C) dont la valeur est estimée à 1 (Ces diagnostics seront également visibles en Annexe C).

On étudie maintenant le même modèle agrémenté de contrainte sur les effets fixes.

4.1.2 Contrainte sur les effets fixes

Le modèle s'écrit de la même manière que précédemment, il n'est juste que légèrement modifié suite à la prise en compte de contrainte :

```
model
{
for( i in 1 : N ) {
for( j in 1 : T ) {
Y[i , j] ~ dnorm(mu[i , j],tau.c)
mu[i , j] <- alpha+ bb[groupe[i]]+beta[x[ j ]]+iac[groupe[i],x[j]]
predict[i,j] <- alpha+ bb[groupe[i]]+beta[x[ j ]]+iac[groupe[i],x[j]]
}
}
alpha ~ dnorm(243,0.001)
bb[2]~dnorm(0,0.001)
bb[1]<--bb[2]

for(l in 2 :5){beta[l] ~ dnorm(0,0.001)}
beta[1] <- -sum(beta[2 :5])

for(l in 2 :5){
iac[2,l]~dnorm(0,0.001)
}

iac[2,1] <- -sum(iac[2, 2 :5] )
for (l in 1 :5)
{
iac[1,l] <- -iac[2,l]
}

tau.c ~ dgamma(0.001,0.001)
sigma <- 1 / sqrt(tau.c)

}
```

100 000 itérations sont à nouveau monitorées, on obtient alors les estimations des paramètres suivantes :

	mean	sd	MC error	2.5%	median	97.5%
alpha	242.0	0.9798	0.003085	240.1	242.0	243.9
bb[1]	-10.25	0.9821	0.003227	-12.17	-10.25	-8.327
bb[2]	10.25	0.9821	0.003227	8.327	10.25	12.17
beta[1]	-90.15	1.967	0.005736	-93.99	-90.15	-86.3
beta[2]	-40.43	1.962	0.006257	-44.28	-40.43	-36.57
beta[3]	2.394	1.961	0.005909	-1.467	2.392	6.259
beta[4]	46.55	1.959	0.0064	42.7	46.55	50.39
beta[5]	81.64	1.966	0.006087	77.76	81.65	85.49
iac[1,1]	3.665	1.967	0.006165	-0.205	3.663	7.545
iac[1,2]	2.646	1.966	0.006651	-1.221	2.649	6.514
iac[1,3]	-0.621	1.96	0.006339	-4.5	-0.61	3.195
...	...	...	...	...	...	...
sigma	12.0	0.7217	0.002471 10.68	11.96	13.5	

On constate que les écart-types des estimations sont très faibles comparés à ceux obtenus dans le modèle précédent. En effet, ils étaient de l'ordre de 30 tandis qu'ici ils sont de l'ordre de 2. Le fait d'avoir un écart-type faible pour chaque estimation engendre un intervalle de confiance plus restreint pour chacun des paramètres, ce qui signifie que l'estimation est plus précise.

On remarque par leur intervalle de confiance que les paramètres *bb* et *beta*, respectivement les facteurs *groupe* et *temps*, ont un effet significatif. $bb[1]$ a un intervalle entièrement négatif, ce qui signifie que la variable réponse est diminuée pour le groupe 1. Concrètement, un rat appartenant au groupe 1 a un poids moins élevé qu'un rat appartenant au groupe 2. Ce qui est logique, vu que le groupe 1 représente les femelles et le groupe 2, les mâles.
Les *beta* (représentant le temps) sont quant à eux de plus en plus élevés, ceci implique que la variable réponse augmente en fonction du *temps*.

Bien que ces deux facteurs ont un effet significatf, ce n'est pas pour autant qu'il en est de même pour leur interaction, en effet, les paramètre $iac[i, j]$ ont majoritairement un intervalle qui contient la valeur 0, ainsi le facteur *temps* n'a pas d'effet significatif. Cela implique que les mâles et les femelles ont la même évolution à travers le temps. Il n'y a donc pas de lien direct entre le $Sexe$ et le *temps*.

A l'aide du logiciel Winbugs, on modélise les densités des lois a posteriori des paramètres du modèle :

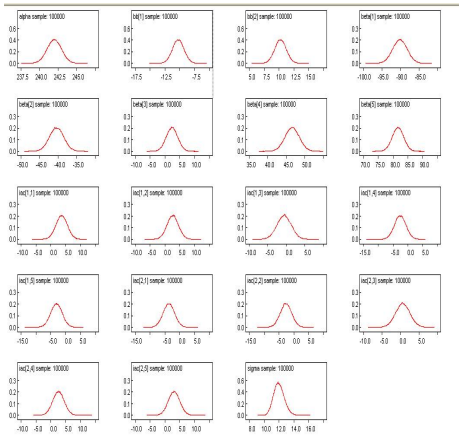


FIGURE 4.2: Densité des lois a posteriori

La qualité des graphes précédents est satisfaisante, en effet, les densités des lois a posteriori sont proches de densités gaussiennes. De plus, en visualisant la fonction d'autocorrélation on remarque que les autocorrélations sont très faibles pour chacun des paramètres (les autocorrélations ne sont pas visibles ici, cependant elles sont équivalentes à celles du modèle 4.1.1 étudiées en Annexe C).

Les deux modèles précédemment étudiés amènent à des résultats semblables en terme de convergence. Par ailleurs, si l'on cherche à voir les prédictions qu'engendrent ces estimations, on se rend compte que les prédictions sont les mêmes entre le modèle avec contraintes et celui sans.

En effet, $\alpha + bb[1] + beta[1] + iac[1,1]$ est égal à 145,21 pour le 1er modèle contre 145,265 pour le 2ème (Les prédictions de chaque modèle ainsi que leurs intervalles de confiance seront exposés à l'Annexe D). Les prédictions sont semblables malgré que les estimations des paramètres soient légèrement différentes. Comme nous l'avons vu précédemment, les écart-types de ces estimations sont loin d'être égales. De faibles écart-types engendrent des estimations plus précises, c'est pour cela que le modèle à préférer ici est celui avec les contraintes sur les effets fixes.

L'écriture de ce modèle étant plus proche de celle d'une écriture fréquentiste, on compare maintenant les résultats obtenus avec ceux des méthodes fréquentistes sur les logiciels SAS et R. Pour utiliser ces logiciels, il nous faut tout d'abord vérifier les hypothèses du modèle ajusté.

4.1.3 Vérification des hypothèses

L'utilisation d'un ajustement de type analyse de la variance nécessite de vérifier les hypothèses suivantes pour les variables erreurs, $\mathcal{E}_{isj}$:
- les erreurs sont indépendantes,
- les erreurs ont même variance σ^2 inconnue,
- les erreurs sont de loi gaussienne.

A l'aide des prédictions obtenues par WinBUGS précédemment, on peut visualiser les résidus, qui sont en fait les réalisations des $\mathcal{E}_{isj}$ et ainsi voir si a priori, ils vérifient les conditions d'utilisation.

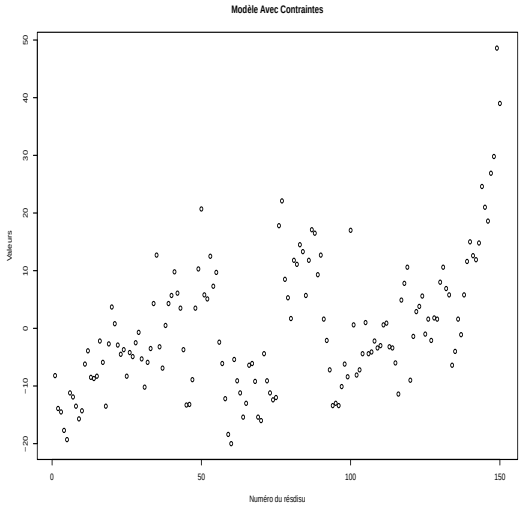


FIGURE 4.3: Résidus du modèle avec contrainte

A première vue, les résidus ne paraîssent pas homoscédastiques, il semble rester un effet observable dans les résidus. Par ailleurs leur étendue est élevée, ils ne semblent donc pas être symétriques par rapport à la valeur 0. La qualité de ces résidus est difficilement acceptable. Cependant, il nous faut étudier ces

hypothèses à l'aide de test. Le test de Shapiro-Wilk testera la normalité des résidus, et le test de Bartlett, leur homoscédasticité. Pour l'hypothèse d'indépendance, on utilisera le test de Durbin-Watson.

La statistique de test de Shapiro-Wilk a ici pour valeur 0,98 et sa p-valeur associée 0,02 . On rejette donc l'hypothèse de normalité des résidus avec un risque de première espèce de 5%. On remarque que l'on ne peut pas ici tester l'égalité des variances (hypothèse d'homoscédasticité), en effet il n'y a qu'une observation par "case". Cependant, pour avoir une idée on peut tester l'égalité des variances des poids selon le groupe, puis selon le temps. On obtient donc deux tests.

Tout d'abord, pour la première égalité des variances, la statistique de test de Bartlett a pour valeur 0,45 et sa p-valeur associée 0,50 . L'égalité des variances des poids selon le groupe est donc mise ici en évidence, avec un risque de première espèce de 5%.
La statistique de test de Bartlett pour l'égalité des variances des poids selon le temps a pour valeur 24,64, et sa p-valeur associée, $5{,}94 \times 10^{-5}$. On rejette donc l'hypothèse d'homoscédasticité avec un risque de première espèce de 5%.

Ces résultats ne nous garantissent pas que les résidus ne sont pas homoscédastiques, mais ce sont de bons indicateurs.

Le test de Durbin-Watson permet de détecter une autocorrélation de la forme :

$$\mathcal{E}_i = \rho\mathcal{E}_{i-1} + \nu_i,\ \nu_i \sim \mathbb{N}(0, \sigma_\nu).$$

Le test d'hypothèse s'écrit :

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

La statistique de test de Durbin Watson est donnée par :

$$DW = \frac{\sum_{t=1}^{n-1}(e_{t+1} - e_t)^2}{\sum_{t=1}^{n} e_t^2},$$

avec e représentant les résidus et n la taille de l'échantillon.

On accepte H_0 si $d_U < DW < 4 - d_U$, on rejette H_0 si $DW < d_l$ ($\rho > 0$) ou $DW > 4 - d_l$ ($\rho < 0$), on est dans un cas d'incertitude si $d_l < DW < d_U$ ou $4 - d_U < DW < 4 - d_l$.
Les valeurs de d_l et d_U sont données par la table de Durbin-Watson qui prend en compte le nombre de régresseurs et la taille. Ici leurs valeurs respectives sont 1,68 et 1,79.

A l'aide du package *car* et de la fonction *durbinWatsonTest()*, on obtient pour la valeur de cette statistique de test 0,36 : $DW < d_l$ donc on rejette H_0, cela implique qu'il y a présence d'autocorrélation.

Ainsi donc, l'hypothèse de normalité, l'hypothèse d'homoscédasticité et l'hypothèse d'indépendance ne semblent pas vérifiées. Avant toute interprétation, il sera nécessaire de corriger ces effets sur les résidus, soit en transformant la variable réponse à l'aide d'une fonction de type $\sqrt{}$, $Arcsin$, $Arcsin(\sqrt{})$..., soit en incluant une variable supplémentaire au prédicteur linéaire, ou en modifiant l'une des variables explicatives.

On se limite ici à la comparaison des résultats obtenus sous les différents logiciels : WinBUGS, R et SAS.

Tout d'abord, on condense les résultats obtenus sous le logiciel Winbugs :

groupe	temps	moyenne	ecarttype
1	1	144.99	3.22
1	2	193.74	3.19
1	3	233.26	3.22
1	4	275.38	3.23
1	5	310.03	3.21
2	1	158.20	2.99
2	2	208.87	3.00
2	3	255.07	2.99
2	4	301.11	2.98
2	5	336.76	3.01

Les 5 premières mesures correspondent aux moyennes et écart-types des prédictions du poids pour les rats appartenant au groupe 1, les femelles, à chaque temps. Les mesures suivantes représentent les mêmes résultats mais pour les rats du groupe 2, les mâles.

Sur R, la fonction *lme* de la librairie *nlme*, permet de réaliser une analyse de la variance à mesures répétées. De plus, cette librairie permet de prendre en compte des hiérarchies multiples. Pour utiliser la fonction *lme*, il nous suffit d'écrire le modèle sous la forme

lme(poids $\sim$ groupe*temps, data=tableau, random=$\sim$ 1|rat)

avec *tableau* qui contient les données *poids* c'est-à-dire les *poids* pour chaque *rat*, avec le *groupe* de celui-ci, à chaque *temps*. On condense les résultats en les affichant par *groupe* et par *temps*.

groupe	temps	moyenne	EcartType
1	1	145.14	7.98
1	2	193.64	7.98
1	3	233.43	7.98
1	4	275.79	7.98
1	5	310.57	7.98
2	1	158.31	10.46
2	2	208.88	10.46
2	3	255.19	10.46
2	4	301.50	10.46
2	5	337.25	10.46

A l'aide d'une PROC MIXED, on modélise ce modèle sur le logiciel SAS. Voici le code utilisé ainsi que les résultats obtenus.

```
PROC MIXED data=Tableau ;
class groupe temps rat ;
model poids = groupe temps groupe*temps /solution ;
repeated /subject=rat(groupe) ;
LSMEANS groupe temps groupe*temps ;
run ;
```

Moyennes des moindres carrés

Effet	groupe	temps	Valeur estimée	Erreur type	DDL	Valeur du test t	Pr > |t|
groupe*temps	1	1	145.14	3.1890	112	45.51	<.0001
groupe*temps	1	2	193.64	3.1890	112	60.72	<.0001
groupe*temps	1	3	233.43	3.1890	112	73.20	<.0001
groupe*temps	1	4	275.79	3.1890	112	86.48	<.0001
groupe*temps	1	5	310.57	3.1890	112	97.39	<.0001
groupe*temps	2	1	158.31	2.9830	112	53.07	<.0001
groupe*temps	2	2	208.87	2.9830	112	70.02	<.0001
groupe*temps	2	3	255.19	2.9830	112	85.55	<.0001
groupe*temps	2	4	301.50	2.9830	112	101.07	<.0001
groupe*temps	2	5	337.25	2.9830	112	113.06	<.0001

FIGURE 4.4: Résultat SAS Modèle 1

On remarque que les résultats obtenus par les logiciels WinBUGS et SAS sont proches, en effet, les résultats obtenus sont équivalents, autant par l'estimation des moyennes que par celles des écart-types. Ceux du logiciel R sont différents du point de vue des écart-types, en effet, dans l'écriture du modèle sur ce logiciel, on a tout de suite considéré que le facteur *rat* était aléatoire, pour prendre en compte les corrélations entre chaque mesure provenant d'un même rat. Tandis que sur le logiciel SAS, le facteur *rat* est défini par *repeated* et non *random*.

On a vu précédemment que la qualité des résidus obtenus était discutable. De plus, on remarque que les résulats obtenus sur les logiciels R et SAS montre que pour le groupe 1 on obtient exactement les mêmes écart-type, il en est de même pour le groupe 2. On étudie alors une autre écriture du modèle.

4.2 Ajout du facteur *rat*

On souhaite se rapprocher de l'écriture fréquentiste qui décrit le mieux les données :

$$Y_{isj} = \mu + \alpha_i + \tau_{s(i)} + \beta_j + (\alpha\beta)_{ij} + \mathcal{E}_{isj}.$$

Sous le logiciel WinBUGS, on écrit alors le modèle de la manière suivante :

```
model
{
for( i in 1 : 150 ) {
poids[i] ~ dnorm(mu[i],tau.c)
mu[i] <- alpha+ b1[rat[i]]+b2[groupe[i]]+b3[temps[ i ]]
+iac[groupe[i],temps[i]]
predict[i]<- alpha+ b1[rat[i]]+b2[groupe[i]]+b3[temps[ i ]]
+iac[groupe[i],temps[i]]
}
alpha ~ dnorm(243,0.001)

for(m in 1 :N){b1[m]~dnorm(0,tau.b1)}

b2[2]~dnorm(0,0.001)
b2[1]<- -b2[2]

for(l in 2 :T){b3[l] ~ dnorm(0,0.001)}
b3[1] <- -sum(b3[2 :5])

for(l in 2 :5)
iac[2,l]~dnorm(0,0.0001)
iac[2,1] <- - sum(iac[2, 2 :5])
for (k in 1 :5)
iac[1,k] <- - iac[2,k]


tau.c ~ dgamma(0.001,0.001)
sigma <- 1 / sqrt(tau.c)

sigma.b1~ dunif(0,100)
tau.b1<- 1/(sigma.b1*sigma.b1)

}
```

En ajoutant le terme $b1[rat[i]]$ au modèle, cela permet de prendre en compte la corrélation entre chaque mesure par rat. Les données étant longitudinales, il n'y a plus d'indépendances entre chaque mesure. Sous WinBUGS, pour prendre en compte l'effet *rat*, on ajoute un hyperprior dans l'écriture du modèle sur la loi de $b1$: $b1[m]\sim dnorm(0, \tau_{b1})$ avec $\tau_{b1} = \frac{1}{\sigma_{b1}^2}$ et $\sigma_{b1}\sim dunif(0, 100)$.

Après avoir monitoré 100 000 itérations, on obtient les estimations suivantes :

	mean	sd	MC error	2.5%	median	97.5%
alpha	241.6	11.3	0.5354	219.7	241.4	264.3
b1[1]	-90.64	11.97	0.5318	-114.5	-90.53	-67.16
b1[2]	-47.47	12.75	0.5596	-72.42	-47.38	-22.73
b1[3]	-4.989	12.75	0.5597	-29.85	-4.912	19.72
b1[4]	38.68	12.74	0.5592	13.8	38.79	63.35
...	...	...	...	...	...	...
b2[1]	3.393	4.174	0.1137	-4.576	3.366	11.53
b2[2]	-3.393	4.174	0.1137	-11.53	-3.366	4.576
b3[1]	-18.86	1.324	0.007573	-21.47	-18.86	-16.25
b3[2]	-7.545	1.305	0.00439	-10.09	-7.547	-4.983
b3[3]	0.2486	1.305	0.004256	-2.318	0.252	2.804
b3[4]	6.346	1.298	0.004546	3.788	6.346	8.892
b3[5]	19.81	1.303	0.004327	17.25	19.81	22.37
iac[1,1]	-0.6292	1.318	0.00569	-3.219	-0.6269	1.951
iac[1,2]	1.703	1.303	0.004057	-0.8474	1.703	4.258
iac[1,3]	1.877	1.303	0.004009	-0.6843	1.876	4.44
iac[1,4]	0.6291	1.303	0.004114	-1.921	0.6251	3.194
...	...	...	...	...	...	...
sigma	7.937	0.5379	0.002408	6.966	7.907	9.072
sigma.b1	65.76	8.986	0.06461	50.8	64.82	86.1

On s'intéresse plus particulièrement au facteur *rat* dans ce modèle. Il nous faut à présent étudier sa variance et non l'intervalle obtenu pour son estimation. Son écart-type est ainsi estimé par la valeur 65,76 (sigma.b1) tandis que l'écart-type du *poids* est estimé par 7,937 (sigma). Comme la variance de $b1$ est grande, cela implique que les poids varient notablement dun rat à lautre.

L'effet *groupe* et l'effet de l'intéraction *groupe* et *temps* ne sont ici pas significatifs, contrairement à l'effet *temps* qui est significatif. Selon la semaine, la variable réponse est ainsi diminuée ou augmentée.

Les paramètres estimés précédemments suivent une loi a posteriori, à l'aide de WinBUGS on modèlise la densité de cette loi pour chaque paramètre :

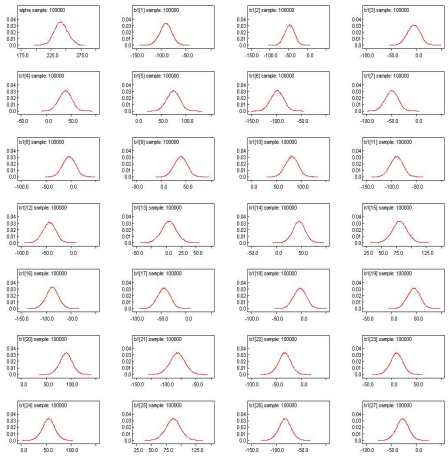


FIGURE 4.5: Densité des lois a posteriori

Les densités de la loi a posteriori semblent proches de densités gaussiennes. Les autocorrélations sont faibles pour les paramètres $b3$ et iac mais sont élevées pour les autres paramètres. Cependant, elles diminuent au fil des itérations. Pour avoir une faible autocorrélation pour tous les paramètres, il suffit de modifier le nombre de *thin* dans Winbugs, en utilisant par exemple $n.thin = 1000$. Ceci permettra de ne prendre en compte que les estimations obtenues toutes les 1000 èmes itérations, ainsi, cela réduira l'autocorrélation vu que les estimations seront plus éloignées.

On étudie à présent ce modèle sur les logiciel R et SAS. Pour cela, il faut alors à nouveau vérifier certaines hypothèses.

4.2.1 Vérification des hypothèses

L'ajustement de type analyse de la variance à mesures répétées,

$$Y_{isj} = \mu + \alpha_i + \tau_{s(i)} + \beta_j + (\alpha\beta)_{ij} + \mathcal{E}_{isj},$$

nécessite de vérifier certaines hypothèses. Tout d'abord, on suppose que les effets aléatoires $\tau_{s(i)}$ sont indépendants et que $\mathcal{L}(\tau_{s(i)}) = \mathcal{N}(0, \sigma^2_{\tau|\alpha})$, pour tout s variant de 1 à 30.

Il nous faut également vérifier les hypothèses suivantes pour les variables erreurs, $\mathcal{E}_{isj}$:
- les erreurs sont indépendantes,
- les erreurs ont même variance σ^2 inconnue,
- les erreurs sont de loi gaussienne.

La condition d'indépendance entre les effets aléatoires $\tau_{s(i)}$ et les erreurs $\mathcal{E}_{isj}$ est également à vérifier. On va ici les supposer comme indépendants.

A l'aide des prédictions obtenues précédemment par WinBUGS, on étudie si a priori les résidus, qui sont en fait les réalisations des $\mathcal{E}_{isj}$, vérifient les conditions d'utilisation.

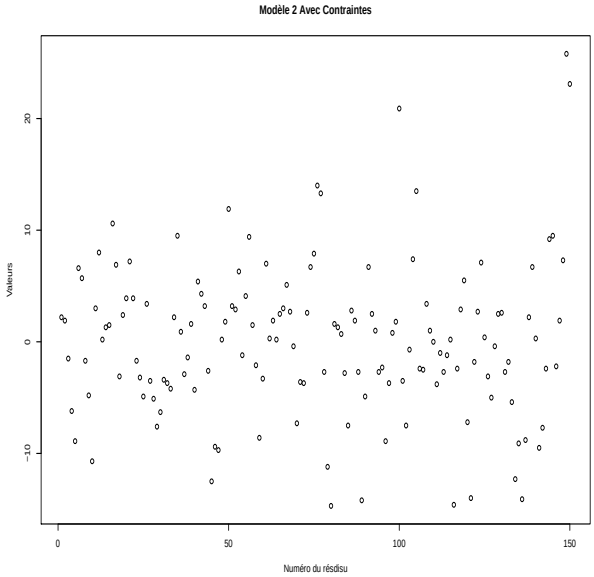


FIGURE 4.6: Residus Modèle 2

Les résidus semblent bien vérifier l'hypothèse d'homoscédasticité, en effet, leur répartition est homogène. De plus, ils semblent être symétriques par rapport à la valeur 0, excepté certaines valeurs. La majorité des résidus est comprise entre -10 et 10, ce qui est élevé. Il faut à présent vérifier les hypothèses à l'aide de test.

La statistique de test de Shapiro-Wilk a ici pour valeur 0,96 et sa p-valeur associée 0,001 . On rejette donc l'hypothèse de normalité des résidus avec un risque de première espèce de 5%. On remarque que l'on ne peut pas ici tester l'égalité des variances (hypothèse d'homoscédasticité), en effet il n'y a qu'une observation par "case". Cependant, pour avoir une idée on peut tester l'égalité des variances des poids selon le groupe, selon le temps puis selon le sujet. On obtient donc trois tests.

Tout d'abord, pour la première égalité des variances, la statistique de test de Bartlett a pour valeur 1,91 et sa p-valeur associée 0,17. L'égalité des variances des poids selon le groupe est donc mise ici en évidence, avec un risque de première espèce de 5%.
La statistique de test de Bartlett pour l'égalité des variances des poids selon le temps a pour valeur 10,55 , et sa p-valeur associée, 0,03. On rejette donc l'hypothèse d'homoscédasticité avec un risque de première espèce de 5%.

Enfin, la statistique de test de Bartlett pour l'égalité des variances des poids selon le sujet a pour valeur 59,87 , et sa p-valeur associée, 0,001. On rejette donc l'hypothèse d'homoscédasticité avec un risque de première espèce de 5%.

Ces résultats ne nous garantissent pas que les résidus ne sont pas homoscédastiques, mais ce sont de bons indicateurs.

La statistique de test de Durbin Watson a ici pour valeur 1,03 et d_l=1,66 d_U=1,80. $DW < d_l$ on rejette donc H_0, cela implique qu'il y a présence d'autocorrélation.

Ainsi donc, l'hypothèse de normalité, l'hypothèse d'homoscédasticité et l'hypothèse d'indépendance ne semblent pas vérifiées. Avant toute interprétation, il sera nécessaire de corriger ces effets sur les résidus, soit en transformant la variable réponse à l'aide d'une fonction de type $\sqrt{}$, $Arcsin$, $Arcsin(\sqrt{})$..., soit en incluant une variable supplémentaire au prédicteur linéaire, ou encore en modifiant l'une des variables explicatives.

Comme précédemment, on se limite ici à la comparaison des résultats obtenus sous les différents logiciels : WinBUGS, R et SAS.

Rat étant un facteur aléatoire, on s'intéresse plus particulièrement à sa variance, c'est donc ce que l'on compare à l'aide des différents logiciels.

On souhaite tout d'abord écrire ce modèle sur le logiciel R. La fonction *lmer* du package *lme*4 permet d'ajuster ce modèle, qui s'écrit alors **lmer(poids∼groupe*temps+(1|sujets))**. Ce package permet par ailleurs d'écrire des Modèles Linéaires Généralisés Mixtes.

En appliquant un *summary* à cette fonction *lmer* on obtient :

Random effects			
Groups	name	Variance	Std.Dev.
rat	(Intercept)	96.57	9.83
Residual		45.81	6.77

Sur WinBUGS, l'estimation de $sigma.b1$ est de 10,3 ce qui n'est pas très loin de 9,83 la valeur obtenue avec le logiciel R. L'estimation de *sigma*, qui correspond à l'écart-type des résidus est égale à 6,8 ce qui est également proche de 6,77 la valeur obtenue avec le logiciel R. Avec une certaine marge d'erreur assez faible, on peut dire que les variances correspondent pour les 2 logiciels.

Le logiciel SAS quant à lui estime cette variance par :

Valeurs estimées des paramètres de covariance	
Param de cov	Valeur estimée
rat(groupe)	96.5686
Residual	45.8083

FIGURE 4.7: Variance du facteur *rat*

Ces résultats sont obtenus grâce à la PROC MIXED de SAS :

```
PROC MIXED data=Tableau ;
class groupe temps rat ;
model poids = groupe|temps / solution ;
random rat(groupe) ;
LSMEANS groupe temps groupe*temps ;
run ;
```

Contrairement à la PROC MIXED précédente, le facteur *rat* est ici bien considéré comme aléatoire par l'utilisation de *random*.

Ainsi, les 3 logiciels obtiennent les mêmes résultats, bien que les écritures soient différentes. On a donc une certaine correspondance entre l'implémentation bayésienne du modèle sous le logiciel WinBUGS et son implémentation fréquentiste sous les logiciels R et SAS.

Chapitre 5

Marquage radioactif

On dispose pour chaque sujet d'un nombre variable de points de mesure (indicés pas "nogg" qui représente le numéro de point de mesure) vu chacun à deux temps (pré et post). La variable réponse Y (*suv* dans le logiciel WinBUGS), est une mesure qui indique l'intensité de fixation d'un marqueur radioactif. On écrit tout d'abord ce modèle sous WinBUGS sans prendre en compte l'effet *nogg*, puis en le prenant en compte. L'objectif étant d'ajuster au mieux un modèle à nos données, il nous est nécessaire pour cela d'étudier différentes écritures du modèle.

Mathématiquement, le modèle de type analyse de la variance à mesures répétées ajusté à nos données s'écrit :

$$Y_{isj} = \mu + \alpha_i + \tau_s + B_{j(s)} + \mathcal{E}_{isj}$$

avec le temps i, allant de 1 à 2, le sujet s, allant de 1 à 18, et le nombre de points de mesure étudié j pour chaque sujet. Le nombre de points de mesure est en fait emboîté dans le facteur *sujets*, il ne peut donc pas y avoir d'interaction entre les deux facteurs. On suppose par ailleurs, que le facteur *sujets* et le facteur *temps* sont indépendants, car on souhaite étudier l'effet du facteur *temps* général et non pour des sujets précis. En effet, si l'on prend en considération l'interaction entre le *temps* et le *sujet*, on interprétera l'effet du *temps* pour chaque *sujet*, et non l'effet du *temps* pour tous les *sujets*.

Y_{isj} représente l'intensité de fixation d'un marqueur radioactif au point de mesure numéro $B_{j(s)}$ au temps i, pour le sujet s.
Les τ_s représentent un échantillon de taille 18 prélevé dans une population importante. On admet que les τ_s sont distribués suivant une loi normale centrée de variance σ^2_τ. Les facteurs τ permettent de prendre en compte la variabilité inter-sujets. Il est nécessaire de rajouter ce facteur, car l'ajustement d'un modèle d'analyse de la variance classique n'est pas possible vu que la condition d'indépendance des mesures n'est pas respectée.

On suppose que les effets aléatoires τ_s sont indépendants et que
$\mathcal{L}(\tau_s) = \mathcal{N}(0, \sigma^2_\tau)$,pour tout s variant de 1 à 18. On suppose également que les effets aléatoires $B_{j(s)}$ sont indépendants et que $\mathcal{L}(B_{j(s)}) = \mathcal{N}(0, \sigma^2_{B|\tau})$, pour tout j variant de 1 à 7.

Pour tout ajustement à un modèle de type analyse de la variance, les hypothèses suivantes doivent être vérifiées pour les $\mathcal{E}_{isj}$:
- les erreurs sont indépendantes,
- les erreurs ont même variance σ^2 inconnue,
- les erreurs sont de loi gaussienne.

Il faut également la condition d'indépendance entre les différents effets aléatoires :
- τ_s et les erreurs $\mathcal{E}_{isj}$
- $B_{j(s)}$ et les erreurs $\mathcal{E}_{isj}$
- τ_s et $B_{j(s)}$.

Toutes ces hypothèses sont à vérifiées pour interpréter des résultats obtenus par méthode fréquentiste. Dans le cas d'une analyse de la variance en bayésien utilisant les MCMC, il n'est pas nécessaire de vérifier toutes ces hypothèses, l'hypothèse nécessaire est la normalité des résidus. Cependant notre objectif est ici d'ajuster au mieux un modèle aux données.

Les données vont alors être ajustées à 2 écritures différentes.

5.1 Première écriture

On ajuste à présent un modèle de type analyse de la variance à mesures répétées. Le modèle sous le logiciel Winbugs s'écrit :

$$\begin{array}{|l|}\hline suv[i] \sim dnorm(mu[i], tau.c) \\ mu[i] \leftarrow alpha[Patient[i]] + beta[nogg1[i]] + temps[prepost[i] + 1] \\ \hline\end{array}$$

pour i allant de 1 à 104, le nombre total de mesures. Le modèle ne comprend pas de terme constant, en effet il n'est pas obligatoire cependant lorsqu'il est présent, il faut faire attention à la surparamétrisation et c'est pour cela que l'on rajoute des contraintes sur les effets fixes. Ainsi, ici on n'inclut pas de contrainte. Le vecteur *prepost* prend la valeur 0 pour le temps *pre* et la valeur 1 pour le temps *post*. On rajoute donc 1 à *prepost* pour que le vecteur *temps* ait bien deux composantes. Le vecteur $nogg1$ représente le numéro du point de mesure étudié, dans la version fréquentiste du modèle, le nombre de points de mesure n'est pas pris en compte explicitement, en fait il se retrouve dans l'indice du facteur $B_{j(s)}$;
$j(s)$ représente le nombre de points de mesure étudié.

100 000 itérations sont à présent monitorées :

	mean	sd	MC error	2.5%	median	97.5%
alpha[1]	0.01838	4.742	0.09162	-10.14	9.938E-5	10.34
alpha[2]	-0.2846	4.778	0.07903	-11.25	-0.04147	9.481
alpha[3]	0.1774	5.025	0.06859	-10.43	0.04099	11.19
...	...	...	...	...	...	...
beta[1]	1.654	6.924	0.1912	-12.12	1.638	15.24
beta[2]	2.648	6.905	0.1907	-16.32	-2.66	10.91
beta[3]	1.217	6.93	0.1912	-12.59	1.218	14.79
...	...	...	...	...	...	...
sigma	4.072	0.4079	0.002344	3.367	4.039	4.963
sigma.alpha	4.118	3.161	0.1061	0.1925	3.444	12.05
temps[1]	8.616	4.626	0.1822	-0.3527	8.586	17.59
temps[2]	2.302	4.627	0.1824	-6.703	2.281	11.25

D'après les intervalles obtenus pour les estimations de *temps*[1] et *temps*[2], le facteur *temps* n'a pas d'effet significatif. Il en est de même pour les autres paramètres.

Les paramètres estimés précédemments suivent une loi a posteriori, à l'aide de WinBUGS on modèlise la densité de cette loi pour chaque paramètre :

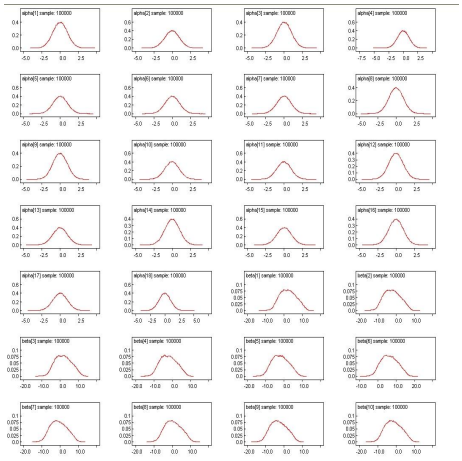


FIGURE 5.1: Densité des lois a posteriori

Les densités des lois a posteriori sont proches de lois gaussiennes. Les autocorrélations sont élevées pour les paramètres *beta* (cf Annexe C) . La technique utilisée précédemment pour réduire l'autocorrélation ne donne pas les résultats escomptés ici. En effet, l'autocorrélation reste élevée après avoir changé le *n.thin* .

5.2 Ajout du facteur *Sujet*

On souhaite maintenant se rapprocher de l'écriture fréquentiste sous le logiciel Winbugs :

$$Y_{isj} = \mu + \alpha_i + \tau_s + B_{j(s)} + \mathcal{E}_{isj},$$

avec la mesure i, allant de 1 à 104, le Patient s allant de 1 à 18, et le nombre de points de mesure étudié $j(s)$ (ce nombre varie en fonction de chaque sujet). Le facteur $B_{j(s)}$ auparavant considéré à effet fixe est supposé à présent à effet aléatoire.

L'écriture du modèle reste exactement la même que précédemment. Cependant, pour prendre en compte l'effet du facteur *nogg*1, il faut agrémenter la loi a priori du paramètre *beta* avec un hyperprior :

```
for (i in 1 :52)
{
beta[i] ~ dnorm(0 , tau.beta)
}
sigma.beta ~ dunif(0,100)
tau.beta<- 1/(sigma.beta*sigma.beta)
```

Après avoir monitoré 50 000 itérations, on obtient les estimations suivantes :

	mean	sd	MC error	2.5%	median	97.5%
alpha[1]	-0.0767	1.231	0.005636	-2.547	-0.06129	2.354
alpha[2]	-1.421	1.449	0.009086	-4.469	-1.341	1.205
alpha[3]	2.082	1.809	0.01339	-1.012	1.938	5.973
...	...	...	...	...	...	...
beta[1]	0.08984	0.6114	0.003652	-1.102	0.0261	1.541
beta[2]	-0.1551	0.6336	0.005537	-1.73	-0.0494	0.9878
beta[3]	0.06441	0.6089	0.002836	-1.167	0.01687	1.486
...	...	...	...	...	...	...
temps[1]	8.819	0.728	0.004971	7.407	8.811	10.28
temps[2]	2.507	0.7291	0.005068	1.079	2.501	3.962
sigma	3.552	0.282	0.001726	3.054	3.535	4.155
sigma.alpha	2.001	0.719	0.008664	0.693	1.953	3.56
sigma.beta	0.5167	0.3865	0.01235	0.0301	0.4393	1.46

Les estimations des paramètres *temps*[1] et *temps*[2] ont des intervalles de confiance entièrement positifs, cela signifie qu'il y a pour ce modèle un effet du facteur *temps*. De plus, comme l'estimation de *temps*[2] est plus faible que celle du paramètre *temps*[1], on en déduit que la mesure de l'intensité de fixation d'un marqueur radioactif est plus faible au temps *post*.

Les écarts-type des densités des paramètres *beta* sont plus faibles que ceux obtenus pour le modèle précédent. Cependant, *sigma.beta* est relativement faible, on en déduit qu'il n'y a pas d'effet du paramètre

beta, et donc que le nombre de points de mesure étudié n'influt pas de façon significative sur la variable réponse.

En visualisant les densités des lois a posteriori, on peut étudier la convergence de la chaîne de Markov. Ci-dessous sont représentés les densités de certains paramètres *alpha* et *beta* ainsi que des paramètres *temps*.

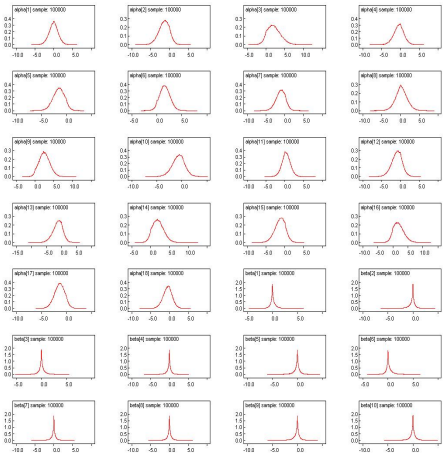


FIGURE 5.2: Densité des lois a posteriori

Les densités des lois a posteriori ressemblent fortement à des densités gaussiennes et contrairement au modèle précédent, l'autocorrélation est faible pour chacun des paramètres.

La simple comparaison des densités dans les deux modèles ne permet pas de mettre l'un d'entre eux en avant. On étudie alors les résidus obtenus par les deux modèles pour pouvoir les comparer.

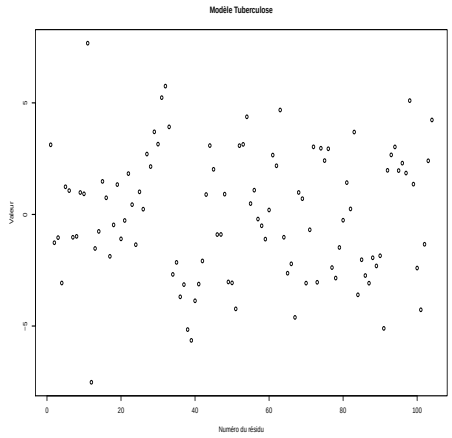


FIGURE 5.3: Residus Modèle 1

Les résidus pour ce modèle, sont symétriques par rapport à la valeur zéro, de plus, ils sont répartis de manière homogène et leurs valeurs sont comprises entre -5 et 5. La condition d'homoscédasticité semble donc ici respectée.

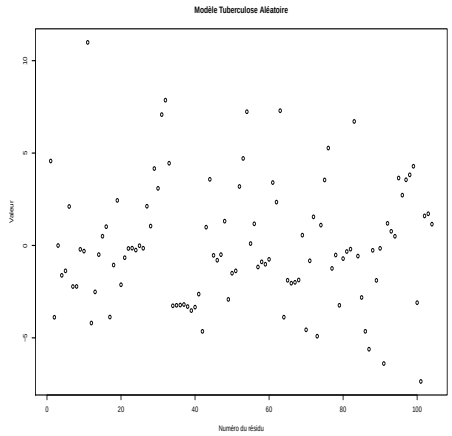


FIGURE 5.4: Résidus Modèle avec $Sujet$

Contrairement au modèle précédent, les valeurs des résidus sont comprises entre -10 et 10, cela engendre une plus grande étendue. La condition d'homoscédasticité semble cependant respectée, mais la non symétrie des résidus par rapport à la valeur zéro, implique une qualité discutable pour ces résidus.

Le modèle où le facteur *beta* n'a pas d'hyperprior est donc à préférer, en vue de la qualité des résidus.

Chapitre 6

Croissance infantile

L'écriture d'un modèle mixte peut se faire de différentes façons. On souhaite ici dans un cas préciser explicitement la matrice de variance-covariance liée aux répétitions par sujet sur le temps, puis dans l'autre, supposer que cette matrice de variance-covariance a pour distribution une loi inverse de Wishart.

6.1 Modèle avec variance explicite

Mathématiquement, notre modèle s'écrit

$$Y_{it} = A_t + \gamma \times sexe_i + (AB)_{it} + \mathcal{E}_{it}$$

i étant le numéro du sujet, i=1,..,27 et t représentant le temps, t=1,..,4 .
Les temps d'observation, 1,2,3 et 4 représentés par le facteur A_t, font référence à l'âge de l'enfant ;
le *temps* 1 représente la mesure faîte à l'âge de 8 ans, le *temps* 2 celle à l'âge de 10 ans, *temps* 3 à l'âge de 12 ans et *temps* 4 à l'âge de 14 ans. L'intervalle de 2 ans entre chaque mesure a été choisi arbitrairement car l'expérimentateur n'avait aucune connaissance a priori. On suppose le facteur A_t à effet aléatoire. $(AB)_{it}$ représente l'interaction entre le *temps* t et le *sujet* i. Le terme $\gamma \times sexe_i$ prend la valeur γ lorsque l'enfant est un garçon, et prend la valeur 0 lorsque l'individu est une fille. Le facteur *sexe* prend en effet la valeur 0 ou 1 selon le sexe de l'individu.

On écrit ainsi sous WinBUGS :

$$\boxed{\begin{array}{l} Y[i,t] \sim dnorm(mu[i,t], tau) \\ mu[i,t] \leftarrow beta[i, temps[t]] + gamma * sex[i] + alpha[temps[t]] \end{array}}$$

avec ici $beta[i, 1:4] \sim dmnorm(M[1:4], T[1:4, 1:4])$. T représente l'inverse de S qui est la matrice de variance-covariance de $beta$, donc T est la précision.

Dans cette première écriture du modèle, on précise la matrice S :

$$S = \begin{pmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 & \sigma^2\rho^3 \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 \\ \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ \sigma^2\rho^3 & \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 \end{pmatrix}$$

avec ρ qui a pour distribution une loi uniforme, $\rho \sim \mathcal{U}(0.01, 0.99)$, $\sigma^2 = 1/prec$ et $prec$ qui a pour distribution une loi gamma de paramètres $\mathcal{G}(0.01, 0.01)$.

On monitore à présent 5000 itérations pour obtenir les estimations de nos paramètres (5000 itérations car beaucoup de paramètres sont à estimer, et l'écriture du modèle implique que le processus est très lent). La matrice de variance-covariance estimée est alors :

$$\hat{S} = \begin{pmatrix} 1.43 & 0.6552 & 0.377 & 0.2452 \\ 0.6552 & 1.43 & 0.6552 & 0.377 \\ 0.377 & 0.6552 & 1.43 & 0.6552 \\ 0.25 & 0.377 & 0.6552 & 1.43 \end{pmatrix}.$$

L'estimation de σ^2, $\hat{\sigma^2}$ est ainsi égale à 1,43.

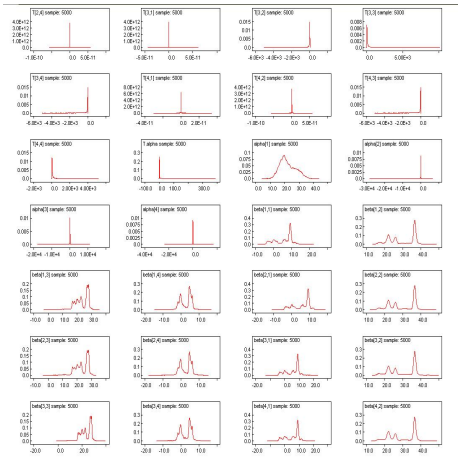


FIGURE 6.1: Densité du modèle mixte

Les densités des lois a posteriori ne semblent pas être gaussiennes, en effet, la plupart d'entre elles ne sont pas symétriques et on aperçoit des variations locales notamment pour les paramètres *beta*.

On étudie alors une autre écriture de ce modèle.

6.2 Modèle Wishart

Sous le logiciel WinBUGS, on garde presque la même écriture que précédemment :

$$\boxed{\begin{array}{l} Y[i,t] \sim dnorm(mu[i,t], tau) \\ mu[i,t] \leftarrow beta[i, temps[t]] + gamma * sex[i] + alpha[temps[t]]. \end{array}}$$

Pour la moyenne du paramètre *beta*, on va cette fois-ci supposer qu'elle est fixe. De plus, la matrice de variance-covariance du paramètre *beta* a pour distribution une loi inverse Wishart, ce qui signifie que son inverse, donc sa précision, a pour distribution une loi de Wishart ; $T[1:4, 1:4] \sim W_4(S[,], 4)$. La moyenne du paramètre *beta* étant également modifiée, le paramètre M représentant cette moyenne n'est ici plus aléatoire mais constant, égal au vecteur $(10, 15, 14, 6)$. Cette modification a été mise en place en vue de comparer le modèle précédent avec le modèle modifié.

Après avoir monitoré 100 000 itérations, on obtient alors comme estimation de la matrice de variance-covariance :

$$\hat{S} = \begin{pmatrix} 3.127 & -0.439 & 0.7205 & 0.05545 \\ -0.439 & 4.614 & 0.9395 & -0.3214 \\ 0.7205 & 0.9395 & 4.378 & -1.239 \\ 0.05545 & -0.3214 & -1.239 & 4.145 \end{pmatrix}.$$

Les estimations de $\sigma^2, \sigma^2\rho, \sigma^2\rho^2 et \sigma^2\rho^3$ sont ici plus faibles que celles obtenues précédemment.
$\hat{S}$ est bien une matrice symétrique, mais les termes de la diagonale ne sont pas tous égaux, ce qui signifie que $Var(\beta[i,1]) \neq Var(\beta[i,2]) \neq Var(\beta[i,3]) \neq Var(\beta[i,4])$ pour tout i.

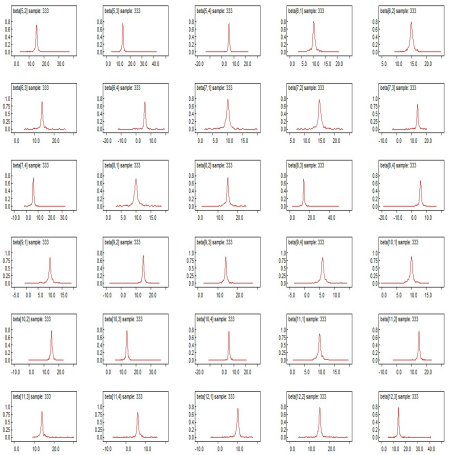


FIGURE 6.2: Densité du modèle mixte avec l'utilisation du n.thin

Ces densités des loi a posteriori ont été obtenues après avoir utilisé $n.thin = 300$. En effet, à cause de certaines valeurs importantes des *beta*, sans l'utilisation du $n.thin$, on obtenait des densités semblables à celles obtenues pour le paramètre T dans le modèle précédent.

Les densités obtenues ici sont donc plus proches de densités gaussiennes que celles obtenues précédemment. Les autocorrélations obtenues sont par ailleurs faibles.

Au vue de la qualité des densités obtenues, le modèle qui ajuste le mieux les données est celui où T a pour distribution une loi de Wishart, c'est-à-dire que la matrice de variance-covariance S, a pour distribution une loi inverse de Wishart.

Chapitre 7

Interprétation des résultats

Un modèle hiérarchique, au sens où nous l'étudions ici, est un modèle à multi-niveaux emboîté. Sur les logiciels R et SAS, cet emboîtement est visible au travers de la notation () ou |. Lorsque l'on modèlise un tel modèle sur le logiciel WinBUGS, il n'est pas nécessaire de faire apparaître explicitement cet emboîtement. Cela nécessite cependant d'entrer les données de manière à prendre en compte implicitement cet emboîtement. A travers l'étude de nos modèles, on a pu constater que, malgré des écritures différentes d'un même modèle, on obtient les mêmes prédictions si les hyperpriors sont respectés dans les écritures.

L'écriture d'un modèle mixte peut se faire de différentes manières sous le logiciel WinBUGS, notamment en modifiant la structure de la matrice des résidus. Cette structure est définie par le contexte et par rapport à l'objectif de l'étude. La loi de Wishart, utilisée ici pour l'inverse de la matrice de variance-covariance, est utilisée de manière récurrente en analyse bayésienne. Cela n'est peut être pas forcément le cas lors d'implémentation fréquentiste.

L'utilisation de la loi normale avec une grande variance a permis de ne pas ajouter d'information supplémentaire et de ne prendre en compte que celle apportée par la vraisemblance. La loi normale utilisée fréquemment comme loi a priori aurait pu être remplacée par d'autres lois, en particulier par la loi uniforme et la loi non-informative de Jeffreys. La loi normale a été choisie pour sa particularité de conjugaison : on s'attend ainsi à observer des densités proches de gaussiennes pour les lois a posteriori.

On a également constaté que les résultats obtenus par méthode fréquentiste étaient semblables à ceux obtenus par méthode bayésienne. Or, on a vu que les hypothèses sur les résidus n'étaient pas vérifiées. On pourrait donc penser que l'analyse bayésienne permet de contourner certaines hypothèses indispensables en analyse fréquentiste. Cependant, de nombreux problèmes se posent par rapport au choix de la loi a priori. Son rôle est fondamentale mais la question est ouverte sur le choix de celle-ci. Il se peut que dans certains cas, le choix de la loi a priori soit naturellement déterminé par le contexte de l'étude. Le cas échéant, par manque d'information, il sera préférable d'utiliser les lois non-informatives. Cela reste controversé. En effet la question est ouverte sur le fait que seul la vraisemblance des données est exploitée.

Annexes

Annexe A

Théorème ergodique

Avant d'introduire le théorème ergodique, il nous faut définir quelques notions qui apparaissent dans cette définition.

La loi de probabilité $\pi(..)$ sur E est **stationnaire** pour la chaîne de Markov homogène $\{X_t\}_{t>=0}$ si

$$\int K(x,y)\pi(x)dx = \pi(y).$$

En d'autres mots, on dit d'un processus (x_t) qu'il est stationnaire, si la distribution de $(x_{t+1},..,x_{t+d})$ est la même que la distribution de $(x_1,,x_d)$ pour tout t et pour tout d.

Soit $K_t(..,..)$ une chaîne de Markov. On définit **le noyau de transition** de cette chaîne par

$$K_t(x,A) = \mathbb{P}(X_{t+1}\in A|X_t = x_t).$$

La chaîne de Markov $\{X_t\}_{t\geq 0}$ est **homogène** si le noyau de transition ne dépend pas de t, c'est-à-dire si :

$$\forall t \geq 0, \mathbb{P}(X_{t+1}\in A|X_t = x_t) = \mathbb{P}(X_1\in A|X_0 = x_0).$$

La chaîne de Markov $\{X_t\}_{t>=0}$ est π**-irréductible** si $\forall x\in E, \forall A\in T(E)$ tel que $\pi(A) > 0$, on a

$$\mathbb{P}_{X_0=x_0}(\inf_t\{X_t\in A\} < +\infty) > 0.$$

Théorème ergodique

Si la chaîne de Markov $\{X_t\}_{t\geq 0}$ est π-irréductible, de loi stationnaire $\pi(.)$ et homogène, alors pour toute fonction f : E $\to \mathbb{R}$ tel que $\int |f(x)|\pi(x)dx < +\infty$, on a

$$\mathbb{P}_{X_0=x_0}\Big(\frac{1}{T}\sum_{t=1}^{T} f(X_t) \xrightarrow[T\to+\infty]{} \int f(x)\pi(x)dx\Big) = 1.$$

En fait, lorsque des chaînes de Markov sont ergodiques, cela signifie que la loi de θ_m converge vers $\pi(.|\underline{x})$ pour presque toute valeur initiale θ_0 ; l'influence de la valeur initiale disparaît.

Annexe B

Loi conjuguée

Une famille de lois de probabilité F est dite conjuguée pour le modèle statistique considéré, si pour toute loi a priori dans F, la loi a posteriori appartient elle aussi à F.

Voici quelques exemples, regroupés dans un taleau, de lois conjuguées :

Vraisemblance $f(y\|paramètres)$	Loi a priori	Loi a posteriori
Binomiale (N,p) avec N connu	Beta(α,β)	Beta($\alpha + n\overline{y}$,$\beta + nN - n\overline{y}$)
Poisson(λ)	Gamma(a,b)	Gamma($a + n\overline{y}$,$b + n$)
Exponentielle(p)	Gamma(a,b)	Gamma($a + n$,$b + n\overline{y}$)
Normal(m,τ) avec τ connu	Normal(μ,ν)	Normal($\frac{\overline{y}n\tau+\mu\nu}{n\tau+\nu}$,$n\tau + \nu$)

Annexe C

Différents Diagnostics

C.1 Autocorrélation

Soit $\{Z_k\}_{k\geq 0}$ un processus stationnaire. Si n est fixé, le coefficient d'autocorrélation de rang n est le coefficient de corrélation entre Z_k et Z_{k+n} qui ne dépend pas de k sous l'hypothèse de stationnarité et qui s'écrit :

$$\rho_n = \frac{Cov(Z_k, Z_{k+n})}{\sqrt{Var(Z_k)Var(Z_{k+n})}}.$$

Lorsque les autocorrélations sont élevées, cela indique que la chaîne de Markov est faiblement mélangeante : Un processus Z est mélangeant si la dépendance entre Z_n et Z_{n+k} tend vers 0 quand k tend vers l'infini.

Ainsi, une chaîne faiblement mélangeante peut impliquer une convergence lente de la chaîne.

C.2 Diagnostic de Gelman et Rubin

Il est nécessaire pour ce diagnostic de simuler plusieurs chaînes de Markov, en effet, son principe est basé sur une comparaison des variances inter et intra chaîne. Si la convergence est acquise, les variances intra et inter-chaîne doivent être proches. Pour évaluer cette convergence, on utilise la statistique de Gelman et Rubin :
Soit w_{ij} le i-ème $(i = 1, ..., n)$ élément de la chaîne j $(j = 1, ..., m)$.

Variance inter-chaînes :

$$B = \frac{n}{m-1}\sum_{j=2}^{m}(\overline{w_j} - \overline{w})^2.$$

où $\overline{w_j}$ est la moyenne des n réalisations de la chaîne j et $\overline{w}$ la moyenne des mn réalisations de toutes les chaînes.

Variance intra-chaînes :

$$W = \frac{1}{m(n-1)} \sum_{i=1}^{n} \sum_{j=2}^{m} (w_{ij} - \overline{w_j})^2.$$

B et W sont des estimateurs convergents de la variance de w :

$$\hat{\sigma}_w^2 = (1 - \frac{1}{n})W + \frac{1}{N}B.$$

D'où la statistique de diagnostic :

$$R = \sqrt{\frac{\hat{\sigma}_w^2}{W}}.$$

Ainsi, si la convergence est acquise, cette statistique doit être proche de 1 :

$$R \xrightarrow[n \to +\infty]{} 1.$$

C.3 Diagnostic d'Heidelberg et de Welch

Dans un premier temps, le Diagnoctic d'Heidelberg et de Welch teste la stationnarité de la chaîne de Markov, la statistique de test utilisée est basée sur la statistique du test de Cramer-von Mises. Lorsque la stationnarité est acceptée, la procédure s'arrête, dans le cas contraire on réitère le test après avoir supprimé les premiers 10% de la chaîne, puis 20% ... jusqu'à 50%. Si à la fin du processus le test est encore rejeté, il faut alors augmenté la taille de la Chaîne de Markov, c'est-à-dire le nombre d'itérations. La portion de chaîne retenue lorsque le test de stationnarité n'a pas été rejeté jusqu'à la fin, est utilisée pour estimer la moyenne et pour effectuer un test à partir d'un intervalle de confiance de celle-ci ; si la demi-largeur de l'intervalle obtenu est plus petite que $\mathcal{E}\hat{\mu}$, avec $\mathcal{E}$ choisi par l'utilisateur, le test conserve H_0 c'est-à-dire l'hypothèse de stationnarité.

Divers diagnostics pourraient encore être exploités ici, tel que le diagnostic de Geweke, mais également le diagnostic de Raftery-Lewis. Le premier a pour objectif de juger de la convergence de la moyenne de chaque paramètre, quant au deuxième, il renvoie le nombre minimum d'itérations requis nécessaire pour estimer les paramètres.

C.4 Utilisation de certains diagnostics

Voici les résultats obtenus de certains diagnostics dans le paragraphe 4.1.1 :
La première figure montre que l'autocorrélation est relativement faible pour chacun des paramètres. La courbe dessinée par Winbugs (figure 2) pour le diagnostic de Gelman et Rubin, montre l'évolution de la statistique de Gelman et Rubin au cours du temps. On remarque ici qu'elle converge vers 1 pour les paramètres des 2 chaînes de Markov simulées. Ainsi donc, la convergence des chaînes est vérifiée, et donc celle de la première chaîne en particulier. Quant au diagnostic de Heidelberg et de Welch (obtenu par la librarie R2WinBUGS de R qui utilise en fait WinBUGS via R), on remarque que les tests de stationnarité sont vérifiés pour chacun des paramètres. La chaîne de Markov est alors bien stationnaire.

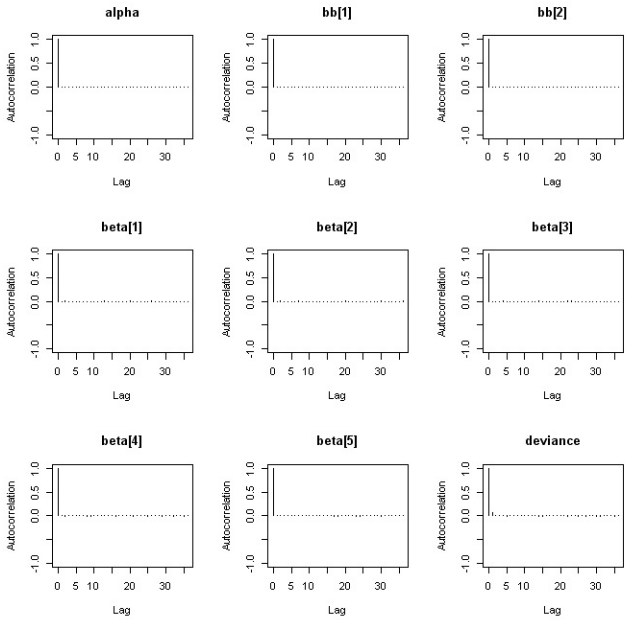


FIGURE C.1: Autocorrélation

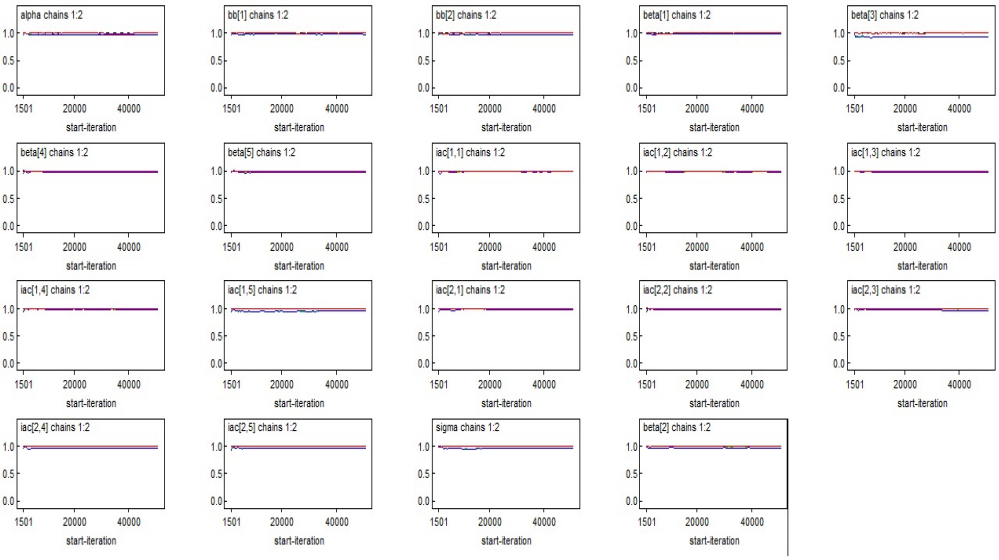


FIGURE C.2: Diagnostic de Gelman et Rubin

	Stationarity	start	p-value
	test	iteration	
alpha	passed	1	0.4079
bb[1]	passed	1	0.6542
bb[2]	passed	1	0.9365
beta[1]	passed	1	0.5494
beta[2]	passed	1	0.1717
beta[3]	passed	1	0.6602
beta[4]	passed	1	0.2193
beta[5]	passed	9501	0.0528
deviance	passed	1	0.3789
iac[1,1]	passed	1	0.8370
iac[1,2]	passed	1	0.1204
iac[1,3]	passed	1	0.7113
iac[1,4]	passed	1	0.1758
iac[1,5]	passed	1	0.2333
iac[2,1]	passed	1	0.7394
iac[2,2]	passed	1	0.4723
iac[2,3]	passed	1	0.6206
iac[2,4]	passed	1	0.4753
iac[2,5]	passed	1	0.1206
	Halfwidth Mean test	Halfwidth	
alpha	passed	89.66	0.1445
bb[1]	passed	41.19	0.1738
bb[2]	passed	48.14	0.1818
beta[1]	passed	2.92	0.1917
beta[2]	passed	11.25	0.2054
beta[3]	passed	18.25	0.2141
beta[4]	passed	25.79	0.2096
beta[5]	passed	31.65	0.1807
deviance	passed	1170.54	0.0271
iac[1,1]	passed	11.36	0.2463
iac[1,2]	passed	51.50	0.2671
iac[1,3]	passed	84.24	0.2459
iac[1,4]	passed	119.02	0.2422
iac[1,5]	passed	147.97	0.2405
iac[2,1]	passed	17.58	0.2440
iac[2,2]	passed	59.76	0.2579
iac[2,3]	passed	99.04	0.2516
iac[2,4]	passed	137.79	0.2525
iac[2,5]	passed	167.66	0.2614

Voici un exemple d'autocorrélations élevées. Ce sont en fait les autocorrélations des paramètres *beta* du paragraphe 5.1 .

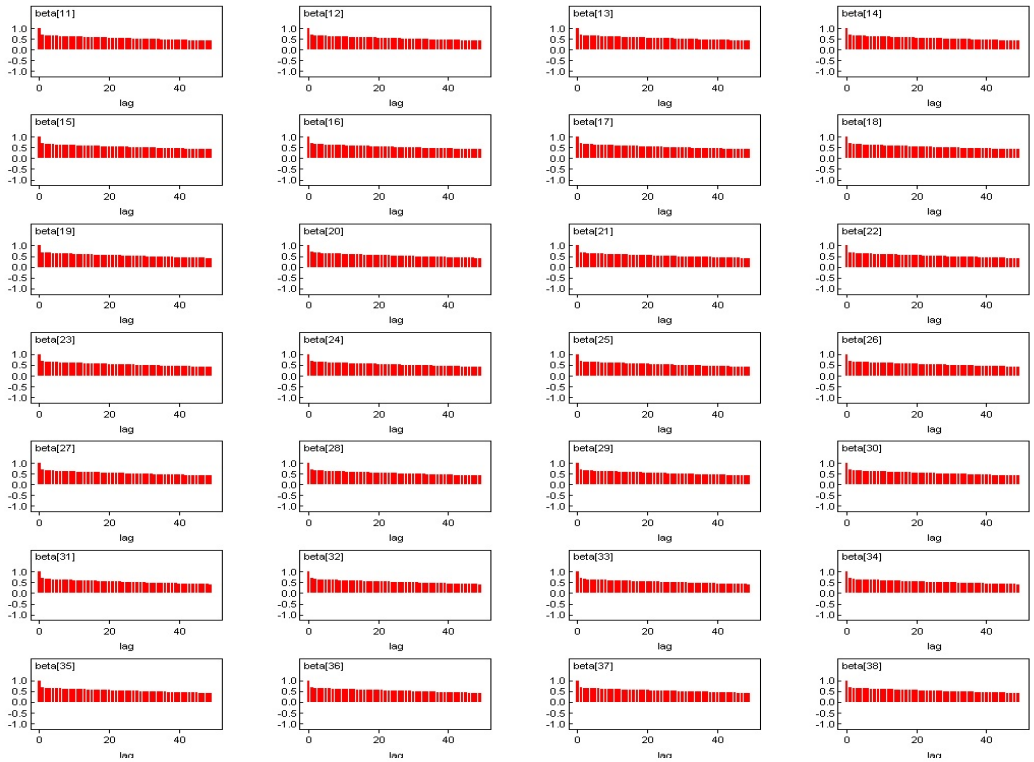


FIGURE C.3: Autocorrélations

Annexe D

Vérification que les prédictions sont égales dans le 4.1

Le tableau suivant regroupe les prédictions obtenues par le logiciel WinBUGS ainsi que leurs intervalles de confiance pour les modèles étudiés au paragraphe 4.1.1 et 4.1.2 sur l'étude de la croissance chez les rats.

On remarque que pour les rats appartenant à un même groupe les prédictions et leurs intervalles de confiance sont à peu de choses près exactement les mêmes. En effet, dans l'écriture de notre modèle le facteur *Rat* n'apparaît pas, c'est pour cela qu'il n'y a pas de différence entre les rats d'un même groupe.

	Modèle	avec	Contrainte	Modèle	sans	Contrainte
node	mean	2.5%	97.5%	mean	2.5%	97.5%
predict[1,1]	145.2	138.9	151.5	145.2	138.9	151.5
predict[1,2]	193.9	187.7	200.2	193.7	187.4	200.0
predict[1,3]	233.5	227.2	239.8	233.4	227.2	239.8
predict[1,4]	275.7	269.4	281.9	275.7	269.4	282.1
predict[1,5]	310.3	303.9	316.6	310.5	304.2	316.8
predict[2,1]	145.2	138.9	151.5	145.2	138.9	151.5
predict[2,2]	193.9	187.7	200.2	193.7	187.4	200.0
predict[2,3]	233.5	227.2	239.8	233.4	227.2	239.8
predict[2,4]	275.7	269.4	281.9	275.7	269.4	282.1
predict[2,5]	310.3	303.9	316.6	310.5	304.2	316.8
predict[3,1]	145.2	138.9	151.5	145.2	138.9	151.5
predict[3,2]	193.9	187.7	200.2	193.7	187.4	200.0
predict[3,3]	233.5	227.2	239.8	233.4	227.2	239.8
predict[3,4]	275.7	269.4	281.9	275.7	269.4	282.1
predict[3,5]	310.3	303.9	316.6	310.5	304.2	316.8
predict[4,1]	145.2	138.9	151.5	145.2	138.9	151.5
predict[4,2]	193.9	187.7	200.2	193.7	187.4	200.0
predict[4,3]	233.5	227.2	239.8	233.4	227.2	239.8
predict[4,4]	275.7	269.4	281.9	275.7	269.4	282.1
predict[4,5]	310.3	303.9	316.6	310.5	304.2	316.8
predict[5,1]	145.2	138.9	151.5	145.2	138.9	151.5
predict[5,2]	193.9	187.7	200.2	193.7	187.4	200.0
predict[5,3]	233.5	227.2	239.8	233.4	227.2	239.8
predict[5,4]	275.7	269.4	281.9	275.7	269.4	282.1
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
predict[28,4]	301.4	295.5	307.3	301.4	295.5	307.3
predict[28,5]	337.0	331.1	342.8	337.2	331.3	343.1
predict[29,1]	158.4	152.5	164.3	158.4	152.5	164.3
predict[29,2]	209.1	203.3	215.0	208.9	203.0	214.8
predict[29,3]	255.2	249.4	261.1	255.2	249.3	261.1
predict[29,4]	301.4	295.5	307.3	301.4	295.5	307.3
predict[29,5]	337.0	331.1	342.8	337.2	331.3	343.1
predict[30,1]	158.4	152.5	164.3	158.4	152.5	164.3
predict[30,2]	209.1	203.3	215.0	208.9	203.0	214.8
predict[30,3]	255.2	249.4	261.1	255.2	249.3	261.1
predict[30,4]	301.4	295.5	307.3	301.4	295.5	307.3
predict[30,5]	337.0	331.1	342.8	337.2	331.3	343.1

Bibliographie

[1] *Exploiter l'approche hiérarchique bayésienne pour la modélisation statistique de structures spatiales*, Sophie Ancelet, 2008

[2] *Applied Mixed Models in Medicine*, Helen Brown and Robin Prescott, 1999

[3] Cours d'analyse bayésienne, Jean-luc Dortet, 2012

[4] *Doing Bayesian Data Analysis*, John K. Kruschke, 2011

[5] *Analyse bayésienne avec WinBUGS et R*, Stéphane Laurent, 2007

[6] *Bayesian Modeling Using Winbugs*, Ioannis Ntzoufras, 2009

[7] *Le choix Bayésien, Principes et pratique*, Christian P.Robert, 2006