Rev. Ivoir. Sci. Technol., 19 (2012) 59 – 71

ISSN 1813-3290, http://www.revist.ci

RÉGRESSION LOGISTIQUE DANS LES ESSAIS CLINIQUES PAR MCMC

Ahlam LABDAOUI * et **Hayet MERABET**

Département de Mathématiques, Université Mentouri-Constantine, Route d'Ain-El Bey, 25000 Constantine, Algérie

*Correspondance, e-mail : *ahlem_stat@live.fr*

RÉSUMÉ

Les méthodes de Monte Carlo par Chaîne de Markov (et abrégée en MCMC) sont apparues il y a 50 ans pour la physique statistique, ont des applications presque illimitées, même si ses performances varient largement, selon la complexité du problème. Elle tire son nom de l'idée que, pour produire des approximations acceptables d'intégrales et d'autres fonctions dépendant d'une loi d'intérêt, il suffit de générer une chaîne de Markov $(\theta^{(m)})$ de loi limite de la loi d'intérêt. Cette idée d'utiliser le comportement limite d'une chaîne de Markov apparaît à la même époque que la technique de Monte Carlo originelle. La simulation de Monte Carlo vise à envisager l'estimation par simulation « aléatoire » plutôt que par analyse algébrique. Notre étude est basée sur l'estimation par les algorithmes de Metropolis-Hasting et l'échantillonnage de Gibbs dans le cadre d'un modèle non linéaire généralisé, le modèle de la régression logistique. Nous avons utilisé le logiciel WinBUGS pour estimer les paramètres, et interpréter les résultats de données réelles.

Mots-clés : *statistique Bayésienne, méthodes MCMC, logiciel WinBUGS.*

ABSTRACT
LOGISTIC REGRESSION CLINICAL TRIALS BY MCMC

The methods of Monte Carlo by Chain of Markov (and abbreviated by MCMC) are emerged 50 years ago to statistical physics, has almost unlimited applications, although its performance varies widely depending on the complexity of the problem. It takes its name from the idea that in order to produce acceptable approximations of integrals and other functions dependent on a statute of interest, simply generate a Markov chain law limits interest law. This idea of using the limit behavior of a Markov chain is at the same time as the original Monte Carlo

technique. The Monte Carlo simulation is to consider the estimation by simulation "random" rather than algebraic analysis. Our study is based on the estimation algorithms of Metropolis-Hasting and Gibbs sampling in the context of a generalized non-linear model, the logistic regression model. We used the software WinBUGS to estimate the parameters, and interpret the results of actual data.

Keywords : *Bayesian statistics, MCMC Methods, WinBUGS software.*

I - INTRODUCTION

Dans le cadre bayésien, il n'existe pas de différence fondamentale entre l'observation et le paramètre d'un modèle statistique, tous deux étant considérés comme quantités variables. Donc, si on note x la donnée, de loi d'échantillonnage f(x\θ), et θ le paramètre du modèle considérés (plus éventuellement, les variables latentes), de loi a priori π une inférence formelle requiert la mise à jour de la distribution conditionnelle f(θ\x) du paramètre. La détermination de $\pi(\theta)$ et de $f(x\backslash\theta)$ donne $f(x,\theta)$ par

$$f(x, \theta) = f(x \backslash \theta) * \pi(\theta) \qquad (1)$$

Après avoir observé x, on peut utiliser le théorème de Bayes pour déterminer la distribution de θ conditionnellement aux données (ou loi a posteriori), (voir [2]).

$$\pi(\propto\backslash x)=\frac{f(x\backslash\theta)*\pi(\theta)}{\int f(x\backslash\theta)*\pi(\theta)\,d(\theta)} \qquad (2)$$

Pour l'approche bayésienne, toutes les caractéristiques de la loi a posteriori sont importantes pour l'inférence : moment, quantile, etc. Ces quantités peuvent souvent être exprimées en termes d'espérance conditionnelle d'une fonction de θ par rapport à la loi a posteriori

$$E(\mathrm{h}(\propto)\backslash x) = \frac{\int \mathrm{h}(\theta) f(x\backslash\theta)*\pi(\theta)\,d(\theta)}{\int f(x\backslash\theta)*\pi(\theta)\,d\theta} \qquad (3)$$

On peut calculer la loi a posteriori directement dans le cas simple ou bien on fait le calcul par simulation MCMC dans le cas où le calcul de l'intégrale est très complexe.

Dans notre travail nous présentons d'abord un modèle de régression non généralisé à savoir le modèle de régression logistique ensuite, nous posons les conditions nécessaires pour l'utilisation des algorithmes de Monte-Carlo par Chaine de Markov (MCMC) ensuite nous introduisons quelques algorithmes MCMC, en

particulier l'algorithme de Metropolis-Hastings, et la méthode d'échantillonnage de Gibbs. Enfin nous présentons les résultats numériques et leurs interprétations.

II - MÉTHODOLOGIE

II-1. Le modèle de régression logistique

II-1-1. Le modèle logit

Un modèle standard de régression qualitative et le modèle de régression logistique ou modèle logit, où la loi de y conditionnelle aux variables explicatives z $\in$ R^p est :

$$P(y=1) = 1 - P(y=0) = \frac{\exp(z^t\gamma)}{1+\exp(z^t\gamma)} \qquad (4)$$

Considérons le cas particulier ou z= (1,x) et γ= (α, β) : les variables aléatoires y_i à valeurs dans {0,1} sont associées à des variables explicatives xi est sont modélisées suivant une loi de Bernoulli de probabilité conditionnelle

$$y_i \backslash x_i \sim B\left(\frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}\right) \qquad (5)$$

Supposons que nos paramètres suivent a priori une loi impropre $\pi(\alpha, \beta) = 1$. La vraisemblance de notre modèle, pour un échantillon $(y_1, x_1),\ldots,(y_n,x_n)$, est égale à : $f(y_1,\ldots,y_n \backslash x_1,\ldots\ldots,x_n,\alpha,\beta) = \prod_{i=1}^{n} \frac{\exp\{(\alpha+\beta x_i)y_i\}}{1+\exp(\alpha+\beta x_i)}$ (6)

La loi a posteriori de (α, β), se déduit alors par application formelle du Théorème de Bayes (voir [10]).

$$f(y_1,\ldots,y_n \backslash x_1,\ldots\ldots,x_n,\alpha,\beta) = \prod_{i=1}^{n} \frac{\exp\{(\alpha+\beta x_i)y_i\}}{1+\exp(\alpha+\beta x_i)} \qquad (7)$$

II-1-2. Définition de la loi a posteriori

L'utilisation de lois a priori impropres, c'est-à-dire de mesure σ finie de masse infinie sur l'espace des paramètres, implique que la dérivation des lois a posteriori par la relation de proportionnalité

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta) \qquad (8)$$

n'est pas nécessairement acceptable pour mettre en œuvre un algorithme de Metropolis- Hastings sur $f(x|\theta)\pi(\theta)$,car la loi correspondante peut ne pas exciter, c'est-à-dire, $f(x|\theta)\pi(\theta)$ n'est pas forcément intégrable. On est confronté à cette difficulté pour l'échantillonnage de Gibbs, par exemple, qui contrairement à

l'algorithme de Metropolis-Hastings, fonctionne avec des lois conditionnelles extraites de $(\theta_1 \ldots\ldots \theta_q)$, elle-même représentée par la relation de proportionnalité ci-dessus. Il peut arriver que ces lois soient clairement définies et simulables, mais qu'elles ne correspondent pas à une loi jointe f, c'est-à-dire que n'est pas intégrable (voir [9], pour des exemples).

II-2. Les méthodes MCMC

II-2-1. L'algorithme de Metropolis-Hasting

L'algorithme de Metropolis-Hastings repose sur l'utilisation d'une densité conditionnelle $q(y|x)$ par rapport à la mesure dominante pour li modèle. Il ne peut être mis en pratique que si $q(.|x)$ est simulable rapidement et est, soit disponible analytiquement à une constante près indépendante de x, soit symétrique, c'est-à-dire tel que $q(y|x) = q(x|y)$. L'algorithme de Metropolis-Hastings associé à la loi objective π et la loi conditionnelle q produit une chaine de Markov $x^{(t)}$ fondé sur la transition suivante :

---

etant donné $x^{(t)}$

1. *générer* $y_t \sim q(y/x^{(t)})$,

2. *prendre* $x^{(t+1)}$

$$= \begin{cases} y_t & \text{avec probabilité} \quad \rho(x^{(t)}, y_t) \\ x^{(t)} & \text{avec probabilité} \quad 1 - \rho(x^{(t)}, y_t) \end{cases}$$

Ou

$$\rho(x^{(t)}, y_t) = \min\left\{\frac{\pi(y_t)}{\pi(x^{(t)})} \quad \frac{q(x^{(t)}/y_t)}{q(y_t/x^{(t)})}, 1\right\}$$

La loi de q est appelée loi instrumentale ou de proposition. Cet algorithme accepte systématiquement les simulations y_t telles que le rapport $\left(\pi(y_t)\middle|q(y_t|x^{(t)})\right)$ est supérieure à la valeur précédente $\left(\pi\left((x^{(t)})\right)\middle|q(x^{(t)}|y_t)\right)$. Ce n'est que dans le cas symétrique que l'acceptation gouvernée par le rapport $\pi(y_t)/\pi(x_t)$. (voir [8]).

II-2-2. L'échantillonnage de Gibbs

L'échantillonnage de Gibbs c'est un algorithme de simulation d'une loi $\pi(x)$ telle que :

- x admet une décomposition de la forme $x = (x_1,\ldots,x_n)$,
- les loi conditionnelles $\pi_i\ (.|(x_1,\ldots,x_{x-1},x_{x+1},\ldots,x_n))$ sont

simulables aisément.(voir [4])

Exemple : $(X,Y)\sim N(0, \sum)$, avec $\sum=\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$.

$X\,|\,Y = y\sim N(\rho x, 1-\rho^2)$

Principe de l'algorithme : Réactualisation "composante par composante".

Initialisation : $x_0 = (x_1^0,\ldots\ldots\ldots, x_n^0)$

A chaque étape $k \geq 0$:

- Simuler $X_1^{k+1}\sim\pi_1(.\,|\,X_2^k,\ldots\ldots\ldots,X_n^k)$
- ...
- Simuler $X_i^{k+1}\sim\pi_i(.\,|\,X_1^k,\ldots\ldots,X_{i-1}^{k+1},X_{i-1}^k,\ldots\ldots\ldots,X_n^k)$
- ...
- Simuler $X_n^{k+1}\sim\pi_1(.\,|\,X_1^{k+1},\ldots\ldots\ldots,X_{n-1}^{k+1})$

III - RÉSULTATS ET DISCUSSION

III-1. Principe

Dans cette section nous allons appliquer la méthode MCMC pour le modèle logit et pour cela nous utilisons le logiciel WinBUGS (Logiciel de simulation Bayésienne MCMC), (voir [5]) pour le traitement des données.

Notre étude est basée sur la comparaison entre une crème antiseptique et un Placebo, tel que le critère de jugement est la guérison d'une infection (voir [1]). Nous cherchons à estimer l'effet de la crème par rapport au placebo, le tableau suivant donne la réponse de 8 centres que nous avons considérés :

Tableau 1 : *les données traitées*

Centre	traitement	Réponse	
		Succès	Echec
1	Crème	11	25
	Placebo	10	27
2	Crème	16	4
	Placebo	22	10
3	Crème	14	5
	Placebo	7	12
4	Crème	2	14
	Placebo	1	16
5	Crème	6	11
	Placebo	0	12
6	Crème	1	10
	Placebo	0	10
7	Crème	1	4
	Placebo	1	8
8	Crème	4	2
	Placebo	6	1

III-1-1. Modèle 1

Certains centres peuvent avoir une plus grande probabilité de guérison, quel que soit le traitement employé.(voir[7]).
Soient r_i^T le nombre de succès dans le groupe T (Placebo ou Crème) dans le centre i, on peut écrire :

* $r_i^T \rightarrow$ Binomial (p_i^T, n_i^T),
* logit $(P_i^p) = \alpha - \beta/2 + u_i$,

* logit (P_i^e) = $\alpha + \beta/2 + u_i$,
* $u_i \rightarrow$ Normal (0,σ_u^2)

Dans ce modèle, l'OR de la crème versus Placebo est constant et égal à e^{β}, et l'hétérogénéité (dans la probabilité de succès) entre les centres est mesuré par σ_2^u.
Le modèle de WinBUGS va s'écrire :

```
Model
    {
    for(i in 1 : 8) {
    rp[i] ~ dbin(pp[i], np[i]) # Vraisemblance pour Placebo.
    rc[i] ~ dbin(pc[i], nc[i]) # Vraisemblance pour Crème.
    logit(pp[i]) <- alpha - beta / 2 + u[i] # Modèle Placebo.
    logit(pc[i]) <- alpha + beta / 2 + u[i] # Modèle crème.
    u[i] ~ dnorm(0.0, tau) # Effet aléatoire.

    }
    alpha ~ dnorm(0.0, 1.0E-6) # A priori vague pour alpha.
    beta ~ dnorm(0.0, 1.0E-6) # A priori vague pour beta.
    tau ~ dgamma(0.1, 0.1) # A priori vague pour la précision.
    sigma <- 1/ sqrt(tau)
    OR <- exp(beta)
    }
```

Nous procédons ensuite à l'estimation, cette fois sur 3 chaînes, avec 100 000 itérations (1 000 suffisaient) chacune, en conservant une itération sur 150. L'effet (suppose homogène) de la crème est estime à 0,757, avec un écart-type de 0,304.
Les sorties de WinBUGS sont les suivantes :

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
OR	2.235	0.7052	0.009858	1.191	2.123	3.88	1	3000
alpha	-0.8362	0.757	0.05594	-2.37	-0.859	0.8177	1	3000
beta	0.7575	0.3043	0.004185	0.1745	0.7528	1.356	1	3000
tau	0.4806	0.7091	0.01471	0.0881	0.3918	1.323	1	3000
u[1]	-0.1132	0.7887	0.05607	-1.787	-0.09136	1.514	1	3000
u[2]	1.893	0.8025	0.05561	0.2114	1.89	3.542	1	3000
u[3]	1.01	0.8064	0.05598	-0.7026	1.008	2.671	1	3000
u[4]	-1.427	0.9	0.05286	-3.341	-1.382	0.2316	1	3000
u[5]	-0.6119	0.8564	0.05491	-2.436	-0.5866	1.111	1	3000
u[6]	-1.871	1.044	0.04911	-4.184	-1.764	-0.05686	1	3000
u[7]	-0.849	0.9642	0.05094	-2.914	-0.7977	0.9901	1	3000
u[8]	1.856	0.9282	0.05278	0.1017	1.806	3.815	1	3000

Nous présentant maintenant la représentation graphique pour les paramètres alpha et beta

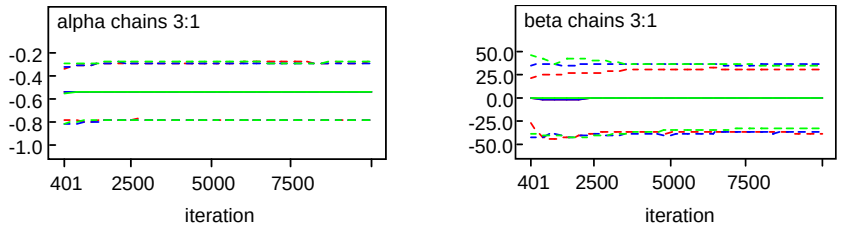

Figure 1 : *Les quantiles*

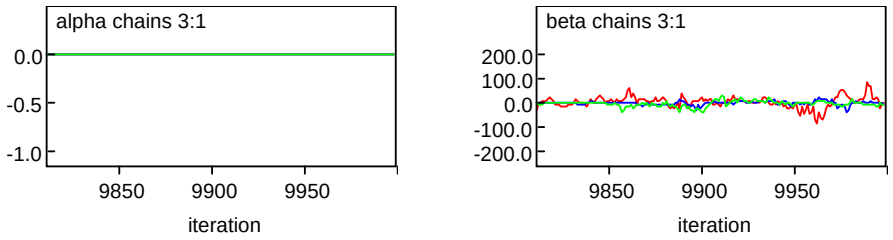

Figure 2 : *Les traces dynamiques*

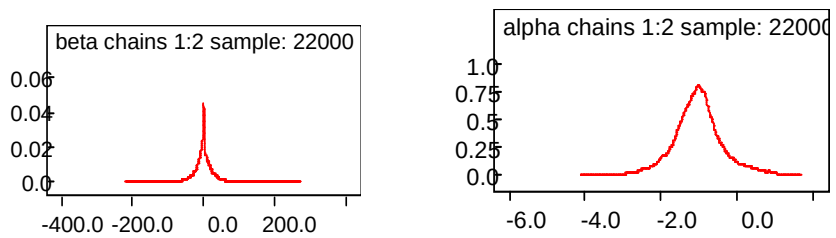

Figure 3 : *La densité de Kernel*

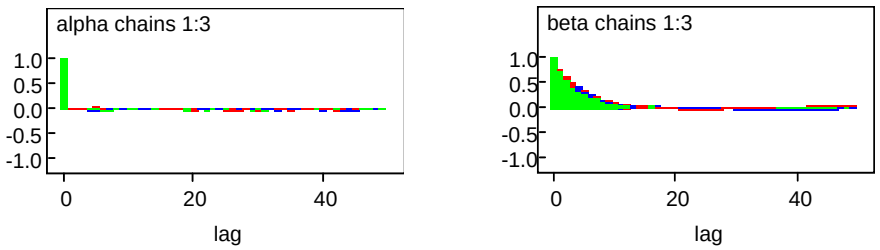

Figure 4 : *La fonction d'Auto corrélation*

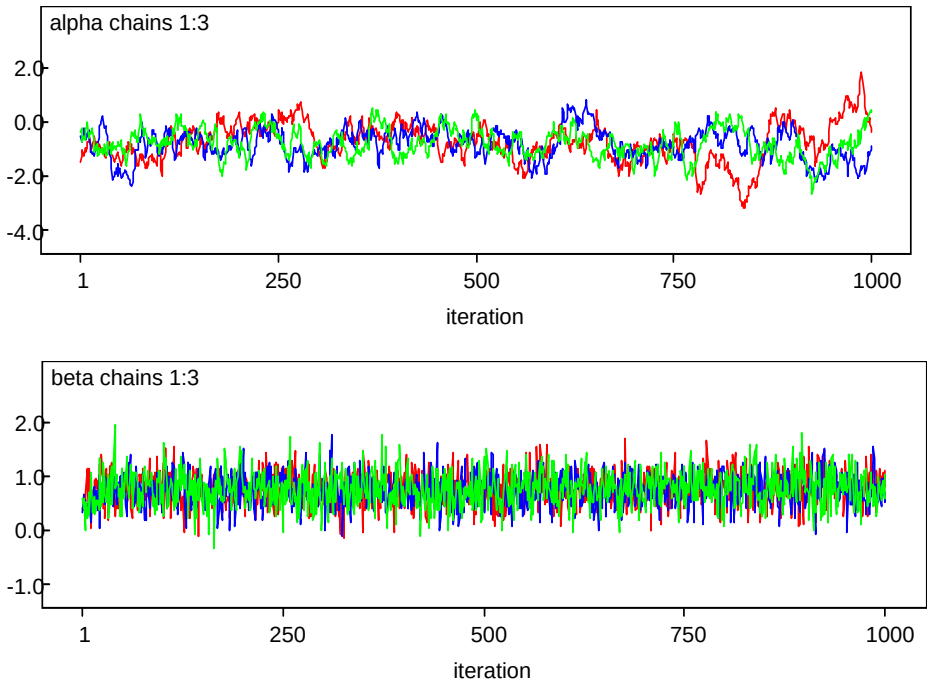

Figure 5 : *les séries temporelles*

III-1-2. Modèle 2

Certains centres peuvent avoir une plus grande probabilité de guérison, quel que soit le traitement employé. De plus, l'effet du traitement peut être hétérogène entre les centres. On complexifie le modèle précédent pour ne pas avoir à faire l'hypothèse d'homogénéité de l'effet de la crème en injectant un effet aléatoire sur β, *on obtient alors* :

* $r_i^T \rightarrow$ Binomial (p_i^T, n_i^T),
* logit (P_i^P) = $\alpha - (\beta + b_i)/2 + u_i$,
* logit (P_i^C) = $\alpha + (\beta + b_i)/2 + u_i$,
* $u_i \rightarrow$ Normal (0,σ_u^2),
* $b_i \rightarrow$ Normal (0,σ_b^2),

Dans ce modèle, l'OR de la crème versus Placebo est constant et égal à e^{β}, et l'hétérogénéité (dans la probabilité de succès) entre les centres est mesuré par σ_2^u.

Le modèle de WinBUGS va s'écrire :

```
model
    {
    for(i in 1 : 8) {
    rp[i] ~ dbin(pp[i], np[i]) # Vraisemblance pour Placebo.
    rc[i] ~ dbin(pc[i], nc[i]) # Vraisemblance pour Crème.
    logit(pp[i]) <- alpha – (beta+b[i]) / 2 + u[i] # Modèle Placebo.
    logit(pc[i]) <- alpha + (beta+b[i] / 2 + u[i] # Modèle crème.
    u[i] ~ dnorm(0.0, tau) # Effet aléatoire sur constante.
    b[i] ~ dnorm(0.0, taub) # Effet aléatoire sur coefficient.
    }
    alpha ~ dnorm(0.0, 1.0E-6) # A priori vague pour alpha.
    beta ~ dnorm(0.0, 1.0E-6) # A priori vague pour beta.
    tau ~ dgamma(0.001, 0.001) # A priori vague pour la précision.
    sigma <- 1/ sqrt(tau)
    taub ~ dgamma(0.001, 0.001)
    sigmab <- 1/ sqrt(taub)
    OR <- exp(beta)
    }
```

Nous procédons ensuite à l'estimation, avec les mêmes paramètres que la précédente.
L'effet (suppose homogène) de la crème est estime à 0,809, avec un écart-type de 0,388.
Les sorties de WinBUGS sont les suivantes :

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
OR	2.439	1.194	0.04907	1.114	2.205	5.073	1	3000
alpha	-1.019	0.6656	0.05523	-2.412	-0.9845	0.206	1	3000
b[1]	-0.1855	0.4092	0.02053	-1.343	-0.06496	0.3428	1	3000
b[2]	-0.06375	0.3686	0.01251	-0.9226	-0.01724	0.6018	1	3000
b[3]	0.1396	0.3834	0.01526	-0.4402	0.04621	1.134	1	3000
b[4]	-0.01424	0.3951	0.008387	-0.903	-3.466E-4	0.8319	1	3000
b[5]	0.2099	0.5129	0.02595	-0.3895	0.05254	1.659	1	3000
b[6]	0.04355	0.4719	0.01307	-0.8845	0.008931	1.142	1	3000
b[7]	-0.003306	0.8799	0.4091	0.009099	-0.9549	1.391E-4	1	3000
b[8]	-0.1955	0.5276	0.02515	-1.701	-0.04612	0.4741	1	3000
beta	0.8093	0.3881	0.01441	0.1083	0.7905	1.624	1	3000
u[1]	0.08317	0.7056	0.05601	-1.237	0.05257	1.577	1	3000
u[2]	2.073	0.7483	0.0573	0.6281	2.052	3.593	1	3000
u[3]	1.187	0.7324	0.055	-0.1551	1.16	2.706	1	3000
u[4]	-1.29	0.8366	0.05211	-2.918	-1.279	0.2839	1	3000
u[5]	-0.5029	0.7608	0.05188	-1.987	-0.5165	1.03	1	3000
u[6]	-1.75	0.9604	0.04458	-3.814	-1.691	0.0156	1	3000
u[7]	-0.7091	0.8792	0.04721	-2.478	-0.7052	1.019	1	3000
u[8]	2.026	0.8818	0.05519	0.3233	2.009	3.814	1	3000

Nous présentons maintenant la représentation graphique pour les paramètres alpha et beta.

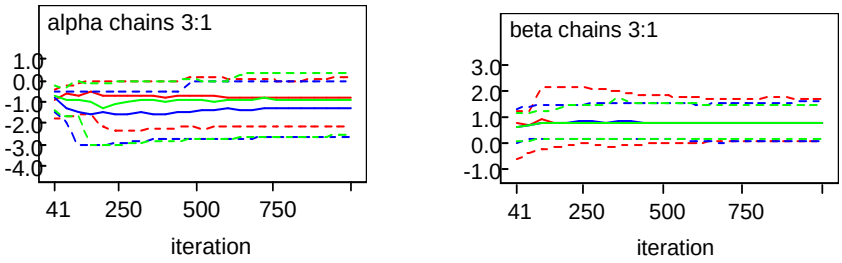

Figure 1 : *Les quantiles*

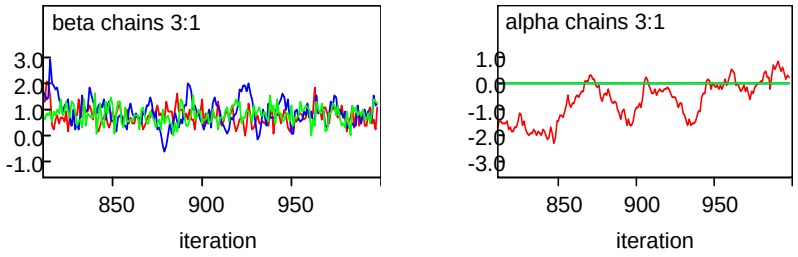

Figure 2 *: Les traces dynamiques*

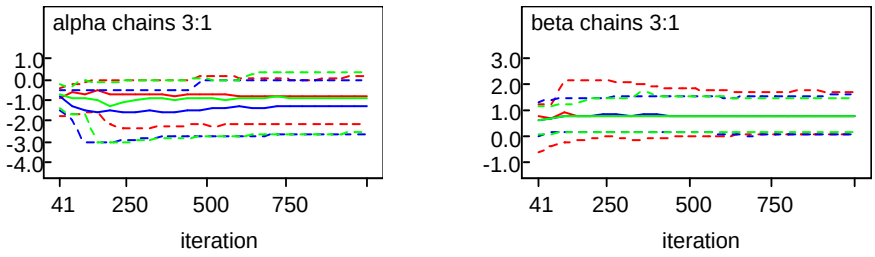

Figure 3 : *La densité de Kernel*

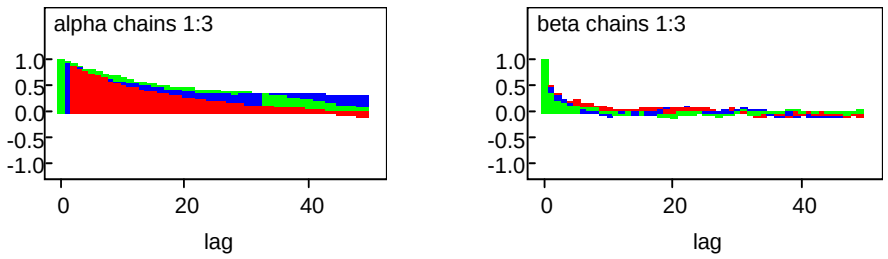

Figure 4 : *la fonction d'Auto corrélation*

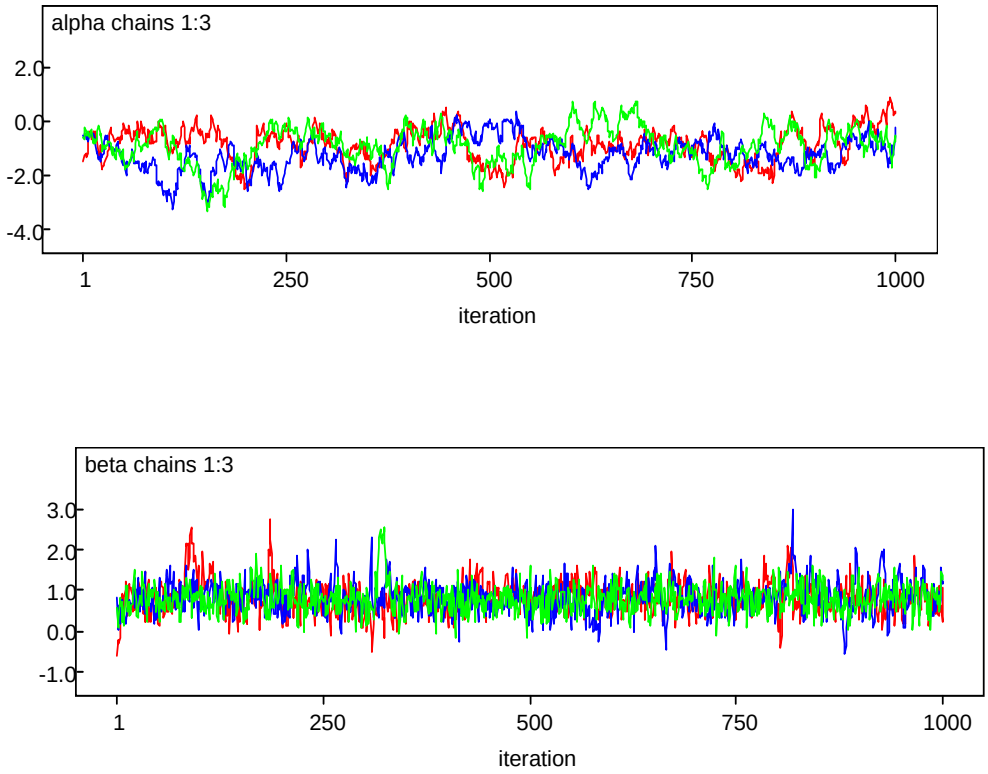

Figure 5 : *Les séries temporelles*

III-2. interprétation

Nous remarquons après la comparaison entre les deux modèles que le paramètre beta et son écart type sont presque égaux, c'est-à-dire l'ajout de l'effet aléatoire sur le coefficient beta ne change rien, ensuite l'OR dans les deux modèles (modèle(1) : OR=2.235, modèle(2) : OR= 2.439) est supérieur à 1 et l'intervalle de crédibilité compris entre 1.191 et 3.88 pour le modèle (1) et 1.114 et 5.07 pour le modèle (2) au niveau de 97.5% pour ces résultats on dit que notre crème antiseptique est efficace dans les deux modèles(1)et(2).

IV - CONCLUSION

Un des mérites de notre travail est d'avoir montré à l'aide de données expérimentales d'essais cliniques qu'un peut modéliser de façon naturelle et en tirer les inférences adéquates, à savoir estimer les paramètres dans le modèle logit avec et sans effet aléatoire, à l'aide des méthodes de Monte Carlo par chaine de

Markov (MCMC) d'autant plus que les performances des ordinateurs, ont rendu faisables des procédés de simulations efficaces et la disponibilité des programmes informatiques a facilité le calcul des probabilités a posteriori, qui étaient jusque là d'une complexité décourageante.

Remerciements

Nous tenons absolument à remercier Monsieur Pierre Druilhet, Professeur à l'université Blaise Pascal de Clermont Ferrand, France, pour son aide et ses fructueux conseils pour la réalisation de ce travail.

RÉFÉRENCES

[1] - Agresti A. *Categorical Data Analysis*. (2002).

[2] - Anas Altaleb, Christian P. Robert, *Analyse bayésienne du modèle logit : algorithme par tranches ou Metropolis-Hastings?*, revue de statistique appliquée, tome 49, n°4 (2001), p. 53-70.

[3] - Christian P, Robert and George Casella, *Monte Carlo Statistical Methods*, Springer, (2004).

[4] - C. Robert et G. Casella, *Monte Carlo Statistical Methods*, Springer, 2nd edition, (2004).

[5] - David J. Lunn, Andrew Thomaa, Nicky Best and David Spiegelhalter *WinBUGS – A Bayesian modeling framework: Concepts, structure, and extensibility, Statistics and Computing* (2000) **10,** 325–337.

[6] - Éric. Parent. Jacques Bernier, *Le raisonnement bayésien*, Springer-Verlag France, Paris, (2007).

[7] - Lionel Riou França, *statistique bayésienne*, INSERM U669, Mai 2009.

[8] - Robert, C.P. and Casella, G. *Monte Carlo Statistical Methods*. New York: Springer Verlag (1999).

[9] - Robert, C.P. *L'analyse statistique bayésienne Economica*, Paris (1992).

[10] - Ton J Cleophas, Aeilko H Zwinderman, Toine F Cleophas, *Statistics Applied to Clinical Trials* (2006).