Université de Montréal

Régression logistique bayésienne : comparaison de densités *a priori*

par

Alexandre Deschênes

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Statistique

août 2015

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé

Régression logistique bayésienne : comparaison de densités *a priori*

présenté par

Alexandre Deschênes

a été évalué par un jury composé des personnes suivantes :

Martin Bilodeau
(président-rapporteur)

Jean-François Angers
(directeur de recherche)

David Haziza
(membre du jury)

Mémoire accepté le:
19 juillet 2015

SOMMAIRE

La régression logistique est un modèle de régression linéaire généralisée (*GLM*) utilisé pour des variables à expliquer binaires. Le modèle cherche à estimer la probabilité de succès de cette variable par la linéarisation de variables explicatives. Lorsque l'objectif est d'estimer le plus précisément l'impact de différents incitatifs d'une campagne marketing (coefficients de la régression logistique), l'identification de la méthode d'estimation la plus précise est recherchée. Nous comparons, avec la méthode MCMC d'échantillonnage par tranche, différentes densités *a priori* spécifiées selon différents types de densités, paramètres de centralité et paramètres d'échelle. Ces comparaisons sont appliquées sur des échantillons de différentes tailles et générées par différentes probabilités de succès. L'estimateur du maximum de vraisemblance, la méthode de Gelman et celle de Genkin viennent compléter le comparatif. Nos résultats démontrent que trois méthodes d'estimations obtiennent des estimations qui sont globalement plus précises pour les coefficients de la régression logistique : la méthode MCMC d'échantillonnage par tranche avec une densité *a priori* normale centrée en 0 de variance 3,125, la méthode MCMC d'échantillonnage par tranche avec une densité Student à 3 degrés de liberté aussi centrée en 0 de variance 3,125 ainsi que la méthode de Gelman avec une densité Cauchy centrée en 0 de paramètre d'échelle 2,5

Mots clés : régression logistique, bayésien, densité *a priori*, simulation MCMC.

SUMMARY

Logistic regression is a model of generalized linear regression (*GLM*) used to explain binary variables. The model seeks to estimate the probability of success of this variable by the linearization of explanatory variables. When the goal is to estimate more accurately the impact of various incentives from a marketing campaign (coefficients of the logistic regression), the identification of the choice of the optimum *prior* density is sought. In our simulations, using the MCMC method of slice sampling, we compare different *prior* densities specified by different types of density, location and scale parameters. These comparisons are applied to samples of different sizes generated with different probabilities of success. The maximum likelihood estimate, Gelman's method and Genkin's method complement the comparative. Our simulations demonstrate that the MCMC method with a normal prior density centered at 0 with variance of 3.125, the MCMC method with a Student prior density with 3 degrees of freedom centered at 0 with variance of 3.125 and Gelman's method with a Cauchy density centered at 0 with scale parameter of 2.5 get estimates that are globally the most accurate of the coefficients of the logistic regression.

Keywords : logistic regression, Bayesian, *prior* density, MCMC simulation.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

REMERCIEMENTS

J'aimerais tout d'abord remercier mon directeur de recherche Jean-François Angers pour qui l'écoute et le soutient considérable étaient au rendez-vous tout au long de ce projet. Sans sa bonne humeur et son positivisme, l'aboutissement de ce mémoire aurait sans aucun doute été beaucoup plus pénible.

J'aimerais aussi remercier l'équipe *Forecasting and Analytics* de TD assurance et plus spécialement Catherine Paradis-Therrien pour avoir eu la chance de développer et de tester pleinement les notions de ce mémoire. Merci pour la confiance que vous m'avez accordée et que vous m'accordez encore.

Je remercie également tout le personnel et mes collègues/étudiants du Département de mathématiques et de statistique de l'Université de Montréal qui ont agrémenté mon séjour dans cette belle institution.

Merci à mes parents Martine et Carol qui m'ont toujours grandement supporté durant les nombreux défis que je me suis donnés au cours de ma vie. Merci à Laurie, Dan, Tom et Market qui amènent un équilibre nécessaire dans ma vie. Je remercie aussi ma copine Katherine pour qui la patience, la compréhension et l'encouragement ne semblent pas avoir de limite.

Finalement, j'aimerais remercier la personne la plus importante : Alex. Merci de constamment sortir de ta zone de confort pour avancer pleinement dans la vie.

INTRODUCTION

Afin d'accroître leur part de marché, des compagnies de toutes sortes utilisent certaines techniques marketing pour attirer de plus en plus de nouveaux clients. C'est le cas des techniques de marketing direct appliquées notamment dans le domaine de l'assurance. En effet, les assureurs (et bien d'autres types d'entreprises) utilisent le marketing direct dans le but d'augmenter leur part de marché. Cette technique de communication sert à cibler de nouveaux clients en leur proposant un message adapté et incitatif afin d'inviter les clients potentiels à soumissionner immédiatement pour un produit d'assurance. Puisque l'effet désiré est immédiat, son influence est plus facile à mesurer que d'autres approches marketing. En plus de vouloir mesurer l'efficacité de campagnes de marketing direct, il est intéressant de comparer simultanément différentes variantes de l'offre marketing sur des groupes d'individus différents afin de trouver l'offre donnant le meilleur coût d'acquisition d'un client. Par exemple, un assureur de dommage peut s'intéresser à déterminer le meilleur incitatif à offrir à un client pour lui faire compléter une soumission. Ainsi, attirer le client à soumissionner en lui offrant comme récompense une participation à un concours serait peut-être plus rentable qu'offrir comme récompense une réduction de prime. Mesurer l'efficacité des campagnes de marketing direct aide donc à faire des choix plus éclairés pour l'entreprise.

Évidemment, pour cibler de nouveaux clients nous devons disposer de leurs informations. Ne faisant pas partie de la banque de clients actifs de l'entreprise, il n'est pas rare de connaître seulement les informations de base des individus à cibler telles que leur nom, leur numéro de téléphone, leur adresse de résidence, leur adresse courriel et la date d'obtention de ces informations. Certaines de ces informations sont obsolètes pour la prédiction de l'efficacité des initiatives marketing, mais nécessaires pour la diffusion directe de l'offre aux clients potentiels et pour la compilation des résultats par la suite. Notre

intérêt s'est donc porté sur l'estimation de l'efficacité d'un incitatif sur la réponse d'individus. Afin de simplifier la conceptualisation, nous appellerons dorénavant une offre marketing un *traitement* et le groupe d'individus ciblés une *cohorte*.

En utilisant un modèle de régression logistique, l'effet de chaque traitement est estimable. Du côté classique, il existe des méthodes d'estimation traitées dans plusieurs livres dont celui de McCullagh et Nelder (1983). Plus récemment du côté bayésien, Genkin et coll. (2007) propose une méthode d'estimation simple et efficace qui utilise l'estimateur du maximum *a posteriori* (MAP). Une densité *a priori* passe-partout applicable pour des modèles de régression logistique a été discutée par Gelman et coll. (2008). Le choix d'une densité passe-partout est intéressant, mais nous aimerions aussi cibler la meilleure densité sous des conditions spécifiques en comparaison avec la méthode classique de l'estimateur du maximum de vraisemblance. C'est ce qui a été fait par Gordovil-Merino et coll. (2012). Des modèles de régression logistique classique et bayésien ont été comparés afin de déterminer l'approche qui conduit aux estimations les plus précises sous différentes conditions. Il compare ces approches sur de petits échantillons en variant certains facteurs sans toutefois faire varier la densité *a priori* employée pour l'approche bayésienne. De notre côté, nous proposons de comparer et de déterminer par simulation la densité *a priori* la plus précise en utilisant les méthodes MCMC dans le cadre d'études de cohortes cherchant à mesurer l'efficacité d'un incitatif sur la réponse d'individus.

L'idée est d'utiliser différents facteurs (type de densité, centralité, variance, nombres d'observations et coefficients de la régression) pour former différentes densités *a priori* comparées sur différents échantillons. Nous évaluons l'influence de ces facteurs sur l'estimation des effets des traitements et déterminons la densité *a priori* qui mène aux estimations les plus précises. Nous n'oublions certainement pas d'incorporer l'estimateur du maximum de vraisemblance (EMV) comme valeur de référence à nos comparaisons en plus des méthodes proposées par Gelman et coll. (2008) et Genkin et coll. (2007).

Relativement au modèle linéaire classique, le chapitre 1 définit les composantes des modèles linéaires généralisés pour ensuite mener au modèle de régression logistique adéquat pour notre situation. Le chapitre 1 se termine avec la méthode d'estimation numérique du maximum de vraisemblance utilisée

pour estimer les coefficients du modèle de régression logistique. Au chapitre 2, nous introduisons l'approche bayésienne. En posant les bases théoriques qui mènent aux méthodes MCMC, nous introduisons la méthode d'échantillonnage par tranche utilisée pour nos simulations. Des diagnostics de convergence y sont aussi discutés. Toujours au chapitre 2, nous voyons les méthodes proposées par Gelman et coll. (2008) et Genkin et coll. (2007) afin de conclure les différentes méthodes utilisées lors de nos simulations. L'influence de la densité *a priori* employée pour les méthodes MCMC et les facteurs influençant potentiellement la précision de l'estimation sont discutés au chapitre 3. Les résultats des simulations sont présentés au chapitre 4. Nous y déterminons la méthode d'estimation qui apporte les estimations les plus précises pour notre cas d'étude de cohortes cherchant à mesurer l'efficacité d'un incitatif sur la réponse d'individus. Enfin, une application sur des données marketing de produits d'assurance est présentée au chapitre 5.

Chapitre 1

RÉGRESSION LOGISTIQUE CLASSIQUE

Lorsqu'il est question de modéliser la relation entre une variable réponse et un ensemble de variables explicatives, une généralisation du modèle linéaire peut s'imposer pour représenter le système adéquatement selon la nature de la variable réponse. En permettant au modèle linéaire d'associer les variables explicatives à la variable réponse par une fonction de lien, la modélisation de l'espérance de la variable réponse avec les variables explicatives est possible.

Afin de mieux s'orienter, la section 1.1 fait le lien entre les modèles linéaires et les modèles linéaires généralisés. Notre cas d'étude de cohortes cherchant à mesurer l'efficacité d'un incitatif sur la réponse d'individus discuté à la section 1.2 nous amène à spécifier un modèle logistique comme modèle linéaire généralisé. Nous terminons ce chapitre avec la section 1.3 où la méthode d'estimation du maximum de vraisemblance des coefficients du modèle logistique est présentée.

1.1. COMPOSANTES DES MODÈLES LINÉAIRES GÉNÉRALISÉS (*GLM*)

Un aperçu des modèles linéaires généralisés est présenté dans cette section en introduisant tout d'abord ses composantes à partir du modèle linéaire.

Soit $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ le vecteur aléatoire de réponses pour n observations. Nous supposons que les observations Y_i sont des réalisations générées de façon indépendante et identiquement distribuée selon une certaine variable aléatoire. Nous définissons $\mathbf{X} = \left(\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T\right)^T$ la matrice des variables explicatives, où les vecteurs $\mathbf{X}_i = (x_{i1}, \ldots, x_{ip})^T$ sont les p variables explicatives associées à Y_i. Nous supposons que toutes les variables explicatives sont connues. Soit $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ le vecteur des coefficients inconnus associé aux vecteurs $\mathbf{X}_i$.

À des fins de simplicité de notation, l'ordonnée à l'origine est incluse dans $\mathbf{X}$. Ainsi, les x_{i1} seront utilisés comme ordonnée à l'origine pour chaque observation ($x_{i1} = 1$).

Pour une régression linéaire classique, nous avons la relation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \text{ où } \boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}),$$

où $\mathcal{N}_n$ représente la loi normale multidimensionnelle et $\mathbf{I}$ la matrice identité. Puisque $\boldsymbol{\epsilon}$ suit une densité normale multidimensionnelle, le vecteur d'observations $\mathbf{Y}$ suit lui aussi une densité normale multidimensionnelle et il en découle que

$$\mathbb{E}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}. \tag{1.1.1}$$

Les observations y_i sont donc indépendantes et identiquement distribuées de densités normales centrées en $\mathbf{X}_i^T\boldsymbol{\beta}$ et de variance σ^2.

Si nous modifions la nature du vecteur aléatoire $\mathbf{Y}$, il est possible que l'équation (1.1.1) ne soit plus valide. Il est probable que le prédicteur linéaire $\mathbf{X}\boldsymbol{\beta}$ ne soit plus égal à l'espérance de $\mathbf{Y}$. Réorganisons l'équation (1.1.1) afin d'introduire une fonction dite *fonction de lien*, notée $g(\cdot)$. La fonction de lien sert à généraliser la relation entre $\mathbb{E}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta})$ et le prédicteur linéaire $\mathbf{X}\boldsymbol{\beta}$ pour former ce que nous appelons les *modèles linéaires généralisés (GLM)*. Ainsi, nous avons

$$g(\mathbb{E}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta})) = \mathbf{X}\boldsymbol{\beta}.$$

Nous notons qu'en appliquant la fonction identité comme fonction de lien, nous retrouvons l'équation (1.1.1) pour la régression linéaire classique.

Les modèles linéaires généralisés permettent deux extensions par rapport au modèle linéaire classique ; premièrement, la distribution des Y_i peut provenir d'une famille exponentielle de loi, et deuxièmement les seules restrictions sur la fonction de lien sont la monotonie et la différentiabilité. Ainsi, Y_i n'est pas nécessairement de loi normale et la fonction de lien n'est pas nécessairement la fonction identité. Définissons ce qu'est la classe de famille exponentielle et les modèles GLM.

Définition 1.1.1 (Famille exponentielle)**.** *Soit* Y *un vecteur de mesures de dimension* n *et* Θ *un espace de paramètres de dimension* k*. Soient* c *et* h *des fonctions, respectivement, de* Y *et de* Θ *dans* $\mathbb{R}^+$ *et soient* R *et* T *des fonctions, respectivement de*

Y *et de* Θ *dans* $\mathbb{R}^k$. *Pour* $y \in Y$ *et* $\theta \in \Theta$, *la classe de lois de densité*

$$f_Y(y|\theta) = c(\theta)h(y)\exp\left\{R(\theta)^T \cdot T(y)\right\},$$

s'appelle une famille exponentielle de lois.

Définition 1.1.2 (Densité d'un modèle GLM)**.** *Soit* Y *une variable aléatoire,* θ *un paramètre réel et* ϕ *un paramètre réel. Soit* a *et* c *des fonctions connues et dérivables et* b *une fonction connue, trois fois dérivable dont sa première dérivée est inversible. La densité d'un modèle GLM prend la forme de*

$$f_Y(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}. \tag{1.1.2}$$

Le paramètre ϕ est généralement appelé paramètre de dispersion. S'il est connu, la densité (1.1.2) fait partie de la famille exponentielle de lois. Sinon, elle n'est pas nécessairement de cette famille. Supposons à partir de maintenant que ϕ est connu. Nous utilisons les relations suivantes pour calculer l'espérance et la variance de la densité d'un modèle GLM :

$$\mathbb{E}\left(\frac{\partial L}{\partial \theta}\right) = 0 \tag{1.1.3}$$

et

$$\mathbb{E}\left(\frac{\partial^2 L}{\partial \theta^2}\right) + \mathbb{E}\left(\left[\frac{\partial L}{\partial \theta}\right]^2\right) = 0, \tag{1.1.4}$$

où L représente la log-densité. À partir de (1.1.2), nous avons

$$L(\theta|y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi),$$

où

$$\frac{\partial L}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)} \tag{1.1.5}$$

et

$$\frac{\partial^2 L}{\partial \theta^2} = -\frac{b''(\theta)}{a(\phi)}. \tag{1.1.6}$$

Avec (1.1.3) et (1.1.5), nous avons

$$0 = \mathbb{E}\left(\frac{\partial L}{\partial \theta}\right) = \frac{\mathbb{E}(Y) - b'(\theta)}{a(\phi)},$$

de sorte que

$$\mathbb{E}(Y) = b'(\theta). \tag{1.1.7}$$

TABLEAU 1.1. Caractéristiques de certaines distributions de la famille exponentielle de loi

	Normal	Poisson	Binomial	Gamma
Notation	$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{P}(\lambda)$	$\mathcal{B}(n, p)/n$	$\mathcal{G}(\alpha, \beta)$
Domaine de y	$\mathbb{R}$	$\mathbb{N}$	$\left\{\frac{m}{n} \middle\vert m \in \mathbb{N}, m \leq n\right\}$	$\mathbb{R}^+$
θ	μ	$\log(\lambda)$	$\log\left(\frac{p}{1-p}\right)$	$-\frac{1}{\alpha}$
ϕ	σ^2	1	$\frac{1}{n}$	$\frac{1}{\beta}$
$b(\theta)$	$\frac{\theta^2}{2}$	e^θ	$\log\left(1+e^\theta\right)$	$-\log(-\theta)$
$c(y\vert\phi)$	$-\frac{1}{2}\left(\frac{y^2}{\phi}+\log(2\pi\phi)\right)$	$-\log(y!)$	$\log\left(\binom{n}{ny}\right)$	$\beta\log(\beta y)-\log(y)-\log(\Gamma(\beta))$
Espérance	θ	e^θ	$\frac{e^\theta}{1+e^\theta}$	$-\frac{1}{\theta}$
Lien canonique	θ	$\log(\theta)$	$\log\left(\frac{\theta}{1-\theta}\right)$	$-\frac{1}{\theta}$

De la même manière, avec (1.1.4), (1.1.5) et (1.1.6), nous avons

$$0 = \mathbb{E}\left(\frac{\partial^2 L}{\partial\theta^2}\right) + \mathbb{E}\left(\left[\frac{\partial L}{\partial\theta}\right]^2\right) = -\frac{b''(\theta)}{a(\phi)} + \mathbb{E}\left(\frac{Y^2 - 2b'(\theta)Y + b'(\theta)^2}{a(\phi)^2}\right),$$

de sorte que

$$\mathbb{V}(Y) = \mathbb{E}\left(Y^2\right) - \mathbb{E}(Y)^2 = b''(\theta)a(\phi). \tag{1.1.8}$$

Par l'équation (1.1.7), nous définissons la fonction suivante.

Définition 1.1.3 (Lien canonique)**.** *Lorsque les observations d'un vecteur de mesures* Y *de dimension* n *suivent une loi de la forme* (1.1.2)*, la fonction de* lien canonique *est définie comme* $g(\mathbb{E}(Y)) = (b')^{-1}(\mathbb{E}(Y))$*, où* $(b')^{-1}$ *est définie comme la fonction inverse de* b'*.*

Pour un modèle GLM, il est nécessaire de spécifier :

(1) la loi de probabilité des observations Y_i

(2) la fonction de lien qui lie l'espérance de $\mathbf{Y}$ au prédicteur linéaire $\mathbf{X\beta}$.

Ainsi, après avoir spécifié la densité du modèle relativement à la nature des observations $\mathbf{Y}$, le choix simple pour une fonction de lien est le lien canonique.

Les caractéristiques de quelques distributions couramment utilisées pour les modèles GLM sont données au tableau 1.1. Les paramétrisations de ces distributions sont laissées en annexe A.1 au tableau A.1.

Pour être en mesure de poursuivre, la spécification du modèle s'impose.

1.2. SPÉCIFICATION DU MODÈLE

Lors d'études de cohortes cherchant à mesurer l'efficacité d'un incitatif sur la réponse d'individus, la variable observée est de nature binaire (un individu répond favorablement ou défavorablement). Nous supposons que chaque individu répond indépendamment les uns des autres.

Supposons que chaque observation de $\mathbf{Y}$ (les réponses de chaque individu) suit une loi de Bernoulli avec probabilité de succès p_i. Ainsi, $\mathbb{E}(Y_i) = p_i$. De cette manière, $\mathbb{E}(Y_i)$ prend seulement des valeurs sur l'intervalle $(0, 1)$ selon la loi de probabilité. Quatre principales fonctions de lien sont considérées dans la littérature pour tenir compte de ce système :

(1) *logit*

$$g(x) = \log\left[\frac{x}{1-x}\right],$$

(2) *probit*

$$g(x) = \Phi^{-1}(x),$$

(3) *complémentaire log-log*

$$g(x) = \log\{-\log[1-x]\},$$

(4) *log-log*

$$g(x) = -\log\{-\log[x]\}.$$

La figure 1.1 montre l'évolution des quatre fonctions de lien précédentes. Nous utilisons à partir de maintenant la fonction logit comme fonction de lien pour ces propriétés théoriques plus simples, mais surtout pour la simplicité de son interprétation. En effet, puisque la fonction de lien sert à unir le prédicteur linéaire au vecteur de réponse, nous avons

$$g(\mathbb{E}(Y_i|\mathbf{X}_i, \boldsymbol{\beta})) = \log\left[\frac{\mathbb{E}(Y_i|\mathbf{X}_i, \boldsymbol{\beta})}{1-\mathbb{E}(Y_i|\mathbf{X}_i, \boldsymbol{\beta})}\right] = \log\left[\frac{p_i}{1-p_i}\right] = \mathbf{X}_i^T\boldsymbol{\beta}.$$

Ainsi, le prédicteur linéaire correspond au logarithme du rapport de cotes. Puisque la fonction de lien est inversible par définition, nous déduisons que

$$\mathbb{E}(Y_i|\mathbf{X}_i, \boldsymbol{\beta}) = p_i = \frac{e^{\mathbf{X}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{X}_i^T\boldsymbol{\beta}}}. \tag{1.2.1}$$

Prenons, par exemple, n individus. Notons par $Y_i = 1$ une réponse favorable de l'individu i où $i \in \{1, 2, \ldots, n\}$. Chaque individu suivra une loi de

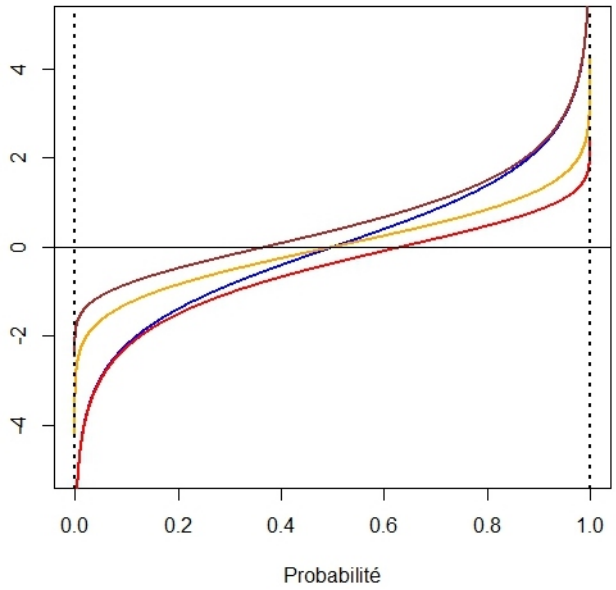


FIGURE 1.1. Fonctions de lien : logit (bleu), probit (jaune), complémentaire log-log (rouge) et log-log (brun).

Bernoulli de paramètre p_i. Avec l'équation (1.2.1), nous avons

$$Y_i \sim \mathcal{B}\left(1, \frac{e^{\mathbf{X}_i^\top \boldsymbol{\beta}}}{1 + e^{\mathbf{X}_i^\top \boldsymbol{\beta}}}\right), \tag{1.2.2}$$

où $\mathcal{B}(n, p)$ représente la loi binomiale pour n épreuves avec probabilité de succès p.

1.3. MÉTHODE D'ESTIMATION

Maintenant que le modèle est adéquatement spécifié, nous cherchons à estimer l'effet des coefficients $\boldsymbol{\beta}$. Pour ce faire, nous cherchons le maximum de vraisemblance.

Puisque maximiser la vraisemblance revient à maximiser la log-vraisemblance, nous avons

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \{L(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X})\} = \arg\max_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \log\left[f(Y_i = y_i|\mathbf{X}_i, \boldsymbol{\beta})\right] \right\},$$

où L représente la log-vraisemblance et $f(Y_i = y_i|\mathbf{X}_i, \boldsymbol{\beta})$ la fonction de densité associée à Y_i. Nous avons

$$\begin{aligned} f(Y_i = y_i|\mathbf{X}_i, \boldsymbol{\beta}) &= \left(\frac{e^{\mathbf{X}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{X}_i^T\boldsymbol{\beta}}}\right)^{y_i}\left(1-\frac{e^{\mathbf{X}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{X}_i^T\boldsymbol{\beta}}}\right)^{1-y_i} \\ &= \frac{e^{\mathbf{X}_i^T\boldsymbol{\beta}y_i}}{1+e^{\mathbf{X}_i^T\boldsymbol{\beta}}}, \end{aligned}$$

pour $i \in \{1, 2, \ldots, n\}$. Ainsi, la log-vraisemblance s'écrit comme

$$\begin{aligned} L(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}) &= \sum_{i=1}^{n} \log\left[\frac{e^{\mathbf{X}_i^T\boldsymbol{\beta}y_i}}{1+e^{\mathbf{X}_i^T\boldsymbol{\beta}}}\right] \\ &= \sum_{i=1}^{n}\left[\mathbf{X}_i^T\boldsymbol{\beta}y_i - \log\left(1+e^{\mathbf{X}_i^T\boldsymbol{\beta}}\right)\right]. \end{aligned}$$

Dérivons maintenant la log-vraisemblance par rapport aux coefficients $\boldsymbol{\beta}$. Nous avons

$$\frac{\partial L}{\partial \beta_r} = \sum_{i=1}^{n} x_{ir}\left[y_i - \frac{e^{\mathbf{X}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{X}_i^T\boldsymbol{\beta}}}\right].$$

Nous remarquons que les dérivées partielles de la log-vraisemblance dépendent de $\boldsymbol{\beta}$. Il est donc impossible de maximiser la log-vraisemblance en $\boldsymbol{\beta}$ de façon directe. Nous devons passer par des méthodes itératives pour arriver aux estimations des coefficients $\boldsymbol{\beta}$.

Nous cherchons à sélectionner itérativement des valeurs de $\boldsymbol{\beta}$ qui maximisent continuellement la log-vraisemblance jusqu'à convergence. Ainsi, nous avons

$$L(\boldsymbol{\beta}^{(k)}|\mathbf{Y}, \mathbf{X}) > L(\boldsymbol{\beta}^{(k-1)}|\mathbf{Y}, \mathbf{X}) \tag{1.3.1}$$

Écrivons dorénavant la log-vraisemblance $L(\boldsymbol{\beta}^{(k)}|\mathbf{Y}, \mathbf{X})$ plus simplement par L_k. Pour trouver un bon candidat à $\boldsymbol{\beta}$, nous devons choisir la direction adéquate. Nous prenons la direction de la plus forte variation définie par le vecteur gradient ∇L tel que $\nabla L = \frac{\partial L}{\partial \boldsymbol{\beta}}$. Pour choisir la longueur du pas, nous utilisons la méthode dite du *score de Fisher* (*Fisher scoring*) utilisée pour les modèles logistiques. Nous la définissons par $-\left(\nabla^2 L\right)^{-1}$, où $\nabla^2 L$ représente la matrice hessienne de la log-vraisemblance (matrice carrée des dérivées partielles secondes de la log-vraisemblance). Il est important de noter que la méthode du score de Fisher coïncide avec la méthode de *Newton-Raphson* lorsque la fonction de lien utilisée est le lien canonique.

À partir d'une valeur initiale $\boldsymbol{\beta}^{(0)}$, nous itérons de la façon suivante :

$$\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta}^{(k-1)} - \left(\nabla^2 L_{k-1}\right)^{-1} \nabla L_{k-1}, \tag{1.3.2}$$

où $\boldsymbol{\beta}^{(k)}$ représente la valeur de $\boldsymbol{\beta}$ à l'itération k et ∇L_{k-1} représente le vecteur gradient évalué en $\boldsymbol{\beta}^{(k-1)}$. Nous rappelons que

$$\frac{\partial L}{\partial \beta_r} = \sum_{i=1}^{n} x_{ir} \left[y_i - \frac{e^{x_i^T \boldsymbol{\beta}}}{1 + e^{x_i^T \boldsymbol{\beta}}} \right]. \tag{1.3.3}$$

Il en découle que

$$\nabla L = \mathbf{X}^T \left[\mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}) \right]. \tag{1.3.4}$$

En dérivant à nouveau l'équation (1.3.3), nous avons

$$\frac{\partial^2 L}{\partial \beta_r \partial \beta_s} = -\sum_{i=1}^{n} x_{ir} x_{is} \frac{e^{x_i^T \boldsymbol{\beta}}}{\left(1 + e^{x_i^T \boldsymbol{\beta}}\right)^2}.$$

Nous obtenons donc

$$\nabla^2 L = -\mathbf{X}^T \mathbf{W} \mathbf{X}, \tag{1.3.5}$$

où $\mathbf{W}$ est une matrice diagonale de poids telle que

$$\mathbf{W} = diag \left\{ \frac{e^{x_i^T \boldsymbol{\beta}}}{\left(1 + e^{x_i^T \boldsymbol{\beta}}\right)^2} \right\}.$$

Lorsque nous combinons les équations (1.3.4) et (1.3.5) avec le modèle itératif de l'équation (1.3.2), nous obtenons

$$\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta}^{(k-1)} + \left(\mathbf{X}^T \mathbf{W}^{(k-1)} \mathbf{X}\right)^{-1} \mathbf{X}^T \left[\mathbf{Y} - \mathbb{E}\left(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}^{(k-1)}\right) \right], \tag{1.3.6}$$

où $\mathbf{W}^{(k-1)}$ et $\mathbb{E}\left(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}^{(k-1)}\right)$ sont respectivement la matrice diagonale $\mathbf{W}$ et le vecteur des espérances $\mathbb{E}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta})$ évalué en $\boldsymbol{\beta}^{(k-1)}$. Multiplions à gauche par $\left(\mathbf{X}^T \mathbf{W}^{(k-1)} \mathbf{X}\right)$ à chaque terme. Nous avons

$$\begin{aligned}
\left(\mathbf{X}^T \mathbf{W}^{(k-1)} \mathbf{X}\right) \boldsymbol{\beta}^{(k)} &= \left(\mathbf{X}^T \mathbf{W}^{(k-1)} \mathbf{X}\right) \boldsymbol{\beta}^{(k-1)} + \mathbf{X}^T \left[\mathbf{Y} - \mathbb{E}\left(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}^{(k-1)}\right) \right] \\
&= \mathbf{X}^T \mathbf{W}^{(k-1)} \left(\mathbf{X} \boldsymbol{\beta}^{(k-1)} + \left(\mathbf{W}^{(k-1)}\right)^{-1} \left[\mathbf{Y} - \mathbb{E}\left(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}^{(k-1)}\right) \right] \right) \\
&= \mathbf{X}^T \mathbf{W}^{(k-1)} \mathbf{Z}^{(k-1)},
\end{aligned}$$

où $\mathbf{Z}^{(k-1)}$ est défini comme

$$\mathbf{Z}^{(k-1)} = \mathbf{X} \boldsymbol{\beta}^{(k-1)} + \left(\mathbf{W}^{(k-1)}\right)^{-1} \left[\mathbf{Y} - \mathbb{E}\left(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}^{(k-1)}\right) \right].$$

Nous obtenons donc

$$\boldsymbol{\beta}^{(k)} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{W}^{(k-1)}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{W}^{(k-1)}\mathbf{Z}^{(k-1)}. \tag{1.3.7}$$

Nous voyons avec l'équation (1.3.7) qu'à l'étape k, l'algorithme revient à faire une régression pondérée de matrice de poids $\mathbf{W}^{(k-1)}$ de $\mathbf{X}$ sur $\mathbf{Z}^{(k-1)}$. L'algorithme est souvent appelé IRLS pour «*iterative reweighted least squares*». La fonction *glm()* du logiciel *R* effectue l'algorithme IRLS sur un jeu de données fourni par l'utilisateur selon le modèle GLM spécifié. Cette méthode d'estimation est notre méthode de référence pour nos simulations expliquées au chapitre 4

À la section 1.1, nous avons tout d'abord introduit les composantes d'un modèle linéaire généralisé par l'extension d'une fonction de lien appliquée au modèle linéaire classique. Par la suite, notre cas d'étude de cohortes cherchant à mesurer l'efficacité d'un incitatif sur la réponse d'individus présenté à la section 1.2 a clos la spécification du modèle linéaire généralisé par un modèle logistique. Finalement, la section 1.3 a exposé la méthode d'estimation du maximum de vraisemblance des coefficients du modèle logistique.

Cherchant à déterminer l'estimation la plus précise pour notre modèle de régression logistique, nous discutons d'une toute autre approche au chapitre 2, soit l'approche bayésienne.

Chapitre 2

RÉGRESSION LOGISTIQUE BAYÉSIENNE

En laissant dorénavant les coefficients de la régression logistique provenir de lois de probabilité et non simplement être de nature déterministe, nous modifions complètement l'approche classique présentée au chapitre 1. Voyons en quoi la régression logistique dite *bayésienne* en diffère.

Nous développons tout d'abord cette pensée bayésienne à la section 2.1. Ensuite, la section 2.2 nous explique la nature et l'utilité des méthodes MCMC (*Markov chain Monte Carlo*) pour l'estimation de nos coefficients. Des diagnostics de convergence pour les méthodes MCMC sont discutés à la section 2.3. Deux autres méthodes sont discutées dans ce chapitre. À la section 2.4, une densité *a priori* faiblement informative introduite par Gelman et coll. (2008) est proposée comme densité conservatrice pour ajuster des modèles de régression logistique. Enfin, nous explorons à la section 2.5 l'approche d'estimation ponctuelle des coefficients de la régression logistique par le maximum *a posteriori*. Cette méthode, utilisée par Genkin et coll. (2007), est construite pour rendre plus rapide et plus efficace l'estimation et la prédiction pour des variables avec événements rares.

2.1. APERÇU DE L'APPROCHE BAYÉSIENNE

Reprenons les mêmes bases théoriques qu'au chapitre 1 à la différence près que le vecteur des coefficients $\boldsymbol{\beta}$ n'est plus supposé fixe, mais suivant une certaine loi de probabilité. Ainsi, nous appelons *densité a priori* la loi de probabilité de $\boldsymbol{\beta}$, notée $\pi(\boldsymbol{\beta})$, cette loi modélise toute l'information connue sur $\boldsymbol{\beta}$ n'étant pas en lien avec les n observations du vecteur $\mathbf{Y}$. Évidemment, l'idée est d'actualiser la densité *a priori* $\pi(\boldsymbol{\beta})$ en introduisant l'information additionnelle contenue dans les observations $\mathbf{Y}$. Nous nommons la loi de probabilité

actualisée, *densité a posteriori*, notée $\pi(\boldsymbol{\beta}|\mathbf{Y})$. Le théorème de Bayes nous aide à faire le lien entre la densité *a priori* et la densité *a posteriori*.

Théorème 2.1.1 (Théorème de Bayes)**.** *Soient* B *et* A *deux événements et* $\{A_i\}_{i\in I}$ *une partition de* A*. Pour tout événement* A_j *de la partition* $\{A_i\}_{i\in I}$*, nous avons*

$$P(A_j|B) = \frac{P(B|A_j)\,P(A_j)}{\sum_{i\in I} P(B|A_i)\,P(A_i)} = \frac{P(B|A_j)\,P(A_j)}{P(B)}.$$

Nous généralisons ce théorème pour des variables aléatoires continues. Soient $\mathbf{Y}$ et $\boldsymbol{\beta}$ deux variables aléatoires. Notons $f(\mathbf{Y}|\boldsymbol{\beta})$ la densité conditionnelle de $\mathbf{Y}|\boldsymbol{\beta}$, $f(\mathbf{Y})$ la densité marginale de $\mathbf{Y}$ et $\pi(\boldsymbol{\beta})$ la densité *a priori* de $\boldsymbol{\beta}$. La densité *a posteriori* $\pi(\boldsymbol{\beta}|\mathbf{Y})$ peut s'exprimer comme

$$\pi(\boldsymbol{\beta}|\mathbf{Y}) = \frac{f(\mathbf{Y}|\boldsymbol{\beta})\,\pi(\boldsymbol{\beta})}{\int_{\boldsymbol{\beta}} f(\mathbf{Y}|\boldsymbol{\beta})\,\pi(\boldsymbol{\beta})\,d\boldsymbol{\beta}} = \frac{f(\mathbf{Y}|\boldsymbol{\beta})\,\pi(\boldsymbol{\beta})}{f(\mathbf{Y})}.$$

L'actualisation de $\pi(\boldsymbol{\beta})$ passe par l'introduction de la vraisemblance des observations $\mathbf{Y}$. Le dénominateur, pour sa part, n'est qu'une constante de normalisation qui force la densité *a posteriori* à intégrer à 1. Rappelons la vraisemblance des observations $\mathbf{Y}$ définie par

$$f(\mathbf{Y}|\mathbf{X},\boldsymbol{\beta}) = \prod_{i=1}^{n} f(Y_i|\mathbf{X}_i,\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{e^{\mathbf{X}_i^T\boldsymbol{\beta}Y_i}}{1+e^{\mathbf{X}_i^T\boldsymbol{\beta}}}. \tag{2.1.1}$$

Puisque les variables explicatives sont connues et restes déterministes, nous avons implicitement

$$\pi(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X}) = \frac{f(\mathbf{Y}|\mathbf{X},\boldsymbol{\beta})\,\pi(\boldsymbol{\beta}|\mathbf{X})}{\int_{\boldsymbol{\beta}} f(\mathbf{Y}|\mathbf{X},\boldsymbol{\beta})\,\pi(\boldsymbol{\beta}|\mathbf{X})\,d\boldsymbol{\beta}} = \frac{f(\mathbf{Y}|\mathbf{X},\boldsymbol{\beta})\,\pi(\boldsymbol{\beta}|\mathbf{X})}{f(\mathbf{Y}|\mathbf{X})}. \tag{2.1.2}$$

Connaissant la vraisemblance des observations $\mathbf{Y}$ pour notre cas d'études de cohortes cherchant à mesurer l'efficacité d'un incitatif sur la réponse d'individus, il nous reste à spécifier la densité *a priori*. Nous parlerons de l'importance du choix de la densité *a priori* au chapitre 3. Pour l'instant, supposons que $\boldsymbol{\beta}$ suit une densité *a priori* de loi normale multivariée telle que

$$\pi(\boldsymbol{\beta}) = \frac{1}{2\pi\,|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}}, \tag{2.1.3}$$

où $\boldsymbol{\Sigma}$ est la matrice de covariance avec σ^2 sur sa diagonale. Nous supposons que les coefficients $\boldsymbol{\beta}$ sont indépendants. Ainsi, les covariances des coefficients $\boldsymbol{\beta}$ seront nulles et $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ avec $\mathbf{I}$ comme étant la matrice identité. L'équation

(2.1.3) s'écrira comme

$$\pi(\boldsymbol{\beta}) = \prod_{i=1}^{p} \left[\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{1}{2\sigma^2}\beta_i^2} \right].$$

Développons l'équation (2.1.2) en excluant tous les termes qui ne dépendent pas de $\boldsymbol{\beta}$. Nous avons

$$\pi(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{\mathbf{X}_i^T \boldsymbol{\beta} Y_i}}{1 + e^{\mathbf{X}_i^T \boldsymbol{\beta}}} \right) e^{-\frac{1}{2\sigma^2}\boldsymbol{\beta}^T \boldsymbol{\beta}}.$$

Il nous suffit de calculer la constante de normalisation pour obtenir la densité *a posteriori* exacte. Il est cependant impossible d'intégrer le dénominateur de l'équation (2.1.2) sur l'ensemble des $\boldsymbol{\beta}$ sans passer par des estimations numériques. Nous parlons donc des méthodes MCMC utilisées pour estimer la densité *a posteriori* à la section suivante.

2.2. MÉTHODE MCMC

Il est maintenant standard pour l'approche bayésienne d'utiliser les méthodes MCMC (*Markov chain Monte Carlo*) pour l'estimation des densités *a posteriori*. Définissons tout d'abord les bases.

Une chaîne de Markov est une suite $X_0, X_1, X_2, \ldots$ de variables aléatoires évoluant à travers le temps selon une probabilité de transition dépendante du passé de la chaîne en question. La récurrence, l'apériodicité et la réversibilité sont trois propriétés nécessaires à une chaîne de Markov pour les méthodes de simulation MCMC.

Une chaîne de Markov est proprement définie à partir d'une fonction de probabilité qui détermine ces transitions. Nous appelons cette fonction *noyau de transition*.

Définition 2.2.1 (Noyau de transition)**.** *Un* noyau de transition *est une fonction* K *définie sur l'espace mesurable* $(\Omega, \mathcal{A})$ *tel que*

(1) $\forall x \in \Omega$, $K(x, \cdot)$ *est une mesure de probabilité,*

(2) $\forall A \in \mathcal{A}$, $K(\cdot, A)$ *est une fonction mesurable.*

Le noyau de transition désigne la densité conditionnelle $K(x, y)$ de la transition $K(x, \cdot)$. Nous avons

$$P(X_{t+1} \in A | X_t = x) = \int_A K(x, y)\, dy.$$

Avec un noyau de transition bien définie, nous pouvons maintenant construire une chaîne de Markov.

Définition 2.2.2 (Chaîne de Markov)**.** *Étant donné un noyau de transition* K *définie sur l'espace mesurable* $(\Omega, \mathcal{A})$*, une suite* $X_0, X_1, X_2, \ldots$ *de variables aléatoires est appelée* chaîne de Markov *(notée* (X_t)*) si* $\forall t \in \mathbb{N}$*,* $\forall x_0, x_1, \ldots, x_t \in \Omega$ *et* $\forall A \in \mathcal{A}$ *nous avons*

$$P(X_{t+1} \in A | X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t) = P(X_{t+1} \in A | X_t = x_t) = \int_A K(x_t, y)\, dy.$$

Ainsi pour une chaîne de Markov, chaque état dépend uniquement de l'état précédent. La chaîne sera définie par

- le noyau de transition K,
- la distribution initiale $X_0 \sim \Pi_0$.

De plus, nous supposons que la chaîne est homogène. En d'autres mots, nous imposons à la chaîne de Markov un mécanisme de transition qui ne change pas à travers le temps. Plus concrètement, une chaîne de Markov homogène sera construite à partir d'un noyau de transition K de telle sorte que $X_{t+1} \sim K(X_t, X_{t+1})\ \forall t$. Prenons par exemple

$$X_t = \theta X_{t-1} + \epsilon_t,$$

où $\theta \in \mathbb{R}$ et $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. Ce système est appelé un modèle AR(1). La distribution normale $\mathcal{N}(\theta X_{t-1}, \sigma^2)$ correspond au noyau de transition de la chaîne (X_t). Si les ϵ_t sont indépendants, X_t est en fait indépendant de $X_{t-2}, X_{t-3}, \ldots, X_0$ conditionnellement à X_{t-1}. La chaîne (X_t) remplit donc la condition d'indépendance des chaînes de Markov et sera aussi homogène.

Pour les méthodes MCMC, notre chaîne de Markov doit atteindre un certain niveau de stabilité. Nous aimerions avoir une distribution de probabilité Π telle que pour $X_t \sim \Pi$, alors $X_{t+1} \sim \Pi$.

Définition 2.2.3 (Mesure invariante)**.** *Une mesure* Π *est dite* invariante *pour le noyau de transition* K *si* $\forall A \in \mathcal{A}$

$$\Pi(A) = \int_{\Omega} K(x, A)\, \Pi(x) dx.$$

Cette mesure de probabilité invariante (aussi appelée *distribution stationnaire*) définit le comportement sans mémoire de la chaîne. Ainsi, il est impossible de retrouver la distribution initiale Π_0 de la chaîne (X_t) et ce, pour toute distribution initiale Π_0. Ainsi, à partir d'un échantillon tiré sur la distribution stationnaire, nous pourrons inférer sur son espérance.

Pour y arriver, notre chaîne de Markov devra avoir les propriétés suivantes :
- récurrence de Harris,
- apériodicité,
- réversibilité.

En résumé, une chaîne de Markov Harris récurrente est suffisante pour obtenir une distribution stationnaire Π unique et indépendante de la distribution initiale $X_0 \sim \Pi_0$. Pour assurer l'existence d'une telle distribution limite Π il est nécessaire d'avoir en plus une chaîne qui est apériodique. En ajoutant la propriété de réversibilité, l'inférence sur la distribution stationnaire Π de notre chaîne de Markov pourra s'effectuer à l'aide du théorème limite centrale. Pour plus de détail sur ces propriétés, nous vous invitons à consulter l'annexe A.2.

Puisque la densité *a posteriori* est souvent impossible à déterminer analytiquement, son estimation passera par la construction d'une chaîne de Markov Harris récurrente, apériodique et réversible en forçant la densité *a posteriori* à coïncider avec la distribution stationnaire. Là est l'essence même des méthodes MCMC.

Définition 2.2.4 (Méthodes MCMC)**.** *Pour la simulation d'une distribution* f*, une* méthode MCMC *(*Markov chain Monte Carlo*) est une méthode qui produit une chaîne de Markov ergodique* (X_t) *avec* f *comme distribution stationnaire.*

Bien que plusieurs algorithmes peuvent être utilisés pour effectuer nos simulations MCMC, telles que l'*algorithme de Metropolis-Hasting* et l'*échantillonnage de Gibbs*, nous utiliserons la méthode d'*échantillonnage par tranche* (*slice sampling*) introduite par Neal (2003). Les détails de la méthode d'échantillonnage

par tranche sont laissés à l'annexe A.3. Nous vous invitons aussi à consulter l'article de Neal (2003).

Le logiciel *JAGS* (un logiciel libre) que nous avons choisi pour nos simulations utilise la méthode d'échantillonnage par tranche. C'est la raison pour laquelle nous avons choisi cette méthode. Pour en savoir davantage sur l'application des méthodes MCMC avec l'aide de *R* et *BUGS* (*JAGS* se veut ouvertement un clone de *BUGS*), nous vous suggérons le livre de Carlin et Louis (2011). Nous suggérons aussi au lecteur de consulter la documentation de la librairie *rjags* (Plummer et Stukalov (2013)) du logiciel *R* qui fournit une interface de *R* à *JAGS*.

Ce qui est important à retenir, ici, est l'utilité de la méthode d'échantillonnage par tranche. Nous avons désormais une méthode qui, à partir d'un certain moment, échantillonne des observations sur la densité *a posteriori*. Puisque la méthode vérifie la propriété de Markov, l'échantillon sera nécessairement corrélé. Tant que l'échantillonnage se fait par rapport à la densité *a posteriori*, la corrélation entre les observations aura un effet marginal. Mais à partir de quel moment l'échantillonnage se fait-il sur la densité *a posteriori* ?

2.3. Diagnostic de convergence

En général pour les méthodes MCMC, il existe deux formes de convergence qui nécessites l'évaluation :

(1) convergence vers la distribution stationnaire,

(2) convergence des moyennes.

La convergence en un échantillonnage i.i.d. est une autre forme de convergence. L'idée générale est de produire un échantillon (quasi-)indépendant par sous-échantillonnage pour réduire la corrélation entre les itérations de la chaîne de Markov. Cependant, même si un échantillon i.i.d. est habituellement recherché, MacEachern et Berliner (1994) montre que nous perdons de l'efficacité par rapport à la convergence des moyennes lors de sous-échantillonnage. C'est pour cette raison que les diagnostics de convergence en un échantillonnage i.i.d. sont laissés tombés. Nous énumérons uniquement dans les sous-sections subséquentes les méthodes de diagnostic de convergence utilisées lors de nos simulations. Pour une revue plus détaillée des diagnostics de convergence,

vous pouvez consulter le chapitre 12 de Robert et Casella (2004).

2.3.1. Convergence en la distribution stationnaire

La convergence en la distribution stationnaire est nécessaire, car nous cherchons à approximer la densité *a posteriori*. Il faut donc s'assurer que notre chaîne atteigne sa distribution stationnaire. Même si la théorie nous indique que la chaîne convergera (voir le théorème A.2.2), il lui faudra un nombre infini d'itérations pour nous en assurer. Dans un environnement de simulation, le facteur de réduction d'échelle proposé par Gelman et Rubin (1992) nous aidera à évaluer l'atteinte de la stationnarité de notre chaîne.

Un petit nombre ($m > 1$) de chaînes sont exécutées en parallèle de façon indépendante. Les m chaînes sont ensuite lancées pendant $2K$ itérations chacune. Afin de diminuer l'influence de la distribution des valeurs initiales, Gelman et Rubin (1992) propose de jeter les K premières itérations de chaque chaîne pour concentrer notre attention sur les K dernières itérations. Pour chaque coefficient β_i à estimer, nous calculons la mesure de convergence appelée *facteur de réduction d'échelle* (*scale reduction factor*), noté $\sqrt{\widehat{R}_i}$.

Le facteur de réduction d'échelle est basé sur la comparaison entre la variance intra chaîne et la variance inter chaîne (similaire à l'analyse de variance classique). Le diagnostic de convergence tente de vérifier si la variation intra chaîne se rapproche de la variation totale, auquel cas les tirages de toutes les chaînes auraient convergé en distribution vers la distribution stationnaire (étant ici la distribution *a posteriori*).

Le facteur de réduction d'échelle associé à β_i est défini par

$$\sqrt{\widehat{R}_i} = \sqrt{\left[\frac{K-1}{K} + \frac{B_i}{K}\left(\frac{m+1}{m}\right)\frac{1}{W_i}\right]\left(\frac{\widehat{df}_i}{\widehat{df}_i - 2}\right)}, \tag{2.3.1}$$

où B_i/K est la variance inter chaîne associée à β_i, W_i est la variance intra chaîne associée à β_i et $\widehat{df}_i$ est le nombre de degrés de liberté d'une loi de Student qui approxime la densité *a posteriori*. Les détails sont donnés à l'annexe A.4.

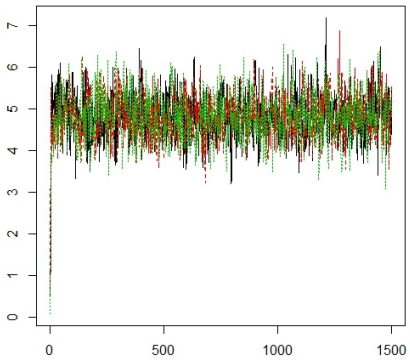
Le facteur de réduction d'échelle représente le facteur pour lequel l'échelle diminue si la chaîne poursuit son échantillonnage indéfiniment. Gelman et Rubin (1992) montre que $\sqrt{\widehat{R}_i} \searrow 1$ lorsque $K \to \infty$. Ainsi, une valeur très proche de 1 indique que les chaînes sont très susceptibles d'avoir convergées en leur distribution stationnaire et que le tirage s'effectue sur la densité *a posteriori*. Les $m \cdot K$ valeurs seront combinées pour inférer sur la distribution *a posteriori*. Si toutefois la valeur est plus grande que 1, nous avons des raisons de croire que davantage d'itérations seraient nécessaires pour approcher la densité *a posteriori*.

L'approche de Gelman et Rubin (1992) est facilement applicable et interprétable pour notre cas d'étude de cohortes cherchant à mesurer l'efficacité d'un incitatif sur la réponse d'individus. Il y aura p tests de convergence (un pour chaque coefficient) à exécuter. Cette méthode gaspille cependant beaucoup d'itérations et peut coûter beaucoup de temps de simulation (nous rejetons la première moitié K de l'échantillon $2K$).

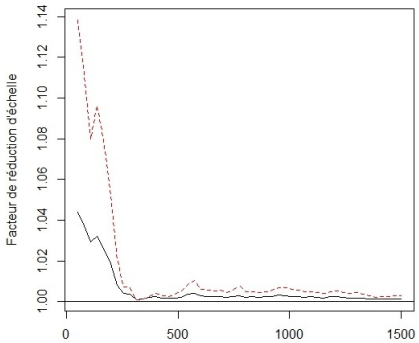
En réduisant le nombre d'itérations rejetées, il faudrait toutefois s'assurer de choisir un échantillon provenant entièrement de la densité *a posteriori*. Supposons que nous effectuons un total de κ itérations par la méthode d'échantillonnage par tranche. Nous rappelons que la méthode est construite telle que la densité *a posteriori* $\pi(\boldsymbol{\beta}|\mathbf{y})$ est en fait la distribution stationnaire. De ce fait, nous obtenons à l'itération k une valeur $\boldsymbol{\beta}^{(k)}$ qui converge en distribution en un tirage provenant de la densité *a posteriori* pour toutes valeurs de k suffisamment grande (disons $k > k_0$). Ainsi, l'ensemble $\left\{ \boldsymbol{\beta}^{(k)} | k = k_0 + 1, \ldots, \kappa \right\}$ est un échantillon provenant de la densité *a posteriori* recherchée. Nous appelons les tirages $k = 0$ à $k = k_0$ la période de chauffe (nous constatons que Gelman et Rubin (1992) utilise la valeur $k_0 = K$ et $\kappa = 2K$ pour leur méthode).

Il est souvent commun d'utiliser en plus un diagnostic graphique pour vérifier la convergence. En superposant chacune des m chaînes exécutées en parallèle sur un graphique, nous examinons visuellement si les m chaînes semblent échantillonnées selon une distribution commune (la distribution stationnaire) ou non.

L'algorithme MCMC d'échantillonnage par tranche est appliqué à des données simulées suivant notre domaine d'application et la figure 2.1(a) suit l'évolution d'un des coefficients. Nous distinguons à peine l'influence des valeurs



(a) Suivi des itérations de l'algorithme MCMC d'échantillonnage par tranche



(b) Facteur de réduction d'échelle (ligne pleine) et son quantile 0.975 (ligne pointillée)

FIGURE 2.1. Diagnostic de convergence graphique d'un coefficient avec $m = 3$ chaînes.

initiales des $m = 3$ chaînes exécutées en parallèle après quelques itérations seulement. La figure 2.1(b) suit pour sa part l'évolution du facteur de réduction d'échelle $\sqrt{\hat{R}_i}$ sur les mêmes données simulées avec une approximation d'une borne supérieure. Nous observons, ici, une courte période de chauffe ($k_0 = 500$) nécessaire pour ce coefficient. L'idée est de tester pour chaque coefficient et de prendre une période de chauffe commune, soit le maximum de celles observées.

2.3.2. Convergence des moyennes

La convergence des moyennes est toute aussi importante que la convergence en la distribution stationnaire. En quoi sert une chaîne de Markov qui a convergé si nous sommes incapables d'obtenir une bonne inférence à partir de celle-ci ? La moyenne empirique

$$\frac{1}{K}\sum_{k=1}^{K} h\left(\beta^{(k)}\right)$$

doit converger vers $\mathbb{E}[h(\beta)]$ pour toute fonction h telle que $\mathbb{E}|h(\beta)| < \infty$. Nous devons nous assurer que la chaîne ait exploré l'ensemble du support de la densité *a posteriori* afin d'inférer adéquatement sur celle-ci. Même si d'un point de vue théorique le théorème ergodique A.2.3 nous assure la convergence en moyenne de la chaîne de Markov, un nombre minimal d'itérations doit être respecté pour atteindre une approximation adéquate de $\mathbb{E}[h(\beta)]$ lors

de nos simulations.

Puisque la méthode de Gelman et Rubin (1992) présentée à la section 2.3.1 n'infère pas sur la convergence des moyennes, nous nous tournons vers une méthode complémentaire introduite par Raftery et Lewis (1992).

Raftery et Lewis (1992) cherche à savoir le temps nécessaire pour qu'un algorithme MCMC obtienne une estimation précise d'un quantile extrême de la densité *a posteriori*. En d'autres mots, pour des valeurs de q, r et s fixés à l'avance, la procédure cherche le nombre minimal K d'observations pour que

$$P\left(|F\left(\beta_q\right) - q| < r\right) = s,$$

où $F\left(\cdot\right)$ est la fonction de répartition de β et β_q est son q^e quantile. Par défaut, la procédure teste pour des valeurs $q = 0{,}025$, $r = 0{,}005$ et $s = 0{,}95$. Raftery et Lewis (1992) construit le processus binaire

$$Z^{(k)} = \mathbb{I}_{\left\{\beta^{(k)} \leq \beta_q\right\}}.$$

Toutefois, le processus $Z^{(k)}$ n'est pas nécessairement une chaîne de Markov. D'un autre côté, nous pouvons définir un nouveau processus avec, cette fois-ci, un délai t tel que

$$Z_t^{(k)} = Z^{(1+(k-1)t)}$$

est approximativement une chaîne de Markov à deux états. Nous cherchons en fait le plus petit t pour lequel la chaîne $Z_t^{(k)}$ est plus près d'une chaîne de Markov de premier ordre (la transition vers $Z_t^{(k)}$ est dépendante de $Z_t^{(k-1)}$) que de deuxième ordre (la transition vers $Z_t^{(k)}$ est dépendante de $Z_t^{(k-1)}$ et $Z_t^{(k-2)}$). En d'autres mots, les observations étant potentiellement fortement dépendantes, l'ajout d'un délai au processus tente de réduire celle-ci en enlevant toute dépendance au-delà de l'ordre un. Vous pouvez trouver les détails de la méthode de comparaison utilisée pour déterminer la valeur de t dans l'article original de Raftery et Lewis (1992).

Après avoir déterminé la valeur de t, nous disposons d'une matrice de transition, notée $\mathbb{P}$, associée à $Z_t^{(k)}$. Nous avons

$$\mathbb{P} = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}.$$

Ainsi, le comportement limite de $Z_t^{(k)}$ sera tel que

$$\lim_{k\to\infty} P\left(Z_t^{(k)} = 1\right) = \frac{a}{a+b}.$$

Or, par construction nous savons que

$$\lim_{k\to\infty} P\left(Z_t^{(k)} = 1\right) = \lim_{k\to\infty} P\left(\mathbb{I}_{\left\{\beta^{(1+(k-1)t)} \leq \beta_q\right\}} = 1\right) = q,$$

et nous en déduisons que

$$\frac{a}{a+b} = q.$$

Sachant q, il est donc facile de déterminer la valeur de a et b à partir de l'échantillon $\left\{Z_t^{(k)}\right\}$.

La moyenne empirique $\overline{Z}_t^{(K)}$ de $Z_t^{(k)}$ est un estimateur sans biais de $P(\beta \leq \beta_q)$ et suit approximativement une loi normale. Nous avons,

$$\overline{Z}_t^{(K)} = \frac{1}{K}\sum_{k=1}^{K} Z_t^{(k)} \sim \mathcal{N}\left(\frac{a}{a+b}; \frac{1}{K}\frac{2ab\,(2-a-b)}{(a+b)^3}\right).$$

Ainsi, pour obtenir $P\left(\left|\overline{Z}_t^{(K)} - q\right| < r\right) = s$, nous devons prendre un minimum de $K = \frac{2ab(2-a-b)}{(a+b)^3}\left(\frac{\Phi\left(\frac{1+s}{2}\right)}{r}\right)^2$ valeurs de $Z_t^{(k)}$, c'est-à-dire tirer un nombre total de tK valeurs de $\beta^{(k)}$.

Pour utiliser cette méthode, il est nécessaire d'exécuter notre algorithme MCMC pour un certain nombre d'itérations puis estimer le nombre d'observations tK nécessaire pour chaque coefficient β_i. De la même façon que pour choisir une période de chauffe à la sous-section 2.3.1, nous utiliserons comme paramètre de simulation le maximum de t et K parmi les p estimations obtenues pour chaque coefficient β_j.

Afin d'établir notre schéma de simulation, les résultats des tests de convergences sont présentés à la section 2.3.3. Nous invitons aussi le lecteur à consulter la documentation de la librairie *CODA* (Plummer et coll. (2006)) du logiciel *R* auquel les diagnostics de convergence présentés dans la présente section ont été implantés.

2.3.3. Application des diagnostics de convergence

Préalablement aux simulations, nous avons testé les diagnostics de convergence présentés à la section 2.3 afin d'établir le nombre d'itérations durant la période de chauffe (k_0) et le nombre d'itérations pour atteindre un bon niveau de précision sur les estimations des coefficients (K).

La convergence en la distribution stationnaire se fait rapidement comme le montre la figure 2.1. Même si la figure 2.1 montre un seul coefficient, nous observons la même tendance chez tous nos coefficients. Une période de chauffe initiale d'au moins $k_0 = 500$ itérations est donc suffisante pour notre modélisation.

Par la suite, la procédure de Raftery et Lewis (1992) indique qu'aucun délai n'est nécessaire afin d'atténuer la dépendance d'ordre deux des observations. De plus, le même diagnostic montre qu'il serait nécessaire d'exécuter la chaîne pendant au moins $K = 11000$ itérations afin d'obtenir une bonne précision sur une estimation d'un quantile éloigné de la distribution *a posteriori*.

En restant prudent, exécuter l'algorithme MCMC d'échantillonnage par tranche avec une seule chaîne pendant une période de chauffe de $k_0 = 1000$ itérations suivie d'un échantillonnage de $K = 12000$ observations est assez pour obtenir la convergence en la distribution stationnaire et la convergence en les moyennes selon le diagnostic de Gelman et Rubin (1992) additionné à celui de Raftery et Lewis (1992). Ce seront les valeurs utilisées pour nos simulations MCMC.

Maintenant que nous avons introduit et choisi la méthode MCMC d'échantillonnage par tranche à la section 2.2, nous aimerions la comparer avec d'autres méthodes bayésiennes d'estimation numérique vue dans la littérature. Nous présentons, ici, deux autres méthodes d'estimation numérique que nous appliquons en comparatif à notre méthode d'estimation MCMC. Évidemment, comme présenté à la section 1.3, l'estimateur du maximum de vraisemblance sera aussi une de nos méthodes de référence.

2.4. DENSITÉ *a priori* FAIBLEMENT INFORMATIVE

En résumé, Gelman et coll. (2008) propose d'ajuster le modèle logistique en utilisant une modification de l'algorithme IRLS (présenté à la section 1.3) afin d'inclure de l'information *a priori*. En utilisant des distributions de Student (centrées en 0 avec ν degrés de liberté et de paramètre d'échelle s) indépendantes comme densités *a priori* pour chaque coefficient, des estimations plus précises que la méthode classique du maximum de vraisemblance sont présentées dans l'article.

Différentes densités *a priori* de Student sont comparées et Gelman et coll. (2008) suggère l'utilisation de leur densité faiblement informative pour l'ajustement de modèles de régression logistique. Cette densité *a priori* est en fait une distribution de Cauchy (Student à 1 degré de liberté) centrée en 0 avec comme paramètre d'échelle 2,5 pour les coefficients β_j (pour $j \in \{2, \ldots, p\}$), alors qu'une Cauchy centrée en 0 avec 10 comme paramètre d'échelle est proposée pour β_1. Cette densité faiblement informative est proposée comme étant un choix conservateur qui la rend applicable comme densité *a priori* par défaut.

Cette méthode d'estimation numérique a l'avantage de toujours donner une réponse même en présence du phénomène appelé *séparation complète* ou *séparation quasi-complète*. Ce phénomène est présent lorsqu'une combinaison linéaire des variables explicatives prédit parfaitement la variable réponse. Ce phénomène rend impossible le calcul de l'estimateur du maximum de vraisemblance.

2.4.1. Séparation complète et quasi-complète

Prenons par exemple le modèle logistique avec comme unique régresseur x. Nous avons

$$\mathbb{E}(Y_i|x_i, \beta_1, \beta_2) = \frac{e^{\beta_1+\beta_2 x_i}}{1 + e^{\beta_1+\beta_2 x_i}},$$

pour $i \in \{1, 2, \ldots, n\}$. Supposons maintenant qu'il existe une valeur x_0 telle que pour chaque observation $(x_i; y_i)$

$$y_i = \begin{cases} 0 & \text{si } x_i < x_0\,; \\ 1 & \text{si } x_i \geq x_0. \end{cases} \tag{2.4.1}$$

Un échantillon respectant la condition de l'équation (2.4.1) présente une séparation complète et l'estimation du maximum de vraisemblance de β_1 et β_2

TABLEAU 2.1. Échantillon présentant une séparation complète.

Obs	X	Y
1	1	0
2	1	0
3	3	0
4	4	0
5	6	1
6	6	1
7	8	1
8	9	1
9	10	1
10	11	1

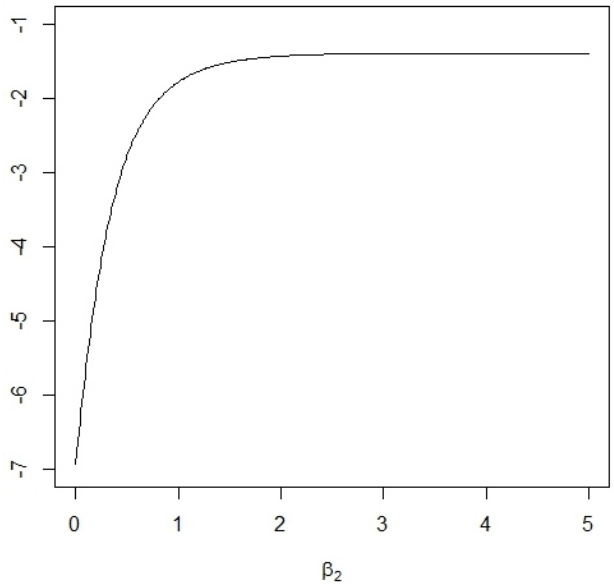


FIGURE 2.2. Log-vraisemblance d'un modèle logistique en présence de séparation complète.

n'existent pas.

Prenons l'échantillon du tableau 2.1, où l'unique façon d'observer une réponse favorable ($y_i = 1$) est d'avoir une valeur $x_i \geq x_0 = 6$. Cet échantillon présente donc une séparation complète. Le graphique 2.2 montre que la log-vraisemblance par rapport à β_2 n'atteindra jamais un maximum. L'estimation sera infinie.

Supposons maintenant que nous modifions la condition de l'équation (2.4.1) par

$$y_i = \begin{cases} 0 & \text{si } x_i < x_0\,; \\ 1 & \text{si } x_i > x_0, \end{cases} \tag{2.4.2}$$

avec $y_i \in \{0, 1\}$ lorsque $x_i = x_0$. Un échantillon respectant la condition de l'équation (2.4.2) présente une séparation quasi-complète. Tout comme la séparation complète, la séparation quasi-complète entraîne l'inexistence du maximum de vraisemblance pour les coefficients de la régression logistique.

Nous pouvons voir le détail de l'inexistence de l'estimateur de maximum de vraisemblance pour la séparation complète et quasi-complète dans l'article de Albert et Anderson (1984). Un résumé de l'idée derrière ce phénomène est laissé à l'annexe A.5.

Nous savons aussi que la méthode MCMC produira une estimation en présence de séparation complète (ou quasi-complète). Notre méthode d'estimation d'échantillonnage par tranche pourra donc être comparée avec la densité faiblement informative de Gelman et coll. (2008), même en présence de séparation complète.

2.5. MAXIMUM *a posteriori* EN PRÉSENCE D'ÉVÉNEMENTS RARES

Pour leur part, Genkin et coll. (2007) propose un algorithme d'estimation ponctuelle des coefficients $\boldsymbol{\beta}$. L'idée est d'introduire une méthode d'estimation rapide et efficace. Il propose de passer par l'estimateur du maximum *a posteriori* (MAP). L'estimateur MAP est donc la valeur $\hat{\boldsymbol{\beta}}$ qui maximise la densité *a posteriori*. Ce qui revient à maximiser

$$r(\boldsymbol{\beta}) = f(\mathbf{Y}|\boldsymbol{\beta})\,\pi(\boldsymbol{\beta})$$

par rapport à $\boldsymbol{\beta}$.

L'approche bayésienne proposée évite le surajustement des données et est efficace pour la prédiction. En utilisant une densité *a priori* Laplace, l'algorithme favorise l'estimation en présence d'événements rares (de certaines variables explicatives) en produisant des modèles de prédictions compactes (c'est-à-dire favoriser l'estimation des coefficients à 0). Cette méthode est appropriée pour des projets de catégorisation de texte (l'accent de l'article) et Genkin

et coll. (2007) arrivent à la conclusion que leur algorithme est au moins aussi efficace que les méthodes standards utilisées dans ce domaine.

De notre côté, nous utiliserons l'algorithme proposé par Genkin et coll. (2007) en utilisant deux densités *a priori* différentes : normale et Laplace. Ces densités seront centrées en 0 et nous fixons les variances à partir des valeurs σ_j^2 calculés selon

$$\sigma_j^2 = \frac{d}{u} = \frac{d}{\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i\|_2^2},$$

où d est le nombre de coefficients à estimer et u est la moyenne de la norme Euclédienne au carré. Nous suivons les recommandations de Genkin et coll. (2007) et utiliserons σ_j^2 comme variance des densités *a priori* normale et $\sqrt{2}/\sigma_j$ comme variance des densités *a priori* Laplace.

Les deux méthodes d'estimation numérique présentées aux sections 2.4 et 2.5 diffèrent de notre approche MCMC. Les méthodes MCMC produisent une estimation de la densité *a posteriori* tandis que ces méthodes proposent des estimations ponctuelles des coefficients. Il est vrai que notre but principal est d'obtenir des estimations ponctuelles et que l'obtention de la densité *a posteriori* n'est donc pas nécessaire. Cependant, nous cherchons à voir, par cette comparaison de méthodes, si l'estimation de la densité *a posteriori* ajoute de la précision aux estimations des coefficients de la régression logistique.

Nous avons commencé ce chapitre en présentant à la section 2.1 la statistique bayésienne comme alternative à la statistique classique. Les bases théoriques d'une chaîne de Markov et ses propriétés nécessaires afin d'utiliser les méthodes MCMC ont été présentées à la section 2.2. Puis, les diagnostics de convergence discutés à la section 2.3 nous permettent de construire un schéma d'échantillonnage afin d'inférer adéquatement sur les coefficients $\boldsymbol{\beta}$ lors de nos simulations. Enfin, deux autres méthodes d'estimations ont été introduites : la méthode de Gelman à la section 2.4 et la méthode de Genkin à la section 2.5.

Afin d'élargir l'horizon de comparaison de la méthode MCMC, une analyse de sensibilité sur la densité *a priori* est discutée au chapitre 3.

Chapitre 3

ANALYSE DE SENSIBILITÉ

Nous avons vu au chapitre 2 qu'il est possible d'estimer les coefficients de la régression logistique sous l'approche bayésienne et que nous avons opté pour la méthode d'échantillonnage par tranche pour le faire. Cependant, un point important reste à déterminer le choix de la densité *a priori*. C'est, ici, le point de ce mémoire. Étudier le comportement de différentes densités *a priori* sur l'estimation des coefficients de la régression, et ce sous différents types d'échantillons.

Rappelons tout d'abord que la densité *a posteriori* est déterminée par un mélange entre la densité *a priori* et la fonction de vraisemblance comme le montre l'équation (2.1.2). Ainsi, ces deux composantes auront une influence sur la densité *a posteriori*. Que ce soit par rapport à la densité *a priori* ou par rapport à l'échantillon utilisé, connaître les facteurs influents nous amènera à déterminer la meilleure densité en terme de précision d'estimation.

Nous cherchons des facteurs qui agissent sur l'estimation des coefficients dans le but d'étudier leur influence. Tout d'abord, les facteurs influents par rapport à la nature de la densité *a priori* sont discutés à la section 3.1, puis les facteurs influents par rapport à l'échantillon sont discutés à la section 3.2. Enfin, nous nous penchons sur l'influence de la fonction de vraisemblance à la section 3.3.

3.1. FACTEURS INFLUENTS PAR RAPPORT À LA DENSITÉ *a priori*

Le choix de la densité *a priori* influencera potentiellement la densité *a posteriori*. Nous pouvons donc opter pour une densité informative qui utilise de l'information *a priori* sur les coefficients à estimer ou plutôt opter pour une

TABLEAU 3.1. Densités *a priori* utilisées

Nom	Paramètres	Espérance	Variance	Densité
Normale	$\mu \in \mathbb{R}$ et $\sigma^2 \in \mathbb{R}^+$	μ	σ^2	$f\left(x\middle\vert \mu;\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$
Student	$\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+$ et $\nu > 2 \in \mathbb{N}$	μ	$\left(\frac{\nu}{\nu-2}\right)\sigma^2$	$f\left(x\middle\vert \mu;\sigma^2;\nu\right) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma^2}}\left(1+\frac{1}{\nu\sigma^2}(x-\mu)^2\right)^{-\frac{\nu+1}{2}}$
Laplace	$\mu \in \mathbb{R}$ et $\sigma^2 \in \mathbb{R}^+$	μ	$2\sigma^2$	$f\left(x\middle\vert \mu;\sigma^2\right) = \frac{1}{2\sqrt{\sigma^2}} e^{-\frac{\lvert x-\mu\rvert}{\sqrt{\sigma^2}}}$
Gumbel	$\mu \in \mathbb{R}$ et $\sigma^2 \in \mathbb{R}^+$	$\mu + \sqrt{\sigma^2}\gamma$	$\frac{\pi^2\sigma^2}{6}$	$f\left(x\middle\vert \mu;\sigma^2\right) = \frac{1}{\sqrt{\sigma^2}} e^{-\left(\frac{x-\mu}{\sqrt{\sigma^2}} + e^{-\frac{x-\mu}{\sqrt{\sigma^2}}}\right)}$

densité non informative qui omet l'information *a priori* pour laisser place à l'objectivité de l'information sur les coefficients.

Nous cherchons des facteurs qui agissent potentiellement sur l'estimation des coefficients. Puisque ces coefficients à estimer sont des éléments de la droite réelle, une densité *a priori* parcourant l'ensemble des réels est nécessaire. Quatre densités continues ont été comparées : normale, Student, Laplace et Gumbel. Le tableau 3.1 indique la paramétrisation utilisée pour chacune des densités, leur espérance ainsi que leur variance. Noter que nous utilisons $\nu = 3$ degrés de liberté pour notre densité Student.

L'épaisseur des queues, le positionnement et l'échelle sont les trois facteurs étudiés pour nos densités *a priori*.

Nous désirons mesurer l'influence des queues de chaque densité. L'indice de queue $\tau(F)$ (*tailed index*) défini au chapitre 10 du livre de Hoaglin et coll. (1983) est tel que

$$\tau(F) = \left[\frac{F^{-1}(0{,}99) - F^{-1}(0{,}50)}{F^{-1}(0{,}75) - F^{-1}(0{,}50)}\right] \Big/ \left[\frac{\Phi^{-1}(0{,}99) - \Phi^{-1}(0{,}50)}{\Phi^{-1}(0{,}75) - \Phi^{-1}(0{,}50)}\right], \tag{3.1.1}$$

où Φ est la fonction de répartition d'une loi normale standard et F est la fonction de répartition de la densité en question. Puisque nous soustrayons la valeur médiane aux 99^e et 75^e percentiles, la formule est invariante par rapport à la position et l'échelle des densités. Ainsi, pour toute densité d'une même famille de position-échelle, l'indice de queue sera le même.

TABLEAU 3.2. Indice de queue

Densité	$\tau_1(F)$	$\tau_2(F)$
Normale	1,000	1,000
Gumbel	1,396	0,792
Laplace	1,636	1,636
Student	1,721	1,721

La densité normale étant la valeur de référence, nous constatons qu'elle aura nécessairement un indice de queue de 1. De plus, une densité avec un indice inférieur à 1 aura des queues moins relevées que la loi normale. Inversement, une densité avec un indice supérieur à 1 aura des queues plus relevées que la loi normale.

Nous ajoutons cependant une variante à l'équation (3.1.1). Cet indice mesure l'épaisseur de la queue à droite seulement. Pour des densités symétriques, la queue à gauche est identique à celle de droite et l'équation (3.1.1) est valide pour ces densités en résumant et classifiant l'épaisseur de leur queue. Cependant, la densité Gumbel que nous utilisons n'est pas symétrique. C'est pourquoi nous calculons l'indice de queue à droite (noté $\tau_1(F) = \tau(F)$) ainsi que l'indice de queue à gauche (noté $\tau_2(F)$) définie réciproquement par

$$\tau_2(F) = \left[\frac{F^{-1}(0{,}50) - F^{-1}(0{,}01)}{F^{-1}(0{,}50) - F^{-1}(0{,}25)}\right] / \left[\frac{\Phi^{-1}(0{,}50) - \Phi^{-1}(0{,}01)}{\Phi^{-1}(0{,}50) - \Phi^{-1}(0{,}25)}\right].$$

Les indices de queues pour les quatre densités utilisées sont indiqués au tableau 3.2. De plus, la figure 3.1 montre la différence entre les queues des différentes densités selon la même position et échelle.

Évidemment, la queue des densités *a priori* n'est pas le seul facteur qui influence la densité *a posteriori*. La position et l'échelle sont deux autres facteurs à considérer. La figure 3.2(a) montre l'évolution de la probabilité de succès d'un modèle logistique simple (sans variable explicative),

$$p = \frac{e^{\beta_1}}{1 + e^{\beta_1}},$$

en fonction de β_1. Nous constatons qu'une valeur au-delà de 6 mène à une probabilité de succès plus grande que 0,9975. Réciproquement, une valeur en deçà de -6 mène à une probabilité de succès plus petite que 0,0025. Ainsi, nous constatons que la grande majeure partie de la variation de β_1 se situe sur l'intervalle $(-6, 6)$. De façon similaire, nous aurons la même variation pour $x_{ij}\beta_j$

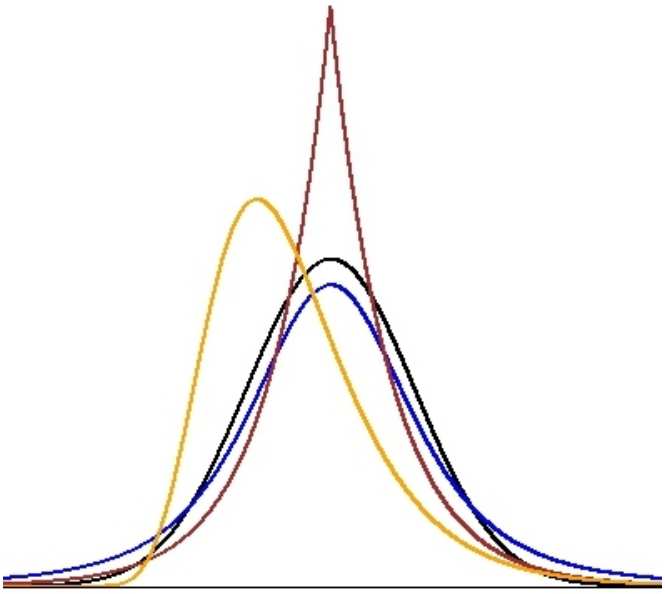


FIGURE 3.1. Comparatif des queues de quatre densités avec la même position et échelle : normale (noir), Student à 3 degrés de liberté (bleu), Laplace (brun) et Gumbel (jaune).

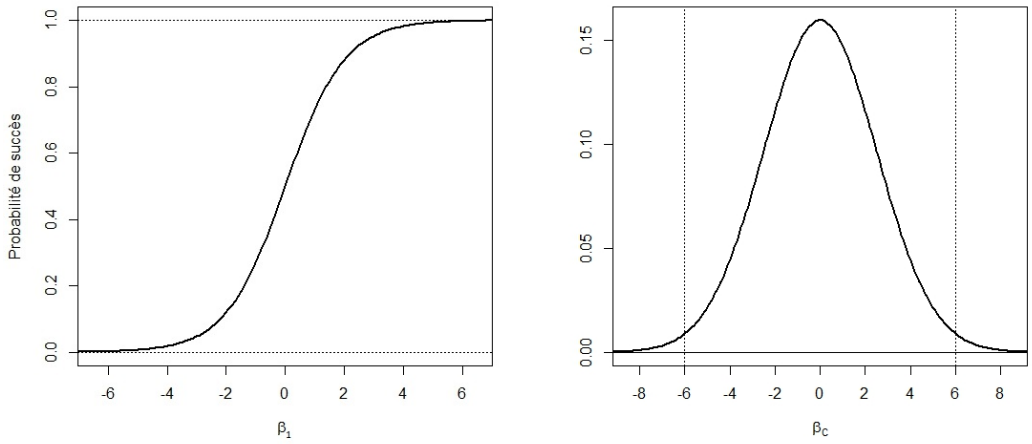
pour tout $j \in \{2, 3, \ldots, p\}$.

Nous voyons dans Gelman (2008) qu'il est important de standardiser les variables explicatives. L'idée derrière cette pensée est la même que l'observation que nous venons de faire à la figure 3.2(a). En standardisant toutes nos variables explicatives, nous nous assurons que β_j (et non pas seulement $x_{ij}\beta_j$) varie en majeure partie sur l'intervalle $(-6, 6)$.

Il est donc préférable de restreindre la majeure partie de notre densité *a priori* sur cet intervalle en s'assurant préalablement que chaque variable explicative le permette. La figure 3.2(b) montre une densité normale centrée en zéro avec une variance de $2{,}5^2$.

Pour généraliser ce cas à tous les autres β_j ($j \in \{2, 3, \ldots, p\}$), nous devrons restreindre les densités *a priori* de telle manière que $\mathbf{X}_i^T\boldsymbol{\beta}$ se retrouve la majeure partie du temps sur l'intervalle $(-6, 6)$.

Est-il plus raisonnable de centrer notre densité *a priori* en zéro avec une variance de $2{,}5^2$ lorsque nous n'avons aucune idée *a priori* de la probabilité de succès ? Inversement, centrer notre densité *a priori* selon notre connaissance antérieure de l'étude améliore-t-il l'estimation de la densité *a posteriori* ? Des



(a) Probabilité de succès en fonction d'un seul coefficient.

(b) Densité normale centrée en zéro avec une variance de $2{,}5^2$.

FIGURE 3.2. Majeure partie de l'étendue du coefficient β_1.

densités centrées en $-2{,}94$, 0 et 2,94 (probabilité de succès respective de 5%, 50% et 95%) avec une variance de $2{,}5^2$ multipliée par des facteurs $\frac{1}{2}$, 1 et 2 ont été étudiées lors de nos simulations.

3.2. FACTEURS INFLUENTS PAR RAPPORT À L'ÉCHANTILLON

En plus de la densité *a priori*, rappelons que la densité *a posteriori* est aussi influencée par la fonction de vraisemblance. Ainsi, les deux derniers facteurs influents que nous étudions sont la taille échantillonnale et la valeur des vrais coefficients.

Afin de tester nos densités *a priori*, nous générons des échantillons basés sur un vrai jeu de données que nous présenterons au chapitre 5. La première variable (notée x_{iC}) indique l'appartenance au groupe contrôle, la seconde (notée x_{iT}) indique l'appartenance au groupe traitement et la troisième (notée x_{iD}) indique le nombre d'années depuis l'obtention de l'opportunité.

Chaque observation appartient à un seul groupe. Une observation du groupe contrôle aura les valeurs $x_{iC} = 1$ et $x_{iT} = 0$. Inversement, une observation du groupe traitement aura les valeurs $x_{iC} = 0$ et $x_{iT} = 1$. Ainsi, pour chaque observation $i \in \{1, 2, \ldots, n\}$, l'ordonnée à l'origine est une combinaison linéaire

de x_{iC} et x_{iT} avec

$$x_{iC} + x_{iT} = x_{i1} = 1.$$

Notre modèle est donc non identifiable. Nous choisissons d'enlever l'ordonnée à l'origine de notre modélisation. Il nous reste alors les variables explicatives $\mathbf{X}_i = (x_{iC}, x_{iT}, x_{iD})^T$ pour chaque observation i.

L'estimation de la densité *a posteriori* avec davantage d'observations sera nécessairement plus précise. Y a-t-il une densité qui estime plus précisément avec un certain nombre d'observations ou même un nombre plus restreint d'observations ? Des tailles de 10, 50 et 100 observations par échantillon ont été étudiées lors de nos simulations. Nous choisissons de partager également le nombre d'observations appartenant à chacun des deux groupes. Ainsi, la moitié des observations d'un échantillon appartiendra au groupe contrôle avec comme vecteur de variables explicatives $\mathbf{X}_i = (1, 0, x_{iD})^T$ et l'autre moitié des observations appartiendra au groupe traitement avec comme vecteur de variables explicatives $\mathbf{X}_i = (0, 1, x_{iD})^T$.

Il ne reste qu'à définir la troisième variable explicative x_{iD}. L'idée est de choisir une loi de probabilité qui représente de façon approximative cette variable explicative. La figure 3.3 montre le diagramme à bâton de la variable x_{iD} pour notre vrai jeu de données. Nous choisissons d'appliquer une loi de Poisson de moyenne $\lambda = 1$ afin de générer notre troisième variable explicative.

Notre moteur de génération de variables explicatives est maintenant défini. Il ne nous reste qu'à déterminer le vecteur des coefficients $\boldsymbol{\beta} = (\beta_C, \beta_T, \beta_D)$ associé aux variables explicatives $\mathbf{X}_i = (x_{iC}, x_{iT}, x_{iD})^T$ pour attribuer une valeur Y_i à chaque observation avec l'équation (1.2.2). Il est intéressant de mesurer l'influence sur l'estimation des différents paramètres pour différentes valeurs de coefficients afin d'évaluer si un certain type de densités estime plus précisément sous certaines combinaisons de vrais coefficients. Nous introduisons ainsi notre dernier facteur influent : la valeur des vrais coefficients.

Puisqu'une observation se retrouve soit dans le groupe contrôle, soit dans le groupe traitement, l'idée est d'évaluer l'influence de l'apport du groupe traitement au groupe contrôle. C'est pourquoi nous fixons le coefficient associé au groupe contrôle à $\beta_C = 0$ (probabilité de succès de 50%) et étudierons différents coefficients associés au groupe traitement. Nous proposons les valeurs

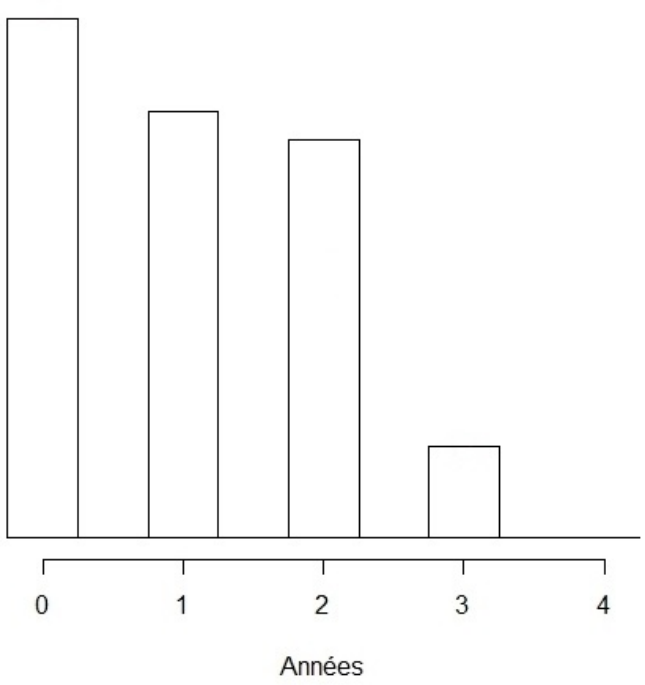


FIGURE 3.3. Distribution de la variable explicative *nombre d'années depuis l'obtention de l'opportunité*.

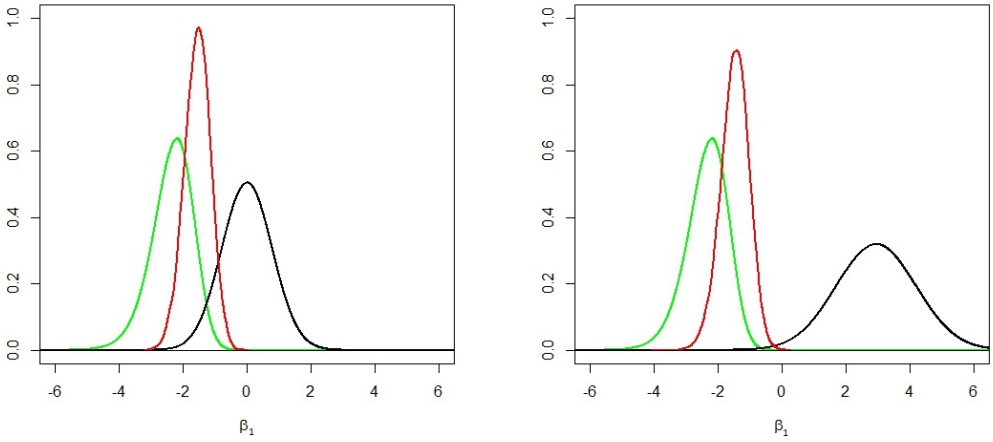
0,85, 1,39, 2,94 (probabilités de succès respectives de 70%, 80% et 95%) pour la vraie valeur du coefficient associée au groupe traitement. De même, nous désirons évaluer l'influence du coefficient associé à la variable x_{iD}. Basés sur l'individu moyen ($x_{iD} = 1$), nous introduirons une petite, moyenne et grande influence qui se traduisent par les valeurs $-0{,}04$, $-0{,}62$, $-1{,}39$ (probabilités de succès respectives de 49%, 35% et 20%).

De cette manière, nous générons des échantillons de différentes tailles et selon différentes probabilités de succès.

3.3. FONCTION DE VRAISEMBLANCE

Nous avons mentionné que la fonction de vraisemblance influence aussi la densité *a posteriori*. Nous avons la fonction de vraisemblance, notée l, pour le modèle logistique simple (sans variable explicative) définie par

$$l(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}) = \prod_{i=1}^{n} \frac{e^{\beta_1 Y_i}}{1 + e^{\beta_1}}.$$



(a) Mélange avec une densité normale centrée en 0 et de variance 1.

(b) Mélange avec une densité normale centrée en 2,94 et de variance $2{,}5^2$.

FIGURE 3.4. Mélange de la fonction de vraisemblance (vert) avec une densité *a priori* normale (noir) pour obtenir la densité *a posteriori* (rouge).

Cette fonction aura l'allure d'une cloche pour des valeurs de n et $\mathbf{Y}$ observées.

Voyons l'allure de la densité *a posteriori* lors du mélange de la fonction de vraisemblance avec la densité *a priori*. Notons que la méthode MCMC choisie (échantillonnage par tranche) pour l'estimation de la densité *a posteriori* est expliquée en détail à la section 2.2. La figure 3.4 montre deux exemples de densités *a priori* normales où 3 succès sur 20 sont observés. À première vue, nous constatons que l'effet de la fonction de vraisemblance sur la densité *a posteriori* semble beaucoup plus important que l'effet de la densité *a priori*.

L'influence de la vraisemblance est-elle trop forte pour distinguer les apports et les différences entre des densités *a priori* ? La figure 3.5 montre les densités *a posteriori* résultantes d'une vraisemblance à 3 succès sur 20 observations avec des densités *a priori* normale, Student à 3 degrés de liberté, Laplace et Gumbel, toutes centrées en 0 de variance $2{,}5^2$.

Malgré la forte influence de la vraisemblance, nous distinguons des différences entre les types de densités. Il est clair, cependant, que les densités *a posteriori* de la figure 3.5 sont plus similaires en comparaison avec les densités *a priori* de la figure 3.1 où les densités sont plus différentes. Des différences

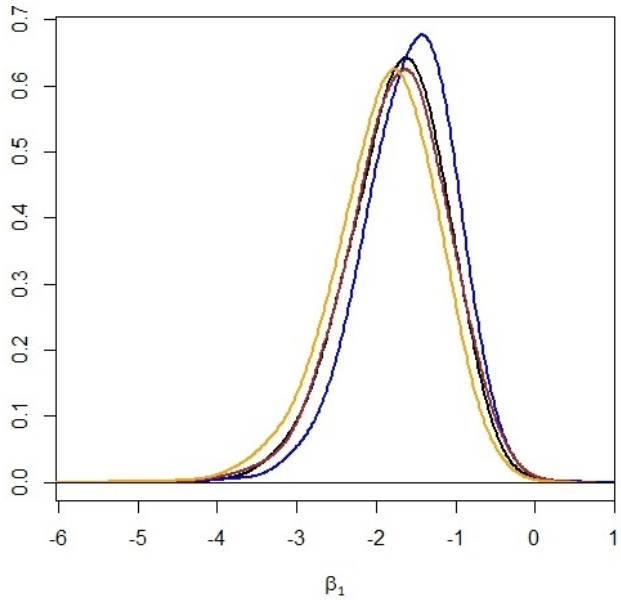


FIGURE 3.5. Comparatif de densités *a posteriori* avec la même vraisemblance (3 succès sur 20 observations) selon quatre densités de même position et échelle : normale (noir), Student à 3 degrés de liberté (bleu), Laplace (brun) et Gumbel (jaune).

sont donc attendues lors de nos simulations, mais reste à voir si celles-ci seront assez prononcées pour y voir un avantage réel lors de l'utilisation d'un certain type de densités *a priori*.

Les facteurs influents étudiés lors de nos simulations ont été introduits aux sections 3.1 et 3.2. Puis, nous avons vu à la section 3.3 l'importance de la vraisemblance sur l'estimation de la densité *a posteriori*.

Nous savons maintenant quoi étudier afin de déterminer la meilleure densité *a priori* pour la méthode MCMC d'échantillonnage par tranche. Nous présentons les résultats de nos comparaisons au chapitre 4.

Chapitre 4

SIMULATIONS

Aux chapitres précédents, nous avons vu différentes méthodes d'estimation telles que la méthode du maximum de vraisemblance, la méthode MCMC d'échantillonnage par tranche, la méthode de Gelman et la méthode de Genkin. Existe-t-il une méthode idéale pour notre cas d'étude de cohortes cherchant à mesurer l'efficacité d'un incitatif sur la réponse d'individus, c'est-à-dire un estimateur qui serait précis pour une variété de circonstances ?

Le résumé du schéma de simulation est décrit à la section 4.1, suivi des résultats à la section 4.2. Nous utilisons le terme *estimateur* pour faire référence aux différentes méthodes d'estimations comparées.

4.1. APERÇU DES SIMULATIONS

Afin d'établir l'estimateur le plus précis, il faut tout d'abord établir ce qu'est « être plus précis ». Nous cherchons en fait l'estimateur qui minimise l'erreur quadratique moyenne (*EQM*) définie par

$$\mathrm{EQM}\,(\theta) = \frac{1}{N}\sum_{l=1}^{N}\left(\hat{\theta}^{(l)} - \theta\right)^2, \tag{4.1.1}$$

où θ est la valeur du paramètre à estimer, $\hat{\theta}^{(l)}$ est l'estimation de θ pour l'itération l et N est le nombre total d'itérations. Un estimateur avec un EQM égal à 0 estimerait parfaitement le paramètre d'intérêt. Donc, plus l'EQM décroît vers 0, plus l'estimateur est précis.

Rappelons les différents niveaux des facteurs influents discutés aux sections 3.1 et 3.2. Nous avons :

(1) le **type de densité *a priori*** : normale, Student à 3 degrés de liberté, Laplace et Gumbel,

(2) la **moyenne** : $-2{,}94$; 0 et $2{,}94$,

(3) la **variance** : $\frac{1}{2} \times 2{,}5^2$; $2{,}5^2$ et $2 \times 2{,}5^2$,

(4) le **nombre d'observations par groupe** : 10 ; 50 et 100,

(5) la **valeur des vrais coefficients** à estimer :

   (a) l'effet β_C du groupe contrôle : 0,

   (b) l'effet β_T du groupe traitement : $0{,}85$; $1{,}39$ et $2{,}94$,

   (c) l'effet β_D du nombre d'années depuis la date d'obtention de l'opportunité : $-0{,}04$; $-0{,}62$ et $-1{,}39$.

Les paramètres des densités *a priori* utilisées pour la méthode MCMC d'échantillonnage par tranche discutée à la section 2.2 sont spécifiés par les trois premiers facteurs influents : le type de densité, la moyenne et la variance. Ainsi, un total de 36 densités *a priori* sont considérées. Nous ajoutons à cela les trois méthodes d'estimations suivantes :

(1) l'estimateur du maximum de vraisemblance présenté à la section 1.3,

(2) la méthode de Gelman avec la densité Cauchy centrée en 0 de paramètre d'échelle $2{,}5$ présentée à la section 2.4,

(3) la méthode de Genkin avec la densités *a priori* normale centrée en 0 avec variance σ_j^2 pour chaque β_j et densité *a priroi* Laplace centrée en 0 avec variance $\sqrt{2}/\sigma_j$ pour β_j présentée à la section 2.5.

Basé sur nos échantillons et suivant les recommandations de Genkin et coll. (2007), nous utilisons $\sigma_C^2 = 6$, $\sigma_T^2 = 6$ et $\sigma_D^2 = 1{,}5$ pour la méthode de Genkin.

Il y a donc un total de 40 estimateurs à comparer. Pour la construction des jeux de données sur lesquels nous testons ces 40 estimateurs, nous utilisons les deux derniers facteurs influents : le nombre d'observations et les coefficients à estimer. Nous obtenons 27 combinaisons différentes pour générer nos différents échantillons. Au total, 1000 échantillons sont générés pour chacune des 27 combinaisons.

Basés sur l'équation (4.1.1), nous obtenons l'équation

$$\mathrm{EQM}\left(\beta_{i_j}\right) = \frac{1}{1000} \sum_{l=1}^{1000} \left(\hat{\beta}_{i_j}^{(l)} - \beta_{i_j}\right)^2, \tag{4.1.2}$$

où β_{i_j} représente la vraie valeur du coefficient β_i associée au type d'échantillon $j \in \{1, 2, \ldots, 27\}$ et $\hat{\beta}_{i_j}^{(l)}$ est l'estimation obtenue du coefficient β_{i_j} pour l'échantillon $l \in \{1, 2, \ldots, 1000\}$ généré à partir de la combinaison j. L'équation (4.1.2) est donc l'EQM par combinaison d'échantillon j et pour chaque coefficient β_i pour $i \in \{C, T, D\}$.

Nous classons chaque estimateur par ordre de précision selon l'EQM pour les différents types d'échantillons à la section 4.2.

4.2. RÉSULTATS

Nous commençons par la présentation des résultats globaux à la sous-section 4.2.1, puis les résultats par taille échantillonnale à la sous-section 4.2.2, les résultats par influence des coefficients $\boldsymbol{\beta}$ à la sous-section 4.2.3 et finalement, les résultats lors de séparation complète à la section 4.2.4.

4.2.1. Résultats globaux

Commençons par regarder les EQM globaux de chaque estimateur afin d'établir l'estimateur le plus précis couvrant une variété de situations. Le tableau 4.1 ordonne les estimateurs du plus petit EQM au plus grand lorsque nous combinons tous les échantillons et tous les coefficients.

Nous voyons que la méthode de Gelman obtient de très bons résultats globaux. Cependant, la méthode MCMC avec les densités *a priori* normale et Student centrées en 0 et de variance 3,125 est plus précise. De plus, toutes les densités *a priori* centrées en 0 de variance 3,125 se classent parmi les estimateurs les plus précis globalement. En fait, plus la variance est petite plus l'EQM des estimateurs est petit. Dans le même ordre d'idées, les densités normales sont plus précises et les densités Gumbel sont globalement moins précises. L'estimateur du maximum de vraisemblance (EMV) est de loin le moins bon estimateur. Nous expliquons ce résultat par le petit nombre d'observations ($N = 10$) pour le tiers des échantillons. Nous discutons ce point à la section 4.2.2. Enfin, la méthode de Genkin, pour sa part, ne rivalise pas avec les meilleurs estimateurs.

TABLEAU 4.1. EQM calculé sur tous les échantillons et tous les coefficients combinés.

	Densité	Moyenne	Variance	EQM
1	Student	0,00	3,125	0,45
2	normale	0,00	3,125	0,46
3	Gelman	0,00	2,5	0,46
4	Gumbel	2,94	3,125	0,53
5	Laplace	0,00	3,125	0,54
6	Student	0,00	6,25	0,57
7	normale	2,94	3,125	0,60
8	normale	0,00	6,25	0,68
9	Gumbel	0,00	3,125	0,69
10	normale	2,94	6,25	0,88
11	Student	0,00	12,5	0,90
12	normale	-2,94	6,25	0,98
13	normale	-2,94	3,125	1,02
14	Laplace	2,94	3,125	1,07
15	Laplace	-2,94	3,125	1,20
16	normale	0,00	12,5	1,20
17	normale	-2,94	12,5	1,32
18	Gumbel	-2,94	3,125	1,37
19	Student	2,94	3,125	1,40
20	Student	2,94	6,25	1,41
21	normale	2,94	12,5	1,47
22	Laplace	0,00	6,25	1,51
23	Student	2,94	12,5	1,66
24	Student	-2,94	6,25	1,72
25	Student	-2,94	3,125	1,85
26	Student	-2,94	12,5	1,88
27	Gumbel	2,94	6,25	1,89
28	Laplace	-2,94	6,25	2,16
29	Laplace	2,94	6,25	2,20
30	Gumbel	0,00	6,25	2,34
31	Genkin - Laplace	0,00	selon β_j	3,16
32	Gumbel	-2,94	6,25	3,19
33	Genkin - normale	0,00	selon β_j	4,02
34	Laplace	0,00	12,5	5,53
35	Laplace	-2,94	12,5	6,22
36	Laplace	2,94	12,5	6,35
37	Gumbel	2,94	12,5	8,02
38	Gumbel	0,00	12,5	8,98
39	Gumbel	-2,94	12,5	10,36
40	EMV	-	-	208,77

Pour la suite de la section, gardons simplement les 10 estimateurs les plus précis du tableau 4.1, l'EMV et la méthode de Genkin par souci de concision. Les résultats complets sont laissés à l'annexe A.6.

En regardant l'EQM calculé par coefficient pour le top 10 des estimateurs au tableau 4.2, nous voyons que ces estimateurs ont sensiblement le même classement. Il n'y a donc pas une tendance très différente à celle observée au tableau 4.1. Cependant, nous voyons que pour la majorité des estimateurs par méthode MCMC, l'EQM sur β_D est plus petit que celui sur β_C, lui-même plus petit que l'EQM sur β_T. Ce qui n'est pas le cas pour la méthode de Genkin, où l'EQM sur β_T est là où l'estimateur est le plus précis. Nous retrouvons les

résultats de tous les estimateurs en annexe aux tableaux A.2, A.3 et A.4.

TABLEAU 4.2. EQM calculé sur tous les échantillons par coefficient.

Densité	Moyenne	Variance	EQM global	EQM β_C	EQM β_T	EQM β_D
Student	0,00	3,125	0.45	0.34	0.75	0.27
normale	0,00	3,125	0.46	0.47	0.59	0.32
Gelman	0,00	2,5	0.46	0.52	0.62	0.23
Gumbel	2,94	3,125	0.53	0.40	0.82	0.37
Laplace	0,00	3,125	0.54	0.52	0.72	0.37
Student	0,00	6,25	0.57	0.55	0.79	0.38
normale	2,94	3,125	0.60	0.63	0.74	0.43
normale	0,00	6,25	0.68	0.78	0.74	0.53
Gumbel	0,00	3,125	0.69	0.64	0.99	0.44
normale	2,94	6,25	0.88	0.81	1.20	0.62
Genkin - Laplace	0,00	selon β_j	3.16	3.09	1.47	4.92
Genkin - normale	0,00	selon β_j	4.02	5.76	1.64	4.66
EMV	-	-	208.77	135.61	366.43	124.28

4.2.2. Taille échantillonnale

Les tableaux 4.3, 4.4 et 4.5 montrent les EQM calculés avec tous les échantillons de taille $N = 10$, $N = 50$ et $N = 100$ (respectivement).

TABLEAU 4.3. EQM calculé sur tous les échantillons de taille $N = 10$ par coefficient.

Densité	Moyenne	Variance	EQM global	EQM β_C	EQM β_T	EQM β_D
Gumbel	2,94	3,125	0.69	0.77	0.71	0.59
Student	0,00	3,125	0.75	0.28	1.46	0.50
Gelman	0,00	2,5	0.80	0.65	1.28	0.46
normale	0,00	3,125	0.81	0.60	1.24	0.60
Laplace	0,00	3,125	0.86	0.54	1.28	0.77
Student	0,00	6,25	0.86	0.53	1.27	0.79
normale	2,94	3,125	1.07	1.51	0.98	0.72
Gumbel	0,00	3,125	1.13	0.96	1.54	0.88
normale	0,00	6,25	1.23	1.13	1.37	1.17
Genkin - Laplace	0,00	selon β_j	1.47	0.11	1.53	2.78
normale	2,94	6,25	1.52	1.76	1.55	1.25
Genkin - normale	0,00	selon β_j	2.27	2.85	0.76	3.21
EMV	-	-	363.96	223.78	544.39	323.71

Nous voyons au tableau 4.3 que l'ordre des estimateurs est sensiblement le même que l'ordre global du tableau 4.1 à l'exception près de la méthode de Genkin.

La méthode de Genkin avec une densité Laplace est de loin la plus précise pour estimer β_C. Aussi, elle n'est pas très loin derrière pour estimer β_T, mais devient très peu précise pour β_D. La méthode de Genkin avec une densité normale, pour sa part, se distingue pour l'estimation de β_T. En combinant tous les

TABLEAU 4.4. EQM calculé sur tous les échantillons de taille $N = 50$ par coefficient.

Densité	Moyenne	Variance	EQM global	EQM β_C	EQM β_T	EQM β_D
Gumbel	2,94	3,125	0.27	0.22	0.47	0.12
Gelman	0,00	2,5	0.28	0.27	0.44	0.12
normale	0,00	3,125	0.29	0.28	0.42	0.15
Laplace	0,00	3,125	0.29	0.22	0.52	0.14
Student	0,00	3,125	0.30	0.20	0.57	0.13
Student	0,00	6,25	0.31	0.25	0.54	0.15
normale	2,94	3,125	0.32	0.30	0.49	0.15
normale	0,00	6,25	0.33	0.32	0.48	0.17
Gumbel	0,00	3,125	0.37	0.33	0.60	0.18
normale	2,94	6,25	0.40	0.33	0.68	0.19
EMV	-	-	2.08	0.34	5.08	0.83
Genkin - Laplace	0,00	selon β_j	2.57	4.06	0.54	3.12
Genkin - normale	0,00	selon β_j	3.64	7.11	0.82	2.99

TABLEAU 4.5. EQM calculé sur tous les échantillons de taille $N = 100$ par coefficient.

Densité	Moyenne	Variance	EQM global	EQM β_C	EQM β_T	EQM β_D
normale	0,00	3,125	0.17	0.15	0.28	0.08
Gelman	0,00	2,5	0.18	0.15	0.31	0.07
Student	0,00	3,125	0.20	0.13	0.40	0.07
Laplace	0,00	3,125	0.20	0.13	0.40	0.07
Student	0,00	6,25	0.21	0.14	0.41	0.08
normale	0,00	6,25	0.21	0.17	0.39	0.09
Gumbel	2,94	3,125	0.22	0.14	0.45	0.07
normale	2,94	3,125	0.22	0.16	0.43	0.08
Gumbel	0,00	3,125	0.25	0.17	0.48	0.09
EMV	-	-	0.27	0.17	0.55	0.10
normale	2,94	6,25	0.28	0.17	0.58	0.09
Genkin - Laplace	0,00	selon β_j	3.23	6.30	0.00	3.38
Genkin - normale	0,00	selon β_j	3.88	8.30	0.01	3.32

coefficients (EQM global), la méthode de Genkin est loin d'être la plus précise.

La méthode MCMC avec densité *a priori* Gumbel excentrée à 2,94 de variance 3,125 devance les trois meilleures densités globales pour les tailles échantillonnales de $N = 10$ et $N = 50$. Elle se fait cependant dépasser en précision par un certain nombre d'estimateurs pour $N = 100$.

La méthode de Gelman obtient des résultats relativement semblables à la méthode MCMC avec densité *a priori* normale (centrée en 0 de variance 3,125) pour les trois tailles échantillonnales.

L'EMV est très peu précis pour des tailles de $N = 10$ et $N = 50$, mais gagne grandement en précision pour une taille de $N = 100$. Pour cette taille échantillonnale, l'EMV rivalise avec les meilleures méthodes.

Sans grande surprise, nous distinguons que plus la taille échantillonnale est petite, moins les estimateurs sont précis (plus grand EQM). À l'exception près de la méthode de Genkin qui est plus précise pour $N = 10$ que pour $N = 50$ ou même $N = 100$.

4.2.3. Coefficients β

Rappelons que les échantillons ont été générés avec différentes combinaisons de coefficients $\boldsymbol{\beta}$. Nous ne présentons pas l'entièreté des cas possibles ici. Cependant, un résultat vaut la peine d'être mentionné.

En combinant tous les échantillons où la valeur $\beta_T = 2{,}94$ a été utilisé pour générer les observations, nous calculons les EQM sur β_T de tous les estimateurs et nous les ordonnons au tableau 4.6.

Nous voyons que les densités *a priori* centrées en 2,94 (sauf pour les densités avec une grande variance de 12,5) se retrouvent davantage dans le peloton de tête dans cette circonstance que pour toute autre situation présentée jusqu'à maintenant. Cependant en regardant pour tous les coefficients combinés (EQM global), seule la méthode MCMC avec densité *a priori* normale centrée en 2,94 de variance 3,125 restent parmi les meilleurs. La méthode de Gelman et la méthode MCMC avec la densité *a priori* normale (centrées en 0 de variance 3,125) sont parmi les estimateurs les plus précis pour tous les coefficients combinés.

Il y a donc effectivement un avantage à centrer la densité *a priori* au bon endroit. En pratique, nous ne connaissons pas le vrai coefficient (il serait inutile d'estimer un coefficient connu !). Nous conseillons de jouer de prudence et opter en tout temps pour une densité *a priori* centrée en 0 avec une variance de 3,125. Le détail du reste des résultats est laissé en annexe aux tables A.8 à A.13.

4.2.4. Séparation complète

Comme discuté à la section 2.4, la séparation complète amène l'inexistence de l'estimateur du maximum de vraisemblance. C'est pourquoi nous ne regardons pas, ici, l'estimateur du maximum de vraisemblance.

TABLEAU 4.6. EQM calculé sur les échantillons générés avec $\beta_T = 2{,}94$.

	Densité	Moyenne	Variance	EQM β_T	EQM global
1	Genkin - normale	0,00	selon β_j	0.19	3.40
2	Student	2,94	3,125	0.36	0.63
3	normale	2,94	3,125	0.52	0.42
4	Laplace	2,94	3,125	0.54	0.60
5	Student	2,94	6,25	0.56	0.68
6	Gumbel	2,94	3,125	0.68	0.40
7	normale	2,94	6,25	0.78	0.58
8	normale	0,00	6,25	0.83	0.54
9	Student	2,94	12,5	0.84	0.81
10	Student	0,00	12,5	0.99	0.59
11	normale	0,00	3,125	0.99	0.51
12	Gelman	0,00	2,5	1.00	0.50
13	normale	0,00	12,5	1.02	0.76
14	Gumbel	0,00	3,125	1.03	0.58
15	normale	-2,94	12,5	1.06	0.79
16	Laplace	0,00	3,125	1.10	0.54
17	Student	0,00	6,25	1.13	0.55
18	Laplace	0,00	6,25	1.24	0.82
19	normale	2,94	12,5	1.24	0.90
20	Genkin - Laplace	0,00	selon β_j	1.29	2.93
21	normale	-2,94	6,25	1.30	0.76
22	Laplace	2,94	6,25	1.38	1.14
23	Gumbel	-2,94	3,125	1.39	0.94
24	Student	0,00	3,125	1.43	0.61
25	Gumbel	2,94	6,25	1.43	0.92
26	Laplace	-2,94	3,125	1.44	0.91
27	Student	-2,94	12,5	1.58	1.00
28	Student	-2,94	6,25	1.69	1.02
29	Laplace	-2,94	6,25	1.71	1.19
30	Gumbel	0,00	6,25	1.90	1.26
31	Student	-2,94	3,125	1.93	1.16
32	normale	-2,94	3,125	2.17	1.06
33	Gumbel	-2,94	6,25	2.58	1.76
34	Laplace	0,00	12,5	3.65	2.50
35	Laplace	-2,94	12,5	4.16	2.84
36	Laplace	2,94	12,5	4.19	2.95
37	Gumbel	2,94	12,5	5.84	3.61
38	Gumbel	0,00	12,5	6.99	4.33
39	Gumbel	-2,94	12,5	8.38	5.23
40	EMV	-	-	289.04	152.69

Le tableau 4.7 montre les EQM calculés uniquement sur des échantillons présentant une séparation complète. Encore une fois, la méthode MCMC avec densité normale centrée en 0 et de variance 3,125 est la plus précise suivie de la méthode de Gelman. Il faut mentionner aussi que la méthode de Genkin avec densité Laplace se positionne 6^e.

Avec toutes ces comparaisons, nous concluons que trois modèles se distinguent des autres sur la majorité des situations. La méthode de Gelman avec une densité *a priori* faiblement informative (Cauchy centrée en 0 avec paramètre d'échelle 2,5) et la méthode MCMC d'échantillonnage par tranche avec des densités *a priori* normale et Student à 3 degrés de liberté centrée en 0 de

TABLEAU 4.7. EQM calculé sur les échantillons présentant une séparation complète.

	Densité	Moyenne	Variance	EQM global	EQM β_C	EQM β_T	EQM β_D
1	normale	0,00	3,125	1.11	1.48	1.04	0.81
2	Gelman	0,00	2,5	1.16	1.81	1.16	0.52
3	Student	0,00	3,125	1.20	1.26	1.66	0.68
4	Laplace	0,00	3,125	1.57	2.10	1.65	0.95
5	normale	2,94	3,125	1.60	1.16	2.18	1.47
6	Genkin - Laplace	0,00	selon β_j	1.72	0.63	1.84	2.68
7	Student	0,00	6,25	1.82	2.25	2.23	0.99
8	normale	0,00	6,25	1.90	2.76	1.61	1.31
9	Gumbel	2,94	3,125	1.92	0.88	3.49	1.37
10	Gumbel	0,00	3,125	2.08	1.98	3.05	1.20
11	Genkin - normale	0,00	selon β_j	2.13	2.78	0.62	3.00
12	normale	2,94	6,25	2.66	1.95	4.19	1.84
13	normale	-2,94	3,125	2.97	4.83	3.05	1.02
14	Laplace	2,94	3,125	2.98	3.20	3.26	2.48
15	normale	-2,94	6,25	3.15	5.94	1.98	1.53
16	Student	0,00	12,5	3.38	3.98	4.48	1.67
17	Laplace	-2,94	3,125	3.56	5.54	3.12	2.01
18	normale	0,00	12,5	3.91	5.06	4.25	2.42
19	Gumbel	-2,94	3,125	4.39	6.25	4.35	2.56
20	Student	2,94	3,125	4.45	7.16	2.35	3.83
21	normale	-2,94	12,5	4.62	8.32	2.93	2.60
22	Student	2,94	6,25	4.68	6.86	3.56	3.63
23	normale	2,94	12,5	5.10	3.80	8.55	2.94
24	Student	2,94	12,5	5.98	7.65	6.27	4.03
25	Laplace	0,00	6,25	6.13	6.69	8.70	3.00
26	Student	-2,94	6,25	6.90	5.83	11.92	2.94
27	Student	-2,94	3,125	7.17	5.40	13.15	2.98
28	Student	-2,94	12,5	8.09	7.16	13.77	3.35
29	Laplace	2,94	6,25	8.35	8.20	12.02	4.81
30	Gumbel	2,94	6,25	8.58	3.97	18.39	3.39
31	Laplace	-2,94	6,25	8.61	9.88	11.64	4.29
32	Gumbel	0,00	6,25	9.85	6.89	18.29	4.39
33	Gumbel	-2,94	6,25	13.11	12.29	20.44	6.58
34	Laplace	0,00	12,5	26.01	22.10	44.45	11.49
35	Laplace	2,94	12,5	29.14	24.25	49.46	13.71
36	Laplace	-2,94	12,5	29.29	25.42	49.52	12.93
37	Gumbel	2,94	12,5	40.49	18.16	88.42	14.90
38	Gumbel	0,00	12,5	43.71	24.19	89.34	17.59
39	Gumbel	-2,94	12,5	49.03	32.60	92.98	21.50

variance 3,125.

Chapitre 5

APPLICATION

Au chapitre 4 nous avons comparé les différentes méthodes d'estimation sous différents angles. Nous appliquons ici nos méthodes d'estimation sélectionnées sur un jeu de données de marketing pour de l'assurance automobile et résidentielle. À des fins de confidentialité, certaines informations telles que les zones et les incitatifs ont été masqués.

Cherchant à augmenter le nombre de soumissions des produits d'assurance automobile et résidentielle, des incitatifs marketing ont été testés. Les clients ciblés ont, ici, ce que nous appelons une *opportunité*. L'opportunité est en fait un produit d'assurance (automobile ou résidentielle) d'un client potentiel ou existant, mais appartenant à un compétiteur. Par exemple, un client avec une opportunité automobile a une automobile assurée chez un compétiteur X. Nous considérons comme variable dans notre modèle le nombre d'années depuis l'obtention de cette opportunité (au plus grand entier inférieur) pour chaque client ciblé.

Les clients choisis habitent aussi dans trois zones jugées profitables (notées zone 1, zone 2 et zone 3). Chaque individu s'est vu attribuer un des trois différents incitatifs marketing (dénommés traitement 1, traitement 2 et traitement 3) ou aucun incitatif (dénommé contrôle).

Nous appliquons sur ces données les trois méthodes ayant démontré les estimations les plus précises globalement ainsi que la méthode du maximum de vraisemblance. Nous utiliserons la méthode MCMC d'échantillonnage par tranche avec des densités *a priori* normale et Student à 3 degrés de liberté centrées en 0 de variance 3,125 et la méthode de Gelman avec une Cauchy centrée

en 0 de paramètre d'échelle 2,5. Le tableau 5.1 montre les estimations des coefficients de la régression selon chaque méthode d'estimation. Il faut mentionner que par identifiabilité, l'ordonnée à l'origine inclut l'effet de la zone 1, du produit automobile et du groupe contrôle.

TABLEAU 5.1. Estimations des coefficients du jeu de données de marketing sur de l'assurance automobile et résidentielle.

Coefficient	Student	Normale	Gelman	EMV
Ordonnée à l'origine	-1.71	-1.63	-1.61	-1.65
Zone 2	-0.92	-0.98	-0.91	-0.90
Zone 3	-0.29	-0.34	-0.25	-0.24
Produit résidentiel	0.24	0.21	0.27	0.33
Traitement 1	-0.06	-0.13	-0.08	-0.04
Traitement 2	-1.14	-1.14	-1.04	-1.16
Traitement 3	-0.34	-0.34	-0.25	-0.23
Année depuis l'obtention de l'opportunité	-0.70	-0.71	-0.58	-0.61

Les conclusions suivantes sont vraies pour les quatre méthodes d'estimations présentées. Nous voyons au tableau 5.1 que la zone 1 à un impact favorable comparée à la zone 3 qui a elle-même un impact favorable comparée à la zone 2. Le produit résidentiel est favorable au produit automobile. Le groupe contrôle génère davantage de soumissions que les trois groupes traitements. Le traitement 1 étant favorable au traitement 3 qui lui même est favorable au traitement 2. Enfin, plus l'année depuis l'obtention de l'opportunité est grande, moins il y a de chance de faire une soumission. Les probabilités de succès (probabilité d'effectuer une soumission) pour chaque combinaison de zone, produit et incitatif marketing sont présentées au tableau 5.2. Par souci de concision, nous fixons l'année depuis l'obtention de l'opportunité à 1 an.

Les probabilités de succès sont assez semblables pour la densité *a priori* normale et Student. De même, les probabilités de succès sont similaires pour la méthode de Gelman et l'estimateur du maximum de vraisemblance. Dans tous les cas, nous obtenons la meilleure probabilité de succès estimée lorsque nous ciblons des clients pour de l'assurance résidentielle sans incitatif marketing dans la zone 1.

Il peut sembler étrange d'observer un effet favorable pour le groupe contrôle en comparaison aux trois traitements. C'est pourtant un phénomène que nous observons de plus en plus depuis les dernières années. La saturation du marché de l'assurance automobile et résidentielle amène une compétitivité avide avec un nombre grandissant de publicités. Nous distinguons un effet négatif

TABLEAU 5.2. Estimations des probabilités de succès par combinaison de zone, produit et incitatif marketing pour des individus avec 1 an depuis l'obtention de l'opportunité.

Zone	Produit	Incitatif	Student	Normale	Gelman	EMV
zone 1	automobile	contrôle	8,22%	8,82%	9,99%	9,45%
		traitement 1	7,80%	7,86%	9,28%	9,12%
		traitement 2	2,78%	3,00%	3,76%	3,17%
		traitement 3	6,00%	6,44%	7,99%	7,63%
	résidentiel	contrôle	10,25%	10,63%	12,71%	12,68%
		traitement 1	9,74%	9,49%	11,83%	12,25%
		traitement 2	3,52%	3,66%	4,88%	4,36%
		traitement 3	7,52%	7,80%	10,22%	10,30%
zone 2	automobile	contrôle	3,44%	3,51%	4,27%	4,05%
		traitement 1	3,26%	3,11%	3,95%	3,90%
		traitement 2	1,13%	1,15%	1,54%	1,31%
		traitement 3	2,48%	2,52%	3,37%	3,23%
	résidentiel	contrôle	4,35%	4,28%	5,52%	5,55%
		traitement 1	4,12%	3,79%	5,11%	5,35%
		traitement 2	1,43%	1,41%	2,02%	1,81%
		traitement 3	3,14%	3,08%	4,37%	4,44%
zone 3	automobile	contrôle	6,27%	6,46%	7,99%	7,61%
		traitement 1	5,94%	5,73%	7,41%	7,34%
		traitement 2	2,09%	2,16%	2,97%	2,52%
		traitement 3	4,55%	4,68%	6,35%	6,12%
	résidentiel	contrôle	7,86%	7,82%	10,22%	10,28%
		traitement 1	7,46%	6,95%	9,49%	9,93%
		traitement 2	2,65%	2,64%	3,85%	3,47%
		traitement 3	5,73%	5,69%	8,17%	8,31%

pour les campagnes de marché direct au point où l'individu moyen semble être dissuadé par l'incitatif marketing lui étant soumis.

Dans ce chapitre, nous avons appliqué et comparé les résultats des trois meilleures méthodes (MCMC avec densité *a priori* normale et Student à 3 degrés de liberté centrées en 0 de variance 3,125 et la méthode de Gelman avec une Cauchy centrée en 0 de paramètre d'échelle 2,5) ainsi que l'estimateur du maximum de vraisemblance sur un jeu de données de marketing pour des produits d'assurance. Comme nous avons pu le voir, aucun incitatif n'a pu générer un taux de soumission plus élevé que le groupe contrôle. Il serait donc raisonnable de réviser la méthode marketing et les incitatifs utilisés avant de produire une campagne marketing de plus grande envergure.

CONCLUSION

Dans ce mémoire, l'évaluation de l'efficacité d'un incitatif sur la réponse d'individus est abordée. À l'aide de la régression logistique, nous cherchions à déterminer l'approche optimale afin d'estimer le plus précisément l'effet de chacun des coefficients. C'est pourquoi, en plus d'utiliser la méthode d'estimation numérique du maximum de vraisemblance expliquée au chapitre 1, nous nous sommes tournés vers l'approche bayésienne et ces méthodes MCMC au chapitre 2. De plus, les méthodes introduites par Gelman et coll. (2008) et Genkin et coll. (2007) sont aussi expliquées au chapitre 2. L'utilisation d'une densité *a priori* amène la possibilité de comparer l'influence de certains facteurs sur l'estimation des coefficients. Ces facteurs se composent de différents types de densité, différents paramètres de centralité, différents paramètres d'échelle, différentes tailles échantillonnales et différentes valeurs des vrais coefficients à estimer et sont étudiés lors de nos simulations au chapitre 3. Les résultats des simulations se trouvent au chapitre 4. Les trois méthodes d'estimations les plus précises ont été appliquées sur un jeu de données de marketing pour des produits d'assurances automobile et résidentiels présentées au chapitre 5.

Suite à nos résultats de simulation, il est clair que la méthode MCMC d'échantillonnage par tranche avec densités *a priori* normale et Student à 3 degrés de liberté centrées en 0 de variance $\sigma^2 = 3{,}125$ sont les plus précises globalement.

Nous ajoutons que globalement, la densité normale est à favoriser tandis que la densité Gumbel est pour sa part la moins précise des quatre types de densités. La plupart du temps, plus la variance des densités *a priori* est petite, plus l'estimateur est précis. Centrer la densité *a priori* au bon endroit mène à de meilleurs résultats. Cependant, s'éloigner de la vraie valeur du coefficient à estimer mène à des résultats bien moins précis. Toutefois, centrée la densité *a priori* en 0 n'est jamais une mauvaise solution. Plus la taille échantillonnale est

grande, plus les estimateurs sont précis et diffèrent moins les uns des autres.

Trois autres méthodes faisaient office de comparatif à la méthode MCMC. La méthode du maximum de vraisemblance est de loin la méthode la moins précise principalement pour les tailles échantillonnales de $N = 10$ et $N = 50$. Le phénomène de séparation complète entraîne aussi l'inexistence de l'estimation. C'est pour toutes ces raisons que nous déconseillons l'utilisation de l'estimateur du maximum de vraisemblance pour 50 observations et moins. La méthode d'estimation ponctuelle de Genkin est globalement moins performante que la méthode MCMC. Elle se distingue cependant pour l'estimation de coefficients à forte influence ($\beta_i = 2{,}94$). Avec une densité normale, la méthode de Genkin est deux fois plus efficace que l'estimateur en deuxième position. Pour sa part, la méthode de Gelman obtient des résultats comparables aux meilleures densités *a priori* utilisées avec la méthode MCMC d'échantillonnage par tranche.

En somme, tester différents incitatifs marketing sous différentes conditions peut faire économiser beaucoup d'argent à une compagnie. Il faut cependant s'assurer d'utiliser des méthodes d'estimations précises. À la lumière de nos résultats de simulation, nous suggérons de jouer de prudence et d'utiliser les trois méthodes qui ont démontrés les résultats les plus précis de façon globale : la méthode MCMC avec les densités *a priori* normale centrée en 0 de variance $\sigma^2 = 3{,}125$, la méthode MCMC avec les densités *a priori* avec les densités *a priori* Student à 3 degrés de liberté centrée en 0 de variance $\sigma^2 = 3{,}125$ ainsi que la méthode de Gelman avec densité Cauchy centrée en 0 de paramètre d'échelle 2,5.

Bibliographie

Albert, A. et Anderson, J. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **71**, 1–10.

Carlin, B. P. et Louis, T. A. (2011). *Bayesian methods for data analysis*. CRC Press.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in medicine*, **27**, 2865–2873.

Gelman, A., Jakulin, A., Pittau, M. G. et Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, **2**, 1360–1383.

Gelman, A. et Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, **7**, 457–472.

Genkin, A., Lewis, D. D. et Madigan, D. (2007). Large-scale bayesian logistic regression for text categorization. *Technometrics*, **49**, 291–304.

Gordovil-Merino, A., Guardia-Olmos, J. et Pero-Cebollero, M. (2012). Estimation of logistic regression models in small samples. A simulation study using a weakly informative default prior distribution. *Psicologica*, **33**, 345–361.

Hoaglin, D. C., Mosteller, F. et Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*, vol. 3. Wiley New York.

Kipnis, C. et Varadhan, S. (1986). Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Communications in Mathematical Physics*, **104**, 1–19.

MacEachern, S. N. et Berliner, L. M. (1994). Subsampling the Gibbs sampler. *The American Statistician*, **48**, 188–190.

McCullagh, P. et Nelder, J. A. (1983). *Generalized linear models*. Chapman and Hall.

Neal, R. M. (2003). Slice sampling. *Annals of statistics*, **31**, 705–741.

Plummer, M., Best, N., Cowles, K. et Vines, K. (2006). Coda : Convergence diagnosis and output analysis for mcmc. *R news*, **6**, 7–11.

Plummer, M. et Stukalov, A. (2013). Package rjags. *update*, **16**, 1.

Raftery, A. E. et Lewis (1992). How many iterations in the Gibbs sampler. *Bayesian statistics*, **4**, 763–773.

Robert, C. P. et Casella, G. (2004). *Monte Carlo statistical methods*, vol. 319. Citeseer.

Tweedie, S. M. R. (1993). Generalized resolvents and Harris recurrence of Markov processes. *Doeblin and modern probability*, **149**, 227.

Annexe A

A.1. Paramétrisation de densités utilisées

TABLEAU A.1. Paramétrisation de densités utilisées

Nom	Notation	Domaine de y	Densité
Normal	$\mathcal{N}(\mu, \sigma^2)$	$\mathbb{R}$	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$
Poisson	$\mathcal{P}(\lambda)$	$\mathbb{N}$	$\frac{\lambda^y}{y!} \exp\{-\lambda\}$
Binomial	$\mathcal{B}(n, p)$	$\{1, \ldots, n\}$	$\binom{n}{y} p^y (1-p)^{n-y}$
Gamma	$\mathcal{G}(\alpha, \beta)$	$\mathbb{R}^+$	$\frac{1}{\Gamma(\beta)} \left(\frac{\beta}{\alpha}\right)^\beta y^{\beta-1} \exp\left\{-\frac{\beta}{\alpha} y\right\}$

A.2. Propriété d'une chaîne de Markov

A.2.1. Récurrence de Harris

Lorsque la chaîne a la capacité de visiter chaque ensemble A de $\mathcal{A}$ en tout temps à partir de n'importe quel état x de Ω, nous l'appelons *irréductible*. Posons $K^1(x, A) = K(x, A)$ où $K(x, A)$ est le noyau de transition. La k^e ($k > 1$) transition du noyau K est donnée par

$$K^k(x, A) = \int_\Omega K^{k-1}(y, A) K(x, y) dy.$$

Définition A.2.1 (Irréductibilité)**.** *Étant donnée une mesure ψ définie sur $\mathcal{A}$, une chaîne de Markov (X_t) avec noyau de transition K est appelée* ψ-irréductible *si $\forall A \in \mathcal{A}$ avec $\psi(A) > 0$, il existe une valeur k telle que pour tous $x \in \Omega$, $K^k(x, A) > 0$.*

Ainsi, une chaîne ψ-irréductible permet des transitions libres tout autour de l'espace d'états. Prenons par exemple

$$X_t = \theta X_{t-1} + \epsilon_t,$$

où $\theta \in \mathbb{R}$ et $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. Ce système est appelé un modèle AR(1). La chaîne de Markov (X_t) sera irréductible selon la *mesure de Lebesgue* λ. En effet, puisque

$X_t|X_{t-1} = x_{t-1} \sim \mathcal{N}(\theta x_{t-1}, \sigma^2)$, nous avons par définition que

$$K(x, A) > 0, \forall x \in \mathbb{R} \text{ et } \forall A \in \mathcal{A} \text{ tel que } \lambda(A) > 0.$$

Si toutefois, pour le modèle AR(1), $\epsilon_t \sim \mathcal{U}_{[-1,1]}$ et $|\theta| > 1$, la chaîne ne sera plus irréductible. Prenons $\theta > 1$. Nous avons

$$X_t - X_{t-1} \geq (\theta - 1) X_{t-1} - 1 \geq 0,$$

si $X_t \geq 1/(\theta - 1)$. La chaîne serait monotone croissante. Il sera donc impossible pour la chaîne de revenir sur ces états précédents. Il nécessite donc que $|\theta| < 1$ pour que la chaîne AR(1) soit irréductible avec $\epsilon_t \sim \mathcal{U}_{[-1,1]}$.

Il faut cependant s'assurer du retour de la chaîne en chacun des états. Notons par

$$\eta_A = \lim_{T \to \infty} \sum_{t=1}^{T} \mathbb{I}_A(X_t),$$

le nombre de passages de la chaîne (X_t) dans A.

Définition A.2.2 (Récurrence de Harris)**.** *Un ensemble* A *est* Harris récurrent *si* $\forall x \in A$, $P_x(\eta_A = \infty) = 1$. *La chaîne* (X_t) *est* récurrente au sens de Harris *s'il existe une mesure* ψ *telle que*

- (X_t) *est* ψ*-irréductible,*
- *pour chaque ensemble* A *avec* $\psi(A) > 0$, A *est Harris récurrent.*

Une chaîne de Markov récurrente au sens de Harris permet donc la transition libre tout autour de l'espace d'états à tout moment et assure le retour en tous ces états. Ce n'est cependant pas suffisant pour les méthodes MCMC. Elles nécessitent aussi une chaîne de Markov qui atteint un certain niveau de stabilité. Nous aimerions avoir une distribution de probabilité Π telle que pour $X_t \sim \Pi$ alors $X_{t+1} \sim \Pi$.

Définition A.2.3 (Mesure ivariante)**.** *Une mesure* Π *est dite* invariante *pour le noyau de transition* K *si* $\forall A \in \mathcal{A}$

$$\Pi(A) = \int_\Omega K(x, A) \Pi(x) dx.$$

S'il existe une mesure invariante Π pour une chaîne ψ-irréductible, Π sera dite une *mesure de probabilité invariante* (aussi appelée *distribution stationnaire*).

Lorsqu'une chaîne de Markov permet une telle mesure de probabilité invariante, cette mesure sera unique. Elle définit aussi le comportement sans mémoire (appelé également ergodicité) de la chaîne qui fait en sorte que la distribution initiale Π_0 de la chaîne (X_t) sera intraçable et ce, pour toute distribution initiale μ.

Toujours en suivant l'exemple précédent du modèle AR(1) avec noyau de transition correspondant à $\mathcal{N}\left(\theta x_{t-1}, \sigma^2\right)$, une distribution normale $\mathcal{N}\left(\eta, \tau^2\right)$ est stationnaire pour la chaîne AR(1) seulement si

$$\eta = \theta\eta \text{ et } \tau^2 = \tau^2\theta^2 + \sigma^2.$$

Ainsi, ces conditions mènent aux valeurs $\eta = 0$ et $\tau^2 = \sigma^2/\left(1-\theta^2\right)$ et l'unique distribution stationnaire sera $\mathcal{N}\left(0, \sigma^2/\left(1-\theta^2\right)\right)$. Il faut noter que $|\theta| < 1$ pour que $\tau^2 > 0$.

A.2.2. Apériodicité

Comment s'assurer que notre chaîne atteindra la distribution stationnaire ? Ajoutons la notion d'*apériodicité* à notre chaîne. Définissons tout d'abord ce qu'est un *petit ensemble* par la *condition de minorisation*.

Définition A.2.4 (Condition de minorisation et petit ensemble)**.** *Un noyau de transition* K *satisfait la* condition de minorisation *s'il existe une constante* $\epsilon > 0$, $m \in \mathbb{N}^*$ *et une mesure de probabilité* ν *tel que nous avons*

$$K^m(x, A) \geq \epsilon\nu(A),$$

pour tout $x \in \Omega$ *et pour tout* $A \in \mathcal{A}$*. En particulier, nous appelons* $C \in \mathcal{A}$ *un* petit ensemble (small set*) si la condition de minorisation est respectée pour tout* $x \in C$*. Plus généralement,* $C \in \mathcal{A}$ *sera un* petit ensemble *s'il existe* $m \in \mathbb{N}^*$ *et une mesure non nulle* ν_m *telle que*

$$K^m(x, A) \geq \nu_m(A),$$

pour tout $x \in C$ *et pour tout* $A \in \mathcal{A}$.

Reprenons l'exemple de la chaîne AR(1) et définissons un petit ensemble. Pour $X_t|X_{t-1} = x_{t-1} \sim \mathcal{N}\left(\theta x_{t-1}, \sigma^2\right)$, le noyau de transition est borné inférieurement par

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\left(x_t^2 - 2\theta x_t \underline{\omega} + \theta^2 \min\left(\bar{\omega}^2, \underline{\omega}^2\right)\right)/2\sigma^2\right\} \text{ si } x_t > 0,$$
$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\left(x_t^2 - 2\theta x_t \bar{\omega} + \theta^2 \min\left(\bar{\omega}^2, \underline{\omega}^2\right)\right)/2\sigma^2\right\} \text{ si } x_t < 0,$$

lorsque $x_{t-1} \in [\underline{\omega}, \bar{\omega}]$. Posons la mesure de probabilité

$$\nu(x) = \frac{1}{\sqrt{2\pi}\sigma} \frac{\exp\left\{-\left(x^2 - 2\theta x \underline{\omega}\right)/2\sigma^2\right\}\mathbb{1}_{\{x>0\}} + \exp\left\{-\left(x^2 - 2\theta x \bar{\omega}\right)/2\sigma^2\right\}\mathbb{1}_{\{x<0\}}}{\Phi\left(-\theta\underline{\omega}/\sigma^2\right)\exp\{\theta^2\underline{\omega}^2\} + \left[1 - \Phi\left(-\theta\underline{\omega}/\sigma^2\right)\right]\exp\{\theta^2\bar{\omega}^2\}},$$

et

$$\epsilon = \Phi\left(-\theta\underline{\omega}/\sigma^2\right)\exp\left\{\theta^2\underline{\omega}^2\right\} + \left[1 - \Phi\left(-\theta\underline{\omega}/\sigma^2\right)\right]\exp\left\{\theta^2\bar{\omega}^2\right\},$$

où $\Phi(\cdot)$ est la fonction de répartition d'une loi normale standard. Ainsi, pour $m = 1$ nous avons

$$\nu_1(x) = \epsilon\nu(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\left(x^2 - 2\theta x\underline{\omega}\right)/2\sigma^2\right\}\mathbb{1}_{\{x>0\}} + \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\left(x^2 - 2\theta x\bar{\omega}\right)/2\sigma^2\right\}\mathbb{1}_{\{x<0\}},$$

qui est toujours inférieur ou égal au noyau de transition K sur l'ensemble $C = [\underline{\omega}, \bar{\omega}]$. Ainsi par construction, l'ensemble C sera un petit ensemble. Cet exemple provient de la section 6.3 du livre de Robert et Casella (2004) en page 215.

Théorème A.2.1. *Prenons une chaîne (X_t) ψ-irréductible. Pour tout $A \in \mathcal{A}$ tel que $\psi(A) > 0$, il existe $m \in \mathbb{N}^*$ et un petit ensemble $C \subset A$ tel que la mesure associée à la condition de minorisation satisfait $\nu_m(C) > 0$. De plus, il est possible de décomposer l'ensemble Ω en une partition dénombrable de petits ensembles.*

La preuve peut être trouvée dans Tweedie (1993). Supposons que C est un petit ensemble pour M. Lorsque la chaîne atteint C, il y aura donc une probabilité non nulle de retourner à C après M itérations puisque $K^M(x, A) \geq \nu_M(A) > 0$.

Définition A.2.5 (Périodicité)**.** *Une chaîne (X_n) ψ-irréductible a un* cycle de longueur d *s'il existe un petit ensemble C pour un entier positif M et une distribution*

(a) Exemple de chaîne de période $d = 2$

(b) Exemple de chaîne apériodique

FIGURE A.1. Chaînes de Markov irréductibles à états discrets.

de probabilité associée ν_M *telle que* d *est le plus grand commun diviseur de l'ensemble* $\{k \geq 1 | \forall x \in C$ *et* $\forall A \in \mathcal{A}, \exists \delta_k > 0$ *tel que* C *est un petit ensemble pour* $\nu_k \geq \delta_k \nu_M\}$.

Le cycle de largeur d étant indépendant du petit ensemble C, d caractérise intrinsèquement la chaîne (X_n). De plus, nous définissons la *période* de la chaîne par le plus grand entier D satisfaisant la définition précédente et (X_n) sera dite *apériodique* pour $D = 1$. Il est plus facile d'illustrer la périodicité pour le cas discret. La période d'une chaîne irréductible à états discrets est donnée par le plus grand commun diviseur d de l'ensemble

$$\left\{k \geq 1 | K^k(x, x) > 0\right\}, \tag{A.2.1}$$

pour n'importe lequel $x \in \Omega$. Une condition suffisante et non nécessaire pour obtenir l'apériodicité est d'avoir la possibilité de revenir à un état $x \in \Omega$ en un seul pas. Prenons l'exemple illustré par les figures A.1. Il est impossible pour la chaîne de la figure A.1(a) de revenir en son état de départ en un seul pas. Du coup, il est seulement possible de revenir en chaque état en $\{2, 4, 6, 8, \ldots\}$ pas. La période de la chaîne est donc $d = 2$. Pour la figure A.1(b), il est toutefois possible pour l'état 2 de revenir sur lui-même en un seul pas. Ainsi, la chaîne sera apériodique.

Pour les chaînes de Markov continues, lorsque le noyau de transition a une composante absolument continue par rapport à la mesure de Lebesgue, avec densité $f(\cdot|x_{t-1})$, une condition suffisante pour l'apériodicité est d'avoir $f(\cdot|x_{t-1})$ positif sur un voisinage de x_{t-1}. La chaîne peut rester dans le voisinage de x_{t-1} pour un certain nombre de temps avant de visiter tout autre ensemble. La chaîne AR(1) avec $\epsilon_t \sim \mathcal{U}_{[-1,1]}$ et $|\theta| < 1$ est apériodique.

A.2.3. Ergodicité

L'apériodicité combinée avec la récurrence de Harris nous assurera l'existence et l'unicité de la distribution stationnaire et sa convergence comme distribution limite. Définissons la norme de variation totale.

Définition A.2.6 (Norme de variation totale)**.** *Pour des mesures* ω_1 *et* ω_2*, la* norme de variation totale *est définie telle que*

$$\|\omega_1 - \omega_2\|_{TV} = \sup_{A \in \mathcal{A}} |\omega_1(A) - \omega_2(A)|.$$

Théorème A.2.2. *Si la chaîne* (X_t) *est récurrente au sens de Harris et apériodique, alors*

$$\lim_{n\to\infty} \left\| \int K^n(x, \cdot)\, \mu(x)\, dx - \Pi \right\|_{TV} = \lim_{n\to\infty} \sup_{A \in \mathcal{A}} \left| \int K^n(x, A)\, \mu(x)\, dx - \Pi(A) \right| = 0$$

pour toute distribution initiale μ*.*

Tweedie (1993) affirme qu'avoir une chaîne de Markov étant récurrente au sens de Harris et apériodique est équivalent à l'ergodicité.

L'estimation de la distribution stationnaire Π se fait de manière peu pratique. En effet, le théorème A.2.2 indique que si la chaîne roule assez longtemps ($t \to \infty$), l'observation résultante proviendra de la distribution stationnaire. Donc, pour estimer cette distribution, il faudrait recommencer un grand nombre de fois l'opération ce qui est inefficace.

Pourquoi ne pas tout simplement garder l'échantillon $(X_1, \ldots, X_t)$ que génère la chaîne et utiliser directement la loi des grands nombres ou même le théorème limite central pour inférer sur la distribution stationnaire ? Deux obstacles s'interposent :

- La dépendance markovienne des observations (X_t est dépendant de X_{t-1}),
- La non-stationnarité de la séquence d'observations (sauf si $X_0 \sim \Pi$).

Il faut donc, contourner le problème en regardant le comportement limite de sommes partielles de la séquence d'observations. Un résultat analogue à la loi des grands nombres appelé théorème d'ergodicité en découle.

Théorème A.2.3 (Théorème d'ergodicité)**.** *Si* (X_t) *est récurrente au sens de Harris avec distribution stationnaire* Π*, alors pour toute fonction h telle que* $\mathbb{E}|h| < \infty$ *nous*

avons

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} h(X_i) = \int h(x)\, d\Pi(x).$$

Ainsi, les observations d'une même chaîne de Markov peuvent être utilisées pour inférer sur la distribution stationnaire.

A.2.4. Réversibilité

Pour obtenir quelque chose d'encore plus puissant comme le théorème limite central, nous devons imposer une dernière propriété à notre chaîne, la réversibilité.

Définition A.2.7 (Réversibilité)**.** *Une chaîne de Markov* (X_t) *est dite* réversible *si la distribution de* X_{t+1} *conditionnellement à* $X_{t+2} = x$ *est la même que la distribution de* X_{t+1} *conditionnellement à* $X_t = x$*.*

Ainsi, la direction du temps n'a pas d'importance. Il existe cependant une façon plus facile de vérifier la réversibilité d'une chaîne.

Définition A.2.8 (Balance détaillée)**.** *Une chaîne de Markov avec noyau de transition* K *satisfait la condition de* balance détaillée *(*detailed balance*) s'il existe une fonction* f *satisfaisant*

$$K(y, x) f(y) = K(x, y) f(x), \forall x, y \in \Omega. \tag{A.2.2}$$

Malgré le fait qu'il n'est pas nécessaire pour f d'être la distribution stationnaire Π associée au noyau de transition K, la balance détaillée reste une condition suffisante pour la réversibilité (souvent plus facile à vérifier que la réversibilité elle-même).

Théorème A.2.4. *Supposons qu'une chaîne de Markov avec noyau de transition* K *vérifie la condition de balance détaillée avec* Π *comme fonction de probabilité. Alors,*

- *la densité* Π *est la densité invariante de la chaîne,*
- *la chaîne est réversible.*

En combinant récurrence de Harris et réversibilité, nous obtenons les conditions nécessaires du *théorème limite central* (*TLC*) pour une chaîne de Markov démontré par Kipnis et Varadhan (1986).

Théorème A.2.5. *Si la chaîne de Markov est récurrente au sens de Harris et réversible, le* théorème limite central

$$\frac{1}{\sqrt{N}}\left(\sum_{t=1}^{N}\left(h\left(X_t\right)-\mathbb{E}^{\Pi}\left[h\left(X_t\right)\right]\right)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0,\gamma_h^2\right),$$

sera applicable lorsque

$$0<\gamma_h^2=\mathbb{E}^{\Pi}\left[\overline{h}^2\left(X_0\right)\right]+2\sum_{t=1}^{\infty}\mathbb{E}^{\Pi}\left[\overline{h}\left(X_0\right)\overline{h}\left(X_t\right)\right]<\infty,$$

où $\overline{h}$ *représente* $h-\mathbb{E}^{\Pi}[h]$ *pour* $\mathbb{E}|h|<\infty$ *et* $\mathbb{E}\left|h^2\right|<\infty$.

Le théorème limite central nous assure donc la convergence en loi de la moyenne échantillonnale de notre chaîne (X_t) vers l'espérance de l'unique distribution stationnaire selon une loi normale et ce, peu importe la distribution initiale $X_0 \sim \mu$.

A.3. ÉCHANTILLONNAGE PAR TRANCHE

Posons $h(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X})$ telle que

$$h(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X})=f(\mathbf{Y}|\boldsymbol{\beta},\mathbf{X})\,\pi(\boldsymbol{\beta}|\mathbf{X})\propto\pi(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X}).$$

Introduisons une variable aléatoire auxiliaire, notée U. Posons la loi conditionnelle

$$U|\boldsymbol{\beta}\sim\mathcal{U}\left(0,h(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X})\right).$$

Puisque la fonction de densité conditionnelle est

$$f(u|\boldsymbol{\beta})=\frac{1}{h(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X})}\mathbb{1}_{\{u<h(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X})\}}, \tag{A.3.1}$$

nous avons

$$f(u,\boldsymbol{\beta})=f(u|\boldsymbol{\beta})\,\pi(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X})\propto\frac{1}{h(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X})}\mathbb{1}_{\{u<h(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X})\}}h(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X})=\mathbb{1}_{\{u<h(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X})\}}.$$

De plus, nous avons

$$f(\boldsymbol{\beta}|u)=\frac{f(u,\boldsymbol{\beta})}{f(u)}=f(u,\boldsymbol{\beta})\propto\mathbb{1}_{\{\beta|u<h(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X})\}}.$$

La méthode d'échantillonnage par tranche consiste à choisir itérativement des valeurs de u et de $\boldsymbol{\beta}$. À partir d'une valeur de départ $\boldsymbol{\beta}^{(0)}$, nous avons les étapes suivantes à l'itération k :

(1) Tirer $u^{(k)}$ uniformément sur l'intervalle $\left(0,h\left(\boldsymbol{\beta}^{(k-1)}|\mathbf{Y},\mathbf{X}\right)\right)$,

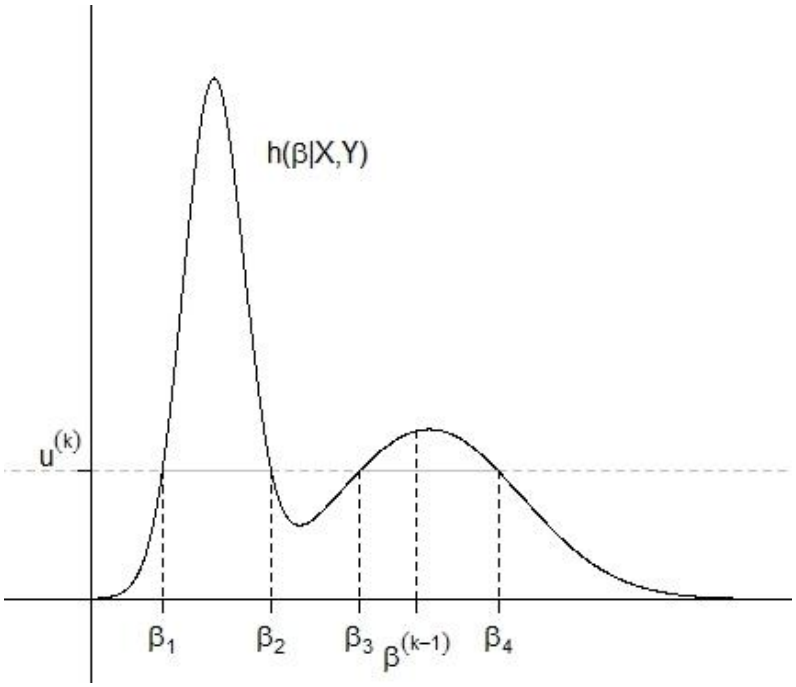


FIGURE A.2. Échantillonnage par tranche pour une loi univariée

(2) Tirer $\boldsymbol{\beta}^{(k)}$ uniformément sur l'ensemble $S = \left\{\boldsymbol{\beta} | h\left(\boldsymbol{\beta} | \mathbf{Y}, \mathbf{X}\right) > u^{(k)}\right\}$.

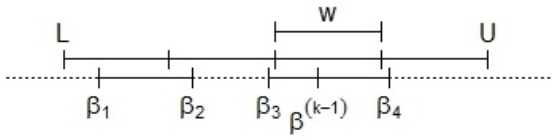
Cette méthode se résume bien graphiquement. Le cas univarié est présenté à la figure A.2. Nous y voyons l'ensemble S comme une *tranche* de $h\left(\boldsymbol{\beta} | \mathbf{Y}, \mathbf{X}\right)$, soit la raison pour laquelle nous nommons la méthode : échantillonnage par tranche.

L'étape (1) est triviale. En effet, après avoir calculé $h\left(\boldsymbol{\beta}^{(k-1)} | \mathbf{Y}, \mathbf{X}\right)$, il est facile d'échantillonner $u^{(k)}$ uniformément sur l'intervalle $\left(0, h\left(\boldsymbol{\beta}^{(k-1)} | \mathbf{Y}, \mathbf{X}\right)\right)$. Il est toutefois plus difficile d'effectuer l'étape (2). Puisque $h\left(\boldsymbol{\beta} | \mathbf{Y}, \mathbf{X}\right)$ n'est pas nécessairement inversible, l'ensemble $S = \left\{\boldsymbol{\beta} | h\left(\boldsymbol{\beta} | \mathbf{Y}, \mathbf{X}\right) > u^{(k)}\right\}$ n'est pas toujours simple à déterminer.

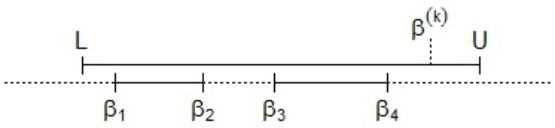
Pour échantillonner sur l'ensemble S, nous appliquons une procédure pas-à-pas. Suivant l'exemple de la figure A.2, la figure A.3 illustre la procédure pas à pas expliquée en détail (voir plus bas) pour le cas univarié au moment du k^e échantillonnage. Les étapes sont :

(1) définir l'intervalle $I = (L, U)$,

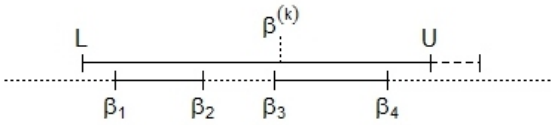
(2) échantillonner uniformément $\boldsymbol{\beta}^{(k)}$ sur I. Choisir la valeur échantillonnée si elle est incluse sur l'ensemble S, sinon passer à l'étape suivante,



(a) Définir l'intervalle $I = (L, U)$.



(b) Échantillonner uniformément $\boldsymbol{\beta}^{(k)}$ sur I jusqu'à ce que la valeur soit incluse dans l'ensemble S.



(c) Réduire l'intervalle I entre les échantillonnages successifs.

FIGURE A.3. Procédure pas-à-pas pour une loi univariée

(3) réduire l'intervalle I, rejeter la valeur $\boldsymbol{\beta}^{(k)}$ et recommencer l'étape précédente.

Pour l'étape (1), nous observons en A.3(a) que pour définir l'intervalle I nous devons tout d'abord choisir un intervalle de longueur arbitraire w autour de $\boldsymbol{\beta}^{(k-1)}$, puis éloigner la borne inférieure d'une distance w jusqu'à ce que celle-ci soit exclue de l'ensemble S. La valeur obtenue composera la borne inférieure L de l'intervalle I. Nous effectuons la même procédure pour définir la borne supérieure U afin d'obtenir l'intervalle $I = (L, U)$ qui, idéalement, seraient tel que $I \setminus S = \emptyset$. Or, cette condition n'est pas nécessaire. Évidemment, il est parfois même impossible d'obtenir $I \setminus S = \emptyset$ lorsque $h(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X})$ n'est pas unimodale comme le montre la figure A.3.

Ensuite à l'étape (2), nous échantillonnons uniformément une valeur de $\boldsymbol{\beta}^{(k)}$ sur l'intervalle I. Nous évaluons la valeur $\boldsymbol{\beta}^{(k)}$. Si cette valeur ne fait pas partie de l'ensemble S comme en A.3(b), nous rejetons $\boldsymbol{\beta}^{(k)}$. Si par contre $\boldsymbol{\beta}^{(k)}$

est incluse dans S, nous gardons cette valeur.

Si nous rejetons la valeur $\boldsymbol{\beta}^{(k)}$ à l'étape (2), nous réduisons l'intervalle I à l'étape (3). Si la valeur $\boldsymbol{\beta}^{(k)}$ rejetée est plus petite que $\boldsymbol{\beta}^{(k-1)}$, L devient $L = \boldsymbol{\beta}^{(k)}$. Si toutefois la valeur $\boldsymbol{\beta}^{(k)}$ rejetée est plus grande que $\boldsymbol{\beta}^{(k-1)}$, U devient $U = \boldsymbol{\beta}^{(k)}$. Comme illustré en A.3(c) nous recommençons l'étape (2) sur le nouvel intervalle I.

Nous remarquons qu'en tout temps $\boldsymbol{\beta}^{(k-1)} \in I$. Ainsi, la procédure pas à pas nous garantit qu'elle prendra fin à un certain moment et que nous obtiendrons une valeur $\boldsymbol{\beta}^{(k)} \in S$ (Neal (2003)). De plus, la procédure devrait s'effectuer sur un nombre plutôt restreint d'itérations puisque l'intervalle I se refermera de plus en plus sur l'ensemble S non vide. Aussi, nous pouvons montrer que chaque itération laisse la distribution stationnaire invariante. Ainsi pour chaque itération après convergence de la chaîne en sa distribution stationnaire, nous échantillonnons sur la densité *a posteriori*. Pour plus de détail sur la preuve, consultez Neal (2003).

La procédure pas à pas est généralisée pour les cas multivariés. Avec le vecteur $\boldsymbol{\beta}^{(k-1)} = \left(\beta_1^{(k-1)}, \ldots, \beta_p^{(k-1)}\right)$, nous itérons chaque composante de $\boldsymbol{\beta}^{(k-1)}$ conditionnellement aux autres. La figure A.4 illustre la procédure pas-à-pas pour une loi bivariée. Ainsi à l'étape k, il y a p tirages au lieu d'un seul pour le cas univarié. Pour chaque composante l de $\boldsymbol{\beta}^{(k-1)}$, l'intervalle I est défini autour de $\left(\beta_1^{(k)}, \ldots, \beta_{l-1}^{(k)}, \beta_l^{(k-1)}, \ldots, \beta_p^{(k-1)}\right)$ afin d'échantillonner $\beta_l^{(k)}$ sur I pour actualiser $\beta_l^{(k-1)}$. Il existe des procédures plus efficaces que la méthode pas-à-pas présentée ici. Pour de plus amples détails, consultez Neal (2003).

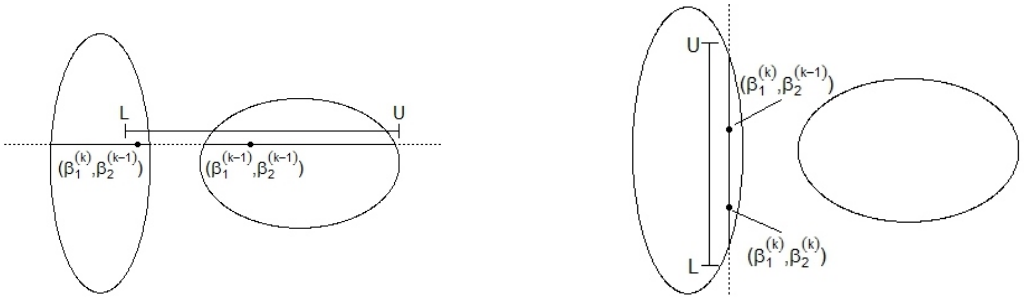
A.4. Facteur de réduction d'échelle

Définissons les notions de variance intra chaîne et variance inter chaîne de façon à décomposer la variance totale. Pour $m > 1$ chaînes, nous avons la variance intra chaîne W_i associée à β_i définie par

$$W_i = \frac{1}{m} \sum_{j=1}^{m} s_{i_j}^2,$$

avec

$$s_{i_j}^2 = \frac{1}{K-1} \sum_{l=K+1}^{2K} \left(\beta_{i_l j} - \overline{\beta}_{i.j}\right)^2,$$



(a) Échantillonner uniformément $\beta_1^{(k)}$ sur $I = (L, U)$ jusqu'à ce que la valeur soit incluse dans l'ensemble S.

(b) Échantillonner uniformément $\beta_2^{(k)}$ sur $I = (L, U)$ jusqu'à ce que la valeur soit incluse dans l'ensemble S.

FIGURE A.4. Procédure pas à pas pour une loi bivariée

où $\overline{\beta}_{i.j}$ est la moyenne des observations de la chaîne j pour le coefficient β_i. Nous avons aussi la variance inter chaîne B_i/K associée à β_i définie par

$$\frac{B_i}{K} = \frac{1}{m-1}\sum_{j=1}^{m}\left(\overline{\beta}_{i.j} - \overline{\beta}_{i..}\right)^2,$$

où $\overline{\beta}_{i..}$ est la moyenne de toutes les observations des m chaînes pour le coefficient β_i. Puisque l'estimation de la variance totale de β_i peut s'estimer par une somme pondérée de la variance intra chaîne et la variance inter chaîne, nous avons

$$\widehat{Var}\left(\beta_i\right) = \left(\frac{K-1}{K}\right)W_i + \frac{B_i}{K}. \tag{A.4.1}$$

De façon conservatrice, pour chaque chaîne il est important de choisir une valeur initiale selon une distribution surdispersée par rapport à la densité *a posteriori*. De cette manière, l'estimateur de la variance de l'équation (A.4.1) ne sous-estimera pas la variance des coefficients β_i. De plus, l'estimateur $\widehat{Var}\left(\beta_i\right)$ sera sans biais pour $Var\left(\beta_i\right)$ sous la stationnarité ou lorsque $K \to \infty$.

La mesure de convergence proposée par Gelman et Rubin (1992) se base sur l'hypothèse de normalité de la distribution stationnaire. Un intervalle de crédibilité est construit à partir d'une loi de Student de moyenne

$$\hat{\mu}_i = \overline{\beta}_{i..},$$

de paramètre d'échelle

$$\sqrt{\hat{V}_i} = \sqrt{\widehat{Var}\left(\beta_i\right) + \left(\frac{1}{m}\right)\frac{B_i}{K}},$$

et de degrés de liberté estimés par la méthode des moments

$$\widehat{df}_i = 2\frac{\hat{V}_i^2}{\widehat{Var}\left(\hat{V}_i\right)},$$

où

$$\widehat{Var}\left(\hat{V}_i\right) = \frac{1}{m}\left(\frac{K-1}{K}\right)^2 \widehat{Var}\left(s_{i_j}^2\right) + \frac{1}{K^2}\left(\frac{m+1}{m}\right)^2 \left(\frac{2}{m-1}\right) B_i^2 + 2\left(\frac{m+1}{m^2}\right)\left(\frac{K-1}{K}\right) \cdot \left[\widehat{Cov}\left(s_{i_j}^2, \overline{\beta}_{i_{\cdot j}}^2\right) - 2\overline{\beta}_{i\cdot\cdot}\,\widehat{Cov}\left(s_{i_j}^2, \overline{\beta}_{i_{\cdot j}}\right)\right].$$

Nous remarquons que $\widehat{df}_i \to \infty$ lorsque $K \to \infty$. Le facteur de réduction d'échelle associé à β_i est définie par

$$\sqrt{\hat{R}_i} = \sqrt{\frac{\hat{V}_i}{W_i}\left(\frac{\widehat{df}_i}{\widehat{df}_i - 2}\right)} = \sqrt{\frac{1}{K}\left[K - 1 + \left(\frac{m+1}{m}\right)\frac{B_i}{W_i}\right]\left(\frac{\widehat{df}_i}{\widehat{df}_i - 2}\right)}.$$

Le facteur de réduction d'échelle $\hat{R}_i$ est donc le ratio entre la variance estimée $\hat{V}_i$ et la variance intra chaîne W_i avec un facteur de correction de variance provenant d'une loi de Student. En plus d'une estimation ponctuelle, une borne supérieure peut être approximé en se basant sur le fait que

$$\frac{B_i}{W_i} \sim \mathcal{F}\left(m-1, \frac{2mW_i^2}{\widehat{Var}\left(s_{i_j}^2\right)}\right).$$

Ainsi, une borne supérieur 97,5% approximée du facteur de réduction peut être calculée. Nous remarquons que l'approximation ignore la variabilité causée par $\widehat{df}_i / \left(\widehat{df}_i - 2\right)$.

A.5. Séparation complète

Prenons le modèle logistique avec comme unique régresseur x. Nous avons

$$\mathbb{E}(Y_i | x_i, \beta_1, \beta_2) = \frac{e^{\beta_1 + \beta_2 x_i}}{1 + e^{\beta_1 + \beta_2 x_i}},$$

pour $i \in \{1, 2, \ldots, n\}$. Supposons maintenant qu'il existe une valeur x_0 telle que pour chaque observation $(x_i; y_i)$

$$y_i = \begin{cases} 0 & \text{si } x_i < x_0\,; \\ 1 & \text{si } x_i \geq x_0. \end{cases}$$

Cherchons les conditions sur β_1 et β_2 telles que

$$\mathbb{E}(Y_i|x_i, \beta_1, \beta_2) \approx \begin{cases} 0 & \text{si } x_i < x_0\,; \\ 1 & \text{si } x_i \geq x_0. \end{cases}$$

Posons $\delta > 0$ tel que le changement soit abrupt dans l'intervalle $(x_0 - \delta; x_0 + \delta)$. Il faut donc chercher les valeurs de β_1 et β_2 telles que

$$\mathbb{E}(Y_i|x_0 - \delta, \beta_1, \beta_2) = \varepsilon, \tag{A.5.1}$$

$$\mathbb{E}(Y_i|x_0 + \delta, \beta_1, \beta_2) = 1 - \varepsilon, \tag{A.5.2}$$

où $\epsilon > 0$ est petit.

En utilisant l'équation (A.5.1), nous obtenons

$$\begin{aligned} &\mathbb{E}(Y_i|x_0 - \delta, \beta_1, \beta_2) = \varepsilon \\ &\quad\iff e^{\beta_1+\beta_2(x_0-\delta)} = \varepsilon\left(1 + e^{\beta_1+\beta_2(x_0-\delta)}\right) \\ &\quad\iff (1-\varepsilon)e^{\beta_1+\beta_2(x_0-\delta)} = \varepsilon \\ &\quad\iff e^{\beta_1+\beta_2(x_0-\delta)} = \frac{\varepsilon}{1-\varepsilon} \\ &\quad\iff \beta_1 + \beta_2(x_0 - \delta) = \log\left(\frac{\varepsilon}{1-\varepsilon}\right) = -\zeta_0. \end{aligned} \tag{A.5.3}$$

De façon similaire, à partir de l'équation (A.5.2), nous obtenons

$$\begin{aligned} &\mathbb{E}(Y_i|x_0 + \delta, \beta_1, \beta_2) = 1 - \varepsilon \\ &\quad\iff e^{\beta_1+\beta_2(x_0+\delta)} = (1-\varepsilon)\left(1 + e^{\beta_1+\beta_2(x_0+\delta)}\right) \\ &\quad\iff \varepsilon e^{\beta_1+\beta_2(x_0+\delta)} = (1-\varepsilon) \\ &\quad\iff e^{\beta_1+\beta_2(x_0+\delta)} = \frac{(1-\varepsilon)}{\varepsilon} \\ &\quad\iff \beta_1 + \beta_2(x_0 + \delta) = \log\left(\frac{(1-\varepsilon)}{\varepsilon}\right) = \zeta_0 \end{aligned} \tag{A.5.4}$$

Ainsi, en additionnant les équations (A.5.3) et (A.5.4), nous obtenons que

$$2 \times (\beta_1 + \beta_2 x_0) = 0$$

et donc $\beta_1 = -\beta_2 x_0$. Maintenant, en choisissant β_2 assez grand, nous aurons une séparation parfaite en utilisant le modèle

$$\mathbb{E}(Y_i|x_i, \beta_2) = \frac{e^{\beta_2(x_i - x_0)}}{1 + e^{\beta_2(x_i - x_0)}}. \tag{A.5.5}$$

A.6. TABLEAUX DES RÉSULTATS

TABLEAU A.2. EQM calculé sur tous les échantillons pour le coefficient β_C.

	Densité	Moyenne	Variance	EQM
1	Student	0,00	3,125	0,34
2	Gumbel	2,94	3,125	0,40
3	normale	0,00	3,125	0,47
4	Laplace	0,00	3,125	0,52
5	Gelman	0,00	2,5	0,52
6	Student	0,00	6,25	0,55
7	normale	2,94	3,125	0,63
8	Gumbel	0,00	3,125	0,64
9	normale	0,00	6,25	0,78
10	normale	2,94	6,25	0,81
11	Student	0,00	12,5	0,90
12	Gumbel	2,94	6,25	1,17
13	Laplace	2,94	3,125	1,18
14	normale	-2,94	3,125	1,23
15	normale	2,94	12,5	1,23
16	normale	0,00	12,5	1,29
17	normale	-2,94	6,25	1,32
18	Laplace	0,00	6,25	1,42
19	Laplace	-2,94	3,125	1,44
20	Student	-2,94	6,25	1,50
21	Gumbel	-2,94	3,125	1,56
22	Student	-2,94	3,125	1,57
23	Student	-2,94	12,5	1,64
24	Student	2,94	6,25	1,72
25	normale	-2,94	12,5	1,75
26	Gumbel	0,00	6,25	1,75
27	Student	2,94	12,5	1,80
28	Student	2,94	3,125	1,87
29	Laplace	2,94	6,25	2,04
30	Laplace	-2,94	6,25	2,15
31	Gumbel	-2,94	6,25	2,75
32	Genkin - Laplace	0,00	?	3,09
33	Gumbel	2,94	12,5	4,19
34	Laplace	0,00	12,5	4,23
35	Laplace	2,94	12,5	4,87
36	Laplace	-2,94	12,5	4,89
37	Gumbel	0,00	12,5	5,28
38	Genkin - normale	0,00	?	5,76
39	Gumbel	-2,94	12,5	6,76
40	EMV	-	-	135,61

TABLEAU A.3. EQM calculé sur tous les échantillons pour le coefficient β_T.

	Densité	Moyenne	Variance	EQM
1	normale	0,00	3,125	0,59
2	Gelman	0,00	2,5	0,62
3	Laplace	0,00	3,125	0,72
4	normale	0,00	6,25	0,74
5	normale	2,94	3,125	0,74
6	Student	0,00	3,125	0,75
7	Student	0,00	6,25	0,79
8	Gumbel	2,94	3,125	0,82
9	normale	-2,94	6,25	0,96
10	Student	2,94	3,125	0,97
11	Gumbel	0,00	3,125	0,99
12	Laplace	2,94	3,125	1,11
13	normale	-2,94	12,5	1,15
14	Student	0,00	12,5	1,15
15	Student	2,94	6,25	1,16
16	normale	2,94	6,25	1,20
17	Laplace	-2,94	3,125	1,34
18	normale	0,00	12,5	1,35
19	normale	-2,94	3,125	1,38
20	Genkin - Laplace	0,00	selon β_j	1,47
21	Gumbel	-2,94	3,125	1,54
22	Genkin - normale	0,00	selon β_j	1,64
23	Student	2,94	12,5	1,65
24	Laplace	0,00	6,25	1,97
25	normale	2,94	12,5	2,11
26	Student	-2,94	6,25	2,66
27	Laplace	-2,94	6,25	2,75
28	Laplace	2,94	6,25	2,77
29	Student	-2,94	12,5	2,87
30	Student	-2,94	3,125	2,99
31	Gumbel	2,94	6,25	3,45
32	Gumbel	0,00	6,25	3,75
33	Gumbel	-2,94	6,25	4,47
34	Laplace	0,00	12,5	8,21
35	Laplace	-2,94	12,5	9,21
36	Laplace	2,94	12,5	9,31
37	Gumbel	0,00	12,5	15,98
38	Gumbel	-2,94	12,5	17,20
39	Gumbel	2,94	12,5	20,18
40	EMV	-	-	366,41

TABLEAU A.4. EQM calculé sur tous les échantillons pour le coefficient β_D.

	Densité	Moyenne	Variance	EQM
1	Gelman	0,00	2,5	0,23
2	Student	0,00	3,125	0,27
3	normale	0,00	3,125	0,32
4	Gumbel	2,94	3,125	0,37
5	Laplace	0,00	3,125	0,37
6	Student	0,00	6,25	0,38
7	normale	2,94	3,125	0,43
8	normale	-2,94	3,125	0,44
9	Gumbel	0,00	3,125	0,44
10	normale	0,00	6,25	0,53
11	normale	2,94	6,25	0,62
12	normale	-2,94	6,25	0,64
13	Student	0,00	12,5	0,64
14	Laplace	-2,94	3,125	0,80
15	Laplace	2,94	3,125	0,91
16	normale	0,00	12,5	0,97
17	Student	-2,94	3,125	0,98
18	Student	-2,94	6,25	0,98
19	Gumbel	-2,94	3,125	1,01
20	Gumbel	2,94	6,25	1,06
21	normale	2,94	12,5	1,06
22	normale	-2,94	12,5	1,07
23	Laplace	0,00	6,25	1,15
24	Student	-2,94	12,5	1,16
25	Student	2,94	6,25	1,36
26	Student	2,94	3,125	1,36
27	Student	2,94	12,5	1,53
28	Gumbel	0,00	6,25	1,53
29	Laplace	-2,94	6,25	1,60
30	Laplace	2,94	6,25	1,80
31	Gumbel	-2,94	6,25	2,37
32	Laplace	0,00	12,5	4,15
33	Laplace	-2,94	12,5	4,57
34	Genkin - normale	0,00	selon β_j	4,66
35	Laplace	2,94	12,5	4,89
36	Genkin - Laplace	0,00	selon β_j	4,92
37	Gumbel	2,94	12,5	5,18
38	Gumbel	0,00	12,5	5,67
39	Gumbel	-2,94	12,5	7,10
40	EMV	-	-	124,27

TABLEAU A.5. EQM calculé sur les échantillons de taille N = 10.

	Densité	Moyenne	Variance	EQM global	EQM β_C	EQM β_T	EQM β_D
1	Gumbel	2,94	3,125	0.69	0.77	0.71	0.59
2	Student	0,00	3,125	0.75	0.28	1.46	0.50
3	Gelman	0,00	2,5	0.80	0.65	1.28	0.46
4	normale	0,00	3,125	0.81	0.60	1.24	0.60
5	Laplace	0,00	3,125	0.86	0.54	1.28	0.77
6	Student	0,00	6,25	0.86	0.53	1.27	0.79
7	normale	2,94	3,125	1.07	1.51	0.98	0.72
8	Gumbel	0,00	3,125	1.13	0.96	1.54	0.88
9	normale	0,00	6,25	1.23	1.13	1.37	1.17
10	Student	0,00	12,5	1.29	0.98	1.40	1.49
11	Genkin - Laplace	0,00	selon β_j	1.47	0.11	1.53	2.78
12	normale	2,94	6,25	1.52	1.76	1.55	1.25
13	normale	-2,94	6,25	1.81	1.58	2.29	1.57
14	normale	-2,94	3,125	2.05	1.75	3.39	1.00
15	Laplace	2,94	3,125	2.15	2.60	1.77	2.09
16	normale	0,00	12,5	2.21	2.00	2.20	2.43
17	Genkin - normale	0,00	selon β_j	2.27	2.85	0.76	3.21
18	Laplace	0,00	6,25	2.30	1.68	2.27	2.95
19	Gumbel	2,94	6,25	2.36	2.12	2.75	2.21
20	normale	-2,94	12,5	2.38	2.03	2.29	2.81
21	Laplace	-2,94	3,125	2.39	2.25	3.01	1.90
22	normale	2,94	12,5	2.59	2.54	2.75	2.49
23	Student	-2,94	6,25	2.63	2.34	3.40	2.16
24	Gumbel	-2,94	3,125	2.66	2.44	3.05	2.48
25	Student	2,94	6,25	2.66	2.74	1.81	3.43
26	Student	-2,94	12,5	2.69	2.21	3.16	2.71
27	Student	2,94	3,125	2.72	3.16	1.70	3.30
28	Student	2,94	12,5	2.98	2.71	2.25	3.99
29	Student	-2,94	3,125	3.00	2.81	4.12	2.08
30	Laplace	-2,94	6,25	3.59	2.86	3.79	4.11
31	Gumbel	0,00	6,25	3.60	2.96	4.13	3.71
32	Laplace	2,94	6,25	3.91	3.42	3.61	4.69
33	Gumbel	-2,94	6,25	5.47	4.17	6.07	6.18
34	Laplace	0,00	12,5	8.20	4.94	8.20	11.46
35	Laplace	-2,94	12,5	9.27	5.81	9.53	12.45
36	Laplace	2,94	12,5	9.96	6.46	10.01	13.42
37	Gumbel	2,94	12,5	10.46	7.19	12.84	11.35
38	Gumbel	0,00	12,5	12.89	8.44	15.66	14.57
39	Gumbel	-2,94	12,5	15.95	10.02	19.04	18.77
40	EMV	-	-	363.96	223.78	544.39	323.71

TABLEAU A.6. EQM calculé sur les échantillons de taille N = 50.

	Densité	Moyenne	Variance	EQM global	EQM β_C	EQM β_T	EQM β_D
1	Gumbel	2,94	3,125	0.27	0.22	0.47	0.12
2	Gelman	0,00	2,5	0.28	0.27	0.44	0.12
3	normale	0,00	3,125	0.29	0.28	0.42	0.15
4	Laplace	0,00	3,125	0.29	0.22	0.52	0.14
5	Student	0,00	3,125	0.30	0.20	0.57	0.13
6	Student	0,00	6,25	0.31	0.25	0.54	0.15
7	normale	2,94	3,125	0.32	0.30	0.49	0.15
8	normale	0,00	6,25	0.33	0.32	0.48	0.17
9	normale	-2,94	6,25	0.35	0.37	0.50	0.17
10	Student	0,00	12,5	0.35	0.30	0.58	0.17
11	Student	2,94	3,125	0.36	0.43	0.46	0.20
12	Gumbel	0,00	3,125	0.37	0.33	0.60	0.18
13	Laplace	2,94	3,125	0.37	0.38	0.55	0.20
14	normale	-2,94	12,5	0.38	0.36	0.57	0.20
15	Student	2,94	6,25	0.38	0.38	0.55	0.20
16	normale	2,94	6,25	0.40	0.33	0.68	0.19
17	Laplace	0,00	6,25	0.40	0.29	0.73	0.19
18	normale	0,00	12,5	0.41	0.35	0.67	0.20
19	Student	2,94	12,5	0.41	0.36	0.67	0.20
20	Laplace	-2,94	3,125	0.43	0.43	0.65	0.22
21	Gumbel	-2,94	3,125	0.44	0.40	0.69	0.22
22	normale	-2,94	3,125	0.45	0.42	0.77	0.17
23	Student	-2,94	12,5	0.47	0.40	0.79	0.22
24	Laplace	2,94	6,25	0.48	0.38	0.82	0.23
25	normale	2,94	12,5	0.49	0.36	0.88	0.22
26	Student	-2,94	6,25	0.49	0.44	0.80	0.23
27	Laplace	-2,94	6,25	0.50	0.40	0.86	0.24
28	Gumbel	2,94	6,25	0.50	0.36	0.92	0.22
29	Gumbel	0,00	6,25	0.52	0.38	0.94	0.24
30	Student	-2,94	3,125	0.54	0.51	0.85	0.26
31	Gumbel	-2,94	6,25	0.54	0.39	0.97	0.25
32	Laplace	0,00	12,5	0.54	0.34	1.04	0.23
33	Laplace	2,94	12,5	0.59	0.39	1.12	0.26
34	Laplace	-2,94	12,5	0.60	0.40	1.13	0.27
35	Gumbel	-2,94	12,5	0.66	0.40	1.29	0.29
36	Gumbel	2,94	12,5	0.66	0.40	1.30	0.28
37	Gumbel	0,00	12,5	0.67	0.40	1.33	0.29
38	EMV	-	-	2.08	0.34	5.08	0.83
39	Genkin - Laplace	0,00	selon β_j	2.57	4.06	0.54	3.12
40	Genkin - normale	0,00	selon β_j	3.64	7.11	0.82	2.99

TABLEAU A.7. EQM calculé sur les échantillons de taille N = 100.

	Densité	Moyenne	Variance	EQM global	EQM β_C	EQM β_T	EQM β_D
1	normale	0,00	3,125	0.17	0.15	0.28	0.08
2	Gelman	0,00	2,5	0.18	0.15	0.31	0.07
3	Student	2,94	3,125	0.19	0.19	0.29	0.09
4	normale	-2,94	6,25	0.19	0.17	0.32	0.08
5	Student	0,00	3,125	0.20	0.13	0.40	0.07
6	Laplace	0,00	3,125	0.20	0.13	0.40	0.07
7	Student	0,00	6,25	0.21	0.14	0.41	0.08
8	normale	0,00	6,25	0.21	0.17	0.39	0.09
9	Laplace	2,94	3,125	0.22	0.18	0.39	0.09
10	Gumbel	2,94	3,125	0.22	0.14	0.45	0.07
11	normale	-2,94	3,125	0.22	0.19	0.39	0.08
12	normale	2,94	3,125	0.22	0.16	0.43	0.08
13	Student	2,94	6,25	0.23	0.18	0.41	0.09
14	Student	0,00	12,5	0.23	0.16	0.46	0.09
15	normale	-2,94	12,5	0.24	0.17	0.44	0.09
16	Gumbel	0,00	3,125	0.25	0.17	0.48	0.09
17	Laplace	-2,94	3,125	0.25	0.19	0.46	0.10
18	Gumbel	-2,94	3,125	0.26	0.18	0.50	0.10
19	Student	2,94	12,5	0.27	0.18	0.53	0.10
20	Laplace	0,00	6,25	0.27	0.15	0.57	0.09
21	EMV	-	-	0.27	0.17	0.55	0.10
22	normale	0,00	12,5	0.27	0.17	0.55	0.10
23	Student	-2,94	12,5	0.28	0.18	0.55	0.10
24	Student	-2,94	6,25	0.28	0.19	0.54	0.10
25	normale	2,94	6,25	0.28	0.17	0.58	0.09
26	Laplace	2,94	6,25	0.29	0.18	0.58	0.10
27	Student	-2,94	3,125	0.29	0.21	0.56	0.11
28	Laplace	-2,94	6,25	0.30	0.18	0.62	0.10
29	Gumbel	-2,94	6,25	0.32	0.18	0.66	0.10
30	Gumbel	0,00	6,25	0.32	0.18	0.68	0.10
31	Laplace	0,00	12,5	0.33	0.17	0.71	0.10
32	Gumbel	2,94	6,25	0.33	0.18	0.71	0.10
33	normale	2,94	12,5	0.33	0.18	0.71	0.10
34	Laplace	2,94	12,5	0.34	0.18	0.73	0.10
35	Laplace	-2,94	12,5	0.35	0.18	0.74	0.11
36	Gumbel	-2,94	12,5	0.36	0.18	0.80	0.11
37	Gumbel	0,00	12,5	0.37	0.18	0.81	0.11
38	Gumbel	2,94	12,5	0.37	0.18	0.83	0.11
39	Genkin - Laplace	0,00	selon β_j	3.23	6.30	0.00	3.38
40	Genkin - normale	0,00	selon β_j	3.88	8.30	0.01	3.32

TABLEAU A.8. EQM calculé sur les échantillons générés avec $\beta_T = 0{,}85$.

	Densité	Moyenne	Variance	EQM global	EQM β_C	EQM β_T	EQM β_D
1	Student	0,00	3,125	0.31	0.31	0.39	0.24
2	normale	0,00	3,125	0.40	0.46	0.46	0.29
3	Gelman	0,00	2,5	0.42	0.51	0.52	0.21
4	Gumbel	2,94	3,125	0.44	0.43	0.63	0.27
5	Laplace	0,00	3,125	0.47	0.49	0.55	0.37
6	Student	0,00	6,25	0.49	0.52	0.56	0.39
7	normale	2,94	3,125	0.61	0.69	0.79	0.36
8	Gumbel	0,00	3,125	0.63	0.64	0.83	0.41
9	normale	0,00	6,25	0.69	0.76	0.75	0.55
10	Student	0,00	12,5	0.82	0.87	0.90	0.70
11	normale	2,94	6,25	0.85	0.86	1.11	0.58
12	normale	-2,94	3,125	0.87	1.13	0.99	0.48
13	normale	-2,94	6,25	0.97	1.23	0.93	0.74
14	Laplace	2,94	3,125	1.16	1.23	1.26	0.99
15	Laplace	-2,94	3,125	1.18	1.39	1.31	0.84
16	normale	0,00	12,5	1.22	1.28	1.31	1.06
17	Gumbel	-2,94	3,125	1.35	1.52	1.47	1.06
18	normale	2,94	12,5	1.35	1.28	1.72	1.06
19	Laplace	0,00	6,25	1.39	1.37	1.50	1.31
20	normale	-2,94	12,5	1.40	1.66	1.27	1.28
21	Student	-2,94	6,25	1.42	1.47	1.83	0.95
22	Gumbel	2,94	6,25	1.42	1.20	2.10	0.95
23	Student	-2,94	3,125	1.49	1.55	2.00	0.90
24	Student	2,94	6,25	1.55	1.59	1.35	1.71
25	Student	-2,94	12,5	1.57	1.60	1.94	1.18
26	Student	2,94	3,125	1.59	1.70	1.34	1.73
27	Student	2,94	12,5	1.74	1.70	1.60	1.91
28	Gumbel	0,00	6,25	1.93	1.76	2.50	1.51
29	Laplace	-2,94	6,25	2.02	2.10	2.19	1.76
30	Laplace	2,94	6,25	2.11	2.01	2.26	2.07
31	Gumbel	-2,94	6,25	2.81	2.74	3.21	2.49
32	Genkin - Laplace	0,00	selon β_j	3.24	3.08	1.82	4.82
33	Genkin - normale	0,00	selon β_j	4.43	5.75	3.06	4.47
34	Laplace	0,00	12,5	4.63	4.13	4.90	4.86
35	Laplace	-2,94	12,5	5.20	4.80	5.58	5.23
36	Laplace	2,94	12,5	5.38	4.72	5.70	5.72
37	Gumbel	2,94	12,5	5.42	4.23	7.61	4.43
38	Gumbel	0,00	12,5	6.46	5.29	8.42	5.68
39	Gumbel	-2,94	12,5	7.88	6.75	9.52	7.36
40	EMV	-	-	150.75	131.97	201.44	118.84

TABLEAU A.9. EQM calculé sur les échantillons générés avec $\beta_T = 1{,}39$.

	Densité	Moyenne	Variance	EQM global	EQM β_C	EQM β_T	EQM β_D
1	Student	0,00	3,125	0.36	0.33	0.52	0.24
2	normale	0,00	3,125	0.41	0.45	0.48	0.30
3	Gelman	0,00	2,5	0.42	0.51	0.54	0.23
4	Laplace	0,00	3,125	0.49	0.50	0.62	0.36
5	Gumbel	2,94	3,125	0.49	0.39	0.75	0.34
6	Student	0,00	6,25	0.50	0.52	0.63	0.36
7	normale	2,94	3,125	0.61	0.64	0.79	0.40
8	Gumbel	0,00	3,125	0.66	0.62	0.94	0.43
9	normale	0,00	6,25	0.67	0.75	0.74	0.52
10	Student	0,00	12,5	0.83	0.86	0.99	0.63
11	normale	2,94	6,25	0.87	0.80	1.21	0.60
12	normale	-2,94	3,125	0.88	1.17	1.04	0.42
13	normale	-2,94	6,25	0.92	1.26	0.85	0.64
14	Laplace	2,94	3,125	1.09	1.17	1.20	0.90
15	Laplace	-2,94	3,125	1.17	1.40	1.30	0.80
16	normale	0,00	12,5	1.19	1.25	1.35	0.98
17	normale	-2,94	12,5	1.30	1.67	1.14	1.09
18	Gumbel	-2,94	3,125	1.34	1.52	1.49	1.01
19	Student	2,94	6,25	1.38	1.58	1.21	1.35
20	Student	2,94	3,125	1.38	1.74	1.07	1.34
21	Laplace	0,00	6,25	1.41	1.36	1.73	1.15
22	normale	2,94	12,5	1.42	1.21	2.00	1.05
23	Student	-2,94	6,25	1.51	1.47	2.12	0.95
24	Student	-2,94	3,125	1.59	1.54	2.29	0.94
25	Student	2,94	12,5	1.61	1.69	1.60	1.52
26	Gumbel	2,94	6,25	1.65	1.13	2.79	1.03
27	Student	-2,94	12,5	1.66	1.60	2.27	1.13
28	Laplace	-2,94	6,25	2.04	2.09	2.43	1.60
29	Laplace	2,94	6,25	2.09	1.94	2.53	1.80
30	Gumbel	0,00	6,25	2.11	1.70	3.13	1.51
31	Gumbel	-2,94	6,25	2.93	2.65	3.80	2.35
32	Genkin - Laplace	0,00	selon β_j	3.09	3.10	1.20	4.98
33	Genkin - normale	0,00	selon β_j	4.04	5.77	1.68	4.68
34	Laplace	0,00	12,5	4.85	4.06	6.31	4.17
35	Laplace	-2,94	12,5	5.46	4.74	7.07	4.58
36	Laplace	2,94	12,5	5.59	4.61	7.27	4.90
37	Gumbel	2,94	12,5	6.49	4.00	10.99	4.48
38	Gumbel	0,00	12,5	7.45	5.03	11.76	5.55
39	Gumbel	-2,94	12,5	8.77	6.43	12.89	6.97
40	EMV	-	-	165.85	128.57	252.03	116.97

TABLEAU A.10. EQM calculé sur les échantillons générés avec $\beta_T = 2{,}94$.

	Densité	Moyenne	Variance	EQM global	EQM β_C	EQM β_T	EQM β_D
1	Gelman	0,00	2,5	0.54	0.55	0.80	0.26
2	normale	0,00	3,125	0.56	0.51	0.82	0.36
3	normale	2,94	3,125	0.58	0.56	0.63	0.54
4	Laplace	0,00	3,125	0.65	0.57	0.98	0.39
5	Gumbel	2,94	3,125	0.65	0.37	1.08	0.50
6	Student	0,00	3,125	0.68	0.38	1.34	0.33
7	normale	0,00	6,25	0.69	0.82	0.73	0.52
8	Student	0,00	6,25	0.73	0.61	1.18	0.40
9	Gumbel	0,00	3,125	0.78	0.65	1.21	0.49
10	normale	2,94	6,25	0.90	0.76	1.27	0.68
11	Laplace	2,94	3,125	0.95	1.15	0.86	0.84
12	normale	-2,94	6,25	1.04	1.48	1.12	0.54
13	Student	0,00	12,5	1.04	0.97	1.57	0.59
14	normale	0,00	12,5	1.20	1.34	1.38	0.88
15	Student	2,94	3,125	1.23	2.19	0.49	1.02
16	Laplace	-2,94	3,125	1.24	1.53	1.43	0.76
17	normale	-2,94	12,5	1.26	1.90	1.03	0.85
18	normale	-2,94	3,125	1.30	1.38	2.10	0.41
19	Student	2,94	6,25	1.30	1.98	0.91	1.01
20	Gumbel	-2,94	3,125	1.42	1.66	1.66	0.95
21	normale	2,94	12,5	1.63	1.21	2.62	1.06
22	Student	2,94	12,5	1.64	2.01	1.74	1.15
23	Laplace	0,00	6,25	1.73	1.51	2.68	0.99
24	Student	-2,94	6,25	2.22	1.56	4.04	1.06
25	Laplace	2,94	6,25	2.40	2.15	3.51	1.52
26	Student	-2,94	12,5	2.42	1.71	4.38	1.16
27	Laplace	-2,94	6,25	2.44	2.25	3.62	1.44
28	Student	-2,94	3,125	2.47	1.62	4.69	1.10
29	Gumbel	2,94	6,25	2.60	1.16	5.45	1.19
30	Gumbel	0,00	6,25	2.98	1.79	5.60	1.56
31	Genkin - Laplace	0,00	selon β_j	3.15	3.09	1.39	4.97
32	Genkin - normale	0,00	selon β_j	3.59	5.75	0.20	4.83
33	Gumbel	-2,94	6,25	3.84	2.87	6.39	2.26
34	Laplace	0,00	12,5	7.10	4.49	13.39	3.42
35	Laplace	-2,94	12,5	8.00	5.13	14.97	3.90
36	Laplace	2,94	12,5	8.08	5.27	14.94	4.04
37	Gumbel	2,94	12,5	12.13	4.34	27.14	4.92
38	Gumbel	0,00	12,5	13.02	5.50	27.77	5.79
39	Gumbel	-2,94	12,5	14.42	7.09	29.21	6.97
40	EMV	-	-	309.70	146.29	645.81	137.02

TABLEAU A.11. EQM calculé sur les échantillons générés avec $\beta_D = -0{,}04$.

	Densité	Moyenne	Variance	EQM global	EQM β_C	EQM β_T	EQM β_D
1	Gelman	0,00	2,5	0.43	0.50	0.59	0.21
2	Student	0,00	3,125	0.44	0.32	0.76	0.25
3	normale	0,00	3,125	0.46	0.46	0.58	0.35
4	Laplace	0,00	3,125	0.53	0.48	0.74	0.38
5	Student	0,00	6,25	0.60	0.51	0.88	0.40
6	normale	0,00	6,25	0.70	0.76	0.76	0.58
7	normale	2,94	3,125	0.76	0.71	0.92	0.64
8	Gumbel	2,94	3,125	0.76	0.53	1.18	0.57
9	Gumbel	0,00	3,125	0.81	0.67	1.14	0.62
10	normale	-2,94	6,25	0.87	1.12	0.94	0.55
11	normale	-2,94	3,125	0.91	1.01	1.34	0.36
12	Student	0,00	12,5	0.95	0.84	1.36	0.66
13	normale	2,94	6,25	1.05	0.87	1.44	0.84
14	Laplace	2,94	3,125	1.11	1.19	1.21	0.94
15	normale	-2,94	12,5	1.16	1.50	1.11	0.89
16	Laplace	-2,94	3,125	1.21	1.35	1.39	0.89
17	normale	0,00	12,5	1.22	1.25	1.42	0.97
18	Student	2,94	3,125	1.24	1.76	0.97	0.98
19	Student	2,94	6,25	1.28	1.63	1.23	0.99
20	Gumbel	-2,94	3,125	1.38	1.49	1.57	1.07
21	Laplace	0,00	6,25	1.55	1.30	2.28	1.06
22	Student	2,94	12,5	1.56	1.71	1.83	1.15
23	normale	2,94	12,5	1.68	1.29	2.50	1.26
24	Student	-2,94	6,25	1.96	1.47	3.07	1.32
25	Student	-2,94	12,5	2.08	1.57	3.27	1.40
26	Laplace	-2,94	6,25	2.18	1.99	2.96	1.57
27	Laplace	2,94	6,25	2.23	1.98	3.11	1.58
28	Student	-2,94	3,125	2.25	1.58	3.76	1.40
29	Genkin - Laplace	0,00	selon β_j	2.42	3.82	1.51	1.92
30	Gumbel	2,94	6,25	2.51	1.35	4.62	1.56
31	Gumbel	0,00	6,25	2.68	1.82	4.42	1.80
32	Gumbel	-2,94	6,25	3.20	2.62	4.68	2.29
33	Genkin - normale	0,00	selon β_j	3.37	6.39	1.78	1.93
34	Laplace	0,00	12,5	5.52	3.76	9.49	3.30
35	Laplace	-2,94	12,5	6.19	4.38	10.40	3.79
36	Laplace	2,94	12,5	6.28	4.45	10.63	3.77
37	Gumbel	2,94	12,5	9.61	4.49	19.04	5.29
38	Gumbel	0,00	12,5	9.87	5.27	18.62	5.72
39	Gumbel	-2,94	12,5	10.49	6.34	18.72	6.41
40	EMV	-	-	172.81	139.91	300.80	77.71

TABLEAU A.12. EQM calculé sur les échantillons générés avec $\beta_D = -0{,}62$.

	Densité	Moyenne	Variance	EQM global	EQM β_C	EQM β_T	EQM β_D
1	Student	0,00	3,125	0.42	0.34	0.70	0.23
2	Gelman	0,00	2,5	0.44	0.52	0.59	0.21
3	normale	0,00	3,125	0.44	0.47	0.56	0.30
4	Gumbel	2,94	3,125	0.47	0.37	0.78	0.26
5	Laplace	0,00	3,125	0.51	0.52	0.69	0.32
6	Student	0,00	6,25	0.54	0.54	0.76	0.33
7	normale	2,94	3,125	0.58	0.61	0.76	0.37
8	normale	0,00	6,25	0.67	0.77	0.74	0.49
9	Gumbel	0,00	3,125	0.68	0.62	1.01	0.40
10	normale	2,94	6,25	0.86	0.78	1.24	0.55
11	Student	0,00	12,5	0.87	0.89	1.18	0.54
12	normale	-2,94	6,25	0.93	1.30	0.91	0.59
13	normale	-2,94	3,125	0.97	1.22	1.28	0.42
14	Laplace	2,94	3,125	1.04	1.16	1.16	0.80
15	Laplace	-2,94	3,125	1.16	1.40	1.33	0.74
16	normale	0,00	12,5	1.17	1.26	1.40	0.86
17	normale	-2,94	12,5	1.26	1.70	1.15	0.93
18	Student	2,94	6,25	1.33	1.73	1.21	1.05
19	Student	2,94	3,125	1.34	1.92	1.01	1.08
20	Gumbel	-2,94	3,125	1.34	1.52	1.58	0.92
21	normale	2,94	12,5	1.44	1.20	2.19	0.92
22	Laplace	0,00	6,25	1.44	1.38	2.03	0.92
23	Student	2,94	12,5	1.56	1.80	1.72	1.17
24	Student	-2,94	6,25	1.72	1.45	2.88	0.84
25	Student	-2,94	3,125	1.80	1.51	3.04	0.85
26	Gumbel	2,94	6,25	1.85	1.12	3.58	0.86
27	Student	-2,94	12,5	1.90	1.58	3.15	0.97
28	Laplace	-2,94	6,25	2.11	2.08	2.92	1.35
29	Laplace	2,94	6,25	2.11	2.00	2.89	1.46
30	Gumbel	0,00	6,25	2.33	1.69	4.00	1.31
31	Genkin - Laplace	0,00	selon β_j	2.94	3.09	1.46	4.28
32	Gumbel	-2,94	6,25	3.15	2.64	4.80	2.01
33	Genkin - normale	0,00	selon β_j	3.86	5.76	1.66	4.16
34	Laplace	0,00	12,5	5.21	4.05	8.55	3.03
35	Laplace	-2,94	12,5	5.94	4.64	9.75	3.43
36	Laplace	2,94	12,5	6.01	4.70	9.73	3.60
37	Gumbel	2,94	12,5	7.88	3.96	16.09	3.60
38	Gumbel	0,00	12,5	8.83	5.00	17.02	4.47
39	Gumbel	-2,94	12,5	10.12	6.36	18.41	5.59
40	EMV	-	-	236.31	149.71	448.89	110.33

TABLEAU A.13. EQM calculé sur les échantillons générés avec $\beta_D = -1{,}39$.

	Densité	Moyenne	Variance	EQM global	EQM β_C	EQM β_T	EQM β_D
1	Gumbel	2,94	3,125	0.36	0.29	0.50	0.28
2	normale	0,00	3,125	0.47	0.49	0.62	0.29
3	normale	2,94	3,125	0.47	0.58	0.54	0.29
4	Student	0,00	3,125	0.50	0.37	0.80	0.33
5	Gelman	0,00	2,5	0.51	0.56	0.68	0.28
6	Laplace	0,00	3,125	0.57	0.57	0.72	0.41
7	Student	0,00	6,25	0.58	0.59	0.72	0.42
8	Gumbel	0,00	3,125	0.59	0.62	0.83	0.31
9	normale	0,00	6,25	0.68	0.80	0.71	0.51
10	normale	2,94	6,25	0.71	0.76	0.90	0.48
11	Student	0,00	12,5	0.87	0.97	0.92	0.73
12	Laplace	2,94	3,125	1.05	1.21	0.96	0.99
13	normale	-2,94	6,25	1.13	1.55	1.04	0.79
14	normale	-2,94	3,125	1.17	1.46	1.51	0.53
15	Laplace	-2,94	3,125	1.22	1.57	1.31	0.78
16	normale	0,00	12,5	1.22	1.36	1.22	1.08
17	normale	2,94	12,5	1.29	1.21	1.65	1.00
18	Gumbel	2,94	6,25	1.31	1.03	2.14	0.74
19	Gumbel	-2,94	3,125	1.40	1.68	1.47	1.03
20	Student	-2,94	6,25	1.47	1.58	2.03	0.79
21	Student	-2,94	3,125	1.49	1.62	2.17	0.69
22	normale	-2,94	12,5	1.54	2.03	1.19	1.40
23	Laplace	0,00	6,25	1.54	1.57	1.60	1.46
24	Student	2,94	6,25	1.62	1.79	1.04	2.03
25	Student	2,94	3,125	1.64	1.94	0.92	2.04
26	Student	-2,94	12,5	1.68	1.76	2.17	1.10
27	Student	2,94	12,5	1.85	1.89	1.39	2.27
28	Gumbel	0,00	6,25	2.01	1.74	2.82	1.48
29	Laplace	-2,94	6,25	2.20	2.37	2.35	1.88
30	Laplace	2,94	6,25	2.26	2.13	2.30	2.35
31	Gumbel	-2,94	6,25	3.24	2.99	3.92	2.80
32	Genkin - Laplace	0,00	selon β_j	4.12	2.35	1.45	8.56
33	Genkin - normale	0,00	selon β_j	4.83	5.12	1.50	7.88
34	Laplace	0,00	12,5	5.85	4.86	6.55	6.12
35	Laplace	-2,94	12,5	6.53	5.64	7.46	6.49
36	Gumbel	2,94	12,5	6.56	4.12	10.62	4.95
37	Laplace	2,94	12,5	6.76	5.45	7.55	7.28
38	Gumbel	0,00	12,5	8.23	5.55	12.31	6.82
39	Gumbel	-2,94	12,5	10.45	7.57	14.48	9.31
40	EMV	-	-	217.19	117.20	349.59	184.79