UNIVERSITÉ DE GENÈVE

FACULTÉ DES SCIENCES ÉCONOMIQUES ET SOCIALES

Mémoire de Master

Prévisions pour les modèles linéaires généralisés robustes :

étude Monte-Carlo d'un nouvel estimateur

Etudiant :
Jaromil Frossard

Directrice de mémoire :
Professeur Eva Cantoni

27 novembre 2012

Table des matières

Introduction

La démarche scientifique donne une importance particulière aux modèles. Ils ont pour but de simplifier la réalité afin d'en faciliter la compréhension. On va supposer que plusieurs phénomènes, variables, sont la cause d'un autre phénomène. Le modèle va spécifier la forme du lien entre les différentes variables. De ce lien de causalité, il est possible d'expliquer la réalité mais aussi de proposer des prévisions. Et c'est un élément essentiel d'un modèle car cela permet de valider le lien de causalité supposé. Ainsi, un scientifique, après avoir élaboré un modèle, désire tester son pouvoir prédictif. Pour cela, il est nécessaire de récolter des données en faisant des observations ou une expérience et de comparer les prévisions du modèle aux données.

Ce travail se centre sur les modèles linéaires généralisés. Ils ont une structure commune avec des paramètres qui nécessitent d'être estimés. On les dénomme "linéaires" car le lien entre les paramètres (et non la variable réponse) est linéaire. Et ils sont généralisés car ils sont adaptés à la modélisation d'un large nombre de variables ; celles issues des distributions de la famille exponentielle. Ils ont l'avantage de pouvoir modéliser des variables discrètes comme les comptages ou les variables binaires.

Les méthodes statistiques utilisées dans ce travail concernent l'estimation des paramètres. Il s'agit de comparer différentes approches d'estimation. Tout d'abord, on s'intéresse à la méthode d'estimation du maximum de vraisemblance. Elle est théoriquement bien développée et elle comporte des propriétés avantageuses ; notamment un comportement asymptotique des estimateurs bien défini. Cependant elle est sensible à la définition du modèle. Si l'on suppose un mauvais modèle, l'estimation est biaisée. Et le statisticien ne commet pas forcément une erreur en posant de mauvaises hypothèses sur le modèle. Un défaut d'informations sur le phénomène à expliquer ou alors des outils mathématiques inexistants peuvent mener à des hypothèses commodes mais fausses. Les estimations robustes remédient à ce problème. Elles permettent de minimiser ces erreurs en réduisant la pondération des données déviant fortement du modèle postulé. Dans ce travail, un nouvel estimateur robuste est développé. Genton et Ronchetti (2008) ont développé pour le modèle linéaire un estimateur aux propriétés robustes qui a montré

de bonnes capacités quand il s'agit de faire des prévisions. Ce nouvel estimateur robuste en est l'adaptation pour le cas généralisé.

On utilise dans ce travail la méthode Monte-Carlo. Elle permet de simuler des expériences. Il s'agit de programmer une expérience et d'enregistrer les données d'intérêts. On évite ainsi les nombreux inconvénients liées à la récolte de données réelles (coût important, durée de l'expérience, organisation), tout en pouvant tester le pouvoir prédictif d'un modèle. De plus, cette méthode permet d'optimiser les conditions de l'expérience. Il faut cependant noter que cette méthode repose sur des conditions idéales et une étude croisée entre les résultats d'une simulation Monte-Carlo et les résultats sur des données réelles permet d'avoir un meilleur jugement de la qualité d'un modèle ou d'une méthode d'estimation.

L'introduction de ce travail est suivie de quatre chapitres. Dans le premier chapitre, une revue de la théorie va permettre de se situer dans le champs des statistiques robustes. On développe autour des ouvrages de Heritier *et al.* (2009) et Hampel *et al.* (1986) la théorie de la statistique robuste. Cette approche générale permet d'expliquer les hypothèses ainsi que les outils majeurs de ces méthodes d'estimation.
Le deuxième chapitre aborde les modèles linéaires généralisés. Les principaux points de la théorie développée dans l'ouvrage de McCullagh et Nelder (1989) sont présentés. Cela concerne notamment une introduction de la famille exponentielle, l'explication de la forme du modèle, l'estimation par maximum de vraisemblance. Ayant une base théorique de statistiques robustes et une présentation des modèles linéaires généralisés, on aborde ensuite l'estimation robuste de ces modèles expliquée notamment par Heritier *et al.* (2009) et Cantoni et Ronchetti (2001). Afin de clore la partie théorique de ce travail, on présente l'estimateur de Genton et Ronchetti (2008) ainsi que son adaptation pour les modèles linéaires généralisés.
Le troisième chapitre présente la simulation Monte-Carlo. Les "conditions d'expérience" sont présentées, les différents choix sont argumentés. Les multiples estimateurs présentés dans le chapitre précédent sont utilisés et comparés. On peut alors vérifier empiriquement les propriétés des estimateurs. Les prévisions produites grâces aux différents estimateurs sont faites selon différentes conditions d'expérience. On peut ainsi déterminer l'influence des estimateurs ainsi que celle des conditions d'expérience sur les prévisions.
Le quatrième chapitre confronte les estimateurs à des données réelles concernant le domaine médicale. Les données proviennent d'un exemple de l'ouvrage Heritier *et al.* (2009) et sont présentées plus en détail dans ce chapitre. On utilise ces données afin d'illustrer les prévisions faites par les différents estimateurs. Il est ainsi possible d'avoir une autre approche des résultats obtenus dans le chapitre précédent.
Les différents résultats importants sont repris dans la conclusion. On y fait

une critique des différents choix, de la méthodologie développée ainsi que des théories statistiques abordées. La conclusion est aussi une mise en perspective du nouvel estimateur, de son rôle pour les prévisions.

Chapitre 1

Inférence robuste

1.1 Formalisation

La formalisation mathématique de la statistique robuste se base sur l'hypothèse que le modèle d'intérêt d'un jeu de données F_θ ne correspond pas à la distribution des observations F_ϵ. Ces dernières sont perturbées par une distribution arbitraire G. Les distributions F_θ et G sont différentes et G génère des données extrêmes. L'hypothèse sous-jacente à l'inférence robuste se formalise ainsi :

$$F_\epsilon = (1-\epsilon)F_\theta + \epsilon G, \tag{1.1}$$

où ϵ correspond à la proportion de données provenant de G (Heritier *et al.*, 2009). La distribution G est qualifiée de contamination du modèle. En posant ces hypothèses, il s'agit en observant F_ϵ de faire de l'inférence sur F_θ. Il est donc important de construire des estimateurs qui ne sont pas influencé par les données provenant de la distribution G.

1.2 La courbe de sensibilité

La courbe de sensibilité est un outil qui permet de juger les propriétés robustes d'un estimateur. Elle montre l'effet de l'observation z sur l'estimateur $\hat{\theta}_n$. Il s'agit donc de comparer les estimateurs $\hat{\theta}_{n-1}(z_1, \ldots, z_{n-1})$ et $\hat{\theta}_n(z_1, \ldots, z_{n-1}, z)$; ce qui correspond au même estimateur évalué sur les échantillons $z_1, \ldots, z_{n-1}$ et $z_1, \ldots, z_{n-1}, z$ respectivement (Heritier *et al.*, 2009; Hampel *et al.*, 1986). Elle s'écrit :

$$SC(z; z_1, \ldots, z_{n-1}, \hat{\theta}_n) = n\left[\hat{\theta}_n(z_1, \ldots, z_{n-1}, z) - \hat{\theta}_{n-1}(z_1, \ldots, z_{n-1})\right]. \tag{1.2}$$

La courbe de sensibilité va mettre en évidence le changement de l'estimé suite à l'observation de z.

Exemple 1. Courbe de sensibilité d'une moyenne

Dans le cas de la moyenne $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} z_i$, la courbe de sensibilité de cet estimateur vaut :

$$\begin{aligned} SC(z; z_1, \ldots, z_{n-1}, \frac{1}{n}\sum_{i=1}^{n} z_i) &= n\left[\frac{1}{n}\left(z + \sum_{i=1}^{n-1} z_i\right) - \frac{1}{n-1}\sum_{i=1}^{n-1} z_i\right] \\ &= z - \frac{1}{n-1}\sum_{i=1}^{n-1} z_i. \end{aligned}$$

La courbe de sensibilité de la moyenne n'est pas bornée. Et l'influence de z sera important si z dévie fortement du reste des données $z_1, \ldots, z_{n-1}$.

Sous l'hypothèse (1.1), l'observation z aurait pu être générée par la distribution G. Un estimateur robuste doit donc avoir la propriété d'avoir une courbe de sensibilité bornée afin que les données provenant de G n'ait qu'un rôle limité dans l'estimation.

1.3 La fonction d'influence

La courbe de sensibilité est un outil puissant afin de mesurer la robustesse d'un estimateur. Elle est cependant limitée. En effet, elle se calcule à travers un échantillonnage $z_1, \ldots, z_n$. Afin d'avoir un outil plus global, la fonction d'influence généralise le concept de la courbe de sensibilité à la population (Heritier *et al.*, 2009; Hampel *et al.*, 1986). Elle se définit par

$$IF(z; \hat{\theta}; F_\theta) = \lim_{\epsilon \to 0} \frac{\hat{\theta}\left((1-\epsilon)F_\theta + \epsilon\Delta_z\right) - \hat{\theta}(F_\theta)}{\epsilon}, \tag{1.3}$$

où $\hat{\theta}$ est un estimateur écrit sous forme de fonctionnelle, F_θ est la distribution du modèle et Δ_z est une distribution ponctuelle en z. Elle peut également se calculer en faisant la dérivée $\frac{\partial}{\partial\epsilon}\hat{\theta}\left((1-\epsilon)F_\theta + \epsilon\Delta_z\right)\Big|_{\epsilon=0}$.
La fonction d'influence peut être interpréter comme l'influence normalisée que produit une contamination infinitésimale en z sur un estimateur ou comme la limite $n \to \infty$ de la courbe de sensibilité (Hampel *et al.*, 1986). Une fonction d'influence bornée, à l'instar de la courbe de sensibilité, signifie que l'influence d'une donnée extrême sur l'estimateur est limitée, donc qu'il possède des propriétés robustes. Ainsi, si $\hat{\theta}$ est un estimateur convergent de θ, la fonction d'influence s'utilise pour calculer le biais dû à une déviation infinitésimale au modèle (Heritier *et al.*, 2009). Ainsi,

$$\text{bias}(\hat{\theta}, F_\theta, \epsilon) = \sup_G \|\hat{\theta}(F_\epsilon) - \hat{\theta}(F_\theta)\| \approx \epsilon \sup_z \|IF(z; \hat{\theta}, F_\theta)\| \tag{1.4}$$

où $\|\cdot\|$ est la norme Euclidienne et $\sup_G(\cdot)$ est le supremum parmi toutes les fonctions G. Cela signifie que sous l'hypothèse d'une déviation infinitésimale au modèle le biais dépend du supremum de la fonction d'influence. Avoir une fonction d'influence bornée permet donc de limiter son supremum et, de ce fait, le biais maximal de l'estimateur $\hat{\theta}$ (Heritier *et al.*, 2009).
De plus, la fonction d'influence est utilisée pour calculer la variance asymptotique $V(\hat{\theta}, F_\theta)$ de l'estimateur $\hat{\theta}$ (Heritier *et al.*, 2009) :

$$V(\hat{\theta}, F_\theta) = \int IF(z; \hat{\theta}; F_\theta) IF^\top(z; \hat{\theta}; F_\theta) dF_\theta(z). \tag{1.5}$$

La fonction d'influence est un outil majeur dans le cadre de la statistique robuste. Elle est notamment utilisée pour la construction de M-estimateurs robustes.

1.4 Le point de rupture

La fonction d'influence considère une contamination infinitésimale, ce qui nous donne qu'un aperçu limité du problème. Le point de rupture ϵ^* est un outil qui permet de rendre compte de la résistance maximale d'un estimateur $\hat{\theta}$ du point de vue de la contamination du modèle. Il se définit par

$$\epsilon^*(\hat{\theta}, F_\theta) = \inf\{\epsilon \quad | \quad \text{bias}(\hat{\theta}, F_\theta, \epsilon) = \infty\}, \tag{1.6}$$

où $\text{bias}(\cdot)$ est le biais de l'estimateur $\hat{\theta}$ et $\inf\{\cdot\}$ est l'infimum. Autrement dit, le point de rupture est la limite inférieure de contamination ϵ qui rend le biais infini (Heritier *et al.*, 2009; Hampel *et al.*, 1986). Ainsi la moyenne possède un point de rupture $\epsilon^*(\overline{z}, F_\theta) = 0$ et la médiane $\epsilon^*(\text{median}\,(z_1, \ldots, z_n), F_\theta) = 0.5$.

1.5 L'estimation robuste

Il existe plusieurs types d'estimateurs robustes. Heritier *et al.* (2009) décrivent les plus couramment utilisés. Cependant, dans ce travail nous allons nous concentrer sur la classe des M-estimateurs $\hat{\boldsymbol{\theta}}_{\boldsymbol{M}}$. Il s'agit d'une généralisation des estimateurs du maximum de vraisemblance. Ils possèdent à l'instar des estimateurs du maximum de vraisemblance plusieurs propriétés utiles. Si $\boldsymbol{z_1}, \ldots, \boldsymbol{z_i}, \ldots, \boldsymbol{z_n}$ sont des observations multivariées du modèle $F_{\boldsymbol{\theta}}$, les M-estimateurs peuvent s'écrire sous la forme (Heritier *et al.*, 2009) :

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{M}} = \arg\max_\theta \sum_{i=1}^{n} \rho(\boldsymbol{z_i}; \boldsymbol{\theta}), \tag{1.7}$$

ce qui donne les équations du premier ordre (équations d'estimation) :

$$\sum_{i=1}^{n} \Psi(\boldsymbol{z_i}; \boldsymbol{\theta}) = \mathbf{0}, \tag{1.8}$$

où $\Psi(\boldsymbol{z};\boldsymbol{\theta}) = \frac{d\rho(\boldsymbol{z};\boldsymbol{\theta})}{d\boldsymbol{\theta}}$. L'estimateur du maximum de vraisemblance correspond donc à $\rho = -\ln f$, où f est la fonction de densité. On peut démontrer que la fonction d'influence d'un M-estimateur peut s'écrire (Heritier *et al.*, 2009) :

$$IF(\boldsymbol{z};\hat{\boldsymbol{\theta}}_{\boldsymbol{M}}, F_{\boldsymbol{\theta}}) = M^{-1}(\Psi; F_{\boldsymbol{\theta}})\Psi(\boldsymbol{z};\boldsymbol{\theta}), \tag{1.9}$$

lorsque

$$M(\Psi; F_{\boldsymbol{\theta}}) = -\int \frac{d}{d\boldsymbol{\theta}}\Psi(\boldsymbol{z};\boldsymbol{\theta})dF_{\boldsymbol{\theta}}(\boldsymbol{x}). \tag{1.10}$$

Cela nous indique que la fonction d'influence d'un M-estimateur est proportionnelle à sa fonction $\Psi(\cdot)$. Donc d'un point de vue de ses propriétés robustes, un M-estimateur possède une fonction d'influence bornée si sa fonction $\Psi(\cdot)$ l'est aussi. Il est donc possible de déduire les propriétés de robustesse d'un M-estimateur directement à partir des équations d'estimations (Heritier *et al.*, 2009).
Ainsi une stratégie à adopter afin de construire un estimateur robuste est ,par exemple, d'appliquer une fonction bornée aux équations de score classiques. Ces fonctions bornées peuvent être, parmi les plus couramment utilisées, la fonction ψ de Huber,

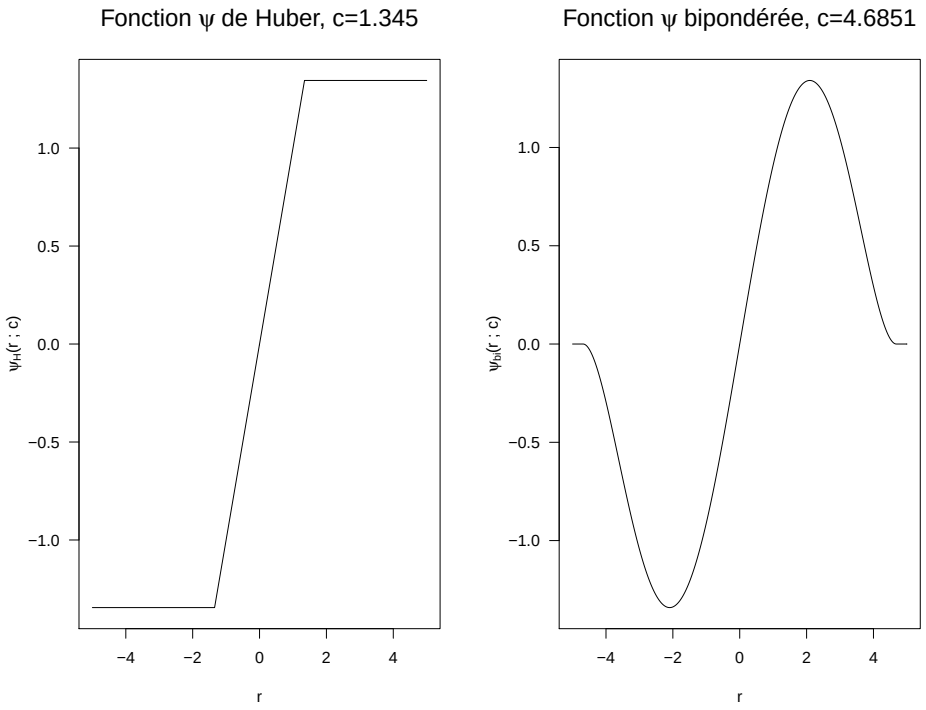
$$\psi_H(r;c) = \min[c, \max[r, -c]] = \begin{cases} c & \text{si } r > c \\ r & \text{si } c > r > -c \\ -c & \text{si } -c > r \end{cases}, \tag{1.11}$$

qui, appliquée à des résidus r_i, va les tronquer quand $|r_i| > c$, ou alors la fonction ψ bi-pondérée,

$$\psi_{bi}(r;c) = \begin{cases} \left(\left(\frac{r}{c}\right)^2 - 1\right)^2 r & \text{si } |r| \leq c \\ 0 & \text{si } |r| > c \end{cases}, \tag{1.12}$$

qui a la particularité d'annuler les résidus trop importants. Ces fonctions sont représentées dans la figure 1.1. On remarque ainsi que pour la fonction ψ de Huber les résidus ayant une trop grande importance sont tronqués à la valeur c ; donc quelque soit la déviation du modèle du résidu, il aura, au pire, une influence limitée dans l'estimation. Pour la fonction ψ bipondérée, les résidus trop grands sont annulés. Cette dernière peut poser cependant certains problèmes lors des calculs numériques car des équations d'estimation utilisant une fonction ψ bipondérée ne possèdent pas de solution unique ; toutes les solutions déviants fortement du modèle ont des équations d'estimation nulles. C'est pour cette raison que la fonction de Huber sera préférée dans le reste du travail. Sa fonction ρ ne possède qu'un seul maximum ce qui est commode lors de l'optimisation numérique.

FIGURE 1.1 – Deux fonctions ψ usuelles



Chapitre 2

Modèles linéaires généralisés

2.1 La famille exponentielle

Les modèles linéaires généralisés (GLM) sont bien définis pour les distributions de la famille exponentielle. Il existe en effet un lien étroit entre ces deux champs. Il est donc nécessaire d'introduire cette famille de distributions avant d'expliquer la construction des modèles linéaires généralisés.

2.1.1 Définition

La famille exponentielle est un ensemble de distributions dont leurs densités peuvent s'écrire sous une même forme. Ainsi la variable aléatoire Y suivra une distribution de la famille exponentielle si sa densité $f_Y(y)$ peut s'écrire sous la forme :

$$f_Y(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right], \tag{2.1}$$

où θ est appelé le paramètre canonique, ϕ est le paramètre de dispersion et les fonctions $a(\cdot)$, $b(\cdot)$ et $c(\cdot)$ sont spécifiques pour chaque distribution de la famille exponentielle (McCullagh et Nelder, 1989) dont font partie notamment les lois gaussienne, de Poisson, Gamma, négative-binomiale, Bernoulli, etc.

2.1.2 Espérance et variance

Il est possible de déduire une forme générale de la variance et de l'espérance des distributions issues de la famille exponentielle. En considérant la log-vraisemblance $l(\theta;y) = \ln(f_Y(y;\theta,\phi))$, on peut écrire :

$$\frac{dl(\theta;y)}{d\theta} = \frac{y - b'(\theta)}{a(\phi)} \tag{2.2}$$

et

$$\frac{d^2 l(\theta; y)}{d\theta^2} = -\frac{b''(\theta)}{a(\phi)}, \tag{2.3}$$

où $b'(\cdot)$ et $b''(\cdot)$ sont respectivement la première et seconde dérivé de $b(\cdot)$. De plus, les relations connues nous sont données par l'annexe A.1,

$$E\left[\frac{dl(\theta; y)}{d\theta}\right] = 0 \tag{2.4}$$

et

$$E\left[\frac{d^2 l(\theta; y)}{d\theta^2}\right] + E\left[\frac{dl^2(\theta; y)}{d\theta}\right] = 0. \tag{2.5}$$

Ainsi grâce à (2.2) et (2.4), on peut écrire :

$$E[Y] = b'(\theta) \tag{2.6}$$

Et grâce à (2.2), (2.3) et (2.5) on trouve :

$$V[Y] = b''(\theta)a(\phi). \tag{2.7}$$

L'espérance et la variance sont ainsi définies pour l'ensemble des distributions appartenant à la famille exponentielle (McCullagh et Nelder, 1989).

2.2 La construction du modèle

Les GLM sont une extension des modèles linéaires quand la variable de réponse Y_i n'est pas issue d'une distribution normale. Ils sont définis pour la famille exponentielle. Ils permettent donc de modéliser des variables discrètes ou asymétriques. Le modèle est construit de cette manière (McCullagh et Nelder, 1989; Heritier *et al.*, 2009) :

$$g(\mu_i) = \eta_i = \boldsymbol{x_i}^\top \boldsymbol{\beta}, \quad \forall i = 1, \ldots, n \tag{2.8}$$

où $\mu_i = E(Y_i)$ est l'espérance des variables de réponse Y_i indépendantes, $\boldsymbol{x_i} = [1 \quad x_{i,2} \quad \ldots \quad x_{i,p}]^\top$ est un vecteur contenant les $p-1$ variables explicatives de l'individu i, $\underset{p\times 1}{\boldsymbol{\beta}}$ sont les paramètres, η_i est appelé le prédicteur linéaire et $g(\cdot)$ est la fonction lien. Cette dernière définit la relation entre le prédicteur linéaire et l'espérance de la variable de réponse. Elle transforme le domaine de définition de μ_i, qui dépend de la distribution de Y_i, en celui de η_i qui est l'ensemble des réels $\mathbb{R}$ (McCullagh et Nelder, 1989; Heritier *et al.*, 2009).
Dans le cas des distributions de la famille exponentielle (2.1), la fonction lien est appelée fonction lien canonique quand $\eta_i = \theta_i$. Ainsi, en développant (2.6), on trouve $E[Y_i] = \mu_i = b'(\theta_i) = b'(\eta_i)$. La fonction lien canonique

dépend donc directement de la distribution de y_i (McCullagh et Nelder, 1989; Heritier *et al.*, 2009) et est donnée par

$$g(\cdot) = b'^{-1}(\cdot). \tag{2.9}$$

Si Y_i suit une distribution normale, la fonction lien canonique est la fonction identité. Dans ce cas, on retrouve la forme des modèles linéaires ; ce qui indique que les GLM sont une généralisation des modèles linéaires.

2.3 L'estimation des paramètres des GLM

L'estimation des $\boldsymbol{\beta}$ se fait par la méthode du maximum de vraisemblance. Grâce aux observations $\boldsymbol{y} = [y_1 \quad \dots \quad y_n]^\top$ et $\boldsymbol{X} = [\boldsymbol{x_1} \quad \dots \quad \boldsymbol{x_n}]^\top$, on écrit la vraisemblance $L(\theta_1, \dots, \theta_n; \boldsymbol{y}, \boldsymbol{X})$ et la log-vraisemblance $l(\theta_1, \dots, \theta_n; \boldsymbol{y}, \boldsymbol{X})$ pour la famille exponentielle :

$$\begin{aligned} L(\theta_1, \dots, \theta_n; \boldsymbol{y}, \boldsymbol{X}) &= \prod_{i=1}^{n} \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right] \\ &= \exp\left[\sum_{i=1}^{n} \frac{y_i\theta_i}{a_i(\phi)} - \sum_{i=1}^{n} \frac{b(\theta_i)}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i, \phi)\right], \end{aligned} \tag{2.10}$$

et

$$\begin{aligned} l(\theta_1, \dots, \theta_n; \boldsymbol{y}, \boldsymbol{X}) &= \sum_{i=1}^{n} \frac{y_i\theta_i}{a_i(\phi)} - \sum_{i=1}^{n} \frac{b(\theta_i)}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i, \phi) \\ &= \sum_{i=1}^{n} l_i. \end{aligned} \tag{2.11}$$

La relation entre le paramètre canonique θ_i et les paramètres $\boldsymbol{\beta}$ des GLM se fait grâce aux relations (2.6) et (2.8). On écrit donc l'estimateur de maximum de vraisemblance :

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\boldsymbol{MLE}} &= \arg\max_{\boldsymbol{\beta}} \left[L(\theta_1, \dots, \theta_n; \boldsymbol{y}, \boldsymbol{X})\right] \\ &= \arg\max_{\boldsymbol{\beta}} \left[l(\theta_1, \dots, \theta_n; \boldsymbol{y}, \boldsymbol{X})\right]. \end{aligned} \tag{2.12}$$

La solution de (2.12) correspond à résoudre le système d'équations de score $\boldsymbol{U} = \frac{d}{d\boldsymbol{\beta}} \sum_{i=1}^{n} l_i = \boldsymbol{0}$, c'est-à-dire :

$$\boldsymbol{U} = \sum_{i=1}^{n} \frac{dl_i}{d\boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{dl_i}{d\theta_i} \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\boldsymbol{\beta}} = \boldsymbol{0}. \tag{2.13}$$

Premièrement selon (2.2) et (2.6), on trouve $\frac{dl_i}{d\theta_i} = \frac{y_i - \mu_i}{a_i(\phi)}$. Deuxièmement, en prenant (2.6) et (2.7) $\frac{d\theta_i}{d\mu_i} = \left(\frac{d\mu_i}{d\theta_i}\right)^{-1} = \left(\frac{V[Y_i]}{a_i(\phi)}\right)^{-1}$. Troisièmement, $\frac{d\mu_i}{d\boldsymbol{\beta}} =$

$\frac{\mu_i}{\eta_i}\frac{\eta_i}{\boldsymbol{\beta}} = \frac{\mu_i}{\eta_i}\boldsymbol{x_i}$. Ainsi, on peut réécrire (2.13) :

$$\boldsymbol{U} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{V[Y_i]} \frac{d\mu_i}{d\eta_i} \boldsymbol{x_i} = \boldsymbol{0}. \tag{2.14}$$

On peut utiliser un algorithme itératif de Newton pour résoudre (2.14). Ainsi à l'itération k on trouve :

$$\begin{aligned} \boldsymbol{\beta}^{[k]} &= \boldsymbol{\beta}^{[k-1]} - \left(\frac{d\boldsymbol{U}}{d\boldsymbol{\beta}^\top}^{[k-1]} \right)^{-1} \boldsymbol{U}^{[k-1]} \\ &\approx \boldsymbol{\beta}^{[k-1]} + \left(\boldsymbol{J}^{[k-1]} \right)^{-1} \boldsymbol{U}^{[k-1]}, \end{aligned} \tag{2.15}$$

où $\boldsymbol{J} = -E[\frac{d\boldsymbol{U}}{d\boldsymbol{\beta}^\top}]$ est la matrice d'information de Fisher dont les éléments m, n sont calculés $\boldsymbol{J}_{m,n} = E[\boldsymbol{U}_m \boldsymbol{U}_n] = \sum_{i=1}^{n} \frac{x_{im} x_{in}}{V[Y_i]} \left(\frac{d\mu_i}{d\eta_i} \right)^2$. L'algorithme utilisant la matrice d'information de Fisher approxime la méthode de Newton et s'appelle l'algorithme de score de Fisher.

2.3.1 Variance asymptotique

Etant un estimateur de maximum de vraisemblance, $\hat{\boldsymbol{\beta}}_{\boldsymbol{MLE}}$ a donc une variance définie selon la théorie de la vraisemblance (Lejeune, 2010). Il suit donc asymptotiquement une loi normale :

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\boldsymbol{MLE}} - \boldsymbol{\beta}) \to N\left(0, \boldsymbol{J}^{-1}(\boldsymbol{\beta})\right), \tag{2.16}$$

où $\boldsymbol{J}(\boldsymbol{\beta})$ est la matrice d'information de Fisher.

2.3.2 Modéliser la sur-dispersion

Certaines distributions, notamment les distributions de Poisson et binomiale, imposent une variance $V[Y_i]$ comme fonction du seul paramètre μ_i. Cependant, les données observées peuvent ne pas satisfaire cette restriction. Dans ce cas, il est donc judicieux de permettre plus de liberté à la variance en ajoutant un paramètre supplémentaire. Ainsi, on suppose $V[Y_i] = \tau V(\mu_i)$ et on parle de sur-dispersion quand $\tau > 1$ et de sous-dispersion quand $\tau < 1$. Dans ce cas, l'estimation se fait en maximisant la quasi-vraisemblance $\sum_{i=1}^{n} Q(y_i, \mu_i)$, car la vraisemblance pour ce type de modèle n'existe pas. Ce qui aboutit au système d'équations d'estimation :

$$\frac{d}{d\boldsymbol{\beta}} \sum_{i=1}^{n} Q(y_i, \mu_i) = \frac{d}{d\boldsymbol{\beta}} \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{\tau V(t)} dt = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\tau V(\mu_i)} \frac{d\mu_i}{d\boldsymbol{\beta}} = \boldsymbol{0}. \tag{2.17}$$

Il est important de noter que le paramètre τ ne va pas changer l'estimation des $\boldsymbol{\beta}$, mais intervient lors du calcul de la variance des estimateurs $V[\hat{\boldsymbol{\beta}}]$.

2.3.3 Déviance et résidus

La déviance mesure la distance entre le modèle estimé et le modèle saturé. Le modèle saturé permet un maximum de paramètres ; ce qui correspond à l'interpolation. Sous l'hypothèse que l'on peut écrire $a(\phi) = \phi/w$ de l'équation (2.1), la déviance $D(\boldsymbol{y}, \boldsymbol{\mu})$ se calcule :

$$D(\boldsymbol{y}, \boldsymbol{\mu}) = 2\phi \left[l(\boldsymbol{y}; \boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) \right] = \phi \sum_{i=1}^{n} 2 \left[l_i(y_i; y_i) - l_i(\hat{\mu}_i; y_i) \right] = \phi \sum_{i=1}^{n} d_i, \tag{2.18}$$

où $\hat{\boldsymbol{\mu}} = [\hat{\mu}_1 \quad \dots \quad \hat{\mu}_n]^\top$ est le vecteur des valeurs ajustées, $\hat{\mu}_i = g^{-1}(\boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}})$, $l(\boldsymbol{y}; \boldsymbol{y})$ et $l(\hat{\boldsymbol{\mu}}; \boldsymbol{y})$ sont les log-vraisemblances des modèles saturés et postulés respectivement (McCullagh et Nelder, 1989; Heritier *et al.*, 2009).
Les GLM admettent plusieurs types de résidus. Les résidus de Pearson sont définis ainsi :

$$r_{Pi} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi} V[\hat{\mu}_i]}}, \tag{2.19}$$

et les résidus de déviance :

$$r_{Di} = \text{sign } (y_i - \hat{\mu}_i) \sqrt{d_i}, \tag{2.20}$$

où les d_i viennent de (2.18). Ces résidus n'ont pas de distribution spécifique. Cependant, on les compare généralement à une distribution normale afin de contrôler la présence de données extrêmes ainsi que s'assurer qu'il n'y ait pas de dépendance temporelle, ou de structures particulières. Il faut cependant noter que des structures apparaissent dans le cas où la variable de réponse est de nature discrète.

2.4 L'estimation robuste

Les estimateurs des $\boldsymbol{\beta}$ pour les GLM robustes reprennent la stratégie expliquée dans le chapitre 1.5 en l'appliquant à l'estimateur décrit précédemment. On transforme les équations d'estimation (2.17) en appliquant une fonction ψ de Huber (1.11) bornée aux résidus et en ajoutant des poids $w(\boldsymbol{x_i})$ (Cantoni et Ronchetti, 2001, 2006; Heritier *et al.*, 2009). On trouve ainsi les équations d'estimation :

$$\sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x_i}; \boldsymbol{\beta}, \phi, c) = \sum_{i=1}^{n} \left[\psi_H(r_i; c) w(\boldsymbol{x_i}) \frac{1}{\sqrt{\phi V(\mu_i)}} \frac{d\mu_i}{d\boldsymbol{\beta}} - a(\beta) \right] = \mathbf{0}, \tag{2.21}$$

où $r_i = (y_i - \mu_i)/\sqrt{\phi V(\mu_i)}$ sont les résidus de Pearson, $w(\boldsymbol{x_i})$ sont des poids prédéfinis, $\psi_H(r_i; c) = \min(c, \max(r_i, -c))$ est la fonction ψ de Huber et $a(\beta) = \frac{1}{n} \sum_{i=1}^{n} E[\psi_H(r_i; c) w(\boldsymbol{x_i}) / \sqrt{\phi V(\mu_i)}]$ est une correction qui assure la consistence de l'estimateur (Cantoni et Ronchetti, 2001, 2006; Heritier *et al.*,

2009).
Il s'agit donc d'un estimateur issu de la classe des M-estimateurs et est appelé un estimateur de type Mallows. Sa paramétrisation usuelle utilise $c = 1.345$ et $w(\boldsymbol{x_i}) = \sqrt{1 - h_{ii}}$, où h_{ii} sont les éléments de la diagonale de la matrice $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$, où $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x_1} & \dots & \boldsymbol{x_i} & \dots & \boldsymbol{x_n} \end{pmatrix}^\top$. Il est possible de définir plusieurs formes pour les poids $w(\boldsymbol{x_i})$, notamment unitaire ou en fonction de la distance de Mahalanobis (Heritier *et al.*, 2009). Cette dernière possibilité pose cependant des problèmes lorsque les variables explicatives comportent des facteurs.
Ces équations d'estimation correspondent à la maximisation de la quasi-vraisemblance de Mallows (Cantoni et Ronchetti, 2001, 2006; Heritier *et al.*, 2009) :

$$\begin{aligned} \sum_{i=1}^{n} Q_M(\mu_i; y_i) = & \int_{\tilde{s}}^{\mu_i} \phi\left(\frac{y_i - t}{\phi v_t}; c\right) w(\boldsymbol{x_i}) \frac{1}{\sqrt{\phi v_t}} dt \\ & - \frac{1}{n} \sum_{j=1}^{n} E\left[\int_{\tilde{s}}^{\mu_i} \phi\left(\frac{y_i - t}{\phi v_t}; c\right) w(\boldsymbol{x_i}) \frac{1}{\sqrt{\phi v_t}} dt\right], \end{aligned} \tag{2.22}$$

où $\frac{d}{d\boldsymbol{\beta}} \sum_{i=1}^{n} Q_M(\mu_i; y_i) = \sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x_i}; \boldsymbol{\beta}, \phi, c)$.
La résolution de (2.21) peut se faire grâce à un algorithme de Newton. Ainsi on peut écrire $\boldsymbol{\beta_R^{[k]}}$ la k-ième itération de l'algorithme de Newton :

$$\begin{aligned} \boldsymbol{\beta_R^{[k]}} = & \boldsymbol{\beta_R^{[k-1]}} - \left[\frac{d}{d\boldsymbol{\beta}^\top}\left(\sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x_i}; \boldsymbol{\beta}, \phi, c)\right)\right]^{-1}\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta_R^{[k-1]}}} \\ & \times \left[\sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x_i}; \boldsymbol{\beta}, \phi, c)\right]\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta_R^{[k-1]}}}. \end{aligned} \tag{2.23}$$

La matrice $\left[\frac{d}{d\boldsymbol{\beta}^\top}\left(\sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x_i}; \boldsymbol{\beta}, \phi, c)\right)\right]^{-1}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta_R^{[k-1]}}}$ peut être facilement approximée grâce à la relation $M(\Psi, F_{\boldsymbol{\beta}}) = -E[\frac{d}{d\boldsymbol{\beta}^\top} \sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x_i}; \boldsymbol{\beta}, \phi, c)] = \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{B}\boldsymbol{X}$, où $\boldsymbol{B}$ est une matrice diagonale dont les éléments s'écrivent $B_{ii} = E[\psi_H(r_i; c)\frac{d}{d\mu_i} \ln h(y_i|\boldsymbol{x_i}, \mu_i)]\frac{1}{\sqrt{\phi V(\mu_i)}} w(\boldsymbol{x_i})(\frac{d\mu_i}{d\eta_i})^2$, où $h(\cdot)$ est la probabilité conditionnelle de $y_i|\boldsymbol{x_i}$ (Cantoni et Ronchetti, 2001).

2.4.1 Variance asymptotique

La variance asymptotique peut être écrite grâce à la fonction d'influence (1.5). De plus, pour un M-estimateur, la fonction d'influence s'écrit en fonction des équations d'estimation $\sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x_i}; \boldsymbol{\beta}, \phi, c)$ et de la matrice $M(\Psi; F)$ définie par l'équation (1.9). La distribution asymptotique de $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ s'écrit donc :

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\boldsymbol{R}} - \boldsymbol{\beta}) \to N\left(0, M^{-1}(\Psi, F_{\boldsymbol{\beta}})Q(\Psi, F_{\boldsymbol{\beta}})M^{-1}(\Psi, F_{\boldsymbol{\beta}})\right), \tag{2.24}$$

où $M^{-1}(\Psi, F_{\boldsymbol{\beta}})Q(\Psi, F_{\boldsymbol{\beta}})M^{-1}(\Psi, F_{\boldsymbol{\beta}})$ est la variance asymptotique de $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$, $M(\Psi, F_{\boldsymbol{\beta}}) = \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{B}\boldsymbol{X}$ est décrit précédemment et $Q(\Psi, F_{\boldsymbol{\beta}}) = \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} - a(\boldsymbol{\beta})a(\boldsymbol{\beta})^\top$, où $\boldsymbol{A}$ est une matrice diagonale dont les éléments a_i s'écrivent $a_i = E[\psi_H(r_i; c)^2]w^2(\boldsymbol{x_i})\frac{1}{\phi V(\mu_i)}(\frac{d\mu_i}{d\eta_i})^2$ (Cantoni et Ronchetti, 2001; Heritier *et al.*, 2009). Le comportement asymptotique de $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ est donc parfaitement décrit.

2.5 L'estimateur bayésien robuste

L'estimateur développé dans ce chapitre est une adaptation pour les GLM de l'estimateur construit par Genton et Ronchetti (2008) pour les modèles linéaires. Il repose sur des hypothèses bayésiennes et a montré de bonnes qualités de prévision. Ce chapitre explique en parallèle le développement de Genton et Ronchetti (2008) (mis en cadre) et l'adaptation pour les GLM.

Genton et Ronchetti (2008) définissent pour le modèle linéaire $y_i = \boldsymbol{x_i}^\top \boldsymbol{\beta} + \epsilon_i$ où $\epsilon_i \sim (0, \sigma^2)$ l'espérance de la distribution à posteriori :

$$\boldsymbol{\beta}_B = E[\boldsymbol{\beta}|y_i, \ldots, y_n] = \frac{\int \boldsymbol{\beta} \exp\left[-nk_R(\boldsymbol{\beta}; y_1, \ldots, y_n)\right] d\boldsymbol{\beta}}{\int \exp\left[-nk_R(\boldsymbol{\beta}; y_1, \ldots, y_n)\right] d\boldsymbol{\beta}},$$

avec

$$k_R(\boldsymbol{\beta}; y_1, \ldots, y_n) = \frac{1}{n}\left[\sum_{i=1}^{n} \ln(\sigma^{-1} g(\frac{y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta}}{\sigma}))\right] - \frac{1}{n}\ln(h(\boldsymbol{\beta})),$$

où $g(\cdot) \propto \exp(-\rho(\cdot))$ pour une fonction $\rho(\cdot)$ bornée et $h(\boldsymbol{\beta})$ est la distribution à priori de $\boldsymbol{\beta}$.

Ainsi pour les GLM, on peut considérer l'adaptation et écrire l'espérance de la distribution à posteriori,

$$\begin{aligned} \boldsymbol{\beta}_B = E[\boldsymbol{\beta}|y_i, \ldots, y_n] &= \frac{\int \boldsymbol{\beta} \exp\left[\sum_{i=1}^n Q_M(y_i, \mu_i)\right] h(\boldsymbol{\beta}) d\boldsymbol{\beta}}{\int \exp\left[\sum_{i=1}^n Q_M(y_i, \mu_i)\right] h(\boldsymbol{\beta}) d\boldsymbol{\beta}} \\ &= \frac{\int \boldsymbol{\beta} \exp\left[-nk_R(\boldsymbol{\beta}; y_1, \ldots, y_n)\right] d\boldsymbol{\beta}}{\int \exp\left[-nk_R(\boldsymbol{\beta}; y_1, \ldots, y_n)\right] d\boldsymbol{\beta}}, \end{aligned} \tag{2.25}$$

où $k_R(\boldsymbol{\beta}; y_1, \ldots, y_n) = -\frac{1}{n}\left[\sum_{i=1}^{n} Q_M(y_i, \mu_i)\right] - \frac{1}{n}\ln(h(\boldsymbol{\beta}))$, $h(\beta)$ est la distribution à priori de $\boldsymbol{\beta}$ et $\sum_{i=1}^n Q_M(y_i, \mu_i)$ est la quasi vraisemblance de

Mallows décrite par l'équation (2.22).

Ainsi, l'idée principale qui permet d'écrire l'équation (2.25) est de remplacer la log-vraisemblance, usuellement utilisée pour la distribution à posteriori, par son équivalent robuste, la quasi-vraisemblance de Mallows (2.22).

Genton et Ronchetti (2008) utilisent l'approximation d'intégrale de Laplace ainsi qu'une hypothèse de normalité de la distribution à priori et trouvent $\hat{\boldsymbol{\beta}}$ qui approxime $\boldsymbol{\beta_B}$ si il vérifie :

$$\sum_{i=1}^{n} \frac{1}{\sigma}\psi(\frac{y_i - \boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}}}{\sigma})\boldsymbol{x_i} - \frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\mu_p}}{\delta^2} = \mathbf{0},$$

où $\psi(r_i) = \frac{d\rho(r_i)}{dr_i}$.

À l'instar de Genton et Ronchetti (2008), l'intégrale (2.25) peut être résolue grâce à l'approximation d'intégrale de Laplace (Annexe A.2). On trouve ainsi $\hat{\boldsymbol{\beta}}$ une approximation d'ordre $\boldsymbol{\beta_B} = \hat{\boldsymbol{\beta}}[1 + O(n^{-1})]$, où $\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} k_R(\boldsymbol{\beta}; y_1, \ldots, y_n)$. En imposant la normalité de la distribution à priori, $h(\boldsymbol{\beta}) \sim N(\boldsymbol{\mu_p}, \delta^2 \boldsymbol{I_p})$, il est possible de trouver $\hat{\boldsymbol{\beta}}$ qui vérifie le système d'équations

$$\sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x_i}; \hat{\boldsymbol{\beta}}, \phi, c) - \frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\mu_p}}{\delta^2} = \mathbf{0}, \tag{2.26}$$

où $\sum_{i=1}^{n} \Psi(\cdot)$ correspond à l'équation (2.21), $\boldsymbol{\mu_p}$ et δ^2 sont respectivement l'espérance et la variance de la distribution à priori de $\boldsymbol{\beta}$.

Genton et Ronchetti (2008), après avoir utilisé l'approximation d'intégrale de Laplace, font une expansion de Taylor et trouvent :

$$\begin{aligned}
\mathbf{0} &\approx \sum_{i=1}^{n} \frac{1}{\sigma}\psi(\frac{y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta}}{\sigma})\boldsymbol{x_i}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta_0}} \\
&\quad + \frac{d}{d\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{1}{\sigma}\psi(\frac{y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta}}{\sigma})\boldsymbol{x_i}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta_0}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) - \frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\mu_p}}{\delta^2} \\
&= \boldsymbol{A} + \delta^{-2}\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) - \frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\mu_p}}{\delta^2},
\end{aligned}$$

où $\boldsymbol{A} = \sum_{i=1}^{n} \frac{1}{\sigma}\psi(\frac{y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta_0}}{\sigma})\boldsymbol{x_i}$ et $\boldsymbol{\Omega} = \delta^2 \frac{d}{d\boldsymbol{\beta}^\top} \sum_{i=1}^{n} \frac{1}{\sigma}\psi(\frac{y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta_0}}{\sigma})\boldsymbol{x_i}$. Il est possible d'isoler $\hat{\boldsymbol{\beta}} = (\boldsymbol{\Omega} + \boldsymbol{I})^{-1}\boldsymbol{\Omega}[\hat{\boldsymbol{\beta}}_{\boldsymbol{R}} + \boldsymbol{\Omega}^{-1}\boldsymbol{\mu_p}]$, où $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}} = \boldsymbol{\beta_0} + \delta^2\boldsymbol{\Omega}^{-1}\boldsymbol{A}$ et $\delta^2\boldsymbol{\Omega}^{-1}\boldsymbol{A}$ correspond donc à un pas supplémentaire de l'algorithme de Newton. On peut donc considérer $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ comme étant un M-estimateur robuste.

Pour les GLM, en faisant l'expansion de Taylor du premier terme de (2.26),

on trouve :

$$\begin{aligned}\mathbf{0} &\approx \sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x_i}; \boldsymbol{\beta}, \phi, c)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta_0}} \\ &\quad + \frac{d}{d\boldsymbol{\beta}} \sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x_i}; \boldsymbol{\beta}, \phi, c)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta_0}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) - \frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\mu_p}}{\delta^2} \\ &= \boldsymbol{A} + \delta^{-2}\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) - \frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\mu_p}}{\delta^2},\end{aligned} \tag{2.27}$$

où $\boldsymbol{A} = \sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x_i}; \boldsymbol{\beta_0}, \phi, c)$ et $\boldsymbol{\Omega} = \delta^2 \frac{d}{d\boldsymbol{\beta}} \sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x_i}; \boldsymbol{\beta_0}, \phi, c)$. On peut donc isoler $\hat{\boldsymbol{\beta}}$ et trouver la solution :

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{\Omega} + \boldsymbol{I})^{-1}\boldsymbol{\Omega}[\hat{\boldsymbol{\beta}}_{\boldsymbol{R}} + \boldsymbol{\Omega}^{-1}\boldsymbol{\mu_p}], \tag{2.28}$$

où $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}} = \boldsymbol{\beta_0} + \delta^2\boldsymbol{\Omega}^{-1}\boldsymbol{A}$ et $\delta^2\boldsymbol{\Omega}^{-1}\boldsymbol{A}$ correspond donc à un pas supplémentaire de l'algorithme de Newton (2.23). $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ est donc un estimateur de $\boldsymbol{\beta}$ robuste défini par (2.21).

La suite du raisonnement de Genton et Ronchetti (2008) est de simplifier $(\boldsymbol{\Omega} + \boldsymbol{I})^{-1}\boldsymbol{\Omega} \to \boldsymbol{I_p}$ en laissant $\delta^2 \to \infty$. Ils trouvent donc $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{R}} + \boldsymbol{\Omega}^{-1}\boldsymbol{\mu_p}$, où $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ est un estimateur robuste du modèle linéaire. L'estimation de $\boldsymbol{\Omega}$ se fait en prenant $\hat{\boldsymbol{\Omega}} = \frac{\delta^2}{\sigma^2} \sum_{i=1}^{n} \psi'(\frac{y_i - \boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}}_{\boldsymbol{R}}}{\sigma}) \boldsymbol{x_i}\boldsymbol{x_i}^\top$ qui est simplifié par $\frac{\delta^2}{\sigma^2} \sum_{i=1}^{n} \boldsymbol{x_i}\boldsymbol{x_i}^\top$ et $\boldsymbol{\mu_p}$ est estimé par $\sum_{i=1}^{n} \sigma\psi_H(\frac{y_i - \boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}}_{\boldsymbol{R}}}{\sigma}; c_2)\boldsymbol{x_i}$, en prenant $c_2 > c$, où c est le paramètre de la ψ-fonction de Huber nécessaire pour estimer $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$; on choisit usuellement $c = 1.345$. On trouve ainsi l'estimateur :

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{R}} + \left(\sum_{i=1}^{n} \boldsymbol{x_i}\boldsymbol{x_i}^\top\right)^{-1} \sum_{i=1}^{n} \hat{\sigma}\psi_H(\frac{y_i - \boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}}_{\boldsymbol{R}}}{\hat{\sigma}}; c_2)\boldsymbol{x_i},$$

où $\delta^2 = 1$.

Dans le cas des GLM, il est aussi possible de simplifier (2.28) en admettant $(\boldsymbol{\Omega} + \boldsymbol{I})^{-1}\boldsymbol{\Omega} \to \boldsymbol{I_p}$ si $\delta^2 \to \infty$ ce qui mène à la forme simple :

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{R}} + \boldsymbol{\Omega}^{-1}\boldsymbol{\mu_p}. \tag{2.29}$$

L'estimation des paramètres $\boldsymbol{\Omega}$ et $\boldsymbol{\mu_p}$ pour ce nouvel estimateur des GLM se fait similairement à l'estimateur décrit par Genton et Ronchetti (2008). Ainsi $\hat{\boldsymbol{\Omega}} = \delta^2 \sum_{i=1}^{n} E[\frac{d}{d\boldsymbol{\beta}} \Psi(y_i, \boldsymbol{x_i}; \hat{\boldsymbol{\beta}}_{\boldsymbol{R}}, \phi, c)]$, où l'espérance correspond à celle décrite par la matrice $-M(\Psi, F_{\boldsymbol{\beta}})$ et $\hat{\boldsymbol{\mu}}_{\boldsymbol{p}} = \sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x_i}; \hat{\boldsymbol{\beta}}_{\boldsymbol{R}}, \phi, c_2)$ où $c_2 > c$. Cet estimateur correspond donc à l'estimateur robuste décrit par l'équation (2.21) corrigé par un terme.

L'analyse de cette correction montre qu'elle n'est composée que d'une somme

d'une fonction des valeurs extrêmes. Cette interprétation peut se faire en analysant l'estimation de $\hat{\boldsymbol{\mu}}_{\boldsymbol{p}} = \sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x_i}; \hat{\boldsymbol{\beta}}_{\boldsymbol{R}}, \phi, c_2)$ en fonction de c_2 ; ce qui revient à analyser l'équation (2.21) en fixant $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ et à considérer c_2 comme la variable d'intérêt. Il est important de noter que deux valeurs, c et c_2, interviennent dans l'estimation de $\boldsymbol{\mu_p}$. c est fixe et sert à estimer $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ et c_2 intervient dans l'évaluation des équations d'estimation en $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ et nécessite d'être calibré.
Tout d'abord, on remarque que dans le cas $c_2 = c$, l'estimé de $\boldsymbol{\mu_p}$ est zéro. Or dans le cas où $c_2 > c$, les données déviant au modèle apporteront des termes non nulles à la somme $\hat{\boldsymbol{\mu}}_{\boldsymbol{p}} = \sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x_i}; \hat{\boldsymbol{\beta}}_{\boldsymbol{R}}, \phi, c_2)$. En effet, dans l'équation (2.21), seule la fonction ψ est dépendante de c. Et il est possible de décomposer la fonction ψ de Huber décrite par l'équation (1.11) ainsi :

$$\psi_H(r_i; c_2) = \begin{cases} c + (c_2 - c) & \text{si } r_i > c_2 \\ c + (r_i - c) & \text{si } c_2 > r_i > c \\ r_i & \text{si } c > r_i > -c \\ -c + (r_i + c) & \text{si } -c > r_i > -c_2 \\ -c + (-c_2 + c) & \text{si } -c_2 > r_i, \end{cases} , \tag{2.30}$$

ce qui équivaut à

$$\psi_H(r_i; c_2) = \psi_H(r_i; c) + \Delta_\psi(r_i; c_2, c), \tag{2.31}$$

où la fonction $\Delta_\psi(r_i; c_2, c)$ est non-nulle pour tous les résidus plus grands que c :

$$\Delta_\psi(r_i; c_2, c) = \begin{cases} c_2 - c & \text{si } r_i > c_2 \\ r_i - c & \text{si } c_2 > r_i > c \\ 0 & \text{si } c > r_i > -c \\ r_i + c & \text{si } -c > r_i > -c_2 \\ -c_2 + c & \text{si } -c_2 > r_i. \end{cases} , \tag{2.32}$$

Il est ainsi possible d'utiliser la décomposition de l'équation (2.31) et de l'appliquer à l'estimation de $\boldsymbol{\mu_p}$ pour trouver $\hat{\boldsymbol{\mu}}_{\boldsymbol{p}} = \sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x_i}; \hat{\boldsymbol{\beta}}_{\boldsymbol{R}}, \phi, c_2) = \sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x_i}; \hat{\boldsymbol{\beta}}_{\boldsymbol{R}}, \phi, c) + \sum_{i=1}^{n} \Delta_\Psi(y_i, \boldsymbol{x_i}; \hat{\boldsymbol{\beta}}_{\boldsymbol{R}}, \phi, c_2, c)$. Or le premier terme est nul et les éléments du second terme sont non-nuls pour les résidus $r_i > c$ (soit les résidus des observations extrêmes). Ainsi seules les données extrêmes servent à l'estimation de $\boldsymbol{\mu_p}$ et donc à la correction de $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$.

2.5.1 Une seconde approche de l'estimateur bayésien

Afin de comprendre l'effet des différentes approximations (expansion de Taylor, $\delta^2 \to \infty$ pour passer de (2.28) à (2.29)) qui mène à la solution définie par l'équation (2.29), l'équation (2.26) est résolue numériquement. On peut donc décrire le k-ième pas de l'algorithme de Newton :

$$\begin{aligned}\boldsymbol{\beta}_{\boldsymbol{B}}^{[\boldsymbol{k}]} &= \boldsymbol{\beta}_{\boldsymbol{B}}^{[\boldsymbol{k-1}]} - \left[\frac{d}{d\boldsymbol{\beta}^\top}\left(\sum_{i=1}^{n}\Psi(y_i,\boldsymbol{x_i};\boldsymbol{\beta},\phi,c) - \frac{\boldsymbol{\beta}-\boldsymbol{\mu_p}}{\delta^2}\right)\right]^{-1}\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{\boldsymbol{B}}^{[\boldsymbol{k-1}]}} \\ &\quad \left[\sum_{i=1}^{n}\Psi(y_i,\boldsymbol{x_i};\boldsymbol{\beta},\phi,c) - \frac{\boldsymbol{\beta}-\boldsymbol{\mu_p}}{\delta^2}\right]\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{\boldsymbol{B}}^{[\boldsymbol{k-1}]}},\end{aligned} \tag{2.33}$$

où l'on peut approximer $\left[\frac{d}{d\boldsymbol{\beta}^\top}\left(\sum_{i=1}^{n}\Psi(y_i,\boldsymbol{x_i};\boldsymbol{\beta},\phi,c) - \frac{\boldsymbol{\beta}-\boldsymbol{\mu_p}}{\delta^2}\right)\right] = \left[\frac{d}{d\boldsymbol{\beta}^\top}\left(\sum_{i=1}^{n}\Psi(y_i,\boldsymbol{x_i};\boldsymbol{\beta},\phi,c)\right) - \delta^{-2}\boldsymbol{I_p}\right]$ par $-\boldsymbol{M}\left(\Psi(y_i,\boldsymbol{x_i};\boldsymbol{\beta},\phi,c)\right) - \delta^{-2}\boldsymbol{I_p}$, similairement à l'algorithme du score de Fisher utilisé pour l'estimation robuste (2.23).
Ainsi deux solutions au même problème pourront être comparées. La première est une approximation analytique qui comporte l'avantage d'avoir une forme simple et la seconde une résolution numérique qui est plus précise mais nécessite plus de calculs.

Les deux estimateurs bayésiens définis par les équations (2.29) et (2.33) nécessitent deux paramètres supplémentaires δ^2 et c_2. Le paramètre δ^2 a une influence sur la force de la correction introduite par l'introduction de l'hypothèse bayésienne et $\hat{\boldsymbol{\beta}} \to \hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ quand $\delta^2 \to \infty$. Il est donc nécessaire d'effectuer un calibrage de ce paramètre afin d'aboutir à une correction optimale. Le paramètre c_2 sera repris de l'analyse faite par Genton et Ronchetti (2008). Ils aboutissent à la conclusion qu'une paramétrisation raisonnable est de fixer $c_2 = 2$.

Chapitre 3

Étude Monte-Carlo

3.1 Le modèle simulé

La simulation se focalise sur un modèle GLM dont la variable de réponse Y_i est issue d'une loi de Poisson. Trois variables $X_1 \sim N(0,1)$, $X_2 \sim$ Bernoulli(0.5) et $X_3 \sim$ Uniform$(0,1)$ servent de variables explicatives. On trouve donc :

$$\begin{aligned} g(E[Y_i]) &= \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} \quad i = 1, \ldots, n \\ &= \eta_i = \boldsymbol{x_i}^\top \boldsymbol{\beta} \end{aligned} \tag{3.1}$$

où $g(\cdot)$ est la fonction lien canonique (le logarithme pour une distribution de Poisson), $\boldsymbol{X} = [\boldsymbol{x_1} \quad \ldots \quad \boldsymbol{x_i} \quad \ldots \quad \boldsymbol{x_n}]^\top$, où $\boldsymbol{x_i} = [1 \quad x_{i,1} \quad x_{i,2} \quad x_{i,3}]^\top$ est le vecteur comprenant les variables explicatives de l'individu i et $\boldsymbol{\beta} = [\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3]^\top$ sont les paramètres définis comme étant $\boldsymbol{\beta} = [0.7 \quad 0.9 \quad 0.8 \quad 0.6]^\top$; bien que choisis arbitrairement, ils ont été préférés relativement petit afin d'assurer la stabilité numérique de la simulation. En effet, $g^{-1}(\cdot)$ est la fonction exponentielle, il est donc préférable de s'assurer que η_i est petit afin d'obtenir des valeurs y_i d'un même ordre de grandeur.
Avec un modèle ainsi défini, la création des jeux de données nécessaires pour l'étude Monte-Carlo a été faite en simulant indépendamment les trois variables explicatives X_1, X_2 et X_3. Il est ainsi possible de calculer le prédicteur linéaire, puis l'espérance $E[Y_i] = g^{-1}(\eta_i)$ grâce à la fonction lien. Connaissant l'espérance, il est possible de simuler les observations de la variable de réponse $\boldsymbol{Y} = [y_1 \quad \ldots \quad y_i \quad \ldots \quad y_n]^\top$.
Pour l'étude Monte-Carlo, on a pris une taille d'échantillon $n = 100$ et on a répliqué le processus 203 fois dont 3 boucles ont été supprimées pour des problèmes de convergence des algorithmes.

3.1.1 Estimation des $\boldsymbol{\beta}$

Les différentes méthodes d'estimation décrites auparavant sont utilisées afin de fournir une estimation pour $\boldsymbol{\beta}$. Ainsi on trouve une estimation par maxi-

mum de vraisemblance, deux estimations robustes, et quatre estimations bayésiennes robustes possibles :

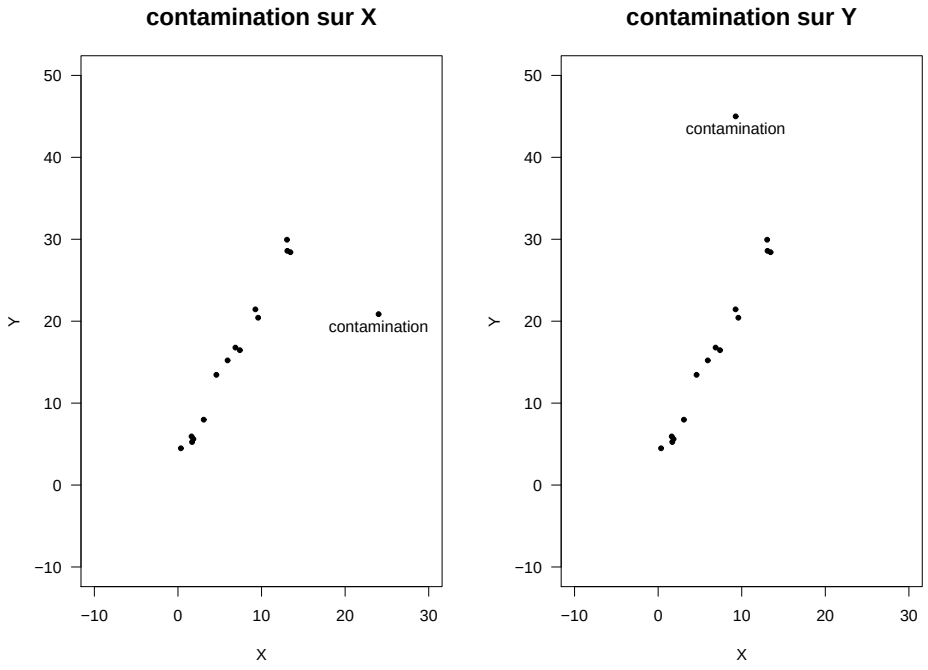
- $\hat{\boldsymbol{\beta}}_{\boldsymbol{MLE}}$ est l'estimateur de maximum de vraisemblance qui est la solution décrite par les équations de score (2.14). Il ne possède pas de propriétés robustes et est censé être fortement influencé par les données extrêmes.
- $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ est l'estimateur robuste qui résout le système d'équations (2.21). On utilisera une fonction ψ de Huber avec $c = 1.345$. Deux versions de cet estimateur seront produites ; l'une avec les poids $w(\boldsymbol{x_i}) = 1$ et l'autre avec des poids $w(\boldsymbol{x_i}) = \sqrt{1 - h_{ii}}$, où h_{ii} sont les éléments de la diagonale de la matrice $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$. On utilisera les abréviations, respectivement, $\hat{\boldsymbol{\beta}}_{\boldsymbol{R1}}$ et $\hat{\boldsymbol{\beta}}_{\boldsymbol{RH}}$ pour décrire ces deux paramétrisations différentes. L'estimation se fera grâce à la fonction `glmrob()` fournie par la librairie `robustbase` (Rousseeuw *et al.*, 2012).
- $\hat{\boldsymbol{\beta}}_{\boldsymbol{B}}$ est l'estimateur bayésien robuste. Les deux résolutions du problème seront traitées, soit l'approximation analytique décrite par (2.29) et la solution de (2.26) par un algorithme de Newton. De ces deux estimateurs, il est possible d'en produire quatre en utilisant des poids $w(\boldsymbol{x_i})$ différents, comme décrit précédemment. Ainsi $\hat{\boldsymbol{\beta}}_{\boldsymbol{BA1}}$ désigne l'estimateur par approximation analytique avec des poids unitaires, $\hat{\boldsymbol{\beta}}_{\boldsymbol{BAH}}$ désigne aussi l'approximation analytique mais avec des poids basées sur $\boldsymbol{H}$ tandis que $\hat{\boldsymbol{\beta}}_{\boldsymbol{BN1}}$ et $\hat{\boldsymbol{\beta}}_{\boldsymbol{BNH}}$ font référence à la résolution par la méthode de Newton, avec des poids unitaires et basés sur $\boldsymbol{H}$ respectivement. L'espérance de la distribution a priori est estimée de la manière décrite dans le chapitre 2.5. On prendra comme paramètre $c_2 = 2$, c'est en effet le calibrage utilisé par Genton et Ronchetti (2008). Le paramètre δ^2, la variance de la distribution a priori sera calibrée par simulation afin de choisir une valeur optimale (voire chapitre 3.4.3.1).

3.2 La modélisation des données extrêmes

Le modèle représenté par (3.1) génère des données propres. Il représente donc F_θ de l'équation (1.1). Il est nécessaire donc de choisir une contamination G du modèle ainsi que la proportion ϵ de données contaminées. Cette dernière a été fixée arbitrairement tel que $\epsilon = 0.1$. Le problème lorsqu'il s'agit de modéliser les données extrêmes est qu'il y a une multitude de manières de proposer des données aberrantes. Il a donc fallu faire le choix de la forme de cette contamination. Deux types de contamination ont été considérés et une troisième en découle, en combinant les deux premières. La première contamination porte sur les variables explicatives. Il s'agit de les transformer, pour une partie des données, sans pour autant changer la variable de réponse. La deuxième va transformer la variable de réponse en créant ainsi une aberration entre le prédicteur linéaire et la réponse. La figure 3.1 en montre l'idée générale (dans le cadre d'une régression unidimensionnelle). Dans le cas de la

contamination sur $\boldsymbol{X}$, la donnée extrême l'est seulement sur la distribution marginale des $\boldsymbol{X}$. Et ainsi en ne connaissant que la distribution marginale de $\boldsymbol{Y}$, il n'est pas possible de la désigner comme étant une donnée extrême. Dans le cas d'une donnée contaminée sur $\boldsymbol{Y}$, c'est l'inverse.
Nous seront intéressés aussi par le cas de données *propres*. Ce sont des données générées uniquement par le modèle décrit par l'équation (3.1) . C'est le cas où $G = F_\theta$, ou alors $\epsilon = 0$. Ce processus est nommé G_0 [1] dans la suite du travail.

FIGURE 3.1 – Exemple de contamination. Les données extrêmes modifient le support de la fonction selon les variables explicatives ou la variable réponse.



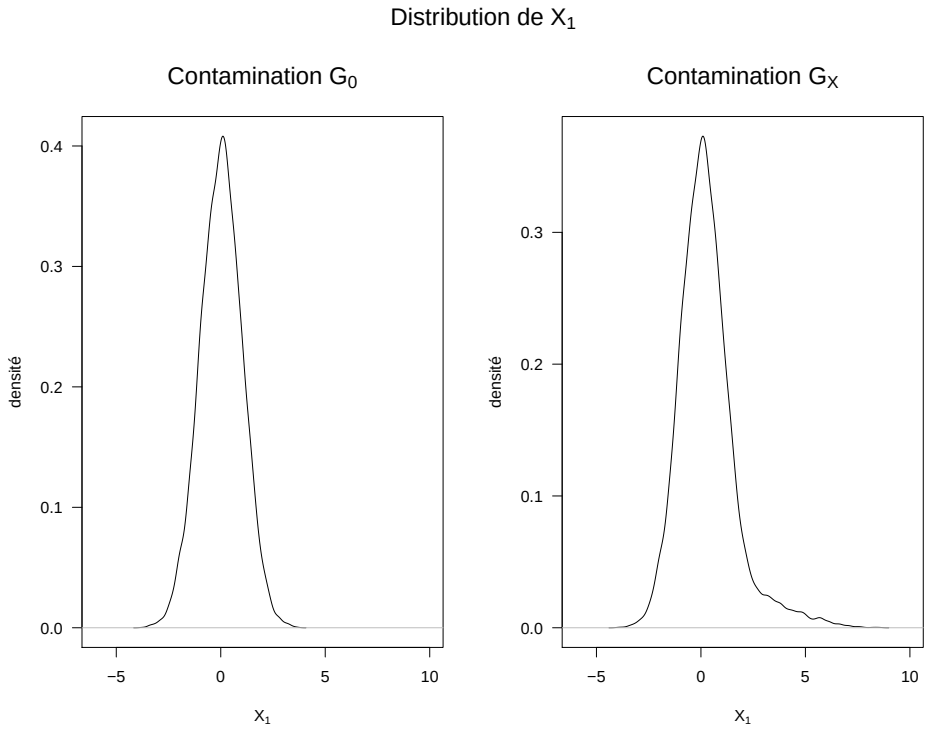
3.2.1 Contamination sur $\boldsymbol{X}$

La contamination sur $\boldsymbol{X}$ va ajouter une perturbation sur la variable X_1, la loi normale. Ainsi, il s'agit de construire un modèle propre en simulant $\boldsymbol{X}$ et $\boldsymbol{Y}$ à partir de $\boldsymbol{X}$, comme décrit par l'équation (3.1). Ensuite, on remplace $\boldsymbol{X} = [\mathbf{1} \quad \boldsymbol{X_1} \quad \boldsymbol{X_2} \quad \boldsymbol{X_3}]$ par $\tilde{\boldsymbol{X}} = [\mathbf{1} \quad \tilde{\boldsymbol{X}}_\mathbf{1} \quad \boldsymbol{X_2} \quad \boldsymbol{X_3}]$. Ainsi une proportion ϵ des données de $\boldsymbol{X_1}$ est modifiée de telle manière que $\tilde{x}_{\epsilon 1} = ax_{\epsilon 1} + b$, ce

1. Les dénominations G_0, G_X, G_Y et G_{XY} seront utilisées abusivement afin de faciliter la lecture de l'analyse. En effet, elles pourront faire référence à la fois au processus de contamination et à l'échantillon ayant subit la contamination. Cependant le contexte permettra au lecteur, sans aucun doute, de comprendre la référence.

qui va aboutir aux observations contaminées $\tilde{\boldsymbol{X}}_{\boldsymbol{1}}$. Ce type de contamination sera appelé G_X.
Le graphique 3.2 montre la distribution de X_1 sous la paramétrisation expliquée dans le chapitre 3.2.4. On remarque que la contamination est minime. Seule une légère asymétrie apparaît. L'effet de cette contamination réside plus dans le fait que la relation entre le prédicteur linéaire et les variables explicatives n'est plus linéaire que dans l'écart à la normalité de X_1.

FIGURE 3.2 – Distribution de la variable explicative X_1 avant et après la contamination.



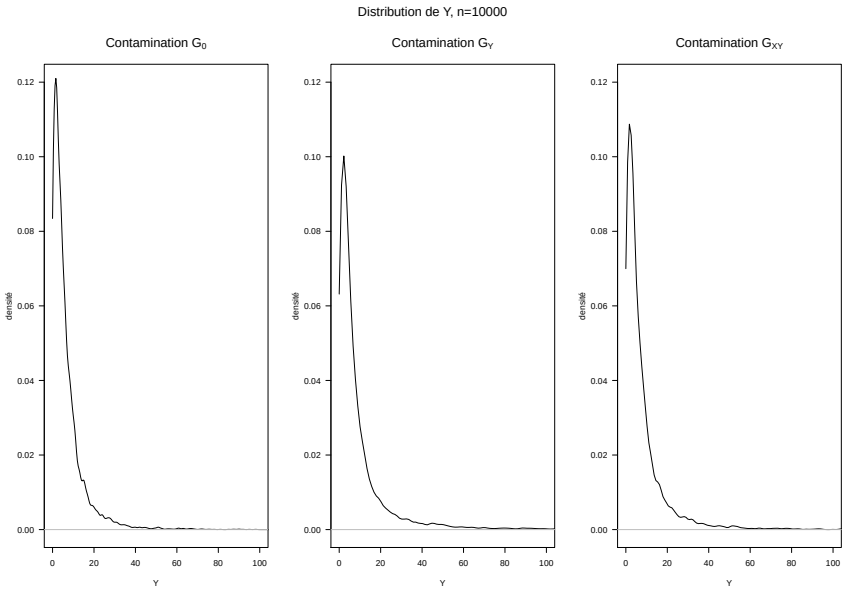
3.2.2 Contamination sur $\boldsymbol{Y}$

Pour les GLM, une contamination sur $\boldsymbol{Y}$ ne peut pas se faire simplement en transformant la variable de réponse. En effet, étant donnée qu'ils couvrent un large nombre de distributions, la contamination se devait d'être compatible pour l'ensemble de la famille exponentielle. Et une transformation de $\boldsymbol{Y}$ en additionnant un réel ne peut pas être appliquée à une distribution Bernoulli. Ainsi dans le cadre des GLM, il est plus judicieux de transformer le prédicteur linéaire. Ainsi, après avoir calculé $\eta_i = \boldsymbol{x_i}^\top \boldsymbol{\beta}$, on transforme une proportion ϵ des prédicteurs linéaires $\tilde{\eta}_\epsilon = \eta_\epsilon + h$. La variables de réponse est simulée $E[\tilde{\boldsymbol{Y}}_{\boldsymbol{\epsilon}}] = g^{-1}(\tilde{\eta}_\epsilon)$ pour les données contaminées et utilise (3.1) pour les données non contaminées. Cette contamination aura l'abréviation G_Y.

3.2.3 Contamination sur X et Y

Ce troisième type de contamination combine les deux premières. Il y a donc une proportion ϵ_X de données qui subissent G_X, une proportion ϵ_Y de données qui subissent G_Y et une proportion ϵ_{XY} qui subissent les deux contaminations de tel sorte que $\epsilon = \epsilon_X + \epsilon_Y + \epsilon_{XY}$. Afin de garantir une certaine quantité de contamination conjointe on a préféré fixer $\epsilon_{XY} = 0.4$ relativement grand. ϵ_X et ϵ_Y ont la même importance donc $\epsilon_X = \epsilon_Y = 0.3$. Cette contamination sera nommée G_{XY}. Le graphique 3.3 montre la densité de la la variable Y selon le modèle décrit par l'équation (3.1). Les processus de contamination G_Y et G_{XY} ont la particularité d'introduire une plus forte densité pour des valeurs extrêmes supérieures à 40.

FIGURE 3.3 – Densité de la variable réponse avant et après contamination.



3.2.4 Paramétrisation des processus de contamination

Les paramètres a, b et h, ainsi que a_{XY}, b_{XY} et h_{XY} pour la contamination G_{XY}, ont été fixés afin de garantir des données extrêmes équivalentes. C'est-à-dire que pour permettre une comparaison entre les différentes contaminations, il faut qu'elles aient une même intensité. Étant donné qu'aucune stratégie ne semblait évidente afin que les contaminations aient un effet comparable, les paramètres ont été choisis tel que $a = 2$, $b = 2.8$, $h = 2$, $a_{XY} = 1$, $b_{XY} = 2.3$ et $h_{XY} = 1.6$, après plusieurs essais et selon le bon sens. Une fois de plus, la contamination se doit de ne pas être trop importante afin de garantir la stabilité numérique des estimations.

3.3 Les prévisions

La prévision est utile dans le cas où l'on connait le modèle et les variables explicatives d'un individu mais que la variable réponse n'est pas observée. Cependant dans le cadre de cette analyse, afin de pouvoir juger la qualité de la prévision, on simulera la valeur de la variable réponse. Les prévisions seront donc faite en supposant les variables réponse inconnues et elles seront analysées en les comparant à la vraie valeur de la réponse.
Il s'agit de faire des prévisions sur un nouvel échantillon de taille $n_{new} = 50$. Les échantillons qui serviront aux prévisions peuvent, eux aussi, être issus de données contaminées. Ainsi, selon notre construction de données extrêmes, seize cas peuvent être analysés ; (4 types d'échantillons pour l'estimation) $\times$ (4 types d'échantillons pour la prédiction). Cependant, seuls deux types de combinaisons (qui dénombrent au total sept cas parmi les seize possibles) semblent dignes d'intérêt et seront traités. Premièrement, les échantillons pour l'estimation et pour la prédiction sont issus de la même distribution. C'est le cas que l'on rencontrerait pour des *vraies* données. En second, on analysera le cas où les échantillons de prévisions viennent de G_0. On a ainsi deux fois quatre cas moins un (quand les deux échantillons proviennent de G_0), soit sept cas possibles par estimateur. Le tableau ci-dessous résume les choix des types de prévisions. Les colonnes indiquent la provenance de l'échantillon d'estimation et les lignes celles de l'échantillon de prévision.

TABLE 3.1 – Combinaison entre échantillon d'estimation et échantillon de prévision analysées

	Estimation G_0	Estimation G_X	Estimation G_Y	Estimation G_{XY}
Prévision G_0	X	X	X	X
Prévision G_X		X		
Prévision G_Y			X	
Prévision G_{XY}				X

Il est ainsi possible de créer des variables explicatives $\boldsymbol{x_{new,i}}$, $i = 1, \ldots, n_{new}$. Il y a donc en tout trois jeux de données $\boldsymbol{x_{new,i}}$; un pour une contamination G_{XY}, un pour G_X et un autre qui peut être utilisé à la fois pour la contamination G_Y et G_0.
Il est possible à partir du jeu de données $\boldsymbol{x_{new,i}}$ de générer les espérances $\mu_{new,i}$ et les variables de réponse $y_{new,i}$ grâce aux modèles décris précédemment. Et similairement on dénombre trois types d'espérance et de variable de réponse ; un pour une contamination G_Y, un pour une contamination G_{XY}, et le même peut être utilisé pour les contamination G_X et G_0. Il est important de noter que le même échantillon servant aux prévisions sera réutilisé

à chaque nouvelle itération de l'analyse Monte-Carlo. Cela a pour avantage de générer une estimation de la distribution des prévisions.
Les prévisions faites grâce aux échantillons de variables explicatives $\boldsymbol{x_{new,i}}$ ont ainsi la forme :

$$\hat{\mu}_i = g^{-1}(\boldsymbol{x_{new,i}}^\top \hat{\boldsymbol{\beta}}), \tag{3.2}$$

où $\boldsymbol{x_{new,i}}$ sont les variables explicatives de l'individu i, $\hat{\boldsymbol{\beta}}$ est l'estimateur produit grâce à l'une des méthodes décrites dans le chapitre 3.1.1 et $g(\cdot)$ est la fonction lien. L'analyse des prévisions $\hat{\mu}_i$ se fera en les comparant avec la variable réponse simulée $y_{new,i}$ mais aussi en comparant les différentes prévisions obtenues grâce aux différentes méthodes d'estimation des $\boldsymbol{\beta}$.

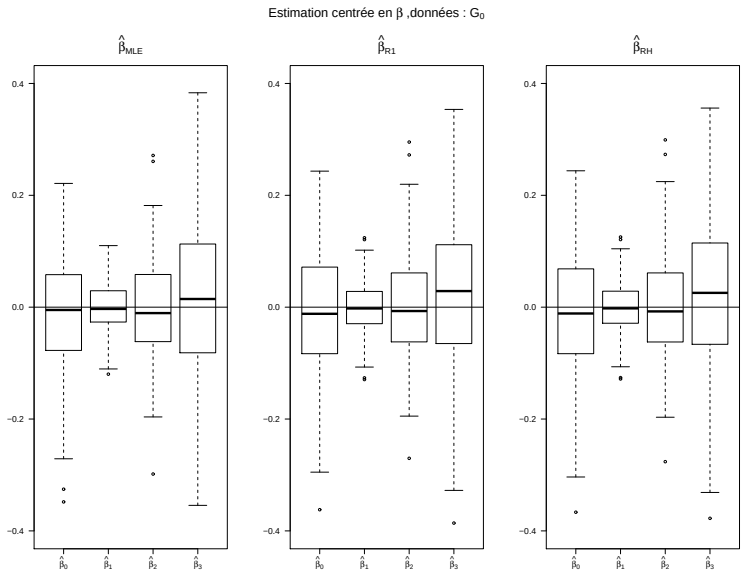
3.4 Les résultats

3.4.1 $\hat{\boldsymbol{\beta}}_{\boldsymbol{MLE}}$ face à des données extrêmes

L'estimation par maximum de vraisemblance donne de bons résultats lorsque le modèle est bien spécifié. On retrouve dans le graphique de gauche de la figure [2] 3.4, les résultats de l'estimateur de maximum de vraisemblance sur des données *propres*. Dans ce cas, l'estimateur $\hat{\boldsymbol{\beta}}_{\boldsymbol{MLE}}$ fait de bonnes performances et elles sont comparables aux estimateurs de type $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$. La variance est relativement faible et on remarque la présence d'un biais résiduel dû à la taille relativement faible de l'échantillon.

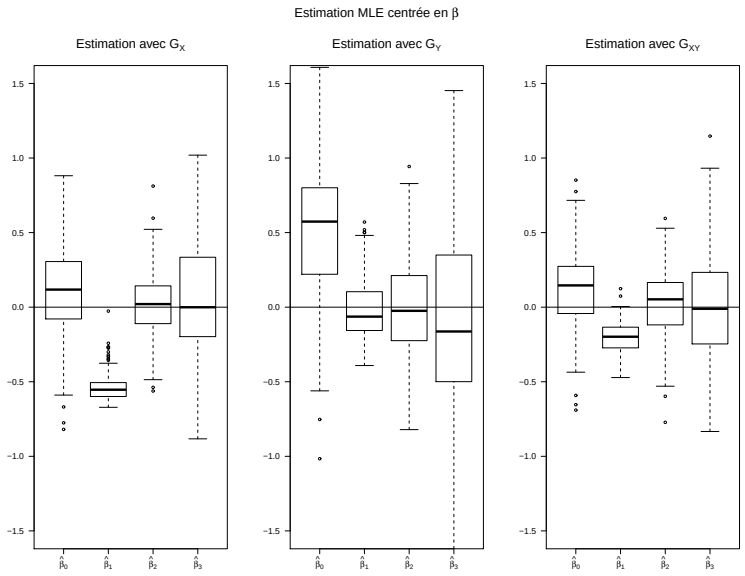
2. Afin de faciliter la lecture et l'analyse des résultats, les estimés sont présentés centrés sur leur vraie valeur. Cela a l'avantage de mettre en évidence ce qui est le plus important - c'est-à-dire le biais et la variance- tout en supprimant une information moins utile pour l'analyse.

FIGURE 3.4



Lorsque le modèle est mal spécifié, l'estimation par maximum de vraisemblance est à éviter. Le graphique 3.5 monte les performances de l'estimateur $\hat{\boldsymbol{\beta}}_{\boldsymbol{MLE}}$ lors des contaminations G_X, G_Y et G_{XY}. On remarque que les biais explosent et que les dispersions des paramètres sont très importantes. On remarque aussi que les différentes contaminations provoquent des biais différents sur les estimateurs. G_X provoque un biais important sur $\boldsymbol{\beta_1}$ et un moindre sur $\boldsymbol{\beta_0}$, G_Y a surtout un effet indésirable sur $\boldsymbol{\beta_0}$, et, sans surprise, G_{XY} est un mélange des biais induits par G_X et G_Y. C'est évidemment un estimateur à éviter lorsqu'il y a des données extrêmes.

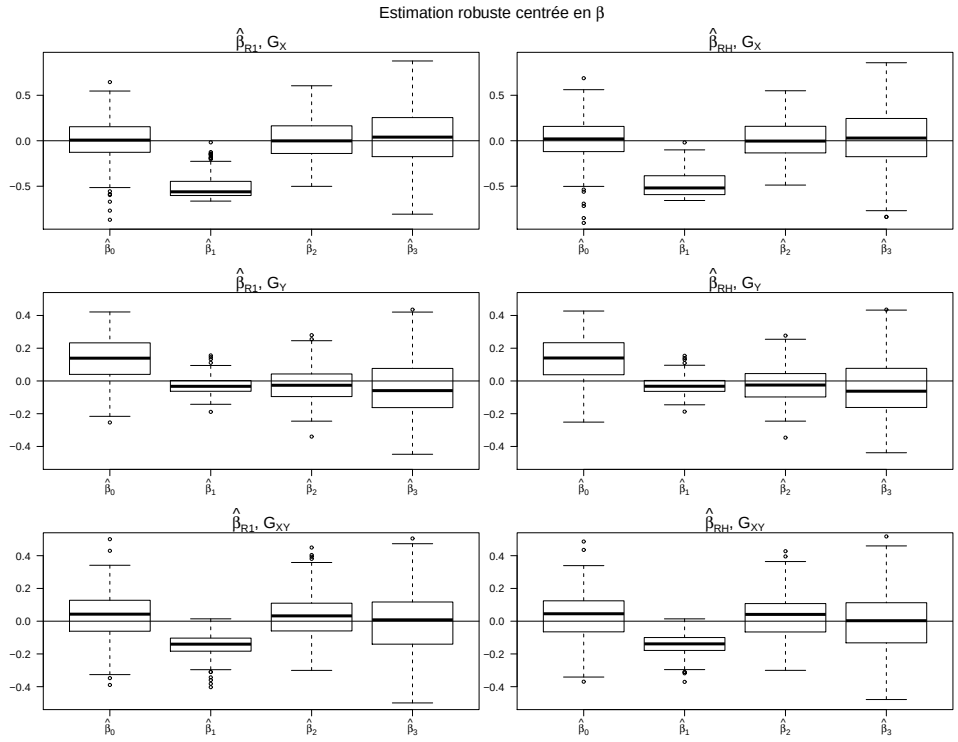
FIGURE 3.5



3.4.2 Les performances de l'estimation robuste

L'estimation robuste fait de bonnes performances lorsqu'il s'agit d'une estimation à partir de données non contaminées. Les graphiques de droites de la figure 3.4 montre que l'on a un résultat comparable à l'estimation par maximum de vraisemblance. Les distributions semblent identiques et l'augmentation de la variabilité expliquée par l'équation (2.24) se voit notamment sur l'estimation de β_0. De plus, l'estimateur robuste va faire une meilleure performance lorsqu'il s'agit de données contaminées. La figure 3.6 montre les estimations $\hat{\boldsymbol{\beta}}_{\boldsymbol{R1}}$ et $\hat{\boldsymbol{\beta}}_{\boldsymbol{RH}}$ lorsque les données sont contaminées. Bien que les boxplots ne sont pas tous centrés en $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}} - \boldsymbol{\beta} = \mathbf{0}$, les biais et les dispersions sont fortement réduits par rapport à l'estimateur $\hat{\boldsymbol{\beta}}_{\boldsymbol{MLE}}$. Théoriquement cela est expliqué par l'équation (1.4) qui montre qu'un estimateur robuste à un biais non nul mais borné.
On remarque toutefois que l'apport des poids $w(\boldsymbol{x_i})$ basés sur la diagonale de la matrice $\boldsymbol{H}$ est quasi indiscernable. On aurait pu s'attendre à une performance légèrement meilleure pour $\hat{\boldsymbol{\beta}}_{\boldsymbol{RH}}$ dans le cas de la contamination G_X, mais ce n'est pas le cas. Les poids en fonction de la distance de Mahalanobis ont été testés mais posaient des problèmes récurrents de convergence.

FIGURE 3.6

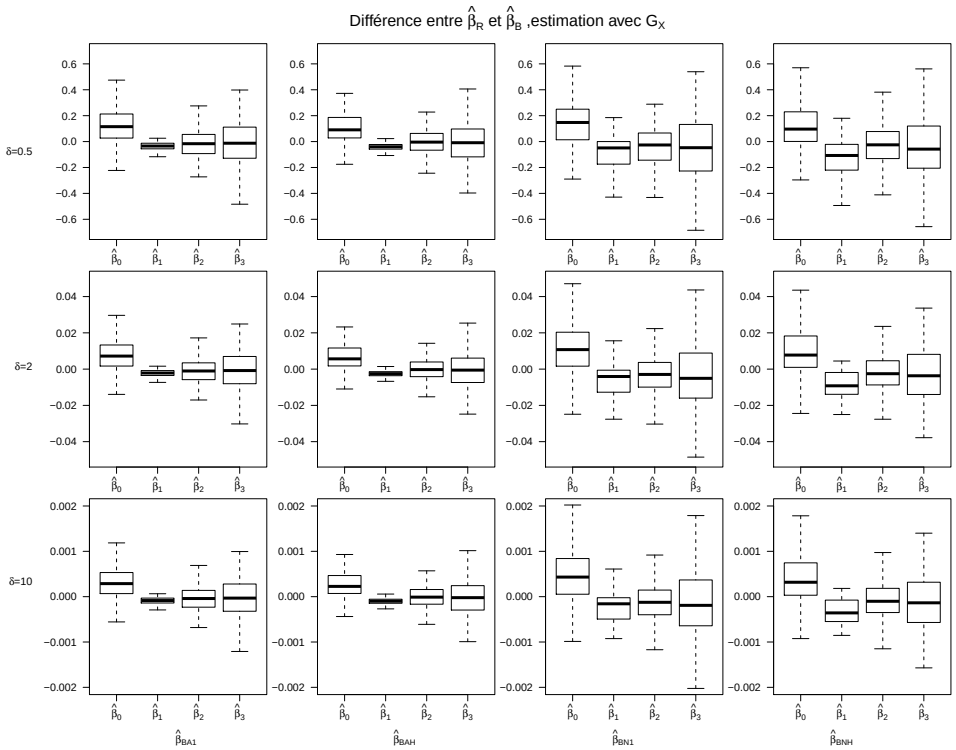


3.4.3 Les performances de l'estimation bayésienne robuste

3.4.3.1 Calibrage de δ^2

Le paramètre δ^2 permet d'influencer la correction apportée par les estimateurs $\hat{\boldsymbol{\beta}}_{\boldsymbol{B}}$. Quand δ^2 est grand cette correction est faible car $\hat{\boldsymbol{\beta}}_{\boldsymbol{B}} \to \hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ quand $\delta \to \infty$. La figure 3.7 montre cette propriété pour une estimation avec un échantillon G_X. En effet la différence $\hat{\boldsymbol{\beta}}_{\boldsymbol{B}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ est déjà infime quand $\delta^2 = 10$. On remarque que la correction qu'introduit $\hat{\boldsymbol{\beta}}_{\boldsymbol{B}}$ est toujours présente mais son intensité devient faible quand $\delta^2 \to \infty$; c'est la raison pour laquelle les boxplots se ressemblent sur une même colonne mais que l'échelle change. En annexe B.1 se trouve le même type de boxplot pour les estimations sur des échantillons G_0, G_Y et G_{XY}.

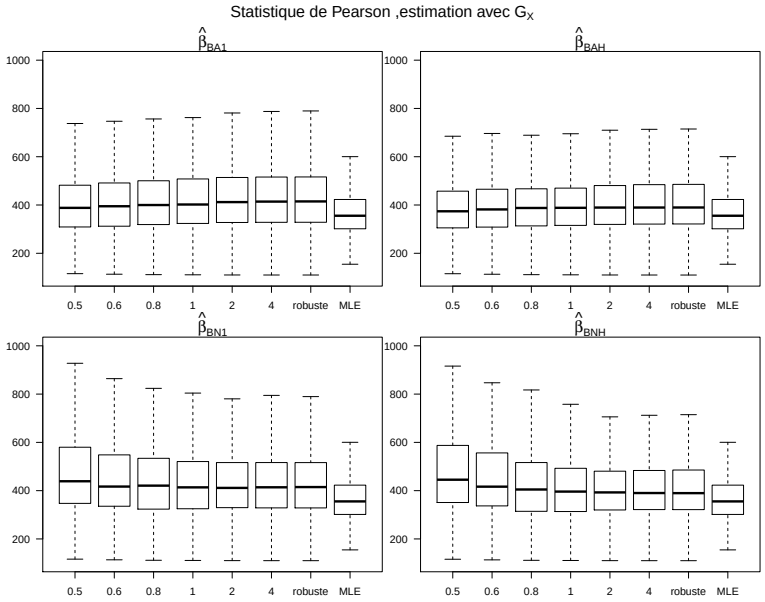
FIGURE 3.7



Au vu de cette propriété, il est donc nécessaire de choisir un δ^2 relativement petit afin qu'il apporte un réel changement par rapport à $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$. Il est cependant nécessaire d'avoir un critère plus précis afin de faire une sélection parmi tous les δ petits.
La statistique de Pearson, $\chi^2 = \sum_{i=1}^{n}[r_{Pi}]^2$, où r_{Pi} sont les résidus de Pearson définis par l'équation (2.19), permet d'avoir une mesure de la qualité du modèle. Elle a été préférée car elle ne dépend pas de la quasi-vraisemblance qui n'a pas été calculée pour les estimateurs bayésiens et elle permet une standardisation en fonction de la variance de l'erreur. En effet, dans le cas d'un modèle où la variable de réponse est issue d'une loi de Poisson, on se retrouve dans un cas d'hétéroscédasticité et la simple somme des erreurs de prévisions $\sum_{i=1}^{n}[y_i - \hat{\mu}_i]^2$ ne prend pas en compte ce phénomène. La statistique χ^2 est donc un critère qui permet de faire une sélection parmi les δ^2 petits ; elle est évaluée sur l'échantillon d'estimation, on peut donc calibrer δ avant de produire les prévisions. Il est important de noter que la statistique de Pearson n'est pas parfaite. En effet, dans le cas d'un modèle de Poisson et pour une erreur de prévision identique, la statistique de Pearson sera plus faible pour une "sur-prévision" plutôt qu'une "sous-prévision" ; cela est dû à la forme des résidus de Pearson décrits par l'équation (2.19).
Les boxplots de la figure 3.8 montrent que l'effet d'un δ^2 petit réduit la

statistique de Pearson pour le cas d'une contamination en G_X. Ce type de comportement est le même pour les autres contaminations (Annexe B.2). Cependant, un problème de convergence de l'algorithme de Newton utilisé lors du calcul de l'estimateur bayésien par résolution numérique $\hat{\boldsymbol{\beta}}_{\boldsymbol{BN}}$ empêche le choix d'un δ^2 trop faible ($\delta < 0.5$). De plus,on remarque que quelque soit le type de contamination, un δ trop faible a une influence négative sur la statistique de Pearson dans le cas des estimateurs bayésiens $\hat{\boldsymbol{\beta}}_{\boldsymbol{BN1}}$ et $\hat{\boldsymbol{\beta}}_{\boldsymbol{BNH}}$. Afin d'avoir une valeur de δ commune, le choix s'est porté sur une valeur relativement faible, comprise entre 0.5 et 1, qui sera utilisé pour la suite de l'analyse. La conclusion de paramétrer $\delta = 0.6$ s'est donc imposée.

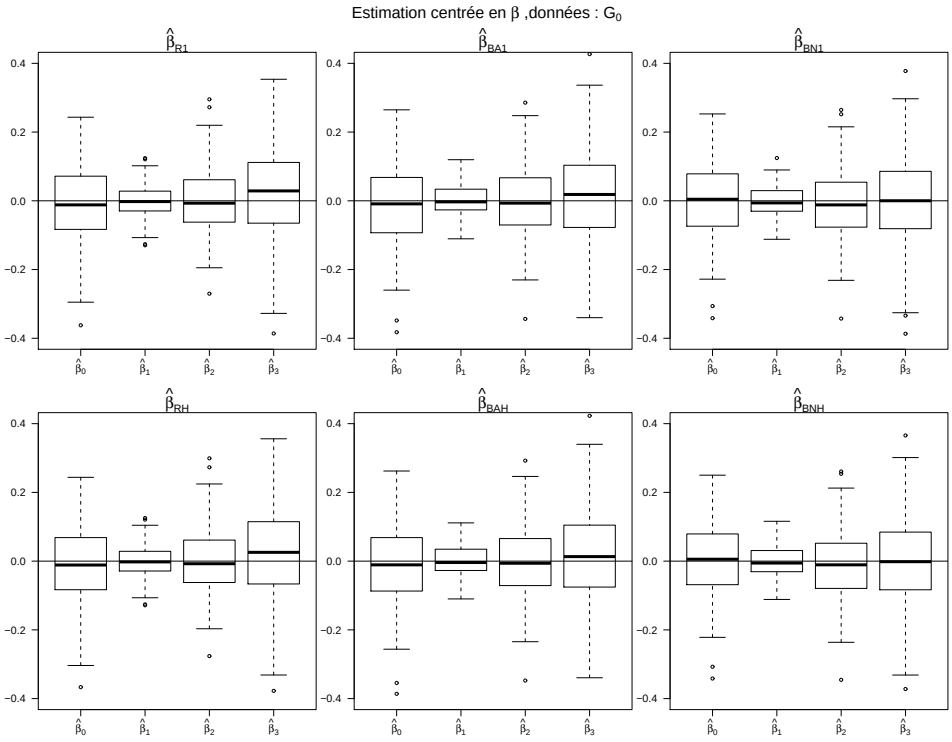
FIGURE 3.8



3.4.3.2 Les estimées de $\hat{\boldsymbol{\beta}}_{\boldsymbol{B}}$

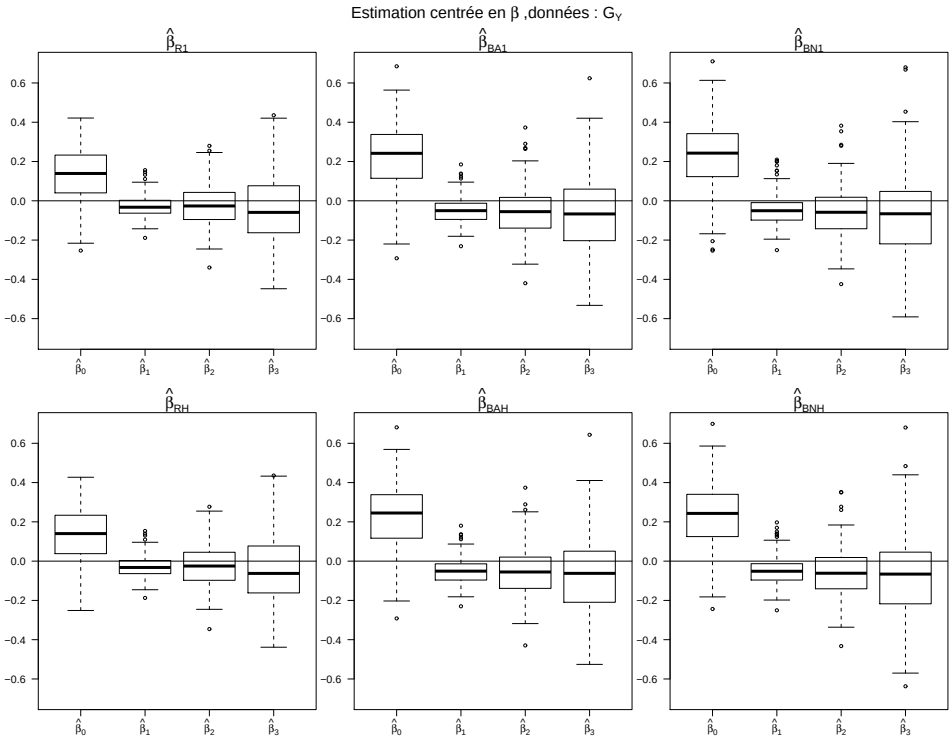
En ayant des données *propres*, les estimateurs $\hat{\boldsymbol{\beta}}_{\boldsymbol{B}}$ produisent des résultats similaires à l'estimation robuste ordinaire. En effet, la correction qu'apporte l'hypothèse bayésienne est en partie dépendante de l'estimation de $\boldsymbol{\mu_p}$ (chapitre 2.5). Or l'importance de $\hat{\boldsymbol{\mu}}_{\boldsymbol{p}}$ est fortement liée à la présence de données extrêmes. Dans le cadre d'une estimation sur des données provenant de G_0, il est donc évident que cette correction sera plus faible. Les boxplots de la figure 3.9 montrent les estimations des coefficients en soustrayant la vraie valeur de $\boldsymbol{\beta}$. On remarque donc que quelque soit la méthode (approximation $\hat{\boldsymbol{\beta}}_{\boldsymbol{BA}}$ analytique ou résolution numérique $\hat{\boldsymbol{\beta}}_{\boldsymbol{BN}}$), la distribution des estimations est similaire à celle produite par l'estimation robuste ordinaire.

FIGURE 3.9



Dans le cas d'une estimation avec des données contaminées la correction est plus importante. La figure 3.10 montre les estimations centrées de $\hat{\boldsymbol{\beta}}_{\boldsymbol{B}}$ avec des données provenant de G_Y. On remarque l'apparition d'un biais plus important pour les estimations des coefficients $\hat{\boldsymbol{\beta}}$. Lors de prévisions ce biais est censé agir comme une compensation à la présence de données extrêmes. On remarque toutefois que les différences entre les estimateurs $\hat{\boldsymbol{\beta}}_{\boldsymbol{BA}}$ et $\hat{\boldsymbol{\beta}}_{\boldsymbol{BN}}$ sont très faibles.

FIGURE 3.10



On retrouve en annexe B.3, les estimations des coefficients pour des données provenant de G_X et G_{XY}. Un comportement similaire est visible pour ces contaminations.

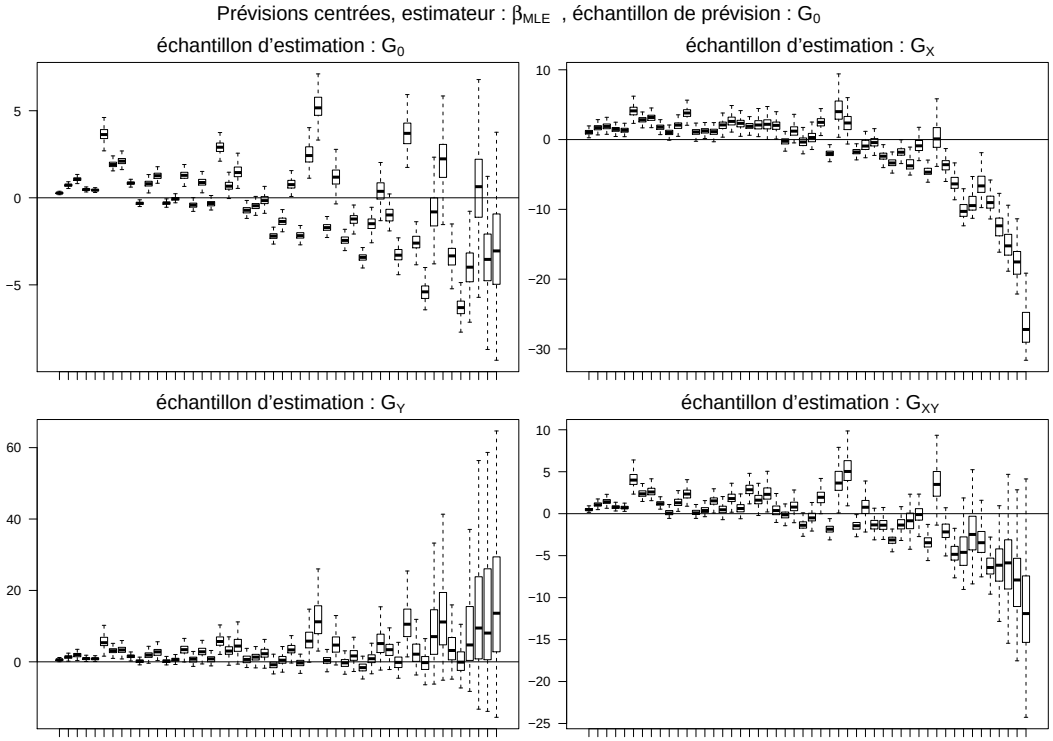
3.4.4 Les prévisions

3.4.4.1 Prévoir des données propres

Dans ce chapitre, les prévisions de données provenant de G_0 seront abordées. Cependant, il est possible d'analyser des prévisions de données propres en ayant des estimateurs évalués sur des données contaminées G_X, G_Y et G_{XY}. Dans la figure 3.11, on peut voir les prévisions centrées en $y_{new,i}$ triées selon $y_{new,i}$. Ainsi ce graphique s'interprète de la manière suivante. Une prévision centrée sur zéro est optimale, une valeur positive correspond à une prévision trop importante et une négative à une prévision trop faible. De plus, la vraie valeur de $y_{new,i}$ sera toujours plus faible (ou égale) à celle de son voisin de droite sans qu'il y ait cependant une échelle mesurable. Les quatre graphiques correspondent aux quatre différents types de contaminations utilisées pour l'estimation. On remarque la tendance générale aux quatre graphiques, c'est-à-dire que la variance des prévisions augmente selon la valeur de $y_{new,i}$; cela reflète la propriété hétéroscedastique des modèles de Poisson. L'estimateur

de maximum de vraisemblance produit de relativement bonnes prévisions lorsqu'on l'utilise avec un échantillon propre. Les biais sont faibles et les variances également. Cependant, lorsqu'on l'utilise avec un échantillon d'estimation contaminées, les prévisions ont un biais et une variance importante. On remarque aussi les effets différents que produisent les contaminations. La contamination G_X sur l'échantillon d'estimation produit des prévisions relativement bonnes en terme de variance, mais pour les fortes valeurs de $y_{new,i}$ les prévisions seront sous-estimées. Dans le cas d'une contamination G_Y, la dégradation des prévisions se situe dans la variabilité de la prévision des grandes valeurs de $y_{new,i}$. Et une contamination G_{XY} sur l'échantillon d'estimation, produit des prévisions sur un échantillon G_0 avec un mélange des problèmes rencontrés sur les deux autres types de contamination.

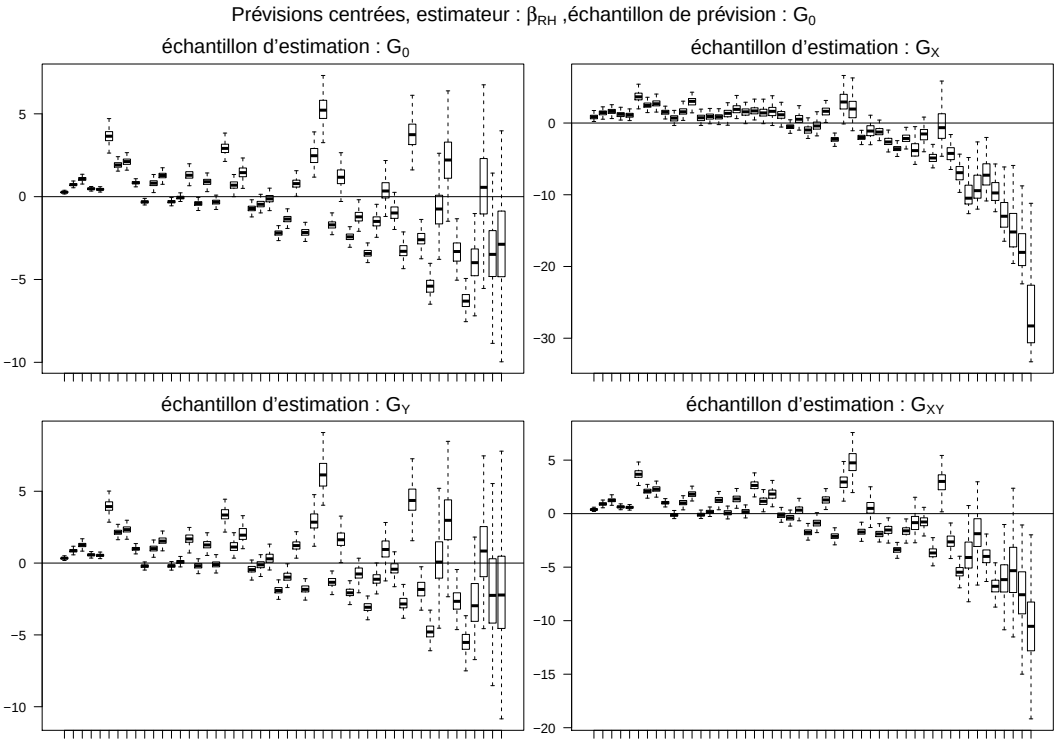
FIGURE 3.11



L'estimateur robuste $\boldsymbol{\beta_R}$ améliore nettement les prévisions lorsque l'on contamine l'échantillon d'estimation par un processus G_Y. On remarque la forte réduction du biais et de la variance des prévisions dans la figure 3.12. Ainsi cet estimateur arrive à contenir le biais et la variance des prévisions dans des limites raisonnables lorsqu'il s'agit d'une contamination G_Y. Cependant les prévisions faites suite à une contamination en G_X de l'échantillon d'estimation reste autant pauvre que celles faites avec un estimateur de maximum

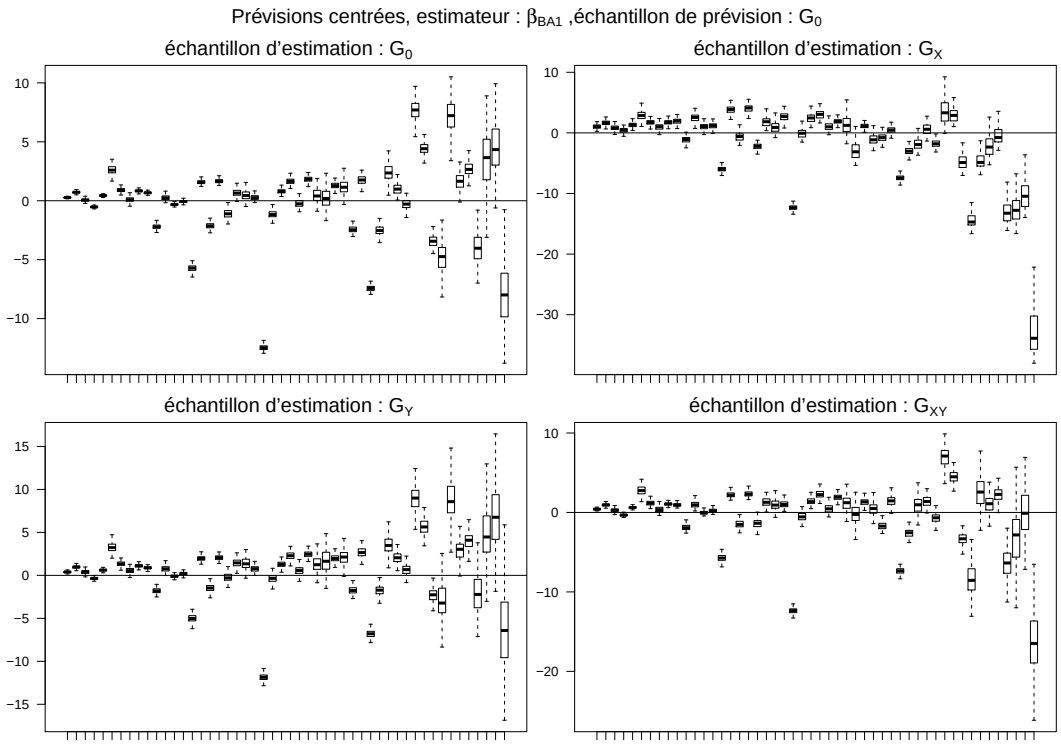
de vraisemblance. Dans ce cas, les deux figures 3.11 et 3.12 sont comparables bien qu'il s'agisse de l'estimateur robuste avec des poids $w(\boldsymbol{x_i})$ basés sur la diagonale de la matrice $\boldsymbol{H}$. Il est possible de voir les prévisions faites par $\hat{\boldsymbol{\beta}}_{\boldsymbol{R1}}$ dans l'annexe B.4 et on ne peut pas conclure que l'ajout de poids $w(\boldsymbol{x_i})$ en fonction de $\boldsymbol{H}$ améliore réellement les prévisions dans le cas d'une contamination G_X. L'estimateur robuste améliore aussi légèrement les prévisions faites sur un échantillon G_0 dans le cas ou l'échantillon d'estimation est issu de G_{XY}. Cependant cela est faible et c'est une conséquence du bon comportement des estimateurs robustes face à une contamination G_Y.

FIGURE 3.12



La figure 3.13 représente les prévisions faites grâce à l'estimateur $\hat{\boldsymbol{\beta}}_{\boldsymbol{BA1}}$. On remarque que les estimateurs bayésiens produisent des prévisions qui semblent moins dégradées lorsque l'échantillon d'estimation a subi une contamination G_X. Dans les autres types de contamination, il est difficile de conclure qu'il est plus performant pour faire des prévisions. Cependant, la dégradation des prévisions en fonction $y_{new,i}$ que l'on remarque dans le cas des estimateurs robustes est moins présente. C'est-à-dire que les prévisions pour des faibles valeurs de $y_{new,i}$ seront meilleures en utilisant un estimateur robuste, alors qu'un estimateur bayésien semble plus performant lorsqu'il s'agit de faire des prévisions lorsque $y_{new,i}$ est grand.

FIGURE 3.13

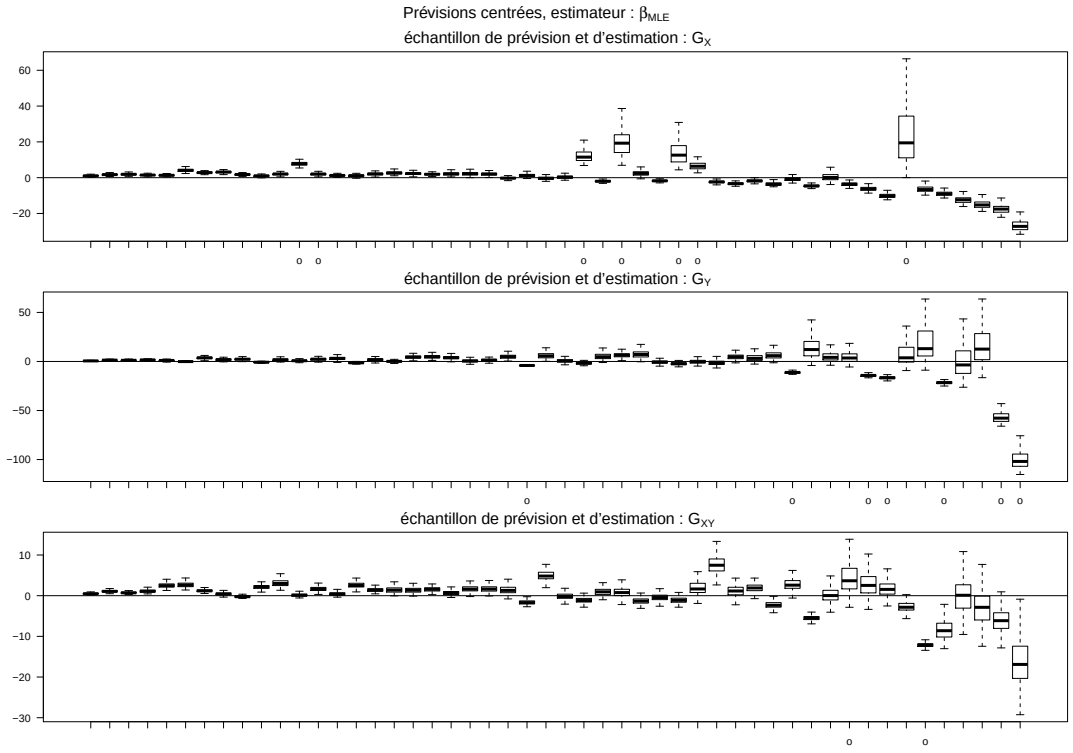


3.4.4.2 Prévoir des données contaminées

Précédemment seules des prévisions provenant d'une distribution non contaminée ont été analysées. Dans cette section, il s'agit de faire des prévisions d'échantillons contaminées. Les cas où les échantillons d'estimation et de prévision sont issus de la même distribution seront étudiés. Les résultats seront présentés comme auparavant, par une série de boxplots qui représentent chacun une donnée à prévoir. On aura soustrait $y_{new,i}$ aux prévisions afin de les représenter centrées en zéro. Elles seront aussi classées par ordre croissant selon $y_{new,i}$ et la lettre "o" indique si la donnée à prévoir à subie une transformation G_X, G_Y ou G_{XY}. Ainsi on peut s'attendre, selon la construction des données extrêmes, que les données extrêmes pour un processus de contamination G_Y auront tendance à être associées à des valeurs de la variable réponse grandes, tandis qu'une donnée extrême d'un processus G_X est indépendante de la valeur de $y_{new,i}$. Cela aura pour effet, dans le cas G_Y, d'avoir une concentration importante de données contaminées ayant une grande valeur de $y_{new,i}$.
La figure 3.14 montre les prévisions faites grâce à $\hat{\boldsymbol{\beta}}_{\boldsymbol{MLE}}$. Dans les cas G_X et G_Y, la qualité de la prévision est fortement dépendante du faite que la donnée soit contaminée ou non. Les prévisions où le biais est important et où

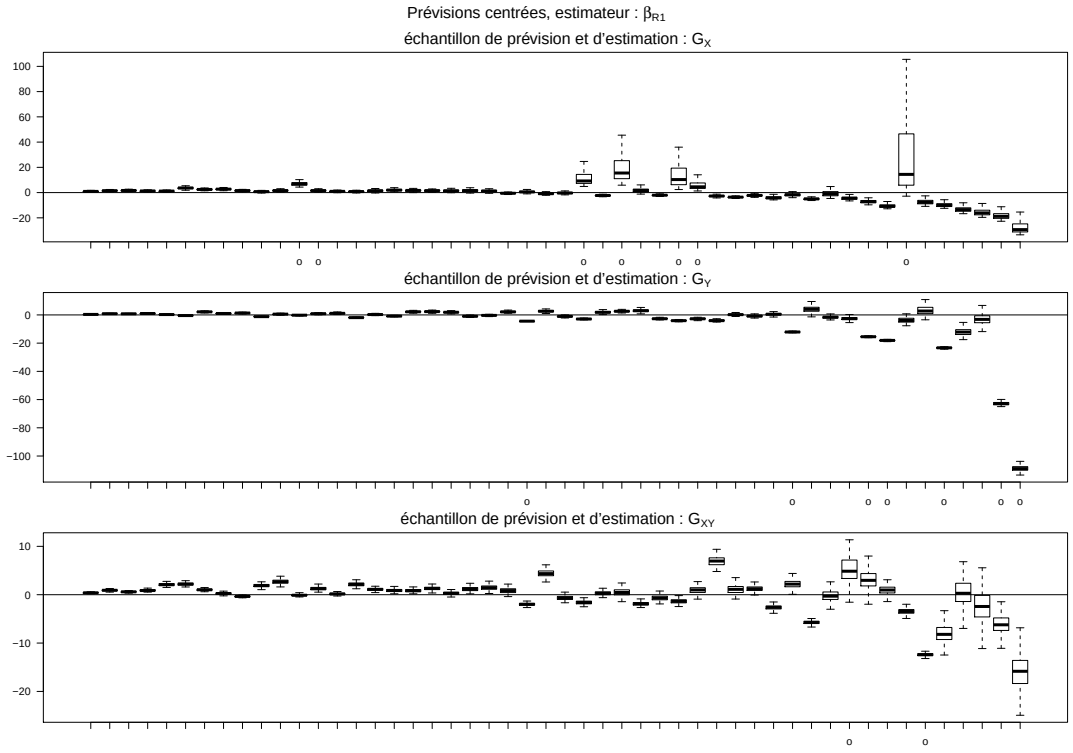
la variance est forte sont régulièrement associées à des données contaminées.

FIGURE 3.14



La figure 3.15 montre les prévisions faites grâce à l'estimateur robuste $\hat{\boldsymbol{\beta}}_{\boldsymbol{R1}}$. On remarque comme pour le cas d'une estimation par maximum de vraisemblance que les prévisions les plus mauvaises sont celles des données contaminées. Dans le cas de la contamination G_Y les prévisions sont de meilleures qualités en terme de variance mais il reste des biais extrêmement importants pour les données contaminées indiquées par "o". De plus, dans le cas d'une contamination G_X, les prévisions sont semblables pour la majorité des données que l'on utilise un estimateur robuste ou celui de maximum de vraisemblance. La même observation peut être faite pour les prévisions provenant d'une contamination G_{XY} qui ne présentent pas d'améliorations perceptibles en comparaison à l'estimateur de maximum de vraisemblance. Les prévisions faites grâce à l'estimateur $\hat{\boldsymbol{\beta}}_{\boldsymbol{RH}}$ sont présentées en annexe B.5 et montrent des résultats comparables.

FIGURE 3.15



Les prévisions faites par les estimateurs bayésiens ne présentent pas des différences importantes qu'il s'agisse du cas de l'approximation analytique ($\hat{\boldsymbol{\beta}}_{\boldsymbol{BA1}}$ et $\hat{\boldsymbol{\beta}}_{\boldsymbol{BAH}}$) ou de la résolution numérique ($\hat{\boldsymbol{\beta}}_{\boldsymbol{BN1}}$ et $\hat{\boldsymbol{\beta}}_{\boldsymbol{BNH}}$). Les figures 3.16 et3.17 représentent les prévisions grâce aux estimateurs $\hat{\boldsymbol{\beta}}_{\boldsymbol{BA1}}$ et $\hat{\boldsymbol{\beta}}_{\boldsymbol{BN1}}$ respectivement.
Dans le cas d'une contamination G_Y, les estimateurs bayésiens font de meilleures prévisions par rapport à un estimateur robuste ou de maximum de vraisemblance. En effet, on remarque que les estimateurs bayésiens permettent de contenir les erreurs de prévisions et les variances dans des limites raisonnables dans le cas d'une contamination G_Y, tandis que les estimateurs robustes et de maximum de vraisemblance font de très pauvres prévisions lorsque les données sont extrêmes. Les estimateurs bayésiens arrivent en effet a garantir une différence entre prévisions et vraie valeur inférieur à 20 et cela même pour les données extrêmes.

Les prévisions faites grâces aux estimateurs $\hat{\boldsymbol{\beta}}_{\boldsymbol{BAH}}$ et $\hat{\boldsymbol{\beta}}_{\boldsymbol{BNH}}$ sont présentées en annexe B.5.

FIGURE 3.16

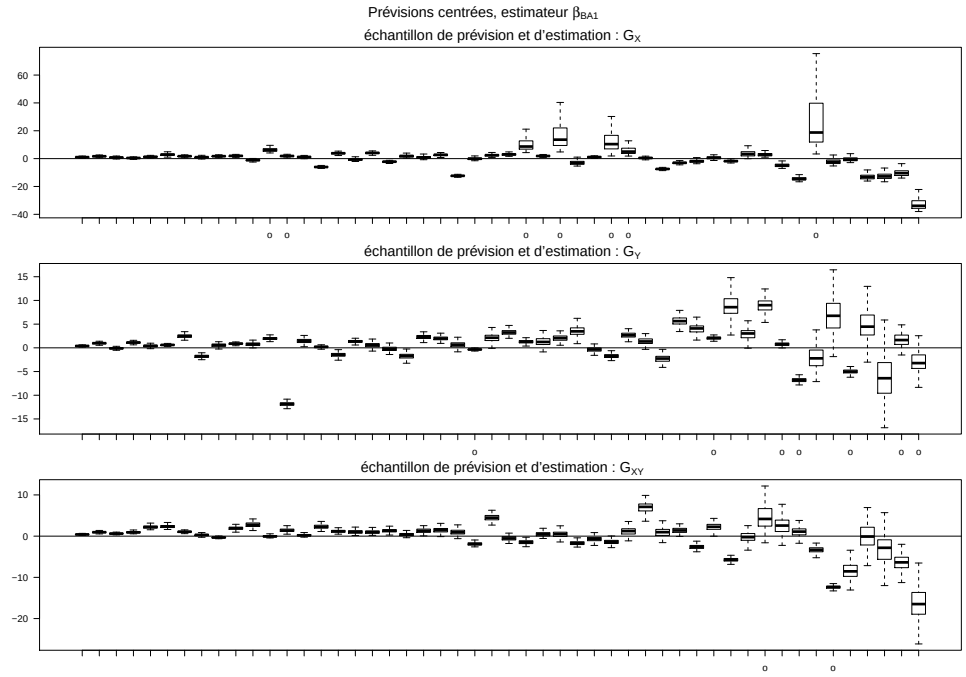
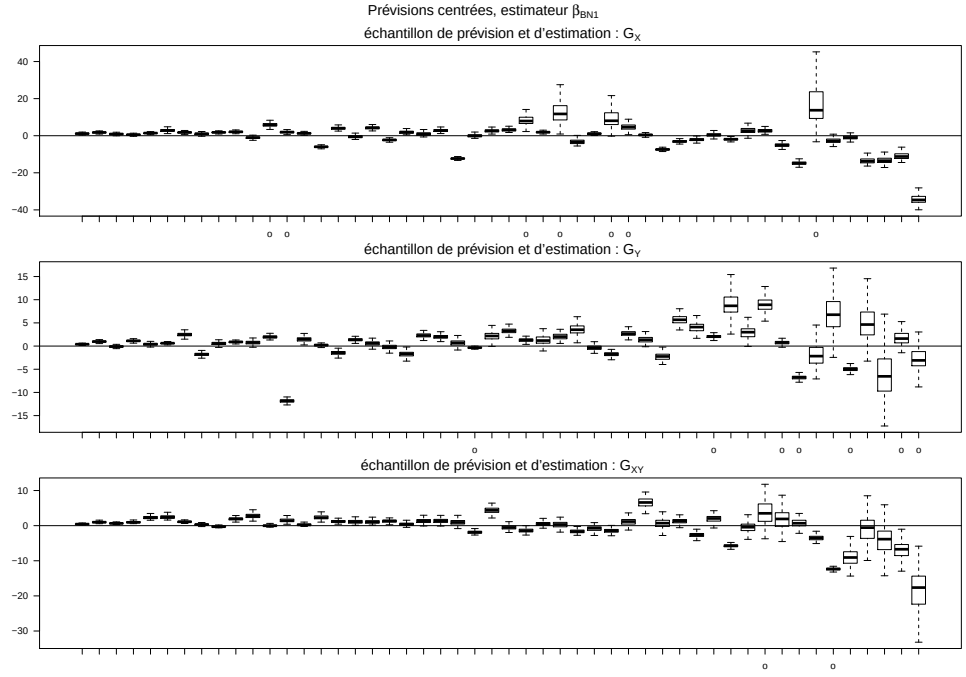


FIGURE 3.17



3.4.4.3 Observations générales de l'analyse

Comme on peut le voir dans le chapitre 3.2, les différentes contaminations ne présentent pas graphiquement des déviations évidentes au modèle. Cependant leur effet est très important qu'il s'agisse de l'estimation ou des prévisions. Les estimateurs robustes arrivent cependant à minimiser l'effet des contaminations. Mais ils ne semblent pas être très performant quand il s'agit d'une contamination G_X.
On remarque aussi que l'effet des poids $w(\boldsymbol{x_i})$ est quasi nul. Les estimations et les prévisions ne semblent pas être trop affectées par le choix de ces poids. Les estimateurs robustes sont performants quand il s'agit de faire des prévisions avec un échantillon de prévision non contaminé. Cependant quand l'échantillon de prévision est contaminé, les prévisions semblent de qualité équivalente qu'il s'agisse de l'estimateur de maximum de vraisemblance ou robuste. L'estimateur bayésien robuste montre dans un cas une nette amélioration ; quand les échantillons d'estimation et de prévision proviennent d'une contamination G_Y. Ainsi le tableau 3.2 résume les méthodes d'estimation que l'étude Monte-Carlo nous suggère. On remarque que l'estimateur développé fait de meilleures performances pour le cas d'une contamination en Y sur l'échantillon d'estimation et celui de prévision. L'estimation robuste classique est plus performant quand les données à prévoir n'ont pas subi de contamination. De plus, aucun estimateur testé n'est réellement bon lorsqu'il s'agit de contamination G_X et par conséquent G_{XY}.

TABLE 3.2 – Résumé des différentes méthodes d'estimation

	Estimation G_0	Estimation G_X	Estimation G_Y	Estimation G_{XY}
Prévision G_0	MLE et robuste	?	robuste	?
Prévision G_X		?		
Prévision G_Y			bayésien	
Prévision G_{XY}				?

Un point de vue plus global des résultats consiste à calculer la statistique de Pearson sur les prévisions en utilisant $\chi^2_{new} = \sum_{i=1}^{n}(y_{new,i} - \hat{\mu}_i)^2/\hat{\mu}_i$, où $\hat{\mu}_i$ est défini par l'équation (3.2). Comme cela a été expliqué auparavant (chapitre 3.4.3.1), cette statistique sera plus faible pour les cas de sur-prévision.

Les figures 3.18 et 3.19 nous montrent cette statistique pour chaque estimateur et les sept cas de prévisions considérés. Les méthodes robustes sont surtout efficaces pour prévoir des données non contaminées. Soit les prévisions sont très proches de celle par maximum de vraisemblance soit elles sont meilleures. C'est en effet l'hypothèse de base des méthodes robustes :

on dispose de données contaminées et on veut inférer sur un modèle non contaminé. La comparaison entre les différentes méthodes robustes (classiques et bayésiennes) est moins claire quand il s'agit de prévoir des données non contaminées. On remarque que les méthodes robustes d'estimation sont surtout efficaces pour réduire l'effet d'une contamination G_Y. Ce sont donc des méthodes intéressantes car on porte usuellement une attention particulière à la variable réponse, ce qui permet de repérer la présence de données extrêmes.
Dans le cas des prévisions de données contaminées, l'estimateur de maximum de vraisemblance semble meilleure. Et dans ce cas aussi les différentes méthodes robustes font des résultats comparables.

FIGURE 3.18

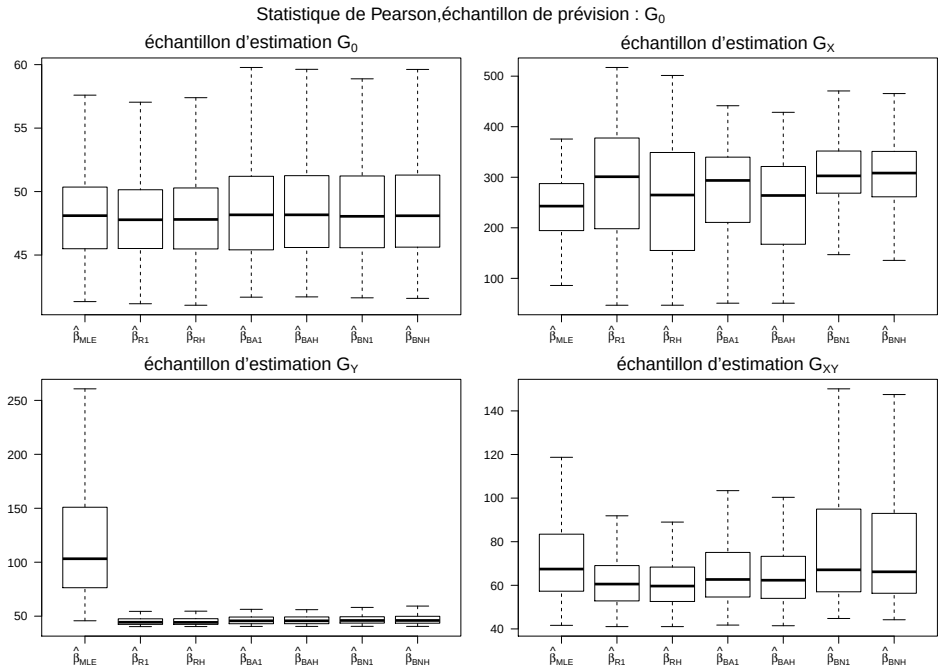
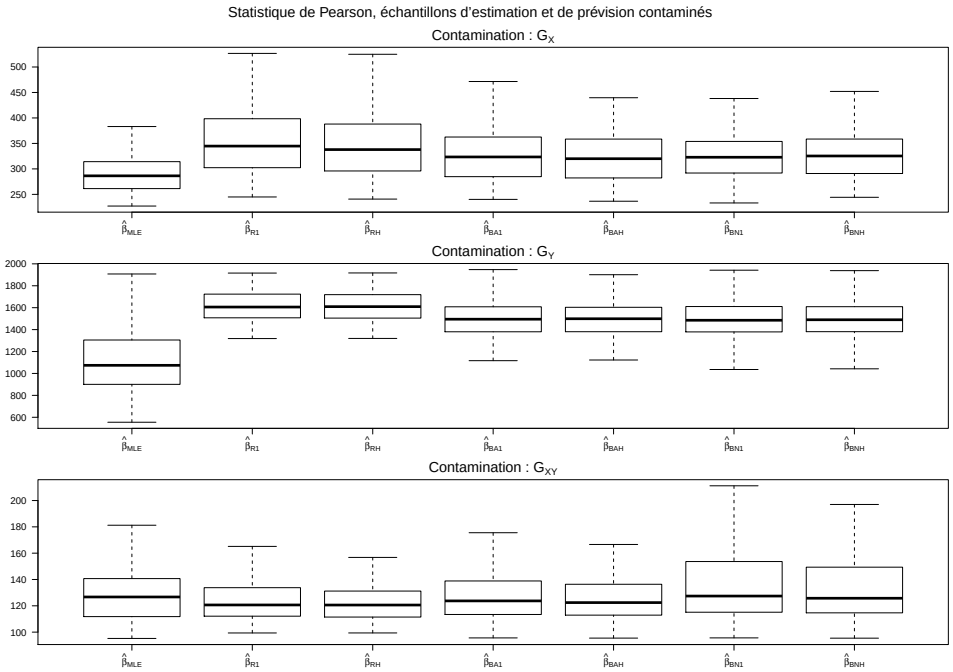


FIGURE 3.19


Statistique de Pearson, échantillons d'estimation et de prévision contaminés
Contamination : G_X
500
450
400
350
300
250
$\hat{p}_{MLE}$
$\hat{p}_{R1}$
$\hat{p}_{RH}$
$\hat{p}_{BA1}$
$\hat{p}_{BAH}$
$\hat{p}_{BN1}$
$\hat{p}_{BNH}$
Contamination : G_Y
2000
1800
1600
1400
1200
1000
800
600
$\hat{p}_{MLE}$
$\hat{p}_{R1}$
$\hat{p}_{RH}$
$\hat{p}_{BA1}$
$\hat{p}_{BAH}$
$\hat{p}_{BN1}$
$\hat{p}_{BNH}$
Contamination : G_{XY}
200
180
160
140
120
100
$\hat{p}_{MLE}$
$\hat{p}_{R1}$
$\hat{p}_{RH}$
$\hat{p}_{BA1}$
$\hat{p}_{BAH}$
$\hat{p}_{BN1}$
$\hat{p}_{BNH}$

Chapitre 4

Analyse d'un jeu de données

4.1 Présentation des données

Les données utilisées pour soumettre à un exemple réel les estimateurs bayésiens construits dans le chapitre 2.5 proviennent de l'ouvrage de Heritier *et al.* (2009). Il s'agit de données produites lors d'une étude menée par le *Health and Retirement Study*. Le jeu de donnée est composé de 3066 individus de plus de 50 ans et a été récolté en 2002. Le but de l'étude est de comprendre l'utilisation des soins médicaux par ces individus. La variable d'intérêt est le nombre de visites à un médecin durant les 2 dernières années. Un grand nombre de variables explicatives [1] ont été récoltées et Heritier *et al.* (2009) ont dû effectuer une sélection des variables. Ils aboutissent au modèle suivant :

$$\begin{aligned} g(E[\texttt{visits}]) = \quad & \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{gender} + \beta_3 \texttt{hispan} + \beta_4 \texttt{arthri} + \beta_5 \texttt{cancer} \\ & + \beta_6 \texttt{hipress} + \beta_7 \texttt{diabet} + \beta_8 \texttt{lung} + \beta_9 \texttt{hearth} + \beta_{10} \texttt{stroke} \\ & + \beta_{11} \texttt{psych} + \beta_{12} \texttt{iadla1} + \beta_{13} \texttt{iadla2} + \beta_{14} \texttt{iadla3} \\ & + \beta_{15} \texttt{adlwa1} + \beta_{16} \texttt{adlwa2} + \beta_{17} \texttt{adlwa3} + \beta_{18} \texttt{edyears} + \beta_{19} \texttt{feduc} \end{aligned} \tag{4.1}$$

La partie de sélection du modèle n'est pas le sujet d'intérêt de ce travail. Le modèle définit par l'équation (4.1) est donc réutilisé pour notre analyse. La description des variables explicatives sélectionnées par Heritier *et al.* (2009) est la suivante :

1. Pour d'avantages d'informations sur le jeu de donnée et l'ensemble des variables explicatives, veuillez vous référer à l'ouvrage de Heritier *et al.* (2009)

TABLE 4.1 – Description des variables

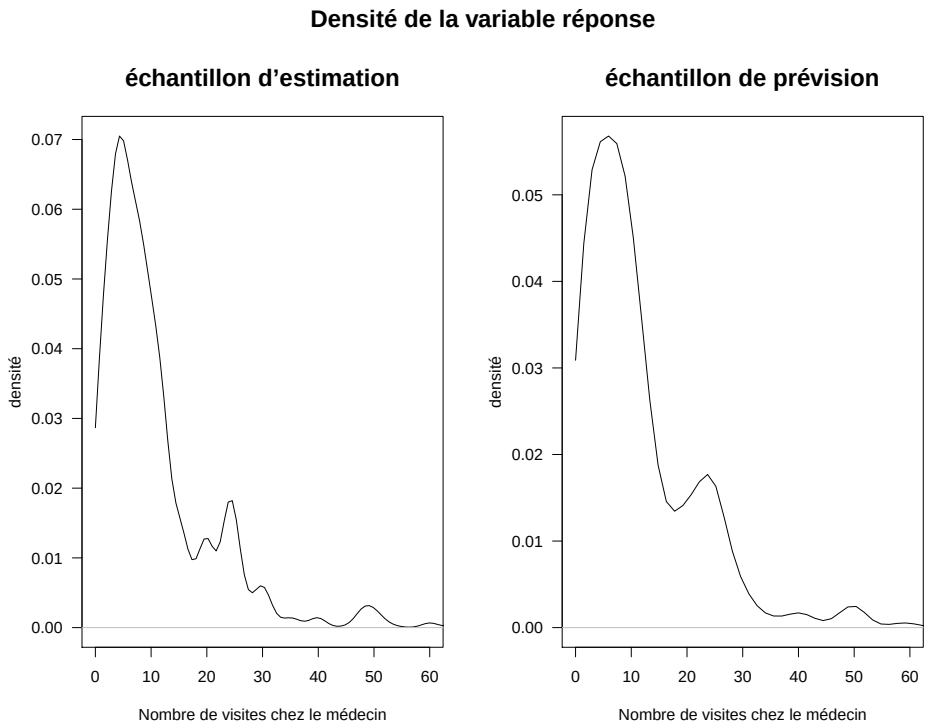
Nom	Description	Mesure
Variable réponse		
`visits`	visites chez le médecin	comptage durant 2 années
Facteurs prédisposants		
`age`	l'âge	année, variable continue
`gender`	le genre	1=femme, 0=homme
`hispan`	origine hispanique	1=hispanique, 0=autre
Besoins de santé	**Le patient a-t-il eu...**	
`arthri`	de l'arthrite ?	1=oui, 0=non
`cancer`	le cancer ?	1=oui, 0=non
`hipress`	de l'hypertension ?	1=oui, 0=non
`diabet`	du diabète ?	1=oui, 0=non
`lung`	une maladie pulmonaire ?	1=oui, 0=non
`hearth`	des problèmes cardiaques ?	1=oui, 0=non
`stroke`	un accident vasculaire cérébrale ?	1=oui, 0=non
`psych`	des problèmes psychiatriques ?	1=oui, 0=non
	Le patient a-t-il des difficultés...	
`iadla1-3`	à suivre des instructions ?	0,1,2,3 où 0=aucun et 3=beaucoup
`adlwa1-3`	dans la vie courante ?	0,1,2,3 où 0=aucun et 3=beaucoup
Accès économique		
`edyears`	formation	année, variable continue
`feduc`	formation paternelle	année, variable continue

4.2 Analyse des données

Afin d'avoir des échantillons d'estimation et de prévision distinctes, il est nécessaire de séparer les données de départ (3066 individus) en deux. L'échantillon d'estimation est de taille $n = 2500$ et l'échantillon de prévision de taille $n_{new} = 566$. Les deux échantillons viennent de la même distribution, nous sommes donc dans le cas de l'analyse Monte-Carlo étudiée dans le chapitre 3.4.4.2 Prévoir des données contaminées.
La variable de réponse `visits` est un comptage. On suppose donc qu'elle suit une loi de Poisson. La figure 4.1 montre la distribution de la variable réponse pour les deux échantillons. On remarque, sous l'hypothèse d'une distribution de Poisson, une forte présence de données extrêmes (entre 18 et 30 visites chez le médecin enregistrées et aux alentours de 50). De plus, les densités des

deux échantillons se ressemblent fortement, supposer qu'elle proviennent de la même distribution est donc fortement plausible.

FIGURE 4.1



4.2.1 Procédure de l'analyse

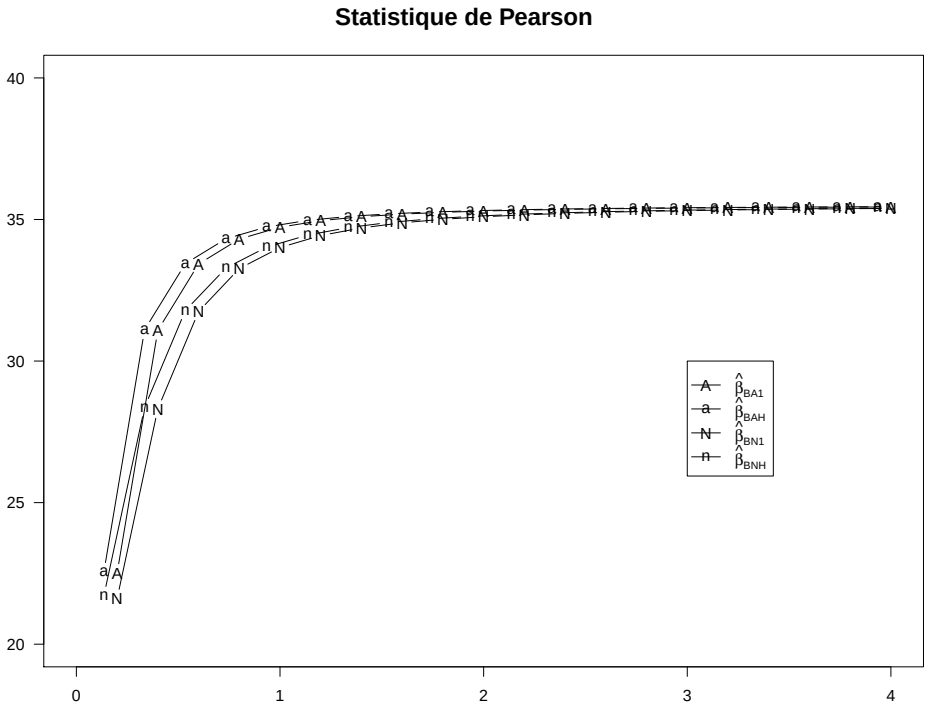
L'analyse de ce jeu de données se fera selon la même procédure que celle menée dans l'étude Monte-Carlo. Le modèle décrit par l'équation 4.1 est donc estimé par les sept estimateurs décrit dans le chapitre 3.1.1, soit $\hat{\boldsymbol{\beta}}_{\boldsymbol{MLE}}$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{R1}}$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{RH}}$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{BA1}}$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{BAH}}$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{BN1}}$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{BNH}}$. Pour les estimateurs bayésiens, la procédure de calibration de δ^2 est faite selon la statistique de Pearson, à l'instar de celle décrite dans le chapitre 3.4.3.1. L'analyse des prévisions est faite en les comparant aux vraies valeurs $y_{new,i}$ et entre les différentes méthodes d'estimation.

4.2.2 Calibrage de δ^2

Le graphique 4.2 nous montre la statistique de Pearson en fonction de δ et selon les différentes méthodes d'estimation bayésienne. Dans ce cas, la procédure n'est répétée qu'une seule fois. Cela a l'inconvénient de ne proposer qu'une seule valeur de la statistique par valeur de δ (et non plus une distribution), mais à l'avantage de demander moins de calculs. On a donc testé

20 valeurs de δ comprises entre 0.2 et 4 par pas de 0.2. La figure 4.2 nous suggère des observations faites auparavant, notamment le fait que l'estimateur bayésien tend vers l'estimateur robuste quand δ devient grand. De plus, aussi dans ce cas, l'ajout des poids en fonction de la matrice $\boldsymbol{H}$ ne semble pas avoir un effet important sur la statistique de Pearson.

FIGURE 4.2



La stratégie à adopter est de choisir un δ qui minimise cette statistique. On remarque que globalement une valeur de δ petite a tendance à la réduire. Dans le cas des estimateurs résolus numériquement ($\hat{\boldsymbol{\beta}}_{\boldsymbol{BN1}}$ et $\hat{\boldsymbol{\beta}}_{\boldsymbol{BNH}}$), le minimum correspond à une optimisation qui n'a pas convergé. Le choix de δ s'est donc reporté sur un autre minimum ayant convergé.

TABLE 4.2 – Choix de δ selon l'estimateur

méthode d'estimation	$\arg\min_{\delta} \frac{1}{n}\sum_{i=1}^{n} \frac{[\boldsymbol{y_i} - g^{-1}(\boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}}_{\boldsymbol{B}})]^2}{g^{-1}(\boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}}_{\boldsymbol{B}})}$	choix de δ
$\hat{\boldsymbol{\beta}}_{\boldsymbol{BA1}}$	0.2	0.2
$\hat{\boldsymbol{\beta}}_{\boldsymbol{BAH}}$	0.2	0.2
$\hat{\boldsymbol{\beta}}_{\boldsymbol{BN1}}$	0.2	0.6
$\hat{\boldsymbol{\beta}}_{\boldsymbol{BNH}}$	0.2	0.6

4.2.3 Estimation

Le tableau suivant résume les valeurs des estimés pour chaque estimateur.

TABLE 4.3 – Estimés des paramètres selon les estimateurs

	$\hat{\boldsymbol{\beta}}_{\boldsymbol{MLE}}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{R1}}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{RH}}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{BA1}}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{BAH}}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{BN1}}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{BNH}}$
`intercept`	3.1600	1.8349	1.8254	3.9454	3.9115	2.0909	2.0655
`age`	-0.0135	-0.0042	-0.0041	-0.0193	-0.0192	-0.0066	-0.0065
`gender`	-0.0304	0.0484	0.0493	-0.1901	-0.1895	-0.0006	0.0005
`hispan`	0.1956	0.3252	0.3249	0.3591	0.3598	0.3485	0.3487
`marital`	-0.0482	-0.0440	-0.0435	-0.0569	-0.0545	-0.0367	-0.0355
`arthri`	0.0568	0.1652	0.1659	0.0166	0.0162	0.1449	0.1455
`cancer`	0.0949	0.2076	0.2080	-0.0117	-0.0120	0.1704	0.1707
`hipress`	0.2258	0.1964	0.1968	0.0989	0.0968	0.1803	0.1801
`diabet`	0.1041	0.1582	0.1580	0.0230	0.0240	0.1235	0.1239
`lung`	0.1427	0.1464	0.1471	0.2912	0.2938	0.1942	0.1952
`hearth`	0.3579	0.3098	0.3097	0.2174	0.2182	0.3059	0.3058
`stroke`	0.0486	0.0820	0.0814	0.2422	0.2426	0.1363	0.1357
`psych`	0.2489	0.1809	0.1799	0.1066	0.1088	0.1801	0.1797
`iadla1`	0.0330	0.0356	0.0361	-0.2775	-0.2795	-0.0308	-0.0308
`iadla2`	0.2059	0.2469	0.2481	0.0562	0.0554	0.2205	0.2209
`iadla3`	0.2385	0.2037	0.2064	0.1663	0.1658	0.2368	0.2382
`adlwa1`	0.3081	0.1837	0.1833	0.1960	0.1945	0.1970	0.1964
`adlwa2`	0.5064	0.2549	0.2546	0.4967	0.4959	0.3331	0.3325
`adlwa3`	0.4698	0.3752	0.3746	0.1054	0.1032	0.2871	0.2860
`edyears`	0.0206	0.0091	0.0091	0.0286	0.0285	0.0144	0.0143
`feduc`	-0.0471	-0.0139	-0.0134	-0.0415	-0.0385	-0.0106	-0.0090

On remarque que l'`intercept` est important. Or comme, on a pu l'apercevoir dans l'analyse Monte-Carlo une contamination sur Y joue un grand rôle dans le biais de l'`intercept`. Il est donc naturel de penser que la contamination perçue dans la figure 4.1 est à l'origine de la valeur importante pour l'`intercept` ainsi que des différences entre les méthodes d'estimation.

4.2.4 Les prévisions

Dans ce cas l'analyse des prévisions doit se faire globalement. En effet il n'est plus possible d'avoir une distribution pour chaque donnée de l'échantillon de prévision. Il s'agit donc de comparer l'ensemble des prévisions. C'est pour cette raison que l'on a préféré un $n_{new} = 566$ grand. La statistique de Pearson évaluée sur l'échantillon de prévision permet donc d'avoir une idée globale de la prévision. Le tableau suivant montre les valeurs de cette statistique. On remarque que l'estimateur de maximum de vraisemblance fait une meilleure performance que les autres estimateurs d'un point de vue de leur prévisions. Cependant les estimateurs $\hat{\boldsymbol{\beta}}_{\boldsymbol{BA1}}$ et $\hat{\boldsymbol{\beta}}_{\boldsymbol{BAH}}$ ont une valeur de la statistique de Pearson très proche. On peut en effet rapprocher le cas de cette analyse de donnée à la simulation Monte-Carlo où l'échantillon de prévision et l'échantillon d'estimation proviennent d'une même distribution ayant subie une contamination en XY ; il est cependant difficile d'observer la part de contamination en X car la majorité des variables explicatives sont catégorielles. La figure 3.19 montre que dans ce type de contamination les estimateurs bayesiens ont une faible valeur de la statistique de Pearson. La différence entre les estimateurs bayésiens par approximation analytique et par résolution numérique peut s'expliquer par le choix de δ qui diffère selon la méthode.

TABLE 4.4 – Statistique de Pearson évaluée sur les prévisions

Estimateur	Statistique de Pearson
$\hat{\boldsymbol{\beta}}_{\boldsymbol{MLE}}$	89.0854
$\hat{\boldsymbol{\beta}}_{\boldsymbol{R1}}$	125.5139
$\hat{\boldsymbol{\beta}}_{\boldsymbol{RH}}$	125.4533
$\hat{\boldsymbol{\beta}}_{\boldsymbol{BA1}}$	89.0970
$\hat{\boldsymbol{\beta}}_{\boldsymbol{BAH}}$	89.0970
$\hat{\boldsymbol{\beta}}_{\boldsymbol{BN1}}$	114.6748
$\hat{\boldsymbol{\beta}}_{\boldsymbol{BNH}}$	114.5116

Conclusion

L'étude Monte-Carlo montre que la méthode d'estimation joue un rôle important pour faire de bonnes prévisions. Lorsque l'on contamine l'échantillon d'estimation mais pas celui de prévision, l'estimation robuste produit globalement de meilleures prévisions. En se replaçant dans le contexte de l'équation (1.1), on remarque que, dans ce cas, on infère sur le modèle non contaminé. Les méthodes robustes sont donc adaptées à ce contexte.
L'estimateur robuste bayésien a montré un comportement intéressant lorsqu'il s'agit de faire des prévisions quand les échantillons d'estimation et de prévision ont été contaminés sur la variable réponse. En effet il garanti un minimum d'erreurs extrêmes alors que les estimateurs de maximum de vraisemblance et robustes classiques peuvent produire des prévisions très éloignées de la vraie valeur. Il apporte donc un avantage certain lorsque l'on veut éviter les erreurs de prévisions extrêmes.

Il semble important d'expliciter les points qui ont abouti à des questions non résolues. En effet dans le cadre d'une recherche sur un sujet proche, ces réflexions seront automatiquement présentes. Les présenter permet donc de préciser les points qui nécessitent une approche particulière.

Tout d'abord il paraît important de discuter des prévisions dans les GLM. En effet, les spécificités des GLM ne permettent pas l'utilisation des outils statistiques développés pour les modèles linéaires. Pour ces derniers, l'hypothèse d'homoscédasticité est forte et il devient aisé de définir une erreur de prévisions. Pour les GLM, il est réellement plus difficile de juger la qualité d'une prévision d'une manière objective. Il est nécessaire de se placer dans un contexte, de savoir quelle part de la distribution nous intéresse, de savoir si l'on est prêt à assumer un biais contre une variance plus faible. Or, ce type de contraintes n'existent pas dans le cas d'une étude Monte-Carlo. De plus, dans le cas particulier de ce travail, il a été impossible d'utiliser la théorie de la vraisemblance. Ainsi pour juger la qualité des prévisions, deux outils majeurs ont été choisis. Le premier est la statistique de Pearson appliquée aux prévisions et le second est l'appréciation graphique des distributions des prévisions. La première technique a le défaut de réduire fortement l'information et de poser les hypothèses qu'une bonne prévision dépend de sa variance mais

aussi qu'il est préférable d'avoir une prévision trop grande plutôt que trop petite. Et la seconde pose tous les problèmes de subjectivité que comporte l'analyse visuelle des graphiques ; et cela est d'autant plus vrai que le nombre de graphiques est important. Il conviendrait de développer un meilleur outil pour juger la qualité des prévisions dans le cadre des GLM.

Ensuite il est nécessaire de s'intéresser aux choix des contaminations. Dans le cadre d'une analyse d'estimateurs robustes, il est important de bien définir la contamination du modèle. Dans ce cas aussi, la comparaison avec le modèle linéaire est à éviter. En effet, la distribution de la variable réponse est, pour le cas d'une variable réponse de Poisson, asymétrique et bornée. Cela implique différentes complications. En effet, il n'est possible de définir des données extrêmes que sur une queue de la distribution. Si l'on produit des données extrêmes trop faibles (plutôt que trop grandes), la variable de réponse aura une distribution avec un excès de zéro et il devient évident qu'un modèle ZIP, par exemple, est plus adapté. L'implémentation des contaminations ne se fait donc pas aussi librement que dans le cas des modèles linéaires. Étudier cette contamination est donc réellement important avant de faire le choix de la méthode robuste.

L'estimateur bayésien développé comporte de nombreux paramètres et il n'a pas été possible dans ce travail de tous les tester. δ est le paramètre qui a nous a particulièrement intéressé. A contrario, le choix de c_2, qui cible les données extrêmes servant à produire la correction, a été repris de l'analyse de Genton et Ronchetti (2008). De plus, les auteurs n'utilisent pas cette correction sur un M-estimateur robuste, mais sur d'autres types d'estimateurs. Il apparaît donc que cette correction n'est pas une forme fixe mais peut être adaptée et ajustée. Ainsi la combinaison entre les différents paramètres doit pouvoir être optimisée pour les GLM.

Pour finir il convient de noter que le choix des poids $w(\boldsymbol{x_i})$ (présents lors des estimations robustes et robustes bayésiennes) n'a que très peu d'influence. Cependant les choisir comme étant 1 ou fonction de la matrice $\boldsymbol{H}$ n'est pas imposé. Ils permettent de pondérer les observations lors de l'estimation afin que les données extrêmes aient une influence plus faible. Et il est apparu que l'estimation de $\boldsymbol{\mu_p}$ définie dans le chapitre 2.5 à la particularité de n'être composée que d'une fonction des données extrêmes. Les équations d'estimations évaluées en $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ en utilisant une constante $c_2 > c$ permettent de détecter les données extrêmes d'un modèle. Il y a donc peut-être dans cette propriété la possibilité de produire des poids $w(\boldsymbol{x_i})$ ayant une plus grande influence sur l'estimation.

Appendices

Annexe A

Appendices théoriques

A.1 Propriétés de la vraisemblance

$$\begin{aligned} E\left[\frac{dl(\theta;x)}{d\theta}\right] &= \int \left[\frac{d}{d\theta}\ln\left(f_X(x;\theta)\right)\right] f_X(x;\theta)dx \\ &= \int \left[\frac{1}{f_X(x;\theta)}\frac{d}{d\theta}f_X(x;\theta)\right] f_X(x;\theta)dx \\ &= \frac{d}{d\theta}\int f_X(x;\theta)dx = 0 \end{aligned} \tag{A.1}$$

$$\begin{aligned} E\left[\frac{d^2l(\theta;y)}{d\theta^2}\right] + E\left[\left(\frac{dl(\theta;y)}{d\theta}\right)^2\right] &= E\left[\frac{d^2l(\theta;y)}{d\theta^2} + \left(\frac{dl(\theta;y)}{d\theta}\right)^2\right] \\ &= \int \left[\frac{d^2l(\theta;y)}{d\theta^2} + \left(\frac{dl(\theta;y)}{d\theta}\right)^2\right] f_X(x;\theta)dx \\ &= \int \left[\frac{d^2l(\theta;y)}{d\theta^2} + \left(\frac{1}{f_X(x;\theta)}\frac{d}{d\theta}f_X(x;\theta)\right)^2\right] f_X(x;\theta)dx \\ &= \int \left[\frac{d^2l(\theta;y)}{d\theta^2}f_X(x;\theta) + \frac{1}{f_X(x;\theta)}\left(\frac{d}{d\theta}f_X(x;\theta)\right)^2\right] dx \\ &= \int \left[\frac{d^2l(\theta;y)}{d\theta^2}f_X(x;\theta) + \left(\frac{d}{d\theta}f_X(x;\theta)\right)\left(\frac{d}{d\theta}l(\theta;x)\right)\right] dx \\ &= \int \left[\frac{d}{d\theta}\left(f_X(x;\theta)\left(\frac{dl(\theta;x)}{d\theta}\right)\right)\right] dx \\ &= \frac{d}{d\theta}\int \frac{dl(\theta;x)}{d\theta}f_X(x;\theta)dx = 0 \end{aligned} \tag{A.2}$$

A.2 Estimation d'intégrale de Laplace

Forme de l'intégrale :

$$I = \int \exp[-\lambda g(\boldsymbol{y})]h(\boldsymbol{y})d\boldsymbol{y}$$

L'expansion de Taylor de $g(\cdot)$ est :

$$g(\boldsymbol{y}) \quad = g(\boldsymbol{y}^*) + (\boldsymbol{y} - \boldsymbol{y}^*)^\top \frac{d^2 g(\boldsymbol{y}^*)}{d\boldsymbol{y} d\boldsymbol{y}^\top} (\boldsymbol{y} - \boldsymbol{y}^*)/2 + \ldots,$$

où $\frac{dg(\boldsymbol{y}^*)}{d\boldsymbol{y}} = \mathbf{0}$. L'expansion de Taylor de $h(\cdot)$ est :

$$h(\boldsymbol{y}) = h(\boldsymbol{y}^*) + (\boldsymbol{y} - \boldsymbol{y}^*)^\top \frac{dg(\boldsymbol{y}^*)}{d\boldsymbol{y}}, \tag{A.3}$$

où $h(\boldsymbol{y})$ est linéaire en $h(\boldsymbol{y}^*)$. On peut ainsi développer l'intégrale I en écrivant :

$$\begin{aligned}
I &= \int \exp[-\lambda g(\boldsymbol{y})] h(\boldsymbol{y}) d\boldsymbol{y} \\
&= \int \exp\left[-\lambda \left(g(\boldsymbol{y}^*) + (\boldsymbol{y} - \boldsymbol{y}^*)^\top \frac{d^2 g(\boldsymbol{y}^*)}{d\boldsymbol{y} d\boldsymbol{y}^\top} (\boldsymbol{y} - \boldsymbol{y}^*)/2 + \ldots \right)\right] \\
&\quad \times \left(h(\boldsymbol{y}^*) + (\boldsymbol{y} - \boldsymbol{y}^*)^\top \frac{dh(\boldsymbol{y}^*)}{d\boldsymbol{y}} \right) d\boldsymbol{y} \\
&\approx \exp\left[-\lambda g(\boldsymbol{y}^*)\right] h(\boldsymbol{y}^*) \int \exp\left[-\lambda (\boldsymbol{y} - \boldsymbol{y}^*)^\top \frac{d^2 g(\boldsymbol{y}^*)}{d\boldsymbol{y} d\boldsymbol{y}^\top} (\boldsymbol{y} - \boldsymbol{y}^*)/2\right] d\boldsymbol{y} \\
&\quad + \exp\left[-\lambda g(\boldsymbol{y}^*)\right] \frac{dg(\boldsymbol{y}^*)}{d\boldsymbol{y}} \int (\boldsymbol{y} - \boldsymbol{y}^*)^\top \exp\left[-\lambda (\boldsymbol{y} - \boldsymbol{y}^*)^\top \frac{d^2 g(\boldsymbol{y}^*)}{d\boldsymbol{y} d\boldsymbol{y}^\top} (\boldsymbol{y} - \boldsymbol{y}^*)/2\right] d\boldsymbol{y}.
\end{aligned}$$

Le premier terme peut être calculé grâce à la densité de la loi normale multivariée. Le second terme peut s'écrire comme étant $k \int \boldsymbol{z} \phi(\boldsymbol{z}) d\boldsymbol{z}$, où k est une constante, on effectue une transformation de variable tel que $\boldsymbol{z} = \boldsymbol{y} - \boldsymbol{y}^*$ et $\phi(\cdot)$ est la densité de la loi normale centrée réduite. Ce second terme est donc nul. Ainsi ce type d'intégrale peut être approximé par :

$$I \approx \exp\left[-\lambda g(\boldsymbol{y}^*)\right] h(\boldsymbol{y}^*) (2\pi/\lambda)^{d/2} \left| \frac{d^2 g(\boldsymbol{y}^*)}{d\boldsymbol{y} d\boldsymbol{y}^\top} \right|^{-1/2}.$$

Annexe B

Appendices graphiques

B.1 Différence entre $\hat{\boldsymbol{\beta}}_{\boldsymbol{B}}$ et $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ selon δ^2

FIGURE B.1

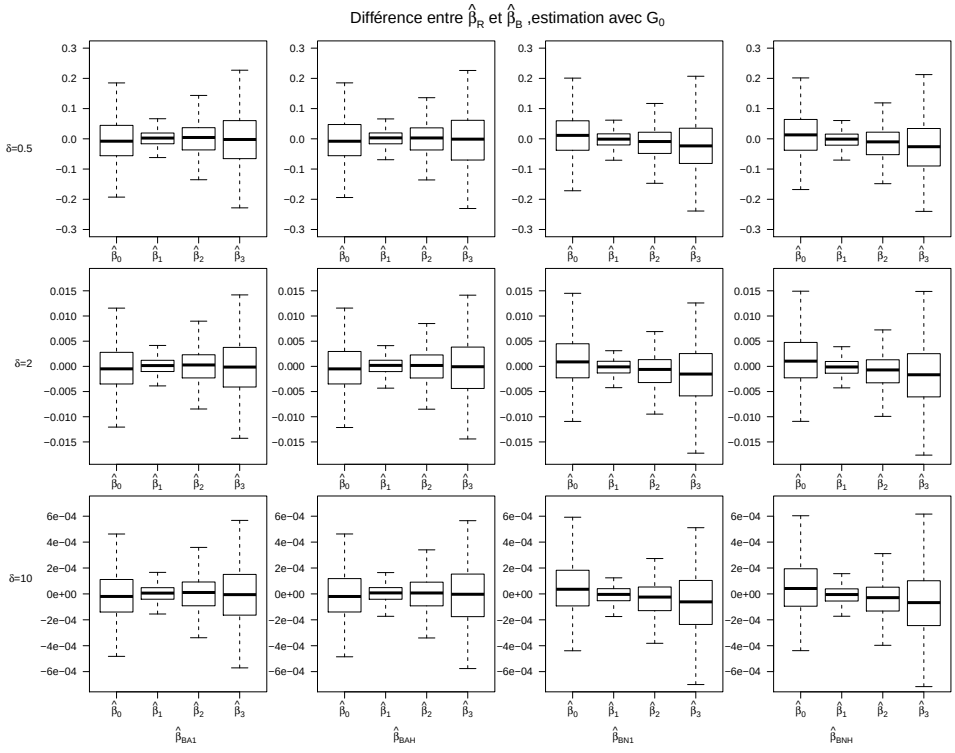


FIGURE B.2

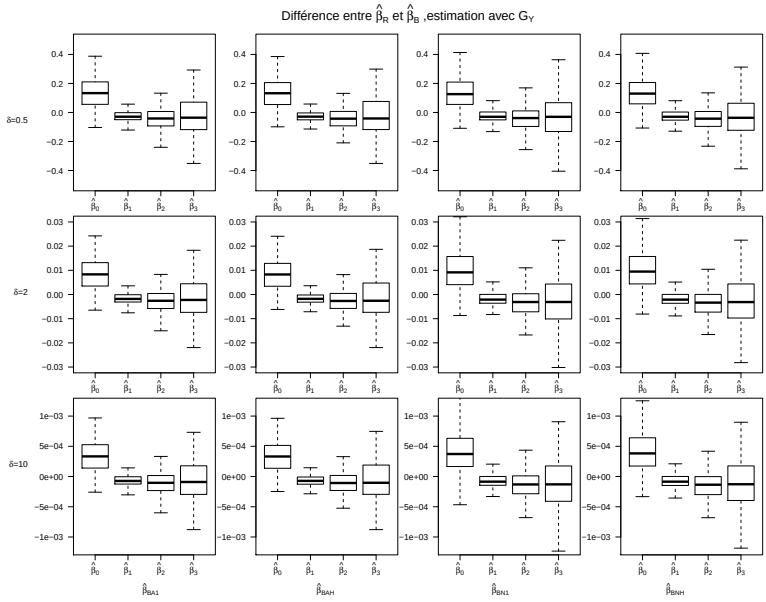
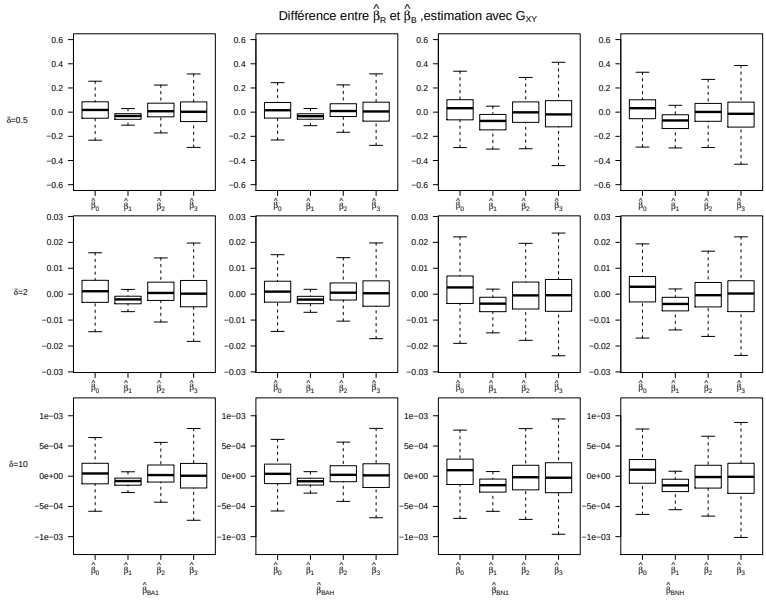


FIGURE B.3



B.2 Erreur de prédiction selon δ^2

FIGURE B.4

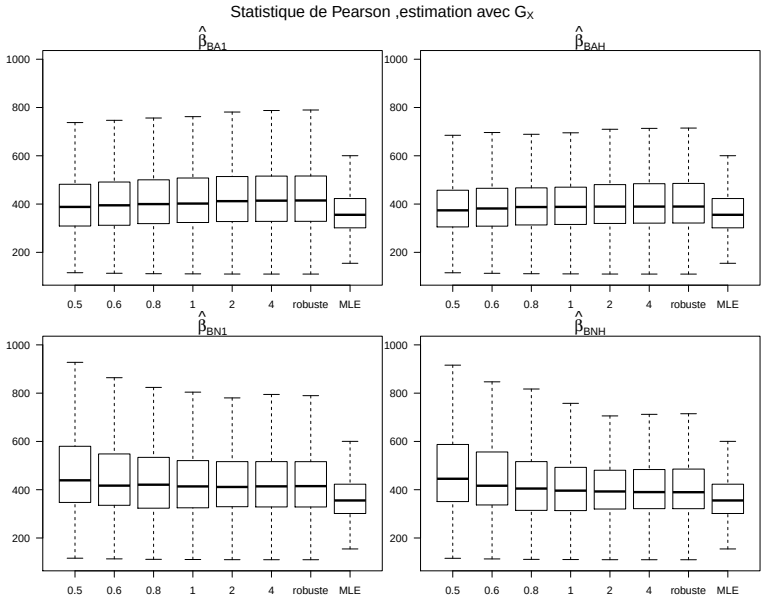


FIGURE B.5

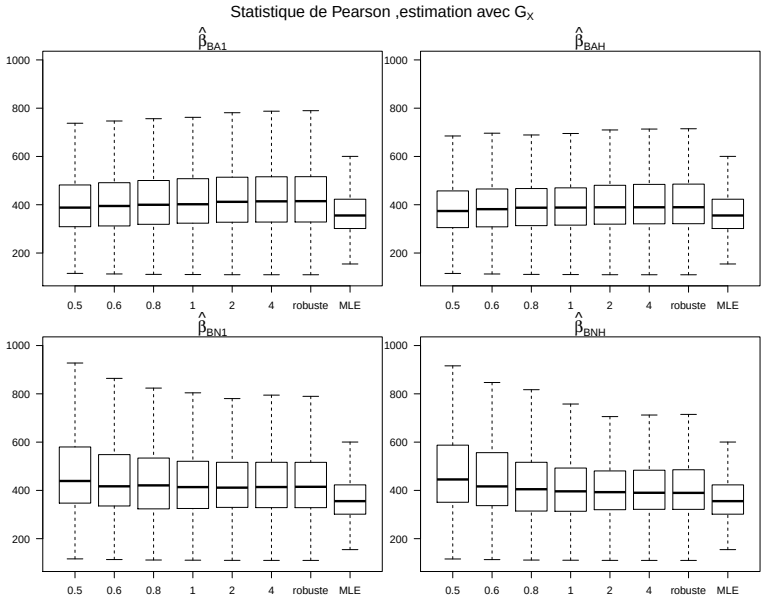
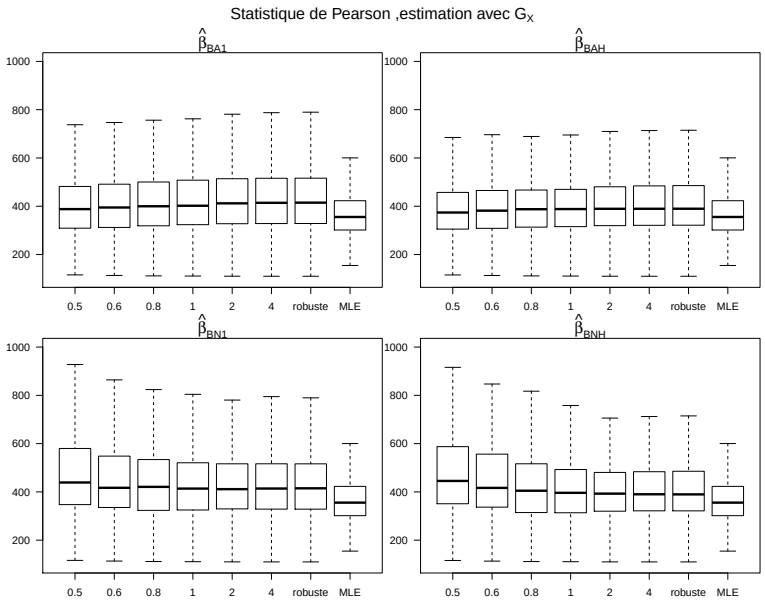


FIGURE B.6



B.3 Estimation bayesienne robuste

FIGURE B.7

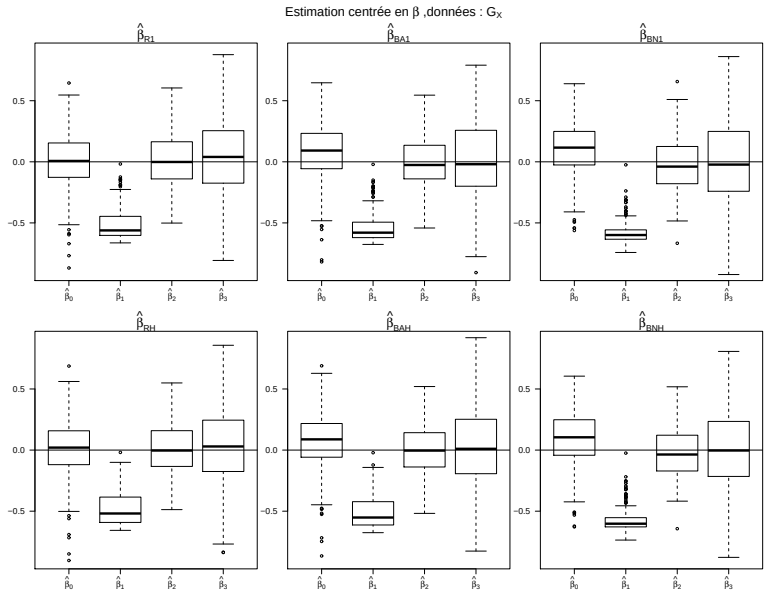
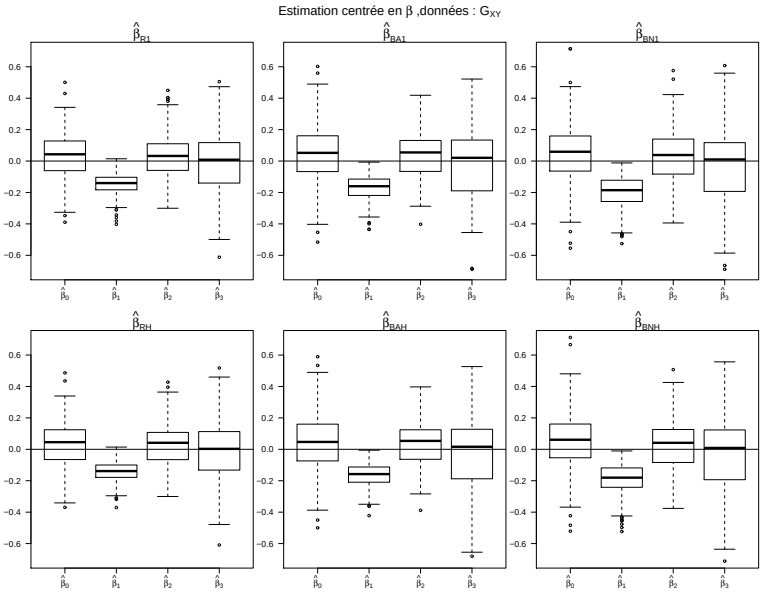


FIGURE B.8



B.4 Prévoir des données propres

FIGURE B.9

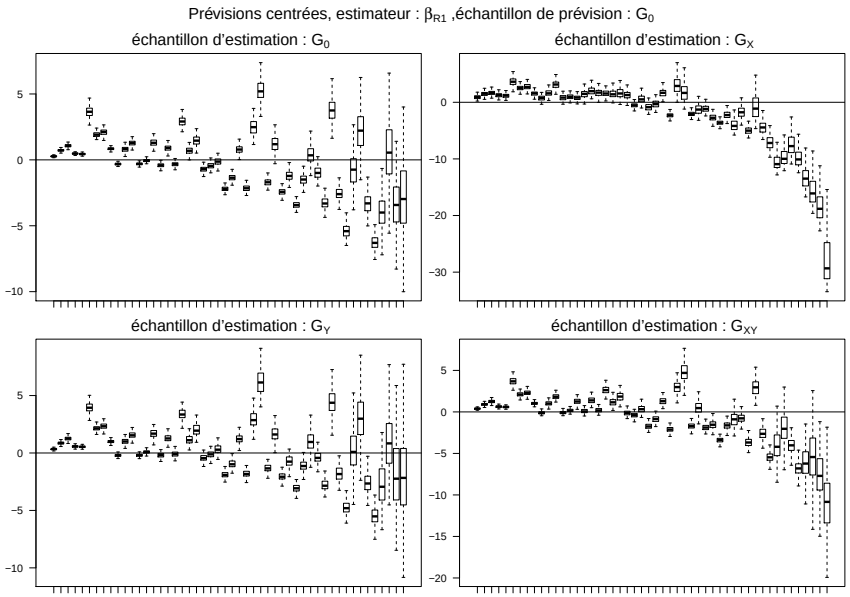


FIGURE B.10

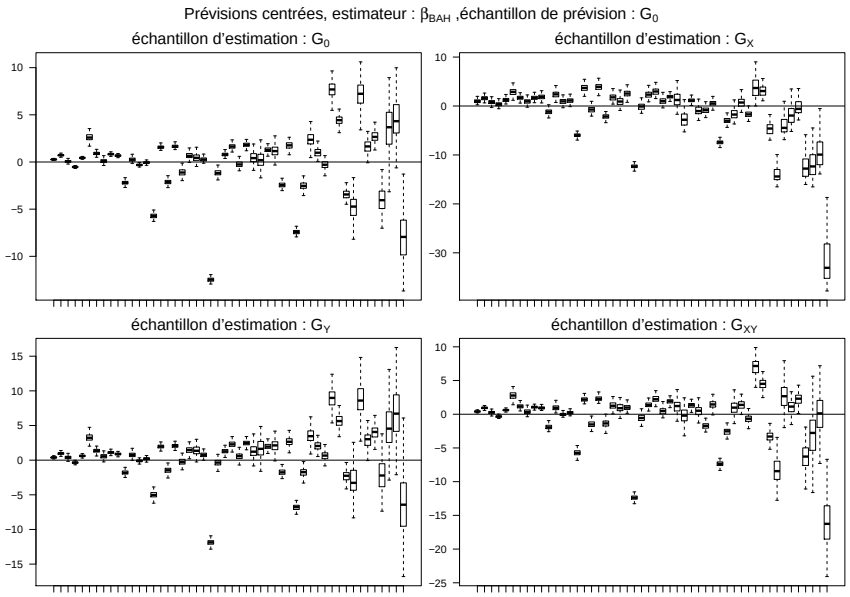


FIGURE B.11

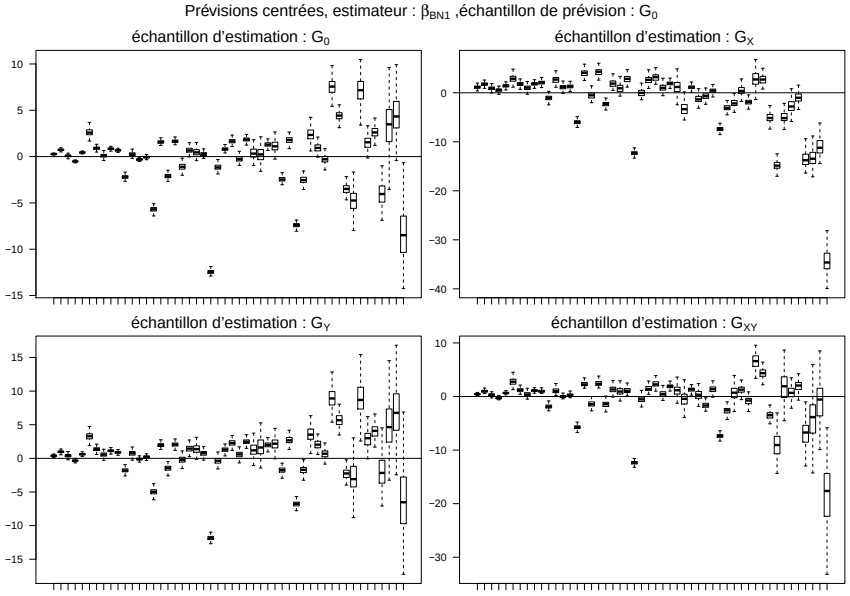
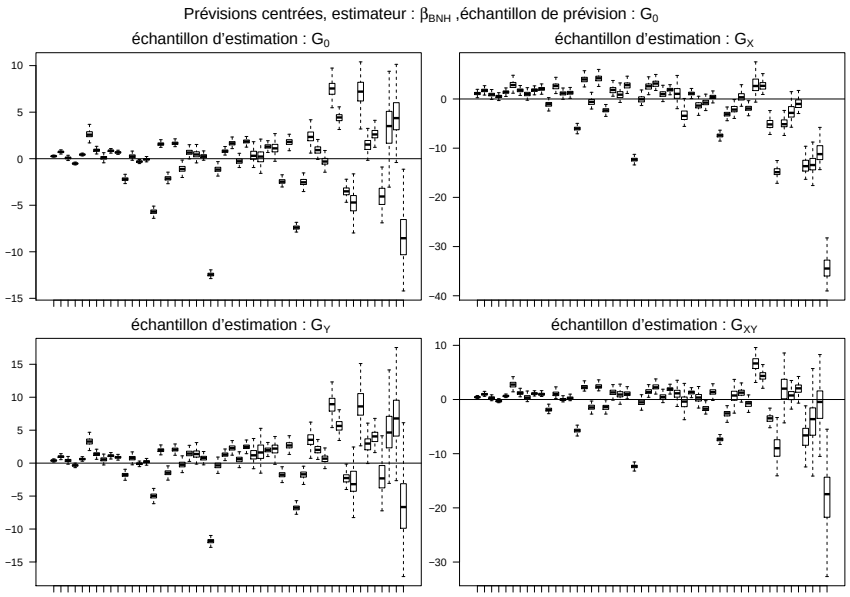


FIGURE B.12



B.5 Prévoir des données contaminées

FIGURE B.13

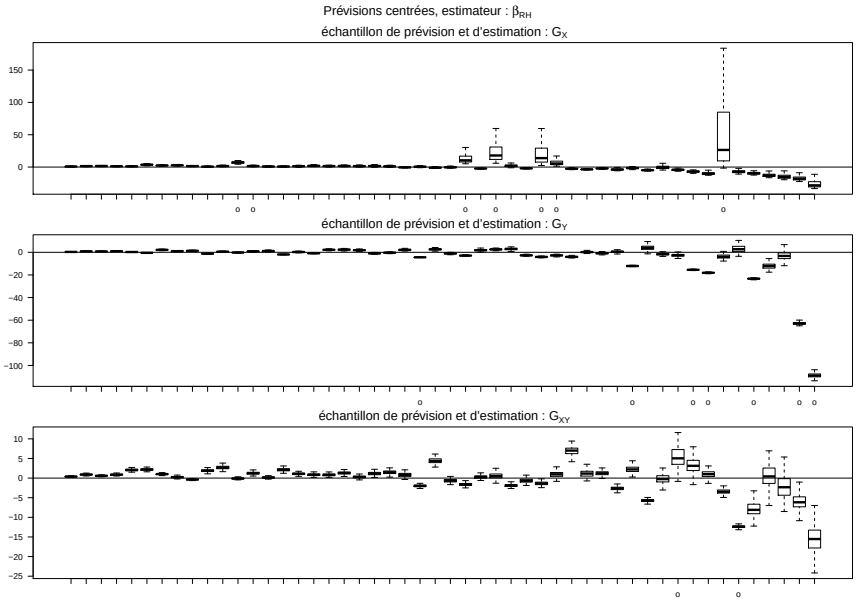


FIGURE B.14

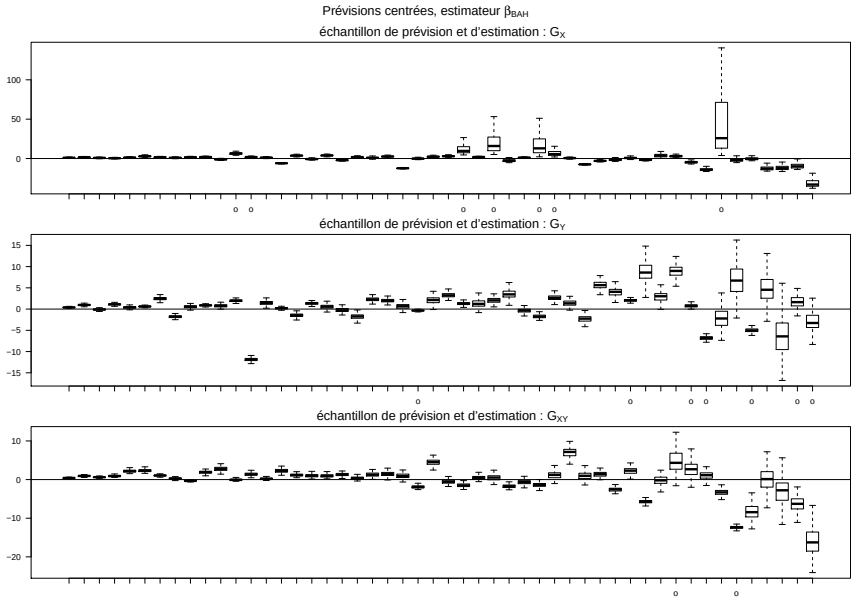
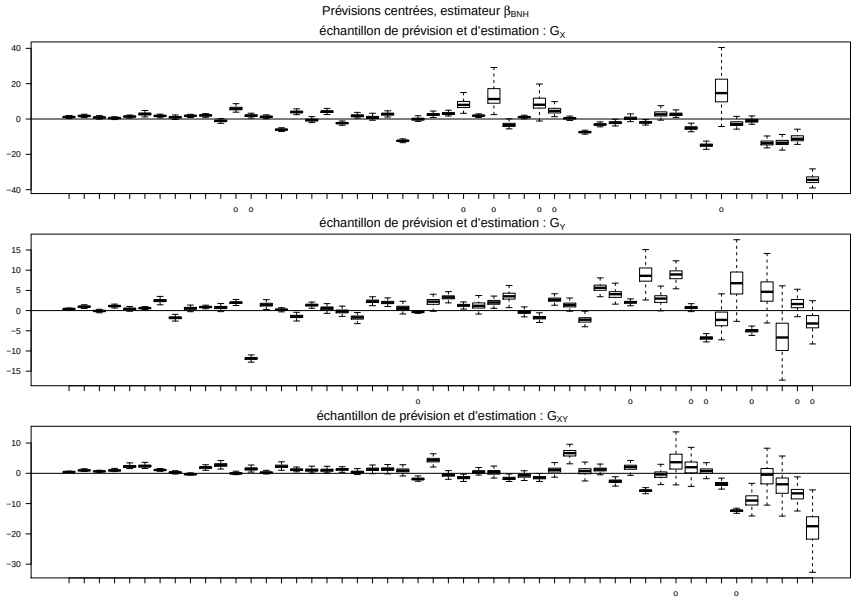


FIGURE B.15



Bibliographie

Cantoni, E. et Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96(455): 1022–1030.

Cantoni, E. et Ronchetti, E. (2006). A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures. *Journal of health economics*, 25(2):198–213.

Genton, M. et Ronchetti, E. (2008). Robust prediction of beta. *Computational Methods in Financial Engineering*, pages 147–161.

Hampel, F., Ronchetti, E., Rousseeuw, P. et Stahel, W. (1986). *Robust statistics : the approach based on influence functions*, volume 114. Wiley.

Heritier, S., Cantoni, E., Copt, S. et Victoria-Feser, M. (2009). *Robust methods in Biostatistics*, volume 838. Wiley.

Lejeune, M. (2010). *Statistique. La théorie et ses applications*. Statistique et probabilités appliquées. Springer.

McCullagh, P. et Nelder, J. (1989). *Generalized linear models*. Chapman & Hall/CRC.

Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M. et Maechler, M. (2012). *robustbase : Basic Robust Statistics*. R package version 0.9-2.