LUCIE VEILLEUX

Modélisation de la trajectoire criminelle de jeunes contrevenants à l'aide de modèles linéaires généralisés mixtes

Mémoire présenté
à la Faculté des études supérieures de l'Université Laval
dans le cadre du programme de maîtrise en statistique
pour l'obtention du grade de Maître ès sciences (M. Sc.)

FACULTÉ DES SCIENCES ET DE GÉNIE
UNIVERSITÉ LAVAL
QUÉBEC

2005

Résumé

La régression linéaire est souvent utilisée en pratique afin de trouver une relation entre une variable réponse et une ou plusieurs variable(s) explicative(s). Une lacune de cette méthode est qu'elle est inappropriée si la variable réponse en est une de dénombrement. Dans un tel cas, la régression de Poisson doit être utilisée.

Ce mémoire décrira de façon détaillée la régression de Poisson. Les propriétés de la loi de Poisson seront énoncées dans le but d'expliquer la régression de Poisson. Les équations d'estimation généralisées (GEE) seront ensuite introduites dans un éventuel but d'élargir la régression de Poisson dans les situations où les données sont corrélées (par exemple, les données longitudinales). Les modèles linéaires généralisés mixtes seront aussi considérés. Les modèles additifs généralisés seront ensuite brièvement expliqués et nous présenterons finalement une étude détaillée d'une base de données sur les trajectoires criminelles de jeunes contrevenants.

Avant-propos

Je tiens tout d'abord à remercier mon directeur de recherche, Thierry Duchesne, professeur au Département de mathématiques et de statistique de l'Université Laval. Toute son aide et tous ses conseils m'ont été d'une grande importance au cours des deux dernières années. Il a de plus toujours su se montrer disponible et m'a encouragée tout au long de mes études, ce qui fut grandement apprécié.

Je dois aussi dire merci à Gaétan Daigle du Service de Consultation Statistique de l'Université. J'ai eu la chance de travailler avec lui à maintes reprises au cours de mes études, et tout ce qu'il m'a appris, autant du côté statistique que du côté programmation SAS, m'a été et me sera toujours d'une grande utilité. Merci également pour l'aide dans l'analyse des données de ce mémoire.

Ensuite, je dois remercier certain(e)s de mes ami(e)s, en particulier Nathalie Savard et Marianne Fournier avec qui j'ai eu la chance et le plaisir d'étudier depuis le début de l'université. Je veux aussi dire merci à Anne Toulouse et à Chantal Poulin, mes amies de la Beauce, avec qui j'entretiens une belle amitié depuis le primaire. Vous avez tous et toutes contribué, peut-être sans le savoir, à l'avancement de mes travaux.

Je ne peux passer sous silence certain(e)s de mes « colocataires préféré(e)s » qui ont fait de mes 4 premières années d'université des années mémorables. Je pense en particulier à Marie-France Proulx, Marie-Josée et Simon Castonguay, Julie Turcotte, Ghislain Hébert, Jérôme Alary et Sébastien Neault. Jamais je n'oublierai le temps passé avec vous : je garde tellement de beaux souvenirs du temps où l'on habitait ensemble !

Il me reste à faire un merci tout particulier à faire à mon père et à ma sœur pour leur encouragement et leur soutien.

À ma mère.

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Introduction

Bien que la régression linéaire soit fréquemment utilisée dans les applications de la statistique, il arrive à certaines occasions que celle-ci ne soit pas appropriée : par exemple, si la variable réponse provient d'une loi binomiale, la régression logistique sera utilisée. Si la variable réponse est une variable de dénombrement, c'est plutôt la régression de Poisson, ou régression logarithmique, qui sera utilisée. Cette dernière méthode d'analyse se veut le sujet principal de ce mémoire.

Des explications très détaillées de l'analyse des modèles où la variable réponse en est une de dénombrement peuvent être trouvées dans les ouvrages d'Agresti (2002) et de Cameron & Trivedi (1998). On y explique, entre autres, comment estimer les p paramètres d'un tel modèle. Cette estimation se fait par l'une des deux méthodes suivantes : méthode du maximum de vraisemblance ou méthode des moments. La validation d'un modèle et l'explication des différents résidus possibles dans une régression non linéaire sont énoncés dans Cameron & Trivedi (1998).

Pour la régression non linéaire (la régression logarithmique étant un type de régression non linéaire, puisque l'effet des covariables dans le modèle est multiplicatif), la variable réponse doit provenir de la famille exponentielle. L'une des composantes d'un modèle linéaire généralisé (GLM) étant que les variables aléatoires Y_i, $i = 1, \ldots, n$, doivent être indépendantes et que leur fonction de densité doit provenir d'une famille exponentielle, une connaissance des GLM s'avère être un atout essentiel à la bonne compréhension de la régression de Poisson. Le manuel de McCullagh & Nelder (1989) est une référence populaire sur les modèles linéaires généralisés.

La régression de Poisson nécessite la prise d'une seule mesure sur chacun des n individus. Cependant, en pratique, il est très courant de faire le suivi des individus

pendant plusieurs semaines, pendant plusieurs mois, voire même pendant plusieurs années, si cela est possible. Donc, plutôt que de prendre une seule mesure sur chacun des n individus, nous prenons des mesures en n_i moments distincts dans le temps. Dans une telle situation, les paramètres du modèle de régression doivent être estimés d'une façon toute particulière puisque l'analyse est longitudinale et puisque les données sont corrélées. Les équations d'estimation généralisées (GEE) ont été introduites à cette fin par Liang & Zeger (1986a) et par Zeger & Liang (1986b). Hardin & Hilbe (2002) ont fait des GEE le sujet principal d'un livre.

Bien que les GEE servent à estimer les p paramètres du modèle, cette méthode ne permet pas de modéliser directement la corrélation entre les temps ou entre les mesures. Afin de réaliser cette tâche, des termes aléatoires sont ajoutés à une ou à plusieurs des p variables du modèle. McCulloch & Searle (2001) décrivent les modèles contenant des termes aléatoires. Quant à Wolfinger & O'Connell (1993), ils expliquent une façon d'estimer les paramètres d'un modèle contenant des effets aléatoires et dont la variable réponse fait partie de la famille exponentielle. Une macro SAS (SAS Institute Inc., 2004), et éventuellement une procédure SAS, toutes deux portant le nom de GLIMMIX, ont été créées dans le but d'analyser des modèles à effets aléatoires.

Si la transformation d'une ou de plusieurs variables est nécessaire afin d'améliorer le modèle, les modèles additifs généralisés peuvent être utilisés. Ceux-ci sont expliqués en détails dans le livre de Hastie & Tibshirani (1990). On y décrit autant les modèles additifs (utilisés dans une régression linéaire ordinaire) que les modèles additifs généralisés (utilisés dans une régression linéaire généralisée), en plus d'énoncer les propriétés respectives de chacun.

1.1 La structure du mémoire

Ce mémoire sera divisé en deux parties : une partie théorique regroupant les chapitres 2 à 6, et une partie pratique rassemblant les chapitres 7 à 9. Le chapitre 2 décrira la loi de Poisson afin de mieux expliquer par la suite la régression de Poisson. Les propriétés de cette méthode d'analyse seront donc énoncées et un exemple complet conclura le chapitre. Quant au chapitre 3, il décrira une façon d'estimer les paramètres du modèle ; les équations d'estimation généralisées (GEE) sont utilisées à cette fin. Le chapitre 4 combinera les chapitres 2 et 3 afin d'expliquer la régression de Poisson longitudinale. Dans le but de modéliser la corrélation entre les mesures, l'ajout de termes aléatoires est nécessaire, et les modèles linéaires généralisés mixtes seront introduits au chapitre 5. Finalement, le chapitre 6 expliquera brièvement les modèles

additifs généralisés.

Le premier chapitre de la partie pratique (chapitre 7) décrira les données sur les jeunes contrevenants. On y expliquera les bases de données, les variables utilisées et les analyses envisagées. Le chapitre 8 utilisera les modèles linéaires généralisés ainsi que les modèles linéaires généralisés longitudinaux afin d'analyser les données décrites au chapitre précédent. Dans le chapitre 9, ce sont les modèles linéaires généralisés mixtes qui seront utilisés pour les analyses. Un chapitre 10 viendra finalement conclure le mémoire.

1.2 Une brève explication de l'analyse effectuée dans la partie pratique

Dans les chapitres 7 à 9, nous étudierons les trajectoires criminelles de jeunes contrevenants. Ainsi, nous voulons prévoir les nombres de crimes commis par les individus une fois qu'ils auront passé l'âge de 18 ans, à partir de ce qu'ils ont commis avant l'âge de 18 ans. Donc, dans le chapitre 8, nous voulons prévoir, à l'aide de modèles linéaires généralisés longitudinaux ou non, les nombres de crimes de type « Violence », « Drogues » et le nombre de crimes de tous types commis après 18 ans à partir de ce qui a été commis avant 18 ans. Les variables explicatives seront ici les nombres de crimes « Violence », « Drogues », « Sexe », « Propriété », « tous types » et « autres types » commis entre 12 et 14 ans, entre 14 et 16 ans, et entre 16 et 18 ans.

L'analyse effectuée dans le chapitre 9 sera quelque peu différente de celle du chapitre précédent. En effet, bien que nous tenterons toujours de prévoir le nombre total de crimes commis après 18 ans à partir de l'information disponible avant 18 ans, cette prévision se fera cette fois à l'aide de modèles linéaires généralisés mixtes. De plus, les variables explicatives seront des variables indicatrices de diagnostics psychiatriques (par exemple, la présence ou l'absence de problèmes de communication, de paranoïa, d'apprentissage, etc.).

Première partie

La théorie

Chapitre 2

La régression de Poisson

Ce premier chapitre présente une introduction à la régression de Poisson. En premier lieu, la famille exponentielle sera décrite. Ensuite, la loi de Poisson et ses différentes propriétés seront énoncées. Les modèles linéaires généralisés seront introduits, pour ensuite faire place au sujet principal de ce chapitre : la régression de Poisson. Un exemple complet sera présenté pour conclure le chapitre.

2.1 La famille exponentielle

On dit d'une loi qu'elle fait partie de la famille exponentielle si sa fonction de densité (ou de probabilité) peut être réexprimée sous la forme

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

où $b(\theta)$ ne dépend pas de y et où $c(y,\phi)$ ne dépend pas du paramètre θ. Les formules d'espérance et de variance d'une loi faisant partie de la famille exponentielle seront présentées à la section A.1 de l'annexe.

Soient $y_1,\ldots,y_n$ les réalisations des variables aléatoires indépendantes $Y_1,\ldots,Y_n$, où l'on suppose que Y_i $(i=1,\ldots,n)$ a comme fonction de densité $f_Y(\theta_i|y_i;\phi)$. La fonction de vraisemblance est définie comme étant

$$\prod_{i=1}^{n} f_Y(\theta_i|y_i;\phi),$$

où n est le nombre d'observations ou d'individus. Ainsi, pour une loi faisant partie de la famille exponentielle,

$$\begin{aligned} L(\theta_i|\boldsymbol{y};\phi) &= \prod_{i=1}^{n} \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} - c(y_i,\phi)\right\} \\ &= \exp\left\{\sum_{i=1}^{n}\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} - \sum_{i=1}^{n} c(y_i,\phi)\right\}, \end{aligned} \tag{2.2}$$

où $\boldsymbol{y} = [y_1, \ldots, y_n]'$. Quant à la fonction de log-vraisemblance, elle s'obtient en prenant le logarithme naturel de la fonction de vraisemblance. Donc,

$$\ell(\theta_i|\boldsymbol{y};\phi) = \ln\{L(\theta_i|\boldsymbol{y};\phi)\} = \sum_{i=1}^{n}\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} - \sum_{i=1}^{n} c(y_i,\phi). \tag{2.3}$$

2.2 La loi de Poisson

On dit que Y suit une loi de Poisson[1] de paramètre μ si sa fonction de probabilité est

$$\mathbb{P}[Y=y] = \begin{cases} e^{-\mu}\dfrac{\mu^y}{y!}, & \text{si } y = 0,1,2,..., \\ 0, & \text{sinon,} \end{cases} \tag{2.4}$$

où μ est un nombre réel positif. De plus, on a que Y a comme fonction de répartition

$$\mathbb{P}[Y\leq y] = \begin{cases} e^{-\mu}\displaystyle\sum_{t=0}^{[y]}\frac{\mu^t}{t!}, & \text{si } y \geq 0, \\ 0, & \text{sinon,} \end{cases}$$

où $[y]$ correspond à la partie entière de y. Afin de démontrer que la loi de Poisson de paramètre μ fait partie de la famille exponentielle, on doit exprimer sa fonction de probabilité sous la forme de l'équation (2.1) en page 5. Suite à une transformation de la fonction de probabilité (2.4), on obtient

$$\begin{aligned} \mathbb{P}[Y=y] &= \frac{e^{-\mu+y\ln(\mu)}}{e^{\ln(y!)}} \\ &= \exp\left\{\frac{y\ln(\mu)-\mu}{1} - \ln(y!)\right\}. \end{aligned}$$

[1] Siméon Denis Poisson (Pithiviers 1781 - Paris 1840)

Alors, les paramètres de la famille exponentielle sont

$$\begin{aligned} \theta &= \ln(\mu) \\ b(\theta) &= \exp(\theta) = \exp(\ln(\mu)) = \mu \\ a(\phi) &= 1 \\ c(y,\phi) &= \ln(y!). \end{aligned}$$

2.2.1 Le lien entre la loi de Poisson et la loi binomiale

Ross (1999, chapitre 4, page 145) mentionne que la loi de Poisson peut être vue comme étant un résultat limite de la loi binomiale : soit $Y \sim$ binomiale(n,p) avec $n \to \infty$ et $p \to 0$, et posons $\mu = np$. Écrivons

$$\begin{aligned} \mathbb{P}[Y=y] &= \binom{n}{y} p^y (1-p)^{n-y} = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y} \\ &= \frac{n!}{y!(n-y)!} \left(\frac{np}{n}\right)^y \left(1-\frac{np}{n}\right)^{n-y} \\ &= \frac{n(n-1)\cdots(n-y+1)}{n^y} \frac{(np)^y}{y!} \frac{(1-\frac{np}{n})^n}{(1-\frac{np}{n})^y}. \end{aligned}$$

En faisant tendre $n \to \infty$ et $p \to 0$ de sorte que $np \to \mu$, on voit alors que

$$\begin{aligned} \left(1-\frac{np}{n}\right)^n \to \left(1-\frac{\mu}{n}\right)^n &\approx e^{-\mu} \\ \frac{n(n-1)\cdots(n-y+1)}{n^y} &\approx 1 \\ \left(1-\frac{np}{n}\right)^y \to \left(1-\frac{\mu}{n}\right)^y &\approx 1. \end{aligned}$$

Donc, $\mathbb{P}[Y=y] = \dfrac{\mu^y}{y!} e^{-\mu}$, qui est la fonction de probabilité d'une loi de Poisson de paramètre μ. □

2.2.2 La fonction génératrice des moments de la loi de Poisson

La fonction génératrice des moments de la loi de Poisson, notée $M_Y(t)$, est utile afin de trouver les moments d'ordre k, où $\mathbb{E}[Y^k] = M_Y^{(k)}(t)\Big|_{t=0}$. On définit cette fonction

génératrice des moments comme étant $\mathbb{E}[e^{tY}]$. On a

$$\begin{aligned} M_Y(t) = \mathbb{E}[e^{tY}] &= \sum_{y=0}^{\infty} e^{ty}\frac{e^{-\mu}\mu^y}{y!} = e^{-\mu}\sum_{y=0}^{\infty}\frac{(\mu e^t)^y}{y!} \\ &= \frac{e^{-\mu}}{e^{-\mu e^t}}\underbrace{\sum_{y=0}^{\infty} e^{-\mu e^t}\frac{(\mu e^t)^y}{y!}}_{=1,\ \text{car Poisson}(\mu e^t)} = e^{-\mu+\mu e^t} \\ &= e^{\mu(e^t-1)}. \end{aligned} \tag{2.5}$$

À l'aide de cette fonction génératrice des moments, l'espérance et la variance de la loi peuvent être calculées :

$$\mathbb{E}[\,Y] = \mu \quad \text{et} \quad \mathbb{V}\text{ar}[\,Y] = \mathbb{E}[Y^2] - \mathbb{E}^2[Y] = \mu.$$

En sachant que la loi de Poisson fait partie de la famille exponentielle, l'espérance et la variance auraient pu être trouvées à l'aide des paramètres de cette famille exponentielle[2]. En effet, on a

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

Alors,

$$\mathbb{E}[Y] = \frac{d}{d\theta}\exp(\theta) = \exp(\theta) = \mu,$$

comme on l'a obtenu précédemment avec la fonction génératrice des moments. De plus,

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

Donc, on obtient

$$\mathbb{V}\text{ar}[Y] = \frac{d^2}{d\theta^2}\exp(\theta) \times 1 = \exp(\theta) = \mu,$$

comme il a également été obtenu avec la fonction génératrice des moments.

On obtient donc une propriété intéressante de la loi de Poisson, appelée propriété d'équidispersion, impliquant que $\mathbb{E}[\,Y] = \mathbb{V}\text{ar}[\,Y]$. Cameron & Trivedi (1998, chapitre 1, page 4) mentionnent qu'une loi est équidispersée dans le cas où son espérance et sa variance sont égales ; elle est surdispersée (sousdispersée) dans le cas où son espérance est inférieure (supérieure) à sa variance. Les propriétés de dispersion des lois discrètes communes sont présentées au Tableau 2.1 de la page suivante.

On mentionne que si $Y_j \sim \text{Poisson}(\mu_j)$ $(j = 1, 2, ...)$, que les Y_j sont des variables aléatoires indépendantes et que $\sum_{j=1}^{\infty}\mu_j < \infty$, alors $Z_Y = \sum_{j=1}^{\infty} Y_j \sim \text{Poisson}\left(\sum_{j=1}^{\infty}\mu_j\right)$.

[2]Les démonstrations des formules 2.6 et 2.7 se trouvent à la section A.1 de l'annexe.

Loi	Espérance	Variance	Propriété
Binomiale(n,p)	np	np(1-p)	Sousdispersion si $0<p\leq 1$ Équidispersion si $p=0$
Binomiale négative(m,p)	$\frac{m}{p}$	$\frac{m(1-p)}{p^2}$	Sousdispersion si $p>\frac{1}{2}$ Équidispersion si $p=\frac{1}{2}$ Surdispersion si $p<\frac{1}{2}$
Poisson(μ)	μ	μ	Équidispersion
Uniforme discrète sur $\{1,\ldots,n\}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	Sousdispersion si $n\leq 6$ Équidispersion si $n=7$ Surdispersion si $n\geq 8$

TAB. 2.1 – *Propriétés des lois discrètes les plus communes.*

2.2.3 Les fonctions de vraisemblance et de log-vraisemblance de la loi de Poisson

Supposons $Y_1,\ldots,Y_n$ des observations mutuellement indépendantes telles que $Y_i \sim$ Poisson(μ_i), $i=1,\ldots,n$. Ainsi, on a $\theta_i=\ln(\mu_i)$, $b(\theta_i)=\mu_i$, $a(\phi)=1$ et $c(y_i,\phi)=\ln(y_i!)$. En substituant ces valeurs dans les équations (2.2) et (2.3) de la page 6, on obtient

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{y};\phi) &= \exp\left\{\sum_{i=1}^{n}\frac{y_i\ln(\mu_i)-\mu_i}{1}-\sum_{i=1}^{n}\ln(y_i!)\right\} \\ &= \exp\left\{\sum_{i=1}^{n}(y_i\ln(\mu_i)-\mu_i-\ln(y_i!))\right\} \\ \text{et } \ell(\boldsymbol{\theta}|\boldsymbol{y};\phi) &= \sum_{i=1}^{n}\left\{y_i\ln(\mu_i)-\mu_i-\ln(y_i!)\right\}, \end{aligned}$$

où $\boldsymbol{\theta}=[\ln(\mu_1),\ldots,\ln(\mu_n)]'=[\theta_1,\ldots,\theta_n]'$.

2.3 Les modèles linéaires généralisés

Un modèle linéaire généralisé (GLM)[3] est une extension du modèle de régression linéaire, mais en permettant à la variable réponse, Y, de suivre n'importe quelle loi faisant partie de la famille exponentielle (section 2.1 de la page 5).

[3] *Generalized Linear Model*

2.3.1 Quand utilise-t-on un GLM plutôt qu'un modèle linéaire ?

Dans le cas où la variable réponse n'est pas continue, ses valeurs attendues doivent aussi être discrètes et suivre la même distribution. Dans ce cas, un modèle linéaire n'est pas approprié.

Une autre raison expliquant le fait qu'un modèle linéaire n'est pas approprié est tout simplement que l'effet entre la variable indépendante et la (les) variable(s) dépendante(s) n'est pas linéaire. Une fonction de lien adéquatement choisie permet de mieux modéliser l'effet des variables exogènes sur la variable réponse.

2.3.2 Les composantes d'un GLM

Dans un GLM, certaines composantes doivent être présentes. En fait, les trois composantes suivantes sont nécessaires :

1. Les variables aléatoires Y_i ($i = 1, \ldots, n$), que l'on suppose indépendantes, ont comme valeur espérée $\mathbb{E}[Y_i] = \mu_i$ et leur fonction de densité (ou de probabilité) fait partie de la famille exponentielle ;
2. Un prédicteur linéaire :
   $\eta_i = \boldsymbol{x_i'}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p'} x_{ip'}$ car $\boldsymbol{x_i'} = [1, x_{i1}, \ldots, x_{ip'}]$ et $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_{p'}]'$;
3. Une fonction de lien, $g(\mu_i)$, décrivant la relation entre $\mathbb{E}[Y_i] = \mu_i$ et $\boldsymbol{x_i'}\boldsymbol{\beta}$. Dans la plupart des cas, $g(\cdot)$ est une fonction de lien monotone, différentiable, connue et non linéaire. On dénote son inverse par $g^{-1}(\cdot)$.

2.4 La régression de Poisson

La régression de Poisson est utilisée dans le cas où la variable réponse, Y_i, est une variable de *dénombrement*. Soit une réponse Y_i et un vecteur de régresseurs $\boldsymbol{x_i}$. On a

$$\mathbb{P}[Y_i = y_i | \boldsymbol{x_i}] = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, i = 1, \ldots, n, \tag{2.8}$$

où $g(\mu_i) = \boldsymbol{x_i'}\boldsymbol{\beta}$ et où le lien logarithmique, c'est-à-dire $\ln(\mu_i) = \boldsymbol{x_i'}\boldsymbol{\beta}$, est le lien le plus commun en régression de Poisson. De plus, $\boldsymbol{\beta}$ est un vecteur de $p = p' + 1$ composantes à estimer.

2.4.1 Les variables *offset*

Souvent, en régression de Poisson, des variables *offset* sont utilisées lorsque la variable de dénombrement, Y, est proportionnelle à une certaine autre variable exogène que l'on veut inclure dans le modèle (dans le prédicteur linéaire). On n'estime pas le coefficient β de cette nouvelle variable exogène : on le force plutôt à prendre la valeur unitaire.

Prenons l'exemple d'une régression où l'on voudrait prévoir le nombre de voitures dans un stationnement. La variable réponse est donc le nombre de voitures retrouvées dans un stationnement en particulier. On pourrait inclure en variable *offset* la superficie du stationnement, z_i, puisque plus le stationnement est grand, plus il est susceptible de loger un grand nombre de voitures. Ainsi, on pourrait utiliser $\mu_i = e^{\boldsymbol{x_i'\beta}+\ln(z_i)}$. Ici, $\boldsymbol{x_i}$ contient toutes les variables exogènes d'intérêt, à l'exception de la superficie du stationnement. Alors, $\mu_i = z_i e^{\boldsymbol{x_i'\beta}} \propto z_i$. Donc, si la superficie du stationnement est multipliée par une constante c quelconque, on obtient $\mu_i = e^{\boldsymbol{x_i'\beta}+\ln(cz_i)} = e^{\boldsymbol{x_i'\beta}+\ln(c)+\ln(z_i)} = ce^{\boldsymbol{x_i'\beta}+\ln(z_i)}$. La moyenne, μ_i, est donc multipliée par c et elle est alors proportionnelle à la grandeur du stationnement, tel qu'il avait été supposé au préalable.

2.4.2 Le modèle de la régression de Poisson

Supposons n observations indépendantes d'une variable réponse Y_i $(i = 1, \ldots, n)$ et p' variables explicatives pour ces n variables réponses. De plus, supposons $Y_i|\boldsymbol{x_i} \sim$ Poisson(μ_i) et que la fonction de lien est $g(\mu_i) = \ln(\mu_i)$.

On tentera donc d'estimer μ_i, l'espérance de la variable réponse. L'estimation des différents β_k $(k = 0, \ldots, p')$ est généralement faite par la méthode du maximum de vraisemblance. Cependant, en pratique, l'emploi de méthodes numériques d'un logiciel sera nécessaire. Avec SAS, la procédure GENMOD, qui s'appuie sur la méthode du maximum de vraisemblance afin d'estimer les paramètres, peut être utilisée.

2.4.3 L'interprétation des paramètres $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$

Les paramètres $\hat{\beta}_k$ ont une interprétation particulière sous le lien logarithmique. Ainsi, $\hat{\beta}_0$ représente le logarithme naturel de l'espérance de la variable réponse lorsque

les p' variables exogènes prennent simultanément la valeur 0 :

$$\begin{aligned} \hat{\mu}_i &= e^{\hat{\beta}_0+(\hat{\beta}_1\times 0)+(\hat{\beta}_2\times 0)+\cdots+(\hat{\beta}_{p'}\times 0)} = e^{\hat{\beta}_0} \\ \Rightarrow \hat{\beta}_0 &= \ln(\hat{\mu}_\mathrm{i}). \end{aligned}$$

Quant aux paramètres $\hat{\beta}_1, \ldots, \hat{\beta}_{p'}$, si on augmente $x_{i\ell}$ ($\ell < p'$) d'une unité et que l'on maintient constante la valeur des autres variables exogènes, alors la valeur moyenne de Y_i est multipliée par $e^{\hat{\beta}_\ell}$:

$$\begin{aligned} \hat{\mu}_i &= e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}+\cdots+\hat{\beta}_{p'} x_{ip'}} \\ \hat{\mu}_i^* &= e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}+\cdots+\hat{\beta}_\ell (x_{i\ell}+1)+\cdots+\hat{\beta}_{p'} x_{ip'}} = e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}+\cdots+\hat{\beta}_{p'} x_{ip'}} e^{\hat{\beta}_\ell} \\ &= \hat{\mu}_i e^{\hat{\beta}_\ell}. \end{aligned}$$

2.4.4 Le problème d'absence d'équidispersion

Il a déjà été mentionné à la sous-section 2.2.2 de la page 7 que l'une des propriétés de la loi de Poisson est l'égalité entre son espérance et sa variance, condition appelée propriété d'équidispersion. Cependant, il arrive fréquemment que la variance de la variable réponse soit supérieure à son espérance. On parle alors d'un problème de surdispersion. Afin de tester si les données sont surdispersées ou équidispersées, l'une des deux statistiques ϕ_P ou ϕ_D, où

$$\begin{aligned} \phi_P &= \frac{\mathcal{X}^2 \text{ de Pearson}}{\text{nombre de degrés de liberté}} \\ \phi_D &= \frac{\text{Statistique de déviance}}{\text{nombre de degrés de liberté}} \end{aligned}$$

doit être calculée. Les statistiques du khi-deux de Pearson et de déviance seront définies à la section 2.6 en page 17. Dans le cas où $\phi >> 1$, il y a surdispersion et si $\phi \approx 1$, il y a équidispersion. Lorsqu'il y a surdispersion dans les données (aussi appelée variabilité extra-poisonnienne), deux solutions s'offrent : continuer les analyses avec la loi binomiale négative, ou tenir compte de la surdispersion en modifiant les résultats obtenus en divisant les statistiques du khi-deux par $\hat{\phi}$ et en multipliant les variances et les covariances par $\hat{\phi}$.

Si nous devons travailler avec une loi binomiale négative, alors $\mathbb{E}[Y|\mu,\alpha] = \mu$ et $\mathbb{V}\mathrm{ar}[Y|\mu,\alpha] = \mu(1+\alpha\mu)$ (Cameron & Trivedi, 1998, chapitre 3, page 62). Donc, le test d'hypothèses suivant peut être fait :

$H_0 : \alpha = 0$ (La loi de Poisson est appropriée, alors il y a équidispersion)
$H_1 : \alpha > 0$ (La loi de Poisson n'est pas appropriée, alors la loi binomiale négative l'est, et il y a surdispersion dans les données). Afin de faire ce test d'hypothèses, un seuil s doit être calculé, où $s = \frac{1}{2}\mathbb{P}[\mathcal{X}_1^2 > q]$, $q = 2(\ell_{BN} - \ell_P)$, ℓ_{BN} est la valeur de la log-vraisemblance obtenue en ajustant un modèle binomial négatif aux données et l_P est la valeur de la log-vraisemblance obtenue en ajustant un modèle de Poisson aux données.

2.5 Les différents types de résidus d'une régression de Poisson

En régression linéaire classique, les résidus sont définis comme étant la différence entre la valeur observée de la variable endogène et sa valeur prédite par le modèle obtenu. Dans ce cas, les résidus sont indépendants et identiquement distribués, de moyenne nulle et de variance constante, σ^2. Cependant, lorsque la variable réponse en est une de dénombrement, les résidus $Y_i - \hat{\mu}_i$ ne sont pas de même variance et proviennent d'une distribution asymétrique. Ainsi, aucun résidu ne provient d'une distribution symétrique et n'est de moyenne nulle et de variance constante. Trois types de résidus ont été définis pour remédier à ce problème : les résidus de Anscombe, les résidus de Pearson et les résidus de déviance. Ces trois types de résidus seront respectivement définis aux sous-sections 2.5.1, 2.5.2 et 2.5.4 suivantes, alors que leur utilisation est plutôt décrite à la section 2.6 en page 17.

2.5.1 Les résidus de Anscombe

Les résidus de Anscombe sont définis comme étant la transformation de Y s'approchant le plus d'une loi normale centrée réduite. Dans le cas où Y suit une loi de Poisson, $Y^{\frac{2}{3}}$ est la transformation se rapprochant le plus de la loi normale centrée réduite (McCullagh & Nelder, 1989, chapitre 2, page 38). Ainsi, on définit les résidus de Anscombe, r_{A_i}, comme étant

$$r_{A_i} = 1.5\frac{(Y_i^{\frac{2}{3}} - \hat{\mu}_i^{\frac{2}{3}})}{\hat{\mu}_i^{\frac{1}{6}}}. \tag{2.9}$$

2.5.2 Les résidus de Pearson

Les résidus de Pearson, r_{P_i}, sont quant à eux utilisés lorsque la propriété qui nous intéresse est l'homoscédasticité. Ceux-ci sont définis comme étant

$$r_{P_i} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{w}_i}}.$$

On note que $\hat{w}_i$ est un estimé de la variance de Y_i. Dans le cas de la loi de Poisson, $\hat{w}_i = \hat{\mu}_i$, et alors

$$r_{P_i} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}. \tag{2.10}$$

Pour de grandes tailles d'échantillons, les résidus r_{P_i} sont de moyenne nulle et de variance unitaire. Cependant, leur distribution est asymétrique. À l'opposé, pour de petites tailles d'échantillons, les résidus de Pearson *studentisés* sont utilisés. Ceux-ci sont définis comme étant

$$r^s_{P_i} \quad = \quad \frac{r_{P_i}}{\sqrt{1 - h_{ii}}} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(1 - h_{ii})}}, \tag{2.11}$$

où h_{ii} est l'élément en position (i, i) de la matrice chapeau (matrice $\mathbf{H}$) définie ci-dessous.

2.5.3 La matrice chapeau, H

Pour des modèles linéaires généralisés, la matrice $\mathbf{H}$ est

$$\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}'\mathbf{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{W}^{\frac{1}{2}}, \tag{2.12}$$

où $\mathbf{W}$ est une matrice diagonale de dimension $n \times n$ ayant w_i comme élément en position (i, i). Nous avons de plus que

$$w_i = \frac{\left(\frac{\partial \mu_i}{\partial \boldsymbol{x_i'}\boldsymbol{\beta}}\right)^2}{\mathbb{V}\mathrm{ar}[Y_i]}.$$

Pour une loi de Poisson, il est mentionné à la sous-section 2.5.2 que $w_i = \mu_i$. Finalement, la matrice $\boldsymbol{X}$ est appelée la matrice de schéma :

$$\boldsymbol{X}_{(n\times p)} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p'} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p'} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p'} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np'} \end{pmatrix}.$$

La matrice chapeau possède les propriétés d'idempotence ($\mathbf{HH} = \mathbf{H}$), de symétrie ($\mathbf{H'} = \mathbf{H}$) ; de plus, sa trace correspond au nombre de paramètres à estimer, p. Les démonstrations de ces propriétés seront faites à la section A.2 de l'annexe. En sachant que la trace de la matrice $\mathbf{H}$ est p, on conclut par définition que

$$\sum_{i=1}^{n} h_{ii} = p.$$

2.5.4 Les résidus de déviance

Les résidus de déviance, r_{D_i}, sont souvent utilisés lorsque la variable réponse, Y, provient d'une loi faisant partie de la famille exponentielle. Ils sont obtenus à l'aide de la fonction de log-vraisemblance et sont définis comme étant

$$r_{D_i} = \text{signe}(Y_i - \hat{\mu}_i)\sqrt{2a(\phi)\{\ell(\boldsymbol{Y}|\boldsymbol{Y};\phi) - \ell(\hat{\boldsymbol{\mu}}_{\boldsymbol{i}}|\boldsymbol{Y};\phi)\}}.$$

Pour une loi de Poisson,

$$\begin{aligned} r_{D_i} &= \text{signe}(Y_i - \hat{\mu}_i)\sqrt{2\{-Y_i + Y_i\ln(Y_i) - \ln(Y_i!) + \hat{\mu}_i - Y_i\ln(\hat{\mu}_i) + \ln(Y_i!)\}} \\ &= \text{signe}(Y_i - \hat{\mu}_i)\sqrt{2\left\{Y_i\ln\left(\frac{Y_i}{\hat{\mu}_i}\right) - (Y_i - \hat{\mu}_i)\right\}}. \end{aligned} \tag{2.13}$$

Tout comme pour les résidus de Pearson, on peut utiliser les résidus de déviance studentisés lorsque la taille d'échantillon est petite. Ceux-ci sont obtenus à l'aide de l'expression suivante :

$$r^s_{D_i} = \frac{r_{D_i}}{\sqrt{1-h_{ii}}} = \frac{\text{signe}(Y_i - \hat{\mu}_i)\sqrt{2\left\{Y_i\ln\left(\frac{Y_i}{\hat{\mu}_i}\right) - (Y_i - \hat{\mu}_i)\right\}}}{\sqrt{1-h_{ii}}}. \tag{2.14}$$

2.5.5 La comparaison des trois types de résidus

McCullagh & Nelder (1989, chapitre 2, page 39) ont réexprimé les 3 types de résidus (Anscombe, Pearson et déviance) sous une forme faisant intervenir $c = Y/\hat{\mu}$:

$$\begin{aligned} r_{A_i} &= \sqrt{\hat{\mu}} \times 1.5(c^{\frac{2}{3}} - 1) \propto 1.5(c^{\frac{2}{3}} - 1) \\ r_{P_i} &= \sqrt{\hat{\mu}} \times c - 1 \propto c - 1 \\ r_{D_i} &= \sqrt{\hat{\mu}} \times \{2(c\ln(c) - c + 1)\}^{\frac{1}{2}} \propto \{2(c\ln(c) - c + 1)\}^{\frac{1}{2}} \end{aligned}$$

Les trois types de résidus en fonction de c sont représentés à la FIGURE 2.1 suivante. On remarque que chacun des 3 types de résidus vaut 0 lorsque $c = 1$ (lorsque $Y = \hat{\mu}$), et ils augmentent lorsque c croît. De cette figure, on voit que les résidus de Anscombe et de déviance prennent des valeurs semblables pour des valeurs données de c. Cependant, les résidus de Pearson prennent des valeurs beaucoup plus grandes pour les mêmes valeurs données de c.

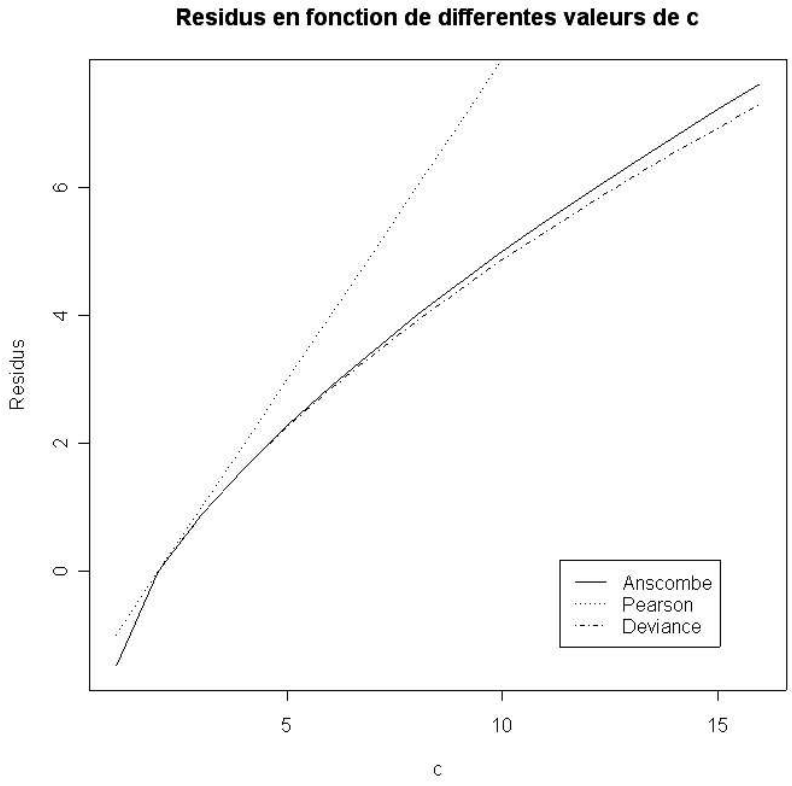

FIG. 2.1 – *Représentation des 3 types de résidus.*

2.6 La validation d'un modèle

2.6.1 La statistique de Pearson

La statistique de Pearson, notée $\mathcal{X}^2$, mesure l'ajustement d'un modèle. Cette statistique est obtenue en sommant les carrés des résidus de Pearson définis à l'équation (2.10) en page 14. Si le modèle que l'on tente d'ajuster aux données est bon, cette statistique devrait suivre approximativement une loi du khi-deux avec $n - p$ degrés de liberté.

La statistique de Pearson peut aussi être utilisée afin de déceler des problèmes avec l'hypothèse d'équidispersion. En fait, si $\mathcal{X}^2 >> n - p$, il y a lieu de penser que les données sont surdispersées.

2.6.2 La statistique de déviance

La statistique de déviance, notée $D(\boldsymbol{Y}, \hat{\boldsymbol{\mu}})$, est obtenue en sommant les carrés des résidus de déviance de l'équation (2.13) en page 15. Elle peut être vue comme une généralisation de la somme des carrés résiduels communément utilisée en régression linéaire. La démonstration sera faite à la section A.3 de l'annexe. Si le modèle à l'étude s'ajuste bien aux données, la statistique de déviance suivra, tout comme la statistique de Pearson, approximativement une loi du khi-deux avec $n-p$ degrés de liberté. Les résidus d'une régression de Poisson somment à 0 si une ordonnée à l'origine est incluse dans le modèle et si la fonction de lien logarithmique est utilisée. Dans ce cas, la statistique de déviance est obtenue plus facilement par

$$\sum_{i=1}^{n} Y_i \ln\left(\frac{Y_i}{\hat{\mu}_i}\right).$$

2.6.3 La validation graphique d'un modèle

Différents graphiques peuvent aussi être faits afin de valider le modèle. McCullagh & Nelder (1989, chapitre 12, page 398) suggèrent de faire un graphique des résidus (de déviance ou de Pearson) en fonction d'une variable exogène incluse dans le prédicteur linéaire afin de voir si cette variable exogène devrait entrer dans le modèle sous une autre

forme. On devrait voir sur ce graphique des résidus de moyenne nulle et de variance constante. Un autre graphique possible est celui des résidus de déviance standardisés en fonction de $2\sqrt{\hat{\mu}_i}$. Encore une fois, on devrait voir des résidus de moyenne nulle et de variance constante.

Cameron & Trivedi (1998, chapitre 5, page 144) proposent quant à eux de représenter les résidus en fonction des valeurs prédites afin de voir si l'ajustement du modèle est bon pour certaines valeurs extrêmes de la variable dépendante.

2.7 La validation croisée

Afin de vérifier le pouvoir prédictif d'un modèle de régression, la validation croisée est souvent utilisée. Cette méthode consiste, pour chaque i ($i = 1, \ldots, n$), à enlever l'observation i de la base de données, à estimer l'espérance de la variable réponse Y_i du modèle obtenu avec les $n-1$ observations restantes (on obtient ainsi $\hat{\mu}_{(i,-i)}$) et à calculer la statistique PRESS[4], où

$$\text{PRESS} = \sum_{i=1}^{n} (Y_i - \hat{\mu}_{(i,-i)})^2.$$

Plus le modèle est en mesure de bien prévoir la valeur espérée de la variable réponse, plus la valeur de $Y_i - \hat{\mu}_{(i,-i)}$ sera faible et donc plus la statistique PRESS sera faible. Ce critère sera utilisé plus tard pour comparer plusieurs modèles entre eux. Un avantage de son utilisation est de permettre la comparaison de plusieurs types de modèles entre eux, tels des modèles de nature différente, imbriqués ou non, etc. Cependant, si les modèles ne font pas intervenir le même nombre d'individus (n différents), nous utiliserons plutôt PRESS/n pour comparer les modèles entre eux.

2.8 Un exemple complet sur la régression de Poisson

L'exemple suivant est tiré de Neter (1990). Dans cet exemple, on cherche à prévoir le nombre de clients chez un détaillant de bois, peinture, appareils électriques et autres

[4] *Predicted Residual Errors Sum of Squares*

(*The Miller Lumber Company*). Pendant une période de 2 semaines, un sondage a été effectué et les adresses des consommateurs ont été obtenues. Ces adresses ont été utilisées afin de connaître le secteur de recensement auquel le consommateur appartient. À la fin de la période d'étude, le nombre total de clients dans chacun des secteurs de recensement a été obtenu, en plus de plusieurs autres variables indépendantes, telles que le nombre total de personnes dans l'ensemble des ménages du secteur de recensement i, le salaire moyen en dollars, l'âge moyen des personnes dans les ménages du secteur de recensement i, la distance au compétiteur le plus près en miles[5] et la distance au magasin en miles. On tentera d'ajuster un modèle de Poisson aux données illustrées au Tableau 2.2.

Secteur de recensement i	Nombre de clients Y	Nombre de personnes X_1	Salaire moyen X_2	Âge moyen X_3	Distance au compétiteur X_4	Distance au magasin X_5
1	9	606	41 393	3	3.04	6.32
2	6	641	23 635	18	1.95	8.89
3	28	505	55 475	27	6.54	2.05
⋮	⋮	⋮	⋮	⋮	⋮	⋮
108	6	817	54 429	47	1.90	9.90
109	4	268	34 022	54	1.20	9.51
110	6	519	52 850	43	2.92	8.62

Tab. 2.2 – *Extrait des données pour l'exemple « The Miller Lumber Company ».*

2.8.1 Une analyse descriptive des données

Le Tableau 2.3 résume les principales statistiques descriptives d'intérêt.

Variable	Moyenne	Écart-type	Minimum	Maximum
Y	11.2	6.64	0	32
X_1	647.76	263.03	19	1289
X_2	48836.78	18531.06	19673	120065
X_3	27.43	16.68	1	58
X_4	3.07	1.50	0.34	6.61
X_5	6.83	2.29	0.87	9.90

Tab. 2.3 – *Statistiques descriptives.*

[5] 1 mile = 1.609 km

2.8.2 L'estimation des paramètres

Étant donné que l'on suppose $Y_1|\boldsymbol{x_1}, \ldots, Y_n|\boldsymbol{x_n}$ indépendantes avec

$$f(Y_i|\boldsymbol{x_i}) = \frac{e^{-\mu_i}\mu_i^{Y_i}}{Y_i!}, i = 1, \ldots, n, \tag{2.15}$$

où $\mu_i = e^{\boldsymbol{x_i'\beta}}$ et où $\boldsymbol{x_i} = [1, x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}]'$, la méthode du maximum de vraisemblance donne

$$\begin{aligned} L(\boldsymbol{\beta}|\boldsymbol{Y}) &= \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{Y_i}}{Y_i!} \\ \Rightarrow \ell(\boldsymbol{\beta}|\boldsymbol{Y}) = \ln(L(\boldsymbol{\beta}|\boldsymbol{Y})) &= \sum_{i=1}^{n}[-\mu_i + Y_i\ln(\mu_i) - \ln(Y_i!)] \\ &= \sum_{i=1}^{n}[-e^{\boldsymbol{x_i'\beta}} + Y_i\boldsymbol{x_i'\beta} - \ln(Y_i!)] \\ &= \sum_{i=1}^{n}[Y_i - e^{\boldsymbol{x_i'\beta}}]\boldsymbol{x}_i. \end{aligned} \tag{2.16}$$

Puisqu'on a un système d'équations non linéaires en $\beta_0, \ldots, \beta_{p'}$, on doit le résoudre avec l'aide d'une méthode numérique. Ceci sera fait à l'aide du logiciel SAS, qui utilise une version de l'algorithme de Newton-Raphson.

2.8.3 L'analyse du jeu de données avec la procédure GENMOD

La procédure GENMOD de SAS (SAS Institute Inc., 2004) est utilisée, entre autres, pour ajuster un modèle décrivant la relation entre une variable réponse de dénombrement et d'autres variables indépendantes à l'aide d'une régression de Poisson. Cette procédure utilise un GLM afin d'estimer, par la méthode du maximum de vraisemblance, les composantes du vecteur $\boldsymbol{\beta}$.

Les résultats obtenus

La première étape est d'ajuster un modèle contenant toutes les variables et de vérifier s'il y a lieu de croire que les données sont surdispersées. Ici, nous obtenons les résultats suivants :

	Poisson			Binomiale négative		
Critère	ddl	Valeur	$\frac{\text{Valeur}}{\text{ddl}}$	ddl	Valeur	$\frac{\text{Valeur}}{\text{ddl}}$
Déviance	104	114.9854	1.1056	104	120.1973	1.1557
$\mathcal{X}^2$ de Pearson	104	101.8808	0.9796	104	107.0003	1.0288
Log-vraisemblance		1898.0224			1898.1121	

TAB. 2.4 – *Résultats obtenus au moyen de SAS.*

Ainsi, $q = 2(l_{BN} - l_P) = 2(1898.1121 - 1898.0224) = 0.0897$ et $s = \frac{1}{2}\mathbb{P}[\mathcal{X}_1^2 > 0.0897] = 0.3823$. Nous concluons donc que les données sont équidispersées. De plus, nous avons que $\phi_P = \frac{101.8808}{104} = 0.9796$ et $\phi_D = \frac{114.9854}{104} = 1.1056$, ce qui implique aussi que les données semblent équidispersées. Les analyses seront donc faites avec la loi de Poisson et les 104 degrés de liberté correspondent aux 110 observations moins les 6 paramètres à estimer (β_0 à β_5). L'étape suivante consiste à sélectionner les variables ayant un effet important. Nous avons, pour les données à l'étude,

Source	ddl	Khi-Carré	Pr> $\mathcal{X}^2$
X_1	1	18.20	<0.0001
X_2	1	31.79	<0.0001
X_3	1	4.38	0.0364
X_4	1	41.66	<0.0001
X_5	1	67.50	<0.0001

TAB. 2.5 – *Seuils associés à chacune des variables de l'exemple.*

Donc, au seuil $\alpha = 5\%$, toutes les variables peuvent entrer dans le modèle. Les interactions et les termes de degrés supérieurs seront donc inclus. Des termes de degrés supérieurs et des interactions sont non significatives et seront donc éliminés jusqu'à l'obtention des résultats suivants :

Paramètre	ddl	Estimé	Erreur Standard	Intervalles de confiance (95%)		Khi-Carré	Pr> $\mathcal{X}^2$
Intercept	1	3.7536	0.3371	3.0929	4.4143	123.99	<.0001
X_1	1	0.0006	0.0001	0.0003	0.0009	19.02	<.0001
X_2	1	-0.0000	0.0000	-0.0000	-0.0000	24.32	<.0001
X_3	1	-0.0052	0.0054	-0.0158	0.0054	0.92	0.3370
X_4	1	0.0797	0.0644	-0.0465	0.2060	1.53	0.2159
X_5	1	-0.2549	0.0413	-0.3359	-0.1738	38.01	<.0001
X_2X_3	1	0.0000	0.0000	0.0000	0.0000	5.95	0.0147
X_3X_4	1	-0.0034	0.0014	-0.0062	-0.0006	5.81	0.0159
X_4X_5	1	0.0278	0.0085	0.0111	0.0444	10.67	0.0011
Scale	0	1.0000	0.0000	1.0000	1.0000		

TAB. 2.6 – *Seuils associés à chacune des variables de l'exemple et aux interactions les faisant intervenir.*

Comme la procédure d'exclusion des variables doit se terminer, nous concluons avec ce modèle :

$$\ln(\hat{\mu}_i) = 3.7536 + 0.0006X_{i1} - 0.0052X_{i3} + 0.0797X_{i4} - 0.2549X_{i5} - 0.0034X_{i3}X_{i4} + 0.0278X_{i4}X_{i5} \tag{2.17}$$

L'interprétation des paramètres obtenus

Selon le modèle (2.17), les interprétations suivantes peuvent être faites :

1. Si les 4 variables impliquées dans le modèle sont nulles, l'espérance du nombre moyen de clients sera $e^{3.7536} \cong 43$.
2. Si X_1 augmente de 1 unité, l'espérance du nombre de clients ne sera pratiquement pas affectée, en autant que les autres variables restent constantes. Cependant, si on augmente X_1 de 100 unités, l'espérance du nombre de clients augmente de 6%, environ.
3. Si X_3 augmente d'une année et que les autres variables restent inchangées, l'espérance du nombre de clients sera multipliée par $e^{-0.0052-0.0034X_4}$. Ainsi, si $X_4 = 6.5$ miles, l'espérance du nombre de clients sera diminuée de 2.7%, environ.
4. Si X_4 augmente de 1 mile et que les autres variables restent constantes, l'espérance du nombre de clients sera multipliée par $e^{0.0797-0.0034X_3+0.0278X_5}$. Ainsi, si $X_3 = 50$ ans et $X_5 = 8.5$ miles, l'espérance du nombre de clients sera augmentée de 15.7%, environ.
5. Si X_5 augmente de 1 mile et que les autres variables ne varient pas, l'espérance

du nombre de clients sera multipliée par $e^{-0.2549+0.0278X_4}$. Ainsi, si $X_4 = 6.5$ miles, l'espérance du nombre de clients sera diminuée de 7.2%, environ.

La validation du modèle obtenu

On se doit maintenant de vérifier la validité du modèle. On obtient une statistique de Pearson de 91.3535 avec 101 degrés de liberté[6]. Donc, $\mathbb{P}[\mathcal{X}^2_{101} > 91.3535] = 0.7435$, ce qui suppose que le modèle est acceptable. De même, la statistique de déviance est de 99.2447 avec 101 degrés de liberté, ce qui confirme que le modèle est raisonnable (seuil de 0.5308). Quant à la statistique PRESS, elle a une valeur de 1169.81. Si aucune interaction n'avait été ajoutée au modèle, une statistique PRESS de 1344.50 aurait été obtenue : le modèle avec interaction prédit donc mieux que le modèle sans interaction.

D'un côté graphique, la FIGURE 2.2 en page suivante montre que le modèle obtenu est bon. En effet, les graphiques des résidus de déviance en fonction des variables exogènes incluses dans le prédicteur linéaire (graphiques 1 à 5) montrent que les variables sont entrées sous la bonne forme et qu'aucune transformation n'aurait été nécessaire. De plus, sur ces graphiques, les points sont placés autour de 0 de façon symétrique. Quant au graphique 6 (résidus de déviance en fonction de $2\sqrt{\hat{\mu}_i}$), on doit y voir un nuage de points centré autour de 0, ce qui est le cas. Finalement, sur le graphique 7 (résidus de déviance en fonction des valeurs prédites), on voit un nuage de points sans patron particulier, ce qui pousse toujours à croire que le modèle trouvé est valide.

2.9 Conclusion du chapitre

Ce premier chapitre a eu pour but d'introduire la régression de Poisson. Tout au long de ce chapitre, il a été supposé que $Y_1|\boldsymbol{x_1}, \ldots, Y_n|\boldsymbol{x_n}$ étaient indépendantes. Cependant, il arrive souvent que l'on cherche à prédire une variable réponse Y_{it} en fonction d'un vecteur $p \times 1$ de plusieurs covariables x_{it} mesurées aux temps $t = 1, \ldots, n_i$ pour les individus $i = 1, \ldots, n$. On parle dans ce cas de régression longitudinale (ou de données longitudinales). Il y a lieu de croire que les observations réalisées sur un même sujet sont corrélées. Dans ce cas, on doit modifier les procédures d'inférences vues dans ce chapitre afin de tenir compte de la corrélation potentielle entre les mesures.

[6] Les 101 degrés de liberté correspondent aux 110 données moins les 9 paramètres à estimer.

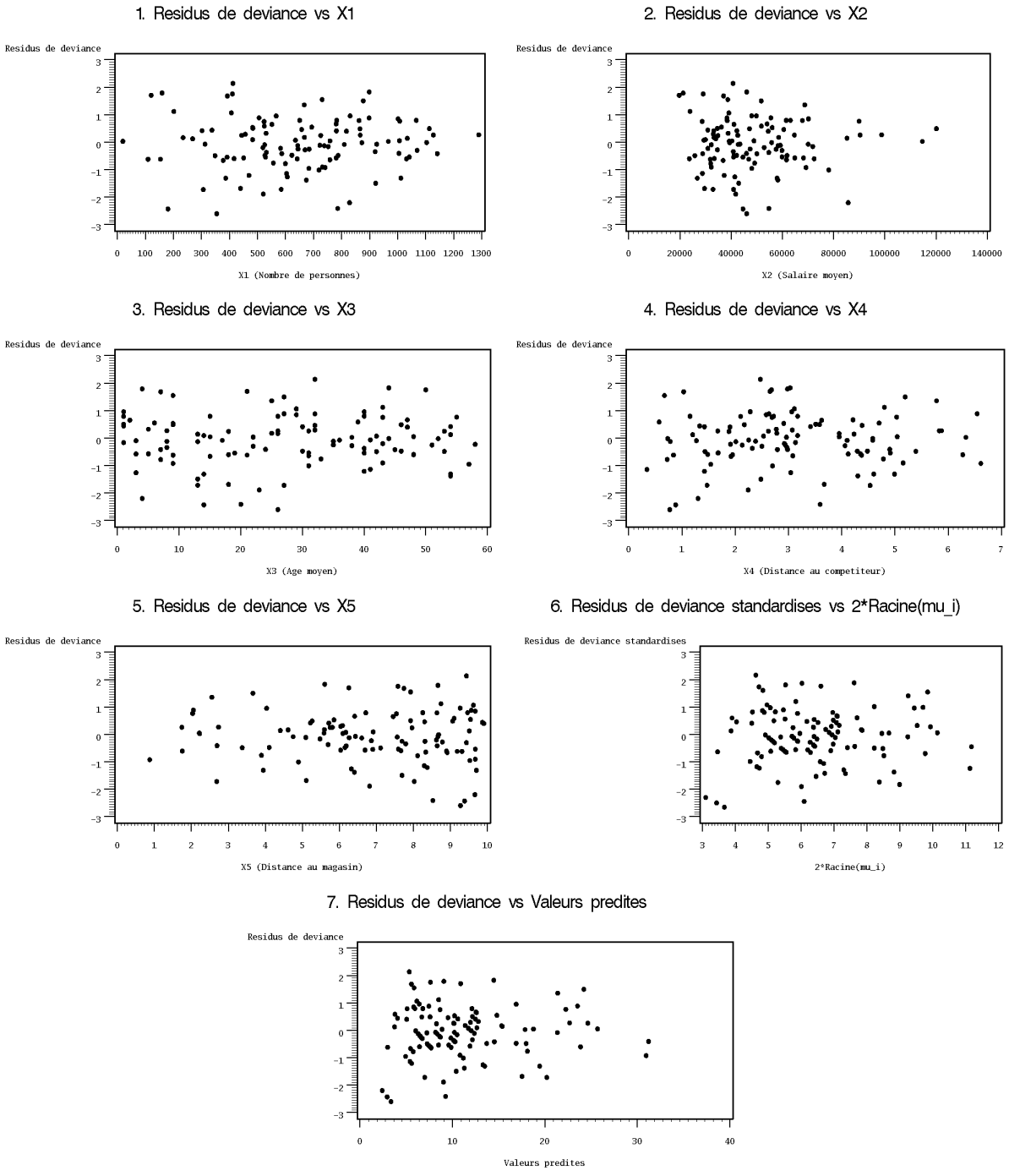

FIG. 2.2 – *Validation graphique du modèle obtenu.*

Au chapitre suivant, les équations d'estimation généralisées (GEE)[7] seront introduites. Celles-ci servent à estimer les paramètres d'un modèle de régression quand les données sont longitudinales ou par grappes.

[7] *Generalized Estimating Equations*

Chapitre 3

Les équations d'estimation généralisées (GEE)

En sciences appliquées, il est fréquent de prendre des mesures sur un individu à plusieurs moments dans le temps, ce qui fait que la corrélation des mesures pour un individu en particuler doit être prise en considération dans les analyses statistiques. Les méthodes décrites au chapitre précédent ne sont alors plus valides, puisqu'elles ne tiennent pas compte de la corrélation entre les observations prises sur un même individu. Dans ce chapitre, la façon d'analyser des données mesurées à travers le temps sera expliquée et nous utiliserons pour ce faire une première approche basée sur les équations d'estimation généralisées (GEE).

L'approche par GEE ne spécifie pas entièrement la distribution conjointe des $\boldsymbol{Y_i}$, mais plutôt une modélisation de la moyenne et une spécification de la structure de corrélation. Dans le contexte longitudinal, différentes formes de travail de la structure de corrélation sont utilisées et les estimateurs sont solutions des GEE. Un élément attrayant de cette approche est que les estimations des paramètres du modèle sont convergentes même dans l'éventualité où la structure de corrélation serait mal spécifiée.

Les données longitudinales sont analysées en utilisant la même fonction de lien et le même prédicteur linéaire que dans un cas où les données seraient indépendantes. Pour une analyse longitudinale, on permet aux variables réponses d'être corrélées entre elles pour un individu donné ; cependant, elles doivent être indépendantes d'un individu à l'autre.

3.1 Des définitions

Soit une fonction de lien $g(\mu_{it}) = \boldsymbol{x'_{it}\beta}$ $(t = 1, \ldots, n_i)$ où $\mu_{it} = b'(\theta_{it}) = \mathbb{E}(Y_{it})$ et où $\eta_{it} = \boldsymbol{x'_{it}\beta}$, et supposons le vecteur de dimension $n_i \times 1$ des variables réponses pour l'individu i $(i = 1, \ldots, n)$, noté $\boldsymbol{Y_i} = [Y_{i1}, \ldots, Y_{in_i}]'$. De plus, chacun des vecteurs $\boldsymbol{Y_i}$ a comme vecteur de moyenne $\boldsymbol{\mu_i} = [\mu_{i1}, \ldots, \mu_{in_i}]'$. Finalement, définissons le vecteur $\boldsymbol{x'_{it}} = [x_{it1}, \ldots, x_{itp}]$ comme étant le vecteur de dimension $p \times 1$ des variables indépendantes ou explicatives pour le i^e individu au temps t.

Une matrice de dimension $n_i \times p$ regroupant l'ensemble des variables explicatives de l'individu i peut être obtenue à partir des différents vecteurs $\boldsymbol{x_{it}}$. On la note alors $\boldsymbol{X_i}$:

$$\boldsymbol{X_i} = [\boldsymbol{x_{i1}}, \ldots, \boldsymbol{x_{in_i}}]' = \begin{pmatrix} x_{i11} & x_{i12} & x_{i13} & \cdots & x_{i1p} \\ x_{i21} & x_{i22} & x_{i23} & \cdots & x_{i2p} \\ x_{i31} & x_{i32} & x_{i33} & \cdots & x_{i3p} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{in_i1} & x_{in_i2} & x_{in_i3} & \cdots & x_{in_ip} \end{pmatrix}.$$

Définissons de plus les matrices et vecteurs suivants :

$$\begin{aligned} \boldsymbol{\Delta_i} &= \operatorname{diag}\left(\frac{d\theta_{it}}{d\eta_{it}}\right) = \operatorname{diag}\left(\frac{d\theta_{it}}{d\boldsymbol{x'_{it}\beta}}\right) \text{ de dimension } n_i \times n_i \\ \boldsymbol{S_i} &= \boldsymbol{Y_i - \mu_i} \text{ de dimension } n_i \times 1 \\ \boldsymbol{A_i} &= \operatorname{diag}(b''(\theta_{it})) = \frac{1}{a(\phi)}\operatorname{diag}(\mathbb{V}\text{ar}(Y_{it})) \text{ de dimension } n_i \times n_i. \end{aligned}$$

Supposons enfin que pour les GEE, la densité marginale de Y_{it} fait partie de la famille exponentielle, c'est-à-dire que la densité de Y_{it} se réexprime comme suit :

$$f(y_{it}|\theta_{it};\phi) = \exp\left\{\frac{Y_{it}\theta_{it} - b(\theta_{it})}{a(\phi)} - c(Y_{it},\phi)\right\}. \tag{3.1}$$

3.2 Les équations d'estimation sous l'indépendance (IEE)

Dans le cas où l'on suppose que les données sont indépendantes, l'estimation des paramètres β_k se fait en posant égale à 0 la fonction score. Cette fonction score est définie comme étant le vecteur dont l'élément en position k $(k = 0, \ldots, p')$ est donné

par

$$U_k(\beta_k) = \frac{\partial}{\partial \beta_k} \ell(\boldsymbol{\beta}|\boldsymbol{Y_i}).$$

De plus, dans le cas où nous sommes en présence d'une loi faisant partie de la famille exponentielle,

$$\begin{aligned}
L(\boldsymbol{\beta}|\boldsymbol{Y_i}) &= \exp\left\{\sum_{i=1}^{n}\sum_{t=1}^{n_i}\left\{\frac{\theta_{it}Y_{it} - b(\theta_{it})}{a(\phi)} - c(Y_{it},\phi)\right\}\right\} \\
\ell(\boldsymbol{\beta}|\boldsymbol{Y_i}) = \ln(L(\boldsymbol{\beta}|\boldsymbol{Y_i})) &= \sum_{i=1}^{n}\sum_{t=1}^{n_i}\left\{\frac{\theta_{it}Y_{it} - b(\theta_{it})}{a(\phi)} - c(Y_{it},\phi)\right\} \\
U_k(\beta_k) = \frac{\partial \ell(\boldsymbol{\beta}|\boldsymbol{Y_i})}{\partial \beta_k} &= \sum_{i=1}^{n}\sum_{t=1}^{n_i}\frac{1}{a(\phi)}\left\{Y_{it}\frac{d\theta_{it}}{d\beta_k} - b'(\theta_{it})\frac{d\theta_{it}}{d\beta_k}\right\} \\
&= \sum_{i=1}^{n}\sum_{t=1}^{n_i}\frac{1}{a(\phi)}\left\{Y_{it}\underbrace{\frac{d\theta_{it}}{d\eta_{it}}}_{\equiv\Delta_{it}}\frac{d\eta_{it}}{d\beta_k} - b'(\theta_{it})\underbrace{\frac{d\theta_{it}}{d\eta_{it}}}_{\equiv\Delta_{it}}\frac{d\eta_{it}}{d\beta_k}\right\} \\
&= \sum_{i=1}^{n}\sum_{t=1}^{n_i}\frac{1}{a(\phi)}\left\{Y_{it}\Delta_{it}x_{itk} - b'(\theta_{it})\Delta_{it}x_{itk}\right\} \\
&= \sum_{i=1}^{n}\sum_{t=1}^{n_i}\frac{1}{a(\phi)}(Y_{it}\Delta_{it}x_{itk} - \mu_{it}\Delta_{it}x_{itk}) \\
&= \sum_{i=1}^{n}\sum_{t=1}^{n_i}\frac{1}{a(\phi)}\Delta_{it}x_{itk}(Y_{it} - \mu_{it}).
\end{aligned}$$

En utilisant la notation matricielle, on peut donc écrire le système d'équations à résoudre comme suit :

$$\boldsymbol{U_{IEE}(\beta)} = \sum_{i=1}^{n}\frac{1}{a(\phi)}\boldsymbol{X_i'\Delta_i(Y_i - \mu_i)}$$

Ainsi, afin de trouver $\hat{\boldsymbol{\beta}}$, résolvons

$$\begin{aligned}
\boldsymbol{U_{IEE}(\beta)} &= \boldsymbol{0} \\
\Rightarrow \sum_{i=1}^{n}\boldsymbol{X_i'\Delta_i(Y_i - \mu_i)} &= \boldsymbol{0} \\
\Rightarrow \sum_{i=1}^{n}\boldsymbol{D_i'V_i^{-1}S_i} &= \boldsymbol{0} \qquad (3.2)
\end{aligned}$$

où $\boldsymbol{D_i} = \boldsymbol{A_i\Delta_i X_i} = \dfrac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}}$ et $\boldsymbol{V_i} = a(\phi)\boldsymbol{A_i^{\frac{1}{2}}R_i(\alpha)A_i^{\frac{1}{2}}}$ avec $\boldsymbol{R_i(\alpha)} = \boldsymbol{I}$.

Donc, dans le cas où nous sommes en présence de données indépendantes, le vecteur $\boldsymbol{\beta}$ est obtenu en résolvant le système d'équations $\boldsymbol{U_{IEE}(\beta)} = \mathbf{0}$, et la matrice de variance-covariance des $\hat{\boldsymbol{\beta}}$ peut être estimée de façon convergente (Liang & Zeger, 1986) par

$$\hat{\boldsymbol{V}} = \left(\sum_{i=1}^{n} \boldsymbol{X_i'\Delta_i A_i \Delta_i X_i}\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{X_i'\Delta_i S_i S_i' \Delta_i X_i}\right) \left(\sum_{i=1}^{n} \boldsymbol{X_i'\Delta_i A_i \Delta_i X_i}\right)^{-1} \Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}.$$

Si les Y_{it} ne sont pas indépendantes et que les équations d'estimation sous l'indépendance sont utilisées, $\hat{\boldsymbol{V}}$ donne un estimé valide de la variance de $\hat{\boldsymbol{\beta}}$. Cependant, si les Y_{it} sont vraiment des données indépendantes, la matrice de variance-covariance peut être estimée de façon plus efficace par la matrice d'information

$$\hat{\boldsymbol{V}} = \left(\sum_{i=1}^{n} \boldsymbol{X_i'\Delta_i A_i \Delta_i X_i}\right)^{-1} \Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}.$$

3.3 Les équations d'estimation généralisées (GEE)

3.3.1 La fonction de quasi-vraisemblance

Zeger & Liang (1986) définissent l'estimation par la quasi-vraisemblance comme étant une méthodologie en régression qui nécessite certaines hypothèses faibles concernant la distribution de la variable dépendante et pouvant être utilisée dans plusieurs situations. Pour la quasi-vraisemblance, il ne faut que spécifier la relation entre la moyenne de la variable réponse et les variables exogènes, et le lien entre la moyenne et la variance de la variable réponse. Dans le cas de données non corrélées, $Y_1|\boldsymbol{x_1}, \ldots, Y_n|\boldsymbol{x_n}$, les estimateurs de quasi-vraisemblance sont la solution de l'équation

$$U_k(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_k} v_i^{-1}(Y_i - \mu_i) = 0, \ k = 0, \ldots, p' \tag{3.3}$$

où v_i représente la variance de $Y_i|\boldsymbol{x_i}$.

3.3.2 L'extension de la fonction de quasi-vraisemblance aux GEE

Les équations d'estimation généralisées ont été développées par Liang & Zeger (1986) afin de traiter les données corrélées entre elles quand celles-ci peuvent être vues marginalement comme un modèle linéaire généralisé (section 2.3 de la page 9).

Posons $\boldsymbol{R_i(\alpha)}$ une matrice de dimension $n_i \times n_i$, où le $(t,t')^e$ élément de cette matrice $(t \neq t')$ est la corrélation entre Y_{it} et $Y_{it'}$, et où le $(t,t)^e$ élément de la matrice est 1. Des choix pour cette matrice sont donnés à la section 3.5 de la page 32. On peut de plus définir une matrice $\boldsymbol{R(\alpha)}$ de dimension $N \times N$ $(N = \sum_{i=1}^{n} n_i)$ de la façon suivante :

$$\boldsymbol{R(\alpha)} = \begin{pmatrix} \boldsymbol{R_1(\alpha)} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R_2(\alpha)} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{R_n(\alpha)} \end{pmatrix}.$$

À l'aide de ces matrices, une matrice de covariance pour $\boldsymbol{Y_i}$ peut être calculée comme suit :

$$\boldsymbol{V_i} = a(\phi)\boldsymbol{A_i^{\frac{1}{2}}}\boldsymbol{R_i(\alpha)}\boldsymbol{A_i^{\frac{1}{2}}}. \tag{3.4}$$

Grâce à l'équation (3.3), les équations d'estimation généralisées peuvent être déterminées par une simple généralisation de cette expression. En fait, nous avons que les équations d'estimation généralisées sont :

$$\boldsymbol{U_{GEE}(\beta)} = \sum_{i=1}^{n} \boldsymbol{D_i'V_i^{-1}(\hat{\alpha})S_i} = \boldsymbol{0}, \tag{3.5}$$

où

$$\boldsymbol{D_i'} = \frac{\partial \boldsymbol{\mu_i'}}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{x_{i11}}{g'(\mu_{i1})} & \frac{x_{i21}}{g'(\mu_{i2})} & \cdots & \frac{x_{in_i1}}{g'(\mu_{in_i})} \\ \frac{x_{i12}}{g'(\mu_{i1})} & \frac{x_{i22}}{g'(\mu_{i2})} & \cdots & \frac{x_{in_i2}}{g'(\mu_{in_i})} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{i1p}}{g'(\mu_{i1})} & \frac{x_{i2p}}{g'(\mu_{i2})} & \cdots & \frac{x_{in_ip}}{g'(\mu_{in_i})} \end{pmatrix}$$

et où $\hat{\boldsymbol{\alpha}}$ est un estimateur convergent de $\boldsymbol{\alpha}$. L'équation (3.5) doit être résolue de façon itérative afin de trouver $\hat{\boldsymbol{\beta}}$ et l'algorithme qui sera utilisé à cette fin est présenté à la section 3.4 qui suit.

Malheureusement, en pratique, la matrice $\boldsymbol{R_i}(\boldsymbol{\alpha})$ est inconnue. Dans un cas où la matrice $\boldsymbol{R_i}(\boldsymbol{\alpha})$ est la vraie matrice de corrélation (ce qui est généralement peu probable), la matrice de variance asymptotique de $\hat{\boldsymbol{\beta}}$ pourrait être estimée par

$$\boldsymbol{V_v} = \left(\sum_{i=1}^{n} \boldsymbol{D_i'} \boldsymbol{V_i^{-1}} \boldsymbol{D_i} \right)^{-1} \Bigg|_{\substack{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}} \\ \boldsymbol{\beta}=\hat{\boldsymbol{\beta}} \\ \boldsymbol{\phi}=\hat{\boldsymbol{\phi}}}} . \tag{3.6}$$

Cependant, étant donné que la matrice $\boldsymbol{R_i}(\boldsymbol{\alpha})$ est au mieux une approximation de $\mathbb{V}\mathrm{ar}(\boldsymbol{Y_i}|\boldsymbol{x_i})$, on peut corriger l'estimateur de la variance de $\hat{\boldsymbol{\beta}}$ obtenu avec l'équation (3.6) comme suit :

$$\begin{aligned} \boldsymbol{V_c} &= \boldsymbol{V_v} \left(\sum_{i=1}^{n} \boldsymbol{D_i'} \boldsymbol{V_i^{-1}} \boldsymbol{S_i} \boldsymbol{S_i'} \boldsymbol{V_i^{-1}} \boldsymbol{D_i} \right) \boldsymbol{V_v} \Bigg|_{\substack{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}} \\ \boldsymbol{\beta}=\hat{\boldsymbol{\beta}} \\ \boldsymbol{\phi}=\hat{\boldsymbol{\phi}}}} \qquad (3.7) \\ &= \left(\sum_{i=1}^{n} \boldsymbol{D_i'} \boldsymbol{V_i^{-1}} \boldsymbol{D_i} \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{D_i'} \boldsymbol{V_i^{-1}} \boldsymbol{S_i} \boldsymbol{S_i'} \boldsymbol{V_i^{-1}} \boldsymbol{D_i} \right) \left(\sum_{i=1}^{n} \boldsymbol{D_i'} \boldsymbol{V_i^{-1}} \boldsymbol{D_i} \right)^{-1} \Bigg|_{\substack{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}} \\ \boldsymbol{\beta}=\hat{\boldsymbol{\beta}} \\ \boldsymbol{\phi}=\hat{\boldsymbol{\phi}}}} . \end{aligned}$$

L'estimateur $\boldsymbol{V_c}$ donné par (3.7) est souvent appelé « estimateur sandwich » de la variance de $\hat{\boldsymbol{\beta}}$.

3.4 L'estimation des paramètres $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$

L'algorithme de Newton-Raphson (Cameron & Trivedi, 1998, chapitre 3, pp.93-94) suivant est proposé afin de trouver la valeur des paramètres $\hat{\beta}_k$ dans le cas où les données sont corrélées ou mesurées à travers le temps :

1. Calculer, selon l'équation (3.2) de la page 28, un estimé initial de $\boldsymbol{\beta}$ à partir d'un modèle linéaire généralisé supposant l'indépendance des observations : dénoter le vecteur obtenu par $\hat{\boldsymbol{\beta}}^{(\boldsymbol{0})}$;
2. Estimer $\hat{\boldsymbol{\alpha}}$ et $\hat{\phi}$ à partir du $\hat{\boldsymbol{\beta}}$ de l'étape précédente et à partir des résidus de Pearson. Le $\hat{\boldsymbol{\alpha}}$ est obtenu selon l'une des expressions (3.9), (3.12), (3.13), (3.14) tandis que le $\hat{\phi}$ est obtenu selon l'une des expressions (3.10) ou (3.11). Obtenir finalement une matrice de corrélation $\boldsymbol{R_i}(\hat{\boldsymbol{\alpha}})$ basée sur la structure de la matrice $\boldsymbol{R_i}(\boldsymbol{\alpha})$ supposée au préalable ;
3. Calculer la matrice de covariance $\boldsymbol{V_i} = a(\hat{\phi}) \boldsymbol{A_i^{\frac{1}{2}}} \boldsymbol{R_i}(\hat{\boldsymbol{\alpha}}) \boldsymbol{A_i^{\frac{1}{2}}}$;

4. Obtenir un nouveau vecteur $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}}^{(\boldsymbol{r+1})} = \hat{\boldsymbol{\beta}}^{(\boldsymbol{r})} + \left(\sum_{i=1}^{n} \boldsymbol{D_i'} \boldsymbol{V_i^{-1}} \boldsymbol{D_i} \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{D_i'} \boldsymbol{V_i^{-1}} \boldsymbol{S_i} \right) \Bigg|_{\substack{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}^{(\boldsymbol{r})} \\ \boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(\boldsymbol{r})} \\ \boldsymbol{\phi}=\hat{\boldsymbol{\phi}}^{(\boldsymbol{r})}}} . \tag{3.8}$$

5. Recommencer les étapes 2 à 4 jusqu'à convergence.

Zheng (2000) mentionne que lorsque le modèle pour μ est bien spécifié, les $\hat{\boldsymbol{\beta}}$ obtenus par la méthode des équations d'estimation généralisées s'approchent des paramètres $\boldsymbol{\beta}$ quand $n \to \infty$, peu importe le choix de $\boldsymbol{R_i}(\boldsymbol{\alpha})$. Cependant, un bon choix de $\boldsymbol{R_i}(\boldsymbol{\alpha})$ assure des estimations plus efficaces de $\boldsymbol{\beta}$ et de sa variance.

3.5 Les types de matrices de corrélation $\boldsymbol{R_i}(\boldsymbol{\alpha})$ les plus communs

3.5.1 La structure autorégressive

Cette première structure de corrélation[1] est utile lorsque l'on suppose une dépendance temporelle des répétitions. Par exemple, si l'on mesure la taille d'un individu en 10 années consécutives, il est possible alors de dire que les mesures sont dépendantes les unes des autres dans le temps et que les répétitions ont un ordre chronologique. De plus, les corrélations entre les mesures diminuent avec le temps, et ce, de façon géométrique. Ce type de matrice nécessite l'estimation d'un seul paramètre. On a

$$\text{corr}(Y_{it}, Y_{it'}) = \alpha^{|t-t'|} \text{ pour } |t-t'| = 0, \ldots, n_i - t,$$

donc,

$$\boldsymbol{R_i}(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & & & & \\ \alpha & 1 & & & \\ \alpha^2 & \alpha & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \alpha^{n_i-1} & \alpha^{n_i-2} & \alpha^{n_i-3} & & 1 \end{pmatrix}.$$

[1] Toutes les formes de matrice et les formules de la section 3.5 sont tirées de la documentation en ligne de SAS.

Dans le but de trouver la valeur de $\hat{\alpha}$, les résidus de Pearson doivent être calculés, et ceux-ci sont définis comme à la sous-section 2.5.2 de la page 14. On calcule ensuite

$$\hat{\alpha} = \frac{1}{(K_1 - p)a(\hat{\phi})} \sum_{i=1}^{n} \sum_{t \leq n_i - 1} r_{P_{it}} r_{P_{i,t+1}} \tag{3.9}$$

où

$$K_1 = \sum_{i=1}^{n} (n_i - 1),$$

et

$$a(\hat{\phi}) = \frac{1}{(\sum_{i=1}^{n} n_i) - p} \sum_{i=1}^{K} \sum_{t=1}^{n_i} r_{P_{it}}^2. \tag{3.10}$$

Une deuxième façon d'exprimer $\hat{\phi}$ est

$$a(\hat{\phi}) = \frac{1}{\sum_{i=1}^{n} n_i} \sum_{i=1}^{n} \sum_{t=1}^{n_i} r_{P_{it}}^2. \tag{3.11}$$

3.5.2 La structure d'équicorrélation

Cette structure est utilisée lorsque les mesures répétées ne dépendent pas du temps et lorsque l'on suppose que les observations ont une corrélation commune. Un exemple de ce type de données peut être les différentes mesures prises chez les individus de la $i^{\text{ème}}$ famille. Un seul paramètre, α, est estimé. On a

$$\text{corr}(Y_{it}, Y_{it'}) = \begin{cases} 1, & \text{si } t = t', \\ \alpha, & \text{si } t \neq t', \end{cases}$$

et donc

$$\boldsymbol{R_i}(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & & & & \\ \alpha & 1 & & & \\ \alpha & \alpha & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \alpha & \alpha & \alpha & \cdots & 1 \end{pmatrix}.$$

Afin d'obtenir l'estimateur de α, les résidus de Pearson vus auparavant doivent être calculés. Ensuite,

$$\hat{\alpha} = \frac{1}{(N^* - p)a(\hat{\phi})} \sum_{i=1}^{n} \sum_{t \neq t'} r_{P_{it}} r_{P_{it'}} \tag{3.12}$$

où

$$N^* = \sum_{i=1}^{n} n_i(n_i - 1).$$

3.5.3 La structure d'indépendance

On utilise ce type de matrice lorsque l'on suppose que le temps n'a pas d'effet sur les mesures prises. En supposant que $\boldsymbol{R_i(\alpha)} = \boldsymbol{I_{n_i}}$, on se ramène aux équations décrites à la section 3.2 en page 27. On a

$$\text{corr}(Y_{it}, Y_{it'}) = \begin{cases} 1, & \text{si } t = t', \\ 0, & \text{si } t \neq t'. \end{cases}$$

3.5.4 La structure m-dépendante

Pour ce type de matrice[2], m paramètres doivent être estimés. On a

$$\text{corr}(Y_{it}, Y_{it'}) = \begin{cases} 1, & \text{si } |t - t'| = 0, \\ \alpha_{|t-t'|}, & \text{si } |t - t'| = 1, 2, \ldots, m, \\ 0, & \text{si } |t - t'| > m, \end{cases}$$

où

$$\hat{\alpha}_{|t-t'|} = \frac{1}{(K_{|t-t'|} - p)a(\hat{\phi})} \sum_{i=1}^{n} \sum_{t \leq n_i - |t-t'|} r_{P_{it}} r_{P_{it'}} \tag{3.13}$$

et

$$K_{|t-t'|} = \sum_{i=1}^{n} (n_i - |t - t'|).$$

Pour une matrice de corrélation de dimension $n_i \times n_i$, il existe $n_i - 1$ choix possibles pour m. Par exemple, si $n_i = 4$, on a

• 1-DÉPENDANTE

$$\boldsymbol{R_i(\alpha)} = \begin{pmatrix} 1 & & & \\ \alpha_1 & 1 & & \\ 0 & \alpha_1 & 1 & \\ 0 & 0 & \alpha_1 & 1 \end{pmatrix}$$

[2]On appelle aussi cette structure *par bandes*.

• **2-DÉPENDANTE**

$$\boldsymbol{R_i}(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & & & \\ \alpha_1 & 1 & & \\ \alpha_2 & \alpha_1 & 1 & \\ 0 & \alpha_2 & \alpha_1 & 1 \end{pmatrix}$$

• **3-DÉPENDANTE**

$$\boldsymbol{R_i}(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & & & \\ \alpha_1 & 1 & & \\ \alpha_2 & \alpha_1 & 1 & \\ \alpha_3 & \alpha_2 & \alpha_1 & 1 \end{pmatrix}$$

3.5.5 Non-structuré

Cette structure de corrélation n'impose aucune structure particulière à la matrice $\boldsymbol{R_i}(\boldsymbol{\alpha})$. Un total de $n_i(n_i-1)/2$ paramètres sont estimés pour ce type de matrice. On pose

$$\text{corr}(Y_{it}, Y_{it'}) = \begin{cases} 1, & \text{si } t = t', \\ \alpha_{tt'}, & \text{si } t \neq t', \end{cases}$$

d'où

$$\boldsymbol{R_i}(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & & & & \\ \alpha_{12} & 1 & & & \\ \alpha_{13} & \alpha_{12} & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \alpha_{1n_i} & \alpha_{2n_i} & \alpha_{3n_i} & \cdots & 1 \end{pmatrix}$$

et on estime $\alpha_{tt'}$ par

$$\hat{\alpha}_{tt'} = \frac{1}{(K-p)a(\hat{\phi})} \sum_{i=1}^{n} r_{P_{it}} r_{P_{it'}}. \tag{3.14}$$

3.6 Conclusion du chapitre

Les équations d'estimation généralisées sont utilisées afin de trouver la valeur des paramètres $\hat{\beta}_k$ d'un modèle de régression dans le cas où plusieurs mesures ont été

prises sur un même individu en différents temps. Ces équations seront appliquées à la régression de Poisson longitudinale au chapitre suivant.

Chapitre 4

La régression de Poisson longitudinale

Au chapitre 2, la régression de Poisson a été introduite dans le cas où les données sont indépendantes les unes des autres. Cependant, il a été vu au chapitre 3 que ce n'est pas toujours le cas : il arrive fréquemment que des mesures soient prises sur un même individu à plusieurs moments distincts dans le temps, ce qui induit de la corrélation. En ajoutant la condition que la variable réponse en soit une de dénombrement, nous obtenons un cas où la régression de Poisson longitudinale doit être considérée. Ce chapitre appliquera l'approche basée sur les GEE dans le cas où la variable réponse en soit une de dénombrement.

4.1 Les données longitudinales

Il a été mentionné auparavant que des données longitudinales surviennent lorsque des mesures sont prises sur des individus à travers le temps. Les deux raisons suivantes peuvent motiver la réalisation d'analyses longitudinales :

1. Accroître la sensibilité lors des comparaisons entre les sujets ;
2. Étudier les changements à travers le temps ;

4.2 Les équations d'estimation généralisées dans le cas d'une loi de Poisson

Dans le cas où la variable réponse est de dénombrement (plus précisément lorsqu'elle suit une loi de Poisson), les équations d'estimation généralisées vues au chapitre 3 peuvent être obtenues de deux façons distinctes : à l'aide de la fonction de quasi-vraisemblance et à l'aide de la fonction de log-vraisemblance.

4.2.1 L'obtention des GEE à l'aide de la fonction de quasi-vraisemblance

Au chapitre 3, nous avons vu que les équations d'estimation généralisées (équation (3.5) en page 30) sont données par

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{D}_{\boldsymbol{i}}' \boldsymbol{V}_{\boldsymbol{i}}^{-1} \boldsymbol{S}_{\boldsymbol{i}} = \boldsymbol{0} \tag{4.1}$$

avec $\boldsymbol{D_i} = \boldsymbol{A_i \Delta_i X_i}$, $\boldsymbol{A_i} = \frac{1}{a(\phi)}\text{diag}(\mathbb{V}\text{ar}(Y_{it}))$, $\boldsymbol{S_i} = \boldsymbol{Y_i - \mu_i}$ et $\boldsymbol{V_i} = a(\phi)\boldsymbol{A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}}}$. Comme les données proviennent d'une loi de Poisson, il s'ensuit que

$$\Delta_{it} = \frac{d\theta_{it}}{d\eta_{it}} = \frac{d(\ln(\mu_{it}))}{d\eta_{it}} = \frac{d(\ln(e^{\boldsymbol{x_{it}'\beta}}))}{d\eta_{it}} = \frac{d\boldsymbol{x_{it}'\beta}}{d\eta_{it}} = \frac{d\eta_{it}}{d\eta_{it}} = 1,$$

et $\boldsymbol{A_i} = \text{diag}(\mu_{it})$. Alors, $\boldsymbol{\Delta_i} = \boldsymbol{I_{n_i}}$ et $\boldsymbol{D_i} = \boldsymbol{A_i X_i}$. L'équation (4.1) se réduit donc à

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{X_i' A_i} \{\boldsymbol{A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}}}\}^{-1} \boldsymbol{S_i} = \boldsymbol{0}. \tag{4.2}$$

4.2.2 L'obtention des GEE à l'aide de la fonction de log-vraisemblance

Il a été mentionné à la sous-section 2.2.3 de la page 9 que la fonction de vraisemblance est obtenue par $\prod_{i=1}^{n} f(Y_i|\boldsymbol{x_i})$. Dans le cas où nous avons des données indépendantes,

la fonction de vraisemblance est plutôt obtenue par $\prod_{i=1}^{n}\prod_{t=1}^{n_i} f(Y_{it}|\boldsymbol{x_{it}})$, avec $f(Y_{it}|\boldsymbol{x_{it}})$ comme à l'équation (4.5). Ainsi, nous avons

$$L(\boldsymbol{\mu_i}|\boldsymbol{Y_i}) = \prod_{i=1}^{n}\prod_{t=1}^{n_i} \frac{e^{-\mu_{it}}\mu_{it}^{Y_{it}}}{Y_{it}!} = \prod_{i=1}^{n}\prod_{t=1}^{n_i} \frac{e^{-e^{\boldsymbol{x'_{it}\beta}}}e^{\boldsymbol{x'_{it}\beta}Y_{it}}}{Y_{it}!}$$

Donc, la fonction de log-vraisemblance est

$$\ell(\boldsymbol{\mu_i}|\boldsymbol{Y_i}) \;=\; \ln(L(\boldsymbol{\mu_i}|\boldsymbol{Y_i})) = \sum_{i=1}^{n}\sum_{t=1}^{n_i}[-e^{\boldsymbol{x'_{it}\beta}} + Y_{it}\boldsymbol{x'_{it}\beta} - \ln(Y_{it}!)]$$

et la fonction score est

$$\begin{aligned}
\boldsymbol{U(\beta)} &= \sum_{i=1}^{n}\sum_{t=1}^{n_i}[-\boldsymbol{x'_{it}}e^{\boldsymbol{x'_{it}\beta}} + Y_{it}\boldsymbol{x'_{it}}] \\
&= \sum_{i=1}^{n}\sum_{t=1}^{n_i}[\boldsymbol{x'_{it}}\{-e^{\boldsymbol{x'_{it}\beta}} + Y_{it}\}] \\
&= \sum_{i=1}^{n}\boldsymbol{X'_i}[-\boldsymbol{\mu_i} + \boldsymbol{Y_i}] \\
&= \sum_{i=1}^{n}\boldsymbol{X'_i}\boldsymbol{S_i}. \qquad (4.3)
\end{aligned}$$

Dans le cas où les données ne seraient pas indépendantes, la fonction score est généralisée ainsi :

$$\sum_{i=1}^{n}\boldsymbol{X'_i}\boldsymbol{A_i}\{\boldsymbol{A_i^{\frac{1}{2}}}\boldsymbol{R_i(\alpha)}\boldsymbol{A_i^{\frac{1}{2}}}\}^{-1}\boldsymbol{S_i}. \qquad (4.4)$$

Donc, en posant $\boldsymbol{R_i(\alpha)} = \boldsymbol{I_{n_i}}$ dans l'équation (4.4), nous nous rapportons à l'équation (4.3), car

$$\begin{aligned}
\boldsymbol{U(\beta)} &= \sum_{i=1}^{n}\boldsymbol{X'_i}\boldsymbol{A_i}\{\boldsymbol{A_i^{\frac{1}{2}}}\boldsymbol{R_i(\alpha)}\boldsymbol{A_i^{\frac{1}{2}}}\}^{-1}\boldsymbol{S_i} \\
&= \sum_{i=1}^{n}\boldsymbol{X'_i}\boldsymbol{A_i}\{\boldsymbol{A_i^{\frac{1}{2}}}\boldsymbol{I_{n_i}}\boldsymbol{A_i^{\frac{1}{2}}}\}^{-1}\boldsymbol{S_i} \\
&= \sum_{i=1}^{n}\boldsymbol{X'_i}\boldsymbol{A_i}\{\boldsymbol{A_i^{\frac{1}{2}}}\boldsymbol{A_i^{\frac{1}{2}}}\}^{-1}\boldsymbol{S_i} = \sum_{i=1}^{n}\boldsymbol{X'_i}\boldsymbol{A_i}\boldsymbol{A_i}^{-1}\boldsymbol{S_i} \\
&= \sum_{i=1}^{n}\boldsymbol{X'_i}\boldsymbol{S_i}.
\end{aligned}$$

□

4.3 Le modèle de la régression de Poisson longitudinale

De la même façon qu'à la section 2.4 du chapitre 2, nous avons toujours

$$f(Y_{it}|\boldsymbol{x_{it}}) = \frac{e^{-\mu_{it}}\mu_{it}^{Y_{it}}}{Y_{it}!}, \quad i = 1, \ldots, n, \ t = 1, \ldots, n_i, \tag{4.5}$$

avec $\mu_{it} = e^{\boldsymbol{x'_{it}\beta}}$. Nous tentons encore d'estimer l'espérance de la variable réponse avec une fonction de lien logarithmique et le même prédicteur linéaire η_{it}. Ainsi, nous voulons estimer $\mu_{it} = e^{\boldsymbol{x'_{it}\beta}} = \exp\{\beta_0 + \beta_1 x_{it1} + \cdots + \beta_{p'} x_{itp'}\}$. Quant au vecteur $\boldsymbol{x_{it}}$, il est défini comme à la section 3.1.

Pour une régression de Poisson longitudinale, la notion de variable *offset* est toujours valide et est utilisée lorsque $\boldsymbol{Y_i}$ est proportionnelle à une variable exogène mesurée. Le coefficient de cette variable exogène est toujours fixé à 1 au lieu d'être estimé comme les p' autres coefficients du modèle. De plus, Zorn (2001) mentionne que l'interprétation des paramètres est toujours la même que sous l'hypothèse d'indépendance des observations.

4.4 Un exemple complet sur la régression de Poisson longitudinale

L'exemple qui suit est tiré de Thall & Vail (1990). Un total de 59 patients âgés entre 18 et 42 ans ont été étudiés ($n = 59$). À chacune des 4 visites ($n_i = 4$), le nombre de crises d'épilepsie survenues depuis la dernière visite a été enregistré. Une période de 2 semaines séparait chacune des 4 visites (la durée totale du traitement était donc de 8 semaines). Les données sont illustrées au Tableau 4.1 de la page suivante.

À la lecture de ce tableau, on peut soupçonner que les données saisies pour le patient 207 sont erronées. En fait, les données pour cet individu sont très différentes de celles pour les 58 autres individus : l'individu 207 sera alors éliminé des analyses. De plus, les variables « Visite 1 » à « Visite 4 » représentent le nombre de crises survenues dans la période de 2 semaines précédant la visite. La variable « Traitement » prend la valeur 0 si le patient reçoit le placebo, et prend la valeur 1 sinon. Quant à la variable « Base », elle représente le nombre de crises survenues dans la période de 8 semaines

ID	Visite 1	Visite 2	Visite 3	Visite 4	Traitement	Base	Âge
104	5	3	3	3	0	11	31
106	3	5	3	3	0	11	30
107	2	4	0	5	0	6	25
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
207	102	65	72	63	1	151	22
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
228	2	3	0	1	1	25	21
232	0	0	0	0	1	13	36
236	1	4	3	2	1	12	37

TAB. 4.1 – *Extrait des données pour l'exemple sur le traitement de l'épilepsie.*

précédant le début du traitement ou du suivi. Vingt-huit patients ont reçu un placebo (traitement = 0), alors que les 30 autres patients ont reçu un traitement contre l'épilepsie (Progabide, traitement=1). Finalement, on a recueilli l'âge du patient et le nombre de crises survenues dans une période de 8 semaines précédant le début du traitement.

Puisque la procédure GENMOD de SAS (SAS Institute Inc., 2004) permet que les données soient corrélées entre elles, celle-ci peut être utilisée. Dans ce cas, cette procédure utilise l'approche par GEE afin d'estimer les paramètres du modèle de régression.

4.4.1 Une analyse descriptive des données

Les TABLEAUX 4.2 et 4.3 résument les principales statistiques descriptives.

Variable	Moyenne	Écart-type	Minimum	Maximum
Visite 1	8.95	14.84	0	102
Visite 2	8.36	10.19	0	65
Visite 3	8.45	14.15	0	76
Visite 4	7.31	9.65	0	63
base	31.22	26.88	6	151
âge	28.34	6.30	18	42

TAB. 4.2 – *Statistiques descriptives (avec le patient 207).*

Variable	Moyenne	Écart-type	Minimum	Maximum
Visite 1	7.34	8.33	0	40
Visite 2	7.38	6.95	0	29
Visite 3	7.34	11.47	0	76
Visite 4	6.34	6.28	0	29
base	29.16	21.89	6	111
âge	28.45	6.30	18	42

TAB. 4.3 – *Statistiques descriptives (sans le patient 207).*

4.4.2 Qu'est-ce qui motive le choix d'un type de matrice de corrélation ?

La structure de corrélation qui sera utilisée dans l'exemple de cette section est celle d'équicorrélation. Mais qu'est-ce qui motive ce choix ? En effet, il a été mentionné au chapitre 3 qu'il soit rare que nous connaissions la vraie structure de matrice de corrélation à utiliser. Ainsi, peu importe le choix de la matrice $\boldsymbol{R_i}(\boldsymbol{\alpha})$, les estimateurs de $\boldsymbol{\beta}$ seraient très similaires d'une structure à l'autre, mais leurs variances seraient corrigées par un estimateur robuste de la variance, donné par l'équation (3.7) de la page 31. Alors, peu importe la structure de corrélation, les estimés de $\boldsymbol{\beta}$ seraient semblables, mais les erreurs standards sont corrigées. Le TABLEAU 4.4 en page 43 démontre ce fait.

Le choix de la structure de la matrice $\boldsymbol{R_i}(\boldsymbol{\alpha})$ est fait selon la connaissance qu'a l'utilisateur de la corrélation entre les n_i mesures. Ici, des 5 visites possibles, la corrélation entre 2 de ces visites est toujours la même, faisant en sorte que nous avons choisi la structure d'équicorrélation.

À ce tableau, les seuils testant l'hypothèse que la valeur du paramètre soit nulle sont inscrits. On voit dans la colonne des seuils que pour certains types de structure de matrices (1-dépendante et 2-dépendante), l'interaction entre la variable x_1 et le traitement n'est pas significative, alors qu'elle l'était en ne tenant pas compte de la corrélation des données (voir TABLEAU 4.5).

4.4.3 Les résultats obtenus avec la procédure GENMOD

Certaines manipulations doivent d'abord être faites à la base de données. En effet, les données pour un individu devront être comme à la page suivante, où la variable

x_1 prend la valeur 0 s'il s'agit de la période précédent le traitement, 1 s'il s'agit de la période de traitement. De plus, un terme *offset* est inclus à la base de données. Celui-ci représente le nombre de semaines entre 2 visites, car plus le temps entre les visites est grand, plus le patient est à risque de faire un grand nombre de crises :

ID	Y	Visite	Traitement	Base	Âge	x_1	ln(temps)
104	11	0	0	11	31	0	2.07944
104	5	1	0	11	31	1	0.69315
104	3	2	0	11	31	1	0.69315
104	3	3	0	11	31	1	0.69315
104	3	4	0	11	31	1	0.69315

Structure	Variable	Estimé	Erreur Standard	Seuil
Autorégressive	Intercept	1.3378	0.1587	<0.0001
	x_1	0.1265	0.0974	0.1939
	trt	-0.1373	0.1980	0.4881
	x_1*trt	-0.3187	0.1571	0.0425
Égale corrélation	Intercept	1.3476	0.1574	<0.0001
	x_1	0.1108	0.1161	0.3399
	trt	-0.1080	0.1937	0.5770
	x_1*trt	-0.3016	0.1712	0.0781
Indépendance	Intercept	1.3476	0.1574	<0.0001
	x_1	0.1108	0.1161	0.3399
	trt	-0.1080	0.1937	0.5770
	x_1*trt	-0.3016	0.1712	0.0781
1-dépendante	Intercept	1.3632	0.1882	<0.0001
	x_1	0.0750	0.1375	0.5854
	trt	-0.1401	0.2321	0.5462
	x_1*trt	-0.2748	0.1862	**0.1399**
2-dépendante	Intercept	1.3121	0.1708	<0.0001
	x_1	0.2064	0.2195	0.3469
	trt	-0.1896	0.2128	0.3728
	x_1*trt	-0.4037	0.3042	**0.1844**
3-dépendante	Intercept	1.3505	0.1656	<0.0001
	x_1	0.1085	0.1690	0.5208
	trt	-0.0890	0.2036	0.6620
	x_1*trt	-0.5041	0.2408	0.0363
4-dépendante	Intercept	1.3512	0.1635	<0.0001
	x_1	0.1034	0.0835	0.2157
	trt	-0.1057	0.2016	0.6002
	x_1*trt	-0.3252	0.1404	0.0205

TAB. 4.4 – *Estimés des paramètres, leurs erreurs standards et les seuils associés au test H_0 : paramètre = 0 (seuil de 10%).*

Les données seront encore une fois analysées avec la procédure GENMOD de SAS. En premier lieu, SAS estime les valeurs des paramètres à l'aide d'un modèle linéaire

généralisé (GLM), sans se soucier de la dépendance entre les observations. Les résultats ainsi obtenus se trouvent au TABLEAU 4.5. Les estimés de paramètres sous l'indépendance serviront de valeurs de départ pour les différentes itérations des GEE (voir étape 1 de l'algorithme de la section 3.4 de la page 31). Les estimés finaux, résultant de plusieurs itérations, sont présentés au TABLEAU 4.6.

Variable	ddl	Estimé	Erreur Standard	Intervalles de confiance (95%)		Khi-Carré	Pr > $\mathcal{X}^2$
Intercept	1	1.3476	0.0341	1.2809	1.4144	1565.44	<0.0001
x_1	1	0.1108	0.0469	0.0189	0.2027	5.58	0.0181
trt	1	-0.1080	0.0486	-0.2034	-0.0127	4.93	0.0264
x_1*trt	1	-0.3016	0.0697	-0.4383	-0.1649	18.70	<0.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

TAB. 4.5 – *Résultats obtenus sous l'hypothèse d'indépendance (un GLM est ici ajusté).*

Variable	Estimé	Erreur Standard	Intervalles de confiance (95%)		Khi-Carré	Pr > $\mathcal{X}^2$
Intercept	1.3476	0.1574	1.0392	1.6560	8.56	<0.0001
x_1	0.1108	0.1161	-0.1168	0.3383	0.95	0.3399
trt	-0.1080	0.1937	-0.4876	0.2716	-0.56	0.5770
x_1*trt	-0.3016	0.1712	-0.6371	0.0339	-1.76	0.0781

TAB. 4.6 – *Résultats obtenus sous l'hypothèse de corrélation des observations (structure d'équicorrélation).*

Il est à noter que les valeurs des estimés sont semblables sous l'hypothèse d'indépendance et sous l'hypothèse de corrélation des observations, alors que les estimés des erreurs standards sous l'indépendance sont inférieurs à ceux sous la corrélation des observations.

Le modèle obtenu est donc

$$\ln(\mu_{it}) = 1.3476 + 0.1108x_{1it} - 0.1080\text{trt}_{it} - 0.3016x_1\text{trt}_{it} + \ln(\text{temps})_{it}, \qquad (4.6)$$

et la matrice de corrélation ainsi obtenue est

$$\boldsymbol{R_i}(\boldsymbol{\alpha}) = \begin{pmatrix} 1.0000 & & & & \\ 0.5941 & 1.0000 & & & \\ 0.5941 & 0.5941 & 1.0000 & & \\ 0.5941 & 0.5941 & 0.5941 & 1.0000 & \\ 0.5941 & 0.5941 & 0.5941 & 0.5941 & 1.0000 \end{pmatrix}.$$

4.4.4 La validation du modèle obtenu

Valider un modèle de régression longitudinale est plus complexe que dans le cas où une seule mesure est prise par individu. En effet, comme la statistique de déviance est obtenue à partir de la log-vraisemblance et qu'aucune vraisemblance (ou log-vraisemblance) n'est calculée dans un cas où les mesures sont répétées, nous ne pouvons nous baser sur cette statistique pour décider si un modèle est acceptable ou non. Cependant, nous pouvons calculer la statistique PRESS qui est ici de 46 015.49. Si les analyses avaient été faites au seuil de 5%, l'interaction entre la variable x_1 et le traitement aurait été enlevée, et la statistique PRESS aurait été de 46 166.08, montrant que le modèle à 10% prévoit mieux le nombre de crises que le modèle à 5%, puisque la statistique PRESS est inférieure.

4.5 Conclusion du chapitre

Dans ce chapitre, les notions vues au chapitre 2 concernant la régression de Poisson ont été modifiées afin de tenir compte de la corrélation entre les mesures prises. Il arrive de plus que certaines variables mesurées aient un effet aléatoire sur la distribution conditionnelle de la variable aléatoire. Dans un tel cas, nous parlerons d'un modèle mixte. Par l'approche GEE, on ne modélise pas directement la corrélation entre les mesures. Cependant, l'ajout d'effets aléatoires au chapitre suivant permettera la modélisation de cette corrélation. Si la variable réponse possède une distribution faisant partie de la famille exponentielle, on parlera alors d'un modèle linéaire généralisé mixte (GLMM[1]). Les GLMM seront vus au chapitre 5.

[1] *Generalized Linear Mixed Model*

Chapitre 5

Les modèles linéaires généralisés mixtes

Cet avant-dernier chapitre théorique se veut une extension des modèles linéaires généralisés dans le cas où certains effets sont aléatoires. Il arrive à l'occasion que l'effet d'une variable sur la distribution conditionnelle de la variable réponse ne soit pas le même pour tous les individus à l'étude. Dans ce cas, cet effet doit être considéré comme étant aléatoire. La façon d'analyser un modèle contenant des effets aléatoires est particulière et sera vue dans ce cinquième chapitre. L'ajout d'effets aléatoires permettera donc de modéliser la corrélation entre les mesures, ce qui n'avait pas été possible par l'approche GEE du chapitre 3.

5.1 Les modèles linéaires généralisés mixtes (GLMM)

5.1.1 Des définitions

On définit un modèle linéaire généralisé mixte (GLMM) comme étant un modèle linéaire généralisé (section 2.3 en page 9), mais en admettant la présence d'effets aléatoires et/ou d'erreurs corrélées. De plus, on suppose que ces effets aléatoires suivent, comme pour le modèle linéaire mixte, une distribution normale. Conditionnellement aux effets aléatoires, la variable réponse suit une loi faisant partie de la famille exponentielle.

Agresti (2002) donne de plus les définitions suivantes[1] :

> *Les termes aléatoires prennent les mêmes valeurs pour chaque observation à l'intérieur d'un groupe, mais prennent des valeurs différentes pour des groupes différents. Ils ne sont pas observés, et, lorsqu'ils varient de façon aléatoire à l'intérieur d'un groupe, on les appelle effets aléatoires. Les variables qui décrivent l'effet d'un facteur dans des modèles linéaires sont appelés effets fixes. Ils s'appliquent à toutes les catégories qui nous intéressent, comme les sexes, les groupes d'âge ou les traitements. Les effets aléatoires s'appliquent quant à eux à un échantillon. Les modèles linéaires généralisés sont des modèles de régression ordinaires, mais en permettant à la variable réponse de suivre une distribution autre que la normale. Ils permettent de plus une fonction de lien pour la moyenne. Les modèles linéaires généralisés mixtes sont une extension des modèles linéaires généralisés, mais en admettant la présence d'effets aléatoires et d'effets fixes à l'intérieur du prédicteur linéaire.*

5.1.2 Les conditions d'utilisation d'un GLMM

Deux conditions distinctes doivent être respectées afin d'ajuster un modèle linéaire généralisé mixte :

1. Si le modèle contient des effets aléatoires, la distribution conditionnelle des données sachant les effets aléatoires est connue, et cette distribution fait partie de la famille exponentielle.
2. L'espérance conditionnelle des données prend la forme d'un modèle linéaire mixte ($\boldsymbol{X\beta} + \boldsymbol{Z\gamma}$) auquel on a appliqué une transformation monotone, $g^{-1}(\cdot)$:

$$\mathbb{E}[\boldsymbol{Y_i}|\boldsymbol{\gamma_i}] = g^{-1}(\boldsymbol{X_i\beta} + \boldsymbol{Z_i\gamma_i}),$$

où $\boldsymbol{X_i}$ et $\boldsymbol{Z_i}$ correspondent respectivement aux coefficients fixes et aux termes aléatoires du $i^{\text{ème}}$ individu ($i = 1, \ldots, n$), et où $\boldsymbol{Z}$ et $\boldsymbol{\gamma}$ seront définis à la sous-section 5.2.1 en page suivante.

[1] Le paragraphe suivant est une traduction libre d'un passage d'Agresti (2002).

5.2 Les modèles

5.2.1 Le modèle d'un GLMM

Dans un modèle linéaire, la valeur de la variable réponse doit être prédite. Pour un modèle linéaire généralisé, c'est plutôt l'*espérance* de cette variable réponse qui est prédite.

Pour un modèle contenant des effets aléatoires, le modèle s'exprime comme suit :

$$\mathbb{E}[\boldsymbol{Y}|\boldsymbol{\gamma}] = g^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma}) = g^{-1}(\boldsymbol{\eta}) = \boldsymbol{\mu} \tag{5.1}$$

où

- $\boldsymbol{Y}$ un vecteur de dimension $n \times 1$ de données observées ;
- $\boldsymbol{X}$ une matrice de dimension $n \times p$ de rang p qui détermine les valeurs prises par les différents coefficients fixes ;
- $\boldsymbol{\beta}$ un vecteur de dimension $p \times 1$ contenant les effets fixes ;
- $\boldsymbol{Z}$ une matrice $n \times r$ qui détermine les valeurs multipliant les différents termes d'effets aléatoires ;
- $\boldsymbol{\gamma}$ un vecteur de dimension $r \times 1$ contenant les effets aléatoires qui ne sont pas observés ;
- $g(\cdot)$ une fonction de lien monotone et différentiable ; $g^{-1}(\cdot)$ en est l'inverse.

Nous pouvons de plus faire l'hypothèse suivante :

$$\boldsymbol{\gamma} \sim \mathcal{N}_r(\mathbf{0}, \mathbf{G}).$$

De plus,

$$\mathbb{V}\mathrm{ar}[\boldsymbol{Y}|\boldsymbol{\gamma}] = a(\phi)\boldsymbol{A}^{\frac{1}{2}}\mathbf{R}\boldsymbol{A}^{\frac{1}{2}}, \tag{5.2}$$

avec $\boldsymbol{A}^{\frac{1}{2}}$ une matrice diagonale comme au chapitre 3, mais contenant $\sqrt{\frac{\mathbb{V}\mathrm{ar}(Y_i)}{a(\phi)}}$ sur la diagonale principale, et $\mathbf{R}$ et $\mathbf{G}$ sont, respectivement, deux matrices de corrélation et de variance dont les formes générales doivent être définies par l'utilisateur.

5.2.2 Le modèle GLMM pour données longitudinales

La principale différence ici est qu'au lieu d'observer chacun des n individus une fois, nous les observons en n_i temps distincts. Donc, le modèle s'exprime comme suit :

$$\mathbb{E}[\boldsymbol{Y_i}|\boldsymbol{\gamma_i}] = g^{-1}(\boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{\gamma_i}) = g^{-1}(\boldsymbol{\eta_i}) = \boldsymbol{\mu_i} \tag{5.3}$$

où

- $\boldsymbol{Y_i} = [Y_{i1}, \ldots, Y_{in_i}]'$ est un vecteur de dimension $n_i \times 1$ de données observées pour l'individu i. Chacun des vecteurs $\boldsymbol{Y_i}$ a toujours comme vecteur de moyenne conditionnelle $\boldsymbol{\mu_i} = [\mu_{i1}, \ldots, \mu_{in_i}]'$.
- $\boldsymbol{X_i}$ est une matrice de dimension $n_i \times p$ qui regroupe l'ensemble des variables explicatives de l'individu i. On a donc, comme au chapitre 3,

$$\boldsymbol{X_i} = [\boldsymbol{x_{i1}}, \ldots, \boldsymbol{x_{in_i}}]' = \begin{pmatrix} x_{i11} & x_{i12} & x_{i13} & \cdots & x_{i1p} \\ x_{i21} & x_{i22} & x_{i23} & \cdots & x_{i2p} \\ x_{i31} & x_{i32} & x_{i33} & \cdots & x_{i3p} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{in_i1} & x_{in_i2} & x_{in_i3} & \cdots & x_{in_ip} \end{pmatrix};$$

- $\boldsymbol{\beta}$ un vecteur de dimension $p \times 1$;
- $\boldsymbol{Z_i}$ une matrice de dimension $n_i \times r$ détermine les valeurs prises par les différents coefficients aléatoires. On a alors

$$\boldsymbol{Z_i} = [\boldsymbol{z_{i1}}, \ldots, \boldsymbol{z_{in_i}}]' = \begin{pmatrix} z_{i11} & z_{i12} & z_{i13} & \cdots & z_{i1r} \\ z_{i21} & z_{i22} & z_{i23} & \cdots & z_{i2r} \\ z_{i31} & z_{i32} & z_{i33} & \cdots & z_{i3r} \\ \vdots & \vdots & \vdots & & \vdots \\ z_{in_i1} & z_{in_i2} & z_{in_i3} & \cdots & z_{in_ir} \end{pmatrix};$$

- $\boldsymbol{\gamma}$ un vecteur de dimension $r \times 1$ contenant les effets aléatoires qui ne sont toujours pas observés.

Finalement, notons

$$\mathbb{V}\text{ar}[\boldsymbol{Y_i}|\boldsymbol{\gamma}] \quad = \quad a(\phi)\boldsymbol{A_i^{\frac{1}{2}}}\mathbf{R_i}\boldsymbol{A_i^{\frac{1}{2}}} \tag{5.4}$$

avec $\boldsymbol{A_i^{\frac{1}{2}}}$ une matrice diagonale comme à la section précédente, et qui contient $\sqrt{\frac{\mathbb{V}\text{ar}(Y_{it})}{a(\phi)}}$ sur la diagonale principale.

5.3 Les formes possibles des matrices G et R

Le Tableau 5.1 en page 51 présente les formes des matrices $\mathbf{G}$ et $\mathbf{R}$ les plus souvent rencontrées, alors que le Tableau 5.2 en page 52 présente des exemples de ces matrices. Dans ces tableaux, t est la dimension de la matrice $\mathbf{G}$ ou de la matrice $\mathbf{R}$.

La structure auto-régressive d'ordre 1 (AR(1)) est à utiliser lorsque l'on suppose que la variance, σ^2, est la même entre les mesures, et lorsque la covariance diminue entre les temps. La structure *« compound symmetry »*, quant à elle, suppose que toutes les observations ont la même variance $(\sigma^2 + \sigma_1)$ et les covariances sont toutes les mêmes (σ_1). La matrice n'imposant aucune structure particulière (UN) est la plus générale : toutes les variances et toutes les covariances sont différentes. La matrice à une bande (UN(1)) nécessite que toutes les covariances soient nulles et que toutes les variances soient différentes. Quant à la structure de composantes de la variance (VC), elle nécessite des groupes de variances pareilles, alors que les covariances entre les mesures sont nulles. Bien qu'il soit difficile de décider quelle structure employer, cette tâche peut être réalisée grâce aux critères AIC, HQIC, BIC ou CAIC présentés à la section 5.5 en page 55.

5.4 L'estimation des paramètres

McCulloch & Searle (2001) mentionnent qu'un effet fixe est considéré comme constant. Le but est de l'estimer. Un effet aléatoire est considéré comme un effet provenant d'une population de plusieurs effets. Nous parlerons donc de l'*estimation* des effets fixes et de la *prévision* des effets aléatoires.

Dans un modèle mixte, l'estimation et la prévision des paramètres sont beaucoup plus complexes que dans le cas d'un modèle où tous les effets seraient fixes. En effet, en plus d'estimer les p composantes du vecteur $\boldsymbol{\beta}$, il faut prévoir les r composantes des n vecteurs $\boldsymbol{\gamma_i}$ et toutes les composantes des matrices $\mathbf{R}$ et $\mathbf{G}$.

Structure	**Description**	**Nombre de paramètres à estimer**	**Élément** (i,j)
ANTE(1)	Ante-dépendance	$2t-1$	$\sigma_i\sigma_j\prod_{k=i}^{j-1}\rho_k$
AR(1)	Autorégressive(1)	2	$\sigma^2\rho^{\|i-j\|}$
ARH(1)	AR(1) Hétérogène	$t+1$	$\sigma_i\sigma_j\rho^{\|i-j\|}$
ARMA(1,1)	Moyenne Mobile Autorégressive	3	$\sigma^2[\gamma\rho^{\|i-j\|-1}I(i\neq j)+I(i=j)]$
CS	Compound Symmetry	2	$\sigma_1+\sigma^2I(i=j)$
CSH	CS Hétérogène	$t+1$	$\sigma_i\sigma_j[\rho I(i\neq j)+I(i=j)]$
FA(q)	Factor Analytic	$\frac{q}{2}(2t-q+1)+t$	$\sum_{k=1}^{\min(i,j,q)}\lambda_{ik}\lambda_{jk}+\sigma_i^2I(i=j)$
FA0(q)	FA sans diagonale	$\frac{q}{2}(2t-q+1)$	$\sum_{k=1}^{\min(i,j,q)}\lambda_{ik}\lambda_{jk}$
FA1(q)	FA diagonale égale	$\frac{q}{2}(2t-q+1)+1$	$\sum_{k=1}^{\min(i,j,q)}\lambda_{ik}\lambda_{jk}+\sigma^2I(i=j)$
HF	Huynh-Feldt	$t+1$	$(\sigma_i^2+\sigma_j^2)/2+\lambda I(i\neq j)$
TOEP	Toeplitz	t	$\sigma_{\|i-j\|+1}$
TOEP(q)	Toeplitz à q bandes	q	$\sigma_{\|i-j\|+1}I(\|i-j\|<q)$
TOEPH	TOEP Hétérogène	$2t-1$	$\sigma_i\sigma_j\rho_{\|i-j\|}$
TOEPH(q)	TOEP Hétérogène à q bandes	$t+q-1$	$\sigma_i\sigma_j\rho_{\|i-j\|}I(\|i-j\|<q)$
UN	Non structurée	$t(t+1)/2$	σ_{ij}
UN(q)	À q bandes	$\frac{q}{2}(2t-q+1)$	$\sigma_{ij}I(\|i-j\|<q)$
UNR	Corrélations sans structure	$t(t+1)/2$	$\sigma_i\sigma_j\rho_{\max(i,j)\min(i,j)}$
UNR(q)	Corrélations à q bandes	$\frac{q}{2}(2t-q+1)$	$\sigma_i\sigma_j\rho_{\max(i,j)\min(i,j)}$
VC	Composantes de la variance	q	$\sigma_k^2I(i=j)$ où i correspond au k^e effet

TAB. 5.1 – *Structures des matrices de variance-covariance* $\mathbf{R}$ *et* $\mathbf{G}$.

Structure	Exemple
ANTE(1)	$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_1\rho_2 \\ \sigma_2\sigma_1\rho_1 & \sigma_2^2 & \sigma_2\sigma_3\rho_2 \\ \sigma_3\sigma_1\rho_2\rho_1 & \sigma_3\sigma_2\rho_2 & \sigma_3^2 \end{pmatrix}$
AR(1)	$\sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}$
ARH(1)	$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho \\ \sigma_3\sigma_1\rho^2 & \sigma_3\sigma_2\rho & \sigma_3^2 \end{pmatrix}$
ARMA(1,1)	$\sigma^2 \begin{pmatrix} 1 & \gamma & \gamma\rho & \gamma\rho^2 \\ \gamma & 1 & \gamma & \gamma\rho \\ \gamma\rho & \gamma & 1 & \gamma \\ \gamma\rho^2 & \gamma\rho & \gamma & 1 \end{pmatrix}$
CS	$\begin{pmatrix} \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 \end{pmatrix}$
CSH	$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho \\ \sigma_3\sigma_1\rho & \sigma_3\sigma_2\rho & \sigma_3^2 \end{pmatrix}$
FA(1)	$\begin{pmatrix} \lambda_1^2+d_1 & \lambda_1\lambda_2 & \lambda_1\lambda_3 \\ \lambda_2\lambda_1 & \lambda_2^2+d_2 & \lambda_2\lambda_3 \\ \lambda_3\lambda_1 & \lambda_3\lambda_2 & \lambda_3^2+d_3 \end{pmatrix}$
HF	$\begin{pmatrix} \sigma_1^2 & \frac{\sigma_1^2+\sigma_2^2}{2}-\lambda & \frac{\sigma_1^2+\sigma_3^2}{2}-\lambda \\ \frac{\sigma_2^2+\sigma_1^2}{2}-\lambda & \sigma_2^2 & \frac{\sigma_2^2+\sigma_3^2}{2}-\lambda \\ \frac{\sigma_3^2+\sigma_1^2}{2}-\lambda & \frac{\sigma_3^2+\sigma_2^2}{2}-\lambda & \sigma_3^2 \end{pmatrix}$

TOEP	$\begin{pmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{pmatrix}$
TOEP(2)	$\begin{pmatrix} \sigma^2 & \sigma_1 & 0 & 0 \\ \sigma_1 & \sigma^2 & \sigma_1 & 0 \\ 0 & \sigma_1 & \sigma^2 & \sigma_1 \\ 0 & 0 & \sigma_1 & \sigma^2 \end{pmatrix}$
TOEPH	$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_2 \\ \sigma_2\sigma_1\rho_1 & \sigma_2^2 & \sigma_2\sigma_3\rho_1 \\ \sigma_3\sigma_1\rho_2 & \sigma_3\sigma_2\rho_1 & \sigma_3^2 \end{pmatrix}$
UN	$\begin{pmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} \\ \sigma_{21} & \sigma_2^2 & \sigma_{32} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix}$
UN(1)	$\begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}$
UNR	$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{21} & \sigma_1\sigma_3\rho_{31} \\ \sigma_2\sigma_1\rho_{21} & \sigma_2^2 & \sigma_2\sigma_3\rho_{32} \\ \sigma_3\sigma_1\rho_{31} & \sigma_3\sigma_2\rho_{32} & \sigma_3^2 \end{pmatrix}$
VC	$\begin{pmatrix} \sigma_B^2 & 0 & 0 \\ 0 & \sigma_B^2 & 0 \\ 0 & 0 & \sigma_{AB}^2 \end{pmatrix}$

TAB. 5.2 – *Exemples des matrices de variance-covariance* $\mathbf{R}$ *et* $\mathbf{G}$.

L'estimation des paramètres est faite, dans le cas des GLMM, de façon itérative. À chacune des itérations, une pseudo variable linéaire est calculée et un modèle mixte pondéré est alors ajusté. Cette pseudo variable linéaire est obtenue en faisant un développement en séries de Taylor[2] d'ordre 1 de $\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) = g^{-1}(\boldsymbol{X\beta} + \boldsymbol{Z\gamma})$ autour de $\widetilde{\boldsymbol{\beta}}$ et de $\widetilde{\boldsymbol{\gamma}}$, où $\widetilde{\boldsymbol{\beta}}$ et $\widetilde{\boldsymbol{\gamma}}$ sont des vecteurs contenant des valeurs données des paramètres $\boldsymbol{\beta}$ et $\boldsymbol{\gamma}$. Cela entraîne

$$\begin{aligned} \boldsymbol{\mu} &\approx g^{-1}(\widetilde{\boldsymbol{\eta}}) + \underbrace{\left(\frac{\partial g^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\right)_{\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\gamma}}}}_{\equiv \widetilde{\boldsymbol{\Delta}}} \boldsymbol{X}(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}) + \underbrace{\left(\frac{\partial g^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\right)_{\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\gamma}}}}_{\equiv \widetilde{\boldsymbol{\Delta}}} \boldsymbol{Z}(\boldsymbol{\gamma} - \widetilde{\boldsymbol{\gamma}}) \\ \Rightarrow \boldsymbol{\mu} &\approx g^{-1}(\widetilde{\boldsymbol{\eta}}) + \widetilde{\boldsymbol{\Delta}}\boldsymbol{X}(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}) + \widetilde{\boldsymbol{\Delta}}\boldsymbol{Z}(\boldsymbol{\gamma} - \widetilde{\boldsymbol{\gamma}}). \end{aligned} \tag{5.5}$$

On peut réarranger l'expression (5.5) de la façon suivante :

$$\begin{aligned} \boldsymbol{\mu} &= g^{-1}(\widetilde{\boldsymbol{\eta}}) + \widetilde{\boldsymbol{\Delta}}(\boldsymbol{X\beta} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}} + \boldsymbol{Z\gamma} - \boldsymbol{Z}\widetilde{\boldsymbol{\gamma}}) \\ \Rightarrow \widetilde{\boldsymbol{\Delta}}^{-1}(\boldsymbol{\mu} - g^{-1}(\widetilde{\boldsymbol{\eta}})) &= \boldsymbol{X\beta} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}} + \boldsymbol{Z\gamma} - \boldsymbol{Z}\widetilde{\boldsymbol{\gamma}} \\ \Rightarrow \widetilde{\boldsymbol{\Delta}}^{-1}(\boldsymbol{\mu} - g^{-1}(\widetilde{\boldsymbol{\eta}})) + \boldsymbol{X}\widetilde{\boldsymbol{\beta}} + \boldsymbol{Z}\widetilde{\boldsymbol{\gamma}} &= \boldsymbol{X\beta} + \boldsymbol{Z\gamma}. \end{aligned} \tag{5.6}$$

Comme $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{\gamma}] = \boldsymbol{\mu}$, il suit que

$$\begin{aligned} \mathbb{E}[(\widetilde{\boldsymbol{\Delta}}^{-1}(\boldsymbol{Y} - g^{-1}(\widetilde{\boldsymbol{\eta}})) + \boldsymbol{X}\widetilde{\boldsymbol{\beta}} + \boldsymbol{Z}\widetilde{\boldsymbol{\gamma}})|\boldsymbol{\gamma}] &= \widetilde{\boldsymbol{\Delta}}^{-1}(\mathbb{E}[\boldsymbol{Y}|\boldsymbol{\gamma}] - g^{-1}(\widetilde{\boldsymbol{\eta}})) + \boldsymbol{X}\widetilde{\boldsymbol{\beta}} + \boldsymbol{Z}\widetilde{\boldsymbol{\gamma}} \\ &= \widetilde{\boldsymbol{\Delta}}^{-1}(\boldsymbol{\mu} - g^{-1}(\widetilde{\boldsymbol{\eta}})) + \boldsymbol{X}\widetilde{\boldsymbol{\beta}} + \boldsymbol{Z}\widetilde{\boldsymbol{\gamma}} \\ &= \boldsymbol{X\beta} + \boldsymbol{Z\gamma} \\ &= g(\boldsymbol{\mu}). \end{aligned} \tag{5.7}$$

De plus, comme $\mathbb{V}\mathrm{ar}[\boldsymbol{Y}|\boldsymbol{\gamma}] = a(\phi)\boldsymbol{A}^{\frac{1}{2}}\mathbf{R}\boldsymbol{A}^{\frac{1}{2}}$, il s'ensuit que

$$\begin{aligned} \mathbb{V}\mathrm{ar}[(\widetilde{\boldsymbol{\Delta}}^{-1}(\boldsymbol{Y} - g^{-1}(\widetilde{\boldsymbol{\eta}})) + \boldsymbol{X}\widetilde{\boldsymbol{\beta}} + \boldsymbol{Z}\widetilde{\boldsymbol{\gamma}})|\boldsymbol{\gamma}] &= a(\phi)\widetilde{\boldsymbol{\Delta}}^{-1}\mathbb{V}\mathrm{ar}[\boldsymbol{Y}|\boldsymbol{\gamma}]\widetilde{\boldsymbol{\Delta}}^{-1} \\ &= a(\phi)\widetilde{\boldsymbol{\Delta}}^{-1}\boldsymbol{A}^{\frac{1}{2}}\mathbf{R}\boldsymbol{A}^{\frac{1}{2}}\widetilde{\boldsymbol{\Delta}}^{-1}. \end{aligned} \tag{5.8}$$

Pour simplifier l'écriture, posons $\mathbf{P} \equiv \widetilde{\boldsymbol{\Delta}}^{-1}(\boldsymbol{Y} - g^{-1}(\widetilde{\boldsymbol{\eta}})) + \boldsymbol{X}\widetilde{\boldsymbol{\beta}} + \boldsymbol{Z}\widetilde{\boldsymbol{\gamma}}$. Alors $\mathbb{E}[\mathbf{P}|\boldsymbol{\gamma}] = \widetilde{\boldsymbol{\Delta}}^{-1}(\boldsymbol{\mu} - g^{-1}(\widetilde{\boldsymbol{\eta}})) + \boldsymbol{X}\widetilde{\boldsymbol{\beta}} + \boldsymbol{Z}\widetilde{\boldsymbol{\gamma}} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma}$ et $\mathbb{V}\mathrm{ar}[\mathbf{P}|\boldsymbol{\gamma}] = a(\phi)\widetilde{\boldsymbol{\Delta}}^{-1}\boldsymbol{A}^{\frac{1}{2}}\mathbf{R}\boldsymbol{A}^{\frac{1}{2}}\widetilde{\boldsymbol{\Delta}}^{-1}$. Nous avons donc le modèle linéaire mixte suivant :

$$\mathbf{P} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon},$$

avec pseudo variable réponse $\mathbf{P}$, effets fixes $\boldsymbol{\beta}$ et effets aléatoires $\boldsymbol{\gamma}$. De plus, $\boldsymbol{\varepsilon}$ est un vecteur de composantes a effets aléatoires, où :

$$\begin{aligned} \mathbb{E}[\boldsymbol{\varepsilon}] &= \mathbf{0} \\ \mathbb{V}\mathrm{ar}[\boldsymbol{\varepsilon}] &= \mathbb{V}\mathrm{ar}[\mathbf{P}|\boldsymbol{\gamma}]. \end{aligned}$$

[2]Un développement en séries de Taylor d'ordre n de la fonction $f(x)$ autour du point $x = a$ est $f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n$.

Wolfinger & O'Connell (1993) proposent les étapes suivantes afin d'estimer les paramètres :

1. Obtenir un estimé de $\boldsymbol{\mu}$, dénoté $\hat{\boldsymbol{\mu}}$. Pour une loi de Poisson, $\hat{\boldsymbol{\mu}} = \boldsymbol{Y} + 0.5$, alors que pour une loi binomiale, $\hat{\boldsymbol{\mu}} = (\boldsymbol{Y} + 0.5)/2$;
2. Calculer $\mathbf{P} = \widetilde{\boldsymbol{\Delta}}^{-1}(\boldsymbol{Y} - g^{-1}(\widetilde{\boldsymbol{\eta}})) + \boldsymbol{X}\widetilde{\boldsymbol{\beta}} + \boldsymbol{Z}\widetilde{\boldsymbol{\gamma}} = \widetilde{\boldsymbol{\Delta}}^{-1}(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}) + g(\hat{\boldsymbol{\mu}})$;
3. Ajuster un modèle linéaire mixte pondéré avec variable réponse $\mathbf{P}$, effets fixes $\boldsymbol{\beta}$ et effets aléatoires $\boldsymbol{\gamma}$ en utilisant les méthodes du maximum de vraisemblance, ML[3], ou du maximum de vraisemblance restreint, REML[4] :
   Définissons $\mathbf{V}(\boldsymbol{\theta}) = \boldsymbol{Z}\mathbf{G}\boldsymbol{Z}' + a(\phi)\widetilde{\boldsymbol{\Delta}}^{-1}\boldsymbol{A}^{\frac{1}{2}}\mathbf{R}\boldsymbol{A}^{\frac{1}{2}}\widetilde{\boldsymbol{\Delta}}^{-1}$ avec $\boldsymbol{\theta}$ un vecteur de dimension $q \times 1$ contenant tous les paramètres inconnus des matrices $\mathbf{R}$ et $\mathbf{G}$.
   On a :
   $$\begin{aligned} \text{ML} : \ell(\boldsymbol{\theta}, \mathbf{p}) &= -\frac{1}{2}\log|\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2}\boldsymbol{r}'\mathbf{V}(\boldsymbol{\theta})^{-1}\boldsymbol{r} - \frac{n}{2}\log(2\pi) \\ \text{REML} : \ell(\boldsymbol{\theta}, \mathbf{p}) &= -\frac{1}{2}\log|\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2}\boldsymbol{r}'\mathbf{V}(\boldsymbol{\theta})^{-1}\boldsymbol{r} - \frac{1}{2}\log|\boldsymbol{X}'\mathbf{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X}| - \frac{n-p}{2}\log(2\pi), \end{aligned}$$
   avec $\boldsymbol{r} = \mathbf{p} - \boldsymbol{X}(\boldsymbol{X}'\mathbf{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{V}^{-1}\mathbf{p}$ et p est le rang de la matrice $\boldsymbol{X}$;
4. Estimer les paramètres $\hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{\gamma}}$ à l'aide des expressions suivantes :
   $$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\mathbf{V}(\hat{\boldsymbol{\theta}})\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{V}(\hat{\boldsymbol{\theta}})\mathbf{p} \\ \hat{\boldsymbol{\gamma}} &= \hat{\mathbf{G}}\boldsymbol{Z}'\mathbf{V}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{r}}; \end{aligned}$$
5. Calculer $\hat{\boldsymbol{\mu}} = g^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma})$;
6. Calculer $\mathbf{P} = \widetilde{\boldsymbol{\Delta}}^{-1}(\boldsymbol{Y} - g^{-1}(\widetilde{\boldsymbol{\eta}})) + \boldsymbol{X}\widetilde{\boldsymbol{\beta}} + \boldsymbol{Z}\widetilde{\boldsymbol{\gamma}}$;
7. Utiliser à nouveau les méthodes ML ou REML afin de trouver un nouveau vecteur $\hat{\boldsymbol{\theta}}$ en maximisant $\ell(\boldsymbol{\theta}, \mathbf{p})$;
8. Comparer les anciens estimés de $\mathbf{G}$ et $\mathbf{R}$ avec les nouveaux estimés. Si la différence est trop importante, passer à l'étape suivante, sinon arrêter. Par défaut, l'algorithme de la procédure GLIMMIX de SAS s'arrête si la différence entre les anciens et les nouveaux estimés est inférieure ou égale à 0.00000001 ;
9. Obtenir de nouveaux estimés des vecteurs $\hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{\gamma}}$ à l'aide des mêmes formules qu'à l'étape 4 ;
10. Reprendre les étapes 5 à 8 jusqu'à convergence.

Les estimés $\hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{\gamma}}$ sont appelés respectivement BLUE[5] et BLUP[6], qui sont les meilleurs estimateurs linéaires sans biais et les meilleurs prédicteurs linéaires sans biais.

[3] *Maximum Likelihood*
[4] *REstricted/REsidual Maximum Likelihood*
[5] *Best Linear Unbiased Estimator*
[6] *Best Linear Unbiased Predictor*

5.5 Le choix d'un modèle

Étant donné que les matrices **R** et **G** ne sont pas connues et que leurs structures doivent être spécifiées, plusieurs combinaisons de celles-ci peuvent être essayées. Mais comment prendre la décision à savoir quelle combinaison est la meilleure ? Plusieurs critères permettent d'aider à prendre cette décision. En fait, 2 critères sont souvent utilisés : ce sont les critères d'information d'Akaike (AIC) et de Schwarz (BIC). D'autres critères sont également disponibles, et le TABLEAU 5.3 suivant présente les différents critères et leur définition.

Critère	Expression
AIC	$-2\ell + 2d$
HQIC	$-2\ell + 2d\ln(\ln(n))$
BIC	$-2\ell + d\ln(n)$
CAIC	$-2\ell + d(\ln(n) + 1)$

TAB. 5.3 – *Critères de décision, où ℓ représente la valeur maximale de la log-vraisemblance ou de la quasi-log-vraisemblance, d est la dimension du modèle (nombre de paramètres du modèle), n est le nombre d'observations.*

Les valeurs de ces critères n'ont que très peu de signification lorsque prises individuellement. Cependant, on s'en sert afin de comparer des modèles entre eux : plus leur valeur est faible, plus le modèle est approprié.

5.6 Les tests d'hypothèses

Supposons que l'on veuille tester une hypothèse de la forme :

$$H_0 : \mathbf{L} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} = \mathbf{d}, \tag{5.9}$$

où **L** est une matrice de dimension $h \times (p + r)$, avec h le nombre d'hypothèses à tester et où **d** est un vecteur de dimension $h \times 1$. La statistique F utilisée afin de faire ce test

d'hypothèses est

$$F = \frac{\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix}' \mathbf{L}'(\mathbf{LCL}')^{-1}\mathbf{L} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} - \mathbf{d}}{\text{rang}(\mathbf{L})}, \tag{5.10}$$

qui, sous H_0, suit une loi de Fisher avec ν_1 et ν_2 degrés de liberté. On a que ν_1 est le rang de la matrice $\mathbf{L}$, alors que ν_2, plus complexe, est obtenu par une approximation comme celle de Satterthwaite, par exemple ; cette méthode sera décrite plus en détails à la sous-section 5.6.1 suivante. De plus, $\mathbf{C}$ est un estimé de la matrice de variance-covariance de $[\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}' - \boldsymbol{\gamma}']'$ donné par

$$\mathbf{C} \;=\; \begin{bmatrix} \boldsymbol{X}'\mathbf{S}^{-1}\boldsymbol{X} & \boldsymbol{X}'\mathbf{S}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\mathbf{S}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\mathbf{S}^{-1}\boldsymbol{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-}$$

avec $\mathbf{S} \equiv \widehat{\mathbb{V}\text{ar}}[\mathbf{P}|\boldsymbol{\gamma}] = a(\phi)\widetilde{\boldsymbol{\Delta}}^{-1}\boldsymbol{A}^{\frac{1}{2}}\mathbf{R}\boldsymbol{A}^{\frac{1}{2}}\widetilde{\boldsymbol{\Delta}}^{-1}$ et où « - » dénote l'inverse généralisé de la matrice. Si $\mathbf{L}$ n'a qu'une seule ligne, la statistique utilisée est plutôt

$$t = \frac{\mathbf{L}\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix}}{\sqrt{\mathbf{LCL}'}}, \tag{5.11}$$

qui, sous H_0, suit approximativement une loi de Student avec $\nu = \frac{2(\mathbf{L}'\mathbf{CL}^2)}{\boldsymbol{g}'\boldsymbol{A}\boldsymbol{g}}$ degrés de liberté.

5.6.1 La méthode de Satterthwaite

Cette approximation nécessite en premier lieu la décomposition spectrale de $\mathbf{LCL}'$= $\mathbf{U}'\mathbf{QU}$, avec $\mathbf{U}$ une matrice orthogonale des vecteurs propres de la matrice $\mathbf{LCL}'$, et $\mathbf{Q}$ une matrice diagonale contenant les ν_1 valeurs propres de la matrice $\mathbf{LCL}'$; les deux matrices $\mathbf{U}$ et $\mathbf{Q}$ sont de dimension $\nu_1 \times \nu_1$. Définissons maintenant $\boldsymbol{b_j}$ comme étant la $j^{\text{ème}}$ ligne de $\mathbf{UL}$. Posons finalement

$$\nu_j = \frac{2(Q_j)^2}{\boldsymbol{g_j'}\boldsymbol{A}\boldsymbol{g_j'}}. \tag{5.12}$$

Dans l'expression (5.12), Q_j est le $j^{\text{ème}}$ élément sur la diagonale de $\mathbf{Q}$ et $\boldsymbol{g_j}$ est le gradient de $\boldsymbol{b_j}\mathbf{C}\boldsymbol{b_j'}$ par rapport à $\boldsymbol{\theta}$ et évalué en $\hat{\boldsymbol{\theta}}$. La matrice $\boldsymbol{A}$ est la matrice de variance-covariance de $\hat{\boldsymbol{\theta}}$. Posons finalement

$$E = \sum_{j=1}^{\text{rang}(\mathbf{L})} \frac{\nu_j}{\nu_j - 2} I(\nu_j > 2), \tag{5.13}$$

où

$$I(\nu_j > 2) = \begin{cases} 1, & \text{si } \nu_j > 2, \\ 0, & \text{sinon.} \end{cases}$$

Les degrés de liberté ν_2 sont donc

$$\nu_2 = \begin{cases} \dfrac{2E}{E - \text{rang}(\mathbf{L})}, & \text{si } E > \text{rang}(\mathbf{L}), \\ 0, & \text{sinon.} \end{cases}$$

Chapitre 6

Les modèles additifs et les modèles additifs généralisés

Lors de la validation d'un modèle, il peut arriver que ce dernier démontre des lacunes et laisse donc place à l'amélioration. Ainsi, afin d'améliorer le modèle en question, la transformation d'une ou de plusieurs variables exogènes peut s'avérer nécessaire. Mais comment savoir quelle transformation appliquer à une variable ? Les méthodes de régression non paramétriques que sont les modèles additifs et les modèles additifs généralisés peuvent aider à trouver une réponse à cette question. Ce chapitre décrira en premier lieu les modèles additifs, et les modèles additifs généralisés suivront.

6.1 Les modèles additifs

Il est possible de définir les modèles additifs comme étant une généralisation du modèle linéaire, mais en admettant une *fonction* d'une ou de plusieurs variables explicatives au lieu d'inclure ces variables de façon linéaire. Ainsi, les modèles linéaires et

les modèles additifs sont respectivement exprimés comme suit :

$$Y_i = \beta_0 + \sum_{j=1}^{p'} \beta_j X_{ij} + \varepsilon_i \tag{6.1}$$

$$Y_i = \beta_0 + \sum_{j=1}^{p'} f_j(X_{ij}) + \varepsilon_i, \tag{6.2}$$

où $i = 1, \ldots, n$ représente le nombre d'observations et $f_j(\cdot)$, $j = 1, \ldots, p'$, sont des fonctions non spécifiées à estimer à partir des données.

6.1.1 La relation entre les modèles linéaires et les modèles additifs

Les modèles linéaires et les modèles additifs partagent la même propriété d'être additifs en ce qui a trait aux variables explicatives. De plus, les deux modèles sont semblables au niveau des postulats. Ainsi, pour les modèles additifs et pour les modèles linéaires, les termes d'erreurs ε_i sont indépendants entre eux, $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ et $\mathbb{V}\text{ar}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I_n}$. De plus, on note que les f_j sont des fonctions univariées. Si le modèle (6.2) est bien spécifié, il s'ensuit que

$$\mathbb{E}\left\{ Y_i - \beta_0 - \sum_{j \neq k} f_j(X_{ij}) | X_{ik} \right\} = f_k(X_{ik}).$$

6.1.2 Un algorithme afin de trouver les $\hat{\boldsymbol{f}}_{\boldsymbol{j}}$

Hastie & Tibshirani (1990) proposent cet algorithme afin d'estimer les fonctions des variables exogènes :

1. Poser $\hat{\beta}_0 = \sum_{i=1}^{n} Y_i/n = \overline{Y}$. Poser de plus $f_j = f_j^0$, $j = 1, \ldots, p'$, un estimé initial des fonctions $f_j(\cdot)$. Un choix raisonnable est $f_j^0(X_{ij}) = \hat{\beta}_j X_{ij}$.
2. Pour $j = 1, \ldots, p', 1, \ldots, p', \ldots,$, calculer $f_j = S_j(\boldsymbol{Y} - \hat{\beta}_0 - \sum_{k \neq j} \boldsymbol{f_k} | \boldsymbol{X_j})$, où S_j est une fonction de lissage de $\boldsymbol{Y}$ sur $\boldsymbol{X_j}$, comme par exemple un estimateur du noyau ou des plus proches voisins (Hastie & Tibshirani, 1990).
3. Répéter la deuxième étape jusqu'à ce que les fonctions f_j individuelles soient constantes (varient très peu).

6.2 Les modèles additifs généralisés

Si les modèles additifs sont une généralisation des modèles linéaires, les modèles additifs généralisés (GAM) en sont une des GLM (section 2.3 de la page 9). De façon analogue, les GAM remplacent la composante linéaire $\beta_0 + \sum_j \beta_j X_{ij}$ par la composante additive $\beta_0 + \sum_j f_j(X_{ij})$. Ainsi, notons

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{p'} f_j(X_{ij}),$$

où $g(\cdot)$ est une fonction de lien monotone, différentiable, connue et non linéaire, et $g^{-1}(\cdot)$ en est l'inverse.

6.2.1 Un algorithme afin de trouver les $\hat{f}_j$

Dans le but de trouver les transformations f_j à appliquer aux variables exogènes, Hastie & Tibshirani (1990) proposent cet algorithme :

1. Poser $\hat{\beta}_0 = g\left(\sum_{i=1}^{n} Y_i/n\right) = g\left(\overline{Y}\right)$ et $f_1^0 = \cdots = f_{p'}^0 = 0$.
2. Calculer la variable dépendante z_i, où

$$z_i = \eta_i^0 + (Y_i - \mu_i^0)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)_0$$

avec $\eta_i^0 = \beta_0^0 + \sum_{j=1}^{p'} f_j^0(X_{ij})$ et $\mu_i^0 = g^{-1}(\eta_i^0)$.
Calculer les poids

$$w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)_0^2 (V_i^0)^{-1}$$

avec $(V_i^0)^{-1}$ la variance de Y_i à μ_i. Ajuster un modèle additif pondéré par les poids w_i à z_i afin d'obtenir les fonctions f_j^1, les prédicteurs additifs η^1 et les valeurs μ_i^1. Calculer le critère de convergence

$$\Delta(\eta^1, \eta^0) = \frac{\sum_{j=1}^{p'} ||f_j^1 - f_j^0||}{\sum_{j=1}^{p'} ||f_j^0||}.$$

3. Répéter la deuxième étape en remplaçant η^0 par η^1, et ce, jusqu'à ce que $\Delta(\eta^1, \eta^0)$ converge.

Les algorithmes présentés dans ce chapitre peuvent être exécutés grâce à la fonction `gam()` de R ou Splus. Une explication et une application de cette fonction se trouvent à la section suivante.

6.3 Exemple sur les modèles additifs généralisés

De la même façon que l'ont fait Breiman & Freidman (1985), générons des vecteurs $\boldsymbol{X_1}, \boldsymbol{X_2}$ et $\boldsymbol{\varepsilon}$ tirés d'une loi normale de moyenne nulle et de variance unité. Posons de plus

$$\ln(Y_i) = 2X_{i1}^2 - 3X_{i2} + \varepsilon_i, \ i = 1, \ldots, 100$$

Nous avons en premier lieu appliqué une transformation aux variables X_1 et X_2 et les seuils associés à ces transformations sont présentés au Tableau 6.1.

Nous testons donc ces deux hypothèses (Chambers & Hastie (1992), chapitre 7) :

H_0 : La transformation $s(X_k)$ est nécessaire ;
H_1 : La transformation $s(X_k)$ n'est pas nécessaire.

Ce tableau montre que chacun des termes a 1 degré de liberté pour la partie linéaire et 3 degrés de liberté pour le partie non linéaire[1]. De plus, d'après l'allure de la Figure 6.1 en page 63, on voit qu'une transformation de X_2 n'est pas nécessaire, car le graphique semble linéaire. On s'assure à l'aide des résultats présentés au Tableau 6.1 que la transformation de X_2 n'est pas nécessaire, car le seuil qui lui est associé est de 0.5430. Ainsi, un nouveau modèle GAM sera ajusté, mais en omettant la fonction de X_2.

Selon la Figure 6.2, l'inclusion de la variable X_1 élevée au carré est nécessaire puisque le seuil présenté au Tableau 6.2 est très fortement significatif ($<< 5\%$).

Ainsi, selon les deux ajustements de modèles GAM, nous concluons que nous devons inclure dans le modèle la variable X_1 mise au carré, alors que la variable X_2 doit être

[1]Le calcul des degrés de liberté est présenté dans Chambers & Hastie (1992), chapitre 7

gardée comme telle. Puisque le vecteur $\boldsymbol{Y}$ a été obtenu de la façon suivante :

$$\ln(Y_i) = 2X_{i1}^2 - 3X_{i2} + \varepsilon_i,\ i = 1, \ldots, 100,$$

où les vecteurs $\boldsymbol{X_1}, \boldsymbol{X_2}$ et $\boldsymbol{\varepsilon}$ sont tirés d'une loi normale de moyenne nulle et de variance unité, nous concluons que les résultats obtenus sont ceux espérés selon la simulation.

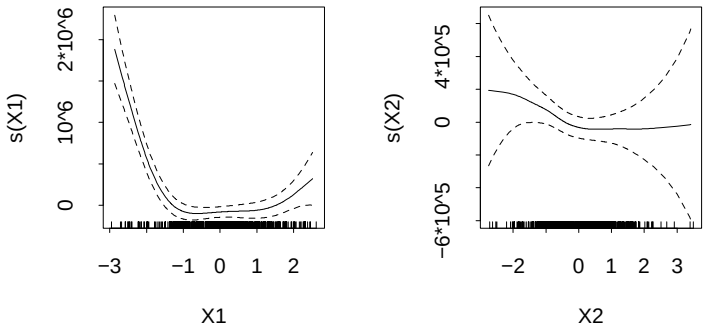

FIG. 6.1 – *Résultats de la première étape.*

	Df	Npar	Df	Npar F	Pr(F)
(Intercept)	1				
s(X1)	1		3	34.50938	0.0000000
s(X2)	1		3	0.71546	0.5430504

TAB. 6.1 – *Seuils associés à chacune des transformations des variables du modèle (première étape).*

	Df	Npar	Df	Npar F	Pr(F)
(Intercept)	1				
s(X1)	1		3	34.08707	0.0000000
s(X2)	1				

TAB. 6.2 – *Seuils associés à chacune des transformations des variables du modèle (deuxième étape).*

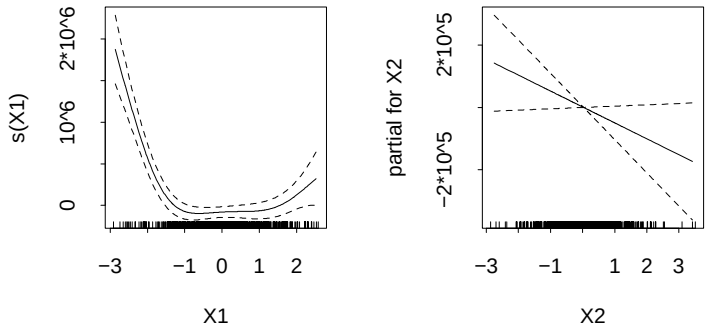

FIG. 6.2 – *Résultats de la deuxième étape.*

6.4 Conclusion du chapitre

Les modèles additifs généralisés ont de belles propriétés qui font en sorte qu'ils aident à trouver une bonne transformation à appliquer aux variables. Ils seront utilisés dans les chapitres suivants afin de tenter d'améliorer les modèles obtenus.

Deuxième partie

La pratique

Chapitre 7

L'explication des données, des variables et des analyses envisagées

Dans les deux chapitres pratiques de ce mémoire (chapitres 8 et 9), les méthodes décrites aux chapitres précédents seront appliquées à une base de données réelle. Ce chapitre se veut donc une explication complète de cette base de données et des variables utilisées pour les analyses. De plus, les analyses envisagées seront expliquées à la fin du chapitre.

7.1 L'explication des bases de données

Deux bases de données sont disponibles pour les analyses. La première contient les informations sur les crimes commis, tandis que l'autre contient les informations sur les variables psychiatriques des 378 individus à l'étude.

Ces 378 individus à l'étude sont des personnes ayant été admises dans un centre pour jeunes de Toronto entre 1986 et 1995. L'âge des individus au moment de leur admission au centre variait entre 16.1 ans et 24.4 ans (moyenne = 17.6 ans, écart-type = 0.9 an). Pendant une période d'environ 10 ans, les crimes commis par ces gens ont été enregistrés. On a entre autres noté le type du crime et la date à laquelle ce crime a été commis. Les individus analysés sont tous nés entre le 17 septembre 1967 et le 16 janvier 1979. Le suivi s'est terminé le 18 mars 2001 pour 334 des 378 individus. Pour les

autres, on les a suivis jusqu'à des dates variant entre le 27 mai 1993 et le 16 septembre 2001. Au moment où le suivi s'est terminé, les individus étaient tous âgés entre 18.8 ans et 33.5 ans (moyenne = 27.8 ans, écart-type = 2.8 ans).

Dans la première base de données, nous disposons des dates de naissance, des dates auxquelles les individus ont été reconnus coupables pour des crimes et du type de crime ayant été commis. Quant aux types de crimes, ils varient de 1 à 31, et la signification des types ainsi que leurs fréquences sont présentées au TABLEAU 7.1 suivant.

Type	Signification	Fréquence d'apparition (%)
1	Propriété	30.11
2	Violence	13.73
3	Drogues	4.23
4	Autres	18.84
5	Propriété + Autres	13.21
6	Violence + Autres	4.59
7	Drogues + Autres	1.24
8	Propriété + Violence	4.49
9	Propriété + Drogues	0.56
10	Violence + Drogues	0.26
11	Propriété + Drogues + Autres	0.50
12	Violence + Drogues + Autres	0.18
13	Propriété + Violence + Drogues	0.16
14	Propriété + Violence + Autres	3.99
15	Propriété + Violence + Drogues + Autres	0.36
16	Sexe	2.00
17	Sexe + Propriété	0.08
18	Sexe + Violence	0.42
19	Sexe + Drogues	0.00
20	Sexe + Autres	0.26
21	Sexe + Propriété + Autres	0.08
22	Sexe + Violence + Autres	0.14
23	Sexe + Drogues + Autres	0.00
24	Sexe + Propriété + Violence	0.10
25	Sexe + Propriété + Drogues	0.00
26	Sexe + Violence +Drogues	0.00
27	Sexe + Propriété + Drogues + Autres	0.02
28	Sexe + Violence + Drogues + Autres	0.04
29	Sexe + Propriété + Violence + Drogues	0.00
30	Sexe + Violence + Propriété + Autres	0.12
31	Sexe + Violence + Propriété + Drogues + Autres	0.02

TAB. 7.1 – *Signification et fréquence des types de crimes.*

La deuxième base de données, très différente de la première, contient les variables de diagnostics psychiatriques des 378 mêmes individus. Ainsi, pour chacun d'eux, nous possédons les informations à propos de la présence ou de l'absence de 19 caractères psychiatriques tels des problèmes de consommation (drogues, alcool), de conduite, de dépression, de paranoïa, de communication et/ou d'apprentissage, etc.

7.1.1 Les buts des analyses

Le but général de l'analyse sera de prévoir le comportement des individus après l'âge de 18 ans, selon certaines variables définies à partir uniquement de l'information disponible avant l'âge de 18 ans. Ainsi, nous utiliserons les modèles linéaires généralisés afin de faire cette prévision.

La première analyse aura donc pour but de déterminer si la trajectoire criminelle d'un individu pour certains types de crimes commis avant 18 ans peut aider à prévoir la trajectoire criminelle de l'individu après 18 ans. Nous tenterons donc de prévoir les nombres de crimes de type « Violence », « Drogues » et le nombre total de crimes (tous types confondus) à partir de plusieurs variables mesurées avant 18 ans. Ces variables sont présentées à la section 7.2 qui suit. Dans un premier temps, nous utiliserons la régression de Poisson longitudinale afin de modéliser l'effet des variables explicatives sur chacune des variables endogènes. Dans un deuxième temps, ce sera la régression de Poisson qui sera utilisée afin de faire les mêmes analyses, mais en considérant les classes d'âge séparément.

Le but de la deuxième analyse est plutôt de voir si le nombre total de crimes commis entre 18 et 20 ans peut être prédit à partir des variables de diagnostics psychiatriques mesurées avant 18 ans. Cette analyse se fera plutôt à l'aide de modèles linéaires généralisés mixtes. Le but de cette nouvelle analyse sera donc de voir si les modèles obtenus précédemment pour prévoir le nombre total de crimes entre 18 et 20 ans peuvent être améliorés en utilisant des effets aléatoires sur l'individu. Les variables de diagnostics psychiatriques utilisées dans cette deuxième analyse sont présentées à la section 7.2 suivante.

7.2 L'explication des variables

L'état de la première base de données ne permettait pas de faire les analyses voulues. Ainsi, nous avons dû transformer le fichier de données afin d'obtenir les variables réponses, explicatives et *offset* qui seront utilisées. Trois variables réponses, vingt-cinq variables explicatives et une variable *offset* ont été créées, et en voici les significations :

Variables réponses

Viol : Représente le nombre de crimes de type « Violence » (types 2-6-8-10-12-13-14-15-18-22-24-26-28-29-30-31). Cette variable est répétée afin de représenter les nombres de crimes de ce type entre 18 et 20 ans, 20 et 22 ans, 22 et 24 ans.

Drug : Représente le nombre de crimes de type « Drogues » (types 3-7-9-10-11-12-13-15-19-23-25-26-27-28-29-31). Cette variable est répétée afin de représenter les nombres de crimes de ce type entre 18 et 20 ans, 20 et 22 ans, 22 et 24 ans.

Off : Représente le nombre total de crimes. Cette variable est répétée afin de représenter le nombre total de crimes entre 18 et 20 ans, 20 et 22 ans, 22 et 24 ans.

Variables explicatives

Age1st : Âge au premier délit de chacun des 378 individus de l'étude.

Viol1218 : Variable indicatrice prenant la valeur 1 si l'individu a commis au moins un crime de type « Violence » entre 12 et 18 ans, 0 sinon.

Drug1218 : Variable indicatrice prenant la valeur 1 si l'individu a commis au moins un crime de type « Drogues » entre 12 et 18 ans, 0 sinon.

Sex1218 : Variable indicatrice prenant la valeur 1 si l'individu a commis au moins un crime de type « Sexe » entre 12 et 18 ans, 0 sinon.

Off1218 : Variable indicatrice prenant la valeur 1 si l'individu a commis au moins un crime entre 12 et 18 ans, 0 sinon.

Prop1218 : Variable indicatrice prenant la valeur 1 si l'individu a commis au moins un crime de type « Propriété » entre 12 et 18 ans, 0 sinon.

Autres1218 : Variable indicatrice prenant la valeur 1 si l'individu a commis au moins un crime de type « Autres » entre 12 et 18 ans, 0 sinon.

Nviol1214, nviol1416, nviol1618 : Représentent le nombre de crimes de type « Violence » commis dans chacune des trois classes d'âge (12-14 ans, 14-16 ans, 16-18 ans).

Ndrug1214, ndrug1416, ndrug1618 : Représentent le nombre de crimes de type « Drogues » commis dans chacune des trois classes d'âge (12-14 ans, 14-16 ans, 16-18 ans).

Nsex1214, nsex1416, nsex1618 : Représentent le nombre de crimes de type « Sexe » (types 16-17-18-19-20-21-22-23-24-25-26-27-28-29-30-31) dans chacune des trois classes d'âge (12-14 ans, 14-16 ans, 16-18 ans).

Noff1214, noff1416, noff1618 : Représentent le nombre total de crimes dans chacune des trois classes d'âge (12-14 ans, 14-16 ans, 16-18 ans).

Nprop1214, nprop1416, nprop1618 : Représentent le nombre de crimes de type « Propriété » (types 1-5-8-9-11-13-14-15-17-21-24-25-27-29-30-31) dans chacune des trois classes d'âge (12-14 ans, 14-16 ans, 16-18 ans).

Nautres1214, nautres1416, nautres1618 : Représentent le nombre de crimes de type « Autres » (types 4-5-6-7-11-12-14-15-20-21-22-23-27-28-30-31) dans chacune des trois classes d'âge (12-14 ans, 14-16 ans, 16-18 ans).

Variable *offset*

t : Variable représentant le temps exact, en années, passé dans chacune des classes d'âge (18-20 ans, 20-22 ans, 22-24 ans) pour chacun des 378 individus. Elle représente donc le temps où les individus étaient à risque de commettre des crimes dans chacune des 3 classes d'âge.

Le Tableau 7.2 en page suivante représente les statistiques descriptives de chacune des variables. Il est clair de ce tableau que la variable ndrug1214 ne peut être utilisée pour les analyses, étant donné qu'aucun individu n'a commis de crime de type « Drogues » entre les âges 12 et 14 ans. De plus, 376 (ou 374) données sont disponibles pour les variables réponse. Cela est dû au fait que la variable *offset* vaut 0 dans certaines classes d'âge, et alors l'observation doit être éliminée des analyses.

	Classes d'âge	N	Moyenne	Écart-Type	Minimum	Maximum
Viol	18-20 ans	376	0.6810	0.9630	0	5
	20-22 ans	376	0.6170	0.9310	0	5
	22-24 ans	374	0.5110	0.9430	0	7
Drug	18-20 ans	376	0.3090	0.6450	0	3
	20-22 ans	376	0.2500	0.6120	0	5
	22-24 ans	374	0.2540	0.5450	0	3
Off	18-20 ans	376	2.6910	2.6850	0	15
	20-22 ans	376	1.8380	2.2070	0	12
	22-24 ans	374	1.4280	1.8020	0	10
Nviol	12-14 ans	378	0.0820	0.3660	0	3
	14-16 ans	378	0.2650	0.5950	0	3
	16-18 ans	378	0.9210	1.2340	0	8
Ndrug	12-14 ans	378	0	0	0	0
	14-16 ans	378	0.0610	0.2710	0	3
	16-18 ans	378	0.3040	0.6310	0	4
Nsex	12-14 ans	378	0.0110	0.1260	0	2
	14-16 ans	378	0.0450	0.2200	0	2
	16-18 ans	378	0.1530	0.4280	0	3
Noff	12-14 ans	378	0.4420	1.0030	0	8
	14-16 ans	378	1.3970	1.8470	0	11
	16-18 ans	378	3.4230	2.3610	0	13
Nprop	12-14 ans	378	0.3070	0.7430	0	5
	14-16 ans	378	0.8920	1.2320	0	7
	16-18 ans	378	1.9020	1.7140	0	9
Nautres	12-14 ans	378	0.1190	0.4600	0	4
	14-16 ans	378	0.4950	1.1610	0	8
	16-18 ans	378	1.5210	1.7280	0	11
Age1st		378	15.4640	1.7450	8.8490	21.2590

TAB. 7.2 – *Statistiques descriptives.*

La FIGURE 7.1 suivante représente quant à elle les variations des moyennes entre chacune des classes d'âge pour chacun des types de crimes. De cette figure, on voit que le nombre moyen de crimes (peu importe le type) augmente de 12 à 18 ans. À partir de l'âge adulte, les individus ont tendance à s'assagir et ainsi à commettre moins de crimes.

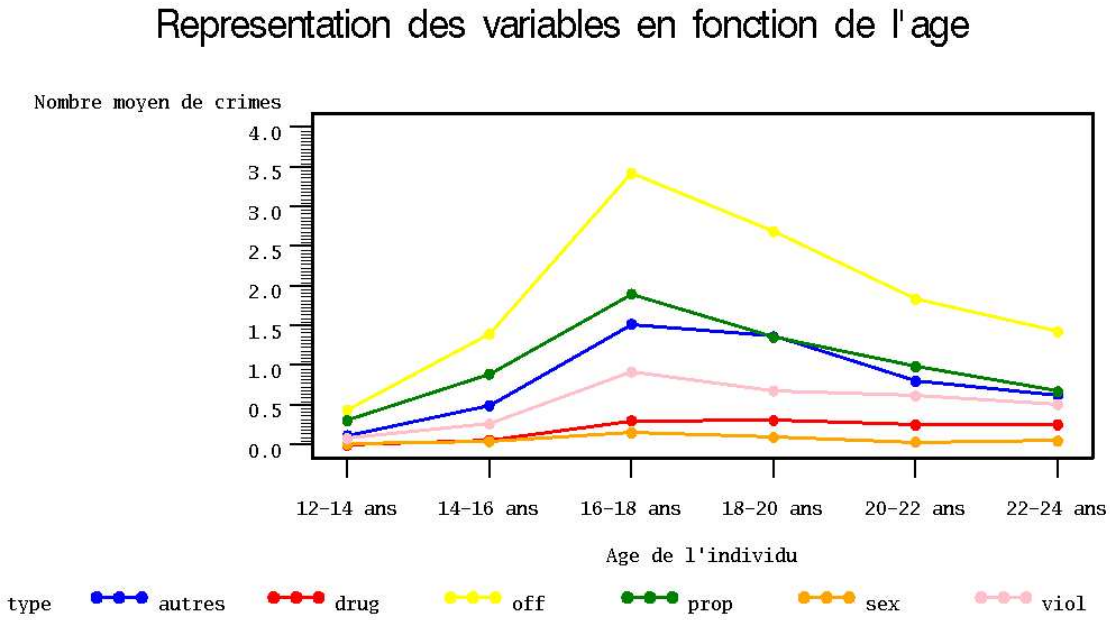

FIG. 7.1 – *Représentation des moyennes de chaque type de crimes en fonction de l'âge.*

Pour la deuxième base de données, dix-neuf variables explicatives sont disponibles. Ces variables ainsi que leur signification sont les suivantes :

adhd : Troubles déficitaires de l'attention avec hyperactivité

adjust : Troubles d'adaptation

anxiety : Troubles anxieux

conduct : Troubles de la conduite

dev : Troubles du développement

dbd : Comportements perturbateurs

dissoc : Troubles dissociatifs

impulse : Troubles du contrôle des impulsions

mood : Troubles de l'humeur

mental : Troubles mentaux

organic : Troubles cognitifs

person : Troubles de la personnalité

subst : Troubles liés à une substance

psychot : Troubles psychotiques

sexual : Troubles sexuels et troubles de l'identité sexuelle

sleep : Troubles du sommeil

somato : Troubles somatoformes

ld : Troubles des apprentissages et tics

comm : Troubles de la communication

De plus, une vingtième variable, l'âge au premier délit (age1st), et une autre variable, age, qui est le point milieu de chacune des classes d'âge, seront incluses dans les analyses.

7.3 Les analyses envisagées

Deux analyses distinctes seront faites dans les chapitres qui suivent. La première utilisera les modèles linéaires généralisés (longitudinaux ou non) vus au chapitre 4 (ou au chapitre 2), alors que la deuxième utilisera les modèles linéaires généralisés longitudinaux mixtes vus au chapitre 5.

La première analyse, présentée au chapitre 8, se divisera en 2 parties : la première partie utilisera les modèles linéaires généralisés longitudinaux afin de trouver un modèle global prédisant le nombre de crimes (de type « Violence », « Drogues » ou nombre de

crimes totaux). Ainsi, le modèle suivant sera ajusté aux données :

$$\ln(Y)_{it} = \beta_0 + \beta_1 \text{age1st}_i + \beta_2 \text{viol1218}_i + \beta_3 \text{drug1218}_i + \beta_4 \text{sex1218}_i + \beta_5 \text{off1218}_i + \beta_6 \text{prop1218}_i + \beta_7 \text{autres1218}_i + \beta_8 \text{nviol1214}_i + \beta_9 \text{nviol1416}_i + \beta_{10} \text{nviol1618}_i + \beta_{11} \text{ndrug1416}_i + \beta_{12} \text{ndrug1618}_i + \beta_{13} \text{nsex1214}_i + \beta_{14} \text{nsex1416}_i + \beta_{15} \text{nsex1618}_i + \beta_{16} \text{noff1214}_i + \beta_{17} \text{noff1416}_i + \beta_{18} \text{noff1618}_i + \beta_{19} \text{nprop1214}_i + \beta_{20} \text{nprop1416}_i + \beta_{21} \text{nprop1618}_i + \beta_{22} \text{nautres1214}_i + \beta_{23} \text{nautres1416}_i + \beta_{24} \text{nautres1618}_i + \ln(t)_{it},$$

$$i = 1, \ldots, n = 378,$$

et

$$t = \begin{cases} \text{classe d'âge 18-20 ans} \\ \text{classe d'âge 20-22 ans} \\ \text{classe d'âge 22-24 ans.} \end{cases}$$

Le modèle ci-haut mentionné sera, dans un premier temps, ajusté en utilisant le nombre de crimes de type « Violence » comme variable endogène. Ensuite, ce sera le nombre de crimes de type « Drogues » et, finalement, le nombre de crimes totaux qui seront les variables endogènes du modèle.

Dans la deuxième partie de la première analyse, ce sont plutôt les modèles linéaires généralisés qui seront utilisés afin de trouver des modèles prédisant les 3 mêmes variables réponses, mais séparément selon les tranches d'âge (18-20 ans, 20-22 ans et 22-24 ans). Donc, au total, 9 modèles seront ajustés à l'aide des mêmes variables explicatives.

Pour la première analyse, la procédure GENMOD sera utilisée et, pour la première partie uniquement, l'énoncé REPEATED de la procédure GENMOD sera utile.

Dans la deuxième analyse, présentée au chapitre 9, des effets aléatoires seront utilisés afin de voir si les modèles peuvent être améliorés. Cette fois, nous tenterons de prédire les nombres de crimes totaux entre 18 et 20 ans à partir des variables de diagnostics psychiatriques et à partir de l'âge au premier délit, en n'utilisant que l'information disponible avant 18 ans. Dans un premier temps, le modèle suivant sera ajusté aux données :

$$\ln(\text{off})_{it^*} = (\beta_0 + \gamma_{i01}) + \beta_1 \text{age}_{it^*} + \beta_2 \text{age}^2_{it^*} + \ln(t)_{it^*}$$

$$i = 1, \ldots, 360^{[1]}$$

avec

$$t^* = \begin{cases} \text{classe d'âge 12-14 ans} \\ \text{classe d'âge 14-16 ans} \\ \text{classe d'âge 16-18 ans} \end{cases}$$

et

$$\text{age} = \begin{cases} 13 & \text{si } t^* = \text{12-14 ans} \\ 15 & \text{si } t^* = \text{14-16 ans} \\ 17 & \text{si } t^* = \text{16-18 ans.} \end{cases}$$

Ce deuxième modèle sera par la suite ajusté aux données :

$$\ln(\text{off})_{it^*} = (\beta_0 + \gamma_{i02}) + \beta_1 \text{age}_{it^*} + \beta_2 \text{age}^2_{it^*} + \beta_3 \text{age1st}_i + \beta_4 \text{adhd}_i + \beta_5 \text{adjust}_i + \beta_6 \text{anxiety}_i + \beta_7 \text{conduct}_i + \beta_8 \text{dev}_i + \beta_9 \text{dbd}_i + \beta_{10} \text{dissoc}_i + \beta_{11} \text{impulse}_i + \beta_{12} \text{mood}_i + \beta_{13} \text{mental}_i + \beta_{14} \text{organic}_i + \beta_{15} \text{person}_i + \beta_{16} \text{subst}_i + \beta_{17} \text{psychot}_i + \beta_{18} \text{sexual}_i + \beta_{19} \text{sleep}_i + \beta_{20} \text{somato}_i + \beta_{21} \text{ld}_i + \beta_{22} \text{comm}_i + \ln(t)_{it^*}.$$

Les variables age et age1st sont forcées à demeurer dans le modèle, alors que les autres variables seront sélectionnées par la méthode d'exclusion avec un seuil de 5%. Pour cette partie d'analyse, la procédure GLIMMIX sera utilisée. L'énoncé RANDOM sera utile afin d'obtenir les effets aléatoires sur l'ordonnée à l'origine. Les deux vecteurs $\boldsymbol{\gamma}$ seront utilisés ultérieurement comme variables explicatives à effet fixe afin d'ajuster ces modèles avec GENMOD et l'énoncé REPEATED :

$$\ln(\text{off})_{it'} = \beta_0 + \beta_1 \gamma_{i01} + \beta_2 \text{age1st}_i + \beta_3 \text{adhd}_i + \beta_4 \text{adjust}_i + \beta_5 \text{anxiety}_i + \beta_6 \text{conduct}_i + \beta_7 \text{dev}_i + \beta_8 \text{dbd}_i + \beta_9 \text{dissoc}_i + \beta_{10} \text{impulse}_i + \beta_{11} \text{mood}_i + \beta_{12} \text{mental}_i + \beta_{13} \text{organic}_i + \beta_{14} \text{person}_i + \beta_{15} \text{subst}_i + \beta_{16} \text{psychot}_i + \beta_{17} \text{sexual}_i + \beta_{18} \text{sleep}_i + \beta_{19} \text{somato}_i + \beta_{20} \text{ld}_i + \beta_{21} \text{comm}_i + \ln(t)_{it'},$$

$$\ln(\text{off})_{it'} = \beta_0 + \beta_1 \gamma_{i02} + \beta_2 \text{age1st}_i + \beta_3 \text{adhd}_i + \beta_4 \text{adjust}_i + \beta_5 \text{anxiety}_i + \beta_6 \text{conduct}_i + \beta_7 \text{dev}_i +$$

[1] 360 données sont utilisées et non pas 378 à cause de la variable *offset*. En effet, comme 18 individus sont entrés dans l'étude après l'âge de 18 ans (age1st $>$ 18 ans), $t = 0$ pour chacune des 3 classes d'âge couvertes par t^* et donc $\ln(t)_{it^*}$ est indéterminé.

$$\beta_8 \text{dbd}_i + \beta_9 \text{dissoc}_i + \beta_{10} \text{impulse}_i + \beta_{11} \text{mood}_i + \beta_{12} \text{mental}_i + \beta_{13} \text{organic}_i + \beta_{14} \text{person}_i + \beta_{15} \text{subst}_i + \beta_{16} \text{psychot}_i + \beta_{17} \text{sexual}_i + \beta_{18} \text{sleep}_i + \beta_{19} \text{somato}_i + \beta_{20} \text{ld}_i + \beta_{21} \text{comm}_i + \ln(t)_{it'},$$

$$\ln(\text{off})_{it'} =$$
$$\beta_0 + \beta_1 \text{age1st}_i + \beta_2 \text{adhd}_i + \beta_3 \text{adjust}_i + \beta_4 \text{anxiety}_i + \beta_5 \text{conduct}_i + \beta_6 \text{dev}_i + \beta_7 \text{dbd}_i + \beta_8 \text{dissoc}_i + \beta_9 \text{impulse}_i + \beta_{10} \text{mood}_i + \beta_{11} \text{mental}_i + \beta_{12} \text{organic}_i + \beta_{13} \text{person}_i + \beta_{14} \text{subst}_i + \beta_{15} \text{psychot}_i + \beta_{16} \text{sexual}_i + \beta_{17} \text{sleep}_i + \beta_{18} \text{somato}_i + \beta_{19} \text{ld}_i + \beta_{20} \text{comm}_i + \ln(t)_{it'},$$

$$i = 1, \ldots, 360$$

et

$$t' = \begin{cases} \text{classe d'âge 12-14 ans} \\ \text{classe d'âge 14-16 ans} \\ \text{classe d'âge 16-18 ans} \\ \text{classe d'âge 18-20 ans.} \end{cases}$$

Pour chacun des modèles de chacune des analyses, la statistique PRESS, présentée à la section 2.7 (voir page 18), sera calculée et ces statistiques seront comparées entre elles afin de décider du meilleur modèle.

Chapitre 8

La première analyse avec des GLM longitudinaux ou non

8.1 La première analyse à l'aide de modèles linéaires généralisés longitudinaux

Tel qu'il a été mentionné au chapitre 7, la première analyse cherchera des modèles servant à prédire les nombres de crimes de type « Violence », « Drogues » et le nombre total de crimes de tous types. Ces variables, mesurées après 18 ans, seront prédites selon plusieurs variables mesurées, quant à elles, avant l'âge de 18 ans. Les modèles linéaires généralisés longitudinaux seront utilisés afin de faire ces prévisions.

Aux sous-sections 8.1.1, 8.1.2 et 8.1.3, le modèle suivant sera ajusté aux données :

$$\ln(Y)_{it} = \beta_0 + \beta_1 \text{age1st}_i + \beta_2 \text{viol1218}_i + \beta_3 \text{drug1218}_i + \beta_4 \text{sex1218}_i + \beta_5 \text{off1218}_i + \beta_6 \text{prop1218}_i + \beta_7 \text{autres1218}_i + \beta_8 \text{nviol1214}_i + \beta_9 \text{nviol1416}_i + \beta_{10} \text{nviol1618}_i + \beta_{11} \text{ndrug1416}_i + \beta_{12} \text{ndrug1618}_i + \beta_{13} \text{nsex1214}_i + \beta_{14} \text{nsex1416}_i + \beta_{15} \text{nsex1618}_i + \beta_{16} \text{noff1214}_i + \beta_{17} \text{noff1416}_i + \beta_{18} \text{noff1618}_i + \beta_{19} \text{nprop1214}_i + \beta_{20} \text{nprop1416}_i + \beta_{21} \text{nprop1618}_i + \beta_{22} \text{nautres1214}_i + \beta_{23} \text{nautres1416}_i + \beta_{24} \text{nautres1618}_i + \ln(t)_{it},$$

$$i = 1, \ldots, 378,$$

et

$$t = \begin{cases} \text{classe d'âge 18-20 ans} \\ \text{classe d'âge 20-22 ans} \\ \text{classe d'âge 22-24 ans.} \end{cases}$$

Le modèle ci-haut mentionné sera, dans un premier temps, ajusté en utilisant le nombre de crimes de type « Violence » comme variable endogène (sous-section 8.1.1). Ensuite, ce sera le nombre de crimes de type « Drogues » (sous-section 8.1.2) et, finalement, le nombre de crimes totaux (sous-section 8.1.3) qui seront les variables endogènes du modèle.

Dans ces modèles, la variable age1st correspond à l'âge au premier délit (variant entre 9 et 22 ans, environ), les variables viol1218, drug1218, sex1218, off1218, prop1218 et autres1218 sont des variables indicatrices de la présence d'au moins un crime des types suivants : violence, drogues, sexe, total, propriété et autres, respectivement. Les autres variables représentent les nombres de crimes des types violence, drogues, sexe, total, propriété et autres commis entre 12 et 14 ans, 14 et 16 ans, 16 et 18 ans. Une variable *offset*, t, est aussi présente dans le modèle et représente le nombre d'années où l'individu était susceptible de commettre des crimes dans chacune des 3 classes d'âge ci-haut mentionnées. Ainsi, pour une personne âgée de 19.4 ans au début de son suivi et âgée de 23.2 ans à la fin de celui-ci a :

$$\begin{aligned} t_{i,18-20} &= 20 - 19.4 = 0.6, \\ t_{i,20-22} &= 2, \\ t_{i,22-24} &= 23.2 - 22 = 1.2. \end{aligned}$$

8.1.1 Crimes de type « Violence »

Les variables retenues sont présentées au TABLEAU 8.1 en page 81. Ces variables ont été sélectionnées selon 6 seuils d'exclusion différents : 1%, 5%, 10%, 15%, 20% et 30%. De plus, 3 structures différentes pour les matrices de corrélation $\boldsymbol{R_i}(\boldsymbol{\alpha})$ (section 3.5, page 32) ont été testées : ce sont les structures auto-régressive (AR(1)), équicorrélation (EXCH) et non structuré (UNSTR).

Dans ce tableau, les modèles et leurs statistiques de validation croisée (PRESS) sont indiqués. Quatre statistiques PRESS y sont inscrites : une pour la tranche d'âge 18-20, une autre pour 20-22, l'autre pour 22-24, et finalement une pour le total des 3 classes d'âge.

Le seul modèle qui sera analysé plus en profondeur est celui ayant le plus petit PRESS (et ce, pour tout choix de la matrice de corrélation de travail), c'est-à-dire le modèle avec les variables nsex1618, nviol1618, ndrug1416, noff1618, nprop1416 et nprop1618.

Nous tenterons d'améliorer le modèle par une transformation d'une ou de plusieurs variables explicatives à l'aide d'un modèle additif généralisé. La FIGURE 8.1 suggère qu'une transformation des variables nsex1618 et nviol1618 n'est pas nécessaire[1]. Ainsi, plusieurs combinaisons de transformations seront essayées sur les trois autres variables et le meilleur modèle sera obtenu en prenant l'inverse du carré de la variable nprop1416 et en gardant les autres variables comme telles. Le PRESS correspondant à ce nouveau modèle est 918.284, et nous pouvons donc retenir ce modèle :

$$\begin{aligned}\ln(\text{viol})_{it} = &-1.6078 - 1.1861\text{nprop1416}_i^{-2} + 1.2616\text{nprop1416}_i^{-1} - 0.0631\text{nprop1618}_i + \\ &0.1510\text{nviol1618}_i - 0.2505\text{ndrug1416}_i - 0.5574\text{nsex1618}_i + \\ &0.0966\text{noff1618}_i + \ln(t)_{it},\end{aligned} \tag{8.1}$$

où $\boldsymbol{R_i}(\boldsymbol{\alpha})$ est de structure d'équicorrélation, et où la corrélation entre les temps est représentée par la matrice suivante :

$$\boldsymbol{R_i}(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & 0.1225 & 0.1255 \\ 0.1225 & 1 & 0.1225 \\ 0.1225 & 0.1225 & 1 \end{pmatrix}, \quad i = 1, \ldots, 378.$$

[1]De cette figure, on voit qu'une transformation de ndrug1416 n'a pas été essayée. La raison est que le logiciel s'attend à ce que les variables à transformer prennent 4 modalités ou plus, ce qui n'est pas le cas avec cette variable. Il faut donc faire attention à la troisième figure de la première ligne, car la courbe qui y est représentée n'est pas interprétable.

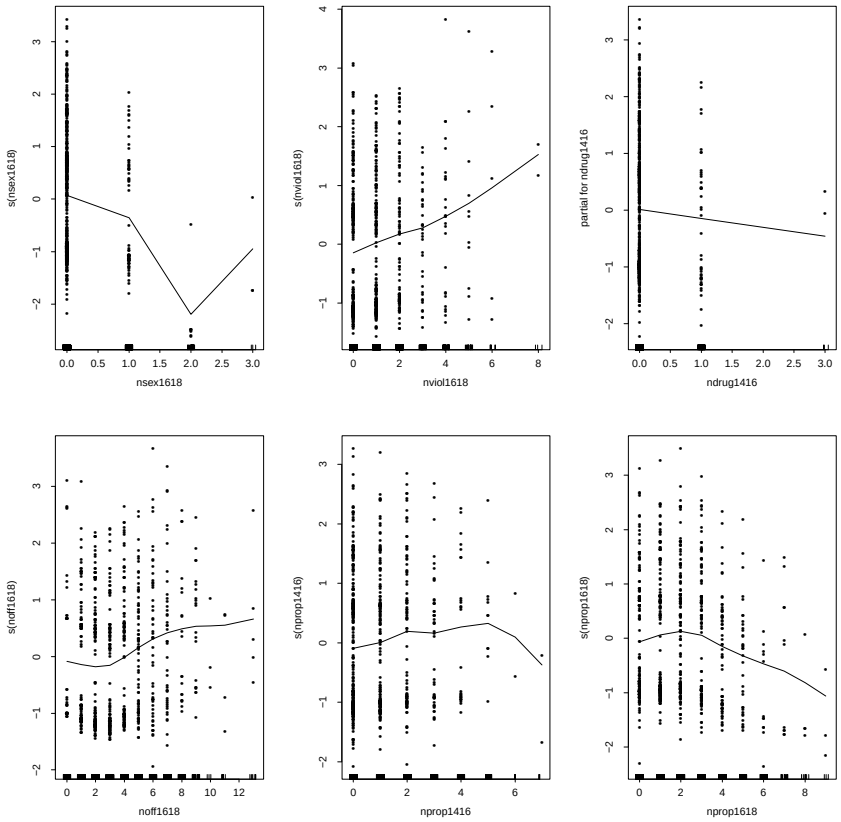

FIG. 8.1 – *Modèle de départ (« Violence », tous âges confondus). Les transformations ont été obtenues par GAM.*

TYPE=UNSTR			
Seuil	Variables retenues	PRESS	
1%	nsex1618 nviol1618	322.043 311.457 300.163	933.662
5% et 10%	nsex1618 nviol1618 noff1618	323.036 305.557 298.154	926.747
15%	nsex1618 nviol1618 noff1618 nprop1416 nprop1618	315.763 310.794 301.143	927.701
20%	ndrug1416 nsex1416 nsex1618 nviol1618 noff1618 nprop1416 nprop1618	313.760 310.370 300.727	924.857
30%	ndrug1416 nsex1416 nsex1618 nviol1214 nviol1618 noff1618 nprop1416 nprop1618	319.086 316.065 303.252	938.403

TYPE=CS			
Seuil	Variables retenues	PRESS	
1% 5% 10%	nsex1618 nviol1618 noff1618	323.037 305.543 298.099	926.680
15% et 20%	nsex1618 nviol1618 ndrug1416 noff1618 nprop1416 nprop1618	314.283 310.280 300.037	**924.600**
30%	nsex1416 nsex1618 nviol1214 nviol1618 ndrug1416 noff1618 nprop1416 nprop1618	319.236 316.076 303.155	938.466

TYPE=AR(1)			
Seuil	Variables retenues	PRESS	
1%	nsex1618 nviol1618	322.053 311.496 299.999	933.548
5% et 10%	nsex1618 nviol1618 noff1618	323.071 305.644 298.192	926.907
15%	nsex1618 nviol1618 noff1618 nprop1416 nprop1618	315.677 310.771 301.108	927.556
20%	ndrug1416 nsex1416 nsex1618 nviol1618 noff1618 nprop1416 nprop1618	313.619 310.358 300.729	924.706
30%	ndrug1416 nsex1416 nsex1618 nviol1214 nviol1618 noff1618 nprop1416 nprop1618	318.743 315.892 303.157	937.792

TAB. 8.1 – *Résultats obtenus pour l'analyse du nombre de crimes « Violence ».*

8.1.2 Crimes de type « Drogues »

Les variables retenues sont présentées au TABLEAU 8.2 en page 84. Le modèle qui sera analysé est celui avec les variables drug1218, sex1218, off1218, prop1218, autres1218, ndrug1416, nviol1416, nviol1618, noff1214, noff1416, noff1618, nprop1214 et nprop1618. Un modèle additif généralisé suggère les transformations présentées à la FIGURE 8.2[2].

De toutes les combinaisons de transformations essayées, seul le logarithme de la variable nviol1618 a pu faire diminuer le PRESS (392.834 comparativement à 393.022). Cependant, comme les deux statistiques PRESS sont pratiquement identiques et qu'une transformation de variable complique l'interprétation du modèle, le modèle de base sera conservé, c'est-à-dire

$$\begin{aligned}\ln(\text{drug})_{it} = &-2.7138 + 0.2953\text{drug1218}_i - 0.4837\text{sex1218}_i + 0.5368\text{off1218}_i - \\ &0.4957\text{prop1218}_i + 0.2175\text{autres1218}_i + 0.2087\text{ndrug1416}_i - \\ &0.2718\text{nviol1416}_i + 0.1018\text{nviol1618}_i - 0.3602\text{noff1214}_i + \\ &0.0851\text{noff1416}_i + 0.1108\text{noff1618}_i + 0.3813\text{nprop1214}_i - \\ &0.0637\text{nprop1618}_i + \ln(t)_{it},\end{aligned} \tag{8.2}$$

où $\boldsymbol{R_i(\alpha)}$ est non-structurée, et où nous pouvons représenter la corrélation entre les temps de la façon suivante :

$$\boldsymbol{R}_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & 0.1160 & 0.0844 \\ 0.1160 & 1 & 0.0366 \\ 0.0844 & 0.0366 & 1 \end{pmatrix}, \quad i = 1, \ldots, 378.$$

[2]Encore une fois, il ne faut pas interpréter les courbes sur les six premiers graphiques, car, la plupart des ces graphiques impliquent des variables dichotomiques, donc des variables ne pouvant être transformées.

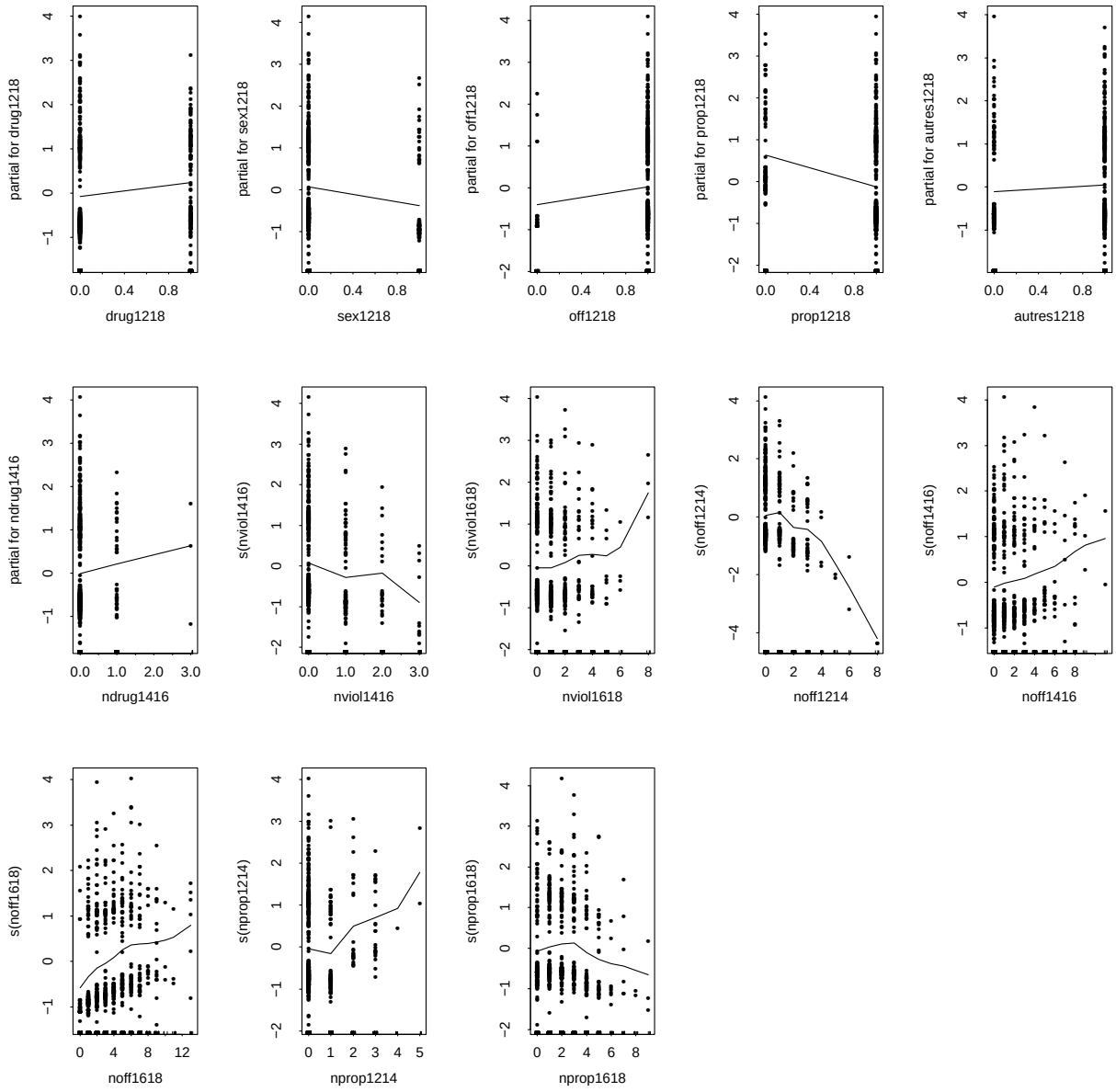

FIG. 8.2 – *Modèle de départ (« Drogues », tous âges confondus). Les transformations ont été obtenues par GAM.*

TYPE=UNSTR			
Seuil	Variables retenues	PRESS	
1%	noff1618	151.187 140.834 108.363	400.385
5% et 10%	drug1218 noff1618	150.534 140.517 107.622	398.673
15%	drug1218 sex1218 prop1218 nviol1416 nviol1618 noff1214 noff1416 noff1618 nprop1214	149.585 139.346 106.484	395.415
20%	drug1218 sex1218 prop1218 autres1218 ndrug1416 nviol1416 nviol1618 noff1214 noff1416 noff1618 nprop1214	150.818 135.263 107.545	393.627
30%	drug1218 sex1218 off1218 prop1218 autres1218 ndrug1416 nviol1416 nviol1618 noff1214 noff1416 noff1618 nprop1214 nprop1618	151.093 134.289 107.641	**393.022**

TYPE=CS			
Seuil	Variables retenues	PRESS	
1%	noff1618	151.205 140.832 108.363	400.401
5% et 10%	drug1218 noff1618	150.556 140.519 107.617	398.691
15%	drug1218 prop1218 autres1218 ndrug1416 nviol1416 nviol1618 noff1214 noff1618 nprop1214 nautres1416	151.009 133.958 108.246	393.213
20%	drug1218 sex1218 prop1218 autres1218 ndrug1416 nviol1416 nviol1618 noff1214 noff1618 nprop1214 nautres1416	150.792 135.178 107.377	393.346
30%	drug1218 sex1218 off1218 prop1218 autres1218 ndrug1416 nviol1416 nviol1618 noff1214 noff1618 nprop1214 nprop1618 nautres1416	150.911 135.162 107.479	393.552

TYPE=AR(1)			
Seuil	Variables retenues	PRESS	
1%	noff1618	151.193 140.849 108.377	400.419
5% et 10%	drug1218 noff1618	150.558 140.521 107.633	398.711
15% et 20%	drug1218 sex1218 prop1218 autres1218 ndrug1416 nviol1416 nviol1618 noff1214 noff1618 nprop1214 nautres1416	150.748 135.319 107.353	393.420
30%	drug1218 sex1218 off1218 prop1218 autres1218 ndrug1416 nviol1416 nviol1618 noff1214 noff1618 nprop1214 nprop1618 nautres1416	150.869 135.264 107.461	393.595

TAB. 8.2 – *Résultats obtenus pour l'analyse du nombre de crimes « Drogues ».*

8.1.3 Nombre total de crimes

On voit au TABLEAU 8.3, en page 87, que le modèle regroupant les variables viol1218, autres1218, ndrug1618, nsex1416, nsex1618, noff1618 et nprop1416 sera analysé. Si un modèle additif généralisé est ajusté à ces données, les graphiques de la FIGURE 8.3[3] sont obtenus.

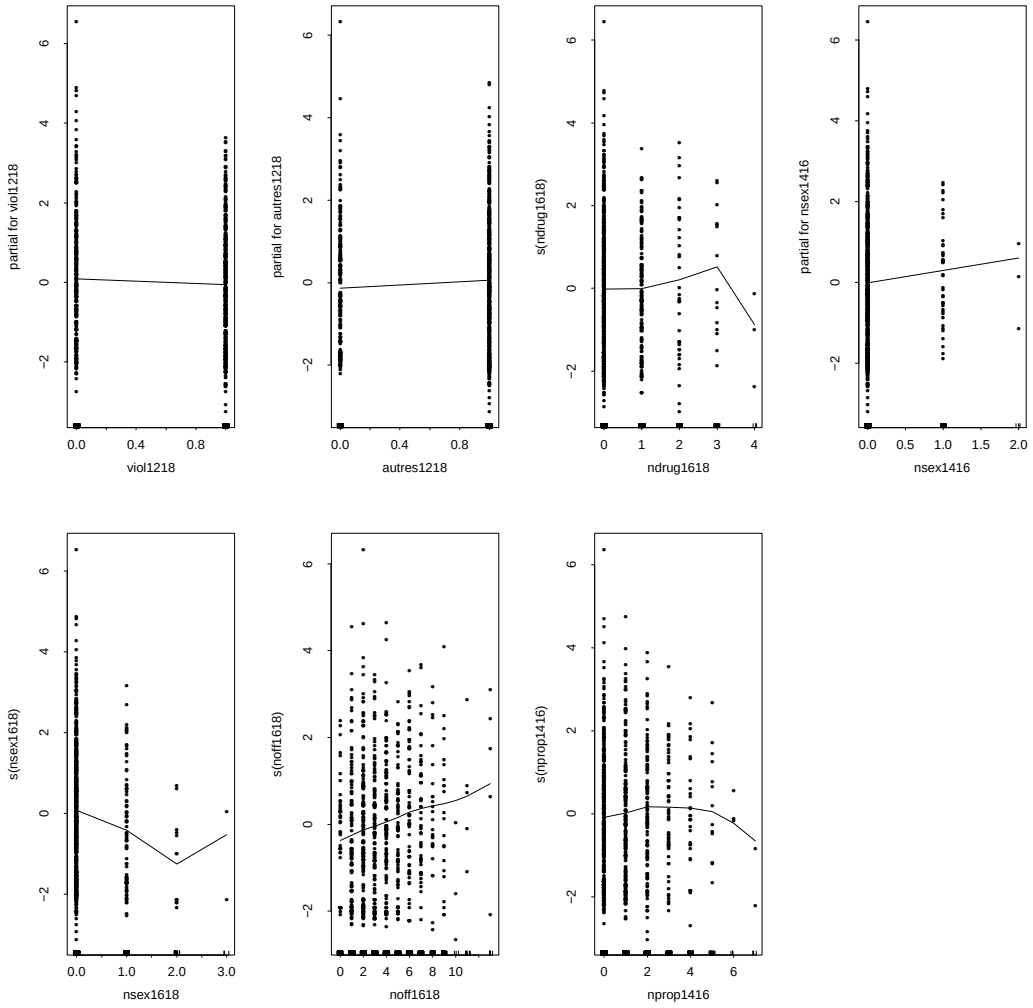

FIG. 8.3 – *Modèle de départ (tous âges confondus). Les transformations ont été obtenues par GAM.*

[3]Attention aux 2 premiers graphiques et au dernier graphique de la première ligne, car ils font tous intervenir une variable ne prenant pas assez de modalités pour que l'on puisse envisager de les transformer.

Afin d'améliorer le modèle de départ, des transformations seront essayées sur la variable nprop1416 uniquement. Le modèle faisant intervenir le carré de la variable nprop1416 est celui améliorant le plus le PRESS de départ (valeur de 5282.39, comparativement à une valeur de 5296.81 avec le modèle de départ). Alors, nous pouvons dire qu'un modèle acceptable est :

$$\begin{aligned}\ln(\text{off})_{it} = &-0.5138 - 0.1709\text{viol1218}_i + 0.2055\text{autres1218}_i + 0.0919\text{ndrug1618}_i + \\ &0.2902\text{nsex1416}_i - 0.4962\text{nsex1618}_i + 0.1019\text{noff1618}_i - \\ &0.0320\text{nprop1416}_i^2 + 0.1847\text{nprop1416}_i + \ln(t)_{it},\end{aligned} \tag{8.3}$$

où $\boldsymbol{R_i(\alpha)}$ est de type auto-régressif, et où :

$$\boldsymbol{R_i}(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & 0.3029 & 0.0917 \\ 0.3029 & 1 & 0.3029 \\ 0.0917 & 0.3029 & 1 \end{pmatrix}, \quad i = 1, \ldots, 378.$$

TYPE=UNSTR			
Seuil	Variables retenues	PRESS	
1%	nsex1618 noff1618	2455.53 1715.60 1208.70	5379.83
5%	nsex1416 nsex1618 noff1618 nprop1416	**2431.64** 1697.97 1217.73	5347.34
10%	viol1218 autres1218 nsex1416 nsex1618 noff1618 nprop1416	2432.39 1672.33 1221.82	5326.53
15%	viol1218 autres1218 ndrug1618 nsex1416 nsex1618 noff1618 nprop1416	2440.14 1650.37 1207.69	5298.20
20%	drug1218 viol1218 autres1218 ndrug1618 nsex1416 nsex1618 nviol1214 noff1214 noff1618 nprop1214 nprop1416	2463.91 1691.46 1225.52	5380.88
30%	drug1218 viol1218 prop1218 autres1218 ndrug1618 nsex1416 nsex1618 nviol1214 noff1214 noff1618 nprop1214 nprop1416 nprop1618	2470.86 1702.72 1225.54	5399.12
TYPE=CS			
Seuil	Variables retenues	PRESS	
1%	nsex1618 noff1618	2471.16 1711.64 1195.74	5378.54
5%	autres1218 nsex1416 nsex1618 noff1618	2457.14 1671.16 1203.37	5331.67
10%	viol1218 autres1218 nsex1416 nsex1618 noff1618 nprop1416	2447.49 1669.74 1208.60	5325.83
15%	viol1218 autres1218 ndrug1618 nsex1416 nsex1618 noff1618 nprop1416	2454.41 1647.59 1194.98	5296.99
20%	drug1218 viol1218 autres1218 ndrug1618 nsex1416 nsex1618 nviol1214 noff1214 noff1618 nprop1214 nprop1416	2474.11 1687.27 1212.34	5373.71
30%	drug1218 viol1218 prop1218 autres1218 ndrug1416 ndrug1618 nsex1416 nsex1618 nviol1214 noff1214 noff1618 nprop1214 nprop1416	2493.13 1659.71 1213.45	5366.29
TYPE=AR(1)			
Seuil	Variables retenues	PRESS	
1%	nsex1618 noff1618	2461.59 1713.58 1203.77	5378.94
5%	nsex1416 nsex1618 noff1618 nprop1416	2438.55 1695.64 1211.96	5346.15
10%	viol1218 autres1218 nsex1416 nsex1618 noff1618 nprop1416	2439.19 1670.03 1215.89	5325.11
15%	viol1218 autres1218 ndrug1618 nsex1416 nsex1618 noff1618 nprop1416	2446.24 1648.20 1202.36	**5296.81**
20%	drug1218 viol1218 autres1218 ndrug1618 nsex1416 nsex1618 nviol1214 noff1214 noff1618 nprop1214 nprop1416	2469.60 1689.38 1220.29	5379.27
30%	drug1218 viol1218 prop1218 autres1218 ndrug1618 nsex1416 nsex1618 nviol1214 noff1214 noff1618 nprop1214 nprop1416 nprop1618	2476.21 1700.63 1220.77	5397.61

TAB. 8.3 – *Résultats obtenus pour l'analyse du nombre total de crimes.*

8.2 La deuxième analyse à l'aide de modèles linéaires généralisés

Aux sous-sections 8.2.1, 8.2.2 et 8.2.3, les modèles suivants sont ajustés aux données :

$$\ln(Y)_{i,18-20} = \beta_0 + \beta_1 \text{age1st}_i + \beta_2 \text{viol1218}_i + \beta_3 \text{drug1218}_i + \beta_4 \text{sex1218}_i + \beta_5 \text{off1218}_i + \beta_6 \text{prop1218}_i + \beta_7 \text{autres1218}_i + \beta_8 \text{nviol1214}_i + \beta_9 \text{nviol1416}_i + \beta_{10} \text{nviol1618}_i + \beta_{11} \text{ndrug1416}_i + \beta_{12} \text{ndrug1618}_i + \beta_{13} \text{nsex1214}_i + \beta_{14} \text{nsex1416}_i + \beta_{15} \text{nsex1618}_i + \beta_{16} \text{noff1214}_i + \beta_{17} \text{noff1416}_i + \beta_{18} \text{noff1618}_i + \beta_{19} \text{nprop1214}_i + \beta_{20} \text{nprop1416}_i + \beta_{21} \text{nprop1618}_i + \beta_{22} \text{nautres1214}_i + \beta_{23} \text{nautres1416}_i + \beta_{24} \text{nautres1618}_i + \ln(t)_{i,18-20}$$

$$\ln(Y)_{i,20-22} = \beta_0 + \beta_1 \text{age1st}_i + \beta_2 \text{viol1218}_i + \beta_3 \text{drug1218}_i + \beta_4 \text{sex1218}_i + \beta_5 \text{off1218}_i + \beta_6 \text{prop1218}_i + \beta_7 \text{autres1218}_i + \beta_8 \text{nviol1214}_i + \beta_9 \text{nviol1416}_i + \beta_{10} \text{nviol1618}_i + \beta_{11} \text{ndrug1416}_i + \beta_{12} \text{ndrug1618}_i + \beta_{13} \text{nsex1214}_i + \beta_{14} \text{nsex1416}_i + \beta_{15} \text{nsex1618}_i + \beta_{16} \text{noff1214}_i + \beta_{17} \text{noff1416}_i + \beta_{18} \text{noff1618}_i + \beta_{19} \text{nprop1214}_i + \beta_{20} \text{nprop1416}_i + \beta_{21} \text{nprop1618}_i + \beta_{22} \text{nautres1214}_i + \beta_{23} \text{nautres1416}_i + \beta_{24} \text{nautres1618}_i + \ln(t)_{i,20-22}$$

$$\ln(Y)_{i,22-24} = \beta_0 + \beta_1 \text{age1st}_i + \beta_2 \text{viol1218}_i + \beta_3 \text{drug1218}_i + \beta_4 \text{sex1218}_i + \beta_5 \text{off1218}_i + \beta_6 \text{prop1218}_i + \beta_7 \text{autres1218}_i + \beta_8 \text{nviol1214}_i + \beta_9 \text{nviol1416}_i + \beta_{10} \text{nviol1618}_i + \beta_{11} \text{ndrug1416}_i + \beta_{12} \text{ndrug1618}_i + \beta_{13} \text{nsex1214}_i + \beta_{14} \text{nsex1416}_i + \beta_{15} \text{nsex1618}_i + \beta_{16} \text{noff1214}_i + \beta_{17} \text{noff1416}_i + \beta_{18} \text{noff1618}_i + \beta_{19} \text{nprop1214}_i + \beta_{20} \text{nprop1416}_i + \beta_{21} \text{nprop1618}_i + \beta_{22} \text{nautres1214}_i + \beta_{23} \text{nautres1416}_i + \beta_{24} \text{nautres1618}_i + \ln(t)_{i,22-24}$$

Les modèles ci-haut mentionnés seront, dans un premier temps, ajustés en utilisant le nombre de crimes de type « Violence » comme variable endogène (sous-section 8.2.1). Ensuite, ce sera le nombre de crimes de type « Drogues » (sous-section 8.2.2) et, finalement, le nombre de crimes totaux (sous-section 8.2.3) qui seront les variables endogènes du modèle. Dans ces modèles, la variable age1st correspond à l'âge au premier délit (variant entre 9 et 22 ans, environ), les variables viol1218, drug1218, sex1218, off1218, prop1218 et autres1218 sont des variables indicatrices de la présence d'au moins un crime des types suivants : violence, drogues, sexe, total, propriété et autres, respectivement. Les autres variables représentent les nombres de crimes des types violence, drogues, sexe, total, propriété et autres commis entre 12 et 14 ans, 14 et 16 ans, 16 et 18 ans. Une variable *offset*, t, est aussi présente dans le modèle et représente le temps exact, en années, que l'individu a passé dans chacune des 3 classes d'âge. Un modèle linéaire généralisé sera donc utilisé. L'approche longitudinale ne sera plus utilisée ici, car le but de cette nouvelle analyse est de trouver des modèles pour chacun des âges de façon séparée.

8.2.1 Crimes de type « Violence »

Les variables retenues pour chacune des classes d'âge et les résultats des validations croisées sont présentées au TABLEAU 8.7 de la page 94.

Entre 18 et 20 ans

Le modèle étudié plus en profondeur est celui regroupant les variables nsex1416, nsex1618, nviol1618, nprop1416, nprop1618 et nautres1618. Si un modèle additif généralisé est ajusté, les graphiques de la FIGURE 8.4[4] sont produits.

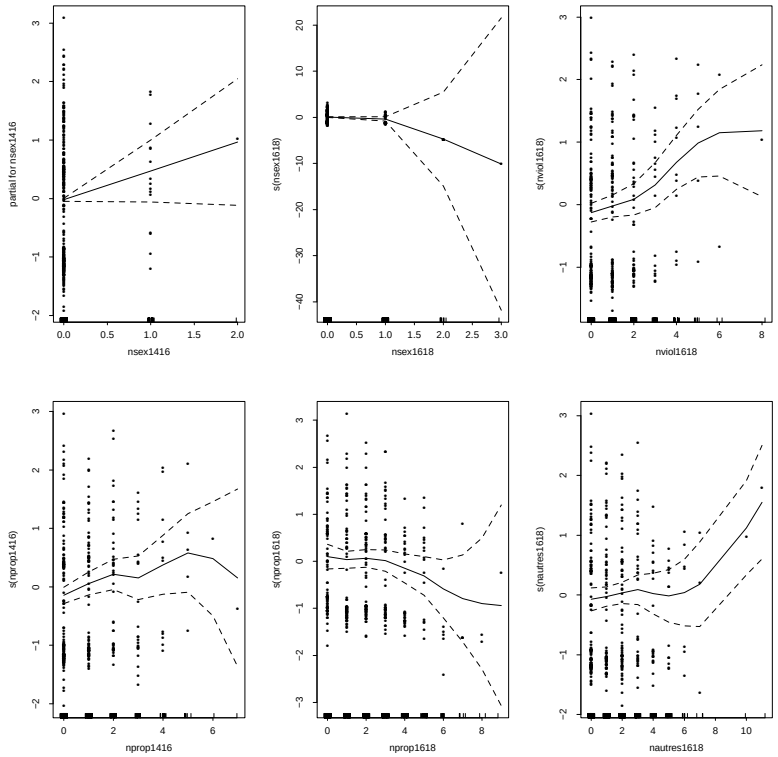

FIG. 8.4 – *Modèle de départ (« Violence », 18-20 ans). Les transformations ont été obtenues par GAM.*

Le TABLEAU 8.4 en page suivante montre les différents seuils associés aux transformations des variables. On y voit que toutes les transformations de variables doivent

[4]Sur les FIGURES 8.4 à 8.13, il faut faire attention de ne pas interpréter les courbes sur les graphiques faisant intervenir une variable ayant 4 modalités ou moins.

être éliminées. Ainsi, nous nous devons de conserver le modèle de départ, c'est-à-dire

$$\begin{aligned}\ln(\text{viol})_{i,18-20} = &-1.2962 + 0.1890\text{nviol1618}_i + 0.5356\text{nsex1416}_i - 0.6165\text{nsex1618}_i + \\ &0.1207\text{nprop1416}_i - 0.1045\text{nprop1618}_i + 0.0798\text{nautres1618}_i + \ln(t)_{i,18-20}.\end{aligned} \tag{8.4}$$

Variable	Seuils				
	Étape 1	Étape 2	Étape 3	Étape 4	Étape 5
nsex1416	Cette variable ne peut être transformée				
nviol1618	**0.4587030**				
nprop1618	0.3876451	**0.4326259**			
nprop1416	0.3113467	0.3691057	**0.3892150**		
nsex1618	0.2774356	0.2709530	0.2741277	**0.2725265**	
nautres1618	0.0779771	0.1405405	0.1418480	0.1383772	**0.1546976**

TAB. 8.4 – *Seuils associés aux transformations des variables (« Violence », 18-20 ans).*

Entre 20 et 22 ans

Le modèle regroupant les dix variables drug1218, ndrug1416, nsex1214, nsex1618, nviol1416, nviol1618, noff1416, noff1618, nautres1214 et nautres1416 sera analysé. Ici, un modèle additif généralisé donne les transformations à la FIGURE 8.5.

Le TABLEAU 8.5 présente les seuils associés aux transformations. Encore une fois, toutes les transformations doivent être éliminées et le modèle suivant sera donc accepté :

$$\begin{aligned}\ln(\text{viol})_{i,20-22} = &-1.6961 + 0.3124\text{drug1218}_i - 0.3555\text{ndrug1416}_i - 17.5070\text{nsex1214}_i - \\ &0.5112\text{nsex1618}_i - 0.2342\text{nviol1416}_i + 0.0851\text{nviol1618}_i + \\ &0.1226\text{noff1416}_i + 0.0909\text{noff1618}_i + 0.1588\text{nautres1214}_i - \\ &0.1319\text{nautres1416}_i + \ln(t)_{i,20-22}.\end{aligned} \tag{8.5}$$

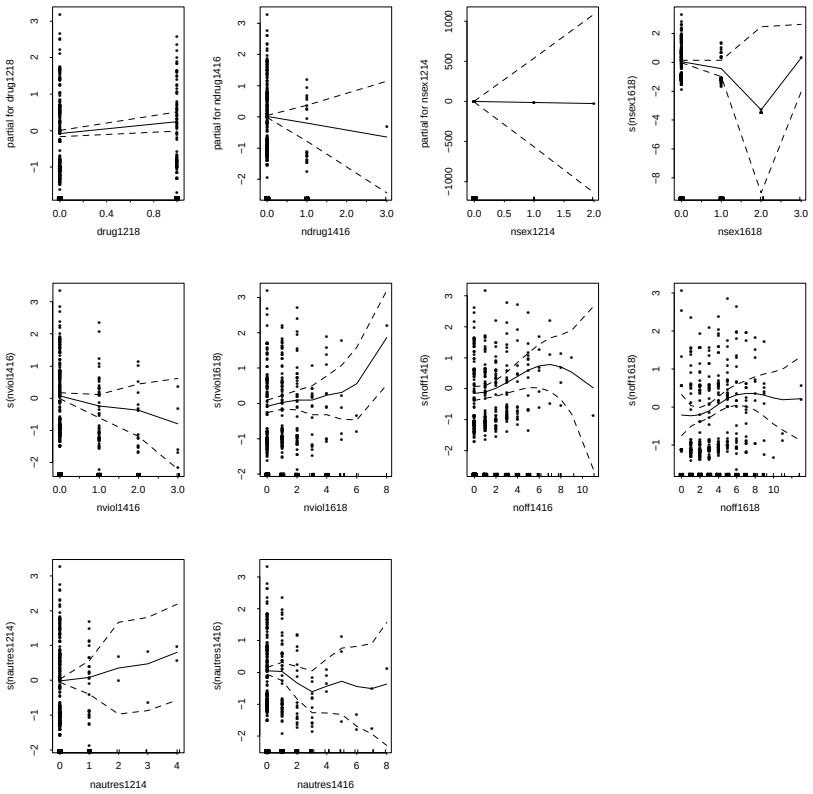

FIG. 8.5 – *Modèle de départ (« Violence », 20-22 ans). Les transformations ont été obtenues par GAM.*

Variable	Seuils						
	Étape 1	Étape 2	Étape 3	Étape 4	Étape 5	Étape 6	Étape 7
drug1218			Cette variable ne peut être transformée				
ndrug1416			Cette variable ne peut être transformée				
nsex1214			Cette variable ne peut être transformée				
nautres1214	**0.9820505**						
nviol1416	0.9188884	**0.9291411**					
noff1416	0.4493012	0.4656156	**0.4627334**				
nsex1618	0.1540934	0.1549819	0.1537707	**0.1572428**			
nautres1416	0.1569353	0.1448168	0.1399181	0.1480815	**0.1459482**		
noff1618	0.2113153	0.2337744	0.2163295	0.1425434	0.1347333	**0.1378600**	
nviol1618	0.1376726	0.1430865	0.1422842	0.1073012	0.1051444	0.1302013	**0.4446904**

TAB. 8.5 – *Seuils associés aux transformations des variables (« Violence », 20-22 ans).*

Entre 22 et 24 ans

Le modèle ayant le plus petit PRESS est celui faisant intervenir les variables nsex1618 et noff1618. Si un modèle additif généralisé est ajusté, les transformations sont données à la FIGURE 8.6.

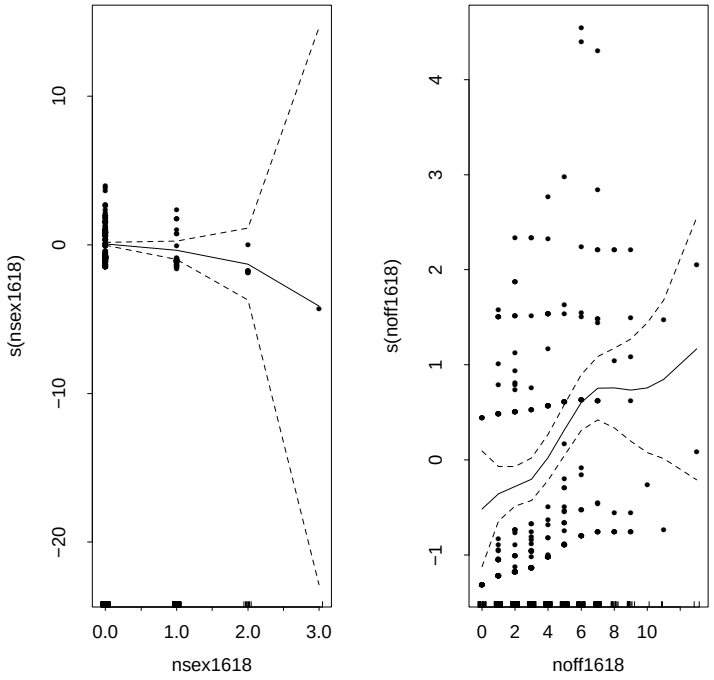

FIG. 8.6 – *Modèle de départ (« Violence », 22-24 ans). Les transformations ont été obtenues par GAM.*

Les seuils associés aux transformations sont présentés au TABLEAU 8.6 de la page suivante. La variable noff1618 sera transformée, et la transformation racine carrée fait diminuer le PRESS jusqu'à une valeur de 312.125, comparativement à une valeur de 312.524 avec le modèle de départ. Ainsi, nous acceptons le modèle suivant :

$$\ln(\text{viol})_{i,22-24} = -2.3293 - 0.5287\text{nsex1618}_i + 0.5781\sqrt{\text{noff1618}_i} + \ln(t)_{i,22-24}. \quad (8.6)$$

Variable	Seuils	
	Étape 1	Étape 2
nsex1618	**0.7776079**	
noff1618	0.0393767	0.03841086

TAB. 8.6 – *Seuils associés aux transformations des variables (« Violence », 22-24 ans).*

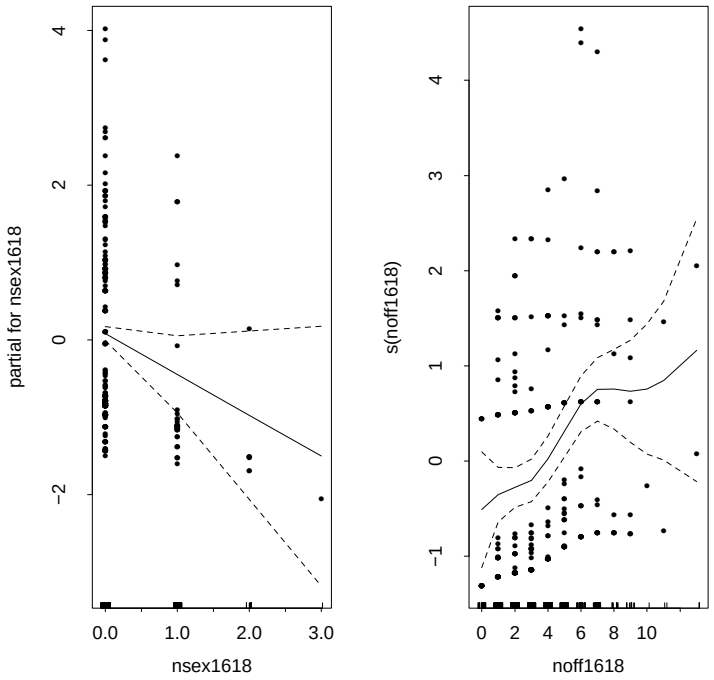

FIG. 8.7 – *Modèle final (« Violence », 22-24 ans). Les transformations ont été obtenues par GAM.*

Entre 18 et 20 ans		
Seuil	Variables retenues	PRESS
1%	nsex1618 nviol1618	324.0088
5%	nsex1618 nviol1618 nprop1416	325.753
10% 15%	nsex1416 nsex1618 nviol1618 nprop1416 nprop1618 nautres1618	**319.178**
20% 30%	drug1218 ndrug1618 nsex1416 nsex1618 nviol1214 nviol1618 nprop1416 nprop1618 nautres1618	327.436

Entre 20 et 22 ans		
Seuil	Variables retenues	PRESS
1%	noff1618	313.395
5%	nsex1618 noff1618	310.269
10% 15%	drug1218 nsex1214 nsex1618 noff1618	309.361
20%	drug1218 nsex1214 nsex1618 nviol1416 nviol1618 noff1416 noff1618	311.558
30%	drug1218 ndrug1416 nsex1214 nsex1618 nviol1416 nviol1618 noff1416 noff1618 nautres1214 nautres1416	**308.958**

Entre 22 et 24 ans		
Seuil	Variables retenues	PRESS
1%	noff1618	314.624
5%	nsex1618 noff1618	**312.524**
10% 15%	drug1218 nsex1214 nsex1618 noff1618	315.673
20%	drug1218 nsex1214 nsex1618 nviol1416 nviol1618 noff1416 noff1618	324.203
30%	drug1218 ndrug1416 nsex1214 nsex1618 nviol1416 nviol1618 noff1416 noff1618 nautres1214 nautres1416	319.701

TAB. 8.7 – *Résultats obtenus pour l'analyse du nombre de crimes « Violence » selon les âges.*

8.2.2 Crimes de type « Drogues »

Le TABLEAU 8.8 en page 98 présente les variables retenues et les résultats des validations croisées pour chacune des tranches d'âge.

Entre 18 et 20 ans

Pour cette première tranche d'âge, le modèle qui sera analysé fait intervenir uniquement la variable nautres1618. La nécessité d'inclure une transformation de variable est évidente à la FIGURE 8.8.

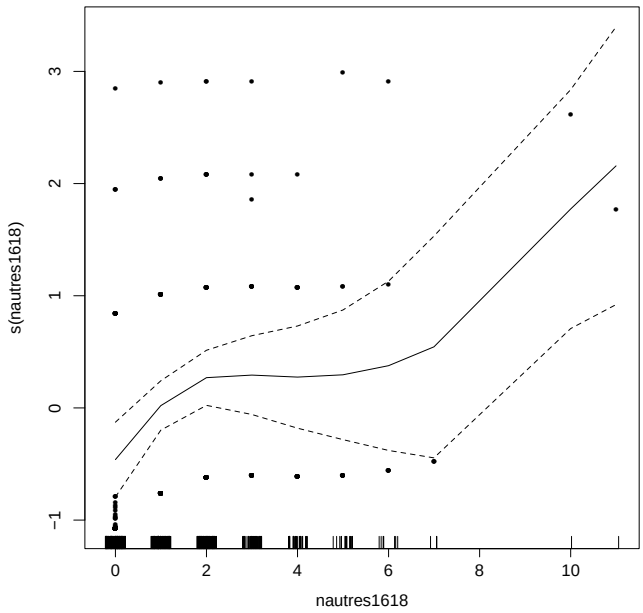

FIG. 8.8 – *Modèle de départ (« Drogues », 18-20 ans). La transformation a été obtenue par GAM.*

La variable impliquée dans le modèle ne peut être transformée. En fait, bien que la transformation ait un seuil relativement faible (0.0685), aucune des transformations essayées n'a pu diminuer le PRESS. Nous conservons donc ce modèle :

$$\ln(\text{drug})_{i,18-20} = -2.1864 + 0.1777\text{nautres1618}_i + \ln(t)_{i,18-20}. \tag{8.7}$$

Comme les intervalles de confiance sur la FIGURE 8.8 sont assez larges, une droite peut facilement passer entre celles-ci, ce qui signifie qu'il peut être raisonnable, malgré le seuil de la transformation, de garder la variable nautres1618 comme telle.

Entre 20 et 22 ans

Le meilleur modèle ici est celui ne faisant intervenir aucune variable, c'est-à-dire que nous acceptons que le nombre de crimes de type « Drogues » commis entre 20 et 22 ans soit fonction d'une constante et du temps où l'individu était à risque de commettre des crimes entre 20 et 22 ans :

$$\begin{aligned} \ln(\text{drug})_{i,20-22} &= -2.0711 + \ln(t)_{i,20-22} \\ \Rightarrow \text{drug}_{i,20-22} &= 0.1260 t_{i,20-22}. \end{aligned} \tag{8.8}$$

Donc, nous attendons d'un individu ayant été à risque pendant 2 ans entre 20 et 22 ans de ne commettre presqu'aucun crime entre les âges 20 et 22 ans.

Entre 22 et 24 ans

Ici, les variables impliquées dans le modèle ayant le meilleur pouvoir prédictif sont les variables sex1218, prop1218 et noff1618. Le seuil de la seule variable pouvant être transformée est de 0.08592228, et la racine carrée du nombre de crimes entre 16 et 18 ans améliore un peu le PRESS (valeur de 106.319 comparativement à 106.622 avec le modèle de départ). Ce modèle sera donc conservé :

$$\begin{aligned} \ln(\text{drug})_{i,22-24} = 0.1415 - 0.2046\text{sex1218}_i - 0.2007\text{prop1218}_i + \\ 0.1824\sqrt{\text{noff1618}_i} + \ln(t)_{i,22-24}. \end{aligned} \tag{8.9}$$

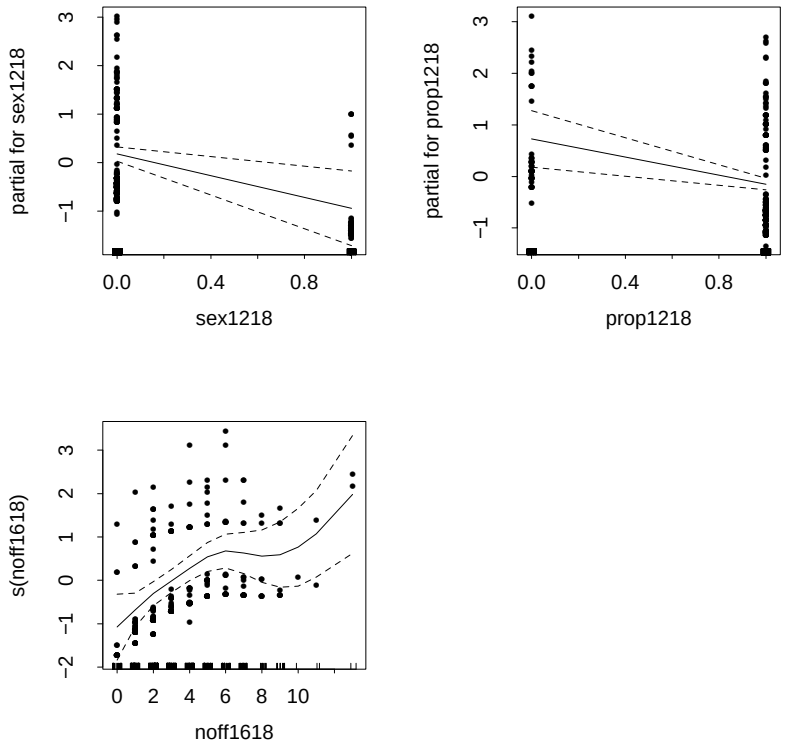

FIG. 8.9 – *Modèle de départ (« Drogues », 22-24 ans). Les transformations ont été obtenues par GAM.*

Entre 18 et 20 ans		
Seuil	Variables retenues	PRESS
1%	nautres1618	**150.663**
5% 10%	off1218 ndrug1618 nautres1618	154.386
15%	sex1218 off1218 prop1218 ndrug1618 nautres1618	155.004
20%	sex1218 off1218 prop1218 ndrug1618 nsex1214 nsex1416 nviol1618 noff1214 nprop1214 nprop1416 nautres1618	158.936
30%	sex1218 off1218 prop1218 autres1218 ndrug1618 nsex1214 nsex1416 nviol1618 noff1214 nprop1214 nprop1416 nautres1618	159.305

Entre 20 et 22 ans		
Seuil	Variables retenues	PRESS
1%	Aucune variable n'est significative au seuil de 1%	**141.473**
5%	nviol1416 noff1214 noff1416 nprop1214 nprop1416 nautres1214	149.381
10%,15% 20%	ndrug1416 nviol1416 nviol1618 noff1214 noff1416 nprop1214 nprop1416 nautres1214	249.795
30%	sex1218 ndrug1416 nsex1618 nviol1416 nviol1618 noff1214 noff1416 nprop1214 nprop1416 nautres1214	264.121

Entre 22 et 24 ans		
Seuil	Variables retenues	PRESS
1%	noff1618	108.767
5%	sex1218 noff1618	107.194
10% 15%	sex1218 prop1218 noff1618	**106.622**
20%	sex1218 prop1218 nviol1214 noff1618 nautres1214 nautres1618	109.679
30%	sex1218 prop1218 nviol1214 noff1618 nprop1416 nprop1618 nautres1214 nautres1618	109.144

TAB. 8.8 – *Résultats obtenus pour l'analyse du nombre de crimes « Drogues » selon les âges.*

8.2.3 Nombre total de crimes

Les variables retenues selon chaque tranche d'âge et selon chaque seuil d'exclusion sont présentées au TABLEAU 8.12 en page 104.

Entre 18 et 20 ans

Le modèle ayant un PRESS de 2294.76, c'est-à-dire celui faisant intervenir les variables viol1218, nsex1416, nsex1618, noff1618 et nprop1416 sera analysé. Un modèle additif généralisé suggère les transformations illustrées à la FIGURE 8.10.

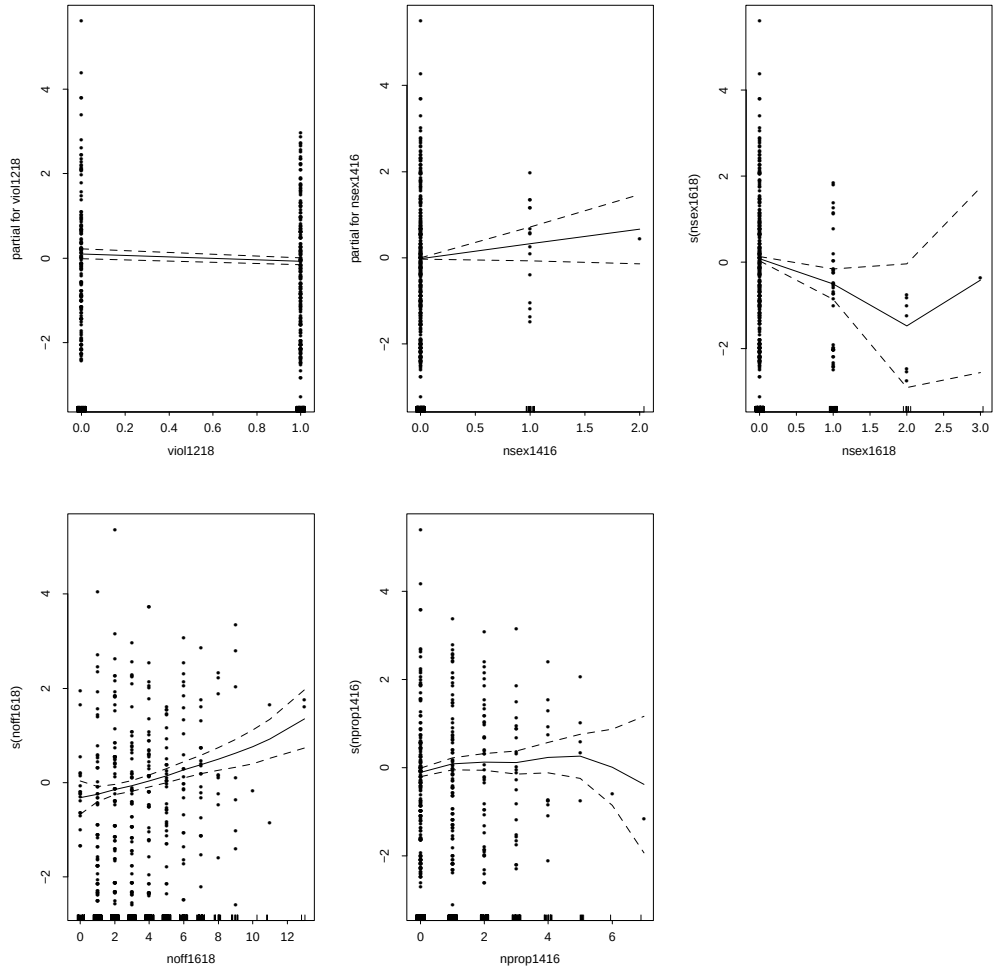

FIG. 8.10 – *Modèle de départ (18-20 ans). Les transformations ont été obtenues par GAM.*

Le TABLEAU 8.9 montre les seuils associés à chacune des transformations. Il suit donc que nous tenterons de transformer la variable nprop1416, et la transformation racine carrée s'avère être celle améliorant le plus le modèle de départ. Nous acceptons donc le modèle suivant :

$$\ln(\text{off})_{i,18-20} = -0.0815 - 0.1943\text{viol1218}_i + 0.3257\text{nsex1416}_i - 0.5780\text{nsex1618}_i + 0.1142\text{noff1618}_i + 0.1661\sqrt{\text{nprop1416}}_i + \ln(t)_{i,18-20}. \quad (8.10)$$

Variable	Seuils		
	Étape 1	Étape 2	Étape 3
viol1218	Cette variable ne peut être transformée		
nsex1416	Cette variable ne peut être transformée		
noff1618	**0.3423764**		
nsex1618	0.1438527	**0.1507058**	
nprop1416	0.0481720	0.0516617	0.04622183

TAB. 8.9 – *Seuils associés aux transformations des variables (18-20 ans).*

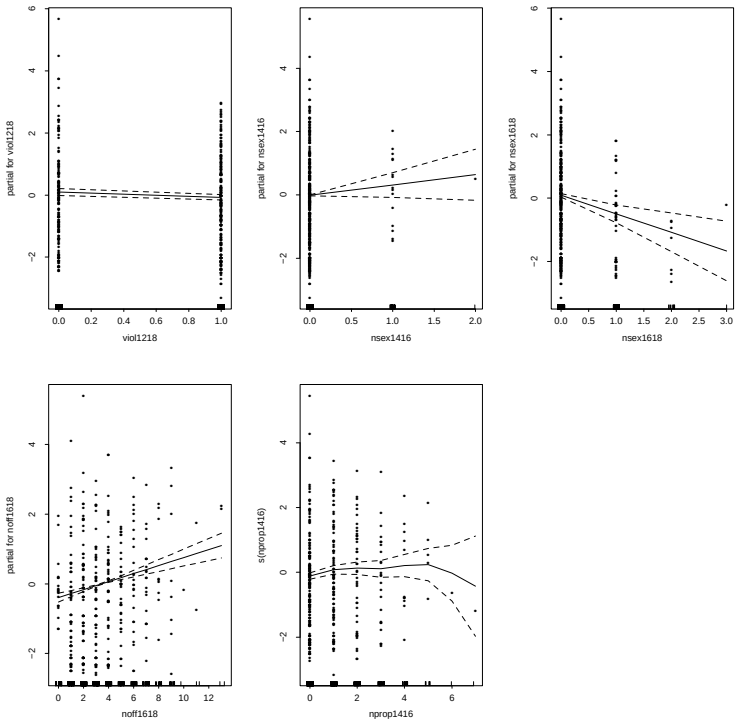

FIG. 8.11 – *Modèle final (entre 18 et 20 ans). Les transformations ont été obtenues par GAM.*

Entre 20 et 22 ans

Le modèle qui sera analysé ici est celui impliquant les variables prop1218, autres1218, ndrug1618, nsex1214, nsex1618, nviol1416, nviol1618, noff1416, noff1618 et nautres1416. Quand un modèle GAM est ajusté, on obtient les graphiques présentés à la FIGURE 8.12.

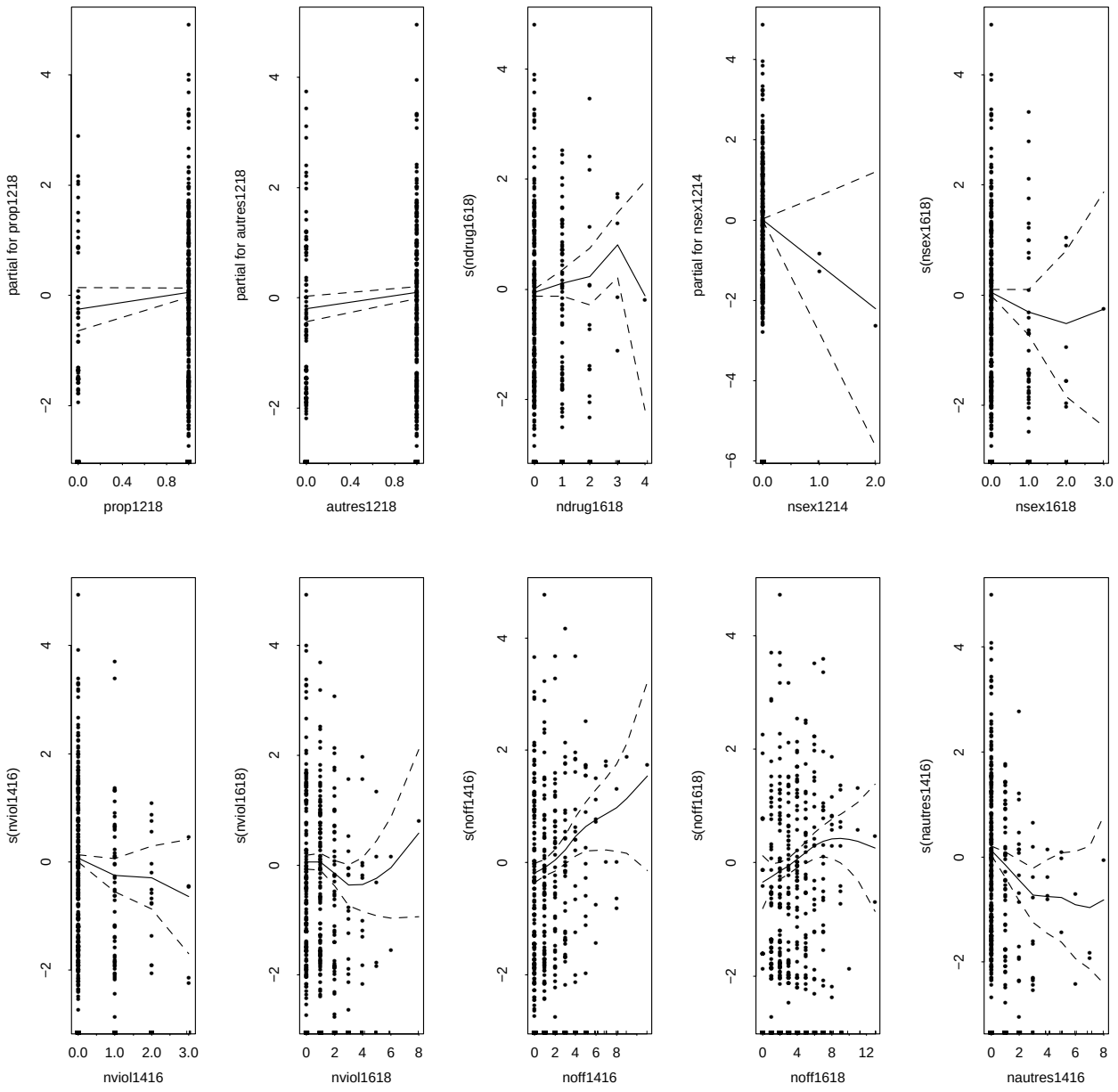

FIG. 8.12 – *Modèle de départ (20-22 ans). Les transformations ont été obtenues par GAM.*

Le TABLEAU 8.10 montre les seuils associés à ces transformations. Toutes les transformations doivent être éliminées, faisant en sorte que l'on doit conserver le modèle de départ, c'est-à-dire

$$\begin{aligned}\ln(\text{off})_{i,20-22} = &-0.9336 + 0.2981\text{prop1218}_i + 0.2628\text{autres1218}_i + 0.1549\text{ndrug1618}_i - \\ &1.0556\text{nsex1214}_i - 0.2602\text{nsex1618}_i - 0.2279\text{nviol1416}_i - \\ &0.0895\text{nviol1618}_i + 0.1511\text{noff1416}_i + 0.1019\text{noff1618} - \\ &0.1926\text{nautres1416} + \ln(t)_{i,20-22}. \end{aligned} \tag{8.11}$$

Variable	Seuils						
	Étape 1	Étape 2	Étape 3	Étape 4	Étape 5	Étape 6	Étape 7
prop1218	Cette variable ne peut être transformée						
autres1218	Cette variable ne peut être transformée						
nsex1214	Cette variable ne peut être transformée						
nsex1618	**0.6435888**						
nviol1416	0.6373136	**0.6371616**					
noff1416	0.4560498	0.4397867	**0.4202859**				
ndrug1618	0.2423160	0.2392897	0.2221160	**0.2495839**			
nautres1416	0.0833778	0.0865182	0.0727718	0.0795177	**0.1104802**		
noff1618	0.1835617	0.1838419	0.1390048	0.1036550	0.1060273	**0.1838760**	
nviol1618	0.0216836	0.0214012	0.0223645	0.0224958	0.0195658	0.0185035	**0.1210953**

TAB. 8.10 – *Seuils associés aux transformations des variables (20-22 ans).*

Entre 22 et 24 ans

Le modèle qui sera ici analysé est celui avec les variables nsex1618 et noff1618. Les transformations obtenues par un modèle additif généralisé sont présentées à la FIGURE 8.13, et les seuils qui leur sont associés sont au TABLEAU 8.11. Comme aucune des transformations n'est conservée, nous gardons le modèle de départ :

$$\ln(\text{off})_{i,22-24} = -0.7503 - 0.5766\text{nsex1618}_i + 0.1371\text{noff1618} + \ln(t)_{i,22-24}. \quad (8.12)$$

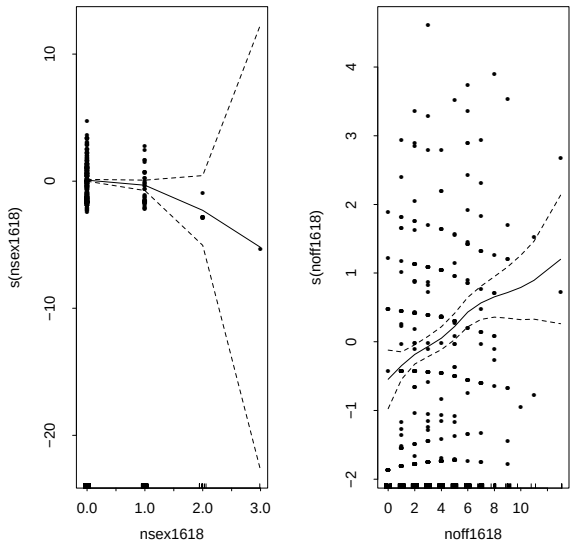

FIG. 8.13 – *Modèle de départ (22-24 ans). Les transformations ont été obtenues par GAM.*

Variable	Seuils	
	Étape 1	Étape 2
noff1618	**0.2059386**	
nsex1618	0.1633683	**0.158931**

TAB. 8.11 – *Seuils associés aux transformations des variables (22-24 ans).*

Entre 18 et 20 ans		
Seuil	Variables retenues	PRESS
1% 5%	nsex1618 noff1618	2298.93
10%	viol1218 nsex1618 noff1618 nprop1416	2309.18
15%	viol1218 nsex1416 nsex1618 noff1618 nprop1416	**2294.76**
20%	drug1218 viol1218 autres1218 ndrug1618 nsex1416 nsex1618 nviol1214 noff1618 nprop1214 nprop1416 nautres1214	2383.59
30%	drug1218 viol1218 off1218 autres1218 ndrug1618 nsex1416 nsex1618 nviol1214 noff1618 nprop1214 nprop1416 nautres1214	2385.14

Entre 20 et 22 ans		
Seuil	Variables retenues	PRESS
1%	noff1618	1733.90
5%	nsex1618 nviol1416 noff1416 noff1618 nautres1416	1677.67
10%	autres1218 nsex1618 nviol1416 noff1416 noff1618 nautres1416	1658.44
15%	autres1218 ndrug1618 nsex1214 nsex1618 nviol1416 nviol1618 noff1416 noff1618 nautres1416	1624.70
20%	prop1218 autres1218 ndrug1618 nsex1214 nsex1618 nviol1416 nviol1618 noff1416 noff1618 nautres1416	**1619.59**
30%	prop1218 autres1218 ndrug1618 nsex1214 nsex1618 nviol1416 nviol1618 noff1416 noff1618 nautres1416 nautres1618	1633.82

Entre 22 et 24 ans		
Seuil	Variables retenues	PRESS
1% 5%	nsex1618 noff1618	**1095.84**
10%	nsex1618 noff1416 noff1618	1096.28
15%	drug1218 ndrug1618 nsex1618 nviol1214 noff1214 noff1416 noff1618 nprop1214	1132.25
20%	drug1218 prop1218 ndrug1618 nsex1618 nviol1214 noff1214 noff1416 noff1618 nprop1214 nprop1416	1144.19
30%	drug1218 viol1218 prop1218 ndrug1618 nsex1618 nviol1214 noff1214 noff1416 noff1618 nprop1214 nprop1416	1145.25

TAB. 8.12 – *Résultats obtenus pour l'analyse du nombre total de crimes selon les âges.*

8.3 Interprétation des modèles (8.1) à (8.12)

L'interprétation des paramètres a été vue à la sous-section 2.4.3, en page 11. Nous avons donc que $\hat{\beta}_0$ représente le logarithme naturel de l'espérance de la variable réponse lorsque les p' variables exogènes prennent simultanément la valeur 0.

Donc, dans le modèle (8.1), si les 6 variables valent 0, l'espérance de la variable réponse doit être de $e^{-1.6078+\ln(t)_{it}} = 0.2t_{it}$. Ainsi, on s'attend d'une personne ayant passé 2 ans dans la classe d'âge 20-22 qu'elle commette $0.2 \times 2 = 0.4 \approx 0$ crime entre 20 et 22 ans. Dans le modèle (8.2), $0.07t_{it}$ crimes sont attendus, alors que dans le modèle (8.3), $0.6t_{it}$ crimes sont attendus d'une personne dont les variables exogènes sont toutes simultanément nulles.

Dans les autres modèles (8.4) à (8.12), ce sont respectivement $0.27t_{i,18-20}$, $0.18t_{i,20-22}$, $0.10t_{i,22-24}$, $0.11t_{i,18-20}$, $0.1260t_{i,20-22}$, $1.15t_{i,22-24}$, $0.92t_{i,18-20}$, $0.39t_{i,20-22}$ et $0.47t_{i,22-24}$ crimes qui sont attendus d'une personne dont les variables exogènes prennent simultanément la valeur 0. Donc, une personne qui a passé 2 ans entre 22 et 24 ans et dont toutes ses variables sont nulles est susceptible de ne commettre environ aucun crime « Violence », 2 crimes « Drogues » et 1 crime de tous types entre 22 et 24 ans.

Quant aux autres paramètres $\hat{\beta}_k$, ils représentent l'effet, sur la variable étudiée, de l'augmentation d'une unité de la variable $x_{i\ell}$ ($\ell < p'$) en gardant constantes toutes les autres variables du modèle. Le Tableau 8.13 en page suivante[5] présente ces effets. Par exemple, l'augmentation d'une unité de la variable ndrug1416 telle qu'elle est présente dans le modèle (8.1) fait diminuer le nombre de crimes de type « Violence » commis après 18 ans de $100 \times (1 - e^{-0.2505}) \approx 22\%$, en autant que les variables nprop1416, nprop1618, nviol1618, nsex1618, noff1618 et t restent constantes.

[5]Dans ce tableau, on voit que la variable noff1618 augmente d'une quantité inconnue dans les modèles (8.6) et (8.9). La raison est que, dans ces deux modèles, la variable noff1618 est transformée de façon non linéaire, faisant en sorte qu'il soit impossible de savoir l'effet exact d'une augmentation de 1 unité de cette variable exogène sur la variable réponse. Il en est de même pour la variable nprop1416 dans les modèles (8.1), (8.3) et (8.10).

	Modèles											
	(8.1)	(8.2)	(8.3)	(8.4)	(8.5)	(8.6)	(8.7)	(8.8)	(8.9)	(8.10)	(8.11)	(8.12)
	Viol.	Drug.	tous types	Violence			Drogues			tous types		
Variables explicatives												
age1st	Cette variable n'est dans aucun des 12 modèles obtenus											
viol1218			↓16%							↓18%		
drug1218		↑34%			↑37%							
sex1218		↓38%							↓19%			
off1218		↑71%										
prop1218		↓39%							↓18%		↑34%	
autres1218		↑24%	↑23%								↑30%	
nviol1214	Cette variable n'est dans aucun des 12 modèles obtenus											
nviol1416		↓24%			↓21%						↓20%	
nviol1618	↑16%	↑11%		↑21%	↑9%						↓9%	
ndrug1416	↓22%	↑23%			↓30%							
ndrug1618			↑10%								↑17%	
nsex1214					↓99%						↓65%	
nsex1416			↑34%	↑71%						↑38%		
nsex1618	↓43%		↓39%	↓46%	↓40%	↓41%				↓44%	↓23%	↓44%
noff1214		↓30%										
noff1416		↑9%			↑13%						↑16%	
noff1618	↑10%	↑12%	↑11%		↑10%	↑			↑	↑12%	↑11%	↑15%
nprop1214		↑46%										
nprop1416	↑		↑	↑13%						↑		
nprop1618	↓6%	↓6%		↓10%								
nautres1214					↑17%							
nautres1416					↓12%						↓18%	
nautres1618				↑8%			↑19%					

TAB. 8.13 – *Effets, sur la variable étudiée, d'une augmentation de 1 unité des variables explicatives, en gardant constantes les autres variables explicatives impliquées dans le modèle.*

8.4 Discussion des résultats obtenus

Des faits intéressants sont à tirer des douze modèles obtenus. En effet, on remarque que le nombre total de crimes commis entre 16 et 18 ans (noff1618) se retrouve très fréquemment dans les modèles, et donc cette variable est très importante pour faire les prévisions désirées. L'interprétation du paramètre β qui lui est associé nous fait conclure que si une personne en particulier avait commis un crime de plus entre 16 et 18 ans, l'espérance du nombre de crimes à l'état adulte d'un type donné augmenterait. Ainsi, plus une personne commet un grand nombre de crimes entre 16 et 18 ans, plus on s'attend à ce que cette personne commette de crimes une fois adulte.

Une autre conclusion similaire peut être tirée avec la variable nsex1618. Le paramètre associé à cette variable nous incite à croire qu'un individu ayant commis un crime de type « Sexe » de plus entre 16 et 18 ans serait porté à commettre moins de crime d'un type donné une fois adulte. La variable nprop1416 a quant à elle l'effet contraire sur la variable réponse : un crime de type « Propriété » commis de plus entre 14 et 16 ans augmente l'espérance de la variable réponse.

On peut voir du Tableau 8.12, en page 104, que la statistique de validation croisée diminue au fur et à mesure que l'âge de l'individu augmente. Cette même constatation peut être faite à partir des Tableaux 8.1, 8.2 et 8.3 des pages 81, 84 et 87, respectivement. Ceci nous porte donc à croire qu'il devient plus facile de prédire ce que fera un individu plus âgé. Comme la Figure 7.1 du chapitre 7 montrait que le nombre de crimes (quelque soit le type) diminue une fois passé l'âge de 18 ans, le fait que la statistique PRESS diminue est probablement dû à cette plus faible variabilité dans les nombres de crimes.

8.5 Conclusion du chapitre

Une autre analyse sera faite au chapitre suivant, mais avec des modèles linéaires généralisés longitudinaux mixtes et avec les variables de diagnostics psychiatriques afin de voir s'il peut être plus facile de prédire le nombre de crimes entre 18 et 20 ans selon cette nouvelle approche.

Chapitre 9

La deuxième analyse avec des GLMM

Le but de ce chapitre est de trouver un modèle servant à prédire le nombre total de crimes commis entre 18 et 20 ans à l'aide de modèles linéaires généralisés longitudinaux mixtes et à l'aide des dix-neuf variables de diagnostics psychiatriques disponibles.

Les trois modèles suivants seront donc ajustés aux données :

$$\ln(\text{off})_{it'} = \beta_0 + \beta_1\gamma_{i01} + \beta_2\text{age1st}_i + \beta_3\text{adhd}_i + \beta_4\text{adjust}_i + \beta_5\text{anxiety}_i + \beta_6\text{conduct}_i + \beta_7\text{dev}_i + \beta_8\text{dbd}_i + \beta_9\text{dissoc}_i + \beta_{10}\text{impulse}_i + \beta_{11}\text{mood}_i + \beta_{12}\text{mental}_i + \beta_{13}\text{organic}_i + \beta_{14}\text{person}_i + \beta_{15}\text{subst}_i + \beta_{16}\text{psychot}_i + \beta_{17}\text{sexual}_i + \beta_{18}\text{sleep}_i + \beta_{19}\text{somato}_i + \beta_{20}\text{ld}_i + \beta_{21}\text{comm}_i + \ln(t)_{it'},$$

$$\ln(\text{off})_{it'} = \beta_0 + \beta_1\gamma_{i02} + \beta_2\text{age1st}_i + \beta_3\text{adhd}_i + \beta_4\text{adjust}_i + \beta_5\text{anxiety}_i + \beta_6\text{conduct}_i + \beta_7\text{dev}_i + \beta_8\text{dbd}_i + \beta_9\text{dissoc}_i + \beta_{10}\text{impulse}_i + \beta_{11}\text{mood}_i + \beta_{12}\text{mental}_i + \beta_{13}\text{organic}_i + \beta_{14}\text{person}_i + \beta_{15}\text{subst}_i + \beta_{16}\text{psychot}_i + \beta_{17}\text{sexual}_i + \beta_{18}\text{sleep}_i + \beta_{19}\text{somato}_i + \beta_{20}\text{ld}_i + \beta_{21}\text{comm}_i + \ln(t)_{it'},$$

$$\ln(\text{off})_{it'} = \beta_0 + \beta_1\text{age1st}_i + \beta_2\text{adhd}_i + \beta_3\text{adjust}_i + \beta_4\text{anxiety}_i + \beta_5\text{conduct}_i + \beta_6\text{dev}_i + \beta_7\text{dbd}_i + \beta_8\text{dissoc}_i + \beta_9\text{impulse}_i + \beta_{10}\text{mood}_i + \beta_{11}\text{mental}_i + \beta_{12}\text{organic}_i + \beta_{13}\text{person}_i + \beta_{14}\text{subst}_i + \beta_{15}\text{psychot}_i + \beta_{16}\text{sexual}_i + \beta_{17}\text{sleep}_i + \beta_{18}\text{somato}_i + \beta_{19}\text{ld}_i + \beta_{20}\text{comm}_i + \ln(t)_{it'},$$

$$i = 1, 2, \ldots, 360$$

et

$$t' = \begin{cases} \text{classe d'âge 12-14 ans} \\ \text{classe d'âge 14-16 ans} \\ \text{classe d'âge 16-18 ans} \\ \text{classe d'âge 18-20 ans.} \end{cases}$$

L'âge au premier crime se trouve toujours dans les modèles (age1st), et les variables adhd, adjust, anxiety, conduct, dev, dbd, dissoc, impulse, mood, mental, organic, person, subst, psychot, sexual, sleep, somato, ld et comm sont les variables indicatrices des diagnostics psychiatriques, et la variable *offset*, t, est toujours présente dans les modèles, et celle-ci représente le temps exact, en années, où l'individu était à risque de commettre des crimes dans chacune des classes d'âge 12-14, 14-16, 16-18 et 18-20 ans.

La variable γ_{i01} a été obtenue en ajustant ce modèle à ordonnée à l'origine aléatoire :

$$\ln(\text{off})_{it^*} = (\beta_0 + \gamma_{i01}) + \beta_1 \text{age}_{it^*} + \beta_2 \text{age}^2_{it^*} + \ln(t)_{it^*},$$

tandis que la variable γ_{i02} a été obtenue en ajustant le modèle suivant, toujours avec une ordonnée à l'origine aléatoire :

$$\begin{gathered} \ln(\text{off})_{it^*} = (\beta_0 + \gamma_{i02}) + \beta_1 \text{age}_{it^*} + \beta_2 \text{age}^2_{it^*} + \beta_3 \text{age1st}_i + \beta_4 \text{adhd}_i + \beta_5 \text{adjust}_i + \\ \beta_6 \text{anxiety}_i + \beta_7 \text{conduct}_i + \beta_8 \text{dev}_i + \beta_9 \text{dbd}_i + \beta_{10} \text{dissoc}_i + \beta_{11} \text{impulse}_i + \beta_{12} \text{mood}_i + \\ \beta_{13} \text{mental}_i + \beta_{14} \text{organic}_i + \beta_{15} \text{person}_i + \beta_{16} \text{subst}_i + \beta_{17} \text{psychot}_i + \beta_{18} \text{sexual}_i + \\ \beta_{19} \text{sleep}_i + \beta_{20} \text{somato}_i + \beta_{21} \text{ld}_i + \beta_{22} \text{comm}_i + \ln(t)_{it^*} \end{gathered}$$

avec

$$t^* = \begin{cases} \text{classe d'âge 12-14 ans} \\ \text{classe d'âge 14-16 ans} \\ \text{classe d'âge 16-18 ans} \end{cases}$$

et

$$\text{age} = \begin{cases} 13 & \text{si classe d'âge 12-14 ans} \\ 15 & \text{si classe d'âge 14-16 ans} \\ 17 & \text{si classe d'âge 16-18 ans.} \end{cases}$$

Chacune des composantes des deux vecteurs $\boldsymbol{\gamma}$ représente les tendances des individus à commettre des crimes et sera utile afin d'ajuster les trois modèles mentionnés à la page précédente.

9.1 Les résultats des analyses effectuées

Nous avons d'abord utilisé l'âge et l'âge au carré afin d'obtenir un premier vecteur $\boldsymbol{\gamma}$, noté $\boldsymbol{\gamma_{01}}$, obtenu en considérant aléatoire l'ordonnée à l'origine. Nous avons ensuite utilisé l'âge, l'âge au carré, l'âge au premier délit (age1st) et les 19 variables psychiatriques afin de construire un modèle avec effets aléatoires sur l'ordonnée à l'origine et obtenir un deuxième vecteur $\boldsymbol{\gamma}$, noté $\boldsymbol{\gamma_{02}}$. Un seuil d'exclusion de 5% a été utilisé afin de sélectionner, parmi les 19 variables psychiatriques, les variables les plus importantes pour la création de ce vecteur $\boldsymbol{\gamma_{02}}$. Les deux variables adjust et organic se sont révélées significatives au seuil de 5%. Nous avons donc obtenu les 2 modèles suivants :

$$\ln(\text{off})_{it^*} = (4.0454 + \gamma_{i01}) - 0.6170\text{age}_{it^*} + 0.02665\text{age}^2_{it^*} + \ln(t)_{it^*}$$
et
$$\ln(\text{off})_{it^*} = (4.7983 + \gamma_{i02}) - 0.5535\text{age}_{it^*} + 0.02560\text{age}^2_{it^*} - 0.09881\text{age1st}_i - 0.2372\text{adjust}_i - 0.3713\text{organic}_i + \ln(t)_{it^*}.$$

Afin d'effectuer ces analyses, les critères AIC, HQIC, BIC et CAIC présentés à la section 5.5 de la page 55 sont utilisés. Le Tableau 9.1 montre que le modèle à utiliser est celui avec une matrice non-structurée, car le modèle ayant les plus petites valeurs de AIC, HQIC, BIC et CAIC s'avère être le meilleur :

	Création du premier vecteur $\boldsymbol{\gamma_{01}}$			
Type	AIC	BIC	CAIC	HQIC
UNSTR	**1374.67**	**1378.56**	**1379.56**	**1376.22**
AR(1)	1376.67	1384.44	1386.44	1379.76
ARMA(1,1)	1378.67	1390.33	1393.33	1383.31
CS	1376.67	1378.56	1379.56	1376.22
VC	1374.67	1378.56	1379.56	1376.22

	Création du deuxième vecteur $\boldsymbol{\gamma_{02}}$			
Type	AIC	BIC	CAIC	HQIC
UNSTR	**1364.10**	**1367.99**	**1368.99**	**1365.65**
AR(1)	1366.10	1373.88	1375.88	1369.19
ARMA(1,1)	1368.10	1379.76	1382.76	1372.74
CS	Problème de convergence avec ce type			
VC	1364.10	1367.99	1368.99	1365.65

Tab. 9.1 – *Statistiques pour le choix du modèle à utiliser.*

Les deux vecteurs $\boldsymbol{\gamma_{01}}$ et $\boldsymbol{\gamma_{02}}$ seront utilisés afin de construire des modèles avec mesures répétées (entre 12 et 20 ans) utilisant les variables de diagnostics psychiatriques

et l'âge au premier délit. Les variables retenues selon les divers seuils d'exclusion et selon les diverses matrices de corrélation sont présentées aux TABLEAUX 9.2, 9.3 et 9.4.

9.1.1 Analyse avec $\boldsymbol{\gamma_{01}}$

Le meilleur modèle servant de prévoir le nombre de crimes entre 18 et 20 ans est celui faisant intervenir les variables age1st, conduct, dbd, impulse, mental, person, psychot, sexual, ld et $\boldsymbol{\gamma_{01}}$. Comme plusieurs variables indicatrices sont impliquées dans ce modèle, nous tenterons de transformer l'âge au premier délit et le $\boldsymbol{\gamma_{01}}$ pour améliorer le modèle. Les résultats de ces transformations sont à la FIGURE 9.1 de la page suivante. La transformation inverse de l'âge et le carré de $\boldsymbol{\gamma_{01}}$ semblent des choix raisonnables afin d'améliorer le modèle et une statistique de validation croisée de 2703.87 a été obtenue avec cette transformation. Nous pouvons donc admettre ce modèle afin de prévoir le nombre de crimes commis entre 18 et 20 ans :

$$\begin{aligned}\ln(\text{off})_{i,18-20} = 1.5570 &- 14.8859\text{age1st}_i^{-1} - 0.0897\text{conduct}_i + 0.2251\text{dbd} - \\ &0.0881\text{impulse}_i - 1.0765\text{mental}_i + 0.1216\text{person}_i + \\ &0.1354\text{psychot}_i - 0.2235\text{sexual}_i - 0.1008\text{ld}_i + 2.2822\gamma_{i01} - \\ &1.2019\gamma_{i01}^2 + \ln(t)_{i,18-20}. \end{aligned} \tag{9.1}$$

TYPE=CS		
Seuil	Variables retenues	PRESS
1% 5%	age1st sexual $\boldsymbol{\gamma_{01}}$	2799.85
10%	age1st impulse sexual $\boldsymbol{\gamma_{01}}$	2788.70
15%	age1st adhd dbd impulse mental sexual $\boldsymbol{\gamma_{01}}$	2768.72
20% 30%	age1st adhd conduct dbd impulse mental person psychot sexual ld $\boldsymbol{\gamma_{01}}$	2762.16
TYPE=AR(1)		
Seuil	Variables retenues	PRESS
1% 5%	age1st sexual $\boldsymbol{\gamma_{01}}$	2788.15
10%	age1st impulse sexual $\boldsymbol{\gamma_{01}}$	2776.40
15%	age1st dbd impulse mental sexual $\boldsymbol{\gamma_{01}}$	2753.04
20%	age1st conduct dbd impulse mental person psychot sexual ld $\boldsymbol{\gamma_{01}}$	**2744.69**
30%	age1st adhd conduct dbd impulse mental person psychot sexual ld $\boldsymbol{\gamma_{01}}$	2746.03

TAB. 9.2 – *Résultats obtenus pour l'analyse avec* $\boldsymbol{\gamma_{01}}$.

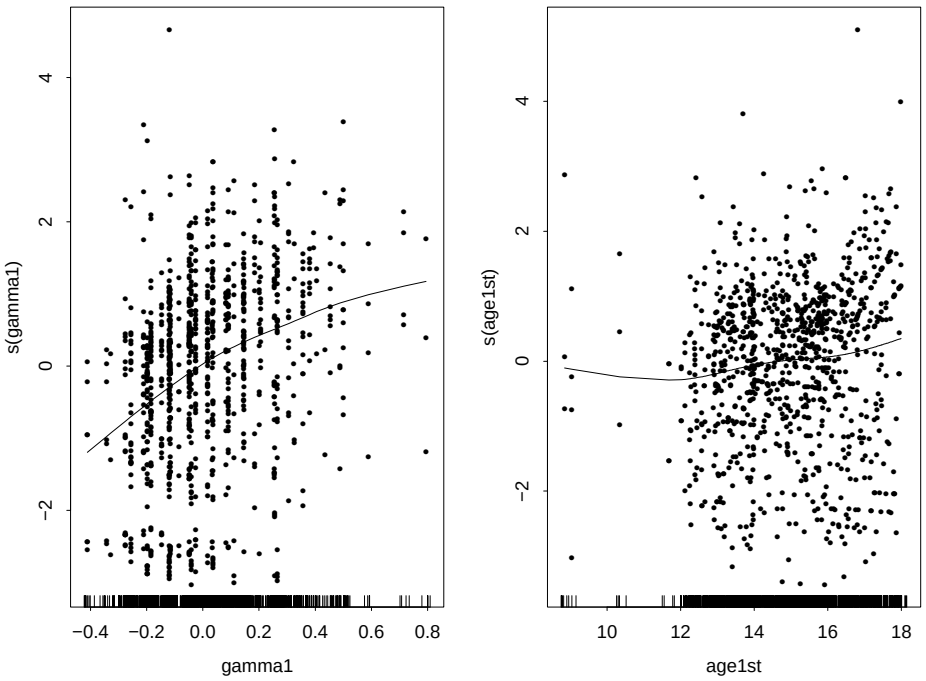

FIG. 9.1 – *Variables âge et* $\boldsymbol{\gamma_{01}}$ *au départ. Les transformations ont été obtenues par GAM.*

9.1.2 Analyse avec $\boldsymbol{\gamma_{02}}$

Il peut être vu au TABLEAU 9.3 de la page suivante que le meilleur modèle est celui regroupant les variables adjust, dbd, impulse, mental, organic, sexual et $\boldsymbol{\gamma_{02}}$. La seule variable pouvant être transformée dans ce modèle est $\boldsymbol{\gamma_{02}}$ et plusieurs transformations seront essayées afin d'améliorer le PRESS de départ. Cependant, le logarithme ou la racine carrée de $\boldsymbol{\gamma_{02}}$ ne peuvent être envisagées étant donné que ce vecteur possède des composantes négatives. Par GAM, la représentation à la FIGURE 9.2 est obtenue. Cette figure montre qu'une transformation n'est pas nécessaire, étant donné qu'aucune forme (ou courbure) particulière ne peut être vue sur ce graphique. Nous acceptons alors de modéliser le nombre de crimes commis entre 18 et 20 ans de la façon suivante :

$$\begin{aligned}\ln(\text{off})_{i,18-20} = {} & 0.5461 - 0.1802\text{adjust}_i + 0.2156\text{dbd}_i - 0.1277\text{impulse}_i - \\ & 1.1557\text{mental}_i - 0.2093\text{organic}_i - 0.2313\text{sexual}_i + \\ & 2.1388\gamma_{i02} + \ln(t)_{i,18-20}.\end{aligned} \tag{9.2}$$

TYPE=CS		
Seuil	Variables retenues	PRESS
1%	sexual $\boldsymbol{\gamma_{02}}$	2741.26
5%	impulse sexual $\boldsymbol{\gamma_{02}}$	2729.21
10%	adhd adjust impulse sexual $\boldsymbol{\gamma_{02}}$	2729.41
15%	adhd adjust dbd impulse mental organic sexual $\boldsymbol{\gamma_{02}}$	2712.32
20% 30%	adhd adjust conduct dbd impulse mental organic person psychot sexual ld $\boldsymbol{\gamma_{02}}$	2705.85
TYPE=AR(1)		
Seuil	Variables retenues	PRESS
1%	sexual $\boldsymbol{\gamma_{02}}$	2720.74
5%	impulse sexual $\boldsymbol{\gamma_{02}}$	2707.91
10%	adjust impulse organic sexual $\boldsymbol{\gamma_{02}}$	2716.32
15%	adjust dbd impulse mental organic sexual $\boldsymbol{\gamma_{02}}$	**2690.77**
20%	age1st adjust conduct dbd impulse mental organic person psychot sexual ld $\boldsymbol{\gamma_{02}}$	2733.84
30%	age1st adhd adjust conduct dbd impulse mental organic person psychot sexual ld $\boldsymbol{\gamma_{02}}$	2734.67

TAB. 9.3 – *Résultats obtenus pour l'analyse avec* $\boldsymbol{\gamma_{02}}$.

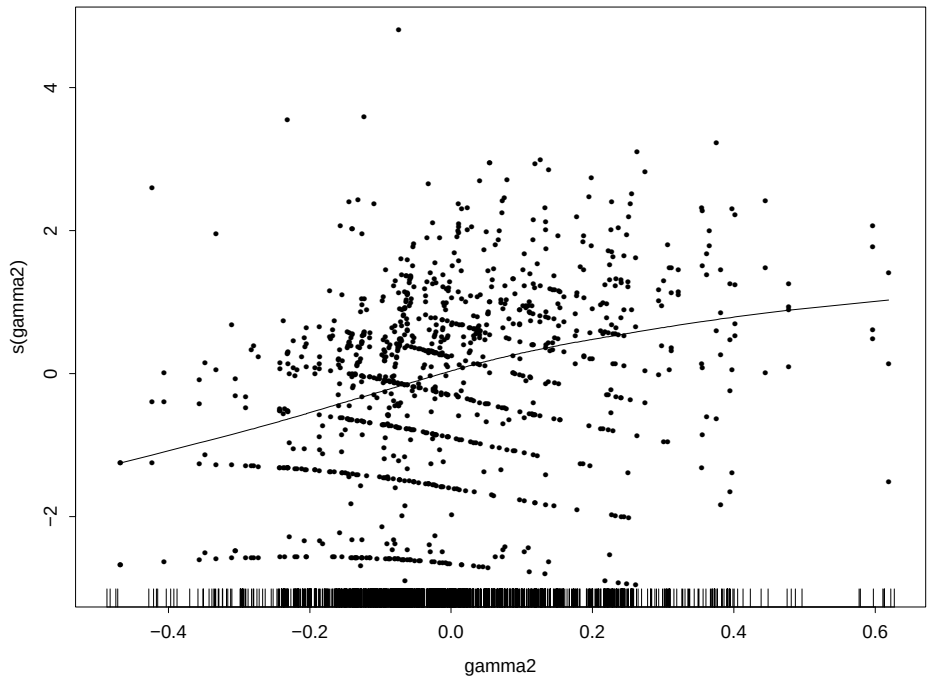

FIG. 9.2 – *Variable* $\boldsymbol{\gamma_{02}}$ *au départ. La transformation a été obtenue par GAM.*

9.1.3 Analyse sans $\boldsymbol{\gamma}$

Les résultats obtenus selon les différents seuils d'exclusion sont présentés au Tableau 9.4. Comme l'âge au premier délit n'est retenu dans aucun des modèles et qu'aucun $\boldsymbol{\gamma}$ ne fait évidemment partie de ceux-ci, nous nous devons de conserver le meilleur modèle tel qu'il est, car il est composé uniquement de variables indicatrices. Alors le modèle servant à prédire le nombre total de crimes commis entre 18 et 20 ans est le suivant :

$$\begin{aligned}\ln(\text{off})_{i,18-20} = 0.5805 - 0.2347\text{adjust}_i - 0.1490\text{mood}_i - 1.3264\text{mental}_i - \\ 0.2037\text{organic}_i - 0.3154\text{sexual}_i - 0.1640\text{ld}_i + \ln(t)_{i,18-20}. \qquad (9.3)\end{aligned}$$

TYPE=CS		
Seuil	Variables retenues	PRESS
1%	sexual	2768.09
5%	sexual ld	2763.69
10% 15%	adjust mood sexual ld	2749.53
20% 30%	adjust mood mental organic sexual ld	**2747.28**
TYPE=AR(1)		
Seuil	Variables retenues	PRESS
1%	sexual	2807.84
5% 10%	sexual ld	2801.56
15%	adjust mood sexual ld	2790.95
20%	adjust mood mental organic sexual ld	2787.60
30%	adjust mood mental organic person sexual ld	2792.18

Tab. 9.4 – *Résultats obtenus pour l'analyse sans $\boldsymbol{\gamma}$.*

9.2 Interprétation des modèles (9.1), (9.2) et (9.3)

Il a été mentionné à la sous-section 2.4.3 de la page 11 que le $\hat{\beta}_0$ représente le logarithme naturel de la variable réponse lorsque toutes les variables explicatives prennent simultanément la valeur 0. Quant aux autres paramètres $\hat{\beta}_k$, ils représentent l'effet, sur la variable réponse, de l'augmentation de 1 unité de la variable $x_{i\ell}$ ($\ell < p'$) en gardant constantes toutes les autres variables du modèle.

Dans le modèle (9.1), il est impossible d'interpréter de façon justifiée la valeur de $\hat{\beta}_0$. Étant donné que ce modèle fait intervenir l'âge au premier crime, il est impossible de poser cette variable égale à 0 puisqu'elle prend des valeurs comprises entre 9 et 22 ans, environ. Quant au modèle (9.2), on s'attend à ce qu'une personne commette $e^{0.5461+\ln(t)_{i,18-20}} = 1.73t_{i,18-20}$ crimes si toutes ses variables explicatives sont nulles. Comme on s'intéresse particulièrement à la classe d'âge 18-20, on s'attend à ce qu'une personne commette $1.73 \times 2 = 3.45 \approx 3$ crimes si elle a été à risque pendant 2 ans entre 18 et 20 ans. Finalement, dans le modèle (9.3), ce sont $1.79t_{i,18-20}$ crimes qui sont attendus d'une personne dont les variables exogènes sont nulles. Ainsi, environ 4 crimes sont attendus d'une personne ayant été à risque pendant 2 ans entre 18 et 20 ans.

Le Tableau 9.5 en page suivante[1] présente l'effet de l'augmentation de 1 unité des variables explicatives sur le nombre total de crimes commis entre 18 et 20 ans. Par exemple, l'augmentation de 1 unité de la variable dbd telle qu'elle est présente dans le modèle (9.1) fait augmenter le nombre total de crimes commis entre 18 et 20 ans de $100 \times (e^{0.2251} - 1) \approx 25\%$, en autant que les variables age1st, conduct, impulse, mental, person, psychot, sexual, ld, $\boldsymbol{\gamma_{01}}$ et t restent constantes.

9.3 Discussion des résultats obtenus

Des 3 meilleurs modèles obtenus, le modèle avec la variable γ_{i02} possède le meilleur pouvoir prédictif du nombre total de crimes entre 18 et 20 ans. Il semble donc que la tendance à commettre des crimes estimée par l'ordonnée à l'origine aléatoire ajoute de l'information utile aux diagnostics psychiatriques. Plusieurs variables reviennent à plusieurs reprises dans les modèles. En effet, le paramètre β associé à la variable dbd est

[1] De ce tableau, on voit que la variable age1st entraîne une augmentation du nombre total de crimes commis entre 18 et 20 ans, mais d'une quantité inconnue. La raison est que, dans le modèle (9.1), la variable age1st est transformée de façon non linéaire.

	Modèles		
	(9.1)	(9.2)	(9.3)
Variables explicatives			
age1st	↑		
adjust		↓16%	↓21%
conduct	↓9%		
dbd	↑25%	↑24%	
impulse	↓8%	↓12%	
mood			↓14%
mental	↓66%	↓69%	↓73%
organic		↓19%	↓18%
person	↑13%		
psychot	↑14%		
sexual	↓20%	↓21%	↓27%
ld	↓10%		↓15%

TAB. 9.5 – *Effets, sur le nombre total de crimes commis entre 18 et 20 ans, d'une augmentation de 1 unité des variables explicatives, en gardant constantes les autres variables explicatives impliquées dans les trois modèles distincts.*

toujours positif, faisant en sorte qu'une personne qui aurait été diagnostiquée avec un problème « dbd » est susceptible de faire augmenter l'espérance du nombre de crimes commis entre 18 et 20 ans.

L'inverse survient pour les variables impulse, mental, sexual, ld, adjust et organic, dont les paramètres β sont tous inférieurs à 0. Ainsi, une personne qui aurait eu un diagnostic positif de l'un de ces problèmes serait susceptible de faire diminuer l'espérance du nombre de crimes commis entre 18 et 20 ans.

Chapitre 10

Conclusions

10.1 Conclusion des chapitres 8 et 9

L'un des buts des chapitres 8 et 9 était de prévoir le nombre total de crimes ayant été commis entre 18 et 20 ans par chacun des individus de l'étude. Dans le chapitre 8, nous avons donc utilisé des modèles linéaires généralisés longitudinaux afin de réaliser cette tâche, et les résultats se trouvent au Tableau 8.3. Les variables exogènes étaient les nombres de crimes des types violence, sexe, drogues, propriété, autres et total commis entre 12 et 14 ans, entre 14 et 16 ans et entre 16 et 18 ans (nviol1214, nviol1416, nviol1618, nsex1214, nsex1416, nsex1618, ndrug1416, ndrug1618, nprop1214, nprop1416, nrop1618, nautres1214, nautres1416, nautres1618, noff1214, noff1416 et noff1618). De plus, les variables indicatrices de la présence d'un de ces types de crimes (viol1218, sex1218, drug1218, prop1218, autres1218 et off1218) ont aussi été utilisées, en plus de l'âge au premier crime (age1st).

Le Tableau 8.13 montre que plusieurs variables se sont avérées très importantes afin de prévoir les nombres de crimes après 18 ans. Par exemple, on voit que la variable nsex1618 se trouve dans 8 des 12 modèles. Le paramètre qui lui est associé est toujours négatif, signifiant qu'une augmentation de 1 unité de cette variable fait diminuer l'espérance de la variable réponse. Il en est de même pour la variable noff1618, qui elle, a été éliminée de 3 modèles seulement. Le paramètre de cette variable porte à croire qu'une augmentation de 1 unité de cette dernière fait augmenter l'espérance de la variable réponse étudiée.

Dans le chapitre 9, nous avons plutôt utilisé des modèles linéaires généralisés mixtes afin de faire les prévisions voulues, et les résultats se trouvent au Tableau 9.3. Les variables indicatrices étaient ici toutes les variables de diagnostics psychiatriques (adhd, adjust, anxiety, conduct, dev, dbd, dissoc, impulse, mood, mental, organic, person, subst, psychot, sexual, sleep, somato, ld et comm).

Du Tableau 9.5, on voit que les variables mental et sexual se retrouvent dans les 3 modèles. On conclut de ces variables qu'une augmentation de 1 unité de celles-ci fait diminuer l'espérance de la variable réponse, puisque le paramètre est négatif.

Du Tableau 8.3, nous voyons que le meilleur modèle afin de prévoir le nombre total de crimes commis entre 18 et 20 ans par les 378 individus est le suivant :

$$\ln(\text{off})_{i,18-20} = -0.4393 + 0.3210\text{nsex1416}_i - 0.5329\text{nsex1618}_i + 0.1191\text{noff1618}_i + 0.0576\text{nprop1416}_i + \ln(t)_{i,18-20}.$$

Nous avons obtenu une statistique de validation croisée (PRESS) de 2431.64. Donc, $2431.64/378 = 6.4329$. Du Tableau 9.3, nous voyons que le meilleur modèle est le suivant :

$$\ln(\text{off})_{i,18-20} = 0.5461 + 2.1388\gamma_{i02} - 0.1802\text{adjust}_i + 0.2156\text{dbd}_i - 0.1277\text{impulse}_i - 1.1557\text{mental}_i - 0.2093\text{organic}_i - 0.2313\text{sexual}_i + \ln(t)_{i18-20}.$$

Ce modèle a une statistique PRESS de 2690.77 pour 360 individus. Nous avons alors $2690.77/360 = 7.4744$. Il est donc clair que le premier modèle a réussi à mieux prévoir le nombre total de crimes commis entre 18 et 20 ans, et ce, même si un nombre inégal d'individus est impliqué dans chacune des deux analyses.

10.2 Conclusion du mémoire

Ce mémoire a couvert de façon détaillée la régression de Poisson. Il a été vu que cette technique d'analyse servait à prévoir une certaine variable de dénombrement. Cependant, cette technique nécessite qu'une seule mesure soit prise sur chacun des n individus indépendants, ce qui n'est pas toujours le cas en pratique. Dans le cas où n_i ($n_i \neq 1$) mesures seraient prises sur les n individus de l'étude, nous avons vu que

les équations d'estimation généralisées peuvent être utilisées dans le but d'estimer de façon convergente les p paramètres du modèle. Il a été mentionné au chapitre 3 que les GEE utilisent des estimateurs robustes pour l'estimation des paramètres du modèle et de leur matrice de variance-covariance.

Nous avons aussi décrit une façon de modéliser la corrélation entre les mesures. Cette façon de faire nécessite l'ajout d'effets aléatoires au modèle. L'ajout d'effets aléatoires à un modèle linéaire généralisé (GLM) donne lieu à un modèle linéaire généralisé mixte (GLMM), le terme *mixte* venant de la présence de terme(s) fixe(s) et de terme(s) aléatoire(s) dans le modèle.

Il a de plus été vu que pour améliorer un modèle linéaire généralisé, un modèle additif généralisé est utilisé afin de trouver une transformation $f(\cdot)$ à appliquer à une ou à plusieurs variables. Cependant, on a vu que si nous sommes en présence de données corrélées entre elles, il est plus difficile de trouver la transformation $f(\cdot)$ à appliquer. En effet, contrairement au cas où une mesure est prise sur chacun des individus, nous ne pouvons nous fier sur les intervalles de confiance pouvant être faits par `plot.gam()`, et nous ne pouvons non plus nous fier sur les statistiques fournies par `summary()`. La nécessité d'inclure une transformation est donc prise « à l'oeil », ce qui implique la possibilité d'une décision différente d'un utilisateur à l'autre.

Beaucoup de travail reste encore à faire sur les données utilisées dans l'étude de cas des chapitres 8 et 9. Par exemple, on aurait pu utiliser comme variables explicatives toutes les variables déjà utilisées (par exemple, les variables de diagnostics psychiatriques et les nombres de crimes des types violence, sexe, drogues, propriété, autres et total commis entre 12 et 14 ans, entre 14 et 16 ans et entre 16 et 18 ans) afin de prévoir les nombres de crimes après 18 ans. Ainsi, des modèles différents auraient été obtenus, et peut-être que ceux-ci auraient pu mieux prévoir les variables étudiées.

Bibliographie

[1] Agresti A. (2002). *Categorical Data Analysis*. Wiley. New York.

[2] Bates D.M., & Watts D.G. (1988). *Nonlinear Regression Analysis and Its Applications*. John Wiley & Sons. New York.

[3] Breiman L., & Friedman J.H. (1985). « Estimating Optimal Transformations for Multiple Regression and Correlation ». *Journal of the American Statistical Association*, **80**(391), 580-598.

[4] Breslow N.E., & Clayton D.G. (1993). « Approximate Inference in Generalized Linear Mixed Models ». *Journal of the American Statistical Association*, **88**(421), 9-25.

[5] Cameron C.A., & Trivedi P.K. (1998). *Regression Analysis of Count Data*. Cambridge University Press. Cambridge.

[6] Chambers J.M., & Hastie T.J. (1992). *Statistical Models in S*. Wadsworth & Brooks. Yonkers.

[7] Chang Y.C. (2000). « Residuals Analysis of the Generalized Linear Models for Longitudinal Data ». *Statistics in Medicine*, **19**(10), 1277-1293.

[8] Hastie T.J., & Tibshirani R.J. (1990). *Generalized Additive Models*. Chapman & Hall. New York.

[9] Hardin J.W., & Hilbe J.M. (2002). *Generalized Estimating Equations*. Chapman & Hall/CRC. Boca Raton, FL.

[10] Liang K.Y., & Zeger S.L. (1986). « Longitudinal Data Analysis using Generalized Linear Models ». *Biometrika*, **73**(1), 13-22.

[11] McCullagh P., & Nelder J.A. (1989). *Generalized Linear Models*. Chapman & Hall. New York.

[12] McCulloch C.E., & Searle S.R. (2001). *Generalized, Linear and Mixed Models*. Wiley. New York.

[13] Myers R.H. (1990). *Classical and Modern Regression With Applications*. Duxbury classical series. Belmont, CA.

[14] Neter J., Kutner M.H., Nachtsheim C.J., & Wasserman W. (1996). *Applied Linear Regression Models*. Irwin. Chicago.

[15] Paternoster R., Brame R., & Farrington D.P. (2001). « On the Relationship Between Adolescent and Adult Conviction Frequencies ». *Journal of Quantitative Criminology*, **17**(3), 201-226.

[16] Pierce D.A., & Schafer D.W. (1986). « Residuals in Generalized Linear Models ». *Journal of the American Statistical Association*, **81**(396), 977-986.

[17] Piquero A.R., & Buka S.L. (2002). « Linking Juvenile and Adult Patterns of Criminal Activity in the Providence Cohort of the National Collaborative Perinatal Project ». *Journal of criminal justice*, **30**(4), 259-272.

[18] Rao P.S.R.S. (1999). *Variance Components Estimation : Mixed Models, Methodologies and Applications*. Chapman & Hall/CRC. London.

[19] Ripley B.D., & Venables W.N. (2002). *Modern Applied Statistics with S*. Springer. New York.

[20] Ross S.M. (1999). *Initiation aux probabilités*. Presses polytechniques et universitaires romandes. Lausanne.

[21] SAS Institute Inc. (2004). **SAS OnlineDoc® 9.1.3.** Cary, NC : SAS Institute Inc.

[22] Seber G.A.F. (1989). *Nonlinear Regression*. Wiley. New York.

[23] Tierney L., & Kadane J.B. (1986). « Accurate Approximations for Posterior Moments and Marginal Densities ». *Journal of the American Statistical Association*, **81**(393), 82-86.

[24] Thall P.F., & Vail S.C. (1990). « Some Covariance Models for Longitudinal Count Data with Overdispersion ». *Biometrics*, **46**(3), 657-671.

[25] Venables W.N., & Ripley B.D. (2002). *Modern Applied Statistics with S-Plus*. Springer. New York.

[26] Wolfinger R., & O'ConnelL R. (1993). « Generalized Linear Mixed Models : A Pseudo-likelihood Approach ». *Journal of Statistical Computation and Simulation*, **48**, 233-243.

[27] Zeger S.L., & Liang K.Y. (1986). « Longitudinal Data Analysis for Discrete and Continuous Outcomes ». *Biometrics*, **42**(1), 121-130.

[28] Zheng B. (2000). « Summarizing the Goodness of Fit of Generalized Linear Models for Longitudinal Data ». *Statistics in Medicine*, **19**(10), 1265-1275.

[29] Zorn C.J.W. (2001). « Generalized Estimating Equation Models for Correlated Data : A Review with Applications ». *American Journal of Political Science*, **45**(2), 470-490.

Troisième partie

Les annexes

Annexe A

Quelques démonstrations

A.1 La démonstration des formules pour l'espérance et la variance d'une loi faisant partie de la famille exponentielle

La fonction génératrice des moments d'une loi faisant partie de la famille exponentielle est

$$\begin{aligned}
\mathrm{M}_Y(t) = \mathbb{E}[e^{ty}] &= \int_{\mathcal{R}} e^{ty+\frac{y\theta-b(\theta)}{a(\phi)}-c(y,\phi)}dy \\
&= e^{\frac{-b(\theta)}{a(\phi)}} \int_{\mathcal{R}} e^{ty+\frac{y\theta}{a(\phi)}-c(y,\phi)}dy \\
&= \frac{e^{\frac{-b(\theta)}{a(\phi)}}}{e^{\frac{-b(a(\phi)t+\theta)}{a(\phi)}}} \int_{\mathcal{R}} e^{\frac{a(\phi)ty+y\theta}{a(\phi)}-c(y,\phi)-\frac{b(a(\phi)t+\theta)}{a(\phi)}}dy \\
&= \frac{e^{\frac{-b(\theta)}{a(\phi)}}}{e^{\frac{-b(a(\phi)t+\theta)}{a(\phi)}}} \underbrace{\int_{\mathcal{R}} e^{\frac{y(a(\phi)t+\theta)-b(a(\phi)t+\theta)}{a(\phi)}-c(y,\phi)}dy}_{=1} \\
&= e^{\frac{-b(\theta)+b(a(\phi)t+\theta)}{a(\phi)}}.
\end{aligned}$$

Ainsi, l'espérance de Y peut être obtenue comme suit :

$$\begin{aligned}
\mathbb{E}[Y] = \mathrm{M}'_Y(t)|_{t=0} &= \left(e^{\frac{-b(\theta)+b(a(\phi)t+\theta)}{a(\phi)}}\right)'\Big|_{t=0} \\
&= e^{\frac{-b(\theta)+b(a(\phi)t+\theta)}{a(\phi)}}\left(\frac{b'(a(\phi)t+\theta)a(\phi)}{a(\phi)}\right)\Big|_{t=0} \\
&= e^{\frac{-b(\theta)+b(\theta)}{a(\phi)}}b'(\theta) \\
&= b'(\theta).
\end{aligned}$$

□

Quant à la variance, on l'obtient de la façon suivante :

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[Y] &= (\mathrm{M}''_Y(t) - (\mathrm{M}'_Y(t))^2)\big|_{t=0}. \\
\text{Or, } \mathrm{M}''_{\mathrm{Y}}(\mathrm{t}) &= \left(e^{\frac{-b(\theta)+b(a(\phi)t+\theta)}{a(\phi)}}\right)''\Big|_{t=0} \\
&= e^{\frac{-b(\theta)+b(a(\phi)t+\theta)}{a(\phi)}}\{(b'(a(\phi)t+\theta))^2 + b''(a(\phi)t+\theta)a(\phi)\}\Big|_{t=0} \\
&= e^{\frac{-b(\theta)+b(\theta)}{a(\phi)}}\{(b'(\theta))^2 + b''(\theta)a(\phi)\} \\
&= (b'(\theta))^2 + b''(\theta)a(\phi) \\
\mathbb{V}\mathrm{ar}[Y] &= (b'(\theta))^2 + b''(\theta)a(\phi) - (b'(\theta))^2 \\
&= b''(\theta)a(\phi).
\end{aligned}$$

□

A.2 La démonstration des propriétés de la matrice H

Idempotence : HH=H

$$\begin{aligned}
\mathbf{HH} &= \mathbf{W}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}'\mathbf{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\underbrace{\mathbf{W}^{\frac{1}{2}}\mathbf{W}^{\frac{1}{2}}}_{=\mathbf{W}}\boldsymbol{X}(\boldsymbol{X}'\mathbf{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{W}^{\frac{1}{2}} \\
&= \mathbf{W}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}'\mathbf{W}\boldsymbol{X})^{-1}\underbrace{\boldsymbol{X}'\mathbf{W}\boldsymbol{X}(\boldsymbol{X}'\mathbf{W}\boldsymbol{X})^{-1}}_{=\mathbf{I}_p}\boldsymbol{X}'\mathbf{W}^{\frac{1}{2}} \\
&= \mathbf{W}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}'\mathbf{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{W}^{\frac{1}{2}} \\
&= \mathbf{H}
\end{aligned}$$

□

Symétrie : $\mathbf{H}' = \mathbf{H}$

$$\begin{aligned}
\mathbf{H}' &= (\mathbf{W}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}'\mathbf{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{W}^{\frac{1}{2}})' \\
&= (\mathbf{W}^{\frac{1}{2}})'(\boldsymbol{X}')'((\boldsymbol{X}'\mathbf{W}\boldsymbol{X})^{-1})'(\boldsymbol{X})'(\mathbf{W}^{\frac{1}{2}})' \\
&= \mathbf{W}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}'\mathbf{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{W}^{\frac{1}{2}} \text{ car } \mathbf{W} \text{ est une matrice diagonale} \\
&= \mathbf{H}
\end{aligned}$$

□

Trace de la matrice $\mathbf{H} = \boldsymbol{p}$

$$\begin{aligned}
\text{trace}(\mathbf{H}) &= \text{trace}(\mathbf{W}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}'\mathbf{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{W}^{\frac{1}{2}}) \\
&= \text{trace}((\boldsymbol{X}'\mathbf{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{W}^{\frac{1}{2}}\mathbf{W}^{\frac{1}{2}}\boldsymbol{X}) \\
&= \text{trace}((\boldsymbol{X}'\mathbf{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{W}\boldsymbol{X}) \\
&= \text{trace}(\mathbf{I}_{\boldsymbol{p}}) \\
&= p
\end{aligned}$$

□

A.3 Une démonstration pour la statistique de déviance

Puisqu'on parle d'un modèle linéaire supposant la normalité des observations, il suit que $Y_i \sim N(\mu_i, \sigma^2)$. Donc,

$$f(\boldsymbol{\mu}|\boldsymbol{Y};\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(Y_i-\mu_i)^2}{2\sigma^2}}. \tag{A.1}$$

La loi normale fait partie de la famille exponentielle, puisqu'il est possible de réexprimer la fonction de densité (A.1) de la façon suivante :

$$\begin{aligned} f(\boldsymbol{\mu}|\boldsymbol{Y};\sigma) &= \exp\left\{\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(Y_i-\mu_i)^2}{2\sigma^2}\right\} \\ &= \exp\left\{-\frac{Y_i^2}{2\sigma^2} + \frac{Y_i\mu_i}{\sigma^2} - \frac{\mu_i^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right\} \\ &= \exp\left\{\frac{Y_i\mu_i - \mu_i^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{Y_i^2}{\sigma^2} + \ln(2\pi\sigma^2)\right)\right\}. \end{aligned}$$

Donc, la loi normale fait partie de la famille exponentielle avec

$$\begin{aligned} \theta_i &= \mu_i, \\ b(\theta_i) &= \frac{\mu_i^2}{2} = \frac{\theta_i^2}{2}, \\ a(\phi) &= \sigma^2, \\ c(Y_i,\phi) &= \frac{1}{2}\left(\frac{Y_i^2}{\sigma^2} + \ln(2\pi\sigma^2)\right). \end{aligned}$$

Avec l'équation (2.3) de la page 6, on peut dire que la log-vraisemblance est

$$\ell(\boldsymbol{\mu}|\boldsymbol{Y};\sigma) = \sum_{i=1}^{n}\left\{\frac{Y_i\mu_i - \mu_i^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{Y_i^2}{\sigma^2} + \ln(2\pi\sigma^2)\right)\right\}.$$

Quant à la déviance, elle est définie comme étant

$$D(\boldsymbol{Y},\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} 2a(\phi)\{\ell(\boldsymbol{Y}|\boldsymbol{Y};\sigma) - \ell(\hat{\boldsymbol{\mu}}|\boldsymbol{Y};\sigma)\}.$$

Puisque la loi normale est ici à l'étude,

$$\begin{aligned}
D(\boldsymbol{Y}, \hat{\boldsymbol{\mu}}) &= 2\sigma^2 \sum_{i=1}^{n} \left\{ \frac{Y_i Y_i - Y_i^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{Y_i^2}{\sigma^2} + \ln(2\pi\sigma^2) \right) - \frac{Y_i\hat{\mu}_i - \hat{\mu}_i^2/2}{\sigma^2} + \frac{1}{2}\left(\frac{Y_i^2}{\sigma^2} + \ln(2\pi\sigma^2) \right) \right\} \\
&= 2\sigma^2 \sum_{i=1}^{n} \left\{ \frac{Y_i^2 - Y_i^2/2}{\sigma^2} - \frac{Y_i\hat{\mu}_i - \hat{\mu}_i^2/2}{\sigma^2} \right\} \\
&= 2\sigma^2 \sum_{i=1}^{n} \left\{ \frac{Y_i^2 - 2Y_i\hat{\mu}_i + \hat{\mu}_i^2}{2\sigma^2} \right\} \\
&= \sum_{i=1}^{n} (Y_i - \hat{\mu}_i)^2.
\end{aligned}$$

□

Annexe B

Des compléments SAS

B.1 La syntaxe de la procédure GENMOD

PROC GENMOD[1] <options> ;
BY variables ;
CLASS variables ;
CONTRAST 'nom_donné_au_contraste' effet valeur<...effet valeur></options> ;
DEVIANCE variable=expression
ESTIMATE 'nom_donné' effet valeur<...effet valeur></options> ;
FWDLINK variable=expression ;
INVLINK variable=expression ;
MODEL dependante=independante(s)</options> ;
OUTPUT OUT=nom MOT-CLE=nom ;
REPEATED SUBJECT=effet</options> ;
VARIANCE variable=expression ;
RUN ;

• **PROC GENMOD** Cet énoncé est nécessaire. Il appelle la procédure utilisée. Certaines options sont disponibles. La principale est DATA=. À défaut de mentionner cette options, SAS utilise la dernière base de données ayant été créée.
• **BY** Cet énoncé produit une analyse séparée pour chaque observation des variables

[1] Cette section n'est qu'un aperçu de la procédure GENMOD. Pour plus d'informations, voir http://www.mat.ulaval.ca/sasdoc/saspdf/stat/chap29.pdf

mentionnées. Lorsque cet énoncé est utilisé, la base de données doit être triée (procédure SORT) selon les variables.

• **CLASS** Spécifie les variables de classification. Cet énoncé doit obligatoirement être avant l'énoncé MODEL.

• **CONTRAST** Sert à tester l'hypothèse $\boldsymbol{L\beta = 0}$. Prenons l'exemple d'une régression faisant intervenir les variables x_1, x_2, x_3, x_4 et x_5, et admettons que l'on veuille tester les hypothèses $3\beta_2 + 4\beta_5 = 0$ et $\beta_0 + 6\beta_4 = 0$. Ainsi,

$$\boldsymbol{L} = \begin{bmatrix} 0 & 0 & 3 & 0 & 0 & 4 \\ 1 & 0 & 0 & 0 & 6 & 0 \end{bmatrix} \text{ et } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}.$$

La syntaxe à utiliser ici est :

CONTRAST 'exemple' x2 3 x5 4, INTERCEPT 1 x4 6 ;

Certaines options sont disponibles. L'une d'elle est E, qui imprime la matrice $\boldsymbol{L}$.

• **DEVIANCE** À l'énoncé MODEL, l'une des options est DIST=. Cependant, d'autres distributions que celles disponibles peuvent être utilisées. Dans ce cas, on se doit de mentionner la fonction de déviance associée à cette distribution. Cet énoncé n'est pas nécessaire si l'on utilise DIST=.

• **ESTIMATE** Très similaire à CONTRAST, à l'exception qu'une seule ligne est disponible dans la matrice $\boldsymbol{L}$. L'option ALPHA= change la valeur de 5% par défaut, E imprime la matrice $\boldsymbol{L}$ et EXP calcule une valeur pour $e^{\boldsymbol{L\beta}}$, en plus de l'erreur standard qui y est associée et un intervalle de confiance pour cette valeur.

• **FWDLINK** et **INVLINK** Spécifient une nouvelle fonction de lien et son inverse, respectivement. Plusieurs fonctions de lien sont disponibles à l'énoncé MODEL, mais d'autres peuvent être définies.

• **MODEL** Indique le modèle à l'étude. Plusieurs options sont disponibles. Les principales sont :

ALPHA= Change la valeur de 5% utilisée par défaut dans les analyses.

DIST= Indique la distribution de la variable réponse.

LINK= Indique la fonction de lien.

MAXITER= Change le nombre maximal d'itérations. Par défaut, ce nombre est fixé à 50.

NOINT Indique que l'ordonnée à l'origine ne doit pas être incluse dans le modèle.

OFFSET= Spécifie le logarithme d'une variable à inclure comme terme d'*offset*. Cette variable ne doit pas apparaître dans l'énoncé CLASS et ne doit pas non plus être la variable réponse ou l'une des variables exogènes.

OBSTATS Imprime, pour chacune des observations, les items suivants :

- Les valeurs de la variable réponse
- Les valeurs des variables exogènes
- Les valeurs prédites par le modèle, $\hat{\boldsymbol{\mu}} = g^{-1}(\boldsymbol{\eta})$
- L'estimation du prédicteur linéaire $\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$
- L'erreur standard de $\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$
- Un intervalle de confiance pour $\hat{\boldsymbol{\mu}}$
- Les résidus $Y_i - \hat{\mu}_i$
- Les résidus de Pearson
- Les résidus de Pearson standardisés
- Les résidus de déviance
- Les résidus de déviance standardisés
- Les résidus de vraisemblance

SCALE=DEVIANCE Peut être utilisé afin de contourner le problème de la surdispersion.

TYPE1 Implique une analyse de Type I. Cette option n'est pas possible pour un GEE (Chapitre 3), car l'analyse de Type I nécessite l'obtention d'une log-vraisemblance, ce qui n'est pas possible pour un GEE.

TYPE3 Demande les statistiques de Type 3 pour chacune des variables du modèle.

• **OUTPUT** Donne un nom à la base de données produite en sortie. Certains des mots-clé disponibles sont donc les suivants :

PREDICTED Afin d'avoir les valeurs prédites.

RESCHI Afin d'avoir les résidus de Pearson.

RESDEV Afin d'avoir les résidus de déviance.

RESLIK Afin d'avoir les résidus de vraisemblance.

STDXBETA Afin d'avoir l'erreur standard de XBETA, qui sera expliquée plus bas.

STDRESCHI Afin d'avoir les résidus de Pearson standardisés.

STDRESDEV Afin d'avoir les résidus de déviance standardisés.

XBETA Afin d'obtenir un estimé du prédicteur linéaire.

• **REPEATED SUBJECT=** Sert à spécifier une structure de covariance pour les réponses multivariées pour un GEE. SUBJECT= est obligatoire et sert à mentionner quel est le sujet dans la base de données. Plusieurs options sont disponibles, mais mentionnons que la principale est TYPE= qui sert à spécifier la structure de corrélation

utilisée afin de modéliser la corrélation entre les réponses. Voir section 3.5 pour les types possibles.
• **VARIANCE** Sert à spécifier une distribution de probabilité autre que celles disponibles à l'énoncé MODEL. Cet énoncé doit apparaître avec l'énoncé DEVIANCE.

B.2 La macro SAS pour la validation croisée lorsque la variable endogène en est une de Poisson

```
%macro validationgenmod(taille,jeu,class,dep,expression,dist,lien,offset,type) ;
ods results off ;
options nonotes ;

proc datasets ;
delete final ;
%do j=1 %to &taille ;
    data analyse ;
    set &jeu ;
    if obs=&j then &dep=. ;
    run ;

    %put "Prediction de l'observation &j" ;
    ods exclude ModelInfo ClassLevels ParmInfo ModelFit ConvergenceStatus Parame-
terEstimates GEEModInfo ConvergenceStatus GEEEmpPEst Type3 ObStats ;
    proc genmod data=analyse ;
    class &class ;
    model &dep=&expression/dist=&dist link=&lien offset=&offset ;
    repeated subject=&class/type=&type ;
    output out=residus pred=Valpredite ;
    run ;

    data stat2 ;
    merge residus &jeu ;
    by obs ;
    if obs=&j ;
    run ;

    proc append data=stat2 base=final ;
    run ;
```

```
%end;
options notes;
ods results on;

    data press2;
    set final;
    res2=(&dep-Valpredite)**2;
    run;

    proc means data=press2 noprint;
    var res2;
    output out=press sum=PRESS;
    run;
    proc print;
    run;
%mend;
```

B.3 La macro GLIMMIX de SAS

La syntaxe générale de la macro est la suivante :

```
%GLIMMIX(DATA=,
    PROCOPT=,
    STMTS=%STR(),
    WEIGHT=,
    FREQ=,
    ERROR=,
    ERRVAR=,
    ERRDEV=,
    LINK=,
    LINKN=,
    LINKND=,
    LINKNI=,
    LINKU=,
    LINKUD=,
    LINKUI=,
    LINKUID=,
    NUMBER=,
```

CF=,
CONVERGE=,
MAXIT=,
OFFSET=,
OUT=,
OUTALPHA=,
OPTIONS=
);

où certains des énoncés sont :
• **DATA** spécifie le jeu de données à l'étude.
• **PROCOPT** contient les options de l'énoncé MIXED de la procédure du même nom.
• **STMTS=%STR()** contient toute la syntaxe de la procédure MIXED[2]. On peut y retrouver les énoncés (et toutes leurs options) CLASS, ID, MODEL, RANDOM, REPEATED, PARMS, CONTRAST, ESTIMATE et LSMEANS. Comme avec la procédure, les énoncés sont séparés par un point-virgule.
• **ERROR** spécifie la distribution de la variable réponse. C'est donc l'équivalent de l'énoncé DIST= de la procédure GENMOD (voir section B.1 de la page 129). On peut donc mentionner les distributions BINOMIAL, NORMAL, POISSON, GAMMA, INVERSE GAUSSIAN ou USER ;
• **ERRVAR** spécifie la fonction de variance dans le cas où ERROR=USER est utilisée ;
• **ERRDEV** spécifie la fonction de déviance dans le cas où ERROR=USER est utilisée ;
• **LINK** mentionne la fonction de lien à utiliser ;
• **LINKN** mentionne une fonction de lien non linéaire ;
• **LINKND** spécifie la dérivée de la fonction de lien non linéaire ;
• **LINKNI** spécifie les valeurs initiales de la fonction de lien non linéaire ;
• **LINKU** spécifie une fonction de lien établie par l'utilisateur ;
• **LINKUD** mentionne la dérivée de la fonction de lien inscrite à l'option LINKUD ;
• **LINKUI** spécifie l'inverse de la fonction de lien inscrite à l'option LINKUD ;
• **LINKUID** mentionne la dérivée de la fonction de lien inscrite à l'option LINKUD ;
• **NUMBER** mentionne la tolérance utilisée afin de dériver les fonctions de lien comme probit et power. Par défaut, cette tolérance est fixée à 0.00001 ;
• **CF** spécifie la correction à apporter aux données afin de calculer un $\hat{\boldsymbol{\mu}}$ initial. Par défaut, cette valeur est de 0.5. Voir étape 1 à la section 5.4 de la page 50.
• **CONVERGE** mentionne le seuil utilisé afin de décider d'arrêter les itérations de la procédure décrite à la section 5.4 de la page 50. Par défaut, cette valeur est fixée à 0.00000001 ;

[2]Pour plus d'informations sur la procédure MIXED, voir
http://www.mat.ulaval.ca/sasdoc/saspdf/stat/chap41.pdf

• **MAXIT** spécifie le nombre maximal d'itérations à faire afin d'estimer les paramètres du GLMM. Par défaut, 20 itérations seront faites ;
• **OFFSET** spécifie la variable *offset*, comme dans la procédure GENMOD ;
• **OUT** crée un jeu de données dans lequel on trouve les valeurs finales de $\boldsymbol{\eta}$, son erreur standard, une borne supérieure et inférieure de ce $\boldsymbol{\eta}$, la valeur de $\boldsymbol{\mu}$, sa dérivée par rapport à $\boldsymbol{\eta}$, son erreur standard, une borne supérieure et inférieure de $\boldsymbol{\mu}$, la variance, les résidus $(Y_i - \hat{\mu}_i)$, les résidus de Pearson, la dérivée de $\boldsymbol{\eta}$ par rapport à $\boldsymbol{\mu}$ et les valeurs de la variable de poids et de la variable dépendante lors du dernier PROC MIXED ;
• **OUTALPHA** est une valeur de α pour calculer les intervalles de confiance de l'option OUT ;
• **OPTIONS** contient plusieurs options...

OPTIONS=MQL calcule les estimateurs par la méthode MQL. Par défaut, la méthode PQL est utilisée.

OPTIONS=NOPRINT empêche toute impression des résultats.

OPTIONS=NOITPRINT empêche toute impression des résultats concernant les différentes itérations.

OPTIONS=NOTES imprime la date et les numéros de page pendant l'exécution de la macro.

OPTIONS=PRINTALL imprime tous les résultats de chaque utilisation de la procédure MIXED.

OPTIONS=PRINTDATA imprime la pseudo variable après chacune des itérations.