Analyse du jeu de données de comptage de crises épileptiques traité initialement par Thall et Vail (1990)

Tamara Cannels

Analyse du Jeu de Données de Comptage de Crises Épileptiques Traité Initialement par Thall et Vail (1990)

CANNELS Tamara
tamara.cannels@etu.unistra.fr
Université de Strasbourg
UFR de Mathématiques et d'Informatique
Master 1 Statistique

26 août 2013

Résumé

Dans ce rapport, on se consacre essentiellement au traitement du jeu de données sur le comptage de crises épileptiques initialement analysé par Thall et Vail (1990).

Après avoir décrit le lancement dans le projet sous chapitre 2, on s'intéresse dans chapitre 3 à la partie théorique sur laquelle les analyses sous chapitre 4 se basent. On va ainsi traiter les modèles utilisées par Breslow et Clayton (1993) : une régression de Poisson naïve, un modèle linéaire généralisé à un effet aléatoire et un modèle linéaire généralisé à deux effets aléatoires. La réponse est le nombre de crises épileptiques compté sur une période de deux semaines et ceci quatre fois pour chacun des 59 patients. L'estimation se fait par maximum de quasi-vraisemblance pénalisée. Ensuite, on traite l'article de Breslow (1996) qui a utilisé une régression de Poisson et une binomiale négative avec la méthode d'estimation des moindres carrées re-pondérés itérativement. La réponse est la somme des quatre comptages de crises épileptiques par patient. Finalement, on refait les calculs de Sinha et Xu (2011) qui ont appliqué une régression logistique à cause du choix binaire de la réponse et des prédicteurs. Au chapitre 5, on se jette dans une discussion sur l'efficacité des différents modèles qu'on a présentés. Vu le nombre énorme d'articles citant Thall et Vail (1990), ils existent beaucoup plus de méthodes pour analyser le jeu de données sur les crises épileptiques qu'on ne peut quand même pas traiter dans ce rapport. Ainsi au chapitre 6, on évoque quelques points à ne pas négliger ou à ne pas oublier lors du traitement de ce jeu de données. Finalement, au chapitre 7 et au chapitre 8 se trouvent respectivement la conclusion finale et la liste de mes références.

Mots-clés : régression de Poisson, quasi-Poisson, binomiale négative, modèles linéaires généralisés, modèles linéaires généralisés à effet(s) mixte(s), données longitudinales

Remerciements

À cette occasion je voulais remercier mon maître de stage Prof. Stephen Senn de m'avoir offert la possibilité de faire un stage au sein de son centre ainsi que pour ses remarques constructives et sa patience infinie.

Je remercie également tout l'équipe du CCMS du CRP-Santé pour leur soutien.

Un grand merci aussi à Madame Ségolen Geffray pour son engagement et temps consacré à me donner une introduction aux modèles linéaires généralisés et aux modèles linéaires généralisés mixtes.

Table des matières

Chapitre 1

Institution d'accueil

1.1. *Le Centre de Recherche Public de la Santé*

Le Centre de Recherche Public de la Santé, court CRP-Santé, est situé au Luxembourg et est l'organisation publique leader dans le domaine de la recherche clinique en santé publique du pays. Fondée en 1988, l'organisation est constamment à la recherche d'acquisition de nouvelles connaissances dans le domaine de la santé publique et possède de vastes départements de recherche. On peut y nommer par exemple le département d'oncologie, virologie ou encore celui des maladies cardiovasculaires.

Le CRP-Santé a de nombreuses collaborations avec d'autres institutions de recherches nationales et internationales comme par exemple des hôpitaux ou de l'« Integrated BioBank of Luxembourg » (IBBL) ainsi que du « Translational Genomics Research Institute » (TGen) en Phoenix, Arizona.

Le CRP-Santé comptait environ 250 employés dont 200 scientifiques en 2012. Durant cette année plus que 154 projets de recherche ont été conduits et plus de 107 articles ont été publiés.

Pour plus d'informations le lecteur est invité à visiter le site internet du Centre de Recherche Public de la Santé (Luxembourg) : http://crp-sante.lu/

1.2. *Competence Center for Methodology and Statistics*

Mon stage de 12 semaines au CRP-Santé s'est déroulé dans le « Competence Center for Methodology and Statistics » (CCMS) qui a été créé en mars 2010. Ce service a comme mission de satisfaire les besoins de connaissances dans le domaine statistique et méthodologique dans les différents départements de recherche du CRP-Santé. Des projets externes de l'organisation pour des organismes nationaux ou internationaux dans le domaine de la santé publique sont également entretenus. L'importance de ce service repose essentiellement sur l'évaluation de données d'études menées par les scientifiques du CRP-Santé et la production de rapports statistiques de qualité.

Chef de service et tuteur de mon stage est Prof. Senn, ancien Professeur de statistiques à l'University of Glasgow, Professeur de statistiques dans le domaine pharmaceutique et de la santé à l'University College London, statisticien du National Health Service en Angleterre et de l'industrie pharmaceutique suisse.

Chapitre 2
Premiers pas

2.1. *Recherche bibliographique*

Pour la recherche bibliographique, le CRP m'a donné accès à la plus grande librairie digitale du Luxembourg ; findit.lu. En novembre 2012, on y comptait environ 50 000 titres de journaux électroniques, 80 000 eBooks électroniques et plus que 350 bases de données spécialisées et plateformes de recherche. Ainsi j'ai pu recourir à une autre base de données Scopus-SciVerse où j'ai trouvé la majorité des articles citant Thall et Vail (1990). La recherche a été complétée par des quêtes sur Google Scholar et findit.lu. Toutes ces références d'articles ont été importées dans EndNote X5. Il s'agit d'un logiciel de gestion de références bibliographiques qui facilite la citation d'articles et ouvrages et permet à l'utilisateur de construire des bases de données personnalisées.

À la fin des importations, j'avais cumulé 236 références d'articles scientifiques dans EndNote (sans doublées) qu'il s'agissait d'examiner afin de pouvoir décider si elles sont utiles pour mon projet ou non. Evoquons ici qu'on a exclu les livres pour raison de sursaturation. Lors de ces lectures, j'ai remarqué qu'il existe effectivement un nombre énorme de différentes méthodes pour analyser ces données. La majorité des articles se concentre sur les différentes méthodes d'estimations de paramètres tandis que d'autres proposent différents modèles pour analyser ces données. En ce qui concerne la sélection d'articles, j'avais des idées précises. Je n'ai par exemple pas retenu les textes où les auteurs ne mentionnent que l'article de Thall et Vail (1990) car il leur a permis de traiter leurs données de façon similaire. Evidemment, les articles citant Thall et Vail (1990) sans traiter les données n'étaient pas jugés utiles.

En ayant fait ce tri (j'en ai retenu 62) et en lisant entièrement quelques articles, j'ai remarqué que presque tous les articles n'ont pas justifié entièrement le choix de leur modèle. Je précise qu'en général les auteurs justifient le choix du modèle généralisé mais dès qu'il s'agit de manipuler des jeux de données réelles, ils n'expliquent plus pourquoi ils ont utilisé exactement ces transformations des prédicteurs et non pas d'autres. Effectivement, en essayant de reproduire par exemple les calculs du texte original de Thall et Vail (1990), j'ai rencontré mes limites. Ils proposent différents modèles avec des structures de matrice de variance-covariance différentes et à la fin ils affichent le tableau avec les estimations des paramètres ainsi que leurs erreurs-standards. Je n'ai pas trouvé comment pouvoir programmer une telle matrice en R ou en SAS. On va en revenir à cette problématique sous 4.2.. Peut-être le problème de la reproduction de leurs calculs est également la raison pourquoi la majorité des scientifiques a travaillé avec des modèles proposés par Breslow et Clayton (1993) et précisent uniquement que ces derniers ont également utilisé quelques modèles de Thall et Vail (1990).

2.2. Le jeu de données

Le jeu de données décrit dans l'article de Thall et Vail (1990) provient d'une étude de Leppik et al. (1985) sur l'effet du médicament Progabide contre des crises partielles d'épilepsie. Je n'ai pas trouvé le texte référencié dans l'article de Thall et Vail (1990) :

Leppik, I. E., et al. (1985). A double-blind crossover evaluation of progabide in partial seizures. *Neurology* **35,** 285.

Pourtant on peut supposer que l'article original est :
Leppik, I. E., F. E. Dreifuss, et al. (1987). "A controlled study of Progabide in partial seizures : methodology and results." Neurology 37(6): 963-968.

D'après l'article de Thall et Vail (1990), il s'agissait d'un essai à double insu avec permutation. Les patients étaient divisés en deux groupes ; un groupe recevant le médicament Progabide et un groupe placebo. De plus, cette étude a été faite parallèlement au traitement standard de chimiothérapie. Par ailleurs, on a récupéré des données d'une période de 8 semaines avant la prise du médicament/placebo. Cet instant de collecte de données va-t-on appeler « visite zéro ». On y récupère le comptage de référence (eng. *baseline count*). Les comptages des crises épileptiques ont été enregistrés tous les deux semaines durant 8 semaines où les patients ont pris le médicament/placebo. Ces données constituent les comptages avant la permutation.

Passons aux données décrites dans l'article de Thall et Vail (1990) (jeu de données affiché sous Annexe I. 31 patients ont reçu le médicament Progabide et 28 patients ont reçu un placebo. Ayant accès à un jeu de données d'une étude clinique à double insu avec permutation, Thall et Vail ont néanmoins qu'utilisé un jeu de données restreint pour faire leurs analyses. Ils se sont contentés alors avec les comptages de la période avant tout traitement médicamenteux et ceux de la période avant la permutation.

En analysant ce jeu de données, on remarque que les identifiants des patients ne commencent pas à 1. Effectivement, en affichant les données dans R `epilepsy` à l'aide du package `{robustbase}`, j'obtiens entre autre la variable `ID` qui indique l'identifiant de chaque patient. En triant ce jeu de données par ordre croissant des `ID`, on voit que les identifiants commencent à 101, s'arrêtent à 147 et recommencent à 201 et vont jusqu'à 238. Pourquoi a-t-on des patients où l'identifiant commence avec 1 et d'autres avec 2 ? Il se peut que ces données aient été récupérées dans deux centres médicaux différents. Le fait d'avoir des lieux de suivis médicaux différents n'est mentionné nul part dans Thall et Vail (1990). Par conséquent, on n'a jamais introduit un tel facteur dans un modèle statistique. En fait, l'article ne nous fournit aucune information sur les lieux où les données ont été récupérées. De plus, on remarque que les numérotations des `ID` ne se succèdent pas. On passe par exemple directement de 118 à 121 ou de 211 à 213. Ce qui nous saute également dans les yeux est que le nombre des patients (59) n'est pas très rond. Ces constatations ainsi que le fait qu'on n'a pas de données incomplètes, nous fait supposer que les données manquantes ont carrément été supprimées. Ceci expliquerait la numérotation bizarre des patients. Effectivement, les données manquantes sont fréquentes dans des études cliniques comme il y a toujours des patients qui oublient de se rendre au rendez-vous de suivi-médicale ou qui abandonnent l'étude à un moment donné pour des raisons différentes importantes à connaître. Souvent, elles compliquent le travail du

statisticien. Le réflexe de les supprimer est quand même faux car ces données censurées doivent être considérées durant l'analyse.

En considérant les modèles utilisés pour ce jeu de données, je remarque que presque tous les scientifiques ont utilisé un modèle respectant la hiérarchie des données et ont appliqué une réponse à deux indices. Ils mettent ainsi en évidence que le nombre de crises épileptiques a été répertorié dans un intervalle de deux semaines. En fait, l'utilité d'une telle réponse peut être mise en question. Effectivement, il n'existe pas de vraie raison pourquoi on a pris des intervalles de deux semaines et non pas d'une semaine ou peut-être d'un mois. Le but principal de l'étude est de déterminer si le médicament antiépileptique Progabide a un effet réducteur sur le nombre de crises épileptiques ou non. On ne s'intéresse pas au moment où le médicament commence à agir ou au meilleur moment pour faire un autre rendez-vous de suivi. Néanmoins, une telle manière de collecte de données nous fournit beaucoup plus d'information que la récupération du nombre de crises avant et après la prise d'un médicament. Les données prises à plusieurs points au long du traitement est également plus utile si on s'intéresse à l'évolution ou à l'effet du médicament au cours du temps. Ici, en gardant en tête qu'on s'intéresse juste à l'effet du traitement sur les crises épileptiques, une méthode qui semble quand même plus logique serait de prendre la somme des crises épileptiques qui ont eu lieu pendant ces 8 semaines de traitement avec le médicament/placebo pour chaque patient. On suppose que les patients sont indépendants entre eux. Ainsi en sommant le nombre de crises épileptiques, on reçoit des réponses qui sont indépendantes entre elles. Par contre, les réponses ne sont pas entièrement indépendantes si on considère un modèle hiérarchique car on compte le nombre de crises épileptiques de chaque patient quatre fois sur une durée de huit semaines à intervalle constant de deux semaines. Ces quatre mesures pour un patient précis sont dépendantes entre elles comme elles proviennent du même patient. Ces prises de mesures répétées au long d'une période de temps constituent des données longitudinales de comptage.

Des raisons pourquoi presque aucun auteur n'a choisi la somme des crises épileptiques sur la période totale de 8 semaines comme réponse revient peut être du fait qu'ils se sont basés entièrement sur le modèle de Thall et Vail (1990). Souvent, le but des auteurs est d'illustrer une méthode d'estimation ou d'obtention de meilleurs résidus qui semblent adapter aux données longitudinales de comptage. En effet, ce jeu de données paraît contenir plus de difficultés venant de la corrélation entre les réponses et assure un nombre plus élevé de données comparé à la somme des crises.

Chapitre 3

Méthodes d'analyse de données longitudinales

3.1. Rappel sur les modèles linéaires gaussiens

Posons le modèle à n observations et p prédicteurs comme suit :

$$Y_i = \beta_0 + \beta_1 X_i^{(1)} + \cdots + \beta_p X_i^{(p)} + \varepsilon_i$$

avec les ε_i étant indépendants et identiquement distribués (i.i.d.) suivant $\mathcal{N}(0, \sigma^2)$ et ceci pour tout $i = 1, \dots, n$.

On peut écrire cette formule également sous forme d'un produit scalaire :

$Y_i = \mathbb{X}_i\beta + \varepsilon_i$ avec les ε_i i.i.d. suivant $\mathcal{N}(0, \sigma^2)$, $\mathbb{X}_i = \left(1, X_i^{(1)}, \dots, X_i^{(p)}\right)$, $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$ et ceci pour tout $i = 1, \dots, n$. De plus, la réponse doit être indépendante des prédicteurs.

3.2. Famille exponentielle

Y de densité f (par rapport à la mesure de Lebesgue pour les Y à loi continue ou par rapport à la mesure de comptage pour les Y à loi discrète) appartient à la famille exponentielle si $f(y, \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$. θ est le paramètre d'intérêt et ϕ un paramètre de nuisance dit de dispersion. Les fonctions a, b et c sont à spécifier, mais la fonction a est généralement de la forme $a(\phi) = \frac{\phi}{w}$ avec w un vecteur de poids des différentes observations. De cette écriture résultent des résultats intéressants :

$$\mathbb{E}[Y] = \mu = b'(\theta) \text{ et } \mathbb{V}[Y] = a(\phi)\, b''(\theta)$$

On appelle fonction de variance $V(\mu) = b''(\theta) = b''(b'^{-1}(\mu))$. Il s'agit d'une fonction qui nous donne la relation entre l'espérance et la variance d'une loi.

On définit également la fonction de lien canonique g à l'aide de l'écriture exponentielle :

$$g(\mu) = \theta = b'^{-1}(\mu)$$

3.3. La loi de Poisson

Une loi de Poisson étant une loi discrète et ne prenant que des valeurs positives convient bien pour des comptages. Effectivement, elle est définie sur l'ensemble des entiers naturels et ne prend donc jamais de valeurs négatives ou décimales.

La loi de Poisson va encore jouer un rôle important, notamment parce qu'elle est une loi qui appartient à la famille exponentielle. Voici la preuve où « ln(.)» est le logarithme népérien :

Pour la loi de Poisson $\mathcal{P}(\mu)$ on a : $f(y,\mu) = \frac{\mu^y}{y!}\exp(-\mu) = \exp(-\mu + y\,ln(\mu) - ln(y!))\ \ (\mu > 0)$
Ainsi on pose : $\theta = \ln(\mu)$, $a(\phi) = 1$, $b(\theta) = \exp(\theta) = \mu$, $c(y,\phi) = -\ln(y!)$
On peut donc définir sa fonction de variance V et sa fonction de lien canonique g :

$$V(\mu) = \mu \text{ et } g(\mu) = \ln(\mu)$$

La fonction log népérien est inversible. Une autre caractéristique importante de la loi de Poisson est l'égalité de son espérance et sa variance. Si $Y \sim \mathcal{P}(\mu)$, on a $\mathbb{E}[Y] = \mu = \mathbb{V}[Y]$.

3.4. La loi de Bernoulli

Si $Y \sim \mathcal{B}(1,p)$, alors on dit que Y suit une loi de Bernoulli de paramètres p. Il s'agit d'une loi discrète qui est idéal pour des données binaires comme Y n'y peut prendre que la valeur 0 ou 1. Cette loi appartient à la famille exponentielle :

Pour $Y \sim \mathcal{B}(1,p)$: $f(y,1,p) = p^y(1-p)^{1-y}$

$$= \exp\left(y \ln\left(\frac{p}{1-p}\right) + \ln(1-p)\right)$$

Ainsi on pose :

$$\theta = \ln\left(\frac{p}{1-p}\right) = logit(p) \Leftrightarrow p = \frac{e^\theta}{1+e^\theta}, \qquad a(\phi) = 1,$$
$$b(\theta) = -\ln(1-p) = \ln\left(1+e^\theta\right), \qquad c(y,\phi) = 0$$

On en déduit la fonction de variance V et la fonction de lien canonique g de la loi binomiale où $\mathbb{E}[Y] = p = \mu$:

$$V(\mu) = b''(\theta) = \mu(1-\mu) \text{ et } g(\mu) = \theta = \ln\left(\frac{\mu}{1-\mu}\right) = logit(\mu)$$

La fonction logit est inversible.

3.5. Recours aux modèles linéaires généralisés

Parfois, les modèles linéaires gaussiens (LNM) ne suffisent plus pour modéliser la réalité correctement comme on suppose la loi de la réponse comme étant continue et souvent gaussienne. Une telle loi n'est pas adaptée au cas où la réponse est par exemple un comptage ou une réponse binaire. Effectivement, ces deux derniers prennent des valeurs positives entières. De plus, contrairement à une loi gaussienne, la distribution de la réponse dans le cas de comptage ou binaire n'est pas forcément symétrique. Ainsi, on introduit les modèles linéaires généralisés (GLM) pour se charger de ces situations. L'extension des modèles

linéaires consiste en admettant à la réponse Y de prendre une loi de la famille exponentielle. Comme tous les modèles marginaux, les GLMs sont également conditionnés par rapport aux prédicteurs fixes.

On part donc des modèles linéaires gaussiens :

$$\begin{cases} \left(Y_i, X_i^{(1)}, \dots, X_i^{(p)}\right) \text{ sont indépendants pour i} = 1, \dots, \text{n} \\ \mathcal{L}(Y_i \mid \mathbb{X}_i) = \mathcal{N}\left(\mu_i, \sigma_i^2\right) \\ \sigma_i^2 = \sigma^2 \text{ pour i} = 1, \dots, \text{n (homoscédasticité)} \\ \mu_i = \mathbb{E}[Y_i \mid \mathbb{X}_i] = \mathbb{X}_i\beta \text{ (application linéaire en β)} \end{cases}$$

et étend cette théorie en obtenant les modèles linéaires généralisés :

$$\begin{cases} \left(Y_i, X_i^{(1)}, \dots, X_i^{(p)}\right) \text{ sont indépendants pour i} = 1, \dots, \text{n} \\ \mathcal{L}(Y_i \mid \mathbb{X}_i) \in \text{famille exponentielle} \\ \quad \text{(où } \mathcal{L}(Y_i \mid \mathbb{X}_i) \text{ représente la loi de } Y_i \text{ conditionnellement aux prédicteurs)} \\ \mu_i = \mathbb{E}[Y_i \mid \mathbb{X}_i] \\ g(\mu_i) = \eta_i = \mathbb{X}_i\beta \text{ (application linéaire en β)} \\ \quad \text{où g(.)est la fonction de lien (inversible)et } \eta_i \text{ est le prédicteur linéaire} \end{cases}$$

Pour la régression de Poisson, un GLM où $\mathcal{L}(Y_i \mid \mathbb{X}_i) = \mathcal{P}(\mu_i)$, on a alors $g(\mu) = \ln(\mu)$ et $V(\mu) = \mu$. De plus, avec ces résultats, on déduit pour l'espérance :

$$\ln(\mathbb{E}[Y_i \mid \mathbb{X}_i]) = \mathbb{X}_i\beta \Leftrightarrow \mathbb{E}[Y_i \mid \mathbb{X}_i] = \mathbb{V}[Y_i \mid \mathbb{X}_i] = \mu_i = \exp(\mathbb{X}_i\beta)$$

Le modèle ne contient pas de terme d'erreur ε_i à cause de la relation entre l'espérance et la variance. Effectivement, une fois l'espérance connue, on a spécifié entièrement le modèle.

En choisissant une loi de Bernoulli, on obtient un GLM où $\mathcal{L}(Y_i \mid \mathbb{X}_i) = \mathcal{B}(1, \pi_i)$ et $\mathbb{E}[Y_i|\mathbb{X}_i] = \mathbb{P}(Y_i = 1|\mathbb{X}_i) = \pi_i$. Ainsi, la fonction de lien canonique vaut $g(\pi) = \text{logit}(\pi)$ et la fonction de variance vaut $V(\pi) = \pi(1 - \pi)$. Ici à nouveau le terme d'erreur devient superflu. Avec ces résultats, on peut calculer l'espérance conditionnellement aux prédicteurs :

$$logit(\mathbb{E}[Y_i \mid \mathbb{X}_i]) = \mathbb{X}_i\beta \Leftrightarrow \mathbb{E}[Y_i \mid \mathbb{X}_i] = \pi_i = \frac{\exp(\mathbb{X}_i\beta)}{1+\exp(\mathbb{X}_i\beta)}.$$

L'estimation des paramètres β n'est plus calculée à l'aide de la méthode des moindres carrées mais par maximum de vraisemblance (ML). Ce changement est dû au fait que pour les GLM, la maximisation de vraisemblance ne revient pas à minimiser la somme des carrées des écarts.

Ainsi pour la régression de Poisson avec $Y_i|\mathbb{X}_i \sim \mathcal{P}(\mu_i)$, la formule de vraisemlance L et de log-vraisemblace ℓ valent :

$$L(Y, \mu) = \prod_{i=1}^{n} \left(\exp(-\mu_i)\frac{\mu_i^{y_i}}{y_i!}\right)$$

$$\ell(Y, \mu) = \ln\big(L(Y, \mu)\big) = \sum_{i=1}^{n} (-\mu_i + y_i \ln(\mu_i) - \ln(y_i!))$$

En dérivant $\ell(Y, \mu)$ par β_j et en l'annulant après, on obtient l'estimateur de β_j, $\hat{\beta}_j$. Une simplification lors de l'estimation est basée sur le fait que maximiser $\ell(Y, \mu)$ revient à maximiser $\sum_{i=1}^{n}(-\mu_i + y_i \ln(\mu_i))$ car $\ln(y_i!)$ est constant.

Pour la loi de Bernoulli avec $Y_i|\mathbb{X}_i \sim \mathcal{B}(1,\pi_i)$, la formule de vraisemblance L et de log-vraisemblance ℓ valent :

$$L(Y,\pi) = \prod_{i=1}^{n}\left(\pi_i^{y_i}(1-\pi_i)^{1-y_i}\right)$$

$$\begin{aligned}\ell(Y,\pi) &= \ln\big(L(Y,\pi)\big)\\ &= \sum_{i=1}^{n}(y_i\ln(\pi_i) - (1-y_i)\ln(1-\pi_i))\\ &= \sum_{i=1}^{n}(y_i\mathbb{X}_i\beta - \ln(1+\exp(\mathbb{X}_i\beta)))\end{aligned}$$

L'estimation des paramètres d'intérêt se fait comme pour la régression de Poisson : on résout l'équation suivante : $\frac{d}{d\beta_j}\ell(Y,\pi) = 0$ pour chaque j=0, …, p.

Pour vérifier s'il s'agit vraiment d'un maximum, il faut qu'on a $-\frac{\partial^2}{\partial\beta_j^2}\ell\big(Y,\hat{\theta}_j\big) > 0$ pour j=0,…, p, si on admet que $\hat{\theta}$ est le vecteur des solutions de l'annulation de la dérivée de la log-vraisemblance.

La fonction `glm(.)` de R calcule les maxima de vraisemblance par la méthode itérative des moindres carrées re-pondérés (IRLS) (eng. *Iteratively reweighted least square method*) (voir Annexe V). Cette méthode est considérée comme étant robuste car elle n'est pas sensible aux outliers. En fait, elle donne des poids faibles aux observations susceptibles d'être atypiques en se basant sur la grande ampleur des résidus à l'étape précédente.

Dans R, pour tester si les prédicteurs ont un effet sur la réponse ou non, un test de Wald est utilisé en calculant les statistiques de Wald (z-valeurs) et leurs p-valeurs correspondantes. On teste alors si le coefficient β_j est significatif ou non avec $H_0 : \beta_j=0$ et $H_a : \beta_j\neq 0$. La statistique de Wald est $Z = \frac{\hat{\beta}_j-\beta_j}{\sigma_j/\sqrt{n}}$ où σ_j est l'écart-type connu de l'estimateur de β_j et sous H_0, $Z\sim\mathcal{N}(0,1)$. De cette façon, l'intervalle de confiance de β_j sera $\left[\hat{\beta}_j \pm z_{1-\alpha/2}\frac{\sigma_j}{\sqrt{n}}\right]$ pour tout j=0, …, p et où z est le $1-\frac{\alpha}{2}$ quantile d'une loi normale centrée réduite.

Dans certains cas, R nous produit aussi des t-valeurs obtenues par une statistique de Student, notamment quand le paramètre de dispersion a été inconnu et a dû être estimé. La statistique de test est alors la même à l'exception de l'écart-type qu'on doit estimer. Cette statistique suit alors une loi de Student à n-1 degrés de liberté. L'intervalle de confiance de β_j est :
$\left[\hat{\beta}_j \pm t_{n-1}^{1-\alpha/2}\frac{\hat{\sigma}_j}{\sqrt{n}}\right]$ où $t_{n-1}^{1-\alpha/2}$ est le $1-\frac{\alpha}{2}$ quantile de Student à n-1 degrés de liberté. On applique le test de Student également dans le cas où on a moins que 30 observations. Pourtant quand la taille de l'échantillon augmente, il n'y a que très peu de différence entre les deux tests.

3.6. Le phénomène de la sur-dispersion

On a déjà mentionné précédemment qu'une caractéristique importante de la loi de Poisson est l'égalité entre l'espérance et la variance. En effet, comme pour la loi binomiale et contrairement à la loi gaussienne, l'espérance et la variance sont liées.

Malheureusement, l'égalité de l'espérance et de la variance n'est pas toujours maintenue ; même pas approximativement. Posons Y la réponse de notre modèle. $Y \sim \mathcal{P}(\mu)$ suit une loi de Poisson de paramètre μ. On appelle sur-dispersion ou encore variabilité extra-poissonnienne si au lieu d'avoir $\mathbb{V}[Y] = \mathbb{E}[Y] = \mu$, nommé équi-dispersion, on observe $\mathbb{V}[Y] > \mu$. Par ailleurs, si on observe $\mathbb{V}[Y] < \mu$, on se trouve dans le cas sous-dispersé.

La sur-dispersion n'est pas un phénomène rare dans les comptages. Elle peut résulter par exemple des données non indépendantes, d'un excès de comptages zéro ou de l'absence d'une variable quand même importante à inclure dans le modèle. Sa détection est facile à partir du rapport entre la déviance résiduelle et son degré de liberté. Si celui-ci vaut environ 1, on est dans le cas équi-dispersé, s'il est par contre supérieur à 1, on se trouve dans le cas sur-dispersé.

Il existe également un test performant pour détecter la sur-dispersion, utilisable dans le cas où on a estimé nos paramètres par la méthode de maximum de vraisemblance. C'est le test de Dean (1992). On va expliquer son principe un peu plus tard dans ce rapport.

Pour prendre en compte la sur-dispersion des données, on introduit un paramètre de dispersion noté ϕ et un vecteur de poids a priori w (eng. *prior weights*) pour obtenir la relation suivante : $\mathbb{V}[Y] = \frac{\phi}{w}\mathbb{E}[Y] = \frac{\phi}{w}\mu$. Il s'agit d'un modèle basé sur la quasi-vraisemblance et dans notre cas poissonien, on parle de quasi-Poisson. Il en résulte que si $\phi > 1$, on a mise en évidence la sur-dispersion des données. Ainsi, on a généralisé le lien entre l'espérance et la variance car si on a $\phi = 1$, on se retrouve dans le modèle de Poisson habituelle. Dans la situation poissonienne, on assume également que les poids a priori sont tous 1. Un modèle basé sur la quasi-vraisemblance s'explique de la façon suivante :

$$\begin{cases} \left(Y_i, X_i^{(1)}, \dots, X_i^{(p)}\right) \text{ sont indépendants pour i} = 1, \dots, \text{n} \\ \mu_i = \mathbb{E}[Y_i \mid \mathbb{X}_i] \\ g(\mu_i) = \eta_i = \mathbb{X}_i\beta \text{ (application linéaire en β)} \\ \quad \text{où g(.)est la fonction de lien et } \eta_i \text{ est le prédicteur linéaire} \\ \mathbb{V}[Y_i|\mathbb{X}_i] = \frac{\phi}{w_i} V(\mu_i) \text{ où V(.) est la fonciton de variance} \end{cases}$$

Comme on obtient les estimateurs et leurs erreurs-types par des méthodes de quasi-vraisemblance et non pas purement vraisemblance, on ne peut pas appliquer les tests de comparaison de modèles se basant sur celle-ci (test de rapport de vraisemblance ; eng. *Likelihood-ratio test*).

Pour estimer le paramètre de dispersion, on peut se servir de la statistique de Pearson qui suit à peu près une loi de chi-deux $\mathcal{X}^2$ à n-p-1 degrés de libertés : $X^2 = \sum_{i=1}^{n} w_i \frac{(y_i - \hat{\mu}_i)^2}{V[\hat{\mu}_i]}$. La valeur du paramètre de dispersion est alors estimée par $\hat{\phi} = \frac{X^2}{n-p-1} = \hat{\sigma}^2$.

Pour obtenir le maximum de quasi-vraisemblance, on a besoin de la fonction de log-quasi-vraisemblance $Q(\mu; y) = \sum_{i=1}^{n} Q_i(\mu_i; y_i)$ où $Q_i(\mu_i; y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)} dt$ où $\phi V(\mu_i)$ est la variance de Y_i. Ceci revient alors dans le cas poissonien à :

$$Q_i(\mu_i; y_i) = \frac{1}{\phi} y_i \ln(\mu_i) - \mu_i + constante \text{ (Preuve : voir Annexe II).}$$

Une autre méthode pour traiter la sur-dispersion est de choisir un modèle respectivement une loi qui est mieux adaptée pour tenir compte de cette variation extra-poissonienne. On va la décrire par la suite.

3.7. Mélange de Poisson et Gamma : la loi binomiale négative

La loi binomiale négative est une loi discrète et peut être considérée comme une généralisation de la loi de Poisson. Plus précisément, on parle souvent d'une mixture de Poisson-Gamma. Effectivement, on peut voir cette loi comme une loi de Poisson avec son seul paramètre étant une variable aléatoire suivant une loi de Gamma.

$$Y|\Theta = \theta \sim \mathcal{P}(\theta) \text{ où } \Theta \sim \mathcal{G}a(\alpha,\beta) \text{ où } \alpha > 0,\ \beta > 0$$

On définie α comme étant le paramètre de forme et β le paramètre d'échelle. Ainsi la densité de la loi de Gamma est :

$$f(\theta;\alpha,\beta) = \frac{1}{\beta^{\alpha}}\frac{1}{\Gamma(\alpha)}\theta^{\alpha-1}\exp\left(-\frac{\theta}{\beta}\right)1\{\theta > 0\}$$

où Γ désigne ici la fonction gamma $\Gamma(x) = \int_0^1 t^{x-1}e^{-t}\,dt$.

Contrairement aux lois discrètes, la loi de gamma ainsi que la loi normale contiennent un paramètre de dispersion et donc en tout deux paramètres normalement capables de gérer la variation extra-poissonienne.

La densité jointe est définie par :

$$\mathbb{P}(Y = y|\Theta = \theta)f_{\Theta}(\theta) = \frac{\theta^{y}}{y!}\exp(-\theta)\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\theta^{\alpha-1}\exp\left(-\frac{\theta}{\beta}\right)1\{\theta > 0\}$$

Ainsi en intégrant cette densité par rapport à θ, on reçoit la probabilité pour Y qui vaut alors

$$\mathbb{P}(Y = y;\alpha,\beta) = \int_0^{\infty}\mathbb{P}(Y = y|\Theta = \theta)f_{\Theta}(\theta)d\theta = \frac{\Gamma(y+\alpha)}{\Gamma(\mathrm{y}+1)\Gamma(\alpha)}\left(1-\frac{1}{1+\beta}\right)^{y}\left(\frac{1}{1+\beta}\right)^{\alpha}$$

Les calculs détaillés se trouvent sous le point Annexe III. Si α n'est pas connue, ce qui est souvent le cas en pratique, cette loi n'appartient pas à la famille exponentielle et ne peut donc pas être utilisée dans un GLM. Par contre si α est connue, on a :

$$\mathbb{P}(Y = y;\ \alpha,\beta) = \frac{\Gamma(y+\alpha)}{\Gamma(\mathrm{y}+1)\Gamma(\alpha)}\left(1-\frac{1}{1+\beta}\right)^{y}\left(\frac{1}{1+\beta}\right)^{\alpha}$$

$$= exp\left(y\ln\left(\frac{\beta}{1+\beta}\right) - \alpha\ln(1+\beta) + \ln\left(\frac{\Gamma(\mathrm{y}+\alpha)}{\Gamma(\mathrm{y}+1)\Gamma(\alpha)}\right)\right)$$

Avec les notations usuelles : $\theta = \ln\left(\frac{\beta}{1+\beta}\right) = logit(\beta) \Leftrightarrow \beta = \frac{\exp(\theta)}{1-\exp(\theta)}$

$a(\phi) = 1\ ;\ b(\theta) = \alpha\ln(1+\beta) = -\alpha\ln((1-exp(\theta))\ \ et\ c(y,\phi) = \frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)}$

En ce qui concerne les espérances et variances, on obtient : $\mathbb{E}[Y] = \alpha\beta = \mu$ et $\mathbb{V}[Y] = \alpha\beta(1+\beta) = \mathbb{E}[Y](1+\beta)$. Les preuves se trouvent en Annexe IV.

Si on pose $\alpha = \frac{\lambda}{\beta}$, alors on obtient la première version (I) de la loi binomiale négative avec $\mathbb{E}[Y] = \lambda \; et \; \mathbb{V}[Y] = \lambda(1+\beta) = \mathbb{E}[Y](1+\beta)$ (variance nommée *scaled variance*).
Si on pose $\beta = \frac{\lambda}{\alpha}$, alors on obtient la seconde version (II) de la loi binomiale négative avec $\mathbb{E}[Y] = \lambda \; et \; \mathbb{V}[Y] = \lambda + \frac{\lambda^2}{\alpha} = \mathbb{E}[Y] + \frac{1}{\alpha}\mathbb{E}^2[Y]$ (variance nommée *quadratic variance*).
En régression, on a toujours :

$$\begin{cases} \left(Y_i, X_i^{(1)}, \dots, X_i^{(p)}\right) \text{ sont indépendants pour i} = 1, \dots, \text{n} \\ \mathcal{L}(Y_i \mid \mathbb{X}_i) = \mathcal{NB} \\ \mu_i = \mathbb{E}[Y_i \mid \mathbb{X}_i] \\ \ln(\mu_i) = \mathbb{X}_i\beta \text{ (application linéaire en β) où ln est la fonction de lien} \end{cases}$$

3.8. Modèles linéaires mixtes généralisées

Dépendant du jeu de données, il devient impossible de modéliser la réalité à l'aide des modèles linéaires ou modèles linéaires généralisés. Un tel cas est par exemple un jeu de données où on a des données corrélées. Ceci peut apparaître si on a des données en cluster ou des mesures répétées. Effectivement, dans cette dernière situation, les mesures prisent sur un individu à plusieurs reprises ne sont plus indépendantes.

On aime alors introduire les modèles de régression linéaire à effets mixtes, donc avec des effets fixes et des effets aléatoires.

Dans le cas de données longitudinales où Y_{ij} est la réponse du $i^{ième}$ sujet au $j^{ième}$ instant et où on a k=1, ..., p prédicteurs $X_{ij}^{(k)}$, on peut écrire modèles linéaire mixtes de la façon suivante :

$$Y_{ij} = \beta_0 + \beta_1 X_{ij}^{(1)} + \cdots + \beta_p X_{ij}^{(p)} + A_i + \varepsilon_{ij}$$

On note A_i l'effet aléatoire dû à l'individu. Cet effet va nous permettre de prendre en compte la variabilité entre les sujets. Les modèles mixtes prennent en compte la corrélation des observations.

En prenant les notations suivantes :

$$\mathbb{Y} = \begin{pmatrix} \mathbb{Y}_1 \\ \vdots \\ \mathbb{Y}_n \end{pmatrix} o\grave{u} \; \mathbb{Y}_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix} \text{ si on a n individus et } n_i \text{ répétitions}$$

$$\mathbb{X} = \begin{pmatrix} \mathbb{X}_1 \\ \vdots \\ \mathbb{X}_n \end{pmatrix} o\grave{u} \; \mathbb{X}_i = \begin{pmatrix} 1 & \mathbb{X}_{i1}^{(1)} & \dots & \mathbb{X}_{i1}^{(p)} \\ \vdots & \vdots & \dots & \vdots \\ 1 & \mathbb{X}_{in_i}^{(1)} & \dots & \mathbb{X}_{in_i}^{(p)} \end{pmatrix} \text{ (Matrice connue)}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \text{ (Vecteur des effets fixes)}$$

$$\mathbb{Z} = \begin{pmatrix} \mathbb{Z}_1 \\ \vdots \\ \mathbb{Z}_n \end{pmatrix} o\grave{u} \; \mathbb{Z}_i = \begin{pmatrix} 0 & & 0 & 1 & 0 & & 0 \\ \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & & 0 & 1 & 0 & & 0 \end{pmatrix} \text{ (les 1 sont à la } i^e \text{ colonne)}$$

$\mathbb{U} = \begin{pmatrix} A_1 \\ \vdots \\ A_n \end{pmatrix}$ (Vecteur des effets aléatoires)

$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ où $\varepsilon_i = \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{in_i} \end{pmatrix}$ (Vecteur de l'erreur résiduelle ; son présence dépend de la du choix de la régression (omis par exemple pour Poisson ou logistique))

Matriciellement, le modèle s'écrit de façon suivante :

$$\mathbb{Y} = \mathbb{X}\beta + \mathbb{Z}\mathbb{U} + \varepsilon$$

Par la suite on va admettre que le nombre de répétitions des mesures est pareil pour tous les individus, c'est-à-dire $\forall i = 1, \dots, n$, on a $n_i = k$ instant de mesures.

En générale, le modèle linéaire généralisé à effets mixtes s'écrit de la façon suivante :

$$\begin{cases} \left(Y_{ij}, X_{ij}^{(1)}, \dots, X_{ij}^{(p)}\right) \text{ sont indépendant pour i} = 1, \dots, \text{n et j} = 1, \dots, \text{k} \\ \qquad \text{conditionellement aux } \mathrm{A_i} \\ \mathcal{L}\left(Y_{ij} \mid X_{ij}^{(l)}, l = 1, \dots, p, A_i\right) \in \text{famille exponentielle} \\ \mu_{ij} = \mathbb{E}\left[Y_{ij} \mid X_{ij}^{(l)}, l = 1, \dots, p, A_i\right] \\ g(\mu_{ij}) = \eta_{ij} = \beta_0 + \beta_1 X_{ij}^{(1)} + \cdots + \beta_p X_{ij}^{(p)} + A_i \\ A_i \sim \mathcal{N}(0, \sigma_A^2) \text{ de façon i. i. d.} \end{cases}$$

Le cas qui va nous intéresser est celui qui s'applique aux données longitudinales. Il s'agit du modèle linéaire généralisée de Poisson à effets mixtes :

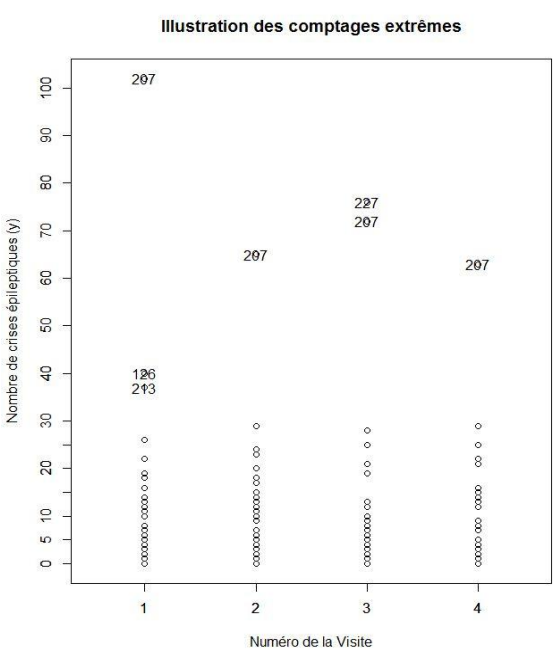
$$\begin{cases} \left(Y_{ij}, X_{ij}^{(1)}, \dots, X_{ij}^{(p)}\right) \text{ sont indépendant pour i} = 1, \dots, \text{n} \\ \qquad \text{et j} = 1, \dots, \text{k conditionellement aux } \mathrm{A_i} \\ \mathcal{L}\left(Y_{ij} \mid X_{ij}^{(l)}, l = 1, \dots, p, A_i\right) = \mathcal{P}(\mu_{ij}) \\ \mu_{ij} = \mathbb{E}\left[Y_{ij} \mid X_{ij}^{(l)}, l = 1, \dots, p, A_i\right] \\ \ln(\mu_{ij}) = \eta_{ij} = \beta_0 + \beta_1 X_{ij}^{(1)} + \cdots + \beta_p X_{ij}^{(p)} + A_i \text{ avec ln la fonction de lien} \\ A_i \sim \mathcal{N}(0, \sigma_A^2) \text{ de façon i. i. d.} \end{cases}$$

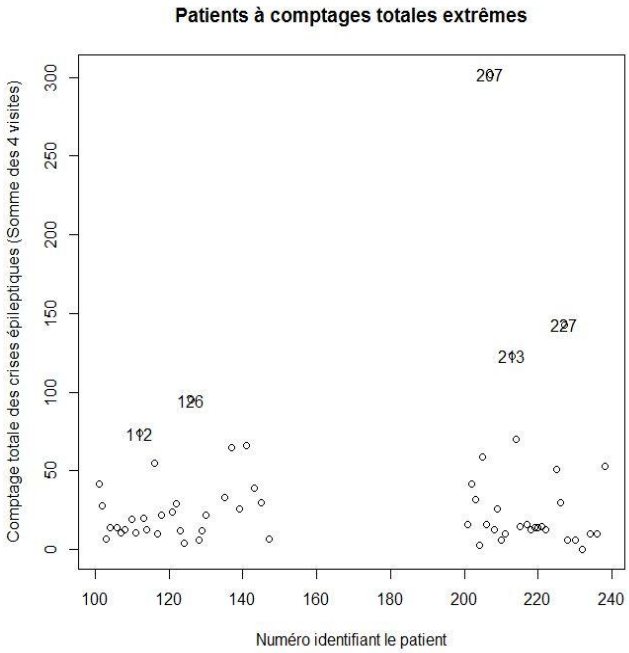
L'estimation des paramètres peut toujours se faire par maximum de vraisemblance, mais souvent ceci est difficilement réalisable à cause de l'introduction des effets aléatoires ce qui rend le modèle très complexe. De plus, la maximisation de la vraisemblance peut produire des estimateurs biaisés. Dans ce cas il vaut mieux d'utiliser la méthode de maximum de vraisemblance restreint (REML) (eng. *restricted maximum likelihood*) qui n'est quand même pas traitée dans ce rapport. La plus simple approche pour se charger de ces problèmes est quand même l'approximation au maximum de vraisemblance. On peut obtenir des approximations par la méthode de maximum de quasi-vraisemblance pénalisée, approximation de Laplace ou quadrature adaptive de Gauss-Hermite. Les méthodes sont énumérées de façon à ce qu'elles donnent des estimations de plus en plus précises. Les principes de ces méthodes d'approximations vont être expliqués brièvement sous Annexe V.

Chapitre 4
Traitement des données

4.1. *Premières observations*

Au premier coup d'œil, on remarque le patient à identifiant `ID` 207 qui possède un nombre élevé de crises épileptiques avant l'étude (période de référence), au long de l'étude et évidemment aussi en totale. Sur les graphiques d'en bas, ces constatations deviennent très visibles. On se demande alors si cette observation n'est pas trop influente et s'il faut envisager de l'exclure du jeu de données. Thall et Vail (1990) par exemple ont fait leurs calculs en l'excluant, mais Breslow et Clayton (1993) ont travaillé avec le jeu de données entier.





De plus, si on compare le nombre de crises épileptiques des 8 semaines précédant l'étude avec le nombre de celles des 8 semaines après le commencement de l'étude, on observe que le nombre de crises de ce patient durant la période avec traitement a doublé par rapport au comptage de référence. Il est donc justifié de se poser la question si les données sont peut-être erronées. La majorité des auteurs d'articles a quand même juste cherché les valeurs extrêmes dans un sens. Effectivement, le patient 232 n'avait aucune crise épileptique durant les huit semaines d'étude active. On va donc procéder aux tests une fois les modèles définies pour voir s'il s'agit d'une observation influente ou aberrante. Pour tester ceci, la distance de Cook ou l'effet de levier d'une observation peuvent être utiles. Pourtant un test de Grubbs pour deux observations situées au sens opposé `grubbs.test` du package R `{outliers}` permet déjà de

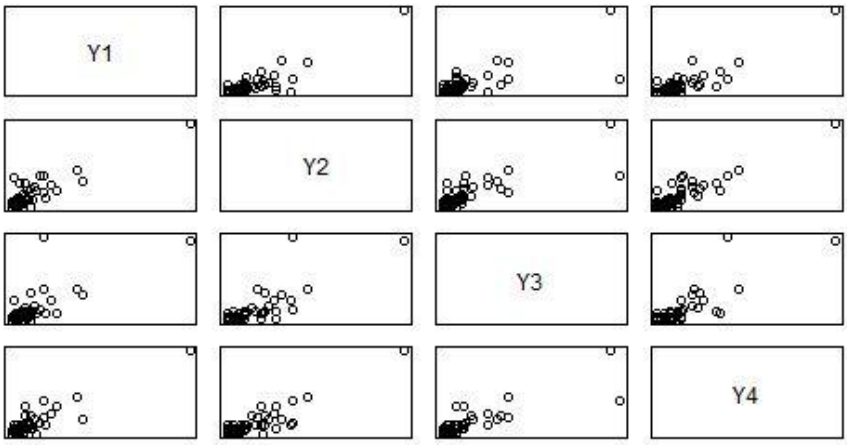
voir avant la spécification d'un modèle s'il s'agit d'outliers positionnés aux côtés extrêmes opposés ou pas. Pour le nombre total de crises épileptiques, on trouve que le comptage de 0 (patient 232) et de 302 (patient 207) sont effectivement des outliers. À ce point, il faut quand même dire qu'ici il est risqué d'exclure un patient car ceci entraîne l'élimination de 4 données. De plus, en ce qui concerne les études cliniques, il est préférable de garder quand même le jeu de données initial sauf forte conviction qu'une erreur d'enregistrement a été survenue. Le médicament est conçu pour aider tous les patients souffrant de crises épileptiques ; forte ou faible, donc il faut garder tous les données récupérées.

En analysant les espérances et variances des deux groupes de traitements et une fois l'ensemble, on voit qu'il faut probablement s'occuper par la suite de la sur-dispersion. Il faut également procéder au test pour pouvoir affirmer qu'il y a de la sur-dispersion, même si dans ce cas, ce phénomène est très prononcé par la différence entre l'espérance et la variance. On remarque également, que l'élimination du patient 207 entraîne une baisse de l'importance de la sur-dispersion. Voici donc les tableaux pour le groupe placebo, groupe Progabide et tout combiné, affiché par visite. `Mean_sans` et `Var_sans` représentent l'espérance respectivement la variance sans le patient 207. On voit clairement, que le patient 207 appartient au groupe Progabide comme il n'y a pas de changement dans le tableau du groupe placebo lors de l'élimination de celui-ci.

```
> placebo
  Visit     Mean       Var Mean_Sans  Var_Sans
1     1 9.357143 102.75661  9.357143 102.75661
2     2 8.285714  66.65608  8.285714  66.65608
3     3 8.785714 215.28571  8.785714 215.28571
4     4 7.964286  58.18386  7.964286  58.18386
> progabide
  Visit     Mean      Var Mean_Sans Var_Sans
1     1 8.580645 332.7183  5.466667 33.22299
2     2 8.419355 140.6516  6.533333 31.42989
3     3 8.129032 193.0495  6.000000 54.34483
4     4 6.709677 126.8796  4.833333 18.35057
> ProgabideEtPlacebo
  Visit     Mean       Var Mean_Sans  Var_Sans
1     1 8.949153 220.08358  7.344828  69.42287
2     2 8.355932 103.78492  7.379310  48.34483
3     3 8.440678 200.18177  7.344828 131.59831
4     4 7.305085  93.11222  6.344828  39.38778
```

Comme le rapport de la variance avec l'espérance n'est pas le même pour chaque visite, on peut se demander si on se trouve peut-être dans le cas où le paramètre de dispersion dépend du temps. De plus, on remarque qu'on n'est probablement pas dans le cas où les variances sont homogènes pour chaque visite.

En ce qui concerne l'indépendance des comptages, il est important à dire que comme il s'agit des données longitudinales, on observe une indépendance partielle entre les comptages. Les quatre mesures prises pour un patient ne sont par exemple pas indépendantes comme elles proviennent du même individu. Voici le graphique de la matrice de corrélation entre les différentes comptages par visite où Y1 représente la première visite, Y2 la deuxième et ainsi de suite. Ce graphique nous suggère qu'il y a effectivement de la corrélation entre les mesures par visite. De plus, la présence de quelques outliers est visible.



Les calculs nous suggèrent également que les comptages par visites ne sont pas indépendants comme on voit bien sur le tableau à gauche :

```
> matrice_corr_sum
          V1        V2        V3        V4
V1 1.0000000 0.8707835 0.7377449 0.8924586
V2 0.8707835 1.0000000 0.8024795 0.8951125
V3 0.7377449 0.8024795 1.0000000 0.8240270
V4 0.8924586 0.8951125 0.8240270 1.0000000
```

```
> matrice_corr_sum_sans
          V1        V2        V3        V4
V1 1.0000000 0.6878522 0.5442376 0.7173347
V2 0.6878522 1.0000000 0.6700603 0.7616303
V3 0.5442376 0.6700603 1.0000000 0.7125445
V4 0.7173347 0.7616303 0.7125445 1.0000000
```

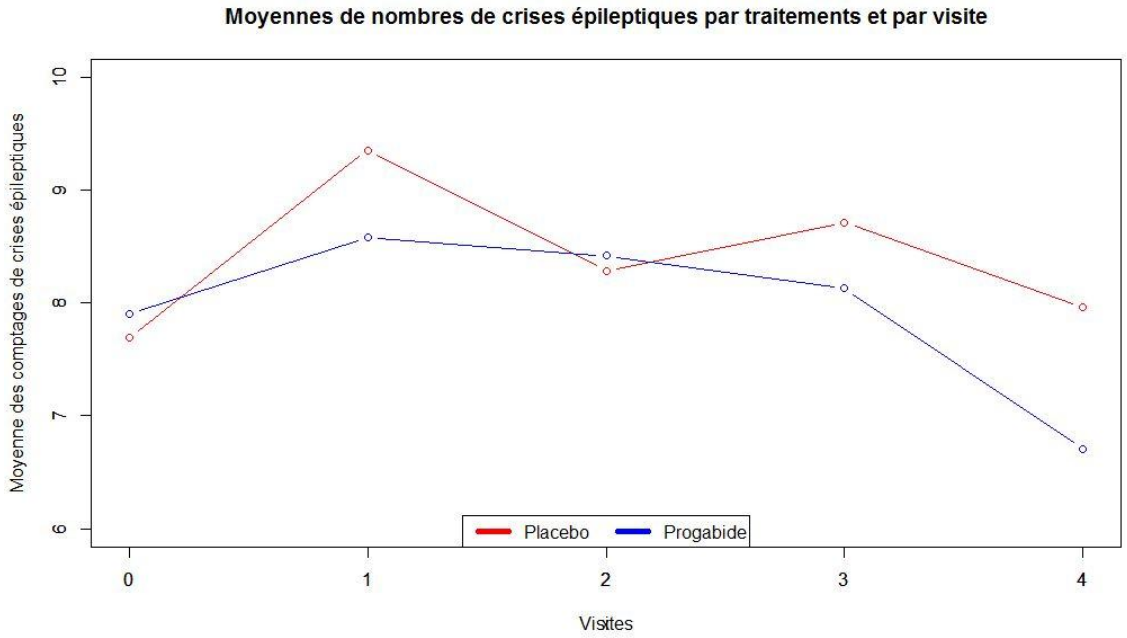
Par contre, les patients sont supposés indépendants. Il n'existe pourtant pas de test pour vérifier ceci.

Dans le tableau de droite, on fait une observation intéressante où on voit que l'exclusion du patient 207 a un effet sur la corrélation entre les visites. Pourtant cette réduction ne suffit pas pour pouvoir affirmer que la corrélation est devenue négligeable.

Pour se faire une idée de l'effet du médicament Progabide durant l'étude, on peut comparer les deux graphes représentant le nombre moyen de crises épileptiques de chacun des deux groupes pour les quatre visites. « o » représente ici la période de référence (eng. *baseline period*). Pour obtenir l'ordonnée de cette visite, on a divisé le nombre de comptages de crises épileptiques de référence par quatre et on a pris la moyenne pour chacun des deux groupes. La division par quatre était nécessaire comme tous les autres comptages ont été faits sur une période de deux semaines tandis que les comptages de référence ont été mesurés initialement sur une période de 8 semaines. Breslow (1996) a construit à peu près le même graphique que vous trouvez sur la page suivante afin de se faire une idée de l'effet du traitement par rapport au placebo et leur évolution au cours de l'étude.

Voici donc les moyennes des comptages de crises épileptiques par traitement et par visite pour illustrer l'effet du médicament au long de l'étude :

```
placebo.1 progabide.1   placebo.2 progabide.2   placebo.3 progabide.3   placebo.4 progabide.4
 9.357143    8.580645    8.285714    8.419355    8.714286    8.129032    7.964286    6.709677
```



La proximité des moyennes à l'instant zéro est évident comme aucun traitement n'a été pris jusqu'à présent. La moyenne des comptages de référence pour les deux groupes est 7,81. Cette valeur est proche de la valeur de placebo dans la quatrième visite. Il est quand même un peu étonnant qu'il y a tant de variations au long de l'étude dans le groupe placebo. On s'attendait plutôt à ce que ces comptages du groupe placebo seraient à peu près constants autour la moyenne des comptages de référence. La variation entre la visite 0 et les autres du groupe placebo semble peut-être si importante parce qu'en sommant les comptages de la période de référence (8 semaines) et en prenant la moyenne pour chacun des deux groupes, des fluctuations éventuellement importantes sont cachées. Pourtant, la moyenne des comptages initiaux est inférieure aux comptages des visites 1, 2 et 3. Ceci nous illustre la plus mauvaise situation en développement de médicament : pas de réduction de la condition du patient prenant le médicament. L'observation sauvante se produit à la visite 4 où la moyenne des crises épileptiques tombe en dessous du seuil de la moyenne des comptages de référence.

On remarque aussi que le groupe Progabide se trouve dès le début de la phase médicamenteuse en dessous du groupe de contrôle sauf pour V2. À la fin de l'étude en V4, on remarque la baisse la plus importante de cette étude pour les deux groupes. À cause de cet événement, Thall et Vail (1990) ont introduit la variable binaire V4 dans leur modèle. Ainsi en regardant le graphique, on pourrait croire que Progabide a effectivement un effet réduisant le nombre de crises. Il est malgré tout important de tester si cette différence est significative. À ce point, j'évoque quand même l'importance de l'échelle choisie. Sur le graphique à la quatrième visite, la différence entre Prograbide et placebo semble peut être beaucoup mais il faut également remarquer qu'il ne s'agit que d'une différence d'une seule crise épileptique en moyenne. Le test le plus simple et adapté qu'on peut effectuer pour tester l'efficacité du traitement est celui de Wilcoxon Mann Whitney (eng. *Wilcoxon rank sum test*). Le traitement des données par des rangs est important parce qu'ainsi la distribution des données ne joue aucun rôle. Le test compare alors les comptages issus du placebo avec ceux issus de Progabide et teste alors si la différence éventuellement existante est significative ou non. Le package R `{coin}` contient ce test nommé `wilcox_test`. Les hypothèses sont les suivantes : H_0 : « il n'y a pas de différence entre les deux groupes » contre H_a : « il y a une différence ». On va prendre

un seuil de test $\alpha = 5\%$ comme pour tous les tests dans ce rapport d'ailleurs. Pour le jeu de données de Breslow et Clayton et pour Breslow, on trouve qu'il n'y a pas la différence statistiquement significative entre le traitement et le placebo et ceci avec des p-valeurs de 0.076 et 0.123 respectivement. Pourtant au niveau de l'interprétation, comme le test se fait avec le plus simple des modèles avec que le facteur traitement, il n'est pas très sensible. Si on trouve par le test qu'il n'y a pas de différence alors on ne peut pas dire qu'on a évidence qu'il n'y vraiment pas de différence. Par contre si on avait trouvé que le traitement a effectivement un effet d'après le test, alors on aurait pu dire que c'est probablement le cas. Pour plus de détails, je propose une leçon de cours de Charlotte Wickham (2013) de l'Oregon State University.

Une autre méthode généralement plus fiable pour détecter l'effet d'un traitement est celle de l'interprétation des coefficients des effets significatifs obtenus par une régression. Le facteur traitement nous intéresse en particulier. Notons que pour les GLMs et GLMMs, on a utilisé une fonction de lien canonique g. Ainsi pour un modèle sans interactions et en fixant tous les prédicteurs autres que traitement, on observerait avec Prograbide en moyenne $g^{-1}(\hat{\beta}_{Traitement})$ fois plus de crises épileptiques comparé au groupe placebo. Dans le cas GLMM sans terme d'interaction, $\hat{\beta}_{Traitement}$ est en fait la somme du coefficient estimé pour les prédicteurs traitement et l'estimation des effets aléatoires. Si g = ln et donc $g^{-1} = exp$, il vient de la définition de la fonction exponentielle qu'ainsi toute valeur négative de $\hat{\beta}_{Traitement}$ représente une réduction du nombre de crises épileptiques. Donc $\hat{\beta}_{Traitement} = g(\mathbb{E}[Y \mid \mathbb{X}\ tq\ X_{Traitement} = 1\ ]) - g(\mathbb{E}[Y \mid \mathbb{X}\ tq\ X_{Traitement} = 0\ ])$ car les autres prédicteurs sont supposés fixés et $g(\mathbb{E}[Y \mid \mathbb{X}]) = \mathbb{X}\beta$. Dans le cas où g = logit, on a $\exp(g(x)) = \frac{x}{1-x}$. L'interprétation se fait alors sur le rapport des côtes (eng. *odds ratio*) : $\exp(\hat{\beta}_{Traitement}) = \frac{\mathbb{E}[Y \mid \mathbb{X}\ tq\ X_{Traitement}=1\ ]}{\mathbb{E}[Y \mid \mathbb{X}\ tq\ X_{Traitement}=0\ ]} \frac{1-\mathbb{E}[Y \mid \mathbb{X}\ tq\ X_{Traitement}=0\ ]}{1-\mathbb{E}[Y \mid \mathbb{X}\ tq\ X_{Traitement}=1\ ]}$. Quand tous les autres prédicteurs à l'exception de celui de traitement sont alors fixés, le phénomène que les patients prenant le traitement Progabide rencontrent un nombre de crises épileptiques plus élevé que le seuil choisi dans le cas binaire sera $\exp(\hat{\beta}_{Traitement})$ fois plus probable de se produire que le cas contraire.

En cas de présence d'interaction, la situation est un peu différente. On va interpréter l'effet du traitement dès qu'on arrive à la modélisation du problème.

4.2. Les modèles initiaux (Thall and Vail 1990)

Dans l'article de Thall and Vail (1990), pour traiter le jeu de données des comptages de crises épileptiques, des modèles de covariance pour données longitudinales ont été définis. Thall et Vail définissent la structure de la matrice variance-covariance des effets aléatoires qui prennent en compte la corrélation entre les 4 mesures d'un patient en donnant la variance et la covariance du modèle. Leur paramétrisation est expliquée sous "2. A covariance matrix" et "4. Other Forms of V". Ils ont estimé les paramètres par la méthode d'équations d'estimation généralisées (GEE) (eng. *generalized estimating equations*) et les paramètres introduits pour définir la matrice de variance-covariance ont été estimés par la méthode des moments.

Malgré tous mes efforts, je n'ai pas réussi à établir comment reproduire ces exactes données. Les difficultés commencent déjà pour trouver une méthode dans SAS ou R pour spécifier cette

même forme de la matrice de variance-covariance. Ils existent plusieurs méthodes pour spécifier la structure de la matrice de corrélation dans SAS et R mais aucune ne correspond à la forme exigée car je n'ai pas réussi à obtenir des résultats cohérents.

Ce problème est très réparti si on essaie de refaire les exemples des articles scientifiques en statistiques. Le manque d'explications sur la réalisation des calculs (logiciels, commandes) ne facilite pas la tâche de reproduire ces résultats. Il s'ajoute alors une toute autre thématique : Est-ce que les auteurs ont fait une erreur dans leurs calculs d'où l'impossibilité de reproduire les mêmes résultats ? Cette problématique pourrait être facilement résolue en déposant un script avec le code utilisé ou simplement en fournissant plus de détails par rapport à la réalisation des calculs. Il faut toujours garder en tête que les exemples servent à illustrer une méthode. Mais qui veut utiliser une méthode où on n'arrive même pas à reproduire les exemples ? Une telle méthode ne semble pas très fiable.

Cette difficulté de reproduction des calculs de Thall et Vail (1990) a été probablement également perçu par d'autres scientifiques car la majorité des articles citant les deux auteurs a également cité Breslow et Clayton (1993) et se sont servis majoritairement de leurs modèles pour reproduire les résultats afin de pouvoir les comparer à leurs nouvelles méthodes.

4.3. Tout autour du Poisson (Breslow and Clayton 1993)

Breslow et Clayton (1993) ont analysé les données de comptage des crises épileptiques en appliquant 4 modèles de Poisson. Dans le package R `{glmmAK}`, il y a le jeu de données nommé « `eplepticBC` », le meilleur adapté car il contient les variables transformées qui ont été utilisé par Breslow et Clayton. Dans cet article, les estimations ont été faites par quasi-vraisemblance pénalisée (PQL) expliquée brièvement en Annexe V c). Leurs résultats obtenus se trouvent en Annexe VI a). Ils ont introduit les éléments suivants avec i=1, …, 59 et t=1, …, 4 :

$Y_{it} = Seizure =$ "nombre de crises épileptiques du patient i à la t$^{\text{ième}}$visite"

$X_i^{(1)} = Trt =$ "Traitement du i$^{\text{ième}}$ patient" $= \begin{cases} 0\ pour\ placebo \\ 1\ pour\ Progabide \end{cases}$

$X_i^{(2)} = Base =$ logarithme népérien d'un quart du comptage sur 8 semaines avant l'étude

$X_i^{(3)} = Age =$ "logarithme népérien des âges des patients"

$X_t^{(4)} = V4 =$ variable binaire pour Visite 4

$= \begin{cases} 0\ si\ le\ comptage\ n'était\ pas\ enregistré\ lors\ de\ la\ 4^{ième}\ visite \\ 1\ sinon \end{cases}$

Pour l'analyse des données, contrairement à Thall et Vail (1990), Breslow et Clayton (1993) n'ont pas exclu le patient à `ID` 207 qui semble pourtant avoir un nombre extrême de crises épileptiques durant l'étude. Ceci peut influencer les estimations et erreurs-standards.

4.3.1. Régression de Poisson simple et naïve

Le **premier modèle (I)** correspond à une régression de Poisson simple :

$$\ln(\mathbb{E}[Y_{it}|X_i^{(1)}, X_i^{(2)}, X_i^{(3)}, X_t^{(4)}, X_i^{(2)}X_i^{(1)}]) = \beta_0 + \beta_1 1_{\{X_i^{(1)}=1\}} + \beta_2 X_i^{(2)} + \beta_3 X_i^{(3)} + \beta_4 1_{\{X_t^{(4)}=1\}} + \beta_5 X_i^{(2)} 1_{\{X_i^{(1)}=1\}}$$

Sous R :

```
I<-glm(Seizure~Age+Base:Trt+Trt+Base+V4,family=poisson,data=BC)
```

On obtient exactement le même résultat que Breslow et Clayton (1993). Tous les effets sont significatifs et la variable `Trt` a même un coefficient négatif. Ce modèle ne tient par contre pas compte de la sur-dispersion résultant peut-être de la dépendance des mesures de comptages d'un patient. Effectivement, en faisant une telle régression de Poisson, on a assumé qu'on a 236 observations indépendantes ce qui revient à avoir des données provenant de 236 personnes ce qui n'est pas du tout la même chose qu'avoir 59 fois 4 observations. Ainsi on a donc sous-estimé les erreurs standards. La sur-dispersion est facilement détectable comme la déviance résiduelle est considérablement plus importante que son degré de liberté. Voici donc le tableau avec nos résultats :

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.75758    0.40746  -6.768 1.31e-11 ***
Age          0.89705    0.11644   7.704 1.32e-14 ***
Trt         -1.34112    0.15674  -8.556  < 2e-16 ***
Base         0.94952    0.04356  21.797  < 2e-16 ***
V4          -0.16109    0.05458  -2.952  0.00316 **
Base:Trt     0.56223    0.06350   8.855  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2521.75  on 235  degrees of freedom
Residual deviance:  869.32  on 230  degrees of freedom
AIC: 1647.3

Number of Fisher Scoring iterations: 5
```

L'estimation à l'aide de PQL est considérée comme une approximation de la méthode du maximum de vraisemblance (ML) avec des effets aléatoires distribués suivant une loi normale. Comme pour notre modèle I sans effets aléatoires, on peut affirmer que la fonction de `glm()` sous R nous fournit les estimateurs souhaités.

Pour tester si nos données sont vraiment sur-dispersées, on fait un test de Dean (1992) ce qui est tout à fait faisable comme nos estimateurs ont été obtenus par ML. Les hypothèses de ce test du score sont les suivantes : $H_0 : \nu = 0$ *contre* $H_a : \nu > 0$ où $\nu = \frac{1}{\alpha}$ de la variance du modèle binomiale négatif II. Effectivement, si $\nu = 0$, on se trouve dans le cas poissonien $\mathbb{V}[Y] = \mathbb{E}[Y]$. La statistique est la suivante :

$$T = \frac{\sum_{j=1}^{236}\left\{\left(y_j - \hat{\mu}_j\right)^2 - y_j\right\}}{\sqrt{2\sum_{j=1}^{236}\hat{\mu}_j^2}} \text{ où } \hat{\mu}_j \textit{ est la moyenne estimée en cas poissonien } (H_0)$$

Sous l'hypothèse nulle, cette statistique suit une loi normale centrée réduite.

Effectivement, ici n=236 car on a supposé avec le modèle de Poisson qu'on a 236 observations indépendante. Le test de Dean n'est qu'adapté dans une telle situation d'indépendance. Ce test est dans le package `{DCluster}` de R et est effectué à l'aide de la commande `DeanB` pour la loi binomiale négative à variance quadratique. Le test nous indique qu'on peut rejeter l'hypothèse nulle (p-valeur $< 2.2 \cdot 10^{-16}$) et donc on se trouve dans le cas sur-dispersé. Il faut quand même être vigilant car ce test est très sensible aux observations extrêmes. Pourtant, les observations sous 4.1. nous montrent qu'il n'y a pas de doute que les données sont sur-dispersées et comme Breslow et Clayton n'excluent pas d'observations, nous allons suivre leur exemple.

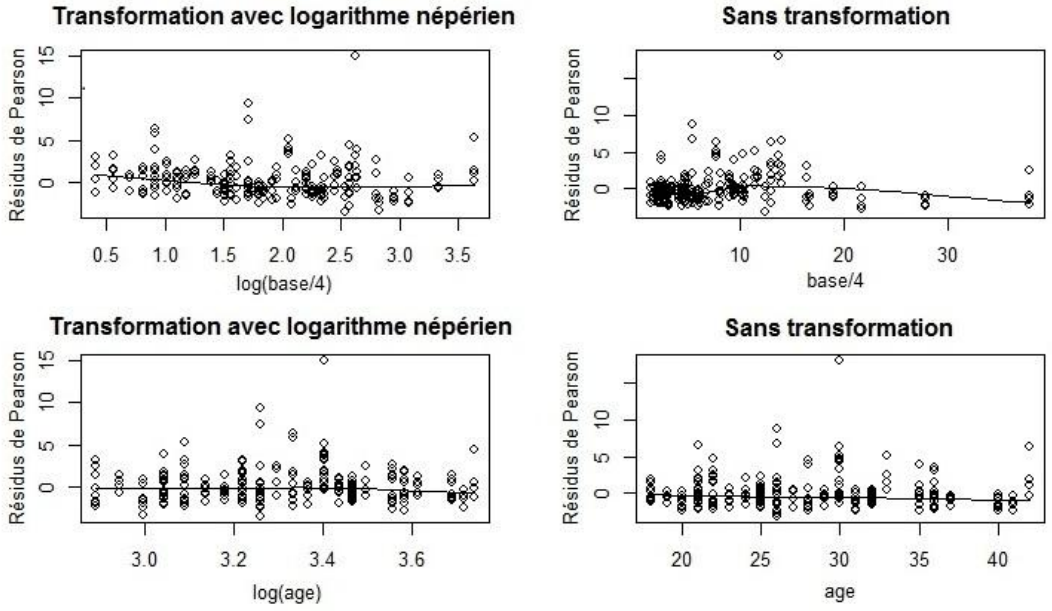
Rappelons que la valeur du paramètre de dispersion est estimée par $\hat{\phi} = \frac{X^2}{n-p-1}$. Dans notre cas, n=236 et p=5. Pour une loi de Poisson, le poids w_i vaut toujours 1. De plus, on remarque que le corps de la somme est en fait la formule des résidus de Pearson au carrées. Il s'impose alors d'utiliser la formule R suivante pour estimer le paramètre de dispersion : `sum((resid(I,typ="pear"))^2)/(236-5-1)`. On obtient $\hat{\phi} = 4.41$. Considérons un modèle quasi-Poisson avec la commande R suivante :

```
Iqp<-glm(Seizure~Age+Base:Trt+Trt+Base+V4,family=quasipoisson,data=BC)
```

Ici, R nous donne directement la valeur du paramètre de dispersion estimée par la méthode évoquée en haut. Effectivement, dans les modèles de quasi-vraisemblance, on a besoin du paramètre de dispersion pour prendre en compte la variation extra-poissonienne.

Par la suite j'ai calculé les résidus de Pearson studentisés pour les covariables continues à partir de deux modèles de Poisson pour trouver des indices expliquant les transformations de variables. L'un des modèles est celui proposé par Breslow et Clayton avec les transformations log népérien de la variable âge et du comptage de référence et l'autre modèle est sans ces transformations logarithmiques. Les graphiques correspondant à l'âge ne montrent aucune amélioration ni une détérioration lors de la transformation qui semble donc inutile. Le graphique sans transformation ayant les comptages initiaux en abscisse nous laisse penser à un problème d'homoscédasticité. Par contre celui avec transformation semble bien avoir réglé ce problème et constitue donc une amélioration. Les points semblent être répartis aléatoirement. Ce qui est également à mentionner est la présence d'outliers comme quelques points semblent très éloignés des autres. De plus, le critère d'information d'Akaike nous indique que le modèle avec transformations (AIC = 1647.3) est le meilleur des deux (AIC du modèle sans transformations vaut 1724.6). Je n'ai pas fait de graphiques pour des variables binaires comme ils ne vont pas être très parlants. Dans le modèle sans transformation, toutes les variables sont considérées comme ayant un effet significatif.

Résidus de Pearson studentisés en fonction des variables continues (transformées ou non) pour modèle I



Si on compare maintenant les estimations du modèle de Poisson avec celui de Quasi-Poisson, on remarque qu'au niveau des estimations il n'y a pas de différence :

```
Call:
glm(formula = Seizure ~ Age + Base:Trt + Trt + Base + V4, family = poisson,
    data = BC)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0949  -1.4271  -0.2763   0.7583  10.7557

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.75758    0.40746  -6.768 1.31e-11 ***
Age          0.89705    0.11644   7.704 1.32e-14 ***
Trt         -1.34112    0.15674  -8.556  < 2e-16 ***
Base         0.94952    0.04356  21.797  < 2e-16 ***
V4          -0.16109    0.05458  -2.952  0.00316 **
Base:Trt     0.56223    0.06350   8.855  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Call:
glm(formula = Seizure ~ Age + Base:Trt + Trt + Base + V4, family = quasipoisson,
    data = BC)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0949  -1.4271  -0.2763   0.7583  10.7557

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.7576     0.8558  -3.222 0.001457 **
Age           0.8971     0.2446   3.668 0.000304 ***
Trt          -1.3411     0.3292  -4.074 6.37e-05 ***
Base          0.9495     0.0915  10.377  < 2e-16 ***
V4           -0.1611     0.1146  -1.405 0.161296
Base:Trt      0.5622     0.1334   4.216 3.58e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 4.411793)
```

Par contre, on observe partout une augmentation des valeurs des erreurs standards en passant du Poisson au Quasi-Poisson. Ceci n'est pas du tout surprenant parce qu'en effectuant une régression de Poisson et en présence de sur-dispersion, on va sous-estimer les erreurs-standards. Ainsi en appliquant un modèle de quasi-vraisemblance, on reçoit des erreurs plus adaptées à la situation. On corrige les erreurs standards obtenues par la régression de Poisson naïve en les multipliant par la racine carrée de la valeur du paramètre de dispersion. De plus, on voit que l'effet de V4 est devenu non significatif ce qui n'est pas non plus surprenant car le niveau de significativité est relié aux erreurs-standards qui sont déjà à la base mauvaises à cause de la sous-estimation.

Regardons maintenant à l'aide du levier si on trouve des points influents. La valeur critique qu'il s'agit de ne pas dépasser est 0.0508 $\left(=\frac{2*(p+1)}{n}=\frac{2*6}{236}\right)$. En utilisant la fonction `hatvalues()` du R package `{stats}`, on voit que le patient 207 avec des hatvalues de 0,204 pour les 3 premières visites et 0,265 pour la dernière possède des valeurs largement au-dessus des autres. À la seconde place se trouve le patient 213 avec les valeurs 0.101 pour les trois premières visites et 0,132 pour la quatrième. Les patients 126 et 205 dépassent par contre que légèrement le seuil critique. Les hatvalues sont de bons indicateurs d'observations influentes. Le procédé standard pour vérifier s'il s'agit vraiment des observations influentes est d'exclure l'observation avec le plus grand effet de levier, puis refaire la régression et regarder si les résultats ont beaucoup changé. Il ne faut quand même pas abuser avec la technique d'élimination d'observations. Pourtant dans notre cas, pour éviter des données incomplètes

comme tous les autres auteurs ayant traité ces données, on va éliminer carrément les observations correspondant au patient.

Effectivement, on remarque que les résultats ont beaucoup changé. L'effet de l'interaction de notre modèle n'est plus significatif et sinon on remarque des fortes différences dans les estimations et des différences légères pour les erreurs standards. Néanmoins, le critère d'information d'Akaike (AIC), indicateur de la validité d'ajustement des données (1565,1) est meilleur que celui de Breslow et Clayton (1647,3) après l'élimination de l'observation suspecte. Comme on n'a rien changé au modèle, la sur-dispersion n'est toujours pas prise en compte et se manifeste.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.1296  -1.3057  -0.3119   0.6139  10.7820

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -2.32747    0.41584  -5.597 2.18e-08 ***
sub_BC$Age              0.76870    0.11925   6.446 1.15e-10 ***
sub_BC$Trt             -0.52145    0.18728  -2.784  0.00536 **
sub_BC$Base             0.95017    0.04354  21.823  < 2e-16 ***
sub_BC$V4              -0.14792    0.05915  -2.501  0.01239 *
sub_BC$Base:sub_BC$Trt  0.13781    0.08544   1.613  0.10677
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1671.48  on 231  degrees of freedom
Residual deviance:  811.66  on 226  degrees of freedom
AIC: 1565.1

Number of Fisher Scoring iterations: 5
```

Une observation intéressante se produit lorsque je veux retrouver le modèle à la main en partant du modèle complet à interactions d'ordre 2. Je retiens le modèle suivant en appliquant la même méthode d'élimination pas-à-pas descendante des variables explicatives.

```
Ib<-glm(Seizure~(Age+Trt+Base+V4)^2-Base:V4-Age:Trt-Age:V4-Trt:V4-Age,
    family=poisson,data=BC)
```

Il s'agit de la même méthode que l'option backward sous SAS en commençant à éliminer les interactions qui sont les moins significatives et en terminant par les effets principaux si le modèle ne possède pas d'interaction significative avec cet effet principal. J'obtiens le modèle possédant l'`Age, Trt, Base, V4`, l'interaction d'`Age` et `Base` et finalement l'interaction entre `Trt` et `Base`. Le modèle retenu diffère donc du modèle proposé par Breslow et Clayton.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0215  -1.4157  -0.2970   0.8903  10.7442

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.91686    1.44589   1.326 0.184929
Age         -0.49196    0.42999  -1.144 0.252579
Trt         -1.57162    0.17271  -9.100  < 2e-16 ***
Base        -1.17610    0.63448  -1.854 0.063789 .
V4          -0.16109    0.05458  -2.952 0.003161 **
Age:Base     0.63079    0.18826   3.351 0.000806 ***
Trt:Base     0.69122    0.07490   9.229  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2521.75  on 235  degrees of freedom
Residual deviance:  858.01  on 229  degrees of freedom
AIC: 1638

Number of Fisher Scoring iterations: 5
```

Comme `Age` est en interaction avec `Base` et comme cette interaction est significative, on n'a pas le droit de supprimer `Age` qui est non significatif. Même si l'intercept et `Age` ne sont pas significatifs, le critère AIC (1638) est légèrement meilleur que celui du modèle de Breslow et Clayton (1647,3).

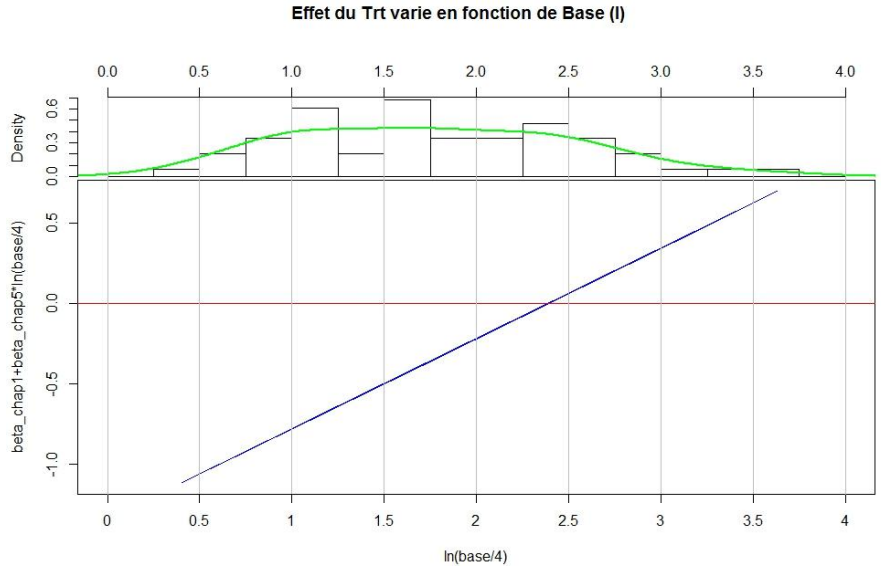
Revenons maintenant aux résultats trouvés pour le modèle I. Il s'agit d'interpréter l'effet du traitement. Comme on a une interaction dans notre modèle, on ne peut pas considérer tous les autres prédicteurs comme étant tous fixes. Effectivement, le facteur traitement va varier en fonction du nombre de log-népérien comptages de référence divisé par 4. On choisit une log-échelle pour représenter ceci pour ne pas nous compliquer les choses. Avec ce choix, en fixant tous les autres prédicteurs, on obtient :

$$\ln(\mathbb{E}[Y \mid \mathbb{X}\ tq\ X_{Traitement} = 1\ ]) = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 X_i^{(2)} + \hat{\beta}_3 X_i^{(3)} + \hat{\beta}_4 1_{\{X_t^{(4)}=1\}} + \hat{\beta}_5 X_i^{(2)}$$

$$-\ln(\mathbb{E}[Y \mid \mathbb{X}\ tq\ X_{Traitement} = 0\ ]) = -\left(\hat{\beta}_0 + \hat{\beta}_2 X_i^{(2)} + \hat{\beta}_3 X_i^{(3)} + \hat{\beta}_4 1_{\{X_t^{(4)}=1\}}\right)$$

$$\Rightarrow \ln(\mathbb{E}[Y \mid \mathbb{X}\ tq\ X_{Traitement} = 1\ ]) - \ln(\mathbb{E}[Y \mid \mathbb{X}\ tq\ X_{Traitement} = 0\ ]) = \hat{\beta}_1 + \hat{\beta}_5 X_i^{(2)}$$

On a $\hat{\beta}_1 = -1.34112$ et $\hat{\beta}_5 = 0.56223$, on peut donc tracer cette droite (en bleu) :



L'histogramme correspond à ln(base/4) et nous indique la fréquence de ces valeurs. On a également tracé la densité associée (en vert) pour rendre ce graphe plus visible. Comme on voit bien sur le graphique, l'effet du traitement varie en fonction des comptages de référence. Effectivement, pour quelqu'un qui a un comptage de référence de 30 crises épileptiques sur une période de huit semaines ($4\exp(2) \cong 30$), le nombre de crises épileptiques d'un patient du groupe Progabide équivaut en moyenne à 0.81 fois ($\exp(-1.34112 + 0.56223 \cdot 2) \cong 0.81$) le nombre de crises d'un patient prenant un placebo. Ainsi on constate une réduction pour ces patients prenant le médicament. On voit également à l'aide des deux graphiques, que la majorité des patients vont expérimenter une réduction. À partir de 44 comptages de référence ($-1.34112 + 0.56223 \cdot x = 0 \Leftrightarrow x \cong 2.39$ d'où $4 \cdot \exp(2.39) \cong 44$), Progabide ne semble plus à être efficace contre les crises épileptiques. 13 des 59 patients (environ 22% de notre échantillon) se trouvent alors dans le cas où le traitement n'est pas effectif.

4.3.2. GLMM Poisson à un effet aléatoire

Le **deuxième modèle (II)** est un GLMM Poisson à un effet aléatoire pour prendre en compte les variations entre les mesures des sujets (eng. *subject-specific random effect*). Il s'écrit :

$$\ln(\mathbb{E}[Y_{it} | X_i^{(1)}, X_i^{(2)}, X_i^{(3)}, X_t^{(4)}, X_i^{(2)}X_i^{(1)}, u_i]) = \beta_0 + \beta_1 1_{\{X_i^{(1)}=1\}} + \beta_2 X_i^{(2)} + \beta_3 X_i^{(3)} + \beta_4 1_{\{X_t^{(4)}=1\}} + \beta_5 X_i^{(2)} 1_{\{X_i^{(1)}=1\}} + u_i$$

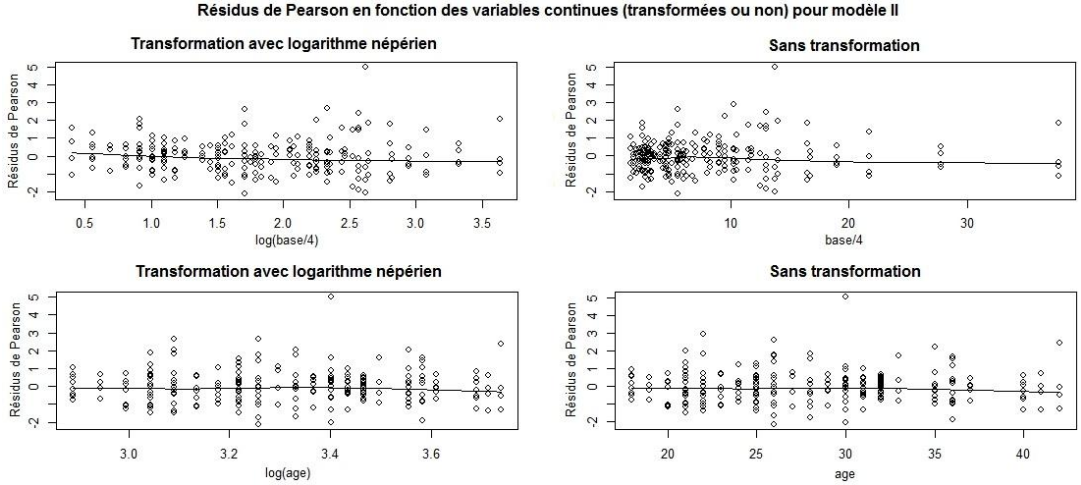
Sous R :

```
II<-glmmPQL(Seizure~Age+Base:Trt+Trt+Base+V4,random=~1|id,family=poisson,data=BC)
```

Avec ce modèle décrivant bien la situation désirée, on n'obtient quand même pas les résultats trouvés par Breslow et Clayton. On observe des différences non négligeables notamment sur l'estimation du coefficient d'`Age` et `Intercept`. Les estimations et écarts-types pour `Base` et l'interaction de `Base` et `Trt` sont proches de celles trouvés par Breslow et Clayton. Les erreurs standards pour `Intercept, Age` et `V4` sont proches tandis que l'estimation du coefficient de `V4` est exactement la même. De plus, on a retrouvé l'estimation et l'erreur-type exacte pour `Trt`.

```
Linear mixed-effects model fit by maximum likelihood
 Data: BC
  AIC BIC logLik
   NA  NA     NA

Random effects:
 Formula: ~1 | id
        (Intercept) Residual
StdDev:   0.4449054 1.399283

Variance function:
 Structure: fixed weights
 Formula: ~invwt
Fixed effects: Seizure ~ Age + Base:Trt + Trt + Base + V4
                Value Std.Error  DF   t-value p-value
(Intercept) -1.4747841 1.1824049 176 -1.247275  0.2140
Age          0.5375809 0.3465146  54  1.551395  0.1266
Trt         -0.9132463 0.4139368  54 -2.206246  0.0316
Base         0.8828134 0.1293711  54  6.823884  0.0000
V4          -0.1610871 0.0773565 176 -2.082398  0.0388
Base:Trt     0.3410064 0.2034656  54  1.675990  0.0995
 Correlation:
         (Intr) Age    Trt    Base   V4
Age      -0.975
Trt       0.054 -0.200
Base     -0.168 -0.038  0.592
V4       -0.014  0.000  0.000  0.000
Base:Trt -0.130  0.267 -0.935 -0.645  0.000

Standardized Within-Group Residuals:
       Min          Q1         Med          Q3         Max
-2.13615625 -0.63779485 -0.08342573  0.42100427  4.97927910

Number of Observations: 236
Number of Groups: 59
```

Si on s'intéresse maintenant à nouveau aux transformations, on trace les graphiques des résidus de Pearson (non studentisés cette fois-ci) en fonction des variables continues log népérienne transformées et non. On remarque de suite que les graphiques représentent une situation similaire que sous le point précédent. À nouveau pour l'âge, la transformation logarithmique népérienne n'était pas nécessaire tandis que pour les comptages de références, la transformation entraîne une amélioration visible. Sans ces transformations, la variable des comptages initiaux et `V4` sont les seules à avoir un effet significatif sur la réponse.



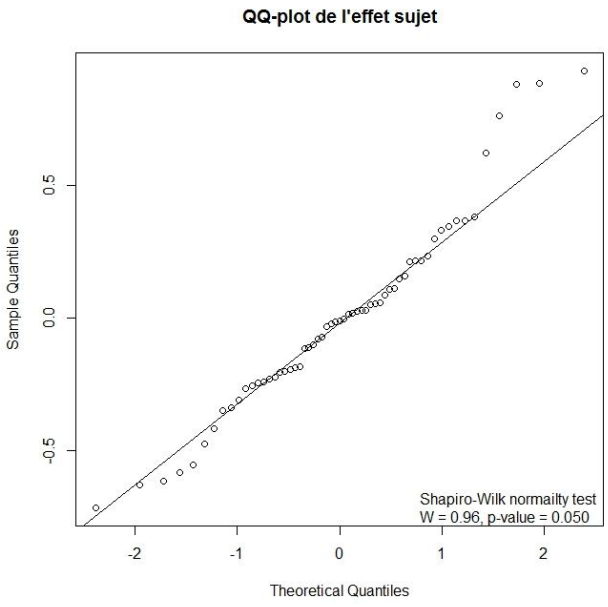
De plus, quand on utilise la fonction `glmer` du package R `{lme4}` qui estime les paramètres par défaut avec l'approximation de Laplace, on peut introduire l'option `nAGQ=k` dans la commande. Cette option va appliquer la quadrature adaptive de Gauss-Hermite à `k` points de quadrature. Cette méthode d'approximation au ML est jugée plus précise que Laplace (les deux méthodes sont expliquées brièvement en Annexe V b) pour l'approximation de Laplace et Annexe V d) pour AGHQ) :

```
II.test<-glmer(Seizure~Age+Base:Trt+Trt+Base+V4+(1|id),family=poisson,
          nAGQ=20, data=BC)
```

```
Generalized linear mixed model fit by the adaptive Gaussian Hermite approximation
Formula: Seizure ~ Age + Base:Trt + Trt + Base + V4 + (1 | id)
   Data: BC
   AIC   BIC logLik deviance
 578.6 602.8 -282.3    564.6
Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 0.25283  0.50282
Number of obs: 236, groups: id, 59

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.33742    1.18036  -1.133  0.25719
Age          0.48423    0.34663   1.397  0.16243
Trt         -0.93300    0.40026  -2.331  0.01975 *
Base         0.88444    0.13110   6.746 1.52e-11 ***
V4          -0.16109    0.05463  -2.948  0.00319 **
Base:Trt     0.33822    0.20315   1.665  0.09594 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
          (Intr) Age    Trt    Base   V4
Age       -0.976
Trt        0.046 -0.191
Base      -0.163 -0.038  0.596
V4        -0.010  0.000  0.000  0.000
Base:Trt -0.118  0.253 -0.930 -0.654  0.000
> |
```

On obtient des résultats qui sont très proches aux originaux (Annexe VI a)) comme on voit sur le tableau qui suit. En regardant les p-valeur du test de Wald, on remarque qu'`Age` n'a plus d'effet sur la réponse et que l'interaction arrive tout juste à être jugée comme ayant un effet significatif.

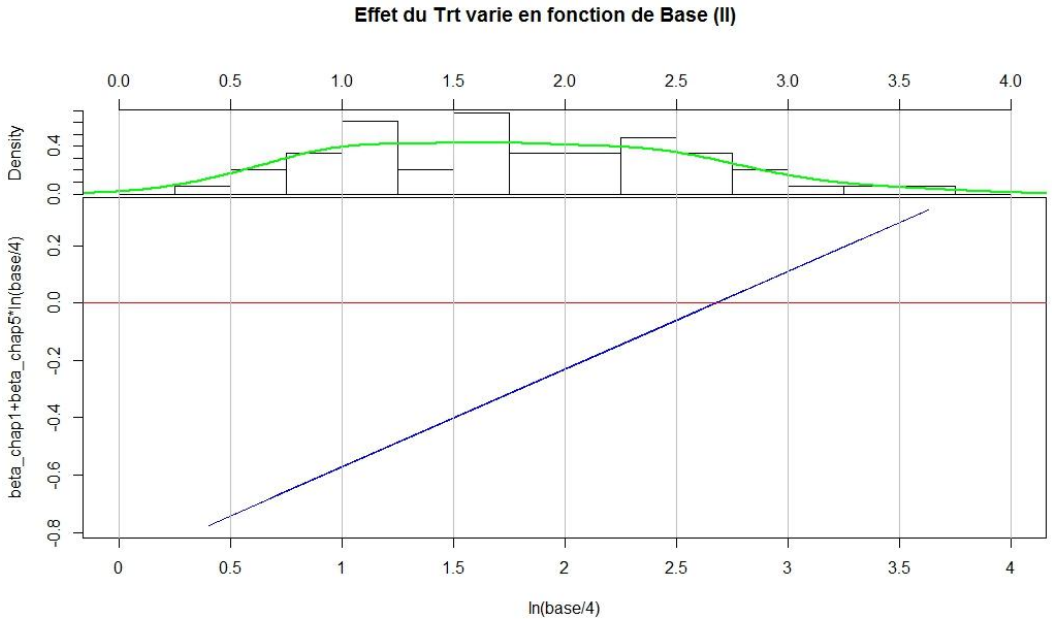
Pour regarder maintenant si l'effet-sujet a une distribution normale, on peut extraire leurs estimations à l'aide de la fonction `ranef` de R. Puis on fait un QQ-plot et effectue le test de Shapiro-Wilk pour vérifier l'hypothèse. Le graphique nous semble suggérer que la normalité n'est pas vérifiée, mais en faisant le test, la valeur de la statistique du test est 0.96 et sa p-valeur correspondante 0.050. Ainsi on peut quand même dire que l'effet aléatoire suit probablement une distribution normale.



En ce qui concerne l'homogénéité des effets-sujets, il est plus difficile à les tester voire impossible.

Parlant maintenant des méthodes de détections d'observations influentes ou aberrantes. En fait si on utilise une des fonctions pour produire des modèles généralisés autre que `glm` ou généralisés mixtes, alors il nous faut le package `{influence.ME}` pour pouvoir calculer la distance de Cook regroupée par individu. Ainsi globalement influent sont jugés les patients à identifiant 135, 207, 227 et 232 car ils dépassent le seuil de $\frac{4}{n-p-1} = \frac{4}{59-5-1} \approx 0.0755$.

En passant à l'interprétation de l'effet du traitement, on fait le même graphique pour les résultats trouvés avant sous 4.3.1..



Effectivement, pour quelqu'un qui a un comptage de référence de 30 crises épileptiques sur une période de huit semaines ($4\exp(2) \cong 30$), le nombre de crises épileptiques d'un patient du groupe Progabide équivaut en moyenne à 0.79 ($\exp(-0.9132463 + 0.3410064 \cdot 2) \cong 0.79$) fois le nombre de crises d'un patient prenant placebo. Progabide aurait donc un effet réducteur de crises. On voit également à l'aide des deux graphiques, que la majorité des patients vont expérimenter une réduction. À partir de 59 comptages de référence, Progabide ne semble plus à être efficace contre les crises épileptiques. 6 des 59 patients (environ 10% de notre échantillon) se trouvent alors dans le cas où le traitement n'est pas effectif.

4.3.3. GLMM Poisson à deux effets aléatoires

Le **troisième modèle (III)** est un GLMM Poisson à deux effets aléatoires indépendants u_i et v_{it}. u_i reste défini comme pour le modèle d'avant et v_{it} est introduit pour modéliser la variation entre les comptages de chaque visite pour chaque sujet. On va le nommer ici l' « effet du temps » ; il peut néanmoins être vu comme un terme d'erreur supplémentaire pour s'en charger de la sur-dispersion qui n'a pas été entièrement gérée par u_i. Le modèle s'écrit :

$$\ln(\mathbb{E}[Y_{it}|X_i^{(1)}, X_i^{(2)}, X_i^{(3)}, X_t^{(4)}, X_i^{(2)}X_i^{(1)}, u_i, v_{it}]) = \beta_0 + \beta_1 1_{\{X_i^{(1)}=1\}} + \beta_2 X_i^{(2)} + \beta_3 X_i^{(3)} + \beta_4 1_{\{X_t^{(4)}=1\}} + \beta_5 X_i^{(2)} 1_{\{X_i^{(1)}=1\}} + u_i + v_{it}$$

Sous R :

```
III<-
glmmPQL(Seizure~Age+Base:Trt+Trt+Base+V4,random=~1|id/Visit,family=poisson,data=BC)
```

`Visit` est une variable qui prend les valeurs {-0.3, -0.1, 0.1, 0.3} pour coder chaque visite, {Visite1, Visite2, Visite3, Visite4} respective.

Voici donc nos résultats obtenus :

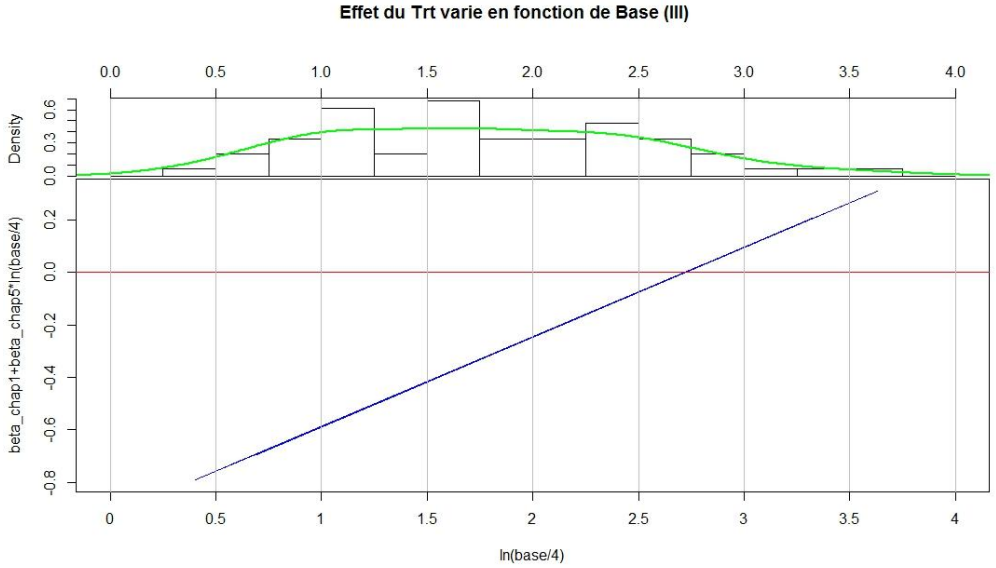
```
Random effects:
 Formula: ~1 | id
        (Intercept)
StdDev:   0.4566422

 Formula: ~1 | Visit %in% id
        (Intercept)  Residual
StdDev:   0.3789165 0.9094633

Variance function:
 Structure: fixed weights
 Formula: ~invwt
Fixed effects: Seizure ~ Age + Base:Trt + Trt + Base + V4
                Value Std.Error  DF   t-value p-value
(Intercept) -1.2679973 1.1602467 176 -1.092869  0.2759
Age          0.4700343 0.3406772  54  1.379706  0.1734
Trt         -0.9298276 0.3914013  54 -2.375637  0.0211
Base         0.8567260 0.1292609  54  6.627882  0.0000
V4          -0.0940711 0.0853002 176 -1.102824  0.2716
Base:Trt     0.3415841 0.1995954  54  1.711383  0.0927
 Correlation:
        (Intr) Age    Trt    Base   V4
Age     -0.976
Trt      0.044 -0.189
Base    -0.163 -0.037  0.597
V4      -0.018  0.001  0.002 -0.003
Base:Trt -0.115  0.250 -0.929 -0.656  0.000

Standardized Within-Group Residuals:
       Min          Q1         Med          Q3         Max
-1.88515519 -0.39973088 -0.03829871  0.33194260  1.27420975

Number of Observations: 236
Number of Groups:
          id Visit %in% id
          59           236
```

Les estimations et erreurs standards sont presque identiques à ceux obtenus par Breslow et Clayton. La plus grande différence trouve-t-on dans l'estimation des variances des effets aléatoires (0.457 ≠ 0.48 et 0.379 ≠ 0.36).

Par ailleurs, si on refait le modèle avec la fonction `glmer` usuelle, l'interaction est quand même jugée à limite significative.

De plus, en procédant par élimination pas-à-pas descendante suivant p-valeur, on obtient le modèle avec que les prédicteurs `Trt`, `V4` et `Base`. L'`intercept` reste non-significatif.

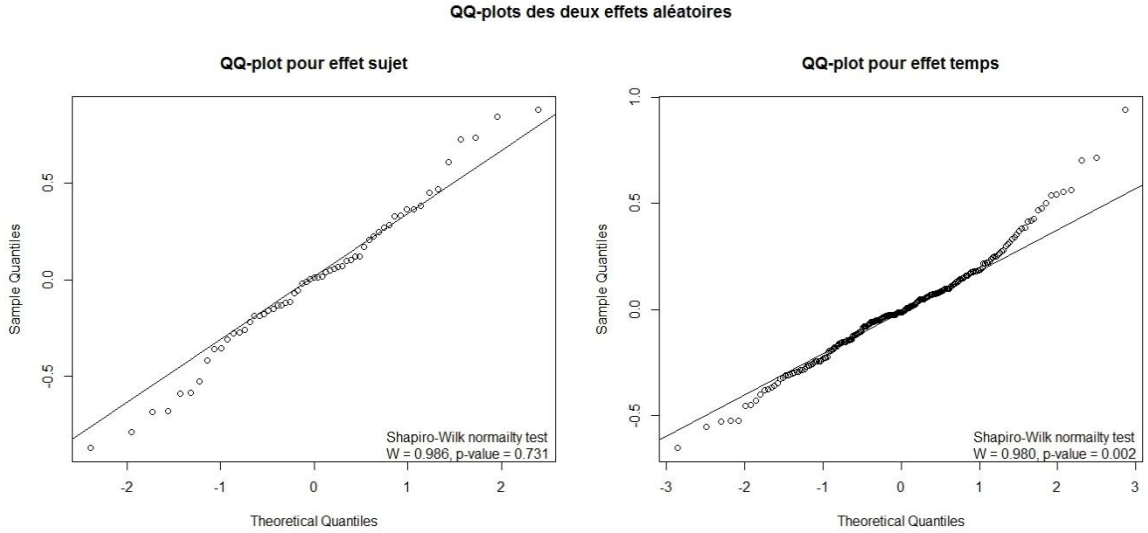
En passant à l'interprétation de l'effet du traitement, on revient à la construction du graphique déjà utilisé avant sous le point précédent.



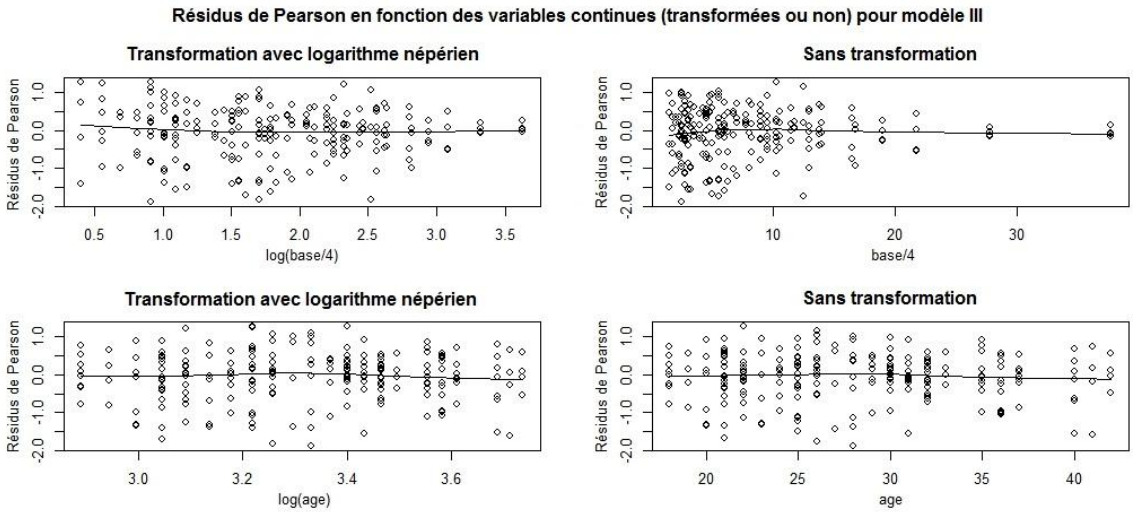
Pour quelqu'un qui a un comptage de référence de 30 $\left(\ln(\frac{30}{4}) \cong 2\right)$ crises épileptiques sur une période de huit semaines, le nombre de crises épileptiques d'un patient du groupe Progabide équivaut en moyenne à 0.78 fois $(\exp(-0.9298276 + 0.3415841 \cdot 2) \cong 0.78)$ le nombre de crises d'un patient prenant un placebo. Progabide aurait donc un effet réducteur de crises. On voit également à l'aide des deux graphiques, que la majorité des patients vont expérimenter une réduction. A partir de 61 comptages de référence, Progabide ne semble plus être efficace

contre les crises épileptiques. 6 des 59 patients (environ 10% de notre échantillon) se trouvent alors dans le cas où le traitement n'est pas effectif.

Comme on suppose que les effets aléatoires suivent tous une loi normale centrée et à variance constante, on peut tester si cette hypothèse est vérifiée. On voit bien sur les graphiques que la normalité à nouveau ne semble pas très évidente. Néanmoins, pour les estimations des effets sujets, le test de Shapiro-Wilk nous suggère qu'ils sont probablement distribués normalement. Par contre le test de normalité pour les estimations de l'effet de temps rejette l'hypothèse nulle de la normalité (W=0.980 et p-valeur=0.002). On n'a donc pas l'évidence que ces effets aléatoires du temps sont distribués normalement mais ça ne veut pas dire qu'on a évidence qu'ils ne le sont pas. Si on voit cette variable aléatoire comme étant les résidus, alors on a le droit de ne pas s'attendre leur normalité. Compte à l'hétérogénéité, il est compliqué à la tester.



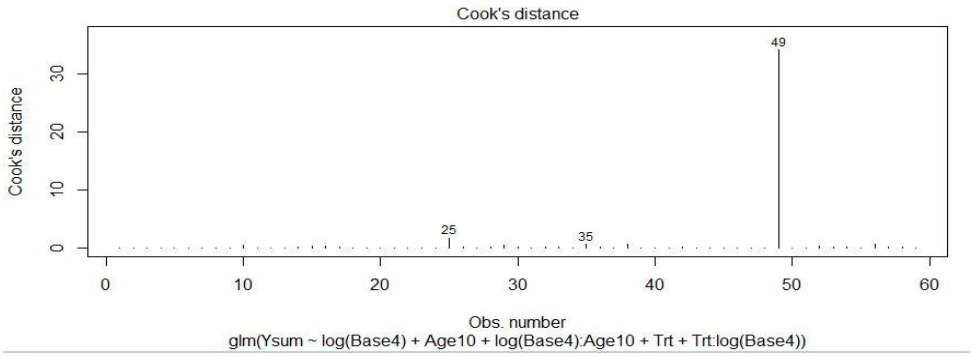
Intéressée à nouveau aux changements entraînés par la transformation des variables, on refait les graphiques des résidus de Pearson en fonction des variables continues transformées et non. La transformation log népérienne améliore à nouveau l'aspect du graphique correspondant aux comptages initiaux même si on pense toujours reconnaitre une espèce de structure d'entonnoir qui suggère de l'hétéroscédasticité. Le graphique lié à la variable correspondant à l'âge ne change gère quelque chose. Intéressant à voir est que la variable des comptages initiaux est la seule à effet significatif dans le modèle sans transformations.



4.4. Changeons de réponse (Breslow 1996)

Maintenant on ne veut plus différencier entre les différents comptages par période de deux semaines. On s'intéresse au total de nombres de crises épileptiques pendant la période de 8 semaines. Ainsi, on prend Y_i comme réponse, définie avec les notations d'avant : $Y_i = \sum_{t=1}^{4} Y_{it}$ pour i=1, …, 59. On se retrouve alors dans le cas où on a 59 observations supposées indépendantes. Les résultats originaux se trouvent dans l'Annexe VI b). Le jeu de données avec lequel on fait nos calculs dans R s'appelle `epilepsy` et provient du package R `{robustbase}`.

Comme on a déjà vu sous 4.1., le patient 207 a toujours les comptages de crises épileptiques les plus extrêmes parmi tous les patients. En faisant un GLM de Poisson naïve, on peut calculer les distances de Cook qui montrent que l'observation 49 (patient à `ID` 207) a un effet influent globalement et d'où la justification de Breslow de l'exclure (Graphique similaire pour régression binomiale négative) :



On a les notations suivantes avec i=1, …, 58 :

$Y_i = Ysum =$ "nombre total de crises épileptiques du patient i"

$X_i^{(1)} = Trt =$ "Traitement du i^{ième} patient" $= \begin{cases} 0\ pour\ placebo \\ 1\ pour\ Progabide \end{cases}$

$X_i^{(2)} = \log(Base4)$
$=$ logarithme népérien d'un quart du comptage sur 8 semaines avant l'étude

$X_i^{(3)} = Age10 =$ "Âges des patients divisés par 10"

4.4.1. Régression de Poisson

Le **premier modèle (breslow_poiss)** est le suivant.

$$\ln(\mathbb{E}[Y_i|\ prédicteurs]) = \beta_0 + \beta_1 1_{\{X_i^{(1)}=1\}} + \beta_2 X_i^{(2)} + \beta_3 X_i^{(3)} + \beta_4 1_{\{X_i^{(1)}=1\}} X_i^{(3)} + \beta_5 X_i^{(2)} X_i^{(3)}$$

On applique une régression de Poisson simple dont l'écriture sous R est :

```
breslow_poiss<-glm(Ysum~log(Base4)+Age10+log(Base4):Age10+ Trt+
                Trt:log(Base4),family=poisson,data=Thall_epilepsy_sans)
```

Ce modèle s'obtient également par élimination pas-à-pas descendante basée sur les p-valeurs. Si on choisit par contre la famille quasi-Poisson, il n'y a presque aucun effet significatif et par la méthode pas-à-pas descendante, on retient le modèle à 2 prédicteurs : `log(Base4)` et `Age10` (voir Annexe VI b)).

En reproduisant les calculs de Breslow (1996), on obtient les mêmes résultats et on observe que `log(Base4)` n'a pas d'effet significatif sur la réponse. Pourtant, on ne pouvait pas l'exclure du modèle parce que ses interactions avec `Age10` et `Trt` sont statistiquement significatives.

```
Call:
glm(formula = Ysum ~ log(Base4) + Age10 + log(Base4):Age10 +
    Trt + Trt:log(Base4), family = poisson, data = Thall_epilepsy_sans)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.3533  -1.5596  -0.4326   0.5097   8.9192

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            3.07922    0.45069   6.832 8.36e-12 ***
log(Base4)            -0.07368    0.20141  -0.366 0.714489
Age10                 -0.51074    0.15329  -3.332 0.000863 ***
Trtprogabide          -0.61042    0.19098  -3.196 0.001392 **
log(Base4):Age10       0.35071    0.06791   5.164 2.42e-07 ***
log(Base4):Trtprogabide  0.20446    0.08796   2.324 0.020099 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1284.72  on 57  degrees of freedom
Residual deviance:  408.41  on 52  degrees of freedom
AIC: 696.12

Number of Fisher Scoring iterations: 5
```

Par défaut, la fonction `glm(.)` estime les coefficients du modèle par la méthode des moindres carrées re-pondérés itérativement (IRLS) (voir Annexe V a) pour une brève explication).

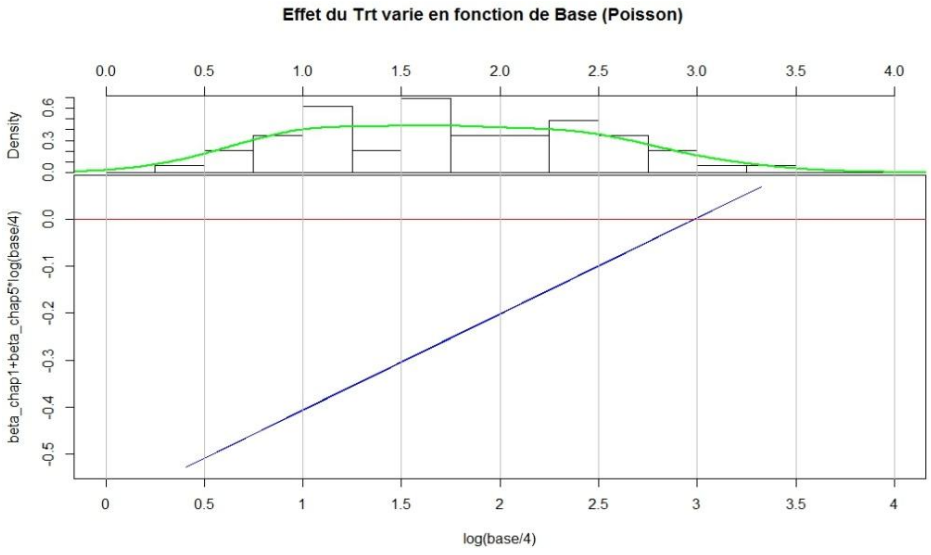
De plus, en prenant le nombre total de crises épileptiques, on s'est occupé du problème de l'indépendance des réponses. Néanmoins, en comparant la déviance résiduelle avec son degré de liberté, on constate toujours un problème de sur-dispersion. Il doit donc exister encore une autre source qui entraîne ce problème.

Testant alors si la présence de sur-dispersion est statistiquement significative en appliquant un test de Dean (1992) comme sous 4.3.1. La statistique vaut cette fois-ci :

$$T = \frac{\sum_{j=1}^{59}\left\{\left(y_j - \hat{\mu}_j\right)^2 - y_j\right\}}{\sqrt{2\sum_{j=1}^{59}\hat{\mu}_j^2}} \text{ où } \hat{\mu}_j = \exp\left(\mathbb{X}_j\beta\right) \ (cas\ poissonien\ sous\ H_0)$$

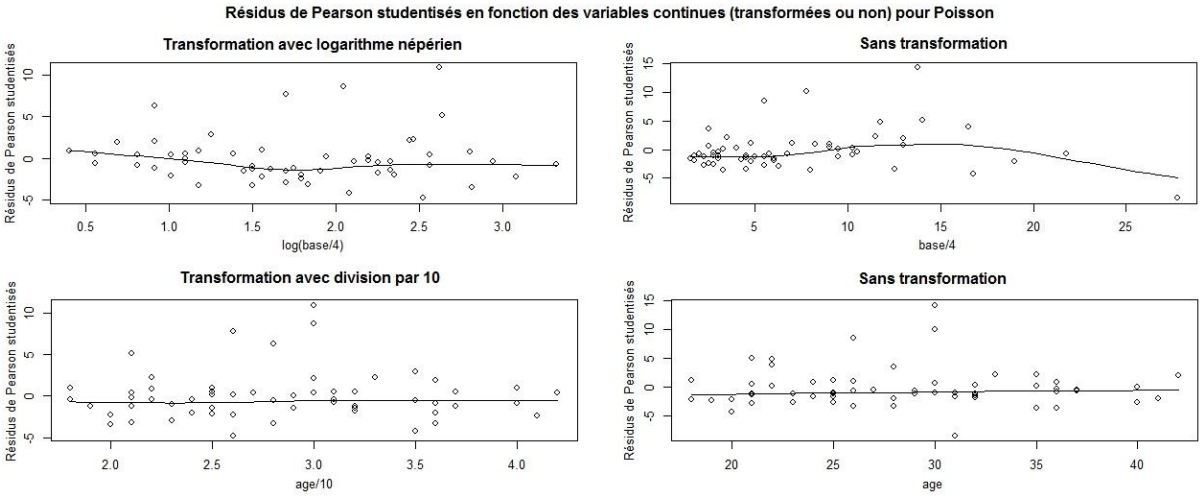
T vaut approximativement 35.9 (p-valeur< $2.2 \cdot 10^{-16}$). L'hypothèse nulle est donc rejetée et on peut affirmer qu'on est très probablement en présence de sur-dispersion. La valeur estimée du paramètre de sur-dispersion $\hat{\phi}$ vaut approximativement 8.78. Un modèle de Poisson pour ces données n'est donc probablement pas un très bon choix.

En ce qui concerne l'effet du traitement, on produit à nouveau le même genre de graphique déjà utilisé plusieurs fois avant.



On constate que l'équation nous donne essentiellement des valeurs négatives ce qui est une bonne nouvelle pour l'effet du traitement. Ainsi pour quelqu'un qui a un comptage de référence de 30 crises épileptiques sur une période de huit semaines ($4\exp(2) \cong 30$), le nombre de crises épileptiques d'un patient du groupe Progabide équivaut en moyenne à 0.82 fois ($\exp(-0.61042 + 0.20446 \cdot 2) \cong 0.82$) le nombre de crises d'un patient prenant un placebo. Progabide aurait donc un effet réducteur de crises. On voit également à l'aide des deux graphiques, presque tous les patients vont expérimenter une réduction. À partir de 80 comptages de référence, Progabide ne semble plus être efficace contre les crises épileptiques. 2 des 58 patients inclus dans notre étude (environ 3.4% de notre échantillon) se trouvent alors dans ce cas où le traitement n'est plus effectif.

Si on regarde les résidus de Pearson studentisés tracés en fonction des variables continues transformées ou non, on remarque que la transformation log népérienne a amélioré l'aspect de la courbe. Néanmoins, on constate qu'on se trouve en présence de quelques outliers. De plus, on voit que la transformation de la variable correspondant à l'âge n'entraîne presque pas de changement d'où elle peut être considérée comme étant inutile. En ce qui concerne les résultats de la régression faite avec le modèle sans transformations, tous les prédicteurs ont été jugés comme ayant un effet significatif sur la réponse. Par contre le modèle avec les transformations est mieux adapté aux données ce qu'on déduit de la valeur du AIC qui est plus petit par rapport à celui du modèle sans transformations (696.12 contre 784.96).



4.4.2. Binomiale négative

Le **deuxième modèle (breslow_nb_glm)** proposé par Breslow exige une loi binomiale négative avec l'idée de se charger de la sur-dispersion à l'aide du paramètre supplémentaire de cette loi. On applique exactement le même modèle qu'auparavant à l'exception du fait que $Y_i|\mathbb{X}_i$ suit maintenant une loi binomiale négative au lieu d'une Poisson. On tape alors dans R :

```
breslow_nb<-glm.nb(formula=Ysum~log(Base4)+Age10+log(Base4):Age10+Trt+
            Trt:log(Base4),data=Thall_epilepsy_sans)
```

La fonction `glm.nb` est basée sur la théorie décrite dans *Venables, W. N. et Ripley, B. D. (2002)* Modern Applied Statistics with S. *Fourth edition. Springer.* et se trouve dans le package R `{MASS}`. En effectuant le `summary`(), cette fonction nous fournit également le paramètre `theta` qui est en fait l'inverse du paramètre de dispersion. Les valeurs de Breslow pour ce paramètre

ne coïncident pas avec les nôtres. On conclut alors que `glm.nb` n'estime pas le paramètre de dispersion par la méthode des moments. Sur la page d'aide de R on nous dit que le paramètre est estimé par des itérations de score et d'information et est retenu quand les deux convergent. Evidemment, comme on a trouvé un autre paramètre de dispersion, nos estimations et erreurs-standards varient par rapport à celles de Breslow.

Dans le package `{MASS}` de R, on peut trouver une fonction `theta.mm` qui estime le `theta` par la méthode des moments. On le sauvegarde dans `inv_disp_para` :

```
inv_disp_para<-theta.mm(Ysum,fitted(breslow_nb),dfr=df.residual(breslow_nb))
```

Il s'agit effectivement de la valeur 3.303, dont l'inverse vaut 0.3027. Comme on connait la valeur de `theta`, on peut maintenant faire appel à la fonction `glm` :

```
breslow_nb_glm<-glm(Ysum~log(Base4)+Age10+log(Base4):Age10+Trt+
              Trt:log(Base4),family=negative.binomial(theta=inv_disp_para),
              data=Thall_epilepsy_sans)
```

Ce modèle n'est quand même pas préservé si on procède par la méthode pas-à-pas descendante basée sur la p-valeur à partir du modèle complet à interactions d'ordre 2. Le modèle final obtenu par cette méthode est constitué de deux prédicteurs sans aucune interaction : `log(Base4)` et `Trt` tous les deux significatif. L'`intercept` lui aussi est significatif.

On se concentre néanmoins sur le modèle de Breslow et on obtient quasi les mêmes résultats que ceux évoqués dans son article :

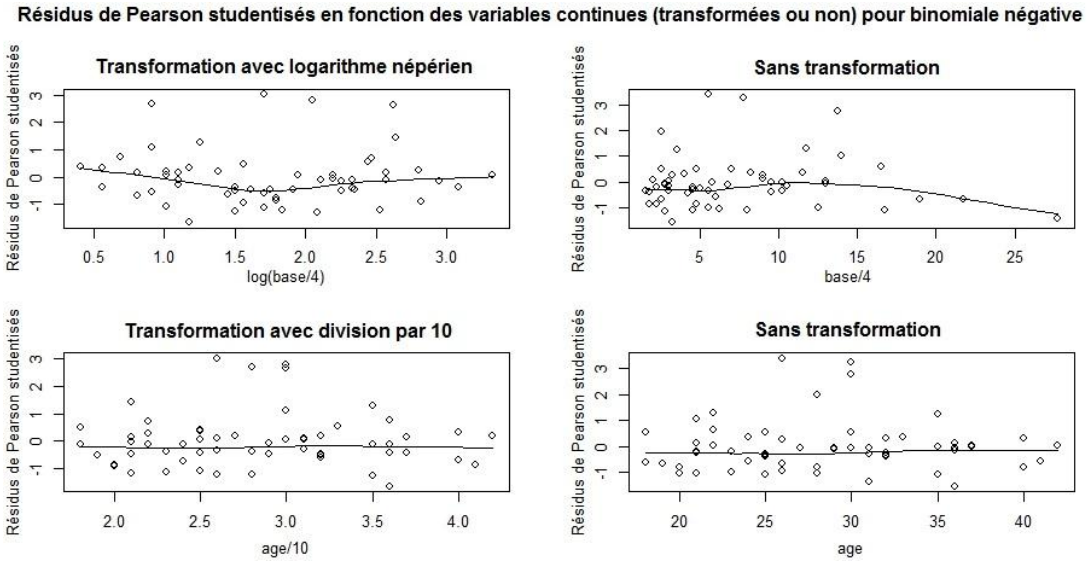
```
Call:
glm(formula = Ysum ~ log(Base4) + Age10 + log(Base4):Age10 +
    Trt + Trt:log(Base4), family = negative.binomial(theta = inv_disp_para),
    data = Thall_epilepsy_sans)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0495  -0.5571  -0.1094   0.2227   2.0977

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.880157   1.098469   2.622   0.0114 *
log(Base4)            0.002866   0.572380   0.005   0.9960
Age10                -0.399131   0.369311  -1.081   0.2848
Trtprogabide         -0.702344   0.452637  -1.552   0.1268
log(Base4):Age10      0.302876   0.191303   1.583   0.1194
log(Base4):Trtprogabide  0.248120   0.240397   1.032   0.3068
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(3.3031) family taken to be 0.9994151)

    Null deviance: 138.439  on 57  degrees of freedom
Residual deviance:  51.728  on 52  degrees of freedom
AIC: 455.29

Number of Fisher Scoring iterations: 5
```

On voit bien que le rapport de la déviance résiduelle et son degré de liberté vaut presque 1 ce qui nous suggère que la loi binomiale négative a bien géré le problème de sur-dispersion. Un aspect négatif de ce modèle est quand même qu'il trouve qu'aucun prédicteur n'a un effet statistiquement significatif. La situation souhaitable est telle qu'au moins le facteur traitement est jugé d'avoir un effet significatif sur la réponse, ce qui n'est quand même pas le cas ici.

Nos résultats nous suggèrent que la prise de Progabide n'a pas d'effet statistiquement significatif sur le nombre de crises épileptiques. Ce modèle est le mieux adapté aux données ce qui nous indique l'AIC et donc on fait plutôt confiance à ce dernier modèle. Ainsi on maintient

notre interprétation : Progabide n'a probablement pas d'effet sur le nombre de crises épileptiques.

Comme déjà sous des points avant, on remarque que le graphique avec transformation des comptages de références constituent une amélioration tandis que le graphique concernant la variable associée à l'âge reste quasi inchangé. Même si on obtient plus d'effets significatifs avec le modèle sans transformations (`intercept`, `Trt` et l'interaction entre base/4 et âge), le critère AIC nous suggère que le modèle avec les transformations est le meilleur car son AIC vaut 455.29 contre 460.61 de celui du modèle sans transformations.



Breslow a choisi d'exclure les observations du patient 207. Si on les garde, avec le modèle utilisant une loi binomiale négative, on obtient que l'`intercept`, `Trt` et l'interaction de `log(Base4)` et `Trt` sont significatifs. Il y a quand même de forte différence en ce qui concerne l'estimation des paramètres, mais des erreurs standards très similaires. La sur-dispersion semble bien prise en compte avec une déviance résiduelle de 53.471 et 53 degrés de libertés.

4.5. La situation binaire (Sinha and Xu 2011)

Sinha et Xu ont analysé le jeu de données des comptages de crises épileptiques avec une réponse binaire et avec des variables explicatives majoritairement binaires. Ils ont introduit les éléments suivants où i=1, … , 59 et t=1, … , 4 :

$$Y_{it} = y_{bin} = \begin{cases} 0 \; si \; le \; nombre \; de \; comptages \; du \; ième \; patient \; à \; la \; visit \; t \leq 10 \\ 1 \; sinon \end{cases}$$

$$X_i^{(1)} = baseline_{bin} = \text{"Comptages de référence binaire (sur 8 semaines) "}$$

$$= \begin{cases} 0 \; si \; comptage \; baseline \; sur \; 8 \; semaines \; \leq 40 \\ 1 \; sinon \end{cases}$$

$$X_i^{(2)} = trt = \text{"Traitement"} = \begin{cases} 0 \; pour \; placebo \\ 1 \; pour \; Progabide \end{cases}$$

$$X_t^{(3)} = visit2 = \text{"Variable indiquant dans les différentes visites"}$$

$$= \begin{cases} 2 \; pour \; indiquer \; la \; deuxième \; semaine \; (donc \; première \; visite) \\ 4 \; pour \; indiquer \; la \; quatrième \; semaine \; (donc \; deuxième \; visite) \\ 6 \; pour \; indiquer \; la \; sixième \; \; semaine \; (donc \; troisième \; visite) \\ 8 \; pour \; indiquer \; la \; huitième \; semaine \; (donc \; quatrième \; visite) \end{cases}$$

$A_i = variable\ aléatoire\ dû\ à\ l'individu$

Les variables aléatoires A_i sont supposées indépendantes et identiquement distribuées suivant une loi normale d'espérance zéro et de variance σ_A^2. On va appliquer une régression logistique.

Les résultats obtenus par Sinha et Xu se trouvent en Annexe VI c) formatés en tableau.

Sous R, on effectue la commande suivante :

```
sinha_m<-glmer(y_bin~baseline_bin+trt+visit2+(1|subject),family=binomial,
         data=data_sinha,nAGQ=20)
```

En effet, la fonction R `glmer` effectue une régression pour des modèles linéaires généralisés à effets mixtes et avec l'option `family=binomial`, on effectue une régression logistique.

```
Generalized linear mixed model fit by the adaptive Gaussian Hermite approximation
Formula: y_bin ~ baseline_bin + trt + visit2 + (1 | subject)
   Data: data_sinha
   AIC   BIC logLik deviance
 140.2 157.5 -65.11    130.2
Random effects:
 Groups  Name        Variance Std.Dev.
 subject (Intercept) 9.7073   3.1156
Number of obs: 236, groups: subject, 59

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.8515     1.3068  -2.947  0.00321 **
baseline_bin   6.9107     1.3454   5.136  2.8e-07 ***
trtprogabide  -0.6347     1.3354  -0.475  0.63461
visit2        -0.2852     0.1349  -2.113  0.03457 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) bsln_b trtprg
baseline_bn -0.530
trtprogabid -0.491  0.042
visit2      -0.433 -0.115  0.010
```

Comme Sinha et Xu l'ont également signalé, le facteur à intérêt principal, `trt`, n'est pas significatif avec ce choix du modèle ; celui-ci nous propose alors la plus mauvaise situation : Le traitement n'a pas d'effet significatif sur le nombre de crises épileptiques.

La problématique principale des modèles à réponse ou prédicteurs binaires à la base numérique est le choix du seuil. Vaut-il mieux prendre la médiane ? La moyenne ?

```
summary(y)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   2.750   4.000   8.254   9.000 102.000
```

```
summary(epilepsy$Base)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 6.00   12.00   22.00   31.22   41.00  151.00
```

Les informations ci-dessus nous montrent qu'en fait, Sinha et Xu n'ont pris aucune valeur de seuil qui peut être lu des résumés sur les valeurs des comptages de crises épileptiques avant ou durant l'étude. Les valeurs sont néanmoins proches du 3^{ième} quartile, mais une solide justification pour ce choix n'existe pas.

Par ailleurs, en nous restreignant à une réponse binaire, on doit s'accommoder du fait qu'on perd une grande partie des informations contenue dans le jeu de données.

De plus, au niveau des interprétations des résultats obtenue par une telle situation binaire dans le domaine du développement de médicaments, il n'est pas du tout évident de tirer des conclusions satisfaisantes et très informative. L'approche « C'est soit vrai soit faux » est mal placée dans ce domaine.

Chapitre 5
Discussion : Comparaison des méthodes

Dans ce chapitre on va comparer les résultats des modèles différents qu'on vient de voir.

`odTest` du R package `{pscl}` effectue un test de khi-deux de rapport de vraisemblance (eng. *Chi square likelihood ratio test*). Il nous montre que pour Breslow (1996), la négative binomiale est le meilleur choix comparé à la régression de Poisson.

```
> odTest(breslow_nb)
Likelihood ratio test of H0: Poisson, as restricted NB model:
n.b., the distribution of the test-statistic under H0 is non-standard
e.g., see help(odTest) for details/references

Critical value of test statistic at the alpha= 0.05 level: 2.7055
Chi-Square Test Statistic =  242.0743 p-value = < 2.2e-16
```

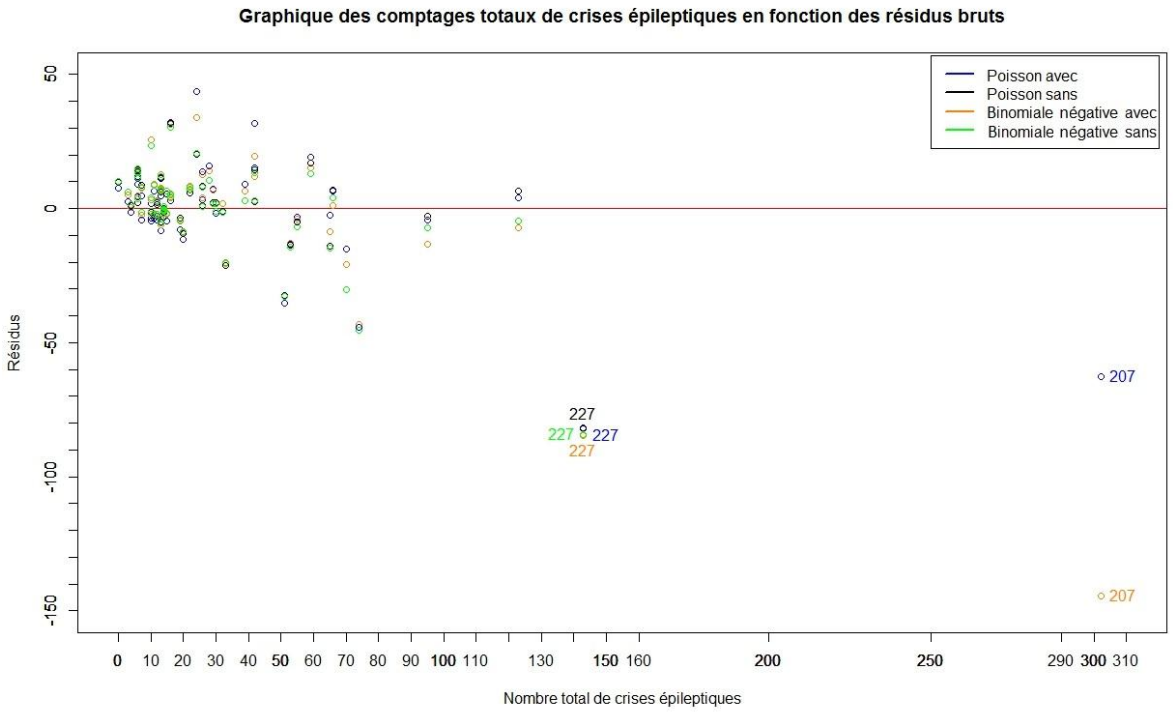
En théorie, la statistique du test est la suivante :

$$LR = -2\ln\left(\frac{\mathcal{L}_{\mathcal{P}}(Y,\mu)}{\mathcal{L}_{\mathcal{NB}}(Y,\mu)}\right) = 2(\ln(\mathcal{L}_{\mathcal{NB}}(Y,\mu)) - \ln(\mathcal{L}_{\mathcal{P}}(Y,\mu)))$$

Où $\mathcal{L}_{\mathcal{NB}}(Y,\mu)$ et $\mathcal{L}_{\mathcal{P}}(Y,\mu)$ sont les vraisemblances du modèle négative binomiale et du modèle de Poisson respective. Avec la fonction `logLik` de `{stats}` de R peut-on déterminer la log-vraisemblance des modèles. La statistique LR suit une loi de chi-deux à 1 degrés de liberté (car la différence entre les degrés de liberté d'une modèle de Poisson et d'un modèle négative binomiale vaut 1). En ce qui concerne le choix du meilleur des deux modèles par le critère d'information d'Akaike (AIC), on observe que celui utilisant une loi binomiale négative est beaucoup plus petit (455.29 contre 696.12 du modèle de Poisson). On peut donc dire que le modèle à loi binomiale négative est probablement le plus adéquat des deux. Intéressant est quand même le fait que aucun effet est statistiquement significatif et qu'il s'agit quand même du meilleur modèle pour ces données. Ainsi on conclut que le traitement n'a pas d'effet statistiquement significatif sur le nombre de crises épileptiques.

Une autre façon de regarder à nos résultats et de comparer les comptages originaux avec les prédictions transformées des modèles. On a fait ceci pour les modèles de Breslow et Clayton (1993) ainsi que pour Breslow (1996). Le livre de Pinheiro and Bates (2000) m'a beaucoup aidé à élaborer ces résultats. Le prédicteur transformé est $\hat{Y} = g^{-1}(\mathbb{X}\hat{\beta} + \mathbb{Z}\hat{\mathbb{U}})$ soit $\hat{Y} = g^{-1}(\mathbb{X}\hat{\beta})$ s'il n'y a pas d'effets aléatoires inclus. Une partie de ces prédicteurs transformés obtenue par les modèles de Breslow et Clayton (1993) (4 comptages prédites de 6 patients) et Breslow (1996) (les 13 patients du groupe placebo et 13 du groupe Progabide) se trouve en Annexe VII.

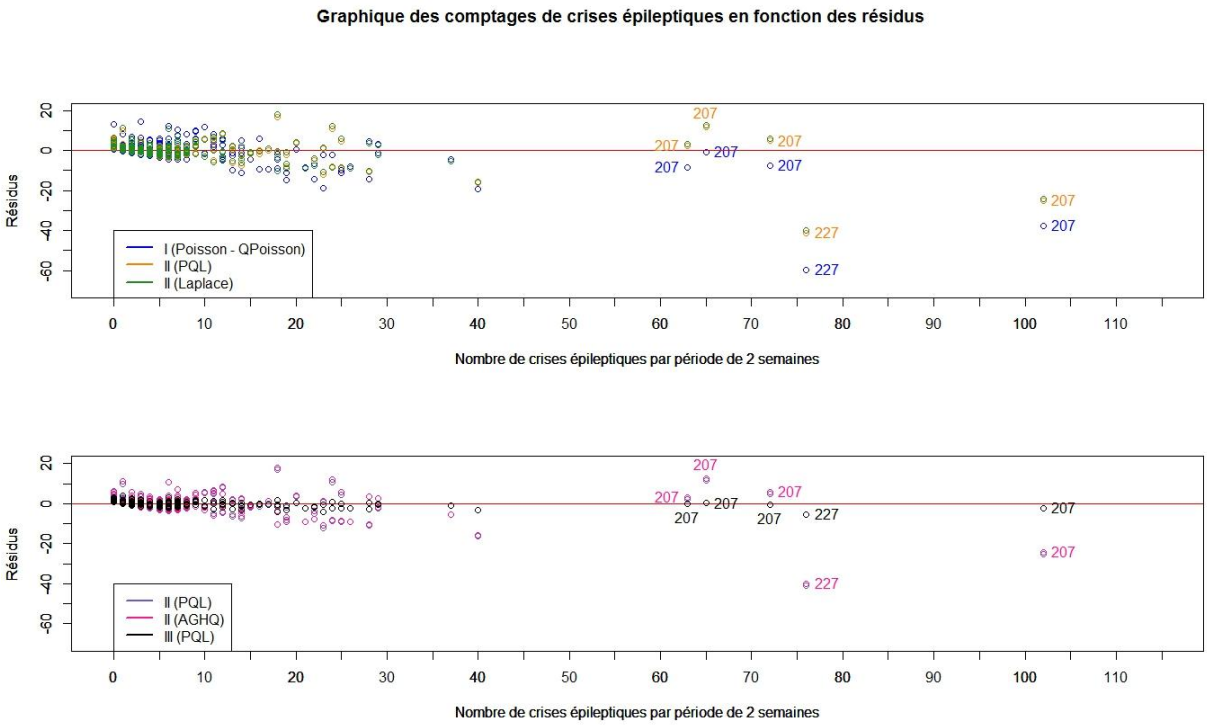
Passons maintenant au tableau des prédictions transformées produites pour les modèles de Breslow (Annexe VII b)). On y retrouve les comptages originaux, les prédictions transformées à l'aide d'un modèle de Poisson et les prédictions transformées obtenues à l'aide d'un modèle utilisant une loi binomiale négative. Les deux premières colonnes à gauche illustrent le jeu de données complet contrairement à la partie à droite où le patient 207 a été exclu. Dans ce tableau il est difficile de conclure comme les valeurs varient parfois légèrement et parfois considérablement des comptages initiaux comme on voit bien sur les graphiques en bas. Ce qui est pourtant évident est que les valeurs avec ou sans le patient à `ID` 207 diffèrent, comme on voit par exemple dans le tableau ligne 4, 11 ou encore 18. Il s'agit donc d'un autre indice de l'influence du patient 207 sur les calculs.



En ce qui concerne les modèles de Breslow et Clayton, il ne s'agit plus d'un secret qu'une régression de Poisson naïve comme sous 4.3.1. n'est pas du tout adaptée à ce jeu de données. La condition d'indépendance entre les 236 données n'est pas assurée d'où une violation d'une condition indispensable. Par contre on observe qu'au fur et à mesure qu'on introduit des variables aléatoires qui se chargent des variations non expliquées par le modèle de Poisson simple, le modèle se rapproche de mieux en mieux à nos données.

À l'aide du tableau en Annexe VII a), on voit bien que les modèles de Poisson et Quasi-Poisson de Breslow et Clayton prédisent très mal les comptages de crises. En revanche, dès qu'on introduit un effet aléatoire, on s'approche de plus en plus des vraies valeurs. La meilleure prédiction est finalement atteint par l'introduction de deux effets aléatoires, les valeurs ne varient que très peu par rapport aux comptages originaux comparé aux autres résultats trouvés. On atteint une bonne approximation pour des grandes et petites valeurs.

On peut observer ceci aussi sur le graphique suivant :



On voit bien qu'au début, pour les petits comptages de crises épileptiques, les prédictions transformées sont à peu près proches de zéro et alternent de l'une et l'autre côté de zéro. Les comptages de zéro sont pourtant toujours surestimés. Néanmoins cette impression généralement bonne pour les petits comptages, on observe une sous-estimation croissante de façon continue s'il s'agit de prédire les comptages de crises épileptiques élevées. Surtout les estimations des comptages obtenues par maximum de vraisemblance à la régression de Poisson (I) sont fortement sur- ou sous-estimées par rapport aux autres méthodes. Par contre les comptages estimés par la méthode de quasi-vraisemblance pénalisée (II PQL), la méthode de Laplace (II Laplace) et la méthode adaptive de la quadrature de Gauss-Hermite (II AGHQ) sont presque identiques. Le meilleur modèle est quand même modèle III qui avec la méthode d'estimation de quasi-vraisemblance pénalisée ne montre que très peu de variation par rapport aux comptages originaux. Les résidus bruts sont tous très proche de zéro indépendamment de la valeur du comptage.

Si on devrait choisir entre l'approximation de Laplace et la méthode de quadrature adaptive de Gauss-Hermite (AGHQ), on choisirait cette dernière. Effectivement, on obtient une meilleure approximation avec AGHQ comme on utilise plusieurs points de quadrature. Malheureusement cette meilleure précision a son prix : perte de rapidité d'effectuer les calculs.

Pour conclure ce chapitre, je choisissais le modèle III de Breslow et Clayton (1993) GLMM à deux effets aléatoires comme étant le meilleur car il s'adapte le mieux à nos données et semble donc bien gérer la sur-dispersion. Ainsi l'effet du traitement serait significatif et donc Progabide un réducteur du nombre de crises épileptiques.

Chapitre 6

Points de départ pour des recherches supplémentaires

En faisant mes recherches je suis tombée encore sur d'autres problématiques de ce jeu de données qui ont été traité par quelques scientifiques. Par manque de temps, je n'ai pas pu approfondir la lecture sur ces sujets néanmoins notables.

En traitant des données de comptages rares, il faut être prudent avec le nombre de zéros. Si ce nombre est élevé, il vaut mieux d'utiliser des modèles modifiés en zéros. Hall and Zhang (2004) par exemple les ont appliqués à des données en cluster. Dans leur article ils ont fait une étude de simulation avec un jeu de données qui avait la même structure que le nôtre. Ils ont procédé par une application des équations d'estimation généralisées (GEE) (eng. *generalized estimation equations*) directe avec des différentes structures de corrélation pour comparer cette méthode à la variante GEE avec un *expectation-solution* (ES) algorithme. Ainsi ils ont choisi un modèle de Poisson modifié en zéro (ZIP) (eng. *zero inflated Poisson*). Au premier coup d'œil, on ne remarque que 23 zéros sur 236 observations respectivement 1 sur 59 si on considère la somme des crises épileptiques des quatre visites. Il est donc difficile à expliquer pourquoi Hall et Zhang se sentaient inspirer du jeu de données sur les crises épileptiques pour tester leur modèle et méthodes pour un nombre de zéros excessif. Un autre modèle pour traiter ces phénomènes rares est modèle binomiale négatif modifié en zéro (ZINB) (eng. *Zero inflated negative binomial*).

Comme on l'a déjà mentionné sous 4.1., on pourrait s'intéresser à analyser si on a un paramètre de sur-dispersion qui dépend du temps. Ye, Yue et al. (2013) se sont consacrés à cette thématique et ont élaboré une théorie comment tester et modéliser cette dépendance. Sous ce même point, on a également évoqué la possibilité de non homogénéité entre les sujets et entre les 4 comptages d'un sujet. On fait donc la référence à Paul and Azad (2012) qui ont évalué des tests d'homogénéité pour des données longitudinales (en cluster) de comptage avec sur-dispersion.

Pour mieux tenir compte des interactions lors de l'interprétation des effets principaux, le centrage des covariables continues incluses dans une interaction peut être proposé. De l'un côté, on s'occupe ainsi en partie de la forte corrélation entre les prédicteurs principaux constituant l'interaction et l'interaction elle-même. Cette corrélation, si ne pas prise en compte, entraine souvent des fausses significativités ou fausses non-significativités des effets principaux. Effectivement, pour le modèle poissonien de Breslow (1996) par exemple, l'effet principal `log(Base4)` n'était pas significatif tandis que son interaction avec `Trt` l'était. Après

le centrage des variables correspondant à l'âge et à la base, j'ai refait les calculs et on trouve que tous les effets sont significatifs. Par ailleurs, on observe qu'à part des deux interactions, les estimations et les erreurs-standards des effets principaux ont changé.

De plus, durant l'interprétation de l'effet principal `Trt`, `log(Base4)` est supposé d'être zéro de même que `Age10`. Pourtant la variable `log(Base4)` ne prend jamais la valeur 1 dans notre jeu de données et aucun patient n'a l'âge 0 d'où ces suppositions ne font pas trop de sens. Le grand avantage du centrage est alors de comparer Progabide et placebo à la moyenne de `log(Base4)` et d'`Age10`. Un article discutant cette problématique en détail est celui de Schielzeth (2010) ou encore un document pdf de Dr. Alan Taylor de Macquarie University, Sydney, Australie avec le titre "Testing and Interpreting Interactions in Regression – In a Nutshell" (Cliquez ici pour accéder au document).

Une autre manière d'analyser ces données longitudinales peut se faire par la théorie sur les séries chronologiques en faisant par exemple attention à l'auto-corrélation ce qui n'a pas été fait dans ce rapport. De plus, on peut essayer de modéliser le jeu de données avec des modèles à différentes structures de la matrice variance-covariance comme par exemple une structure autorégressive du premier ordre (AR(1)) (eng. *first order autoregressive*) ou une structure de symétrie composée (CS) (eng. *compound symmetry*). En SAS, on trouve un grand choix de ces structures dans les fonctions GENMOD ou MIXED. Je fais donc référence à Al-Rawwash and Pourahmadi (2006) qui se sont essentiellement consacrés à cette thématique et ont traité le jeu de données des crises épileptiques.

Analyser la problématique par une méthode bayésienne a été également souvent proposé. Notamment, en appliquant la méthode de Monte Carlo par chaîne de Markov (MCMC). Je cite ici Gamerman (1997) qui s'est intéressé à cette approche et l'a appliqué également au jeu de données des comptages de crises épileptiques.

On pourrait également s'intéresser à tester si la fonction de lien canonique est vraiment un bon choix pour ce jeu de données. Breslow (1996) a proposé un test pour faire ceci dans son article sous le point 7.

Ils existent bien sûr encore beaucoup d'autres façons intéressantes à explorer pour analyser ce jeu de données de Thall et Vail. Dans ce chapitre je n'ai nommé que quelques-unes utilisées par des scientifiques.

Chapitre 7
Conclusions

Pour conclure, on peut dire qu'il est impossible de tirer des vraies conclusions sur l'effet du médicament Progabide à partir de l'article initial de Thall et Vail (1990) ou même des articles le citant. Effectivement, comme le jeu de données et les informations le concernant sont incomplets, on ne fait pas vraiment confiance à nos résultats concernant l'effet du traitement. De plus, le centrage des variables continues pour pouvoir améliorer l'interprétation en présence d'interactions me semble vital.

De même, on a pu observer que la majorité des articles n'ont qu'utilisé le jeu de données des crises épileptiques pour avoir des données à manipuler et ainsi pour pouvoir illustrer leurs méthodes qu'ils croient performant pour ce genre de données longitudinales. Leur but principal n'était pas de mettre en évidence l'effet du médicament. Néanmoins, citant un exemple pour prouver qu'une méthode marche n'est pas très réaliste. Il fallait l'appliquer très souvent à différents jeux de données similaires pour vraiment pouvoir affirmer sa robustesse.

En outre, on a constaté de grandes différences concernant les valeurs d'estimations et erreurs standards lorsqu'on applique différents modèles. La sur-dispersion doit absolument être prise en compte comme le montrent les grands changements des résultats des modèles qui s'en chargent et ceux qui ne le font pas. Les effets aléatoires ont amélioré le modèle qui s'adapte de mieux en mieux aux données avec le nombre croissant d'effets aléatoires. Les transformations des variables continues ont entrainé une amélioration d'adaptation aux données. Le modèle à loi binomiale négative semble une bonne manière de se charger de la sur-dispersion, même si elle n'arrive pas à s'approcher si bien aux données que les modèles mixtes.

Par ailleurs, on a vu que la vérification des conditions d'application des modèles n'est presque jamais traitée dans un article scientifique. Il s'agit d'une pratique commune dans le domaine du développement de médicaments. En général, on ne s'intéresse plus à tester les hypothèses parce qu'on part du principe qu'on s'est préoccupé au long et large avec ce médicament et la mène de l'étude clinique. On devrait donc savoir comment analyser les données et ceci même avant la récolte des données. On suppose qu'on sait ce qu'on veut tester statistiquement et qu'on sait ce qui est la meilleure méthode pour le réaliser. Ceci explique également le comportement envers les justifications manquantes du choix du modèle ou les transformations des prédicteurs qui rendent la reproduction des résultats illustrés difficile. Les principes comment mener une telle étude clinique ainsi que l'évaluation statistique des données sont consignés par écrit dans le document « ICH E9 : Statistical Principles for Clinical Trials » publié par International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH).

Chapitre 8

Références

Al-Rawwash, M. and M. Pourahmadi (2006). "Gaussian estimation and joint modeling of dispersions and correlations in longitudinal data." Computer Methods and Programs in Biomedicine **82**(2): 106-113.

Breslow, N. (1996). "Generalized linear models: checking assumptions and strengthening conclusions." Statistica Applicata **8**: 23-41.

Breslow, N. E. and D. G. Clayton (1993). "Approximate inference in generalized linear mixed models." Journal of the American Statistical Association **88**(421): 9-25.

Dean, C. B. (1992). "Testing for overdispersion in Poisson and binomial regression models." Journal of the American Statistical Association **87**(418): 451-457.

Gamerman, D. (1997). "Sampling from the posterior distribution in generalized linear mixed models." Statistics and Computing **7**(1): 57-68.

Hall, D. B. and Z. Zhang (2004). "Marginal models for zero inflated clustered data." Statistical Modelling **4**(3): 161-180.

Leppik, I., F. Dreifuss, et al. (1987). "A controlled study of progabide in partial seizures Methodology and results." Neurology **37**(6): 963-963.

McCullagh, P. and J. A. Nelder (1989). "Generalized linear models (Monographs on statistics and applied probability 37)." Chapman Hall, London.

Paul, S. and K. Azad (2012). "Testing homogeneity in clustered (longitudinal) count data regression model with over-dispersion." Journal of Statistical Planning and Inference **142**(6): 1608-1618.

Pawitan, Y. (2001). In all likelihood: statistical modelling and inference using likelihood, Oxford University Press.

Pinheiro, J. C. and D. M. Bates (2000). Mixed-effects models in S and S-Plus, Springer-Verlag New York.

Schielzeth, H. (2010). "Simple means to improve the interpretability of regression coefficients." Methods in Ecology and Evolution **1**(2): 103-113.

Sinha, S. K. and X. Xu (2011). "Sequential D-optimal designs for generalized linear mixed models." Journal of Statistical Planning and Inference **141**(4): 1394-1402.

Sun, L. (2011). "Comparison of Different Estimation Methods for Linear Mixed Models and Generalized Linear Mixed Models." 1-30.

Tang, X. (2012). "Two-level lognormal frailty model and competing risks model with missing cause of failure."

Thall, P. F. and S. C. Vail (1990). "Some covariance models for longitudinal count data with overdispersion." Biometrics: 657-671.

Wickham, C. (2013). Wilcoxon Rank Sum Test. Oregon State University.

Ye, F., C. Yue, et al. (2013). "Modeling time-dependent overdispersion in longitudinal count data." Computational Statistics and Data Analysis **58**(1): 257-264.

Annexe I

Successive two-week seizure counts for 59 epileptics. Covariates are adjuvant treatment (0 = placebo, 1 = progabide), eight-week baseline seizure counts, and age (in years).

ID	Y_1	Y_2	Y_3	Y_4	Trt	Base	Age
104	5	3	3	3	0	11	31
106	3	5	3	3	0	11	30
107	2	4	0	5	0	6	25
114	4	4	1	4	0	8	36
116	7	18	9	21	0	66	22
118	5	2	8	7	0	27	29
123	6	4	0	2	0	12	31
126	40	20	23	12	0	52	42
130	5	6	6	5	0	23	37
135	14	13	6	0	0	10	28
141	26	12	6	22	0	52	36
145	12	6	8	4	0	33	24
201	4	4	6	2	0	18	23
202	7	9	12	14	0	42	36
205	16	24	10	9	0	87	26
206	11	0	0	5	0	50	26
210	0	0	3	3	0	18	28
213	37	29	28	29	0	111	31
215	3	5	2	5	0	18	32
217	3	0	6	7	0	20	21
219	3	4	3	4	0	12	29
220	3	4	3	4	0	9	21
222	2	3	3	5	0	17	32
226	8	12	2	8	0	28	25
227	18	24	76	25	0	55	30
230	2	1	2	1	0	9	40
234	3	1	4	2	0	10	19
238	13	15	13	12	0	47	22
101	11	14	9	8	1	76	18
102	8	7	9	4	1	38	32
103	0	4	3	0	1	19	20
108	3	6	1	3	1	10	30
110	2	6	7	4	1	19	18
111	4	3	1	3	1	24	24
112	22	17	19	16	1	31	30
113	5	4	7	4	1	14	35
117	2	4	0	4	1	11	27
121	3	7	7	7	1	67	20
122	4	18	2	5	1	41	22
124	2	1	1	0	1	7	28
128	0	2	4	0	1	22	23
129	5	4	0	3	1	13	40
137	11	14	25	15	1	46	33
139	10	5	3	8	1	36	21
143	19	7	6	7	1	38	35
147	1	1	2	3	1	7	25
203	6	10	8	8	1	36	26
204	2	1	0	0	1	11	25
207	102	65	72	63	1	151	22
208	4	3	2	4	1	22	32
209	8	6	5	7	1	41	25
211	1	3	1	5	1	32	35
214	18	11	28	13	1	56	21
218	6	3	4	0	1	24	41
221	3	5	4	3	1	16	32
225	1	23	19	8	1	22	26
228	2	3	0	1	1	25	21
232	0	0	0	0	1	13	36
236	1	4	3	2	1	12	37

Annexe II

Démontrons :

$$Q_i(\mu_i; y_i) = \frac{1}{\phi} y_i \ln(\mu_i) - \mu_i + constante$$

Par définition de quasi-vraisemblance, on a :

$$\begin{aligned} Q_i(\mu_i; y_i) &= \frac{1}{\phi} \int_{y_i}^{\mu_i} \frac{y_i - t}{t} dt \\ &= \frac{1}{\phi} \left([y_i \ln(t)]_{y_i}^{\mu_i} - [t]_{y_i}^{\mu_i} \right) \\ &= \frac{1}{\phi} (y_i \ln(\mu_i) - y_i \ln(y_i) - \mu_i + y_i) \\ &= \frac{1}{\phi} (y_i \ln(\mu_i) - \mu_i + C) \end{aligned}$$

où C est une constante. On a introduit y_i et $-y_i \ln(y_i)$ dans la partie constante pour alléger l'écriture et principalement parce que ces deux termes vont disparaître lors de la maximisation car il s'agit de termes à valeur constante.

Annexe III

Démontrons :

$$\mathbb{P}(Y = y) = \int_0^\infty \mathbb{P}(Y = y | \Theta = \theta) f_\Theta(\theta) d\theta = \frac{\Gamma(y + \alpha)}{\Gamma(y+1)\Gamma(\alpha)} \left(1 - \frac{1}{1+\beta}\right)^y \left(\frac{1}{1+\beta}\right)^\alpha$$

On a :

$$Y|\Theta = \theta \sim \mathcal{P}(\theta) \text{ où } \Theta \sim \mathcal{G}a(\alpha, \beta)$$

Sa densité jointe est définie par :

$$\mathbb{P}(Y = y | \Theta = \theta) f_\Theta(\theta) = \frac{\theta^y}{y!} \exp(-\theta) \frac{1}{\Gamma(\alpha)\beta^\alpha} \theta^{\alpha-1} \exp\left(-\frac{\theta}{\beta}\right) 1\{\theta > 0\}.$$

Donc la probabilité non conditionnelle vaut :

$$\begin{aligned} \mathbb{P}(Y = y) &= \int_0^\infty \mathbb{P}(Y = y | \Theta = \theta) f_\Theta(\theta) d\theta \\ &= \frac{1}{\Gamma(\alpha)\, \beta^\alpha\, y!} \int_0^\infty \theta^{y+\alpha-1} \exp\left(-\theta \left(1 + \frac{1}{\beta}\right)\right) d\theta \\ &= \frac{1}{\Gamma(\alpha)\Gamma(y+1)\beta^\alpha} \frac{\Gamma(y+\alpha)}{\left(1 + \frac{1}{\beta}\right)^{y+\alpha}} \\ &\left(car\ y \in \mathbb{N}\ et \int_0^\infty \frac{\left(1 + \frac{1}{\beta}\right)^{y+\alpha}}{\Gamma(y+\alpha)} \theta^{y+\alpha-1} \exp\left(-\theta \left(1 + \frac{1}{\beta}\right)\right) d\theta = 1 \right) \end{aligned}$$

$$= \frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)} \frac{\beta^{y+\alpha}}{\beta^{\alpha}(1+\beta)^{y+\alpha}}$$

$$= \frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)} \frac{\beta^{y}}{(1+\beta)^{y+\alpha}}$$

$$= \frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)} \left(1-\frac{1}{1+\beta}\right)^{y} \left(\frac{1}{1+\beta}\right)^{\alpha} \quad c.q.f.d.$$

Annexe IV

Démonstration pour :

$$Quand\ Y \sim \mathcal{NB}(\alpha, p)\ alors\ \mathbb{E}[Y] = \frac{\alpha(1-p)}{p}\ et\ \mathbb{V}[Y] = \frac{r(1-p)}{p^2}$$

Une autre façon de définir la loi binomiale négative est atteinte par la réécriture suivante :

$$\mathbb{P}(Y=y) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)}(1-p)^{y}p^{r} \quad \text{quand } Y \sim \mathcal{NB}(r,p)$$

On a donc posé $r = \alpha\ et\ p = \frac{1}{1+\beta} \Leftrightarrow \beta = \frac{1-p}{p}\ quand\ p \neq 0$.

Ainsi comme il s'agit d'une loi discrète :

$$\mathbb{E}[Y] = \sum_{k=0}^{+\infty} k\ \mathbb{P}(Y=k)$$

$$= \sum_{k=1}^{+\infty} k \frac{\Gamma(k+r)}{\Gamma(k+1)\Gamma(r)}(1-p)^{k}p^{r}$$

$$= \sum_{k=1}^{+\infty} \frac{\Gamma(k+r)}{(k-1)!\,\Gamma(r)}(1-p)^{k}p^{r}\ \ car\ \Gamma(k+1) = k!\ comme\ k \in \mathbb{N}$$

$$= \frac{r(1-p)}{p} \sum_{k=1}^{+\infty} k \frac{\Gamma(k-1+r+1)}{(k-1)!\,\Gamma(r+1)}(1-p)^{k-1}p^{r+1}$$

Posons maintenant j=k-1 et on obtient :

$$\mathbb{E}[Y] = \frac{r(1-p)}{p} \sum_{j=0}^{+\infty} \frac{\Gamma(j+r+1)}{j!\,\Gamma(r+1)}(1-p)^{j}p^{r+1}$$

$$= \frac{r(1-p)}{p}$$

Effectivement $\sum_{j=0}^{+\infty} \frac{\Gamma(j+r+1)}{j!\Gamma(r+1)}(1-p)^{j}p^{r+1} = 1$ car si X suit une loi discrète (dans notre cas $X \sim \mathcal{NB}(r+1,p)$), alors par définition de la probabilité, on a $\sum_{j=0}^{+\infty} \mathbb{P}(X=j) = 1$.

En posant ensuite $= \frac{1}{1+\beta}$, on obtient le résultat mentionné dans la partie théorique du rapport: $\mathbb{E}[Y] = \frac{\alpha(1-p)}{p} = \alpha\beta$

On va également établir la démonstration pour l'expression de la variance, mais calculons d'abord $\mathbb{E}[Y^2]$.

$$\mathbb{E}[Y^2] = \sum_{k=0}^{+\infty} k^2\, \mathbb{P}(Y = k)$$

$$= \sum_{k=1}^{+\infty} k^2 \frac{\Gamma(k+r)}{\mathrm{k!}\,\Gamma(\mathrm{r})}(1-p)^k p^r$$

$$= \sum_{k=1}^{+\infty} k \frac{\Gamma(k+r)}{(\mathrm{k}-1)!\,\Gamma(\mathrm{r})}(1-p)^k p^r \; avec\; k = k-1+1$$

$$= \sum_{k=2}^{+\infty} (k-1) \frac{\Gamma(k+r)}{(\mathrm{k}-1)!\,\Gamma(\mathrm{r})}(1-p)^k p^r + \sum_{k=1}^{+\infty} (\frac{\Gamma(k+r)}{(\mathrm{k}-1)!\,\Gamma(\mathrm{r})}(1-p)^k p^r$$

$$= \sum_{k=2}^{+\infty} \frac{\Gamma(k+r)}{(\mathrm{k}-2)!\,\Gamma(\mathrm{r})}(1-p)^k p^r + \mathbb{E}[Y]$$

$$= \frac{(1-p)^2\, r(r+1)}{p^2} \sum_{k=2}^{+\infty} \frac{\Gamma(k-2+r+2)}{(\mathrm{k}-2)!\,\Gamma(\mathrm{r}+2)}(1-p)^{k-2} p^{r+2} + \mathbb{E}[Y]$$

$$= \frac{(1-p)^2\, r(r+1)}{p^2} + \frac{r(1-p)}{p}$$

En posant j=k-2, $\sum_{k=2}^{+\infty} \frac{\Gamma(k-2+r+2)}{(\mathrm{k}-2)!\Gamma(\mathrm{r}+2)}(1-p)^{k-2}p^{r+2} = \sum_{j=0}^{+\infty} \frac{\Gamma(j+r+2)}{\mathrm{j!}\Gamma(\mathrm{r}+2)}(1-p)^j p^{r+2}$ et on obtient à nouveau avec $X \sim \mathcal{NB}(\mathrm{r}+2,\mathrm{p})$ et par la définition de la probabilité :

$\sum_{j=0}^{+\infty} \mathbb{P}(X = j) = 1.$

Ainsi : $\mathbb{V}[Y] = \mathbb{E}[Y^2] - \mathbb{E}^2[Y]$

$$= \frac{(1-p)^2\, r(r+1)}{p^2} + \frac{r(1-p)}{p} - \frac{r^2(1-p)^2}{p^2}$$

$$= \frac{(1-p)r[(r+1)(1-p) + p - r(1-p)]}{p^2}$$

$$= \frac{(1-p)r(r - rp + 1 - p + p - r + rp)}{p^2}$$

$$= \frac{(1-p)r}{p^2}$$

En reparamétrisant avec $r = \alpha$ et $p = \frac{1}{1+\beta} \Leftrightarrow \beta = \frac{1-p}{p}$ quand $\mathrm{p} \neq 0$, on obtient ce qu'on vient de trouver dans la partie théorique de ce rapport sous 3.7. : $\mathbb{V}[Y] = \frac{(1-p)r}{p^2} = \alpha\beta(1+\beta)$ c.q.f.d.

Annexe V

Introduisons maintenant les différentes méthodes d'approximation de maximum de vraisemblance pour les GLMM. On suppose, comme pour tout le rapport d'ailleurs, que les variables aléatoires sont i.i.d. suivant une loi normale centrée de variance constante i.e. $A_i \sim \mathcal{N}(0, \sigma_A^2)$ de façon i. i. d.. Pour généraliser l'écriture, on dit que :

$$Y_{ij}|\mathbb{X}_{ij}, A_i \sim \mathcal{L} \in famille\ exponentielle\ tel\ que$$

$$f_{Y_{ij}|\mathbb{X}_{ij},A_i}(y_{ij}|\mathbb{X}_{ij}, A_i) = \exp\left(\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{a(\phi)} + c(y_{ij}, \phi)\right)$$

Ainsi pour obtenir la loi marginale de $Y_{ij}|\mathbb{X}_{ij}, A_i$, il faut intégrer la loi jointe par A_i comme les $\mathbb{X}_{ij}$ étant connues. On obtient :

$$f_{Y_{ij}|\mathbb{X}_{ij}}(y_{ij}|\mathbb{X}_{ij}) = \int_{-\infty}^{+\infty} f_{Y_{ij}|\mathbb{X}_{ij},\,A_i}(y_{ij}|\mathbb{X}_{ij}, A_i)\ f_{A_i}(a_i)\, da_i$$

Ainsi, la quasi-vraisemblance s'écrit :

$$L(Y) = \prod_i \int_{-\infty}^{+\infty} \exp\left(\sum_j \frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{a(\phi)} + c(y_{ij}, \phi)\right) \frac{\exp\left(-\frac{a_i^2}{2\sigma_A^2}\right)}{\sqrt{2\pi\sigma_A^2}}\, da_i$$

Pour obtenir les maximums de vraisemblance, il s'agit maintenant de maximiser L(Y). On voit bien que maximiser une intégrale de cette forme est difficile et c'est exactement à cause de ceci que quelques scientifiques ont cherché un remède. Ils ont donc trouvé plusieurs moyens pour approximer la valeur de cette intégrale.

a) Moindres carrés re-pondérés itérativement (IRLS)

La méthode d'estimation par défaut de la fonction `glm` de R est maximum de vraisemblance qui est pourtant obtenue par la méthode des moindres carrées re-pondérés itérativement (IRLS) (eng. *Iterative re-weighted least square*).

On va juste décrire le principe de cette méthode sans la démontrer. Si le lecteur est quand même intéressé à la preuve, je l'invite à jeter un coup d'œil dans le livre *Generalized Linear Models* de McCullagh and Nelder (1989) (voir Chapitre 8 Références pour plus de détails) page 40 à 43.

Expliquons donc maintenant le principe de la méthode des moindres carrés re-pondérés itérativement. On désigne Z comme étant la nouvelle variable dépendante. Elle est définie à partir de la variable initiale Y à laquelle on applique la fonction de lien g :

$$g(Y_i) \cong g(\mu_i) + (Y_i - \mu_i)g'(\mu_i) \quad (par\ le\ théorème\ de\ Taylor)$$

$$= \eta_i + (Y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i} = Z_i$$

Comme pour tout procès itératif, il faut commencer avec des estimations initiales du prédicteur linéaire $\hat{\eta}^{(0)}$ et les valeurs ajustées $\hat{\mu}^{(0)}$ obtenues par la relation $g(\mu) = \eta$.

Normalement, on prend le modèle nul (eng. *NULL model* or *intercept-only model*) et donc $\hat{\beta}_0^{(0)}$(l'intercept) est prise comme étant la moyenne de la réponse et tous les autres coefficients sont choisis comme étant zéro. Ainsi la variable dépendante ajustée initiale s'écrit $Z^{(0)} = \hat{\eta}^{(0)} + \left(Y - \hat{\mu}^{(0)}\right)\left(\frac{\partial \eta}{\partial \mu}\right)^{(0)}$ où $\left(\frac{\partial \eta}{\partial \mu}\right)^{(0)}$ représente la dérivée de η par rapport à μ évaluée en $\hat{\mu}^{(0)}$. Les poids ont été choisis comme étant de la forme quadratique et sa valeur initiale se calcule par la formule suivante : $W_0^{-1} = \left(\left(\frac{\partial \eta}{\partial \mu}\right)^{(0)}\right)^2 V(\hat{\mu}^{(0)})$ où V est la fonction de variance. D'ailleurs, la méthode a son caractère itératif visiblement du fait que les Z et les poids W dépendent de $\hat{\mu} = g^{-1}(\hat{\eta})$. Par la suite, on va essayer de minimiser $\sum_{i=1}^{n} W_i^{(0)}\left(Z_i^{(0)} - \eta_i\right)$ pour obtenir les nouveaux estimateurs de $\hat{\beta}^{(1)}$ et donc les nouveaux $\hat{\eta}^{(1)}$ pour pouvoir recommencer l'algorithme. Ainsi, on a effectué une régression de $Z^{(0)}$ sur les prédicteurs $X^{(1)}, \ldots, X^{(p)}$ avec le poids $W^{(0)}$. On procède de cette manière tant que les changements dans les estimations sont encore assez grands.
Toutes les méthodes de moindres carrées pondérées sont considérées comme étant robuste car les estimateurs sortant ne sont pas influencés par des outliers éventuels à cause de l'attribution d'un poids faible à ces observations.

b) Approximation de Laplace

Laplace a introduit une méthode pour approximer des intégrales de la forme $\int_a^b \exp\big(M \cdot f(x)\big)\, dx$ où M est un nombre à grande valeur et f une fonction qui possède un maximum globale en x_o. Si on fait croître M maintenant vers l'infinie, Laplace a remarqué qu'il suffit en effet de se concentrer sur $f(x_o)$ et que l'intégrale après avoir appliqué la formule de Taylor, est gaussienne. Cette conclusion n'a quand même pas été faite sans hypothèses. Il fallait donc supposer que $a \neq x_0$ *et* $b \neq x_0$ ne sont pas une valeur d'une borne de l'inégrale et que la dérivée seconde de f en x_o est négative.

Par Taylor, on obtient : $f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2 + h_2(x - x_0)^2$ où $\lim_{x \to x_0} h_2(x) = 0$. Dans notre cas, comme x_o est un maximum globale, $f'(x_0) = 0$, d'où $f(x) \approx f(x_0) - \frac{|f''(x_0)|}{2}(x - x_0)^2$ ce qui nous permet de dire qu'on a l'approximation : $\int_a^b \exp\big(M\, f(x)\big)\, dx \approx \exp\big(M\, f(x_0)\big) \int_a^b \exp\left(-\frac{M|f''(x_0)|}{2}(x - x_0)^2\right) dx$

Si on prend maintenant $a = -\infty$ et $b = +\infty$ comme c'est le cas dans notre problème initial de trouver le maximum de la fonction de vraisemblance, alors l'intégrale incorpore la partie principale d'une densité d'une loi normale centrée en x_o et à terme de variance $(M|f''(x_0)|)^{-1}$. Ainsi on a : $\int_{-\infty}^{+\infty} \exp\left(-\frac{M|f''(x_0)|}{2}(x - x_0)^2\right) dx = \sqrt{\frac{2\pi}{M|f''(x_0)|}}$.

c) Maximum de quasi-vraisemblance pénalisée (PQL)

Le principe de l'estimation par quasi-vraisemblance pénalisée est d'appliquer l'approximation de Laplace à la fonction de quasi-vraisemblance qu'on vient d'intégrer. D'après Breslow and Clayton (1993), on a alors avec nos notations :

$\mathbb{E}[Y_i|\mathbb{X}_i, A_i] = \mu_i$ *et* $\mathbb{V}[Y_i|\mathbb{X}_i, A_i] = \phi a_i V(\mu_i)$ *et où* $A = \begin{pmatrix} A_1 \\ \vdots \\ A_I \end{pmatrix} \sim \mathcal{N}_I(0, D(\theta))$ (iid) :

$$\exp\big(ql(\beta, y)\big) \propto |D(\theta)|^{-\frac{1}{2}} \int \exp\left(-\frac{1}{2\phi} \sum_{i=1}^{n} d_i(y_i; \mu_i) - \frac{1}{2} A^t D(\theta)^{-1} A \right) dA$$

Où $ql(\beta, y)$ est la log-quasi-vraisemblance pénalisée et $d_i(y, \mu) = -2 \int_y^\mu \frac{y-u}{a_i V(u)} du$

Ainsi en reparamétrisant, on reçoit une formule qui nous est utile pour ensuite appliquer l'approximation de Laplace :

$$\begin{aligned} \exp\big(ql(\beta, y)\big) &\propto |D(\theta)|^{-\frac{1}{2}} \int \exp\left(-\frac{1}{2\phi} \sum_{i=1}^{n} d_i(y_i; \mu_i) - \frac{1}{2} A^t D(\theta)^{-1} A \right) dA \\ &\propto c|D(\theta)|^{-\frac{1}{2}} \int \exp\big(-\kappa(A)\big)\, dA \end{aligned}$$

d) Quadrature adaptive de Gauss-Hermite (AGHQ)

La quadrature de Gauss-Hermite approche des intégrales de la forme $\int_{-\infty}^{+\infty} \exp(-x^2)\ f(x)\, dx$ par une somme pondérée en n points $\sum_{i=1}^{n} w_i f(x_i)$. x_i est appelé nœud de la quadrature, il s'agit ici des racines du $n^{\text{ième}}$ polynôme orthogonal de Hermite $H_n(x) = (-1)^n \exp(x^2) \frac{\partial^n}{\partial x^n} \exp(-x^2)$. w_i est choisi comme suit dans la méthode de quadrature de Gauss-Hermite : $w_i = \frac{2^{n+1} n! \sqrt{\pi}}{(2n\, H_{n-1}(x_i))^2}$.

La version adaptive des quadratures, donc dans notre cas la version de quadrature adaptive de Gauss-Hermite, repose dans la transformation de la variable à intégrer en une variable centrée réduite. On a alors l'approximation suivante obtenue par changement de variable $t = \frac{x-\hat{\mu}}{\sqrt{2}\hat{\sigma}}$ où $\hat{\mu}$ est le maximum globale de f(x) et $\hat{\sigma}^2 = -\left(\frac{d^2}{dx^2} \ln\big(f(x)\big)\right)^{-1}|_{x=\hat{\mu}}$:

$$\begin{aligned} \int_{-\infty}^{+\infty} f(x)dx &= \int_{-\infty}^{+\infty} f\big(\sqrt{2}\hat{\sigma}t + \hat{\mu}\big)\sqrt{2}\hat{\sigma} \exp(-t^2) \exp(t^2)\, dt \\ &\simeq \sum_{i=1}^{n} f\big(\sqrt{2}\hat{\sigma}t_i + \hat{\mu}\big) \exp\big(t_i^2\big)\sqrt{2}\hat{\sigma}w_i \\ &= \sum_{i=1}^{n} f(u_i)v_i \ \ avec\ u_i = \sqrt{2}\hat{\sigma}t_i + \hat{\mu}\ et\ v_i = \exp\big(t_i^2\big)\sqrt{2}\hat{\sigma}w_i \end{aligned}$$

On voit alors que les nœuds et les poids dépendent de x ce qui revient à dire qu'ils dépendent de la moyenne et de l'écart-type de x ce qui assure une meilleure approximation. Par défaut, si on utilise seulement n=1 point, alors on applique l'approximation de Laplace. Par contre en choisissant n élevé, de nombreuses études ont pu montrer que la précision des estimations va augmenter. Dans la fonction `glmer` de R, on peut choisir le nombre de points par l'option `nAGQ=`.

Annexe VI

a) Résultats de Breslow et Clayton (1993)

Table 4. PQL Model Fits to Thall and Vail's Epilepsy Data

	Model			
	I	*II*	*III*	*IV*
Variable	$\hat{\beta} \pm SE$	$\hat{\beta} \pm SE$	$\hat{\beta} \pm SE$	$\hat{\beta} \pm SE$
Fixed effects				
Constant	−2.76 ± .41	−1.25 ± 1.2	−1.27 ± 1.2	−1.27 ± 1.2
Base	.95 ± .04	.87 ± .14	.86 ± .13	.87 ± .14
Trt	−1.34 ± .16	−.91 ± .41	−.93 ± .40	−.91 ± .41
Base × Trt	.56 ± .06	.33 ± .21	.34 ± .21	.33 ± .21
Age	.90 ± .12	.47 ± .36	.47 ± .35	.46 ± .36
V4	−.16 ± .05	−.16 ± .05	−.10 ± .09	—
Visit/10	—	—	—	−.26 ± .16
Subject level random effects				
Constant ($\sqrt{\sigma_{11}}$)	—	.53 ± .06	.48 ± .06	.52 ± .06
Visit/10 ($\sqrt{\sigma_{22}}$)	—	—	—	.74 ± .16
Covariance (σ_{12})	—	—	—	−.01 ± .03
Unit level random effects				
Constant ($\sqrt{\sigma_{00}}$)	—	—	.36 ± .04	—

b) Résultats de Breslow (1996)

Table 1. Log-linear Poisson regression fit to the epilepsy data

Coefficient	Value	Stnd. Error	*t*-statistic
(Intercept)	3.079	0.451	6.833
log(Base.Cnt/4)	-0.073	0.201	-0.366
Age/10	-0.511	0.153	-3.332
log(Base.Cnt/4):Age/10	0.351	0.068	5.164
Progabide	-0.610	0.191	-3.197
Progabide:log(Base.Cnt/4)	0.204	0.088	2.325

Deviance=408.41; Pearson χ^2=456.52; DF=52

Table 3. Negative binomial regression fit to the epilepsy data

Coefficient	Value	Stnd. Error	*t*-statistic
(Intercept)	2.879	1.094	2.631
log(Base.Cnt/4)	0.003	0.570	0.005
Age/10	-0.399	0.368	-1.084
log(Base.Cnt/4):Age/10	0.303	0.191	1.588
Progabide	-0.702	0.451	-1.558
Progabide:log(Base.Cnt/4)	0.248	0.240	1.037

$\hat{\phi}$=0.302 by method of moments

- **Quasi-Poisson avec le jeu de données complet :**

```
Call:
glm(formula = Somme ~ log(Bdiv4) + Agediv10 + log(Bdiv4):Agediv10 +
    trait + trait:log(Bdiv4), family = quasipoisson, data = Thall_epilepsy_rename)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.1377  -1.5986  -0.5294   1.2061   8.9098

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.1941     1.4484   1.515  0.13575
log(Bdiv4)                 0.2904     0.6376   0.455  0.65068
Agediv10                  -0.2031     0.4856  -0.418  0.67752
traitprogabide            -1.5240     0.5412  -2.816  0.00681 **
log(Bdiv4):Agediv10        0.2247     0.2128   1.056  0.29586
log(Bdiv4):traitprogabide  0.6624     0.2327   2.847  0.00627 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 10.09628)

    Null deviance: 2122.73  on 58  degrees of freedom
Residual deviance:  476.08  on 53  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

- **Quasi-Poisson avec le jeu de données restreint (sans patient 207) :**

```
Call:
glm(formula = Ysum ~ log(Base4) + Age10 + log(Base4):Age10 +
    Trt + Trt:log(Base4), family = quasipoisson, data = Thall_epilepsy_sans)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.3533  -1.5596  -0.4326   0.5097   8.9192

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               3.07922    1.33546   2.306   0.0251 *
log(Base4)               -0.07368    0.59682  -0.123   0.9022
Age10                    -0.51074    0.45422  -1.124   0.2660
Trtprogabide             -0.61042    0.56590  -1.079   0.2857
log(Base4):Age10          0.35071    0.20124   1.743   0.0873 .
log(Base4):Trtprogabide   0.20446    0.26063   0.784   0.4363
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 8.780303)

    Null deviance: 1284.72  on 57  degrees of freedom
Residual deviance:  408.41  on 52  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

c) Résultats de Sinha et Xu (2011)

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = \beta_0 + \beta_1\, \text{base}_i + \beta_2\, \text{trt}_i + \beta_3\, \text{visit}_t + u_i,$$

Paramètres	Estimations	Erreurs-standards
$\hat{\beta}_0$	-3.849	1.439
$\hat{\beta}_1$	6.912	1.862
$\hat{\beta}_2$	-0.635	1.214
$\hat{\beta}_3$	-0.286	0.137
$\hat{\sigma}$	3.121	0.945

où $\hat{\sigma}$ est l'estimateur de l'écart-type du effet aléatoire.

Annexe VII

a) Prédicteurs transformés pour Breslow et Clayton (1993)

> BC_predictors

	Original	Poisson I	QuasiPoisson_I	II PQL	II Laplace	II AGHQ	III PQL
2	11	19.0175248	19.0175248	11.7008028	11.2267933	11.2251456	11.1302557
3	14	19.0175248	19.0175248	11.7008028	11.2267933	11.2251456	13.1639229
4	9	19.0175248	19.0175248	11.7008028	11.2267933	11.2251456	9.8405646
5	8	16.1880576	16.1880576	9.9599328	9.5564463	9.5550370	8.8861731
7	8	11.1744213	11.1744213	7.7419985	7.4304602	7.4295250	7.7263415
8	7	11.1744213	11.1744213	7.7419985	7.4304602	7.4295250	7.1627690
9	9	11.1744213	11.1744213	7.7419985	7.4304602	7.4295250	8.3082662
10	4	9.5118675	9.5118675	6.5901277	6.3249400	6.3241394	5.3276682
12	0	2.5706169	2.5706169	2.4337344	2.1484208	2.1473441	1.6263819
13	4	2.5706169	2.5706169	2.4337344	2.1484208	2.1473441	2.7011511
14	3	2.5706169	2.5706169	2.4337344	2.1484208	2.1473441	2.3946324
15	0	2.1881551	2.1881551	2.0716383	1.8287740	1.8278562	1.5102898
177	11	12.9790194	12.9790194	5.9931480	5.0390717	5.0353361	8.2651505
178	0	12.9790194	12.9790194	5.9931480	5.0390717	5.0353361	3.0388772
179	0	12.9790194	12.9790194	5.9931480	5.0390717	5.0353361	3.0388772
180	5	11.0479737	11.0479737	5.1014749	4.2893475	4.2861647	4.8230416
182	102	64.2806394	64.2806394	76.8138659	77.7079664	77.7110408	99.7775156
183	65	64.2806394	64.2806394	76.8138659	77.7079664	77.7110408	65.2255118
184	72	64.2806394	64.2806394	76.8138659	77.7079664	77.7110408	71.6816980
185	63	54.7168313	54.7168313	65.3853382	66.1464038	66.1489743	62.8932640
187	4	4.8909698	4.8909698	3.8941587	3.5951026	3.5941346	3.7693549
188	3	4.8909698	4.8909698	3.8941587	3.5951026	3.5941346	3.3865948
189	2	4.8909698	4.8909698	3.8941587	3.5951026	3.5941346	3.0295208
190	4	4.1632811	4.1632811	3.3147777	3.0602153	3.0593892	3.5587420

Il s'agit des prédicteurs transformés pour 4 comptages associés à 6 patients (les 4 mesures pour un patient correspondent aux 4 lignes de la case respective)

b) Prédicteurs transformés pour Breslow (1996)

	Original	Poisson_Avec_207	Binomiale Négative Avec_207	Poisson_Sans_207	Binomiale_Négative_Sans_207
16	16	48.177960	47.346383	47.859792	46.460836
17	6	20.255362	20.962855	20.395388	20.956264
18	123	126.993392	115.685860	129.554654	118.227794
19	15	21.378063	21.680536	20.532551	21.434563
20	16	19.971737	21.282606	21.615562	21.547517
21	14	14.013041	14.792240	13.935992	14.743184
22	14	10.866790	12.142277	12.732416	12.936975
23	13	20.179630	20.537817	19.338061	20.276099
24	30	28.344270	29.000476	28.936895	28.827505
25	143	61.104282	58.442410	61.041740	58.663927
26	6	10.444276	10.631850	8.282186	9.663942
27	10	11.768836	13.197674	14.180435	14.178499
28	53	39.669689	39.969985	39.451358	38.507865
29	42	73.749232	61.383148	44.402141	44.863794
30	28	43.983064	42.135589	38.683418	38.382300
31	7	11.572644	14.631716	15.549011	15.096655
32	13	4.718546	6.926399	7.541787	7.713435
33	19	11.237334	14.352964	15.438328	14.878460
34	11	17.391749	20.039481	19.795831	19.531723
35	74	29.726651	30.664760	28.754907	28.644283
36	20	8.483023	10.934717	10.835036	11.282942
37	10	5.470036	7.911081	8.846227	8.857895
38	24	67.737910	57.990217	44.379553	44.441821
39	29	36.270693	35.767979	31.348349	31.018016
40	4	2.682466	4.443642	5.266888	5.340320
41	6	15.001980	17.850191	18.031869	17.729553

Les observations 16 à 28 sont des patients prenant du placebo et les observations 29 à 41 constituent des patients prenant du Progabide.