Université **G**aston **B**erger de **S**aint-**L**ouis

UFR de **S**ciences **A**ppliquées et de **T**echnologie

MÉMOIRE

pour obtenir le diplôme de

Master en Mathématiques Appliquées

Soutenue publiquement le 03 Août 2023 par

Ibrahim LAWANI

spécialité : Sciences des Données et Applications

Option : Statistiques

MODÈLES DE HURDLE POUR LES DONNÉES DE COMPTAGE SURDISPERSÉES

Encadreur : Prof. Aliou DIOP

Co-encadreur : Dr. Essoham ALI

Composition du Jury

Aboubakary DIAKHABY	Professeur Titulaire, Université Gaston Berger de Saint-Louis, Sénégal	Président
Ahmadou Bamba SOW	Professeur Titulaire, Université Gaston Berger de Saint-Louis, Sénégal	Examinateur
Aliou DIOP	Professeur Titulaire, Université Gaston Berger de Saint-Louis, Sénégal	Encadreur
El Hadji DEME	Maître de Conférences, Université Gaston Berger de Saint-Louis, Sénégal	Examinateur

Dédicaces

À mon père Abrassizou LAWANI,
À ma mère Abiba NAPO,

À toute ma famille.

Remerciements

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner toute ma gratitude.

Je voudrais adresser toute ma reconnaissance à mon encadreur le Prof Aliou DIOP pour sa patience, sa disponibilité, et surtout ses conseils judicieux, qui ont contribués à alimenter ma réflexion. J'ai acquis pas mal de connaissances en vous cotoyant et grâce à votre générosité sans limite tant pédagogique que sociale.

Je voudrais également remercier mon co-encadreur le Dr Essoham ALI pour sa disponibilité, sa patience et ses conseils. Sans vous, ce mémoire n'aurais pas abouti.

Je désire remercier les professeurs de l'Université Gaston Berger de Saint-Louis en particulier les professeurs de l'unité de formation et de recherche en sciences appliquées et technologie, qui m'ont fourni des outils nécessaires à la réussite de mes études universitaires.

Je ne saurais terminer cette partie sans remercier ma famille, mes ami(e)s, mes frères et soeurs. Un merci spécial à la famille pour le soutient inconditionnel, à la fois moral et financier, qui m'a permis de réaliser les études que je voulais et par conséquent ce mémoire.

Abréviations & Notations

$\mathbb{P}(A)$	: La probabilité de l'événement A.
$\mathbb{E}(Y)$	: L'espérance mathématique de la variable aléatoire Y.
$\mathbb{V}(Y)$	: La variance de la variable aléatoire Y.
$\mathbb{E}(X\|Y)$	: Espérance conditionnelle de X sachant Y.
$\mathrm{Cov}(X, Y)$	: Covariance des variables aléatoires X et Y.
i.i.d	: Indépendantes et identiquement distribuées.
$\mathbb{N}^*$	: Ensemble des entiers naturels non nuls.
$\mathbb{R}$	: Ensemble des réels et $\mathbb{R}^d = \underbrace{\mathbb{R} \times \ldots \times \mathbb{R}}_{d \text{ fois}}$.
$X^\top$	: Transposée du vecteur X.
$\|X\|$	: Norme du vecteur X.
I_p	: Matrice identité d'ordre p.
diag	: Diagonale d'une matrice
$\mathcal{P}(\lambda)$	: Loi de Poisson de paramètre λ.
$\mathcal{B}(\mathrm{p})$	: Loi Bernoulli de paramètre p.
EMV	: Estimateur du Maximum de Vraisemblance.
GLM	: Generalised Linear Model
ZIP	: Zero-Inflated Poisson
ZIB	: Zero-Inflated Binomial
ZINB	: Zero-Inflated Negative Binomial
HP	: Hurdle Poisson
HB	: Hurdle Bell
HPLN	: Hurdle Poisson Log-Normale
HNB	: Hurdle Negative Binomial

Table des matières

Liste des tableaux

Résumé

Les données de comptage comportant un nombre excessif de zéros sont fréquemment rencontrées dans plusieurs domaines de la vie pratique. Le modèle de Hurdle est un modèle qui comprend deux parties. Un modèle binaire pour évaluer si l'observation est nulle ou non et un modèle de comptage tronqué pour modéliser l'autre partie (les comptes positifs). C'est un modèle conçu pour modéliser les données de comptage surdispersées où les modèles Poisson et binomiale négative sont inefficaces. Vu l'importance cruciale de ces modèles, l'étude de trois modèles de hurdle ont fait l'objet de ce mémoire. Ce mémoire recouvre plusieurs aspect à savoir: la modélisation statistique, létude mathématiques des modèles proposés ainsi que l'inférence statistique, létudes de simulations numériques pour évaluer la performance de ces modèles. Les paramètres sont estimés par la méthode du maximum de vraisemblance.

Mots clés: Données de comptage, modèles de Hurdle, excès de zéros, simulations, tronquée, maximum de vraisemblance.

Abstract

Counting data with an excessive number of zeros is frequently encountered in many areas of pratical life. The hurdle model is model that comprise two parts. A binary model to evaluate whether the observation in zero or not and truncated counting model to model the other part (positive account). It is a model designed to model over-disperded count data where the Poisson and negative binomial models are ineffective. Given the crucial importance of these models, the study of three hurdle models was the subject of this dissertation. This thesis covers several aspects, namely: statistical modelling, mathematical study of the proposed models and statistical inference, numerical simulation studies to assess the performance of these models. The estimators are estimated by the maximum likelihood method.

Keywords: Asymptotic normality, consistency, counting data, hurdle models, excess zeros, simulations, truncated, maximum likelihood.

Introduction générale

Une revue de la littérature

Au cours de ces dernières années, nous avons pu observer dans la recherche scientifique, la modélisation statistique des données de comptage. Les méthodes de modélisation adaptées à ce type de données ont été largement explorées dans la littérature. Les modèles linéaires généralisés (Mc Cullagh et Nelder (1989, [23])) fournissent un cadre puissant pour l'analyse de ces données. Mais les applications de la vie réelle soulèvent continuellement de nouveaux problèmes et un énorme travail a été fait pour étendre la portée de ces modèles. Les données de comptage se produisent généralement dans toutes les disciplines. Une approche pour modéliser ces types de données est la régression de Poisson (Consul et Famoye, (1992, [7])). La régression de Poisson suppose que la moyenne est égale à la variance. Ce qui n'est pas toujours le cas dans de nombreux scénarios de la vie réelle. Si l'on néglige l'hypothèse d'égalité de la moyenne et de la variance, la régression de poisson produit des estimations biaisées et des résultats trompeurs (Winkelmann et Zimmermann, (1995, [35])) dû à la surdispersion. Ce phénomène (surdispersion) a été largement et diversement étudié dans la littérature (Dean et Lanless, (1989, [9])). Une solution pour corriger la surdispersion est le modèle négative binomiale (Hilbe,J.M. (2011, [19])). Lorsqu'une surdisperdion se produit en raison d'une grande proportion de zéros, les modèles Poisson et NB deviennent inefficaces. C'est dans ce sens que les modèles ZI et Hurdle ont montré leurs utilités et sont devenus très populaires dans la modélisation des données de comptage avec un excès de zéros car ces données sont fréquemment rencontrées dans la vie pratique. Par exemple Dans l'utilisation des services de santé, le nombre d'utilisation des services comprend souvent un grand nombre de zéros représentant les patients qui n'ont pas utilisé les services pendant la période d'étude.

Les chercheurs ont montré que si le zéro excessif n'est pas pris en compte, il en résultera un ajustement déraisonnable à la fois pour les zéros et les comptes nuls (Perumean-chaney et al. (2013, [29])). En effet on distingue deux types de zé-

ros (Lambert (1992, [22])); les zéros structurels et les zéros d'échantillonnage. les sujets qui sont exposés au résultat mais qui n'ont pas fait l'expérience du résultat ou qui ne l'ont pas signalé pendant la période d'étude sont appellés zéros d'échantillonnages et les sujets qui produisent toujours les comptes nuls sont appelés zéros structurels (Cindy Xin Feng (2021, [13])). Ne pas prendre en compte ce facteur peut conduire à un cas particulier de surdispersion: l'inflation de zéros (Lambert (1992 [22]), Mullahy (1997, [25])). Pour résoudre ce problème, deux approches ont été proposées. L'une est le modèle Hurdle développé par Mullahy (1986, [24]) et l'autre est le modèle ZI développé par Lambert (1992, [22]). Ainsi les modèles ZI et Hurdle ont été développés pour traiter les données comportant un excès de zéros où les modèles de Poisson et NB ne sont pas efficaces. Les deux modèles (ZI et Hurdle) ont été employés avec succès dans beaucoup de domaines à savoir: la recherche en santé publique (Neelon et al. (2016, [27])), en épidémiologie (Osei, Stein et Andreo (2022, [28])), en médecine Rose et al. (2006, [30])), en sécurité routière (Ndilimeke Shiyuka (2018, [33])), en toxicomanie (Buu et al. (2012, [4])), études écologiques et environnementales (Feng et Dean (2012, [14])).

Vu l'importance capitale de ces types de modèles, il est nécessaire de pouvoir les connaître, de les comprendre et de savoir les utiliser pour modéliser les données de comptage avec un excès de zéros. C'est dans cette perspective que notre intérêt a été porté sur le thème intitulé: «**Modèle de Hurdle pour les données de comptage surdispersées**».

Dans notre mémoire nous allons nous concentrer sur l'étude de trois modèles de Hurdle à savoir: le modèle de Hurdle Poisson, le modèle de Hurdle Bell et le modèle de Hurdle Poisson log-normal. En ce qui concerne le modèle HP et le modèle HB, nous avons fait des simulations sur des échantillons de tailles finis pour évaluer la performance des estimateurs par la méthode du maximum de vraisemblance. Pour le modèle HPLN nous nous sommes limités à une étude théorique du modèle.
Pour mieux appréhender ce mémoire, nous l'avons subdivisé en trois chapitres. Dans le chapitre premier, nous avons présenté quelques rappels sur les modèles linéaires généralisés, dans le deuxième chapitre, nous avons parlé des données de comptage et l'inflation zéros et enfin dans le chapitre 3 nous avons fait des études de simulation du modèle Hurdle pour les données de comptage surdispersées.

1 Modèles linéaires généralisés

Résumé

Dans ce chapitre, nous énonçons quelques rappels essentiels sur les modèles linéaires généralisés. Nous définissons les différentes méthodes d'estimation des modèles linéaires généralisés, les propriétés asymptotiques, les différents tests d'hypothèses et le choix entre les différents modèles.

Sommaire

1.1 Introduction

Présentés pour la première fois par Nelder et Wedderburn (1972, [26]) et exposés de manière complète par Mc Cullagh et Nelder (1989, [23]), les modèles linéaires généralisés (GLMs) permettent d'étudier la liaison entre une variable réponse Y et un ensemble de variables explicatives $X_1, \cdots, X_k$. Ils sont conçus pour traiter les problèmes où la variable réponse est décrite par un modèle paramétrique. Les modèles linéaires généralisés englobent: le modèle linéaire général (régression multiple, analyse de la variance et analyse de la covariance), le modèle log-linéaire, la régression logistique, la régression de Poisson. L'objet de ce chapitre est de définir les concepts communs à ces modèles à savoir: la famille exponentielle, estimation du maximum de vraisemblance, propriétés asymptotiques, tests, la qualité de l'ajustement, le choix entre différents modèles et l'étude des résidus.

1.2 Présentation des modèles linéaires généralisés

Les modèles linéaires généralisés sont caractérisés par trois composantes:
- La composante aléatoire: identifie la distribution de probabilité de la variable à expliquer;
- La composante déterministe (prédicteur linéaire): il s'agit des variables $X_1, \cdots, X_k$ utilisées comme prédicteurs dans le modèle;
- La fonction de lien: décrit la relation fonctionnelle entre la combinaison linéaire des variables $X_1, \cdots, X_k$ et l'espérance mathématique de la variable réponse Y.

1.2.1 Composante aléatoire

Notons $(Y_1, \cdots, Y_n)$ un échantillon aléatoire de taille n de variable réponse Y, les variables aléatoires $Y_1, \cdots, Y_n$ étant supposées indépendantes et identiquement distribuées (i.i.d) admettent des distributions issues d'une structure exponentielle (voir Nelder et Wedderburn (1972, [26])). Cela signifie que les lois de ces variables sont données par une même mesure dite de référence. La mesure de référence change d'une structure exponentielle à une autre, la mesure de Lebesgue pour une loi continue, une mesure discrète combinaison de masse de Dirac pour une loi discrète. Pour une présentation générale de la famille exponentielle, des propriétés asymptotiques, des estimations et leurs paramètres voir Antoniadis et al. (1992, [2]). La famille de leurs densités par rapport à cette mesure se met sous la forme:

$$f(y_i, \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\} \tag{1.1}$$

où $\theta_i \in \mathbb{R}$ est un paramètre canonique et $\phi \in \mathbb{R}_+^*$ est un paramètre de dispersion. Les fonctions b et c sont spécifiques pour chaque distribution. La fonction b est supposée deux fois dérivables de dérivée première inversible, dérivable et d'inverse dérivable. La fonction a_i s'écrit $a_i = \frac{\phi}{\omega_i}$ où les poids ω_i sont les poids connus des observations, fixés ici à 1 pour simplifier. Parmi les lois appartenant à la famille exponentielle on rencontre les lois classiques telles que la loi binomiale, la loi poisson, la loi normale, la loi gamma $\cdots$. L'expression de la famille exponentielle se met alors sous la forme canonique en posant:

$Q(\theta) = \frac{\theta}{\phi}$, $\quad d(\theta) = \exp\left(\frac{-b(\theta)}{\phi}\right)$, $\quad e(y) = \exp\big(c(y, \phi)\big)$.

On obtient :

$$f(y_i, \theta_i) = d(\theta_i)e(y_i)\exp\big(y_iQ(\theta_i)\big) \tag{1.2}$$

1.2.2 Prédicteur linéaire

Les covariables interviennent linéairement dans la modélisation, comme dans les modèles linéaires classiques. Elles sont organisées dans la matrice X. Soit β un vecteur de p paramètre, le prédicteur linéaire, composante déterministe de modèle, est le vecteur à n composantes:

$\eta = X\beta$

1.2.3 Fonction de lien

La troisième composante exprime une relation fonctionnelle entre l'espérance Y_i et la i^{eme} composante du prédicteur linéaire c'est-à-dire pour tout $i = 1, \cdots, n$, on a: $u_i = \mathbb{E}(Y_i)$. On pose $\eta_i = g(u_i), i = 1, \cdots, n$ où g appelée fonction de lien est supposée monotone et différentiable. Parmi toutes les fonctions de lien, celle qui permet d'égaler le prédicteur linéaire et le paramètre canonique est appelée la fonction de lien canonique. Puisqu'on a la relation $\eta_i = g\big(b'(\theta_i)\big)$, la fonction associée à une distribution donnée vérifie $g = (b')^{-1}$. Les fonctions de lien canonique associées aux lois classiques sont indiquées dans McCullagh et Nelder (1989, [23]).

1.2.4 Exemples

1.2.4.1 Loi Gaussienne

Dans le cas d'un échantillon gaussien, les densités d'une famille de $\mathcal{N}(\mu_i, \sigma^2)$ s'écrit:

$$\begin{aligned} f(y_i, \mu_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right\} \\ &= \exp\left\{-\ln\sqrt{2\pi\sigma^2} - \frac{(y_i^2 - 2y_i\mu_i + \mu_i^2)}{2\sigma^2}\right\} \\ &= \exp\left\{-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\frac{y_i^2}{\sigma^2} + \frac{y_i\mu_i}{\sigma^2} - \frac{1}{2}\frac{\mu_i^2}{\sigma^2}\right\} \\ &= \exp\left\{-\frac{1}{2}\frac{\mu_i^2}{\sigma^2}\right\} \exp\left\{-\frac{1}{2}\frac{y_i^2}{\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right\} \exp\left\{\frac{y_i\mu_i}{\sigma^2}\right\} \end{aligned}$$

En posant:

$$\begin{aligned} Q(\theta_i) &= \frac{\mu_i}{\sigma^2} \\ d(\theta_i) &= \exp\left\{-\frac{1}{2}\frac{\mu_i^2}{\sigma^2}\right\} \\ e(y_i) &= \exp\left\{-\frac{1}{2}\frac{y_i^2}{\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right\} \end{aligned}$$

La famille gaussienne se met sous la forme canonique (1.2) qui en fait une famille exponentielle de paramètre de dispersion $\phi = \sigma^2$ et de paramètre naturel $\theta_i = \mathbb{E}(y_i) = \mu_i$ et donc de fonction de lien canonique, la fonction identité.

1.2.4.2 Loi Bernoulli

Considérons *n* variables aléatoires binaires indépendantes z_i de probabilité de succès π_i et donc d'espérance $\mathbb{E}(z_i) = \pi_i$. les fonctions de densité de ces variables sont éléments de la famille:

$$\begin{aligned} f(z_i, \pi_i) &= \pi_i^{z_i} \left(1 - \pi_i^{1-z_i}\right) \\ &= (1 - \pi_i) \exp\left\{ z_i \ln \frac{\pi_i}{1 - \pi_i} \right\} \end{aligned}$$

qui est la forme canonique d'une structure exponentielle de paramètre naturel: $\theta_i = \ln \frac{\pi_i}{1-\pi_i}$.
Cette relation définit la fonction logit pour la fonction de lien canonique associée à ce modèle. La loi binomiale conduit à des résultats identiques en considérant les sommes de n_i (n_i connus) variables de Bernoulli.

1.2.4.3 Loi Poisson

On considère *n* variables indépendantes y_i de la loi de Poisson de paramètre $\mu_i = \mathbb{E}(y_i)$. Les y_i sont par exemples les effectifs d'une table de contingence. les variables admettant pour densités:

$$\begin{aligned} f(y_i, \mu_i) &= \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \\ &= \exp\left\{ y_i \ln \mu_i - \ln(y_i!) - \mu_i \right\} \\ &= \exp\left\{ -\mu_i \right\} \frac{1}{y_i!} \exp\left\{ y_i \ln \mu_i \right\} \end{aligned}$$

qui sont issues d'une structure exponentielle et mis sous la forme canonique de paramètre naturel: $\theta_i = \ln \mu_i$ définissant comme fonction de lien canonique le logarithme pour ce modèle.

1.3 Estimation

1.3.1 Expression des moments

L'espérance et la variance de Y_i s'expriment en fonction des paramètres θ_i et ϕ, et sont liées. En effet, soit $l_i(\theta_i, \phi, y_i) = \ln(f(y_i, \theta_i, \phi))$ la contribution de la *ième* ob-

servation à la log-vraisemblance. On a $\forall\ i \in 1, \cdots, n$

$$\frac{\partial l_i}{\partial \theta_i} = \frac{[y_i - b'(\theta_i)]}{a_i(\phi)} \qquad et \qquad \frac{\partial^2 l_i}{\partial \theta_i^2} = \frac{-b''(\theta_i)}{a_i(\phi)}$$

Pour les lois issues de structure exponentielle, les conditions de régularités vérifiées permettent d'écrire:

$$\mathbb{E}\left(\frac{\partial l_i}{\partial \theta_i}\right) = 0 \qquad et \qquad \mathbb{E}\left(\frac{\partial^2 l_i}{\partial \theta_i^2}\right) = -\mathbb{E}\left(\left(\frac{\partial^2 l_i}{\partial \theta_i^2}\right)^2\right)$$

Alors $\mathbb{E}(y_i) = b^{'}(\theta_i)$ et $\mathbb{V}(y_i) = b^{''}(\theta_i)a_i(\phi)$. Cette dernière expression justifiant ainsi l'appellation du paramètre de dispersion pour ϕ lorsque a_i est la fonction identité. On a donc une relation directe entre l'espérance de y_i et sa variance.
$\mathbb{V}(y_i) = a_i(\phi)b^{''}((b^{'})(\mu_i)) = \phi b^{''}((b^{'-1})(\mu_i))$

Preuve:

Montrons que $\mathbb{E}(Y) = b^{'}(\theta)$ et $\mathbb{V}(Y) = b^{''}(\theta)a(\phi)$
Notons Z le support de f_Y; f_Y étant une densité de probabilité, elle vérifie

$$\int_Z f_Y(y, \theta, \phi)dy = 1$$

Pour alléger l'écriture, nous omettons d'écrire le paramètre ϕ dans $f_Y(y, \theta, \phi)$. Sous les conditions de dérivabilité sous le signe somme (la dérivée existe et est dominée uniformément par une fonction intégrable), on a:
- En dérivant par rapport à θ:

$$\begin{aligned}
\frac{\partial}{\partial \theta}\int_Z f_Y(y,\theta)d_y &= \int_Z \frac{\partial}{\partial \theta} f_Y(y,\theta)dy \\
&= \int_Z \frac{1}{f_Y(y,\theta)}\frac{\partial}{\partial \theta} f_Y(y,\theta) f_Y(y,\theta)dy \\
&= \int_Z \frac{\partial}{\partial \theta}\log f_Y(y,\theta) f_Y(y,\theta)dy \\
&= \int_Z \frac{1}{a(\phi)}\left[y - b^{'}(\theta)\right] f_Y(y,\theta)dy \\
&= \frac{1}{a(\phi)}\left(\mathbb{E}[Y] - b^{'}(\theta)\right) \qquad \textbf{or} \qquad \frac{\partial}{\partial \theta}\int_Z f_Y(y,\theta)dy = 0
\end{aligned}$$

ce qui donne $\mathbb{E}[Y] = b^{'}(\theta)$
- En dérivant une seconde fois, on obtient facilement que:

$$\int_Z \frac{\partial^2}{\partial^2\theta} \log f_Y(y,\theta)dy + \int_Z \frac{\partial}{\partial\theta} \left(\log f_Y(y,\theta)\right)^2 f_Y(y,\theta)dy = 0$$

Pour obtenir la variance, calculons les deux termes du membre de gauche de l'égalité, notés respectivement A et B. le premier terme vaut

$$\begin{aligned} A &= \frac{1}{a(\phi)} \int_Z \frac{\partial}{\partial\theta} \left[y - b^{'}(\theta)\right] f_y(y,\theta)dy \\ &= \frac{-1}{\partial\theta} \int_Z b^{''}(\theta) f_Y(y,\theta)dy \end{aligned}$$

Le second terme vaut:

$$\begin{aligned} B &= \frac{1}{a(\phi)} \int_Z \left[y - b^{'}(\theta)\right]^2 f_y(y,\theta)dy \\ &= \frac{1}{\partial(\phi)^2} \mathbb{V}(Y) \end{aligned}$$

D'après l'expression de $\mathbb{E}[Y] = b^{'}(\theta)$ la preuve se termine en annulant la somme de A et B.

1.3.2 Équation de vraisemblance

Considérons p variables explicatives dont les observations sont rangées dans la matrice de plan d'espérance X, β un vecteur de p paramètres et le prédicteur linéaire à n composantes

$$\eta = X\beta.$$

La fonction de lien g est supposée monotone différentiable telle que: $\eta_i = g(\mu_i)$;
c'est la fonction lien canonique si: $g(\mu_i) = \theta_i$.
Pour n observations supposées indépendantes et en tenant compte que θ_i dépend de β, la log-vraisemblance s'écrit:

$$L(\beta) = \sum_{i=1}^{n} \ln f(y_i, \theta_i, \phi) = \sum_{i=1}^{n} l_i(\theta_i, \phi, y_i).$$

Calculons:

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}.$$

Comme:

$$\begin{aligned}
\frac{\partial l_i}{\partial \theta_i} &= \frac{[y_i - b^{'}(\theta_i)]}{a_i(\phi)} = \frac{(y_i - \mu_i)}{a(\phi)} \\
\frac{\partial \theta_i}{\partial \mu_i} &= 1/\frac{\partial \mu_i}{\partial \theta_i} = 1/b^{''}(\theta_i) = \frac{a_i(\phi)}{\mathbb{V}(y_i)} \\
\frac{\partial \mu_i}{\partial \eta_i} & \quad \text{dépend de la fonction de lien : } g(\mu_i) = \eta_i \\
\frac{\partial \eta_i}{\partial \beta_j} &= X_{ij} \quad \textbf{car} \qquad \eta_i = x_i^T \beta
\end{aligned}$$

Alors les équations de la vraisemblance sont:

$$U_j = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\mathbb{V}(y_i)}\frac{\partial \mu_i}{\eta_i} = 0 \qquad j = 1, \cdots, p.$$

Ce sont des équations non-linaires en β dont la résolution requiert des méthodes itératives dans lesquelles interviennent le Hessien (pour Newton-Raphson) ou la matrice d'information (pour les scores de fisher).

1.3.3 Algorithme de Newton-Raphson

L'algorithme de Newton-Raphson approxime le logarithme de la fonction de vraisemblance dans un voisinage du paramètre initial par une fonction polynomiale qui a la forme d'une parabole concave. Elle a la même pente et la même courbure dans les conditions initiales que la log-fonction de vraisemblance. Il est facile de déterminer le maximum de ce polynôme d'approximation. Ce maximum fournit la seconde étape du processus d'estimation et l'on reprend la procédure décrite précédemment. Les approximations successives convergent rapidement vers les estimations au sens du maximum de vraisemblance. Chaque étape de l'algorithme de Newton-Raphson constitue un ajustement de type moindres carrés pondérés.
Ceci est une généralisation des moindres carrés ordinaires qui prend en compte la

non constance de la variance de Y dans les modèles. les observations recueillies en des points où la variabilité est plus faible sont affectées d'un poids plus important dans la détermination des paramètres.
L'algorithme de Newton-Raphson utilise la matrice d'information de Fisher. Cette matrice contient l'information concernant la courbure de la fonction de log-vraisemblance au point d'estimation. Plus grande est la courbure, plus l'information apportée au sujet des paramètres du modèle est importante. En effet, les écarts-types des estimations sont des racines carrées des éléments diagonaux de l'inverse de la matrice d'information de Fisher. Plus la courbure de la log-vraisemblance est importante, plus les écarts-types sont petits. Ceci est raisonnable dans la mesure où une grande courbure implique que le logarithme de la vraisemblance diminue rapidement quand on s'éloigne de l'estimation $\hat{\beta}$. En conséquence les données observées ont plus de chance d'apparaître si β prend la valeur $\hat{\beta}$ qu'une valeur éloignée de $\hat{\beta}$.

1.3.4 Algorithme des scores de Fisher

Pour traiter la plupart des équations non linéaires, on utilise l'algorithme de Newton-Raphson décrit dans la section (1.3.3). dans le cas des modèles linaires généralisés, on utilise l'algorithme de Fisher (voir Jennrich et Sampson (1976, [20])).
Soit $\beta^{(k)}$ la k^{ieme} approximation pour l'estimateur du maximum de vraisemblance $\hat{\beta}$.
Dans la méthode de Newton-Raphson on a:
$\beta^{(k+1)} = \beta^k - (H^{(k)})^{-1} q^{(k)}$
Où H est la matrice hessienne ayant pour élément $\frac{\partial^2 \ln \beta}{\partial \beta_j \partial \beta_h}$, q est le vecteur des dérivées ayant pour élément $\frac{\partial \ln(\beta)}{\partial \beta_s}$; $H^{(k)}$ et $q^{(k)}$ sont évalués en $\beta = \beta^{(k)}$.
La formule de l'algorithme des scores de Fisher s'écrit comme suit:
$\beta^{(k+1)} = \beta^k + (I_f(\beta^{(k)}))^{-1} q^{(k)}$ où $I_f(\beta^{(k)})$, d'élément $\left(-\mathbb{E}\left(\frac{\partial^2 \ln(\beta)}{\partial \beta_j \partial \beta_h}\right)\right)$, est la $k - ieme$ approximation de la matrice d'information de Fisher évaluée en $\beta = \beta^{(k)}$. On itère la procédure employée jusqu'à obtenir la stabilité.
La méthode de Fisher scoring peut s'interpréter comme une succession des moindres carrés, pondérés par des poids qui changent à chaque itération. L'estimation de la variance-covariance est un sous-produit de cette méthode. Pour cette raison l'algorithme est appelé *Moindre carrés repondérés itératifs*. La matrice d'information de Fisher d'un GLM s'obtient comme suit:

$$\begin{aligned} I_{jh} &= \mathbb{E}(U_j U_h) \\ &= \mathbb{E}\left\{ \sum_{i=1}^{n} \left[\frac{(y_i - \mu_i) x_{ij}}{\mathbb{V}(y_i)} \frac{\partial \mu_i}{\partial \eta_i} \right] \sum_{i=1}^{n} \left[\frac{(y_l - u_l) x_{lh}}{\mathbb{V}(y_l)} \frac{\partial \mu_l}{\eta_l} \right] \right\} \\ &= \sum_{i=1}^{n} \frac{\mathbb{E}\left[(y_i - \mu_i)^2\right] x_{ij} x_{ih}}{\left[\mathbb{V}(y_i)\right]^2} \left(\frac{\partial \mu_i}{\eta_i} \right)^2 \\ &= \sum_{i=1}^{n} \frac{x_{ij} x_i h}{\mathbb{V}(y_i)} \left(\frac{\partial \mu_i}{\eta_i} \right)^2 \end{aligned}$$

Car l'indépendance des y_i implique que $\mathbb{E}\left[(y_i - \mu_i)(y_l - u_l)\right] = 0$ pour $i \neq l$, et $\mathbb{E}\left[(y_i - \mu_i)^2\right] = \mathbb{V}(y_i)$. sinon.
Alors la matrice d'information de Fisher est:

$$I_F = X^T W X$$

où W est la matrice diagonale de pondération:

$$W_{ii} = \frac{1}{\mathbb{V}(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2.$$

La méthode d'itération suggère à l'étape (k+1) la procédure suivante:
- Choix d'une valeur initiale $\beta^{(0)}$ proche de β (exemple: par la méthode des moments)
- calcul des quantités:

$$\beta^{(k+1)} = \beta^{(k)} + \left[I_F^{(k)} \right]^{-1} U^{(k)} \qquad et \qquad I_F^{(k)} \beta^{(k+1)} = I_F^{(k)} \beta^{(k)} + U^{(k)}$$

La composante h de $I_F^{(k)} \beta^{(k)} + U^{(k)}$ évaluée en $\beta^{(k)}$ est :

$$\sum_{h=1}^{p} \sum_{i=1}^{n} \frac{x_{ij} x_{ih}}{\mathbb{V}(y_i)} \left(\frac{\partial \mu_i}{\eta_i} \right)^2 \beta_h^k + \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{\mathbb{V}(y_i)} \left(\frac{\partial \mu_i}{\eta_i} \right)$$

L'expression ci-dessous s'écrit sous une forme matricielle comme suit:

$$X^T W^{(k)} Z^{(k)},$$

où le vecteur $Z^{(k)}$ a pour composante:

$$Z_i^{(k)} = \sum_{h=1}^{p} x_{ih}\beta_h^{(k)} + (y_i - \mu_i)\left(\frac{\partial \mu_i}{\partial \eta_i}\right)$$

et $W^{(k)}$ la matrice diagonale des poids W_{ii} évaluée au point $(\mu_i^{(k)}, \eta_i^{(k)})$ avec $\eta_i^{(k)} = X_i^T\beta^{(k)}$ et $\mu_i^{(k)} = g^{-1}\left(\eta_i^{(k)}\right)$.
Alors à l'itération (k+1) l'algorithme se réécrit alors:

$$\beta^{(k+1)} = (X^T W^(k) X)^{-1}(X^T W^{(k)} Z^{(k)})$$

- Critère d'arrêt: pour ε petit, en pratique de l'ordre de 10^{-4}:

$$\left\|\beta^{(k+1)} - \beta^{(k)}\right\| < \varepsilon \quad \text{ou} \quad \left\|\ln\left(\beta^{(k+1)}\right) - \ln\left(\beta^{(k)}\right)\right\| < \varepsilon.$$

Remarque:

Cas particulier de la fonction de lien canonique: $\eta_i = \theta_i = x_i^T\beta$

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\theta_i} = b''(\theta_i);$$

On a une simplification de l'équation du score:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{var(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right) = \sum_{i=1}^{n}(y_i - \mu_i)x_{ij} = 0,\ j = 1, .., p.$$

d'où $X^T y = X^T \mu$. Pour le modèle linéaire $\mu = X^T\beta$, la solution par maximisation de la vraisemblance $\hat{\beta} = (X^T X)^{-1}(X^T y)$ coincide avec la solution pour minimisation des moindres carrés. (Diallo Alpha Oumar (2017, [10])).

1.3.5 Quasi-vraisemblance

La quasi-vraisemblance est une fonction des paramètres évaluée aux observations, à l'instar de la vraisemblance. Elle est définie comme suit:

$$Q(\mu, y) = \int_y^{\mu} \frac{y - t}{a_i(\phi)\mathbb{V}(t)} dt$$

Où Y est un vecteur d'observation y $i.i.d$ de moyenne $\mu = g^{-1}(X\beta)$ et de variance $\mathbb{V}(\mu)$. Cette fonction possède trois propriétés communes avec la log-vraisemblance

d'une loi exponentielle GLM:

$$\begin{aligned} a- \quad \mathbb{E}\left(\frac{\partial Q}{\partial \mu}\right) &= 0 \\ b- \quad \mathbb{E}\left(\frac{\partial^2 Q}{\partial \mu^2}\right) &= \frac{-1}{\phi \mathbb{V}(\mu)} \\ c- \quad \mathbb{V}\left(\frac{\partial Q}{\partial \mu}\right) &= \frac{1}{\phi \mathbb{V}(\mu)} \end{aligned}$$

Où $\mathbb{V}(\mu)$ est la fonction de variance associée à la loi exponentielle GLM ou à la quasi-vraisemblance. La quasi-vraisemblance possède les mêmes propriétés que la vraisemblance sur l'expérience des deux premières dérivées et sur la variance de la première dérivée. Donc choisir une loi exponentielle GLM ou une quasi-vraisemblance revient aux mêmes estimations. Pour plus de détails voir McCullagh et Nelder (1989, [23])

1.4 Propriétés asymptotiques

On définit: $H_n(\beta) = \frac{\partial^2 \ln(\beta)}{\partial \beta \partial \beta^T}$; $I_n(\beta) = -\mathbb{E}\left(H_n(\beta)\right) = X^T W_\beta^{-1} X$ avec $I_n(\beta)$ la matrice d'information de Fisher où $W_\beta = diag\left(\mathbb{V}(Y_1)g^T(\mu_1)^2; \cdots; \mathbb{V}(Y_n)g^T(\mu_n)^2\right)$. $I_n^{1/2}(\beta)$, la matrice d'ordre n, symétrique définie positive vérifiant: $I_n^{1/2}(\beta) I_n^{1/2}(\beta^T) = I_n(\beta)$ et l'inverse noté $I_n^{-1/2}(\beta)$.

On note encore β_0 la vraie valeur inconnue du paramètre, I_p la matrice identité en dimension p, $\lambda_{min}(A)$ la plus petite valeur propre de la matrice carrée. A et $\lambda_{min}(A)$ sa plus grande valeur propre.
Supposons les conditions suivantes vérifiées:
i- il existe deux constantes c et ε strictement positives, un entier n_1 et un voisinage $\mathbb{V}(\beta_0)$ tel que pour tout $\beta \in \mathbb{V}(\beta)$ et $\forall\ n \geqslant n_1$:

$$\lambda_{min}(I_n(\beta)) \geqslant c\lambda_{min}^{1/2+\varepsilon}(I_n(\beta_0));$$

ii- $\lambda_{min}(I_n(\beta_0)) \to +\infty$ lorsque $n \to +\infty$
iii- $\forall\ \sigma > 0$, $\forall\ \lambda \in \mathbb{R}$ tel que $\|\lambda\| = 1$, en posant $\beta = \beta_0 + \sigma I_n^{-\frac{1}{2}}(\beta_0)'\lambda$ pour n assez grand, on a:

$$\max \left\| I_n^{-\frac{1}{2}}(\beta_0)' - I_p \right\| \to 0 \text{ en probabilité lorsque } n \to +\infty$$

où $\mathbb{V}_n(\sigma)$ est le voisinage de β_0 défini par: $\mathbb{V}_n(\sigma) = \left\{ \beta \in \Theta \quad \text{ouvert de} \quad \mathbb{R}^p / \left\| I_n^{\frac{1}{2}}(\beta_0)'(\beta_0 - \beta) \right\| \leqslant \sigma \right.$

Théorème 1.1 ***Existence et consistance***
Si les hypothèses (i) et (ii) sont vérifiées, il existe une suite $(\hat{\beta}_n)_{n\geqslant 1}$ *de variables aléatoires et une variable aléatoire* n_2 *à valeurs entières telles que :*
1- $\forall n \geqslant n_2, P(\nabla \ln(\hat{\beta}_n) = 0) = 1$
2- la suite $(\hat{\beta}_n)_{n\geqslant 1}$ *converge presque surement vers* β_0

Théorème 1.2 ***Normalité asymptotique***

Si les hypothèses (ii) et (iii) sont vérifiées, la suite des estimateurs de maximum de vraisemblance $(\hat{\beta}_n)_{n\geqslant 1}$ *est asymptotiquement gaussienne, c'est-à-dire:* $I_n^{1/2}(\beta_0)^T(\hat{\beta}_n - \beta_0) \to \mathbb{N}(0, I_p)$ *en loi. L'estimateur du maximum de vraisemblance* $\hat{\beta}_n$ *est donc asymptotiquement gaussien.*

Pour la démonstration de ces théorèmes, voir Fahrmeir et Kaufmann (1985, [15]). Pour une étude des propriétés asymptotiques des estimateurs, voir Antoniadis et al (1992, [2]).

1.5 Qualité de l'ajustement

Il s'agit d'évaluer la qualité d'ajustement du modèle sur la base des différences entre observations et estimation. Plusieurs critères sont proposés:

1.5.1 Déviance

Le modèle estimé est comparé avec le modèle dit saturé, c'est-à-dire le modèle possède autant de paramètres que d'observations et estimant donc exactement les données. Cette comparaison est basée sur l'expression de la déviance $\mathcal{D}$ des log-vraisemblances $\mathcal{L}$ et $\mathcal{L}_{sat}$:

$$\begin{aligned} \mathcal{D} &= 2\left[\sum_{i=1}^{n}\left(l_{sat}(y_i) - l(y_i, \beta^2, \phi)\right)\right] \geqslant 0 \qquad \forall \qquad \phi \qquad \text{fixé} \\ &= 2(\mathcal{L}_{sat} - \mathcal{L}) \geqslant 0 \end{aligned}$$

qui est le logarithme du carré du rapport des vraisemblances. Ce rapport remplace l'usage des sommes de carrés propres au cas gaussien et donc à l'estimation par

moindres carrés.
On montre qu'asymptotiquement, $\mathcal{D}$ suit une loi du $\mathcal{X}^2$ à $n-p$ degré de liberté, ce qui permet de construire un test de rejet ou l'acceptation du modèle selon que la déviance est jugée significativement ou non importante.

Remarque 1.3 *L'approximation de la loi du $\mathcal{X}^2$ peut être douteuse. De plus, dans le cas des données non groupées (modèle binomiale), le cadre asymptotique n'est plus adapté car le nombre de paramètres estimés tend également vers l'infini avec n et il ne faut plus se fier à ce test.*

1.5.2 La Statistique du khi-deux de Pearson

Un test du $\mathcal{X}^2$ est également utilisé pour comparer les valeurs observées y_i à leur prévision pour le modèle. La statistique du test est définie par:

$$\mathcal{X}^2 = \sum_{i=1}^{I} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mathbb{V}}(\hat{\mu}_i)}$$

(μ_i est remplacé par $n_i\pi_i$ dans le cas binomial) et on montre qu'elle admet asymptotiquement la même loi que la déviance.
En pratique ces deux approches conduisent à des résultats peu différents et dans le cas contraire, c'est une indication de mauvaise approximation de la loi asymptotique. Sachant que l'espérance d'une loi de $\mathcal{X}^2$ est son nombre de degré de liberté et connaissant les aspects approximatifs des tests construits, l'usage est souvent de comparer les statistiques avec le nombre de degré de liberté. Le modèle peut être jugé satisfaisant pour un rapport $\frac{\mathcal{D}}{ddl}$ plus petit que 1.

1.6 Les tests d'hypothèses

Pour les GLMs, la loi de l'estimateur du maximum de vraisemblance n'est connue que de manière asymptotique (pour n suffisamment grand). Cela entraine que toute la démarche d'analyse (tests de comparaison de modèles, intervalles de confiance des valeurs des paramètres,$\cdots$) est conduite dans un cadre asymptotique.
Cette section présente différents tests d'hypothèses qui vont permettre d'examiner les qualités du modèle, de déterminer si les différents variables explicatives présentes dans le modèles sont pertinentes ou non.

1.6.1 Test de modèles emboités

Le test de comparaison des modèles permet de déterminer si un sous ensemble des variables explicatives suffit à expliquer la variable Y. On rappelle que les modèles M_1 et M_0 respectivement définis par $g(\mu) = X_1\beta_1$ et $g(\mu) = X_0\beta_0$ sont dits emboités si le modèle M_0 est un cas particulier du modèle général M_1, c'est dire si le sous-espace engendré par les colonnes de X_0 est inclus dans le sous espace linéaire engendré par les colonnes de X_1. Le test des hypothèses $H_0 : \{M_0\}$ contre $H_1 : \{M_1\}$ est alors réalisé à l'aide d'un test du rapport de vraisemblance dont la statistique de test s'écrit:

$$\begin{aligned} T &= -2\log\frac{\mathcal{L}(y,\hat{\beta}_0)}{\mathcal{L}(y,\hat{\beta}_1)} \\ &= -2\left(l(y,\hat{\beta}_0) - l(y,\hat{\beta}_1)\right) \end{aligned}$$

Où $\hat{\beta}_0$ et $\hat{\beta}_1$ sont respectivement les estimateurs du maximum de vraisemblance de $\hat{\beta}$ dans les modèles M_0 et M_1.
Sous certaines hypothèses, on peut montrer que cette statistique de test converge en loi vers une loi du $\mathcal{X}^2$ à $q_1 - q_0$ degrés de liberté, où q_0 et q_1 sont respectivement les dimensions des espaces engendrés par les colonnes de X_0 et X_1. Ainsi si on effectue le test au niveau α, on rejettera H_0 au profit de H_1 si $T \geqslant \mathcal{X}^2_{1-\alpha,q_1-q_0}$ où $\mathcal{X}^2_{1-\alpha,q_1-q_0}$ est le quantile d'ordre $1-\alpha$ de la loi du $\mathcal{X}^2$ à $q_1 - q_0$ degrés de liberté.
Ce même test est parfois présenté sous une forme légèrement différente reposant sur la déviance, ce qui est l'écart entre le logarithme de la vraisemblance du modèle d'intérêt M et celui du modèle le plus complet possible, appelé modèle saturé, noté M_s. Le modèle saturé est le modèle comparant n paramètres, c'est-à-dire autant que l'observation. La déviance du modèle M est alors définie par:

$$D(M) = -2\left(l(y,\hat{\beta}) - l(y,\hat{\beta}_s)\right)$$

La statistique de test T du rapport de vraisemblance peut être réécrite en terme de déviance sous la forme:

$$T = D(M_o) - D(M_1)$$

Le test global du modèle consiste à tester $H_0 = \{g(\mu_i) = x_i\beta\}$ contre $H_1 = \{g(\mu_i) \neq x_i\beta\}$ à l'aide du test du rapport de vraisemblance. Il permet de savoir si toutes les

variables sont utiles pour expliquer la variable réponse Y.

1.6.2 Test de Wald

Si la réponse au test global est positive, la suite logique consiste à tester quelles sont les variables ou facteurs qui ont une influence. La connaissance de la loi asymptotique de $\hat{\beta}$ permet de construire des tests sur les paramètres β, sur des combinaisons linéaires de β ou encore de μ_i ainsi que des intervalles de confiance. Tous ces résultats sont asymptotiques. On souhaite tester l'hypothèse $H_0 : \{\beta_i = \beta_k\}$ contre $H_1 : \{\beta_i \neq \beta_k\}$ où β_k est une valeur définie à priori. Sous l'hypothèse H_0, $T_i = I(\beta, \phi)_{ii}(\hat{\beta}_i - \hat{\beta}_k)^2$ converge en loi vers un $\mathcal{X}^2$ et $P(T_i > t_i)$ donne une p-valeur asymptotique du test. En général on utilise ce test avec $\beta_k = 0$ afin de déterminer si le paramètre β_1 est significativement non nul.
En pratique, comme il a été précisé plus haut, l'information de Fisher est calculée non pas en les vrais paramètres qui sont inconnus mais en $\hat{\beta}$ et $\hat{\phi}$. La statistique de test est donc $T_i = I_n(\hat{\beta}, \hat{\phi})_{ii}(\hat{\beta}_i - \hat{\beta}_k)^2$.
Attention: Le test de Wald approximatif peut ne pas être précis si le nombre d'observations est faible.
Pour plus d'informations concernant ces statistiques, voir McCullagh et Nelder (1989, [23]) ou Dobson (2018, [12]).

1.7 Choix entre différents modèles

Les critères de choix de modèles tel que le AIC ou le BIC sont souvent utilisés pour comparer entre eux des modèles qui ne sont pas forcément emboités les uns dans les autres. voir Akaike (1973, [1]) et Schwarz (1978 [32]).
-Le AIC (Akaike Information Criterion) pour un modèle à p paramètres est défini par:

$$AIC(M(\hat{\beta})) = -2l(y, \hat{\beta}) + 2p$$

-Le BIC (Bayesian Information Criterion) pour un modèle à p paramètres estimé sur n observations est défini par:

$$BIC(M(\hat{\beta})) = -2l(y, \hat{\beta}) + np$$

L'estimation de ces critères est simple. Pour chaque modèle concurrent le critère de choix de modèle est calculé et le modèle qui présente le plus faible critère est sélectionné.

1.8 Diagnostics, résidus

De nombreux indicateurs, comme dans le cas de la régression linéaire, sont proposés afin d'évaluer la qualité ou la robustesse des modèles estimés. Ils concernent la détection valeurs influentes et l'étude graphique des résidus.

1.8.1 Effet levier

On construit la matrice de projection

$$H = W^{1/2}X(X^{'}WX)^{-1}X^{'}W^{1/2}$$

relative au produit scalaire de la matrice W (matrice de pondération), sur le sous-espace engendré par les variables explicatives. les termes diagonaux de cette matrice supérieure à $(3p/n)$ indiquent les valeurs potentiellement influentes. Le graphe représentant les points d'ordonnés h_{ii} d'abscisse le numéro de l'observation les visualise.

1.8.2 Résidus

Avec des erreurs centrées, additives, c'est-à-dire dans le cas du modèle gaussien utilisant la fonction lien identité, il est naturel de définir les résidus par:

$$\varepsilon_i = y_i - \mathbb{E}(y_i) = y_i - \mu_i$$

comme dans le cas du modèle linaire. Ce cadre est ici inadapté au cas général et différents substituts sont proposés. Chacun possède par ailleurs une version standardisée et une version studentisée.

-Person:

Les résidus obtenus en comparant valeurs observées y_i et valeurs prédites $\hat{y}_i$ sont pondérés par leur précision estimée par l'écart-type: s_i de $\hat{y}_i$. Ceci définit les résidus

de Pearson:

$$rp_i = \frac{y_i - \hat{y}_i}{s_i}$$

dont la somme des carrées conduit à la statistique de même nom. Ces résidus mesurent donc la contribution de chaque observation à la significativité du test découlant de cette statistique. Par analogie du modèle linaire, on vérifie que ce sont également les résidus de la projection par la matrice H.
Ces résidus ne sont pas de variance unité et sont donc difficiles à interpréter. Une estimation de leurs écart-types conduit à la définition des résidus de Pearson standardisés:

$$rp_{si} = \frac{y_i - \hat{y}_i}{s_i\sqrt{h_{ii}}}$$

faisant intervenir le terme général de la matrice H.
De plus, prenant en compte que les estimations des écart-types s_i dépendent de la i^{eme} observation et sont donc biaisées, des résidus studentisés sont obtenus en approchant au premier ordre le paramètre de dispersion $s_{(i)}$ calculé sans la i^{eme} observation

$$rp_{si} = \frac{y_i - \hat{y}_i}{s_{(i)}\sqrt{h_{ii}}}$$

-Déviance

Ces résidus mesurent la contribution de chaque observation à la déviance du modèle par rapport au modèle saturé. Des versions standardisées et studentisées en sont définies comme ceux de Pearson Voir (L.Bel et al. (2016, [3])).

Les applications de ces modèles à des données réelles ont révélé que les données de comptages possèdent des distributions ayant des caractéristiques particulières comme la non normalité, l'hétérogénéité des variances ainsi qu'un nombre important de zéros (voir Hilbe (2011, [19])). Il est donc nécessaire de connaitre les données de comptages, d'utiliser les modèles appropriés à ces données afin d'obtenir les résultats non biaisés.

2

Données de comptage et inflation de zéros

Résumé

Dans ce chapitre, nous parlons des données de comptages, de l'inflation de zéros qui est l'une des causes de la surdispersion et des différents modèles qui permettent de les traiter.

Sommaire

2.1 Introduction

En statistique, les données de comptage sont un type de données statistiques qui prennent des valeurs entières non négatives et dont les valeurs proviennent d'un processus de comptage (processus de dénombrement). Ces données se produisent généralement dans toutes les disciplines. Une approche pour modéliser ces types de données est la régression de poissons. La régression de poisson fonctionne sous l'hypothèse d'égalité de la variance et de la moyenne. Ce qui n'est pas toujours le cas dans de nombreux scénarios de la vie réelle. Si l'on néglige cette hypothèse de l'égalité, une surdispersion se produit. Cette surdispersion est due à l'hétérogénéité ou à un nombre important de zéros (Hilbe (2011, [19])). Pour palier à ce problème, la régression binomiale négative est proposée comme une alternative pour corriger le problème d'équi-dispersion caractérisant le modèle de poisson. Elle est une généralisation de la régression de poisson qui assouplit l'hypothèse restrictive que la variance est égale à la moyenne.
Dans ce chapitre nous allons étudier d'abord le modèle de régression Poisson ensuite le phénomène de surdispersion qui intervient dans les données de comptage et les différentes méthodes pour résoudre ce problème.

2.2 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 des 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 :

$$f(y,\mu) = \frac{\mu^y}{y!}\exp(-\mu) = \exp\left(-\mu + yln(\mu) - ln(y!)\right)$$

avec ($\mu > 0$) 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 $\mathbb{V}$ et sa fonction de lien canonique g:

$$\mathbb{V}(\mu) = \mu \qquad g(\mu) = \ln(\mu)$$

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

Cependant le modèle de régression de Poisson fait face à une surdispersion de la variance puisque l'hypothèse forte de la loi de Poisson $\mathbb{E}(Y) = \mathbb{V}(Y) = \mu$ est assez difficile à vérifier dans la réalité. Dans la suite, nous allons étudier la surdispersion et les modèles de régression adaptés pour gérer cette dernière.

2.3 La surdispersion

Une caractéristique importante de la loi de Poisson est l'égalité entre l'espérance et la variance. Malheureusement, l'égalité de l'espérance et de la variance n'est pas toujours maintenue; même pas approximativement. Ce qui entraine la dispersion. La dispersion se définit comme une mesure quantifiable de la variabilité des données autour d'une valeur centrale. Posons Y la réponse de notre modèle. $Y \sim \mathcal{P}(\mu)$ suit une loi de Poisson de paramètre μ. On appelle surdispersion ou variabilité extra-poissonnienne si au lieu d'avoir $\mathbb{V}(Y) = \mathbb{E}(Y) = \mu$, nommée équi-dispersion, on observe $\mathbb{V}(Y) > \mu$. Par ailleurs dans le cas où $\mathbb{V}(Y) < \mu$, on se trouve dans le cas de la sous-dispersion.
Avec une variabilité dite comptée, la surdispersion est un phénomène très courant dans les données de comptage. Elle peut résulter par exemple des données non indépendantes, d'un excès de comptage zéro ou de l'absence d'une variable quand même importante à inclure dans le modèle. La présence de la surdispersion est détectable 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 supérieur à 1 on se trouve dans le cas sur-dispersé.
Il existe également un test performant pour détecter la surdispersion, utilisable dans le cas où on a estimé nos paramètres par la méthode du maximum de vraisemblance. Il s'agit du test de Dean (1989, [9])
Pour prendre en compte la surdispersion des données, on introduit un paramètre de dispersion noté ϕ et un vecteur de poids noté w pour obtenir la relation suivante: $\mathbb{V}(Y) = \frac{\phi}{w}\mathbb{E}(Y) = \frac{\phi}{w}\mu$. Il s'agit d'un paramètre basé sur la quasi-vraisemblance et dans notre cas, on parle de quasi-poisson. Il en résulte que si $\phi > 1$, on a mis en évidence, la surdispersion 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 poisson habituel. Un modèle basé sur la quasi vraisemblance s'explique de la manière suivante:

$$\begin{cases} (Y_i, X_i^1, ..., X_i^p) \quad \text{sont indépendantes pour } i=1...n \\ \mu_i = \mathbb{E}[Y_i/X_i] \quad \text{oú} \quad X_i = X_1...X_n \\ g(\mu_i) = \eta_i = X_i\beta \quad \text{application linéaire en } \beta \text{ où g(.) est une fonction de lien et} \\ \eta_i \quad \text{est le prédicteur linéaire} \\ \mathbb{V}[Y_i/X_i] = \frac{\phi}{w_i}\mathbb{V}(\mu_i) \quad \text{où} \quad \mathbb{V}(.) \text{ est la fonction de variance} \end{cases}$$

Étant donnée qu'on obtient les estimateurs et leurs erreurs-types par des méthodes de quasi-vraisemblance et non par pure vraisemblance, on ne peut pas appliquer des tests de comparaison de modèles se basant sur celle-ci.
Pour estimer le paramètre de dispersion, on peut se servir de la statistique de Pearson qui suit à peu près une loi du Chi-deux $\mathcal{X}^2$ à $n-p-1$ dégrées de liberté:

$$\mathcal{X}^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\mathbb{V}[\hat{\mu}_i]} w_i$$

La valeur du paramètre de dispersion est alors estimé par $\hat{q} = \frac{\mathcal{X}}{m-p-1} = \hat{\sigma}^2$.
Pour obtenir le maximum de log-quasi-vraisemblance $Q(\mu, y) = \sum_{i=1}^{n} Q_i(\mu_i, y_i)$ où $Q_i(\mu_i, y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi\mathbb{V}(t)} dt$ où $\phi\mathbb{V}(\mu_i)$ est la variance de y_i. Ceci revient alors dans le cas poissonnien à:

$$Q_i(\mu_i, y_i) = \frac{1}{\phi} y_i \ln(\mu_i) - \mu_i + C \quad \text{avec C une constante}$$

Preuve
Par définition de quasi-vraisemblance, on a:

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

Avec $C = y_i - y_i \ln(y_i)$

Une autre méthode pour traiter la surdispersion est de choisir un modèle respectivement une loi qui est mieux adaptée pour tenir compte de cette variation extra-poissonnienne.

2.4 Modèle de régression binomiale négative

En régression binomiale négative la moyenne de Y est déterminée par le temps d'exposition t et un ensemble de n variables de régression. L'expression reliant ces quantités est:
$\lambda_i = \exp(\ln(t_i) + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni})$ souvent $x_1 = 1$, auquel cas β_1 est appelé l'ordonnée à l'origine. Les coefficients de régression $\beta_1, \beta_2, \cdots, \beta_n$ sont des paramètres inconnus qui sont estimés à partir d'un ensemble de données. Leurs estimations sont symbolisées par $b_1, b_2, \cdots, b_n$.
On suppose que $Y \mid \Theta \sim \mathcal{P}(\lambda\Theta)$, où Θ suit une loi Gamma de paramètre identique α (de telle sorte que $\mathbb{E}(\Theta) = 1$), on obtient la loi binomiale négative:

$$\mathbb{P}(Y = k) = \frac{\Gamma(k + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(k+1)} \left(\frac{1}{1+\alpha\lambda}\right)^{\alpha^{-1}} \left(\frac{\alpha\lambda}{1+\alpha\lambda}\right)^{k}$$

On peut réécrire cette loi, en posant $r = \alpha^{-1}$ et $p = \frac{1}{1+\alpha\lambda}$

$$f(y) = \binom{y}{y+r-1} p[1-p]^y, \forall k \in \mathbb{N}$$

ou encore

$$f(y) = \exp\left[y\ln(1-p) + r\ln p + \ln\binom{y}{y+r-1}\right], \forall k \in \mathbb{N}$$

qui est une loi de la famille exponentielle, en posant: $\theta = \ln(1-p)$, $b(\theta) = -r\ln(p)$ et $a(\phi) = 1$
si on calcule la moyenne, on obtient

$$\mathbb{E}(Y) = b^{'}(\theta) = \frac{\partial b}{\partial p}\frac{\partial p}{\partial \theta} = \frac{r(1-p)}{p} = \lambda$$

et si on calcule la variance, on obtient:

$$\mathbb{V}(Y) = b^{''}(\theta) = \frac{\partial^2 b}{\partial p^2}\left(\frac{\partial p}{\partial}\right)^2 + \frac{\partial b}{\partial p}\frac{\partial^2 p}{\partial \theta^2} = \frac{r(1-p)}{p^2}$$

Autrement dit:

$$\mathbb{V}(Y) = \frac{1}{p}\mathbb{E}(Y) = [1 + \alpha - \lambda]$$

Pour une régression binomiale négative de type 2 (NB2) on a:

$$\mathbb{E}(Y) = \lambda = \mu \quad \text{et} \quad \mathbb{V}(Y) = \lambda + \alpha\lambda^2$$

Le lien canonique est $g(\lambda) = \theta$, c'est-à-dire:

$$g(\mu) = \ln\left(\frac{\alpha\mu}{1+\alpha\mu}\right)$$

NB:
Si $\alpha = 0$, On a une loi de Poisson, Si $\alpha = 1$, on a une loi géométrique.
La régression binomiale de type 1 (NB1) permet d'avoir une régression géométrique c'est-à-dire

$$\mathbb{E}(Y) = \lambda = \mu \quad \text{et} \quad \mathbb{V}(Y) = \lambda + \lambda^2$$

Le lien canonique est alors $g(\lambda) = \theta$, c'est-à-dire $g(\mu) = \ln\left(\frac{\mu}{1+\mu}\right)$

La régression de Poisson et la régression binomiale négative sont des méthodes adaptées à la modélisation d'une variable dépendante discrète avec une distribution asymétrique. Toutefois la régression de Poisson est susceptible d'être confrontée à un problème de surdispersion. La surdispersion des données cause la sous-estimation des erreurs standards et la sur-estimation des tests statistiques. La régression binomiale négative est souvent la solution préférable dans cette situation. Cependant la surdispersion peut être liée à un nombre excessif de zéros. Les modèles ZIP (modèle de Poisson avec inflation de zéros), les modèles ZINB (modèle binomiale négative à inflation de zéros), les modèles ZIB (modèle binomiale à inflation de zéros) et les modèles de Hurdle sont couramment utilisés pour traiter ce phénomène statistique.

2.5 Modèle de régression à inflation de zéros

Dans cette partie, nous allons nous intéresser à un cas particulier de surdispersion appelée **inflation de zéros**. Ce phénomène que nous définissons plus précisément dans la suite, intervient lorsqu'on observe un nombre excessif de zéros dans les données de comptage. Il existe plusieurs méthodes pour modéliser ces types de données. Nous allons nous intéresser particulièrement à une classe de modèle appelé **modèles à inflation de zéros** qui est un mélange entre une masse en 0 et un modèle classique de comptage.

2.6 Modèle de régression ZIP

2.6.1 Définition

Soit Y une variable de comptage sur un échantillon de n individus. On note Y_i l'observation de Y. La probabilité pour qu'un individu i soit dans le groupe des zéros est notée ϕ_i. La variable Y_i est modélisée par un ZIP par:

$$\mathbb{P}(Y_i = y_i|X_i, \Phi_i) = \begin{cases} \phi_i + (1-\phi_i)\exp^{-\mu_i} & \text{si} \quad y_i = 0 \\ (1-\phi_i)\frac{\exp^{-\mu_i}\mu_i^{y_i}}{y_i!} & \text{si} \quad y_i = 1, 2, \cdots \end{cases} \tag{2.1}$$

où ϕ_i et μ_i sont fonctions respectivement des vecteurs de covariables $\Phi_i = (\Phi_{il}, \cdots, \Phi_{iq})^T$ et $X_i = (X_{il}, \cdots, Xip)^T$. Dans la régression ZIP, la probabilité de mélange ϕ_i et le paramètre μ_i, sont généralement modélisés par des modèles logistiques et log-linéaires respectivement, c'est-à-dire:

$$logit(\phi_i) = \gamma^T \Phi_i \quad \text{et} \quad \log(\mu_i) = \beta^T X_i \tag{2.2}$$

où $\beta = (\beta_1, \cdots, \beta_p)^T$ et $\gamma = (\gamma_1, \cdots, \gamma_p)^T$ sont des vecteurs de paramètres inconnus. On peut synthétiser le modèle sous la forme suivante:

$$\forall i = 1, \cdots, n, \begin{cases} Y_i \sim \phi_i \delta_0 + (1-\phi_i)\mathcal{P}(\mu_i) \\ logit(\phi_i) = \gamma^T \Phi_i \\ log(\mu_i) = \beta^T X_i \end{cases} \tag{2.3}$$

Conditionnellement à X_i et ϕ_i, l'espérance et la variance de Y_i sont données par: $\mathbb{E}(Y_i|X_i, \Phi_i) = (1-\phi_i)\mu_i$ et $\mathbb{V}(Y_i|X_i, \Phi_i) = (1+\phi_i\mu_i)(1-\phi_i)\mu_i$

2.6.2 Estimation dans le modèle ZIP

Les méthodes d'estimations dans un contexte de régression de Poisson avec inflation de zéros ont été proposées par plusieurs auteurs. Généralement, l'estimation du maximum de vraisemblance est utilisée pour estimer de tels modèles (voir Lambert (1992, [22]), Czado et al. (2007, [8])). Cependant, il est bien connu que l'EMV est très sensible à la présence des valeurs aberrantes et peut devenir instable lorsque les composantes du mélange sont mal séparées. Pour pallier à ce problème, Hall et Shen (2010, [18]) ont suggéré une nouvelle procédure d'estimation du modèle (2.1) dite *robust expectation-solution (RES) estimation* ou tout simplement l'algorithme ES (*expectation solution*). Cet algorithme est une modification de l'algorithme *expectation maximization (EM)* (voir Dempster et al (1997, [11])) avec la propriété de robustesse. Dans cette partie, nous discutons brièvement de cet algorithme ES et des

propriétés asymptotiques de l'estimateur sous certaines conditions. On considère également que tous les individus n'ont pas forcément la même probabilité d'appartenir au groupe des zéros.

La log-vraisemblance du modèle, basée sur les observations $(Y_i, X_i, \Phi_i)\ i = 1, \cdots, n$ est :

$$l_n(y, \gamma, \beta) = \sum_{i=1}^{n} \left\{ \lambda_i \log[e^{\gamma^T \Phi_i} + \exp(-e^{\beta^T X_i})] \right\} + \sum_{i=1}^{n} \left\{ (1 - \lambda_i)(y_i \beta^T X_i - e^{\beta^T X_i} - \log(y!)) \right\} - \sum_{i=1}^{n} \left\{ \log(1 + e^{\gamma^T \Phi_i}) \right\},$$

où $\lambda_i = 1$ si $y_i = 0$ et $\lambda_i = 0$ sinon.
En particulier, supposons que l'on observe la variable indicatrice v telle que $v_i = 1$ si y_i provient de l'ensemble des zéros (distribution dégénérée) et $v_i = 0$ si y_i résulte du zéro aléatoire (distribution non dégénérée). Alors la log-vraisemblance pour les données complètes (y, v) est donnée par:

$$\begin{aligned} l_n^c(y, v, \gamma, \beta) &= \sum_{i=1}^{n} \left\{ v_i \gamma^T \Phi_i - \log(1 + e^{\gamma^T \Phi_i}) \right\} + \sum_{i=1}^{n} (v_i) \left\{ y_i \beta^T X_i - e^{\beta^T X_i} - \log(y_i!) \right\} \\ &= l_\gamma^c(\gamma, y, v) + l_\beta^c(\beta, y, v), \end{aligned} \tag{2.4}$$

où $v = (v_1, \cdots, v_n)^T$.

Cette log-vraisemblance est sous une forme appropriée car l_γ^c et l_β^c peuvent être maximisés séparément. Le principe de cet algorithme consiste à maximiser la fonction $l_n^c(y, v, \gamma, \beta)$ de manière itérative en commençant par une valeur initiale $(\beta^{(0)T}, \gamma^{(0)T})^T$. A l'itération $r + 1$, nous avons les deux étapes suivantes:
1.étape E: estimer la variable v_i par son espérance conditionnelle aux observations $v_i^{(r)}$ sous les estimations courantes des paramètres $\beta^{(r)}$ et $\gamma^{(r)}$]. Cette espérance est donnée par:

$$v_i^r = \begin{cases} \left[1 + \exp\left(-\gamma^{(r)T} \phi_i - e^{\beta(r)^T X_i}\right)\right]^{-1} & \text{si} \quad y_i = 0 \\ 0 & \text{si} \quad y_i > 0 \end{cases} \tag{2.5}$$

2.étape M: trouver $\beta^{(r+1)}$ et $\gamma^{(r+1)}$ en maximisant respectivement les fonctions $l_\gamma^c(\gamma, y, v^{(r)})$ et $l_\beta^c(\beta, y, v^{(r)})$.
Hall et Shen (2010, [18]) ont montré que maximiser ces deux fonctions revient à résoudre respectivement les deux équations suivantes:

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ v_i^{(r)} - \phi_i \right\} \Phi = 0 \tag{2.6}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(1-v_i^{(r)}\right)\left\{y_i-e^{\beta^T X_i}\right\}=0 \tag{2.7}$$

Dans l'approche RES, Hall et Shen (2010, [18]) proposent de remplacer les équations (2.6) et (2.7) par les estimations des fonctions robustes. Essentiellement, ils proposent de pondérer les observations qui se situent dans la queue extrême supérieure et inférieure de la distribution de Poisson dans la fonction d'estimation. Sous les conditions de régularité de Rosen et al. (2000, [31]) liées à l'algorithme ES et de Caroll et al. (1995, [5]), Hall et Shen (2010, [18]) ont montré le résultat suivant (qui généralise le théorème 1 dans Czado et al. (2007, [8]) dans le sens où ψ=$(\beta^T,\gamma^T)^T \quad \in \mathbb{R}^{p+q}$):

Théorème 2.1 *si l'algorithme RES converge, alors il existe une suite de variables aléatoires* $\hat{\psi}_n$ *telle que:*
i- $\hat{\psi}_n \to \psi_0$ *en probabilité quand* $n \to \infty$ *(consistance),*
ii- $\sqrt{n}(\hat{\psi}_n-\psi_0) \to \mathcal{N}(0,\mathbb{V}(\psi_0))$ *en loi quand* $n \to \infty$ *(normalité asymptotique).*

où l'expression $\mathbb{V}(\psi_0)$ de la variance asymptotique est donnée dans Hall et Shen (2010, [18]).

2.7 Modèle de régression ZINB

Le phénomène d'inflation de zéros a été constaté pour la première fois sur les données de comptage. Il a donc fallu la mise en place de nouveaux outils plus adaptés, comme les modèles de régression ZIP et ZINB pour traiter ce type de problème. En effet pour modéliser les données de comptage, la régression de poisson suppose que la moyenne conditionnelle est égale à la variance conditionnelle, ce qui peut ne pas être valable dans certaines situations. Si les données ont une variance supérieure à celle proposée par le modèle de poisson, une surdispersion se produirait. Le modèle binomiale négatif avec inflation de zéro est une solution pour résoudre le problème.
Pour une variable réponse Y_i, $i=1,\cdots,n$ on dira que Y_i est modélisée par un ZINB si sa distribution est donnée par:

$$\mathbb{P}(Y_i=y_i|X_i,\Phi_i)=\begin{cases}\phi_i+(1-\phi_i)\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha} & \text{si} \quad y_i=0\\ (1-\phi_i)\frac{\Gamma(y_i+1/\alpha)}{\Gamma(1/\alpha)y_i!}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}\left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha} & \text{si} \quad y_i=1,2,\cdots\end{cases} \tag{2.8}$$

avec $\mathbb{E}(y_i|X_i,\Phi_i)=(1-\phi_i)\mu_i$ et $\mathbb{V}(Y_i|X_i,\Phi_i)=(1-\phi_i)\mu_i(1+(\alpha\phi_i)\mu_i)$
où α est un paramètre de surdispersion. Dans les deux cas ϕ_i représente la probabilité d'inflation de zéros. Comme pour les modèles Poissons et binomiale négative, le modèle ZINB tend vers le modèle ZIP lorsque α tend vers zéro.

L'étude des propriétés asymptotiques dans le modèle ZINB peut se faire de manière similaire à celle effectuée précédemment dans le modèle ZIP. Pour plus de détails voir Hilbe (2011, [19]), Czado et al. (2007, [8]).

2.8 Modèle de régression ZIB

Le modèle de régression binomiale zéro-inflaté a été utilisé en premier par Kemp et Kemp (1988, [21]), mais ce n'est que vers les années (2000) que Hall (2000, [17]) et Vieria et al. (2000, [34]) l'ont introduit de manière beaucoup plus claire et ont donné quelque application détaillées dans le cadre des données réelles. En considérant les mêmes notations que Hall (2000, [17]), le modèle ZIB a deux états définis comme suit:

$$Y_i \sim \begin{cases} 0 & \text{avec une probabilité} \quad P_i \\ \text{Binomiale}(n_i,\pi_i) & \text{avec une probabilité} \quad 1-P_i \end{cases}$$

ce qui implique

$$Y_i = \begin{cases} 0 & \text{avec une probabilité} P_i+(1-p_i)(1-\pi_i)^{n_i} \\ k & \text{avec une probabilité}(1-P_i)\begin{pmatrix} ni \\ k \end{pmatrix}\pi_i^{n_i}(1-\pi_i)^{n_i-k}, \quad k=1,2,\cdots,n_i \end{cases}$$

avec $\mathbb{E}(Y_i)=(1-P_i)n_i\pi_i$ et $\mathbb{V}(Y_i)=(1-P_i)n_i\pi_i(1-\pi_i(1-P_in_i))$

Les deux probabilités peuvent également être exprimées conjointement comme une distribution de Bernoulli généralisée donnant la vraisemblance suivante:

$$l_n(\beta,\gamma)=\prod_{i=1}^{n}(P_i(1-P_i)(1-\pi_i)^{n_i})^{J_i}.\left((1-P_i)\begin{pmatrix} n_i \\ k \end{pmatrix}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i}\right)^{1-J_i}$$

Les paramètres $P=(P_1,\cdots,P_n)$ et $\pi=(\pi_1,\cdots,\pi_n)$ sont respectivement modélisés via une fonction de lien logit,

$$logit(P)=\beta^T X_i \quad \text{et} \quad logit(\pi)=\gamma^T\omega$$

où $\omega \in \mathbb{R}^q$ et $X \in \mathbb{R}^p$ sont les vecteurs des covariables n est le nombre d'individus, p et q sont respectivement le nombre de covariables dans la partie inflation de zéro, $\gamma \in \mathbb{R}^q$ et $\beta \in \mathbb{R}^p$ sont les paramètre de régression. La log-vraisemblance du modèle basé sur les observations (Y_i, X_i, Φ_i), $i = 1, \cdots, n$ est donnée par:

$$l_n(\theta) = \sum_{i=1}^{n} \left\{ J_i \log\left(e^{\gamma^T \omega} + (1 + e^{\beta^T X_i})\right)^{-m_i} - \log(1 + e^{\gamma^T \omega}) + (1 - J_i)\left[Y_i \beta^T X_i - m_i \log(1 + e^{\beta^T X_i})\right] \right\}$$

où $J_i = 1_{\{Y_i = 0\}}$.
Les estimations des paramètres de γ et β peuvent être déterminées via la méthode du maximum de vraisemblance ou via l'algorithme EM comme décrit dans le modèle ZIP précédemment.

2.9 Modèle de régression de Hurdle

Les modèles de Hurdle (Mullahy 1986, [24]) peuvent être considérés comme un modèle de mélange à deux composantes comprenant une masse nulle et la composante des observations positives suivant une distribution de comptage tronquée, telle que la distribution de Poisson tronquée ou la distribution Binomiale Négative tronquée.
Soit Y_i la réponse de la ième observation, $i = 1, \cdots, n$ ou n représente le nombre total d'observations. La structure générale d'un modèle Hurdle est donnée par:

$$\mathbb{P}(Y_i = y_i) = \begin{cases} \pi_i & y_i = 0 \\ (1 - \pi_i) \frac{\mathbb{P}(y_i, \mu_i)}{1 - \mathbb{P}(y_i = 0, \mu_i)} & y_i > 0 \end{cases}$$

où π_i est la probabilité qu'un sujet appartienne à la composante zéro. $\mathbb{P}(y_i, \mu_i)$ représente une fonction de masse de probabilité (PMF) pour une distribution de comptage régulière avec un vecteur de paramètre μ_i et $1 - \mathbb{P}(y_i = 0, \mu_i)$ est la distribution évaluée à zéro. On constate par une distribution de comptage positif est régi une distribution de comptage régulière évaluée à zéro.
Par exemple, si la distribution des comptes suit une distribution de Poisson, la distribution de probabilité pour le modèle HP (Hurdle Poisson) à seuil s'écrit comme suit:

$$\mathbb{P}(Y_i = y_i) = \begin{cases} \pi_i & y_i = 0 \\ (1 - \pi_i) \frac{e^{-\mu_i} \mu_i^{y_i} / y_i!}{1 - e^{-\mu_i}} & y_i > 0 \end{cases}$$

Alternativement, la composante de comptage non nulle peut suivre d'autres distributions pour prendre en compte la surdispersion. Dans ce cas le modèle HBN (Hurdle Binomial Negative) est plus couramment utilisé. Le modèle est alors donné par:

$$\mathbb{P}(Y_i = y_i) = \begin{cases} \pi_i & y_i = 0 \\ \frac{1-\pi_i}{1-\left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha}} \frac{\Gamma(y_i+1/\alpha)}{\Gamma(1/\alpha)y_i!} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha} & y_i > 0 \end{cases} \tag{2.9}$$

Comme pour un modèle ZI, les covariables peuvent entrer dans la probabilité d'un zéro π_i et la fonction moyenne μ_i pour un modèle Hurdle. Par conséquent, le modèle Hurdle peut être écrit comme suit:

$$\log(\mu_i) = \alpha^\intercal X_i \quad , \quad logit(\pi_i) = \beta^\intercal Z_i$$

où α et β sont les coefficients de régression pour les covariables X_i et Z_i respectivement. Pour plus de détails sur le modèle Hurdle (estimation et propriété asymptotique) voir Mullahy (1986, [24]).

Les données de comptage sont des données qu'on rencontre dans plusieurs domaines de la vie. Il existe des modèles tels que le modèle de Poisson ou encore binomiale pour les traiter. Le constat est que; ces modèles sont vites dépassés lorsqu'on rencontre un nombre excessifs de zéros dans les données (inflation de zéros). Il faut donc se tourner vers les modèles plus robustes qui prennent en compte l'excès de zéros telque le modèle de Hurdle pour ne pas avoir des estimations biaisées.

3

Étude de simulation du modèle de Hurdle pour les données de comptage surdispersées

Résumé

Dans ce chapitre, nous nous sommes focalisés essentiellement sur le modèle de Hurdle et ses sous modèles tels que le modèle de Hurdle Poisson (HP), de Hurdle Bell (HB) et le modèle de Hurdle Poisson Log-Normale (HPLN). Nous avons fait une étude numérique sur les modèles HP et HB pour évaluer la perfomance (biais, erreur quadratique moyenne, erreur standard, déviation standard et la longueur de l'intervalle de confiance) de l'estimateur du maximum de vraisemblance. Pour le modèle HPLN nous n'avons pas fait de simulations mais nous nous sommes focalisés sur la partie théorique du modèle

Sommaire

3.1 Introduction

Le modèle de Hurdle est une classe de modèle statistique où une variable aléatoire est modélisés à l'aide de deux parties, la première est la probabilité d'avoir des zéros et la seconde modélise la probabilité des valaurs non nulles. L'utilisation du modèle est motivée par un excès de zéros dans les données de comptage, qui n'est pas suffisamment pris en compte dans les modèles statistiques plus standards.

3.2 Modèle de Hurdle

Le modèle de régression de Hurdle, est un modèle éfficace pour traiter les données de comptage à inflation de zéros. Ce modèle a deux processus de génération des données. Un processus pour générer les valeurs nulles et un autre pour générer les comptes positifs. Le premier processus évalue si les valeurs nulles se produisent. Il est noté 0 lorsque l'évènement (valeurs nulles) se produit avec une probabilité π et est noté 1 lorsque l' évènement ne se produit pas avec une probabilité de $1-\pi$. Lorsque l'évènement étudié se réalise, nous rentrons dans le deuxième processus, c'est-à-dire combien de fois l'évènement se produit. Dans ce processus l'occurence de l'évènement se conforme par exemple à une distribution de Poisson, binomiale négative, Bell ou Poisson log-normale. Toutefois dans ce processus, la valeur de l'évènement doit prendre une valeur positive et l'évènement doit se produire au moins une fois, ce qui est basé sur le premier processus. Cela nous conduit au modèle de Hurdle définit par:

$$\mathbb{P}(Y=y)=\begin{cases} g(0)=\pi & \text{si} \quad y=0 \\ (1-g(0))\frac{h(y)}{1-h(0)} & \text{si} \quad y>0 \end{cases} \tag{3.1}$$

où $g(y)$ et $h(y)$ sont les fonctions de densité du premier processus et du deuxième processus respectivement. $g(0)=\pi$ est la probabilité que zéro se produise et $\frac{h(y)}{1-h(0)}$ est la densité tronquée.
L'espérance du modèle de Hurdle est donnée par:

$$\mathbb{E}(Y)=\mathbb{P}(Y>0)\times\mathbb{E}_{Y>0}(Y|Y>0)=\frac{1-\pi}{1-h(0)}\times\nu \tag{3.2}$$

où ν est la moyenne de la densité non tronquée $h(y)$.
Le moment d'ordre deux est donnée par

$$\mathbb{E}(Y^2) = \frac{1-\pi}{1-h(0)} \times \sigma^2$$

où σ^2 est la variance de la densité non tronquée $h(y)$. Ainsi, nous pouvons définir la variance comme suit:

$$\mathbb{V}(Y) = \frac{1-\pi}{1-h(0)}\sigma^2 - \big(\frac{1-\pi}{1-h(0)}\nu\big)^2 \tag{3.3}$$

(Voir Zuo et al (2021, [36])).
La vraisemblance est définie comme suit:

$$\ell^{HM} = \prod_{i=1}^{n} (\pi_i)^{J_i} \left[(1-\pi_i)\frac{h(y_i)}{1-h(0)}\right]^{1-J_i}$$

où $J_i = 1_{\{y_i=0\}}$
La log-vraisemblance du modèle (3.1) est donnée par :

$$\ell^{HM} = \sum_{i=1}^{n} \{J_i \log(\pi_i) + (1-J_i)\left[\log h(y_i) - \log(1-h(0))\right]\} \tag{3.4}$$

3.3 Modèle de Hurdle Poisson

Dans la section (3.2) nous avons définie le principe de base du modèle de Hurdle. D'après ce principe si le deuxième processus suit une distribution de Poisson, alors on peut définir le modèle de régression Hurdle Poisson comme suit:

$$\mathbb{P}(Y_i = y_i) = \begin{cases} \pi_i & y_i = 0 \\ (1-\pi_i)\frac{e^{-\lambda_i}\lambda^{y_i}/y_i!}{1-e^{-\lambda_i}} & y_i > 0 \end{cases}$$

En utilisant les équations (3.2) et (3.3) la moyenne et la variance sont données par:

$$\mathbb{E}(Y) = \frac{\lambda(1-\pi)}{1-e^{-\lambda}}$$

$$\mathbb{V}(Y) = \frac{\lambda(1-\pi)}{1-e^{-\lambda}} - \lambda^2\big(\frac{1-\pi}{1-e^{-\lambda}}\big)^2$$

Comme pour les modèles ZI définient dans le chapitre 2, les paramètres λ_i et π_i sont modélisés respectivement par un log-linéaire et un logit. On a:

$$\log(\lambda_i) = \beta^\intercal X_i \quad \text{et} \quad logit(\pi_i) = \gamma^\intercal W_i \tag{3.5}$$

Où $\beta = (\beta_1, \cdots, \beta_p)^\intercal$ et $\gamma = (\gamma_1, \cdots, \gamma_q)^\intercal$ sont des vecteurs de paramètres inconnus, $\beta \in \mathbb{R}^p$ et $\gamma \in \mathbb{R}^q$, avec $p+q < n$ et $X_i = (1, X_{i1}, X_{i2}, \cdots, X_{ip})$, $W_i = (1, W_{i1}, W_{i2}, \cdots, W_{iq})$ sont des vecteurs de covariables.
D'après (3.5) on a:

$$\begin{aligned} logit(\pi_i) &= \gamma^\intercal W_i = \gamma_1 + \gamma_2 W_{i2} + \cdots + \gamma_q W_{iq} && (3.6) \\ \Rightarrow \pi_i &= \frac{\exp(\gamma^\intercal W_i}{1 + \exp(\gamma^\intercal W_i)} && (3.7) \end{aligned}$$

et

$$\begin{aligned} \ln(\lambda_i) &= \beta^\intercal X_i = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} && (3.8) \\ \Rightarrow \lambda_i &= \exp(\beta^\intercal X_i) && (3.9) \end{aligned}$$

D'après les équations (3.4), (3.5), (3.9) et (3.7) la log-vraisemblance du modèle de Hurdle Poisson est donnée par:

$$\ell_{(\theta)}^{HPM} = \sum_{i=1}^{n} J_i(-\log(1 + e^{\gamma^\intercal W_i}) + \gamma^\intercal W_i) +$$

$$\sum_{i=1}^{n}(1 - J_i)\left[-\log(1 + e^{\gamma^\intercal W_i}) - e^{\beta^\intercal X_i} + y_i \beta^\intercal X_i - \log(y_i!) - \log(1 - e^{-e^{\beta^\intercal X_i}})\right]$$

$$\ell_{(\theta)}^{HPM} = \sum_{i=1}^{n}\left\{J_i \gamma^\intercal W_i + (1 - J_i)\left[y_i \beta^\intercal X_i - e^{\beta^\intercal X_i} - \log(1 - e^{-e^{\beta^\intercal X_i}}) - \log(y_i!)\right] - \log(1 + e^{\gamma^\intercal W_i})\right\}$$

3.3.1 Étude de simulations du modèle HP

Dans cette partie nous parlons du processus de génération des données, de la proportion des zéros dans l'ensemble des données et du nombre d'échentillon utilisés. Nous évaluons la performance de l'EMV par le biais d'expérience de Monte Carlo.

3.3.1.1 Expérience numérique par simulation

Les données sont simulées à partir du modèle de régression de Hurdle Poisson tel que:

$$log(\lambda_i) = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5}$$

et

$$logit(\pi_i) = \gamma_1 + \gamma_2 W_{i2} + \gamma_3 Wi3 + \gamma_4 W_{i4} + \gamma_5 W_{i5}$$

Nous avons pris $\beta = (0.5, 0.9, 0.2, -0.25, -0.6)^\intercal$. Nous avons considéré trois valeurs $\gamma = (-0.5, 0.01, 0.15, 0, 0)^\intercal$, $\gamma = (0.7, -0.1, 0, 0.01, 0)^\intercal$ et $\gamma = (-0.45, 0.7, 0.8, -0.4, 0)^\intercal$ avec ces valeurs, la proportion moyenne c de données à inflation de zéro dans les ensembles de données simulées est respectivement de 42%, 60% et 90% de zéros. Les covariables $X_{i2}, ..., X_{i5}$ et $W_{i2}, ..., W_{i5}$ sont générées de la manière suivante: $X_{i2} \sim \mathcal{N}(0, 1)$, $X_{i3} \sim \mathcal{B}(1, 0.4)$, $X_{i4} \sim \mathcal{N}(1, 1.15)$, $X_{i5} \sim \mathcal{U}[-2, 2]$ et $W_{i2} \sim \mathcal{U}[2, 5]$, $W_{i3} \sim \mathcal{E}(1)$, $W_{i4} \sim \mathcal{N}(-1, 1)$, $W_{i5} \sim \mathcal{U}[2, 5]$. On considère les tailles d'échantillons $n = 500, 1000$ et 1500. Nous simulons $N = 1000$ réplications et nous calculons l'estimateur $\hat{\theta}$. Les simulations sont éffectuées á l'aide du logiciel R. Nous utilisons le package maxLik pour résoudre l'équation du score via l'algorithme de Newton-Raphson.

3.3.1.2 Résultats

Pour chaque scénario de simulation et chaque estimateur $\hat{\beta}$ et $\hat{\gamma}$, nous calculons le biais moyen, l'écart-type, l'erreur standard moyenne et l'erreur quadratique moyenne de l'estimation sur les n échantillons simulés. Nous obtenons également la longueur moyenne des intervalles de confiance de wald à 95% pour les estimateurs $\hat{\beta}$ et $\hat{\gamma}$. Les tableaux ci-dessous présentent respectivement les résulats pour $n = 500, 1000$ et 1500

L'erreur standard (SE) et la déviation Standard (SD) sont des mésures de variabilité. Le SD décrit la variabilité au sein d'un même échantillon et le SE estime la variabilté entre plusieurs échantillon d'une population. Les deux mesures ont des valeurs proches. Le biais est la mesure de distance entre l'estimateur et la valeur à estimer, si le modèle est bien spécifié le biais est faible. Le RMSE est la racine carrée des moyennes des prévisions et les valeurs observées. Ce sont des indicateurs de performance d'un modèle statistique. D'après les résultats, le biais, les deux mésures de variabilités et la longueur des intervalles de confiance diminuent à mesure que la taille de l'échantillon augmente. Pour une taille d'échantillon fixé, on observe que les performances de l'estimateur $\hat{\beta}$ restent stables lorsque la proportion d'inflation de zéros atteint des valeurs faibles à modérées (pour notre cas de 42% à 60%) et se détériorent lorsque l'inflation de zéros atteint des valeurs élévées. En ce qui concerne l'estimateur $\hat{\gamma}$, les performances s'améliorent et se détériorent lorsque la proportion d'inflation de zéros augmente. Les observations nous montrent que pour avoir une estimation précise dans un modèle à inflation de zéros il est nécessaire d'avoir dans les données une quantité suffisante d'observations nulles et non nulles pour estimer avec précision des probabilités d'inflation de zéros et le sous modèle de comptage.

n		$\hat{\beta}_n$					$\hat{\gamma}_n$				
		$\hat{\beta}_{1,n}$	$\hat{\beta}_{2,n}$	$\hat{\beta}_{3,n}$	$\hat{\beta}_{4,n}$	$\hat{\beta}_{5,n}$	$\hat{\gamma}_{1,n}$	$\hat{\gamma}_{2,n}$	$\hat{\gamma}_{3,n}$	$\hat{\gamma}_{4,n}$	$\hat{\gamma}_{5,n}$
500											
	bias	-0.0114	0.0019	0.0001	-0.0008	-0.0042	0.7988	-0.0044	-0.0366	-0.01108	0.0021
	SD	0.0847	0.0450	0.0753	0.0359	0.0449	0.4623	0.1069	0.1001	0.2334	0.0920
	SE	0.0853	0.0456	0.0805	0.0358	0.0435	0.4472	0.1068	0.0968	0.2320	0.0928
	RMSE	0.1207	0.641	0.1102	0.0507	0.0634	1.0255	0.1512	0.1439	0.3292	0.1307
	ℓ(CI)	0.3334	0.1769	0.3149	0.1394	0.1699	1.7524	0.4187	0.3786	0.9087	0.3637
1000											
	bias	-0.0030	0.0008	-0.0006	-0.0009	-0.0003	0.7801	-0.0033	-0.0354	0.0006	-0.0024
	SD	0.0589	0.0304	0.0552	0.0253	0.0313	0.3143	0.0758	0.0653	0.1574	0.0649
	SE	0.0593	0.0313	0.0559	0.0247	0.0301	0.3146	0.0.0751	0.0677	0.1631	0.0652
	RMSE	0.0836	0.0436	0.0785	0.0353	0.0434	0.8979	0.1067	0.1005	0.2266	0.0920
	ℓ(CI)	0.2320	0.1218	0.2188	0.0964	0.1179	1.2328	0.2943	0.2652	0.6390	0.2554
1500											
	bias	-0.0023	0.0006	-0.0029	0.0010	-0.0004	0.7889	-0.0050	-0.0396	0.0009	-0.0001
	SD	0.0487	0.0258	0.0457	0.0199	0.0238	0.2576	0.0613	0.0567	0.1334	0.0525
	SE	0.0479	0.0249	0.0452	0.0199	0.0244	0.2561	0.0612	0.0551	0.1326	0.0531
	RMSE	0.0383	0.0359	0.0643	0.0281	0.0341	0.8625	0.0868	0.0884	0.1880	0.0746
	ℓ(CI)	0.1874	0.0972	0.1772	0.0778	0.0955	1.0036	0.2399	0.2158	0.5196	0.2081

Tableau 3.1 – Résultats de la simulation pour $n = 500, 1000$ et 1500 avec une proportion moyenne de 42% de zéros pour $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^\intercal = (0.5, 0.9, 0.2, -0.25, -0.6)^\intercal$ et $\gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5)^\intercal = (-0.5, 0.01, 0.15, 0, 0)^\intercal$. SD : écart-type empirique. SE : erreur type moyenne. RMSE : erreur quadratique moyenne. ℓ(CI) : longueur moyenne des intervalles de confiance.

n		$\hat{\beta}_n$					$\hat{\gamma}_n$				
		$\hat{\beta}_{1,n}$	$\hat{\beta}_{2,n}$	$\hat{\beta}_{3,n}$	$\hat{\beta}_{4,n}$	$\hat{\beta}_{5,n}$	$\hat{\gamma}_{1,n}$	$\hat{\gamma}_{2,n}$	$\hat{\gamma}_{3,n}$	$\hat{\gamma}_{4,n}$	$\hat{\gamma}_{5,n}$
500											
	bias	-0.0118	0.0029	0.0020	-0.0019	-0.0025	0.5254	0.0157	0.0077	-0.0001	0.0027
	SD	0.1045	0.0554	0.0964	0.0421	0.0526	0.4907	0.1203	0.1064	0.2483	0.1042
	SE	0.1021	0.0553	0.0969	0.0433	0.0522	0.4899	0.1162	0.1026	0.2523	0.1009
	RMSE	0.1465	0.0784	0.1367	0.0604	0.0741	0.8698	0.1680	0.1480	0.3539	0.1450
	ℓ(CI)	0.3988	0.2141	0.3787	0.1682	0.2036	1.9191	0.4554	0.4009	0.9878	0.3951
1000											
	bias	-0.0062	0.0009	-0.0019	-0.0028	0.5330	0.5330	0.0213	0.0011	-0.0109	-0.0003
	SD	0.0703	0.0375	0.0647	0.0291	0.0339	0.3452	0.0821	0.0709	0.1745	0.0684
	SE	0.0687	0.0363	0.0649	0.0287	0.0350	0.3379	0.0803	0.0701	0.1743	0.0696
	RMSE	0.0984	0.0522	0.0916	0.0409	0.0488	0.7193	0.1167	0.0997	0.2468	0.0976
	ℓ(CI)	0.2688	0.1411	0.2540	0.1119	0.1369	1.3243	0.3147	0.2744	0.6828	0.2727
1500											
	bias	-0.0027	-0.0004	-0.0002	-0.0008	-0.0005	0.5091	0.0192	0.0044	-0.0092	-0.0018
	SD	0.0571	0.0294	0.0560	0.0241	0.0284	0.2969	0.0700	0.0598	0.1475	0.0564
	SE	0.0571	0.0298	0.0538	0.0238	0.0290	0.2801	0.0666	0.0580	0.1441	0.0576
	RMSE	0.0806	0.0418	0.0776	0.0339	0.0406	0.6525	0.0984	0.0834	0.2064	0.0806
	ℓ(CI)	0.2233	0.1160	0.2107	0.0929	0.1136	1.0975	0.2609	0.2272	0.5648	0.2257

Tableau 3.2 – Résultats de la simulation pour $n = 500, 1000$ et 1500 avec une proportion moyenne de 60% de zéros pour $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^\intercal = (0.5, 0.9, 0.2, -0.25, -0.6)^\intercal$ et $\gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5)^\intercal = (0.7, -0.1, 0, 0.1, 0)^\intercal$. SD : écart-type empirique. SE : erreur type moyenne. RMSE : erreur quadratique moyenne. ℓ(CI) : longueur moyenne des intervalles de confiance.

n		$\hat{\beta}_n$					$\hat{\gamma}_n$				
		$\hat{\beta}_{1,n}$	$\hat{\beta}_{2,n}$	$\hat{\beta}_{3,n}$	$\hat{\beta}_{4,n}$	$\hat{\beta}_{5,n}$	$\hat{\gamma}_{1,n}$	$\hat{\gamma}_{2,n}$	$\hat{\gamma}_{3,n}$	$\hat{\gamma}_{4,n}$	$\hat{\gamma}_{5,n}$
500											
	bias	-0.0573	0.0220	0.0006	-0.0056	-0.0113	0.5895	-0.0340	0.0317	-0.0521	-0.0095
	SD	0.2483	0.1542	0.2317	0.1131	0.1296	0.8991	0.2280	0.3217	0.5498	0.1805
	SE	0.2376	0.1494	0.2332	0.1090	0.1258	0.9104	0.2296	0.3208	0.5416	0.1830
	RMSE	0.3483	0.2155	0.3287	0.1571	0.1809	1.3912	0.3236	0.4553	0.7733	0.2572
	ℓ(CI)	0.9107	0.5551	0.8915	0.4100	0.4795	3.5444	0.8953	1.2333	2.0712	0.7149
1000											
	bias	-0.0175	-0.0004	-0.0055	-0.0012	-0.0056	0.5887	-0.0325	-0.0056	-0.0052	0.0081
	SD	0.1494	0.0823	0.1507	0.0661	0.0781	0.6261	0.1603	0.2208	0.3672	0.1314
	SE	0.1512	0.0860	0.1452	0.0649	0.0784	0.6233	0.1586	0.2180	0.3619	0.1272
	RMSE	0.2132	0.1189	0.2093	0.0926	0.1108	1.0614	0.2278	0.3102	0.5155	0.1830
	ℓ(CI)	0.5879	0.3294	0.5638	0.2505	0.3043	2.4361	0.6204	0.8467	1.4039	0.4977
1500											
	bias	-0.0146	0.0002	0.0020	-0.0002	-0.0066	0.5832	-0.0327	-0.0110	-0.0216	-0.0021
	SD	0.1268	0.0709	0.1150	0.0503	0.0625	0.5230	0.1287	0.1870	0.3055	0.1024
	SE	0.1196	0.0656	0.1137	0.0514	0.0617	0.5051	0.1288	0.1765	0.2930	0.1033
	RMSE	0.1748	0.0966	0.1617	0.0719	0.0881	0.9319	0.1850	0.2573	0.4237	0.1454
	ℓ(CI)	0.4660	0.2530	0.4434	0.1988	0.2404	1.9761	0.5042	0.6870	1.1404	0.4044

Tableau 3.3 – Résultats de la simulation pour $n = 500, 1000$ et 1500 avec une proportion moyenne de 90% de zéros pour $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^\intercal = (0.5, 0.9, 0.2, -0.25, -0.6)^\intercal$ et $\gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5)^\intercal = (-0.45, 0.7, 0.8, -0.4, 0)^\intercal$. SD : écart-type empirique. SE : erreur type moyenne. RMSE : erreur quadratique moyenne. ℓ(CI) : longueur moyenne des intervalles de confiance.

3.4 Modèle de Hurdle Bell

Le modèle de Hurdle suivant une distribution de Bell est peu étudié dans la littérature. Nous proposons un modèle de hurdle Bell pour modéliser les données de comptage à inflation de zéros. dans cette partie nous avons établi proprement le modèle HB, calculé l'espérance et la variance. Nous avons établi la log-vraisemblance du modèle et pour finir une étude numérique du modèle pour évaluer la performance de l'estimateur du maximum de vraisemblance a été proposée.

La variable discrète Y suit une distribution de Bell si la fonction de masse de probabilité est donnée par:

$$\mathbb{P}(Y=y)=\exp(1-e^{Z(\lambda)})\frac{Z(\lambda)^y B_y}{y!},\quad y=0,1,2,\cdots \tag{3.10}$$

où $\lambda>0$, $Z(.)$ est la fonction de Lambert et B_y est le nombre de Bell défini par:

$$B_y=\frac{1}{e}\sum_{d=1}^{\infty}\frac{d^y}{d!}$$

Le modèle de regression de Hurdle Bell se défini comme suit:

$$\mathbb{P}(Y_i=y_i)=\begin{cases}\pi_i & y_i=0\\ (1-\pi_i)\frac{\exp(1-e^{Z(\lambda)})Z(\lambda)^Y B_y}{Y!(1-\exp(1-e^{Z(\lambda)}))} & y_i>0\end{cases}$$

La variance et la moyenne d'une variable $Y\sim Bell(Z(\lambda))$ sont données par:

$$\mathbb{E}(Y)=Z(\lambda)e^{Z(\lambda)}\qquad\text{et}\qquad\mathbb{V}(Y)=Z(\lambda)(1+Z(\lambda))e^{Z(\lambda)} \tag{3.11}$$

(Voir Castellares et al (2017, [6])).

D'après les équations (3.2), (3.3) et (3.11) la moyenne et la variance du modèle Hurdle Bell sont données par:

$$\mathbb{E}(Y)=\frac{(1-\pi)Z(\lambda)e^{Z(\lambda)}}{1-\exp(1-e^{Z(\lambda)})} \tag{3.12}$$

$$\mathbb{V}(Y)=\frac{(1-\pi)Z(\lambda)(1+Z(\lambda))e^{Z(\lambda)}}{1-\exp(1-e^{Z(\lambda)})} \tag{3.13}$$

De la même façon que le modèle de Hurdle Poisson, le paramètre λ_i et la probabilité π_i sont modélisés respectivement par un log-linéaire et par un logit:

$$\log(\lambda_i)=\beta^\intercal X_i\quad\text{et}\quad logit(\pi_i)=\gamma^\intercal W_i$$

Où $\beta = (\beta_1, \cdots, \beta_p)^\intercal$ et $\gamma = (\gamma_1, \cdots, \gamma_q)^\intercal$ sont des vecteurs de paramètres inconnus, $\beta \in \mathbb{R}^p$ et $\gamma \in \mathbb{R}^q$, avec $p+q < n$ et $X_i = (1, X_{i1}, X_{i2}, \cdots, X_{ip})$, $W_i = (1, W_{i1}, W_{i2}, \cdots, W_{iq})$ sont des vecteurs de covariables.
D'après les équations (3.11), (3.4), (3.5), (3.9) et (3.7), la log-vraisemblance du modèle HB est définie comme suit:

$$\begin{aligned} \ell_\theta^{HBM} &= \sum_{i=1}^{n} J_i \gamma^\intercal W_i - \log(1 + e^{\gamma^\intercal W_i}) + (1 - J_i)(Y_i \log(Z(e^{\beta^\intercal X_i}))) \\ &- \exp(Z(e^{\beta^\intercal X_i})) - \log(1 - \exp(1 - \exp(Z(e^{\beta^\intercal X_i})))) \end{aligned}$$

3.4.1 Étude de simulations

Dans cette partie nous parlons du processus de génération des données, de la proportion des zéros dans l'ensemble des données et du nombre d'échntillon utilisés. Nous évaluons la performance de l'EMV par le biais d'expérience de Monte Carlo.

3.4.1.1 Expérience numérique par simulation

3.4.1.2 Expérience numérique

Les données sont simulées à partir du modèle de régression Hurdle Bell tel que:

$$log(\lambda_i) = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i4} + \beta_5 X_{i5}$$

et

$$logit(\pi_i) = \gamma_1 + \gamma_2 W_{i2} + \gamma_3 Wi3 + \gamma_4 W_{i4} + \gamma_5 W_{i5}$$

Nous avons pris $\beta = (0.5, 0.9, 0.2, -0.25, -0.6)^\intercal$. Nous avons considéré une valeur de $\gamma = (1.72, 0.7, -0.2, 0, 0)^\intercal$ avec une proportion de 50% de zéros. Les covariables $X_{i2}, ..., X_{i5}$ et $W_{i3}, ..., W_{i5}$ sont générées de la manière suivante: $X_{i2} \sim \mathcal{N}(0, 1)$, $X_{i3} \sim \mathcal{B}(1, 0.5)$, $X_{i4} \sim \mathcal{N}(1, 1.15)$, $X_{i5} \sim \mathcal{U}[-2, 2]$ et $W_{i3} \sim \mathcal{E}(1)$, $W_{i4} \sim \mathcal{B}(0.8)$, $W_{i5} \sim \mathcal{N}(-1, 1)$. Les prédicteurs linéaires $log(\lambda_i(\beta))$ et $logit(\pi_i(\gamma))$ peuvent utiliser les termes communs. On a $W_{i2} = X_{i5}$. On considère les tailles d'échantillons $n = 500$ et 750. Nous simulons $N = 1000$ réplications et nous calculons l'estimateur $\hat{\theta}$. Les simulations sont éffectuées á l'aide du logiciel R. Nous utilisons le package maxLik pour résoudre l'équation du score via l'algorithme de Newton-Raphson.

3.4.1.3 Résultats

Pour chaque scénario de simulation et chaque estimateur $\hat{\beta}$ et $\hat{\gamma}$, nous calculons le biais moyen, l'écart-type, l'erreur standard moyenne et l'erreur quadratique

moyenne de l'estimation sur les n échantillons simulés. Nous obtenons également la longueur moyenne des intervalles de confiance de wald à 95% pour les estimateurs $\hat{\beta}$ et $\hat{\gamma}$. Les tableaux ci-dessous présentent respectivement les résulats pour $n = 500$ et 750

Tout comme le modèle Hurdle de Poisson, le biais, les deux mésures de variabilités et la longueur des intervalles de confiance diminuent à mesure que la taille de l'échantillon augmente. Pour une taille d'échantillon fixée, on observe que les performances de l'estimateur $\hat{\beta}$ sont bonnes pour une proportion de 50% de l'inflation de zéros. En ce qui concerne l'estimateur $\hat{\gamma}$, les performances sont bonnes également. Les observations nous montrent que pour avoir une estimation précise dans un modèle à inflation de zéro il est nécessaire d'avoir dans les données une quantité suffisante d'observations nulles et non nulles pour estimer avec précision des probabilités d'inflation de zéros et le sous modèle de comptage.

n		$\hat{\beta}_n$					$\hat{\gamma}_n$				
		$\hat{\beta}_{1,n}$	$\hat{\beta}_{2,n}$	$\hat{\beta}_{3,n}$	$\hat{\beta}_{4,n}$	$\hat{\beta}_{5,n}$	$\hat{\gamma}_{1,n}$	$\hat{\gamma}_{2,n}$	$\hat{\gamma}_{3,n}$	$\hat{\gamma}_{4,n}$	$\hat{\gamma}_{5,n}$
500											
	bias	-0.0069	0.0021	0.0001	-0.0002	-0.0009	0.5158	0.1168	-0.0184	0.0111	0.0035
	SD	0.0888	0.0293	0.0572	0.0241	0.0170	0.3522	0.0786	0.1120	0.2731	0.1141
	SE	0.0877	0.0290	0.0557	0.0245	0.0165	0.3509	0.0791	0.0.1128	0.2786	0.1112
	RMSE	0.1250	0.0412	0.0798	0.0343	0.0237	0.7162	0.1615	0.1600	0.3902	0.1593
	ℓ(CI)	0.3425	0.1127	0.2177	0.0954	0.0642	1.3727	0.3094	0.4411	1.0908	1.4353
750											
	bias	-0.0001	-0.0009	-0.0003	-0.0010	-0.0002	0.5030	0.1124	0.0220	-0.0010	0.0008
	SD	0.0692	0.0231	0.0446	0.0193	0.0128	0.2857	0.0645	0.0910	0.2282	0.0907
	SE	0.0704	0.0230	0.0447	0.0195	0.0131	0.2839	0.0641	0.0907	0.2259	0.0902
	RMSE	0.0987	0.0326	0.0631	0.0275	0.0181	0.6444	0.1446	0.1303	0.3210	0.1279
	ℓ(CI)	0.2752	0.0896	0.1749	0.0761	0.0751	1.1110	0.2509	0.3551	0.8848	0.3534

Tableau 3.4 – Résultats de la simulation pour $n = 500$ et 750 avec une proportion moyenne de 50% de zéros pour $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^\intercal = (0.5, 0.9, 0.2, -0.25, -0.6)^\intercal$ et $\gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5)^\intercal = (1.72, 0.7, -0.2, 0, 0)^\intercal$. SD : écart-type empirique. SE : erreur type moyenne. RMSE : erreur quadratique moyenne. ℓ(CI) : longueur moyenne des intervalles de confiance.

3.5 Modèle de Hurdle Poisson log-normale

Le modèle de Hurdle Poisson log-normale est rarement utilisé pour traiter les données. Nous présentons dans cette partie le modèle de Hurdle Poisson log-normale, nous calculons l'espérance et la variance du modèle enfin nous établissons la log-vraisemblance du modèle. Une étude numérique pour évaluer l'estimateur du maximum de vraisemblance ne sera pas réalisée.

La densité de probabilité de la loi Poisson log-normale est donnée par:

$$\mathbb{P}(Y_i = y_i) = \int_0^\infty \frac{\lambda^{y_i} \exp^{-\lambda}}{y_i!} \left\{ \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln\lambda - \mu)^2}{2\sigma^2} \right) \right\} d\lambda$$ voir Farzaneh et al (2017, [16])

L'espérance et la variance sont données par:

$$\mathbb{E}(Y_i) = \mu \exp(0.5\sigma^2) \tag{3.14}$$

$$\mathbb{V}(Y_i) = \mathbb{E}(Y_i) + \mathbb{E}(Y_i)^2[\exp(\sigma^2) - 1] = \psi \tag{3.15}$$

Le modèle de régression de Hurdle Poisson log-normale se définit comme suit:

$$\mathbb{P}(Y = y) = \begin{cases} \pi & \text{si} \quad y = 0 \\ (1-\pi)\frac{h(y)}{1-h(0)} & \text{si} \quad y > 0 \end{cases} \tag{3.16}$$

où

$$h(y) = \int_0^\infty \frac{\lambda^{y_i} \exp^{-\lambda}}{y_i!} \left\{ \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln\lambda - \mu)^2}{2\sigma^2} \right) \right\} d\lambda$$

et

$$h(0) = \int_0^\infty e^{-\lambda} \left\{ \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln\lambda - \mu)^2}{2\sigma^2} \right) \right\} d\lambda$$

D'après les équations (3.2), (3.3), (3.15) et (3.15), la moyenne et la variance du modèle de Hurdle Poisson log-normale sont données par:

$$\mathbb{E}(Y) = \mu \exp(0.5\sigma^2) \frac{1-\pi}{1-h(0)} \quad \text{et} \quad \mathbb{V}(Y) = \frac{1-\pi}{1-h(0)} \psi^2 - (\mathbb{E}(Y))^2$$

Le modèle de Hurdle Poisson log-normale est clairement défini comme suit:

$$\begin{cases} logit(p_i) = \gamma^\intercal W_i \\ \log\lambda_i = \beta^\intercal X_i + \mu_i = \nu_i \\ \exp(\mu_i) \sim lognormal(0, \sigma^2) \end{cases}$$

On note $\theta = (\beta^\intercal, \gamma^\intercal, \sigma)^\intercal$ le paramètre à estimer. La log-vraisemblance du modèle HPLN est définie par:

$$\begin{aligned} \ell_{(\theta)}^{HPLN} = \sum_{i=1}^{n} \Bigg\{ & J_i\big(\gamma^\intercal W_i - \log\left(1 + e^{\gamma^\intercal W_i}\right)\big) \\ & + \big(1 - J_i\big) \left[-\log\left(1 + e^{\gamma^\intercal W_i}\right) + \log \int_0^\infty \frac{e^{y_i \nu_i}}{y!} h_i(\beta, \sigma) de^{\nu_i} - \log\left(1 - \int_0^\infty h_i(\beta, \sigma) de^{\nu_i}\right) \right] \Bigg\} \end{aligned}$$

Avec $h_i(\beta, \sigma) = \frac{e^{e^{\nu_i}}}{e^{\nu_i}\sigma\sqrt{2\pi}} \exp\left[-\frac{(\beta^\intercal X_i)^2}{2\sigma^2} \right]$

Remarque 3.1 *Le modèle HPLN est un modèle mixte dont les paramètres sont plus complexes à estimer que les modèles HP ou HB d'autant plus que la fonction de vraisemblance n'est pas facile à manipuler. Néanmoins il peut être définit comme nous l'avons fait dans la section (3.5). Il est alors difficile de calculer l'estimateur par la méthode du maximum de vraisemblance. Toutefois il existe des méthodes numériques pour pouvoir approcher l'intégrale et faciliter l'estimation.*

Conclusion et perspectives

Ce mémoire porte sur le modèle Hurdle et ses sous modèles.

Pour faciliter la lecture du document, nous avons rappelé dans le chapitre 1 quelques concepts fondamentaux sur les modèles linéaires généralisés.

Dans le chapitre 2, nous avons défini les données de comptage, nous avons étudié le phénomène de la surdispersion qui survient dans les données de comptage causé par l'inflation de zéros. Nous avons étudié également quelques modèles qui permettent de traiter les données de comptage à inflation de zéros à savoir: les modèles ZIP, ZINB, ZIB, HP, HNB.

Enfin, dans le chapitre 3, Nous avons clairement défini les modèles HP, HB et HPLN. Nous avons fait une étude numérique sur les modèles HP et HB pour évaluer la perfomance (biais, erreur quadratique moyenne, erreur standard, déviation standard et longueur moyenne de l'intervalle de confiance) de l'estimateur du maximum de vraisemblance. Pour le modèle HPLN nous n'avons pas fait de simulations mais nous nous sommes focalisé sur la partie théorique du modèle en établissant le modèle et en calculant la log-vraisemblance.

Après les études numériques réalisées sur le modèle HP et le modèle HB, nous pouvons aisément conclure que les deux modèles sont bien éfficaces pour traiter les données de comptage avec inflation de zéros.

Dans l'avenir, nous proposons de comparer les modèles HP et HB sur les données réelles afin de choisir le modèle le plus performant c'est-à-dire, le modèle qui estimera le mieux les paramètres. Nous proposons également de mener des études numériques sur le modèle HPLN pour évaluer la performance de l'estimateur du maximum de vraisemblance.

Bibliographie

[1] Akaike, H., 1973. Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*

[2] Antoniadis, A.B., Jacques, C., 1992. Régression non linéaire et applications. *Economica.*

[3] Bel, Jean, J., Etienne, Lebarbier, M., Mary-Huard, E., Robin, T., Vuillet, C. 2016. Le Modèle Linéaire et ses Extensions. 1-328.

[4] Buu, Anne and Li, Runze and Tan, Xianming and Zucker, Robert, A. Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. *Statistics in medicine*, 4074-4086.

[5] Carroll,R. J., Ruppert, D., and Stefanski,L. A. 1995. Measurement error in nonlinear models. *Chapman and Hall, New York.*

[6] Castellares F., Ferrari L.P., Lemonte A., 2017. On the distribution and its associated regression model for count data. *Applied mathematical modeling*

[7] Consul, P.C., 1992. Famoye,F. Generalized Poisson regression model.*Communications in Statistics Theory and Method*, 21,89-109.

[8] Czado, C., Erhardt, V., Min, A., Wagner, S. 2007. Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. *Statistical Modelling* 7(2), 125-153.

[9] Dean, C., Lawless, J.F. , 1989. Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association*, 84: 467-471.

[10] Diallo A., 2017. Inférence statistique dans des modèles de comptage à inflation de zéro. Applications en économie de la santé. *Rennes, INSA.*

[11] Dempster,A., Laird,N., and Rubin, D. 1977. Maximum likelihood from incomplete data via the em algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B*, 39 :1-38.

[12] Dobson,A., Bernet, J., and Adrian, G., 2018. An introduction to generalized linear models. *Chapman and Hall/CRC.*

[13] Feng, C.X., 2021. A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *Journal of statistical distributions and applications* 8, 1-19.

[14] Feng, C.X., 2012. Joint analysis of multivariate spatial count and zero-heavy count outcomes using common spatial factor models. *Environmetrics* 23, 493-508

[15] Fahrmeir, L. et Kaufmann, H., 1985. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Annals of Statistics,* 13 :342-368.

[16] Farzaneh N., Bahrampour A., Yunes Y., 2017. Comparing zero-inflated Poisson, Poisson gamma, and Poisson lognormal regression models in dental health data. *Journal of Biostatistics and Epidemiology*, 41-48.

[17] Hall, DB.,2000. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics.*,56(4),1030-1039.

[18] Hall, D.B., Shen, J.2010. Robust estimation for zero-inflated poisson regression. *Scand. J. Statist.*,37 :237-252

[19] Hilbe, J. M., 2011. Negative Binomial Regression. *2nd ed. Cambridge : Cambridge University Press.*

[20] Jennrich, R. and Sampson, P.F., 1976. Newton-Raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics.*

[21] Kemp, C.D., and Kemp,A.W., 1988. Rapid estimation for discrete distributions. *The Statistician,* 37 :243-255.

[22] Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*,34:1-14.

[23] McCullagh, P., Nelder, J. A., 1989. Generalized linear models (Second edition). *Monographs on Statistics and Applied Probability. Chapman & Hall, London.*

[24] Mullahy, J., 1986. Specification and testing of some modified count data models. *Journal of Econometrics*,33:341-365.

[25] Mullahy, J., 1997. Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics*,12(3) :337-350.

[26] Nelder, J.A. , Wedderburn, R.W.M., 1972. Generalized Linear Models. *Journal of the Royal Statistical Society, Series B*, 56(1) :61-69

[27] Neelon, B., O'Malley, A., Valerie A., 2016. Modeling zero-modified count and semicontinuous data in health services research part 1: background and overview. *Statistics in Medicine*, 70-93

[28] Osei F.B., Stein, A., Andreo, V., 2022. A zero-inflated mixture spatially varying coefficient modeling of cholera incidences. *Spatial statistics* 48.

[29] Perumean-Chaney and al., 2013. Zero-inflated and overdispersed: what's one to do? *Journal of Statistical Computation and Simulation* 83, 1671-1683.

[30] Rose, C.E., Martin, S.W., Wannemuehler, K.A., Plikaytis, B.D., 2006 On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of biopharmaceutical statistics* 16:463-481.

[31] Rosen, O., Jiang,W. X., and Tanner,M. A., 2000. *Mixtures of marginal models. Biometrika,* 87 :391-404.

[32] Schwarz G., 1978. Estimating thedimension of a model. Annals of Statistics, 6 (2), 461–464.

[33] Shiyuka, N. 2018. Hurdle negative binomial model for motor vehicle crash injuries in Namibia. *University of Namibia*

[34] Vieira, AMC., Hinde, JP., Demetrio, CGB. , 2000. Zero-inflated proportion data models applied to a biological control assay. *J Appl Stat.*,27(3):373?389.

[35] Winkelmann, R., and Zimmermann, K.F. 1995. Recent developments in count data modelling: theory and application. *Journal of economic surveys* 9 1-24

[36] Zuo, G., Fu, K., Dai, X. and Zhang, L. 2021. Generalized Poisson Hurdle model for count data and its application in ear disease. *Entropy*, 23(9), 1206.