UNIVERSITE
GASTON BERGER
L'excellence au service du développement

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

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

MÉMOIRE DE MASTER

pour l'obtention du diplôme de

Master en Mathématiques Appliquées

Soutenue publiquement le 03 Août 2023 par

Agnondji GNON SIYA

Mention : Mathématiques Appliquées

Spécialité : Sciences de Données et Applications

MODÈLES POISSON-GAMMA FLEXIBLE À INFLATION DE ZÉROS

Encadreur de Mémoire : Prof Aliou DIOP
Co-encadreur de Mémoire : Dr Essoham ALI

Composition du Jury

Aboubakary DIAKHABY	Professeur Titulaire , UGB de Saint-Louis, Sénégal	Président
Aliou DIOP	Professeur Titulaire, UGB de Saint-Louis, Sénégal	Encadreur de Mémoire
Abdou Kâ DIONGUE	Professeur Titulaire ,UGB de Saint-Louis, Sénégal	Examinateur
El Hadji DEME	Maître de Conférences, UGB de Saint-Louis, Sénégal	Rapporteur

Dédicaces

À la mémoire de mon père Feu Tchigrr K. GNON SIYA ,
À ma mère Affouwa BALEMA,
À toute la famille GNON SIYA.

Remerciements

Ce travail a été dirigé par le professeur Aliou DIOP de l'Université Gaston Berger de Saint-Louis . Cher professeur je tiens à vous exprimer ma profonde gratitude pour m'avoir permis de réaliser ce travail . Votre soutien tout au long de ce processus ont été inestimables et ont contribué de manière significative à la réussite de ce sujet de mémoire. Votre expertise dans le domaine a été d'une grande valeur pour moi. Vos conseils éclairés et vos recommandations pertinentes m'ont permis de mieux cerner mon sujet, d'approfondir mes recherches et d'élaborer une analyse approfondie. Votre disponibilité pour discuter des idées, répondre à mes questions et corriger mes erreurs a été extrêmement appréciée Je suis extrêmement reconnaissant d'avoir eu honneur de travailler avec vous et de bénéficier de votre expertise. Votre soutien indéfectible et votre mentorat ont été essentiels à ma réussite académique.

Je ne saurais trouver les mots adéquats pour remercier M. Essoham ALI,Ph.D .Laboratoire de Mathématiques de Bretagne Atlantique (LMBA) Université de Bretagne-Sud en France. M. Essoham ALI votre contribution en tant que co-directeur a grandement enrichi mon travail. Vos idées, suggestions et commentaires éclairés ont apporté une perspective supplémentaire à mon étude et ont contribué à son amélioration. Votre présence et votre implication active lors des réunions ont été extrêmement appréciées, et j'ai pu bénéficier de vos connaissances approfondies dans le domaine.votre patience et votre soutien inconditionnel ont joué un rôle essentiel dans la réalisation de mon mémoire.Je vous dois énormement. Merci pour TOUT. Je tiens également à remercier sincèrement tout le corps professoral et le personnel

du département UFR-SAT pour leur précieux enseignement et leur soutien tout au long de mon parcours académique au sein de cet université.

Je dédie une mention spéciale et toute particulière à mes parents en particulier à ma très chère mère surtout pour ses encouragements, à ma famille et à mes amis

les plus proches. Un grand merci à mes adorables et aimables soeurs Nestorine, Ornéla, Nativité et Adéle , à mes chers fréres Razack, Ulrich, Donatien mon cher ami Dominique YABA. Votre confiance en mes capacités et votre croyance en ma réussite m'ont donné la force de persévérer même dans les moments les plus difficiles, je tiens à partager avec vous cette réussite qui est aussi la vôtre.

Il m'est bien sûr impossible de terminer ces remerciements sans avoir une pensée à tous mes camarades de promotion, chères amis merci pour la chaleur fraternelle qui a prévalu entre nous. Je vous souhaite une bonne continuation

Abréviations & Notations

$\mathbb{P}(A)$	: La probabilité de l'événement A.
$\mathbf{Var}(X)$	: La variance de la variable aléatoire X.
$\mathbb{E}(X\|Y)$	: Espérance conditionnelle de X sachant Y.
$X_n \xrightarrow{P} Y$	: La suite de variables aléatoires $(X_n)_n$ converge en probabilité vers 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.
Diag	: Diagonale d'une matrice
$\mathcal{P}(\lambda)$	: Loi de Poisson de paramètre λ.
$\mathcal{B}(\mathrm{p})$	: Loi Bernoulli de paramètre p.
$\mathrm{NB}(\mathrm{r}, \mathrm{p})$	: Loi binomiale négative de paramètres r et p.
EMV	: Estimateur du Maximum de Vraisemblance.
GLM	: Generalised Linear Model
ZIP	: Zero-Inflated Poisson
ZIB	: Zero-Inflated Binomial
ZINB	: Zero-Inflated Negative Binomial
ZIPLG	Zero-Inflated Poissson-log-gamma

Table des matières

Liste des tableaux

Résumé

Les modèles de comptage à inflation de zéros sont des techniques statistiques utilisées pour modéliser des données de comptage avec une proportion élevée de zéros. Ces modèles de comptages sont basés essentiellement sur des distributions de poisson, de binomiale négativ et de gamma. Ces modèles sont largement utilisés dans de nombreux domaines, tels que la biologie, l'économie, l'épidémiologie etc. Ils permettent de modéliser de manière plus précise les données en prenant compte à la fois la surabondance de zéros et la surdispersion.

L'objectif de ce mémoire est d'expliquer ces modèles, qui reposent encore sur des hypothèses mathématiques. Dans la suite de notre travail, nous introduirons un nouveau modèle de comptage basé sur la régresion poisson log-gamma qui est très peux discuter dans la littérature. Pour finir nous nous intéressons à une étude de simulation exhaustive sur des échantillons de tailles finies pour évaluer la cohérence de nos résultats.

Mots clés: Normalité asymptotique, consistance, données de comptage, excès de zéros, simulations, l'estimateur du maximum de vraisemblance, poisson log-gamma.

Abstract

Zero-inflated count models are statistical techniques used to model count data with a high proportion of zeros. These count models are primarily based on Poisson, negative binomial, and gamma distributions. They are widely used in various fields, such as biology, economics, epidemiology, etc. And allow for a more accurate modeling of data by accounting for both zero-inflation and overdispersion.

The objective of this dissertation is to explain these models, which still rely on mathematical assumptions. In the continuation of our work, we will introduce a novel count model based on Poisson log-gamma regression, which has received little discussion in the literature. Finally, we will conduct an extensive simulation study on finite-sized samples to evaluate the consistency of our results.

Keywords: Asymptotic normality, consistency, count data, excess zeros, simulations, maximum likelihood estimator, poisson log-gamma.

Contexte autour des Modèles de comptage

Une revue de la littérature

Les modèles de comptage font référence à une famille de modèles statistiques utilisés pour modéliser et prédire le nombre d'occurrences d'un évènement donné dans une période de temps ou dans une zone géographique donnée. La modélisation de ces données de comptage est un problème très répandue dans divers domaines comme l'épidémiologie, la banque, les assurances, l'économétrie, la médecine, la sécurité routière ou encore l'écologie. Par exemple dans le domaine de l'épidémiologie, les modèles de comptage peuvent être utilisés pour prévoir le nombre de cas de maladies infectieuses dans une région donnée. Dans le domaine de la finance, ils peuvent être utilisés pour prévoir le nombre de transactions boursières ou le nombre de sinistres dans une compagnie d'assurance.

Les modèles de comptage sont souvent utilisés pour modéliser des données qui ont une distribution de Poisson ou une distribution binomiale négative, qui sont des distributions de probabilité discrètes.

Ainsi, les méthodes de modélisation adaptées à ce type de données ont fait l' objet de nombreuses discutions dans plusieurs revues littéraires ces dernières années. Pendant longtemps, la régression de Poisson a été le dernier recours standard dans ce genre de situation, cependant son application à des cas réels ont mis à nu certaines insuffisances et limites ; par exemple un excès de zéros dans le comptage, la sur-dispersion des données. Dans la plupart des ensembles de données de comptage la variance est beauucoup plus grande que la moyenne, souvent beaucoup plus superieure que celle attendue sous le modèle de Poisson. Ce phénomène est connu sous le nom de sur-dispersion (1983, Cox [7]) ce qui peut entrainer une sous estimation de la variance et le rejet d'une hypothèse lorsqu'il ne le faut pas. Les causes sont multiples. Une inflation de zéros est une autre cause majeure de la sur-dispersion que l'on rencontre dans les données de comptage, d'où la nécessité de trouver des alternatives qui permettent de gérer les problèmes sur-dispersion et les excès de zéros induits par les mécanismes du phénomène étudie. Certains auteurs comme

(Mullahy1987, [26]) ,(Lambert1992, [20]) et Hall2000, [16] , ont posé les bases et ouvert la voie à de nombreux travaux méthodologiques, théoriques et appliqués , en proposant des modèles de régression à inflation basé uniquement sur le modèle Binomial et le modèle de Poisson. Parmi les alternatives existantes, nous pouvons citer les Zero-Inflated Poisson (ZIP), Zero-Inflated Poisson Gamma (ZIPG),Binomial Negative (NB), et Zero-Inflated Binomial Negative (ZINB) . Ces modèles répondent de manière spécifique aux problèmes des zéros en excès tout en gérant la sur-dispersion des données. Les travaux de recherche sur la généralisation de ces modèles , ainsi que leurs mises en application sont nombreux. Par exemple dans le domaine de la médecine, il existe en générale deux types d'approches pour traiter les types de zéros dans les données de comptage de séquences du Microbiome. Le NB a été utilisé par Stanikov pour évaluer les différences d' abondance de séquençage et pour détecter les caractéristiques différentielles dans léchantillon métagénomique. L' opportunité d'utiliser un modèle à excès de zéro dans l'étude du microbiome a été évaluer par des simulations approfondies. (Consul et Famoye 1992,[30]) proposent un modèle de règression de Poisson généŕalisée, avec l' introduction d' un nouveau paramètre dans le modèle standard pour gérer la sur-dispersion. La particularité du modèle ZIP est qu' il existe deux processus sous-jacent qui d' eterminent si un compte est nul ou non nul . Une fois qu'un compte est non nul le processus de Poisson régulier prend le relai pour déterminer sa valeur réelle. L'insuffisance de ce modèle ZIP est qu'il ne peut pas prendre en charge la sur-dispersion qui est souvent largement observée dans les données de comptage. Ainsi donc Ridout et Al proposent le modèle de régression Binomial Negative à Inflation de Zéros (ZINB)(Ridout2001,[29]). Ce modèle n'étant pas capable de prendre en compte tous les paramètres qui influent sur les variables d'intérêt très tôt va révéler certaines limites et pousser des chercheurs à l'améliorer. (Mwalili2008, [23]), ont examiner un modèle binomial négatif et illustrer comment la correction des erreurs de classification peut être obtenue. Remarquons que la distribution prédite à l'aide des modèle ZINB par les packages de R dans les données du microbiome(Zeileis2008,[33]) ne prend pas en compte la dispersion dans les données réelle. Pour remédier à ces limites ROULAN J, XIANG Z et TIANYING W (2022,[31]) proposent un modèle Poisson Gamma Zero Inflated (ZIPG) qui fournie une modélisation flexible qui relie à la fois l'abondance moyenne et la dispersion à différent ensemble. Dans presque tous les modèles prè citer, pour l'estimation des paramètres on a toujours fait recoure à la Méthode de l' Estimation du Maximum de Vaisemblence(EMV) pour estimer les paramètres du modèle car elle fournie des estimateurs convergent. Dans le cas des modèles à inflation de zéros . Lambert a montré que les estimations du maximum de vraisemblance (EMV) sont approximativement normales dans les grands échantillons et les intervalles de confiance peuvent être construits en inversant les tests du rapport de vraisem-

blance ou en utilisant la normalité approximative des EMV (1992, [20]), mais il est nécessaire de savoir que EMV est très sensible en présence des valeurs aberrantes tels que les excès de zéros. Hall et Shen(2010,[15]) ont suggéré une nouvelle procédure d'estimation dit algorithme Expectation-Maximization (EM) qui constitue une approche pratique pour le calcule de l'EMV. Pour adapter au modèle d'inflation. d'autres auteurs procèdent à la maximisation directe qui n'est pas commode dans toutes les situations, dans le cas ZIPG la maximisation directe peut poser un problème pour distinguer les zéros de la partie PG et l'autre partie d'inflation zéros du modèle ZIPG Par conséquent on utilise l'algorithme EM.

1

Méthodes d'analyse des données de comptage à excès de zéros

Résumé

Dans ce chapitre, nous ferons quelques rappels essentiels sur les modèles linéaires généraliser, nous passerons en révu quelque lois de famille exponentielle et nous discuterons de la notion de sur-dispersion. Puis, nous énoncerons quelques modèles à inflation de zéros, les méthodes d'estimations puis les propriétés asymptotiques.

Sommaire

Sommaire

1.1 Introduction

La méthode d'analyse des données de comptage à excès de zéros est une approche statistique utilisée pour analyser des ensembles de données où les valeurs observées sont principalement des zéros, avec un petit nombre d'observations non nulles. Cette méthode est couramment utilisée dans des domaines tels que la biologie, l'épidémiologie, l'écologie et la finance, où les données de comptage sont fréquentes.

L'analyse des données de comptage à excès de zéros vise à modéliser et à comprendre les mécanismes qui génèrent ces excès de zéros. Elle repose sur l'hypothèse que les zéros peuvent être générés par deux processus distincts : un processus de comptage nul (absence de l'évènement d'intérêt) et un processus de comptage positif (présence de l'événement d'intérêt).

Il existe plusieurs méthodes pour analyser les données de comptage parmi lesquelles on retrouve :

Les modéles de comptage classiques : Ces modéles, tels que le modéle de Poisson ou le modéle binomial négatif, sont utilisés lorsque les excès de zéros sont attribués uniquement au processus de comptage nul. Cependant, ces modèles ne prennent pas en compte le processus de comptage positif.

Les modéles de comptage à excès de zéros : Ces modéles plus avancés, tels que le modèle de comptage à excès de zéros de Hurdle ou le modèle de comptage à excès de zéros à deux composantes, prennent en compte à la fois le processus de comptage nul et le processus de comptage positif. Ils permettent ainsi de modéliser les excès de zéros de manière plus précise.

Les modèles mixtes : Ces modèles combinent à la fois des composantes fixes et aléatoires pour tenir compte de la variabilité observée dans les données de comptage à excès de zéros. Ils sont utiles lorsque les observations sont regroupées dans des clusters ou présentent des dépendances spatiales ou temporelles.

En résumé la méthode d'analyse des données de comptage à excès de zéros est une approche statistique qui permet de modéliser et d'analyser des ensembles de données comportant principalement des zéros. Elle prend en compte à la fois le processus de comptage nul et le processus de comptage positif, et utilise des modèles spécifiques pour estimer les paramètres et effectuer des inférences sur les facteurs qui influencent les événements d'intérêt.

1.2 Rappels sur les modèles linéaires gaussiens

Définition 1 *Le modèle linéaire est un ensemble de techniques statistiques qui permettent d'analyser et d'expliquer une variable Y (appelée variable dépendante) par p variables explicatives $X = (X_1, ..., X_p)$.*

Les modèles linéaires gaussiens sont une classe de modèles statistiques largement utilisée pour analyser les relations entre des variables continues. Ils sont basés sur l'hypothèse selon laquelle les variables dépendantes et indépendantes suivent une distribution normale (gaussienne). Ces modèles sont particuliérement utiles lorsque l'on souhaite comprendre et quantifier les relations linéaires entre les variables et effectuer des prédictions.

L'un des modèles linéaires gaussiens les plus couramment utilisés est la régression linéaire, qui cherche à modéliser la relation linéaire entre une variable dépendante continue et un ensemble de variables indépendantes.

Pour ajuster un modèle linéaire gaussien, la méthode des moindres carrés ordinaires (MCO) est souvent utilisée. L'objectif est de trouver les coefficients de régres-

sion qui minimisent la somme des carrés des différences entre les valeurs observées et prédites.

Les modèles linéaires gaussiens offrent plusieurs avantages, tels que leur interprétabilité, leur simplicité et leur large applicabilité. Ils peuvent être utilisés dans divers domaines tels que l'économie, les sciences sociales, la biostatistique, la finance et bien d'autres.

Il convient de noter que les modèles linéaires gaussiens reposent sur certaines hypothèses, notamment l'indépendance des erreurs, l'homoscédasticité (variance constante des erreurs) et la normalité des résidus. Si ces hypothèses ne sont pas respectées, des transformations de variables ou des extensions du modéle peuvent être nécessaires.

En résumé, les modèles linéaires gaussiens constituent une approche puissante et largement utilisée pour modéliser les relations linéaires entre des variables continues. Ils offrent une compréhension précise des effets des variables indépendantes sur la variable dépendante, permettant ainsi d'effectuer des prédictions et de tirer des conclusions statistiques importantes.

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

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

avec les ε_i sont indépendants et identiquement distribués (i.i.d.) suivant $\mathcal{N}(0, \sigma^2)$ et ceci pour tout $i = 1, \cdots, n$. On peut écrire cette formule également sous forme d'un produit scalaire:

$$Y_i = \mathbb{X}\beta + \varepsilon_i$$

avec les $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \mathbb{X} = (1, X_i^{(1)}, \cdots, X_i^{(p)}), \beta = (\beta_0, \cdots, \beta_p)^\top$ et ceci pour tout $i = 1, \cdots, n$. De plus, la réponse doit être indépendante des prédicteurs.
Pour illustrer un exemple supposons n réalisations indépendantes $(X_1, Y_1), ..., (X_n, Y_n)$ du couple (X, Y). Le but est de modéliser la dépendance de la variable réponse Y par rapport aux variables explicatives $X_1, ..., X_p$. Des études diverses et détaillées ont été effectuées sur le modèle linéaire (Azais 2006 [5]). Le modèle linéaire standard traduit la dépendance linéaire de l'espérance en:

$$\beta = (\beta_0, ..., \beta_p)^\top$$

Dans cette expression β_0 représente l'ordonnée à l'origine et les valeurs β_i sont les coefficients de régression des variables calculés à partir des données, β un paramètre inconnu non-contraint de $\mathbb{R}^{p+1}$ et il s'écrit :

$$Y_i = X_i\beta^\top + \varepsilon_i, i = 1, ..., n$$

ε_i est le terme d'erreur dans le modèle ; c'est une variable aléatoire réelle non observée. Le modèle linéaire est caractérisé par les propriétés suivantes:

1. les variables $\varepsilon_i \rightsquigarrow N(0, \sigma^2)$ et sont iid $\forall i = 1, \cdots, n$

2. $\mathbb{E}(\varepsilon_i | X_i) = 0$ les erreurs sont centées

3. $Var(\varepsilon_i) = \sigma^2 \ \forall i = 1, \cdots, n$ la variance des erreurs est constante (homoscèdasticité).

Il est à retenir que certaines données ne peuvent pas être modélisé avec succès par une simple équation linéaire, pour deux raisons:

I) : La distribution de la variable dépendante, elle peut être discrète, et par conséquent, les valeurs prévues doivent l'être également. Par exemple, un chercheur peut s'intéresser à prévoir trois résultats discrets possibles (par exemple, le choix d'un produit parmi trois possibles). Dans ce cas, la variable dépendante ne peut prendre que 3 valeurs distinctes, et la distribution de la variable dépendante est dite multinomiale.
Supposez aussi que nous cherchions à modéliser la structure de nombre d'enfants que les familles souhaitent, en fonction du revenu et d'autres indicateurs socio-économiques. La variable dépendante "nombre d'enfants" est discrète (c'est-à-dire, qu'une famille peut avoir 1, 2, ou 3 enfants etc., mais pas 2,4 enfants), et la distribution de cette variable est très certainement asymétrique (c'est-à-dire, que la plupart des familles souhaitent 1, 2 ou 3 enfants, peu en voudront 4 ou 5, très peu en voudront 6 ou 7, etc.). Dans ce cas il serait raisonnable de penser que la variable dépendante suit une distribution de Poisson.
Si Y est qualitative ou factoriel c'est-à-dire possède un nombre fini de modalités (sexe, couleur,absence ou présence d'un microbe,mort ou vivant) , on ne parlera plus de régression mais de discrimination car cela revient à expliquer l'appartenance de Y à un groupe.

II) Du fait que l'effet des prédicteurs sur la variable dépendante peut ne pas être linéaire. Par exemple, la relation entre l'âge d'une personne et divers indicateurs de santé n'est probablement pas de nature linéaire. Dans ces cas de figure , le modèle linéaire généralisé peut être utilisé pour une modélisation plus efficace des variables dépendantes suivant des distributions discrètes liées de façon non linéaire aux prédicteurs.

1.3 Famille exponentielle

Les familles exponentielles présentent certaines propriétés algébriques et inférentielles remarquables. La caractérisation d'une loi en famille exponentielle per-

met de reformuler la loi à l'aide de ce qu'on appelle des paramètres naturels.
En statistique fréquentiste, ces familles permettent d'obtenir facilement des statistiques d'échantillonnage, à savoir les statistiques suffisantes naturelles de la famille, qui résument un échantillon de données à l'aide d'un nombre réduit de valeurs, constituant les variables de décision en statistique inférentielle.
La famille exponentielle comprend un grand nombres de lois parmi les plus courantes on peut citer entre autres: normale, exponentielle,gamma, chi-deux, bêta, Dirichlet, Bernoulli, Bernoulli multinomiale, Poisson, Wishart, Wishart inverse, etc. D'autres lois courantes ne forment une famille exponentielle que si certains paramètres sont fixes et de valeur connue, telles les lois binomiale et multinomiale (pour un nombre de tirages fixe dans les deux cas), et binomiale négative (pour un nombre d'échecs fixe). Parmi les lois d'usage courant qui ne sont pas de famille exponentielle, on peut citer la loi t de Student, la plupart des mixtures(on appelle densité mélange, ou loi mélange une fonction de densité qui est issue d'une combinaison convexe de plusieurs fonctions de densité , ainsi que la famille des lois uniformes de bornes non fixées.

Définition 2 *Soit Y une variable admettant une densité f Pour une mesure dominante adaptée(mesure de Lebesgue pour une loi continue, mesure discrète combinaison de masses de Dirac pour une loi discrète) supposons $Y_i(i = 1, ..., n)$ les composantes de Y on dira que la densité f de Y appartient a la famille exponentielle si :*

$$f_Y(\theta, \phi) = \exp\left\{\frac{Y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \tag{1.1}$$

où θ est le paramètre d'intérêt et ϕ un paramètre de nuisance dit de dispersion. Les fonctions $a(.), b(.), c(.)$ sont a spécifier , la fonction $a(.)$ est généralement de la forme $a(\phi) = \frac{\phi}{\omega}$ avec ω un vecteur de poids des différents observations, on peut donc en déduire l'espérance et la variance de la variable Y noté respectivement par:

$$\mathbb{E}(Y) = \mu = b'(\theta) \quad \text{et} \quad \mathbb{V}ar(Y) = a(\phi)b''(\theta)$$

1.4 Exemple de lois appartenant à la famille exponentielle

1.4.1 Loi Gaussienne

$y \sim \mathcal{N}(\mu, \sigma^2)$

$$f(y, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(\frac{(y-\mu)}{\sigma})^2\right\}$$

On va identifier $a(\theta)$ et $b(\theta)$ en appliquant le logarithme et en développent la fonction de densité on obtient:

$$f(y,\mu,\sigma^2)=\exp\{\frac{-y^2+2\mu y-\mu^2}{2\sigma^2}+\log(\frac{1}{\sigma\sqrt{2\pi}})\}$$

Par identification

$$\frac{2y\mu}{2}=y\theta\Rightarrow\theta=\mu\quad b(\theta)\frac{\mu^2}{2}=\frac{\theta^2}{2}\quad a(\phi)=\phi,$$

soit

$$\theta=\mu,\quad \phi=\sigma^2,\quad a(\phi)=\phi,\quad b(\theta)=\frac{\theta^2}{2}.$$

1.4.2 Loi Gamma

(incluant la loi exponentielle) de moyenne μ et de variance α^{-1}

$$f(y|\mu,\alpha)=\frac{y^{\mu-1}}{\Gamma(\alpha)\alpha^{\mu}}e^{-\frac{y}{\alpha}},\quad y\in\mathbb{R}$$

est également de la famille exponentielle à condition que $\theta=-\frac{1}{\mu}$.
et on a

$$a(\phi)=\phi,\ b(\theta)=-\log(-\theta),\phi=\alpha^{-1}$$

et

$$c(y,\phi)=\left(\frac{1}{\phi}-1\right)\log(y)-\log\left(\Gamma\left(\frac{1}{\phi}\right)\right).$$

1.4.3 Loi de Poisson

Définition 3 *En thèorie des probabilités et en statistiques, la loi de Poisson est une loi de probabilité discrète qui dècrit le comportement du nombre d'événements se produisant dans un intervalle de temps fixé, si ces événements se produisent avec une fréquence moyenne ou espérance connue, et indépendamment du temps écoulé depuis l'événement précédent.*

Cette loi a été introduite en 1838 par Denis Poisson (1781-1840), dans son ouvrage *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*. Elle s'applique généralement aux phénomène accidentelle où la probabilité p est très faible ou aux phénomène sans mémoire (panes de machines accident de voitures , erreur de fabrication etc $\cdots$) Si le nombre moyen de réalisations dans un intervalle fixé est λ,alors la probabilité que le nombre Y de réalisations sur cet intervalle soit égal à $k,k\in\mathbb{N}$ est:

$$\mathbb{P}(Y=k)=\exp(-\lambda)\frac{\lambda^k}{k!}\tag{1.2}$$

avec $\lambda \in \mathbb{R}_+$
On dit alors que Y suit la loi de Poisson de paramètre λ, et on note $Y \sim \mathcal{P}(\lambda)$ on peut facilement montrer que la loi de poisson appartient à la famille exponentielle . Soit $f(y|\lambda)$ la densité de probabilité en appliquant le $\log$ on a

$$\begin{aligned} f(y|\lambda) &= e^{-\lambda}\frac{\lambda^y}{y!} \\ &= \exp\{-\lambda + y\log(\lambda) - \log(y!)\} \\ &= \exp\{\frac{y\log(\lambda) - \lambda}{1} - \log(y!)\}\ y \in \mathbb{N} \end{aligned}$$

de cette dernière expression on identifie
$\theta = \log(\lambda)$, $a(\phi) = 1$, $b(\phi) = \exp(\theta) = \lambda$
une autre caractéristique importante de la loi de poisson est l'égalité entre l'espérance et la variance
$\mathbb{E}(Y) = \mathbb{V}ar(Y) = \lambda$.
On définit la fonction de variance et la fonction de lien canonique comme suit :

$$V(\lambda) = \lambda$$

et

$$g(\lambda) = \log(\lambda)$$

La loi de Poisson est appelée la loi des événements rares , car elle exprime une répartition binomiale pour un grand nombre d'expériences indépendantes et une probabilité infime de l'événement ,pour cette raison dans certaines conditions, elle peut également être définie comme limite d'une loi binomiale (notamment lorsque $n > 50$ et $np < 5$

1.4.4 Loi de Bernoulli

La loi de Bernoulli est la loi de probabilité d'une variable aléatoire discrète qui prend la valeur 1 avec la probabilité p et 0 avec la probabilité $q = 1 - p$.
Une variable aléatoire suivant la loi de Bernoulli est appelée variable de Bernoulli. Plus formellement, une variable aléatoire Y suit la loi de Bernoulli de probabilité p et on note $\mathcal{B}(\mathrm{P})$ si:

$$\mathbb{P}(Y = y) = p^y(1-p)^{y-1} \tag{1.3}$$

La loi de Bernoulli appartient a la famille exponentielle et cela se démontre comme suit:

notons $f(Y|\lambda)$ la densité de la variable de Bernoulli on a

$$\begin{aligned} f(Y|\lambda) &= p^y(1-p)^{y-1} \\ &= \exp\{\frac{y(\log p - \log(1-p)) + \log(1-p)}{1}\} \end{aligned}$$

on en déduit soit $\theta = \log(\frac{p}{1-p}), \phi = 1, a(\phi) = 1$, $b(\theta) = \log(1+e^{\theta})$ avec
$p = \frac{e^{\theta}}{1+e^{\theta}}$, $c(y,\phi) = 0$
$\mathbb{E}(Y) = p = \lambda$ et $\mathbb{V}ar(Y) = p(1-P)$
On définit également :
$V(\lambda) = b''(\theta) = \lambda(1-\lambda)$
et $g(\lambda) = \theta = \log(\frac{\lambda}{1-\lambda}) = logit(\lambda)$
qui sont respectivement la fonction variance et la fonction de lien canonique de la loi Binomiale

1.5 Recours aux Modèles linéaire Généralisés

Dans cette partie, nous énonçons quelques notions essentielles sur la théorie des modèles de comptage. Ainsi, nous présentons brièvement des résultats essentiels rencontrés dans la littérature. Nous définissons les notions de modèle linéaire généralisé puis présentons les méthodes d'estimation les plus couramment utilisé dans ces modèles. Pour la description de cette section nous nous sommes en grande partie basé en sur (McCullagh et Nelder 1989[22]) et Diallo 2017[3].

Définition 4 *Le modèle linéaire généralisé est une généralisation souple de la régression linéaire simple elle est souvent noté (GLM).*

Parfois, les modèles linaires gaussiens (LNM) ne suffisent plus pour modéliser la réalité correctement comme on suppose la loi de la réponse comme étant continue et souvent gaussienne. Une telle loi n'est pas adaptée au cas où la réponse est par exemple un comptage ou une réponse binaire. Effectivement, ces deux derniers prennent des valeurs positives entières. De plus, contrairement à une loi gaussienne, la distribution de la réponse dans le cas de comptage ou binaire n'est pas forcément symétrique. Ainsi, on introduit les modèles linéaires généralisés (GLM) pour se charger de ces situations. L'extension des modèles linéaires consiste en admettant à la réponse Y de prendre une loi de la famille exponentielle. Comme tous les modèles marginaux, les GLMs sont également conditionnés par rapport aux prédicteurs fixes. On part donc des modèles linéaires gaussiens :

$$\begin{cases} (Y_i, X_i^{(1)}; \cdots; X_i^{(p)}) \text{sont indépendant pour } i = 1, \cdots; n \\ \mathcal{L}(Y_i|\mathbb{X}_i) = \mathcal{N}(\mu_i, \sigma_i^2) \\ \sigma_i^2 = \sigma^2 \text{pour } i : 1 \cdots n \text{(homoscédasticité)} \\ \mu_i = \mathbf{E}(Y_i|\mathbb{X}_i) = \mathbb{X}_i\beta \text{une application linéaire en } \beta \end{cases} \tag{1.4}$$

et étend cette théorie en obtenant les modèles linéaires généralisés :

$$\begin{cases} (Y_i, X_i^{(1)}; \cdots; X_i^{(p)}) \text{sont indépendant pour } i = 1, \cdots; n \\ \mathcal{L}(Y_i|\mathbb{X}_i) \in \text{famille exponentielle} \\ (\text{où } \mathcal{L}(Y_i|\mathbb{X}_i) \text{ répresente la loi de } Y_i \text{ conditionnellement au predicteur }) \\ \mu_i = \mathbf{E}(Y_i|\mathbb{X}_i) \\ g(\mu_i) = \eta_i = \mathbb{X}_i\beta (\text{application linéaire en } \beta) \\ \text{où } g(.) \text{ est une fonction de lien inversible et } \eta_i \text{ est le predicteur linéaire} \end{cases} \tag{1.5}$$

Pour la régression de Poisson, un GLM où $\mathcal{L}(Y_i|\mathbb{X}_i) = \mathcal{P}(\mu_i)$ on a alors :

$$g(\mu) = \log(\mu)$$

et

$$Var(\mu) = \mu$$

et de ces résultats on déduit l'espérance:

$$\log(\mathbf{E}(Y_i|\mathbb{X}_i)) = \mathbb{X}_i\beta \Leftrightarrow \mathbf{E}(Y_i|\mathbb{X}_i) = Var(Y_i|\mathbb{X}_i) = \mu_i = \exp(\mathbb{X}_i\beta)$$

Le modèele ne contient pas de terme d'erreur ε_i à cause de la relation entre l'espérance et la variance. Effectivement, une fois l'espérance connue, on a spécifié entièrement le modèle. En choisissant une loi de Bernoulli, on obtient un GLM où $\mathcal{L}(Y_i|\mathbb{X}_i) = \mathcal{B}(1, \pi_i)$ et $\mathbf{E}(Y_i|\mathbb{X}_i) = \mathbb{P}(Y_i = 1|\mathbb{X}_i) = \pi_i$.
Ainsi la fonction de lien canonique vaut: $g(\pi) = \log(\pi)$ et la fonction de variance vaut $Var(\pi) = \pi(1 - \pi)$. Ici à nouveau le terme d'erreur devient superflu.
Avec ces résultats, on peut calculer l'espérance conditionnellement aux prédicteurs:

$$logit(\mathbf{E}(Y_i|\mathbb{X}_i)) = \mathbb{X}_i\beta \Leftrightarrow \mathbf{E}(Y_i|\mathbb{X}_i) = \pi = \frac{\exp(\mathbb{X}_i\beta)}{1 + \exp(\mathbb{X}_i\beta)}$$

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

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

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

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

En dérivant $\ell(Y|\mu)$ par β_j et en l'annulant, on obtient l'estimateur $\beta_j\hat{\beta}_j$.
Une simplification lors de l'estimation est basée sur le fait que maximiser $\ell(Y|\mu)$ revient à maximiser $\sum_{i=1}^{n}(-\mu_i + y_i\log(\mu_i))$ car $\log(y_i!)$ est constante.
Pour vérifier s'il s'agit vraiment d'un maximum, il faut qu'on a $-\frac{\partial^2\ell(Y|\hat{\theta}_j)}{\partial\beta_j^2} > 0$ pour $j = 0\cdots p$.
Si on admet que $\hat{\theta}$ est le vecteur des solutions de l'annulation de la dérivée de la log-vraisemblance.
La fonction$glm(.)$ de R calcule les maxima de vraisemblance par la méthode itérative des moindres carrées re-pondérés. Le modèle GLM est très utilisé dans le cas des données discrètes mais aussi parfois pour des données continue pour lesquelles la loi normale n'est pas très adapté, (1972, Nelder[24]) et (1989, McCullagh [25]) en font une présentation détaillée.
Le phénomène connu sous le nom d'inflation de Zeros a été à l'origine de l'extension du modèle linéaire gaussien en modèle linéaire Généralisé.

1.5.1 Caractérisation d'un modèle linéaire généralisé

Si nous disposons d'une variable à expliquer et des variables explicatives, un modèle linéaire généralisé GLM est caractérisé par les trois hypothèses suivantes:

1.5.1.1 La distribution de la variable à expliquer

Soit un échantillon statistique composé de n variables aléatoires
$Y_i(i = 1, ..., n)$, indépendant admettant des distributions issu d'une structure exponentielle au sens de (Nelder et Wedderburn 1972, [24]). Pour une mesure dominante adaptée (mesure de lebesgue pour une loi continue et mesure discrète combine de masse de Dirac pour une loi discrète) la fonction de densité de la variable aléatoireY_i s'ecrit sous la forme

$$f_{Y_i}(y_i, \theta_i) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\} \tag{1.6}$$

ou $\theta_i \in \mathbb{R}$ est un paramètre canonique (ou de position) et $\phi \in \mathbb{R}_+^\star$ un paramètre de dispersion (ou d'échelle) b et c sont des fonctions spécifiques à chaque distribution,

la fonction b est supposé deux fois dérivable et la fonction a_i est donnée par :
$a_i(\phi) = \frac{\phi}{\omega_i}$ où ω_i est un poids connu associer à la réalisation de y_i
tableaux :Exemple

Distribution	θ	$b(\theta)$	$\phi \neq a(\phi)$	$C(y, \phi)$
$\mathcal{B}(n, \pi)$	$\log(\pi/1-\pi)$	$n\log(1+e^{\theta})$	1	$\log(C_n^y)$
$\mathcal{P}(\lambda)$	$\log(\lambda)$	e^{θ}	1	$-\log(y!)$
$\mathcal{N}(\mu, \sigma^2)$	μ	$\theta^2/2$	σ^2	$\frac{1}{2}\left[\log(2\pi\sigma^2) + \frac{y^2}{\sigma^2}\right]$
$\mathcal{NB}(\mu, \alpha)$	$\log(\frac{\mu\alpha}{1+\mu\alpha})$	$-\frac{1}{\alpha}\log(1-e^{\theta})$	1	$\log(\Gamma(y+\frac{1}{\alpha})) - \log(y!\Gamma(\frac{1}{\alpha}))$
$\gamma(\alpha, \beta)$	$-\frac{1}{\alpha}$	$-\log(-\theta)$	$\frac{1}{\beta}$	$\beta\log[y\beta] - \log[y\Gamma(\beta)]$

Quelque exemple de distribution de famille exponentielle.

1.5.1.2 Le prédicteur Linéaire

Comme dans un modèle linéaires ,les variables interviennent linéairement dans la modélisation
Soit X une matrice de variables explicatives,$X = X_1, X_2, \cdots, X_n)^\top$ La matrice X est d'ordre $n \times p$ ($n > p$) où p est le nombre de variables explicatives. Soit β un vecteur de p paramètres . Le prédicteur linéaire est définit par :

$$\eta = X\beta^\top \tag{1.7}$$

C'est le composant déterministe du modèle , si rang(X) = p alors le modèle est dit régulier

1.5.1.3 Fonction de lien

Définition 5 *La fonction de lien exprime la relation fonctionnelle entre l'espérance de Y_i et la i-ème composante du predicteur linéaire. Elle est noté $\forall i = 1, ..., n$ on a*

$$\eta_i = g(E(Y_i)) \tag{1.8}$$

où g est supposée monotone et la différentiable , g est donc appelée fonction de lien .
La fonction de lien pour laquelle $\eta_i = \theta_i$ est appelée **fonction de lien canonique**.
La fonction de lien canonique associée à une distribution donnée vérifie $g = b'^{-1}$
Voici un exemple de fonctions de liens canonique associées à quelque loi classique

loi	$\mathcal{P}(\lambda)$	$\mathcal{B}(n,\pi)$	$\mathcal{N}(\mu,\sigma^2)$	$\gamma(\alpha,\beta)$
$g(x)$	$\log(x)$	$\log(\frac{x}{1-x})$	x	$\frac{1}{x}$

Remarque 1.1 *La fonction de lien permettant d'etablire une égalié entre le prédicteur linéaire et le paramètre canonique est appelée fonction de lien canonique.car $\eta_i = g(b'(\theta_i))$, la fonction de lien associée à une distribution donnée vérifie $g = b'^{-1}$.*

1.5.1.4 Expression des moments

Soit Y une variable à expliquer qui a des observations indépendantes $y_i, \cdots, y_n$ provenant d'une distribution ayant une une densité de probabilité ou une fonction de masse Y_i de la forme (1.6)

L'espérance et la Variance de Y_i son liées et s'expriment en fonction des paramètres θ_i et ϕ , on peut donc écrire :
soit $l_i(\theta_i, \phi, y_i) = \log f_{y_i}(y_i, \theta_i, \phi)$ la fonction de log-vraisemblance de la iem de y_i on à

$$l_i(\theta_i, \phi, y_i) = \frac{\{y_i\theta_i - b(\theta_i)\}}{a_i(\phi)} + c(y_i, \theta_i) \tag{1.9}$$

$$\frac{\partial l_i(\theta, \phi, y_i)}{\partial \theta_i} = \frac{\{y_i\theta_i - b'(\theta_i)\}}{a_i(\phi)} \tag{1.10}$$

et

$$et\ \frac{\partial^2 l_i(\theta, \phi, y_i)}{\partial \theta_i^2} = \frac{-b''(\theta_i)}{a_i(\phi)} \tag{1.11}$$

Sous certaines conditions de régularités vérifiées par les lois issues de structures exponentielles, on a les relations suivantes :

$$\mathbb{E}(\frac{\partial l_i(\theta, \phi, y_i)}{\partial \theta_i}) = 0$$

et

$$-\mathbb{E}(\frac{\partial^2 l_i(\theta, \phi, y_i)}{\partial \theta_i^2}) = \mathbb{E}((\frac{\partial^2 l_i(\theta, \phi, y_i)}{\partial \theta_i^2})^2)$$

et donc on en déduit alors que
$\mathbb{E}(Y_i) = b''(\theta_i)$ et $Var(Y_i) = b''(\theta_i)a_i(\phi)$
On peut aussi dire qu'il existe une relation entre l'espérance et la variance de Y_i

$$Var(Y_i) = a_i(\phi)b''(b'^{-1}(\mathbb{E}(Y_i))) = \frac{\phi}{w_i}b''(b'^{-1}(\mathbb{E}(Y_i))) \tag{1.12}$$

La variance est le produit de deux fonctions l'une dépendant du paramètre canonique θ et l'autre dépendant de ϕ qui est le paramètre de dispersion, la fonction $a(\phi) = \frac{\phi}{w}$.
Pour une loi normale de moyenne m on peut donc écrire $a(\phi) = \frac{\sigma^2}{m}$ et donc $m = w$

1.5.1.5 Estimation des paramètres de régression

Dans cette partie, nous nous intéressons à l'estimation du paramètre β, l'estimation des paramètres de régression consiste à chercher $\widehat{\beta}$ de β qui maximise la vraisemblance .
Pour définir les éequations de vraisemblance, nous allons donc supposer que ϕ est connue et que la fonction de lien utilisé est une fonction de lien canonique. Les équations sont alors définis par:

$$l = \{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\} \tag{1.13}$$

L'équation du score est donnée par la formule des chaine ci-dessous :

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

En effet

$$\begin{aligned}
\frac{\partial l_i}{\partial \beta_j} &= \frac{y_i - b(\theta_i)}{a_i(\phi)} \\
\frac{\partial \theta_i}{\partial \mu_i} &= 1/\frac{\partial \mu_i}{\theta_i} = \frac{1}{b''(\theta_i)} = \frac{a_i(\phi)}{var(Y_i)} \\
\frac{\partial \mu_i}{\partial \eta_i} &= \eta_i \text{dépend de la fonction de lien } g(\mu_i) \\
\frac{\partial \eta_i}{\partial \beta_j} &= X_{ij} \text{ car } X_i^\top \beta = \eta_i
\end{aligned}$$

On en dédiut les équation de la vraisemblance pour le paramètre β_j

$$\sum_{i=1}^{n} X_i = \frac{y_i - \mu_i}{var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i} = 0. \forall j = 1, \cdots, p \tag{1.14}$$

l'équation 1.14 est lineaire en β.
En effet, elle dépend de β au travers μ_i de $\eta_i \forall i = 1, \cdots, n$.
Remarquons aussi que cette équation n'admet pas de solution analytique. Cependant des méthodes itératives comme l'algorithme de Newton-Raphson et de Ficher-Scoring sont utilisées pour approcher l'estimateur du maximum de vraisemblance .

D'une manière plus géneral pour les modèles linéaires généralisés, (Fahrmeir et Kaufmann, 1985 [12]) ont démontré différents résultats dont, en particulie, le théorème sur la normalité asymptotique de $(\hat{\beta}_n)$, solution des équations du maximum de vraisemblance pour un échantillon de taille n ce théorèmerepose principalement sur des hypothèses concernant les matrices hessiennes et d'information de Fishe

Théorème 1.2 *Sous les conditions émises par (Fahrmeir et Kaufmann 1985[12]), la suite $(\hat{\beta}_n)$ des estimateurs du maximum de vraisemblance $\hat{\beta}_n$ converge en probabilité vers β et $\mathcal{I}_n(\hat{\beta}_n)^{\frac{1}{2}}(\hat{\beta}_n - \beta_0)$ converge en loi vers le vecteur gaussien $\mathcal{N}(0, I_p)$, où $\mathcal{I}_n(\beta) = \frac{-\partial \ell_n(\beta)}{\partial \beta \partial \beta^\top}$ où β_0 est une valeur inconnu du paramètre.*

1.6 Phénomène de Sur-dispersion

On a déjà mentionné précédemment mentionner qu'une caractéristique importante de la loi de Poisson est l'égalité entre l'espérance et la variance. En effet, comme pour la loi binomiale et contrairement à la loi gaussienne, l'espérance et la variance sont liées. Malheureusement, l'égalité de l'espérance et de la variance n'est pas toujours maintenue; même pas approximativement. La sur-dispersion est un phénomène couramment rencontré en analyse statistique des données de comptage.
Elle survient dans de nombreux domaines comme assurance, économie , épidémiologie etc $\cdots$. Ses causes sont variées, par exemple, la présence d'une hétérogénéité inobservée entre individus ou l'inflation de zéros, la non normalité(pour plus de détaille voir Hilbe [19]) . La loi de Poisson a toujours servi de référence pour le modèle de comptage. l'une des caractéristique la plus importante de ladite loi est que la variance est egale à l'espérance $Var(Y) = \mathbb{E}(Y) = \mu$ Pour les lois gaussiennes l'espérance et la variance sont liées.
Force est de constater que l'égalité entre la variance et l'espérance n'est pas toujours vérifié pour certaines données .
Soit Y la variable réponse d'un modèle de Poisson , c'est à dire $Y \sim \mathcal{P}(\mu)$ On dira que il y a sur - dispersion si $Var(Y) > \mu$.
Elle peut être lier à plusieurs facteurs. Un excès de comptages zéro est l'une des causes principales de cette sur-dispersion , les données non indépendant des variables sont toute fois importants à inclure dans le modèle. Sa détection est facile à partir du rapport entre la déviance résiduelle et son degré de liberté. Il existe également un test performant pour détecter la sur-dispersion , utilisable dans le cas où on a estimé nos paramètres par la méthode de maximum de vraisemblance. Dean (1992) a proposé une statistique de test permettant d'évaluer la sur-dispersion liée à un modèle de Poisson qui est basé sur le test du score. Pour prendre en compte la sur-dispersion des données, on introduit un paramètre de dispersion noté ϕ et un vecteur de poids à priori noté ω pour obtenir la relation suivante :

$$Var(Y) = \frac{\phi}{\omega}\mathbb{E}(Y) = \frac{\phi}{\omega}\mu$$

Il s'agit d'un modèle 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 sur-dispersion des données. Ainsi on a généralisé le lien entre l'espérance et la variance car si on a $\phi = 1$, on se retrouve dans le modèle de Poisson habituelle. Dans la situation poissonienne, on assume également que les poids à priori sont tous égale à 1. Notons aussi que l'intérêt majeur de l'estimation par quasi-vraisemblance est d'étendre les techniques d'estimations classiques à des modèles ayant des fonctions

variances n'appartenant pas à des familles exponentielles ainsi qu'à des modèles dont la distribution n'est pas totalement spécifiée [18]. En effet, cette technique d'estimation ne nécessite que la spécification des deux premiers moments de la distribution ainsi que la fonction de lien entre ces deux moments. Un modèle basé sur la quasi-vraisemblance s'explique de la façon suivante:

$$\begin{cases} Y_i, X_i^{(1)}, ..., X_i^p & \text{sont independent pour } i =, ..., n \\ \mu_i = \mathbb{E}(Y_i|X_i) & \\ g(\mu_i) = \eta_i = X_i\beta & \text{(application lineaire en } \beta) \\ & g(.)\text{est la fonction de lien } \eta_i \text{ est le predicteur lineaire} \\ Var(y_i|X_i) = \frac{\phi}{\omega_i}V(\mu_i) & \text{où } V(.) \text{ est la fonction de variance} \end{cases} . \tag{1.15}$$

A partir d'un modèle de la quasi-vraisemblance on a obtenue les estimateurs et leurs erreurs on ne peut pas appliquer les tests de comparaison des modèles en se basant sur le test de rapport de vraisemblance car les estimateurs et leurs erreurs ne sont pas obtenue à partir de la pure vraisemblance.
Donc pour estimer les paramètres de dispersion, on peut utiliser le test de Pearson qui suit à peut près la loi de chi-deux X^2 à $n-p-1$ degrés de libertés: on a

$$X^2 = \sum_{i=1}^{n} \omega_i \frac{(y_i - \hat{\mu_i})^2}{\hat{Var(\mu_i)}}$$

la valeur du paramètre de dispersion estimer est alors :

$$\hat{\phi} = \frac{X^2}{n-p-1} = \hat{\sigma}^2$$

Pour obtenir le maximum de quasi-vraisemblance , on doit utiliser la fonction log-quasi-vraisemblance

Définition 6 *La quasi-vraisemblance est une fonction des paramètres évaluée aux observations, à l'instar de la vraisemblance. Elle est définie comme suit:*

$$Q(\mu, y) = \sum_{1}^{n} Q_i(\mu_i, y_i)$$

où

$$Q_i(\mu_i, y_i) = \int_{y_i}^{\mu_i} \frac{(y_i - t)}{\phi Var(t)} dt$$

avec $\phi V(\mu_i)$ est la variance de y_i et dans le cadre poisson on obtient

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

Comme y_i et $y_i \log(y_i)$ sont des termes à valeur constante on pose:

$$y_i - y_i \log(y_i) = c$$

Alors

$$Q_i(\mu_i, y_i) = \frac{1}{\phi}(y_i \log(\mu_i) - \mu_i + c)$$

Y est un vecteur d'observation et $y_i iid$ de moyenne $\mu = g^{-}1(X\beta)$ et de variance $V(\mu)$. Cette fonction possèd trois propriétés commune avec la log-vraisemblance d'une loi exponentielle GLM

- $\mathbb{E}(\frac{\partial Q}{\partial \mu}) = 0$
- $\mathbb{E}(frac\partial^2 Q \partial \mu^2) = -\frac{1}{\phi V(\mu)}$
- $Var(\frac{\partial Q}{\partial \mu}) = \frac{1}{\phi V(\mu)}$

1.7 Mélange poisson Gamma: la loi Binomial Négative

La loi binomiale négative est une loi discrète et peut être considéré comme une généralisation de la loi Poisson. Plus précisément on parle d'une mixture de Poisson Gamma . Effectivement on peut voir cette loi comme une loi de poisson avec sont seul paramètre étant une variable aléatoire suivant une loi Gamma
$Y|\Theta = \theta \sim \mathcal{P}(\theta)$ où $\Theta \sim \mathcal{G}(\alpha, \beta)$ où $\alpha > 0$, $\beta > 0$
On définie θ comme un paramètre de forme et β comme un paramètre d'échelle .
Ainsi la densité de la loi Gamma est:

$$f(\theta, \alpha, \beta) = \frac{1}{\beta^\alpha} \frac{1}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\frac{\theta}{\beta})$$

où Γ désigne la fonction gamma définie par

$$\Gamma(x) = \int_0^1 t^{x-1} e^{-t} dt.$$

Contrairement aux lois discrètes , la loi gamma ainsi que la normal contiennent un paramètre de dispersion et donc en tout, deux paramètres normalement capables de gérer les variables extra poissonnière. La densité jointe est définie par

$$\mathbb{P}(Y = y|\Theta = \theta)f_{\Theta}(\theta) = \frac{\theta^y}{y!}\exp(-\theta)\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\theta^{\alpha-1}\exp(-\frac{\theta}{\beta}) \tag{1.16}$$

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

$$\mathbb{P}(Y = y, \alpha, \beta) = \int_0^{\infty} \mathbb{P}(Y = y|\Theta = \theta)f_{\Theta}(\theta)d\theta$$

$$\mathbb{P}(Y = y, \alpha, \beta) = \frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)}\left(1 - \frac{1}{1+\beta}\right)^y\left(\frac{1}{1+\beta}\right)^{\alpha} \tag{1.17}$$

preuve Annexs

Pour l' espérances et la variances on obtient :

$$\mathbb{E}(Y) = \alpha\beta = \mu$$

et

$$\mathbb{V}ar(Y) = \mu(1+\beta) = \mathbb{E}(Y)(1+\beta)$$

preuve Annexs

2

Modèle de régression à inflation de zéros.

Résumé

Dans ce chapitre, nous nous concentrerons sur un phénomène spécifique connu sous le nom d'inflation de zéros, qui est une cause courante de sur-dispersion. Celui-ci se produit lorsque nous observons un nombre "excessif" de zéros dans des données de comptage. Différentes approches peuvent être utilisées pour modéliser ce type de données. Dans cette section, nous nous pencherons sur une classe particulière de modèles appelés "modèles à inflation de zéros". Ces modèles sont une combinaison d'une masse de Dirac en zéro et d'un modèle de comptage classique. Nous discuterons des modèles de régression telles que poisson et binomiale négative puis énoncerons quelques propriétés asymptotiques du modèle ZIP.

Sommaire

2.1 Introduction

Les modèles de base pour les données de comptages sont les modèles de Poisson et binomial négatif. Le modèle de régression de Poisson suppose que les données sont dispersées de manière égale, elle sous-estimer la probabilitè d'absence d'évenement c'est-à-dire, que la variance est égale à la moyenne, ce qui n'est pas toujours le cas. La loi binomiale négative quant à elle permet de ne pas sous-estimer cette probabilitè avec l'inclusion du paramétre de sur-dispersion, ce qui permet d'augmenter la variance conditionnelle sans modifier la moyenne conditionnelle.
Les modèles de régression à inflation de zéros sont des modèles souvent utilisés pour modéliser des données de comptage sur-dispersées lorsque la sur-dispersion est liée à la présence d'une grande proportion de zéros. Ces modèles ont démontré leur utilité dans divers domaines comme l'épidémiologie, l'économie , l'assurance, l'agriculture, l'industrie, l'écologie,la santé publique, la psychologie, la sociologie,etc. La régression de Poisson à excès de zéro (ZIP) et la régression binomiale négative à excès de zéro (ZINB),le régression ZIPG, sont beaucoup utiles pour modéliser ces données.
Lors d'un comptage, les zéros ont souvent un statut particulier qui peut prêter à confusion (Ridout et al.2001,[28]). En effet, on distingue deux types de zéros: ceux qui sont dûs à l'échantillonnage (zéros aléatoires) et ceux qui sont dûs à la structure (zéros structurels)
Il est important de tenir compte de ce facteur au cas contraire nous serons confronter a une situation de de sur-dispersion, l'inflation de zéros (Fong and Yip1995, [13]) ; (Mullahy 1997,[26]) ;(Diop et al. 2011[11])

2.2 Modèle de Régression ZI

2.2.1 Modèles de régression de Poisson et binomial négatif

Pour expliquer une variable quantitative Y à valeurs entières le modèle de régression de Poisson est souvent utilise. La probabilité que la variable Y prenne la

valeur $y_i (y_i = 0, 1, 2 \cdots)$est donnée par:

$$\mathbb{P}(Y_i = y_i | X_i = x_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i} \tag{2.1}$$

où le paramètre λ_i dépend du vecteur de X_i par une équation log-linéaire à savoir: $\log(\lambda_i) = \beta^\top X_i$, où $\beta = (\beta_0, \cdots, \beta_p)$ est le vecteur des coefficients à estimer.

On vérifie aisément que dans le modèle 2.1, l'espérance est égale à la variance $\mathbb{E}(Y_i|X_i = x_i) = \mathbb{V}(Y_i|X_i = x_i) = \lambda_i = \exp(\beta^\top X_i)$. La forme de la fonction exponentielle assure la non-négativité du paramètre de la moyenne λ_i. Dans la pratique, l'abondance de valeurs nulles ou des valeurs extrêmes, qui entraine la sur-dispersion, (Cox et David R.1983, [7]) de la variable Y, peut remettre en cause l'utilisation de ce modèle.
Cette situation ne permet pas une bonne estimation des paramètres d'où l'idée d'utiliser un modèl alternative basé sur le binomiale négative qui prend en compte la sur-dispersion et l'introduction d'un nouveau paramètre α qui gère l'hétérogénéité inobservée de la variable endogène.
On definit la probabilité d'une variable Y qui prend la valeur y_i dans un modèle Binomiale négatif comme suit:

$$\mathbb{P}(Y_i = y_i | X_i = x_i) = \frac{\Gamma(y + \frac{1}{\alpha})}{y_i! \Gamma(\frac{1}{\alpha})} \left(\frac{1}{1 + \alpha\lambda_i}\right)^{\frac{1}{\alpha}} \left(\frac{\lambda_i}{\frac{1}{\alpha} + \lambda_i}\right)^{y_i}$$

Dans ce modèle le paramètre α permet de mesure le degré de sur-dispersion . La loi Binomiale négative tend vers la loi de Poisson lorsque α tend vers zéro.
Si $\alpha > 0$, le modèle de poisson est rejeté au profil du modèle binomial négatif. La sur-dispersion peut être testée :

- soit par le ratio $D/(n-p)$,où D désigne la déviance, n le nombre d'observations et p le nombre de para'etres dans le modèl
- soit par le ratio $\chi^2/(n-p)$ où χ^2 est la statistique de chi-deux de person.

Le modèle Zero-Inflated (ZI) est utilisé pour capturer ce phéenomeène d'inflation zéro. Il modélise à la fois la probabilité d'observer une valeur zéro et la distribution des valeurs non nulles . Le modèle ZI permet donc de prendre en compte cette sur-dispersion et d'ajuster la probabilité d'observer une valeur zéro de manière distincte.Le phénomène d'inflation zéro est couramment observé dans les données de comptage ce qui a permit la mise en place des modèles tels que ZIP, ZINB

2.2.2 Modèle de Régression ZIP , ZIB & ZINB

2.2.2.1 Modèle de Régression ZIP

Soit $Y_i, i = 1, ..., n$ une variable de comptage positive , on dira que Y_i est modelisée par ZIP si sa distribution est de la forme:

$$P(Y_i = y_i) = \begin{cases} \pi + (1-\pi)\exp(-\lambda_i) & \text{si } y_i = 0 \\ (1-\pi)\frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} & \text{si } y_i > 0 \end{cases} \tag{2.2}$$

On note $Y_i \sim$ ZIP (λ, π).
Conditionnellement à X_i et W_i ,l' espérance et la variance sont donnés par:
$\mathbb{E}(Y_i|X_i; W_i) = (1-\pi)\lambda_i$ et $\mathbb{V}ar(Y_i|X_i; W_i) = (1-\pi)(1+\pi\lambda_i)\lambda_i = (1+\pi)\mathbb{E}(Y|X, W)$

Preuve

2.3 Modèle de Régression ZIB

2.3.1 Régression binomiale

La régression binomiale ou logistique est une technique statistique utilisée pour modéliser des variables de réponse binaires (absence ou présence , succès ou échec) ou des proportions qui suivent une distribution binomiale. Elle est particulièrement utile lorsque les données sont regroupées en catégories ou lorsqu'il y a des essais de Bernoulli répétés.

Soit $i \cdots I$ différent valeurs fixées , $x_{i1} \cdots x_{ip}$ des variable explicative , $X_1 \cdots X_p$, pour chaque groupes, on réalise n observations $n = \sum_{i=1}^{n_i}$. Supposons qu' à l'intérieure de chaque groupe tout les individus ont la même probabilité de succée ; si la variable Y est distribuer selon la loi binomiale $\mathbb{B}(n, \pi)$ alors sa densité est donner par:

$$\mathbb{P}(Y = y_i) = \binom{n}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} \tag{2.3}$$

où la probabilité de succès π_i est modélisée par une fonction de lien $logit$:
$logit(\pi_i) = \beta^\top X_i$.
Dans la pratique, il n'est pas toujours garanti que les observations binaires successives soient ind épendantes, la modélisation pourrait s'étendre dans le cas où la variable réponse Y prend K (K > 2) modalités (Agresti2002, [1]).

Le modèle "Zero-Inflated Binomial" (ZIB) est un modèle statistique utilisé pour modèliser des données binaires ou de comptage avec une proportion importante de

zéros. Dans ce modèle la probabilité des zéros est modélisée par une composante qui modélise inflation de zéro, tandis que la probabilité des autres valeurs non nulles suit une distribution binomiale. Ce modèle est définit en remplaçant la loi de Poisson par la loi binomiale. Notons $\mathbb{B}(n,\pi)$ la loi binomiale de paramètre (n,π) avec $n \in \mathbb{N}$ et $p \in (0,1)$ on définit alors le modèle Binomiale à inflation de zéros (ZIB). On dit qu'une variable aléatoire Y_i est modelisé par ZIB si

$$Y \sim \begin{cases} 0 & \text{avec une probabilité } p_i \\ \mathcal{B}(n,\pi_i) & \text{avec une probabilité } 1-p_i \end{cases} \tag{2.4}$$

D'une manère explicite on dit qu'une variable $Y_i \sim$ ZIB (n,π_i) si sa loi de probabilité s'écrit sous la forme:

$$Y_i = \begin{cases} 0 & \text{avec une probabilié } p_i + (1-p_i)(1-\pi)^{n_i} \\ k & \text{avec une probabilité} (1-p_i)\binom{n_i}{k}\pi_i^k(1-\pi_i)^{n_i-k}, k = 1,2,\cdots n_i \end{cases} \tag{2.5}$$

les probabilitès π et p sont généralement données par une régression logistiques donnant le modèle suivant

$$\begin{cases} Y_i \sim \pi_i\delta_0 + (1-\pi_i)\mathcal{B}(n,\pi_i) \\ logit(\pi) = \beta^\top X_i \\ logit(p_i) = \gamma^\top W_i \end{cases} \tag{2.6}$$

où $W \in \mathbb{R}^p$ et $X \in \mathbb{R}^q$ sont des vecteurs de covariables ,n étant le nombre d'individus p et q sont respectivement le nombre de covariables dans le modèle de régression binomial et le nombre de covariables dans la partie inflation de zéros, $\beta \in \mathbb{R}^q$ et $\gamma \in \mathbb{R}^p$ sont des paramètres de régression.
l'espérance et la variance du modèle sont respectivement donner par:

$$\mathbf{E}(Y_i) = (1-p_i)n_i\pi_i.$$
$$\mathbf{V}(Y_i) = (1-p_i)n_i\pi_i(1-\pi_i(1-p_in_i)).$$

2.3.2 La vraisemblance dans le modèle ZIB

Les deux probabilités de l'équation (2.5) peuvent être exprimées conjointement comme une distribution de Bernoulli généralisée donnant la vraisemblance suivante:

$$L_n(\beta,\gamma) = \prod_{i=1}^{n}\left(\pi_i + (1-\pi_i)\,(1-p_i)^n)^{j_i}\,(1-\pi_i)\binom{n}{y_i}p_i^n\,(1-p_i)^{n_i-y_i}\right)^{1-j_i} \tag{2.7}$$

La log vraisamblance du modèle est alors donnée en appliquant le logarithme a la vraisemblance à l'équation 2.7 on a alors :

$$\log(L_n(\beta,\gamma)) = \sum_{i=1}^{n} \Big(j_i \log\left(\exp(\gamma^\top W_i) + (1+\exp(\beta^\top X_i)^{-n_i}\right) - \log(1+\exp\gamma^\top W_i)\Big) + \Big((1-j_i)\times(y\beta^\top X_i - n_i\log(1+\exp\beta^\top X_i))\Big) \tag{2.8}$$

Les estimations des paramètres de β et γ peuvent être déterminées en utilisant la méthode du maximum de vraisemblance en se basant sur l'algorithme EM .

2.4 Modèle de Régression ZINB

Le modèle ZINB est une extension du modèle ZIP (Zero-Inflated Poisson) qui permet de modéliser des données avec une distribution de comptage qui ne suit pas exactement une distribution de Poisson, mais qui présente une surdispersion supplémentaire. La distribution utilisée pour modéliser les valeurs non nulles dans un modèle ZINB est la distribution de la loi binomiale négative (Negative Binomial).
il existe deux étapes dans la régression ZINB, à savoire l'emergence des valeurs nulles dans la variable de répose 2011,[14]. Le premier est l'état zéro avec une probabilité π_i . Le deuxième état est un état binomial négatif où la variable réponse a une valeur suivant une distribution binomial négative avec une moyenne et avec une chance d'occurrence $(1-\pi_i)$. La densité de probabilité du modèle ZINB avec k variable prédictive est donnée par:

Yi est modèlisée par un ZINB si sa distribution s'exprime comme suit:

$$P(Y_i=y_i) = \begin{cases} \pi + (1-\pi)(\frac{1}{1+\alpha\lambda_i})^\alpha & \text{si } y_i = 0 \\ (1-\pi)\frac{\Gamma(y_i+\frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})y_i!}(\frac{\alpha\lambda_i)}{1+\alpha\lambda_i})^{y_i}(\frac{1}{1+\alpha\lambda_i})^{\frac{1}{\alpha}} & \text{si } y_i > 0 \end{cases} \tag{2.9}$$

avec $Y_i \sim \mathbf{NB}(\lambda_i,\alpha)$ et les covariable W_i et X_i sont données par : $W_i = (W_{i1},...,W_{iq})^\top$ et $X_i = (X_{i1},...,X_{ip})$.
où $X_i \in \mathbb{R}^p$ et $W_i \in \mathbb{R}^q$

L'espérence et la variance sont donnée respectivement donné par:
$\mathbb{E}(Y_i|X_i;W_i) = (1-\pi_i)\lambda_i$ et $\mathbb{V}ar(Y_i|X_i;W_i) = (1-\pi_i)\lambda_i(1+(\alpha\pi_i)\lambda_i)$

$$\begin{cases} \lambda_i = \exp(\beta^\top X_i) \\ \pi_i = \frac{\exp\gamma^\top W_i}{1+\exp\gamma^\top W_i} \\ (1-\pi_i) = \frac{1}{1+\exp\gamma^\top W_i} \end{cases} \tag{2.10}$$

Dans le modèle (2.9) si α tend vers zero alors le modèle ZINB tend vers le modèle ZIP
On peut donc décomposer,le modèle sous la forme suivante:

$$\forall i = 1, \cdots, n \qquad \begin{cases} Y_i \sim \pi_i \delta_0 + (1-\pi_i)\mathcal{NB}(\lambda_i, \alpha) \\ logit(\pi) = \gamma^\top W_i \\ \lambda_i = \exp(\beta^\top X_i) \end{cases} \tag{2.11}$$

La vraisemblance du modèle est alors:

$$Ln(\beta,\gamma) = \prod_{i=1}^{n} \left[\pi + (1-\pi)\left(\frac{1}{1+\alpha\lambda_i}\right)^{\frac{1}{\alpha}}\right]^{j_i} \times \left[(1-\pi_i)\left(\frac{\Gamma(y_i+\frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})y_i!}\right)\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right)^{y_i}\left(\frac{1}{1+\alpha\lambda_i}\right)^{\frac{1}{\alpha}}\right]^{1-j_i} \tag{2.12}$$

La log-vraisemblance est alors :

$$\ell_n(\alpha,\beta,\gamma) = \sum_{i=1}^{n} j_i \log\left[\frac{e^{\gamma^\top W_i}}{1+e^{\gamma^\top W_i}} + \frac{1}{1+e^{\gamma^\top W_i}}\frac{1}{(1+\alpha e^{\beta^\top X_i})^{\frac{1}{\alpha}}}\right] + \sum_{i=1}^{n}(1-j_i) ln\left[\frac{1}{1-e^{\gamma^\top W_i}}\frac{\Gamma(y_i+\frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(y_i+1)}\left(\frac{\alpha e^{\beta^\top X_i}}{1+\alpha e^{\beta^\top X_i}}\right)^{y_i}\left(\frac{1}{1+\alpha e^{\beta^\top X_i}}\right)^{\frac{1}{\alpha}}\right] \tag{2.13}$$

$$\begin{aligned} \ell_n(\alpha,\beta,\gamma) &= \sum_{i=1}^{n}\Bigg\{ j_i \log\left(\frac{1}{1+e^{\gamma^\top W_i}}(e^{\gamma^\top W_i} + (1+\alpha e^{\beta^\top X_i})^{-\frac{1}{\alpha}}\right) \\ &\quad + (1-j_i)\log\left(\frac{1}{1+e^{\gamma^\top W_i}}\frac{\Gamma(y_i+\frac{1}{\alpha}}{\Gamma(\frac{1}{\alpha})y!}\left(\frac{\alpha e^{\beta^\top X_i}}{1+\alpha e^{\beta^\top X_i}}\right)^{y_i}\left(\frac{1}{1+\alpha e^{\beta^\top X_i}}\right)^{\frac{1}{\alpha}}\right)\Bigg\} \end{aligned}$$

$$\begin{aligned} \ell_n(\alpha,\beta,\gamma) &= \sum_{i=1}^{n}\Bigg\{ j_i\left(\log(e^{\gamma^\top W_i} + (1+\alpha e^{\beta^\top X_i})^{-\frac{1}{\alpha}} - \log(1+e^{\gamma^\top W_i})\right) \\ &\quad + (1-j_i)\left(-\log(1+e^{\gamma^\top W_i}) + \log(\Gamma(y_i+\frac{1}{\alpha})) - \log(\Gamma(\frac{1}{\alpha})) - \log(y!)\right. \\ &\quad \left. + y_i\log(\alpha e^{\beta^\top X_i}) - y_i\log(1+\alpha e^{\beta^\top X_i}) - \frac{1}{\alpha}\log(1+\alpha e^{\beta^\top X_i})\right)\Bigg\} \end{aligned}$$

$$\begin{aligned}\ell_n(\alpha,\beta,\gamma) &= \sum_{i=1}^{n}\Big\{ j_i\left(\log(e^{\gamma^\top W_i}+(1+\alpha e^{\beta^\top X_i})^{-\frac{1}{\alpha}}-\log(1+e^{\gamma^\top W_i})\right)\\ &\quad +(1-j_i)\left(-\log(1+e^{\gamma^\top W_i})+\log(\Gamma(y_i+\frac{1}{\alpha}))-\log(\Gamma(\frac{1}{\alpha}))\right.\\ &\quad \left.+y_i\log(\alpha e^{\beta^\top X_i})-y_i\log(1+\alpha e^{\beta^\top X_i})-\frac{1}{\alpha}\log(1+\alpha e^{\beta^\top X_i})\right)\Big\}\end{aligned}$$

2.5 Estimation et propriétés Asymptotiques du modèle ZI

Les modèles ZI sont couramment utilisés pour modéliser des variables avec un excès de zéros par rapport à la distribution de Poisson ou binomiale négative. L'estimation des paramètres d'un modèle ZI peut être effectuée à l'aide de diverses méthodes statistiques(D.Lambert1992 [20]),(C.zardo et al. 2007, [8]) telles que la méthode du maximum de vraisemblance (EMV) ou des méthodes d'estimation bayésienne. Le choix de la méthode dépend souvent des caractéristiques des données et des préférences de lánalyste.

2.5.1 Estimation et propriétés Asymptotiques du modèle ZIP

Dans cette étude d'estimation et propriétés du modèle ZIP , nous avons établi notre travail en nous inspirant des travaux de (ALI, 2021[4]) et de (D.Lambert, 1992 [20])
En générale, l'estimation du maximum de vraisemblance est utilisée pour estimer de tels modèles. Supposons que l'on dispose d'un échantillon $(Y_i, X_i, W_i)...(Y_n, X_n, W_n)$ d'observations indépendantes du modèle (2.2). (Pour alléger la notation nous ne ferons pas apparaître le terme d'exploitation). Le paramètre $\theta = (\beta^\top, \gamma^\top)^\top$ du modèle peut être estimé par la méthode du maximum de vraisemblance. La vraisemblance se calcule comme suit :

$$L_n(\theta) = \prod_{i=1}^{n}(\pi_i + (1-\pi_i)e^{-\lambda_i})^{\mathbb{1}\{y_i=0\}}.\big((1-\pi_i)e^{-\lambda_i}\frac{\lambda_i^{y_i}}{y_i!}\big)^{\mathbb{1}\{y_i>0\}}$$

On en deduit la log- vraisemblance $\ell_n(\theta) = \log(L_n(\theta))$:

$$\begin{aligned}\ell_n(\theta) = &\sum_{i=1}^{n} \mathbb{1}\{y_i = 0\} \log(\pi_i + (1-\pi_i)e^{-\lambda_i}) \\ &+ \sum_{i=1}^{n} \mathbb{1}\{y_i > 0\}(\log(1-\pi) - \lambda_i + y_i \log(\lambda_i) - \log(y_i!))\end{aligned} \tag{2.14}$$

Pour mieux estimer nos paramètres nous allons écrire la log-vraisemblance sous sa forme la plus simple et nous utiliserons un algorithme pour estimer les paramètres

$$\begin{aligned}\log(\theta) = &\sum (J_i) \log\left(e^{\gamma^\top W_i} + e^{-\exp(\beta^\top X_i)}\right) \\ &+ (1-J_i)\left(y_i \beta^\top X_i - e^{\beta^\top X_i} - \log(y_i)\right) - \log\left(1 + e^{\gamma^\top W_i}\right)\end{aligned} \tag{2.15}$$

En particulier, supposons que l'on observe la variable indicatrice s telle que $s_i = 1$ si y_i provient de l' ensemble des zéros (distribution dégénérée) et $s_i = 0$ si y_i résulte du zéro aléatoire (distribution non dégénérée). Alors la log-vraisemblance (dit complète)au vu des observations $(Y_1, s_1, X_1, W_1)...(Y_n, s_n, X_1, W_n)$ se calcule comme suit:

$$\begin{aligned}
\ell_n^c(\theta) &= \log\left[\prod_{i=1}^{n} \mathbb{P}(Y_i = 0, s_i = 1|X_i, W_i)^{s_i} \times \mathbb{P}(Y_i = y_i, s_i = 0|X_i, W_i)^{1-s_i}\right] \\
&= \log\left[\prod_{i=1}^{n} \mathbb{P}(Y_i = 0|s_i = 1, X_i, W_i)\mathbb{P}(s_i = 1|X_i, W_i)^{s_i} \times \mathbb{P}(Y_i = y_i|s_i = 0, X_i, W_i)\mathbb{P}(s_i = 1|X_i, W\right. \\
&= \log\left[\prod_{i=1}^{n} \pi_i^{s_i} \times (e^{-\lambda}\frac{\lambda_i^{y_i}}{y_i!})^{1-s_i}(1-\pi_i)^{1-s_i}\right] \\
&= \sum_{i=1}^{n} [s_i \log \pi_i + (1-s_i)\log(1-\pi_i) - (1-s_i)\lambda_i + y_i(1-s_i)\log\lambda_i - (1-s_i)\log(Y_i!)] \\
&= \sum_{i=1}^{n} \left[s_i \gamma^\top W_i - \log(1 + \exp(\gamma^\top W_i)\right] + \sum_{i=1}^{n}(1-s_i)\left[Y_i \xi^\top X_i - \exp(\beta^\top X_i) - \log(Y_i!)\right].
\end{aligned}$$

On remarque que la log-vraisemblance se décompose en deux termes l'une dépendant uniquement de γ et l'autre de β uniquement. Nous pouvons donc écrire la log-vraisemblance complète comme suit:

$$\ell_n^c(\theta) = \ell_\gamma^c(\gamma, y, s) + \ell_\beta^c(\beta, y, s) \tag{2.16}$$

Sous la forme (2.16) la log-vraisemblance peut-être maximiser soit suivant le paramètre γ ou β.

Pour approcher $\hat{\theta_n}$ l'auteur propose d'utiliser l'algorithme EM (Dempster et al. 1977, [10]). Cet Algorithme consiste à maximiser la fonction $\ell_n^c(\theta)$ de manière itérative en commençant par des valeurs initiales $(\theta^{(0)\top}, \beta^{(0)\top})^\top$ et en alternant les étapes 1 et 2 suivant.

– L'étape E: Elle consiste à calculer l'espérance conditionnelle $\mathbb{E}(\ell_n^c(\theta)|\mathcal{O}_n;\theta^{(k)})$ de la log-vraisemblance complété sachant les observations sous la loi du paramètre $\theta^{(k)}$ (où $\theta^{(k)}$ désigne l'estimation de θ obtenu à la *k-ième* itération de l'algorithme),

$$\begin{aligned}\mathbb{E}(\ell_n^c(\theta)|\mathcal{O}_n;\psi^{(k)}) = &\sum_{i=1}^{n}\left[\mathbb{E}(s_i|\mathcal{O}_n,\theta^{(k)})\gamma^\top W_i - \ell_n(1+\exp(\gamma^\top W_i)\right] \\ &+\sum_{i=1}^{n}(1-\mathbb{E}(s_i|\mathcal{O}_n,\theta^{(k)}))\left[Y_i\beta^\top X_i - \exp(\beta^\top X_i) - \ell_n(Y_i!)\right]\end{aligned} \tag{2.17}$$

– L'étape M: On cherche la valeur de :
$\theta^{(k+1)}$ =$argmax_\theta\mathbb{E}(\ell_n^c(\theta)|\mathcal{O}_n;\theta^{(k)})$ ie trouver $\gamma^{(k+1)}$ et $\beta^{(k+1)}$ en maximisant respectivement $\ell_\gamma^c(\gamma,y,s^{(k)})$ et $\ell_\theta^c(\beta,y,s^{(k)})$.
(Hall et schen, 2010 [15]) ont montré que maximiser ces deux fonctions revient a résoudre les deux équations suivantes

$$\frac{1}{n}\sum_{i=1}^{n}\{s^r-\pi_i\}W_i = 0 \tag{2.18}$$

$$\frac{1}{n}\sum_{i=1}^{n}(1-s^r)\{y_i - e^{(\beta^\top X_i)}\}X_i = 0 \tag{2.19}$$

compte tenu de la sensibiliser de EMV en présence des données aberrantes (Hall et shen [15]) proposons une approche alternative.
Dans l'approche RES, (Hall et Shen 2010) proposent de remplacer les équations (2.18) et (2.19) par des estimations de 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 des conditions de régularité de (Rosen et al2000[32]) liées à l'algorithme ES et de (Carroll et al. 1995, [6]), (Hall & Shen, 2010,[6]) sous les hypothèses suivantes.

– H_1 : $\frac{n}{\lambda_{min}(V_n)} \leq C_1 \forall n \geq 1$ où C_1 est une constante positive V_n la matrice d'information de Fischer et λ_{min} sa plus petite valeur propres .

– H_2 les variables explicatives sont uniformément bornées ie $\exists C_2 < \infty : \|X\| < C_2$

– H_3 soit $\mathcal{B}$ un ensemble ouvert de $\mathbb{R}^p$ et $\theta_0 = (\beta_0^\top, \pi_0)^\top$ la vraie valeur de $\theta = (\beta^\top, \pi)^\top$ un point intérieure de $\mathcal{B}\times[0,1]$.
Ont montré le résultat suivant plus général (dans le sens où $\theta = (\beta^\top, \gamma^\top)^\top \in \mathbb{R}^{p+q}$ ce qui se résume dans le théorème suivant:

Théorème 2.1 *Si l'algorithme RES converge, alors il existe une suite de variables aléatoires $\hat{\theta}$ telles que:*
Il existe une suite de variables aléatoire $\hat{\theta}_n$ telle que :

1. $\mathbb{P}(s_n(\hat{\theta}) = 0) \longrightarrow 1$ *quand* $n \to \infty$*(existence asymptotique)*
2. $\hat{\theta} \xrightarrow{\mathbb{P}} \theta_0$ *quant* $n \to \infty$ *(Consistance)*
3. $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbf{V}(\theta_0))$ *quand* $n \to \infty$ *(Normalité Asymptotique) où* $s_n(\theta) = \frac{\partial \ell_n(L_n)}{\partial \theta}$ *et* $V = -\mathbb{E}\big[\frac{\partial^2 \ell_n(L_n(\theta))}{\partial \theta \partial \theta^\top}\big]$
avec $\ell_n(L_n)$ la log-vraisemblance du modèle ZIP.

où l'expression $V(\theta_0)$ de la variance asymptotique est donnée dans (Hall et Shen 2010,[15]).

2.5.2 Algorithme de Newton-Raphson

L'algorithme de Newton-Raphson résout par itération des équations non linéaires. Par exemple, pour déterminer le point où une fonction prend son maximum, on se donne une valeur initiale puis on obtient une seconde valeur en approchant la fonction à maximiser dans le voisinage de la valeur initiale par un polynôme du second degré puis en trouvant la valeur maximisant ce polynôme. Cela fait appel à la matrice hessienne, matrice des dérivées secondes de la log-vraisemblance(Lange 2004[21]). Puis on réitère le même procédé en approchant la fonction à maximiser dans le voisinage de la seconde valeur obtenue et ainsi de suite, jusqu'à ce qu'un critère de convergence soit satisfait.
Recherchant la solution de l'équation de vraisemblance qui est telle que:

$$\frac{\partial \ell_n(\beta)}{\partial \beta} = \frac{\partial \ell_n(\beta)}{\partial \beta}\Big|_{\beta - \hat{\beta}_n} = 0$$

l'algorithme de Newton-Raphson se présente comme suit:

$$\frac{\partial \ell_n(\beta^{(k+1)})}{\partial \beta} = \frac{\partial \ell_n(\beta^k)}{\partial \beta} + \frac{\partial^2 \ell_n(\beta^k)}{\partial \beta \partial \beta^\top}(\beta^{k+1} - \beta^k)$$

On obtient:

$$\beta^{k+1} = \beta^k - \left(\frac{\partial^2 \ell_n(\beta^k)}{\partial \beta \partial \beta^\top}\right)^{-1} \frac{\partial \ell_n(\beta^k)}{\partial \beta} \tag{2.20}$$

Partant d'une valeur initialeβ_0, on réitère la formule 2.20 jusqu'à ce qu'un critère de convergence soit satisfait (de la stabilité) par exemple la norme $\left\|\beta^{k+1} - \beta^k\right\|$ de

la différence entre deux approximations successives devient plus petite qu'un seuil $\varepsilon > 0$ fixé d'avance).

2.6 Modèles de régression ZIPLG

La loi log-gamma est une distribution de probabilité d'une variable dont le logarithme suit une distribution gamma. Elle est utiliser pour modéliser des variables aléatoir qui ont une asymétrie positive et une queue de distribution longue. ont définit la loi log-gamma comme suit [35]

$$f(x) = \frac{\exp(\beta x)\exp(-\exp(\frac{x}{\alpha}))}{\alpha^{\beta}\Gamma(\beta)} \tag{2.21}$$

où x est la variable aléatoire , α est le paramètre de forme et β le paramètre d'echelle.
On écrit alors $x \sim log - gamma(\alpha, \beta)$
en se basant sur l'équation 2.21 on peut définir le modèle poisson log-gamma.
La distribution de Poisson log-gamma est une distribution de probabilité qui combine la distribution de Poisson et la distribution log-gamma. Elle est utilisée pour modéliser des phénomènes où les événements sont rares et peuvent varier dans le temps .elle est utilisée lorsque la moyenne d'une distribution de poisson est également incertaine et peut varier.Elle permet de prendre en compte l'incertitude de la moyenne et ajuster la distribution de poisson en conséquence.
Soit Y une variable aléatoire tels que

$$\begin{cases} Y \sim \mathcal{P}(\lambda) \\ \lambda \sim \log - gamma(\alpha, \beta) \end{cases} \tag{2.22}$$

alors sa densité de probabilité est définie par :

$$\begin{aligned} \mathbb{P}(y, \alpha, \beta) &= \int_0^{\infty} \frac{\lambda^y \exp(-\lambda)}{y!} \frac{\exp(\beta\lambda)\exp(-\exp(\frac{\lambda}{\alpha}))}{\alpha^{\beta}\Gamma(\beta)} d\lambda \\ &= \frac{1}{y!\alpha^{\beta}\Gamma(\beta)} \int_0^{\infty} \lambda^y \exp(-\lambda)\exp(\beta\lambda)\exp(-\exp(\frac{\lambda}{\alpha}))d\lambda \\ &= \frac{1}{y!\alpha^{\beta}\Gamma(\beta)} \int_0^{\infty} \lambda^y \exp\left((\beta-1)\lambda - \exp\left(\frac{\lambda}{\alpha}\right)\right) d\lambda \end{aligned}$$

Le terme $\exp(-\exp(\frac{\lambda}{\alpha}))$ est difficile à intégrer directement, nous allons donc approximer l'intégrale à l'aide d'une méthode d'intégration numérique, par exemple, la méthode de Simpson.

L'integrale devient alors:

$$\frac{1}{y!\alpha^{\beta}\Gamma(\beta)}\int_0^{\infty}\lambda^{y}\exp\left((\beta-1)\lambda-\exp\left(\frac{\lambda}{\alpha}\right)\right)d\lambda\approx\frac{1}{y!\alpha^{\beta}\Gamma(\beta)}\sum_{i=0}^{m}w_i\cdot f(\lambda_i) \tag{2.23}$$

ou n est le nombre d'intervalles pour la méthode de Simpson, w_i sont les poids correspondants et λ_i sont les points d'évaluation.
Pour chaque point d'évaluation λ_i, nous évaluons la fonction.
$f(\lambda)=\lambda^{y}\exp\left((\beta-1)\lambda-\exp\left(\frac{\lambda}{\alpha}\right)\right)$. Cela peut être fait en utilisant une méthode d'approximation numérique telle que la méthode des différences finies ou la méthode de Newton-Cotes. Par exemple, avec la méthode des différences finies,posons:

$$f(\lambda_i)\approx\lambda_i^{y}\exp\left((\beta-1)\lambda_i-\exp\left(\frac{\lambda_i}{\alpha}\right)\right)$$

En utilisant les valeurs approchées de $f(\lambda_i)$, nous pouvons calculer l'intégrale approximée :

$$\frac{1}{y!\alpha^{\beta}\Gamma(\beta)}\sum_{i=0}^{m}w_i\cdot f(\lambda_i) \tag{2.24}$$

Cela nous donne une approximation numérique de l'intégrale donnée.

La forme généle des modéle ZI est definit par:

$$\mathbf{P}(Y_i=y_i)=\begin{cases}\pi_i+(1-\pi_i)\mathbb{P}(y_i=0,\lambda_i) & \text{si } y_i=0\\ (1-\pi_i)\mathbb{P}(y_i,\lambda_i) & \text{si } y_i>0\end{cases} \tag{2.25}$$

qui se compose d'une distribution dégénérée à zéro et d'une distribution de comptage non tronquée avec un vecteur de paramètres λ_i. Si la distribution de comptage suit une distribution de Poisson, le moèle Zéro Inflated poisson Log-gamma est alors définit par :

$$\mathbb{P}(Y_i=y_i)=\begin{cases}\pi_i+(1-\pi_i)\int_0^{\infty}\exp(-\lambda_i)\frac{\exp(\beta\lambda_i)\exp(-\exp(\frac{\lambda_i}{\alpha}))}{\alpha^{\beta}\Gamma(\beta)}d\lambda_i & \text{si } y_i=0\\ (1-\pi_i)\int_0^{\infty}\frac{\lambda_i^{y}\exp(-\lambda_i)}{y!}\frac{\exp(\beta\lambda_i)\exp(-\exp(\frac{\lambda_i}{\alpha}))}{\alpha^{\beta}\Gamma(\beta)}d\lambda_i & \text{si } y_i>0\end{cases} \tag{2.26}$$

où y_i est une variable aléatoire qui suit une distribution zi poisson log-gamma π_i est la probabilité d'observé des zéros , λ_i represente la moyenne de la distribution de poisson α est un paramètre de forme et β un paramètre d'échelle de distribution

log-gamma
La vraisemblance du modèle est alors donner par:

pour faciliter la lecture posons: $w_i \cdot f(\lambda_i) = \Delta_i$

$$L_n(\gamma, \alpha, \beta) = \prod_{i=1}^{n} (\frac{1}{1 + e^{\gamma^\top W_i}} (e^{\gamma^\top W_i} + \frac{\sum_{i=1}^{m} \Delta_i}{\alpha^\beta \Gamma(\beta)})^{j_i} \times (\frac{\sum_{i=1}^{m} \Delta_i}{(1 + e^{\gamma^\top W_i})(y_i! \alpha^\beta \Gamma(\beta))})^{1-j_i})$$

$$L_n(\gamma, \alpha, \beta) = \sum_{i=1}^{n} j_i \log((\frac{1}{1 + e^{\gamma^\top W_i}} (e^{\gamma^\top W_i} + \frac{\sum_{i=1}^{m} \Delta_i}{\alpha^\beta \Gamma(\beta)}))) + (1 - j_i) \log((\frac{\sum_{i=1}^{m} \Delta_i}{(1 + e^{\gamma^\top W_i})(y_i! \alpha^\beta \Gamma(\beta))}))$$

$$\begin{aligned} \ell_n(\gamma, \alpha, \beta) \quad = \quad & \sum_{i=1}^{n} \Bigg\{ j_i \log(e^{\gamma^\top W_i} + \frac{\sum_{i=1}^{m} \Delta_i}{\alpha^\beta \Gamma(\beta)}) \\ & + (1 - j_i) \left[\log(\sum_{i=1}^{m} \Delta_i) - \log(1 + e^{\gamma^\top W_i}) - \log(y_i!) - \beta \log(\alpha) - \log(\Gamma(\beta) \right] \Bigg\}. \end{aligned}$$

3

Une étude basée sur la simulation de la régression avec divers sous-modèles à inflation de zéros

Résumé

Dans ce chapitre l'étude propose une approche basée sur la simulation de la régression ZIP pour examiner différents sous-modèles à inflation de zéros. L'objectif est de comprendre comment ces sous-modèles peuvent être utilisés pour capturer et modéliser efficacement les caractéristiques des données présentant une surabondance de zéros.
Les données simulées serons ensuite utilisées pour ajuster différents sous-modèles à inflation de zéros, afin de voir comment ces modèles se comportent et comment ils peuvent être appliqués à des ensembles de données réelles.
L'étude vise à évaluer les performances des différents sous-modèles ZIP en termes de capacité à estimer les paramètres réels des données, à capturer la surabondance de zéros et à prédire les valeurs non nulles.
Ces simulations vont être utilisées pour examiner les performances (biais, erreur quadratique moyenne, probabilités de couverture et calculs de l'erreur standard) de l'EMV.

Sommaire

3.1 Études de simulations

Dans cette partie , nous évaluons la performance de EMV en échantillon fini par le biais d'expériences de Newton-Raphson
Soit Y_i une variable de comptage d'un événement $\forall i(i = 1; \cdots; n)$ et soit $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^\top$ et $W = (wi1, wi2 \cdots; wiq)^\top$ des vecteurs de covariables de dimension respectif p et q on definit le modèle ZIP par :

$$Y \sim \begin{cases} 0 & \text{avec une probabilité } p_i \\ \mathcal{P}(\lambda_i) & \text{avec une probabilité } 1 - p_i \end{cases} \tag{3.1}$$

où$0 < \pi_i < 1$ est la probabilité d'inflation zéros , on peut se placer dans le cas d'une sur ou sou- dispersion (Conssul et Famoye1992[30]).
La densité de probabilié de la fonction ZIP est donnée par :

$$P(Y_i = y_i) = \begin{cases} \pi + (1 - \pi)\exp(-\lambda_i) & \text{si } y_i = 0 \\ (1 - \pi)\frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} & \text{si } y_i > 0 \end{cases} \tag{3.2}$$

L'espérance et la variance sont respectivement donné par : $\mathbb{E}(Y_i) = (1 - \pi)\lambda_i$ et $\mathbb{V}ar(Y_i) = (1 - \pi)(1 + \pi\lambda_i)\lambda_i$
Notons également que la sur-dispersion peut-être causer par une abondance de valeurs nulles ou une hétérogénéité non observer.
Les paramètres dans ce modèle sont estimer par la régression logistique suivante:

$logit(\pi_i) = \gamma^\top W_i$ et $\lambda_i = e^{\beta^\top X_i}$ où $\beta = (\beta_1; \cdots, \beta_p)^\top \in \mathbb{R}^p$ et $\gamma = (\gamma_1; \cdots, \gamma_p)^\top \in \mathbb{R}^q$ qui sont des paramètres inconnus à estimer. suposons que nous observons n valeurs indépendante $(Y_i, X_i, W_i)_{i=1;\cdots;n}$ soit $\theta = (\beta^\top, \gamma^\top)^\top$, le paramètre à estimer, alors la vraisemblance du modèle est donné par:
$L_n(\theta) = \prod_{i=1}^n \mathbb{P}(Y_i = yy_i|X_i, W_i)^{j_i}\mathbb{P}(Y_i > 0y_i|X_i, W_i)^{1-j}$
la log vraisemblance est alors définit par:

$$\begin{aligned} \log(\psi) = &\sum (J_i) \log\left(e^{\gamma^\top W_i} + e^{-\exp(\xi^\top X_i)}\right) \\ &+ (1 - J_i)\left(y_i\beta^\top X_i - e^{\beta^\top X_i} - \log(y_i)\right) - \log\left(1 + e^{\gamma^\top W_i}\right) \end{aligned} \tag{3.3}$$

EMV $\hat{\theta} = (\hat{\beta}_n^\top, \hat{\gamma}_n^\top)^\top$ est obtenu en résolvant l'équation du score $\partial\ell_n(\theta)/\partial\theta = 0$. Cette equation peut -être résolut par une optimisation.

Dans notre travail,toutes les estimation sont obtenu en utilisant la fonction "MaxLink" de R(HENNINGSEN & TOOMET 2011, [17]) qui implémente l'algorithme de Newton. cette fonction nous donne la matrice hessienne de ℓ_n qui est très importante dans l'estimation de la variance de EMV. La matrice de variance covariance de $\hat{\theta}$ est donné par $\hat{\sum}_n = \left[-\partial\ell_n(\hat{\theta})/\partial\theta\partial\theta^\top\right]^{-1}$. L'erreur standard du paramètre estimer est obtenu en calculant les racine carre des éléments de la diagonale de la matrice des variance-covariance. Laissant de côté la théorie distributionnelle, nous proposons d'étudier ces propriéés au moyen de simulation.

3.1.1 Expériences numériques par simulation pour le modèle ZIP

Nous génerons les données à partir du modèle de régression ZIP suivant

$$logit(\pi_i) = \beta_1 X_1 + \cdots + \beta_6 X_6$$
$$logit(\lambda_i) = \gamma_1 W_1 + \cdots + \gamma_4 W_4$$

où $X_{i1} = 1$ et $X_{i2}, X_{i3}, X_{i4}, X_{i5}$ sont des covariables indépendante de loi $\mathbb{N}(n,0,1), \mathbb{B}(n,1,0.4), \exp(n), \mathbb{U}(n,-2,2), \mathbb{N}(n,-1,1)$ respectivement, Nous laissons $W_{i1} = 1$ et W_{i2}, W_{i3}, W_{i4} et de distribution $\exp(n)\mathbb{B}(n,1,0.8)\mathbb{U}(n,2,5)$ respectivement. Les prédicteurs $logit(\pi_i)$ et $logit(\lambda_i)$ peuvent partager quelques termes en communs en laissant$W_{i2} = X_{i3}$.
Le paramètre de régression β est choisi comme$\beta = (-0.5, 0.75, -1.2, 0.5, -0.25, 0)^\intercal$ le paramètre γ est choisi comme suit:

cas 1 $\gamma = (-0.7, -1.2, -0.8, 0)^\intercal$
cas 2 $\gamma = (1.2, 0.45, -0.9, -0.7)^\intercal$
cas 3 $\gamma = (-0.15, -0.25, 0.5, 0.75)^\intercal$
cas 4 $\gamma = (-.7, 1, -0.8, 0.9)^\intercal$

On considère quelque échantillons de taille $n = 500, 1000, 2000$ pour chaque échantillons nous allons simuler des proportions de $10\%, 20\%, 50\%, 90\%$ d'inflation de zéros.
Le modèle ZINB est aussi définit comme le modèle ZIP. En cosiderant (3.1) si nous remplaçons la distribution de poisson par la loi binomiale négative($\mathcal{B}(n,\lambda)$) on obtien un noubeau modèl , le ZINB , sa densité de probabilité est alors donné par :

$$P(Y_i = y_i) = \begin{cases} \pi + (1-\pi)(\frac{1}{1+\alpha\lambda_i})^{\frac{1}{\alpha}} & \text{si } y_i = 0 \\ (1-\pi)\frac{\Gamma(y_i+\frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})y_i!}(\frac{\alpha\lambda_i)}{1+\alpha\lambda_i})^{y_i}(\frac{1}{1+\alpha\lambda_i})^{\frac{1}{\alpha}} & \text{si } y_i > 0 \end{cases} \tag{3.4}$$

avec π_i la probabilité d'inflation de zéros , $\lambda_i > 0$,,,, et α un réel positif qui est le paramètre de dispersion, $\mathbb{E}(Y_i) = (1-\pi_i)\lambda_i$ et $\mathbb{V}ar(Y_i) = (1-\pi_i)\lambda_i(1+(\alpha\pi_i)\lambda_i)$ sont

respectivement l'espérance et la variance du modèle ZINB. On modélise la probabilité du risque par :$logit(\pi_i) = \gamma^\top W_i$ et le paramètre $\lambda_i = e^{\beta^\top X_i}$ où $\beta \in \mathbb{R}^p$ et $\gamma^\top \in \mathbb{R}^q$ sont des paramètres inconnus à estimer.
Alors EMV $\theta = (\beta^\top, \gamma^\top, \alpha)^\top$ est aussi calculer comme dans le cas ZIP. La log-vraisemblance $\ell_n(\theta)$ est donné par:

$$\ell_n(\alpha,\beta,\gamma) = \sum_{i=1}^{n} j_i \log\left[\frac{e^{\gamma^\top W_i}}{1+e^{\gamma^\top W_i}} + \frac{1}{1+e^{\gamma^\top W_i}}\frac{1}{(1+\alpha e^{\beta^\top X_i})^{\frac{1}{\alpha}}}\right] + \sum_{i=1}^{n}(1-j_i) ln\left[\frac{1}{1-e^{\gamma^\top W_i}}\frac{\Gamma(y_i+\frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(y_i+1)}\left(\frac{\alpha e^{\beta^\top X_i}}{1+\alpha e^{\beta^\top X_i}}\right)^{y_i}\left(\frac{1}{1+\alpha e^{\beta^\top} X_i}\right)^{\frac{1}{\alpha}}\right] \quad (3.5)$$

L'EMV , $\hat{\theta}_n = (\hat{\beta}_n^\top, \hat{\gamma}_n^\top, \hat{\alpha}_n)^\top$ est obtenu en resolvant l'équation du score $\partial\ell_n(\theta)/\partial\theta = 0$ qui nécessite une optimisation numérique. Les propriétés de l'EMV sont étudier par des simulations dans la section suivante.Comme pour le modèle ZIP, nous calculons l'erreur standard sous la forme $\sqrt{diag(\hat{\sum}_n)}$ où $\hat{\sum}_n = \left[-\partial\ell_n(\hat{\theta})/\partial\theta\partial\theta\right]^{-1}$

L'optimisation numerique est obtenu via" maxLiK". Les valeurs du départ pour tout les paramètres du modèle sont obtenu en ajustant le modèle ZINB au donnés simuler avec "Zeroinfl".

3.2 Expériences numériques par simulation du modèle ZINB

On simule des donnés qui suivent une distribution binomiale négative et

$$logit(\lambda_i) = \beta_1 X_{i1} + \cdots + \beta_1 X_{i4}$$

$$logit(\pi_i) = \gamma_1 W + \gamma_2 W + \cdots + \gamma_5 W$$

où $X_{i1} = W_{i1} = 1$ et $X_{i1}, \cdots, X_{i6}, W_{i4}X_{i5}$ suivent des lois normale $\mathcal{N}(0,1)$,Bernoulli $\mathcal{B}(0.3)$, normale $\mathcal{N}(1, 2.25)$, exponentiel $\mathcal{E}(1)$,uniforme $\mathcal{U}(2,5)$,normale $\mathcal{N}(-1,1)$ et Bernoulli $\mathcal{B}(0.5)$ respectivement comme dans le modèle ZIP, $logit(\lambda_i)$ et $logit(\pi_i)$ peuvent peuvent partager les même covariable dans notre cas on choisie $W_{i2} = X_{i2}$ et on donne $\beta = (0.7, 0.1, 0.4, 0.85, -0.5, 0)^\top$ et $\gamma = (-0.9, -0.65, -0.2, 0.65, 0)^\top$ et nous fixons le paramètre de dispersion $\alpha = 0.5$ et une proportion de zéro à 20%. les tableaux de l'estimation du modèle ZINB fournissent, les même mesures récapitulatives que pour le moèle ZIP.

3.2.1 Description

Dans cette partie nous avons crée des données simulées à l'aide de la fonction "rzip" et "rnegbin" de R respectivement pour les modèles ZIP et ZINB qui permettent de simuler des données à partir d'une distribution zero inflated poisson et binomiale. Ensuite nous avons défini une fonction "MaxLik " qui prend en entrées les paramètres de la distribution et les données et calculer la log-vraisemblance de la distribution pour ces paramètres. Puis nous estimons le patamètre θ

Pour la simulation on effectue 1000 réplications et pour chaque taille n d'échantillons plusieurs mesures sommaires ont été obtenues, plus précisément le biais relatif moyen, l'erreur type moyenne, l'écart type empirique, la longueur de l'intervalle de confiance, l'erreur quadratique moyenne et la probabilité de couverture empirique correspondante pour chaque paramètre dans le modèle (on considére des intervalles de confiance de type Wald à 95%)

c		$\hat{\beta}_n$						$\hat{\gamma}_n$			
		$\hat{\beta}_{1,n}$	$\hat{\beta}_{2,n}$	$\hat{\beta}_{3,n}$	$\hat{\beta}_{4,n}$	$\hat{\beta}_{5,n}$	$\hat{\beta}_{6,n}$	$\hat{\gamma}_{1,n}$	$\hat{\gamma}_{2,n}$	$\hat{\gamma}_{3,n}$	$\hat{\gamma}_{4,n}$
10%											
	bias	-0.0020	-0.0039	0.0006	0.0008	-0.0012	-0.0013	-0.3561	-1.8745	-0.5260	0.7402
	SD	0.1011	0.0461	0.1151	0.0306	0.0412	0.0459	59.7064	23.8877	10.0656	12.4745
	SE	0.1054	0.0466	0.1195	0.0308	0.0403	0.0458	5.9682	2.4795	4.2714	1.4395
	RMSE	0.1460	0.0656	0.1659	0.0434	0.0576	0.0648	60.0638	24.0756	10.9418	12.5729
	CP	0.9532	0.9623	0.9646	0.9463	0.9509	0.9395	0.9315	0.9144	0.9017	0.9532
	ℓ(CI)	0.4094	0.1816	0.4674	0.1170	0.1577	0.1789	11.3796	5.0704	6.2433	2.7711
20%											
	bias	-0.0135	0.0017	-0.0034	-0.0004	-0.0018	0.0000	0.0697	0.0471	-0.0526	-0.0597
	SD	0.1193	0.0565	0.1353	0.0429	0.0460	0.0548	1.3348	0.2131	0.8166	0.3654
	SE	0.1187	0.0547	0.1417	0.0440	0.0471	0.0531	1.1058	0.2035	0.5763	0.3341
	RMSE	0.1688	0.0786	0.1959	0.0615	0.0659	0.0763	1.7342	0.2983	1.0005	0.4986
	CP	0.9519	0.9459	0.9649	0.9539	0.9509	0.9429	0.9549	0.9639	0.9529	0.9519
	ℓ(CI)	0.4639	0.2133	0.5539	0.1686	0.1844	0.2076	4.1894	0.7721	2.1172	1.2430
50%											
	bias	-0.2121	0.0439	-0.1284	0.0357	-0.0098	0.0018	-0.2226	0.0233	-0.0286	0.0623
	SD	0.5716	0.2553	0.6068	0.2198	0.2088	0.2419	1.5692	0.2593	0.8435	0.3747
	SE	0.5190	0.2364	0.5540	0.1822	0.1908	0.2233	1.3317	0.2363	0.6229	0.3481
	RMSE	0.8004	0.3506	0.8314	0.2877	0.2829	0.3291	2.0695	0.3515	1.0486	0.5151
	CP	0.9308	0.9298	0.9418	0.9268	0.9308	0.9358	0.9569	0.9639	0.9649	0.9539
	ℓ(CI)	1.9445	0.8797	2.0784	0.6137	0.7154	0.8290	4.7805	0.8383	2.3356	1.2806
90%											
	bias	-0.2660	0.0499	-0.1721	0.0602	-0.0050	0.0193	-0.4114	0.2176	-0.6385	0.2499
	SD	0.7129	0.3029	0.9024	0.8065	0.2311	0.2748	10.6327	1.3426	5.7731	4.7024
	SE	0.5923	0.2723	0.9247	0.6364	0.2158	0.2521	2.2742	0.7343	1.7392	0.6108
	RMSE	0.9640	0.4103	1.3031	1.0288	0.3162	0.3733	10.8754	1.5451	6.0602	4.7460
	CP	0.9021	0.9255	0.9521	0.9106	0.9426	0.9266	0.9596	0.9543	0.9606	0.9596
	ℓ(CI)	2.2288	1.0210	2.5128	2.1668	0.8220	0.9513	6.8057	2.4953	4.2704	1.5204

Tableau 3.1 – Résultats de la simulation pour $n = 500$. c : proportion moyenne d'inflation de zéro. SD : écart-type empirique. SE : erreur type moyenne. CP : probabilité de couverture empirique des intervalles de confiance à 95 %. ℓ(CI) : longueur moyenne des intervalles de confiance.

c		$\hat{\beta}_n$						$\hat{\gamma}_n$			
		$\hat{\beta}_{1,n}$	$\hat{\beta}_{2,n}$	$\hat{\beta}_{3,n}$	$\hat{\beta}_{4,n}$	$\hat{\beta}_{5,n}$	$\hat{\beta}_{6,n}$	$\hat{\gamma}_{1,n}$	$\hat{\gamma}_{2,n}$	$\hat{\gamma}_{3,n}$	$\hat{\gamma}_{4,n}$
10%											
	bias	-0.0053	0.0011	-0.0060	0.0014	-0.0009	0.0007	-0.1306	-0.3268	-0.1297	0.0023
	SD	0.0730	0.0320	0.0873	0.0214	0.0284	0.0304	2.5908	1.3897	1.5650	0.6882
	SE	0.0723	0.0322	0.0836	0.0201	0.0279	0.0315	2.5427	1.1510	1.1173	0.6355
	RMSE	0.1028	0.0454	0.1210	0.0293	0.0399	0.0438	3.6315	1.8333	1.9266	0.9365
	CP	0.9472	0.9545	0.9389	0.9275	0.9482	0.9534	0.9638	0.9306	0.9814	0.9669
	ℓ(CI)	0.2823	0.1259	0.3273	0.0770	0.1094	0.1232	6.5308	3.0327	3.1270	1.6763
20%											
	bias	-0.0075	0.0021	-0.0045	0.0004	0.0003	-0.0004	1.9278	1.6744	-0.1180	-0.7309
	SD	0.0864	0.0373	0.0987	0.0300	0.0311	0.0365	0.7138	0.1336	0.3599	0.2149
	SE	0.0819	0.0378	0.0989	0.0290	0.0327	0.0368	0.7156	0.1306	0.3590	0.2123
	RMSE	0.1193	0.0531	0.1397	0.0418	0.0451	0.0518	2.1765	1.6848	0.5218	0.7909
	CP	0.936	0.955	0.939	0.942	0.964	0.962	0.206	0.000	0.944	0.028
	ℓ(CI)	0.3205	0.1476	0.3870	0.1120	0.1279	0.1440	2.7874	0.5073	1.3989	0.8208
50%											
	bias	-0.0752	0.0128	-0.0543	0.0120	-0.0038	0.0048	-0.1111	0.0111	0.0043	0.0298
	SD	0.2946	0.1333	0.3535	0.0859	0.1122	0.1289	0.7864	0.1388	0.4034	0.2244
	SE	0.2955	0.1342	0.3295	0.0823	0.1100	0.1248	0.8004	0.1335	0.3967	0.2188
	RMSE	0.4239	0.1895	0.4862	0.1195	0.1571	0.1794	1.1273	0.1928	0.5656	0.3148
	CP	0.953	0.949	0.944	0.942	0.952	0.937	0.956	0.945	0.947	0.951
	ℓ(CI)	1.1390	0.5137	1.2668	0.2998	0.4234	0.4779	3.1260	0.5183	1.5444	0.8534
90%											
	bias	-0.0948	0.0208	-0.0924	-0.0026	-0.0101	0.0022	-0.0765	0.0617	-0.1385	0.0444
	SD	0.3816	0.1711	0.4427	0.3442	0.1362	0.1615	1.3183	0.3929	0.9840	0.2630
	SE	0.3562	0.1610	0.4048	0.3155	0.1338	0.1525	1.0877	0.3766	0.7414	0.2471
	RMSE	0.5304	0.2358	0.6068	0.4668	0.1911	0.2221	1.7103	0.5476	1.2394	0.3635
	CP	0.9476	0.9436	0.9466	0.9436	0.9547	0.9335	0.9436	0.9396	0.9597	0.9517
	ℓ(CI)	1.3717	0.6203	1.5584	1.1297	0.5180	0.5879	3.9787	1.3991	2.4168	0.9599

Tableau 3.2 – Résultats de la simulation pour $n = 1000$. c : proportion moyenne d'inflation de zéro. SD : écart-type empirique. SE : erreur type moyenne. CP : probabilité de couverture empirique des intervalles de confiance à 95%. ℓ(CI) : longueur moyenne des intervalles de confiance.

c		$\hat{\beta}_n$						$\hat{\gamma}_n$			
		$\hat{\beta}_{1,n}$	$\hat{\beta}_{2,n}$	$\hat{\beta}_{3,n}$	$\hat{\beta}_{4,n}$	$\hat{\beta}_{5,n}$	$\hat{\beta}_{6,n}$	$\hat{\gamma}_{1,n}$	$\hat{\gamma}_{2,n}$	$\hat{\gamma}_{3,n}$	$\hat{\gamma}_{4,n}$
10%											
	bias	-0.0043	0.0010	-0.0027	0.0004	-0.0010	-0.0013	-0.0393	-0.1253	-0.0234	0.0109
	SD	0.0500	0.0220	0.0584	0.0132	0.0198	0.0220	0.9521	0.5382	0.4486	0.2420
	SE	0.0500	0.0223	0.0585	0.0133	0.0196	0.0220	0.9928	0.5743	0.4671	0.2516
	RMSE	0.0708	0.0314	0.0827	0.0187	0.0278	0.0312	1.3758	0.7968	0.6479	0.3492
	CP	0.9499	0.9499	0.9549	0.9469	0.9399	0.9469	0.9629	0.9459	0.9719	0.9709
	ℓ(CI)	0.1957	0.0874	0.2293	0.0510	0.0766	0.0863	3.7188	1.7962	1.7312	0.9501
20%											
	bias	-0.0031	0.0019	-0.0010	-0.0002	0.0003	0.0006	0.0087	0.0038	0.0052	-0.0131
	SD	0.0564	0.0266	0.0694	0.0182	0.0219	0.0262	0.4714	0.0873	0.2448	0.1433
	SE	0.0569	0.0262	0.0692	0.0192	0.0228	0.0257	0.4897	0.0882	0.2458	0.1438
	RMSE	0.0802	0.0374	0.0980	0.0265	0.0316	0.0367	0.6796	0.1241	0.3469	0.2033
	CP	0.956	0.940	0.950	0.961	0.957	0.952	0.955	0.954	0.950	0.962
	ℓ(CI)	0.2226	0.1026	0.2711	0.0743	0.0894	0.1005	1.9152	0.3445	0.9614	0.5607
50%											
	bias	-0.0321	0.0063	-0.0280	0.0032	-0.0055	-0.0005	-0.0457	0.0047	-0.0036	0.0138
	SD	0.1891	0.0877	0.2231	0.0453	0.0708	0.0790	0.5651	0.0929	0.2777	0.1593
	SE	0.1888	0.0846	0.2150	0.0459	0.0700	0.0794	0.5525	0.0904	0.2737	0.1515
	RMSE	0.2691	0.1220	0.3110	0.0646	0.0998	0.1119	0.7914	0.1297	0.3898	0.2203
	CP	0.950	0.948	0.945	0.951	0.945	0.942	0.949	0.957	0.944	0.939
	ℓ(CI)	0.7333	0.3273	0.8334	0.1714	0.2723	0.3076	2.1624	0.3530	1.0706	0.5927
90%											
	bias	-0.0475	0.0031	-0.0286	0.0082	-0.0022	-0.0007	-0.0448	0.0467	-0.0772	0.0269
	SD	0.2375	0.1046	0.2692	0.1887	0.0920	0.1053	0.6685	0.2354	0.4478	0.1730
	SE	0.2331	0.1055	0.2682	0.1833	0.0889	0.1008	0.6775	0.2376	0.4120	0.1691
	RMSE	0.3361	0.1486	0.3810	0.2631	0.1279	0.1457	0.9526	0.3376	0.6132	0.2434
	CP	0.944	0.948	0.950	0.944	0.937	0.942	0.957	0.962	0.951	0.943
	ℓ(CI)	0.9073	0.4099	1.0448	0.6883	0.3467	0.3917	2.6449	0.9210	1.5911	0.6610

Tableau 3.3 – Résultats de la simulation pour $n = 2000$. c : proportion moyenne d'inflation de zéro. SD : écart-type empirique. SE : erreur type moyenne. CP : probabilité de couverture empirique des intervalles de confiance à 95%. ℓ(CI) : longueur moyenne des intervalles de confiance.

c		$\hat{\beta}_n$						$\hat{\gamma}_n$					$\hat{\alpha}_n$
		$\hat{\beta}_{1,n}$	$\hat{\beta}_{2,n}$	$\hat{\beta}_{3,n}$	$\hat{\beta}_{4,n}$	$\hat{\beta}_{5,n}$	$\hat{\beta}_{6,n}$	$\hat{\gamma}_{1,n}$	$\hat{\gamma}_{2,n}$	$\hat{\gamma}_{3,n}$	$\hat{\gamma}_{4,n}$	$\hat{\gamma}_{5,n}$	$\hat{\alpha}_{1,n}$
20%													
	bias	0.5013	-0.0150	-0.0459	-0.0931	0.0578	0.0004	0.4116	0.0971	-0.0882	-0.1323	0.0083	-0.3588
	SD	0.1837	0.0411	0.0804	0.0333	0.0563	0.0435	0.2924	0.1916	0.3760	0.1743	0.3392	0.0390
	SE	0.1611	0.0362	0.0737	0.0273	0.0461	0.0404	0.2912	0.1804	0.3689	0.1791	0.3301	0.0315
	RMSE	0.5576	0.0568	0.1183	0.1026	0.0929	0.0593	0.5827	0.2805	0.5339	0.2827	0.4733	0.3623
	ℓ(CI)	0.6283	0.1409	0.2874	0.1065	0.1795	0.1574	1.1378	0.7027	1.4382	0.6980	1.2913	0.1217

Tableau 3.4 – Résultats de la simulation du modèle ZINB pour $n = 300$. c : proportion moyenne d'inflation de zéro. SD : écart-type empirique. SE : erreur type moyenne. CP : probabilité de couverture empirique des intervalles de confiance à 95 %. ℓ(CI) : longueur moyenne des intervalles de confiance.

c		$\hat{\beta}_n$						$\hat{\gamma}_n$					$\hat{\alpha}_n$
		$\hat{\beta}_{1,n}$	$\hat{\beta}_{2,n}$	$\hat{\beta}_{3,n}$	$\hat{\beta}_{4,n}$	$\hat{\beta}_{5,n}$	$\hat{\beta}_{6,n}$	$\hat{\gamma}_{1,n}$	$\hat{\gamma}_{2,n}$	$\hat{\gamma}_{3,n}$	$\hat{\gamma}_{4,n}$	$\hat{\gamma}_{5,n}$	$\hat{\alpha}_{1,n}$
20%													
	bias	0.5039	-0.0143	-0.0427	-0.0936	0.0586	0.0005	0.4277	0.1157	-0.0988	-0.1462	0.0001	-0.3557
	SD	0.1424	0.0311	0.0636	0.0246	0.0437	0.0343	0.2185	0.1351	0.2715	0.1359	0.2728	0.0300
	SE	0.1241	0.0279	0.0568	0.0210	0.0353	0.0311	0.2220	0.1360	0.2792	0.1351	0.2516	0.0245
	RMSE	0.5382	0.0442	0.0953	0.0990	0.0812	0.0463	0.5291	0.2239	0.4017	0.2410	0.3710	0.3578
	ℓ(CI)	0.4852	0.1091	0.2220	0.0821	0.1380	0.1217	0.8689	0.5319	1.0924	0.5281	0.9855	0.0954

Tableau 3.5 – Résultats de la simulation du modèle ZINB pour $n = 500$. c : proportion moyenne d'inflation de zéro. SD : écart-type empirique. SE : erreur type moyenne. CP : probabilité de couverture empirique des intervalles de confiance à 95 %. ℓ(CI) : longueur moyenne des intervalles de confiance.

c		$\hat{\beta}_n$						$\hat{\gamma}_n$					$\hat{\alpha}_n$
		$\hat{\beta}_{1,n}$	$\hat{\beta}_{2,n}$	$\hat{\beta}_{3,n}$	$\hat{\beta}_{4,n}$	$\hat{\beta}_{5,n}$	$\hat{\beta}_{6,n}$	$\hat{\gamma}_{1,n}$	$\hat{\gamma}_{2,n}$	$\hat{\gamma}_{3,n}$	$\hat{\gamma}_{4,n}$	$\hat{\gamma}_{5,n}$	$\hat{\alpha}_{1,n}$
20%													
	bias	0.5152	-0.0161	-0.0456	-0.0937	0.0600	-0.0010	0.4198	0.1219	-0.0784	-0.1596	0.0019	-0.3510
	SD	0.0736	0.0154	0.0319	0.0126	0.0213	0.0172	0.1083	0.0686	0.1348	0.0677	0.1259	0.0151
	SE	0.0620	0.0139	0.0284	0.0104	0.0176	0.0155	0.1093	0.0667	0.1369	0.0662	0.1238	0.0124
	RMSE	0.5241	0.0263	0.0625	0.0951	0.0661	0.0232	0.4471	0.1550	0.2074	0.1856	0.1765	0.3516
	ℓ(CI)	0.2422	0.0543	0.1112	0.0409	0.0689	0.0607	0.4281	0.2612	0.5363	0.2595	0.4850	0.0485

Tableau 3.6 – Résultats de la simulation du modèle ZINB pour $n = 2000$. c : proportion moyenne d'inflation de zéro. SD : écart-type empirique. SE : erreur type moyenne. CP : probabilité de couverture empirique des intervalles de confiance à 95 %. ℓ(CI) : longueur moyenne des intervalles de confiance.

3.3 Interprétation des Résultats

Pour chaque configuration (Taille de l'échantillons $\times$ proportion de zéros) des paramètres de simulation, nous calculons le biais (en pourcentage) des estimateurs $\hat{\beta}_{i,n}$ et $\hat{\gamma}_{j,k}$ sur les N échantillons simulés.
On observe également une erreur standard moyenne SE calculer à partir de la matrice des variance covariance, l'écart-type empirique(SD)et l'erreur quadratique moyenne (RMSE) pour chaque estimateurs $\hat{\beta}_{i,n}$ et $\hat{\gamma}_{j,k}$ enfin nous proposons la couverture de probabilié de l'intervalle de confiance de niveaux 95%. Pour les βi et γ_j, les resultats sont données dans les tableau (3.1, 3.2, 3.3) pour le modèle ZIP et (3.4 3.5, 3.6) pour le modèle ZINB, Pour les différents taille d'échantillons.
À partire de ces résultats nous observons comme prévue que le biais, SE, SD, RMSE, et les longueurs moyenne des intervalles de confiance diminuent lorsque la taille de l'échantillon augmente et la proportion d'inflation de zéros diminue. Les couvertures de probabilité sont proche du niveau de confiance nominale. Comme on peut également s'y attendre pour chaque taille l'estimateur est plus performant lorsque la proportion d'inflation nul diminue. Ce qui signifie que $\hat{\sum}_n$(la matrice variance covariance) est un estimateur adéquate de la variance de l'EMV dans les modèles ZIP et ZINB. Globalement ces résultats indique la bonne performance de l'EMV.

Conclusion et perspectives

Ce sujet de mémoire porte sur les modèles de comptage avec inflation de zéro dans ce travail, nous nous sommes intéressés au problème de l'inférence statistique dans des modèles de comptage à inflation de zéro. Pour faciliter la lecture du document, nous avons rappelé dans le chapitre 1 quelques concepts fondamentaux sur les modèles linéaires et les familles exponentielles, les modèles lineaire généralisés et de sur-dispersion dans les données de comptage.

Dans le chapitre 2, dans un premier temps, nous avons établi rigoureusement l'existence et les propriétés asymptotiques (consistance, normalité asymptotique, estimation convergente de la variance asymptotique) de l'estimateur du maximum de vraisemblance (EMV) des paramètres du modèle de régression binomial à inflation de zéro.

Enfin, dans le chapitre 3, pour compléter cette étude théorique, nous avons menê une étude de simulations exhaustive sur des échantillons finis, et qui a permis d'étudier les propriétés des estimateurs du MV.
Nous avons pu montrer également que les fonctions de lien alternatives proposées sont assez flexibles et surpassent la fonction de lien standard.

Au terme de cet étude, dans l'avenir, nous proposons de faire une application de ces modèles sur des données réelles. Dons le chapitre 2, en s'inspirant du modèle poisson-gamma nous avons proposé un nouveau modèle pour données de comptages avec inflation de zéro appelé modèle ZIPLG dont nous avons fait une brève description, vu la complexité de ce modèl, nous pensons étudier dans la suite de nos travaux pour établires certains propriétés d'estimation et son application sur des données réels. Ce modèle devra également prendre en compte la nature imparfaite des comptages observés, et en particulier, la présence de censure.

Annexe: Démonstrations

Annexe: preuve de l'esperance et de la variance

On note $Y_i \sim$ ZIP (λ, π).
Où λ et π dependent respectivement de X et W Conditionnellement à X_i et W_i, l' espérance et la variance sont donnés par:
$\mathbb{E}(Y_i|X_i;W_i) = (1-\pi)\lambda_i$ et $\mathbb{V}ar(Y_i|X_i;W_i) = (1-\pi)(1+\pi\lambda_i)\lambda_i = (1+\pi)\mathbb{E}(Y|X,W)$

Preuve

$$\begin{aligned}
\mathbb{E}(Y|X;W) &= \sum_{y=0}^{\infty} y\mathbb{P}(Y=y|X;W) \\
&= \sum_{y=0}^{\infty} y(1-\pi)\frac{\exp(-\lambda)\lambda^y}{y!} \\
&= (1-\pi)\sum_{y=1}^{\infty}\frac{\exp(-\lambda)\lambda^y}{(y-1)!} \\
&= (1-\pi)\lambda\exp(-\lambda)\sum_{y=1}^{\infty}\frac{\lambda^{y-1}}{(y-1)!} \\
&= (1-\pi)\lambda\exp(-\lambda)\exp(-\lambda) \\
\mathbb{E}(Y|X;W) &= (1-\pi)\lambda
\end{aligned}$$

Calculons la variance en utilisant le theoreme de young

$\mathbb{V}ar(Y|X;W) = \mathbb{E}(Y^2|X;W) - (\mathbb{E}(Y|X;W))^2$

$$\begin{aligned}
\mathbb{V}ar(Y|X;W) &= \sum_{y=0}^{\infty} y^2 \mathbb{P}(Y=y|X;W) - (\mathbb{E}(Y|X;W))^2 \\
&= \sum_{y=0}^{\infty} y^2 (1-\pi)\frac{\exp(-\lambda)\lambda^y}{y!} - ((1-\pi)\lambda)^2 \\
&= (1-\pi)\sum_{y=0}^{\infty} y^2 \frac{\exp(-\lambda)\lambda^y}{y!} - ((1-\pi)\lambda)^2 \\
&= (1-\pi)(\lambda+\lambda^2) - ((1-\pi)\lambda)^2 \\
&= (1-\pi)\left((\lambda+\lambda^2) - (1-\pi)\lambda^2\right) \\
&= (1-\pi)\lambda(1+\pi\lambda) \\
\mathbb{V}ar(Y|X;W) &= \mathbb{E}(Y|X;W)(1+\pi\lambda)
\end{aligned}$$

Dnsité jointe

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

Donc la probabilité non conditionnelle vaut:

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

(car $y \in \mathbb{N}$ et $\int_0^\infty \frac{(1+\frac{1}{\beta})^{y+\alpha}}{\Gamma(y+\alpha)}\theta^{y+\alpha-1}\exp(-\theta(1+\frac{1}{\beta}))d\theta = 1$

$$\begin{aligned}
&= \frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)}\frac{\beta^{y+\alpha}}{\beta^\alpha(1+\beta)^{y+\alpha}} \\
&= \frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)}\frac{\beta^{y+\alpha}}{(1+\beta)^{y+\alpha}} \\
&= \frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)}(1-\frac{1}{1+\beta})^y(\frac{1}{1+\beta})^y(\frac{1}{1+\beta})^\alpha
\end{aligned}$$

Si α n'est pas connue, ce qui est souvent le cas en pratique, cette loi n'appartient pas a la famille exponentielle et ne peut donc pas être utiliser dans un cas GLM.

Par contre si α est connue on a :

$$\begin{aligned}\mathbb{P}(Y=y,\alpha,\beta) &= \frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)}\left(1-\frac{1}{1+\beta}\right)^{y}\left(\frac{1}{1+\beta}\right)^{\alpha}\\ &= \exp(y\log(\frac{\beta}{1+\beta})-\alpha\log(1+\beta)+\log(\frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)})\end{aligned}$$

Avec la notation usuelle $\theta=\log(\frac{\beta}{1+\beta})=logit(\beta)\Leftrightarrow\beta=\frac{\exp(\theta)}{1-\exp(\theta)}$, $a(\phi)=1$
$b(\theta)=\alpha\log(1+\beta)=-\log(1-\exp(\theta))$ et $c(y,\phi)=\frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)}$
Espérance et variance de la loi binomiale négative
$\mathbb{P}(Y=y)=\frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)}(1-p)^{y}p^{r}$ quand $Y\sim\mathcal{NB}(r,p)$
posons :$r=\alpha$ et $p=\frac{1}{1+\beta}\Leftrightarrow\beta=\frac{1-p}{p}$ avec $p\neq 0$
Ainsi comme il s'agit d'une loi Discrète :

$$\begin{aligned}\mathbb{E}(Y) &= \sum_{k=0}^{\infty}k\mathbb{P}(Y=k)\\ &= \sum_{k=1}^{\infty}k\frac{\Gamma(k+r)}{\Gamma(k+1)\Gamma(r)}(1-p)^{k}p^{r}\\ &= \sum_{k=1}^{\infty}\frac{\Gamma(k+r)}{(k-1)\Gamma(r)}(1-p)^{k}p^{r}\ car\Gamma(k+1)=k!\\ &= \frac{r(r-p)}{p}\sum_{k=1}^{\infty}k\frac{\Gamma(k-1+r+1)}{(k-1)!\Gamma(r+1)}(1-p)^{k-1}p^{r+1}\\ &= \frac{r(r-p)}{p}\sum_{j=0}^{\infty}\frac{\Gamma(j+r+1)}{j!\Gamma(r+1)}(1-p)^{j}p^{r+1}\end{aligned}$$

Nous avons poser $j=k-1$ et on a obtenue la dernière égalité
effectivement $\sum_{j=0}^{\infty}\frac{\Gamma(j+r+1)}{j!\Gamma(r+1)}(1-p)^{j}p^{r+1}=1$
car si X suit une loi discrète (dans notre cas $X\sim NB(r+1,p)$ alors par définition de la probabilité on a :

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

en posant en suite $p=\frac{1}{\beta+1}$ on obtient le résultat
$\mathbb{E}(Y)=\frac{\alpha(1-p)}{p}=\alpha\beta$
calcule de la Variance $\mathbb{V}ar(Y)=\mathbb{E}(Y^{2})-\mathbb{E}^{2}(Y)$
calculons d'abord $\mathbb{E}(Y^{2})$

$$\begin{aligned}
\mathbb{E}(Y^2) &= \sum_{k=0}^{\infty} k^2 \mathbb{I}(Y=k) \\
&= \sum_{k=1}^{\infty} k^2 \frac{\Gamma(k+r)}{k!\Gamma(r)}(1-p)^k p^r \\
&= \sum_{k=1}^{\infty} k^2 \frac{\Gamma(k+r)}{(k-1)!\Gamma(r)}(1-p)^k p^r \text{ avec } k = k-1+1 \\
&= \sum_{k=2}^{\infty}(k-1)(\frac{\Gamma(k+r)}{(k-1)!\Gamma(r)}(1-p)^k p^r + \sum_{k=1}^{\infty}(\frac{\Gamma(k+r)}{(k-1)!\Gamma(r)}(1-p)^k p^r \\
&= \sum_{k=2}^{\infty}(k-1)(\frac{\Gamma(k+r)}{(k-2)!\Gamma(r)}(1-p)^k p^r + \mathbb{E}(Y) \\
&= \frac{(1-p)^2 r(r+1)}{p^2}\sum_{k=2}^{\infty}\frac{\Gamma(k-2+r+2)}{(k-2)!\Gamma(r+2)}(1-p)^{k-2}p^{r+2} + \mathbb{E}(Y) \\
&= \frac{(1-p)^2 r(r+1)}{p^2} + \frac{r(1-p)}{p}
\end{aligned}$$

en posant j= k-2

$$\sum_{k=2}^{\infty}\frac{\Gamma(k-2+r+2)}{(k-2)!\Gamma(r+2)}(1-p)^{k-2}p^{r+2} = \sum_{j=0}^{\infty}\frac{\Gamma(j+r+2)}{j!\Gamma(r+2)}(1-p)^j p^{r+2}$$

on obtient à nouveaux $X \sim \mathcal{NB}(r+2,p)$ et par définition de probabilité $\sum_j^{\infty}\mathbb{P}(X=j)=1$
Ainsi donc

$$\begin{aligned}
\mathbb{V}ar(Y) &= \mathbb{E}(Y^2) - \mathbb{E}^2(Y) \\
&= \frac{(1-p)^2 r(r+1)}{p^2} + \frac{r(1-p)}{p} - \frac{r^2(1-p)^2}{p^2} \\
&= \frac{(1-p)r[(r+1)(r-p)+p-r(1-p)]}{p^2} \\
&= \frac{(1-P)r}{P^2}
\end{aligned}$$

en paramètrent avec $r=\alpha$ et $p=\frac{1}{1+\beta} \Leftrightarrow \beta = \frac{1-p}{p}$ quand $p \neq 0$ on obtient $Var(Y) = \frac{(1-p)r}{p^2} = \alpha\beta(1+\beta)$.
Si on pose $\alpha = \frac{\lambda}{\beta}$ alors on obtient la première version (I) de la loi Binomiale Négative avec
$\mathbb{E}(Y)=\lambda$ et $\mathbb{V}ar(Y) = \lambda(1+\beta) = \mathbb{E}(Y)(1+\beta)$ (nommé scaled Variance)
Si on pose $\beta = \frac{\lambda}{\beta}$ alors on obtient la seconde Version (II) de la loi Binomiale Négative avec $\mathbb{E}(Y)=\lambda$ et $Var(Y) = \lambda + \frac{\lambda^2}{\alpha} = \mathbb{E}(Y) + \frac{1}{\alpha}\mathbb{E}^2(Y)$ (Variance nomme quadratique)
en régression on a toujours

On calcule l'espérance et la variance de ce modèle ZIB comme suit:

$$\begin{aligned}
\mathbb{E}(Y|X;W) &= \sum_{y}^{n} y(1-\pi)\binom{n}{y}p^y(1-p)^{n-y} \\
&= (1-\pi)\sum_{y}^{n} y\frac{n!}{y!(n-y)!}p^y q^{n-y} \\
&= (1-\pi)p\sum_{y}^{n} \frac{n(n-1)!}{(y-1)!(n-y)!}p^{(y-1)}q^{n-y} \\
&= (1-\pi)np\sum_{y}^{n} \frac{(n-1)!}{(y-1)!(n-y)!}p^{(y-1)}q^{n-y} \\
&= (1-\pi)np\sum_{j=0}^{n-1} \frac{(n-1)!}{j!(n-j-1)!}p^{j}q^{n-1-j} \\
&= (1-\pi)np\sum_{j=0}^{n-1} \binom{n-1}{j}p^{j}q^{n-1-j} \\
&= np(p+q)^{n-1} \\
\mathbb{E}(Y|X;W) &= (1-\pi)np
\end{aligned}$$

Calcule de la variance $\mathbb{V}ar(Y|X;W) = \mathbb{E}(Y^2|X;W) - \mathbb{E}(Y|X;W)$
Nous allons dans un permier temps calculer $\mathbb{E}(Y^2|X;W)$

$$\mathbb{E}(Y^2|X;W) = \mathbb{E}(Y(Y-1)X;W) + \mathbb{E}(Y|X;W)$$

$$\begin{aligned}
\mathbb{E}(Y^2|X;W) &= \mathbb{E}(Y(Y-1)|X;W) + \mathbb{E}(Y|X;W) \\
&= \sum_{y=0}^{n} y(y-1)(1-\pi)\frac{n!}{y!(n-y)!}p^y q^{n-y} + \\
&= (1-\pi)n(n-1)p^2\sum_{y=2}^{n} \frac{(n-2)!}{(y-1)!(n-y)!}p^{y-2}q^{n-y} + (1-\pi)np \\
&= (1-\pi)n(n-1)p^2\sum_{j=0}^{n-2} \frac{(n-2)!}{(y-2)!(n-y)!}p^{y-2}q^{n-y} + (1-\pi)np \\
&= (1-\pi)n(n-1)p^2\sum_{j=0}^{n-2} \binom{n-2}{j}p^{j}q^{n-2-j} + (1-\pi)np \\
&= (1-\pi)n(n-1)p^2(p+q)^{n-2} + (1-\pi)np \\
&= (1-\pi)\left(n(n-1)p^2 + np\right) \\
\mathbb{E}(Y^2|X;W) &= (1-\pi)\left[n(n-1)p^2 + np\right]
\end{aligned}$$

On en deduit alors la variance de Y par:

$$\begin{aligned}
\mathbb{V}ar(Y|X;W) &= (1-\pi)\left[n(n-1)p^2+np\right]+\left[(1-\pi)np\right]^2 \\
&= (1-\pi)\left[n(n-1)p^2+np\right]+\left[(1-\pi)np\right]^2 \\
&= (1-\pi)\left[n(n-1)p^2+np+(1-\pi)(np)^2\right] \\
\mathbb{V}ar(Y|X;W) &= (1-p)n\pi\left[1-\pi(1-pn)\right]
\end{aligned}$$

Bibliographie

[1] Akaike H. IEEE transactions on automatic control. *A new look at the statistical model identification AC-(4) :716-723, 1974.1*

[2] Diallo A.O. ,Diop. A , and J. F. Dupuy .,2007. Asymptotic properties of the maximum likelihood estimator in zero-inflated binomial regression. *Communications in Statistics-Theory and Methods, 46 (20) : 9930-9948, 2017. 104, 106, 111, 112, 123.*

[3] A. O.Dialo, (2017). *Inférence statistique dans des modèles de comptage à inflation de zéro. Applications en économie de la santé. Theses, INSA de Rennes*.

[4] *ali, E. (2021). Modéles de régression marginaux pour des données de comptage à excèes de zéros. Thèse de doctorat, IRMAR-INSA de Rennes & LERSTAD-UGB de Saint-Louis*

[5] Azais et Bardet 2006. *the lionear model :Regression, analys of variance 2006.5*

[6] *Carroll R.J, Ruppert D. et Stefanski L. A. . Measurement error in nonlinear models. Chapman and Hall, New York, 1995. 23*

[7] Cox., David R.,1983. Some remarks on overdispersion. *Journal of Biometrika Trust* 70(1):269-274.

[8] zardo,C. 2007 *Zero-inflated generalized poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcingrates. Statist. Model,7(2) :125-153, 2007. 24, 26, 27*

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

[10] *Dempster A.,Laird N. et Rubin D. Maximum likelihood from incomplete data via the em algorithm (with discussion).J.Roy. Statist. Soc. Ser. B, vol. 39, pages 1-38, 1977. 23*

[11] Diop A., Diop A., Dupuy J.-F., 2011. Maximum likelihood estimation in the logistic regression model with a cure fraction *Electronic Journal of Statistics*,5, 460-483.

[12] L.fahrmeir, et H.kufmann (1985). *Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. The Annals of Statistics, 13(1):342-368.*

[13] Fong D.Y.T., Yip., 1995. A note on information loss in analysing a mixture model of count data. *Comm. Statist. Theory Methods*,24:3197-3209.

[14] Garray, A. M., Hashimoto, E. M., Ortega, E.M.M., dan Lachos, V. H. (2011) *On Estimation and Influence Diagnostics For Zero Inflated Negative Binomial Regression Models. Computational Statistics and Data Analysis, 55 (3), p.1304-1318.*

[15] D.B. Hall ., J. Shen.,(2010) Robust estimation for zero-inflated poisson regression.. *Scand. J. Statist* 37 :237-252, 25, 26, 27.

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

[17] Henningsen *" maxLik : A package for maximum likelihood estimation in R. " Computational Statistics, n 26 (201)) : 443-458.*

[18] Heyde (1997) *Quasi-likelihood and its application : a general approach to optimal parameter estimation 1997*

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

[20] Lambert D.,(1992). Zero-inflated poisson regression models with an application to defects in manufacturing. *Technometrics* 34 :1-14, 21, 24, 34, 66, 103 .

[21] *Lange, K. (2004). Computational statistics and optimization theory at ucla. The American Statistician, 58(1):9-11.*

[22] *P.mccullagh , J. nelder,(1989). Generalized Linear Models, Second Edition. Chapman & Hall / CRC Monographs on Statistics and Applied Probability. Taylor & Francis.*

[23] Mwalili, SM., Lesaffre, E., Declerck, D., Demetrio, C.G.B. The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. *Stat Methods Med Res.*,17(2):123-139, 2008.

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

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

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

[27] Mullahy, J., 1986. Specification and testing of some modified count data models. *Journal of Econometrics*,33:341-365. J.A. Nelder and R. W. M. Wedderburn. Generalized linear models. *J. Roy. Statist. Soc. Ser. A* 135, 370-384.

[28] M. Ridout., C. G. B. Demetrio., and J Hinde., Models for count data with many zeros. *Invited paper presented at the Nineteenth In Bio Conf, Cape Town, South Africa, pages* 179-190,21.

[29] M. Ridout., J. Hinde., and C. G. B. Demetrio.,2001. A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. *Biometrics* 57 :219-223, 35, 66, 104.

[30] P. C. Consul et F. Famoye. Communication in Statistics-Theory and Methods . *Generalized Poisson regression model* 21, 89-109.

[31] Roulan Jiang., Xiang Zhan.,et Tianing Wng(2022) a flexible zero-inflated poisson-gamma model with application to microbiome read counts *arXiv preprint arXiv* :2207.07796, 2022.

[32] *O. Rosen, W. X. Jiang et M. A. Tanner. Mixtures of marginal models. Biometrika, vol. 87, pages 391-404, 2000. 23*

[33] A.Zeileis, C.Kleiber and S.Jacman. Regression Models for Count Data in R *Journal of Statistical Software* 2.

[34] Winkelmann.,R. Econometric Analysis of Count Data. *Springer, Berlin, 5th edition, 2008*

[35] du 21/07/2023 *http:/ /www.math.wm.edu /leemis /chart /UDR /UDR.html*