Modèles de régression pour les géométriques Poisson-Tweedie des données ultra-dispersées de comptage

Rahma Abid[1], Célestin C. Kokonendji[2] & Afif Masmoudi[3]

[1] *Université de Sfax, Laboratoire de Probabilités et Statistique, Tunisie.*
rahma.abid.ch@gmail.com
[2] *Université Bourgogne Franche-Comté, Laboratoire de Mathématiques de Besançon, France.*
celestin.kokonendji@univ-fcomte.fr
[3] *Université de Sfax, Laboratoire de Probabilités et Statistique, Tunisie.*
afif.masmoudi@fss.rnu.tn

Résumé. Une nouvelle classe de mélange Poisson-exponentiel-Tweedie (PET) est introduite dans le cadre des modèles linéaires généralisés pour l'analyse des données de comptage ultra-surdispersées. Le modèle proposé est équivalent aux modèles exponentiels-Poisson-Tweedie issus des sommes géométriques de variables Poisson-Tweedie. À cet égard, les modèles PET englobent les versions géométriques des modèles Hermite, de Neyman Type A, de Pólya-Aeppli, de négative binomiale et de Poisson inverse Gaussienne. La géométrique décalée à zéro est considérée comme la distribution de référence. Des propriétés, dans les modèles PET, des nouveaux indices relatifs des phénomènes de dispersion et de zéro-inflation sont alors établies. Les modèles de régression correspondants sont ajustés par l'approche de quasi-vraisemblance. Les performances de ces modèles sont illustrées sur des données réelles dans les domaines de la fiabilité et de l'assurance.

Mots-clés. Indice de dispersion relatif, modèle linéaire généralisé, quasi-vraisemblance, zéro-inflation relatif

Abstract. A new class of Poisson-exponential-Tweedie (PET) mixture is introduced in the framework of generalized linear models for analysing ultra-overdispersed count data. The proposed model is equivalent to the exponential-Poisson-Tweedie models arising from geometric sums of Poisson-Tweedie random variables. In this respect, the PET models encompass the geometric versions of Hermite, Neyman Type A, Pólya-Aeppli, negative binomial and Poisson inverse Gaussian models. The zero-shifted geometric is presented as the reference count distribution. Some properties, into the PET models, of new relative indexes of dispersion and zero-inflation phenomena are established. The corresponding regression models are fitted by the quasi-likelihood approach. Illustrative practical applications on real count datasets in the reliability and insurance fields are analyzed.

Keywords. Generalized linear model, quasi-likelihood, relative dispersion index, relative zero-inflation

1 Introduction

Dans l'analyse des données de comptage, la présence d'au moins une surdispersion ou d'un excès de zéros mérite une attention particulière pour le choix du modèle de comptage (par exemple, Hinde et Demétrio, 1998). Le phénomène de surdispersion se manifeste généralement par le zéro-inflation et/ou la queue lourde. Les deux mesures sont effectuées par rapport à la distribution de Poisson et à la négative binomiale qui sont utilisées pour modéliser ces types de jeux de données de comptage. De nombreux modèles ont été construits en combinant et en mélangeant la distribution de Poisson ; voir, par exemple, Kokonendji et al. (2004) et Bonat et al. (2018). Le but de cet article est de fournir des réponses aux trois questions suivantes. Comment procéder lorsque le degré ou le niveau de surdispersion est très élevé ? Devrions-nous relativiser sa mesure par rapport à une autre distribution de comptage de référence que la distribution de Poisson ? Comment construire une nouvelle famille de modèles de comptage ultra-surdispersés ?

De manière analogue aux modèles exponentiels de dispersion (Jørgensen, 1997), les modèles géométriques et discrets de dispersion ont été introduits ultérieurement par Jørgensen-Kokonendji (2011, 2016) en tant que modèles de dispersion pour les sommes géométriques et pour les variables de comptage, respectivement. Une classe importante basée sur ces modèles est représentée comme

$$Y = \sum_{\ell=1}^{G} \mathrm{PT}_\ell \tag{1}$$

où $\mathrm{PT}_1, \mathrm{PT}_2, \ldots$ sont indépendants et identiquement distribués (iid) comme PT, une variable aléatoire Poisson-Tweedie (Kokonendji et al., 2004), et G est une variable géométrique, indépendante de PT. Abid et al. (2018b) ont caractérisé les sommes géométriques des modèles de Tweedie pour l'analyse de données continues et semi-continues. La représentation (1), également considérée comme un mélange exponentiel de modèles PT, conduit à appeler de tels modèles exponentiel-Poisson-Tweedie ou aussi Poisson-exponentiel-Tweedie (PET), en indiquant que ce type de modèle traite les données de comptage. La classe de PET présente une variance de la forme

$$\mathrm{Var}Y = m + m^2 + \phi m^p, \tag{2}$$

où $m = EY$, ϕ et p sont les paramètres de Tweedie.

Dans ce travail, nous proposons une nouvelle classe de modèles linéaires généralisés (McCullagh et Nedler, 1989) pour traiter les données de comptage ultra-surdispersées. La section 2 est consacrée aux modèles PET pour $p \geq 1$ et aux propriétés des indices de dispersion et de zéro-inflation relatifs à la géométrique décalée à zéro. La section 3 présente le modèle de régression pour les PET. La section 4 résume des applications intéressantes du modèle considéré et une conclusion est donnée en Section 5.

2 Modèles Poisson-exponentiel-Tweedie (PET)

Nous étudions les propriétés des modèles Poisson-exponential-Tweedie (PET), en tenant compte du modèle mélange équivalent. La famille PET est donnée par la formulation hiérarchique suivante

$$Y|Z \sim \text{Poisson}(Z), \;\; Z \sim \text{Tw}_p(Xm, X^{1-p}\phi) \;\text{ et }\; X \sim \text{Exp}(1). \tag{3}$$

où $\text{Exp}(\lambda)$, $\text{Poisson}(\lambda)$ et $\text{Tw}_p(\lambda, \psi)$ désignent la distribution exponentielle de paramètre λ, la distribution de Poisson de paramètre λ et la distribution Tweedie (Tweedie, 1984) de moyenne λ, de paramètre de dispersion ψ et de paramètre de puissance p, respectivement. Les versions géométriques (1) de la classe PT se réduisent à une représentation de mélange exponentiel (Abid et al., 2019a ; proposition 2.4) en tant que

$$X \sim \text{Exp}(1), \;\; [Y|X]|Z \sim \text{Poisson}(Z) \;\text{ et }\; Z \sim \text{Tw}_p(Xm, X^{1-p}\phi). \tag{4}$$

Pour les deux modèles (3) et (4), il faut $p \geq 1$ pour que Z soit positive. La proposition suivante souligne le fait que ces deux processus aléatoires génèrent le même modèle que nous allons explorer dans le reste du document.

Proposition 2.1 *Soient Y_1 et Y_2 deux variables aléatoires définies par (3) et (4), respectivement. Alors:*

(i) Y_1 et Y_2 ont la même distribution donnée par

$$\mathbb{P}(Y_1 = y) = \int_0^\infty \int_0^\infty \frac{\exp\{-zx\}z^y}{y!} \text{Tw}_p(mx, \phi x^{1-p})(z)dzdx = \mathbb{P}(Y_2 = y).$$

(ii) Nous avons (2) pour $Y = Y_1 = Y_2$.

Cette classe est notée $\text{PETw}_p(m, \phi)$ pour $p \in \{0\} \cup (1, \infty)$ et $m \in (0, \infty)$. Un cas particulier inclut la négative binomiale pour $p = 2$ avec la relation variance-moyenne $m + (1 + \phi)m^2$.

Calculer l'indice de P-dispersion, en relation avec la distribution de Poisson (P), constitue généralement la première étape dans l'analyse des données de comptage. Un autre indice caractéristique est le P-zéro-inflation, défini comme la proportion des zéros observés par rapport à ceux de la distribution de P.

$$\text{P-DI} = \frac{\text{Var}Y}{\mathbb{E}Y} \quad \text{et} \quad \text{P-ZI} = \mathbb{E}Y + \log \mathbb{P}(Y = 0). \tag{5}$$

La P-DI indique une surdispersion pour P-DI > 1, une sous-dispersion pour P-DI < 1 et une équidispersion pour P-DI = 1. Le P-zéro-inflation indique une inflation pour P-ZI > 0, une déflation pour P-ZI < 0 et P-pas excès de zéros pour P-ZI = 0. Cependant, comme nous nous intéressons à un phénomène d'ultra-surdispersion, il est naturel

de considérer la négative binomiale unitaire qui est la géométrique décalée à zéro (G_0) comme modèle de référence alternatif. En effet, nous définissons les indices G_0-dispersion (G_0- DI) et G_0-zéro-inflation (G_0-ZI) comme mesure de départ par rapport au modèle G_0 par

$$G_0\text{-DI} = \frac{\text{Var}Y}{\mathbb{E}Y + (\mathbb{E}Y)^2} \quad \text{et} \quad G_0\text{-ZI} = \log(1 + \mathbb{E}Y) + \log \mathbb{P}(Y = 0).$$

Au vu de ces indices, la distribution G_0 correspond toujours à G_0-DI = 1, tandis qu'une distribution avec G_0-DI > 1 est G_0-surdispersée et une distribution avec G_0-DI > 1 est G_0- sousdispersée. De plus, les valeurs G_0-ZI > 0 expriment G_0-zéro-inflation, alors que G_0-ZI < 0 implique G_0-zéro-déflation.

Proposition 2.2 *Le PET est surdispersé et zéro-inflaté par rapport à P et G_0, respectivement.*

Tous les modèles de régression PET sont conçus pour traiter les données G_0-surdispersées pour $\phi > 0$. La seule restriction à avoir un modèle significatif est que $\text{Var}Y > 0$, c'est-à-dire $\phi > -m^{2-p} - m^{1-p}$. En conséquence, les modèles PET peuvent être étendus pour traiter les données G_0-sousdispersées.

Référence	Puissance	Caractéristiques	Dispersion
P / G_0	-	Equi / Equi	−
Géométrique Hermite	$p = 0$	Sur, sous	$\phi \leqslant 0$
[N'existe pas]	$0 < p < 1$	−	−
Géométrique Neyman Type A	$p = 1$	Sur, sous, ZI	$\phi \leqslant 0$
Géométrique Poisson Poisson composé	$1 < p < 2$	Sur, sous, ZI	$\phi \leqslant 0$
Géométrique Pólya-Aeppli	$p = 1.5$	Sur, sous, ZI	$\phi \leqslant 0$
Négative binomiale	$p = 2$	Sur, sous	$\phi \leqslant 0$
Géométrique Poisson positif stable	$p > 2$	Sur, HT	$\phi > 0$
Géométrique Poisson inverse Gaussienne	$p = 3$	Sur, HT	$\phi > 0$

Table 1: Modèles de référence et caractéristiques dominantes par rapport aux P et G_0. HT désigne la queue lourde.

3 Modèles de régession PET

Considérons un jeu des données transversales, $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, où y_i sont des observations iid de $Y_i \sim \text{PETw}_p(m_i, \phi)$ et $\mathbf{x}_i$ est un vecteur $(Q \times 1)$ de covariables connues. Ainsi, l'espérance et la variance sont données respectivement par

$$\mathbb{E}(Y_i) = m_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \tag{6}$$

$$Var(Y_i) = m_i + m_i^2 + \phi m_i^p = V_i, \tag{7}$$

où $\boldsymbol{\beta}$ est un vecteur de coefficients de régression inconnus. On utilise ici la fonction lien logarithme dans (6), mais toute autre fonction de lien appropriée pourrait être adoptée. Le modèle de régression PET est alors paramétré par $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$, avec $\boldsymbol{\gamma} = (\phi, p)$.

D'après Jørgensen-Knudsen (2004), la fonction de quasi-score pour $\boldsymbol{\beta}$ a la forme

$$\psi_\beta(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \left[\sum_{i=1}^{n} \frac{\partial m_i}{\partial \beta_1} V_i^{-1}(y_i - m_i), \ldots, \sum_{i=1}^{n} \frac{\partial m_i}{\partial \beta_Q} V_i^{-1}(y_i - m_i) \right],$$

avec $\partial m_i / \partial \beta_j = m_i$ pour $j = 1, \ldots, Q$. La fonction d'estimation de Pearson pour ϕ et p est donnée par

$$\psi_\gamma(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \left[-\sum_{i=1}^{n} \frac{\partial V_i^{-1}}{\partial \phi}\{(y_i - m_i)^2 - V_i\}, -\sum_{i=1}^{n} \frac{\partial V_i^{-1}}{\partial p}\{(y_i - m_i)^2 - V_i\} \right].$$

Soit $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$ la solution du système d'équations $\psi_\gamma(\boldsymbol{\beta}, \boldsymbol{\gamma}) = 0$ et $\psi_\beta(\boldsymbol{\beta}, \boldsymbol{\gamma}) = 0$, l'algorithme est fourni par : $\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - S_\beta^{-1}\psi_\beta(\boldsymbol{\beta}^{(i)}, \boldsymbol{\gamma}^{(i)})$ et $\boldsymbol{\gamma}^{(i+1)} = \boldsymbol{\gamma}^{(i)} - \alpha S_\gamma^{-1}\psi_\gamma(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\gamma}^{(i)})$, où

$$S_{\beta_{jk}} = \mathbb{E}\frac{\partial}{\partial \beta_k}\psi_{\beta_j}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = -\sum_{i=1}^{n} m_i x_{ij} V_i^{-1} x_{ik} m_i \text{ et } S_{\gamma_{jk}} = \mathbb{E}\frac{\partial}{\partial \gamma_k}\psi_{\gamma_j}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = -\sum_{i=1}^{n} \frac{\partial V_i^{-1}}{\partial \gamma_j} V_i \frac{\partial V_i^{-1}}{\partial \gamma_k} V_i,$$

tels que γ_j et γ_k désignent ϕ ou p, et donnant

$$S_\gamma = \begin{pmatrix} -\sum_{i=1}^{n}(m_i^p/V_i)^2 & -\sum_{i=1}^{n}\{\phi m_i^{2p}\log(m_i)\}/V_i^2 \\ -\sum_{i=1}^{n}\{\phi m_i^{2p}\log(m_i)\}/V_i^2 & -\sum_{i=1}^{n}\{\phi m_i^p \log(m_i)/V_i\}^2 \end{pmatrix}.$$

4 Simulations et exemples d'applications

Pour explorer la flexibilité du modèle de régression PET, nous présenterons une étude de simulation. Les résultats montrent que le biais et l'erreur standard tendent vers 0 lorsque la taille de l'échantillon augmente. Aussi, pour tous les scénarios de simulation, les taux de confiances empiriques sont proches du niveau nominal.

Nous présenterons deux exemples pour illustrer l'application des modèles PET. Le premier concerne le nombre d'accidents automobiles. Les indices de dispersion estimés indiquent une surdispersion par rapport à la distribution de P et une équidispersion par rapport à la distribution G_0. Le deuxième exemple porte sur le nombre d'actions d'entretien pour des bâtiments. Ces données illustrent les cas de surdispersion et d'inflation élevées par rapport à la distribution de P.

5 Remarques finales

Dans ce travail, nous avons présenté la nouvelle distribution de PET pour traiter les données de comptage ultra-surdispersées. Nous avons ensuite introduit les modèles de régression PET dans le cadre des modèles linéaires généralisés. De plus, nous avons adopté une approche par fonction d'estimation pour l'estimation et l'inférence basée uniquement sur des hypothèses de moment de second ordre. Etendre le modèle proposé aux données de comptage multivariées ultra-surdispersées, avec de nombreuses applications pour l'analyse de données longitudinales et spatiales est en préparation.

Bibliographie

Abid, R., Kokonendji, C.C. and Masmoudi, A. (2019a). Geometric dispersion models with real quadratic v-functions, *Statistics and Probability Letters*, 145, pp.197-204.

Abid, R., Kokonendji, C.C. and Masmoudi, A. (2019b). Geometric-Tweedie regression models for continuous and semicontinuous data with variation phenomenon, *AStA Advances in Statistical Analysis*, DOI : 10.1007/s10182-019-00350-8.

Bonat, W.H., Jørgensen, B., Kokonendji, C.C., Hinde, J. and Demétrio, C.G.B. (2018). Extended Poisson-Tweedie: properties and regression models for count data, *Statistical Modelling*, 18, pp.24-49.

Hinde, J. et Demétrio, C.G.B. (1998), *Overdispersion: Models and Estimation*, Associacao Brasileira de Estatistica, Sao Paulo.

Jørgensen, B. (1997), *The Theory of Dispersion Models*, Chapman and Hall, London.

Jørgensen, B. and Knudsen, S.J. (2004). Parameter orthogonality and bias adjustment for estimating functions, *Scandinavian Journal of Statistics*, 31, pp.93-114.

Jørgensen B. and Kokonendji, C.C. (2011), Dispersion models for geometric sums, *Brazilian Journal of Probability and Statistics*, 25, pp.263-293.

Jørgensen, B. and Kokonendji, C.C. (2016). Discrete dispersion models and their Tweedie asymptotics. *AStA Advances in Statistical Analysis*, 100, pp.43-78.

Kokonendji, C.C., Dossou-Gbété, S. and Demétrio, C.G.B. (2004). Some discrete exponential dispersion models: Poisson-Tweedie and Hinde-Demetrio classes, *Statistics and Operations Research Transactions*, 28, pp.201-214.

McCullagh, P. et Nelder, J. (1989), *Generalized Linear Models, 2nd edition*, Chapman & Hall, London.

Tweedie, M.C K. (1984), An index which distinguishes between some important exponential families, In: Ghosh JK, Roy J (eds), *Statistics: Applications and New Directions*, Proceedings of the Indian Statistical Golden Jubilee International Conference, Calcutta, 579-604.