Inférence statistique dans des modèles de comptage à inflation de zéro. Applications en économie de la santé

Alpha Oumar Diallo

Thèse

THESE INSA Rennes
sous le sceau de l'Université Bretagne Loire
pour obtenir le titre de
DOCTEUR DE L'INSA RENNES
Spécialité : Mathématiques et leurs interactions

présentée par
ALPHA OUMAR DIALLO
ECOLE DOCTORALE : *ED MATHSTIC et ED ST/UGB*
LABORATOIRE : *IRMAR et LERSTAD*

Inférence statistique dans des modèles de comptage à inflation de zéro. Applications en économie de la santé

Thèse soutenue le 27-11-2017
devant le jury composé de :

Chirstophe RAULT
Professeur, Université d'Orléans / Président du jury
Anne-Françoise YAO
Professeur, Université Clermont-Ferrand II / Rapporteur
Célestin C KOKONENDJI
Professeur, Université Bourgogne Franche-Comté / Rapporteur
Valérie MONBET
Professeur, Université Rennes 1 / Examinatrice
Aliou DIOP
Professeur, Université Gaston Berger (Sénégal) / Co-directeur de thèse
Jean-François DUPUY
Professeur, INSA de Rennes (France) / Co-directeur de thèse.

Inf¶rence statistique dans des mod· les de comptage ¯ inflation de z¶ro. Applications en ¶conomie de la sant¶.

Alpha Oumar DIALLO

En partenariat avec

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

UNIVERSITE GASTON BERGER SAINT-LOUIS DU SENEGAL

THÈSE

En cotutelle, présentée pour obtenir
LE GRADE DE DOCTEUR

DE L'INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES
ET DE L'UNIVERSITÉ GASTON BERGER DE SAINT-LOUIS

Spécialité : Mathématiques et leurs interactions

Par **Alpha Oumar DIALLO**

Inférence statistique dans des modèles de comptage à inflation de zéros. Applications en économie de la santé

Soutenue publiquement le 27/11/2017 devant le jury :

Anne-François YAO	**Professeur, Clermont-Ferrand II**	**Rapporteur**
Célestin C KOKONENDJI	**Professeur, Université Bourgogne Franche-Comté**	**Rapporteur**
Christophe RAULT	**Professeur, Université d'Orléans**	**Examinateur**
Valérie MONBET	**Professeur, Université Rennes 1**	**Examinatrice**
Aliou DIOP	**Professeur, Université Gaston BERGER**	**Co-directeur de Thèse**
Jean-François DUPUY	**Professeur, INSA de Rennes**	**Co-directeur de Thèse**

Thèse préparée au Laboratoire d'Études et de Recherches en Statistique et Développement et à l'Institut de Recherche Mathématique de Rennes

Remerciements

Pour souscrire à la tradition, j'adresse ces quelques mots de remerciements à l'endroit de toutes les personnes qui ont contribué, à divers titres, à l'élaboration de ce travail.

Je tiens à remercier tout particulièrement mes deux directeurs de thèse les professeurs Aliou DIOP et Jean-françois DUPUY pour m'avoir offert l'opportunuité de réaliser cette thèse dans un cadre fort agréable et d'avoir accompagné mon travail sans stress. Je vous remercie également pour les discussions fructueuses et enrichissantes que nous avons eues : elles m'ont beaucoup appris et m'ont aidé à avancer dans ce milieu nouveau pour moi. J'espère que notre collaboration ne s'arrêtera pas à cette thèse.

J'adresse mes sincères remerciements aux professeurs Anne-François YAO et Célestin C KOKONENDJI en acceptant d'examiner mes travaux en qualité de rapporteurs malgré leurs nombreuses charges respectives. Je remercie aussi les professeurs Christophe RAULT et Valérie MONBET de me faire l'honneur de participer à mon jury de thèse. Merci d'avoir accepté d'examiner mes travaux.

Je suis extrêmement reconnaissant vis-à-vis du SCAC (Service de Coopération et d'Action Culturelle) de l'ambassade de France à Dakar, du CEA-MITIC (Centre d'Excellence Africain en Mathématiques, Informatique et TIC) et de l'IRMAR (Institut de Recherche Mathématique de Rennes) d'avoir participé au financement de ma thèse et ma participation à des conférences de niveau international.

Ma gratitude va également à James LEDOUX qui m'a accueilli et m'a permi de bénéficier des moyens logistiques de fonctionnement des doctorants du Département de mathématiques de l'INSA de Rennes .

Je remercie les membres des écoles doctorales ED-MATISSE et ED-ST, au personnel administratif et plus particulièrement à Martine FIXOT, Aurore GOUIN et Claire DURAND pour leur aide administrative et gentillesse.

Je tiens à remercier tous les membres de l'Unité d'Epidémiologie et des Maladies Infectieuses, de l'équipe G4BBM (Groupe à 4ans Bio-informatique Biostatistique et Modélisation) de l'Institut Pasteur de Dakar et je témoigne ma reconnaissance particulière au Docteur Cheikh Loucoubar.

Pendant ces années de thèse, j'ai aussi eu le plaisir d'enseigner dans le cadre de mon monitorat. Je remercie donc en particulier Laurent MONIER, Jean-Louis MERRIEN et Olivier LEY pour m'avoir fait confiance et m'avoir accompagné dans mes premiers pas d'enseignant. Je n'oublie pas de remercier Pierrette CHAGNEAU, Mohamed CAMAR-EDDINE, Mounir HADDOU pour leurs aides, conseils, encouragements et leur sympathie.

J'exprime ma profonde sympathie à mes collgues du laboratoire et je leur souhaite une bonne continuation dans leurs travaux de recherches.

Pour finir, c'est à mes parents, à ma femme, à mes frères, à mes sœurs et belles sœurs que j'adresse un grand merci et dédie ce travail pour m'avoir toujours soutenu et encouragé.

Table des matières

Résumé

Les modèles de régressions à inflation de zéros constituent un outil très puissant pour l'analyse de données de comptage avec excès de zéros, émanant de divers domaines tels que l'épidémiologie, l'économie de la santé ou encore l'écologie. Cependant, l'étude théorique dans ces modèles attire encore peu d'attention. Ce manuscrit s'intéresse au problème de l'inférence dans des modèles de comptage à inflation de zéro.

Dans un premier temps, nous revenons sur la question de l'estimateur du maximum de vraisemblance dans le modèle binomial à inflation de zéro. D'abord nous montrons l'existence de l'estimateur du maximum de vraisemblance des paramètres dans ce modèle. Ensuite, nous démontrons la consistance de cet estimateur, et nous établissons sa normalité asymptotique. Puis, une étude de simulation exhaustive sur des tailles finies d'échantillons est menée pour évaluer la cohérence de nos résultats. Et pour finir, une application sur des données réelles d'économie de la santé a été conduite.

Dans un deuxième temps, nous proposons un nouveau modèle statistique d'analyse de la consommation de soins médicaux. Ce modèle permet, entre autres, d'identifier les causes du non-recours aux soins médicaux. Nous avons étudié rigoureusement les propriétés mathématiques du modèle. Ensuite nous avons mené une étude numérique approfondie à l'aide de simulations informatiques et enfin, nous l'avons appliqué à l'analyse d'une base de données recensant la consommation de soins de plusieurs milliers de patients aux USA.

Un dernier aspect de ces travaux de thèse a été de s'intéresser au problème de l'inférence dans le modèle binomial à inflation de zéro dans un contexte de don-

nées manquantes sur les covariables. Dans ce cas nous proposons la méthode de pondération par l'inverse des probabilités de sélection pour estimer les paramètres du modèle. Ensuite, nous établissons la consistance et la normalité asymptotique de l'estimateur proposé. Enfin, une étude de simulation sur plusieurs échantillons de tailles finies est conduite pour évaluer le comportement de l'estimateur.

Mots clés

Normalité asymptotique, consistance, données de comptage, excès de zéros, simulations, utilisation de soins de santé, logit multinomial, pondération par l'inverse de la probabilité de sélection.

Abstract

The zero-inflated regression models are a very powerful tool for the analysis of counting data with excess zeros from various areas such as epidemiology, health economics or ecology. However, the theoretical study in these models attracts little attention. This manuscript is interested in the problem of inference in zero-inflated count models.

At first, we return to the question of the maximum likelihood estimator in the zero-inflated binomial model. First we show the existence of the maximum likelihood estimator of the parameters in this model. Then, we demonstrate the consistency of this estimator, and let us establish its asymptotic normality. Then, a comprehensive simulation study finite sample sizes are conducted to evaluate the consistency of our results. Finally, an application on real health economics data has been conduct.

In a second time, we propose a new statistical analysis model of the consumption of medical care. This model allows, among other things, to identify the causes of the non-use of medical care. We have studied rigorously the mathematical properties of the model. Then, we carried out an exhaustive numerical study using computer simulations and finally applied to the analysis of a database on health care several thousand patients in the USA.

A final aspect of this work was to focus on the problem of inference in the zero inflation binomial model in the context of missing covariate data. In this case we propose the weighting method by the inverse of the selection probabilities to estimate the parameters of the model. Then, we establish the consistency and asymptotic normality of the estimator offers. Finally, a simulation study on several

samples of finite sizes is conducted to evaluate the behavior of the estimator.

Key words

Asymptotic normality, consistency, count data, excess of zeros, simulations, health-care utilization, multinomial logit, inverse-probability-weighting.

Abréviations & Notations

$\mathbf{P}(A)$	: Probabilité de l'événement A.
$\mathbb{E}(X)$	: L'espérance mathématique de la variable aléatoire X.
$\mathrm{var}(X)$	: Variance de la variable aléatoire X.
$\mathrm{cov}(X, Y)$	: La covariance des variables aléatoires X et Y.
$X_n \Longrightarrow Y$	: La suite de variables aléatoires $(X_n)_{n\geq 0}$ converge faiblement vers Y.
$X_n \xrightarrow{p.s.} Y$	: La suite de variables aléatoires $(X_n)_{n\geq 0}$ converge presque sûrement vers Y.
$\mathbb{N}$	: Ensemble des entiers naturels.
$\mathbb{N}^*$	: 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.
i.i.d.	: indépendantes et identiquement distribuées.
$\mathscr{M}(n \times p)$	: Ensemble des matrices réelles à n lignes et p colonnes.
I_p	: Matrice identité d'ordre p.
EMV	: Estimateur du maximum de vraisemblance.
GLMs	: Modèles Linéaires Généralisés
ZIB	: Zero-Inflated Binomial.
ZIP	: Zero-Inflated Poisson.
ZINB	: Zero-Inflated Negative Binomial.
ZIPO	: Zero-Inflated Proportional Odds.

CHAPITRE 1

Introduction générale

Depuis longtemps, les statisticiens ont reconnu les difficultés de travailler avec des données de comptages et/ou les données de présence/absence. Et les modèles de régressions de Poisson (voir [Nelder and Wedderburn, 1972, Frome et al., 1973, Griffith and Haining, 2006]) et binomial (voir [Nelder and Wedderburn, 1972, Nerlove and Press, 1973, Cox and Snell, 1989]) sont les outils usuels pour analyser, respectivement, les données de comptages et les données de présence/absence. Ils sont appliquées à de nombreux domaines tels que l'économie de la santé, l'épidémiologie ou encore les sciences de l'environnement. Mais ce sont des modèles qui ont une caractéristique particulière puisqu'ils n'ont qu'un paramètre à estimer : la *moyenne*. La variance est quant à elle considérée comme étant une fonction de la moyenne. Ces relations entre variance et moyenne trouvent tout leur sens dans un contexte théorique où par exemple l'ensemble des processus aboutissant à la présence/absence ou l'abondance de l'espèce est bien modélisé et est mesuré sans erreurs [Fisher, 1941, Hinde and Demétrio, 1998, Boes, 2010]. En pratique, il est illusoire de penser toutes les variables pouvant affecter ces processus puissent être intégrées dans le modèle, soit parce qu'elles sont inconnues, soit parce qu'elles sont impossibles à mesurer sur le terrain. Et cela peut conduire à une mauvaise spécification de la variance, la moyenne restant théoriquement peu affectée. La situation la plus fréquente est la *surdispersion*, c'est à dire lorsque la variance des observations est supérieure à celle suggérée par le modèle. De nombreuses solutions ont été proposées pour tenir compte de cette surdispersion (voir Hinde and Demétrio [1998] pour une vue d'ensemble) parmi lesquelles le déve-

loppement de nouveaux modèles, la régression binomiale négative et les modèles à inflation de zéros. Ces derniers modèles a pour l'instant attiré peu d'attention. L'objet central de cette thèse est de proposer une étude complète (théorique et numérique) de l'estimateur du maximum de vraisemblance dans des modèles de comptages à inflation de zéro.

La première partie est consacrée à de brefs rappels sur les modèles GLMs et les modèles de régression à inflation de zéro. Dans un premier chapitre de cette première partie, nous revenons sur la théorie des modèles linéaires généralisés. Nous nous intéressons en particulier à la spécification d'un modèle linéaire généralisé, à l'expression des moments et à l'estimation des paramètres de régression. Puis, nous rappelons les principaux résultats concernants les propriétés de l'estimateur du maximum de vraisemblance dans ces modèles. Et dans un deuxième chapitre de cette partie, nous définissons les modèles zéro- inflatés existants. Ensuite, nous présentons les résultats d'existence d'estimateur du maximum de vraisemblance et les propriétés asymptotiques de ces estimateurs.

La deuxième partie de cette thèse renferme les contributions originales de ce manuscrit où nous nous sommes intéressés au problème de l'inférence statistique dans des modèles à inflation de zéro. Dans le quatrième chapitre nous nous focalisons à l'étude des propriétés asymptotiques de l'estimateur du maximum de vraisemblance dans le modèle ZIB, qui jusqu'à présent était un aspect important ignoré. Pour cela, nous démontrons l'existence et la consistance de l'estimateur du maximum de vraisemblance et sa normalité asymptotique dans ce modèle. Et pour illustration nous effectuons des simulations sur des tailles finies d'échantillons et une application sur des données réelles d'économie de la santé. Le cinquième chapitre est une généralisation des travaux entamés dans le chapitre précédent. Mais aussi c'est un nouveau outil qui permet de prendre en compte simultanément plusieurs mesures de consommation de soins médicaux. Nous définissons le modèle

et nous établissons rigoureusement sa propriété d'identifiabilité et les propriétés asymptotiques de l'estimateur du maximum de vraisemblance proposé. Puis, nous proposons une application de ce modèle à l'évaluation de la demande de soins médicaux et à l'étude du renoncement aux soins. Dans le sixième chapitre nous nous intéressons au problème de l'inférence statistique dans le modèle ZIB en présence de données manquantes sur les covariables. Dans cette situation nous proposons une nouvelle approche d'estimation, basée sur la méthode de pondération par l'inverse de la probabilité de sélection. Une étude théorique et numérique des propriétés asymptotiques de l'estimateur suggéré a été établi.

La dernière partie comporte les différents scripts R de simulations et applications décrit dans la thèse.

Première partie

Rappels sur quelques modèles

CHAPITRE 2

Rappels sur les modèles linéaires généralisés

Sommaire

2.1 Introduction

Les **modèles linéaires généralisés** (GLMs) sont une classe de modèles statistiques de régression conçus pour traiter les problèmes où la loi de la variable réponse est décrite par un modèle paramétrique. Cette classe, formulée sous ce nom par Nelder and Wedderburn [1972] et popularisée par McCullagh and Nelder [1989], généralise le modèle linéaire classique qui est caractérisé par deux contraintes fortes : la linéarité de sa composante déterministe et la normalité de sa composante aléatoire. Dans ce chapitre, nous présentons d'abord dans la section 2.2 la théorie des GLMs (spécification, estimation, propriétés asymptotiques des estimateurs, qualité d'ajustement, tests et choix entre différents modèles). Puis, dans la section 2.3 nous donnons quelques compléments utiles en cas de surdispersion des données.

2.2 Théorie des modèles linéaires généralisés

2.2.1 Spécification d'un modèle linéaire généralisé

Disposant d'une variable à expliquer et de variables explicatives, trois composantes permettent de caractériser un GLM : la *distribution* de la variable à expliquer, le *prédicteur linéaire* et la *fonction de lien*.

2.2.1.1 Distribution du vecteur aléatoire observé

On suppose que l'échantillon statistique est constitué de n variables aléatoires $\{Y_i; i = 1, \ldots, n\}$ indépendantes admettant des distributions issues d'une structure exponentielle (voir McCullagh and Nelder [1989] ou Antoniadis et al. [1992] pour plus de détails). Pour une mesure dominante adaptée (mesure de *Lebesgue* pour une loi continue, mesure discrète combinaison de masses de *Dirac* pour une loi discrète) la famille de leurs densités par rapport à cette mesure s'écrit sous la

forme :

$$f_{Y_i}(y_i, \theta_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\} \tag{2.1}$$

où $\theta_i \in \mathbb{R}$ est appelé paramètre naturel, ou de position (ou paramètre canonique) et $\phi \in \mathbb{R}_+^*$ est un paramètre de dispersion, ou de nuisance (ou paramètre d'échelle). Les fonctions b et c sont spécifiques à chaque type de distribution. La fonction b est supposée deux fois dérivable de dérivée première inversible, dérivable et d'inverse dérivable. La fonction a_i est de la forme $a_i(\phi) = \frac{\phi}{w_i}$ pour certaines lois avec $w_i \in \mathbb{R}_+^*$ un poids connu associé à l'observation i. Dans la suite w_i sera fixé à 1 pour tout $i = 1, \ldots, n$ pour simplifier. Parmi les lois appartenant à la famille exponentielle on peut citer pour les lois discrètes Poisson, binomiale et pour les lois continues, les lois normales, Gamma, ...

Pour chacune des lois énumérées ci-dessus, le tableau suivant décrit, en fonction des paramètres usuels de la loi, les expressions du paramètre canonique θ_i, du paramètre ϕ et des fonctions b et a_i associées. La fonction c n'intervenant pas dans la suite, nous ne la mentionnons pas ici. Pour simplifier la lecture du tableau nous omettons l'indice i.

Loi	θ	$b(\theta)$	$\phi = a(\phi)$
$\mathscr{P}(\lambda)$	$\log(\lambda)$	e^θ	1
$\mathscr{B}(n,p)$	$\log\big(p/(1-p)\big)$	$n\log(1+e^\theta)$	1
$\mathscr{N}(\mu,\sigma^2)$	μ	$\theta^2/2$	σ^2
$\gamma(a,b)$	$-b$	$-a\log(-\theta)$	1

TABLE 2.1 – Exemples de modèle linéaire généralisé

2.2.1.2 Prédicteur linéaire

Comme dans les modèles linéaires, les variables interviennent linéairement dans la modélisation. Nous supposons des variables explicatives organisées dans la matrice $\mathbf{X}$, matrice du plan d'expérience d'ordre $n \times p$, où p ($p \leq n$) est le nombre de variables explicatives. Soit β un vecteur de p paramètres. Le prédicteur linéaire, **composante déterministe** du modèle, est le vecteur à n composantes :

$$\eta = X\beta^{\top}.$$

Le modèle sera dit **régulier** si la matrice $\mathbf{X}$ est de rang plein, c'est-à-dire si $rang(\mathbf{X}) = p$.

2.2.1.3 Fonction de lien

La troisième composante des GLMs exprime une relation fonctionnelle entre l'espérance de Y_i et la i-ème composante du prédicteur linéaire, c'est à dire pour tout $i = 1, \ldots, n$ on a :

$$\eta_i = g(\mathbb{E}(Y_i))$$

où g appelée fonction de lien est supposée monotone et différentiable.
Parmi toutes les fonctions de lien, celle qui permet d'égaler le prédicteur linéaire et le paramètre canonique est appelée **fonction de lien canonique**. Puisqu'on a la relation $\eta_i = g(b'(\theta_i))$, la fonction de lien canonique associée à une distribution donnée vérifie $g = b'^{-1}$. Dans le tableau suivant nous avons indiqué les fonctions de liens canoniques associées à quelques lois classiques .

Loi	$\mathscr{P}(\lambda)$	$\mathscr{B}(n,p)$	$\mathscr{N}(\mu,\sigma^2)$	$\gamma(a,b)$
$g(x)$	$\log(x)$	$\log(\frac{x}{1-x})$	x	$\frac{1}{x}$

Notons que dans les modèles linéaires gaussiens la fonction de lien canonique n'apparaît pas car elle est égale à l'identité.

2.2.2 Expression des moments

L'espérance et la variance de Y_i s'expriment en fonction des paramètres θ_i et ϕ et sont liées. En effet, soit $\ell_{[i]}(\theta_i, \phi, y_i) = \log\left(f(y_i, \theta_i, \phi)\right)$ la contribution de la i-ème observation à la log-vraisemblance. On a $\forall i \in \{1, \ldots, n\}$

$$\frac{\partial \ell_{[i]}}{\partial \theta_i} = \left[y_i - b^{'}(\theta_i)\right] / a_i(\phi) \quad et \quad \frac{\partial^2 \ell_{[i]}}{\partial \theta_i^2} = b^{''}(\theta_i)/a_i(\phi).$$

Pour les lois issues de structures exponentielles, les conditions de régularités vérifiées permettent d'écrire

$$\mathbb{E}\left(\frac{\partial \ell_{[i]}}{\partial \theta_i}\right) = 0 \quad et \quad \mathbb{E}\left(\frac{\partial^2 \ell_{[i]}}{\partial \theta_i^2}\right) = -\mathbb{E}\left(\left(\frac{\partial^2 \ell_{[i]}}{\partial \theta_i^2}\right)^2\right).$$

Alors

$$\mathbb{E}\left(Y_i\right) = b^{'}(\theta_i) \quad et \quad \mathrm{var}\left(Y_i\right) = b^{''}(\theta_i) a_i(\phi).$$

Cette dernière expression justifiant ainsi l'appellation de paramètre de dispersion pour ϕ lorsque a_i est la fonction identité. On a donc une relation directe entre l'espérance de Y_i, que nous noterons μ_i dans tout ce qui suit, et sa variance

$$\mathrm{var}\left(Y_i\right) = a_i(\phi) b^{''}((b^{'-1})(\mu_i)) = \phi\, b^{''}((b^{'-1})(\mu_i)).$$

2.2.3 Estimation des paramètres de régression

Dans cette partie nous nous intéressons à l'estimation du paramètre de régression β. La procédure consiste à rechercher la valeur $\widehat{\beta}$ de β qui maximise la vraisemblance ou plus précisément son logarithme noté ℓ_n . Nous supposerons que ϕ est connu, et que la fonction de lien utilisée est la fonction de lien canonique. Nous aborderons d'abord la question des équations de vraisemblance, nous donnerons ensuite un algorithme de résolution de ces équations.

2.2.3.1 Equations de vraisemblance

Sous l'hypothèse d'indépendance des coordonnées de Y et en tenant compte que θ dépend de β, la log-vraisemblance s'écrit :

$$\ell_n(\beta) = \sum_{i=1}^{n} \ell_{[i]}(\theta_i, \phi; y_i).$$

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

$$U_j = \frac{\partial \ell_n}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{\partial \ell_{[i]}}{\partial \beta_j}\right] = \sum_{i=1}^{n} \left[\frac{\partial \ell_{[i]}}{\partial \theta_i}.\frac{\partial \theta_i}{\partial \mu_i}.\frac{\partial \mu_i}{\partial \eta_i}.\frac{\partial \eta_i}{\partial \beta_j}\right].$$

Or

$$\begin{aligned}
&\frac{\partial \ell_{[i]}}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a_i(\phi)} \\
&\frac{\partial \theta_i}{\partial \mu_i} = 1/\frac{\partial \mu_i}{\partial \theta_i} = 1/b''(\theta_i) = \frac{a_i(\phi)}{\mathrm{var}(Y_i)} \\
&\frac{\partial \mu_i}{\partial \eta_i} \quad \text{dépend de la fonction de lien : } \; g(\mu_i) = \eta_i \\
&\frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \;\; car \;\; \eta_i = x_i^\top \beta
\end{aligned}$$

Alors les équations de la vraisemblance sont :

$$U_j = \sum_{i=1}^{n} \frac{(y_i - \mu_i)\, x_{ij}}{\mathrm{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \;\; j = 1, \cdots, p. \tag{2.2}$$

La résolution de 2.2 se fait d'une manière numérique faisant intervenir le Hessien (pour Newton-Raphson) ou la matrice d'information (pour les scores de Fisher). Nous allons nous restreindre à l'algorithme de Fisher appelé ici IRLS (Iterative Reweighted Least Square).

2.2.3.2 Algorithme de Fisher (Fisher scoring)

La procédure employée dans la plupart des résolutions d'équations non linéaires est basée sur l'algorithme de Newton-Raphson. Dans le cas des GLMs, on

utilise souvent une autre procédure, celle de l'**algorithme de Fisher** (voir Jennrich and Sampson [1976] pour plus de détails).

Soit $\beta^{(k)}$ la k-ième approximation pour l'estimateur du maximum de vraisemblance $\hat{\beta}$. Dans la méthode de Newton-Raphson on a :

$$\beta^{(k+1)} = \beta^{(k)} - \left(H^{(k)}\right)^{-1} q^{(k)}$$

où H est la matrice hessienne ayant pour élément $\frac{\partial^2 \ell_n(\beta)}{\partial \beta_j \partial \beta_h}$, q est le vecteur des dérivées ayant pour éléments $\frac{\partial \ell_n(\beta)}{\partial \beta_s}$; $H^{(k)}$ et $q^{(k)}$ sont évalués en $\beta = \beta^{(k)}$.
La formule de l'algorithme de Fisher scoring s'écrit comme suit :

$$\beta^{(k+1)} = \beta^{(k)} + \left(I_F(\beta^{(k)})\right)^{-1} q^{(k)}$$

où $I_F(\beta^{(k)})$, d'éléments $\left(-\mathbb{E}\left(\frac{\partial^2 \ell_n(\beta)}{\partial \beta_j \partial \beta_h}\right)\right)$, est la k-ème approximation de la matrice d'information de Fisher évaluée en $\beta = \beta^{(k)}$. On itère la procédure employée jusqu'à obtenir la stabilité. Par exemple jusqu'au moment où la valeur absolue de la différence entre les valeurs calculées pour le logarithme de la vraisemblance à deux étapes successives soit en deçà d'un seuil fixé à l'avance.
La méthode de Fisher scoring peut s'interpréter comme une succession de moindres carrés, pondérés par des poids qui changent à chaque itération. L'estimation de la variance-covariance est un sous-produit de cette méthode. Pour cette raison l'algorithme est appelée "moindres carrés repondérés itératifs" (ou IRLS en anglais), algorithme utilisé pour la résolution de nos équations (2.2). La matrice d'information de Fisher d'un GLM s'obtient comme suit :

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

car l'indépendance des Y_i implique que $\mathbb{E}\left[(Y_i - \mu_i)(Y_l - \mu_l)\right] = 0$ pour $i \neq l$, et $\mathbb{E}\left[(Y_i - \mu_i)^2\right] = \mathrm{var}(Y_i)$ sinon.
Alors la matrice d'information de Fisher est :

$$I_F = \mathbf{X}^\top \mathbf{W} \mathbf{X}$$

où $\mathbf{W}$ est la matrice diagonale de "pondération" : $\mathbf{w}_{ii} = \frac{1}{\mathrm{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$.
La méthode itérative suggère à l'étape $(k+1)$ la procédure suivante :

— Choix d'une valeur initiale $\widehat{\beta}^{(0)}$ proche de β (ex par la méthode des moments),

— Calcul des quantités

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

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

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

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

$$\mathbf{X}^\top \mathbf{W}^{(k)} z^{(k)},$$

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

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

et $\mathbf{W}^{(k)}$ la matrice diagonale des poids $\mathbf{w}_{ii}$ évaluée au point $\left(\mu_i^{(k)}, \eta_i^{(k)}\right)$ avec $\eta_i^{(k)} = \mathbf{X}_i^\top \beta^{(k)}$ et $\mu_i^{(k)} = g^{-1}\left(\eta_i^{(k)}\right)$.
Alors à l'itération $(k+1)$ l'algorithme se réécrit alors :

$$\beta^{(k+1)} = \left(\mathbf{X}^\top \mathbf{W}^{(k)} \mathbf{X}\right)^{-1} \left(\mathbf{X}^\top \mathbf{W}^{(k)} z^{(k)}\right).$$

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

$$\|\beta^{(k+1)} - \beta^{(k)}\| < \varepsilon \;\; ou \;\; \|\ell_n(\beta^{(k+1)}) - \ell_n(\beta^{(k)})\| < \varepsilon.$$

Remarque 2.2.1 *Cas particulier de fonction de lien canonique :* $\eta_i = \theta_i = x_i^\top \beta$

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

on a une simplification de l'équation du score :

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

d'où $\mathbf{X}^\top y = \mathbf{X}^\top \mu$. *Pour le modèle linéaire* $\mu = \mathbf{X}^\top \beta$, *la solution par maximisation de la vraisemblance* $\hat{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \left(\mathbf{X}^\top y\right)$ *coïncide avec la solution par minimisation des moindres carrés.*

2.2.4 Propriétés asymptotiques des estimateurs

Dans cette section nous présentons brièvement les résultats d'existence, de consistance et de normalité asymptotique de l'estimateur du maximum de vraisemblance d'un modèle linéaire généralisé régulier. Avant de poser les conditions de régularités nous donnons quelques notations et définitions.

Notations et définitions

— $H_n(\beta)$ désigne la matrice hessienne définie par $H_n(\beta) = \frac{\partial^2 \ell_n(\beta)}{\partial \beta \partial \beta^\top}$;

— $I_n(\beta)$ la matrice d'information de Fisher définie par :

$$I_n(\beta) = -\mathbb{E}\big(H_n(\beta)\big) = \mathbf{X}^\top W_\beta^{-1} \mathbf{X},$$

où $W_\beta = \mathrm{diag}\Big(\mathbb{V}(Y_1) g^\top(\mu_1)^2, \ldots, \mathbb{V}(Y_n) g^\top(\mu_n)^2\Big)$;

— $I_n^{1/2}(\beta)$, la matrice d'ordre n, symétrique définie positive vérifiant : $I_n^{1/2}(\beta) I_n^{1/2}(\beta)^\top = I_n(\beta)$ et d'inverse notée $I_n^{-1/2}(\beta)$.

On note encore β_0 la vraie valeur inconnue du paramètre, I_p la matrice identité en dimension p, $\lambda_{min}(A)$ la plus petite valeur propre de la matrice carrée A et $\lambda_{max}(A)$ sa plus grande valeur propre.

Supposons les conditions suivantes vérifiées :

(C1) il existe deux constantes c et ε strictement positives, un entier n_1 et un voisinage $V(\beta_0)$ de β_0 tels que pour tout $\beta \in V(\beta_0)$ et pour tout $n \geq n_1$:

$$\lambda_{min}(I_n(\beta)) \geq c\lambda_{max}^{\frac{1}{2}+\varepsilon}(I_n(\beta_0));$$

(C2) $\lambda_{min}(I_n(\beta_0)) \underset{n\to+\infty}{\longrightarrow} +\infty$;

(C3) $\forall \delta > 0, \forall \lambda \in \mathbb{R}^p$ tel que $\|\lambda\| = 1$, en posant $\beta_n = \beta_0 + \delta I_n^{-1/2}(\beta_0)'\lambda$ pour n assez grand, on a :

$$\max_{\beta \in V_n(\delta)} \|I_n^{-1/2}(\beta_0) I_n(\beta) I_n^{-1/2}(\beta_0)' - I_p\| \underset{n\to+\infty}{\longrightarrow} 0 \text{ en probabilité}$$

où $V_n(\delta)$ est le voisinage de β_0 défini par :

$$V_n(\delta) = \{\beta \in \mathscr{O} \text{ ouvert de } \mathbb{R}^p / \ \|I_n^{1/2}(\beta_0)'(\beta_0 - \beta)\| \leq \delta\}.$$

Théorème 2.2.1 *(Existence et consistance)*
*Si les hypothèses **(C1)** et **(C2)** sont vérifiées, il existe une suite $\left(\widehat{\beta}_n\right)_{n\geq 1}$ de variables aléatoires et une variable aléatoire n_2 à valeurs entières telles que :*

1. $\forall n \geq n_2$, $\mathbf{P}(\nabla \ell_n(\widehat{\beta}_n) = \mathbf{0}) = 1$
2. *la suite $\left(\widehat{\beta}_n\right)_{n\geq 1}$ converge presque sûrement vers β_0.*

Théorème 2.2.2 *(Normalité asymptotique)*
*Si les hypothèses **(C2)** et **(C3)** sont vérifiées, la suite des estimateurs de maximum de vraisemblance $\left(\widehat{\beta}_n\right)_{n\geq 1}$ est asymptotiquement gaussienne, c'est-à-dire :*

$$I_n^{1/2}(\beta_0)^\top(\widehat{\beta}_n - \beta_0) \underset{n\to+\infty}{\longrightarrow} \mathcal{N}(0, I_p) \text{ en loi.}$$

L'estimateur du maximum de vraisemblance $\widehat{\beta}_n$ est donc asymptotiquement gaussien.

Pour la démonstration de ces théorèmes le lecteur peut consulter Fahrmeir and Kaufmann [1985]. On peut retrouver aussi une étude des propriétés asymptotiques des estimateurs dans Antoniadis et al. [1992].

2.2.5 Qualité d'ajustement, tests et choix entre différents modèles

Une fois que les paramètres de la loi choisie sont estimés, leurs propriétés étudiées, il est fondamental d'étudier la qualité d'ajustement, de vérifier les hypothèses concernant les coefficients du modèle. Dans le cas où plusieurs modèles concurrents sont maintenant en compétition, des critères de choix du modèle le plus adéquat sont proposés.

2.2.5.1 Qualité d'ajustement

Deux statistiques sont souvent utilisées pour juger l'adéquation du modèle aux données :

- **la déviance normalisée** (scaled deviance) définie comme suit

$$\begin{aligned} D^* &= 2\left[\sum_{i=1}^{n}\left(\ell_{sat}(y_i) - \ell(y_i, \widehat{\beta}, \phi)\right)\right] \geq 0, \quad \text{pour } \phi \text{ fixé}, \\ &= 2\left(L_{sat} - L\right) \geq 0, \end{aligned}$$

  où L_{sat} est la log-vraisemblance du modèle saturé, c'est-à-dire le modèle possédant autant de paramètres que de variables ou observations, L est la log-vraisemblance du modèle estimé.

  Dans le cadre général des GLMs, si $a_i(\phi) = \phi/\omega_i$, alors la déviance est donnée par la formule ci-dessous :

$$\begin{aligned} D &= \sum_{i=1}^{n} 2\omega_i\left(y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - b(\widetilde{\theta}_i) + b(\widehat{\theta}_i)\right)/\phi, \\ &= D\left(y, \widehat{\mu}\right)/\phi, \end{aligned}$$

où $\widetilde{\theta}_i = \theta(y)$, $\widehat{\theta}_i = \theta(\widehat{\mu})$ et $D(y, \widehat{\mu})$ la déviance du modèle courant,

- **la statistique du Khi-deux de Pearson** est définie par :

$$\chi^2 = \sum_{i=1}^{n} (y_i - \widehat{\mu}_i)^2 / var(\widehat{\mu}_i)$$

Le Khi-deux normalisé de Pearson est égal à χ^2/ϕ.

Remarque 2.2.3 *Lorsque le modèle étudié est exact, ces deux statistiques suivent approximativement une loi du Khi-deux à $n - p$ degrés de liberté.*

2.2.5.2 Tests

La statistique de Wald et le rapport de vraisemblance sont les critères habituels utilisés pour tester la significativité des effets du modèle. Pour plus d'informations concernant ces statistiques, le lecteur pourra se reporter à McCullagh and Nelder [1989] ou Dobson [1990].

- **Le test de Wald.** Ce test est basé sur la forme quadratique faisant intervenir la matrice de covariance des paramètres, l'inverse de la matrice d'information observée. Si la matrice L, dite *contraste*, définit l'ensemble H_0 des hypothèses à tester sur les paramètres :

$$L^\top \beta = 0,$$

on montre que **la statistique de Wald** définie par

$$\chi^2_w = (L^\top \widehat{\beta})^\top (L \widehat{V} L^\top)^{-1} (L^\top \widehat{\beta}) \ converge \ en \ loi \ vers \ \chi^2(p),$$

avec $\widehat{V} = var(\widehat{\beta}) = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1}$.

Attention : le test de Wald approximatif peut ne pas être précis si le nombre d'observations est faible.

- **Le test du rapport de vraisemblance.** C'est un test qui permet de faire des comparaisons entre deux modèles M_1 (p paramètres) et M_2 (q paramètres)

emboités ($q \leq p$). Cela revient à tester l'hypothèse de **nullité** de q paramètres du modèle :

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_q \text{ avec } q \leq p$$

Sous l'hypothèse alternative, au moins un des paramètres $\beta_1, \ldots, \beta_q$ est non-nul.

L'hypothèse nulle peut être testée au moyen de la statistique :

$$\chi^2_L = (D_q - D_p), \text{ où } D_p \text{ et } D_q \text{ déviances respectives dans } M_1 \; et \; M_2,$$

qui suit asymptotiquement une loi de $\chi^2(p-q)$ sous H_0 pour les lois à un paramètre (binomial, Poisson, exponentielle) et une loi de Fisher pour les lois à deux paramètres (gaussienne). Ceci permet de tester la significativité de la diminution de la déviance par l'ajout de variables explicatives ou la prise en compte d'interactions.

Remarque 2.2.2 *La validité de la loi et donc du test n'est qu'asymptotique, il est nécessaire d'avoir un peu de recul quant aux conclusions. Lorsque les données sont binaires le test de* ***Hosmer et Lemeshow*** *est conseillé (voir Hosmer and Lemeshow [2000]).*

2.2.5.3 Choix entre différents modèles

Les critères de choix de modèles tels l'AIC (Akaike [1974]) ou le BIC (Schwarz [1978]) sont souvent utilisés pour comparer entre eux des modèles qui ne sont pas forcément emboîtés les uns dans les autres.

- **L'AIC (Akaike Information Criterion)** pour un modèle à p paramètres est défini par

$$AIC = -2\ell_n + 2p.$$

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

$$BIC = -2\ell_n + p\log(n).$$

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

2.3 Compléments sur les modèles GLMs

La plupart des situations rencontrées, dans le cas où la variable réponse est supposée binomiale, Poisson ou exponentielle, sont caractérisées par une *sur-dispersion* des données c'est-à-dire la variance observée est supérieure à la variance théorique. Dans ce cas on maximise la formule de quasi-vraisemblance pour estimer les paramètres.

Définition 2.3.1 *La quasi-vraisemblance est une fonction des paramètres évaluée aux observations, à l'instar de la vraisemblance. Elle est définie comme suit*

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

où Y est un vecteur d'observations y $i.i.d$ de moyenne $\mu = g^{-1}(\mathbf{X}\beta)$ et de variance $V(\mu)$.

Cette fonction possède trois propriétés communes avec la log-vraisemblance d'une loi exponentielle GLM :

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

où $V(\mu)$ est la fonction de variance associée à la loi exponentielle GLM où à la quasi-vraisemblance. La quasi-vraisemblance possède les mêmes propriétés que la vraisemblance sur l'espérance des deux premières dérivées et sur la variance de la première dérivée. Donc choisir une loi exponentielle GLM ou une quasi-vraisemblance revient aux mêmes estimations.

Cependant, des applications à des données réelles ont révélé que les données de comptages possèdent des distributions ayant des caractéristiques particulières comme la non-normalité, l'hétérogénéité des variances ainsi qu'un nombre important de zéros (voir Hilbe [2007]). Il est donc nécessaire d'utiliser des modèles appropriés afin d'obtenir des résultats non biaisés. D'où la naissance des modèles à inflation de zéros.

CHAPITRE 3

Modèles de régression à inflation de zéros

Sommaire

3.1 Introduction

Les modèles de régression à inflation de zéros sont des modèles souvent utilisés pour modéliser des données de comptage surdispersées lorsque la surdispersion 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 de la santé, l'assurance, l'agriculture, l'industrie, l'écologie. Aussi, les modèles adaptés à ce type de données ont été largement explorés dans la littérature. Lors d'un comptage, les zéros ont souvent un statut particulier qui peut prêter à confusion (Ridout et al. [1998]). 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). Ne pas tenir compte de ce facteur peut conduire à un cas particulier de surdispersion, l'inflation de zéros (voir Lambert [1992] ; Fong and Yip [1995] ; Mullahy [1997] ; Ridout et al. [1998] ; Tu [2002] ; Diop et al. [2011] ; Preisser et al. [2012]). Ce phénomène a particulièrement été mis en évidence dans le cas de la régression de Poisson et a conduit au développement de plusieurs outils pour en tenir compte. Pour traiter ce problème des approches ont été proposées parmi lesquelles la modélisation en deux parties (*hurdle model*, Mullahy [1986] ; *two-part models*, Heilbron [1994]) et l'autre approche est de considérer un mélange de deux modèles au lieu de les modéliser séparément. Cette dernière approche donne lieu aux modèles dits *zéro-excès* dont la version la plus commune est le modèle *zéro-inflated* (Lambert [1992] ; Greene [1994]). Plusieurs autres améliorations et extensions de ces modèles ont été documentées (voir Lukusa et al. [2016] ; Diop et al. [2011]). De manière générale, un modèle à inflation de zéros est un mélange entre une distribution dégénérée en zéro et une distribution de comptage standard (par exemple Poisson, Binomial, Binomial négatif).

3.2 Modèles de régression ZIP et ZINB

Les modèles de base pour données de comptages sont les modèles de *Poisson* et *binomial négatif*.

3.2.1 Modèles de régression de Poisson et binomial négatif

Le modèle de régression de Poisson (régression log-linéaire) Hilbe [2007] est souvent retenu pour expliquer une variable quantitative Y (par exemple un nombre d'événements) à valeurs entières. La probabilité que la variable Y prenne la valeur y_i $(y_i = 0, 1, 2, \ldots)$ est donnée par

$$\mathbf{P}(Y_i = y_i | \mathbf{X}_i = x_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots \tag{3.1}$$

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

On vérifie aisément que dans le modèle 3.1, l'espérance est égale à la variance $\mathbb{E}(Y_i|\mathbf{X}_i = x_i) = \mathrm{var}(Y_i|\mathbf{X}_i = x_i) = \lambda_i = e^{\beta^\top \mathbf{X}_i}$. La forme de la fonction exponentielle assure la non-négativité du paramètre de la moyenne λ_i.

L'hypothèse d'équidispersion dans ce modèle est très restrictive. Dans la pratique, du fait d'une abondance de valeurs nulles et/ou de la présence de quelques valeurs extrêmes, la variance est souvent supérieure à la moyenne. Dans ce cas, on parle d'une *sur-dispersion* (voir Cox [1983] ; Hinde and Demétrio [1998]) de la variable Y. Cette situation peut remettre en cause l'utilisation de ce modèle, par une sous-estimation des variances des paramètres du modèle. D'où l'idée d'utiliser un modèle de comptage alternatif, basé sur la loi binomiale négative, qui prend en compte cette sur-dispersion par l'introduction d'un paramètre supplémentaire α qui permet de capter l'hétérogénéité inobservée de la variable endogène (qui peut impliquer la sur-dispersion inobservée).

Dans un modèle de régression binomial négatif, on définit la probabilité pour que Y prenne la valeur y_i par

$$\mathbf{P}(Y_i = y_i|\mathbf{X}_i = x_i) = \frac{\Gamma(y_i + 1/\alpha)}{y_i!\Gamma(1/\alpha)}\Big(\frac{1}{1+\alpha\lambda_i}\Big)^{1/\alpha}\Big(\frac{\lambda_i}{1/\alpha+\lambda_i}\Big)^{y_i} \tag{3.2}$$

où α est un paramètre auxiliaire mesurant le degré de sur-dispersion. Cette loi a une moyenne conditionnelle λ_i et une variance conditionnelle $\lambda_i(1+\alpha\lambda_i)$. 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 paramètres dans le modèle,

— soit par le ratio $\chi^2/(n-p)$, où χ^2 correspond à la statistique du chi-deux de Pearson.

Si ces ratios sont supérieurs à 1, les données présentent une sur-dispersion (et une sous-dispersion si ces ratios sont inférieurs 1).

3.2.2 Modèles de régression ZIP et ZINB

Le phénomène d'inflation de zéro a été constaté pour la première fois sur des données de comptage. D'où la mise en place de nouveaux outils plus adaptés, comme les modèles de régression ZIP et ZINB, pour traiter ce genre de problème.

Pour une variable réponse $Y_i, i = 1, \ldots, n$, on dira que :
– Y_i est modélisée par un ZIP si sa distribution s'exprime comme suit :

$$\mathbf{P}(Y_i = y_i|\mathbf{X}_i, \mathbf{Z}_i) = \begin{cases} \pi_i + (1-\pi_i)\exp(-\lambda_i) & \text{si } y_i = 0 \\ \\ (1-\pi_i)\frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} & \text{si } y_i > 0 \end{cases} \tag{3.3}$$

avec

$$\mathbb{E}(Y_i|\mathbf{X}_i,\mathbf{Z}_i)=(1-\pi_i)\lambda_i \quad \text{et} \quad \text{var}(Y_i|\mathbf{X}_i,\mathbf{Z}_i)=(1-\pi_i)\lambda_i(1+\pi_i\lambda_i).$$

– Y_i est modélisée par un ZINB si sa distribution est donnée par :

$$\mathbf{P}(Y_i=y_i|\mathbf{X}_i,\mathbf{Z}_i)=\begin{cases}\pi_i+(1-\pi_i)(\frac{1}{1+\alpha\lambda_i})^{\alpha} & \text{si } y_i=0\\ (1-\pi_i)\frac{\Gamma(y_i+1/\alpha)}{\Gamma(1/\alpha)y_i!}\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right)^{y_i}\left(\frac{1}{1+\alpha\lambda_i}\right)^{1/\alpha} & \text{si } y_i>0\end{cases} \tag{3.4}$$

avec

$$\mathbb{E}(Y_i|\mathbf{X}_i,\mathbf{Z}_i)=(1-\pi_i)\lambda_i \quad \text{et} \quad \text{var}(Y_i|\mathbf{X}_i,\mathbf{Z}_i)=(1-\pi_i)\lambda_i(1+(\alpha+\pi_i)\lambda_i),$$

où α est un paramètre de sur-dispersion. Dans les deux cas π_i représente la probabilité d'inflation de zéro.
Comme pour les modèles de Poisson et Binomial Négatif, le modèle ZINB tend vers le modèle ZIP lorsque α tend vers zéro. Pour ces deux modèles (3.3)-(3.4), on suppose que la probabilité π_i et la moyenne conditionnelle λ_i sont respectivement modélisées par $logit(\pi_i)=\gamma^{\top}\mathbf{Z}_i$ et par $\log(\lambda_i)=\beta^{\top}\mathbf{X}_i$. Les vecteurs $\mathbf{X}_i\in\mathbb{R}^p$ et $\mathbf{Z}_i\in\mathbb{R}^q$ sont les covariables. $\beta\in\mathbb{R}^p$ et $\gamma\in\mathbb{R}^q$ sont les vecteurs des paramètres inconnus. Les covariables $\mathbf{X}_i$ et $\mathbf{Z}_i$ peuvent ou non avoir des composantes communes [Pradhan and Leung, 2006, Diop et al., 2011].

3.2.3 Estimation et propriétés asymptotiques du modèle ZIP

De nombreux auteurs ont proposé des méthodes d'estimation dans un contexte de régression de Poisson avec inflation de zéros. En règle générale, l'estimation du maximum de vraisemblance est utilisée pour estimer de tels modèles (voir Lambert [1992], Czado et al. [2007]). Cependant, il est bien connu que l'EMV est très sensible à la présence de valeurs aberrantes et peut devenir instable lorsque les composantes du mélange sont mal séparées. Pour pallier à ce problème,

Hall and Shen [2010] ont suggéré une nouvelle procédure d'estimation du modèle (3.3) dite "robust expectation-solution (RES) estimation" ou tout simplement l'*algorithme ES* (expectation-solution). Cet algorithme est une modification de l'algorithme *expectation-maximization* (EM) (pour plus de détails voir Dempster et al. [1977]) avec la propriété de robustesse. Dans cette partie, nous discutons brièvement de cet algorithme ES et des propriétés asymptotiques de l'estimateur sous certaines conditions. Nous considérons également que tous les individus n'ont pas forcément la même probabilité π d'appartenir au groupe des zéros.

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

$$\begin{aligned}\ell_n(y, \gamma, \beta) &= \sum_{i=1}^{n} \left\{ u_i \log \left[e^{\gamma^\top Z_i} + exp\left(-e^{\beta^\top X_i} \right) \right] \right\} \\ &\quad + \sum_{i=1}^{n} \left\{ (1 - u_i) \left(y_i \beta^\top X_i - e^{\beta^\top X_i} - \log(y!) \right) \right\} \\ &\quad - \sum_{i=1}^{n} \left\{ \log\left(1 + e^{\gamma^\top Z_i} \right) \right\}, \end{aligned} \tag{3.5}$$

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

$$\begin{aligned}\ell_n^c(y, v, \gamma, \beta) &= \sum_{i=1}^{n} \left\{ v_i \gamma^\top Z_i - \log(1 + e^{\gamma^\top Z_i}) \right\} \\ &\quad + \sum_{i=1}^{n} (v_i) \left\{ y_i \beta^\top X_i - e^{\beta^\top X_i} - \log(y!) \right\} \\ &= \ell_\gamma^c(\gamma, y, v) + \ell_\beta^c(\beta, y, v), \end{aligned} \tag{3.6}$$

où $v = (v_1, \ldots, v_n)^\top$.
Cette log-vraisemblance est sous une forme appropriée car ℓ_γ^c et ℓ_β^c peuvent être

maximisés séparément. Le principe de cet algorithme consiste à maximiser la fonction $\ell_n^c(y, v, \gamma, \beta)$ de manière itérative en commençant par une valeur initiale $(\beta^{(0)\top}, \gamma^{(0)\top})^\top$. À l'itération $r+1$, nous avons les deux étapes suivantes :

1. **Étape E** : estimer la variable v_i par son espérance conditionnelle aux observations $v_i^{(r)}$ sous les estimations courantes des paramètres $\beta^{(r)}$ et $\gamma^{(r)}$. Cette espérance est donnée par

$$v_i^{(r)} = \begin{cases} \left[1 + exp\left(-\gamma^{(r)\top} Z_i - e^{\beta^{(r)\top} X_i}\right)\right]^{-1} & \text{si } y_i = 0, \\ 0 & \text{si } y_i > 0, \end{cases} \tag{3.7}$$

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

$$\frac{1}{n}\sum_{i=1}^{n}\left\{v_i^{(r)} - \pi_i\right\} Z_i = 0. \tag{3.8}$$

$$\frac{1}{n}\sum_{i=1}^{n}(1 - v_i^{(r)})\left\{y_i - e^{\beta^\top X_i}\right\} X_i = 0. \tag{3.9}$$

Dans l'approche RES, Hall and Shen [2010] proposent de remplacer les équations (3.8) et (3.9) 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 al. [2000] liées à l'algorithme ES et de Carroll et al. [1995], Hall and Shen [2010] ont montré le résultat suivant (qui généralise le `théorème 1` dans Czado et al. [2007] dans le sens où $\psi = (\beta^\top, \gamma^\top)^\top \in \mathbb{R}^{p+q}$) :

Théorème 3.2.1 *Si l'algorithme RES converge, alors il existe une suite de variables aléatoires $\widehat{\psi}_n$ telles que*

(i) $\widehat{\psi}_n \xrightarrow{\mathbf{P}} \psi_0$ *quand* $n \to \infty$ *(consistance),*

(ii) $\sqrt{n}(\widehat{\psi}_n - \psi_0) \xrightarrow{\mathscr{L}} \mathscr{N}(0, \boldsymbol{V}(\psi_0))$ *quand* $n \to \infty$ *(normalité asymptotique).*

où l'expression $\mathbf{V}(\psi_0))$ de la variance asymptotique est donnée dans Hall and Shen [2010].
Des auteurs comme Lam et al. [2006] et He et al. [2010] ont étendu ce modèle ZIP respectivement dans le cadre semi-paramétrique et doublement semi-paramétrique et ont établi les résultats de consistance et de normalité asymptotique des estimateurs proposés.

L'étude des propriétés asymptotiques dans le modèle ZINB peut se faire de manière similaire à celle effectuée précédemment dans le modèle ZIP. Pour plus de détails le lecteur intéressé peut se reporter à Hilbe [2007], Czado et al. [2007] et Mwalili et al. [2014].

3.3 Le modèle de régression ZIB

3.3.1 Régression binomiale

La régression binomiale ou régression logistique binaire fut la première méthode utilisée, notamment en marketing pour le *scoring* et *épidémiologie*, pour aborder la modélisation d'une variable binaire (nombre de succès pour n_i essais) ou de Bernoulli (avec $n_i = 1$) : absence ou présence d'une pathologie, ...
Ce modèle appartient à la famille des modèles linéaires généralisés et partagent à ce titre beaucoup d'aspects : estimation par maximisation de la vraisemblance, statistiques de test suivant asymptotiquement des lois du chi-deux, calcul des résidus, critère pénalisé pour la sélection de modèle, ...

On considère, pour $i = 1 \ldots, I$, diffèrentes valeurs fixées $x_{i1}, \ldots, x_{ip}$ des variables

explicatives $X_1, \ldots, X_p$. Ces dernières peuvent être des variables quantitatives ou qualitatives. Pour chaque groupe, on réalise n_i observations ($n = \sum_{i=1}^{I} n_i$) de la variable réponse binaire Y (succès-échec). On suppose que toutes les observations sont indépendantes et qu'à l'intérieur d'un même groupe, tous les individus ont la même probabilité de succès. La variable $Y = \sum_{i=1}^{n_i} y_i$ est distribuée selon une loi binomiale $\mathscr{B}(n_i, \pi_i)$ dont la fonction de densité s'écrit

$$\mathbf{P}(Y = y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}.$$

où la probabilité π_i de succès est modélisée par une fonction de lien `logit` :

$$logit(\pi_i) = \beta^\top x_i, \quad i = 1, \ldots, I.$$

Le vecteur des paramètres est estimé par maximisation de la log-vraisemblance. Il n'y a pas de solution analytique, celle-ci est obtenue par des méthodes numériques itératives (algorithme de Newton Raphson).

Dans la pratique, il n'est pas toujours garanti que les observations binaires successives soient indépendantes. La modélisation pourrait s'étendre dans le cas où la variable réponse Y prend $K (K > 2)$ modalités (voir Agresti [2002]).

3.3.2 Spécification du modèle ZIB

Le modèle de régression Binomial zéro-inflaté (ZIB) a été utilisé en premier par Kemp and Kemp [1988], mais ce n'est que vers les années 2000 que Hall [2000] et Vieira et al. [2000] l'ont introduit de manière beaucoup plus claire et ont donné quelques applications détaillées dans le cadre de données réelles. En considérant les mêmes notations que Hall [2000], le modèle ZIB est une distribution à deux états défini comme suit :

$$Y_i \sim \begin{cases} 0 & \text{avec une probabilité } p_i \\ \text{Binomiale}(n_i, \pi_i) & \text{avec une probabilité } 1 - p_i. \end{cases} \tag{3.10}$$

Ce qui implique que

$$Y_i = \begin{cases} 0 & \text{avec une probabilité } p_i + (1-p_i)(1-\pi_i)^{n_i} \\ k & \text{avec une probabilité } (1-p_i)\binom{n_i}{k}\pi_i^{n_i}(1-\pi_i)^{n_i-k}, k = 1, 2, ..., n_i, \end{cases} \tag{3.11}$$

avec

$$\mathbb{E}(Y_i) = (1-p_i)n_i\pi_i, \quad \text{et} \quad \text{var}(Y_i) = (1-p_i)n_i\pi_i\Big(1-\pi_i(1-p_in_i)\Big).$$

Les paramètres $p = (p_1, \ldots, p_n)$ et $\pi = (\pi_1, \ldots, \pi_n)$ sont respectivement modélisés via une fonction de lien `logit`, $logit(p) = \gamma^\top \mathbf{W}$ et $logit(\pi) = \beta^\top \mathbf{X}$ où $\mathbf{W} \in \mathbb{R}^q$ et $\mathbf{X} \in \mathbb{R}^p$ sont les vecteurs de covariables, n est 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, $\gamma \in \mathbb{R}^q$ et $\beta \in \mathbb{R}^p$ sont les paramètres de régression. La log-vraisemblance du modèle basée sur les observations $(Y_i, \mathbf{X}_i, \mathbf{W}_i)$, $i = 1, \ldots, n$, est donnée par

$$\begin{aligned} l_n(\psi) &= \sum_{i=1}^{n} \Big\{ J_i \log\Big(e^{\gamma^\top \mathbf{W}_i} + (1+e^{\beta^\top \mathbf{X}_i})^{-m_i}\Big) - \log\Big(1+e^{\gamma^\top \mathbf{W}_i}\Big) \\ &\qquad + (1-J_i)\Big[Y_i\beta^\top \mathbf{X}_i - m_i \log\Big(1+e^{\beta^\top \mathbf{X}_i}\Big)\Big]\Big\}, \\ &:= \sum_{i=1}^{n} l_{[i]}(\psi), \end{aligned} \tag{3.12}$$

où $J_i := 1_{\{Y_i=0\}}$ (voir Hall, 2000)

Les estimations des paramètres de γ et β peuvent être déterminées via la méthode du maximum de vraisemblance ou via l'algorithme EM comme décrit dans le modèle ZIP précédemment.

Les conditions d'identifiabilité dans un mélange de régression binomiales ont été données par Teicher [1960], Teicher [1963], Blischke [1978] et Margolin et al. [1989].

3.4 Le modèle de régression ZIPO

il s'agit d'un modèle développé par Kelley and Anderson [2008] pour modéliser des données sur la consommation d'alcool. La spécification du modèle et son estimation sont similaires à celles étudiées ci-dessus dans les modèles ZIP, ZINB et ZIB. Soit $Y_i, 1 \leq i \leq n$ une variable ordinale à J niveaux (Agresti [2002]). Les probabilités cumulatives sont données par $\gamma_j = \mathbf{P}(Y_i \leq j), j = 0, 1, \ldots, J$. On a

$$Y_i \sim \begin{cases} 0 & \text{avec } p_i \\ \text{Multinomiale}(1, \gamma_{0,i}, \ldots, \gamma_{J,i}) & \text{avec } 1 - p_i, \end{cases} \tag{3.13}$$

d'où

$$Y_i = \begin{cases} 0 & \text{avec } p_i + (1 - p_i)\gamma_{0,i} \\ j & \text{avec } (1 - p_i)(\gamma_{j,i} - \gamma_{j-1,i}), \end{cases} \tag{3.14}$$

où $p = (p_1, \ldots, p_n)$ et $\gamma = (\gamma_0, \ldots, \gamma_J)$ sont modélisés respectivement par

$$logit(p) = \theta^\top \mathbf{Z} \quad \text{et} \quad logit(\pi) = \beta^\top \mathbf{X}.$$

Deuxième partie

Contributions originales de cette thèse

CHAPITRE 4

Propriétés asymptotiques de l'estimateur du maximum de vraisemblance dans le modèle de régression ZIB.

Sommaire

Dans ce chapitre, nous établissons rigoureusement les propriétés asymptotiques de l'estimateur du maximum de vraisemblance des paramètres d'un modèle de régression ZIB. L'existence et la normalité asymptotique sont démontrées. Un estimateur consistant de la matrice de variance covariance est également fourni. Une étude de simulation approfondie est menée pour évaluer les propriétés des estimateurs proposés sur des tailles finies d'échantillons. Les résultats obtenus dans cette étude confirment les propriétés mathématiques établies théoriquement. Enfin, une application sur des données issues d'un problème en économie de la santé est proposée pour illustration.

Ce chapitre est publié dans la revue *Communications in Statistics-Theory and Methods*, Vol 46, No. 20, pages 9930-9948, 2017.

Authors : DIALLO A. O, DIOP A., and DUPUY J.-F.,
Asymptotic properties of the maximum likelihood estimator in zero-inflated binomial regression.
Communications in Statistics-Theory and Methods 46, No. 20, 9930-9948, 2017.

Abstract

The zero-inflated binomial (ZIB) regression model was proposed by Hall [2000] to account for excess zeros in binomial regression. Since then, the model has been applied in various fields, such as ecology and epidemiology. In these applications, maximum likelihood estimation (MLE) is used to derive parameter estimates. However, theoretical properties of the MLE in ZIB regression have not yet been rigorously established. The current paper fills this gap and thus provides a rigorous basis for applying the model. Consistency and asymptotic normality of the MLE in ZIB regression are proved. A consistent estimator of the asymptotic variance-covariance matrix of the MLE is also provided. Finite-sample behavior of the estimator is assessed via simulations. Finally, an analysis of a data set in the field of health economics illustrates the paper.

keywords **: asymptotic normality, consistency, count data, excess of zeros, simulations.**

4.1 Introduction

Zero-inflated regression models have attracted a great deal of attention over the past two decades. These models account for excess zeros in count data by mixing a degenerate distribution with point mass of one at zero with a standard count regression model, such as Poisson, negative binomial or binomial. The zero-inflated Poisson (ZIP) regression model was proposed by Lambert [1992] and fur-

ther developed by Dietz and Böhning [2000], Li [2011], Lim et al. [2014] and Monod [2014], among many others. Zero-inflated negative binomial (ZINB) regression was proposed by Ridout et al. [2001], see also Moghimbeigi et al. [2008], Mwalili et al. [2014], Garay et al. [2011]. The zero-inflated binomial (ZIB) regression model was discussed by Hall [2000], Vieira et al. [2000] and Hall and Berenhaut [2002]. Since their introduction, these models have been applied in numerous fields, such as agriculture, econometrics, epidemiology, insurance, species abundance, terrorism study, traffic safety research... In particular, ZIB regression model was recently used in dental caries epidemiology [Gilthorpe et al., 2009, Matranga et al., 2013]. This increasing interest for zero-inflated models renders necessary to establish theoretical properties for their parameter estimates. So far, however, mathematical considerations in zero-inflated models (such as asymptotic properties of maximum likelihood estimates) have attracted much less attention than applications. Moreover, the existing literature essentially focuses on the ZIP regression model. See, for example, Min and Czado [2010] who establish asymptotic properties of maximum likelihood estimates (MLE) in a zero-modified generalized Poisson regression model. But to the best of our knowledge, no asymptotic results have been provided for the zero-inflated binomial regression model. In this paper, we investigate this issue.

In the ZIB model proposed by Hall [2000], the individual observation is a bounded count which can be thought of as the number of successes occurring out of a finite number of trials. The mixing probabilities and success probabilities are assumed to follow logistic regression models with parameters γ and β respectively. We provide rigorous proofs of consistency and asymptotic normality of the maximum likelihood estimators of γ and β. We also conduct a simulation study to evaluate finite-sample performance of these estimators. All these results provide a firm basis for making statistical inference in the zero-inflated binomial regression model.

The remainder of this paper is organized as follows. In Section 4.2, we recall the definition of the ZIB model, we describe maximum likelihood estimation and we introduce some useful notations. In Section 4.3, we state some regularity conditions and establish consistency and asymptotic normality of the maximum likelihood estimator in ZIB regression. Section 4.4 reports results of the simulation study. An application of ZIB model to the analysis of health-care utilization by elderlies in United States is described in Section 4.5. A discussion and some perspectives are provided in Section 4.6.

4.2 Zero-inflated binomial regression model

In this section, we briefly recall the definition of the ZIB model, we describe maximum likelihood estimation in ZIB regression and we introduce some useful notations.

4.2.1 Model and estimation

Let $(Z_i, \mathbf{X}_i, \mathbf{W}_i)$, $i = 1, \ldots, n$ be independent random vectors defined on the probability space $(\Omega, \mathscr{C}, \mathbb{P})$. For every $i = 1, \ldots, n$, the response variable Z_i is generated from the following two-state process :

$$Z_i \sim \begin{cases} 0 & \text{with probability } p_i, \\ \mathscr{B}(m_i, \pi_i) & \text{with probability } 1 - p_i, \end{cases} \tag{4.1}$$

where $\mathscr{B}(m, \pi)$ denotes the binomial distribution with size m and success (or event) probability π. Thus, Z_i follows a standard binomial distribution with probability $1 - p_i$. The first state (also called zero state) occurs with probability p_i. No success can occur in the zero state. The ZIB model reduces to a standard binomial distribution if $p_i = 0$, while $p_i > 0$ leads to zero-inflation. In Hall [2000], the mixing probabilities p_i and event probabilities π_i $(i = 1, \ldots, n)$ are modeled by the

logistic regression models

$$\text{logit}(p_i) = \gamma^\top \mathbf{W}_i \tag{4.2}$$

and

$$\text{logit}(\pi_i) = \beta^\top \mathbf{X}_i \tag{4.3}$$

respectively, where $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})^\top$ and $\mathbf{W}_i = (1, W_{i2}, \ldots, W_{iq})^\top$ are random vectors of predictors or covariates (both categorical and continuous covariates are allowed) and $\top$ denotes the transpose operator. Let $\psi = (\beta^\top, \gamma^\top)^\top$ be the unknown k-dimensional ($k := p + q$) parameter in models (4.1)-(4.3). The log-likelihood of ψ, based on observations $(Z_i, \mathbf{X}_i, \mathbf{W}_i)$, $i = 1, \ldots, n$, is

$$\begin{aligned} l_n(\psi) &= \sum_{i=1}^{n} \Big\{ J_i \log\left(e^{\gamma^\top \mathbf{W}_i} + (1 + e^{\beta^\top \mathbf{X}_i})^{-m_i}\right) - \log\left(1 + e^{\gamma^\top \mathbf{W}_i}\right) \\ &\qquad + (1 - J_i)\left[Z_i \beta^\top \mathbf{X}_i - m_i \log\left(1 + e^{\beta^\top \mathbf{X}_i}\right)\right]\Big\}, \\ &:= \sum_{i=1}^{n} l_{[i]}(\psi), \end{aligned} \tag{4.4}$$

where $J_i := 1_{\{Z_i = 0\}}$ (see Hall, 2000). The maximum likelihood estimator $\widehat{\psi}_n := (\widehat{\beta}_n^\top, \widehat{\gamma}_n^\top)^\top$ of ψ is the solution of the k-dimensional score equation

$$\dot{l}_n(\psi) := \frac{\partial l_n(\psi)}{\partial \psi} = 0. \tag{4.5}$$

In what follows, we establish consistency and asymptotic normality of $\widehat{\psi}_n$. First, we need to introduce some further notations.

4.2.2 Some further notations

Define first the $(p \times n)$ and $(q \times n)$ matrices

$$\mathbb{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1p} & X_{2p} & \cdots & X_{np} \end{pmatrix} \qquad \text{and} \qquad \mathbb{W} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ W_{12} & W_{22} & \cdots & W_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ W_{1q} & W_{2q} & \cdots & W_{nq} \end{pmatrix},$$

and let $\mathbb{V}$ be the $(k \times 2n)$ block-matrix defined as

$$\mathbb{V} = \begin{bmatrix} \mathbb{X} & 0_{p,n} \\ 0_{q,n} & \mathbb{W} \end{bmatrix},$$

where $0_{a,b}$ denotes the $(a \times b)$ matrix whose components are all equal to zero. Let also $C(\psi) = (C_j(\psi))_{1 \leq j \leq 2n}$ be the $2n$-dimensional column vector defined by

$$C(\psi) = (A_1(\psi), \ldots, A_n(\psi), B_1(\psi), \ldots, B_n(\psi))^\top,$$

where for every $i = 1, \ldots, n$,

$$\begin{aligned} A_i(\psi) &= -J_i \frac{m_i e^{\beta^\top \mathbf{x}_i}}{e^{\gamma^\top \mathbf{w}_i}(h_i(\beta))^{m_i+1} + h_i(\beta)} + (1 - J_i)\left(Z_i - \frac{m_i e^{\beta^\top \mathbf{x}_i}}{h_i(\beta)}\right), \\ B_i(\psi) &= \frac{J_i e^{\gamma^\top \mathbf{w}_i}(h_i(\beta))^{m_i}}{e^{\gamma^\top \mathbf{w}_i}(h_i(\beta))^{m_i} + 1} - \frac{e^{\gamma^\top \mathbf{w}_i}}{1 + e^{\gamma^\top \mathbf{w}_i}}, \end{aligned}$$

and $h_i(\beta) := 1 + e^{\beta^\top \mathbf{x}_i}$. Finally, let $k_i(\psi) := e^{\gamma^\top \mathbf{w}_i}(h_i(\beta))^{m_i+1} + h_i(\beta)$, $i = 1, \ldots, n$. Then, some simple algebra shows that the score equation (4.5) can be rewritten as

$$\dot{l}_n(\psi) = \mathbb{V}C(\psi) = 0.$$

If $A = (A_{ij})_{1 \leq i \leq a, 1 \leq j \leq b}$ denotes some $(a \times b)$ matrix, we will denote by $A_{\bullet j}$ its j-th column $(j = 1, \ldots, b)$ that is, $A_{\bullet j} = (A_{1j}, \ldots, A_{aj})^\top$. Then, it will be useful to rewrite the score vector as

$$\dot{l}_n(\psi) = \sum_{j=1}^{2n} \mathbb{V}_{\bullet j} C_j(\psi).$$

We shall further denote by $\ddot{l}_n(\psi)$ the $(k \times k)$ matrix of second derivatives of $l_n(\psi)$ that is, $\ddot{l}_n(\psi) = \partial^2 l_n(\psi)/\partial\psi\partial\psi^\top$. Let $\mathbb{D}(\psi) = (\mathbb{D}_{ij}(\psi))_{1 \leq i,j \leq 2n}$ be the $(2n \times 2n)$ block matrix defined as

$$\mathbb{D}(\psi) = \begin{bmatrix} \mathbb{D}_1(\psi) & \mathbb{D}_3(\psi) \\ \mathbb{D}_3(\psi) & \mathbb{D}_2(\psi) \end{bmatrix},$$

where $\mathbb{D}_1(\psi), \mathbb{D}_2(\psi)$ and $\mathbb{D}_3(\psi)$ are $(n \times n)$ diagonal matrices, with i-th diagonal elements ($i = 1, \ldots, n$) respectively given by

$$\mathbb{D}_{1,ii}(\psi) = \frac{J_i m_i e^{\beta^\top \mathbf{X}_i}}{(k_i(\psi))^2} \left(k_i(\psi) - e^{\beta^\top \mathbf{X}_i} \left[e^{\gamma^\top \mathbf{W}_i} (m_i + 1)(h_i(\beta))^{m_i} + 1 \right] \right) + \frac{m_i (1 - J_i) e^{\beta^\top \mathbf{X}_i}}{(h_i(\beta))^2},$$

$$\mathbb{D}_{2,ii}(\psi) = \frac{J_i e^{\gamma^\top \mathbf{W}_i} (h_i(\beta))^{m_i+1}}{(k_i(\psi))^2} \left(e^{\gamma^\top \mathbf{W}_i} (h_i(\beta))^{m_i+1} - k_i(\psi) \right) + \frac{e^{\gamma^\top \mathbf{W}_i}}{\left(1 + e^{\gamma^\top \mathbf{W}_i}\right)^2},$$

$$\mathbb{D}_{3,ii}(\psi) = -\frac{J_i m_i e^{\beta^\top \mathbf{X}_i + \gamma^\top \mathbf{W}_i} (h_i(\beta))^{m_i+1}}{(k_i(\psi))^2}.$$

Then, some tedious albeit not difficult algebra shows that $\ddot{l}_n(\psi)$ can be expressed as

$$\ddot{l}_n(\psi) = -\mathbb{V}\mathbb{D}(\psi)\mathbb{V}^\top.$$

Note that $C(\psi), \mathbb{V}$ and $\mathbb{D}(\psi)$ depend on n. However, in order to simplify notations, n will not be used as a lower index for these quantities.

In the next section, we establish rigorously the existence, consistency and asymptotic normality of the maximum likelihood estimator $\widehat{\psi}_n$ in the ZIB models (4.1)-(4.3).

4.3 Regularity conditions and asymptotic properties of the MLE

We first state some regularity conditions that will be needed for proving our asymptotic results :

C1 The covariates are bounded that is, there exist compact sets $\mathscr{X} \subset \mathbb{R}^p$ and $\mathscr{W} \subset \mathbb{R}^q$ such that $\mathbf{X}_i \in \mathscr{X}$ and $\mathbf{W}_i \in \mathscr{W}$ for every $i = 1, 2, \ldots$ For every $i = 1, 2, \ldots$, $j = 2, \ldots, p$ and $\ell = 2, \ldots, q$, var$[X_{ij}] > 0$ and var$[W_{i\ell}] > 0$. For every $i = 1, 2, \ldots$, the X_{ij} $(j = 1, \ldots, p)$ are linearly independent and the $W_{i\ell}$ $(\ell = 1, \ldots, q)$ are linearly independent.

C2 The true parameter value $\psi_0 := (\beta_0^\top, \gamma_0^\top)^\top$ lies in the interior of some known compact set $\mathscr{B} \times \mathscr{G} \subset \mathbb{R}^p \times \mathbb{R}^q$.

C3 The Hessian matrix $\ddot{l}_n(\psi)$ is negative definite and of full rank, for every $n = 1, 2, \ldots$, and $\frac{1}{n}\ddot{l}_n(\psi)$ converges to a negative definite matrix. Let λ_n and Λ_n be respectively the smallest and largest eigenvalues of $\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top$. There exists a finite positive constant c_1 such that $\Lambda_n/\lambda_n < c_1$ for every $n = 1, 2, \ldots$ The matrix $\mathbb{V}\mathbb{V}^\top$ is positive definite for every $n = 1, 2, \ldots$ and its smallest eigenvalue $\widetilde{\lambda}_n$ tends to $+\infty$ as $n \to \infty$.

C4 For every $i = 1, \ldots, n$, $m_i \in \{2, \ldots, M\}$ for some finite integer value M.

In what follows, the space $\mathbb{R}^k$ of k-dimensional (column) vectors will be provided with the Euclidean norm $\|\cdot\|_2$ and the space of $(k \times k)$ real matrices will be provided with the norm $|||A|||_2 := \max_{\|x\|_2=1} \|Ax\|_2$ (for notations simplicity, we will use $\|\cdot\|$ for both norms). Recall that for a symmetric real $(k \times k)$-matrix A with eigenvalues $\lambda_1, \ldots, \lambda_k$, $\|A\| = \max_i |\lambda_i|$.

We first prove existence and consistency of $\widehat{\psi}_n$:

Theorem 4.3.1 (Existence and consistency) *Under conditions C1-C4, the maximum likelihood estimator $\widehat{\psi}_n$ exists almost surely as $n \to \infty$ and converges almost surely to ψ_0.*

Proof of Theorem 4.3.1. The proof is inspired by the proof of consistency of the MLE in usual logistic regression [Gouriéroux and Monfort, 1981] but technical details are different. We first prove an intermediate technical lemma.

Lemma 4.3.2 *Let $\phi_n : \mathbb{R}^k \longrightarrow \mathbb{R}^k$ be defined as : $\phi_n(\psi) = \psi + (\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top)^{-1}\dot{l}_n(\psi)$. Then there exists an open ball $B(\psi_0, r)$ (with $r > 0$) and a constant c $(0 < c < 1)$ such that :*

$$\left\|\phi_n(\psi) - \phi_n(\widetilde{\psi})\right\| \leq c\|\psi - \widetilde{\psi}\| \text{ for all } \psi, \widetilde{\psi} \in B(\psi_0, r). \tag{4.6}$$

Proof of Lemma 4.3.2. The property (4.6) holds if we can prove that $\left\|\frac{\partial \phi_n(\psi)}{\partial \psi^\top}\right\| \leq c$ for all $\psi \in B(\psi_0, r)$.

Letting I_k be the identity matrix of order k, we have :

$$\begin{aligned}
\left\|\frac{\partial \phi_n(\psi)}{\partial \psi^\top}\right\| &= \left\|I_k + (\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top)^{-1}\ddot{l}_n(\psi)\right\| \\
&= \left\|I_k - (\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top)^{-1}\mathbb{V}\mathbb{D}(\psi)\mathbb{V}^\top\right\| \\
&= \left\|(\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top)^{-1}\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{V}^\top\right\| \\
&\leq \left\|(\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top)^{-1}\right\| \left\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{V}^\top\right\| \\
&= \lambda_n^{-1} \left\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{V}^\top\right\|.
\end{aligned}$$

Now, let $\mathscr{I}$ denote the set of indices $\{(i,j) \in \{1,2,\ldots,2n\}^2 \text{ such that } \mathbb{D}_{ij}(\psi_0) \neq 0\}$. Then the following holds :

$$\begin{aligned}
\left\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{V}^\top\right\| &= \left\|\sum_{i=1}^{2n}\sum_{j=1}^{2n} \mathbb{V}_{\bullet i}\mathbb{V}_{\bullet j}^\top(\mathbb{D}_{ij}(\psi) - \mathbb{D}_{ij}(\psi_0))\right\| \\
&\leq \sum_{(i,j)\in\mathscr{I}} \left\|\mathbb{V}_{\bullet i}\mathbb{V}_{\bullet j}^\top \mathbb{D}_{ij}(\psi_0)\right\| \left|\frac{\mathbb{D}_{ij}(\psi) - \mathbb{D}_{ij}(\psi_0)}{\mathbb{D}_{ij}(\psi_0)}\right|.
\end{aligned}$$

From C1 and C2, there exists a constant c_2 ($c_2 > 0$) such that $|\mathbb{D}_{ij}(\psi_0)| > c_2$ for every $(i,j) \in \mathscr{I}$. For example, consider the case where $\mathbb{D}_{ij}(\psi_0)$ coincides with some $\mathbb{D}_{3,\ell\ell}(\psi_0)$, for $\ell \in \{1,\ldots,n\}$. For every $\psi \in \mathscr{B} \times \mathscr{G}$, we have :

$$|\mathbb{D}_{3,\ell\ell}(\psi)| = \frac{m_\ell e^{\beta^\top \mathbf{x}_\ell + \gamma^\top \mathbf{w}_\ell}(1 + e^{\beta^\top \mathbf{x}_\ell})^{m_\ell - 1}}{\left(1 + e^{\gamma^\top \mathbf{w}_\ell}(1 + e^{\beta^\top \mathbf{x}_\ell})^{m_\ell}\right)^2} > \frac{m_{\mathbf{X}}^{m_\ell} m_{\mathbf{W}}}{(1 + M_{\mathbf{W}}(1 + M_{\mathbf{X}})^{m_\ell})^2},$$

where $m_{\mathbf{X}} := \min_{\beta,\mathbf{X}} e^{\beta^\top \mathbf{X}}$, $m_{\mathbf{W}} := \min_{\gamma,\mathbf{W}} e^{\gamma^\top \mathbf{W}}$, $M_{\mathbf{X}} := \max_{\beta,\mathbf{X}} e^{\beta^\top \mathbf{X}}$, $M_{\mathbf{W}} := \max_{\gamma,\mathbf{W}} e^{\gamma^\top \mathbf{W}}$. By C1, C2 and C4, there exists a positive constant d_3 such that $\frac{m_{\mathbf{X}}^{m_\ell} m_{\mathbf{W}}}{(1+M_{\mathbf{W}}(1+M_{\mathbf{X}})^{m_\ell})^2} > d_3$. Using similar arguments, we obtain that for every $\psi \in \mathrm{B} \times \mathrm{G}$, $|\mathbb{D}_{1,\ell\ell}(\psi)| > d_1$ and $|\mathbb{D}_{2,\ell\ell}(\psi)| > d_2$ for some $d_1, d_2 > 0$. Letting $c_2 = \min_{1\leq i\leq 3} d_i$, we conclude that $|\mathbb{D}_{ij}(\psi_0)| > c_2$ for every $(i,j) \in \mathscr{I}$. Moreover, $\mathbb{D}_{ij}(\cdot)$ is uniformly continuous on $\mathscr{B} \times \mathscr{G}$ (by Heine theorem) thus for every $\varepsilon > 0$, there exists a positive number r such that for all $\psi \in B(\psi_0, r)$,

$|\mathbb{D}_{ij}(\psi) - \mathbb{D}_{ij}(\psi_0)| < \varepsilon$. It follows that

$$\begin{aligned}
\left\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{V}^\top\right\| &\leq \frac{\varepsilon}{c_2} \sum_{(i,j)\in\mathscr{I}} \left\|\mathbb{V}_{\bullet i}\mathbb{V}_{\bullet j}^\top \mathbb{D}_{ij}(\psi_0)\right\| \\
&\leq \frac{\varepsilon}{c_2} \text{trace}\left(\sum_{(i,j)\in\mathscr{I}} \mathbb{V}_{\bullet i}\mathbb{V}_{\bullet j}^\top \mathbb{D}_{ij}(\psi_0)\right) \\
&= \frac{\varepsilon}{c_2} \text{trace}\left(\sum_{i=1}^{2n}\sum_{j=1}^{2n} \mathbb{V}_{\bullet i}\mathbb{V}_{\bullet j}^\top \mathbb{D}_{ij}(\psi_0)\right) \\
&= \frac{\varepsilon}{c_2} \text{trace}\left(\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top\right) \\
&\leq \frac{\varepsilon}{c_2} k\Lambda_n.
\end{aligned}$$

This in turn implies that $\left\|\frac{\partial\phi_n(\psi)}{\partial\psi^\top}\right\| \leq \frac{\varepsilon k\Lambda_n}{c_2\lambda_n} < \frac{\varepsilon k c_1}{c_2}$. Now, choosing $\varepsilon = c\frac{c_2}{kc_1}$ with $0 < c < 1$, we get that $\left\|\frac{\partial\phi_n(\psi)}{\partial\psi^\top}\right\| \leq c$ for all $\psi \in B(\psi_0, r)$, which concludes the proof.

We now turn to proof of Theorem 4.3.1. Define the function $\psi \mapsto \eta_n(\psi)$ by $\eta_n(\psi) := \psi - \phi_n(\psi) = -(\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top)^{-1}\dot{l}_n(\psi)$. Then $\eta_n(\psi_0)$ converges almost surely to 0 as $n \to \infty$. To see this, note that

$$\eta_n(\psi_0) = (\ddot{l}_n(\psi_0))^{-1} \cdot \dot{l}_n(\psi_0) = \left(\frac{1}{n}\ddot{l}_n(\psi_0)\right)^{-1} \cdot \left(\frac{1}{n}\dot{l}_n(\psi_0)\right).$$

By C3, $\left(\frac{1}{n}\ddot{l}_n(\psi_0)\right)^{-1}$ converges to some matrix Σ. Moreover,

$$\frac{1}{n}\dot{l}_n(\psi_0) = \frac{1}{n}\mathbb{V}C(\psi_0) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_{i1}A_i(\psi_0) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n X_{ip}A_i(\psi_0) \\ \frac{1}{n}\sum_{i=1}^n W_{i1}B_i(\psi_0) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n W_{iq}B_i(\psi_0) \end{pmatrix}$$

converges to 0 almost surely as $n \to \infty$. To see this, note that for every $i = 1, \ldots, n$

and $j = 1, \ldots, p$:

$$\mathbb{E}[X_{ij}A_i(\psi_0)] = \mathbb{E}[\mathbb{E}[X_{ij}A_i(\psi_0)|\mathbf{X}_i, \mathbf{W}_i]] = \mathbb{E}[X_{ij}\mathbb{E}[A_i(\psi_0)|\mathbf{X}_i, \mathbf{W}_i]],$$

and

$$\begin{aligned}
\mathbb{E}[A_i(\psi_0)|\mathbf{X}_i, \mathbf{W}_i] &= \mathbb{E}\left[-J_i\frac{m_i e^{\beta_0^\top \mathbf{X}_i}}{e^{\gamma_0^\top \mathbf{W}_i}(h_i(\beta_0))^{m_i+1} + h_i(\beta_0)} + (1-J_i)\left(Z_i - \frac{m_i e^{\beta_0^\top \mathbf{X}_i}}{h_i(\beta_0)}\right)\middle| \mathbf{X}_i, \mathbf{W}_i\right] \\
&= -\frac{m_i e^{\beta_0^\top \mathbf{X}_i}}{e^{\gamma_0^\top \mathbf{W}_i}(h_i(\beta_0))^{m_i+1} + h_i(\beta_0)}\mathbb{E}\left[J_i|\mathbf{X}_i, \mathbf{W}_i\right] + \mathbb{E}\left[(1-J_i)Z_i|\mathbf{X}_i, \mathbf{W}_i\right] \\
&\quad -\frac{m_i e^{\beta_0^\top \mathbf{X}_i}}{h_i(\beta_0)}\mathbb{E}\left[1-J_i|\mathbf{X}_i, \mathbf{W}_i\right].
\end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{E}\left[J_i|\mathbf{X}_i, \mathbf{W}_i\right] &= \mathbb{P}(Z_i = 0|\mathbf{X}_i, \mathbf{W}_i) \\
&= p_i + (1-\pi_i)^{m_i}(1-p_i) \\
&= \frac{e^{\gamma_0^\top \mathbf{W}_i}}{1 + e^{\gamma_0^\top \mathbf{W}_i}} + \frac{1}{(h_i(\beta_0))^{m_i}\left(1 + e^{\gamma_0^\top \mathbf{W}_i}\right)} \\
&= \frac{e^{\gamma_0^\top \mathbf{W}_i}(h_i(\beta_0))^{m_i+1} + h_i(\beta_0)}{(h_i(\beta_0))^{m_i+1}\left(1 + e^{\gamma_0^\top \mathbf{W}_i}\right)}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[(1-J_i)Z_i|\mathbf{X}_i, \mathbf{W}_i\right] &= m_i(1-p_i)\pi_i \\
&= \frac{m_i e^{\beta_0^\top \mathbf{X}_i}}{h_i(\beta_0)\left(1 + e^{\gamma_0^\top \mathbf{W}_i}\right)}.
\end{aligned}$$

Thus

$$\begin{aligned}
\mathbb{E}[A_i(\psi_0)|\mathbf{X}_i,\mathbf{W}_i] &= -\frac{m_i e^{\beta_0^\top \mathbf{X}_i}}{(h_i(\beta_0))^{m_i+1}\left(1+e^{\gamma_0^\top \mathbf{W}_i}\right)} + \frac{m_i e^{\beta_0^\top \mathbf{X}_i}}{h_i(\beta_0)\left(1+e^{\gamma_0^\top \mathbf{W}_i}\right)} - \frac{m_i e^{\beta_0^\top \mathbf{X}_i}}{h_i(\beta_0)} \\
&\quad + \frac{m_i e^{\beta_0^\top \mathbf{X}_i}}{h_i(\beta_0)} \times \frac{e^{\gamma_0^\top \mathbf{W}_i}(h_i(\beta_0))^{m_i+1} + h_i(\beta_0)}{(h_i(\beta_0))^{m_i+1}\left(1+e^{\gamma_0^\top \mathbf{W}_i}\right)} \\
&= \frac{m_i e^{\beta_0^\top \mathbf{X}_i}}{h_i(\beta_0)\left(1+e^{\gamma_0^\top \mathbf{W}_i}\right)} - \frac{m_i e^{\beta_0^\top \mathbf{X}_i}}{h_i(\beta_0)} + \frac{m_i e^{\beta_0^\top \mathbf{X}_i} e^{\gamma_0^\top \mathbf{W}_i}}{h_i(\beta_0)\left(1+e^{\gamma_0^\top \mathbf{W}_i}\right)} \\
&= \frac{m_i e^{\beta_0^\top \mathbf{X}_i}}{h_i(\beta_0)\left(1+e^{\gamma_0^\top \mathbf{W}_i}\right)}\left[1-(1+e^{\gamma_0^\top \mathbf{W}_i})\right] + \frac{m_i e^{\beta_0^\top \mathbf{X}_i} e^{\gamma_0^\top \mathbf{W}_i}}{h_i(\beta_0)\left(1+e^{\gamma_0^\top \mathbf{W}_i}\right)} \\
&= -\frac{m_i e^{\beta_0^\top \mathbf{X}_i} e^{\gamma_0^\top \mathbf{W}_i}}{h_i(\beta_0)\left(1+e^{\gamma_0^\top \mathbf{W}_i}\right)} + \frac{m_i e^{\beta_0^\top \mathbf{X}_i} e^{\gamma_0^\top \mathbf{W}_i}}{h_i(\beta_0)\left(1+e^{\gamma_0^\top \mathbf{W}_i}\right)} \\
&= 0.
\end{aligned}$$

It follows that $\mathbb{E}[X_{ij}A_i(\psi_0)] = 0$. Similarly, for every $i = 1,\ldots,n$ and $\ell = 1,\ldots,q$, we have :

$$\mathbb{E}[W_{i\ell}B_i(\psi_0)] = \mathbb{E}[\mathbb{E}[W_{i\ell}B_i(\psi_0)|\mathbf{X}_i,\mathbf{W}_i]] = \mathbb{E}[W_{i\ell}\mathbb{E}[B_i(\psi_0)|\mathbf{X}_i,\mathbf{W}_i]]$$

and

$$\begin{aligned}
\mathbb{E}[B_i(\psi_0)|\mathbf{X}_i,\mathbf{W}_i] &= \frac{e^{\gamma_0^\top \mathbf{W}_i}(h_i(\beta_0))^{m_i}}{e^{\gamma_0^\top \mathbf{W}_i}(h_i(\beta_0))^{m_i}+1}\mathbb{E}\left[J_i|\mathbf{X}_i,\mathbf{W}_i\right] - \frac{e^{\gamma_0^\top \mathbf{W}_i}}{1+e^{\gamma_0^\top \mathbf{W}_i}} \\
&= 0,
\end{aligned}$$

thus $\mathbb{E}[W_{i\ell}B_i(\psi_0)] = 0$. Moreover, for every $i = 1,\ldots,n$ and $\ell = 1,\ldots,q$,

$$\begin{aligned}
\text{var}(W_{i\ell}B_i(\psi_0)) &= \mathbb{E}[\text{var}(W_{i\ell}B_i(\psi_0)|\mathbf{X}_i,\mathbf{W}_i)] + \text{var}(\mathbb{E}[W_{i\ell}B_i(\psi_0)|\mathbf{X}_i,\mathbf{W}_i]) \\
&= \mathbb{E}[W_{i\ell}^2\text{var}(B_i(\psi_0)|\mathbf{X}_i,\mathbf{W}_i)] \\
&= \mathbb{E}\left[W_{i\ell}^2\left(\frac{e^{\gamma_0^\top \mathbf{W}_i}(h_i(\beta_0))^{m_i}}{e^{\gamma_0^\top \mathbf{W}_i}(h_i(\beta_0))^{m_i}+1}\right)^2 \text{var}(J_i|\mathbf{X}_i,\mathbf{W}_i)\right] \\
&\leq \mathbb{E}\left[W_{i\ell}^2\left(\frac{e^{\gamma_0^\top \mathbf{W}_i}(h_i(\beta_0))^{m_i}}{e^{\gamma_0^\top \mathbf{W}_i}(h_i(\beta_0))^{m_i}+1}\right)^2\right].
\end{aligned}$$

Therefore, by C1, C2 and C4, there exists a finite constant c_3 such that $\text{var}(W_{i\ell}B_i(\psi_0)) \leq c_3$. Similarly, there exists a finite constant c_4 such that $\text{var}(X_{ij}A_i(\psi_0)) \leq c_4$ for every $i = 1, \ldots, n$ and $j = 1, \ldots, p$. It follows that

$$\sum_{i=1}^{\infty} \frac{\text{var}(W_{i\ell}B_i(\psi_0))}{i^2} \leq c_3 \sum_{i=1}^{\infty} \frac{1}{i^2} < \infty$$

and

$$\sum_{i=1}^{\infty} \frac{\text{var}(X_{ij}A_i(\psi_0))}{i^2} \leq c_4 \sum_{i=1}^{\infty} \frac{1}{i^2} < \infty.$$

Kolmogorov's strong law of large numbers (see Jiang [2010], Theorem 6.7) implies that for every $j = 1, \ldots, p$,

$$\frac{1}{n} \sum_{i=1}^{n} \{X_{ij}A_i(\psi_0) - \mathbb{E}\left[X_{ij}A_i(\psi_0)\right]\} = \frac{1}{n} \sum_{i=1}^{n} X_{ij}A_i(\psi_0)$$

converges almost surely to 0. Similarly, for every $\ell = 1, \ldots, q$, $\frac{1}{n}\sum_{i=1}^{n} W_{i\ell}B_i(\psi_0)$ converges almost surely to 0. Finally, $\frac{1}{n}\dot{l}_n(\psi_0)$ and $\eta_n(\psi_0)$ converge almost surely to 0 as $n \to \infty$.

Now, let ε be an arbitrary positive value. Almost sure convergence of $\eta_n(\psi_0)$ implies that for almost every $\omega \in \Omega$, there exists an integer $n(\varepsilon, \omega)$ such that for any $n \geq n(\varepsilon, \omega)$, $\|\eta_n(\psi_0)\| \leq \varepsilon$ or equivalently, $0 \in B(\eta_n(\psi_0), \varepsilon)$. In particular, let $\varepsilon = (1-c)s$ with $0 < c < 1$ such as in Lemma 4.3.2. Since ϕ_n satisfies the Lipschitz condition (4.6), Lemma 2 of Gouriéroux and Monfort [1981] ensures that there exists an element of $B(\psi_0, s)$ (let $\widehat{\psi}_n$ denote this element) such that $\eta_n(\widehat{\psi}_n) = 0$ that is,

$$(\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^{\top})^{-1}\dot{l}_n(\widehat{\psi}_n) = 0.$$

Condition C3 implies that $\dot{l}_n(\widehat{\psi}_n) = 0$ and that $\widehat{\psi}_n$ is the unique maximizer of l_n.

To summarize, we have shown that for almost every $\omega \in \Omega$ and for every $s > 0$, there exists an integer value $n(s, \omega)$ such that if $n \geq n(s, \omega)$, then the maximum

likelihood estimator $\widehat{\psi}_n$ exists, and $\|\widehat{\psi}_n - \psi_0\| \leq s$ (that is, $\widehat{\psi}_n$ converges almost surely to ψ_0). □

We now turn to asymptotic normality of the maximum likelihood estimator in the ZIB regression model.

Theorem 4.3.3 (Asymptotic normality) *Let $\widehat{\Sigma}_n := \mathbb{VD}(\widehat{\psi}_n)\mathbb{V}^\top$. Then, under conditions C1-C4, $\widehat{\Sigma}_n^{\frac{1}{2}}(\widehat{\psi}_n - \psi_0)$ converges in distribution, as $n \to \infty$, to the Gaussian vector $\mathscr{N}(0, I_k)$.*

Proof of Theorem 4.3.3. A Taylor expansion of the score function yields

$$0 = \dot{l}_n(\widehat{\psi}_n) = \dot{l}_n(\psi_0) + \ddot{l}_n(\widetilde{\psi}_n)(\widehat{\psi}_n - \psi_0),$$

where $\widetilde{\psi}_n$ lies between $\widehat{\psi}_n$ and ψ_0. Thus, $\dot{l}_n(\psi_0) = -\ddot{l}_n(\widetilde{\psi}_n)(\widehat{\psi}_n - \psi_0)$.
Letting $\widetilde{\Sigma}_n := -\ddot{l}_n(\widetilde{\psi}_n) = \mathbb{VD}(\widetilde{\psi}_n)\mathbb{V}^\top$ and $\Sigma_{n,0} := \mathbb{VD}(\psi_0)\mathbb{V}^\top$, we have :

$$\widehat{\Sigma}_n^{\frac{1}{2}}(\widehat{\psi}_n - \psi_0) = \left[\widehat{\Sigma}_n^{\frac{1}{2}}\widetilde{\Sigma}_n^{-\frac{1}{2}}\right]\left[\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}}\right]\Sigma_{n,0}^{-\frac{1}{2}}\left(\widetilde{\Sigma}_n(\widehat{\psi}_n - \psi_0)\right). \tag{4.7}$$

The terms $[\widehat{\Sigma}_n^{\frac{1}{2}}\widetilde{\Sigma}_n^{-\frac{1}{2}}]$ and $[\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}}]$ in (4.7) converge almost surely to I_k. To see this, we show for example that $\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}} - I_k\| \xrightarrow{a.s.} 0$ as $n \to \infty$. First, note that

$$\left\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}} - I_k\right\| \leq \Lambda_n^{\frac{1}{2}}\left\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\right\|\left\|\Lambda_n^{-\frac{1}{2}}\left(\Sigma_{n,0}^{\frac{1}{2}} - \widetilde{\Sigma}_n^{\frac{1}{2}}\right)\right\|, \tag{4.8}$$

and

$$\Lambda_n^{-1}\left\|\Sigma_{n,0} - \widetilde{\Sigma}_n\right\| = \Lambda_n^{-1}\left\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\widetilde{\psi}_n))\mathbb{V}^\top\right\|.$$

By Theorem 4.3.1, $\widetilde{\psi}_n$ converges almost surely to ψ_0. Let $\omega \in \Omega$ be outside the negligible set where this convergence does not hold. By the same arguments as in proof of Lemma 4.3.2, for every $\varepsilon > 0$, there exists $n(\varepsilon, \omega) \in \mathbb{N}$ such that if $n \geq n(\varepsilon, \omega)$, then $\Lambda_n^{-1}\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\widetilde{\psi}_n))\mathbb{V}^\top\| \leq \varepsilon$. Thus $\Lambda_n^{-1}\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\widetilde{\psi}_n))\mathbb{V}^\top\|$ converges almost surely to 0. By continuity of the map $A \mapsto A^{\frac{1}{2}}$, $\|\Lambda_n^{-\frac{1}{2}}(\Sigma_{n,0}^{\frac{1}{2}} - \widetilde{\Sigma}_n^{\frac{1}{2}})\|$ converges almost surely to 0. Moreover, for n sufficiently large, there exists

$0 < c_5 < \infty$ such that almost surely, $\Lambda_n^{\frac{1}{2}} \|\widetilde{\Sigma}_n^{-\frac{1}{2}}\| \leq c_5 \Lambda_n^{\frac{1}{2}} / \lambda_n^{\frac{1}{2}} < c_5 c_1^{\frac{1}{2}}$ (by condition C3). Thus $\|\widetilde{\Sigma}_n^{-\frac{1}{2}} \Sigma_{n,0}^{\frac{1}{2}} - I_k\|$ converges almost surely to 0. Almost sure convergence of $\|\widehat{\Sigma}_n^{\frac{1}{2}} \widetilde{\Sigma}_n^{-\frac{1}{2}} - I_k\|$ to 0 follows by similar arguments.

It remains us to show that $\Sigma_{n,0}^{-\frac{1}{2}}(\widetilde{\Sigma}_n(\widehat{\psi}_n - \psi_0))$ converges in distribution to the Gaussian vector $\mathscr{N}(0, I_k)$. Note that $\Sigma_{n,0}^{-\frac{1}{2}}(\widetilde{\Sigma}_n(\widehat{\psi}_n - \psi_0)) = \Sigma_{n,0}^{-\frac{1}{2}} \sum_{j=1}^{2n} \mathbb{V}_{\bullet j} C_j(\psi_0)$. Thus, by Eicker [1963], this convergence holds if we can check that the following conditions are fulfilled :

1. $\max_{1 \leq j \leq 2n} \mathbb{V}_{\bullet j}^\top (\mathbb{V}\mathbb{V}^\top)^{-1} \mathbb{V}_{\bullet j} \to 0$ as $n \to \infty$,
2. $\sup_{1 \leq j \leq 2n} \mathbb{E}[C_j(\psi_0)^2 1_{\{|C_j(\psi_0)| > c\}}] \to 0$ as $c \to \infty$,
3. $\inf_{1 \leq j \leq 2n} \mathbb{E}[C_j(\psi_0)^2] > 0$.

Condition 1. follows by noting that

$$0 < \max_{1 \leq j \leq 2n} \mathbb{V}_{\bullet j}^\top (\mathbb{V}\mathbb{V}^\top)^{-1} \mathbb{V}_{\bullet j} \leq \max_{1 \leq j \leq 2n} \|\mathbb{V}_{\bullet j}\|^2 \|(\mathbb{V}\mathbb{V}^\top)^{-1}\| = \max_{1 \leq j \leq 2n} \|\mathbb{V}_{\bullet j}\|^2 / \widetilde{\lambda}_n$$

and that $\|\mathbb{V}_{\bullet j}\|$ is bounded, by C1. Moreover, $1/\tilde{\lambda}_n$ tends to 0 as $n \to \infty$ by C3. Condition 2. follows by noting that the $C_j(\psi_0)$, $j = 1, \ldots, 2n$ are bounded under C1, C2, C4. Finally, we note that $\mathbb{E}[C_j(\psi_0)^2] = \text{var}(C_j(\psi_0))$ since $\mathbb{E}[C_j(\psi_0)] = 0$, $j = 1, \ldots, 2n$. If $j \in \{n+1, \ldots, 2n\}$, $C_j(\psi_0) = B_{j'}(\psi_0)$, with $j' = j - n$. Then $\text{var}(C_j(\psi_0)) = \text{var}(B_{j'}(\psi_0)) = \mathbb{E}[\text{var}(B_{j'}(\psi_0)|\mathbf{X}_{j'}, \mathbf{W}_{j'})] + \text{var}(\mathbb{E}[B_{j'}(\psi_0)|\mathbf{X}_{j'}, \mathbf{W}_{j'}]) = \mathbb{E}[\text{var}(B_{j'}(\psi_0)|\mathbf{X}_{j'}, \mathbf{W}_{j'})]$. Now,

$$\begin{aligned}
\text{var}(B_{j'}(\psi_0)|\mathbf{X}_{j'}, \mathbf{W}_{j'}) &= \left(\frac{e^{\gamma_0^\top \mathbf{W}_{j'}} (h_{j'}(\beta_0))^{m_{j'}}}{e^{\gamma_0^\top \mathbf{W}_{j'}} (h_{j'}(\beta_0))^{m_{j'}} + 1} \right)^2 \text{var}(J_{j'}|\mathbf{X}_{j'}, \mathbf{W}_{j'}) \\
&= \left(\frac{e^{\gamma_0^\top \mathbf{W}_{j'}} (h_{j'}(\beta_0))^{m_{j'}}}{e^{\gamma_0^\top \mathbf{W}_{j'}} (h_{j'}(\beta_0))^{m_{j'}} + 1} \right)^2 \mathbb{P}(Z_{j'} = 0|\mathbf{X}_{j'}, \mathbf{W}_{j'})(1 - \mathbb{P}(Z_{j'} = 0|\mathbf{X}_{j'}, \mathbf{W}_{j'})) \\
&= \left(\frac{e^{\gamma_0^\top \mathbf{W}_{j'}} (h_{j'}(\beta_0))^{m_{j'}}}{e^{\gamma_0^\top \mathbf{W}_{j'}} (h_{j'}(\beta_0))^{m_{j'}} + 1} \right)^2 (p_{j'} + (1 - \pi_{j'})^{m_{j'}} (1 - p_{j'}))(1 - p_{j'}) \\
&\qquad \times (1 - (1 - \pi_{j'})^{m_{j'}}),
\end{aligned}$$

and thus, $\text{var}(B_{j'}(\psi_0)|\mathbf{X}_{j'}, \mathbf{W}_{j'}) > 0$ for every $j' = 1, \ldots, n$ by C1, C2, C4. It follows that $\text{var}(C_j(\psi_0)) > 0$ for every $j = n+1, \ldots, 2n$. By similar arguments,

var$(C_j(\psi_0)) > 0$ for every $j = 1, \ldots, 2n$ and condition 3. is satisfied.

To summarize, we have proved that $\Sigma_{n,0}^{-\frac{1}{2}}(\widetilde{\Sigma}_n(\widehat{\psi}_n - \psi_0))$ converges in distribution to $\mathscr{N}(0, I_k)$. This result combined with Slutsky's theorem and equation (4.7) implies that $\widehat{\Sigma}_n^{\frac{1}{2}}(\widehat{\psi}_n - \psi_0)$ converges in distribution to $\mathscr{N}(0, I_k)$. □

4.4 Simulation study

In this section, we assess finite-sample properties of the maximum likelihood estimator $\widehat{\psi}_n$.

4.4.1 Study design

We generate data from the following ZIB regression model :

$$\text{logit}(\pi_i) = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7}$$

and

$$\text{logit}(p_i) = \gamma_1 W_{i1} + \gamma_2 W_{i2} + \gamma_3 W_{i3} + \gamma_4 W_{i4} + \gamma_5 W_{i5},$$

where $X_{i1} = 1$ and the $X_{i2}, \ldots, X_{i7}$ are independently drawn from normal $\mathscr{N}(0, 1)$, uniform $\mathscr{U}(2, 5)$, normal $\mathscr{N}(1, 1.5)$, exponential $\mathscr{E}(1)$, binomial $\mathscr{B}(1, 0.3)$ and normal $\mathscr{N}(-1, 1)$ distributions respectively. We let $W_{i1} = 1$ and W_{i4} and W_{i5} be independently drawn from normal $\mathscr{N}(-1, 1)$ and binomial $\mathscr{B}(1, 0.5)$ distributions respectively. The linear predictors in logit(π_i) and logit(p_i) are allowed to share some common terms by letting $W_{i2} = X_{i2}$ and $W_{i3} = X_{i6}$. The regression parameter β is chosen as $\beta = (-0.3, 1.2, 0.5, -0.75, -1, 0.8, 0)^\top$. The regression parameter γ is chosen as :

— case 1 : $\gamma = (-0.55, -0.7, -1, 0.45, 0)^\top$

— case 2 : $\gamma = (0.25, -0.4, 0.8, 0.45, 0)^\top$

We consider several sample sizes, namely $n = 150, 300, 500$. The numbers m_i are allowed to vary across subjects, with $m_i \in \{4, 5, 6\}$.
Let $(n_4, n_5, n_6) = (\text{card}\{i : m_i = 4\}, \text{card}\{i : m_i = 5\}, \text{card}\{i : m_i = 6\})$.
For $n = 150$, we let $(n_4, n_5, n_6) = (60, 50, 40)$. For $n = 300$, we let
$(n_4, n_5, n_6) = (120, 100, 80)$ and for $n = 500$, we let $(n_4, n_5, n_6) = (200, 170, 130)$.
Using these values, in case 1 (respectively case 2), the average proportion of zero-inflation in the simulated data sets is 25% (respectively 50%). For each combination of the simulation design parameters (sample size and zero-inflation proportion), we simulate $N = 5000$ samples and we calculate the maximum likelihood estimate $\widehat{\psi}_n$.

Computational aspects of maximum likelihood estimation in ZIB regression are discussed by Hall [2000]. There, the author develops an EM algorithm for estimating ψ. Alternatively, he also suggests to use Newton-Raphson algorithm for solving (4.5). In his paper, Hall [2000] motivated his preference for the EM algorithm by programming simplicity. Since then, numerous `R` packages [Team, 2013] have been developed for maximizing log-likelihoods such as (4.4) or for solving likelihood equations such as (4.5). In our simulation study, we use the `R` package `maxLik` [Henningsen and Toomet, 2011].

4.4.2 Results

For each configuration `sample size` $\times$ `zero-inflation proportion` of the simulation design parameters, we calculate the average bias of the estimates $\widehat{\beta}_{j,n}$ and $\widehat{\gamma}_{k,n}$ of the β_j and γ_k over the N estimates. Based on the N simulated samples, we also obtain the average standard error (SE) and empirical standard deviation (SD) for each estimator $\widehat{\beta}_{j,n}$ $(j = 1, \ldots, 7)$ and $\widehat{\gamma}_{k,n}$ $(k = 1, \ldots, 5)$. Finally, we obtain 95%-level confidence intervals for the β_j and γ_k. We provide their empirical coverage probability (CP) and average length ℓ(CI). Results are given in Table 4.1 (case 1) and Table 4.2 (case 2).

Finally, in order to assess the quality of the Gaussian approximation stated in Theorem 4.3.3, we provide normal Q-Q plots of the estimates (see figures 4.1 and 4.2 for $n = 300$ in case 1 and figures 4.3 and 4.4 for $n = 300$ in case 2.
Plots for $n = 150$ and $n = 500$ yield similar observations and are thus omitted).

From these results, it appears as expected that the bias, SE, SD and ℓ(CI) of all estimators decrease as the sample size increases. The bias stays moderate provided that the sample size is large enough (say, $n \geq 300$). The empirical coverage probabilities are close to the nominal confidence level, even when the sample size is moderate. As may also be expected, we observe that the maximum likelihood estimator of the β_js (respectively γ_ks) performs better when the zero-inflation proportion decreases (respectively increases). Finally, it appears from normal Q-Q plots that the Gaussian approximation of the distribution of the maximum likelihood estimator in ZIB regression is reasonably satisfied, even when the sample size is moderate.

n		$\widehat{\beta}_n$							$\widehat{\gamma}_n$				
		$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\beta}_{7,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$	$\widehat{\gamma}_{5,n}$
150													
	bias	-0.0116	0.0297	0.0132	-0.0230	-0.0317	0.0286	0.0009	-0.0753	-0.0504	-0.0736	0.0512	0.0074
	SD	0.5686	0.1673	0.1484	0.1018	0.1635	0.2755	0.1278	0.5208	0.3583	0.7332	0.3150	0.5887
	SE	0.5546	0.1635	0.1454	0.0993	0.1596	0.2706	0.1241	0.5038	0.3441	0.8796	0.3061	0.5771
	CP	0.9446	0.9419	0.9436	0.9440	0.9459	0.9486	0.9459	0.9609	0.9534	0.9659	0.9576	0.9586
	ℓ(CI)	2.1648	0.6375	0.5682	0.3877	0.6210	1.0572	0.4840	1.9387	1.3256	2.8945	1.1760	2.2280
300													
	bias	-0.0105	0.0173	0.0072	-0.0100	-0.0150	0.0097	-0.0015	-0.0359	-0.0206	-0.0732	0.0242	0.0115
	SD	0.3887	0.1140	0.1012	0.0689	0.1128	0.1887	0.0863	0.3411	0.2281	0.5015	0.2078	0.3838
	SE	0.3793	0.1120	0.0996	0.0681	0.1088	0.1849	0.0845	0.3304	0.2255	0.4953	0.1998	0.3794
	CP	0.9467	0.9499	0.9489	0.9487	0.9427	0.9457	0.9423	0.9503	0.9501	0.9595	0.9479	0.9559
	ℓ(CI)	1.4843	0.4380	0.3900	0.2663	0.4251	0.7240	0.3304	1.2894	0.8791	1.8958	0.7782	1.4824
500													
	bias	0.0009	0.0092	0.0027	-0.0066	-0.0096	0.0081	0.0010	-0.0170	-0.0094	-0.0395	0.0157	0.0010
	SD	0.2925	0.0856	0.0766	0.0518	0.0828	0.1418	0.0664	0.2545	0.1740	0.3687	0.1554	0.2922
	SE	0.2910	0.0858	0.0764	0.0521	0.0832	0.1418	0.0647	0.2497	0.1702	0.3631	0.1508	0.2874
	CP	0.9490	0.9498	0.9480	0.9508	0.9518	0.9500	0.9436	0.9514	0.9470	0.9548	0.9506	0.9510
	ℓ(CI)	1.1397	0.3359	0.2993	0.2041	0.3255	0.5555	0.2531	0.9766	0.6653	1.4107	0.5891	1.1246

TABLE 4.1 – Simulation results (case 1). SE : average standard error. SD : empirical standard deviation. CP : empirical coverage probability of 95%-level confidence intervals. ℓ(CI) : average length of confidence intervals. All results are based on $N = 5000$ simulated samples.

n		$\widehat{\beta}_n$							$\widehat{\gamma}_n$				
		$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\beta}_{7,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$	$\widehat{\gamma}_{5,n}$
150													
	bias	-0.0396	0.0607	0.0278	-0.0378	-0.0589	0.0414	-0.0008	-0.0019	-0.0034	0.0507	0.0430	-0.0108
	SD	0.7568	0.2220	0.1976	0.1384	0.2257	0.4291	0.1719	0.4184	0.2593	0.4788	0.2377	0.4307
	SE	0.7228	0.2115	0.1909	0.1313	0.2154	0.4078	0.1639	0.4084	0.2490	0.4679	0.2298	0.4278
	CP	0.9395	0.9451	0.9445	0.9409	0.9443	0.9467	0.9431	0.9549	0.9517	0.9515	0.9527	0.9559
	ℓ(CI)	2.8115	0.8214	0.7435	0.5105	0.8330	1.5810	0.6358	1.5947	0.9704	1.8286	0.8948	1.6738
300													
	bias	-0.0180	0.0263	0.0121	-0.0160	-0.0278	0.0249	-0.0018	-0.0011	-0.0046	0.0255	0.0214	0.0043
	SD	0.4954	0.1447	0.1311	0.0896	0.1462	0.2784	0.1088	0.2808	0.1694	0.3184	0.1581	0.3014
	SE	0.4851	0.1415	0.1282	0.0881	0.1430	0.2707	0.1088	0.2780	0.1672	0.3174	0.1550	0.2918
	CP	0.9466	0.9482	0.9496	0.9492	0.9428	0.9448	0.9508	0.9498	0.9460	0.9522	0.9490	0.9446
	ℓ(CI)	1.8953	0.5524	0.5011	0.3441	0.5572	1.0566	0.4245	1.0883	0.6541	1.2429	0.6058	1.1431
500													
	bias	-0.0075	0.0151	0.0073	-0.0109	-0.0163	0.0142	-0.0016	-0.0027	-0.0030	0.0160	0.0108	0.0022
	SD	0.3707	0.1101	0.0982	0.0679	0.1083	0.2094	0.0837	0.2133	0.1298	0.2448	0.1175	0.2237
	SE	0.3684	0.1075	0.0974	0.0670	0.1083	0.2053	0.0824	0.2125	0.1273	0.2423	0.1178	0.2231
	CP	0.9492	0.9472	0.9502	0.9516	0.9498	0.9446	0.9458	0.9502	0.9492	0.9484	0.9516	0.9510
	ℓ(CI)	1.4413	0.4204	0.3813	0.2621	0.4229	0.8029	0.3224	0.8323	0.4984	0.9492	0.4613	0.8744

TABLE 4.2 – Simulation results (case 2). SE : average standard error. SD : empirical standard deviation. CP : empirical coverage probability of 95%-level confidence intervals. ℓ(CI) : average length of confidence intervals. All results are based on $N = 5000$ simulated samples.

4.5 An application of ZIB model to health economics

4.5.1 Data description and modelling

In this section, we describe an application of ZIB regression to the analysis of health-care utilization by elderlies in the United States. This application is based on data obtained from the National Medical Expenditure Survey (NMES) conducted in 1987-1988. This data set was first described by Deb and Trivedi [1997]. It provides a comprehensive picture of how Americans (aged 66 years and over) use and pay for health services. Several measures of health-care utilization were reported in this study, including the number of visits to a doctor in an office setting (denoted by *ofd* in what follows), the number of visits to a non-doctor health professional (such as a nurse, optician, physiotherapist...) in an office setting (*ofnd*), the number of visits to a doctor in an outpatient setting, the number of visits to a non-doctor in an outpatient setting, the number of visits to an emergency service and the number of hospital stays. A feature of these data is the high proportion of zero counts observed for some of the health-care utilization measures. In addition to health services utilization, the data set also contains information on health status, sociodemographic characteristics and economic status. Deb and Trivedi [1997] analyse separately each measure of health-care utilization by fitting zero-inflated count data models to each type of health-care usage in turns. Here, we consider the following issue. Consider a patient who decides to visit a health professional in an office setting. We wish to identify factors that explain patient's choice between a visit to a doctor and a visit to a non-doctor. For our study, we consider patients in the NMES data set who have a total number of office consultations comprised between 2 and 25. Among these $n = 3227$ patients, frequencies of zero in *ofnd* and *ofd* counts are 62.1% and 1.21% respectively. Let Z_i and m_i be respectively the number of non-doctor office visits and the total number of office visits for the i-th patient ($i = 1, \ldots, 3227$). Given m_i, one may model

Z_i as a $\mathscr{B}(m_i, \pi_i)$ distribution. However, the high frequency of zero in *ofnd* count suggests that Z_i is affected by zero-inflation. Therefore, we suggest to use a ZIB model for Z_i. Several covariates are available in the NMES data set, including : i) socio-economic variables : gender (1 for female, 0 for male), age (in years, divided by 10), marital status (1 if married, 0 if not married), educational level (number of years of education), income (in ten-thousands of dollars), ii) various measures of health status : number of chronic conditions (cancer, arthritis, diabete...) and a variable indicating self-perceived health level (poor, average, excellent) and iii) a binary variable indicating whether individual is covered by medicaid or not (medicaid is a US health insurance for individuals with limited income and resources, we code it as 1 if the individual is covered and 0 otherwise). Self-perceived health is re-coded as two dummy variables denoted by "health1" (1 if health is perceived as poor, 0 otherwise) and "health2" (1 if health is perceived as excellent, 0 otherwise). As mentioned above, we wish to identify determinants of patients choice between a doctor and a non-doctor visit. We model zero-inflation and event probabilities p_i and π_i by (4.2) and (4.3) respectively, where $\mathbf{X}_i$ and $\mathbf{W}_i$ are the set of covariates listed above. First, we fitted a ZIB regression model incorporating all available covariates in (4.2) and (4.3), *i.e.*, letting $\mathbf{X}_i = \mathbf{W}_i$ for every i. Then, Wald tests were used to select relevant covariates in submodels (4.2) and (4.3). However, this procedure can be cumbersome when the number of covariates is large. Thus, we propose an alternative procedure (here, both procedures yield the same final set of significant predictors). In a first stage, we fit a standard logistic regression model with all available covariates to binary indicators $1_{\{Z_i=0\}}$, $i = 1, \ldots, n$. The resulting model is not a model for zero-inflation since some of the 0 may arise from the binomial distribution $\mathscr{B}(m_i, \pi_i)$. However, we expect that this rough procedure will still select a relevant subset of covariates, that will be used in a second stage in the logistic sub-model (4.2) for p_i. Using this procedure and Wald testing, we identify four significant predictors : "health1" dummy variable, gender, educational level and medicaid status, that are included in p_i while all covariates are

included in π_i. Results for the resulting ZIB model are displayed in Table 4.3.

4.5.2 Results

In Table 4.3, we report estimate, standard error (s.e.) and significance level (as : not significant, significant or very significant) of Wald test of nullity for each parameter.

As mentioned above, gender, educational level, medicaid status and "health1" dummy variable are identified as the most influencing factors of the decision of never resorting to non-doctor health professionals, with a probability of never resorting which increases when health level degradates (one reason is that patients whose health declines may tend to favor visits to a doctor). Medicaid recipients are more likely to renounce non-doctor office visits. One explanation is that patients with medicaid coverage may limit their consultations to those necessary, that is, to doctor visits only (recall that medicaid is a health insurance for poor people). The probability of never resorting to non-doctor office consultations decreases with the number of years of education. This is coherent with previous findings, *e.g.*, Deb and Trivedi [1997], who postulate that education may make individuals more informed consumers of medical care services. More informed patients may tend to diversify their health-care utilization. For patients who eventually consult non-doctor health professionals in an office setting, ZIB model suggests that health status variables (number of chronic conditions and self-perceived health) are the most influencing factors of the choice between doctor and non-doctor visit. ZIB model also suggests that patients with poor health will favor visits to doctors over non-doctors, which seems a natural finding. Perhaps surprisingly, marital status has a significant effect on the choice of doctor *vs* non-doctor visit (being married increases the probability of visiting a non-doctor health professional). One explanation is that marital status may capture some income effect leading married patients to diversify their health-care utilization.

parameter	variable	estimate	s.e.	Wald test of $H_0 : \beta_j = 0$
β_1	intercept	-0.2095	0.2983	NS
β_2	health1	-0.3459	0.0750	VS
β_3	health2	0.2642	0.0816	VS
β_4	chronic	-0.0939	0.0167	VS
β_5	age	-0.0566	0.0360	NS
β_6	gender	0.0687	0.0487	NS
β_7	marital status	0.1372	0.0476	VS
β_8	educational	-0.0031	0.0067	NS
β_9	income	-0.0069	0.0064	NS
β_{10}	medicaid	-0.0911	0.0924	NS
γ_1	intercept	1.1095	0.1549	VS
γ_2	health1	0.3338	0.1284	S
γ_3	gender	-0.3220	0.0873	VS
γ_4	educational	-0.0746	0.0124	VS
γ_5	medicaid	0.4519	0.1621	VS

TABLE 4.3 – Health-care data analysis (NS : not significant at the 5% level, S : significant at level between 1% and 5%, VS (very significant) : significant at level less than 1%).

4.6 Discussion

Zero-inflated binomial regression is now commonly used for investigating count data with zeros excess. In this paper, we provide a rigorous basis for maximum likelihood inference in this model. Precisely, we establish consistency and asymptotic normality of the maximum likelihood estimator in ZIB regression. Moreover, our simulation study suggests that the maximum likelihood estimator performs well under a wide range of conditions pertaining to sample size and proportion of zero-inflation.

We consider here the basic ZIB regression model. Hall [2000] proposes to incorporate random effects to this model when the count data are correlated. Several other generalizations of ZIB regression may be developed to account for the increasing complexity of experimental data. For example, one may use partially linear link functions for the mixing and/or success probabilities (such as in the ZIP model, see for example Lam et al. [2006] and He et al. [2010]). Asymptotic properties of the statistical inference in these generalizations are still unknown and their rigorous derivation remains an open problem. This is a topic for our future work.

Acknowledgements

Authors are grateful to the referees and associate editor for their comments and suggestions on an earlier version of this paper. Authors acknowledge financial support from the "Service de Coopération et d'Action Culturelle" of the French Embassy in Senegal and logistical support from Campus France (French national agency for the promotion of higher education, international student services, and international mobility). Authors also acknowledge grants from CEA-MITIC (an African Center of Excellence in Mathematics, Informatics and ICT), implemented by Gaston Berger University (Senegal).

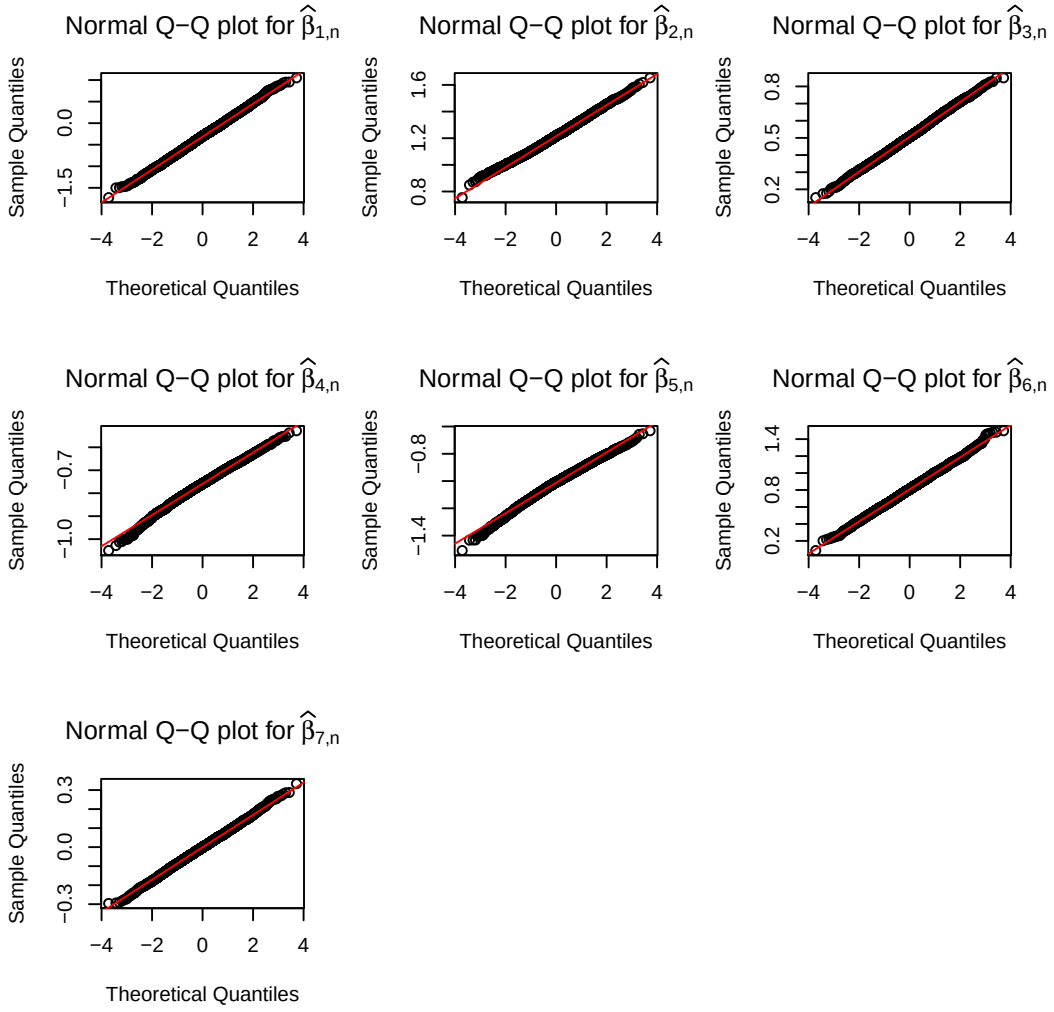

FIGURE 4.1 – Normal Q-Q plots for $\hat{\beta}_{1,n}, \ldots, \hat{\beta}_{7,n}$ with $n = 300$ (case 1).

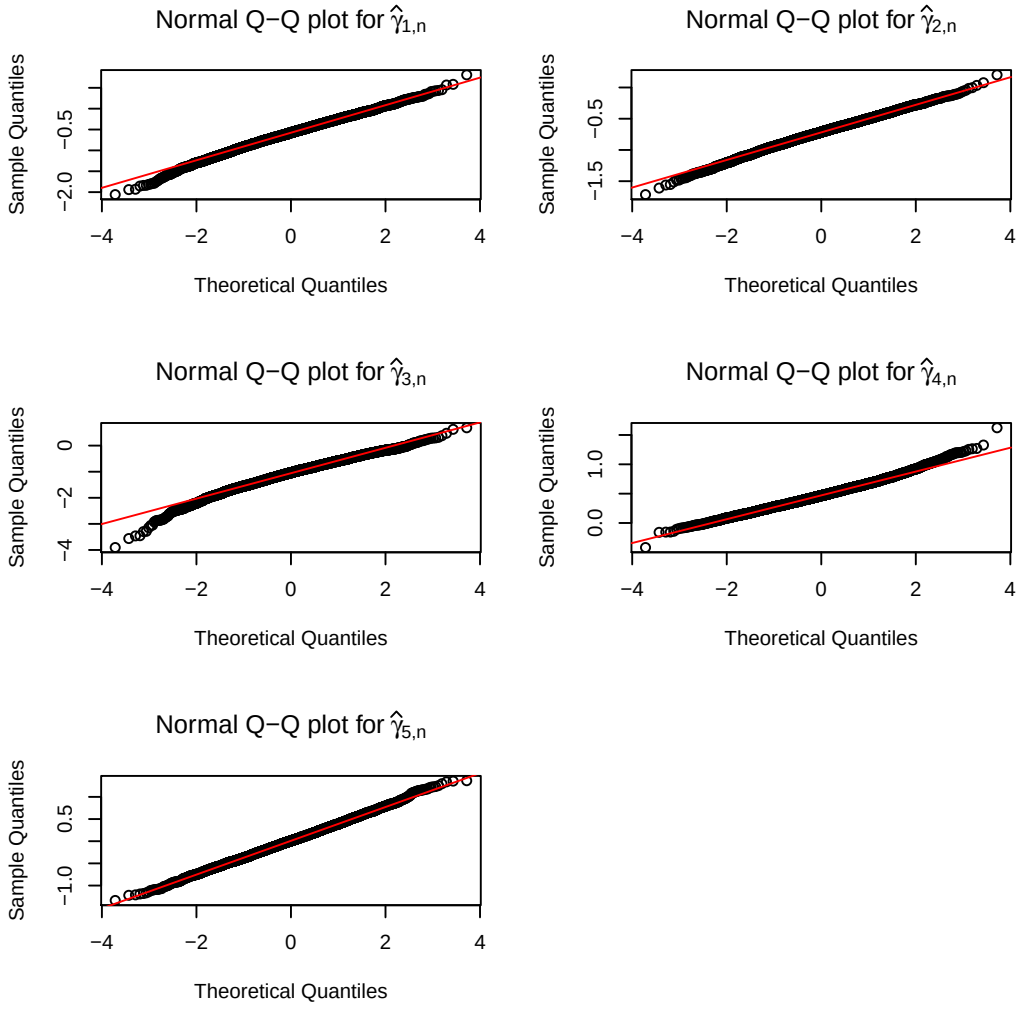

FIGURE 4.2 – Normal Q-Q plots for $\widehat{\gamma}_{1,n}, \ldots, \widehat{\gamma}_{5,n}$ with $n = 300$ (case 1).

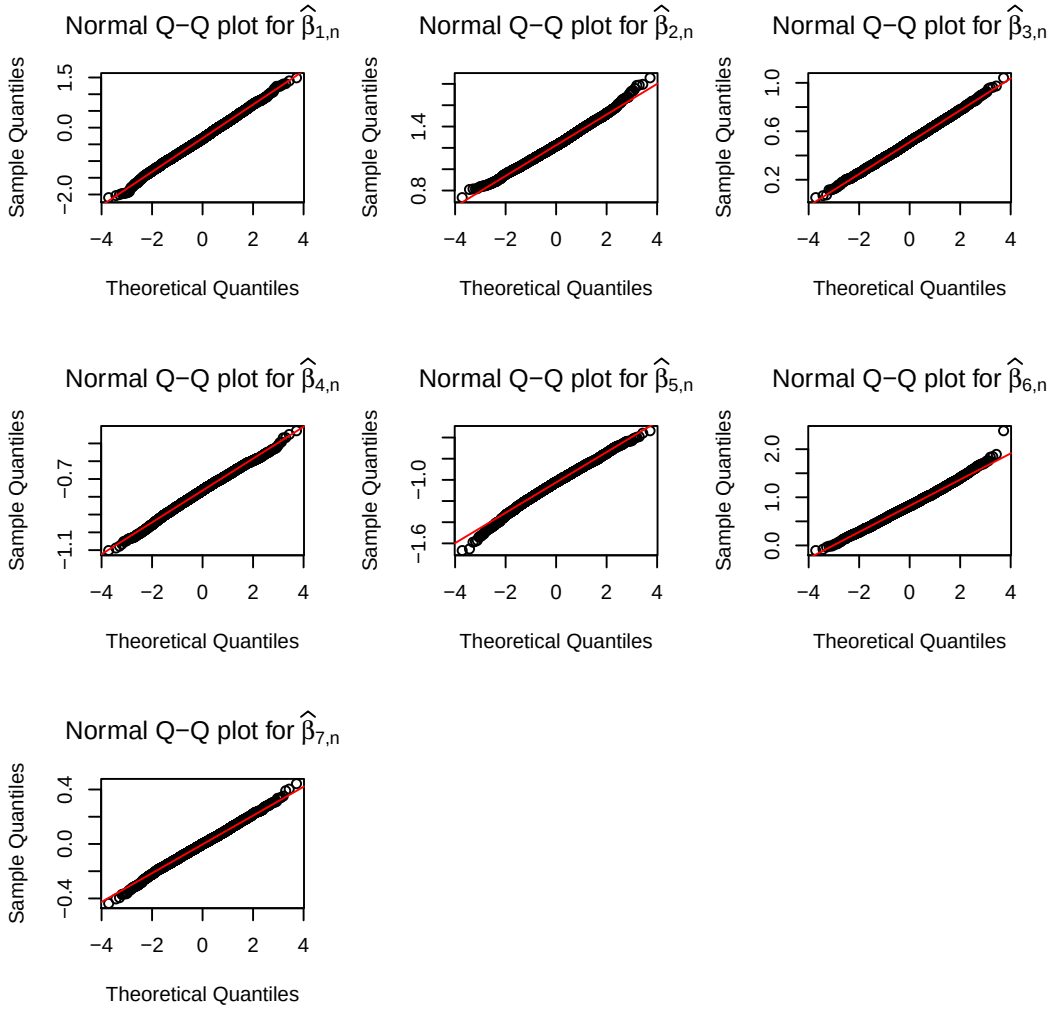

FIGURE 4.3 – Normal Q-Q plots for $\widehat{\beta}_{1,n}, \ldots, \widehat{\beta}_{7,n}$ with $n = 300$ (case 2).

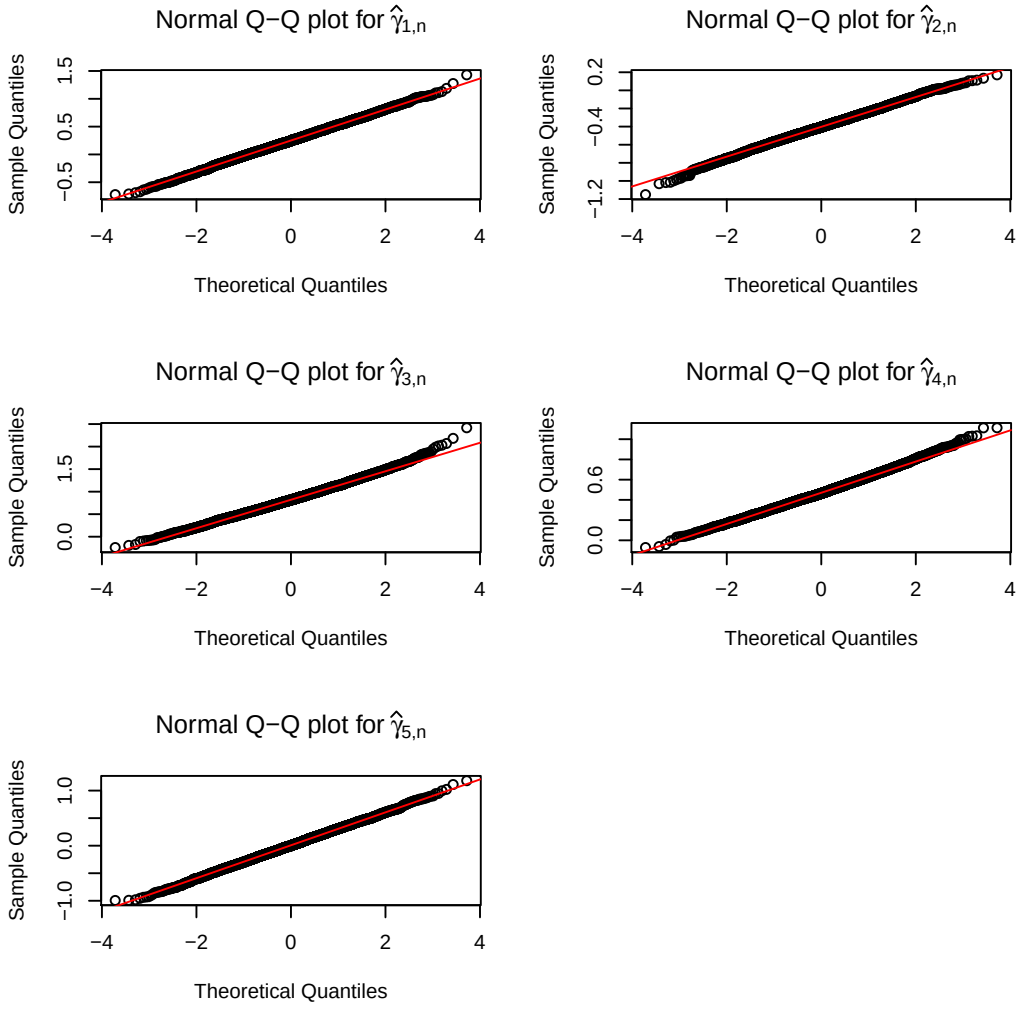

FIGURE 4.4 – Normal Q-Q plots for $\widehat{\gamma}_{1,n}, \ldots, \widehat{\gamma}_{5,n}$ with $n = 300$ (case 2).

CHAPITRE 5

Données multinomiales avec une inflation conjointe de zéro. Application en économie de la santé

Sommaire

Dans le chapitre 4, nous avons présenté une étude mathématique rigoureuse et complète de l'estimateur du maximum de vraisemblance dans le modèle ZIB (existence, consistance forte et normaité asymptotique). L'étude de simulation que nous avons ensuite conduite a confirmé ces bonnes propriétées.

Dans ce cinquième chapitre, nous avons poursuivi notre travail sur la modélisation statistique de données de comptage en présence d'inflation de zéros. En particulier, un nouveau modèle de régression multinomial à inflation de zéros a été mis en œuvre. Ce travail généralise ce qui a été entamé dans le chapitre précédent. Mais aussi, c'est un modèle qui permet de prendre en compte simultanément plusieurs mesures de la consommation de soins. Nous nous sommes intéressés aux propriétés asymptotiques dans ce modèle. Nous avons procédé à une simulation numérique de l'estimateur proposé sur différents échantillons de tailles finies. Les résultats obtenus sont cohérents. Pour finir, nous avons proposé une application de ce nouveau modèle à l'évaluation de la demande de soins médicaux et à l'étude du renoncement aux soins. Pour cela le modèle a été confronté à un jeu de données issu d'une étude de santé publique américaine. Notre modèle a permis de conforter des hypothèses émises dans la littérature et relatives aux déterminants du renoncement aux soins de santé.

Ce chapitre a fait l'objet d'un article accepté pour publication dans la revue *Journal of Statistical Planning and Inference*, 194, 85-105
Authors : DIALLO A. O, DIOP A., and DUPUY J.-F.,
Analysis of multinomial counts with joint zero-inflation, with an application to health economics.

Abstract

Zero-inflated regression models for count data are often used in health economics to analyse demand for medical care. Indeed, excess of zeros often affects health-care utilization data. Much of the recent econometric literature on the topic has focused on univariate health-care utilization measures, such as the number of doctor visits. However, health service utilization is usually measured by a number of different counts (*e.g.*, numbers of visits to different health-care providers). In this case, zero-inflation may jointly affect several of the utilization measures. In this paper, a zero-inflated regression model for multinomial counts with joint zero-inflation is proposed. Maximum likelihood estimators in this model are constructed and their properties are investigated, both theoretically and numerically. We apply the proposed model to an analysis of health-care utilization.

keywords **: excess zeros, health-care utilization, multinomial logit..**

5.1 Introduction

Statistical modeling of count data with zero inflation has become an important issue in numerous fields and in particular, in econometrics. The zero inflation (or excess zeros) problem occurs when the proportion of zero counts in the observed sample is much larger than predicted by standard count models. In health economics, this issue often arises in analysis of health-care utilization, as measured

by the number of doctor visits [Sarma and Simpson, 2006, Sarma, 2009, Staub and Winkelmann, 2013]. The present work is also motivated by an econometric analysis of health-care utilization and is illustrated by a data set described by Deb and Trivedi [1997].

Deb and Trivedi [1997] investigate the demand for medical care by elderlies in the United States. Their analysis is based on data from the National Medical Expenditure Survey (NMES) conducted in 1987 and 1988. These data provide a comprehensive picture of how Americans (aged 66 years and over) use and pay for health services. Six measures of health-care utilization were reported in this study, namely the number of visits to a doctor in an office setting, the number of visits to a non-doctor health professional (such as a nurse, optician, physiotherapist...) in an office setting, the number of visits to a doctor in an outpatient setting, the number of visits to a non-doctor in an outpatient setting, the number of visits to an emergency service and the number of hospital stays. A feature of these data is the high proportion of zero counts observed for some of the health-care utilization measures, that is, there is a high proportion of non-users of the corresponding health-care service over the study period. In addition to health services utilization, the data set also provides information on health status, sociodemographic characteristics and economic status. Deb and Trivedi [1997] analyse separately each measure of health-care utilization by fitting models for zero-inflated count data to each type of health-care usage in turns. However, several studies suggest that health-care utilization measures are not independent [Gurmu and Elder, 2000, Wang, 2003]. Therefore, we suggest to analyse jointly the various health-care utilization measures by fitting a multinomial logistic regression model to the data.

For illustrative purpose, and in order to keep notations simple, we will illustrate our model and methodology by considering three out of the six measures of health-care utilization, namely the : i) number Z_1 of consultations with a non-doctor in an office setting (denoted by *ofnd* in what follows), ii) number Z_2 of consultations with a non-doctor in an outpatient setting (*opnd*) and iii) number

Z_3 of consultations with a doctor in an office setting (*ofd*). If m_i denotes the total number of consultations for the i-th individual and $\mathbf{X}_i$ is a vector of covariates for this individual, we let
$Z_i = (Z_{1i}, Z_{2i}, Z_{3i})$ and we assume that Z_i has a multinomial distribution mult$(m_i, \mathbf{p}_i)$, where $\mathbf{p}_i = (p_{1i}, p_{2i}, p_{3i})$, $p_{1i} = \mathbb{P}(Z_{1i} = 1|\mathbf{X}_i)$ is the probability that a consultation is of type *ofnd*, $p_{2i} = \mathbb{P}(Z_{2i} = 1|\mathbf{X}_i)$ is the probability that a consultation is of type *opnd* and $p_{3i} = \mathbb{P}(Z_{3i} = 1|\mathbf{X}_i)$ is the probability that a consultation is of type *ofd*. We consider individuals in the NMES data set who have a total number of consultations less than or equal to 25. Among these 3224 individuals, frequencies of zero in variables *ofnd*, *opnd* and *ofd* are 62.7%, 81.3% and 1.5% respectively. Frequencies of zeros occuring simultaneously in variables of pairs (*ofnd* and *opnd*), (*ofnd* and *ofd*) and (*opnd* and *ofd*) are 51.7%, 0.24% and 1% respectively. That is, 51.7% of the surveyed subjects did not use any services associated with counts Z_1 and Z_2. This high frequency and the very low frequency of zero counts for *ofd* suggest that there may exist some permanent non-users of *ofnd* and *opnd*, *i.e.*, individuals who would never use these health-care services. In other words, there may exist an excess of observations of the form $(0, 0, m_i)$ in the data set.

To accommodate these observations, we propose to define, for each individual i, a zero-inflated multinomial regression model as the mixture

$$\pi_i \cdot \delta_{(0,0,m_i)} + (1 - \pi_i) \cdot \text{mult}(m_i, \mathbf{p}_i) \tag{5.1}$$

of the multinomial distribution mult$(m_i, \mathbf{p}_i)$ with a degenerate distribution $\delta_{(0,0,m_i)}$ at $(0, 0, m_i)$. π_i represents the probability that the i-th individual is a permanent non-user of health-care services of the type *ofnd* and *opnd*.

Mixture models for zero-inflated count data date back to early '90s. Zero-inflated Poisson (ZIP) regression was proposed by Lambert [1992] and further developed by Dietz and Böhning [2000], Li [2011], Lim et al. [2014] and Monod [2014], among many others. Zero-inflated negative binomial (ZINB) regression was proposed by Ridout et al. [2001], see also Moghimbeigi et al. [2008],

Mwalili et al. [2014], Garay et al. [2011]. Hall [2000] and Vieira et al. [2000] introduced the zero-inflated binomial (ZIB) regression model, see also Diop et al. [2016]. But to the best of our knowledge, and although some related models can be found in Kelley and Anderson [2008] and Bagozzi [2015], the zero-inflated multinomial model (5.1) has not been yet considered. Kelley and Anderson [2008] [respectively Bagozzi, 2015] propose a model for a discrete ordinal (respectively nominal) dependent variable with levels $\{0, 1, \ldots, J\}$ and zero-inflation. However, authors do not report any systematic investigation of their models (such as model identifiability or estimation). In the present paper, we aim at providing a rigorous study of model (5.1) that will serve as a basis for future application of the model to real-data problems. We derive maximum likelihood estimators of parameters π_i and $\mathbf{p}_i$, we establish their asymptotic properties (consistency and asymptotic normality) and we assess their finite-sample behaviour using simulations. Then, we illustrate the model on the health-care utilization data set described above.

The remainder of the paper is organized as follows. In Section 5.2, we specify precisely the model and we address the estimation of π_i and $\mathbf{p}_i$. In Section 5.3, we report results of our simulation study. Section 5.4 describes the health-care data analysis. A conclusion and some perspectives are provided in Section 5.5. All proofs are postponed to an appendix.

5.2 Zero-inflated multinomial regression model

In this section, we describe the zero-inflated multinomial (ZIM) regression model. We consider two cases : i) π_i is fixed (that is, $\pi_i = \pi$ for every individual) and ii) π_i depends on covariates. In section 5.2.3, identifiability of the ZIM model and asymptotics of the maximum likelihood estimator are described for fixed π but results can be generalized to case ii) without major difficulty. Moreover, for notational simplicity, we consider the case where the multinomial response Z_i has $K = 3$ outcomes. Proofs for a general K proceed similarly.

5.2.1 Model and estimation with fixed π

Let $(Z_i, \mathbf{X}_i)$, $i = 1, \ldots, n$ be independent random vectors defined on the probability space $(\Omega, \mathscr{C}, \mathbb{P})$. For every i, we assume that given the total $Z_{1i} + Z_{2i} + Z_{3i} = m_i$, the multivariate response $Z_i = (Z_{1i}, Z_{2i}, Z_{3i})$ is generated from the model

$$Z_i \sim \begin{cases} (0, 0, m_i) & \text{with probability } \pi, \\ \text{mult}(m_i, \mathbf{p}_i) & \text{with probability } 1 - \pi, \end{cases} \tag{5.2}$$

where $\mathbf{p}_i = (p_{1i}, p_{2i}, p_{3i})$ and $p_{1i} + p_{2i} + p_{3i}$=1. This model reduces to the standard multinomial distribution (with three modalities, here) if $\pi = 0$, while $\pi > 0$ leads to simultaneous zero-inflation in the first two modalities. We model probabilities p_{1i}, p_{2i} and p_{3i} $(i = 1, \ldots, n)$ via multinomial logistic regression :

$$p_{1i} = \frac{e^{\beta_1^\top \mathbf{X}_i}}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}}, p_{2i} = \frac{e^{\beta_2^\top \mathbf{X}_i}}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}} \text{ and } p_{3i} = \frac{1}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}}, \tag{5.3}$$

where $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})^\top$ is a vector of predictors or covariates (both categorical and continuous covariates are allowed) and $\top$ denotes the transpose operator.
Let $\psi = (\pi, \beta_1^\top, \beta_2^\top)^\top$ be the unknown k-dimensional parameter of ZIM model ($k := 1 + 2p$). For $i = 1, \ldots, n$, let $J_i := 1_{\{Z_i \neq (0,0,m_i)\}}$ and $h_i(\beta) = 1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}$, where $\beta = (\beta_1^\top, \beta_2^\top)^\top$. Then, the log-likelihood of ψ based on observations $(Z_1, \mathbf{X}_1), \ldots, (Z_n, \mathbf{X}_n)$ is :

$$\begin{aligned} l_n(\psi) &= \sum_{i=1}^{n} \left\{ (1 - J_i) \log \left(\pi + (1 - \pi) \frac{1}{(h_i(\beta))^{m_i}} \right) \right. \\ &\quad \left. + J_i \left[\log \left(\frac{m_i}{Z_{1i} Z_{2i} Z_{3i}} \right) - m_i \log h_i(\beta) + Z_{1i} \beta_1^\top \mathbf{X}_i + Z_{2i} \beta_2^\top \mathbf{X}_i + \log(1 - \pi) \right] \right\}, \\ &:= \sum_{i=1}^{n} l_{[i]}(\psi). \end{aligned}$$

The maximum likelihood estimator $\widehat{\psi}_n := (\widehat{\pi}, \widehat{\beta}_1^\top, \widehat{\beta}_2^\top)^\top$ of ψ is the solution of the k-dimensional score equation

$$\dot{l}_n(\psi) := \frac{\partial l_n(\psi)}{\partial \psi} = 0. \tag{5.4}$$

Solving this (non-linear) equation is relatively straightforward using standard mathematical softwares. In our simulation study and real-data analysis, we use `R` package `maxLik`

[Henningsen and Toomet, 2011], which provides efficient computational tools for solving likelihood equations such as (5.4).

We need to introduce some further notations and a few regularity assumptions before stating asymptotic properties of $\widehat{\psi}_n$.

5.2.2 Some further notations

It will be useful to define the $(p \times n)$ and $(k \times 3n)$ matrices :

$$\mathbb{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1p} & X_{2p} & \cdots & X_{np} \end{pmatrix} \text{ and } \mathbb{V} = \begin{pmatrix} \mathbb{1}_{(1,n)} & 0_{(1,n)} & 0_{(1,n)} \\ 0_{(p,n)} & \mathbb{X} & 0_{(p,n)} \\ 0_{(p,n)} & 0_{(p,n)} & \mathbb{X} \end{pmatrix},$$

where $\mathbb{1}_{(1,n)}$ denotes the n-dimensional raw vector $(1, 1, \ldots, 1)$ and $0_{(a,b)}$ denotes the $(a \times b)$ matrix whose components are all equal to zero (with a and b two positive integers). Let also $C(\psi) = (C_j(\psi))_{1 \leq j \leq 3n}$ be the $3n$-dimensional column vector defined by

$$C(\psi) = (A_1(\psi), \ldots, A_n(\psi), B_{1,1}(\psi), \ldots, B_{n,1}(\psi), B_{1,2}(\psi), \ldots, B_{n,2}(\psi))^\top,$$

where for every $i = 1, \ldots, n$,

$$A_i(\psi) = \frac{(h_i(\beta))^{m_i} - 1}{\pi\left[(h_i(\beta))^{m_i} - 1\right] + 1}(1 - J_i) - \frac{1}{1-\pi} J_i,$$
$$B_{i,\ell}(\psi) = -(1-\pi)\frac{m_i e^{\beta_\ell^\top \mathbf{x}_i}}{k_i(\psi)}(1 - J_i) + \left(-\frac{m_i e^{\beta_\ell^\top \mathbf{x}_i}}{h_i(\beta)} + Z_{\ell i}\right) J_i, \quad \ell = 1, 2,$$

and $k_i(\psi) = \pi\left[(h_i(\beta))^{m_i+1} - h_i(\beta)\right] + h_i(\beta)$. Then, some simple algebra shows that the likelihood equation (5.4) can be rewritten as

$$\dot{l}_n(\psi) = \mathbb{V} C(\psi) = 0.$$

If $A = (A_{ij})_{1 \leq i \leq a, 1 \leq j \leq b}$ is some $(a \times b)$ matrix, we denote its j-th column $(j = 1, \ldots, b)$ by $A_{\bullet j}$. That is, $A_{\bullet j} = (A_{1j}, \ldots, A_{aj})^\top$. Then, it will be useful to rewrite the score vector as

$$\dot{l}_n(\psi) = \sum_{j=1}^{3n} \mathbb{V}_{\bullet j} C_j(\psi).$$

We shall further denote by $\ddot{l}_n(\psi) = \partial^2 l_n(\psi)/\partial\psi\partial\psi^\top$ the $(k\times k)$ matrix of second derivatives of $l_n(\psi)$. Let $\mathbb{D}(\psi) = (\mathbb{D}_{ij}(\psi))_{1\leq i,j\leq 3n}$ be the $(3n \times 3n)$ block matrix defined as

$$\mathbb{D}(\psi) = \begin{pmatrix} \mathbb{D}_1(\psi) & \mathbb{D}_4(\psi) & \mathbb{D}_5(\psi) \\ \mathbb{D}_4(\psi) & \mathbb{D}_2(\psi) & \mathbb{D}_6(\psi) \\ \mathbb{D}_5(\psi) & \mathbb{D}_6(\psi) & \mathbb{D}_3(\psi) \end{pmatrix},$$

where $\mathbb{D}_1(\psi)$ to $\mathbb{D}_6(\psi)$ are $(n \times n)$ diagonal matrices, with i-th diagonal elements respectively given by

$$\begin{aligned}
\mathbb{D}_{1,ii}(\psi) &= \left(\frac{(h_i(\beta))^{m_i} - 1}{\pi\left[(h_i(\beta))^{m_i} - 1\right] + 1}\right)^2 (1 - J_i) + \frac{1}{(1-\pi)^2} J_i, \\
\mathbb{D}_{\ell+1,ii}(\psi) &= \frac{(1-\pi)(1-J_i)e^{\beta_\ell^\top \mathbf{X}_i}\left((k_i(\psi) - e^{\beta_\ell^\top \mathbf{X}_i}(\pi[(m_i+1)(h_i(\beta))^{m_i} - 1] + 1)\right)}{(k_i(\psi))^2} \\
&\quad - \frac{m_i J_i e^{\beta_\ell^\top \mathbf{X}_i}\left(h_i(\beta) - e^{\beta_\ell^\top \mathbf{X}_i}\right)}{(h_i(\beta))^2}, \quad \ell = 1, 2, \\
\mathbb{D}_{\ell+3,ii}(\psi) &= -\frac{(1-J_i)m_i e^{\beta_\ell^\top \mathbf{X}_i}(h_i(\beta))^{m_i+1}}{(k_i(\psi))^2}, \quad \ell = 1, 2, \\
\mathbb{D}_{6,ii}(\psi) &= -\frac{(1-\pi)(1-J_i)m_i e^{\beta_1^\top \mathbf{X}_i} e^{\beta_2^\top \mathbf{X}_i}\left(\pi[(h_i(\beta))^{m_i} - 1] + 1\right)}{(k_i(\psi))^2} - \frac{J_i m_i e^{\beta_1^\top \mathbf{X}_i} e^{\beta_2^\top \mathbf{X}_i}}{(h_i(\psi))^2}.
\end{aligned}$$

Then, some tedious albeit not difficult algebra shows that $\ddot{l}_n(\psi)$ can be expressed as $\ddot{l}_n(\psi) = -\mathbb{V}\mathbb{D}(\psi)\mathbb{V}^\top$. Note that $C(\psi), \mathbb{V}$ and $\mathbb{D}(\psi)$ depend on n. However, in order to simplify notations, n will not be used as a lower index for these quantities. In the next section, we state some regularity conditions and asymptotic properties of the maximum likelihood estimator $\widehat{\psi}_n$.

5.2.3 Regularity conditions, identifiability and asymptotic results

The following conditions are somewhat classical in the framework of generalized linear regression models and are adapted to our setting.

C1 Covariates X_{ij} are bounded and $\text{var}[X_{ij}] > 0$, for every $i = 1, 2, \ldots$ and $j = 2, \ldots, p$. The X_{ij} $(j = 1, \ldots, p)$ are linearly independent, for every $i = 1, 2, \ldots$

C2 The true parameter value $\psi_0 := (\pi_0, \beta_{1,0}^\top, \beta_{2,0}^\top)^\top$ lies in the interior of some known compact set $\mathbf{K} \subset [0,1] \times \mathbb{R}^p \times \mathbb{R}^p$ (in what follows, we will also note $\beta_0 := (\beta_{1,0}^\top, \beta_{2,0}^\top)^\top$).

C3 The Hessian matrix $\ddot{l}_n(\psi)$ is negative definite and of full rank, for every $n = 1, 2, \ldots$ and $\frac{1}{n}\ddot{l}_n(\psi)$ converges to a negative definite matrix. Let λ_n and Λ_n be respectively the smallest and largest eigenvalues of $\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top$. There exists a finite positive constant c_1 such that $\Lambda_n/\lambda_n < c_1$ for every $n = 1, 2, \ldots$ The matrix $\mathbb{V}\mathbb{V}^\top$ is positive definite for every $n = 1, 2, \ldots$ and its smallest eigenvalue $\widetilde{\lambda}_n$ tends to $+\infty$ as $n \to \infty$.

Next condition will be useful for proving identifiability of the ZIM model (*i.e.*, distinct parameter values yield distinct values of the likelihood function).

C4 For every $i = 1, \ldots, n$, we have $m_i \geq 2$ (that is, in our application, we consider individuals who had at least two visits of all type).

Then, the following result holds for a fixed probability of zero-inflation (proof is given in Appendix A) :

Theorem 5.2.1 (Identifiability) *Under conditions C1-C4, the ZIM model* (5.2)-(5.3) *is identifiable, that is,* $l_{[i]}(\psi) = l_{[i]}(\psi^*)$ *almost surely implies* $\psi = \psi^*$.

Now, we state asymptotic properties of the estimator $\widehat{\psi}_n$. Proofs are outlined in Appendix B. Main steps are similar to proofs of asymptotics in the logistic regression model [*e.g.*, Gouriéroux and Monfort, 1981]. However, specific technical difficulties arise in the ZIM model. In particular, observations $(Z_i, \mathbf{X}_i)$ are not identically distributed (the number m_i of visits of all types varies across individuals).

In what follows, the space $\mathbb{R}^k$ of k-dimensional vectors is equipped with the Euclidean norm $\|\cdot\|$. The space of $(k \times k)$ real matrices is equipped with the norm $\|A\|_2 := \max_{\|x\|=1} \|Ax\|$ (for notations simplicity, we use $\|\cdot\|$ for both norms). Recall that for a symmetric real $(k \times k)$-matrix A with eigenvalues $\lambda_1, \ldots, \lambda_k$, $\|A\| := \|A\|_2 = \max_i |\lambda_i|$. Finally, we let I_k denote the identity matrix of order k. Our results are as follows :

Theorem 5.2.2 (Existence and consistency) *The maximum likelihood estimator* $\widehat{\psi}_n$ *exists almost surely as* $n \to \infty$ *and converges almost surely to* ψ_0.

Moreover, $\widehat{\psi}_n$ is asymptotically Gaussian :

Theorem 5.2.3 (Asymptotic normality) *Let* $\widehat{\Sigma}_n := \mathbb{VD}(\widehat{\psi}_n)\mathbb{V}^\top$. *Then, as* $n \to \infty$, $\widehat{\Sigma}_n^{\frac{1}{2}}(\widehat{\psi}_n - \psi_0)$ *converges in distribution to the Gaussian vector* $\mathscr{N}(0, I_k)$.

In the next section, we describe briefly the ZIM model with covariate-dependent probability of zero-inflation.

5.2.4 Model and estimation with covariate-dependent π_i

We assume that the probability π_i of $(0, 0, m_i)$-inflation for individual i depends on some observed q-dimensional covariate $\mathbf{W}_i$ ($\mathbf{W}_i$ may overlap with $\mathbf{X}_i$ or be distinct from $\mathbf{X}_i$. This issue is discussed in the application). We model π_i via logistic regression :

$$\pi_i = \frac{e^{\gamma^\top \mathbf{W}_i}}{1 + e^{\gamma^\top \mathbf{W}_i}}. \tag{5.5}$$

The log-likelihood of $\psi = (\gamma^\top, \beta_1^\top, \beta_2^\top)^\top$, based on observations $(Z_i, \mathbf{X}_i, \mathbf{W}_i)$, $i = 1, \ldots, n$ is similar to (5.4), with π_i replacing π, and the maximum likelihood estimator of ψ is defined similarly as above. Identifiability of the ZIM model (5.2)-(5.3)-(5.5) can be proved along the same lines as Theorem 5.2.1 under the following additional regularity condition : covariates W_{ij} are bounded and var$[W_{ij}] > 0$, for every $i = 1, 2, \ldots$ and $j = 2, \ldots, q$. The W_{ij} $(j = 1, \ldots, q)$ are linearly independent, for every $i = 1, 2, \ldots$

5.3 A simulation study

In this section, we assess finite-sample properties of the maximum likelihood estimator $\widehat{\psi}_n$ with fixed π and covariate-dependent π_i.

Case (i) : fixed probability of zero-inflation. We simulate data from a ZIM model defined by :

$$p_{1i} = \frac{e^{\beta_1^\top \mathbf{X}_i}}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}}, \quad p_{2i} = \frac{e^{\beta_2^\top \mathbf{X}_i}}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}} \quad \text{and} \quad p_{3i} = 1 - p_{1i} - p_{2i},$$

where $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{i7})^\top$ and $X_{i2}, \ldots, X_{i7}$ are independent covariates simulated from normal $\mathscr{N}(0, 1)$, uniform $\mathscr{U}(2, 5)$, normal $\mathscr{N}(1, 1.5)$, exponential $\mathscr{E}(1)$, binomial

$\mathscr{B}(1, 0.3)$ and normal $\mathscr{N}(-1, 1)$ distributions respectively. Parameters β_1 and β_2 are chosen as

$\beta_1 = (0.3, 1.2, 0.5, -0.75, -1, 0.8, 0)^\top$ and $\beta_2 = (0.5, 0.5, 0, -0.5, 0.5, -1.1, 0)^\top$. Several sample sizes n are considered : $n = 150, 300$ and 500. Numbers m_i are allowed to vary across subjects, with $m_i \in \{3, 4, 5\}$. Let $(n_3, n_4, n_5) = (\text{card}\{i : m_i = 3\}, \text{card}\{i : m_i = 4\}, \text{card}\{i : m_i = 5\})$. For $n = 150$, we let $(n_3, n_4, n_5)) = (50, 50, 50)$. For $n = 300$, we let $(n_3, n_4, n_5) = (120, 100, 80)$ and for $n = 500$, we let $(n_3, n_4, n_5) = (230, 170, 100)$. Zero-inflation is simulated from a Bernoulli variable with parameter π, with $\pi = 0.15, 0.25$ and 0.5.

Case (ii) : covariate-dependent probability of zero-inflation. In a second set of simulation scenarios, zero-inflation is allowed to depend on covariates. Simulation design is essentially similar as above, except that : i) $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{i5})^\top$ and $X_{i2}, \ldots, X_{i5}$ are simulated independently from uniform $\mathscr{U}(2, 5)$, normal $\mathscr{N}(1, 1.5)$, exponential $\mathscr{E}(1)$ and binomial $\mathscr{B}(1, 0.3)$ distributions respectively and ii) for each individual i, zero-inflation is simulated from a Bernoulli random variable with parameter π_i, where $\text{logit}(\pi_i) = \gamma^\top \mathbf{W}_i$ and $\mathbf{W}_i$ is taken as $\mathbf{W}_i = (1, X_{i2}, X_{i5}, W_{i4})^\top$ with $W_{i4} \sim \mathscr{N}(-1, 1)$. Parameters β_1 and β_2 are taken as $\beta_1 = (0.3, 0.5, -0.75, -1, 0)^\top$ and $\beta_2 = (0.5, 0, -0.5, 1.5, -1.1)^\top$.
The parameter vector $\gamma \in \mathbb{R}^4$ is chosen to yield various average proportions of zero-inflation within each sample, namely : $0.15, 0.25$ and 0.5.

Results. For each combination `sample size` $\times$ `zero-inflation proportion`, we simulate $N = 5000$ samples and for each of them, we calculate the maximum likelihood estimate $\widehat{\psi}_n$ of (π, β_1, β_2) (case (i)) and $(\gamma, \beta_1, \beta_2)$ (case (ii)). Several authors developed EM-type algorithms for estimation in zero-inflated models [*e.g.*, Wang, 2003, Kelley and Anderson, 2008]. Other authors proceed to direct maximization using Newton-Raphson or related algorithms [*e.g.*, Staub and Winkelmann, 2013]. Here, we use Newton-Raphson-like algorithm implemented in the `R` package `maxLik` developed by Henningsen and Toomet [2011]. We did not observe convergence problem, when taking initial values of the algorithm close to 0.

Based on the N estimates, we obtain, for each simulation scenario, the : i) empirical bias and relative bias of each estimator (the relative bias is calculated as the absolute value

of the ratio (bias/true parameter value) $\times 100$), ii) average standard error (SE) and empirical standard deviation (SD) of each estimator, iii) empirical coverage probability (CP) and average length ℓ(CI) of 95%-level confidence interval for each parameter. Let $\widehat{\beta}_{1,1}^{(k)}$ be the maximum likelihood estimate of the first component $\beta_{1,1}$ of β_1, obtained on the k-th simulated sample ($k = 1, \ldots, N$). Let $\bar{\beta}_{1,1} := N^{-1} \sum_{k=1}^{N} \hat{\beta}_{1,1}^{(k)}$. Then, the empirical bias and SD are calculated as $\bar{\beta}_{1,1} - \beta_{1,1}$ and $\sqrt{N^{-1} \sum_{k=1}^{N} (\widehat{\beta}_{1,1}^{(k)} - \bar{\beta}_{1,1})^2}$ respectively. From Theorem 5.2.3, the maximum likelihood estimate $\widehat{\beta}_{1,1}$ is asymptotically distributed as a Gaussian random variable whose variance can be consistently estimated by the corresponding element of $\widehat{\Sigma}_n^{-1}$. Let denote by $\widehat{\sigma}^2_{\widehat{\beta}_{1,1}^{(k)}}$ this element, calculated on the k-th simulated sample. Then the empirical SE is calculated as $N^{-1} \sum_{k=1}^{N} \widehat{\sigma}_{\widehat{\beta}_{1,1}^{(k)}}$. For case (i), results are given in Table 5.1 ($\pi = 0.15$), Table 5.2 ($\pi = 0.25$) and Table 5.3 ($\pi = 0.5$). For case (ii), results are given in Table 5.4 (average sample proportion of zero-inflation equal to 0.15), Table 5.5 (average sample proportion of zero-inflation equal to 0.25) and Table 5.6 (average sample proportion of zero-inflation equal to 0.5).

From these tables, the bias, SE, SD and ℓ(CI) of all estimators decrease as sample size increases. The relative bias of the $\widehat{\beta}_j$'s is generally smaller than 5%, except when n is small and the proportion of zero-inflation is high ($n = 150$ and $\pi = 0.5$, see Table 5.3 and Table 5.6), and the relative bias of the $\widehat{\gamma}_j$'s is generally smaller than 5% except when the sample size is small ($n = 150$). Overall, the relative bias is generally slightly larger for the $\widehat{\gamma}_j$'s than for the $\widehat{\beta}_j$'s. The empirical coverage probabilities are close to the nominal confidence level. Maximum likelihood seems to provide an efficient method for estimating ZIM model, even when the number of parameters is quite large.

We assess empirically the Gaussian approximation stated in Theorem 5.2.3 by plotting normal Q-Q plots of the estimates. Figures 5.1, 5.2 and 5.3 provide plots for case (ii) with $n = 300$ and an average sample proportion of zero-inflation equal to 0.25 (plots for the other simulation scenarios are similar and are thus omitted). From these plots, the distribution of the maximum likelihood estimator in the ZIM model is reasonably approximated by the Gaussian distribution. This, in particular, will allow Wald tests of covariate effects to be performed in ZIM model.

One question of interest is robustness of multinomial regression to presence of zero-

inflation. Indeed, poor performance of multinomial regression on zero-inflated data will constitute an additional argument in favor of the introduction of ZIM model. To answer this issue, we conduct a short simulation study. We simulate N samples of zero-inflated multinomial data and we fit to them the multinomial regression model mult$(m_i, \mathbf{p}_i)$ (with $\mathbf{p}_i = (p_{1i}, p_{2i}, p_{3i})$ given by (5.3)) and the proposed ZIM model. Results for case (ii) (covariate-dependent probabilities of zero-inflation) with $n = 500$ and an average sample proportion of zero-inflation equal to 0.15 are provided in Table 5.7. In the lower panel, we observe that the estimates of β_1 and β_2 obtained from multinomial regression ignoring zero-inflation are highly biased. This results in confidence intervals for the components of β_1, β_2 having very poor coverage probabilities. Unreported simulations run with other values for n and π yield similar observations.

n	$\widehat{\pi}$	$\widehat{\beta}_1$							$\widehat{\beta}_2$						
		$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{1,6}$	$\widehat{\beta}_{1,7}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$	$\widehat{\beta}_{2,6}$	$\widehat{\beta}_{2,7}$
150															
bias	-0.0020	-0.0073	0.0389	0.0157	-0.0225	-0.0282	0.0259	-0.0028	-0.0072	0.0224	0.0036	-0.0149	0.0158	-0.0284	-0.0035
rel. bias	1.3034	2.4213	3.2379	3.1492	2.9947	2.8156	3.2383	-	1.4416	4.4816	-	2.9886	3.1625	2.5785	-
SD	0.0320	0.6687	0.1728	0.1749	0.1171	0.2110	0.3208	0.1475	0.6542	0.1630	0.1735	0.1131	0.1577	0.3751	0.1469
SE	0.0316	0.6470	0.1689	0.1689	0.1109	0.2046	0.3153	0.1446	0.6434	0.1614	0.1699	0.1085	0.1531	0.3716	0.1465
CP	0.9300	0.9462	0.9478	0.9424	0.9404	0.9460	0.9498	0.9466	0.9436	0.9500	0.9486	0.9430	0.9474	0.9482	0.9500
ℓ(CI)	0.1234	2.5286	0.6597	0.6604	0.4330	0.7988	1.2326	0.5644	2.5137	0.6299	0.6643	0.4236	0.5939	1.4493	0.5718
300															
bias	-0.0011	0.0006	0.0182	0.0075	-0.0113	-0.0142	0.0135	0.0002	-0.0033	0.0102	0.0030	-0.0083	0.0080	-0.0131	0.0007
rel. bias	0.7582	0.1969	1.5165	1.4913	1.5122	1.4175	1.6835	-	0.6672	2.0311	-	1.6648	1.6031	1.1885	-
SD	0.0224	0.4495	0.1210	0.1185	0.0782	0.1445	0.2180	0.1014	0.4473	0.1141	0.1192	0.0768	0.1061	0.2590	0.1025
SE	0.0224	0.4527	0.1178	0.1181	0.0774	0.1422	0.2204	0.1007	0.4497	0.1123	0.1187	0.0758	0.1050	0.2583	0.1020
CP	0.9408	0.9538	0.9432	0.9496	0.9476	0.9428	0.9506	0.9508	0.9560	0.9476	0.9458	0.9470	0.9476	0.9498	0.9478
ℓ(CI)	0.0878	1.7719	0.4610	0.4624	0.3028	0.5563	0.8631	0.3939	1.7599	0.4393	0.4648	0.2965	0.4095	1.0105	0.3990
500															
bias	-0.0009	0.0006	0.0115	0.0041	-0.0068	-0.0073	0.0065	0.0017	-0.0019	0.0064	0.0010	-0.0054	0.0049	-0.0093	0.0001
rel. bias	0.5812	0.1925	0.9613	0.8290	0.9009	0.7256	0.8075	-	0.3890	1.2886	-	1.0884	0.9823	0.8463	-
SD	0.0175	0.3519	0.0929	0.0914	0.0604	0.1115	0.1739	0.0804	0.3506	0.0884	0.0922	0.0579	0.0830	0.2065	0.0799
SE	0.0175	0.3534	0.0918	0.0922	0.0603	0.1107	0.1719	0.0784	0.3512	0.0875	0.0927	0.0590	0.0814	0.2012	0.0794
CP	0.9456	0.9524	0.9462	0.9502	0.9506	0.9514	0.9434	0.9442	0.9496	0.9510	0.9516	0.9522	0.9452	0.9440	0.9516
ℓ(CI)	0.0685	1.3841	0.3594	0.3611	0.2360	0.4336	0.6732	0.3069	1.3752	0.3426	0.3632	0.2310	0.3183	0.7878	0.3109

TABLE 5.1 – Simulation results (case (i), $\pi = 0.15$). SE : average standard error. SD : empirical standard deviation. CP : empirical coverage probability of 95%-level confidence intervals. ℓ(CI) : average length of confidence intervals. All results are based on $N = 5000$ simulated samples.

n	$\widehat{\pi}$	$\widehat{\beta}_1$							$\widehat{\beta}_2$						
		$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{1,6}$	$\widehat{\beta}_{1,7}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$	$\widehat{\beta}_{2,6}$	$\widehat{\beta}_{2,7}$
150															
bias	-0.0044	-0.0014	0.0418	0.0180	-0.0314	-0.0297	0.0152	0.0013	0.0113	0.0220	0.0023	-0.0247	0.0185	-0.0554	0.0013
rel. bias	1.7592	0.4549	3.4793	3.5926	4.1922	2.9665	1.9008	-	2.2654	4.3973	-	4.9417	3.6989	5.0360	-
SD	0.0379	0.7106	0.1877	0.1874	0.1243	0.2278	0.3487	0.1614	0.7164	0.1815	0.1884	0.1228	0.1718	0.4151	0.1659
SE	0.0377	0.7003	0.1826	0.1828	0.1205	0.2201	0.3403	0.1562	0.6972	0.1751	0.1842	0.1183	0.1655	0.4023	0.1588
CP	0.9392	0.9492	0.9454	0.9432	0.9408	0.9460	0.9476	0.9446	0.9460	0.9474	0.9448	0.9418	0.9438	0.9454	0.9466
ℓ(CI)	0.1474	2.7343	0.7127	0.7142	0.4704	0.8584	1.3294	0.6092	2.7216	0.6828	0.7196	0.4616	0.6414	1.5672	0.6191
300															
bias	-0.0024	-0.0061	0.0206	0.0095	-0.0157	-0.0146	0.0082	-0.0020	0.0014	0.0114	0.0003	-0.0115	0.0106	-0.0238	-0.0032
rel. bias	0.9510	2.0491	1.7208	1.9087	2.0899	1.4634	1.0232	-	0.2725	2.2891	-	2.3021	2.1201	2.1663	-
SD	0.0267	0.4926	0.1299	0.1270	0.0855	0.1545	0.2414	0.1103	0.4930	0.1235	0.1303	0.0833	0.1133	0.2795	0.1123
SE	0.0267	0.4865	0.1269	0.1271	0.0836	0.1525	0.2366	0.1083	0.4844	0.1214	0.1281	0.0821	0.1133	0.2778	0.1099
CP	0.9454	0.9456	0.9420	0.9502	0.9474	0.9480	0.9466	0.9488	0.9464	0.9438	0.9444	0.9452	0.9508	0.9498	0.9478
ℓ(CI)	0.1048	1.9036	0.4965	0.4974	0.3272	0.5965	0.9263	0.4234	1.8950	0.4748	0.5014	0.3210	0.4418	1.0861	0.4297
500															
bias	-0.0009	0.0044	0.0146	0.0042	-0.0093	-0.0071	0.0060	0.0006	0.0036	0.0077	-0.0005	-0.0064	0.0081	-0.0167	0.0013
rel. bias	0.3567	1.4647	1.2175	0.8392	1.2452	0.7136	0.7533	-	0.7233	1.5320	-	1.2787	1.6135	1.5187	-
SD	0.0210	0.3847	0.0984	0.1010	0.0654	0.1162	0.1847	0.0853	0.3771	0.0942	0.1002	0.0646	0.0880	0.2166	0.0868
SE	0.0208	0.3797	0.0988	0.0991	0.0651	0.1185	0.1848	0.0843	0.3781	0.0945	0.0999	0.0639	0.0876	0.2167	0.0856
CP	0.9488	0.9480	0.9508	0.9442	0.9460	0.9538	0.9504	0.9490	0.9506	0.9494	0.9474	0.9510	0.9496	0.9458	0.9492
ℓ(CI)	0.0816	1.4867	0.3869	0.3882	0.2548	0.4640	0.7238	0.3300	1.4802	0.3699	0.3912	0.2500	0.3422	0.8482	0.3349

TABLE 5.2 – Simulation results (case (i), $\pi = 0.25$). SE : average standard error. SD : empirical standard deviation. CP : empirical coverage probability of 95%-level confidence intervals. ℓ(CI) : average length of confidence intervals. All results are based on $N = 5000$ simulated samples.

n	$\widehat{\pi}$	$\widehat{\beta}_1$							$\widehat{\beta}_2$						
		$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{1,6}$	$\widehat{\beta}_{1,7}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$	$\widehat{\beta}_{2,6}$	$\widehat{\beta}_{2,7}$
150															
bias	-0.0044	0.0018	0.0748	0.0270	-0.0501	-0.0479	0.0379	-0.0040	0.0056	0.0453	0.0053	-0.0360	0.0282	-0.0723	-0.0018
rel. bias	0.8814	0.6009	6.2302	5.3927	6.6748	4.7923	4.7384	-	1.1229	9.0521	-	7.2004	5.6449	6.5744	-
SD	0.0430	0.9224	0.2484	0.2445	0.1631	0.2952	0.4615	0.2110	0.9107	0.2426	0.2454	0.1620	0.2332	0.5426	0.2144
SE	0.0430	0.9060	0.2377	0.2364	0.1566	0.2842	0.4410	0.2030	0.9061	0.2295	0.2390	0.1543	0.2182	0.5262	0.2070
CP	0.9462	0.9454	0.9408	0.9446	0.9424	0.9495	0.9426	0.9450	0.9521	0.9386	0.9468	0.9384	0.9434	0.9529	0.9460
ℓ(CI)	0.1685	3.5248	0.9238	0.9210	0.6087	1.1039	1.7160	0.7885	3.5223	0.8908	0.9305	0.5995	0.8383	2.0361	0.8033
300															
bias	-0.0025	-0.0022	0.0393	0.0132	-0.0209	-0.0267	0.0223	-0.0001	-0.0037	0.0211	0.0031	-0.0137	0.0144	-0.0306	0.0004
rel. bias	0.5068	0.7479	3.2737	2.6451	2.7891	2.6654	2.7856	-	0.7318	4.2187	-	2.7471	2.8800	2.7807	-
SD	0.0304	0.6204	0.1669	0.1618	0.1089	0.1985	0.3078	0.1388	0.6099	0.1588	0.1630	0.1068	0.1474	0.3579	0.1420
SE	0.0304	0.6140	0.1609	0.1604	0.1059	0.1922	0.2989	0.1370	0.6125	0.1546	0.1619	0.1043	0.1442	0.3523	0.1393
CP	0.9460	0.9470	0.9416	0.9462	0.9426	0.9466	0.9450	0.9516	0.9516	0.9430	0.9480	0.9480	0.9494	0.9544	0.9470
ℓ(CI)	0.1190	2.3985	0.6285	0.6268	0.4136	0.7502	1.1683	0.5347	2.3919	0.6035	0.6326	0.4072	0.5600	1.3742	0.5436
500															
bias	-0.0007	-0.0069	0.0205	0.0096	-0.0147	-0.0103	0.0059	-0.0007	-0.0039	0.0134	0.0035	-0.0114	0.0117	-0.0242	-0.0001
rel. bias	0.1398	2.2894	1.7044	1.9224	1.9555	1.0298	0.7322	-	0.7717	2.6854	-	2.2882	2.3352	2.1995	-
SD	0.0236	0.4795	0.1259	0.1253	0.0834	0.1521	0.2333	0.1083	0.4812	0.1213	0.1267	0.0822	0.1127	0.2756	0.1100
SE	0.0235	0.4747	0.1243	0.1240	0.0819	0.1478	0.2311	0.1057	0.4736	0.1194	0.1252	0.0807	0.1101	0.2717	0.1074
CP	0.9446	0.9496	0.9468	0.9510	0.9452	0.9416	0.9484	0.9472	0.9456	0.9454	0.9470	0.9450	0.9498	0.9478	0.9432
ℓ(CI)	0.0923	1.8573	0.4863	0.4853	0.3205	0.5779	0.9046	0.4134	1.8525	0.4670	0.4900	0.3156	0.4295	1.0624	0.4200

TABLE 5.3 – Simulation results (case (i), $\pi = 0.50$). SE : average standard error. SD : empirical standard deviation. CP : empirical coverage probability of 95%-level confidence intervals. ℓ(CI) : average length of confidence intervals. All results are based on $N = 5000$ simulated samples.

n	$\widehat{\gamma}$				$\widehat{\beta}_1$					$\widehat{\beta}_2$				
	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$	$\widehat{\gamma}_4$	$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$
150														
bias	-0.0815	0.0160	-0.0982	0.0492	0.0161	0.0077	-0.0185	-0.0166	-0.0063	0.0073	0.0001	-0.0149	0.0447	-0.0258
rel. bias	4.0961	5.3854	16.5987	6.9716	5.1057	1.5184	2.4504	1.5224	-	1.5428	-	2.9690	3.0396	2.4189
SD	1.0758	0.2874	0.6583	0.2867	0.6900	0.1858	0.1192	0.3212	0.3173	0.6565	0.1796	0.1149	0.2508	0.3285
SE	1.1344	0.3004	0.7909	0.2774	0.6652	0.1790	0.1174	0.3177	0.3152	0.6453	0.1754	0.1119	0.2524	0.3273
CP	0.9686	0.9665	0.9703	0.9502	0.9420	0.9430	0.9486	0.9490	0.9490	0.9480	0.9457	0.9476	0.9554	0.9490
ℓ(CI)	4.3992	1.1665	2.5944	1.0756	2.5969	0.6992	0.4582	1.2409	1.2317	2.5200	0.6854	0.4368	0.9826	1.2784
300														
bias	-0.0594	0.0106	-0.0540	0.0214	-0.0040	0.0062	-0.0091	-0.0026	-0.0012	0.0035	0.0000	-0.0087	0.0261	-0.0163
rel. bias	2.9725	3.5427	9.0045	3.0545	1.3362	1.2364	1.2116	0.2647	-	0.6907	-	1.7311	1.7418	1.4854
SD	0.7920	0.2110	0.4482	0.1942	0.4662	0.1258	0.0827	0.2255	0.2259	0.4485	0.1230	0.0788	0.1770	0.2307
SE	0.7785	0.2059	0.4431	0.1890	0.4662	0.1253	0.0820	0.2219	0.2208	0.4528	0.1229	0.0781	0.1760	0.2294
CP	0.9487	0.9493	0.9594	0.9497	0.9505	0.9479	0.9505	0.9483	0.9449	0.9505	0.9515	0.9521	0.9481	0.9531
ℓ(CI)	3.0378	0.8041	1.7149	0.7374	1.8246	0.4904	0.3206	0.8685	0.8645	1.7720	0.4812	0.3056	0.6878	0.8981
500														
bias	-0.0405	0.0072	-0.0350	0.0077	0.0035	0.0026	-0.0063	-0.0004	-0.0013	-0.0002	0.0006	-0.0051	0.0158	-0.0082
rel. bias	1.9923	2.3617	5.8320	1.1149	1.2029	0.5019	0.8367	0.0402	-	0.0125	-	1.0210	1.0496	0.7454
SD	0.6003	0.1589	0.3379	0.1477	0.3673	0.0995	0.0647	0.1735	0.1774	0.3539	0.0964	0.0617	0.1387	0.1806
SE	0.5957	0.1576	0.3344	0.1438	0.3668	0.0986	0.0642	0.1742	0.1736	0.3564	0.0968	0.0612	0.1381	0.1802
CP	0.9536	0.9536	0.9562	0.9466	0.9486	0.9466	0.9480	0.9518	0.9454	0.9538	0.9464	0.9508	0.9502	0.9504
ℓ(CI)	2.3281	0.6161	1.3020	0.5622	1.4364	0.3862	0.2515	0.6824	0.6802	1.3957	0.3791	0.2396	0.5403	0.7060

TABLE 5.4 – Simulation results (case (ii), average sample proportion of zero-inflation equal to 0.15). SE : average standard error. SD : empirical standard deviation. CP : empirical coverage probability of 95%-level confidence intervals. ℓ(CI) : average length of confidence intervals. All results are based on $N = 5000$ simulated samples.

	$\widehat{\gamma}$				$\widehat{\beta}_1$					$\widehat{\beta}_2$				
n	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$	$\widehat{\gamma}_4$	$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$
150														
bias	-0.0135	-0.0272	0.0265	-0.0026	0.0100	0.0098	-0.0224	-0.0130	-0.0015	-0.0013	0.0022	-0.0153	0.0453	-0.0227
rel. bias	4.5044	5.4456	2.6547	-	3.3181	1.9535	2.9847	1.2987	-	0.2506	-	3.0529	3.0169	2.0664
SD	1.0894	0.3109	0.5223	0.2544	0.7035	0.1850	0.1222	0.3216	0.3466	0.6805	0.1807	0.1157	0.2592	0.3604
SE	1.0610	0.3037	0.5125	0.2522	0.6884	0.1825	0.1191	0.3216	0.3355	0.6637	0.1774	0.1134	0.2566	0.3526
CP	0.9534	0.9566	0.9604	0.9570	0.9488	0.9480	0.9426	0.9560	0.9454	0.9488	0.9474	0.9446	0.9512	0.9516
ℓ(CI)	4.1293	1.1798	1.9964	0.9812	2.6895	0.7132	0.4650	1.2562	1.3108	2.5928	0.6932	0.4425	0.9985	1.3772
300														
bias	0.0055	-0.0158	0.0150	0.0007	-0.0033	0.0070	-0.0099	-0.0130	-0.0001	-0.0001	0.0014	-0.0072	0.0183	-0.0137
rel. bias	1.8310	3.1672	1.5024	-	1.0957	1.4021	1.3263	1.3016	-	0.0156	-	1.4353	1.2184	1.2463
SD	0.7349	0.2109	0.3517	0.1751	0.4906	0.1298	0.0826	0.2282	0.2337	0.4731	0.1263	0.0800	0.1821	0.2465
SE	0.7227	0.2064	0.3487	0.1719	0.4820	0.1277	0.0830	0.2252	0.2347	0.4647	0.1241	0.0790	0.1790	0.2467
CP	0.9536	0.9540	0.9552	0.9518	0.9480	0.9482	0.9522	0.9460	0.9542	0.9498	0.9496	0.9482	0.9504	0.9524
ℓ(CI)	2.8243	0.8060	1.3642	0.6718	1.8862	0.4997	0.3247	0.8813	0.9188	1.8184	0.4858	0.3089	0.6992	0.9653
500														
bias	-0.0080	-0.0068	0.0075	-0.0018	-0.0010	0.0039	-0.0063	-0.0037	0.0014	-0.0013	0.0010	-0.0043	0.0143	-0.0083
rel. bias	2.6646	1.3537	0.7487	-	0.3473	0.7880	0.8379	0.3745	-	0.2582	-	0.8692	0.9520	0.7584
SD	0.5491	0.1560	0.2688	0.1346	0.3754	0.1001	0.0667	0.1770	0.1894	0.3643	0.0972	0.0629	0.1410	0.1926
SE	0.5538	0.1578	0.2669	0.1315	0.3799	0.1006	0.0652	0.1767	0.1846	0.3665	0.0979	0.0621	0.1406	0.1940
CP	0.9548	0.9556	0.9492	0.9478	0.9486	0.9484	0.9442	0.9526	0.9442	0.9480	0.9494	0.9466	0.9508	0.9530
ℓ(CI)	2.1669	0.6172	1.0451	0.5145	1.4873	0.3940	0.2552	0.6919	0.7228	1.4352	0.3833	0.2430	0.5499	0.7597

TABLE 5.5 – Simulation results (case (ii), average sample proportion of zero-inflation equal to 0.25). SE : average standard error. SD : empirical standard deviation. CP : empirical coverage probability of 95%-level confidence intervals. ℓ(CI) : average length of confidence intervals. All results are based on $N = 5000$ simulated samples.

	$\widehat{\gamma}$				$\widehat{\beta}_1$					$\widehat{\beta}_2$				
n	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$	$\widehat{\gamma}_4$	$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$
150														
bias	-0.0701	0.0260	-0.0455	0.0059	0.0090	0.0259	-0.0551	-0.0245	-0.0021	0.0053	0.0041	-0.0416	0.1009	-0.0467
rel. bias	7.0132	5.1997	5.6880	5.9408	2.9927	5.1829	7.3499	2.4477	-	1.0558	-	8.3119	6.7252	4.2454
SD	0.7962	0.2210	0.4025	0.1898	1.0438	0.2950	0.1915	0.4979	0.4833	1.0219	0.2926	0.1868	0.4000	0.4972
SE	0.7861	0.2167	0.3987	0.1838	0.9957	0.2818	0.1831	0.4815	0.4638	0.9731	0.2786	0.1747	0.3818	0.4721
CP	0.9504	0.9482	0.9532	0.9498	0.9422	0.9444	0.9428	0.9492	0.9446	0.9464	0.9440	0.9390	0.9456	0.9398
ℓ(CI)	3.0757	0.8475	1.5599	0.7189	3.8652	1.0931	0.7087	1.8688	1.8068	3.7779	1.0815	0.6752	1.4704	1.8383
300														
bias	-0.0497	0.0175	-0.0199	0.0034	-0.0061	0.0141	-0.0242	-0.0099	-0.0054	-0.0072	0.0022	-0.0168	0.0547	-0.0211
rel. bias	4.9747	3.4938	2.4931	3.4063	2.0351	2.8171	3.2227	0.9883	-	1.4330	-	3.3580	3.6490	1.9210
SD	0.5532	0.1521	0.2801	0.1280	0.6947	0.1948	0.1267	0.3296	0.3212	0.6748	0.1918	0.1221	0.2621	0.3316
SE	0.5450	0.1501	0.2759	0.1273	0.6740	0.1905	0.1233	0.3268	0.3169	0.6588	0.1887	0.1175	0.2576	0.3224
CP	0.9480	0.9486	0.9494	0.9526	0.9450	0.9490	0.9488	0.9542	0.9498	0.9472	0.9476	0.9458	0.9504	0.9458
ℓ(CI)	2.1344	0.5877	1.0806	0.4984	2.6307	0.7435	0.4806	1.2756	1.2392	2.5716	0.7365	0.4579	1.0019	1.2606
500														
bias	-0.0177	0.0061	-0.0082	0.0013	-0.0085	0.0097	-0.0134	-0.0042	-0.0024	-0.0059	0.0021	-0.0111	0.0330	-0.0113
rel. bias	1.7730	1.2194	1.0287	1.2563	2.8342	1.9363	1.7895	0.4181	-	1.1859	-	2.2237	2.1968	1.0276
SD	0.4136	0.1141	0.2110	0.0990	0.5222	0.1491	0.0958	0.2543	0.2514	0.5198	0.1471	0.0931	0.2009	0.2590
SE	0.4188	0.1152	0.2120	0.0976	0.5236	0.1477	0.0953	0.2533	0.2465	0.5128	0.1465	0.0909	0.1994	0.2510
CP	0.9556	0.9554	0.9534	0.9500	0.9534	0.9516	0.9512	0.9486	0.9432	0.9460	0.9518	0.9482	0.9480	0.9430
ℓ(CI)	1.6409	0.4514	0.8305	0.3824	2.0475	0.5777	0.3723	0.9906	0.9651	2.0052	0.5731	0.3550	0.7781	0.9826

TABLE 5.6 – Simulation results (case (ii), average sample proportion of zero-inflation equal to 0.50). SE : average standard error. SD : empirical standard deviation. CP : empirical coverage probability of 95%-level confidence intervals. ℓ(CI) : average length of confidence intervals. All results are based on $N = 5000$ simulated samples.

	$\widehat{\gamma}$				$\widehat{\beta}_1$					$\widehat{\beta}_2$				
	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$	$\widehat{\gamma}_4$	$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$
bias	-0.0243	0.0047	-0.0356	0.0140	-0.0015	0.0037	-0.0054	-0.0048	-0.0002	0.0020	0.0003	-0.0049	0.0122	-0.0093
rel. bias	1.2150	1.5667	5.9333	2.0000	0.5000	0.7400	0.7200	0.4800	-	0.4000	-	0.9800	0.8133	0.8455
SD	0.5856	0.1546	0.3365	0.1447	0.3669	0.0989	0.0656	0.1749	0.1772	0.3600	0.0966	0.0619	0.1366	0.1803
SE	0.5944	0.1573	0.3339	0.1439	0.3667	0.0986	0.0643	0.1742	0.1735	0.3562	0.0968	0.0612	0.1378	0.1801
CP	0.9566	0.9580	0.9584	0.9538	0.9478	0.9534	0.9458	0.9506	0.9454	0.9442	0.9490	0.9500	0.9506	0.9488
ℓ(CI)	2.3238	0.6154	1.3001	0.5625	1.4358	0.3862	0.2516	0.6820	0.6796	1.3947	0.3789	0.2398	0.5392	0.7055
bias					-0.1604	-0.2553	0.2738	-0.3744	0.4879	-0.0823	-0.2259	0.2754	-0.6350	0.6208
rel. bias					53.4667	51.0600	36.5067	37.4400	-	16.4600	-	55.0800	42.3333	56.4364
SD					0.3936	0.1067	0.0689	0.1819	0.1903	0.3825	0.1023	0.0626	0.1669	0.1855
SE					0.2896	0.0758	0.0466	0.1352	0.1381	0.2596	0.0697	0.0415	0.0764	0.1361
CP					0.8162	0.1566	0.0050	0.2682	0.1206	0.8076	0.1856	0.0016	0.0068	0.0300
ℓ(CI)					1.1344	0.2968	0.1823	0.5294	0.5412	1.0170	0.2731	0.1625	0.2982	0.5334

TABLE 5.7 – Robustness of multinomial regression to zero-inflation : simulation results (case (ii), $n = 500$, average sample proportion of zero-inflation equal to 0.15). Upper panel : ZIM model. Lower panel : multinomial regression model. All results are based on $N = 5000$ simulated samples.

5.4 An application in health economics

5.4.1 Data description and competing models

In this section, we apply the proposed ZIM model to health-care utilization data obtained from the National Medical Expenditure Survey conducted in 1987-1988. This data set was first described by Deb and Trivedi [1997]. We consider jointly three health-care utilization measures : the number *ofnd* of consultations with a non-doctor in an office setting, the number *opnd* of consultations with a non-doctor in an outpatient setting and the number *ofd* of consultations with a doctor in an office setting.

The sample contains 3224 individuals with at least two consultations of all types among *ofnd*, *opnd* and *ofd*. Frequencies of individuals with zero occuring simultaneously in (*ofnd* and *opnd*), (*ofnd* and *ofd*) and (*opnd* and *ofd*) are 51.7%, 0.24% and 1% respectively. The high frequency of zeros in (*ofnd*, *opnd*) and low frequencies of zero counts in the other two pairs of health services suggest that there may exist permanent non-users of the combination (*ofnd* and *opnd*). That is, there may exist an excess of observations of the type $(0, 0, m_i)$, where m_i denotes the total number of consultations for individual i. Hence we propose to use ZIM model to investigate the determinants of health-care utilization in this data set.

Several covariates were recorded on each individual. They include : i) socio-economic variables : gender (1 for female, 0 for male), age (in years, divided by 10), marital status (1 if married, 0 if not married), educational level (number of years of education), income (in ten-thousands of dollars), ii) various measures of health status : number of chronic conditions (cancer, arthritis, diabete...) and a variable indicating self-perceived health level (poor, average, excellent) and iii) a binary variable indicating whether individual is covered by medicaid or not (medicaid is a US health insurance for individuals with limited income and resources, we code it as 1 if the individual is covered and 0 otherwise). Self-perceived health is re-coded as two dummy variables denoted by "health1" (1 if health is perceived as poor, 0 otherwise) and "health2" (1 health is perceived as average, 0 otherwise).

We fit the following three models : i) a multinomial logistic regression model mult$(m_i, \mathbf{p}_i)$,

where $\mathbf{p}_i = (p_{1i}, p_{2i}, p_{3i})$ and the p_{ji} are specified as in (5.3), ii) the ZIM model with fixed probability π of $(0, 0, m_i)$-inflation (denoted by ZIM^a thereafter) and iii) the ZIM model with covariate-dependent probability π_i of $(0, 0, m_i)$-inflation (denoted by ZIM^b), where π_i is as in (5.5). Selection of regressors for inclusion in π_i requires some care. Indeed, it was previously observed in various other zero-inflated models that including all available regressors in both count and zero-inflation probabilities can yield lack of identification of model parameters. See for example Diop et al. [2011] and Staub and Winkelmann [2013], who suggest to solve this issue by letting at least one of the covariates included in the count model to be excluded from the zero-inflation model (or the converse). Such condition is not required in the ZIM model. However, in order to avoid numerical problems, we propose a two-stage procedure for covariate selection in the zero-inflation and count models. In a first stage, we fit a standard logistic regression model with all available covariates to the binary indicator $\delta_{(0,0,m_i)}$. The resulting model is not a model for zero-inflation since some of the $(0, 0, m_i)$ may arise from the multinomial model mult$(m_i, \mathbf{p}_i)$. However, we expect that this rough procedure will still select a relevant subset of covariates, that will be used in a second stage in the logistic sub-model (5.5) for π_i. Using this procedure and Wald testing, we identify five significant predictors : age, gender, educational level, number of chronic conditions and medicaid status, that are included in π_i. Note that using this two-stage procedure and initial values for Newton-Raphson algorithm close to 0, we did not observe any convergence problem on this data set.

5.4.2 Results

Results for the three models (standard multinomial, ZIM^a and ZIM^b) are displayed in Table 5.8. We report estimate, standard error and significance level (as : not significant, significant or very significant) of Wald test for each parameter. Given the large number of observations of the type $(0, 0, m_i)$ and the lack of robustness of the multinomial model to zero-inflation (see Table 5.7), one should prefer ZIM model to fit these data. Now, in order to compare ZIM^a and ZIM^b models, we report their log-likelihood and AIC. Likelihood-based comparison of both models is meaningul since ZIM^a is nested in ZIM^b. To see this, note that if $\gamma^\top \mathbf{W}_i = \gamma_1 + \gamma_2 W_{i2} + \ldots + \gamma_q W_{iq}$ reduces to γ_1, ZIM^a with zero-inflation

probability $\pi_i = \pi$ (for $i = 1, \ldots, n$) coincides with ZIM^b with zero-inflation probability $\pi_i = \frac{e^{\gamma_1}}{1+e^{\gamma_1}}$ (for $i = 1, \ldots, n$). Here, ZIM^b appears as the best model in terms of both likelihood and AIC (in an unreported analysis, we also fitted ZIM^b with various other subsets of covariates in π_i. The smallest AIC is achieved for the subset selected by our two-stage procedure).

Among 1667 non-users of both *ofnd* and *opnd*, 41.5% are identified as permanent non-users by ZIM^a. Gender, educational level and medicaid status are identified by ZIM^b as the most influencing factors for being a permanent non-user, with medicaid recipients being more likely to be permanent non-users. The three models identify the same subset of influent factors for *opnd* utilization, with similar parameter estimates except for medicaid status : ZIM^a and ZIM^b suggest that probability of using *opnd* is less sensitive to medicaid status than suggested by standard multinomial regression. Moreover, for ZIM^a and ZIM^b, medicaid status does not affect *ofnd* utilization. These findings are coherent with the fact that part of the decision of (not) using *ofnd* and *opnd* by medicaid recipients was captured in the model for π_i. All this suggests that medicaid recipients tend to favor doctor visits in an office setting over non-doctor visits in either office or outpatient settings. This confirms previous findings that patients with medicaid insurance coverage have less non-doctor health professional visits. This feature is also captured by the standard multinomial regression model but ZIM model additionally confirms that medicaid recipients are more likely to decide to never use *ofnd* and *opnd* services. From ZIM^b, educational level is an important determinant of the decision of being a permanent non-user of both *ofnd* and *opnd*. But once an individual has chosen to use eventually these health-care services (with a probability that increases with level of education), ZIM^b suggests that schooling does not tend to favor a specific kind of health-care service. Income does not affect utilization of medical care. This is consistent with previous findings [*e.g.*, Deb and Trivedi, 1997] and is explained in the literature by the fact that income may affect intensity and quality of care rather than visits number. Marital status has a strong effect on *ofnd* and *opnd* utilization, with similar magnitude but opposite sign. Married patients are more likely to visit a doctor in an office setting, which may be due to couples having more financial resources than single individuals. Deb and Trivedi [1997] report that an increase in the number of chronic conditions increases utilization of each form of medical care. We find here that chronic

condition does not affect *opnd* and affects negatively *ofnd*. Thus, *ofd* utilization increases with the number of chronic conditions. Contradiction with conclusions by Deb and Trivedi [1997] is only apparent. By considering simultaneously *ofnd*, *opnd* and *ofd*, we are able to rank the various forms of medical care by order of utilization. Our observation reflects the fact that as the number of chronic conditions increases, doctor visits are preferred to non-doctor visits, which seems natural.

		multinomial model			ZIM^a			ZIM^b		
parameter	variable	est.	s.e.	test	est.	s.e.	test	est.	s.e.	test
$\beta_{1,1}$	intercept	-1.6311	0.2511	VS	-0.8986	0.2887	VS	-0.9331	0.3883	S
$\beta_{1,2}$	health1	-0.8457	0.0919	VS	-0.7275	0.1058	VS	-0.7308	0.1043	VS
$\beta_{1,3}$	health2	-0.3143	0.0681	VS	-0.3089	0.0793	VS	-0.3072	0.0790	VS
$\beta_{1,4}$	chronic	-0.0903	0.0141	VS	-0.1243	0.0161	VS	-0.1270	0.0164	VS
$\beta_{1,5}$	age	-0.0287	0.0301	NS	0.0023	0.0349	NS	0.0214	0.0445	NS
$\beta_{1,6}$	gender	0.3155	0.0407	VS	0.2058	0.0462	VS	0.1839	0.0475	VS
$\beta_{1,7}$	marital status	0.2160	0.0414	VS	0.2028	0.0468	VS	0.2031	0.0473	VS
$\beta_{1,8}$	educational	0.0405	0.0055	VS	0.0152	0.0064	S	0.0071	0.0068	NS
$\beta_{1,9}$	income	-0.0084	0.0061	NS	-0.0098	0.0065	NS	-0.0093	0.0065	NS
$\beta_{1,10}$	medicaid	-0.3406	0.0757	VS	-0.1217	0.0908	NS	-0.0276	0.0893	NS
$\beta_{2,1}$	intercept	1.0023	0.4982	S	1.8090	0.5235	VS	1.7695	0.4370	VS
$\beta_{2,2}$	health1	0.4011	0.1838	S	0.5185	0.1807	VS	0.5102	0.1891	VS
$\beta_{2,3}$	health2	0.4084	0.1611	S	0.4063	0.1567	VS	0.4051	0.1718	S
$\beta_{2,4}$	chronic	-0.0036	0.0232	NS	-0.0339	0.0246	NS	-0.0363	0.0249	NS
$\beta_{2,5}$	age	-0.5980	0.0593	VS	-0.5741	0.0627	VS	-0.5539	0.0519	VS
$\beta_{2,6}$	gender	0.0870	0.0698	NS	-0.0079	0.0729	NS	-0.0301	0.0747	NS
$\beta_{2,7}$	marital status	-0.2317	0.0708	VS	-0.2407	0.0732	VS	-0.2407	0.0754	VS
$\beta_{2,8}$	educational	0.0185	0.0096	NS	-0.0100	0.0105	NS	-0.0180	0.0104	NS
$\beta_{2,9}$	income	0.0113	0.0095	NS	0.0112	0.0095	NS	0.0116	0.0094	NS
$\beta_{2,10}$	medicaid	-0.6667	0.1385	VS	-0.4809	0.1522	VS	-0.3905	0.1605	S
π					0.4150	0.0107				
γ_1	intercept							-0.5814	1.3793	NS
γ_2	chronic							-0.0345	0.0339	NS
γ_3	age							0.1661	0.1706	NS
γ_4	gender							-0.2711	0.0994	VS
γ_5	educational							-0.0763	0.0150	VS
γ_6	medicaid							0.5784	0.1788	VS
log-lik					-14183.48			**-14142.65**		
AIC					28408.97			**28337.31**		

TABLE 5.8 – Health-care data analysis : estimates, log-likelihood and AIC values from multinomial, ZIM^a and ZIM^b models (NS : not significant at the 5% level, S : significant at level between 1% and 5%, VS (very significant) : significant at level less than 1%).

5.4.3 Some further numerical considerations

In order to assess stability of the estimates given in Table 5.8, we obtain $B = 500$ bootstrap samples of size 3224 from the initial data set and we estimate models ZIM^a and ZIM^b on each of them. Then, we calculate the average value of the B estimates, for each parameter (values are reported in Table 5.9). We compare these values to the maximum likelihood estimates of Table 5.8. We observe that the average bootstrap estimates are close to the maximum likelihood estimates, which indicates the good stability property of the estimation procedure described in Section 5.2.

From Theorem 5.2.3, the maximum likelihood estimator of ψ in ZIM model is asymptotically multivariate normal. Multivariate normality can be assessed using a generalization [recently proposed by Villasenor-Alva and Gonzalez-Estrada, 2000] of Shapiro-Wilk's test for univariate normality. For illustrative purpose, we use this test for testing multivariate normality of $\widehat{\psi}_n$ in the ZIM^a and ZIM^b models fitted to the health-care utilization data, based on the B bootstrap samples. The test is available from the `R` package "mvShapiro-Test" [Gonzalez-Estrada and Villasenor-Alva, 2013]. The p-value obtained for ZIM^a model (respectively ZIM^b) is equal to 0.0892 (respectively 0.1582) and thus, there is no reason to doubt of multivariate normality of $\widehat{\psi}_n$, in either model.

parameter	variable	ZIM[a]	ZIM[b]
$\beta_{1,1}$	intercept	-0.8996	-1.0090
$\beta_{1,2}$	health1	-0.7326	-0.7303
$\beta_{1,3}$	health2	-0.3030	-0.2887
$\beta_{1,4}$	chronic	-0.1236	-0.1275
$\beta_{1,5}$	age	0.0007	0.0271
$\beta_{1,6}$	gender	0.2038	0.1899
$\beta_{1,7}$	marital status	0.2062	0.2070
$\beta_{1,8}$	educational	0.0158	0.0079
$\beta_{1,9}$	income	-0.0113	-0.0094
$\beta_{1,10}$	medicaid	-0.1334	-0.0038
$\beta_{2,1}$	intercept	1.7966	1.7231
$\beta_{2,2}$	health1	0.5150	0.4837
$\beta_{2,3}$	health2	0.4347	0.4149
$\beta_{2,4}$	chronic	-0.0326	-0.0365
$\beta_{2,5}$	age	-0.5774	-0.5552
$\beta_{2,6}$	gender	-0.0017	-0.0068
$\beta_{2,7}$	marital status	-0.2489	-0.2519
$\beta_{2,8}$	educational	-0.0102	-0.0155
$\beta_{2,9}$	income	0.0126	0.0138
$\beta_{2,10}$	medicaid	-0.5060	-0.4088
π		0.4141	
γ_1	intercept		-0.6102
γ_2	chronic		-0.0317
γ_3	age		0.1713
γ_4	gender		-0.2673
γ_5	educational		-0.0785
γ_6	medicaid		0.5712

TABLE 5.9 – Health-care data analysis : bootstrap average estimates (all estimates are based on B bootstrap samples).

5.5 Conclusion

In this paper, we introduce a model for multivariate count data with excess zeros when zero-inflation affects jointly several component counts. Maximum likelihood estimation is shown to perform well in this model, under a range of scenarios. Moreover, in our analysis of health-care utilization, the proposed model provides plausible explanations and interpretations and gives useful insight into the decision of using or not available health-care services. Several issues now deserve attention, such as derivation of a formal test for zero-inflation in multinomial counts. Generalizing the proposed model to more complex settings (*e.g.*, cluster correlation, longitudinal or hierarchical data) is also desirable and constitutes the topic for our future work.

Appendix A. Proof of identifiability.

Suppose that $l_{[i]}(\psi) = l_{[i]}(\psi^*)$ almost surely. Under C1 and C2, there exists $\varepsilon > 0$ such that for every $\mathbf{X}_i$ and $\psi \in \mathbf{K}$, $\varepsilon < \mathbb{P}\left(Z_i \neq (0, 0, m_i)|\mathbf{X}_i\right) = (1-\pi)\left(1 - (h_i(\beta))^{-m_i}\right)$. Therefore, we can find $\omega \in \Omega$, with ω outside the negligible set where $l_{[i]}(\psi) \neq l_{[i]}(\psi^*)$, such that $Z_i(\omega) \neq (0, 0, m_i)$. For such ω, $J_i = 1$ and thus, $l_{[i]}(\psi) = l_{[i]}(\psi^*)$ becomes :

$$Z_{1i}(\beta_1 - \beta_1^*)^\top \mathbf{X}_i + Z_{2i}(\beta_2 - \beta_2^*)^\top \mathbf{X}_i = \log\left[\left(\frac{h_i(\beta)}{h_i(\beta^*)}\right)^{m_i} \times \left(\frac{1-\pi^*}{1-\pi}\right)\right]. \tag{5.6}$$

The right-hand side of (5.6) does not depend on $Z_i = (Z_{1i}, Z_{2i}, Z_{3i})$. Therefore, the left-hand side of (5.6) should be constant for two different values of Z_i. Consider for example $Z_i = (z_{1i}, z_{2i}, m_i - z_{1i} - z_{2i})$ and $Z_i = (z_{1i}, z_{2i} - 1, m_i - z_{1i} - z_{2i} + 1)$, with $z_{1i}, z_{2i} \geq 1$ (which is possible since $m_i \geq 2$ by C4). Then we obtain $(\beta_2 - \beta_2^*)^\top \mathbf{X}_i = 0$. A similar argument yields $\beta_1 = \beta_1^*$ and finally, $\pi = \pi^*$, which concludes the proof. □

Appendix B. Proofs of asymptotic results.

An intermediate technical lemma is first proved.

Lemma 5.5.1 *Let $\phi_n : \mathbb{R}^k \longrightarrow \mathbb{R}^k$ be defined as $\phi_n(\psi) = \psi + (\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top)^{-1}\dot{l}_n(\psi)$. Then there exists an open ball $B(\psi_0, r)$ (with $r > 0$) and a constant $0 < c < 1$ such that :*

$$\left\|\phi_n(\psi) - \phi_n(\widetilde{\psi})\right\| \leq c \left\|\psi - \widetilde{\psi}\right\| \text{ for all } \psi, \widetilde{\psi} \in B(\psi_0, r). \tag{5.7}$$

Proof of Lemma 5.5.1. Property (5.7) holds if we can prove that $\left\|\frac{\partial \phi_n(\psi)}{\partial \psi^\top}\right\| \leq c$ for all $\psi \in B(\psi_0, r)$. We have :

$$\begin{aligned}
\left\|\frac{\partial \phi_n(\psi)}{\partial \psi^\top}\right\| &= \left\|I_k + (\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top)^{-1}\ddot{l}_n(\psi)\right\| \\
&= \left\|(\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top)^{-1}\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{V}^\top\right\| \\
&\leq \left\|(\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top)^{-1}\right\| \left\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{V}^\top\right\| \\
&= \lambda_n^{-1} \left\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{V}^\top\right\|.
\end{aligned}$$

Now, let $\mathscr{I}$ denote the set of indices $\{(i,j) \in \{1, 2, \ldots, 3n\}^2$ such that $\mathbb{D}_{ij}(\psi_0) \neq 0\}$. Then the following holds :

$$\begin{aligned}
\left\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{V}^\top\right\| &= \left\|\sum_{i=1}^{3n}\sum_{j=1}^{3n} \mathbb{V}_{\bullet i}\mathbb{V}_{\bullet j}^\top(\mathbb{D}_{ij}(\psi) - \mathbb{D}_{ij}(\psi_0))\right\| \\
&\leq \sum_{(i,j)\in\mathscr{I}} \left\|\mathbb{V}_{\bullet i}\mathbb{V}_{\bullet j}^\top\mathbb{D}_{ij}(\psi_0)\right\| \left|\frac{\mathbb{D}_{ij}(\psi) - \mathbb{D}_{ij}(\psi_0)}{\mathbb{D}_{ij}(\psi_0)}\right|.
\end{aligned}$$

Under conditions C1 and C2, there exists a constant c_2 ($c_2 > 0$) such that $|\mathbb{D}_{ij}(\psi_0)| > c_2$ for every $(i,j) \in \mathscr{I}$. For example, consider the case where $\mathbb{D}_{ij}(\psi_0)$ coincides with some $\mathbb{D}_{4,\ell\ell}(\psi_0)$, for $\ell \in \{1, \ldots, n\}$. Then

$$|\mathbb{D}_{4,\ell\ell}(\psi)| = \frac{m_\ell e^{\beta_1^\top \mathbf{x}_\ell}(h_\ell(\beta))^{m_\ell - 1}}{(\pi[(h_\ell(\beta))^{m_\ell} - 1] + 1)^2} > \frac{m_{\mathbf{X}}}{(1 + 2M_{\mathbf{X}})^{2m_\ell}},$$

where $m_{\mathbf{X}} := \min_{\beta,\mathbf{X}} e^{\beta^\top \mathbf{X}}$ and $M_{\mathbf{X}} := \max_{\beta,\mathbf{X}} e^{\beta^\top \mathbf{X}}$. Under C1, C2, C4, there exists a positive constant d_4 such that $\frac{m_{\mathbf{X}}}{(1+2M_{\mathbf{X}})^{2m_\ell}} > d_4$. Using similar arguments, we obtain that $|\mathbb{D}_{i,\ell\ell}(\psi)| > d_i$ for some $d_i, i = 1, \ldots, 6$. If $c_2 = \min_{1\leq i\leq 6} d_i$, we obtain $|\mathbb{D}_{ij}(\psi_0)| > c_2$ for

every $(i,j) \in \mathscr{I}$. Moreover, $\mathbb{D}_{ij}(\cdot)$ is uniformly continuous on $\mathbf{K}$ thus for every $\varepsilon > 0$, there exists a positive r such that for all $\psi \in B(\psi_0, r)$, $|\mathbb{D}_{ij}(\psi) - \mathbb{D}_{ij}(\psi_0)| < \varepsilon$. It follows that

$$\begin{aligned}\left\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{V}^\top\right\| &\leq \frac{\varepsilon}{c_2} \sum_{(i,j)\in\mathscr{I}} \left\|\mathbb{V}_{\bullet i}\mathbb{V}_{\bullet j}^\top \mathbb{D}_{ij}(\psi_0)\right\| \\ &= \frac{\varepsilon}{c_2} \text{trace}\left(\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top\right) \\ &\leq \frac{\varepsilon}{c_2} k\Lambda_n.\end{aligned}$$

This in turn implies that $\left\|\frac{\partial \phi_n(\psi)}{\partial \psi^\top}\right\| \leq \frac{\varepsilon k \Lambda_n}{c_2 \lambda_n} < \frac{\varepsilon k c_1}{c_2}$. Now, choosing $\varepsilon = c\frac{c_2}{kc_1}$ with $0 < c < 1$, we get that $\left\|\frac{\partial \phi_n(\psi)}{\partial \psi^\top}\right\| \leq c$ for all $\psi \in B(\psi_0, r)$, which concludes the proof. □

Proof of Theorem 5.2.2. Let the function η_n be defined as $\eta_n(\psi) := \psi - \phi_n(\psi) = -(\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top)^{-1}\dot{l}_n(\psi)$. Then $\eta_n(\psi_0)$ converges almost surely to 0 as $n \to \infty$. To see this, note that

$$\eta_n(\psi_0) = \left(\frac{1}{n}\ddot{l}_n(\psi_0)\right)^{-1} \cdot \left(\frac{1}{n}\dot{l}_n(\psi_0)\right).$$

By C3, $\left(\frac{1}{n}\ddot{l}_n(\psi_0)\right)^{-1}$ converges to some matrix Σ. Moreover,

$$\frac{1}{n}\dot{l}_n(\psi_0) = \frac{1}{n}\mathbb{V}C(\psi_0) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n A_i(\psi_0) \\ \frac{1}{n}\sum_{i=1}^n X_{i1}B_{i,1}(\psi_0) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n X_{ip}B_{i,1}(\psi_0) \\ \frac{1}{n}\sum_{i=1}^n X_{i1}B_{i,2}(\psi_0) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n X_{ip}B_{i,2}(\psi_0) \end{pmatrix}$$

converges to 0 almost surely as $n \to \infty$. To see this, note that for every $i = 1, \ldots, n$, $\mathbb{E}[A_i(\psi_0)] = \mathbb{E}[\mathbb{E}[A_i(\psi_0)|\mathbf{X}_i]]$, where

$$\mathbb{E}[A_i(\psi_0)|\mathbf{X}_i] = \frac{(h_i(\beta_0))^{m_i} - 1}{\pi_0[(h_i(\beta_0))^{m_i} - 1] + 1}\mathbb{E}\left[1 - J_i|\mathbf{X}_i\right] - \frac{1}{1-\pi_0}\mathbb{E}\left[J_i|\mathbf{X}_i\right].$$

Now,

$$\mathbb{E}\left[J_i|\mathbf{X}_i\right] = \mathbb{P}(Z_i \neq (0,0,m_i)|\mathbf{X}_i) = (1-\pi_0)\left(1 - \frac{1}{(h_i(\beta_0))^{m_i}}\right),$$

thus

$$\begin{aligned}\mathbb{E}[A_i(\psi_0)|\mathbf{X}_i] &= \frac{(h_i(\beta_0))^{m_i}-1}{\pi_0[(h_i(\beta_0))^{m_i}-1]+1}\left[\pi_0+(1-\pi_0)\frac{1}{(h_i(\beta_0))^{m_i}}\right]-\left(1-\frac{1}{(h_i(\beta_0))^{m_i}}\right)\\ &= 0.\end{aligned}$$

It follows that $\mathbb{E}[A_i(\psi_0)]=0$. Next, for every $i=1,\ldots,n$,

$$\begin{aligned}\text{var}(A_i(\psi_0)) &= \mathbb{E}[\text{var}(A_i(\psi_0)|\mathbf{X}_i)]+\text{var}(\mathbb{E}[A_i(\psi_0)|\mathbf{X}_i])\\ &= \mathbb{E}[\text{var}(A_i(\psi_0)|\mathbf{X}_i)]\\ &\leq c_3 := \mathbb{E}\left[\left(\frac{(h_i(\beta_0))^{m_i}}{(1-\pi_0)\{\pi_0[(h_i(\beta_0))^{m_i}-1]+1\}}\right)^2\right].\end{aligned}$$

Conditions C1, C2, C4 ensure that $c_3<\infty$ and thus

$$\sum_{i=1}^{\infty}\frac{\text{var}(A_i(\psi_0))}{i^2}\leq c_3\sum_{i=1}^{\infty}\frac{1}{i^2}<\infty.$$

Kolmogorov's strong law of large numbers applies to terms $A_i(\psi_0)$ and implies that

$$\frac{1}{n}\sum_{i=1}^{n}\{A_i(\psi_0)-\mathbb{E}\left[A_i(\psi_0)\right]\}=\frac{1}{n}\sum_{i=1}^{n}A_i(\psi_0)$$

converges almost surely to 0.

Similarly, for every $i=1,\ldots,n$ and $j=1,\ldots,p$, we have $\mathbb{E}[X_{ij}B_{i,1}(\psi_0)]=\mathbb{E}[X_{ij}\mathbb{E}[B_{i,1}(\psi_0)|\mathbf{X}_i]]$, where

$$\mathbb{E}[B_{i,1}(\psi_0)|\mathbf{X}_i]=-(1-\pi_0)\frac{m_i e^{\beta_{1,0}^\top\mathbf{X}_i}}{k_i(\psi_0)}\mathbb{E}\left[1-J_i|\mathbf{X}_i\right]-\frac{m_i e^{\beta_{1,0}^\top\mathbf{X}_i}}{h_i(\beta_0)}\mathbb{E}\left[J_i|\mathbf{X}_i\right]+\mathbb{E}\left[J_iZ_{1i}|\mathbf{X}_i\right]$$

and

$$\mathbb{E}\left[J_iZ_{1i}|\mathbf{X}_i\right]=(1-\pi_0)m_i\frac{e^{\beta_{1,0}^\top\mathbf{X}_i}}{h_i(\beta_0)}.$$

Straightforward calculations yield $\mathbb{E}[B_{i,1}(\psi_0)|\mathbf{X}_i]=0$ and thus $\mathbb{E}[X_{ij}B_{i,1}(\psi_0)]=0$. Moreover,

$$\sum_{i=1}^{\infty}\frac{\text{var}(X_{ij}B_{i,1}(\psi_0))}{i^2}<\infty$$

by similar arguments as above. Therefore, $\frac{1}{n}\sum_{i=1}^{n}X_{ij}B_{i,1}(\psi_0)$ converges almost surely to 0. Similar result holds for $\frac{1}{n}\sum_{i=1}^{n}X_{ij}B_{i,2}(\psi_0)$. Finally, $\frac{1}{n}\dot{l}_n(\psi_0)$ and $\eta_n(\psi_0)$ converge almost surely to 0 as $n\to\infty$.

Now, let ε be an arbitrary positive value. Almost sure convergence of $\eta_n(\psi_0)$ implies that for almost every $\omega \in \Omega$, there exists an integer $n(\varepsilon, \omega)$ such that for any $n \geq n(\varepsilon, \omega)$, $\|\eta_n(\psi_0)\| \leq \varepsilon$ or equivalently, $0 \in B(\eta_n(\psi_0), \varepsilon)$. In particular, let $\varepsilon = (1-c)s$ with $0 < c < 1$ such as in Lemma 5.5.1. Since ϕ_n satisfies Lipschitz condition (5.7), Lemma 2 of Gouriéroux and Monfort [1981] ensures that there exists an element of $B(\psi_0, s)$ (let denote this element by $\widehat{\psi}_n$) such that $\eta_n(\widehat{\psi}_n) = 0$ that is, $(\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top)^{-1}\dot{l}_n(\widehat{\psi}_n) = 0$. Condition C3 implies that $\dot{l}_n(\widehat{\psi}_n) = 0$ and that $\widehat{\psi}_n$ is the unique maximizer of l_n.

To summarize, we have shown that for almost every $\omega \in \Omega$ and for every $s > 0$, there exists an integer value $n(s, \omega)$ such that if $n \geq n(s, \omega)$, then the maximum likelihood estimator $\widehat{\psi}_n$ exists and $\|\widehat{\psi}_n - \psi_0\| \leq s$ (that is, $\widehat{\psi}_n$ converges almost surely to ψ_0). □

Proof of Theorem 5.2.3. A Taylor expansion of the score function yields

$$0 = \dot{l}_n(\widehat{\psi}_n) = \dot{l}_n(\psi_0) + \ddot{l}_n(\widetilde{\psi}_n)(\widehat{\psi}_n - \psi_0),$$

where $\widetilde{\psi}_n$ lies between $\widehat{\psi}_n$ and ψ_0. Thus, $\dot{l}_n(\psi_0) = -\ddot{l}_n(\widetilde{\psi}_n)(\widehat{\psi}_n - \psi_0)$. Letting $\widetilde{\Sigma}_n := -\ddot{l}_n(\widetilde{\psi}_n) = \mathbb{V}\mathbb{D}(\widetilde{\psi}_n)\mathbb{V}^\top$ and $\Sigma_{n,0} := \mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top$, we have :

$$\widehat{\Sigma}_n^{\frac{1}{2}}(\widehat{\psi}_n - \psi_0) = \left[\widehat{\Sigma}_n^{\frac{1}{2}}\widetilde{\Sigma}_n^{-\frac{1}{2}}\right]\left[\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}}\right]\Sigma_{n,0}^{-\frac{1}{2}}\left(\widetilde{\Sigma}_n(\widehat{\psi}_n - \psi_0)\right). \tag{5.8}$$

Terms $[\widehat{\Sigma}_n^{\frac{1}{2}}\widetilde{\Sigma}_n^{-\frac{1}{2}}]$ and $[\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}}]$ in (5.8) converge almost surely to I_k. To see this, we show for example that $\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}} - I_k\| \xrightarrow{a.s.} 0$ as $n \to \infty$. First, note that

$$\left\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}} - I_k\right\| \leq \Lambda_n^{\frac{1}{2}}\left\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\right\|\left\|\Lambda_n^{-\frac{1}{2}}\left(\Sigma_{n,0}^{\frac{1}{2}} - \widetilde{\Sigma}_n^{\frac{1}{2}}\right)\right\|, \tag{5.9}$$

and

$$\Lambda_n^{-1}\left\|\Sigma_{n,0} - \widetilde{\Sigma}_n\right\| = \Lambda_n^{-1}\left\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\widetilde{\psi}_n))\mathbb{V}^\top\right\|.$$

By Theorem 5.2.2, $\widetilde{\psi}_n$ converges almost surely to ψ_0. Let $\omega \in \Omega$ be outside the negligible set where this convergence does not hold. By the same arguments as in proof of Lemma **??**, for every $\varepsilon > 0$, there exists $n(\varepsilon, \omega) \in \mathbb{N}$ such that if $n \geq n(\varepsilon, \omega)$, then $\Lambda_n^{-1}\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\widetilde{\psi}_n))\mathbb{V}^\top\| \leq \varepsilon$. Thus $\Lambda_n^{-1}\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\widetilde{\psi}_n))\mathbb{V}^\top\|$ converges almost surely to 0. By continuity of the map $A \mapsto A^{\frac{1}{2}}$, $\|\Lambda_n^{-\frac{1}{2}}(\Sigma_{n,0}^{\frac{1}{2}} - \widetilde{\Sigma}_n^{\frac{1}{2}})\|$ converges almost surely to 0. Moreover, for n sufficiently large, there exists $0 < c_4 < \infty$ such that almost surely,

$\Lambda_n^{\frac{1}{2}}\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\| \leq c_4\Lambda_n^{\frac{1}{2}}/\lambda_n^{\frac{1}{2}} < c_4 c_1^{\frac{1}{2}}$ (by condition C3). Thus $\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}} - I_k\|$ converges almost surely to 0. Almost sure convergence of $\|\widehat{\Sigma}_n^{\frac{1}{2}}\widetilde{\Sigma}_n^{-\frac{1}{2}} - I_k\|$ to 0 follows by similar arguments.

It remains us to show that $\Sigma_{n,0}^{-\frac{1}{2}}(\widetilde{\Sigma}_n(\widehat{\psi}_n - \psi_0))$ converges in distribution to the Gaussian vector $\mathscr{N}(0, I_k)$. Note that $\Sigma_{n,0}^{-\frac{1}{2}}(\widetilde{\Sigma}_n(\widehat{\psi}_n - \psi_0)) = \Sigma_{n,0}^{-\frac{1}{2}}\sum_{j=1}^{3n}\mathbb{V}_{\bullet j}C_j(\psi_0)$. Thus, by Eicker [1963], this convergence holds if we can check that the following conditions are fulfilled :

1. $\max_{1\leq j\leq 3n}\mathbb{V}_{\bullet j}^\top(\mathbb{V}\mathbb{V}^\top)^{-1}\mathbb{V}_{\bullet j} \to 0$ as $n \to \infty$,
2. $\sup_{1\leq j\leq 3n}\mathbb{E}[C_j(\psi_0)^2 1_{\{|C_j(\psi_0)|>c\}}] \to 0$ as $c \to \infty$,
3. $\inf_{1\leq j\leq 3n}\mathbb{E}[C_j(\psi_0)^2] > 0$.

Condition 1) follows by noting that

$$0 < \max_{1\leq j\leq 3n}\mathbb{V}_{\bullet j}^\top(\mathbb{V}\mathbb{V}^\top)^{-1}\mathbb{V}_{\bullet j} \leq \max_{1\leq j\leq 3n}\|\mathbb{V}_{\bullet j}\|^2\|(\mathbb{V}\mathbb{V}^\top)^{-1}\| = \max_{1\leq j\leq 3n}\|\mathbb{V}_{\bullet j}\|^2/\tilde{\lambda}_n$$

and that $\|\mathbb{V}_{\bullet j}\|$ is bounded, by C1. Moreover, $1/\tilde{\lambda}_n$ tends to 0 as $n \to \infty$ by C3. Condition 2) follows by noting that the $C_j(\psi_0)$, $j = 1, \ldots, 3n$ are bounded under C1, C2, C4. Finally, we note that $\mathbb{E}[C_j(\psi_0)^2] = \text{var}(C_j(\psi_0))$ since $\mathbb{E}[C_j(\psi_0)] = 0$, $j = 1, \ldots, 3n$. If $j \in \{1, \ldots, n\}$, $C_j(\psi_0) = A_j(\psi_0)$. Then $\text{var}(C_j(\psi_0)) = \text{var}(A_j(\psi_0)) = \mathbb{E}[\text{var}(A_j(\psi_0)|\mathbf{X}_j)]$. Now,

$$\begin{aligned}
\text{var}(A_j(\psi_0)|\mathbf{X}_j) &= \left(\frac{(h_j(\beta_0))^{m_j}}{(1-\pi_0)[\pi_0((h_j(\beta_0))^{m_j}-1)+1]}\right)^2 \text{var}(J_j|\mathbf{X}_j) \\
&= \left(\frac{(h_j(\beta_0))^{m_j}}{(1-\pi_0)[\pi_0((h_j(\beta_0))^{m_j}-1)+1]}\right)^2 \mathbb{P}(Z_j \neq (0,0,m_j)|\mathbf{X}_j)(1-\mathbb{P}(Z_j \neq (0,0,m_j)|\mathbf{X}_j)) \\
&= \left(\frac{(h_j(\beta_0))^{m_j}}{(1-\pi_0)[\pi_0((h_j(\beta_0))^{m_j}-1)+1]}\right)^2 \left((1-\pi_0)(1-\frac{1}{(h_j(\beta_0))^{m_j}})\right)\left(\pi_0 + (1-\pi_0)\frac{1}{(h_j(\beta_0))^{m_j}}\right),
\end{aligned}$$

and thus, $\text{var}(A_j(\psi_0)|\mathbf{X}_j) > 0$ for every $j = 1, \ldots, n$ by C1, C2, C4. It follows that $\text{var}(C_j(\psi_0)) > 0$ for every $j = 1, \ldots, n$. By similar arguments, $\text{var}(C_j(\psi_0)) > 0$ for every $j = 1, \ldots, 3n$ and condition 3) is satisfied.

To summarize, we have proved that $\Sigma_{n,0}^{-\frac{1}{2}}(\widetilde{\Sigma}_n(\widehat{\psi}_n - \psi_0))$ converges in distribution to $\mathscr{N}(0, I_k)$. This result combined with Slutsky's theorem and equation (5.8) implies that $\widehat{\Sigma}_n^{\frac{1}{2}}(\widehat{\psi}_n - \psi_0)$ converges in distribution to $\mathscr{N}(0, I_k)$. □

Acknowledgements

Authors acknowledge financial support from the "Service de Coopération et d'Action Culturelle" of the French Embassy in Senegal and logistical support from Campus France (French national agency for the promotion of higher education, international student services, and international mobility). Authors also acknowledge grants from CEA-MITIC, an African Center of Excellence in Mathematics, Informatics and ICT implemented by Gaston Berger University (Senegal).

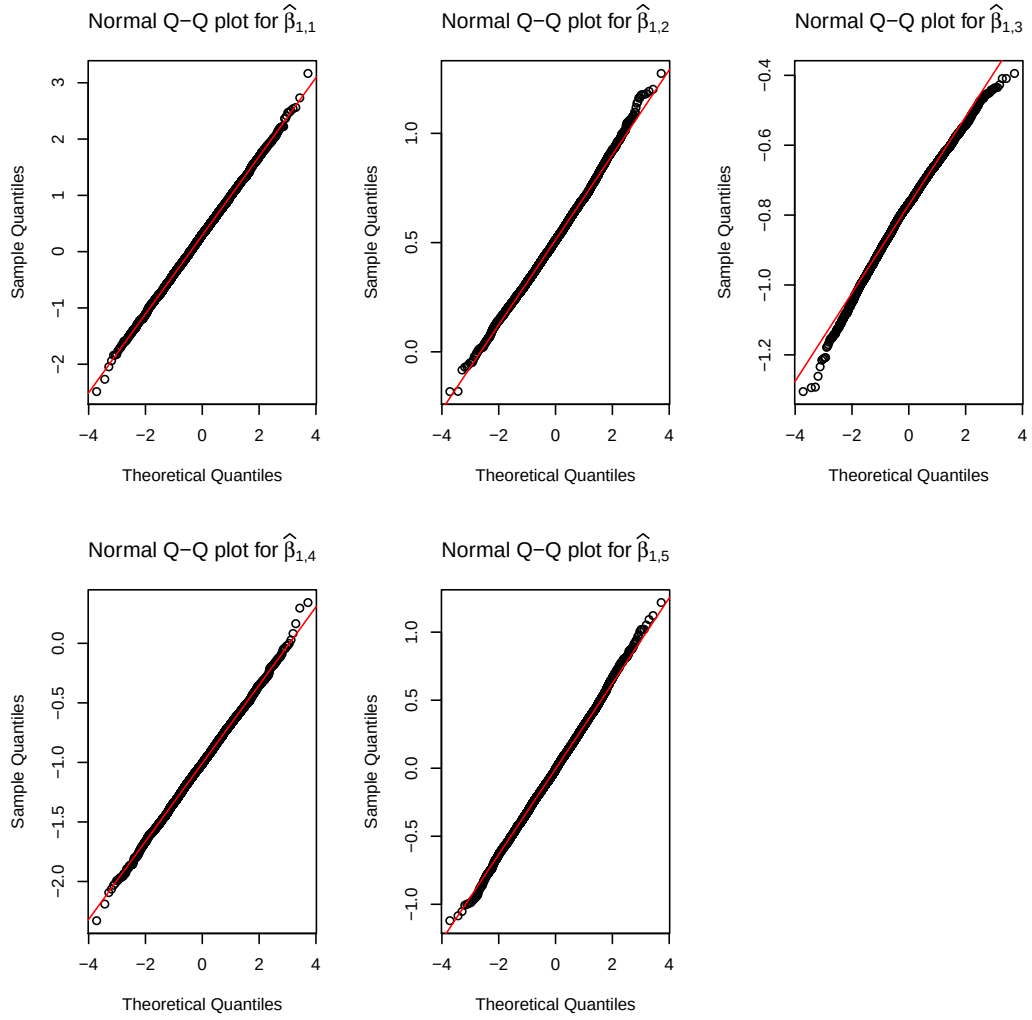

FIGURE 5.1 – Normal Q-Q plots of $\hat{\beta}_{1,1}, \ldots, \hat{\beta}_{1,5}$, for case (ii) with $n = 300$ and average sample proportion of zero-inflation equal to 0.25.

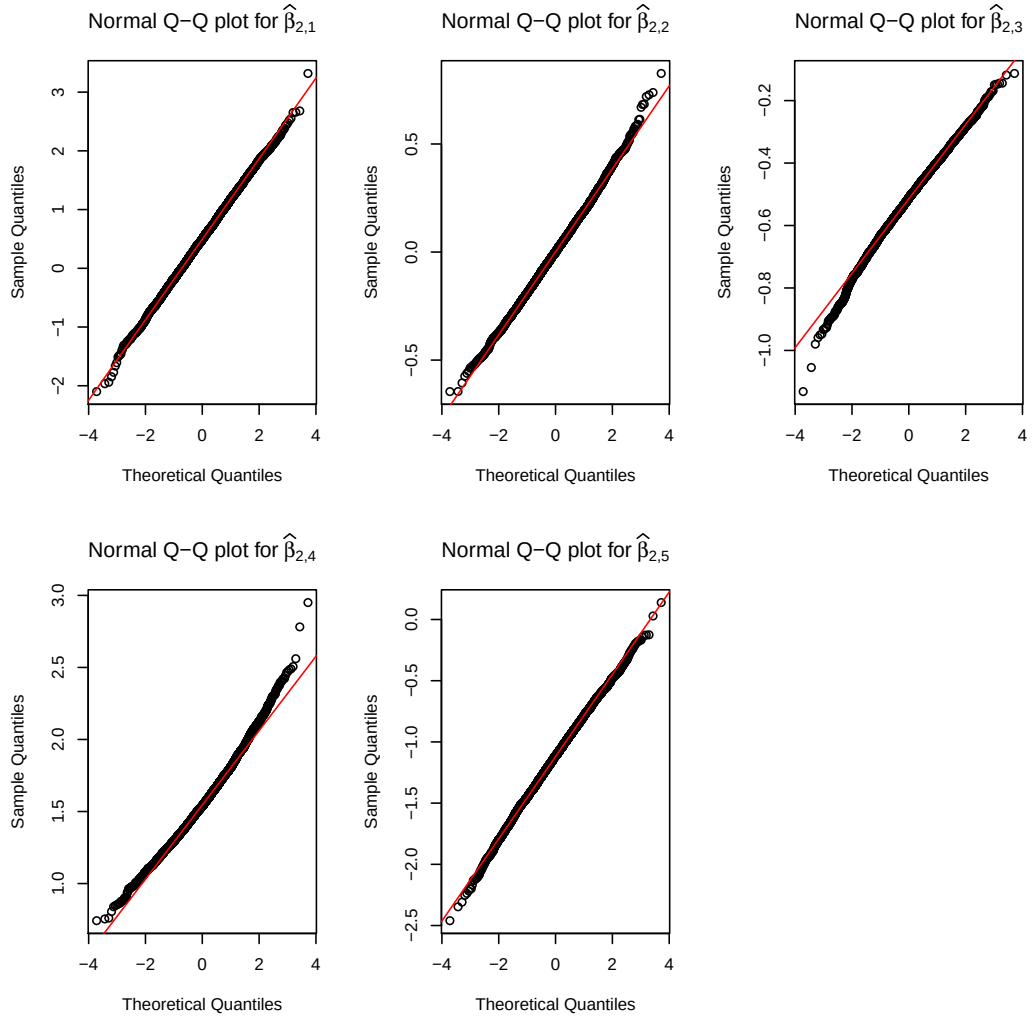

FIGURE 5.2 – Normal Q-Q plots of $\hat{\beta}_{2,1}, \ldots, \hat{\beta}_{2,5}$, for case (ii) with $n = 300$ and average sample proportion of zero-inflation equal to 0.25.

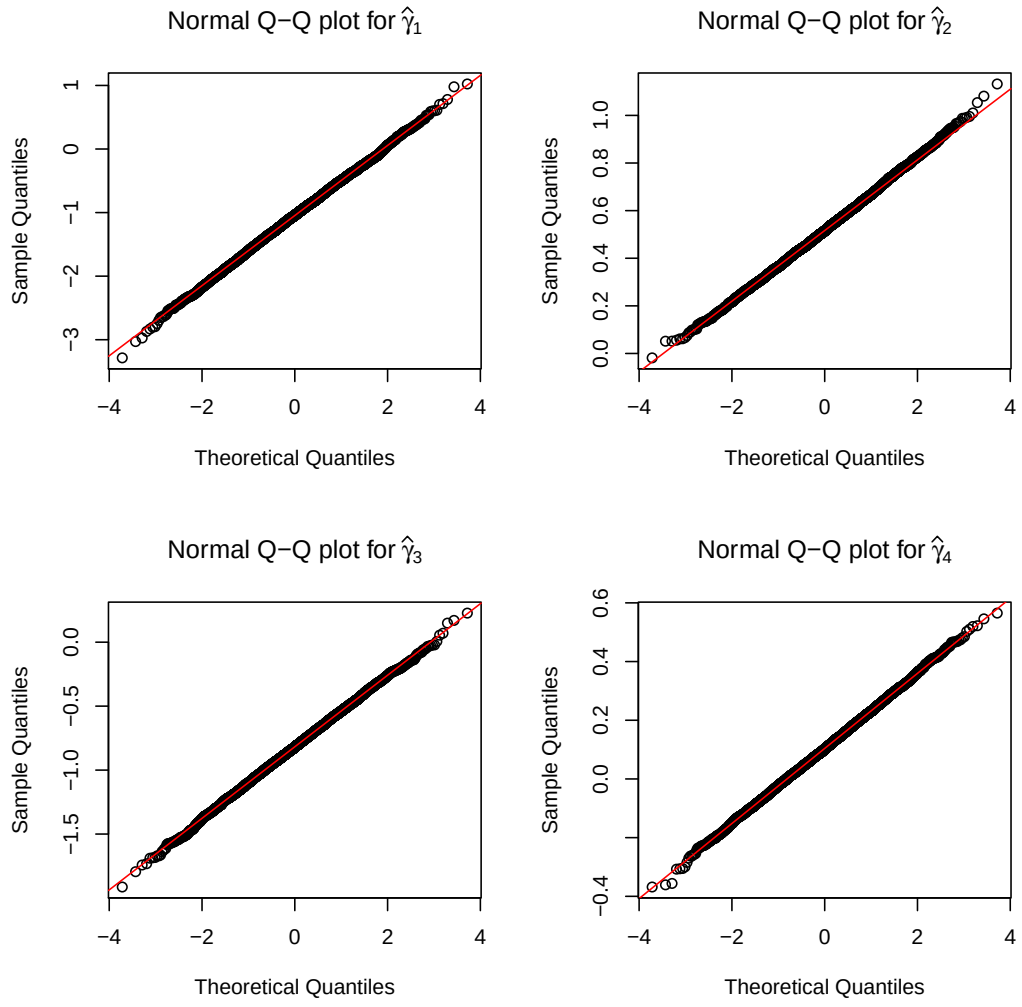

FIGURE 5.3 – Normal Q-Q plots of $\hat{\gamma}_1, \ldots, \hat{\gamma}_4$, for case (ii) with $n = 300$ and average sample proportion of zero-inflation equal to 0.25.

CHAPITRE 6

Estimation du modèle de régression binomial à inflation de zéro avec données manquantes

Sommaire

Dans le premier chapitre de la seconde partie de cette thèse nous avons établi rigoureusement les propriétés asymptotiques (consistance, normalité asymptotique, estimation convergente de la variance asymptotique) de l'estimateur du maximum de vraisemblance des paramètres du modèle de régression binomial à inflation de zéro. Une étude de simulation exhaustive suivie d'une application du modèle sur des données réelles a complété cette étude théorique.

Ce sixième chapitre s'intéresse au problème de l'inférence statistique dans le modèle binomial à inflation de zéro en présence de données manquantes sur les covariables. L'ennui des données manquantes est un problème récurrent de l'analyse statistique. Nous avons suggéré une méthode d'estimation adaptée à cette situation, en utilisant le principe de la pondération par l'inverse de la probabilité de sélection. Une investigation théorique et numérique des propriétés asymptotiques de l'estimateur proposé a été établie à nouveau.

Le chapitre présenté ci-après est soumis pour publication dans la revue *Statistics* mais des modifications ne sont pas exclues.

Authors : DIALLO A. O, DIOP A., and DUPUY J.-F.,

Estimation in zero-inflated binomial regression with missing covariates.

Abstract

The zero-inflated binomial (ZIB) regression model was recently proposed to account for excess zeros in binomial regression. Since then, the model has been applied in various domains, such as dental epidemiology and health economics. In practice, it often arises that some covariates involved in ZIB regression have missing values. Assuming that the missingness probability can be estimated parametrically, we propose an inverse-probability-weighted estimator of the parameters of a ZIB model with missing-at-random covariates. Consistency and asymptotic normality of the proposed estimator are established. The finite-sample behavior of the estimator is assessed via simulations.

keywords **: Asymptotics, count data, excess of zeros, inverse-probability-weighting**

6.1 Introduction

Count data with excess zeros arise in many disciplines, such as agriculture, economics, epidemiology, industry, insurance, terrorism study, traffic safety research...

Excess of zeros refers to the situation where the number of observed zeros is larger than predicted by standard models for count data. Zero-inflated regression models, which are obtained by mixing a degenerate distribution at zero with a standard count regression model (such as Poisson, negative binomial or binomial) have been developed to analyze such data. For example, the zero-inflated Poisson (ZIP) regression model was proposed by Lambert [1992] and further developed by Dietz and Böhning [2000], Lim et al. [2014] and Monod [2014], among many others. Recent variants of ZIP regression include random-effects ZIP models [Hall, 2000, Min and Agresti, 2005], semi-varying coefficient ZIP mo-

dels [Zhao et al., 2015] and semiparametric ZIP models [Lam et al., 2006, Feng and Zhu, 2011]. The zero-inflated negative binomial (ZINB) regression model was proposed by Ridout et al. [2001], see also Moghimbeigi et al. [2008], Mwalili et al. [2014], Garay et al. [2011]. When counts have an upper bound, ZIP and ZINB regression models are no longer appropriate. Hall [2000] and Vieira et al. [2000] thus introduced the zero-inflated binomial (ZIB) regression model, see also Diop et al. [2016]. ZIB regression was recently used in dental caries epidemiology [Gilthorpe et al., 2009, Matranga et al., 2013] and in health economics [Diallo et al., 2017].

In addition, missing data arise in a wide variety of disciplines. In the past decades, there has been an enormous literature on estimation in regression models with missing covariates, including missing covariates in linear models, generalized linear models, generalized linear mixed models, survival models... Despite this interest, only a few papers have focused on missing covariates in zero-inflated regression models. Chen and Fu [2011] develop a model selection criterion for zero-inflated regression models with missing covariates. Lukusa et al. [2016] consider estimation in ZIP regression with missing-at-random covariates. Motivated by this work, we investigate estimation in ZIB regression when some covariate values are missing for some of the sample individuals.

We propose an inverse-probability-weighted estimator of the parameters of a ZIB model with missing covariates. Inverse-probability-weighting (IPW) is a general estimation method under missing data. It was originally proposed by Horvitz and Thompson [1952] and further developed by Zhao and Lipsitz [1992]. The basic idea of IPW is to correct for missing data by giving extra weight to subjects with fully observed data. This idea has proved useful in a variety of models, such as the logistic regression model [Hsieh et al., 1952], proportional hazards regression model [Qi et al., 2005] and single-index model [Li and Hu, 2010], for example.

The rest of the paper is organized as follows. In Section 6.2, we first provide a brief review of ZIB regression, including model formulation and maximum likelihood estimation without missing data. Then we introduce a IPW estimator of the parameters of a ZIB regression model with missing-at-random covariates. In Section 6.3, we establish consistency and asymptotic normality of the proposed estimator. Section 6.4 reports results of a simulation study. A discussion and some perspectives are provided in Section 6.5.

6.2 ZIB regression with missing covariates

We first provide a brief review of ZIB regression and maximum likelihood estimation in ZIB model with complete data.

6.2.1 A brief review of ZIB regression

Let Z_i denote the random count of interest for individual i, $i = 1, \ldots, n$. Individuals are assumed to be independent. The ZIB distribution is a mixture of a degenerate distribution at zero and a binomial distribution. It is given as follows :

$$Z_i \sim \begin{cases} 0 & \text{with probability } p_i, \\ \mathscr{B}(m_i, \pi_i) & \text{with probability } 1 - p_i, \end{cases} \tag{6.1}$$

where p_i is a mixing probability for the accommodation of extra zeros and $\mathscr{B}(m, \pi)$ denotes the binomial distribution with size m and event probability π. The ZIB distribution reduces to a standard binomial distribution when $p_i = 0$.

In ZIB regression, the mixing probabilities p_i and event probabilities π_i are usually modeled via logistic regression models : $\text{logit}(p_i) = \gamma^\top \mathbf{W}_i$ and $\text{logit}(\pi_i) = \beta^\top \mathbf{X}_i$, where $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})^\top$ and $\mathbf{W}_i = (1, W_{i2}, \ldots, W_{iq})^\top$ are random vectors of predictors or covariates (both categorical and continuous covariates are allowed) and $\top$ denotes the transpose operator. Vectors $\mathbf{X}_i$ and $\mathbf{W}_i$ may either have some common components or be distinct (note that some caution is required in the special case where $m_i = 1$ for all $i = 1, \ldots, n$, see Diop et al., 2011). Here, β and γ are respectively p and q-dimensional vectors of unknown regression parameters to be estimated.

Let $\{(Z_i, \mathbf{X}_i, \mathbf{W}_i), i = 1, \ldots, n\}$ be a sample of independent observations and $\psi = (\beta^\top, \gamma^\top)^\top$ denote the whole unknown k-dimensional ($k := p + q$) parameter. The log-likelihood function $\ell\ell_n(\psi)$ based on the observed sample is :

$$\begin{aligned} \ell\ell_n(\psi) &= \sum_{i=1}^{n} \Big\{ J_i \log\left(e^{\gamma^\top \mathbf{W}_i} + (1 + e^{\beta^\top \mathbf{X}_i})^{-m_i}\right) - \log\left(1 + e^{\gamma^\top \mathbf{W}_i}\right) \\ &\qquad + (1 - J_i)\left[Z_i \beta^\top \mathbf{X}_i - m_i \log\left(1 + e^{\beta^\top \mathbf{X}_i}\right)\right]\Big\}, \\ &:= \sum_{i=1}^{n} \ell_i(\psi), \end{aligned}$$

where $J_i := 1_{\{Z_i=0\}}$ (see Hall, 2000). The maximum likelihood estimator (MLE) $\widehat{\psi}_n := (\widehat{\beta}_n^\top, \widehat{\gamma}_n^\top)^\top$ of ψ is obtained by solving the score equation $U_n(\psi) = 0$, where

$$U_n(\psi) = \frac{1}{\sqrt{n}} \frac{\partial \ell\ell_n(\psi)}{\partial \psi} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \ell_i(\psi)}{\partial \psi} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_i(\psi). \tag{6.2}$$

This estimating equation can be solved by using the expectation-maximization (EM) algorithm [Hall, 2000] or by a direct maximization of $\ell\ell_n(\psi)$ [Diallo et al., 2017]. The MLE $\widehat{\psi}_n$ is a consistent and asymptotically normal estimator of the true ψ, see Diallo et al. [2017].

The next section describes the problem and the proposed estimator.

6.2.2 ZIB regression with missing covariates : the proposed estimator

In this work, we assume that some components of $\mathbf{X}_i$ may be missing for some individuals. Decompose $\mathbf{X}_i$ as $\mathbf{X}_i = (\mathbf{X}_i^{(obs),\top}, \mathbf{X}_i^{(miss),\top})^\top$, where $\mathbf{X}_i^{(obs)}$ and $\mathbf{X}_i^{(miss)}$ contain the observed and missing components of $\mathbf{X}_i$ respectively (we assume that the same components of $\mathbf{X}_i$ may be missing for all individuals). Let δ_i be a dummy variable indicating whether $\mathbf{X}_i$ is fully observed ($\delta_i = 1$) or not ($\delta_i = 0$). Finally, let $\mathbf{S}_i := (Z_i, \mathbf{X}_i^{(obs),\top}, \mathbf{W}_i^\top)^\top$ denote the vector of variables that are always observed on each individual. Then $\{Z_i, \mathbf{X}_i, \mathbf{W}_i\} = \{\mathbf{S}_i, \mathbf{X}_i^{(miss)}\}$. Let d denote the dimension of $\mathbf{S}_i$.

We assume that $\mathbf{X}_i^{(miss)}$ is missing at random (MAR, see Rubin, 1976) : the probability that some components of $\mathbf{X}_i$ are missing depends only on the observed variables. The MAR assumption can be expressed in terms of the missingness (or selection) probability $\mathbb{P}(\delta_i = 1|\mathbf{S}_i, \mathbf{X}_i^{(miss)})$, as :

$$\mathbb{P}(\delta_i = 1|\mathbf{S}_i, \mathbf{X}_i^{(miss)}) = \mathbb{P}(\delta_i = 1|\mathbf{S}_i).$$

Under missing data, we propose to estimate ψ in ZIB model (6.1) by using the IPW method. Originally proposed by Horvitz and Thompson [1952], IPW has recently been used for estimating various regression models with missing or mismeasured covariates. Basic idea is to inversely weight the observed data by the selection probability $\mathbb{P}(\delta_i = 1|\mathbf{S}_i)$, so as to reduce the bias due to incomplete cases deletion. Recall that without missing data,

ψ in model (6.1) can be estimated by solving the score equation (6.2). Under missing data, we propose to estimate ψ by solving the following estimating equation, derived from (6.2) by weighting individuals with fully observed data by the inverse of their selection probability :

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{\mathbb{P}(\delta_i=1|\mathbf{S}_i)}\dot{\ell}_i(\psi)=0. \tag{6.3}$$

In practice, however, selection probabilities $\mathbb{P}(\delta_i=1|\mathbf{S}_i)$ are usually unknown and need to be estimated. Several estimation procedures can be used. In this work, we consider the case where the unknown $\mathbb{P}(\delta_i=1|\mathbf{S}_i)$, $i=1,\ldots,n$, can be estimated parametrically. In this case, logistic regression is the most frequently used option. Let $r_i(\alpha):=\mathbb{P}(\delta_i=1|\mathbf{S}_i)$ be defined as :

$$r_i(\alpha)=\frac{\exp(\alpha^\top\mathbf{S}_i)}{1+\exp(\alpha^\top\mathbf{S}_i)}, \tag{6.4}$$

where α is a d-dimensional vector of unknown regression parameters. We need to estimate α before solving the weighted score equation (6.3). Maximum likelihood estimation can be used for that purpose. The MLE $\widehat{\alpha}_n=\arg\max_\alpha\prod_{i=1}^n\{r_i(\alpha)^{\delta_i}(1-r_i(\alpha))^{1-\delta_i}\}$ in model (6.4) is known to be consistent and asymptotically Gaussian (Gouriéroux and Monfort, 1981). Once $\widehat{\alpha}_n$ is available, one can estimate ψ by solving the estimated weighted score equation $U_{w,n}(\psi,\widehat{\alpha}_n)=0$, where

$$U_{w,n}(\psi,\widehat{\alpha}_n)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{r_i(\widehat{\alpha}_n)}\dot{\ell}_i(\psi)=0.$$

In what follows, the resulting estimator of ψ will be denoted by $\widehat{\psi}_n$. Asymptotic properties of this estimator are established in Section 6.3. First, we need to introduce some further notations.

6.2.3 Some further notations

Define first the $(k\times d)$, $(k\times k)$ and $(d\times d)$ matrices

$$B(\psi,\alpha)=\lim_{n\to\infty}\left(-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\delta_i\frac{1-r_i(\alpha)}{r_i(\alpha)}\dot{\ell}_i(\psi)\mathbf{S}_i^\top\right]\right),$$

$$J(\psi,\alpha)=\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\dot{\ell}_i(\psi)\dot{\ell}_i(\psi)^\top}{r_i(\alpha)}\right]\right),$$

and

$$\Sigma(\alpha) = \mathbb{E}\left[\mathbf{S}_i\mathbf{S}_i^\top r_i(\alpha)(1-r_i(\alpha))\right]. \tag{6.5}$$

For every $i = 1, \ldots, n$, let $U_i = (U_{i1}, \ldots, U_{ik})^\top$ denote the k-dimensional column vector $U_i := B(\psi_0, \alpha_0)\Sigma(\alpha_0)^{-1}\mathbf{S}_i$. Then, define the $(p \times n)$, $(q \times n)$ and $(k \times n)$ matrices

$$\mathbb{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1p} & X_{2p} & \cdots & X_{np} \end{pmatrix}, \qquad \mathbb{W} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ W_{12} & W_{22} & \cdots & W_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ W_{1q} & W_{2q} & \cdots & W_{nq} \end{pmatrix},$$

and

$$\mathbb{U} = \begin{pmatrix} U_{11} & U_{21} & \cdots & U_{n1} \\ U_{12} & U_{22} & \cdots & U_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ U_{1k} & U_{2k} & \cdots & U_{nk} \end{pmatrix} = \begin{bmatrix} \mathbb{U}_1 \\ \mathbb{U}_2 \end{bmatrix},$$

where $\mathbb{U}_1$ is the $(p \times n)$ sub-matrix of $\mathbb{U}$ consisting of the first p rows of $\mathbb{U}$ and $\mathbb{U}_2$ is the $(q \times n)$ sub-matrix of $\mathbb{U}$ consisting of the last q rows of $\mathbb{U}$. Let $\mathbb{V}$ be the $(k \times 3n)$ block-matrix defined as

$$\mathbb{V} = \begin{bmatrix} \mathbb{X} & 0_{p,n} & \mathbb{U}_1 \\ 0_{q,n} & \mathbb{W} & \mathbb{U}_2 \end{bmatrix},$$

and $C(\psi, \alpha) = (C_j(\psi, \alpha))_{1 \le j \le 3n}$ be the $3n$-dimensional column vector defined by

$$C(\psi, \alpha) = \left(\frac{\delta_1}{r_1(\alpha)}A_1(\psi), \ldots, \frac{\delta_n}{r_n(\alpha)}A_n(\psi), \frac{\delta_1}{r_1(\alpha)}B_1(\psi), \ldots, \frac{\delta_n}{r_n(\alpha)}B_n(\psi), \delta_1 - r_1(\alpha), \ldots, \delta_n - r_n(\alpha)\right)^\top,$$

where $0_{a,b}$ denotes the $(a \times b)$ matrix whose components are all equal to zero and for every $i = 1, \ldots, n$,

$$A_i(\psi) = -J_i \frac{m_i e^{\beta^\top \mathbf{x}_i}}{e^{\gamma^\top \mathbf{w}_i}(h_i(\beta))^{m_i+1} + h_i(\beta)} + (1 - J_i)\left(Z_i - \frac{m_i e^{\beta^\top \mathbf{x}_i}}{h_i(\beta)}\right) \tag{6.6}$$

and

$$B_i(\psi) = \frac{J_i e^{\gamma^\top \mathbf{w}_i}(h_i(\beta))^{m_i}}{e^{\gamma^\top \mathbf{w}_i}(h_i(\beta))^{m_i} + 1} - \frac{e^{\gamma^\top \mathbf{w}_i}}{1 + e^{\gamma^\top \mathbf{w}_i}}, \tag{6.7}$$

with $h_i(\beta) := 1 + e^{\beta^\top \mathbf{x}_i}$.

If $A = (A_{ij})_{1\leq i\leq a, 1\leq j\leq b}$ is a $(a \times b)$ matrix, $A_{\bullet j}$ will denote its j-th column $(j = 1, \ldots, b)$ that is, $A_{\bullet j} = (A_{1j}, \ldots, A_{aj})^\top$.

Finally, under model (6.4), it is known that the MLE $\widehat{\alpha}_n$ verifies

$$\sqrt{n}(\widehat{\alpha}_n - \alpha_0) = \Sigma(\alpha_0)^{-1} M_n(\alpha_0) + o_{\mathbb{P}}(1), \tag{6.8}$$

where $\Sigma(\alpha_0)$ is given by (6.5) and $M_n(\alpha) = n^{-1/2} \sum_{i=1}^n \mathbf{S}_i(\delta_i - r_i(\alpha))$.

In the next section, we establish rigorously the consistency and asymptotic normality of the proposed IPW-MLE $\widehat{\psi}_n$.

6.3 Asymptotic results

We first state some regularity conditions that will be needed for proving our asymptotic results.

6.3.1 Regularity conditions and consistency

C1 Covariates are bounded, that is, there exists a finite positive constant c_0 such that $|X_{ij}| \leq c_0$ and $|W_{i\ell}| \leq c_0$ for every $i = 1, \ldots, n$, $j = 2, \ldots, p$ and $\ell = 2, \ldots, q$. For every $i = 1, \ldots, n$, $j = 2, \ldots, p$ and $\ell = 2, \ldots, q$, var$[X_{ij}] > 0$ and var$[W_{i\ell}] > 0$. For every $i = 1, \ldots, n$, the X_{ij} $(j = 1, \ldots, p)$ are linearly independent and the $W_{i\ell}$ $(\ell = 1, \ldots, q)$ are linearly independent.

C2 The true parameter value $\psi_0 := (\beta_0^\top, \gamma_0^\top)^\top$ lies in the interior of some known compact set of $\mathbb{R}^p \times \mathbb{R}^q$. The true α_0 belongs to the interior of some known compact set of $\mathbb{R}^d$.

C3 As $n \to \infty$, $n^{-1} \sum_{i=1}^n \mathbb{E}\left[\frac{\delta_i}{r_i(\alpha_0)} \frac{\partial^2 \ell_i(\psi)}{\partial \psi \partial \psi^\top}\right]$ converges to some invertible matrix $A(\psi, \alpha_0)$ and the smallest eigenvalue λ_n of $\mathbb{V}\mathbb{V}^\top$ tends to $+\infty$.

C4 For every $i = 1, \ldots, n$, we have $m_i \in \{2, \ldots, M\}$ for some finite integer value M.

In what follows, the space $\mathbb{R}^k$ of k-dimensional (column) vectors will be provided with the Euclidean norm $\|\cdot\|_2$ and the space of $(k \times k)$ real matrices will be provided with the norm $|||A|||_2 := \max_{\|x\|_2=1} \|Ax\|_2$ (for notations simplicity, we will use $\|\cdot\|$ for both norms).

We first prove consistency of $\widehat{\psi}_n$:

Theorem 6.3.1 *Assume that conditions C1-C4 hold. Then, as $n \to \infty$, $\widehat{\psi}_n$ converges in probability to ψ_0.*

Proof of Theorem 6.3.1. To prove consistency of $\widehat{\psi}_n$, we verify the conditions of the inverse function theorem of Foutz [1977]. These conditions are proved in a series of technical lemmas.

Lemma 6.3.2 *$\partial U_{w,n}(\psi, \widehat{\alpha}_n)/\partial \psi^\top$ exists and is continuous in an open neighborhood of ψ_0.*

Proof of Lemma 6.3.2. The $\ell_i(\psi), i = 1, \ldots, n$ are twice differentiable with respect to ψ. Continuity of $\frac{\partial^2 \ell_i(\psi)}{\partial \psi \partial \psi^\top}$ is straightforward and is omitted. □

Lemma 6.3.3 *As $n \to \infty$, $n^{-1/2} U_{w,n}(\psi_0, \widehat{\alpha}_n)$ converges in probability to 0.*

Proof of Lemma 6.3.3. Decompose $n^{-1/2} U_{w,n}(\psi_0, \widehat{\alpha}_n)$ as :

$$n^{-1/2} U_{w,n}(\psi_0, \widehat{\alpha}_n) = \left(n^{-1/2} U_{w,n}(\psi_0, \widehat{\alpha}_n) - n^{-1/2} U_{w,n}(\psi_0, \alpha_0) \right) + n^{-1/2} U_{w,n}(\psi_0, \alpha_0), \qquad (6.9)$$

and consider the first term on the right-hand side of this decomposition. We have :

$$\begin{aligned} n^{-1/2} U_{w,n}(\psi_0, \widehat{\alpha}_n) - n^{-1/2} U_{w,n}(\psi_0, \alpha_0) &= \frac{1}{n} \sum_{i=1}^n \delta_i \left(\frac{1}{r_i(\widehat{\alpha}_n)} - \frac{1}{r_i(\alpha_0)} \right) \dot{\ell}_i(\psi_0), \\ &= \frac{1}{n} \sum_{i=1}^n \delta_i \left(\frac{1}{e^{\widehat{\alpha}_n^\top \mathbf{S}_i}} - \frac{1}{e^{\alpha_0^\top \mathbf{S}_i}} \right) \dot{\ell}_i(\psi_0). \end{aligned}$$

By a Taylor expansion of $1/e^{\widehat{\alpha}_n^\top \mathbf{S}_i}$ around α_0,

$$n^{-1/2} U_{w,n}(\psi_0, \widehat{\alpha}_n) - n^{-1/2} U_{w,n}(\psi_0, \alpha_0) = \frac{1}{n} \sum_{i=1}^n \delta_i (\alpha_0 - \widehat{\alpha}_n)^\top \mathbf{S}_i \frac{1}{e^{\alpha_*^\top \mathbf{S}_i}} \dot{\ell}_i(\psi_0),$$

where α_* is on the line segment between $\widehat{\alpha}_n$ and α_0. Then, we have :

$$\begin{aligned} \| n^{-1/2} U_{w,n}(\psi_0, \widehat{\alpha}_n) - n^{-1/2} U_{w,n}(\psi_0, \alpha_0) \| &\leq \frac{1}{n} \sum_{i=1}^n \left\| \delta_i (\alpha_0 - \widehat{\alpha}_n)^\top \mathbf{S}_i \frac{1}{e^{\alpha_*^\top \mathbf{S}_i}} \dot{\ell}_i(\psi_0) \right\|, \\ &\leq \frac{1}{n} \sum_{i=1}^n \left| \delta_i (\alpha_0 - \widehat{\alpha}_n)^\top \mathbf{S}_i \frac{1}{e^{\alpha_*^\top \mathbf{S}_i}} \right| \left\| \dot{\ell}_i(\psi_0) \right\|, \\ &\leq \frac{1}{n} \sum_{i=1}^n \| \alpha_0 - \widehat{\alpha}_n \| \, \| \mathbf{S}_i \| \frac{1}{e^{\alpha_*^\top \mathbf{S}_i}} \left\| \dot{\ell}_i(\psi_0) \right\|, \end{aligned}$$

where the second to third line comes from Cauchy-Schwarz inequality. Now, straightforward calculations show that

$$\dot{\ell}_i(\psi_0) = (\mathbf{X}_i^\top, 0_q^\top)^\top \cdot A_i(\psi_0) + (0_p^\top, \mathbf{W}_i^\top)^\top \cdot B_i(\psi_0),$$

where $0_p := 0_{p,1}$ is the p-dimensional column vector having all its components equal to 0 and $A_i(\psi_0), B_i(\psi_0)$ are given by (6.6) and (6.7) respectively. Thus we have :

$$\| \dot{\ell}_i(\psi_0) \| \leq \| (\mathbf{X}_i^\top, 0_q^\top)^\top \| \cdot |A_i(\psi_0)| + \| (0_p^\top, \mathbf{W}_i^\top)^\top \| \cdot |B_i(\psi_0)|.$$

Under conditions C1, C2 and C4, it is easy to see that $|A_i(\psi_0)|$ and $|B_i(\psi_0)|$ are bounded above. Thus, there exists a finite constant c_2 such that $n^{-1}\sum_{i=1}^{n}\|\dot{\ell}_i(\psi_0)\| \leq c_2$. Note that $\|\mathbf{S}_i\|$ and $1/e^{\alpha_*^\top \mathbf{S}_i}$ are also bounded, by conditions C1 and C2. Therefore, there exists some finite constant c_3 such that $\|n^{-1/2}U_{w,n}(\psi_0,\widehat{\alpha}_n) - n^{-1/2}U_{w,n}(\psi_0,\alpha_0)\| \leq c_3\|\alpha_0 - \widehat{\alpha}_n\|$. Finally, the convergence of $\widehat{\alpha}_n$ to α_0 imply that $\|n^{-1/2}U_{w,n}(\psi_0,\widehat{\alpha}_n) - n^{-1/2}U_{w,n}(\psi_0,\alpha_0)\|$ converges in probability to 0 as $n \to \infty$. Next, consider the term $n^{-1/2}U_{w,n}(\psi_0,\alpha_0)$ in decomposition (6.9). Some simple algebra yields :

$$n^{-1/2}U_{w,n}(\psi_0,\alpha_0) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{r_i(\alpha_0)}X_{i1}A_i(\psi_0) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{r_i(\alpha_0)}X_{ip}A_i(\psi_0) \\ \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{r_i(\alpha_0)}W_{i1}B_i(\psi_0) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{r_i(\alpha_0)}W_{iq}B_i(\psi_0) \end{pmatrix}.$$

We prove that $n^{-1/2}U_{w,n}(\psi_0,\alpha_0)$ converges in probability to 0 as $n \to \infty$. To see this, note first that for every $i = 1,\ldots,n$ and $\ell = 1,\ldots,q$:

$$\begin{aligned} \mathbb{E}\left[\frac{\delta_i}{r_i(\alpha_0)}W_{i\ell}B_i(\psi_0)\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{\delta_i}{r_i(\alpha_0)}W_{i\ell}B_i(\psi_0)\middle|\,\mathbf{S}_i\right]\right], \\ &= \mathbb{E}\left[\frac{1}{r_i(\alpha_0)}W_{i\ell}\mathbb{E}\left[\,\delta_i B_i(\psi_0)|\,\mathbf{S}_i\right]\right]. \end{aligned}$$

Given $\mathbf{S}_i$, $B_i(\psi_0)$ is a function of $\mathbf{X}_i^{(miss)}$ only. Thus, by the MAR assumption, $B_i(\psi_0)$ and δ_i are independent given $\mathbf{S}_i$. It follows that :

$$\begin{aligned} \mathbb{E}\left[\frac{\delta_i}{r_i(\alpha_0)}W_{i\ell}B_i(\psi_0)\right] &= \mathbb{E}\left[\frac{1}{r_i(\alpha_0)}W_{i\ell}\mathbb{E}\left[\,\delta_i|\,\mathbf{S}_i\right]\mathbb{E}\left[\,B_i(\psi_0)|\,\mathbf{S}_i\right]\right], \\ &= \mathbb{E}\left[W_{i\ell}\mathbb{E}\left[\,B_i(\psi_0)|\,\mathbf{S}_i\right]\right], \\ &= \mathbb{E}\left[W_{i\ell}B_i(\psi_0)\right]. \end{aligned}$$

Diallo et al. [2017] proved that $\mathbb{E}\left[W_{i\ell}B_i(\psi_0)\right] = 0$ for every $i = 1,\ldots,n$ and $\ell = 1,\ldots,q$. Thus, $\mathbb{E}\left[\frac{\delta_i}{r_i(\alpha_0)}W_{i\ell}B_i(\psi_0)\right] = 0$ for every $i = 1,\ldots,n$ and $\ell = 1,\ldots,q$. Similarly, for every $i = 1,\ldots,n$ and $j = 1,\ldots,p$, we have :

$$\mathbb{E}\left[\frac{\delta_i}{r_i(\alpha_0)}X_{ij}A_i(\psi_0)\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\delta_i}{r_i(\alpha_0)}X_{ij}A_i(\psi_0)\middle|\,\mathbf{S}_i\right]\right].$$

Two cases should be considered, namely : $i)$ X_{ij} is a component of $\mathbf{X}_i^{(miss)}$ and $ii)$ X_{ij} is a component of $\mathbf{X}_i^{(obs)}$. In case $i)$, we have :

$$\mathbb{E}\left[\mathbb{E}\left[\frac{\delta_i}{r_i(\alpha_0)}X_{ij}A_i(\psi_0)\middle|\,\mathbf{S}_i\right]\right] = \mathbb{E}\left[\frac{1}{r_i(\alpha_0)}\mathbb{E}\left[\,\delta_i X_{ij}A_i(\psi_0)|\,\mathbf{S}_i\right]\right].$$

Given $\mathbf{S}_i$, $X_{ij}A_i(\psi_0)$ is a function of $\mathbf{X}_i^{(miss)}$ only. Thus, by the MAR assumption,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{r_i(\alpha_0)}\mathbb{E}\left[\delta_i X_{ij}A_i(\psi_0)\middle|\mathbf{S}_i\right]\right] &= \mathbb{E}\left[\frac{1}{r_i(\alpha_0)}\mathbb{E}\left[\delta_i\middle|\mathbf{S}_i\right]\mathbb{E}\left[X_{ij}A_i(\psi_0)\middle|\mathbf{S}_i\right]\right]\\ &= \mathbb{E}\left[X_{ij}A_i(\psi_0)\right].\end{aligned}$$

Diallo et al. [2017] proved that $\mathbb{E}\left[X_{ij}A_i(\psi_0)\right]=0$ for every $i=1,\ldots,n$ and $j=1,\ldots,p$. Therefore, $\mathbb{E}\left[\frac{\delta_i}{r_i(\alpha_0)}X_{ij}A_i(\psi_0)\right]=0$. In case ii),

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[\frac{\delta_i}{r_i(\alpha_0)}X_{ij}A_i(\psi_0)\middle|\mathbf{S}_i\right]\right] &= \mathbb{E}\left[\frac{1}{r_i(\alpha_0)}X_{ij}\mathbb{E}\left[\delta_i A_i(\psi_0)\middle|\mathbf{S}_i\right]\right],\\ &= \mathbb{E}\left[\frac{1}{r_i(\alpha_0)}X_{ij}\mathbb{E}\left[\delta_i\middle|\mathbf{S}_i\right]\mathbb{E}\left[A_i(\psi_0)\middle|\mathbf{S}_i\right]\right],\\ &= \mathbb{E}\left[X_{ij}A_i(\psi_0)\right],\\ &= 0,\end{aligned}$$

where the first to second line comes from the fact that under MAR, $A_i(\psi_0)$ and δ_i are independent given $\mathbf{S}_i$. Finally, in case ii), we also have : $\mathbb{E}\left[\frac{\delta_i}{r_i(\alpha_0)}X_{ij}A_i(\psi_0)\right]=0$.
Now, for every $i=1,\ldots,n$ and $\ell=1,\ldots,q$, we have :

$$\text{var}\left(\frac{\delta_i}{r_i(\alpha_0)}W_{i\ell}B_i(\psi_0)\right)\leq\mathbb{E}\left[\frac{\delta_i}{r_i^2(\alpha_0)}W_{i\ell}^2B_i^2(\psi_0)\right].$$

By C1, C2, C4, there exists finite constants c_4 and c_5 such that $1/r_i^2(\alpha_0)\leq c_4$ and $B_i^2(\psi_0)\leq c_5$ for every $i=1,\ldots,n$. Therefore,

$$\text{var}\left(\frac{\delta_i}{r_i(\alpha_0)}W_{i\ell}B_i(\psi_0)\right)\leq c_6:=c_0^2c_4c_5.$$

It follows that

$$\sum_{i=1}^{\infty}\frac{\text{var}\left(\frac{\delta_i}{r_i(\alpha_0)}W_{i\ell}B_i(\psi_0)\right)}{i^2}\leq c_6\sum_{i=1}^{\infty}\frac{1}{i^2}<\infty.$$

One can easily show that a similar result holds for $\text{var}\left(\frac{\delta_i}{r_i(\alpha_0)}X_{ij}A_i(\psi_0)\right)$. By Kolmogorov's law of large numbers (see for example Jiang [2010], Theorem 6.7),

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\delta_i}{r_i(\alpha_0)}W_{i\ell}B_i(\psi_0)-\mathbb{E}\left[\frac{\delta_i}{r_i(\alpha_0)}W_{i\ell}B_i(\psi_0)\right]\right\}=\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{r_i(\alpha_0)}W_{i\ell}B_i(\psi_0),\quad\ell=1,\ldots,q$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\delta_i}{r_i(\alpha_0)}X_{ij}A_i(\psi_0)-\mathbb{E}\left[\frac{\delta_i}{r_i(\alpha_0)}X_{ij}A_i(\psi_0)\right]\right\}=\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{r_i(\alpha_0)}X_{ij}A_i(\psi_0),\quad j=1,\ldots,p$$

converge in probability to 0 as $n\to\infty$ and thus, $n^{-1/2}U_{w,n}(\psi_0,\alpha_0)$ converges in probability to 0. This implies that $n^{-1/2}U_{w,n}(\psi_0,\widehat{\alpha}_n)$ converges to 0, which concludes the proof. □

Lemma 6.3.4 *As $n \to \infty$, $n^{-1/2}\partial U_{w,n}(\psi, \widehat{\alpha}_n)/\partial\psi^\top$ converges in probability to a fixed function $A(\psi, \alpha_0)$, uniformly in an open neighborhood of ψ_0.*

Proof of Lemma 6.3.4. Let $\widetilde{U}_{w,n}(\psi, \alpha) := n^{-1/2}\partial U_{w,n}(\psi, \alpha)/\partial\psi^\top$ and $\mathscr{V}_{\psi_0}$ be an open neighborhood of ψ_0. Let $\psi \in \mathscr{V}_{\psi_0}$. Using similar arguments as in proof of Lemma 6.3.3, we have :

$$\|\widetilde{U}_{w,n}(\psi, \widehat{\alpha}_n) - \widetilde{U}_{w,n}(\psi, \alpha_0)\| \leq \frac{1}{n}\sum_{i=1}^{n} \|\alpha_0 - \widehat{\alpha}_n\| \, \|\mathbf{S}_i\| \, \frac{1}{e^{\alpha_*^\top \mathbf{S}_i}} \left\|\ddot{\ell}_i(\psi)\right\|,$$

where α_* is on the line segment between $\widehat{\alpha}_n$ and α_0. From this, one easily proves that $\|\widetilde{U}_{w,n}(\psi, \widehat{\alpha}_n) - \widetilde{U}_{w,n}(\psi, \alpha_0)\|$ converges in probability to 0 as $n \to \infty$. Details are omitted. Now, consider the (ℓ, j)-th element of $\widetilde{U}_{w,n}(\psi, \alpha_0)$, namely :

$$\left(\widetilde{U}_{w,n}(\psi, \alpha_0)\right)_{(\ell,j)} = \frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{\delta_i}{r_i(\alpha_0)} \frac{\partial^2 \ell_i(\psi)}{\partial\psi_\ell \partial\psi_j} \right\}.$$

We have :

$$\left(\widetilde{U}_{w,n}(\psi, \alpha_0)\right)_{(\ell,j)} = \frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{\delta_i}{r_i(\alpha_0)} \frac{\partial^2 \ell_i(\psi)}{\partial\psi_\ell \partial\psi_j} - \mathbb{E}\left[\frac{\delta_i}{r_i(\alpha_0)} \frac{\partial^2 \ell_i(\psi)}{\partial\psi_\ell \partial\psi_j} \right] \right\} + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\frac{\delta_i}{r_i(\alpha_0)} \frac{\partial^2 \ell_i(\psi)}{\partial\psi_\ell \partial\psi_j} \right].$$

Now,

$$\begin{aligned} \text{var}\left(\frac{\delta_i}{r_i(\alpha_0)} \frac{\partial^2 \ell_i(\psi)}{\partial\psi_\ell \partial\psi_j} \right) &\leq \mathbb{E}\left(\frac{\delta_i}{r_i^2(\alpha_0)} \left\{ \frac{\partial^2 \ell_i(\psi)}{\partial\psi_\ell \partial\psi_j} \right\}^2 \right) \\ &\leq c_4 \mathbb{E}\left(\left\{ \frac{\partial^2 \ell_i(\psi)}{\partial\psi_\ell \partial\psi_j} \right\}^2 \right). \end{aligned}$$

We prove that $\text{var}\left(\frac{\delta_i}{r_i(\alpha_0)} \frac{\partial^2 \ell_i(\psi)}{\partial\psi_\ell \partial\psi_j} \right)$ is bounded. Some tedious albeit easy calculations show that $\frac{\partial^2 \ell_i(\psi)}{\partial\psi_\ell \partial\psi_j}$ is the (ℓ, j)-th element of the $(k \times k)$ matrix $(-\mathbf{V}_i \mathbf{U}_i(\psi) \mathbf{V}_i^\top)$, where $\mathbf{V}_i$ is the $(k \times 2)$ matrix defined as

$$\mathbf{V}_i = \begin{pmatrix} \mathbf{X}_i & 0_p \\ 0_q & \mathbf{W}_i \end{pmatrix}$$

and

$$\mathbf{U}_i(\psi) = \begin{pmatrix} \mathbf{U}_{i,1}(\psi) & \mathbf{U}_{i,2}(\psi) \\ \mathbf{U}_{i,2}(\psi) & \mathbf{U}_{i,3}(\psi) \end{pmatrix}$$

is the (2×2) symmetric matrix defined by

$$\begin{aligned} \mathbf{U}_{i,1}(\psi) &= \frac{J_i m_i e^{\beta^\top \mathbf{X}_i}}{(k_i(\psi))^2} \left(k_i(\psi) - e^{\beta^\top \mathbf{X}_i} \left[e^{\gamma^\top \mathbf{W}_i} (m_i + 1)(h_i(\beta))^{m_i} + 1 \right] \right) + \frac{m_i (1 - J_i) e^{\beta^\top \mathbf{X}_i}}{(h_i(\beta))^2}, \\ \mathbf{U}_{i,2}(\psi) &= -\frac{J_i m_i e^{\beta^\top \mathbf{X}_i + \gamma^\top \mathbf{W}_i} (h_i(\beta))^{m_i+1}}{(k_i(\psi))^2}, \\ \mathbf{U}_{i,3}(\psi) &= \frac{J_i e^{\gamma^\top \mathbf{W}_i} (h_i(\beta))^{m_i+1}}{(k_i(\psi))^2} \left(e^{\gamma^\top \mathbf{W}_i} (h_i(\beta))^{m_i+1} - k_i(\psi) \right) + \frac{e^{\gamma^\top \mathbf{W}_i}}{\left(1 + e^{\gamma^\top \mathbf{W}_i}\right)^2}, \end{aligned}$$

with $k_i(\psi) := e^{\gamma^\top \mathbf{W}_i}(h_i(\beta))^{m_i+1} + h_i(\beta)$, $i = 1, \ldots, n$. Using these notations, it is easy to see that

$$\begin{aligned}\frac{\partial^2 \ell_i(\psi)}{\partial\psi_\ell \partial\psi_j} &= -\left(\mathbf{V}_{i,(\ell,1)}\mathbf{U}_{i,1}(\psi) + \mathbf{V}_{i,(\ell,2)}\mathbf{U}_{i,2}(\psi)\right)\mathbf{V}_{i,(j,1)} \\ &\quad -\left(\mathbf{V}_{i,(\ell,1)}\mathbf{U}_{i,2}(\psi) + \mathbf{V}_{i,(\ell,2)}\mathbf{U}_{i,3}(\psi)\right)\mathbf{V}_{i,(j,2)}, \qquad (6.10)\end{aligned}$$

where $\mathbf{V}_{i,(a,b)}$ denotes the (a,b)-th element of matrix $\mathbf{V}_i$. For a given row ℓ ($\ell = 1, \ldots, k$), exactly one of $\mathbf{V}_{i,(\ell,1)}$ and $\mathbf{V}_{i,(\ell,2)}$ must be equal to 0 (this is straightforward from the expression of $\mathbf{V}_i$). Suppose for example that $\mathbf{V}_{i,(\ell,1)} = 0$ and $\mathbf{V}_{i,(j,2)} = 0$ (other combinations of null and non-null values among $(\mathbf{V}_{i,(\ell,1)}, \mathbf{V}_{i,(\ell,2)})$ and $(\mathbf{V}_{i,(j,1)}, \mathbf{V}_{i,(j,2)})$ can be treated similarly). Then (6.10) reduces to :

$$\frac{\partial^2 \ell_i(\psi)}{\partial\psi_\ell \partial\psi_j} = -\mathbf{V}_{i,(\ell,2)}\mathbf{U}_{i,2}(\psi)\mathbf{V}_{i,(j,1)}.$$

Let $M_{\mathbf{X}} := \max_{\beta,\mathbf{X}} e^{\beta^\top \mathbf{X}}$ and $M_{\mathbf{W}} := \max_{\gamma,\mathbf{W}} e^{\gamma^\top \mathbf{W}}$. Under conditions C1, C2 and C4, we have :

$$|\mathbf{U}_{i,2}(\psi)| \leq M^* := M \cdot M_{\mathbf{X}} \cdot M_{\mathbf{W}} \cdot (1 + M_{\mathbf{X}})^{M+1} < \infty,$$

which implies

$$\mathbb{E}\left(\left\{\frac{\partial^2 \ell_i(\psi)}{\partial\psi_\ell \partial\psi_j}\right\}^2\right) \leq c_0^4 M^{*2},$$

and finally,

$$\text{var}\left(\frac{\delta_i}{r_i(\alpha_0)}\frac{\partial^2 \ell_i(\psi)}{\partial\psi_\ell \partial\psi_j}\right) \leq c_4 c_0^4 M^{*2} < \infty.$$

It follows that

$$\sum_{i=1}^{\infty} \frac{\text{var}\left(\frac{\delta_i}{r_i(\alpha_0)}\frac{\partial^2 \ell_i(\psi)}{\partial\psi_\ell \partial\psi_j}\right)}{i^2} \leq c_4 c_0^4 M^{*2} \sum_{i=1}^{\infty} \frac{1}{i^2} < \infty.$$

Therefore, Kolmogorov's law of large numbers implies that

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\delta_i}{r_i(\alpha_0)}\frac{\partial^2 \ell_i(\psi)}{\partial\psi_\ell \partial\psi_j} - \mathbb{E}\left[\frac{\delta_i}{r_i(\alpha_0)}\frac{\partial^2 \ell_i(\psi)}{\partial\psi_\ell \partial\psi_j}\right]\right\}$$

converges in probability to 0 as $n \to \infty$ and by condition C3, $(\widetilde{U}_{w,n}(\psi, \alpha_0))_{(\ell,j)}$ converges in probability to the (ℓ, j)-th element of the matrix $A(\psi, \alpha_0)$. Finally, $\widetilde{U}_{w,n}(\psi, \widehat{\alpha}_n)$ converges in probability to $A(\psi, \alpha_0)$. Under conditions C1, C2 and C4, the derivative of $\widetilde{U}_{w,n}(\psi, \widehat{\alpha}_n)$ with respect to ψ is bounded, for every n. Therefore, the sequence $(\widetilde{U}_{w,n}(\psi, \widehat{\alpha}_n))_n$ is equicontinuous. It follows that the convergence of $\widetilde{U}_{w,n}(\psi, \widehat{\alpha}_n)$ to $A(\psi, \alpha_0)$ is uniform on $\mathscr{V}_{\psi_0}$. □

Having now verified the conditions of Foutz [1977] inverse function theorem, we conclude that $\widehat{\psi}_n$ converges in probability to ψ_0. □

6.3.2 Asymptotic normality

Our second main result asserts that the IPW-MLE $\widehat{\psi}_n$ is asymptotically Gaussian.

Theorem 6.3.5 *Assume that conditions C1-C4 hold. Then $\sqrt{n}(\widehat{\psi}_n - \psi_0)$ is asymptotically normally distributed with mean zero and covariance matrix Δ, where*

$$\Delta := A(\psi_0,\alpha_0)^{-1}\{J(\psi_0,\alpha_0) - B(\psi_0,\alpha_0)\Sigma(\alpha_0)^{-1}B(\psi_0,\alpha_0)^\top\}\left[A(\psi_0,\alpha_0)^{-1}\right]^\top.$$

Proof of Theorem 6.3.5. A Taylor series expansion of $U_{w,n}(\widehat{\psi}_n,\widehat{\alpha}_n)$ at (ψ_0,α_0) yields

$$0 = U_{w,n}(\widehat{\psi}_n,\widehat{\alpha}_n) = U_{w,n}(\psi_0,\alpha_0) + \frac{\partial U_{w,n}(\psi_0,\alpha_0)}{\partial \psi^\top}(\widehat{\psi}_n - \psi_0) + \frac{\partial U_{w,n}(\psi_0,\alpha_0)}{\partial \alpha^\top}(\widehat{\alpha}_n - \alpha_0) + o_{\mathbb{P}}(1).$$

Let $\check{U}_{w,n}(\psi,\alpha) := n^{-1/2}\partial U_{w,n}(\psi,\alpha)/\partial\alpha^\top$. Then we have :

$$0 = U_{w,n}(\psi_0,\alpha_0) + \widetilde{U}_{w,n}(\psi_0,\alpha_0)\sqrt{n}(\widehat{\psi}_n - \psi_0) + \check{U}_{w,n}(\psi_0,\alpha_0)\sqrt{n}(\widehat{\alpha}_n - \alpha_0) + o_{\mathbb{P}}(1). \tag{6.11}$$

Now, straightforward calculations yield

$$\check{U}_{w,n}(\psi,\alpha) = -\frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{1-r_i(\alpha)}{r_i(\alpha)}\dot{\ell}_i(\psi)\mathbf{S}_i^\top,$$

and it can be proved that $\check{U}_{w,n}(\psi_0,\alpha_0)$ converges in probability to $B(\psi_0,\alpha_0)$ (arguments are similar to those in proof of Lemma 6.3.4 and are thus omitted). Combining this with (6.8), we can re-express (6.11) as :

$$0 = U_{w,n}(\psi_0,\alpha_0) + \widetilde{U}_{w,n}(\psi_0,\alpha_0)\sqrt{n}(\widehat{\psi}_n - \psi_0) + B(\psi_0,\alpha_0)\Sigma(\alpha_0)^{-1}M_n(\alpha_0) + o_{\mathbb{P}}(1),$$

and it follows that :

$$\sqrt{n}(\widehat{\psi}_n - \psi_0) = -\widetilde{U}_{w,n}(\psi_0,\alpha_0)^{-1}\left(U_{w,n}(\psi_0,\alpha_0) + B(\psi_0,\alpha_0)\Sigma(\alpha_0)^{-1}M_n(\alpha_0)\right) + o_{\mathbb{P}}(1).$$

Using notations introduced in Section 6.2.3, we finally obtain :

$$\begin{aligned}\sqrt{n}(\widehat{\psi}_n - \psi_0) &= -\widetilde{U}_{w,n}(\psi_0,\alpha_0)^{-1}n^{-1/2}\mathbb{V}C(\psi_0,\alpha_0) + o_{\mathbb{P}}(1),\\ &= -\widetilde{U}_{w,n}(\psi_0,\alpha_0)^{-1}\sum_{j=1}^{3n}\mathbb{V}_{\bullet j}C_{j,n}(\psi_0,\alpha_0) + o_{\mathbb{P}}(1),\end{aligned}$$

where $C_{j,n}(\psi_0,\alpha_0) = n^{-1/2}C_j(\psi_0,\alpha_0)$. Let $\mathbb{C}_n^2 = \text{var}\left(U_{w,n}(\psi_0,\alpha_0) + B(\psi_0,\alpha_0)\Sigma(\alpha_0)^{-1}M_n(\alpha_0)\right)$. Then, by Eicker [1966], the random linear form $\mathbb{C}_n^{-1}\sum_{j=1}^{3n}\mathbb{V}_{\bullet j}C_{j,n}(\psi_0,\alpha_0)$ converges in distribution to the k-dimensional standard Gaussian distribution if the following conditions are satisfied : a) $\max_{1\leq j\leq 3n}\mathbb{V}_{\bullet j}^\top(\mathbb{V}\mathbb{V}^\top)^{-1}\mathbb{V}_{\bullet j} \to 0$ as $n\to\infty$, b) $\sup_{1\leq j\leq 3n}\mathbb{E}[C_{j,n}^2(\psi_0,\alpha_0)1_{\{|C_{j,n}(\psi_0,\alpha_0)|>c\}}] \to 0$ as $c\to\infty$, c) $\inf_{1\leq j\leq 3n}\mathbb{E}[C_{j,n}^2(\psi_0,\alpha_0)] > 0$. Note first that

$$0 < \max_{1\leq j\leq 3n}\mathbb{V}_{\bullet j}^\top(\mathbb{V}\mathbb{V}^\top)^{-1}\mathbb{V}_{\bullet j} \leq \max_{1\leq j\leq 3n}\|\mathbb{V}_{\bullet j}\|^2\|(\mathbb{V}\mathbb{V}^\top)^{-1}\| = \max_{1\leq j\leq 3n}\|\mathbb{V}_{\bullet j}\|^2/\lambda_n.$$

Since $\|\mathbb{V}_{\bullet j}\|$ is bounded, condition C3 implies that condition a) is satisfied. Condition b) follows by noting that $C_{j,n}(\psi_0, \alpha_0)$ (for $j = 1, \ldots, 3n$) are bounded under C1, C2, C4. Finally, under C1, C2 and C4, we have $\mathbb{E}[C_{j,n}^2(\psi_0, \alpha_0)] > 0$ for every
$j = 1, \ldots, 3n$.
Moreover,

$$\mathbb{C}_n^2 = \text{var}\left(U_{w,n}(\psi_0, \alpha_0)\right) + \text{var}\left(B(\psi_0, \alpha_0)\Sigma(\alpha_0)^{-1}M_n(\alpha_0)\right) + 2\text{cov}\left(U_{w,n}(\psi_0, \alpha_0), B(\psi_0, \alpha_0)\Sigma(\alpha_0)^{-1}M_n(\alpha_0)\right).$$

Straightforward calculations yield : $\text{var}\left(U_{w,n}(\psi_0, \alpha_0)\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{\dot{\ell}_i(\psi_0)\dot{\ell}_i(\psi_0)^\top}{r_i(\alpha_0)}\right]$,
$\text{var}\left(M_n(\alpha_0)\right) = \Sigma(\alpha_0)$ and

$$\text{cov}\left(U_{w,n}(\psi_0, \alpha_0), B(\psi_0, \alpha_0)\Sigma(\alpha_0)^{-1}M_n(\alpha_0)\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\dot{\ell}_i(\psi_0)\delta_i\frac{(1 - r_i(\alpha_0))}{r_i(\alpha_0)}\mathbf{S}_i^\top\right]\Sigma(\alpha_0)^{-1}B(\psi_0, \alpha_0)^\top.$$

Hence, $\mathbb{C}_n^2$ converges to $J(\psi_0, \alpha_0) - B(\psi_0, \alpha_0)\Sigma(\alpha_0)^{-1}B(\psi_0, \alpha_0)^\top$. It follows that
$\sum_{j=1}^{3n}\mathbb{V}_{\bullet j}C_{j,n}(\psi_0, \alpha_0)$ converges in distribution to a k-dimensional Gaussian vector with mean zero and variance $J(\psi_0, \alpha_0) - B(\psi_0, \alpha_0)\Sigma(\alpha_0)^{-1}B(\psi_0, \alpha_0)^\top$. Finally, using Lemma 6.3.4, condition C3 and Slutsky's theorem, $\sqrt{n}(\widehat{\psi}_n - \psi_0)$ converges in distribution to a mean-zero Gaussian vector with variance Δ, where Δ is defined in Theorem 6.3.5. □

Remark. A consistent estimator of Δ is given by

$$\widehat{\Delta}_n := A_n(\widehat{\psi}_n, \widehat{\alpha}_n)^{-1}\{J_n(\widehat{\psi}_n, \widehat{\alpha}_n) - B_n(\widehat{\psi}_n, \widehat{\alpha}_n)\Sigma_n(\widehat{\alpha}_n)^{-1}B_n(\widehat{\psi}_n, \widehat{\alpha}_n)^\top\}\left[A_n(\widehat{\psi}_n, \widehat{\alpha}_n)^{-1}\right]^\top,$$

where

$$A_n(\psi, \alpha) = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{r_i(\alpha)}\frac{\partial^2\ell_i(\psi)}{\partial\psi\partial\psi^\top},$$

$$B_n(\psi, \alpha) = -\frac{1}{n}\sum_{i=1}^n \delta_i\frac{1 - r_i(\alpha)}{r_i(\alpha)}\dot{\ell}_i(\psi)\mathbf{S}_i^\top,$$

$$J_n(\psi, \alpha) = \frac{1}{n}\sum_{i=1}^n \frac{\dot{\ell}_i(\psi)\dot{\ell}_i(\psi)^\top}{r_i(\alpha)},$$

$$\Sigma_n(\alpha) = \frac{1}{n}\sum_{i=1}^n \mathbf{S}_i\mathbf{S}_i^\top r_i(\alpha)(1 - r_i(\alpha)).$$

The proof proceeds along the same lines as proof of Lemma 6.3.4 and is therefore omitted.

6.4 Simulation study

In this section, we investigate the finite-sample performances of the IPW estimator under various conditions.

6.4.1 Simulation design

The following ZIB regression model is used to simulate data :

$$\text{logit}(\pi_i) = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6}$$

and

$$\text{logit}(p_i) = \gamma_1 W_{i1} + \gamma_2 W_{i2} + \gamma_3 W_{i3} + \gamma_4 W_{i4},$$

where $X_{i1} = W_{i1} = 1$ and the $X_{i2}, \ldots, X_{i6}$ and W_{i4} are independently drawn from normal $\mathscr{N}(0,1)$, uniform $\mathscr{U}(2,5)$, normal $\mathscr{N}(1,1.5)$, exponential $\mathscr{E}(1)$, binomial $\mathscr{B}(1,0.3)$ and normal $\mathscr{N}(-1,1)$ distributions respectively. Linear predictors in $\text{logit}(\pi_i)$ and $\text{logit}(p_i)$ are allowed to share common terms by letting $W_{i2} = X_{i2}$ and $W_{i3} = X_{i6}$. The regression parameter β is chosen as $\beta = (-0.3, 1.2, 0.5, -0.75, -1, 0.8)^\top$. The regression parameter γ is chosen as :

— case 1 : $\gamma = (-0.55, -0.7, -1, 0.45)^\top$

— case 2 : $\gamma = (0.25, -0.4, 0.8, 0.45)^\top$

We consider the following sample sizes, $n = 500, 1000$. The numbers m_i are allowed to vary across subjects, with $m_i \in \{4, 8, 10, 15\}$. Let $n_j = \text{card}\{i : m_i = j\}$, for $j = 4, 8, 10, 15$. For $n = 500$, we let $(n_4, n_8, n_{10}, n_{15}) = (125, 125, 125, 125)$ and for $n = 1000$, we let $(n_4, n_8, n_{10}, n_{15}) = (250, 250, 250, 250)$.

Using these values, in case 1 (respectively case 2), the average percentage of zero-inflation in the simulated data sets is 25% (respectively 50%). Missingness indicators δ_i are simulated from a logistic regression model with selection probability $r_i(\alpha) := \mathbb{P}(\delta_i = 1|\mathbf{S}_i) = \text{logit}^{-1}(\alpha^\top \mathbf{S}_i)$, with $\mathbf{S}_i := (1, Z_i, X_{i2}, W_{i4})$. The regression parameter α is chosen to yield average missingness proportions in the simulated samples successively equal to 0.2 and 0.4. Finally, for each combination of the simulation design parameters (sample size, proportions of zero-inflation and missing data), we simulate $N = 1000$ samples and we calculate the IPW estimate $\widehat{\psi}_n$. Simulations are carried out using the statistical software `R`. We use the package `maxLik` [Henningsen and Toomet, 2011] to solve the estimating equation (6.2.2).

6.4.2 Results

For each configuration [`sample size` $\times$ `zero-inflation proportion` $\times$ `proportion of missing data`] of the simulation parameters, we calculate the average absolute relative bias (as a percentage) of the estimates $\widehat{\beta}_{j,n}$ and $\widehat{\gamma}_{k,n}$ over the N simulated samples (for example, the absolute relative bias of $\widehat{\beta}_{j,n}$ is obtained as $N^{-1}\sum_{t=1}^{N}|(\widehat{\beta}_{j,n}^{(t)}-\beta_j)/\beta_j|\times 100$, where $\widehat{\beta}_{j,n}^{(t)}$ denotes the IPW estimate of β_j in the t-th simulated sample). We also obtain the average standard error SE (calculated from (6.3.2)), empirical standard deviation (SD) and root mean square error (RMSE) for each estimator $\widehat{\beta}_{j,n}$ $(j=1,\ldots,6)$ and $\widehat{\gamma}_{k,n}$ $(k=1,\ldots,4)$. Finally, we provide the empirical coverage probability (CP) of 95%-level confidence intervals for the β_j and γ_k. Results are given in Table 6.1 (case 1, $n=500$), Table 6.2 (case 1, $n=1000$), Table 6.3 (case 2, $n=500$), Table 6.4 (case 2, $n=1000$). For purpose of comparison, we also provide results for the usual MLE, which could be obtained if there were no missing covariates. This estimator solves the score equation $U_n(\psi)=0$ given by (6.2) (in what follows, we refer this estimator to as the "full data" - or FD - estimator). The complete-case estimator of ψ can be obtained by solving the score equation (6.2), based on complete-cases only. However, this estimator is generally highly biased. For example, in our simulations, its relative biais can reach up to 200%, resulting in very low covarage probabilities. Therefore, we do not provide results for this estimate.

From these results, we observe, as expected, that the bias, SE, SD and RMSE of the IPW estimator decrease as the sample size increases and the proportion of individuals with missing covariates decreases. Moreover, the bias of the IPW estimator stays moderate and its empirical coverage probabilities are close to the nominal confidence level, even when the sample size is moderate ($n=500$). As may also be expected, for a given proportion of missing data, we observe that the IPW estimator of the β_js (respectively γ_ks) performs better when the zero-inflation proportion decreases (respectively increases). The FD estimator obviously performs better than the IPW estimator, but FD analysis cannot be performed when missing data are present. Overall, these numerical results indicate the good performance of the IPW method for estimating a ZIB regression model under missing data.

Finally, in order to assess the quality of the Gaussian approximation stated in Theorem 6.3.5, we provide normal Q-Q plots of the estimates (see figures 6.1 and 6.2 for $n=500$ in case 2 and a fraction of missing data equal to 0.4 and figures 6.3 and 6.4 for $n=1000$ in case 2 and a fraction of missing data equal to 0.2 (plots for the other simulated scenarios yield similar observations and are thus omitted). From these figures, it appears that the Gaussian approximation of the distribution of the IPW estimator is reasonably satisfied, even when the sample size is moderate and the proportion of individuals with missing covariates is as high as 0.4.

average fraction of missing data			$\widehat{\beta}_n$						$\widehat{\gamma}_n$			
			$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$
FD		rel. bias	3.4726	0.2880	0.1150	0.3161	0.5336	0.8380	1.8986	0.9012	1.9091	1.5690
		SD	0.1944	0.0611	0.0519	0.0361	0.0575	0.0964	0.1785	0.1509	0.3360	0.1361
		SE	0.1979	0.0591	0.0536	0.0370	0.0592	0.0981	0.1854	0.1513	0.3264	0.1372
		RMSE	0.2775	0.0850	0.0746	0.0517	0.0827	0.1376	0.2575	0.2138	0.4687	0.1933
		CP	0.9629	0.9459	0.9569	0.9609	0.9619	0.9559	0.9609	0.9559	0.9429	0.9599
IPW	0.2											
		rel. bias	2.7202	0.4861	0.0590	0.4251	0.6387	0.6311	3.6159	0.9854	4.0350	2.9869
		SD	0.2103	0.0649	0.0554	0.0386	0.0643	0.1047	0.2135	0.1965	0.4653	0.1865
		SE	0.2150	0.0640	0.0582	0.0408	0.0666	0.1057	0.2167	0.1925	0.4466	0.1799
		RMSE	0.3008	0.0913	0.0804	0.0563	0.0927	0.1488	0.3048	0.2751	0.6460	0.2594
		CP	0.9620	0.9550	0.9610	0.9650	0.9610	0.9570	0.9700	0.9560	0.9520	0.9490
IPW	0.4											
		rel. bias	0.4867	0.3730	0.7212	0.6061	0.7341	0.8895	3.4735	2.5081	5.9316	3.5603
		SD	0.2491	0.0817	0.0664	0.0462	0.0759	0.1234	0.2453	0.2391	0.5275	0.2178
		SE	0.2537	0.0802	0.0684	0.0476	0.0760	0.1261	0.2569	0.2450	0.5602	0.2249
		RMSE	0.3555	0.1146	0.0954	0.0665	0.1077	0.1765	0.3556	0.3427	0.7716	0.3134
		CP	0.9479	0.9550	0.9530	0.9550	0.9479	0.9550	0.9690	0.9700	0.9690	0.9670

TABLE 6.1 – Simulation results (case 1, $n = 500$). SD : empirical standard deviation. SE : average standard error. RMSE : empirical root mean square error. CP : empirical coverage probability of 95%-level confidence intervals.

average fraction of missing data			$\widehat{\beta}_n$						$\widehat{\gamma}_n$			
			$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$
FD												
		rel. bias	0.6276	0.0786	0.1061	0.2259	0.0748	0.4154	1.1894	0.2426	0.7370	0.2883
		SD	0.1346	0.0416	0.0370	0.0242	0.0423	0.0708	0.1279	0.1040	0.2341	0.0954
		SE	0.1389	0.0415	0.0376	0.0260	0.0414	0.0690	0.1300	0.1057	0.2270	0.0958
		RMSE	0.1933	0.0588	0.0528	0.0356	0.0592	0.0989	0.1824	0.1482	0.3261	0.1352
		CP	0.9570	0.9510	0.9570	0.9640	0.9449	0.9479	0.9560	0.9499	0.9540	0.9560
IPW	0.2											
		rel. bias	0.2160	0.1263	0.1224	0.2970	0.0850	0.3158	1.4117	0.6978	2.5321	0.7454
		SD	0.1460	0.0446	0.0399	0.0259	0.0465	0.0771	0.1472	0.1323	0.3032	0.1267
		SE	0.1505	0.0448	0.0408	0.0284	0.0475	0.0748	0.1511	0.1357	0.3050	0.1269
		RMSE	0.2096	0.0632	0.0570	0.0385	0.0665	0.1074	0.2110	0.1895	0.4308	0.1794
		CP	0.9590	0.9450	0.9540	0.9670	0.9500	0.9500	0.9610	0.9600	0.9590	0.9550
IPW	0.4											
		rel. bias	0.7322	0.3466	0.2252	0.2982	0.5650	0.2990	0.7791	0.9170	3.8227	2.8852
		SD	0.1712	0.0537	0.0462	0.0330	0.0512	0.0868	0.1665	0.1647	0.3638	0.1514
		SE	0.1774	0.0556	0.0480	0.0334	0.0532	0.0875	0.1728	0.1625	0.3537	0.1490
		RMSE	0.2465	0.0774	0.0666	0.0470	0.0741	0.1233	0.2400	0.2314	0.5087	0.2128
		CP	0.9620	0.9560	0.9600	0.9520	0.9610	0.9580	0.9590	0.9500	0.9540	0.9440

TABLE 6.2 – Simulation results (case 1, $n = 1000$). SD : empirical standard deviation. SE : average standard error. RMSE : empirical root mean square error. CP : empirical coverage probability of 95%-level confidence intervals.

average fraction of missing data			$\widehat{\beta}_n$						$\widehat{\gamma}_n$			
			$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$
FD												
		rel. bias	0.8028	0.7069	0.5104	0.7715	0.9714	0.0022	0.0209	0.0103	0.2573	0.6092
		SD	0.2427	0.0749	0.0677	0.0461	0.0736	0.1352	0.1711	0.1162	0.2346	0.1118
		SE	0.2475	0.0738	0.0675	0.0469	0.0759	0.1410	0.1669	0.1164	0.2293	0.1110
		RMSE	0.3466	0.1054	0.0957	0.0660	0.1061	0.1953	0.2389	0.1645	0.3280	0.1575
		CP	0.9559	0.9349	0.9529	0.9529	0.9499	0.9549	0.9499	0.9529	0.9449	0.9439
IPW	0.2											
		rel. bias	1.0135	0.7771	0.5658	0.8772	1.0211	0.1457	1.2474	0.3254	0.2989	0.8908
		SD	0.2524	0.0787	0.0705	0.0477	0.0754	0.1386	0.1833	0.1277	0.2522	0.1204
		SE	0.2563	0.0771	0.0698	0.0487	0.0788	0.1453	0.1833	0.1289	0.2551	0.1207
		RMSE	0.3596	0.1106	0.0992	0.0684	0.1095	0.2007	0.2592	0.1814	0.3586	0.1705
		CP	0.9500	0.9310	0.9510	0.9450	0.9550	0.9580	0.9440	0.9600	0.9550	0.9530
IPW	0.4											
		rel. bias	1.3807	0.9649	0.5492	1.0489	0.9995	0.5640	1.6315	4.0769	2.0604	5.0802
		SD	0.2771	0.0872	0.0773	0.0545	0.0938	0.1549	0.2181	0.1845	0.3650	0.1812
		SE	0.2888	0.0854	0.0789	0.0559	0.0928	0.1631	0.2182	0.1805	0.3557	0.1778
		RMSE	0.4001	0.1226	0.1105	0.0784	0.1323	0.2249	0.3085	0.2586	0.5099	0.2548
		CP	0.9640	0.9460	0.9590	0.9550	0.9460	0.9610	0.9510	0.9470	0.9440	0.9540

TABLE 6.3 – Simulation results (case 2, $n = 500$). SD : empirical standard deviation. SE : average standard error. RMSE : empirical root mean square error. CP : empirical coverage probability of 95%-level confidence intervals.

	average fraction of missing data		$\widehat{\beta}_n$						$\widehat{\gamma}_n$			
			$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$
FD												
		rel. bias	0.8391	0.1674	0.0338	0.4053	0.1128	0.0499	1.9813	1.7535	1.1359	0.9103
		SD	0.1688	0.0518	0.0461	0.0316	0.0552	0.0969	0.1194	0.0837	0.1668	0.0801
		SE	0.1731	0.0514	0.0472	0.0328	0.0528	0.0985	0.1175	0.0818	0.1615	0.0780
		RMSE	0.2417	0.0730	0.0660	0.0456	0.0764	0.1382	0.1676	0.1172	0.2323	0.1119
		CP	0.9549	0.9529	0.9599	0.9619	0.9379	0.9579	0.9449	0.9339	0.9409	0.9409
IPW	0.2											
		rel. bias	0.7979	0.1926	0.0279	0.4838	0.1741	0.1785	1.2374	1.6033	1.6918	0.9563
		SD	0.1738	0.0534	0.0474	0.0322	0.0566	0.1004	0.1272	0.0911	0.1811	0.0866
		SE	0.1814	0.0552	0.0495	0.0343	0.0556	0.1031	0.1272	0.0893	0.1823	0.0858
		RMSE	0.2512	0.0768	0.0686	0.0472	0.0793	0.1439	0.1799	0.1277	0.2572	0.1220
		CP	0.9570	0.9520	0.9570	0.9630	0.9540	0.9640	0.9550	0.9440	0.9520	0.9490
IPW	0.4											
		rel. bias	2.6218	0.2938	0.2344	0.5926	0.1376	0.4597	4.6748	3.3595	0.6924	3.6510
		SD	0.1964	0.0584	0.0537	0.0374	0.0666	0.1105	0.1536	0.1246	0.2435	0.1308
		SE	0.2005	0.0596	0.0547	0.0391	0.0647	0.1133	0.1516	0.1251	0.2482	0.1262
		RMSE	0.2807	0.0835	0.0766	0.0542	0.0928	0.1583	0.2161	0.1770	0.3476	0.1824
		CP	0.9540	0.9480	0.9570	0.9600	0.9470	0.9560	0.9500	0.9480	0.9540	0.9420

TABLE 6.4 – Simulation results (case 2, $n = 1000$). SD : empirical standard deviation. SE : average standard error. RMSE : empirical root mean square error. CP : empirical coverage probability of 95%-level confidence intervals.

6.5 Discussion

Zero-inflated binomial regression is now commonly used for investigating count data with excess of zero ; see, for example, Gilthorpe et al. [2009], Matranga et al. [2013], Diallo et al. [2017]. In this paper, we extend the scope of ZIB regression by considering the situation where some covariates are missing at random. In this setting, we propose an inverse-probability-weighted-type estimator by assuming that missingness probabilities can be modeled parametrically. Consistency and asymptotic normality of the proposed estimator are established and a consistant variance estimator is constructed. Our simulation study suggests that the IPW estimator performs well under a wide range of conditions.

Now, several issues deserve attention. First, the proposed estimator is valid if the parametric model for missingness probabilities $\mathbb{P}(\delta_i = 1|\mathbf{S}_i)$ is correctly specified. Misspecifying this model may lead to a biased IPW estimator. Several solutions to this issue might be investigated. For example, one may consider nonparametric estimation of the missingness probabilities. An alternative approach relies on the so-called augmented IPW method, which is robust to a misspecification of the selection probabilities. Robustness of these various approaches to a violation of the MAR assumption also constitutes a topic of interest for future research. Another stimulating topic for future work is as follows. In this paper, we consider missing covariates in the basic ZIB regression model (6.1). The same issue could be investigated in various generalizations of ZIB regression (such as the random-effects ZIB model proposed by Hall [2000], or semi-parametric ZIB models, for example).

Overall, the present work constitutes a first step towards the analysis of zero-inflated binomial counts with missing data. Further research is now needed to extend this contribution to more complex ZIB models and more sophisticated and robust estimation methods.

Acknowledgements

Authors acknowledge financial support from the "Service de Coopération et d'Action Culturelle" of the French Embassy in Senegal and logistical support from Campus France (French national agency for the promotion of higher education, international student services, and international mobility). Authors also acknowledge grants from CEA-MITIC, an African Center of Excellence in Mathematics, Informatics and ICT implemented by Gaston Berger University (Senegal).

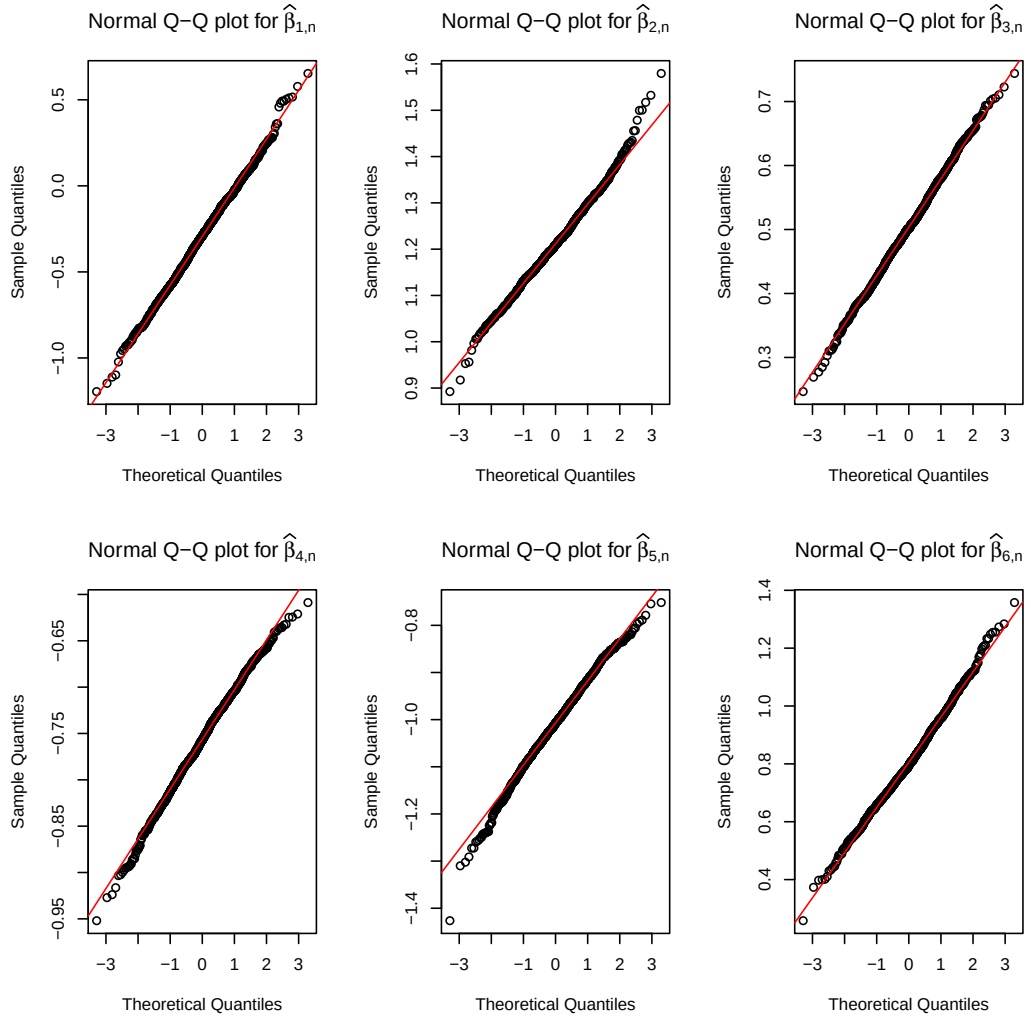

FIGURE 6.1 – Normal Q-Q plots for $\hat{\beta}_{1,n}, \ldots, \hat{\beta}_{6,n}$ with $n = 500$ (case 2) and a fraction of missing data equal to 0.4.

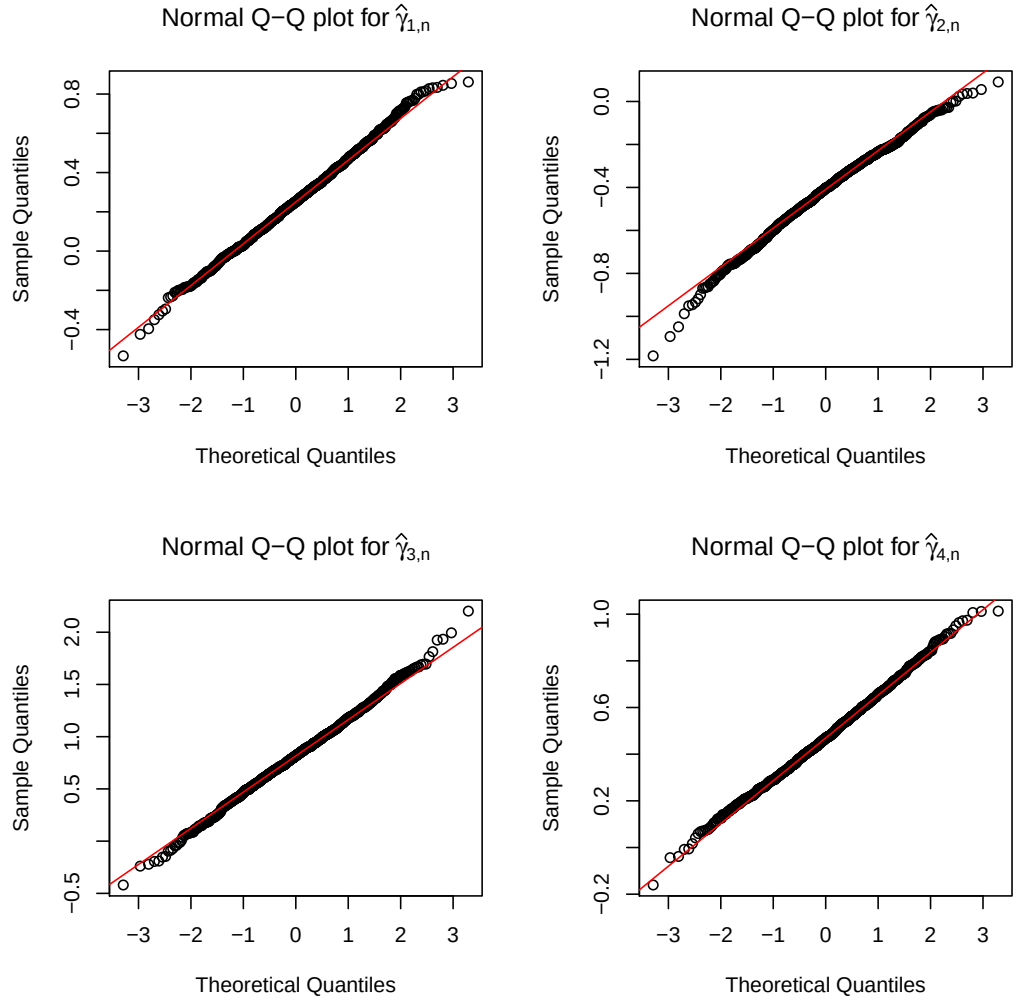

FIGURE 6.2 – Normal Q-Q plots for $\hat{\gamma}_{1,n}, \ldots, \hat{\gamma}_{4,n}$ with $n = 500$ (case 2) and a fraction of missing data equal to 0.4.

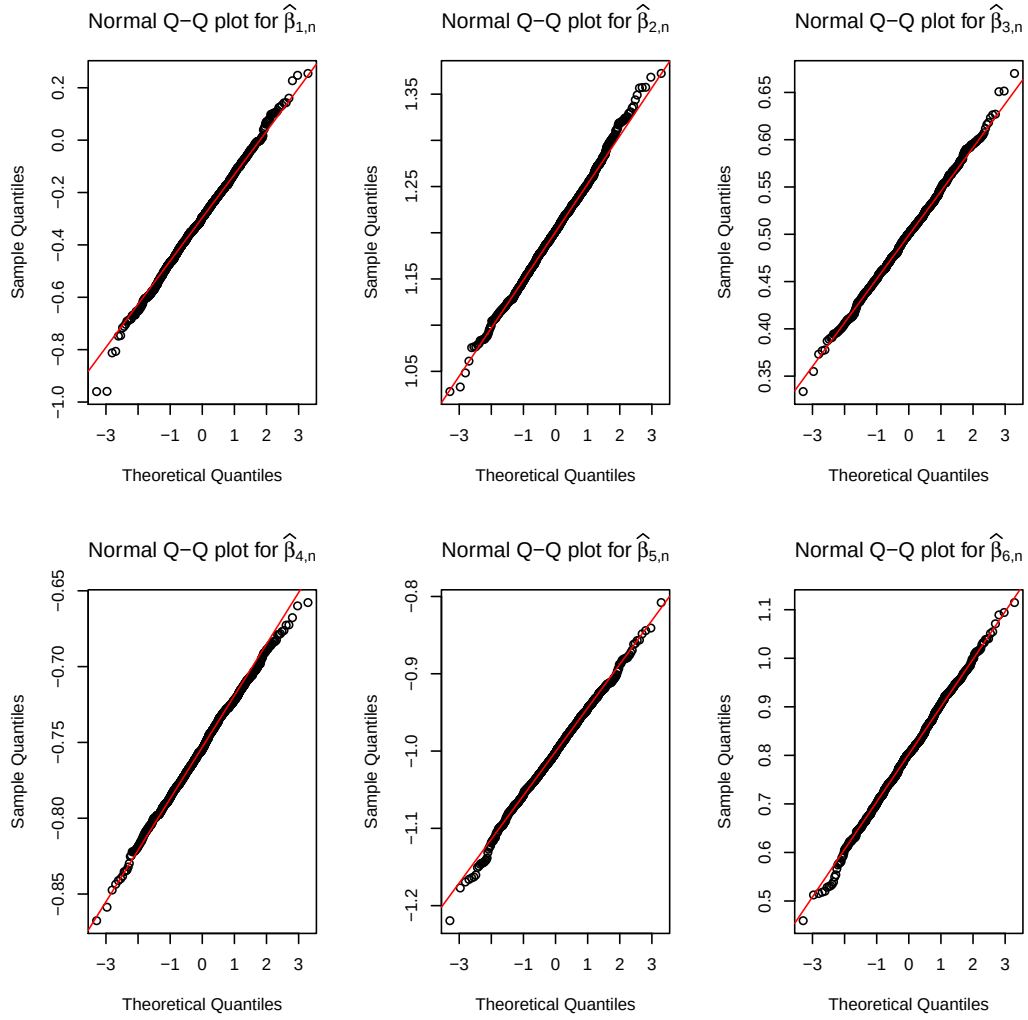

FIGURE 6.3 – Normal Q-Q plots for $\hat{\beta}_{1,n}, \ldots, \hat{\beta}_{6,n}$ with $n = 1000$ (case 2) and a fraction of missing data equal to 0.2.

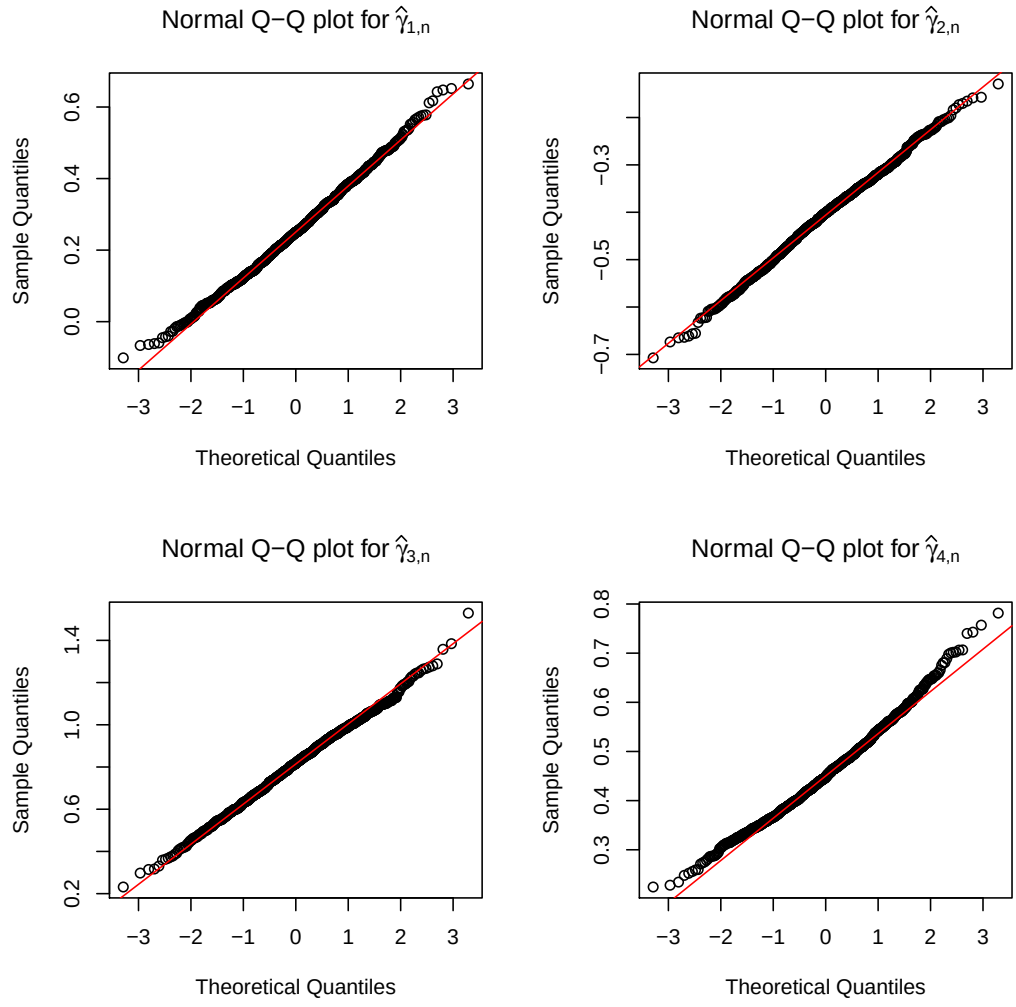

FIGURE 6.4 – Normal Q-Q plots for $\hat{\gamma}_{1,n}, \ldots, \hat{\gamma}_{4,n}$ with $n = 1000$ (case 2) and a fraction of missing data equal to 0.4.

Conclusion et perspectives

Dans ce travail, nous nous somme intéressés au problème de l'inférence statistique dans des modèles de comptage à inflation de zéro. C'est un travail qui s'articule autour de trois contributions.

D'abord 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 (MV) des paramètres du modèle de régression binomial à inflation de zéro, question théorique qui jusqu'à présent était ignorée. Pour compléter cette étude théorique, nous avons mené une étude de simulations exhaustive, et qui a permis d'étudier les propriétés à distance finie des estimateurs du MV. Une application du modèle sur des données réelles d'économie de la santé, issues d'une étude portant sur la consommation de soins médicaux des patients âgés aux USA, a ensuite été réalisée.

Ensuite, nous avons proposé un nouveau modèle pour données de comptages avec inflation de zéro. C'est un nouvel outil qui permet de modéliser la survenue d'un excès de zéro dans des données de comptage multinomiales. Il permet donc de considérer simultanément plusieurs comptages dépendant les uns des autres. Ce modèle permet en particulier d'étudier les déterminants du renoncement aux soins dans une population confrontée à une offre de soins multiples. Nous avons construit des estimateurs du type MV des paramètres de ce modèle et en avons établi rigoureusement les propriétés asymptotiques. Nous avons également réalisé des études de simulations approfondies, qui nous ont permis de valider les résultats théoriques. Pour finir, une application de ce modèle sur données réelles nous a permis d'obtenir des résultats beaucoup plus concluants que les analyses précédemment réalisées, dans la littérature, sur cette même base de données.

Et enfin, nous nous sommes intéressés aussi au problème de l'inférence statistique dans le modèle ZIB en présence de données manquantes sur les covariables. Dans cette situation nous avons proposé une méthode d'estimation adaptée en utilisant le principe de la « *pondération par l'inverse de la probabilité de sélection* ». Nous avons à nouveau établi l'asymptotique des estimateurs proposés et en avons étudié, par simulations, les propriétés pour des tailles finies d'échantillons. Nous avons en particulier mis en évidence l'avantage des estimateurs proposés, par rapport aux seuls estimateurs actuellements disponibles dans le modèle ZIB à données manquantes (estimateurs dit « *cas-complets* »).

Cependant plusieurs axes de recherche sont envisagés. Le premier consisterait à explorer le modèle ZIB avec effets aléatoires. Le second axe consisterait à généraliser le modèle ZIB selon la complexité croissante des données expérimentales. Nous nous intéressons au problème de l'inférence dans le modèle ZIB dans le cas où les probabilités de mélanges et/ou de succès sont modélisées par une régression semi-paramétrique.

Troisième partie

Annexes

ANNEXE A

Listing R

Nous présentons ci-dessous les programmes d'implémentation décrits dans le chapitre 4 et le chapitre 6 pour les simulations et dans le chapitre 5 pour ce qui concerne l'application sur données réelles.

A.1 Script de simulations du chapitre 4

```
library(maxLik)

################################
#
# Initialisations paramètres du modèle
#
################################

n=500 # sample size

b=c(-.3,1.2,0.5,-0.75,-1,0.8,0) # vecteur beta
g=c(-.55,-.7,-1,.45,0)  # vecteur gamma avec g=c(-.55,-.7,-1,.45,0) => 0.25 de zéro-inflation
                        # g=c(.25,-.4,.8,.45,0) => 0.50 de zéro-inflation

phi=c(b,g)

################################
# Simulation/estimation
################################
```

```
N=5000 # nombre de replications

# vecteurs de récupération

estimates=matrix(rep(0,N*(length(phi))),ncol=N) # récupération des N vecteurs d'estimations
                                                # de tous les paramètres
est.cov=estimates

mi=c(rep(4,200),rep(5,170),rep(6,130)) ### ATTENTION: CHANGER EN FONCTION DE n !!!!!

# pour n=300: mi=c(rep(4,120),rep(5,100),rep(6,80)) pour n=150, moitié de chaque

for (i in 1:N){
print(i)
# simulation des données

inter=rep(1,n)
X1=rnorm(n,0,1)
X2=runif(n,2,5)
X3=rnorm(n,1,1.5)
X4=rexp(n)
X5=rbinom(n,1,.3)
X6=rnorm(n,-1,1)
W1=rnorm(n,-1,1)
W2=rbinom(n,1,.5)

X=rbind(inter,X1,X2,X3,X4,X5,X6)
W=rbind(inter,X1,X5,W1,W2)

pi1=exp(t(b)%*%X)/(1+exp(t(b)%*%X))
pi2=exp(t(g)%*%W)/(1+exp(t(g)%*%W))

S=rbinom(n,1,pi2)  # indicateur de la zéro-inflation
Z=0*(S==1)+rbinom(n,mi,pi1)*(S!=1)
```

```
J=as.integer(Z==0)

# programmation de la fonction d'estimation

loglikfun=function(param){
b=param[1:(length(b))]
g=param[(length(b)+1):length(phi)]
#
sum(J*log(exp(t(g)%*%W)+(1+exp(t(b)%*%X))^{-mi})-log(1+exp(t(g)%*%W))
+(1-J)*(Z*t(b)%*%X-mi*log(1+exp(t(b)%*%X))))
}
# fin fonction

# estimation de phi=c(b,g) sur l'échantillon simulé

mle=maxLik(logLik=loglikfun,start=c(rep(1,length(phi))))
estimates[,i]=mle$est
est.cov[,i]=diag(vcov(mle))
}
```

A.2 Script de simulations du chapitre 6

```
library(maxLik)

N=1000 # nombre de réplications

# vecteurs de récupération

estimatesFD=matrix(rep(0,N*10),ncol=N) # récupération des N vecteurs d'estimations
                                       # de tous les paramètres
est.covFD=estimatesFD
estimatesIPW=estimatesFD
```

```
est.covIPW=estimatesFD
estimatesCC=estimatesFD
est.covCC=estimatesFD

estimatesCC2=estimatesFD
est.covCC2=estimatesFD

missing=rep(0,N)
zeroinflp=rep(0,N)

for (ind in 1:N){

#######################################
# Initialisations paramètres du modèle
#######################################

n=500 # sample size

b=c(-.3,1.2,0.5,-0.75,-1,0.8) # vecteur beta
g=c(.25,-.4,.8,.45)           # vecteur gamma
phi=c(b,g)

eta=c(.6,.6,.4,-.35)

#### valeurs de eta pour 25% de zéro-inflation (g=c(-.55,-.7,-1,.45))
#
# eta=c(.7,.6,-.3,.5) assure environ 20% de missing value;
# eta=c(.2,.2,.5,.4) assure environ 40% de missing value;
#
#### valeurs de eta pour 50% de zéro-inflation (g=c(.25,-.4,.8,.45))
#
# eta=c(.6,.6,.4,-.35) assure environ 20% de missing value;
# eta=c(0.1,.6,-.4,.6) assure environ 40% de missing value;
#
#########################################################
```

```
# mi=c(rep(4,250),rep(8,250),rep(10,250),rep(15,250)) ## ATTENTION: CHANGER EN FONCTION DE n !!
mi=c(rep(4,125),rep(8,125),rep(10,125),rep(15,125))

print(ind)

# simulation des données

inter=rep(1,n)
X1=rnorm(n,0,1)
X2=runif(n,2,5)
X3=rnorm(n,1,1.5)
X4=rexp(n)
X5=rbinom(n,1,.3)
# X6=rnorm(n,-1,1)
W1=rnorm(n,-1,1)
# W2=rbinom(n,1,.5)

X=rbind(inter,X1,X2,X3,X4,X5)
W=rbind(inter,X1,X5,W1)

pi1=exp(t(b)%*%X)/(1+exp(t(b)%*%X))
pi2=exp(t(g)%*%W)/(1+exp(t(g)%*%W))

S=rbinom(n,1,pi2)  # indicateur de la zéro-inflation
Z=0*(S==1)+rbinom(n,mi,pi1)*(S!=1)
J=as.integer(Z==0)

So=rbind(inter,Z,X1,W1)
pr=exp(t(eta)%*%So)/(1+exp(t(eta)%*%So))
delta=rbinom(n,1,pr) # indicateur de la valeur manquante: delta = 1 si données complètes
                     # et 0 si données manquantes
missing[ind]=1-mean(delta)
zeroinflp[ind]=mean(S)
```

```
# fonction d'estimation FD

loglikfunFD=function(param){
b=param[1:(length(b))]
g=param[(length(b)+1):length(phi)]
#
sum(J*log(exp(t(g)%*%W)+(1+exp(t(b)%*%X))^{-mi})-log(1+exp(t(g)%*%W))
+(1-J)*(Z*t(b)%*%X-mi*log(1+exp(t(b)%*%X))))
}

# fonction d'estimation CC

loglikfunCC=function(param){
b=param[1:(length(b))]
g=param[(length(b)+1):length(phi)]
sum(delta*(J*log(exp(t(g)%*%W)+(1+exp(t(b)%*%X))^{-mi})-log(1+exp(t(g)%*%W))
+(1-J)*(Z*t(b)%*%X-mi*log(1+exp(t(b)%*%X)))))
}

# estimation FD de phi=c(b,g)

mleFD=maxLik(logLik=loglikfunFD,start=c(rep(1,length(phi))))
estimatesFD[,ind]=mleFD$est
est.covFD[,ind]=diag(vcov(mleFD))

# estimation CC de phi=c(b,g)

mleCC=maxLik(logLik=loglikfunCC,start=c(rep(1,length(phi))))
estimatesCC[,ind]=mleCC$est
est.covCC[,ind]=diag(vcov(mleCC))

# estimation IPW de phi=c(b,g)

modele=glm(delta~Z+X1+W1,family=binomial)
pr.est=predict(modele,type="response")
```

```
loglikfunIPW=function(param){
b=param[1:(length(b))]
g=param[(length(b)+1):length(phi)]
#
# sum(na.omit((delta/pr.est)*(J*log(exp(t(g)%*%W)+(1+exp(t(b)%*%X))^{-mi})
-log(1+exp(t(g)%*%W))+(1-J)*(Z*t(b)%*%X-mi*log(1+exp(t(b)%*%X))))))
sum(((delta/pr.est)*(J*log(exp(t(g)%*%W)+(1+exp(t(b)%*%X))^{-mi})-log(1+exp(t(g)%*%W))
+(1-J)*(Z*t(b)%*%X-mi*log(1+exp(t(b)%*%X))))))
}

mleIPW=maxLik(logLik=loglikfunIPW,start=c(rep(1,length(phi))))
estimatesIPW[,ind]=mleIPW$est
# est.covIPW[,ind]=diag(vcov(mleIPW)) INAPPROPRIE CAR NE PREND PAS EN COMPTE
                                    # LA VARIABILITE DUE A ALPHA

################################################################################
# calcul de la variance asymptotique IPW
################################################################################

An=1/n*mleIPW$hessian

bn=matrix(rep(0,(length(phi))*(length(eta))),ncol=(length(eta)))
jn=matrix(rep(0,(length(phi))*(length(phi))),ncol=(length(phi)))

b=estimatesIPW[1:length(b)]
g=estimatesIPW[(length(b)+1):length(phi)]

for (i in 1:n){
#
Ai=-J[i]*mi[i]*exp(t(b)%*%X[,i])/(exp(t(g)%*%W[,i])*(1+exp(t(b)%*%X[,i]))^(mi[i]+1)
+(1+exp(t(b)%*%X[,i])))+(1-J[i])*(Z[i]-mi[i]*exp(t(b)%*%X[,i])/(1+exp(t(b)%*%X[,i])))
Bi=J[i]*exp(t(g)%*%W[,i])*(1+exp(t(b)%*%X[,i]))^(mi[i])/(1+exp(t(g)%*%W[,i])
*(1+exp(t(b)%*%X[,i]))^(mi[i]))-exp(t(g)%*%W[,i])/(1+exp(t(g)%*%W[,i]))
dotli=rbind(cbind(X[,i],rep(0,length(X[,i]))),cbind(rep(0,length(W[,i])),W[,i]))
```

```
%*%rbind(Ai,Bi)
#
bn=bn + dotli%*%t(So[,i])*(1-pr.est[i])/pr.est[i]*delta[i]
jn=jn + delta[i]*dotli%*%t(dotli)/(pr.est[i]^2)
#
}
#
Bn=-bn/n
Jn=jn/n
Sig=solve(n*vcov(modele))

Delta=solve(An)%*%(Jn-Bn%*%solve(Sig)%*%t(Bn))%*%t(solve(An))
est.covIPW[,ind]=diag(Delta/n)

} # fin boucle sur ind
```

A.3 Script R d'application sur données réelles du chapitre 5

```
##############################################################################
##############################################################################
#                                  #
#                              préparation des données  #
#                                  #
##############################################################################
##############################################################################

setwd("C:/")
load("DebTrivedi.rda")
head(DebTrivedi)
```

```
# OFP Number of physician office visits
# OFNP Number of nonphysician office visits
# OPP Number of physician outpatient visits
# OPNP Number of nonphysician outpatient visits
# EMR Number of emergency room visits
# HOSP Number of hospitalizations


par(mfrow=c(3,2))
for(i in 1:6){
  plot(table(DebTrivedi[,i]),ylab="",main=dimnames(DebTrivedi)[[2]][i])
}
attach(DebTrivedi)
dim(DebTrivedi)

############################################################
#  statistiques descriptives
############################################################
#
# proportion de 0 dans ofnp (données de base)
#
length(which(ofnp==0))/dim(DebTrivedi)[1] # proportion de 0 dans ofnp
length(which(opnp==0))/dim(DebTrivedi)[1] # proportion de 0 dans ofnp
length(which(ofp==0))/dim(DebTrivedi)[1] # proportion de 0 dans ofnp

ub=25
set_ind=which(ofnp+opnp+ofp==0|ofnp+opnp+ofp==1|ofnp+opnp+ofp>ub)
data=DebTrivedi[-set_ind,c("ofp","ofnp","opnp","health","numchron","age","gender",
"married","school","faminc","privins","medicaid")]
dim(data)
attach(data)

#
# proportion de 0 dans ofnp (données tronquées à ub)
#
```

```
length(which(ofnp==0))/dim(data)[1] # proportion de 0 dans ofnp
length(which(opnp==0))/dim(data)[1] # proportion de 0 dans ofnp
length(which(ofp==0))/dim(data)[1] # proportion de 0 dans ofnp

par(mfrow=c(1,3))
for(i in 1:3){
  plot(table(data[,i]),ylab="",main=dimnames(data)[[2]][i])
}

############################################################
#  pour sélectionner un sous-échantillon
############################################################

#set.seed(3)
#sub_sample=sample(1:dim(data)[1], 800, replace = FALSE)
#data=data[sub_sample,]

############################################################
#  fin sélection
############################################################

dim(DebTrivedi)
dim(data)
length(set_ind)

attach(data)
set_ind2=which(ofp+ofnp+opnp==0|ofp+ofnp+opnp==1) # vérif: aucune observation
                                                  # en Z_i=(0,0,0) et Z_i=(0,0,1)
set_ind3=which(ofnp+opnp==0) # indices des Z_i=(0,0,m_i)
length(set_ind3)/dim(data)[1] # proportion de Z_i=(0,0,m_i) dans les données

# mise en forme des données (ordre des variables, création des dummies)

h1=model.matrix(~0+factor(health))[,1] # 1 si poor
h2=model.matrix(~0+factor(health))[,2] # 1 si average
```

```
f=model.matrix(~0+factor(gender))[,1] # 1 si female
sta=model.matrix(~0+factor(married))[,2] # 1 if married
priv=model.matrix(~0+factor(privins))[,2] # 1 if yes
med=model.matrix(~0+factor(medicaid))[,2] # 1 if yes

# données définitives

d=cbind(ofnp,opnp,ofp,h1,h2,age,f,sta,faminc,priv,med)
d
dev.off()

library(plot3D)
z=table(ofnp,opnp)
hist3D(z=z, border="black",theta=45,phi=45)

#########################################################################
#                                  #
#                               estimation du modèle (pi fixe) #                                  #
#########################################################################

library(maxLik)
library(Matrix)

#########################################################
#
# programmation des fonctions d'estimation
#
# avec zéro-inflation
#
loglikfun=function(param){
  p=param[1:length(p)]
  b1=param[(length(p)+1):(length(b1)+length(p))]
  b2=param[(length(b1)+length(p)+1):length(phi)]
  #
  sum((Y[1,]+Y[2,]==0)*log(p+(1-p)*(1/(1+exp(t(b1)%*%X)+exp(t(b2)%*%X))^mi)))+
```

```
    sum((Y[1,]+Y[2,]!=0)*(log(1-p)+Y[1,]*t(b1)%*%X+Y[2,]*t(b2)%*%X-mi*
log(1+exp(t(b1)%*%X)+exp(t(b2)%*%X))))
}
#
# sans zéro-inflation
#
loglikfun2=function(param){
  b1=param[1:length(b1)]
  b2=param[(length(b1)+1):(2*length(b1))]
  #
  sum(Y[1,]*t(b1)%*%X+Y[2,]*t(b2)%*%X-mi*log(1+exp(t(b1)%*%X)+exp(t(b2)%*%X)))
}
#
# fin programmation des fonctions d'estimation
#
###########################################################
#
# préparation des données et initialisations
#
mi=ofnp+opnp+ofp

Y=rbind(ofnp,opnp,mi)
inter=rep(1,length(mi))
X=rbind(inter,h1,h2,numchron,age,f,sta,school,faminc,med)

# rankMatrix(X)

# X=rbind(inter,age,f,sta,faminc)

b1=rep(1,dim(X)[1])
b2=rep(1,dim(X)[1])
p=0.5          # pi
phi=c(p,b1,b2)
#
# fin préparation des données et initialisations
```

```
#
#########################################################
#
# exécution des fonctions
#
mle=maxLik(logLik=loglikfun,start=c(rep(0.25,1+2*dim(X)[1]))) # avec 0-inflation
mle2=maxLik(logLik=loglikfun2,start=c(rep(0.25,2*dim(X)[1]))) # sans 0-inflation
#
# fin exécution des fonctions
#
#########################################################
#
# récupération et affichage des résultats
#
est.cov=diag(vcov(mle))
est.cov2=diag(vcov(mle2))
#
cat("proportion de 0-inflation=",mle$est[1]) # proportion de 0-inflation
#
#
cat("beta_1 avec 0-infl=",mle$est[2:(1+dim(X)[1])])
#
cat("beta_1 sans 0-infl=",mle2$est[1:dim(X)[1]])
#
#
#
cat("beta_2 avec 0-infl=",mle$est[(2+dim(X)[1]):(1+2*dim(X)[1])])
#
cat("beta_2 sans 0-infl=",mle2$est[(1+dim(X)[1]):(2*dim(X)[1])])
#
#
#
cat("beta_1 stand avec 0-infl=",mle$est[2:(1+dim(X)[1])]/sqrt(est.cov)[2:(1+dim(X)[1])])
#
cat("beta_1 stand sans 0-infl=",mle2$est[1:dim(X)[1]]/sqrt(est.cov2)[1:dim(X)[1]])
```

```
#
#
#
cat("beta_2 stand avec 0-infl=",mle$est[(2+dim(X)[1]):(1+2*dim(X)[1])]/sqrt(est.cov)
[(2+dim(X)[1]):(1+2*dim(X)[1])])
#
cat("beta_2 stand sans 0-infl=",mle2$est[(1+dim(X)[1]):(2*dim(X)[1])]/sqrt(est.cov2)
[(1+dim(X)[1]):(2*dim(X)[1])])

#
#############################################################################
#                                  #
#                       estimation du modèle (pi fonction des covariables)  #
#                                  #
#############################################################################
#
# programmation des fonctions d'estimation
#
# avec zéro-inflation
#
loglikfun=function(param){
  g=param[1:length(g)]
  b1=param[(length(g)+1):(length(b1)+length(g))]
  b2=param[(length(b1)+length(g)+1):length(phi)]
  #
  sum((Y[1,]+Y[2,]==0)*log(exp(t(g)%*%W)+(1/(1+exp(t(b1)%*%X)+exp(t(b2)%*%X))^mi)))+
    sum((Y[1,]+Y[2,]!=0)*(Y[1,]*t(b1)%*%X+Y[2,]*t(b2)%*%X-mi*log(1+exp(t(b1)%*%X)
+exp(t(b2)%*%X))))-sum(log(1+exp(t(g)%*%W)))
}
#
# fin programmation des fonctions d'estimation
#
#########################################################
#
# préparation des données et initialisations
```

```
#
mi=ofnp+opnp+ofp

Y=rbind(ofnp,opnp,mi)
inter=rep(1,length(mi))
X=rbind(inter,h1,h2,numchron,age,f,sta,school,faminc,med)
# W=rbind(inter,h1,h2,numchron,age,f,sta,school,faminc)
#W=rbind(inter,numchron,age,f,school,med)
W=rbind(inter,age,f,school,med)
# rankMatrix(X)

# X=rbind(inter,age,f,sta,faminc)

b1=rep(1,dim(X)[1])
b2=rep(1,dim(X)[1])
g=rep(1,dim(W)[1])          # pi
phi=c(g,b1,b2)
#
# fin préparation des données et initialisations
#
#########################################################
#
# exécution des fonctions
#
mle3=maxLik(logLik=loglikfun,start=c(rep(0.25,length(g)+2*dim(X)[1])))# avec 0-inflation
#
# fin exécution des fonctions
#
#########################################################
#
# récupération et affichage des résultats
#
est.cov3=diag(vcov(mle3))
#
cat("gamma=",mle3$est[1:dim(W)[1]]) # gamma (probabilité de zéro-inflation)
```

```
#
cat("beta_1 avec 0-infl et cov=",mle3$est[(dim(W)[1]+1):(dim(W)[1]+dim(X)[1])])
#
cat("beta_2 avec 0-infl et cov=",mle3$est[(dim(W)[1]+dim(X)[1]+1):(dim(W)[1]+2*dim(X)[1])])
#
cat("beta_1 stand avec 0-infl et cov=",mle3$est[(dim(W)[1]+1):(dim(W)[1]+dim(X)[1])]
/sqrt(est.cov3)[(dim(W)[1]+1):(dim(W)[1]+dim(X)[1])])
#
cat("beta_2 stand avec 0-infl et cov=",mle3$est[(dim(W)[1]+dim(X)[1]+1):(dim(W)[1]+
2*dim(X)[1])]/sqrt(est.cov3)[(dim(W)[1]+dim(X)[1]+1):(dim(W)[1]+2*dim(X)[1])])
#
#########################################################
#
#     logl-lik des trois modèles
#
cat("loglik modele sans 0-infl=",mle2$maximum)
#
cat("loglik modele avec 0-infl sans cov=",mle$maximum)
#
cat("loglik modele avec 0-infl avec cov=",mle3$maximum)
#
#########################################################
#
#     AIC des trois modèles (à minimiser)
#
cat("AIC modele sans 0-infl=",2*length(mle2$estimate)-2*mle2$maximum)
#
cat("AIC modele avec 0-infl sans cov=",2*length(mle$estimate)-2*mle$maximum)
#
cat("AIC modele avec 0-infl avec cov=",2*length(mle3$estimate)-2*mle3$maximum)
#
#
#########################################################
#
#     p-value (en vrac)
```

```
#
#
round(2*(1-pnorm(abs(mle3$est[1:dim(W)[1]]/sqrt(est.cov3)[1:5]))),4) # gamma

round(2*(1-pnorm(abs(mle2$est[1:dim(X)[1]]/sqrt(est.cov2)[1:dim(X)[1]]))),4)
round(2*(1-pnorm(abs(mle2$est[(1+dim(X)[1]):(2*dim(X)[1])]/sqrt(est.cov2)
[(1+dim(X)[1]):(2*dim(X)[1])]))),4)   # beta_2 sans 0-infl

round(2*(1-pnorm(abs(mle$est[2:(1+dim(X)[1])]/sqrt(est.cov)[2:(1+dim(X)[1])]))),4)
round(2*(1-pnorm(abs(mle$est[(2+dim(X)[1]):(1+2*dim(X)[1])]/sqrt(est.cov)[(2+dim(X)[1])
:(1+2*dim(X)[1])]))),4) # beta_2 avec 0-infl

round(2*(1-pnorm(abs(mle3$est[(dim(W)[1]+1):(dim(W)[1]+dim(X)[1])]/sqrt(est.cov3)
[(dim(W)[1]+1):(dim(W)[1]+dim(X)[1])]))),4) # beta_1 avec 0-infl et cov
round(2*(1-pnorm(abs(mle3$est[(dim(W)[1]+dim(X)[1]+1):(dim(W)[1]+2*dim(X)[1])]
/sqrt(est.cov3)[(dim(W)[1]+dim(X)[1]+1):(dim(W)[1]+2*dim(X)[1])]))),4) # beta_2 avec 0-infl et



################################################################################
#                                 #
#                        sélection des covariables dans le modèle de 0-inflation #
#                                 #
################################################################################

izi=as.integer(ofnp+opnp==0)
fitzi=glm(izi~h1+h2+numchron+age+f+sta+school+faminc+med,family=binomial(link=logit))
summary(fitzi)
```

Bibliographie

A. Agresti. *Categorical data analysis*. John Wiley & Sons, Inc., 2002. 28, 30

H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, AC-(4) :716–723, 1974. 17

A. Antoniadis, J. Berruyer, and R. Carmona. Régression non linéaire et applications. *Economica, Paris*, 1992. 6, 15

B. E. Bagozzi. The baseline-inflated multinomial logit model for international relations research. *Conflict Management and Peace Science*, doi 10.1177/0738894215570422 (to appear), 2015. 67

W.R. Blischke. Mixtures of distributions. *Intertional Encyclopedia of Statistics*, Vol.1, W.H. Kruskal and J.M. Tanur (Eds.). New York : The Free Press :174–180, 1978. 29

S. Boes. Count data models with correlated unobserved heterogeneity. *Scandinavian Journal of Statistics*, 37 :382–402, 2010. 1

R. J. Carroll, D. Ruppert, and L. A. Stefanski. *Measurement error in nonlinear models*. Chapman and Hall, New York, 1995. 26

X.-D. Chen and Y.-Z. Fu. Model selection for zero-inflated regression with missing covariates. *Computational Statistics & Data Analysis*, 55(1) :765–773, 2011. 104

D.R. Cox. Some remarks on overdispersion. *Biometrika Trust*, 70(1) :269–274, 1983. 22

D.R. Cox and E.J. Snell. Analysis of binary data. *Chapman and Hall New York*, 2, 1989. 1

C. Czado, V. Erhardt, A. Min, and S. Wagner. Zero-inflated generalized poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. *Statist. Model.*, 7(2) :125–153, 2007. 24, 26, 27

P. Deb and P. K. Trivedi. Demand for medical care by the elderly : a finite mixture approach. *Journal of Applied Econometrics*, 12(3) :313–336, 1997. 53, 55, 65, 83, 85, 86

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B*, 39 :1–38, 1977. 25

A. O. Diallo, A. Diop, and J. F. Dupuy. 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

E. Dietz and D. Böhning. On estimation of the poisson parameter in zero-modified poisson models. *Comput. Statist. Data Anal.*, 34 :441–459, 2000. 35, 66

E. Dietz and D. Böhning. On estimation of the poisson parameter in zero-modified poisson models. *Computational Statistics & Data Analysis*, 34(4) :441–459, 2000. 103

A. Diop, A. Diop, and J.-F. Dupuy. Maximum likelihood estimation in the logistic regression model with a cure fraction. *Electronic J. Statist.*, 5 :460–483, 2011. 21, 24, 84, 105

A. Diop, A. Diop, and J.-F. Dupuy. Simulation-based inference in a zero-inflated bernoulli regression model. *Communications in Statistics - Simulation and Computation*, 45(10) :3597–3614, 2016. 67, 104

A.J. Dobson. *An introduction to generalized linear models*. Chapman and Hall, Londo., 1990. 16

F. Eicker. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Ann. Math. Statist.*, 34 :447–456, 1963. 47, 96

F. Eicker. A multivariate central limit theorem for random linear vector forms. *Ann. Math. Statist.*, 37 :1825–1828, 1966. 115

L. Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear model. *Ann. Statist.*, 13 :342–368, 1985. 15

J. Feng and Z. Zhu. Semiparametric analysis of longitudinal zero-inflated count data. *Journal of Multivariate Analysis*, 102 :61–72, 2011. 104

R.A. Fisher. The negative binomial distribution. *Annals od Eugenics*, 11(1) :182–187, 1941. 1

D.Y.T. Fong and Yip. A note on information loss in analysing a mixture model of count data. *Comm. Statist. Theory Methods*, 24 :3197–3209, 1995. 21

R. V. Foutz. On the unique consistent solution to the likelihood equations. *Journal of the American Statistical Association*, 72 :147–148, 1977. 109, 114

E.L. Frome, M.H. Kutner, and J.J. Beauchamp. Regression analysis of poisson distributed data. *Journal of the American Statistical Association*, 63(244 :935–940, 1973. 1

A. M. Garay, E. M. Hashimoto, E. M. M. Ortega, and V. H. Lachos. On estimation and influence diagnostics for zero-inflated negative binomial regression models. *Computational Statistics & Data Analysis*, 55(3) :1304–1318, 2011. 35, 67, 104

M. S. Gilthorpe, M. Frydenberg, Y. Cheng, and V. Baelum. Modelling count data with excessive zeros : The need for class prediction in zero-inflated models and the issue of data generation in choosing between zero-inflated and generic mixture models for dental caries data. *Statistics in Medicine*, 28 :3539–3553, 2009. 35, 104, 123

E. Gonzalez-Estrada and J. A. Villasenor-Alva. mvshapirotest : Generalized shapiro-wilk test for multivariate normality. r package version 1.0. *https ://CRAN.R-project.org/package=mvShapiroTest*, 2013. 88

G. Gouriéroux and A. Monfort. Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. *J. Econom.*, 17 :83–97, 1981. 40, 45, 71, 95, 107

W. Greene. Accounting for excess zeros and sample selection in poisson and negative binomial regression models. *Working Paper, Dep of Eco, New York University*, 1994. 21

D.H. Griffith and R. Haining. Beyond mule kicks the poisson distribution in geographical analysis. *Geographical Analysis*, 38 :123–139, 2006. 1

S. Gurmu and J. Elder. Generalized bivariate count data regression models. *Economics Letters*, 68 : 31–36, 2000. 65

D. Hall. Zero-inflated poisson and binomial regression with random effects : a case study. *Biometrics*, 56 (4) :1030–1039, 2000. 28, 29, 34, 35, 36, 37, 49, 57, 67, 103, 104, 106, 123

D. B. Hall and K. S. Berenhaut. Score tests for heterogeneity and overdispersion in zero-inflated poisson and binomial regression models. *The Canadian Journal of Statistics*, 30(3) :415–430, 2002. 35

D.B. Hall and J. Shen. Robust estimation for zero-inflated poisson regression. *Scand. J. Statist.*, 37 :237–252, 2010. 25, 26, 27

X. He, H. Xue, and N.Z. Shi. Sieve maximum likelihood estimation for doubly semiparametric zero-inflated poisson models. *J. Multivar. Anal.*, 101(9) :2026–2038, 2010. 27, 57

D.C. Heilbron. Zero-alterned and other regression models for count data with added zeros. *Biometrical Journal*, 36 :531–547, 1994. 21

A. Henningsen and O. Toomet. maxlik : A package for maximum likelihood estimation in R. *Computational Statistics*, 26 :443–458, 2011. 49, 69, 73, 117

J.M. Hilbe. *Negative Binomial Regression*. Cambridge University Press, 2007. 19, 22, 27

J. Hinde and C. G. B. Demétrio. Overdispersion : models and estimation. *Computational Statistics & Data Analysis*, 27 :151–170, 1998. 1, 22

D.G. Horvitz and D.J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47 :663–685, 1952. 104, 106

D.W. Hosmer and S. Lemeshow. *Applied logistic regression*. Editions Wiley, 2000. 17

S. H. Hsieh, S. M. Lee, and P. S. Shen. Logistic regression analysis of randomized response data with missing covariates. *Journal of Statistical Planning and Inference*, 140(4) :927–940, 1952. 104

R.I. Jennrich and P.F. Sampson. Newton-raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics*, 18 :11–17, 1976. 11

J. Jiang. Large sample techniques for statistics. *Springer, New York*, 2010. 45, 112

M. E. Kelley and S. J. Anderson. Zero inflation in ordinal data : incorporating susceptibility to response through the use of a mixture model. *Statist. Med.*, 27 :3674–3688, 2008. 30, 67, 73

C.D. Kemp and A.W. Kemp. Rapid estimation for discrete distributions. *The Statistician*, 37 : 243–255, 1988. 28

K.F. Lam, H. Xue, and Y.B. Cheung. Semiparametric analysis of zero-inflated count data. *Biometrics*, 62 :996–1003, 2006. 27, 57, 104

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

C.-S. Li. A lack-of-fit test for parametric zero-inflated poisson models. *Journal of Statistical Computation and Simulation*, 81(9) :1081–1098, 2011. 35, 66

T. Li and Y. Hu. Inverse probability weighted estimators for single-index models with missing covariates. *Communications in Statistics. Theory and Methods*, 45(5) :1199–1214, 2010. 104

H. K. Lim, W. K. Li, and P. L. H. Yu. Zero-inflated poisson regression mixture model. *Computational Statistics & Data Analysis*, 71 :151–158, 2014. 35, 66, 103

T. M. Lukusa, S.-M. Lee, and C.-S. Li. Semiparametric estimation of a zero-inflated Poisson regression model with missing covariates. *Metrika*, 79(4) :457–483, 2016. 21, 104

B.H. Margolin, B.S. Kim, and K.J. Risko. The ames salmonella/microsome mutagenicity assay : issues of inference and validation. *J. Amer. Statist. Assoc.*, 84 :651–661, 1989. 29

D. Matranga, A. Firenze, and A. Vullo. Can bayesian models play a role in dental caries epidemiology ? evidence from an application to the belcap data set. *Community Dentistry and Oral Epidemiology*, 41(5) :473–480, 2013. 35, 104, 123

P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, chapman and hall, london edition, 1989. 6, 16

A. Min and C. Czado. Testing for zero-modification in count regression models. *Statistica Sinica*, 20(1) :323–341, 2010. 35

Y. Min and A. Agresti. Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, 5(1), 2005. 103

A. Moghimbeigi, M. R. Eshraghian, K. Mohammad, and B. McArdle. Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. *Journal of Applied Statistics*, 35(9) :1193–1202, 2008. 35, 66, 104

A. Monod. Random effects modeling and the zero-inflated poisson distribution. *Communications in Statistics. Theory and Methods*, 43(4) :664–680, 2014. 35, 66, 103

J. Mullahy. Specification and testing of some modified count data models. *Journal of Econometrics*, 33 :341–365, 1986. 21

J. Mullahy. Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics*, 12(3) :337–350, 1997. 21

S. M. Mwalili, E. Lesaffre, and D. Declerck. The zero-inflated negative binomial regression model with correction for misclassification : an example in caries research. *Statistical Methods in Medical Research*, 17(2) :123–139, 2014. 27, 35, 67, 104

J.A. Nelder and R. W. M. Wedderburn. Generalized linear models. *J. Roy. Statist. Soc. Ser. A*, 135 : 370–384, 1972. 1, 6

M. Nerlove and S.J. Press. Univariate and multivariate log-linear and logistic models. *Rand corporation, Santa Monica*, 1973. 1

N.C. Pradhan and P. Leung. A poisson and negative binomial regression model of sea turtle interactions in hawaii's longline fishery. *Fish. Res.*, 78 :309–322, 2006. 24

J. S. Preisser, D. L. Long, and M. E. Kincade. Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. *Caries Research*, 54(4) : 413–423, 2012. 21

L. Qi, C. Y. Wang, and R. L. Prentice. Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, 100(472) :1250–1263, 2005. 104

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, 1998. 21

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

O. Rosen, W. X. Jiang, and M. A. Tanner. Mixtures of marginal models. *Biometrika*, 87 :391–404, 2000. 26

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3) :581–592, 1976. 106

N. Sarma. Physical inactivity and its impact on healthcare utilization. *Health Economics*, 18(8) : 885–901, 2009. 65

S. Sarma and W. Simpson. A microeconometric analysis of canadian health care utilization. *Health Economics*, 15(3) :219–239, 2006. 65

G. Schwarz. Estimating the dimension of a model. *Annals of statistics*, 6 :461–464, 1978. 17

K. E. Staub and R. Winkelmann. Consistent estimation of zero-inflated count models. *Health Economics*, 22(6) :673–686, 2013. 65, 73, 84

R Core Team. R foundation for statistical computing. *Vienna, Austria http ://www.R-project.org/*, 2013. 49

H. Teicher. On the mixture of distributions. *Ann. Math. Statist.*, 31 :55–73, 1960. 29

H. Teicher. Identifiability of finite mixtures. *Ann. Math. Statist.*, 34 :1265–1269, 1963. 29

W. Tu. Zero-inflated data. *Encyclopedia of Environmetrics*, 4 :2387–2391, 2002. 21

A. M. C. Vieira, J. P. Hinde, and C. G. B. Demetrio. Zero-inflated proportion data models applied to a biological control assay. *Journal of Applied Statistics*, 27(3) :373–389, 2000. 28, 35, 67, 104

J. A. Villasenor-Alva and E. Gonzalez-Estrada. A generalization of shapiro-wilk's test for multivariate normality. *Communications in Statistics : Theory and Methods*, 38(11) :31870–1883, 2000. 88

P. Wang. A bivariate zero-inflated negative binomial regression model for count data with excess zeros. *Economics Letters*, 78 :373–378, 2003. 65, 73

L. P. Zhao and S. Lipsitz. Designs and analysis of two-stage studies. *Statistics in Medicine*, 11(6) : 769–782, 1992. 104

W. Zhao, R. Zhang, J. Liu, and Y. Lv. Semi varying coefficient zero-inflated generalized poisson regression model. *Communications in Statistics. Theory and Methods*, 44(1) :171–185, 2015. 104

INSA de RENNES
Service des Formations

AVIS DU JURY SUR LA REPRODUCTION DE LA THESE SOUTENUE

Titre de la thèse:
Inférence statistique dans des modèles de comptage à inflation de zéros. Applications en économie de la santé

Nom Prénom de l'auteur : DIALLO ALPHA OUMAR

Membres du jury :
- Monsieur KOKONENDJI Célestin
- Monsieur DIOP Aliou
- Monsieur DUPUY Jean-François
- Madame MONBET Valérie
- Madame YAO Anne-Françoise
- Monsieur RAULT Christophe

Président du jury : Monsieur RAULT Christophe

Date de la soutenance : 27 Novembre 2017

Reproduction de la these soutenue

☒ Thèse pouvant être reproduite en l'état
☐ Thèse pouvant être reproduite après corrections suggérées

Fait à Rennes, le 27 Novembre 2017

Signature du président de jury

Le Directeur,

M'hamed DRISSI

Résumé

Les modèles de régressions à inflation de zéros constituent un outil très puissant pour l'analyse de données de comptage avec excès de zéros, émanant de divers domaines tels que l'épidémiologie, l'économie de la santé ou encore l'écologie.
Cependant, l'étude théorique dans ces modèles attire encore peu d'attention. Ce manuscrit s'intéresse au problème de l'inférence dans des modèles de comptage à inflation de zéro.

Dans un premier temps, nous revenons sur la question de l'estimateur du maximum de vraisemblance dans le modèle binomial à inflation de zéro. D'abord nous montrons l'existence de l'estimateur du maximum de vraisemblance des paramètres dans ce modèle. Ensuite, nous démontrons la consistance de cet estimateur, et nous établissons sa normalité asymptotique. Puis, une étude de simulation exhaustive sur des tailles finies d'échantillons est menée pour évaluer la cohérence de nos résultats. Et pour finir, une application sur des données réelles d'économie de la santé a été conduite.

Dans un deuxième temps, nous proposons un nouveau modèle statistique d'analyse de la consommation de soins médicaux. Ce modèle permet, entre autres, d'identifier les causes du non-recours aux soins médicaux. Nous avons étudié rigoureusement les propriétés mathématiques du modèle. Ensuite nous avons mené une étude numérique approfondie à l'aide de simulations informatiques et enfin, nous l'avons appliqué à l'analyse d'une base de données recensant la consommation de soins de plusieurs milliers de patients aux USA.

Un dernier aspect de ces travaux de thèse a été de s'intéresser au problème de l'inférence dans le modèle binomial à inflation de zéro dans un contexte de données manquantes sur les covariables. Dans ce cas nous proposons la méthode de
pondération par l'inverse des probabilités de sélection pour estimer les paramètres du modèle. Ensuite, nous établissons la consistance et la normalité asymptotique de l'estimateur proposé. Enfin, une étude de simulation sur plusieurs échantillons de tailles finies est conduite pour évaluer le comportement de l'estimateur.

Mots clés
Normalité asymptotique, consistance, données de comptage, excès de zéros, simulations, utilisation de soins de santé, logit multinomial, pondération par l'inverse de la probabilité de sélection.

Abstract

The zero-inflated regression models are a very powerful tool for the analysis of counting data with excess zeros from various areas such as epidemiology, health economics or ecology. However, the theoretical study in these models attracts little
attention. This manuscript is interested in the problem of inference in zero-inflated count models.

At first, we return to the question of the maximum likelihood estimator in the zero-inflated binomial model. First we show the existence of the maximum likelihood estimator of the parameters in this model. Then, we demonstrate the consistency of this estimator, and let us establish its asymptotic normality. Then, a comprehensive simulation study finite sample sizes are conducted to evaluate the consistency of our results. Finally, an application on real health economics data has been conduct.

In a second time, we propose a new statistical analysis model of the consumption of medical care. This model allows, among other things, to identify the causes of the non-use of medical care. We have studied rigorously the mathematical properties
of the model. Then, we carried out an exhaustive numerical study using computer simulations and finally applied to the analysis of a database on health care several thousand patients in the USA.

A final aspect of this work was to focus on the problem of inference in the zero inflation binomial model in the context of missing covariate data. In this case we propose the weighting method by the inverse of the selection probabilities to estimate the parameters of the model. Then, we establish the consistency and asymptotic normality of the estimator offers. Finally, a simulation study on several samples of finite sizes is conducted to evaluate the behavior of the estimator.

Key words
Asymptotic normality, consistency, count data, excess of zeros, simulations, healthcare utilization, multinomial logit, inverse-probability-weighting.

INSA RENNES

N° d'ordre : 17ISAR 31 / D17 - 31
Institut National des Sciences Appliquées de Rennes
20, Avenue des Buttes de Coëmes • CS 70839 • F-35708 Rennes Cedex 7
Tel : 02 23 23 82 00 - Fax : 02 23 23 83 96