RÉPUBLIQUE DE CÔTE D'IVOIRE
Union-Discipline-Travail

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

Institut National Polytechnique
Félix HOUPHOUËT-BOIGNY

EDP
École Doctorale Polytechnique

THÈSE

Pour l'obtention du grade de
Docteur de l'Institut National Polytechnique Félix HOUPHOUËT-BOIGNY

Mention : **Mathématiques Appliquées**

Spécialité : **Statistique**

SUJET :

INFÉRENCE STATISTIQUE DANS LE MODÈLE DE POISSON BIVARIÉ À INFLATION DE ZÉROS

Présentée par **KOUAKOU Konan Jean Geoffroy**

Numéro d'ordre de la thèse : 127/2022

Soutenue publiquement le 07 Octobre 2022 devant le jury composé de **:**

M. OUATTARA Sié	Professeur Titulaire, INP-HB	Président
M. HILI Ouagnina	Professeur Titulaire, INP-HB	Directeur de thèse
M. DIONGUE Abdou Kâ	Professeur Titulaire, UGB	Rapporteur
M. MONSAN Vincent	Maître de Conférences, UFHB	Rapporteur
M. YODE Armel	Maître de Conférences, UFHB	Examinateur

Dédicaces

À mes parents,
ceux qui m'ont fait naître dans ce monde,
et ceux qui m'ont aidé à y grandir.

A cœur vaillant rien d'impossible
A conscience tranquille tout est accessible
Quand il y a la soif d'apprendre
Tout vient à point à qui sait attendre
Quand il y a le souci de réaliser un dessein
Tout devient facile pour arriver à nos fins
Malgré les obstacles qui s'opposent
En dépit des difficultés qui s'interposent
Les études sont avant tout notre unique et seul atout
Elles représentent la lumière de notre existence
L'étoile brillante de notre réjouissance
Espérant des lendemains épiques
Un avenir glorieux et magique.

Remerciements

Au terme de ce passionnant projet, je saisis l'occasion qui m'est donnée, pour remercier toutes les personnes ayant contribué à sa réalisation.

Mes vifs remerciements s'adressent en premier lieu à mon directeur de thèse, M. HILI Ouagnina, Professeur Titulaire à l'Institut National Polytechnique Félix Houphouët-Boigny de Yamoussoukro. Il est difficile de résumer en quelques mots ces années de travail, mais je vous remercie sincèrement pour votre écoute, votre patience, votre disponibilité, votre soutien et la confiance que vous m'avez accordée. Merci de m'avoir encouragé, laissé expérimenter et guidé. J'ai énormément progressé et appris grâce à vous et j'espère être un jour aussi expérimenté et clairvoyant que vous. Merci pour TOUT.

Je tiens à remercier tout particulièrement M. DUPUY Jean-François, Professeur Titulaire à l'Institut de Recherche des Sciences Appliquées (INSA) de Rennes en France pour sa précieuse contribution à la réalisation de ce projet de thèse. Vous m'avez tellement appris. Je ne trouve pas les mots adéquats pour vous remercier. Sans vous Professeur, ce travail n'aurait pas abouti. Merci pour votre écoute et gentillesse. J'espère un jour être aussi qualifié que vous. Ce fut un réel plaisir de travailler avec vous et je souhaite vivement que cette collaboration continuera.

Je voudrais exprimer ma gratitude à Monsieur OUATTARA Sié, Professeur Titulaire à l'Institut National Polytechnique Félix Houphouët-Boigny de Yamoussoukro. pour l'intérêt qu'il a bien voulu apporter à ce travail et avoir accepté de présider mon jury de soutenance. Que vous soyez assuré de mon entière reconnaissance pour avoir été mon professeur.

Je tiens tout particulièrement à témoigner ma reconnaissance à M. DIONGUE Abdou Kâ, Professeur Titulaire à l'Université Gaston Berger de Saint-Louis, qui m'a

fait le plaisir de rapporter ma thèse. Il a montré un grand intérêt pour ma thèse. Sa connaissance approfondie des techniques statistiques constitue un atout indéniable pour juger ce travail.

Je suis très honoré que M. MONSAN Vincent, Maître de Conférences à l'Université Félix Houphouët Boigny de Cocody ait accepté de rapporter cette thèse. Merci pour le temps, l'attention et les suggestions pertinentes que vous apportez à ce travail.

Mes remerciements vont également à M. YODÉ Armel, Maître de Conférences à l'Université Félix Houphouët Boigny de Cocody pour avoir accepté de faire partie de mon jury de thèse en tant qu'examinateur de cette thèse. Votre savoir-faire et savoir-être nous ont subjugués au point de nous transmettre la passion de la Statistique.

Ce travail de thèse s'est principalement déroulé au sein de l'UMRI (Unité Mixte de Recherche et d'Innovation) MNTI (Mathématiques et Nouvelles Technologies de l'Information) de l'EDP-INPHB. Je remercie tous ses doctorants, en particulier ceux de l'équipe Statistique, Probabilité et Recherche Opérationnelle. Merci pour votre sympathie durant ces années de dur labeur. Je remercie également tous les enseignants du Département MI (Mathématiques & Informatiques) de l'INPHB, en particulier, son Directeur M. Safidine.

Je ne saurais achever cette partie sans remercier ma famille, mes parents, mes ami(es), mes frères et sœurs. Vous avez su m'accompagner, me soutenir, m'encourager et me stimuler depuis le début de ma scolarisation jusqu'à maintenant, en vous privant parfois de ma présence durant toutes ces années d'études et de recherche.

Table des matières

Résumé

La modélisation conjointe de deux ou plusieurs données de comptage a fait l'objet d'une attention particulière ces dernières années au sein de la communauté scientifique. En effet, les données de comptage multivariées se présentent dans un contexte très large. Particulièrement, les modèles de comptage à deux variables sont utilisés dans les cas où deux variables de comptage sont corrélées et doivent être estimées conjointement. Les modèles de régression de Poisson bivarié à inflation de zéros sont le plus largement utilisés pour les données de comptages bivariées qui ont un nombre important de $(0; 0)$. Cependant, les propriétés théoriques dans ces modèles n'ont pas encore été assurées. C'est dans ce cadre que s'inscrit ce travail qui a pour objectif de combler ce déficit et de fournir à cet effet, une base rigoureuse pour l'application du modèle.

Dans la première partie, nous présentons quelques notions utiles à la compréhension de ce manuscrit. Pour ce faire, nous rappelons des notions sur les modèles linéaires généralisés (formalisme des modèles, construction d'estimateurs et de tests, aspects numériques). Ensuite, nous énonçons quelques modèles à inflation de zéros, les méthodes d'estimations puis les propriétés asymptotiques qui sont le plus souvent rencontrés dans la littérature.

Dans la deuxième partie, nous nous penchons singulièrement sur le modèle de régression de Poisson bivarié à inflation de zéros. Nous étudions d'abord, les propriétés asymptotiques de son estimateur du maximum de vraisemblance. Ensuite, nous menons une étude de simulations sur plusieurs échantillons de tailles finies pour évaluer les performances de l'estimateur proposé. Enfin, nous proposons une application de ce modèle pour évaluer la demande et le renoncement aux soins médicaux de plusieurs milliers de patients aux USA.

Dans la dernière partie de notre travail, nous nous intéressons à un problème fréquemment rencontré dans la pratique. Plus précisément, au problème de l'inférence

dans un contexte où les covariables qui interviennent dans le modèle de régression de Poisson bivarié à inflation de zéros sont partiellement observées. À cet effet, nous proposons des méthodes de pondération par l'inverse des probabilités de sélection et d'imputation multiple pour estimer les paramètres de notre modèle lorsque des données sont manquantes sur des covariables. En outre, nous établissons les propriétés asymptotiques des estimateurs proposés. Nous réalisons une étude de simulation exhaustive sur des tailles finies d'échantillons afin d'évaluer la cohérence de nos résultats. Pour finir, nous présentons une application des méthodes proposées sur des données d'économie de la santé.

Mots clés : Données de comptage, propriétés asymptotiques, imputation multiple, estimateurs par pondération, non paramétrique, inflation de zéros, manquant de manière aléatoire.

Abstract

The joint modeling of two or more count data has received a lot of attention in recent years in the scientific community. Indeed, multivariate count data occur in a very broad context. In particular, bivariate count models are used in cases where two count variables are correlated and need to be estimated jointly. Bivariate Poisson regression models with zero inflation are most widely used for bivariate count data that have a large number of $(0;0)$. However, the theoretical properties in these models have not yet been assured. It is in this context that this work is carried out, with the aim of filling this gap and providing a rigorous basis for the application of the model.

First, we present some notions useful for the understanding of this manuscript. To do so, we recall some notions on generalized linear models (formalism of models, construction of estimators and tests, numerical aspects). Then, we describe some models with inflation of zeros, the estimation methods and the asymptotic properties which are most often met in the literature.

In the second part, we focus on the zero-inflated bivariate Poisson regression model. First, we study the asymptotic properties of its maximum likelihood estimator. Then, we conduct a simulation study on several finite sample sizes to evaluate the performance of the proposed estimator. Finally, we propose an application of this model to evaluate the demand and the renunciation of medical care of several thousand patients in the USA.

In the last part of our work, we are interested in a problem frequently encountered in practice. More precisely, the problem of inference in a context where the covariates involved in the bivariate Poisson regression model with zero inflation are partially observed. In this context, we propose inverse selection probability weighting and multiple imputation methods for estimating the parameters of model when data are missing on covariates. In addition, we establish the asymptotic properties of the

proposed estimators. We perform an exhaustive simulation study on finite sample sizes to assess the consistency of our results. Finally, we present an application of the proposed methods on health economics data.

Key words : Count data, asymptotic properties, multiple imputation, weighting estimators, nonparametric, zero-inflation, missing at random.

Table des figures

Liste des tableaux

Abréviations & Notations

Abréviations

$\mathcal{B}(\mathrm{p})$	: Loi Bernoulli de paramètre p.
$\mathcal{NB}(r,p)$	: Loi binomiale négative de paramètres r et p.
$\mathcal{P}(\lambda)$	: Loi de Poisson de paramètre λ.
EMV	: Estimateur du Maximum de Vraisemblance.
GLM	: Modèle Linéaire Généralisé.
GLMs	: Modèles Linéaires Généralisés.
TCL	: Théorème central limite.
ZIB	: Zero-Inflated Binomial.
ZINB	: Zero-Inflated Negative Binomial.
ZIP	: Zero-Inflated Poisson.
ZIM	: Zero-Inflated Multinomial.
ZIBP	: Zero-Inflated Bivariate Poisson.

Notations d'ordre général

$\mathbb{N}$	: 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.
$\|X\|$	: Norme euclidienne du vecteur X.
$\mathcal{M}(n \times p)$	: Ensemble des matrices réelles à n lignes et p colonnes.
I_p	: Matrice identité d'ordre p.
$a^{\otimes 2} = aa^\top$	: Pour tout vecteur colonne a.
1_A	: Fonction indicatrice de A définie par: $1_A(w)$ vaut 1 si $w \in A$ et 0 sinon.
s.e.	: Standard error.

Variables aléatoires et modes de convergence

$\mathbb{E}(X)$	: L'espérance mathématique de la variable aléatoire X.
$\text{var}(X)$	: La variance de la variable aléatoire X.
$\mathbb{E}(X\|Y)$	: Espérance conditionnelle de X sachant Y.
$\text{cov}(X,Y)$	: Covariance des variables aléatoires X et Y.
$X_n \xrightarrow{p.s} X$	: La suite de variables aléatoires $(X_n)_n$ converge presque sûrement vers X.
$X_n \xrightarrow{\mathbb{P}} X$	: La suite de variables aléatoires $(X_n)_n$ converge en probabilité vers X.
$X_n \xrightarrow{d} X$	: La suite de variables aléatoires $(X_n)_n$ converge en distribution vers X.

Notations de Landau

$a_n = O(1)$	: Signifie que a_n est reste bornée quand n tend vers l'infini.
$a_n = O(b_n)$	: Signifie que $a_n/b_n = O(1)$, c'est-à-dire tend vers 0 quand n tend vers l'infini.
$a_n = o(1)$	: Signifie que a_n tend vers 0 quand n tend vers l'infini.
$a_n = o(b_n)$	: Signifie que a_n/b_n tend vers 0 quand n tend vers l'infini.
$\mathbf{O}(a_n)$	: Désigne un vecteur colonne et une matrice dont les composantes sont uniformément $O(a_n)$.
$X_n = o_{\mathbb{P}}(1)$	: Signifie que X_n est bornée en probabilité.
$X_n = o_{\mathbb{P}}(1)$	: Signifie que X_n converge en probabilité vers 0.
$X_n = O_{\mathbb{P}}(Y_n)$	: Signifie que $X_n/Y_n = O_{\mathbb{P}}(1)$.
$X_n = o_{\mathbb{P}}(Y_n)$	: Signifie que $X_n/Y_n = o_{\mathbb{P}}(1)$.
$\mathbf{o}_{\mathbb{P}}(1)$	: Désigne un vecteur colonne et une matrice dont les composantes convergent en probabilité vers 0.

Introduction générale

Au cours de ces trois dernières décennies, la communauté scientifique s'est beaucoup intéressée à l'analyse des données de comptage. Ce grand intérêt pour la modélisation des données de comptage réside dans le fait qu'on les rencontre dans un grand nombre de domaines tels que l'économie, l'assurance, l'épidémiologie, les sciences de l'environnement, le sport, l'étude du terrorisme et bien d'autres. Cependant, ces données contiennent le plus souvent un nombre excessif de zéros, c'est-à-dire qui ne peuvent être expliqués par les modèles de comptage classiques (binomial, binomial négatif, Poisson,...). Dans les études de Sarma et Simpson (2006), Sari (2009) et Diallo *et al.* (2017), il a été montré que dans une étude économique portant sur la consommation et le renoncement aux soins médicaux, une variable réponse sujette à l'inflation de zéros peut-être le nombre de fois qu'une personne consulte un médecin dans un intervalle de temps donné.

Ridout *et al.* (1998) affirment que les zéros ont souvent un statut particulier qui peut prêter à confusion. En effet, on distingue deux types de zéros : les zéros aléatoires (ceux qui sont dûs à l'échantillonnage) et les zéros structurels (ceux qui sont dûs à la structure). À titre d'exemple, en assurance automobile un excès de zéros peut affecter le nombre d'accidents responsables déclarés par les assurés d'un portefeuille. En effet, l'observation d'un zéro qui signifie qu'aucun sinistre responsable n'est déclaré dans l'année peut correspondre à deux situations. Premièrement, l'assuré considéré n'a eu aucun sinistre. Deuxièmement, l'assuré a été responsable d'un sinistre mais ne l'a pas déclaré. La non déclaration d'un sinistre responsable peut être dûe à l'existence du système de bonus-malus, dans la mesure où un assuré va préférer supporter le coût des réparations plutôt que de perdre le bénéfice de son bonus (voir Dupuy (2018)). Ainsi, il faut tenir compte de ces deux types de zéros lors de la modélisation afin d'éviter un cas particulier de surdispersion, appelé, l'inflation de zéros (voir Lambert (1992), Mullahy (1997), Ridout *et al.* (1998), Diop *et al.* (2011) et Preisser *et al.* (2012)). Pour traiter cette difficulté, des approches ont été proposées.

Parmi celles-ci, la modélisation en deux parties "*hurdle model* de Mullahy (1986) et *two-part models* de Heilbron (1994)" ainsi qu'une autre approche qui est de considérer un mélange de deux modèles. Cette dernière approche conduit aux modèles dits *zéro-excès* communément appelé modèles à inflation de zéros. Dans son travail, Feng (2021) a montré que les modèles à obstacle ou *hurle models* ajustent mieux les données que les modèle à inflation de zéros lorsqu'il y'a déflation de zéros ou lorsque la proportion de zéros en excès est faible, voire nulle. Par contre, les modèles à inflation de zéros sont suggérés lorsque la proportion de zéros en excès est importante. En outre, les modèles à obstacle supposent qu'il n'y a qu'un seul processus par lequel un zéro peut être produit, tandis que les modèles à inflation de zéros supposent qu'il existe deux (02) processus différents qui peuvent produire un zéro. Les modèles à obstacle supposent deux (02) types de sujet : (i) ceux qui ne connaissent jamais le résultat et (ii) ceux qui connaissent toujours le résultat au moins une fois. Les modèles à inflation de zéros conceptualisent les sujets comme (i) ceux qui n'expérimentent jamais le résultat et (ii) ceux qui peuvent expérimenter le résultat mais ne le font pas toujours. Ainsi, l'interprétation des zéros diffère des modèles à obstacle et à inflation de zéros.

Plusieurs travaux ont été réalisés sur les modèles de régression univariés à inflation de zéros. Parmi ceux-ci, nous pouvons d'abord citer Lambert (1992), Li (2011), Monod (2014) et Ali (2022) qui ont évalué le modèle de régression de Poisson à inflation de zéros (modèle "ZIP" pour zero-inflated Poisson). Ensuite Ridout *et al.* (2001), Moghimbeigi *et al.* (2008) et Mwalili *et al.* (2008) se sont penchés sur le modèle de régression binomial négatif à inflation de zéros (modèle "ZINB" pour zero-inflated negative binomial). Quant à Famoye et Singh (2006), ils se sont intéressés au modèle de Poisson généralisé avec inflation de zéros (modèle "ZIGP" pour zero-inflated generalized Poisson). Enfin Hall (2000), Hall et Berenhaut (2002), Diallo *et al.* (2017) et Diallo *et al.* (2019) ont porté leur intérêt au modèle de régression binomial à inflation de zéros (modèle "ZIB" pour zero-inflated binomial). Depuis lors, plusieurs autres extensions et améliorations de ces modèles ont été réalisées par Diop *et al.* (2011), Ali *et al.* (2020) et Lee *et al.* (2021). Mais les modèles ZIP, ZINB, ZIPG et ZIB ne sont pas appropriés aux réponses bivariées. En effet, l'application de deux régressions indépendantes de comptage à des données de comptage d'événements conjoints appariés conduit à des estimateurs incohérents et inefficaces. Les événements de comptage appariés présentant une corrélation doivent être estimés conjointement; et les modèles de régression de comptage à deux variables sont conçus pour traiter de tels cas. Ainsi, quelques modèles ont été proposés pour les données de comptage bivariées à inflation de zéros. Par exemple, pour les modèles binomiaux négatifs bivariés à in-

flation de zéros (voir Wang (2003), Faroughi et Ismail (2017)) et pour les modèles de Poisson bivariés, Li *et al.* (1999), Karlis et Ntzoufras (2003), AlMuhayfith *et al.* (2016) et Yang *et al.* (2016) entre autres.

En inférence statistique, plusieurs études théoriques et numériques dans les modèles de régression univariés à inflation de zéros ont été menées. Elles reposent généralement sur la méthode du maximum de vraisemblance (voir McCullagh et Nelder (1989), Czado et Min (2005), Wang *et al.* (2021)). Des études récentes dans ce domaine ont été mené dans ce sens par Ali *et al.* (2020) et Lee *et al.* (2021). Contrairement au cas univariés, l'inférence statistique sur les modèles multivariés à inflation de zéros a pour l'instant attiré peu d'attention bien que la modélisation des données de comptage multivariées corrélées est plus que jamais d'actualité.

Cependant, dans de nombreuses circonstances, on est souvent confronté au fait qu'un ensemble de données contient des données manquantes. En effet, les données manquantes sont un problème très répandu dans de nombreuses disciplines, notamment l'économie, la sociologie, les sciences médicales, les sciences politiques, les transports et la communication, et d'autres domaines. Il existe plusieurs raisons à ce problème, comme le fait que les répondants ne répondent pas aux questions posées ou fournissent des réponses confuses, etc. Pour plus de détails sur les circonstances qui peuvent être à l'origine des données manquantes, on peut se référer à Schafer et Graham (2002).

Les données manquantes constituent une menace sérieuse pour la validité de l'inférence ou de la prise de décision dans de nombreuses applications. Compte tenu de l'intérêt de ce problème, diverses méthodes ont été proposées pour traiter les données manquantes dans les modèles de régression au cours des dernières décennies. Pour plus d'informations sur ces méthodes, nous renvoyons le lecteur intéressé à Rubin (1976), Little (1992), Zhao et Lipsitz (1992), Robins *et al.* (1994), Reilly et Pepe (1995), Clayton *et al.* (1998), Creemers *et al.* (2012), Lukusa *et al.* (2016) et Lee *et al.* (2021). Le mode de gestion le plus courant du problème de données manquantes consiste à restreindre les analyses aux sujets pour lesquels l'ensemble des variables sont entièrement renseignées (analyse dite de cas complet). Cette méthode, généralement appliquée par défaut, induit un risque potentiel de biais dans les estimations. Une alternative au cas complet en cas de données manquantes est l'Imputation Multiple (MI). L'Imputation Multiple consiste à remplacer chaque donnée manquante par un ensemble de données estimées à partir de données observées. Chacune des bases complètes ainsi obtenues fournit alors une estimation du paramètre d'intérêt, puis un estimateur unique est obtenu en calculant la moyenne de ces estimations. Une autre

méthode de traitement des données manquantes est la pondération par inverse de la probabilité de sélection (IPW, pour "Inverse Probability Weighting"). Introduite par Horvitz et Thompson (1952) puis développée par Zhao et Lipsitz (1992), la méthode IPW est basée sur la création de pseudo-populations de cas complets dans lesquelles le biais de sélection dû aux données manquantes est éliminé par des poids. La détermination de ces poids nécessite un modèle pour évaluer de la probabilité qu'un individu ait des données complètes. On peut se référer à Diallo *et al.* (2019) et Seaman et White (2013) pour plus d'information sur cette méthode.

Certaines approches ont été développées pour traiter les problèmes de covariables manquantes dans les modèles zéro-inflatés lorsque les covariables sont manquantes de manière aléatoire (MAR, pour Missing At Random). Notamment, Lukusa *et al.* (2016); Lukusa et Phoa (2020) ont proposé des méthodes d'estimation de pondération par l'inverse des probabilités de sélection (IPW) semi-paramétrique pour un modèle de régression de Poisson (ZIP) à inflation de zéros avec des covariables manquantes. Diallo *et al.* (2019) ont proposé une méthode d'estimation IPW paramétrique pour un modèle de régression binomial à inflation de zéros (ZIB) avec covariables MAR. Lee *et al.* (2020) ont proposé des méthodes d'estimation d'imputation multiple non-paramétrique pour le modèle de régression ZIP. Et plus récemment, Lee *et al.* (2021) se sont intéressés à l'estimation des paramètres dans le modèle de régression de Bernoulli à inflation de zéros (ZIBer).

Bien qu'il existe de nombreuses études sur les modèles zéros inflatés (ZI) avec covariables manquantes, ces études se sont limitées aux cas où les variables réponses sont univariées. Á notre connaissance, il n'existe pas de travaux portant sur les modèles de comptage multivariés dans un contexte de covariables manquantes. Notre travail a aussi pour but de combler cet important déficit et de fournir à cet effet, une base rigoureuse pour l'application de ces modèles.

L'objectif de ce travail est de proposer des études théoriques et numériques dans les modèles de Poisson bivarié à inflation de zéros afin d'apporter une base solide au problème de l'inférence statistique dans ces modèles.

Dans un souci de compréhension de ce travail, la thèse est structurée de la manière suivante.

Dans le chapitre 1, nous donnons des outils mathématiques qui seront nécessaires dans ce travail de thèse. Dans un premier temps, nous présentons quelques rappels essentiels sur les modèles linéaires généralisés (GLMs). D'abord, il s'agit de revenir sur la théorie des modèles linéaires généralisés. Ensuite, nous nous intéressons en particulier à la spécification d'un modèle linéaire généralisé et nous rappelons les principaux résultats concernant les propriétés de l'estimateur du maximum de vraisemblance dans ces modèles. Dans un second temps, nous présentons quelques modèles à inflation de zéros et des propriétés asymptotiques de leur estimateur obtenu par la méthode du maximum de vraisemblance. Dans un troisième temps, nous rappelons quelques notions sur les données manquantes. D'abord, nous décrivons les mécanismes de données manquantes. Nous présentons ensuite des méthodes de traitement de données manquantes.

Le chapitre 2 de ce manuscrit renferme notre première contribution dans cette thèse. D'abord, nous examinons les propriétés asymptotiques de l'estimateur du maximum de vraisemblance du modèle de Poisson bivarié à inflation de zéros. Ensuite, par le biais de la simulation, nous étudions le comportement en échantillon fini de l'estimateur proposé. Enfin, une application aux données réelles a été réalisée pour analyser les données d'utilisation des soins médicaux.

Dans le chapitre 3, nous nous intéressons à l'épineux problème de données manquantes. Nous proposons plusieurs méthodes d'estimation des paramètres du modèle de régression de Poisson bivarié avec inflation de zéros lorsque des covariables qui interviennent dans la régression sont partiellement observées. Nous examinons théoriquement ces méthodes d'estimation proposées. En outre, une étude de simulations numériques est réalisée pour évaluer la cohérence de nos résultats. Pour finir, nous appliquons les méthodes proposées à un jeu de données de l'économie de la santé.

1

Rappels sur les modèles de comptages et les données manquantes

Résumé

Dans ce chapitre, d'abord nous rappelons quelques notions essentielles sur les modèles linéaires généralisés. Ensuite, nous présentons quelques modèles à inflation de zéros et leurs propriétés asymptotiques. Enfin, nous décrivons les mécanismes de gestion des données manquantes et quelques méthodes de traitement.

Sommaire

1.1 Modèles linéaires généralisés

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

1.1.1 Présentation

Attribués à l'origine à Nelder et Wedderburn (1972) et exposés de façon complète par McCullagh et Nelder (1989), les modèles linéaires généralisés constituent une synthèse et une extension remarquables des modèles de régression familiers tels que les modèles linéaires. Les modèles linéaires généralisés permettent la modélisation de variables réponses (ou variables à expliquer) dont la loi appartient à la famille exponentielle. Ces variables réponses peuvent être de différents types : binaires (présence/absence), ordinales (très mauvais/indifférent/excellent), de comptage (nombre de sinistres au cours de l'année), ou exponentielles (durée de vie d'une lampe) par exemple. Les modèles linéaires généralisés occupent une place importante dans la modélisation statistique, car ils trouvent leur intérêt dans de nombreux domaines

d'application.

Dans la suite de cette section, nous présentons brièvement la théorie des modèles linéaires généralisés (structure, estimation, propriétés asymptotiques des estimateurs, qualité d'ajustement, tests et choix entre différents modèles). Les ouvrages de Dobson et Barnett (2018), McCullagh et Nelder (1989) et de Agresti (2015) proposent une étude plus approfondie sur les modèles linéaires généralisés, leurs méthodes d'estimation et leurs domaines d'application.

1.1.2 Structure des modèles linéaires généralisés

Un modèle linéaire généralisé (GLM) est caractérisé par trois éléments : une *composante aléatoire*, un *prédicteur linéaire* et une *fonction de lien*.

1.1.2.1 Composante aléatoire

La composante aléatoire détermine la distribution de probabilité de la variable réponse (ou sa distribution conditionnelle sachant les variables explicatives, si des variables explicatives sont présentes). Dans les modèles linéaires généralisés, on suppose que cette distribution appartient à une *famille exponentielle*, voir Nelder et Wedderburn (1972) pour plus de détails. Ainsi, on suppose que l'échantillon des observations est constitué de n variables aléatoires $Y_1, \ldots, Y_n$ indépendantes et que la densité de Y_i (par rapport à la mesure dominante appropriée : mesure de Lebesgue sur $\mathbb{R}$ ou mesure de comptage sur $\mathbb{N}$) est de la forme :

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

où $\theta_i \in \mathbb{R}$ est appelé un paramètre canonique (ou paramètre naturel) et $\phi \in \mathbb{R}_+^*$ est un paramètre de dispersion (ou de nuisance). Les fonctions b et c sont spécifiées à chaque distribution. De plus, la fonction $b(.)$ est supposée deux fois dérivable de dérivée première inversible, dérivable et d'inverse dérivable. En outre, la fonction a_i s'écrit $a_i(\phi) = \frac{\phi}{\omega_i}$ où ω_i est un poids connu associé à l'observation i (différent de 1 lorsque les données ont été groupées). Le tableau 1.1 présente quelques exemples classiques de distributions qui appartiennent à des familles exponentielles. Nous précisons les trois fonctions a, b et c ainsi que les paramètres canonique et de dispersion. Pour simplifier la lecture du tableau, nous omettons l'indice i.

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

Tableau 1.1 – Exemples de distributions de familles exponentielles.

1.1.2.2 Prédicteur linéaire

Considérons des variables explicatives organisées dans la matrice $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)^\top$ (avec $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$ pour $i = 1, \ldots, n$) appelée matrice du plan d'expérience (ou matrice *design*) d'ordre $n \times p$ où p est le nombre de variables explicatives et $(p \leq n)$. Supposons $\beta \in \mathbb{R}^p$ un vecteur de p paramètres, le prédicteur linéaire, *composante déterministe* du modèle est le vecteur à n composantes :

$$\eta = \mathbf{X}\beta.$$

Il faut noter que cette combinaison linéaire peut inclure des transformations des variables explicatives initiales (par exemple, $X_{i4} = \ln X_{i2}$) ou des interactions (par exemple, $X_{i5} = X_{i2} \times X_{i3}$). Le vecteur β est un vecteur de paramètres inconnus de taille p.

1.1.2.3 Fonction de lien

La troisième composante d'un modèle linéaire généralisé est la *fonction de lien*. Elle explicite la relation entre la composante aléatoire et la composante déterministe. Plus précisément, la *fonction de lien* spécifie comment l'espérance mathématique de la variable Y notée $\mathbb{E}(Y)$ est liée au prédicteur linéaire construit à partir des variables explicatives. On obtient

$$\eta_i = g(\mathbb{E}(Y_i)), \quad i = 1, \ldots, n,$$

où g est une fonction monotone et différentiable.
La *fonction de lien* qui permet d'égaler le prédicteur linéaire et le paramètre canonique est appelée *fonction lien canonique*. La *fonction lien canonique* $g = \left(\partial b/\partial\theta\right)^{-1}$ est souvent utilisée et dans ce cas, on a $\theta_i = g(\eta_i)$. Le tableau 1.2 présente des fonctions de lien canonique associées à quelques lois classiques. Pour plus d'informations, on peut se référer à McCullagh et Nelder (1989).

Loi	$\mathcal{P}(\lambda)$	$\mathcal{B}(n,p)$	$\mathcal{N}(\mu,\sigma^2)$	$\gamma(\mu,\nu)$
$g(x)$	$\log(x)$	$\log\left(\frac{x}{1-x}\right)$	x	$\frac{1}{x}$

Tableau 1.2 – Exemples de fonctions de lien canonique associées à des distributions classiques.

Dans la section suivante, nous nous intéressons à l'estimation des paramètres β d'un modèle linéaire généralisé.

1.1.3 Estimation du paramètre de régression

L'estimation du paramètre de régression β consiste à rechercher la valeur $\widehat{\beta}$ de β qui maximise la vraisemblance. Nous débutons notre section par la présentation des expressions des moments d'ordre un et deux de la variable aléatoire Y.

Considérons une variable à expliquer Y ayant des observations indépendantes $(y_1,\ldots,y_n)$ provenant d'une distribution ayant une densité de probabilité ou une fonction de masse pour Y_i de la forme (1.1). Soit $\ell_i(\theta_i,\phi;y_i) = \log(f(y_i;\theta_i,\phi)$ désigne la contribution de la i-ème observation y_i à la log-vraisemblance.
Pour tout $i=1,\ldots,n$, on a :

$$\ell_i(\theta_i,\phi,y_i) \;=\; \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi),$$

$$\frac{\partial \ell_i(\theta,\phi,y_i)}{\partial \theta_i} = \frac{y_i\theta_i - \dot{b}(\theta_i)}{a_i(\phi)} \quad \text{et} \quad \frac{\partial^2 \ell_i(\theta,\phi,y_i)}{\partial \theta_i^2} = \frac{-\ddot{b}(\theta_i)}{a_i(\phi)},$$

où $\dot{b}(\theta_i)$ et $\ddot{b}(\theta_i)$ désignent respectivement la dérivée première et la dérivée seconde de la fonction b par rapport à θ.
Sous certaines conditions de régularités vérifiées par les lois issues de structures exponentielles, on a les relations suivantes :

$$\mathbb{E}\Big(\frac{\partial \ell_i(\theta,\phi,y_i)}{\partial \theta_i}\Big) = 0 \quad \text{et} \quad -\mathbb{E}\Big(\frac{\partial^2 \ell_i(\theta,\phi,y_i)}{\partial \theta_i^2}\Big) = \mathbb{E}\Big(\Big(\frac{\partial \ell_i(\theta,\phi,y_i)}{\partial \theta_i}\Big)^2\Big).$$

Alors, on montre que

$$\mathbb{E}(Y_i) = \dot{b}(\theta_i) \quad \text{et} \quad \mathbf{var}(Y_i) = \ddot{b}(\theta_i)a_i(\phi).$$

On peut donc remarquer qu'il existe une relation directe entre l'espérance de Y_i et sa variance :

$$\mathbf{var}(Y_i) = a_i(\phi)\ddot{b}(\dot{b}^{-1}(\mathbb{E}(Y_i))) = \frac{\phi}{\omega_i}\ddot{b}(\dot{b}^{-1}(\mathbb{E}(Y_i))). \tag{1.2}$$

1.1.3.1 Equations de la vraisemblance

Supposons p variables explicatives dont les observations sont rangées dans la matrice de plan d'expérience $\mathbf{X}$, $\beta \in \mathbb{R}^p$ un vecteur de paramètres de p paramètres et le prédicteur linéaire à n composantes $\eta = \mathbf{X}\beta$. Soit g la fonction de lien telle que: $\eta_i = g(\mu_i)$ où $\mu_i = \mathbb{E}(Y_i)$.
Considérons n observations indépendantes de Y et supposons que θ dépend de β, la log-vraisemblance du vecteur des paramètres canoniques θ pour les données observées s'écrit :

$$\ell_n(\beta) = \sum_{i=1}^{n} \log f_i(y_i, \theta_i, \phi) = \sum_{i=1}^{n} \ell_i(\theta_i, \phi, y_i). \tag{1.3}$$

L'équation du score est donnée donné par :

$$\frac{\ell_i(\beta)}{\partial \beta_j} = \frac{\ell_i(\beta)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}.$$

D'autre part,

$$\begin{aligned}
\frac{\ell_i(\beta)}{\partial \theta_i} &= \frac{y_i - \dot{b}(\theta_i)}{a_i(\phi)}, \\
\frac{\partial \theta_i}{\partial \mu_i} &= 1/\frac{\partial \mu_i}{\partial \theta_i} = 1/\ddot{b}(\theta_i) = \frac{a_i(\phi)}{\mathbf{var}(Y_i)} \\
\frac{\partial \mu_i}{\partial \eta_i} & \quad \text{depend de la fonction de lien:} \quad g(\mu_i) = \eta_i \\
\frac{\partial \eta_i}{\partial \beta_j} &= X_{ij} \quad \mathbf{car} \quad \eta_i = X_i^\top \beta.
\end{aligned}$$

On en déduit des équations d'estimation ou équations de la vraisemblance pour les paramètres β_j :

$$\sum_{i=1}^{n} X_{ij} \frac{(y_i - \mu_i)}{\mathbf{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1, \ldots, p. \tag{1.4}$$

Les équations (1.4) sont non-linéaires en β. En effet, elles dépendent de β au travers de μ_i et de η_i, $i = \ldots, n$ (nous n'avons pas fait apparaître explicitement cette dépendance, afin de conserver des notations simples). De manière générale, les équations (1.4) n'admettent pas de solution analytique. Cependant, des méthodes itératives comme les algorithmes de types Newton-Raphson et de Fisher-scoring sont utilisées pour approcher l'estimateur du maximum de vraisemblance.

1.1.3.2 Algorithme de Newton-Raphson

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

Recherchant la solution de l'équation de vraisemblance qui est telle que

$$\frac{\partial \ell_n(\beta)}{\partial \beta} = \left.\frac{\partial \ell_n(\beta)}{\partial \beta}\right|_{\beta=\widehat{\beta}_n} = 0,$$

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

On commence par approcher la fonction $\ell_n(\beta)/\partial\beta$ par son développement de Taylor à l'ordre 1.
Si $\beta^{(k)}$ désigne l'approximation de $\widehat{\beta}_n$ obtenue à la k-ième itération de l'algorithme, on cherche $\beta^{(k+1)}$ tel que :

$$0 = \frac{\partial \ell_n(\beta^{k+1})}{\partial \beta} = \frac{\partial \ell_n(\beta^{(k)})}{\partial \beta} + \frac{\partial^2 \ell_n(\beta^{(k)})}{\partial \beta \partial \beta^\top}(\beta^{(k+1)} - \beta^{(k)}).$$

On obtient:

$$\beta^{(k+1)} = \beta^{(k)} - \left(\frac{\partial^2 \ell_n(\beta^{(k)})}{\partial \beta \partial \beta^\top}\right)^{-1} \frac{\partial \ell_n(\beta^{(k)})}{\partial \beta}. \tag{1.5}$$

Partant d'une valeur initiale $\beta^{(0)}$, on réitère la formule (1.5) jusqu'à ce qu'un critère de convergence soit satisfait (de la stabilité) par exemple, la norme $\|\beta^{(k+1)} - \beta^{(k)}\|$ de la différence entre deux approximations successives devient plus petite qu'un seuil $\varepsilon > 0$ fixé d'avance).

1.1.3.3 Algorithme des scores de Fisher

L'algorithme des scores de Fisher est aussi une méthode itérative qui permet de résoudre des équations de vraisemblance (voir Jennrich et Sampson (1976) pour plus de détails). Sa procédure est proche de celle de l'algorithme de Newton-Raphson,

la différence provenant de la matrice hessienne. L'algorithme des scores de Fisher utilise l'espérance de cette matrice, appelée information espérée, tandis que celui de Newton-Raphson utilise la matrice même, aussi appelée information observée.
Soit $\beta^{(k)}$ la k-ième approximation pour l'estimateur du maximum de vraisemblance $\widehat{\beta}$ de β, la procédure de l'algorithme des scores de Fisher se présente comme suit :

$$\beta^{(k+1)} = \beta^{(k)} - \left[\mathbb{E}\left(\frac{\partial^2 \ell_n(\beta)}{\partial\beta\partial\beta^\top}\right)^{(k)}\right]^{-1} \frac{\partial \ell_n(\beta^{(k)})}{\partial\beta},$$

où la matrice d'information de Fisher est évaluée en $\beta = \beta^{(k)}$.
Ainsi, on réitère cette procédure jusqu'à obtenir la stabilité comme dans la procédure de l'algorithme de Newton-Raphson.

1.1.4 Propriétés asymptotiques

Dans le cadre général des modèles linéaires généralisés, Fahrmeir et Kaufmann (1985) ont démontré différents résultats dont, en particulier, le théorème sur la normalité asymptotique de $\widehat{\beta}_n$, solution des équations du maximum de vraisemblance pour un échantillon de taille n. Ce théorème repose principalement sur des hypothèses concernant les matrices hessiennes et d'information de Fisher.

Théorème 1.1 *Sous les conditions émises par Fahrmeir et Kaufmann (1985), la suite $(\widehat{\beta}_n)$ des estimateurs du maximum de vraisemblance $\widehat{\beta}_n$ converge en probabilité vers β et $\mathcal{I}_n(\widehat{\beta}_n)^{1/2}(\widehat{\beta}_n - \beta_0)$ converge en loi vers le vecteur gaussien $\mathcal{N}(0, I_p)$, où $\mathcal{I}_n(\beta) = -\frac{\partial^2 \ell_n(\beta)}{\partial\beta\partial\beta^\top}$ et β_0 est la vraie valeur inconnue du paramètre.*

1.1.5 Qualité d'ajustement, test et choix entre différents modèles

Après avoir estimé les paramètres de la loi choisie, puis étudiées leurs propriétés, il est fondamental d'étudier la qualité d'ajustement, de vérifier les hypothèses concernant les coefficients du modèle. Dans le cas où il existe plusieurs modèles concurrents, des critères permettant de choisir le modèle le plus approprié sont proposés.

1.1.5.1 Qualité d'ajustement

La qualité d'ajustement se fait sur la base des différences entre observations et estimations. Le plus souvent, deux statistiques sont utilisées pour juger l'adéquation du modèle aux données.

1. **La déviance** mesure l'écart entre la log-vraisemblance obtenue en β et celle obtenue avec un modèle parfait (dit saturé), c'est-à-dire le modèle possédant autant de paramètres que de variables. La déviance est donc définie par :

$$Dev \quad = \quad 2\phi \times \left[\ell_{n,sat} - \widehat{\ell}_n\right],$$

où $\ell_{n,sat}$ est la log-vraisemblance du modèle saturé, c'est-à-dire le modèle possédant autant de paramètres que de variable, $\widehat{\ell}_n$ est la log-vraisemblance du modèle estimé.

2. **La statistique de test du Khi-deux de Pearson ou test du** χ^2 est également utilisé pour comparer les valeurs observées y_i à leur prévision par le modèle. La statistique du test est définie par :

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

où $\widehat{\mu}$ est l'estimateur du maximum de vraisemblance des $\mu = \mathbb{E}(Y_i|\mathbf{X}_i)$ dans le modèle saturé, $i = 1, \ldots, n$.
Le Khi-deux normalisé est égal à χ^2/ϕ.

Remarque 1.1 *Lorsque le modèle étudié est exact, ces deux statistiques suivent approximativement une loi du Khi-deux à $n - p$ degrés de liberté. Dans la pratique ces deux approches conduisent à des résultats peu différents et dans le cas contraire, c'est une indication de mauvaise approximation de la loi asymptotique.*

1.1.5.2 Tests

La statistique de Wald et le rapport de vraisemblance sont deux critères couramment proposés pour aider au choix de modèle.

1. **Test de Wald** est un test paramétrique qui permet de tester la non-significativité des coefficients de regression. En pratique, si nous voulons tester la non-significativité de la j-ième variable explicative dans le prédicteur linéaire $\beta^\top \mathbf{X}_i$.
   Soit l'hypothèse nulle $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$. Sous H_0, la statistique de Wald $\widehat{\beta}_{n,j}/\widehat{\sigma}_j$ converge en loi vers la loi normale centrée réduite, où $\widehat{\beta}_{n,j}$ est la j-ième composante de $\widehat{\beta}_n$ et $\widehat{\sigma}_j^2$ est le j-ième terme diagonal de $\mathcal{I}_n(\widehat{\beta}_n)^{-1}$. Ainsi, on peut prendre pour région de rejet de H_0 au niveau asymptotique α :

$$\mathcal{R}_\alpha = \Big\{ \Big| \frac{\widehat{\beta}_{n,j}}{\widehat{\sigma}_j} \Big| \geq u_{1-\alpha/2} \Big\},$$

   où $u_{1-\alpha/2}$ désigne le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$.

2. **Test du rapport de vraisemblance** est un test qui permet de faire des comparaisons entre deux modèles *emboîtés*. Pratiquement, en supposant que l'on veuille tester la nullité simultanée de q paramètres (sans perte de généralité et pour simplifier les notations, supposons que l'on souhaite tester la nullité des q premiers coefficients de β). Alors, le problème de test peut s'écrire :

$$H_0 \ : \ \beta_1 = \beta_2 = \ldots = \beta_q = 0 \text{ contre } H_1 : \text{il existe } i \in \{1, \ldots, q\} \text{ tel que } \beta_i \neq 0.$$

   Ce test consiste à comparer les vraisemblances (ou log-vraisemblance) sous H_0 et sous H_1 et à accepter H_0 si elles sont "proches". La statistique de test est donnée par :

$$D_n = 2\big(\ell_n(\widehat{\beta}_n) - \ell_n(\widehat{\beta}_{n,H_0})\big),$$

   où $\ell_n(\widehat{\beta}_{n,H_0})$ est la log-vraisemblance sous H_0 (lorsqu'on retire du modèle les q premières variables explicatives). Lorsque H_0 est vraie, D_n converge en loi vers un $\chi^2(q)$ lorsque n tend vers l'infini. Ainsi, une région de rejet de H_0, au niveau asymptotique α est :

$$\mathcal{R}_\alpha^D = \Big\{ D_n \geq c_q(1-\alpha) \Big\},$$

   où $c_q(1-\alpha)$ désigne le quantile d'ordre $1-\alpha$ de la loi de $\chi^2(q)$.

1.1.6 Choix entre différents modèles

Pour comparer des modèles qui ne sont pas forcément emboîtés les uns dans les autres, on utilise les plus souvent les critères de choix de modèles tels l'AIC (Akaike (1974)) ou le BIC (Schwarz (1978)). Ces critères peuvent être utilisés pour réaliser une sélection de variables. Supposons que nous avons un modèle statistique de certaines données. Désignons par $\ell_n(\widehat{\beta}_n)$ la valeur maximale de la fonction de log-vraisemblance du modèle.

– L'**AIC** (*Akaike Information Criterion*) pour un modèle à p paramètres estimés sur n observations est défini par :

$$AIC = -2 \times \ell_n(\widehat{\beta}_n) + 2p.$$

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

$$BIC = -2 \times \ell_n(\widehat{\beta}_n) + p\log(n).$$

En pratique, on calcule le critère de choix de modèle (AIC ou BIC) pour les modèles concurrents et le modèle qui présente le plus faible critère est choisi comme étant le meilleur.

1.2 Rappels sur la modélisation des données de comptage

La description de cette section est basée en grande partie sur Diallo (2017), Dupuy (2018) et Ali (2021).

1.2.1 Modèles de régression de Poisson et binomial Négatif

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

1.2.1.1 Le modèle de régression de Poisson

La distribution de Poisson est souvent retenu pour expliquer une variable quantitative Y à valeurs entières positives (par exemple un nombre d'occurrences d'un événement). La probabilité que Y prenne la valeur y_i ($y_i = 0, 1, 2, \cdots$) est donnée par

$$\mathbb{P}(Y_i = y_i | \mathbf{X}_i = x_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} \quad y_i = 0, 1, 2, \cdots$$

où $\mathbb{P}$ est la probabilité, Y_i est une variable de comptage observée (un nombre d'événements) pour l'individu i et $\mathbf{X}_i$ est un vecteur de p variables explicatives linéairement indépendantes observées pour l'individu i. On peut aussi montrer que

$$\mathbb{E}(Y = y_i | \mathbf{X}_i = x_i) = \mu_i = \exp(\beta^\top \mathbf{X}_i),$$

où $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^\top$ est un vecteur de paramètres de dimension appropriée $p + 1$. La forme de la fonction exponentielle assure la non-négativité du paramètre de la moyenne μ. La fonction log-vraisemblance du modèle est donnée par l'équation suivante :

$$\ell_n(\beta) = \sum_{i=1}^{n} \{Y_i \beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} - \log(Y_i!)\}. \tag{1.6}$$

Les paramètres sont choisis de façon à maximiser la valeur de la fonction log-vraisemblance. Les conditions de premier ordre sont :

$$\sum_{i=1}^{n} \mathbf{X}_{ij}(Y_i - e^{\beta^\top \mathbf{X}_i}) = 0, j = 1, \cdots, p.$$

Le modèle de régression de Poisson est trop restrictif pour les données de comptage, ce qui a incité les chercheurs à recourir à des modèles alternatifs comme le modèle binomial négatif qui permet la surdispersion.

1.2.1.2 Le modèle de regression binomial négatif

Le modèle de régression binomial négatif est le plus souvent utilisé pour modéliser les données de comptage surdispersées. Ce modèle est une généralisation du modèle de Poisson qui permet de prendre en compte la surdispersion des données de comptage par l'introduction d'un paramètre qui mesure le degré de surdispersion. Dans un modèle de régression binomial négatif, conditionnellement à $\mathbf{X}_i$ la probabilité pour que la variable Y_i prenne la valeur y_i est définie par

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

où α est un paramètre auxiliaire qui mesure le degré de sur-dispersion. Cette distribution a une moyenne conditionnelle égale à μ_i et une variance conditionnelle égale à $\mu_i(1 + \alpha\mu_i)$. On peut aussi remarquer que $\mathbb{E}(Y_i|\mathbf{X}_i = x_i) \leq \text{var}(Y_i|\mathbf{X}_i = x_i)$ car $\alpha > 0$. La loi binomiale négative tend vers la loi de Poisson lorsque α tend vers zéro.

Dans la section suivante, nous présentons la notion d'*excès de zéros* ou d'*inflation de zéros* et nous expliquons en quoi les modèles de comptages classiques (tels que les modèles de Poisson, binomial, binomial négatif) ne sont pas adaptés à cette situation. Les modèles à inflation de zéros en particuliers les modèles de régression à inflation de zéros (ZIP) et binomial (ZIB) qui permettent de répondre plus efficacement aux problèmes posés par l'excès de zéros sont décrits en détail dans la section suivante.

1.3 Modèles de régression à inflation de zéros

1.3.1 Introduction

Dans cette section , nous nous intéressons à une cause particulière de sur-dispersion, appelée inflation de zéros. Ce phénomène que nous définissons plus précisément dans la suite, intervient lorsque l'on observe un nombre "*excessif*" de zeros dans des données de comptage. Il existe plusieurs modélisations possibles de ce type de données. Nous nous intéressons ici à une classe particulière de modèles, dits "modèles à inflation de zéros", qui se présentent comme des mélanges entre une masse de Dirac en zero et un modèle classique de comptage (typiquement, un modèle de Poisson, ou Poisson genéralisé, ou binomial, etc $\cdots$).

1.3.2 Le modèle de regression ZIP

1.3.2.1 Définition

Soit Z une variable de comptage sur un échantillon de n individus. On note Z_i l'observation de Z sur un individu i. La probabilité pour que l' individu i soit dans le groupe des zéros est notée π_i. La variable Z_i est modélisée par un ZIP lorsque

$$\mathbb{P}(Z_i = z_i|\mathbf{X}_i, \mathbf{W}_i) = \begin{cases} \omega_i + (1-\omega_i)e^{-\mu_i} & \text{si} \quad z_i = 0, \\ (1-\omega_i)\frac{e^{-\mu_i}\mu_i^{z_i}}{z_i!} & \text{si} \quad z_i = 1, 2, \cdots \end{cases} \tag{1.7}$$

où ω_i et μ_i sont fonctions respectivement des vecteurs de covariables $\mathbf{W}_i = (W_{i1}, \cdots, W_{iq})^\top$ et $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})^\top$. Dans la régression ZIP, la probabilité de mélange ω_i et le paramètre μ_i sont généralement modélisés par des modèles logistiques et log-linéaires respectivement, c'est-à-dire :

$$\text{logit}(\omega_i) = \gamma^\top \mathbf{W}_i \quad \text{et} \quad \log(\mu_i) = \beta^\top \mathbf{X}_i, \tag{1.8}$$

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

$$\forall i = 1, \cdots, n, \begin{cases} Z_i \sim \omega_i \delta_0 + (1-\omega_i)\mathcal{P}(\mu_i) \\ \text{logit}(\omega_i) = \gamma^\top \mathbf{W}_i \\ \log(\mu_i) = \beta^\top \mathbf{X}_i \end{cases} \tag{1.9}$$

Conditionnellement à $\mathbf{X}_i$ et $\mathbf{W}_i$, l'espérence et la variance de Z_i sont données par :

$$\mathbb{E}(Z_i|\mathbf{X}_i, \mathbf{W}_i) = (1-\omega_i)\mu_i \quad \text{et} \quad \text{var}(Z_i|\mathbf{X}_i, \mathbf{W}_i) = (1+\omega_i\mu_i)(1-\omega_i)\mu_i.$$

1.3.2.2 Estimation dans le modèle ZIP

Dans la littérature, plusieurs 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) et Czado *et al.* (2007) pour plus de détails. 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 spécifiées. Pour pallier à ce problème, Hall et Shen (2010) ont suggéré une nouvelle procédure d'estimation du modèle (1.7) dite "robust expectation-solution (RES) estimation" ou tout simplement l'algorithme ES (expectation-solution). Cet algorithme est une modification de l'algorithme expectation-maximization (EM), voir Dempster *et al.* (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 ω d'appartenir au groupe des zéros.
Supposons que nous observons n vecteurs indépendants $(Z_1, \mathbf{X}_1, \mathbf{W}_1), \cdots, (Z_n, \mathbf{X}_n, \mathbf{W}_n)$ à partir des modèles (1.7)-(1.8), tous définis sur l'espace de probabilité $(\Omega, \mathcal{C}, \mathbb{P})$. Sur la base de ces observations, la log-vraisemblance de $\theta = (\beta^\top, \gamma^\top)^\top$ est égale à

$$\begin{aligned}\ell_n(\theta) &= \sum_{i=1}^{n} \Big\{ J_i \log\left[e^{\gamma^\top \mathbf{W}_i} + \exp(-e^{\beta^\top \mathbf{X}_i})\right] \\ &\quad + (1-J_i)\left[Z_i \beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} - \log(Z_i!)\right] - \log(1+e^{\gamma^\top \mathbf{W}_i})\Big\}\end{aligned}$$

où $J_i = 1_{\{Z_i=0\}}$.
En particulier, supposons que l'on observe la variable indicatrice s telle que $S_i = 1$ si z_i provient de l'ensemble des zéros (distribution dégénérée) et $S_i = 0$ si z_i résulte du zéro aléatoire (distribution non dégénérée). Alors la log-vraisemblance pour les données complètes $(z; S)$ est donnée par

$$\begin{aligned}\ell_n^C(z, S; \theta) &= \sum_{i=1}^{n} \Big\{ \left[S_i \gamma^\top \mathbf{W}_i - \log\left(1 + e^{\gamma^\top \mathbf{W}_i}\right)\right] \\ &\quad + (1-S_i)\left[Z_i \beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} - \log(Z_i!)\right]\Big\} \\ &= \tilde{\ell}_{n,1}(\gamma) + \tilde{\ell}_{n,2}(\beta),\end{aligned}$$

où $S = (S_1, \cdots, S_n)^\top$.
Cette log-vraisemblance n'est pas calculable, puisque les $S_i, i = 1, \cdots, n$ ne sont pas observés. Avec l'algorithme EM (voir Dempster *et al.* (1977) pour plus de détails), la log-vraisemblance est maximisée de manière itérative en commençant par une va-

leur initiale $(\beta_0^\top, \gamma_0^\top)^\top$ et en alternant les étapes suivantes :

Étape E : estimer la variable S_i par son espérance conditionnelle $S_i^{(r)}$ sous les estimations courantes des paramètres $\beta^{(r)}$ et $\gamma^{(r)}$.

Étape M : trouver $\beta^{(r+1)}$ et $\gamma^{(r+1)}$ en maximisant respectivement les fonctions $\tilde{\ell}_{n,1}(\gamma)+\tilde{\ell}_{n,2}(\beta)$. Hall et Shen (2010) ont montré que maximiser ces deux fonctions revient à résoudre respectivement les deux équations suivantes

$$\frac{1}{n}\sum_{i=1}^{n}\{S_i^{(r)} - \omega_i\}W_i = 0. \tag{1.10}$$

$$\frac{1}{n}\sum_{i=1}^{n}(1 - S_i^{(r)})\{z_i - e^{\beta^\top \mathbf{x}_i}\}X_i = 0. \tag{1.11}$$

Dans l'approche RES, Hall et Shen (2010) proposent de remplacer les équations (1.10) et (1.11) par des fonctions d'estimation robustes. Sous des conditions de régularité de Rosen *et al.* (2000) liées à l'algorithme ES et de Carroll *et al.* (2006), Hall et Shen (2010) ont montré le résultat suivant qui généralise le celui Czado *et al.* (2007).

Théorème 1.2 *Si l'algorithme RES converge, alors il existe une suite de variables aléatoires $\hat{\theta}_n$ telle que :*

1. $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$ *quand* $n \to \infty$ *(consistance),*
2. $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{V}(\theta_0))$ *quand* $n \to \infty$ *(normalité asymptotique),*

où l'expression $\mathbf{V}(\theta_0)$ de la variance asymptotique est donnée dans Hall et Shen (2010). Des extensions de modèle ZIP ont été étudiés. Citons entre autres Lam *et al.* (2007) , He *et al.* (2010), Nguyen et Dupuy (2019) ont étendu ce modèle ZIP respectivement dans le cadre semi-paramétrique, doublement semiparamétrique et dans le cas des données censurées. Ces auteurs ont établi les résultats de consistance et de normalité asymptotique des estimateurs proposés.

1.3.3 Le modèle de régression ZINB

Pour une variable réponse $Z_i, i = 1, \cdots, n$, on dira que Z_i est modélisée par un ZINB si sa distribution est donnée par :

$$\mathbb{P}(Z_i = z_i|\mathbf{X}_i, \mathbf{W}_i) = \begin{cases} \omega_i + (1-\omega_i)\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha} & \text{si} \quad z_i = 0, \\ (1-\omega_i)\frac{\Gamma(z_i+1/\alpha)}{\Gamma(1/\alpha)z_i!}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{z_i}\left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha} & \text{si} \quad z_i = 1, 2, \cdots \end{cases} \tag{1.12}$$

avec

$$\mathbb{E}(Z_i|\mathbf{X}_i, \mathbf{W}_i) = (1-\omega_i)\mu_i \quad \text{et} \quad \text{var}(Z_i|\mathbf{X}_i, \mathbf{W}_i) = (1-\omega_i)\mu_i(1+(\alpha\omega_i)\mu_i),$$

où α est un paramètre de surdispersion et ω_i représente la probabilité d'inflation de zéros. Comme pour les modèles de Poisson et Binomial Négatif, le modèle ZINB tend vers le modèle ZIP lorsque α tend vers zéro.

L'étude des propriétés asymptotiques dans le modèle ZINB peut se faire de manière similaire à celle effectuée précédemment dans le modèle ZIP. Le lecteur est renvoyé à Hilbe (2007), Czado *et al.* (2007) et Mwalili *et al.* (2008) pour plus de détails.

1.3.4 Le modèle de regression ZIB

Le modèle de régression Binomial à inflation de zéros (ZIB) a été utilisé en premier par Kemp et 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 se défini comme suit :

$$Y_i \sim \begin{cases} 0 & \text{avec une probabilité} \quad p_i, \\ \mathcal{B}(n_i, \pi_i) & \text{avec une probabilité} \quad 1-p_i, \end{cases} \tag{1.13}$$

ce qui implique

$$Y_i \sim \begin{cases} 0 & \text{avec une probabilité} \quad p_i + (1-p_i)(1-\pi_i)^{n_i}, \\ k & \text{avec une probabilité} \quad (1-p_i)\binom{n_i}{k}\pi_i^{n_i}(1-\pi_i)^{n_i-k}, k = 1, 2, \cdots, n_i \end{cases} \tag{1.14}$$

avec $\mathbb{E}(Y_i) = (1-p_i)n_i\pi_i$ et $\text{var}(Y_i) = (1-p_i)n_i\pi_i(1-\pi_i(1-p_in_i))$. La vraisemblance du modèle ZIB peut s'écrire de la manière suivante :

$$L_n(\beta,\gamma) = \prod_{i=1}^{n}\left(p_i(1-p_i)(1-\pi_i)^{n_i}\right)^{J_i}.\left((1-p_i)\binom{n_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i}\right)^{1-J_i} \tag{1.15}$$

où $J_i := 1_{\{Z_i=0\}}$.
Les paramètres $p = (p_1,\cdots,p_n)$ et $\pi = (\pi_1,\cdots,\pi_n)$ sont respectivement modélisés via une fonction de lien logit,

$$\text{logit}(p) = \beta^\top \mathbf{W} \quad \text{et} \quad \text{logit}(\pi) = \gamma^\top \mathbf{X}, \tag{1.16}$$

où $\mathbf{W} \in \mathbb{R}^q$ et $\mathbf{X} \in \mathbb{R}^\ell$ 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 $\beta \in \mathbb{R}^\ell$ sont les paramètres de régression. La log-vraisemblance du modèle basée sur les observations $(Y_i, X_i, W_i), i = 1,\cdots,n$, est donnée par

$$\begin{aligned}\ell_n(\theta) \;=\; \sum_{i=1}^{n}\Big\{ & J_i \log\left(e^{\gamma^\top \mathbf{W}_i} + (1+e^{\beta^\top \mathbf{X}_i})^{-m_i}\right) - \log\left(1+e^{\gamma^\top \mathbf{W}_i}\right) \\ & +(1-J_i)\left[Y_i\beta^\top \mathbf{X}_i - m_i \log\left(1+e^{\beta^\top \mathbf{X}_i}\right)\right]\Big\}.\end{aligned}$$

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.

1.4 Données manquantes

1.4.1 Introduction

Dans les ensembles de données réelles, il est assez courant d'avoir des observations avec des valeurs manquantes pour une ou plusieurs caractéristiques d'entrée. Ces absences d'observations compliquent l'analyse de ces données. Les données manquantes apparaissent dans divers contextes, notamment dans les enquêtes, les essais cliniques et les études épidémiologiques. Avec ou sans données manquantes, le but d'une analyse statistique est de faire des inférences valides et efficaces sur une population d'intérêt. La question des valeurs manquantes complique ce processus. Ainsi, nous allons d'abord décrire les mécanismes de données manquantes et ensuite présenter quelques méthodes de traitement de ces données. La description de cette section est basée en grande partie sur Little et Rubin (1987), Tsiatis (2006) et Héraud

(2012).

1.4.2 Mécanismes des données manquantes

En statistique, on parle de données manquante lorsqu'on n'a pas d'observations sur une variable donnée pour un individu. Pour une analyse efficiente des jeux de données, Little et Rubin (1987) recommandent d'identifier le mécanisme qui induit l'observation ou l'absence de la donnée. Le mécanisme qui induit l'observation ou l'absence de la donnée est souvent appelé mécanisme de données manquantes. Rubin (1976) et Rubin (2004) ont introduit une classification statistique des données manquantes permettant de différencier trois types de données manquantes. Considérons δ une variable qui indique si la variable Z est observée ou non. δ prend la valeur 1 si la variable Z est observée et 0 sinon. Notons par Z_i^{obs} l'ensemble des données observées et par Z_i^{mis} celui des données manquantes. Soit ϕ le paramètre du modèle des données manquantes. L'expression générale du modèle des données manquantes est alors

$$\mathbb{P}(\delta|Z^{obs}, Z^{mis}, \phi).$$

- Les données Z sont *manquantes complètement au hasard* (MCAR, pour Missing Completely At Random) si

  $$\mathbb{P}(\delta = 1|Z^{obs}, Z^{mis}, \phi) = \mathbb{P}(\delta = 1|\phi)$$

  ce qui signifie que la probabilité d'observé Z ne dépend pas des données $Z = (Z^{obs}, Z^{mis})$ mais dépend uniquement de certains paramètres ϕ. Ce qui suggère que les causes des données manquantes ne sont pas liées aux données.

- Les données Z sont *Manquantes au hasard* (MAR, pour Missing At Random) si

  $$\mathbb{P}(\delta = 1|Z^{obs}, Z^{mis}, \phi) = \mathbb{P}(\delta = 1|Z^{obs}, \phi)$$

  ce qui signifie que la probabilité que la variable Z ne soit pas observée dépend de l'information observée, y compris des facteurs de conception.

- Les données Z sont *manquantes non au hasard* (MNAR, pour Missing Non At Random) si

  $$\mathbb{P}(\delta = 1|Z^{obs}, Z^{mis}, \phi) \quad \text{ne se simplifie pas}$$

  ce qui signifie que la probabilité que la variable Z ne soit pas observée dépend aussi de l'information non observée ou bien des variables qui n'auraient pas été collecté.

1.4.3 Méthodes classiques de traitement des données manquantes

La littérature présente diverses méthodes de traitement des données manquantes. Nous présentons dans la suite de la section trois des méthodes les plus populaires.

Analyse des cas-complets

L'analyse des cas-complets est une méthode simple d'analyse de données manquantes. Cette méthode consiste à restreindre l'analyse aux individus pour lesquels toutes les variables sont entièrement renseignées. Dans le traitement des données manquantes, les logiciels statistiques l'appliquent par défaut. Certes, cette méthode est facile à appliquer mais peut conduire à d'importants biais dans les estimations.

Imputation Multiple

L'imputation multiple a été présentée officiellement par Rubin (1978). Le concept d'imputation multiple consiste à remplacer chaque valeur manquante par une valeur générée suivant un modèle d'imputation. L'imputation est répétée M fois, ce qui permet d'avoir M ensembles de données complètes. L'objectif est de combiner la simplicité des stratégies d'imputation, avec l'absence de biais dans les estimations ponctuelles et les mesures de précision. Les références bibliographiques pour cette méthode sont les livres de Little et Rubin (1987), Little et Rubin (2002). Plusieurs autres sources, telles que Rubin et Schenker (1986), Tanner et Wong (1987) et les ouvrages Schafer (1997) et van Buuren (2012), donnent d'excellentes descriptions de cette technique.

Procédure générale

La procédure d'imputation multiple peut être résumée en trois étapes principales.

Étape 1 : M ensembles complets de données sont générés en remplaçant les valeurs manquantes par des valeurs plausibles qui sont tirées suivant une distribution spécifique.

Étape 2 : On applique la méthode qui aurait été utilisée si les données avaient été complètes sur de chacun des m ($m = 1, \ldots, M$) ensembles imputés pour obtenir M estimations du paramètre.

Étape 3 : Les M estimations du paramètre sont combinées en une seule estimation.

La figure 1.1 ci-dessous illustre ces trois étapes :

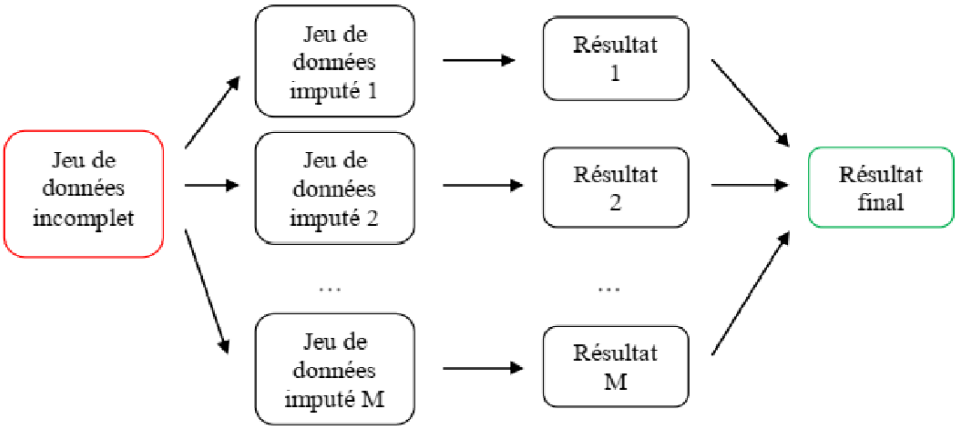

Figure 1.1 – Schéma d'une imputation multiple.

Règle de Rubin

Soit β le paramètre inconnu à estimer. En supposant que la procédure générale de Rubin est appliquée comme décrit précédemment, on obtient M estimations $\widehat{\beta}_m$ de β. L'estimation finale du paramètre β est la moyenne arithmétique des $\widehat{\beta}_m$ ($m = 1, \ldots, M$) définie par $\widehat{\beta}^*_M = \frac{1}{M}\sum_{m=1}^{M} \widehat{\beta}_m$.

Pour calculer la matrice de covariance de l'estimateur $\widehat{\beta}^*_M$ on utilise les matrices de covariances "*intra-imputation*" notées V^{intra} et "*inter-imputation*" notées V^{inter} définies de façon intuitive par Rubin :

- $\mathrm{V}^{intra} = \frac{1}{M}\sum_{m=1}^{M} V_m$ est une mesure classique de la variabilité car nous prenons un échantillon plutôt que la population,
- $\mathrm{V}^{inter} = \frac{1}{M-1}\sum_{m=1}^{M} \left(\widehat{\beta}_m - \widehat{\beta}^*_M\right)\left(\widehat{\beta}_m - \widehat{\beta}^*_M\right)^\top$ mesure la variabilité due au fait que l'on impute des échantillons aléatoirement.

La matrice de covariance de $\widehat{\beta}^*_M$ est donnée par

$$V = \mathrm{V}^{intra} + \left(1 + \frac{1}{M}\right)\mathrm{V}^{inter}.$$

Pondération par l'inverse de la probabilité de sélection

La méthode d'estimation par pondération par l'inverse de la probabilité de sélection (IPW, pour "Inverse Probability Weighting") est une méthode couramment utilisée pour analyser les données manquantes. Cette méthode a été introduite par Horvitz et Thompson (1952), elle consiste à corriger les données manquantes en donnant un poids supplémentaire aux sujets dont les données sont entièrement observées. Plus précisément, en désignant par $\widehat{\pi}$ l'estimation de la probabilité d'observer

un cas complet $\mathbb{P}(\delta = 1|\mathcal{O})$, la méthode IPW consiste à pondérer les cas complets par l'inverse de $\widehat{\pi}$ où $\mathcal{O}$ est un vecteur qui contient l'ensemble des variables entièrement observées. Pour plus de détails sur la méthode voir Rubin (2002).

2

Estimation dans le modèle de Poisson bivarié à inflation de zéros avec une application aux données portant sur l'utilisation des services de santé

Résumé

Les données sur la demande de soins médicaux sont généralement mesurées au moyen d'un certain nombre de comptes différents. Ces données de comptage sont le plus souvent corrélées et sujettes à de fortes proportions de zéros. Cependant, l'excès de zéros et la dépendance entre ces données peuvent affecter conjointement plusieurs de ces mesures d'utilisation. Dans ce chapitre, le modèle de régression de Poisson bivarié à inflation de zéros (ZIBP) est utilisé pour analyser les données d'utilisation des services de santé. Tout d'abord, nous étudions sur le plan théorique les propriétés asymptotiques de l'estimateur du maximum de vraisemblance (EMV) de ce modèle. Ensuite, une étude de simulation est réalisée pour évaluer le comportement de l'estimateur dans des échantillons finis. Enfin, une application du modèle ZIBP à des données de demandes de soins médicaux est fournie à titre d'illustration.

Sommaire

2.1 Introduction

Les modèles de comptage bivariées sont utilisés dans les situations où deux variables de comptage dépendantes sont corrélées et doivent être modélisées conjointement. Les données bivariées sont observées dans de nombreux domaines, notamment le marketing (nombre d'achats de différents produits), la recherche médicale (nombre de crises d'épilepsie avant et après traitement), l'épidémiologie (incidence de différentes maladies dans une série de districts), l'analyse des accidents (nombre d'accidents sur un site avant et après modification des infrastructures), l'économétrie (nombre de changements d'emploi volontaires et involontaires), le sport (nombre de buts marqués par chacune des deux équipes adverses au football), pour n'en citer que quelques-uns. Dans la plupart des cas, les données bivariées sont modélisées par des modèles de Poisson bivariés.

Cependant, dans de nombreuses applications, les données de comptage contiennent un excès de zéros, c'est-à-dire un nombre de zéros qui ne peut être expliqué par les modèles standard. Un grand nombre d'outils statistiques ont été développés pour résoudre ce problème, tels que les modèles de régression à inflation de zéros (ZI). Ces modèles tiennent compte de l'excès de zéros dans les données de comptage en mélangeant une masse de Dirac en zéro avec un modèle de régression de comptage standard (Poisson, binomiale ou binomiale négative, lorsque la variable réponse est univariée et Poisson bivariée, binomiale négative bivariée, lorsque la variable réponse est bivariée, etc.). Plusieurs travaux ont été réalisés sur les modèles de régression univariés à inflation de zéros, tels que Lambert (1992), Lim *et al.* (2014), Monod (2014) et Ali (2022) pour le modèle de régression ZIP (Zero-inflated Poisson); Ridout *et al.* (2001), Moghimbeigi *et al.* (2008), Mwalili *et al.* (2008) et Garay *et al.* (2011) pour le modèle ZINB (Zero-inflated negative binomial), puis Hall (2000), Hall et Berenhaut (2002), Diallo *et al.* (2017) et Diallo *et al.* (2019) pour le modèle de régression ZIB (Zero-inflated binomial). Mais les modèles ZIP, ZIB et ZINB ne sont pas adaptés aux réponses bivariées. Ainsi, plusieurs modèles ont été proposés pour les données biva-

riées de comptage avec inflation de zéros. Par exemple, pour les modèles binomiaux négatifs bivariés à inflation de zéros, voir Wang *et al.* (2003), Faroughi et Ismail (2017), puis pour les modèles de Poisson bivariés à inflation de zéros, on peut se référer à Karlis et Ntzoufras (2003), AlMuhayfith *et al.* (2016), Yang *et al.* (2016), entre autres. Dans ce chapitre, nous considérons le modèle de régression de Poisson bivarié avec inflation de zéros. Ce modèle permet d'analyser conjointement deux variables de comptage corrélées et de prendre en compte le grand nombre d'observations $(0, 0)$ dans la série de données. Depuis son introduction par Li *et al.* (1999), le modèle ZIBP (Zero-inflated bivariate Poisson) a été appliqué dans des contextes variés tels que le marketing, l'épidémiologie, l'analyse des accidents, la recherche médicale, le sport, l'économétrie, etc. D'où la nécessité de considérer l'estimation dans le modèle ZIBP. Ce travail est également motivé par des données issues de l'économie de la santé. En économétrie de la santé, on examine le plus souvent des données sur l'utilisation des services de santé. Deb et Trivedi (1997) ont analysé les données d'utilisation des services de santé pour les personnes âgées de 65 ans et plus. Ces données proviennent de la National Medical Expenditure Survey (NMES) menée en 1987 et 1988. Ce jeu de données est connu sous le nom de NMES1988. Ces données contiennent des mesures de l'utilisation des services de santé telles que le nombre de visites à un professionnel de santé non-médecin (tel qu'un opticien, un physiothérapeute, ...) en cabinet médical, le nombre de consultations externes d'un professionnel de santé non-médecin. Les proportions de zéros sont élevées dans ces mesures d'utilisation des services de santé. Cela signifie qu'au cours de la période étudiée, ces services de santé correspondants n'ont pas été utilisés par un grand nombre de personnes. De plus, selon Gurmu et Elder (2000) et Wang (2003), les mesures d'utilisation des services de santé sont dépendantes. Par conséquent, une analyse univariée de ces données ne serait pas appropriée. Pour résoudre ce problème, Diallo *et al.* (2018) ont proposé le modèle de régression ZIM (Zero-inflated multinomial) et l'ont appliqué aux données de l'enquête NMES1988. Cependant, le modèle de régression ZIM est restrictif, il ne convient que pour les composantes bornées. Ainsi, dans ce travail nous nous intéressons au modèle ZIBP, qui prend en compte tous les individus de la population et tient compte de la corrélation entre les données de demande de soins étudiées. Dans un premier temps, nous nous intéressons aux propriétés asymptotiques du modèle ZIBP. Deuxièmement, une application du modèle ZIBP a permis d'évaluer la demande de soins médicaux et d'identifier les principaux déterminants du renoncement à certains soins médicaux.

Le chapitre est organisé comme suit. Dans la section 2.2, nous présentons le modèle ZIBP et décrivons l'estimateur du maximum de vraisemblance. Dans la section 2.3, nous donnons d'abord quelques notations utiles, puis nous indiquons cer-

taines conditions de régularité et enfin nous établissons la consistance et la normalité asymptotique de l'EMV dans la régression ZIBP. La section 2.4 présente les résultats d'une étude de simulations. La section 2.5 décrit une application du modèle ZIBP à l'analyse des soins de santé utilisés par les personnes âgées aux États-Unis. Pour conclure, une discussion et quelques perspectives sont fournies dans la section 2.6.

2.2 Le modèle de régression de Poisson bivarié à inflation de zéros

Le modèle de Poisson bivarié a été proposé par Holgate (1964) et présenté par Johnson et Kotz (1969). Ce modèle est utilisé pour la modélisation de variables de comptage corrélées.
Considérons l'espace de probabilité $(\Omega, \mathcal{C}, \mathbb{P})$. Un vecteur aléatoire de Poisson bivarié (Y_1, Y_2) peut être construit à partir de trois variables aléatoires de Poisson indépendantes

$$Z_1 \sim \mathcal{P}(\lambda_1), \quad Z_2 \sim \mathcal{P}(\lambda_2) \quad \text{et} \quad U \sim \mathcal{P}(\mu).$$

En posant :

$$Y_1 = Z_1 + U \qquad \text{et} \qquad Y_2 = Z_2 + U.$$

Les variables marginales Y_1 et Y_2 (lois marginales) sont des lois de Poisson $Y_1 \sim \mathcal{P}(\lambda_1 + \mu)$ et $Y_2 \sim \mathcal{P}(\lambda_2 + \mu)$. La loi du vecteur de Poisson bivarié (Y_1, Y_2) est déterminée par les probabilités $\mathbb{P}(Y_1 = y_1, Y_2 = y_2)$, pour $y_1, y_2 \in \mathbb{N}$.
Notons $y_1 \wedge y_2 := \min(y_1, y_2)$. On a :

$$\begin{aligned}
\mathbb{P}(Y_1 = y_1, Y_2 = y_2) &= \mathbb{P}(Z_1 + U = y_1, Z_2 + U = y_2) \\
&= \sum_{s=0}^{y_1 \wedge y_2} \mathbb{P}(U = s, Z_1 = y_1 - s, Z_2 = y_2 - s) \\
&= \sum_{s=0}^{y_1 \wedge y_2} \mathbb{P}(U = s)\mathbb{P}(Z_1 = y_1 - s)P(Z_2 = y_2 - s) \\
&= e^{-\mu - \lambda_1 - \lambda_2}\varphi(y_1, y_2)
\end{aligned}$$

où

$$\varphi(y_1, y_2) = \sum_{s=0}^{y_1 \wedge y_2} \frac{\mu^s}{s!} \frac{\lambda_1^{y_1 - s}}{(y_1 - s)!} \frac{\lambda_2^{y_2 - s}}{(y_2 - s)!}.$$

En utilisant l'indépendance de Z_1, Z_2 et U, on a $\text{cov}(Y_1, Y_2) = \text{cov}(Z_1 + U, Z_2 + U) = \text{var}(U) = \mu$. Par conséquent, μ est une mesure de la dépendance entre Y_1 et Y_2. Lorsque $\mu = 0$, la distribution de Poisson bivarié se réduit au produit de deux dis-

tributions de Poisson indépendantes (appelée distribution de double-Poisson).

Il est bien connu que les distributions de Poisson univariées ne sont pas appropriées pour modéliser les données de comptages qui ont un excès de zéros. Dans ce cas, on propose le plus souvent des modèles de régression à inflation de zéros, voir Lambert (1992). Il en va de même pour le cas bivarié lorsque les données contiennent une forte proportion de couples $(0,0)$. Le modèle de Poisson bivarié à inflation de zéros a été introduit par Li *et al.* (1999). Depuis, ce modèle a été utilisé, entre autres par Wang *et al.* (2003) pour analyser deux types d'accidents du travail, par Bermúdez (2009) dans le domaine de l'assurance automobile et par Yang *et al.* (2016) pour modéliser des données bivariées en économie de la santé. Selon Li *et al.* (1999), un modèle ZIBP est un mélange d'une distribution de Poisson bivariée et d'une masse ponctuelle en $(0,0)$. Ainsi, le modèle ZIBP est spécifié par la fonction de probabilité définie par :

$$f_{ZIBP}(y_1,y_2;\pi,\lambda_1,\lambda_2,\mu)=\begin{cases}\pi+(1-\pi)\exp(-\mu-\lambda_1-\lambda_2), & (y_1,y_2)=(0,0)\\ (1-\pi)\exp(-\mu-\lambda_1-\lambda_2)\varphi(y_1,y_2), & (y_1,y_2)\neq(0,0),\end{cases} \tag{2.1}$$

où $0<\pi<1$. Lorsque des covariables sont présentes, le modèle (2.1) peut être étendu à un modèle de régression. À cette fin, pour chaque $i=1,...,n$, on considère respectivement $\mathbf{W}_i=(W_{i1},....,W_{iq})^\top$ et $\mathbf{X}_i=(X_{i1},...,X_{ip})^\top$ des vecteurs aléatoires de covariables où $W_{i1}=X_{i1}=1$. La probabilité de mélange π_i est généralement modélisée par une régression logistique, à savoir :

$$\text{logit}(\pi_i)=\gamma^\top\mathbf{W}_i \tag{2.2}$$

et les paramètres de Poisson λ_{1i}, λ_{2i} et μ_i sont modélisés par :

$$\log(\lambda_{1i})=\beta_1^\top\mathbf{X}_i,\quad \log(\lambda_{2i})=\beta_2^\top\mathbf{X}_i,\quad \text{et}\quad \log(\mu)=\eta \tag{2.3}$$

où $\gamma\in\mathbb{R}^q$, $\beta_1,\beta_2\in\mathbb{R}^p$ et $\eta\in\mathbb{R}$ sont des paramètres de régression inconnus et le symbole $\top$ désigne l'opérateur de transposition. On pourrait également modéliser μ en fonction des covariables, par exemple par $\log(\mu)=\beta_3^\top\mathbf{X}_i$. Cette généralisation n'a aucun intérêt théorique (elle rend seulement les calculs plus "pénibles"), nous pensons également qu'en termes d'interprétation, il est plus pertinent d'avoir une covariance "fixe".

Soit $\theta=(\gamma^\top,\beta_1^\top,\beta_2^\top,\eta)^\top$ le vecteur de dimension k ($k=2p+q+1$) des paramètres du modèle ZIBP (2.1)-(2.2)-(2.3). La vraisemblance de θ basée sur un échantillon de n

observations indépendantes $(Y_{1i}, Y_{2i}, \mathbf{X}_i, \mathbf{W}_i),\ i = 1, \ldots, n$, est donnée par :

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} \left\{ \Big(\pi_i + (1-\pi_i) f_{BP}(0,0,\lambda_{1i},\lambda_{2i},\mu)\Big)^{a_i} \times \Big((1-\pi_i) f_{BP}(Y_{1i},Y_{2i},\lambda_{1i},\lambda_{2i},\mu)\Big)^{1-a_i} \right\}, \\ &= \prod_{i=1}^{n} \left\{ \left(\frac{e^{\gamma^\top \mathbf{W}_i} + e^{-(e^{\eta}+e^{\beta_1^\top \mathbf{X}_i}+e^{\beta_2^\top \mathbf{X}_i})}}{1+e^{\gamma^\top \mathbf{W}_i}} \right)^{a_i} \left(\frac{1}{1+e^{\gamma^\top \mathbf{W}_i}} \times \right. \right. \\ &\qquad \left. \left. e^{-(e^{\eta}+e^{\beta_1^\top \mathbf{X}_i}+e^{\beta_2^\top \mathbf{X}_i})} \times \sum_{s=0}^{Y_{1i}\wedge Y_{2i}} \frac{(e^{\eta})^s}{s!} \frac{(e^{\beta_1^\top \mathbf{X}_i})^{Y_{1i}-s}}{(Y_{1i}-s)!} \frac{(e^{\beta_2^\top \mathbf{X}_i})^{Y_{2i}-s}}{(Y_{2i}-s)!} \right)^{1-a_i} \right\}, \end{aligned}$$

où $a_i := 1_{(Y_{1i}=0, Y_{2i}=0)}$. D'où, la log-vraisemblance $\ell\ell_n(\theta) = \log\big(L_n(\theta)\big)$ est donnée par

$$\begin{aligned} \ell\ell_n(\theta) &= \sum_{i=1}^{n} \left\{ a_i \log\big(e^{\gamma^\top \mathbf{W}_i} + h_i(\theta)\big) - (1-a_i)\big(e^{\eta} + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}\big) \right. \\ &\qquad \left. + (1-a_i) \log\left(\sum_{s=0}^{Y_{1i}\wedge Y_{2i}} \frac{(e^{\eta})^s}{s!} \frac{(e^{\beta_1^\top \mathbf{X}_i})^{y_{1i}-s}}{(y_{1i}-s)!} \frac{(e^{\beta_2^\top \mathbf{X}_i})^{y_{2i}-s}}{(y_{2i}-s)!} \right) - \log(1+e^{\gamma^\top \mathbf{W}_i}) \right\}, \\ \ell\ell_n(\theta) &:= \sum_{i=1}^{n} \ell_i(\theta), \end{aligned}$$

où $h_i(\theta) = e^{-(e^{\eta}+e^{\beta_1^\top \mathbf{X}_i}+e^{\beta_2^\top \mathbf{X}_i})}$.

L'estimateur du Maximum de Vraisemblance (EMV) $\widehat{\theta}_n = \big(\widehat{\gamma}^\top, \widehat{\beta}_1^\top, \widehat{\beta}_2^\top, \widehat{\eta}\big)^\top$ de θ est la solution de l'équation du score de dimension k

$$U_n(\theta) = \frac{1}{\sqrt{n}} \frac{\partial \ell\ell_n(\theta)}{\partial \theta} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \ell_i(\theta)}{\partial \theta} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_i(\theta) = 0. \tag{2.4}$$

Les composantes du vecteur gradient sont de la forme suivante

$$\frac{\partial \ell_i(\theta)}{\partial \gamma_j} = \left(a_i \frac{e^{\gamma^\top \mathbf{W}_i}}{e^{\gamma_j^\top \mathbf{W}_i} + h_i(\theta)} - \frac{e^{\gamma_j^\top \mathbf{W}_i}}{1+e^{\gamma_j^\top \mathbf{W}_i}} \right) W_{ij},$$

$$\frac{\partial \ell_i(\theta)}{\partial \beta_{1,\ell}} = \left(- a_i \frac{e^{\beta_{1,\ell}^\top \mathbf{X}_i} h_i(\theta)}{e^{\gamma^\top \mathbf{W}_i} + h_i(\theta)} - (1-a_i) e^{\beta_{1,\ell}^\top \mathbf{X}_i} + (1-a_i) \frac{\sum_{s=0}^{Y_{1i}\wedge Y_{2i}} (Y_{1i}-s) g_i(s,\theta)}{\sum_{s=0}^{Y_{1i}\wedge Y_{2i}} g_i(s,\theta)} \right) X_{i\ell},$$

$$\frac{\partial \ell_i(\theta)}{\partial \beta_{2,\ell}} = \Bigg(- a_i \frac{e^{\beta_{2,\ell}^\top \mathbf{X}_i} h_i(\theta)}{e^{\gamma^\top \mathbf{W}_i} + h_i(\theta)} - (1-a_i) e^{\beta_{2,\ell}^\top \mathbf{X}_i} + (1-a_i) \frac{\sum_{s=0}^{Y_{1i} \wedge Y_{2i}} (Y_{2i} - s) g_i(s,\theta)}{\sum_{s=0}^{Y_{1i} \wedge Y_{2i}} g_i(s,\theta)} \Bigg) X_{i\ell},$$

et

$$\frac{\partial \ell_i(\theta)}{\partial \eta} = -a_i \frac{e^{\eta} h_i(\theta)}{e^{\gamma^\top \mathbf{W}_i} + h_i(\theta)} - (1-a_i) e^{\eta} + (1-a_i) \frac{\sum_{s=0}^{Y_{1i} \wedge Y_{2i}} s g_i(s,\theta)}{\sum_{s=0}^{Y_{1i} \wedge Y_{2i}} g_i(s,\theta)},$$

avec

$$g_i(s,\theta) = \frac{(e^{\eta})^s}{s!} \times \frac{(e^{\beta_1^\top \mathbf{X}_i})^{Y_{1i}-s}}{(Y_{1i}-s)!} \times \frac{(e^{\beta_2^\top \mathbf{X}_i})^{Y_{2i}-s}}{(Y_{2i}-s)!},$$

pour tout $i = 1, \ldots, n$, $j = 1, \ldots, q$ et $\ell = 1, \ldots, p$. En outre, l'équation d'estimation (2.4) peut être résolue par un algorithme de type Newton-Raphson.
Dans la section suivante, nous établissons la consistance et la normalité asymptotique de $\widehat{\theta}_n$.

2.3 Propriétés asymptotiques de l'EMV

Dans cette section, nous donnons d'abord quelques notations supplémentaires que nous utilisons dans la suite de ce chapitre. Ensuite, nous énonçons certaines conditions de régularité. Enfin, nous présentons les propriétés asymptotiques de l'estimateur $\widehat{\theta}_n$ de θ.

2.3.1 Notations et hypothèses de régularité

Pour chaque $i = 1, \ldots, n$, posons

$$A_i(\theta) = a_i \frac{e^{\gamma^\top \mathbf{W}_i}}{e^{\gamma^\top \mathbf{W}_i} + h_i(\theta)} - \frac{e^{\gamma^\top \mathbf{W}_i}}{1 + e^{\gamma^\top \mathbf{W}_i}},$$

$$B_{1,i}(\theta) = -a_i \frac{e^{\beta_1^\top \mathbf{X}_i} h_i(\theta)}{e^{\gamma^\top \mathbf{W}_i} + h_i(\theta)} - (1-a_i) e^{\beta_1^\top \mathbf{X}_i} + (1-a_i) \frac{\sum_{s=0}^{Y_{1i} \wedge Y_{2i}} (Y_{1i} - s) g_i(s,\theta)}{\sum_{s=0}^{Y_{1i} \wedge Y_{2i}} g_i(s,\theta)},$$

$$B_{2,i}(\theta) = -a_i \frac{e^{\beta_2^\top \mathbf{X}_i} h_i(\theta)}{e^{\gamma^\top \mathbf{W}_i} + h_i(\theta)} - (1-a_i) e^{\beta_2^\top \mathbf{X}_i} + (1-a_i) \frac{\sum_{s=0}^{Y_{1i} \wedge Y_{2i}} (Y_{2i} - s) g_i(s, \theta)}{\sum_{s=0}^{Y_{1i} \wedge Y_{2i}} g_i(s, \theta)},$$

et

$$C_i(\theta) = -a_i \frac{e^{\eta} h_i(\theta)}{e^{\gamma^\top \mathbf{W}_i} + h_i(\theta)} - (1-a_i) e^{\eta} + (1-a_i) \frac{\sum_{s=0}^{Y_{1i} \wedge Y_{2i}} s g_i(s, \theta)}{\sum_{s=0}^{Y_{1i} \wedge Y_{2i}} g_i(s, \theta)}.$$

Maintenant, nous énonçons les hypothèses de régularité sous lesquelles nous établissons les propriétés asymptotiques de l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$.

(A1) La vraie valeur du paramètre $\theta_0 := (\gamma_0^\top, \beta_{1,0}^\top, \beta_{2,0}^\top, \eta_0)^\top$ se trouve à l'intérieur d'un certain ensemble compact connu $\Theta \subset \mathbb{R}^k$.

(A2) $\mathbb{E}\Big[\big(\dot{\ell}_i(\theta)\big)\big(\dot{\ell}_i(\theta)\big)^\top\Big]$ est définie positive dans un voisinage de θ_0.

(A3) Dans un voisinage de θ_0, les dérivées première et seconde de $U_n(\theta)$ par rapport à θ sont uniformément bornées supérieurement par une fonction de $(Y_1, Y_2, \mathbf{X}, \mathbf{W})$, dont les espérances existent.

(A4) Pour chaque $i = 1, \ldots, n$, $\mathbb{E}\Big[\frac{\partial^2 \ell_i(\theta)}{\partial\theta\partial\theta^\top}\Big]$ est fini et est défini négatif dans un voisinage de θ_0. De plus, $-\frac{1}{\sqrt{n}} \frac{\partial U_n(\theta)}{\partial \theta^\top}$ converge vers une matrice définie positive $\Sigma(\theta)$ lorsque n tend vers l'infini.

Les hypothèses **(A1)** - **(A4)** sont classiques dans les modèles de régression à inflation de zéros (cf. Lukusa *et al.* (2016), Diallo *et al.* (2018), Lee *et al.* (2020)).
Dans ce qui suit, l'espace $\mathbb{R}^k$ des vecteurs colonnes de dimension k sera muni de la norme euclidienne $\|\cdot\|_2$ et l'espace des matrices réelles $(k \times k)$ sera muni de la norme $|||A|||_2 := max_{\|x\|_2=1} \|Ax\|_2$ (pour simplifier les notations, nous utiliserons $\|\cdot\|$ pour les deux normes).

Après avoir donné les hypothèses de régularité, nous sommes maintenant en mesure d'énoncer nos résultats.

2.3.2 Résultats asymptotiques pour l'EMV

Théorème 2.1 *Supposons que les hypothèses (**A1**) à (**A4**) sont vérifiées. Alors lorsque n tend vers l'infini, $\widehat{\theta}_n$ converge en probabilité vers θ_0.*

Preuve du théorème 2.1. Pour justifier la consistance de $\hat{\theta}_n$, on vérifie les conditions du théorème de la fonction inverse de Foutz (1977). Ces conditions sont prouvées dans une série de lemmes techniques.

D'abord, montrons que $\partial \ell\dot{\ell}_n(\theta)/\partial\theta^\top$ existe et est continue dans un voisinage ouvert de θ_0.

Pour justifier cela, on peut remarquer que $\ell\ell_n(\theta)$ est deux fois différentiable par rapport à θ et ses dérivées secondes sont continues. Donc $\frac{\partial^2 \ell\ell_n(\theta)}{\partial\theta\partial\theta^\top}$ existe et est continue dans un voisinage ouvert de θ_0.
La condition 1 est donc vérifiée. ■

Ensuite, montrons que $\frac{1}{n}\ell\dot{\ell}_n(\theta_0) = \frac{1}{n}\frac{\partial \ell\ell_n(\theta_0)}{\partial\theta}$ converge en probabilité vers 0 lorsque n tend vers l'infini. Pour justifier cela, on note

$$\begin{aligned}\frac{1}{n}\ell\dot{\ell}_n(\theta_0) \;=\; &\Big(\frac{1}{n}\sum_{i=1}^n W_{i1}A_i(\theta_0),\ldots,\frac{1}{n}\sum_{i=1}^n W_{iq}A_i(\theta_0),\frac{1}{n}\sum_{i=1}^n X_{i1}B_{1,i}(\theta_0),\ldots,\frac{1}{n}\sum_{i=1}^n X_{ip}B_{1,i}(\theta_0),\\ &\frac{1}{n}\sum_{i=1}^n X_{i1}B_{2,i}(\theta_0),\ldots,\frac{1}{n}\sum_{i=1}^n X_{ip}B_{2,i}(\theta_0),\frac{1}{n}\sum_{i=1}^n C_i(\theta_0)\Big)^\top.\end{aligned}$$

Puisque le vecteur de score est centré, il s'ensuit que $\operatorname{var}\Big[W_{i\ell}A_i(\theta_0)\Big] = \mathbb{E}\Big[W_{i\ell}^2A_i^2(\theta_0)\Big]$.
De plus, pour chaque $i=1,\ldots,n$, $\mathbb{E}\big[-\frac{\partial^2\ell_i(\theta_0)}{\partial\theta\partial\theta^\top}\big] = \mathbb{E}\big[\big(\dot{\ell}_i(\theta_0)(\dot{\ell}_i(\theta_0)\big)^\top\big]$.
Ainsi, par (**A2**), il s'ensuit que $\operatorname{var}\big(W_{i\ell}A_i(\theta_0)\big) < \infty$.
Par conséquent, par la loi faible des grands nombres, nous avons $\frac{1}{n}\sum_{i=1}^n W_{i\ell}A_i(\theta_0)$ converge en probabilité vers 0 lorsque $n\to\infty$, pour tout $\ell = 1,\ldots,q$.

Par des arguments similaires, nous montrons que pour tout $j=1,...,p$, $t\in\{1,2\}$, $\frac{1}{n}\sum_{i=1}^n X_{ij}B_{t,i}(\theta_0)$ et $\frac{1}{n}\sum_{i=1}^n C_i(\theta_0)$ converge en probabilité vers 0, lorsque n tend vers l'infini.

Ainsi, nous pouvons conclure que $\frac{1}{n}\ell\dot{\ell}_n(\theta_0)$ converge en probabilité vers $0_{(k,1)}$, lorsque n tend vers l'infini.
La condition 2 est donc vérifiée. ■

Troisièmement, montrons que $-\frac{1}{n}\frac{\partial^2 \ell\ell_n(\theta)}{\partial\theta\partial\theta^\top}$ converge uniformément en probabilité vers la fonction $\Sigma(\theta)$ dans un voisinage ouvert de θ_0.

Pour le vérifier, prenons $\mathcal{V}_{\theta_0}$ comme un voisinage ouvert de θ_0 et prenons $\theta \in \mathcal{V}_{\theta_0}$. Soit $H_{i,(j,\ell)}(Y_1, Y_2, \mathbf{X}_i, \mathbf{W}_i, \theta) = -\frac{\partial^2 \ell_i(\theta)}{\partial\theta_j \partial\theta_\ell}$ pour $j, \ell \in \{1, \dots 2p+q+1\}$.
Par l'hypothèse **(A3)**, il existe une fonction $N_i(.)$ telle que pour $\theta, \widetilde{\theta} \in \mathcal{V}_{\theta_0}$, $i \in \{1, \dots, n\}$, on a

$$\left|H_{i,(j,\ell)}(Y_{1i}, Y_{2i}, \mathbf{X}_i, \mathbf{W}_i, \theta) - H_{i,(j,\ell)}(Y_{1i}, Y_{2i}, \mathbf{X}_i, \mathbf{W}_i, \widetilde{\theta})\right| \leq N_i(Y_{1i}, Y_{2i}, \mathbf{X}_i, \mathbf{W}_i)\|\theta - \widetilde{\theta}\|,$$

aussi nous avons $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[N_i(Y_{1i}, Y_{2i}, \mathbf{X}_i, \mathbf{W}_i)] = O(1)$. De plus, par l'hypothèse **(A4)**, $-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ell_i(\theta)}{\partial\theta_j\partial\theta_\ell}$ converge en probabilité vers $\Sigma_{(j,\ell)}(\theta)$ lorsque n tend vers l'infini, où $\Sigma_{(j,\ell)}(\theta)$ désigne le (j,ℓ)-ième élément de $\Sigma(\theta)$, pour $j, \ell \in \{1, \dots, 2p+q+1\}$. Par conséquent, en utilisant le corrolaire 3.1 de Newey (1991) sous les hypothèses **(A1)** - **(A4)**, il s'ensuit que $-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ell_i(\theta)}{\partial\theta\partial\theta^\top}$ converge uniformément en probabilité vers une matrice définie positive $\Sigma(\theta)$ sur $\mathcal{V}_{\theta_0}$.
La condition 3 est donc vérifiée. ■

Les trois conditions du théorème de la fonction inverse de Foutz (1977) sont vérifiées. Ainsi, nous concluons que $\widehat{\theta}_n$ converge en probabilité vers θ_0.

Nous abordons maintenant la normalité asymptotique de l'estimateur du maximum de vraisemblance dans le modèle de régression ZIBP.

Théorème 2.2 *Supposons que les hypothèses **(A1)** à **(A4)** sont vérifiées. Alors* $\sqrt{n}(\hat{\theta}_n - \theta_0)$ *est distribué asymptotiquement comme une normale multivariée centrée et une matrice de covariance* $\Sigma(\theta_0)^{-1}$. *De plus, un estimateur consistant de la variance asymptotique est donné par* $(n^{-1/2}\partial\dot{\ell}\ell_n(\hat{\theta}_n)/\partial\theta^\top)^{-1}$.

Preuve du Théorème (2.2)

Pour prouver la normalité asymptotique, nous effectuons un développement de Taylor de la fonction score au voisinage de θ_0. On obtient

$$0 = \ell\dot{\ell}_n(\widehat{\theta}_n) = \ell\dot{\ell}_n(\theta_0) + \ell\ddot{\ell}_n(\widetilde{\theta}_n)(\widehat{\theta}_n - \theta_0),$$

où $\ell\ddot{\ell}_n(\theta) = \partial\ell\dot{\ell}_n/\partial\theta^\top$, $\widetilde{\theta}_n$ se situe entre $\widehat{\theta}_n$ et θ_0.

Ainsi,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \Big(-\frac{1}{n}\ell\ddot{\ell}_n(\widetilde{\theta}_n)\Big)^{-1} \times \Big(\frac{1}{\sqrt{n}}\ell\dot{\ell}_n(\theta_0)\Big).$$

Par le théorème (2.1), on a $\widehat{\theta}_n$ qui converge en probabilité vers θ_0. Comme $\widetilde{\theta}_n$ reste entre θ_0 et $\widehat{\theta}_n$, $\Big(-\frac{1}{n}\ell\ddot{\ell}_n(\widetilde{\theta}_n)\Big)$ converge en probabilité vers $\Sigma(\theta_0)$.

De plus, par le théorème central limite, on a

$$\frac{1}{\sqrt{n}}\ell\dot{\ell}_n(\theta_0) \xrightarrow{loi} \mathcal{N}(0, \Sigma(\theta_0).$$

Ainsi, en appliquant le lemme de Slutsky, on a:

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{loi} \mathcal{N}(0, \Sigma(\theta_0)^{-1}).$$

2.4 Etudes de simulations

Dans cette section, nous évaluons les performances de l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$ en échantillon fini.

2.4.1 Expériences numériques par simulation

Nous générons des données à partir du modèle de régression ZIBP suivant :

$$\begin{cases} \text{logit}(\pi_i) = \gamma^\top \mathbf{W}_i \\ \log(\lambda_{1i}) = \beta_1^\top \mathbf{X}_i, \;\; \log(\lambda_{2i}) = \beta_2^\top \mathbf{X}_i, \;\; \text{et} \;\; \log(\mu) = \eta, \end{cases}$$

avec $\mathbf{X}_i = (1, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6})^\top$ et $\mathbf{W}_i = (1, W_{i2}, W_{i3}, W_{i4}, W_{i5})$ où

- les covariables $X_{i1} = 1$ et $X_{i2}, \ldots, X_{i6}$ sont générées à partir de la distribution normale $\mathcal{N}(0, 0.1)$, la distribution uniforme $\mathcal{U}_{[-1,1]}$, la distribution exponentielle $\mathcal{E}(1)$, la distribution de Bernoulli $\mathcal{B}(1, 0.8)$ et la distribution de Bernoulli $\mathcal{B}(1, 0.4)$ respectivement.
- $W_{i1} = 1$ et W_{i3}, W_{i4}, W_{i5} sont générées à partir de la distribution de Bernoulli $\mathcal{B}(1, 0.3)$, la distribution normale $\mathcal{N}(-1, 1)$ et la distribution normale $\mathcal{N}(1, 0.5)$ respectivement,
- on prend $W_{i2} = X_{i2}$.
- Les paramètres de regression β_1, β_2 et η sont choisis comme suit: $\beta_1 = (-0.3, 0.85, 0.1, 0.25, -0.1, -0.05)^\top$, $\beta_2 = (0.8, -0.74, -0.1, -0.1, 0.15, -0.1)^\top$,

$\eta = 0.4$.

– Nous considérons trois (03) cas pour le paramètre de régression γ:

Cas 1 $\gamma = (-0.55, -0.75, -1, 0.45, 0)^\top$ pour 25% d'inflation de zéros,

Cas 2 $\gamma = (-0.25, -0.4, 0.8, 0.45, 0)^\top$ pour 50% d'inflation de zéros.

Cas 3 $\gamma = (0.2, -0.27, 0.85, 0.23, 0.91)^\top$ pour 65% d'inflation de zéros.

En utilisant les valeurs de γ des cas 1, cas 2 et cas 3, les proportions moyennes d'inflation de zéros dans les ensembles de données simulées sont respectivement de 25 %, 50 % et 65%.

Pour chaque combinaison `[taille de l'échantillon × proportion de l' inflation zéro]` des paramètres de conception, nous simulons $N = 1000$ échantillons et nous calculons l'estimation du maximum de vraisemblance (EMV) $\widehat{\theta}_n$ de $\theta = (\gamma^\top, \beta_1^\top, \beta_2^\top, \eta)^\top$. Plusieurs auteurs ont développé des algorithmes d'estimation de type EM dans des modèles d'inflation de zéros (voir Wang *et al.* (2003)). D'autres auteurs effectuent une maximisation directe à l'aide d'algorithmes de Newton-Raphson, voir Diallo *et al.* (2017) et Diallo *et al.* (2019). Dans notre étude de simulation, nous utilisons un algorithme de type Newton-Raphson mis en œuvre dans le `R` package `maxLik` développé par Henningsen et Toomet (2011).

2.4.2 Résultats

Pour chaque combinaison `taille de l'échantillon × proportion d'inflation de zéros` des paramètres de simulation, nous calculons l'EMV $\widehat{\theta}_n$, le biais moyen et le biais relatif moyen (exprimé en pourcentage) des estimations $\widehat{\gamma}_{i,n}$, $\widehat{\beta}_{1,j,n}$, $\widehat{\beta}_{2,k,n}$ et $\widehat{\eta}_n$ sur les N échantillons simulés. Par exemple, le biais relatif de $\widehat{\gamma}_{i,n}$ est obtenu par $\frac{1}{N}\sum_{t=1}^{N}\frac{\widehat{\gamma}_{i,n}^{(t)}-\gamma_i}{\gamma_i} \times 100$ où $\widehat{\gamma}_{i,n}^{(t)}$ désigne l'EMV de γ_i dans le t-ième échantillon simulé. Nous calculons également l'erreur standard (SE) moyenne, l'écart type (SD) empirique et l'erreur quadratique moyenne (RMSE) pour chaque $\widehat{\gamma}_{i,n}$ $(i = 1, \ldots, 5)$, $\widehat{\beta}_{1,j,n}$, $\widehat{\beta}_{2,k,n}$ $(j, k = 1, \ldots, 6)$ et $\widehat{\eta}_n$. Le SE est calculé comme la moyenne des erreurs standard sur les N échantillons simulés. Par exemple, pour $\widehat{\gamma}_{i,n}$, la SE est obtenue par $\frac{1}{N}\sum_{t=1}^{N} s.e.\big(\widehat{\gamma}_{i,n}^{(t)}\big)$, tandis que SD (respectivement RMSE) est la racine carrée de la variance empirique (respectivement MSE) de $\big(\widehat{\gamma}_{i,n}^{(1)}, \ldots, \widehat{\gamma}_{i,n}^{(N)}\big)$. De plus, nous fournissons la probabilité de couverture (CP) empirique et la longueur moyenne des intervalles de confiance de Wald à 95 % pour les estimateurs de γ_i, $\beta_{1,j}$, $\beta_{2,k}$ et η. Les résultats sont donnés dans les tableau 2.1 (pour le cas 1), le tableau 2.2 (pour le cas 2) et le tableau 2.3 (pour le cas 3). Les tableaux 2.1, 2.2 et 2.3 fournissent les résultats pour 25%, 50% et 65% d'inflation de zéros respectivement. Dans chaque configuration, on

prendra comme taille des échantillons $n = 500$ et $n = 2000$.

D'après les résultats obtenus, nous observons que le biais et le biais relatif sont assez faibles. Ensuite, le biais, le biais relatif, le SE, le SD, le RMSE et le ℓ(CI) de tous les estimateurs diminuent lorsque la taille de l'échantillon augmente. En outre, pour γ_i, $\beta_{1,j}$, $\beta_{2,k}$ et η, les probabilités de couverture empirique sont proches du niveau de confiance nominal dans toutes les configurations. D'autre part, nous observons que l'EMV des paramètres $\beta_{1,j}$s, $\beta_{2,k}$s et η_n (respectivement, γ_is) est plus performante lorsque la proportion d'inflation de zéros diminue (respectivement, augmente).

Pour voir l'impact du paramètre η sur la qualité des estimations, nous avons réalisé des simulations pour les tailles des échantillons $n = 500$, 50% d'inflation de zéros avec $\eta = 0.18$, $\eta = 0.4$ et $\eta = 0.75$ respectivement. Le tableau 2.4 fournit les résultats de ces simulations. L'analyse du tableau 2.4, révèle qu'autant la corrélation entre les variables réponses est forte autant les estimations sont meilleures.

Pour évaluer la qualité de l'approximation gaussienne énoncée dans le théorème (2.2), nous fournissons des graphiques Q-Q normaux des estimations et des histogrammes des estimations normalisées de $(\widehat{\gamma}_{i,n} - \gamma_i)/\text{s.e.}(\widehat{\gamma}_{i,n})$, $j = 1, \ldots, 5$, $(\widehat{\beta}_{1,j,n} - \beta_{1,j})/\text{s. e.}(\widehat{\beta}_{1,j,n})$, $j = 1, \ldots, 6$, $(\widehat{\beta}_{2,k,n} - \beta_{2,k})/\text{s.e. } (\widehat{\beta}_{2,k,n})$, $k = 1, \ldots, 6$ et $(\widehat{\eta}_n - \eta_l)/\text{s.e.}(\widehat{\eta}_n)$. Nous fournissons ces graphiques pour $n = 2000$ et une proportion moyenne d'inflation de zéros dans l'échantillon égale à 25%. Les figures 2.1 à 2.4 fournissent des graphiques Q-Q plot pour $(\widehat{\gamma}_{1,n}, \ldots, \widehat{\gamma}_{5,n})$, $(\widehat{\beta}_{1,1,n}, \ldots, \widehat{\beta}_{1,6,n})$, $(\widehat{\beta}_{2,1,n}, \ldots, \widehat{\beta}_{2,6,n})$, $\widehat{\eta}_n$, respectivement. De plus, les figures 2.5 à 2.8 fournissent les histogrammes des valeurs normalisées de $(\widehat{\gamma}_{1,n}, \ldots, \widehat{\gamma}_{5,n})$, $(\widehat{\beta}_{1,1,n}, \ldots, \widehat{\beta}_{1,6,n})$, $(\widehat{\beta}_{2,1,n}, \ldots, \widehat{\beta}_{2,6,n})$, $\widehat{\eta}_n$, respectivement). Les graphiques des autres scénarios simulés sont similaires et ne sont donc pas présentés. D'après ces figures, il apparaît que l'approximation gaussienne de la distribution de l'EMV dans le ZIBP est raisonnablement satisfaite, même lorsque la taille de l'échantillon est modérée et que la proportions d'inflation de zéros est importante (atteint 65%).

Tableau 2.1 – Résultats de la simulation pour $N = 1000$ replications, tailles des échantillons $n = 500$ (au dessus) et $n = 2000$ (en dessous) avec 25% d'inflation de zéros.

		$\widehat{\gamma}$					$\widehat{\beta_1}$						$\widehat{\beta_2}$						$\widehat{\eta}$
n		$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$	$\widehat{\gamma}_4$	$\widehat{\gamma}_5$	$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{1,6}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$	$\widehat{\beta}_{2,6}$	$\widehat{\eta}$
	biais	-0.0112	0.0232	-0.0362	0.0086	0.0098	-0.0602	-0.0149	0.0010	-0.0040	0.0266	0.0094	-0.0170	-0.0340	-0.0023	-0.0067	0.0113	0.0002	0.0052
	biais rel.	2.0278	3.0954	3.6227	1.9176	-	20.0556	1.7503	1.0411	1.5938	26.6454	18.7259	2.1222	4.5889	2.2930	6.7037	7.5364	0.2397	1.3125
	SD	0.1972	1.1554	0.3052	0.1184	0.2318	0.2426	0.8371	0.1498	0.0625	0.2016	0.1721	0.1275	0.4329	0.0731	0.0493	0.1179	0.0917	0.0786
500	SE	0.1951	1.2121	0.2987	0.1190	0.2278	0.2380	0.8632	0.1445	0.0618	0.2042	0.1707	0.1243	0.4410	0.0748	0.0498	0.1161	0.0892	0.0765
	RMSE	0.2775	1.6743	0.4285	0.1681	0.3251	0.3451	1.2023	0.2080	0.0879	0.2881	0.2426	0.1788	0.6188	0.1046	0.0704	0.1658	0.1279	0.1098
	CP	0.9560	0.9530	0.9440	0.9560	0.9450	0.9460	0.9610	0.9490	0.9520	0.9580	0.9520	0.9430	0.9480	0.9500	0.9590	0.95300	0.9350	0.9500
	ℓ(CI)	0.7639	4.6888	1.1650	0.4658	0.8925	0.9237	3.3416	0.5632	0.2378	0.7926	0.6655	0.4854	1.7225	0.2928	0.1944	0.4535	0.3493	0.2977
	biais	-0.0043	0.0182	-0.0064	0.0026	0.0032	-0.0095	-0.0185	-0.0001	-0.0007	0.0043	-0.0049	0.0013	-0.0105	-0.0010	-0.0025	-0.0018	-0.0019	0.0020
	biais rel.	0.7788	-2.4311	0.6356	0.5684	-	3.1738	-2.1723	-0.0677	-0.2777	-4.2934	9.8046	0.1637	1.4257	0.9654	2.4769	-1.2300	1.8964	0.5013
	SD	0.0942	0.5721	0.1456	0.0571	0.1094	0.1152	0.3886	0.0698	0.0279	0.0977	0.0849	0.0605	0.2169	0.0361	0.0241	0.0560	0.0437	0.0377
2000	SE	0.0962	0.5929	0.1448	0.0585	0.1124	0.1121	0.4126	0.0696	0.0282	0.0962	0.0822	0.0607	0.2163	0.0369	0.0242	0.0566	0.0440	0.0379
	RMSE	0.1347	0.8239	0.2054	0.0818	0.1568	0.1610	0.5669	0.0985	0.0396	0.1371	0.1183	0.0856	0.3064	0.0516	0.0342	0.0796	0.0621	0.0535
	CP	0.9470	0.9530	0.9510	0.9640	0.9560	0.9320	0.9560	0.9530	0.9520	0.9440	0.9400	0.9440	0.9500	0.9550	0.9480	0.9530	0.9530	0.9530
	ℓ(CI)	0.3770	2.3005	0.5668	0.2293	0.4406	0.4383	1.6088	0.2725	0.1098	0.3762	0.3220	0.2376	0.8466	0.1447	0.0947	0.2218	0.1725	0.1485

SD: écart-type empirique. SE: erreur type moyenne. RMSE: racine carrée de l'erreur type moyenne. CP: probabilité de couverture empirique des intervalles de confiance à 95 %. ℓ(CI): longueur moyenne des intervalles de confiance.

Tableau 2.2 – Résultats de la simulation pour $N = 1000$ replications, tailles des échantillons $n = 500$ (au dessus) et $n = 2000$ (en dessous) avec 50% d'inflation de zéros.

		$\widehat{\gamma}$					$\widehat{\beta_1}$						$\widehat{\beta_2}$						$\widehat{\eta}$
n		$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$	$\widehat{\gamma}_4$	$\widehat{\gamma}_5$	$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{1,6}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$	$\widehat{\beta}_{2,6}$	$\widehat{\eta}$
	biais	0.0041	-0.0002	0.0132	0.0078	0.0019	-0.0759	0.0825	0.0071	-0.0052	0.0193	-0.0043	-0.0210	0.0077	-0.0070	-0.0114	0.0134	-0.0005	0.0076
	biais rel.	1.6350	0.0431	1.6454	1.7245	-	25.3001	9.7067	7.0622	-2.0873	-19.2686	8.5766	-2.6278	-1.0451	6.9572	11.4178	8.9470	0.5032	1.8895
	SD	0.1827	1.0000	0.2103	0.1026	0.1852	0.3116	1.0775	0.1819	0.0800	0.2651	0.2292	0.1642	0.5596	0.0942	0.0645	0.1461	0.1114	0.0995
500	SE	0.1757	0.9939	0.2115	0.1000	0.1903	0.3075	1.1625	0.1864	0.0823	0.2633	0.2210	0.1577	0.5655	0.0945	0.0636	0.1473	0.1127	0.0962
	RMSE	0.2535	1.4096	0.2985	0.1435	0.2655	0.4441	1.5868	0.2605	0.1148	0.3740	0.3183	0.2286	0.7954	0.1336	0.0913	0.2079	0.1585	0.1386
	CP	0.9420	0.9480	0.9560	0.9380	0.9620	0.9460	0.9470	0.9670	0.9590	0.9500	0.9440	0.9430	0.9410	0.9470	0.9570	0.9570	0.9570	0.9330
	ℓ(CI)	0.6885	3.8698	0.8286	0.3918	0.7458	1.1868	4.3976	0.7240	0.3137	1.0169	0.8583	0.6146	2.1952	0.3696	0.2473	0.5742	0.4411	0.3723
	biais	-0.0024	0.0174	0.0028	0.0003	-0.0012	-0.0192	0.0017	-0.0038	0.0003	0.0040	-0.0024	-0.0039	-0.0054	-0.0002	-0.0026	0.0021	0.0001	0.0031
	biais rel.	-0.9420	-4.3542	0.3477	0.0727	-	6.3964	0.1977	-3.7725	0.1000	-4.0352	4.7782	-0.4815	0.7335	0.1866	2.6356	1.3974	-0.0507	0.7705
	SD	0.0838	0.4657	0.0992	0.0505	0.0937	0.1471	0.4906	0.0921	0.0362	0.1242	0.1041	0.0757	0.2692	0.0472	0.0315	0.0714	0.0550	0.0472
2000	SE	0.0870	0.4904	0.1046	0.0495	0.0943	0.1417	0.5365	0.0878	0.0360	0.1218	0.1038	0.0761	0.2751	0.0463	0.0304	0.0712	0.0552	0.0474
	RMSE	0.1208	0.6764	0.1441	0.0707	0.1329	0.2051	0.7269	0.1273	0.0510	0.1739	0.1470	0.1074	0.3849	0.0661	0.0438	0.1008	0.0779	0.0670
	CP	0.9630	0.9600	0.9490	0.9470	0.9500	0.9390	0.9610	0.9320	0.9440	0.9480	0.9510	0.9570	0.9560	0.9440	0.9390	0.9440	0.9530	0.9530
	ℓ(CI)	0.3410	1.9134	0.4098	0.1939	0.3696	0.5534	2.0763	0.3436	0.1398	0.4758	0.4060	0.2980	1.0741	0.1814	0.1191	0.2788	0.2162	0.1855

SD: écart-type empirique. SE: erreur type moyenne. RMSE: racine carrée de l'erreur type moyenne. CP: probabilité de couverture empirique des intervalles de confiance à 95 %. ℓ(CI): longueur moyenne des intervalles de confiance.

Tableau 2.3 – Résultats de la simulation pour $N = 1000$ replications, tailles des échantillons $n = 500$ (au dessus) et $n = 2000$ (en dessous) avec 65% d'inflation de zéros.

		$\widehat{\gamma}$					$\widehat{\beta_1}$						$\widehat{\beta_2}$						$\widehat{\eta}$
n		$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$	$\widehat{\gamma}_4$	$\widehat{\gamma}_5$	$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{1,6}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$	$\widehat{\beta}_{2,6}$	$\widehat{\eta}$
	biais	-0.0003	-0.0178	0.0128	0.0058	0.0010	-0.1202	0.1121	0.0065	0.0032	0.0254	-0.0014	-0.0307	-0.0160	-0.0029	-0.0095	0.0124	-0.0036	0.0126
	biais rel.	-0.1343	-7.7933	1.2208	1.1706	1.1706	40.0577	13.1832	6.5331	1.2888	-25.4025	2.7836	-3.8357	2.1614	2.8981	9.5081	8.2636	3.6006	3.1549
	SD	0.1818	0.9745	0.2315	0.1027	0.2111	0.3764	1.3514	0.2376	0.1019	0.3327	0.2665	0.1929	0.6726	0.1173	0.0779	0.1785	0.1341	0.1126
500	SE	0.1783	1.0117	0.2323	0.1010	0.2009	0.3803	1.3516	0.2280	0.1007	0.3265	0.2689	0.1894	0.6628	0.1135	0.0764	0.1771	0.1347	0.1141
	RMSE	0.2515	1.4045	0.2284	0.1414	0.2615	0.5483	1.9141	0.3292	0.1433	0.4667	0.3785	0.2720	0.9442	0.1632	0.1095	0.2517	0.1901	0.1608
	CP	0.9470	0.9630	0.9610	0.9480	0.9460	0.9560	0.9450	0.9460	0.9540	0.9610	0.9630	0.9490	0.948	0.9470	0.9410	0.9540	0.9480	0.9230
	ℓ(CI)	0.6984	3.9563	0.9096	0.3958	0.7872	1.4548	5.1733	0.8818	0.3818	1.2471	1.0406	0.7360	2.5816	0.4436	0.2962	0.6881	0.5269	0.4396
	biais	0.0010	0.0172	0.0044	0.0015	-0.0009	-0.0131	0.0177	0.0015	-0.0018	0.0057	-0.0023	-0.0051	0.0061	-0.0011	-0.0016	0.0033	-0.0010	0.0000
	biais rel.	0.4174	-6.8058	0.3440	0.0257	0.0652	4.3763	2.0854	1.4641	-0.7037	-5.7171	4.5432	-0.6417	-0.8204	1.1100	1.5653	2.2083	0.9606	0.007
	SD	0.0833	0.4156	0.0953	0.0491	0.0934	0.1424	0.5056	0.0901	0.0361	0.1208	0.1061	0.0781	0.2768	0.0475	0.0294	0.0721	0.0574	0.0480
2000	SE	0.0871	0.4860	0.1046	0.0490	0.0941	0.1412	0.5319	0.0875	0.0358	0.1213	0.1035	0.0763	0.2741	0.0463	0.0303	0.0713	0.0552	0.0477
	RMSE	0.1207	0.6704	0.1435	0.0697	0.1324	0.2009	0.7339	0.1256	0.0508	0.1713	0.1482	0.1092	0.3895	0.0663	0.0423	0.1014	0.0797	0.0677
	CP	0.9430	0.9520	0.9450	0.9500	0.9480	0.9550	0.9520	0.9470	0.9530	0.9520	0.9510	0.9410	0.9470	0.9380	0.9530	0.9420	0.9440	0.9430
	ℓ(CI)	0.3409	1.8970	0.4090	0.1938	0.3696	0.5516	2.0563	0.3424	0.1389	0.4739	0.4048	0.2986	1.0698	0.1813	0.1186	0.2791	0.2164	0.1865

SD: écart-type empirique. SE: erreur type moyenne. RMSE: racine carrée de l'erreur type moyenne. CP: probabilité de couverture empirique des intervalles de confiance à 95 %. ℓ(CI): longueur moyenne des intervalles de confiance.

Tableau 2.4 – Résultats de la simulation pour $N = 1000$ replications, tailles des échantillons $n = 500$, 50% d'inflation de zéros avec $\eta = 0.18$ (en dessus), $\eta = 0.4$ (au milieu) et $\eta = 0.75$ (en dessous)

		$\widehat{\gamma}$					$\widehat{\beta_1}$						$\widehat{\beta_2}$						$\widehat{\eta}$
n		$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$	$\widehat{\gamma}_4$	$\widehat{\gamma}_5$	$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{1,6}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$	$\widehat{\beta}_{2,6}$	$\widehat{\eta}$
	biais	-0.0009	-0.0001	0.0308	0.0076	-0.0002	-0.0002	-0.0792	0.0233	0.0116	0.0000	0.0289	-0.0089	-0.0201	-0.0079	0.0016	-0.0126	0.0156	-0.0050
	SD	0.1885	0.9774	0.2111	0.1048	0.1996	0.3028	1.0384	0.1903	0.0772	0.2641	0.2103	0.1562	0.5632	0.0955	0.0629	0.1508	0.1116	0.1178
$\eta = 0.18$	SE	0.1765	0.9925	0.2124	0.1007	0.1912	0.3021	1.0721	0.1793	0.0792	0.2570	0.2125	0.1562	0.5484	0.0930	0.0626	0.1457	0.1112	0.1160
	RMSE	0.2581	1.3927	0.3010	0.1455	0.2764	0.4349	1.4924	0.2616	0.1106	0.3696	0.2991	0.2218	0.7860	0.1333	0.0896	0.2102	0.1576	0.1655
	CP	0.9560	0.9530	0.9440	0.9560	0.9450	0.9460	0.9610	0.9490	0.9520	0.9580	0.9520	0.9430	0.9480	0.9500	0.9590	0.95300	0.9350	0.9500
	biais	0.0041	-0.0002	0.0132	0.0078	0.0019	-0.0759	0.0825	0.0071	-0.0052	0.0193	-0.0043	-0.0210	0.0077	-0.0070	-0.0114	0.0134	-0.0005	0.0076
	SD	0.1827	1.0000	0.2103	0.1026	0.1852	0.3116	1.0775	0.1819	0.0800	0.2651	0.2292	0.1642	0.5596	0.0942	0.0645	0.1461	0.1114	0.0995
$\eta = 0.4$	SE	0.1757	0.9939	0.2115	0.1000	0.1903	0.3075	1.1625	0.1864	0.0823	0.2633	0.2210	0.1577	0.5655	0.0945	0.0636	0.1473	0.1127	0.0962
	RMSE	0.2535	1.4096	0.2985	0.1435	0.2655	0.4441	1.5868	0.2605	0.1148	0.3740	0.3183	0.2286	0.7954	0.1336	0.0913	0.2079	0.1585	0.1386
	CP	0.9420	0.9480	0.9560	0.9380	0.9620	0.9460	0.9470	0.9670	0.9590	0.9500	0.9440	0.9430	0.9410	0.9470	0.9570	0.9570	0.9570	0.9330
	biais	0.0010	-0.0002	0.0102	0.0064	0.0012	-0.0743	0.0817	0.0059	-0.0041	0.0173	-0.0040	-0.0083	0.0072	-0.0061	-0.0113	0.0112	-0.0008	0.0048
	SD	0.1731	0.9002	0.1703	0.1021	0.1783	0.2709	1.0003	0.1764	0.0733	0.2369	0.2187	0.1597	0.5431	0.0942	0.0645	0.1461	0.1114	0.0995
$\eta = 0.75$	SE	0.1698	0.9836	0.2034	0.9275	0.1873	0.3009	1.1561	0.1748	0.0820	0.2631	0.2206	0.1561	0.5630	0.0939	0.0629	0.1462	0.1121	0.0960
	RMSE	0.2528	1.3978	0.2981	0.1429	0.2649	0.4440	1.5863	0.2601	0.1142	0.3732	0.3179	0.2276	0.7950	0.1334	0.0912	0.2071	0.1582	0.1381
	CP	0.9450	0.9490	0.9550	0.9420	0.9600	0.9470	0.9480	0.9670	0.9570	0.9500	0.9460	0.9470	0.9460	0.9490	0.9560	0.9540	0.9560	0.9340

SD: écart-type empirique. SE: erreur type moyenne. RMSE: racine carrée de l'erreur type moyenne. CP: probabilité de couverture empirique des intervalles de confiance à 95 %.

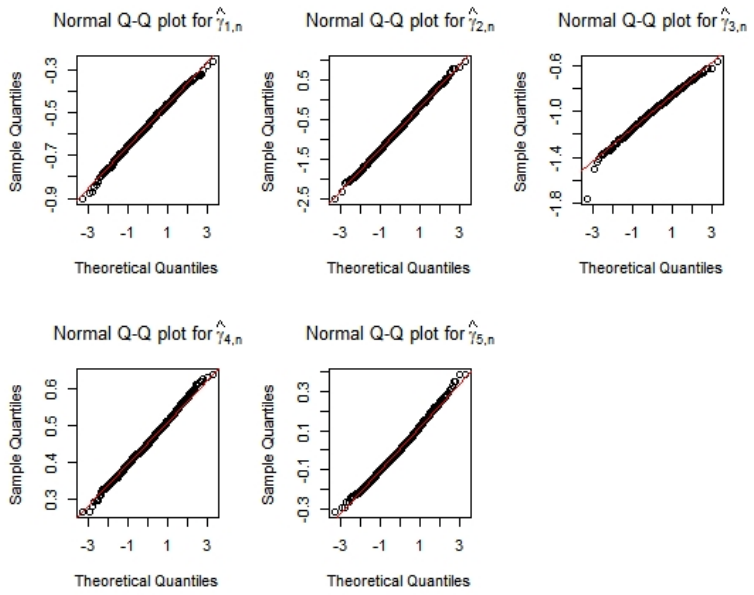

Figure 2.1 – QQ-plot normaux pour $\widehat{\gamma}_{1,n}, \ldots, \widehat{\gamma}_{5,n}$ avec $n = 2000$ et 25% d'inflation de zéros.

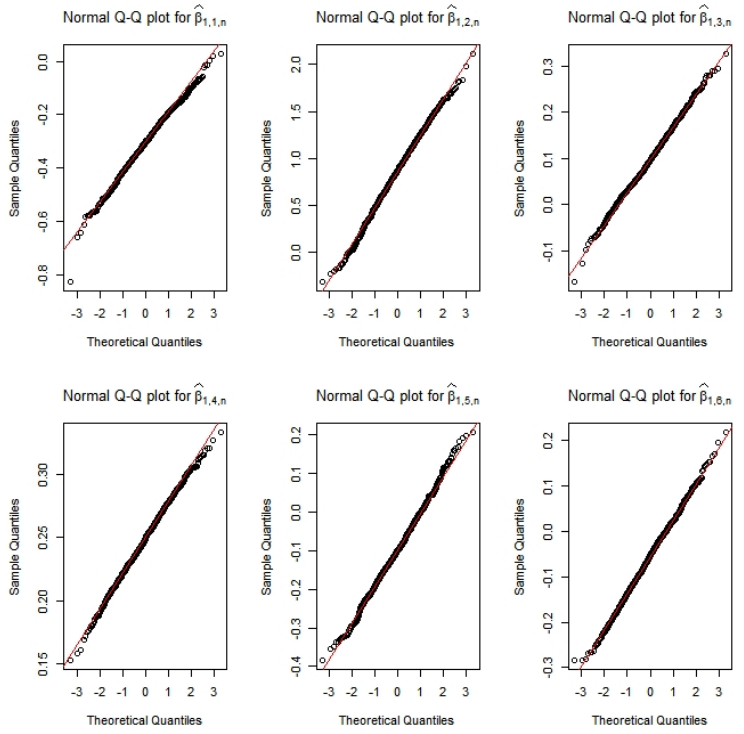

Figure 2.2 – QQ-plot normaux pour $\widehat{\beta}_{1,1,n}, \ldots, \widehat{\beta}_{1,6,n}$ avec $n = 2000$ et 25% d'inflation de zéros.

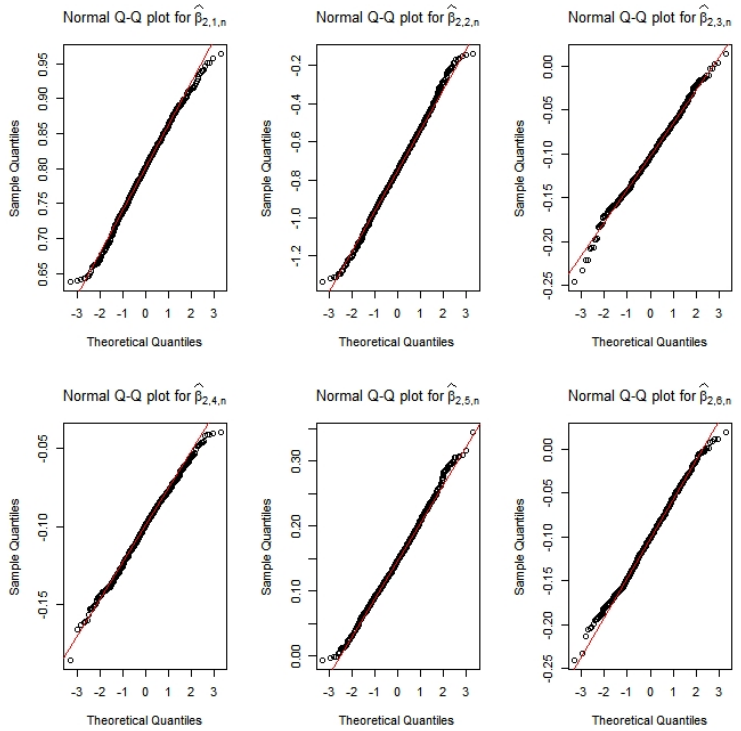

Figure 2.3 – QQ-plot normaux pour $\widehat{\beta}_{2,1,n}, \ldots, \widehat{\beta}_{2,6,n}$ avec $n = 2000$ et 25% d'inflation de zéros.

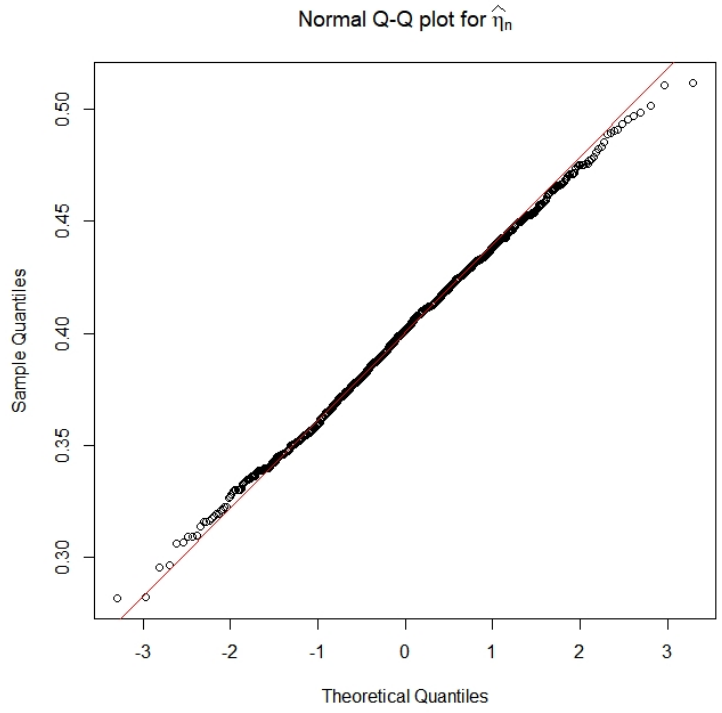

Figure 2.4 – QQ-plot normal pour $\widehat{\eta}_n$ avec $n = 2000$ et 25% d'inflation de zéros.

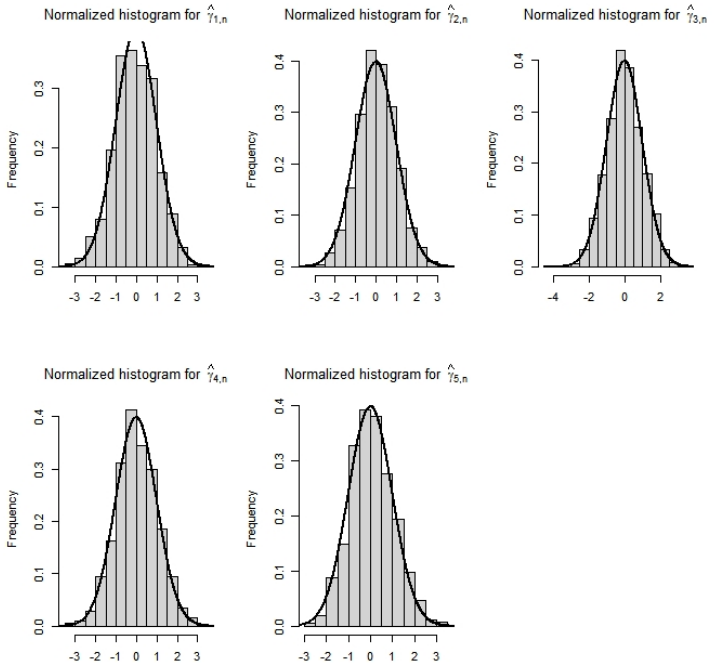

Figure 2.5 – Histogrammes des estimations normalisées $(\widehat{\gamma}_{j,n} - \gamma_j)/\text{s.e.}(\widehat{\gamma}_{j,n})$, $j = 1, \ldots, 5$ avec $n = 2000$ et 25% d'inflation de zéros.

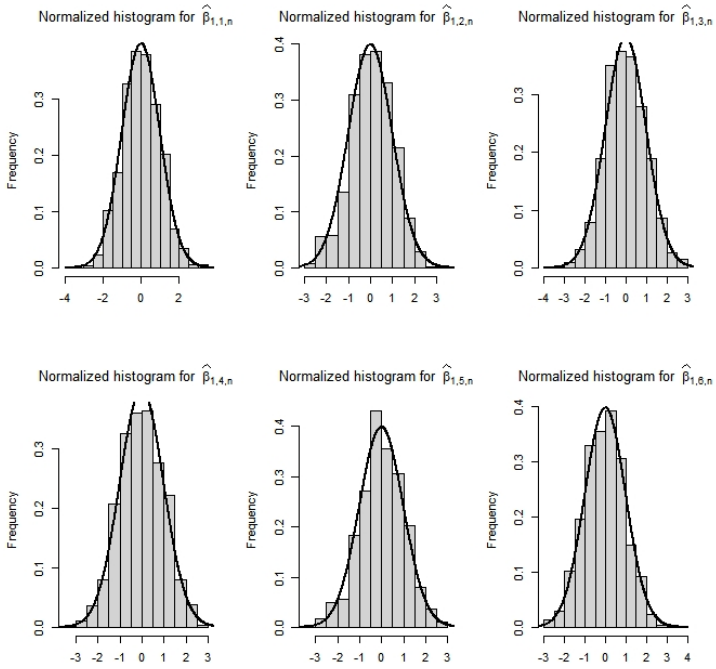

Figure 2.6 – Histogrammes des estimations normalisées $(\widehat{\beta}_{1,j,n} - \beta_{1,j})/\text{s.e.}(\widehat{\beta}_{1,j,n})$, $j = 1, \ldots, 6$ avec $n = 2000$ et 25% d'inflation de zéros.

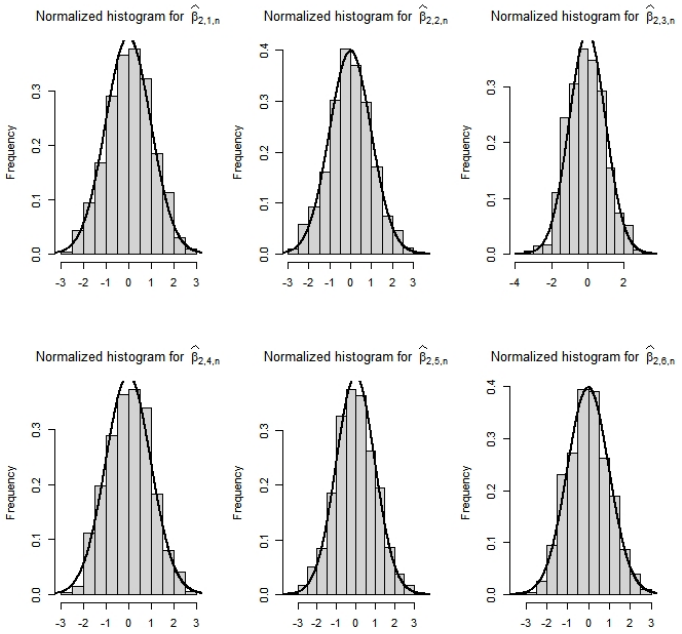

Figure 2.7 – Histogrammes des estimations normalisées $(\widehat{\beta}_{2,j,n} - \beta_{2,j})/\text{s.e.}(\widehat{\beta}_{2,j,n})$, $j = 1, \ldots, 6$ avec $n = 2000$ et 25% d'inflation de zéros.

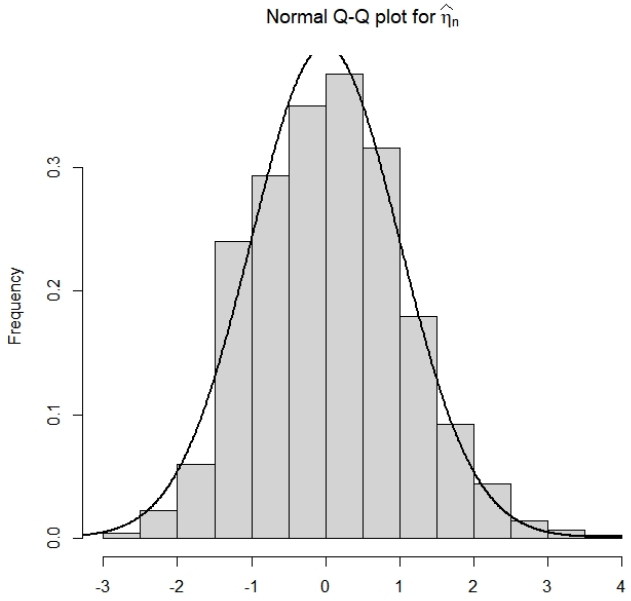

Figure 2.8 – Histogramme des estimations normalisées $(\widehat{\eta}_n - \eta_j)/\text{s.e.}(\widehat{\eta}_n)$, avec $n = 2000$ et 25% d'inflation de zéros.

2.5 Application

2.5.1 Description et modélisation des données

Dans cette section, nous décrivons une application du modèle de régression ZIBP pour analyser les données d'utilisation des services de santé par les personnes âgées aux États-Unis. Nous utilisons les données de l'enquête nationale sur les dépenses médicales (NMES) menée en 1987-1988 aux États-Unis. Ces données d'enquête de santé contiennent un ensemble de 4406 observations d'individus âgés de 66 ans et plus. Cet ensemble de données a été examiné par Deb et Trivedi (1997), Diallo *et al.* (2017), Ali *et al.* (2020) et est disponible dans le package `AER` du logiciel `R` sous le nom de "NMES1988". Dans ces données, nous considérons conjointement deux mesures d'utilisation des services de santé: le nombre `ofnd` de consultations avec un professionnel de santé non-médecin en cabinet et le nombre `opnd` de consultations avec un professionnel de santé non-médecin en ambulatoire.

Soient Y_1 le nombre de consultations d'un professionnel de santé non médecin en cabinet et Y_2 le nombre de consultations externes d'un professionnel de santé non médecin. Les distributions des fréquences des variables Y_1 et Y_2 sur cet échantillon sont données par les figures 2.9 et 2.10 respectivement. Le tableau 2.5 donne la proportion de zéros observés pour chacune des deux variables. Les fortes proportions de zéros indiqués dans le Tableau 2.5 pour les variables Y_1 et Y_2 et les Figures 2.9 et 2.10 suggèrent une situation d'inflation de zéros. Le tableau 2.5 donne aussi la proportion d'observations $(0;0)$ pour le couple de variables (Y_1, Y_2). Ce tableau indique qu'une proportion importante de patients n'ont pas eu recours à un professionnel de santé non-médecin durant la période de l'enquête. Plus précisément, la fréquence des individus avec zéro consultation simultanément dans (`ofnd` et `opnd`) est de 59,03314%. Une représentation graphique de la distribution des fréquences du couple (Y_1, Y_2) est donnée par la figure 2.11.

variable	Y_1 (ofnp)	Y_2 (opnp)	(Y_1, Y_2) (ofnp, opnp)
proportion de zéros en %	68.17975	83.84022	59.03314

Tableau 2.5 – Proportion de zéros observés pour Y_1, Y_2 et (Y_1, Y_2).

Les tests effectués avec la fonction `cor.test` du `package stats` du logiciel `R` sur les variables `ofnd` et `opnd` montre qu'elles sont corrélées. Nous proposons donc d'utiliser le modèle ZIBP pour étudier les déterminants de l'utilisation des soins de santé dans ce jeu de données. Certaines covariables ont été enregistrées sur chaque individu. Elles comprennent : (i) des variables socio-économiques : le sexe (1 pour les

femmes, 0 pour les hommes, noté `gender`), l'âge (en années, divisé par 10, noté `age`), l'état civil (1 si marié, 0 sinon, noté `status`), le niveau d'éducation (la durée de la scolarité, noté `school`), le revenu familial (en dix mille dollars, noté `income`); (ii) diverses mesures de l'état de santé: nombre de maladies chroniques (cancer, diabète, arthrite, ..., dénoté par `chronic`) et la perception qu'a le patient de son état de santé (excellent, moyen, mauvais) recodé comme deux variables binaires: `health`1 qui prend la valeur 1 si l'état de santé est perçu comme mauvais et valeur 0 sinon; `health`2 qui prend la valeur 1 si l'état de santé est excellent comme mauvais et valeur 0 sinon. (iii) `medicaid`, une variable binaire qui indique si la personne est couverte par medicaid ou non. Nous la codons comme 1 si la personne est couverte et 0 sinon. Nous avons ajusté un modèle de régression ZIBP incorporant toutes les covariables disponibles dans (2.2)-(2.3) pour chaque individu. Nous avons ensuite utilisé les tests de Wald pour sélectionner les covariables significatives. La covariable la moins significative "au niveau 5%" est éliminée et le modèle est ajusté à nouveau, jusqu'à ce que toutes les covariables restantes soient significatives; le critère BIC diminue à chaque étape de cette procédure. Le tableau 2.3 présente le modèle final du ZIBP.

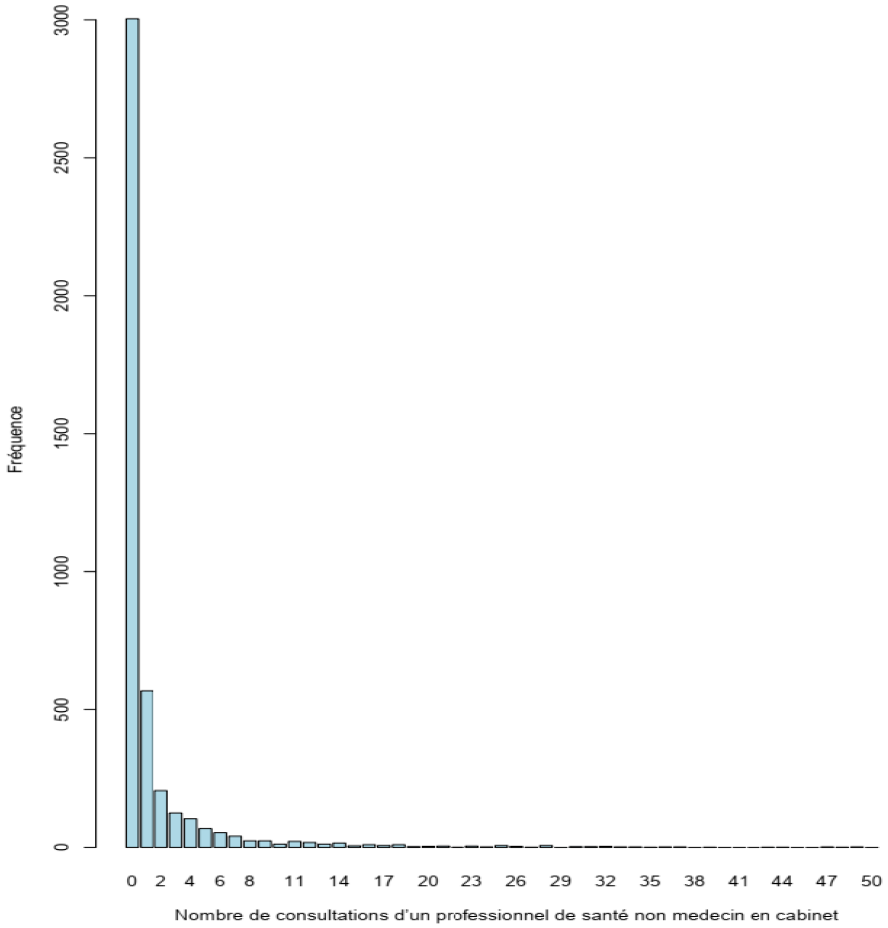

Figure 2.9 – Diagramme en barres du nombre de consultations d'un professionnel de santé non médecin en cabinet

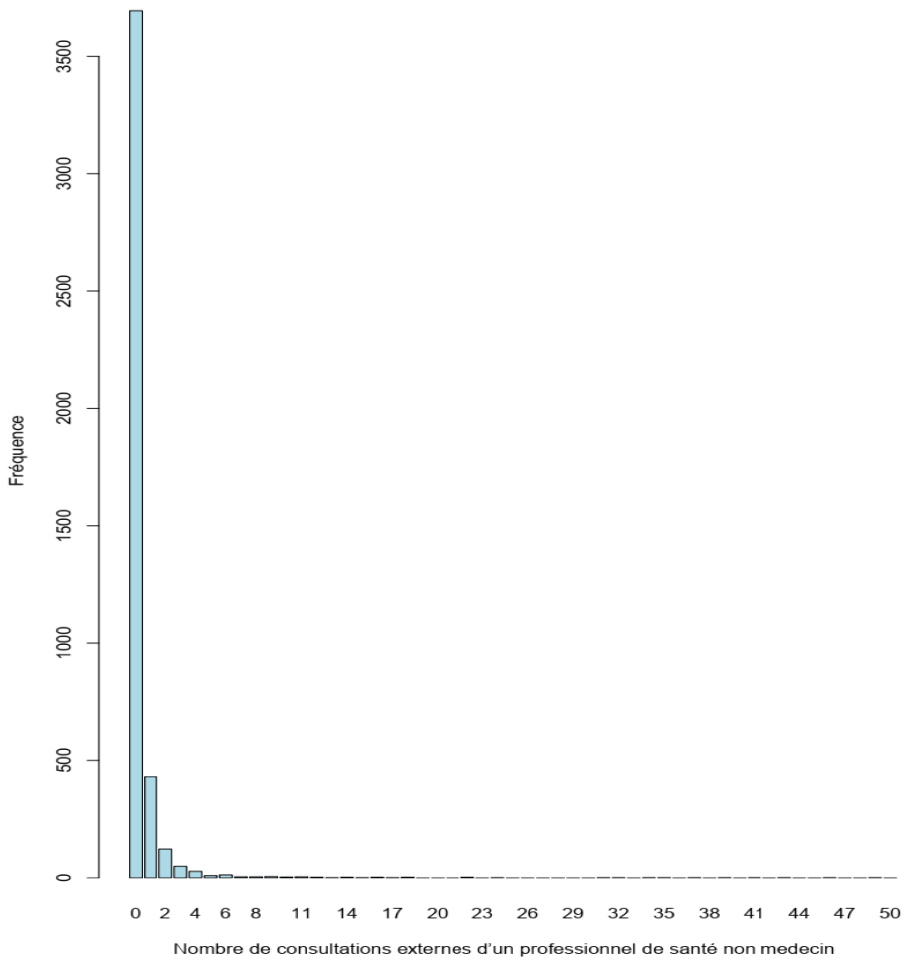

Figure 2.10 – Diagramme en barres du nombre de consultations externes d'un professionnel de santé non médecin

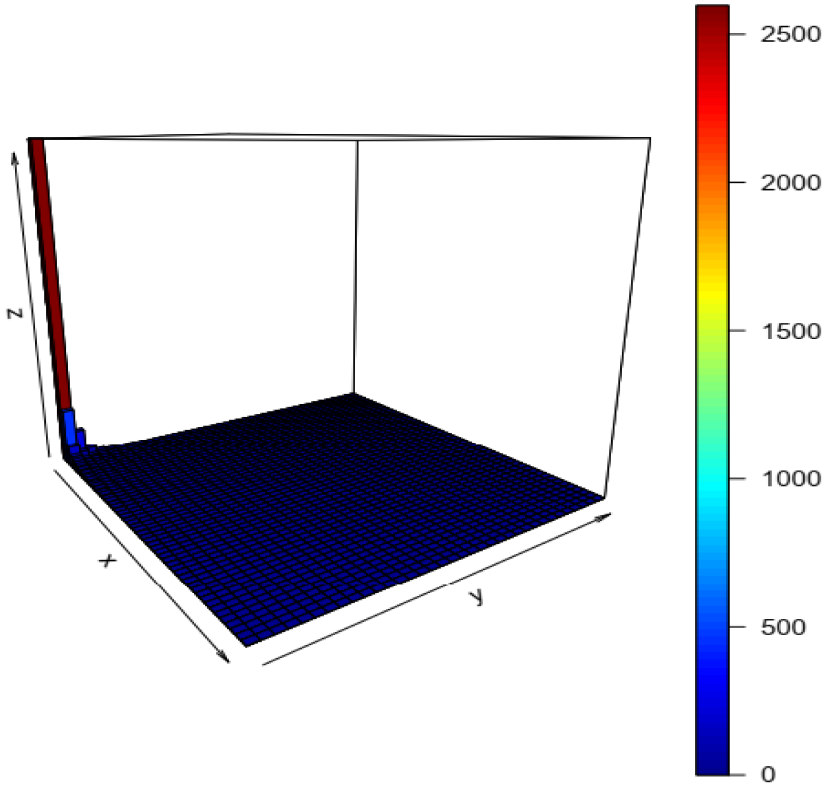

Figure 2.11 – Représentation de la distribution des fréquences du couple $(\text{opnp}, \text{ofnp})$.

2.5.2 Résultats

Les résultats du modèle ZIBP résultant sont présentés dans le tableau 2.3. L'estimation, l'erreur standard (s.e.) et le niveau de significativité (comme : non significatif, significatif ou très significatif) du test de nullité de Wald pour chaque paramètre sont reportés dans le tableau 2.3. Les résultats montrent que le nombre de maladies chroniques, le sexe, le niveau d'éducation et le statut de bénéficiaire de Medicaid sont identifiés par le ZIBP comme les facteurs qui influencent le plus la décision de ne jamais recourir à un professionnel de santé non médecin (en cabinet et en consultation externe). Il apparaît que la probabilité de ne jamais recourir à des consultations non médicales diminue avec le nombre de pathologies chroniques. Cela se justifie par le fait que plus l'état du patient est chronique, plus il est susceptible de privilégier les visites chez les spécialistes. Ensuite, la probabilité de ne jamais avoir recours à des consultations avec un non-médecin diminue avec le nombre d'années d'éducation. En effet, l'éducation peut faire des individus des consommateurs plus informés des services de soins de santé. Ces résultats confirment également ceux de Deb et Trivedi (1997). Les bénéficiaires de Medicaid ont tendance à renoncer aux consul-

tations avec un non-médecin. En effet, Medicaid étant une assurance maladie pour les personnes pauvres, les bénéficiaires sont limités dans leur choix de consultations. Ils sont limités aux seules visites chez le médecin. On constate également que les femmes sont plus susceptibles d'être non-utilisatrices de `ofnd` et de `opnd`. La probabilité de consulter un non-médecin diminue avec l'âge. Cela peut s'expliquer par plusieurs facteurs, tels que la diminution de la mobilité associée au vieillissement (les patients âgés auront tendance à limiter leurs consultations à celles considérées comme les plus nécessaires, c'est-à-dire aux visites chez le médecin).

paramètre	variable	estimation	s.e.	$\Pr(>t)$	signif.
$\widehat{\gamma}_1$	intercept	0.667400	0.202808	0.000999	$\star\star\star$
$\widehat{\gamma}_2$	chronic	-0.165918	0.023503	$1.67e^{-12}$	$\star\star\star$
$\widehat{\gamma}_3$	gender	0.314814	0.064692	$1.14e^{-06}$	$\star\star\star$
$\widehat{\gamma}_4$	education	-0.088953	0.008967	$< 2e^{-16}$	$\star\star\star$
$\widehat{\gamma}_5$	medicaid	0.397012	0.117432	0.000723	$\star\star\star$
$\widehat{\beta}_{1,1}$	intercept	1.381952	0.209195	$3.95e^{-11}$	$\star\star\star$
$\widehat{\beta}_{1,2}$	health1	0.083504	0.041901	0.046273	$\star$
$\widehat{\beta}_{1,3}$	health2	0.133144	0.046970	0.004588	$\star\star$
$\widehat{\beta}_{1,4}$	chronic	0.025524	0.009689	0.008428	$\star\star$
$\widehat{\beta}_{1,5}$	age	-0.124267	0.021390	6.26^{e-09}	$\star\star\star$
$\widehat{\beta}_{1,6}$	gender	-0.007129	0.027981	0.798901	
$\widehat{\beta}_{1,7}$	marital statuts	0.004958	0.028733	0.862999	
$\widehat{\beta}_{1,8}$	education	0.032632	0.003950	$< 2^{e-16}$	$\star\star\star$
$\widehat{\beta}_{1,9}$	income	-0.018566	0.004775	0.000101	$\star\star\star$
$\widehat{\beta}_{1,10}$	medicaid	0.205903	0.051070	$5.54e^{-05}$	$\star\star\star$
$\widehat{\beta}_{2,1}$	intercept	7.573439	0.377543	$< 2e^{-16}$	$\star\star\star$
$\widehat{\beta}_{2,2}$	health1	-0.168968	0.063093	0.007404	$\star\star$
$\widehat{\beta}_{2,3}$	health2	-0.788500	0.144316	$4.66e^{-08}$	$\star\star\star$
$\widehat{\beta}_{2,4}$	chronic	0.112977	0.015864	$1.07e^{-12}$	$\star\star\star$
$\widehat{\beta}_{2,5}$	age	-0.491438	0.040842	$< 2e^{-16}$	$\star\star\star$
$\widehat{\beta}_{2,6}$	gender	0.214599	0.048058	7.99^{e-06}	$\star\star\star$
$\widehat{\beta}_{2,7}$	marital statuts	-0.116671	0.050193	0.020101	$\star$
$\widehat{\beta}_{2,8}$	education	-0.103404	0.006346	$< 2e^{-16}$	$\star\star\star$
$\widehat{\beta}_{2,9}$	income	-0.024264	0.010076	0.016033	$\star$
$\widehat{\beta}_{2,10}$	medicaid	-1.758261	0.061534	$< 2e - 16$	$\star\star\star$

Tableau 2.3 – Analyse des donnés d'utilisation des services de santé NMES1988

Parmi les patients qui n'ont pas systématiquement renoncé à consulter un professionnel de santé non médecin, la probabilité d'avoir recours à une consultation `ofnp` ou `opnp` diminue avec l'âge et le revenu. Le niveau de revenu d'un patient va influencer la nature et la qualité des soins qu'il recherche, plutôt que le nombre de consultations, ce qui est cohérent avec Deb et Trivedi (1997). La probabilité de recou-

rir à une consultation `opnp` diminue lorsque le patient estime que sa santé n'est plus excellente, qu'elle s'est dégradée. Les patients mariés semblent renoncer à consulter un non-médecin en ambulatoire. Bien que les patients mieux informés aient tendance à diversifier leur recours aux soins, ils semblent délaisser le service de santé `opnp` au profit du service `ofnp`.

Par conséquent, en considérant les variables `ofnd` et `opnd` simultanément, tout en tenant compte de la corrélation entre elles, le modèle ZIBP permet une meilleure compréhension des éléments qui justifient le recours à différentes formes de soins médicaux par ordre d'utilisation. Nous avons constaté que les patients à la santé fragile, âgés et couverts par l'assurance Medicaid préfèrent les consultations avec des médecins que celles avec des non-médecins.

2.6 Conclusion

Dans ce chapitre, nous avons évalué théoriquement et numériquement les performances de l'estimateur du maximum de vraisemblance dans le modèle ZIBP. Une application du modèle ZIBP au jeu de données NMES1988 a permis de mieux comprendre les facteurs qui favorisent le renoncement ou l'utilisation de certains services de soins de santé. Maintenant, plusieurs extensions de ce travail devraient être développées pour étendre son champ d'application. Par exemple, nous pouvons d'abord envisager l'étude théorique et numériques des estimateurs dans les modèles de régression de Poisson multivariés avec inflation de zéros. Aussi, nous pouvons étudier les propriétés des estimateurs dans les modèles ZIBP avec des valeurs censurées aléatoirement à droite ou à gauche ou par intervalles. Ou encore, nous intéresser à l'estimation des paramètres dans les modèles de régression de Poisson bivariée à inflation de zéros lorsque certaines covariables qui interviennent dans la régression sont partiellement observées. Ce dernier axe de recherche fait l'objet de notre étude dans le chapitre suivant.

Annexe A

Théorème de Foutz [1977]

Théorème : Soit $X_1, X_2, \cdots, X_n$, n observations indépendantes de X avec $f(x, \theta)$ sa densité de probabilité, où $\theta \in \Theta$ un sous ensemble de $\mathbb{R}^d$. Sur la base de ces observations, la log-vraisemblance de θ peut s'écrire comme suit: $\mathcal{L}_n(\theta) = \sum_{i=1}^{n} \log f(X_i, \theta)$. De plus, si

- $\frac{\partial^2 \log f(X,\theta)}{\partial \theta_k \partial \theta_j}$, $k, j = 1, \cdots, d$, les dérivées secondes de $\log f(X, \theta)$ existent et sont continues;
- $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(X_i,\theta_0)}{\partial \theta}$ converge en probabilité vers $\mathbf{0}$ lorsque n tend vers l'infini.
- dans un voisinage de θ_0, $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(X_i,\theta)}{\partial \theta \partial \theta^\top}$ converge uniformément en probabilité vers $\sigma(\theta)$ une matrice définie négative lorsque n tend vers l'infini.

Alors, il existe une suite $\{\widehat{\theta}_n\}_n$ telle que $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(X_i,\widehat{\theta}_n)}{\partial \theta} = \mathbf{0}$ avec la probabilité 1 et $\{\widehat{\theta}_n\}$ converge en probabilité vers θ_0 lorsque n tend vers l'infini.

3

Estimation d'une régression de Poisson bivariée à inflation de zéros avec covariables manquantes.

Résumé

Les modèles de régression de Poisson bivariés à inflation de zéros (ZIBP) sont appliqués à des données de comptage bivariées corrélées. Ces modèles ont été appliqués dans divers domaines, tels que la recherche médicale, l'économie de la santé, l'assurance, le sport, etc. Dans la pratique, certaines des covariables impliquées dans la modélisation sont souvent partiellement observées. Nous proposons des méthodes pour estimer les paramètres du modèle de régression ZIBP avec des covariables manquantes au hasard. En supposant que la probabilité de sélection est inconnue et estimée de manière paramétrique ou non paramétrique (par un estimateur à noyau), nous proposons des méthodes de pondération par l'inverse des probabilités de sélection (IPW) et d'imputation multiple (MI) pour estimer les paramètres du modèle de régression ZIBP lorsque des covariables sont manquantes au hasard (MAR). Les propriétés asymptotiques des estimateurs proposés sont étudiées sous certaines conditions de régularité. En outre, nous avons réalisé une étude de simulation exhaustive sur des tailles finies d'échantillons afin d'évaluer la cohérence de nos résultats. Pour finir, une application pratique des méthodologies proposées est illustrée sur des données recensant l'utilisation des services de santé de milliers de patients aux USA.

Sommaire

3.1 Introduction

Le modèle de Poisson bivarié à inflation de zéros (ZIBP pour Zero-inflated bivariate Poisson) a été appliqué dans de nombreux contextes, notamment le marketing (nombre d'achats de différents produits), la recherche médicale (nombre de crises d'épilepsie avant et après traitement), l'épidémiologie (incidence de différentes maladies dans une série de districts), l'analyse des accidents (nombre d'accidents sur un site avant et après modification des infrastructures), l'économétrie (nombre de changements d'emploi volontaires et involontaires), le sport (nombre de buts marqués par chacune des deux équipes adverses au football), pour n'en citer que quelques-uns.

Dans de nombreuses circonstances, on est souvent confronté au fait qu'un ensemble de données contient des données manquantes. En effet, les données manquantes sont un problème très répandu dans de nombreuses disciplines, notamment l'économie, la sociologie, les sciences médicales, les sciences politiques, les transports, la communication et d'autres domaines. Les données manquantes constituent une menace sérieuse pour la validité de l'inférence ou de la prise de décision dans de nombreuses applications. Compte tenu de l'intérêt de ce problème, diverses méthodes ont été proposées pour traiter les données manquantes dans les modèles de régression au cours des dernières décennies. Nous renvoyons le lecteur intéressé à Rubin (1976), Little (1992), Zhao et Lipsitz (1992), Robins *et al.* (1994), Reilly et Pepe (1995), Clayton *et al.* (1998), Creemers *et al.* (2012), Lukusa *et al.* (2016) et Lee *et al.* (2021) pour plus de détails. L'approche commune et simple du problème de données manquantes est appelée le cas complet (CC). Cette méthode consiste à exclure les individus dont les données sont manquantes. Cependant, la méthode du cas complet peut induire

des biais et une forte augmentation de la variance. Une alternative au cas complet en cas de données manquantes est l'imputation multiple (MI). L'imputation multiple est une approche largement utilisée pour résoudre les problèmes de données manquantes. Cette méthode exige que les échantillons soient complets. Les données manquantes sont donc remplacées par des valeurs plausibles générées à partir d'un modèle d'imputation. Cette imputation est répétée M fois, générant ainsi M ensembles de données complètes. Chaque échantillon complet obtenu est analysé et un estimateur global est obtenu en combinant les estimations de ses M échantillons complets. Une autre méthode de traitement des données manquantes est la pondération par inverse de la probabilité de sélection (IPW, pour "Inverse Probability Weighting"). Introduite par Horvitz et Thompson (1952) puis développée par Zhao et Lipsitz (1992), la méthode IPW est basée sur la création de pseudo-populations de cas complets dans lesquelles le biais de sélection dû aux données manquantes est éliminé par des poids. La détermination de ces poids nécessite un modèle pour évaluer la probabilité qu'un individu ait des données complètes. On peut se référer à Diallo *et al.* (2019) et Seaman et White (2013) pour plus d'information sur cette méthode.

Cependant, certaines approches ont été développées pour traiter les problèmes de covariables manquantes dans les modèles zéro-inflatés (ZI) lorsque les covariables sont manquantes de manière aléatoire (MAR, pour missing at random). Par exemple, Lukusa *et al.* (2016) et Lukusa et Phoa (2020) ont proposé des méthodes d'estimation par pondération de inverse des probabilités de sélection (IPW) semi-paramétrique pour un modèle de régression de Poisson (ZIP) à inflation de zéros avec des covariables manquantes. Diallo *et al.* (2019) ont proposé une méthode d'estimation IPW paramétrique pour un modèle de régression binomial à inflation de zéros (ZIB) avec covariables MAR. Lee *et al.* (2020) ont proposé des méthodes d'estimation d'imputation multiple non-paramétrique pour le modèle de régression ZIP. Et plus récemment, Lee *et al.* (2021) se sont intéressés à l'estimation des paramètres dans le modèle de régression de Bernoulli à inflation de zéros (ZIBer).

Bien qu'il existe de nombreuses études sur les modèles zéros-infaltés (ZI) avec covariables manquantes, ces études se sont limitées aux cas où les variables réponses sont univariées. À notre connaissance, il n'existe pas de travaux portant sur les modèles de comptage multivariés dans un contexte de covariables manquantes. Ce chapitre a pour but de combler cet important déficit.

Dans ce chapitre, nous étendons les travaux antérieurs réalisés sur des modèles de comptage univariés au cas où les variables réponses sont bivariées et corrélées. D'abord, nous estimons les probabilités de sélection par une méthode paramétrique, puis nous proposons une estimation IPW paramétrique des paramètres du modèle de

Poisson bivarié à inflation de zéros (ZIBP) lorsque les covariables sont manquantes. Ensuite, étant donné qu'une mauvaise spécification du modèle utilisé pour la probabilité de sélection peut produire des estimations biaisées, nous proposons des estimations plus robustes de la probabilité de sélection basées sur une modélisation non paramétrique. Pour ce faire, nous utilisons un noyau au lieu de la fonction indicatrice dans Lukusa *et al.* (2016) et Lee *et al.* (2020) afin de considérer des covariables continues et catégorielles. À la suite, nous proposons respectivement des méthodes estimation par pondération IPW semi-paramétrique et deux méthodes d'imputation multiple (MI) pour l'estimation des paramètres du modèle ZIBP. Sachant qu'il n'est pas toujours évident d'avoir un bon modèle paramétrique pour générer des données manquantes, dans nos deux méthodes MI nous imputons les données manquantes de manière non paramétrique en nous inspirant des travaux de Wang et Chen (2009) et de Lee *et al.* (2020). La première méthode utilise l'idée de Rubin (2004) où les estimations obtenues à partir de tous les ensembles de données imputées sont résumées en une estimation finale. L'autre méthode met en œuvre l'idée de Fay (1996) où les équations d'estimation pour différents ensembles de données imputées sont combinées en une seule équation d'estimation à optimiser. Par des études théoriques et numériques, nous avons montré les bonnes propriétés des estimateurs proposés pour l'estimation des paramètres du modèle de ZIBP lorsque certaines covariables qui interviennent dans la régression sont manquantes de manière aléatoire. En un mot, dans cette partie du travail, nous mettons l'accent sur l'estimation des modèles de régression de Poisson bivariés à inflation de zéros dans le contexte des données manquantes dans les covariables. Nous supposons que les covariables utilisées dans le modèle de régression sont mixtes. Ainsi, nous considérons un mélange de variables catégorielles, discrètes et continues. Nous proposons respectivement, l'IPW paramétrique, l'IPW semi-paramétrique et deux méthodes d'imputation multiple dans les modèles de régression ZIBP lorsque certaines covariables sont manquantes de manière aléatoire (MAR).

Le reste de ce chapitre est organisé comme suit. Dans la section 3.2, nous décrivons brièvement le modèle de régression ZIBP et son estimation dans le cas des données complètes. Les sections 3.3 et 3.4 présentent et décrivent les propriétés asymptotiques des estimateurs proposés dans le modèle ZIBP lorsque les covariables sont manquantes aléatoirement. Dans la section 3.5, nous réalisons une étude de simulation pour étudier les performances des estimateurs proposés. La section 3.6, présente une application des méthodes proposées. Pour clore, une discussion et quelques perspectives sont fournies dans la section 3.7.

3.2 Modèle de régression de Poisson bivarié à inflation de zéros et estimation

Considérons des variables aléatoires Z_1, Z_2 et U qui suivent des distributions de Poisson indépendantes avec des paramètres λ_1, λ_2 et μ respectivement. Alors les variables aléatoires $Y_1 = Z_1 + U$ et $Y_2 = Z_2 + U$ suivent conjointement une distribution de Poisson bivariée BP$(\lambda_1, \lambda_2, \mu)$. Soit $y_1 \wedge y_2 := \min(y_1, y_2)$. La distribution conjointe du vecteur de Poisson bivarié (Y_1, Y_2) est donnée par:

$$f_{BP}(y_1, y_2; \lambda_1, \lambda_2, \mu) \quad = \quad \exp(-\mu - \lambda_1 - \lambda_2)\varphi(y_1, y_2)$$

où

$$\varphi(y_1, y_2) = \sum_{s=0}^{\min(y_1,y_2)} \frac{\mu^s}{s!} \frac{\lambda_1^{y_1-s}}{(y_1-s)!} \frac{\lambda_2^{y_2-s}}{(y_2-s)!}.$$

Pour les données de comptage bivariées qui ont une forte proportion de $(0, 0)$, Li et al. Li *et al.* (1999) proposent le modèle de Poisson bivarié avec inflation de zéros (ZIBP). Ce modèle est obtenu en mélangeant une distribution dégénérée à $(0, 0)$ avec un modèle de régression de Poisson bivarié. Un modèle ZIBP donne la distribution de (Y_{1i}, Y_{2i}) comme suit:

$$\left(Y_{1i}, Y_{2i}\right) \sim \begin{cases} (0,0) & \text{avec la probabilité } \varepsilon_i \\ \text{BP}(\lambda_{1i}, \lambda_{2i}, \mu_i) & \text{avec la probabilité } 1-\varepsilon_i \end{cases} \tag{3.1}$$

où BP$(\lambda_{1i}, \lambda_{2i}, \mu_i)$ représente le modèle de Poisson de paramètres λ_{1i}, λ_{2i}, μ_i et ε_i désigne la proportion de mélange qui est telle que $0 < \varepsilon_i < 1$.
Le modèle (3.1) permet à λ_{1i}, λ_{2i} et ε_i de dépendre des covariables par le biais des relations

$$\text{logit}(\varepsilon_i) \quad = \quad \gamma^\top \mathbf{W}_i \tag{3.2}$$

et

$$\lambda_{1i} = \exp(\beta_1^\top \mathbf{X}_i), \quad \lambda_{2i} = \exp(\beta_2^\top \mathbf{X}_i) \text{ et } \mu = \exp(\alpha), \tag{3.3}$$

où $\mathbf{W}_i = (W_{i1}, ..., W_{iq})^\top$ et $\mathbf{X}_i = (X_{i1}, ..., X_{ip})^\top$ sont des covariables pour la probabilité de mélange et pour les paramètres de Poisson, respectivement $\gamma \in \mathbb{R}^q$, $\beta_1, \beta_2 \in \mathbb{R}^p$ et $\alpha \in \mathbb{R}$ sont les paramètres de régression correspondants; $\top$ désigne l'opérateur de transposition. On note $\Theta = (\gamma^\top, \beta_1^\top, \beta_2^\top, \alpha)^\top$ le vecteur de paramètres de dimension $k = 2p + q + 1$.
Supposons que nous observons n vecteurs indépendants $(Y_{1i}, Y_{2i}, \mathbf{X}_i, \mathbf{W}_i),\ i = 1, ..., n$

à partir des modèles (3.1)-(3.2)-(3.3), tous définis sur l'espace de probabilité $(\Omega, \mathcal{C}, \mathbb{P})$. Sur la base de ces observations, la log-vraisemblance de peut s'écrire comme suit :

$$\begin{aligned}\ell\ell_n(\Theta) &= \sum_{i=1}^{n}\Bigg\{J_i \log\left(e^{\gamma^\top \mathbf{W}_i} + h_i(\Theta)\right) - (1-J_i)\left(e^{\alpha} + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}\right) \\ &\qquad +(1-J_i)\log\left(\sum_{s=0}^{Y_{1i}\wedge Y_{2i}} \frac{(e^{\alpha})^s}{s!}\frac{(e^{\beta_1^\top \mathbf{X}_i})^{y_{1i}-s}}{(y_{1i}-s)!}\frac{(e^{\beta_2^\top \mathbf{X}_i})^{y_{2i}-s}}{(y_{2i}-s)!}\right) - \log(1+e^{\gamma^\top \mathbf{W}_i})\Bigg\} \\ &:= \sum_{i=1}^{n}\ell_i(\Theta),\end{aligned}$$

où $J_i := 1_{(Y_{1i}=0, Y_{2i}=0)}$ et $h_i(\Theta) = e^{-(e^{\alpha}+e^{\beta_1^\top \mathbf{X}_i}+e^{\beta_2^\top \mathbf{X}_i})}$.
Supposons que toutes les covariables du modèle de régression ZIBP soient entièrement observées. L'estimateur du maximum de vraisemblance $\widehat{\Theta}_{F,n} = \left(\widehat{\gamma}^\top, \widehat{\beta}_1^\top, \widehat{\beta}_2^\top, \widehat{\alpha}\right)^\top$ de Θ est obtenu en résolvant l'équation de score $U_{F,n}(\Theta) = 0$, où

$$U_{F,n}(\Theta) = \frac{1}{\sqrt{n}}\frac{\partial \ell\ell_n(\Theta)}{\partial \Theta} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ell_i(\Theta)}{\partial \Theta} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{\ell}_i(\Theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Phi_i(\Theta). \tag{3.4}$$

La résolution de l'équation de score peut être effectuée à l'aide de l'algorithme EM (voir Ntzoufras et Karlis (2005)) ou par maximisation directe de $\ell\ell_n(\Theta)$ en utilisant un algorithme de type Newton Raphson. De plus, il faut souligner que le vecteur de score est centré. Dans le chapitre 2 ont montré que l'EMV $\widehat{\Theta}_{F,n}$ de Θ dans le modèle ZIBP est un estimateur consistant et asymptotiquement normal.

Après avoir présenté le modèle de régression ZIBP et son estimateur EMV lorsque toutes les covariables sont observées, nous nous intéressons dans la suite de ce chapitre à une situation récurrente, le cas où certaines covariables utilisées dans la régression sont partiellement observées. Dans les deux sections suivantes, nous proposons des méthodes pour estimer les paramètres du modèle de régression ZIBP lorsque certaines des covariables sont manquantes au hasard (MAR, missing at random).

3.3 Méthodes de pondération par l'inverse des probabilités

La méthode de pondération par l'inverse des probabilités (IPW) a été proposée par Horvitz et Thompson (1952) pour estimer des modèles de régression avec des covariables manquantes. Soient $\mathbf{X}^{obs}$ les composantes de $\mathbf{X}$ qui sont entièrement observées et $\mathbf{X}^{mis}$ les composantes de $\mathbf{X}$ qui ont au moins une donnée manquante. De

même, on note respectivement $\mathbf{W}^{obs}$ et $\mathbf{W}^{mis}$ les composantes de $\mathbf{W}$ qui respectivement sont entièrement observées et ont au moins une donnée manquante. Notons $\mathcal{O} = (Y_1, Y_2, \mathbf{X}^{(obs),\top}, \mathbf{W}^{(obs),\top})^\top$ le vecteur des variables qui sont toujours observées sur chaque individu. Soit δ une variable indicatrice qui prend la valeur 1 lorsque toutes les covariables $\{\mathbf{X}, \mathbf{W}\}$ chez un individu sont observées; 0 sinon. Nous supposons tout au long de ce chapitre que les données $\mathcal{Z} = \{\mathbf{X}^{(mis)}, \mathbf{W}^{(mis)}\}$ sont manquantes au hasard (MAR). Dans le cadre du mécanisme MAR, la probabilité de sélection ou la probabilité de manquant se réduit à

$$\pi(\mathcal{O}_i) = \mathbb{P}(\delta_i = 1|\mathcal{O}_i, \mathbf{X}_i^{(mis)}, \mathbf{W}_i^{(mis)}) = \mathbb{P}(\delta_i = 1|\mathcal{O}_i). \tag{3.5}$$

Par conséquent, sous l'hypothèse MAR, l'estimateur du maximum de vraisemblance IPW de Θ dans le modèle (3.1) est la solution de l'équation pondérée

$$U_{W,n}(\Theta, \pi) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\delta_i}{\mathbb{P}(\delta_i = 1|\mathcal{O}_i)} \Phi_i(\Theta) = 0. \tag{3.6}$$

Les probabilités de sélection $\pi(\mathcal{O}_i)$ sont généralement inconnues. Par conséquent, il nous faut donc les estimer pour l'estimation de Θ. En utilisant des méthodes d'estimation paramétriques puis non-paramétriques pour l'estimation de $\pi(O_i)$, nous proposons respectivement deux méthodes d'estimation pondérées pour estimer les paramètres du modèle de régression ZIBP lorsque des covariables MAR.

3.3.1 Estimateur IPW paramétrique

Supposons que la probabilité de sélection $\mathbb{P}(\delta_i = 1|\mathcal{O}_i)$ puisse être spécifiée par un modèle paramétrique $\pi(\omega, \mathcal{O}_i)$, où ω est un vecteur de paramètres de régression inconnus avec la valeur vraie ω_0. Soit $\widehat{\omega}_n$ l'estimation du maximum de vraisemblance de ω_0. Sous certaines conditions de régularité que nous définissons dans la suite, $\widehat{\omega}_n$ peut être obtenu comme suit

$$\widehat{\omega}_n = \arg\max_{\omega} \prod_{i=1}^{n} \pi(\omega, \mathcal{O}_i)^{\delta_i} \big(1 - \pi(\omega, \mathcal{O}_i)\big)^{1-\delta_i}.$$

Par des calculs, on peut montrer que

$$\sqrt{n}(\hat{\omega}_n - \omega_0) = n^{-1/2} \sum_{i=1}^{n} \frac{\delta_i - \pi(\omega_0, \mathcal{O}_i)}{\pi(\omega_0, \mathcal{O}_i)\big(1 - \pi(\omega_0, \mathcal{O}_i)\big)} \Omega^{-1}(\omega_0) \dot{\pi}(\omega, \mathcal{O}_i) + \mathbf{o}_{\mathbb{P}}(1),$$

où

$$\dot{\pi}(\omega, \mathcal{O}_i) = \frac{\partial \pi(\omega, \mathcal{O}_i)}{\partial \omega} \text{ et } \Omega(\omega) = \mathbb{E}\Big[\frac{\dot{\pi}(\omega, \mathcal{O})\big(\dot{\pi}(\omega, \mathcal{O})\big)^\top}{\pi(\omega, \mathcal{O})(1 - \pi(\omega, \mathcal{O}))}\Big].$$

Après avoir donné l'estimateur EMV de ω, nous pouvons maintenant déterminer l'estimateur $\widehat{\Theta}_{W,n}$ de Θ en résolvant l'équation suivante

$$U_{W,n}(\Theta, \hat{\omega}_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\delta_i}{\pi(\hat{\omega}_n, \mathcal{O}_i)} \Phi_i(\Theta) = 0.$$

Hypothèses de régularité et résultats asymptotiques

Les hypothèses de régularité dont nous avons besoin pour établir les propriétés asymptotiques des estimateurs $\widehat{\Theta}_{W,n}$ sont les suivantes :

(B1) La vraie valeur du paramètre Θ_0 notée $\Theta_0 := (\gamma_0^\top, \beta_{1,0}^\top, \beta_{2,0}^\top, \alpha_0)^\top$ se situe à l'intérieur d'un ensemble compact connu $\mathcal{G} \subset \mathbb{R}^{2p+q+1}$. De plus, ω_0 appartient à l'intérieur d'un ensemble compact connu $\mathbf{K}$.

(B2) Pour toute valeur $o \in \mathcal{O}$ et pour tout $\omega \in \mathbf{K}$, il existe C tel que la probabilité de sélection $\pi(\omega, \mathcal{O}) > C > 0$.

(B3) La fonction $\pi(\omega, \mathcal{O})$ est différentiable par rapport à ω, pour tout $o \in \mathcal{O}$. Pour tout $\omega,\ \widetilde{\omega} \in \mathbf{K}$, $\left|\pi(\omega, \mathcal{O}) - \pi(\widetilde{\omega})\right| \leq \kappa(\mathcal{O})\left\|\omega - \widetilde{\omega}\right\|$ par une certaine fonction bornée κ avec $\mathbb{E}[\kappa(\mathcal{O})] = \rho$.

(B4) Le score $\Phi_1(\Theta)$ a un moment d'ordre 2 fini défini dans un voisinage de Θ_0. De plus, $\mathbb{E}\left[(\Phi_1(\Theta))^{\otimes 2}\right]$ est définie positive dans un voisinage de Θ_0, où pour tout vecteur colonne a, $a^{\otimes 2} = aa^\top$.

(B5) Dans un voisinage de Θ_0, les dérivées première et seconde de $U_{W,n}(\Theta, \omega)$ par rapport à Θ sont uniformément bornées supérieurement par une fonction de $(Y_1, Y_2, \mathbf{X}, \mathbf{W})$ dont les espérances existent. De plus, $-\frac{1}{\sqrt{n}} \frac{\partial U_{W,n}(\Theta,\omega)}{\partial \Theta^\top}$ converge vers une certaine matrice définie positive $\Sigma_1(\Theta, \omega)$ lorsque n tend vers l'infini dans un voisinage de Θ_0.

Le théorème suivant donne les propriétés asymptotiques de $\widehat{\Theta}_{W,n}$.

Théorème 3.1 *Supposons que les hypothèses de régularité (**B1**) à (**B5**) sont vérifiées. Alors $\widehat{\Theta}_{W,n}$ est un estimateur consistant de Θ. De plus $\sqrt{n}(\widehat{\Theta}_{W,n} - \Theta_0)$ est distribué asymptotiquement suivant une loi normale multivariée avec une moyenne nulle et une matrice de covariance Σ_W, où*

$$\Sigma_W := \left[\Sigma_1(\Theta_0)^{-1}\right]\left\{\Sigma_2(\Theta_0,\omega_0) - \Sigma_3(\Theta_0,\omega_0)\Omega(\omega_0)^{-1}\Sigma_3(\Theta_0,\omega_0)^\top\right\}[\Sigma_1(\Theta_0)^{-1}]^\top,$$

avec

$$\begin{aligned}\Sigma_2(\Theta,\omega) &= \mathbb{E}\Big[\frac{\delta}{\pi^2(\omega,\mathcal{O})}\Big[\Phi(\Theta)\Big]^{\otimes 2}\Big],\\ \Sigma_3(\Theta,\omega) &= -\mathbb{E}\Big[\frac{\delta}{\pi^2(\omega,\mathcal{O})}\Phi(\Theta)\dot{\pi}(\omega,\mathcal{O})^\top\Big].\end{aligned}$$

Remarque 3.1 *Dans la pratique, on utilise un estimateur consistant de la matrice de covariance Σ_W. Un estimateur consistant de Σ_W est donné par*

$$\begin{aligned}\widehat{\Sigma}_{W,n} &:= \Big[\Sigma_{1,n}(\widehat{\Theta}_n,\widehat{\omega}_n)^{-1}\Big]\Big\{\Sigma_{2,n}(\widehat{\Theta}_n,\widehat{\omega}_n) - \Sigma_{3,n}(\widehat{\Theta}_n,\widehat{\omega}_n)\Omega_n(\widehat{\Theta}_n,\widehat{\omega}_n)^{-1}\Sigma_{3,n}(\widehat{\Theta}_n,\widehat{\omega}_n)^\top\Big\}\\ &\quad\times\Big[\Sigma_{1,n}(\widehat{\Theta}_n,\widehat{\omega}_n)^{-1}\Big]^\top,\end{aligned}$$

où

$$\begin{aligned}\Sigma_{1,n}(\Theta,\omega) &= -\frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi(\omega,\mathcal{O}_i)}\frac{\partial^2\ell_i(\Theta)}{\partial\Theta\partial\Theta^\top},\\ \Sigma_{2,n}(\Theta,\omega) &= \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi^2(\omega,\mathcal{O}_i)}\Big[\Phi_i(\Theta)\Big]^{\otimes 2},\\ \Sigma_{3,n}(\Theta,\omega) &= -\frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi(\omega,\mathcal{O}_i)}\Phi_i(\Theta)\dot{\pi}(\omega,\mathcal{O}_i)^\top,\\ \Omega_n(\Theta,\omega) &= \frac{1}{n}\sum_{i=1}^n \frac{\dot{\pi}(\omega,\mathcal{O}_i)\big(\dot{\pi}(\omega,\mathcal{O}_i)\big)^\top}{\pi(\omega,\mathcal{O}_i)(1-\pi(\omega,\mathcal{O}_i))}.\end{aligned}$$

Preuve du Théorème 3.1

La consistance de $\widehat{\Theta}_{W,n}$, peut être prouvée en vérifiant que les conditions du théorème de la fonction inverse de Foutz (1977) sont satisfaites. Tout d'abord, montrons que $\frac{\partial U_{W,n}(\Theta,\hat{\omega}_n)}{\partial\Theta^\top}$ existe et est continue dans un voisinage ouvert de Θ_0.

Pour justifier cela, on peut remarquer que $U_{W,n}(\Theta, \hat{\omega}_n)$ est deux fois différentiable par rapport à Θ et que ses dérivées secondes sont continues.

Montrons que $n^{-1/2}U_{W,n}(\Theta_0, \hat{\omega}_n)$ converge en probabilité vers 0 lorsque $n \to \infty$. De l'égalité

$$n^{-1/2}U_{W,n}(\Theta_0, \hat{\omega}_n) = \big(n^{-1/2}U_{W,n}(\Theta_0, \hat{\omega}_n) - n^{-1/2}U_{W,n}(\Theta_0, \omega_0)\big) + n^{-1/2}U_{W,n}(\Theta_0, \omega_0).$$

Nous obtenons

$$\begin{aligned}
\Big\|n^{-1/2}U_{w,n}(\Theta_0, \hat{\omega}_n) - n^{-1/2}U_{w,n}(\Theta_0, \omega_0)\Big\| &= \Big\|\frac{1}{n}\sum_{i=1}^{n}\delta_i\Big(\frac{1}{\pi(\widehat{\omega}_n, \mathcal{O}_i)} - \frac{1}{\pi(\omega_0, \mathcal{O}_i)}\Big)\Phi_i(\Theta_0)\Big\|, \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\Big\|\frac{\pi(\omega_0, \mathcal{O}_i) - \pi(\hat{\omega}_n, \mathcal{O}_i)}{\pi(\omega_0, \mathcal{O}_i)\pi(\hat{\omega}_n, \mathcal{O}_i)}\Phi_i(\Theta_0)\Big\|, \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\Big\|\pi(\omega_0, \mathcal{O}_i) - \pi(\hat{\omega}_n, \mathcal{O}_i)\Big\|\Big\|\Phi_i(\Theta_0)\Big\| \\
&\quad \times \Big|\pi(\omega_0, \mathcal{O}_i)\pi(\hat{\omega}_n, \mathcal{O}_i)\Big|^{-1}.
\end{aligned}$$

Les hypothèses **(B2)** et **(B4)** garantissent l'existence d'une une constante positive finie k_1 telle que

$$\Big\|n^{-1/2}U_{w,n}(\Theta_0, \hat{\omega}_n) - n^{-1/2}U_{w,n}(\Theta_0, \omega_0)\Big\| \leq \frac{k_1}{n}\sum_{i=1}^{n}\Big\|\pi(\omega_0, \mathcal{O}_i) - \pi(\hat{\omega}_n, \mathcal{O}_i)\Big\|.$$

Par ailleurs, l'hypothèse **(B3)** implique que

$$\begin{aligned}
\Big\|n^{-1/2}U_{w,n}(\Theta_0, \hat{\omega}_n) - n^{-1/2}U_{w,n}(\Theta_0, \omega_0)\Big\| &\leq \frac{k_1}{n}\sum_{i=1}^{n}\kappa(\mathcal{O}_i)\big\|\hat{\omega}_n - \omega_0)\big\|, \\
&\leq k_1\big(\rho + o_{\mathbb{P}}(1)\big)\big\|\hat{\omega}_n - \omega_0\big\|.
\end{aligned}$$

Donc, la convergence de $\hat{\omega}_n$ vers ω_0 implique que $n^{-1/2}U_{w,n}(\Theta_0, \hat{\omega}_n) - n^{-1/2}U_{w,n}(\Theta_0, \omega_0)$ converge vers 0 lorsque n tend vers l'infini.

D'autre part, on a

$$\begin{aligned} n^{-1/2}U_{W,n}(\Theta_0,\omega_0) &= \Big(\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}W_{i1}A_i(\Theta_0),\ldots,\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}W_{iq}A_i(\Theta_0),\\ &\quad \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}X_{i1}B_{1i}(\Theta_0)\ldots,\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}X_{ip}B_{1i}(\Theta_0),\\ &\quad \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}X_{i1}B_{2i}(\Theta_0),\ldots,\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}X_{ip}B_{2i}(\Theta_0),\\ &\quad \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}C_i(\Theta_0))\Big)^\top. \end{aligned}$$

Nous montrons par la suite que $n^{-1/2}U_{W,n}(\Theta_0,\omega_0)$ converge en probabilité vers **0** lorsque $n\to\infty$. Pour s'en convaincre, notons d'abord que pour chaque $i=1,\ldots,n$, on a

$$\begin{aligned} \mathbb{E}\left[\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}C_i(\Theta_0)\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}C_i(\Theta_0)|\mathcal{O}_i\right]\right],\\ &= \mathbb{E}\left[\frac{1}{\pi(\omega_0,\mathcal{O}_i)}\mathbb{E}[\delta_i C_i(\Theta_0)|\mathcal{O}_i]\right]. \end{aligned}$$

Sous l'hypothèse MAR, $C_i(\Theta_0)$ et δ_i sont indépendants étant donné $\mathcal{O}_i$. Il s'ensuit que

$$\begin{aligned} \mathbb{E}\left[\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}C_i(\Theta_0)\right] &= \mathbb{E}\left[\frac{1}{\pi(\omega_0,\mathcal{O}_i)}\mathbb{E}[\delta_i|\mathcal{O}_i]\mathbb{E}[C_i(\Theta_0)|\mathcal{O}_i]\right],\\ &= \mathbb{E}[\mathbb{E}[C_i(\Theta_0)|\mathcal{O}_i]],\\ &= \mathbb{E}[C_i(\Theta_0)]. \end{aligned}$$

Puisque, nous avons $\mathbb{E}[C_i(\Theta_0)]=0$ pour chaque $i=1,\ldots,n$.
Il s'ensuit que, $\mathbb{E}\left[\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}C_i(\Theta_0)\right]=0$ pour chaque $i=1,\ldots,n$.
Par l'hypothèse (**B5**), on a

$$\text{var}\left(\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}C_i(\Theta_0)\right)=\mathbb{E}\left[\frac{\delta_i}{\pi^2(\omega_0,\mathcal{O})}C_i^2(\Theta_0)\right]<\infty$$

pour tout $i=1,\ldots n$.
Donc, par la loi faible des grands nombres, $\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}C_i(\Theta_0)$ converge en probabilité vers 0 lorsque $n\to\infty$.

Ensuite, pour tout $i=1,\ldots,n$ et $\ell=1,\ldots,q$, nous avons

$$\mathbb{E}\left[\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}W_{i\ell}A_i(\Theta_0)\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}W_{i\ell}A_i(\Theta_0)|\mathcal{O}_i\right]\right].$$

Deux cas doivent être considérés: dans le premier cas (a), $W_{i\ell}$ est une composante de $\mathbf{W}_i^{(mis)}$ et dans le second cas (b) $W_{i\ell}$ est une composante de $\mathbf{W}_i^{(obs)}$.
Dans le premier cas (a), nous avons

$$\mathbb{E}\left[\mathbb{E}\left[\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}W_{i\ell}A_i(\Theta_0)|\mathcal{O}_i\right]\right]=\mathbb{E}\left[\frac{1}{\pi(\omega_0,\mathcal{O}_i)}\mathbb{E}[\delta_i W_{i\ell}A_i(\Theta_0)|\mathcal{O}_i]\right].$$

Étant donné $\mathcal{O}_i$, $W_{i\ell}A_i(\Theta_0)$ est une fonction de $(\mathbf{W}_i^{(mis)},\mathbf{X}_i^{(mis)})$. Par conséquent, sous l'hypothèse MAR, $W_{i\ell}A_i(\Theta_0)$ et δ_i sont indépendants étant donné $\mathcal{O}_i$, il suit que

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\pi(\omega_0,\mathcal{O}_i)}\mathbb{E}[\delta_i W_{i\ell}A_i(\Theta_0)|\mathcal{O}_i]\right] &= \mathbb{E}\left[\frac{1}{\pi(\omega_0,\mathcal{O}_i)}\mathbb{E}[\delta_i|\mathcal{O}_i]\mathbb{E}[W_{i\ell}A_i(\Theta_0)|\mathcal{O}_i]\right],\\ &= \mathbb{E}[W_{i\ell}A_i(\Theta_0)].\end{aligned}$$

Or, on a $\mathbb{E}[W_{i\ell}A_i(\Theta_0)]=0$ pour tout $i=1,\ldots,n$ et $\ell=1,\ldots,q$.
D'où, pour tout $i=1,\ldots,n$ et $\ell=1,\ldots,q$, $\mathbb{E}[\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}W_{i\ell}A_i(\Theta_0)]=0$.
Dans le second cas (b),

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}W_{i\ell}A_i(\Theta_0)|\mathcal{O}_i\right]\right] &= \mathbb{E}\left[\frac{1}{\pi(\omega_0,\mathcal{O}_i)}W_{i\ell}\mathbb{E}[\delta_i A_i(\Theta_0)|\mathcal{O}_i]\right],\\ &= \mathbb{E}\left[\frac{1}{\pi(\omega_0,\mathcal{O}_i)}W_{i\ell}\mathbb{E}[\delta_i|\mathcal{O}_i]\mathbb{E}[A_i(\Theta_0)|\mathcal{O}_i]\right],\\ &= \mathbb{E}[W_{i\ell}A_i(\Theta_0)],\\ \mathbb{E}\left[\mathbb{E}\left[\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}W_{i\ell}A_i(\Theta_0)|\mathcal{O}_i\right]\right] &= 0,\end{aligned}$$

ce qui implique que dans le second cas, nous avons aussi $\mathbb{E}[\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}W_{i\ell}A_i(\Theta_0)]=0$.
De plus, par l'hypothèse (**B5**), on a

$$\text{var}\left(\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}W_{i\ell}A_i(\Theta_0)\right)=\mathbb{E}\left[\frac{\delta_i}{\pi_i^2(\omega_0,\mathcal{O}_i)}W_{i\ell}^2A_i^2(\Theta_0)\right]<\infty,$$

pour tout $i=1,\ldots,n$ et $\ell=1,\ldots,q$.
Ainsi, par la loi faible des grands nombres, nous avons $\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}W_{i\ell}A_i(\Theta_0)$ converge en probabilité vers 0 lorsque $n\to\infty$.

Par des arguments similaires, on montre que pour tout $j=1,...,p$, $t\in\{1,2\}$, $\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\omega_0,\mathcal{O}_i)}X_{ij}B_{ti}(\Theta_0)$ converge en probabilité vers 0 quand n tend vers l'infini.

Finalement, $n^{-1/2}U_{W,n}(\Theta_0,\omega_0)$ converge en probabilité vers **0** lorsque $n\to\infty$. Par conséquent, nous avons $n^{-1/2}U_{W,n}(\Theta_0,\hat{\omega}_n)$ converge vers **0** lorsque n tend vers l'infini.

Enfin, nous montrons que $n^{-1/2}\frac{\partial U_{W,n}(\Theta,\hat{\omega}_n)}{\partial\Theta^\top}$ converge uniformément en probabilité vers une fonction donnée $\Sigma_1(\Theta,\omega_0)$ dans un voisinage ouvert de Θ_0 lorsque $n\to\infty$.

Pour prouver cela, prenons $\Theta \in \mathcal{V}_{\Theta_0}$. Nous avons

$$n^{-1/2}\frac{\partial U_{W,n}(\Theta,\hat{\omega}_n)}{\partial \Theta^\top} = \left(n^{-1/2}\frac{\partial U_{W,n}(\Theta,\hat{\omega}_n)}{\partial \Theta^\top} - n^{-1/2}\frac{\partial U_{W,n}(\Theta,\omega_0)}{\partial \Theta^\top}\right) + n^{-1/2}\frac{\partial U_{W,n}(\Theta,\omega_0)}{\partial \Theta^\top}.$$

Sous les hypothèses **(B2)** à **(B5)** et la consistance de $\hat{\omega}_n$ vers ω_0, on peut montrer que $\left(n^{-1/2}\frac{\partial U_{W,n}(\Theta,\hat{\omega}_n)}{\partial \Theta^\top} - n^{-1/2}\frac{\partial U_{W,n}(\Theta,\omega_0)}{\partial \Theta^\top}\right)$ converge en probabilité vers $\mathbf{0}$.
Ensuite, en utilisant le corollaire 3.1 de Newey (1991) sous les hypothèses **(B1)** et **(B5)**, $n^{-1/2}\frac{\partial U_{W,n}(\Theta,\hat{\omega}_n)}{\partial \Theta^\top}$ converge uniformément en probabilité vers $\Sigma_1(\Theta,\omega_0)$ sur $\mathcal{V}_{\Theta_0}$ lorsque $n \to \infty$.

Les conditions du théorème de la fonction inverse de Foutz (1977) sont vérifiées. Ainsi, nous concluons que $\widehat{\Theta}_{W,n}$ converge en probabilité vers Θ_0.

Prouvons maintenant la normalité asymptotique de $\hat{\Theta}_{W,n}$. Pour ce faire, nous faisons un développement de Taylor de $U_{W,n}(\hat{\Theta}_n,\hat{\omega}_n)$ en (Θ_0,ω_0). On obtient

$$\mathbf{0} = U_{W,n}(\hat{\Theta}_n,\hat{\omega}_n) = U_{W,n}(\Theta_0,\omega_0) + \frac{\partial U_{W,n}(\Theta_0,\omega_0)}{\partial \Theta^\top}(\hat{\Theta}_n - \Theta_0) + \frac{\partial U_{W,n}(\Theta_0,\omega_0)}{\partial \omega^\top}(\hat{\omega}_n - \omega_0) + \mathbf{O}_{\mathbb{P}}(1).$$

Il s'en suit que

$$\mathbf{0} = U_{W,n}(\Theta_0,\omega_0) + n^{-1/2}\frac{\partial U_{W,n}(\Theta_0,\omega_0)}{\partial \Theta^\top}\sqrt{n}(\hat{\Theta}_n - \Theta_0) + n^{-1/2}\frac{\partial U_{W,n}(\Theta_0,\omega_0)}{\partial \omega^\top}\sqrt{n}(\hat{\omega}_n - \omega_0) + \mathbf{O}_{\mathbb{P}}(1).$$

En outre, $n^{-1/2}\frac{\partial U_{W,n}(\Theta_0,\omega_0)}{\partial \omega^\top}$ converge en probabilité vers $\Sigma_3(\Theta_0,\omega_0)$. Ainsi, on a

$$\begin{aligned}
\sqrt{n}(\hat{\Theta}_n - \Theta_0) &= -\left(\frac{1}{\sqrt{n}}\frac{\partial U_{W,n}(\Theta_0,\omega_0)}{\partial \Theta^\top}\right)^{-1}\left(U_{W,n}(\Theta_0,\omega_0) + \frac{1}{\sqrt{n}}\frac{\partial U_{W,n}(\Theta_0,\omega_0)}{\partial \omega^\top}\sqrt{n}(\hat{\omega}_n - \omega_0)\right) \\
&\quad + \mathbf{o}_{\mathbb{P}}(1). \\
&= -\left(\frac{1}{\sqrt{n}}\frac{\partial U_{W,n}(\Theta_0,\omega_0)}{\partial \Theta^\top}\right)^{-1}\left(U_{W,n}(\Theta_0,\omega_0) + n^{-1/2}\sum_{i=1}^{n}\frac{\delta_i - \pi(\omega,\mathcal{O}_i)}{\pi(\omega,\mathcal{O}_i)\big(1-\pi(\omega,\mathcal{O}_i)\big)}\right. \\
&\quad \left. \times \Sigma_3(\Theta_0,\omega_0)\Omega^{-1}(\omega_0)\dot{\pi}(\omega,\mathcal{O}_i)\right) + \mathbf{O}_{\mathbb{P}}(1).
\end{aligned}$$

Nous avons $-\frac{1}{\sqrt{n}}\frac{\partial U_{W,n}(\Theta_0,\omega_0)}{\partial \Theta^\top}$ converge en probabilité vers Σ_1. De plus, on a

$$\text{var}\left[\frac{\delta - \pi(\omega_0,\mathcal{O})}{\pi(\omega_0,\mathcal{O})\big(1-\pi(\omega_0,\mathcal{O})\big)}\dot{\pi}(\omega_0,\mathcal{O})\right] = \Omega(\omega_0).$$

Par conséquent, nous avons

$$\mathbf{var}\Big[\frac{\delta-\pi(\omega_0,\mathcal{O})}{\pi(\omega_0,\mathcal{O})\big(1-\pi(\omega_0,\mathcal{O})\big)}\Sigma_3(\Theta_0,\omega_0)\Omega^{-1}(\omega_0)\dot{\pi}(\omega_0,\mathcal{O})\Big]=\Sigma_3(\Theta_0,\omega_0)\Omega^{-1}(\omega_0)\big[\Sigma_3(\Theta_0,\omega_0)\big]^\top.$$

En outre,

$$\begin{aligned}
&\mathbf{cov}\Big[\frac{\delta_i}{\pi(\Theta_0,\omega_0)}\Phi_i(\Theta_0),\frac{\delta_i-\pi(\omega_0,\mathcal{O}_i)}{\pi(\omega_0,\mathcal{O}_i)\big(1-\pi(\omega_0,\mathcal{O}_i)\big)}\Sigma_3(\Theta_0,\omega_0)\Omega^{-1}(\omega_0)\dot{\pi}(\omega_0,\mathcal{O}_i)\Big]\\
&=\mathbb{E}\Big[\frac{\delta_i}{\pi^2(\omega_0,\mathcal{O}_i)}\Phi_i(\Theta_0)\big(\dot{\pi}(\omega_0,\mathcal{O}_i)\big)^\top\frac{\delta_i-\pi(\omega_0,\mathcal{O}_i)}{1-\pi(\omega,\mathcal{O}_i)}\Big]\Omega^{-1}(\omega_0)\big[\Sigma(\Theta_0,\omega_0)\big]^\top,\\
&=\mathbb{E}\Big[\frac{\delta_i}{\pi^2(\omega_0,\mathcal{O}_i)}\Phi_i(\Theta_0)\big(\dot{\pi}(\omega_0,\mathcal{O}_i)\big)^\top\Big]\Omega^{-1}(\omega_0)\big[\Sigma_3(\Theta_0,\omega_0)\big]^\top,\\
&=-\Sigma_3(\Theta_0,\omega_0)\Omega^{-1}(\omega_0)\big[\Sigma_3(\Theta_0,\omega_0)\big]^\top.
\end{aligned}$$

De plus,

$$\mathbf{var}\Big(U_{W,n}(\Theta_0,\omega_0)\Big)=\Sigma_2(\Theta_0,\omega_0).$$

Donc,

$$\begin{aligned}
&\mathbf{var}\Big(U_{W,n}(\Theta_0,\omega_0)+n^{-1/2}\sum_{i=1}^n\frac{\delta_i-\pi(\omega,\mathcal{O}_i)}{\pi(\omega,\mathcal{O}_i)\big(1-\pi(\omega,\mathcal{O}_i)\big)}\Sigma_3(\Theta_0,\omega_0)\Omega^{-1}(\omega_0)\dot{\pi}(\omega,\mathcal{O}_i)\Big)\\
&=\Sigma_2(\Theta_0,\omega_0)-\Sigma_3(\Theta_0,\omega_0)\Omega(\omega_0)^{-1}\big[\Sigma_3(\Theta_0,\omega_0)\big]^\top.
\end{aligned}$$

Par conséquent, en utilisant le théorème centrale limité et le théorème de Slutsky, nous avons $\sqrt{n}(\hat{\Theta}_{W,n}-\Theta_0)$ converge en distribution vers une loi normale de moyenne nulle et de matrice de variance covariance Σ_W.
Cette première approche permet d'obtenir des estimations consistantes des coefficients de régression du modèle, à condition que π soit correctement spécifié. Cependant, une mauvaise spécification peut conduire à une estimation biaisée des coefficients de régression. Pour surmonter cette difficulté, nous considérons un estimateur non paramétrique à noyau pour les probabilités de sélection. Ainsi, dans les sections suivantes, nous considérons un modèle non paramétrique pour estimer les probabilités de sélection.

3.3.2 Estimateur IPW semi-paramétrique

Dans ce qui suit, nous supposons que $\mathcal{O}$ est un vecteur contenant un mélange de variables discrètes, catégorielles et continues. Soit $\mathcal{O}=\big(\mathcal{O}^b,\mathcal{O}^c\big)$, où $\mathcal{O}^c\in\mathbb{R}^d$ est un vecteur de variables aléatoires continues de dimension d et $\mathcal{O}^b$ est un vecteur de variables aléatoires discrètes et catégorielles de dimension b.
De plus, nous rappelons que les variables qui contiennent au moins une donnée man-

quante $\mathcal{Z} := \left\{\mathbf{X}^{(mis)}, \mathbf{W}^{(mis)}\right\}$ sont MAR. Soit K une fonction noyau d'ordre r de d variables avec un support fini et soit $K_h(.) = K(./h)$, où h est la fenêtre de lissage qui satisfait les conditions que nous donnons dans la suite. Un estimateur à noyau consistant de $\pi(\mathcal{O})$ est donné par

$$\widehat{\pi}(o) = \widehat{\pi}(o^b, o^c) = \sum_{k=1}^{n} \frac{\delta_k K_h(\mathcal{O}_k^b = o^b, \mathcal{O}_k^c - o^c)}{\sum_{j=1}^{n} K_h(\mathcal{O}_j^b = o^b, \mathcal{O}_j^c - o^c)}.$$

où $\widehat{\pi}(\mathcal{O}) - \pi(\mathcal{O}) = \eta_n$ avec $\eta_n = \{nh^{2r} + 1/(nh^{2d})\}^{1/2}$, pour plus de détails sur l'estimateur $\widehat{\pi}(\mathcal{O})$ on peut se référer aux travaux de Wang *et al.* (1997) et Wang et Wang (2001).
Soit $f(o)$ la fonction de densité de probabilité de $\mathcal{O}$ en $\mathcal{O} = o$. Considérons ensuite $\widehat{f}(o) = \frac{1}{h^d} \sum_{k=1}^{n} K_h(\mathcal{O}_k^b = o^b, \mathcal{O}_k^c - o^c)$ un estimateur de $f(o)$.
L'estimateur IPW semi-paramétrique $\widehat{\Theta}_{Ws}$ de Θ est la solution de l'équation pondérée

$$U_{Ws,n}(\Theta, \widehat{\pi}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\delta_i}{\widehat{\pi}(\mathcal{O}_i)} \Phi_i(\Theta) = 0. \tag{3.7}$$

À présent, nous donnons les hypothèses de régularité suivantes pour prouver les propriétés asymptotiques de l'estimateur $\widehat{\Theta}_{Ws}$.

Hypothèses de régularité

B6 Pour tout $o^b \in \mathcal{O}^b$ et o^c dans le support de $\mathcal{O}^c$, il existe C_1 tel que la probabilité de sélection $\pi(o^b, o^c) > C_1 > 0$. De plus, $\pi(o^b, o^c)$ a r dérivées partielles continues et bornées par rapport aux composantes continues de $(\mathcal{O}^d, \mathcal{O}^c)$ presque partout.

B7 La fonction $K(.)$ est un noyau d'ordre r, c'est-à-dire $\int K(u)du = 1$, $\int u^m K(u)du = 0$ pour $m = 1, ..., (r-1)$, $\int u^r K(u)du \neq 0$ et $\int K^2(u)du < \infty$. De plus, h est tel que $nh^{2r} \to 0$ et $nh^{2d} \to \infty$, lorsque $n \to \infty$.

B8 La fonction de densité $f(.)$ de $(\mathcal{O}^b, \mathcal{O}^c)$ est toujours bornée au voisinage de zéro et possède r dérivées partielles par rapport aux composantes continues de $(\mathcal{O}^b, \mathcal{O}^c)$ qui sont continues et bornées presque partout.

B9 $\mathbb{E}\left[\frac{\Phi_1(\Theta)(\Phi_1(\Theta))^\top}{\pi(\mathcal{O}_1^b, \mathcal{O}_1^c)}\right]$ est finie et définie positive dans un voisinage de Θ_0.

B10 Les dérivées premières de $U_{Ws,n}(\Theta, \pi)$ par rapport à Θ existent presque sûrement dans un voisinage de Θ_0. De plus, dans un tel voisinage, les dérivées premières sont uniformément bornées supérieurement par une fonction de $\mathcal{O} = (\mathcal{O}^b, \mathcal{O}^c)$ dont les espérances existent.

Théorème 3.2 *Supposons que les hypothèses* (***B1***), (***B4***) *et* (***B6***) *à* (***B10***) *sont vérifiées. Alors* $\widehat{\Theta}_{Ws}$ *est un estimateur consistant de* Θ *et* $\sqrt{n}(\widehat{\Theta}_{Ws} - \Theta_0)$ *est distribué asymptotiquement suivant une normale multivariée centrée de matrice de covariance* Σ_{Ws}*, où*

$$\Sigma_{Ws} := \left[\Sigma_1(\Theta_0)^{-1}\right]\Big[\Sigma_2(\Theta_0,\pi) - \left[\Sigma_2^*(\Theta_0,\pi) - \Sigma_0^*(\Theta_0,\pi)\right]\Big]\left[\Sigma_1(\Theta_0)^{-1}\right]^\top,$$

avec

$$\begin{aligned}
\Sigma_2(\Theta,\pi) &= \mathbb{E}\Big[\frac{\left[\Phi_1(\Theta)\right]^{\otimes 2}}{\pi(\mathcal{O}^b,\mathcal{O}^c)}\Big], \qquad \Sigma_2^*(\Theta,\pi) = \mathbb{E}\Big[\frac{\left[\Phi^*(\Theta)^{\otimes 2}\right]}{\pi(\mathcal{O}^b,\mathcal{O}^c)}\Big],\\
\Sigma_0^*(\Theta,\pi) &= \mathbb{E}\Big[\left[\Phi_1^*(\Theta)\right]^{\otimes 2}\Big] \quad \textit{et} \quad \Phi_1^*(\Theta) = \mathbb{E}\Big[\Phi_1(\Theta)|\mathcal{O}^b,\mathcal{O}^c\Big].
\end{aligned}$$

Remarque 3.2 *Un estimateur consistant de* Σ_{Ws} *est défini par:*

$$\begin{aligned}
\widehat{\Sigma}_{Ws,n} := &\left[\Sigma_{1,n}(\Theta_{Ws},\widehat{\pi})^{-1}\right]\Big[\Sigma_{2,n}(\widehat{\Theta}_{Ws},\widehat{\pi}) - \left[\widehat{\Sigma}_{2,n}^*(\widehat{\Theta}_{Ws},\widehat{\pi}) - \widehat{\Sigma}_{0,n}^*(\widehat{\Theta}_{Ws},\widehat{\pi})\right]\Big]\\
&\times\left[\Sigma_{1,n}(\widehat{\Theta}_{Ws},\widehat{\pi})^{-1}\right]^\top,
\end{aligned}$$

où

$$\begin{aligned}
\widehat{\Sigma}_{2,n}^*(\Theta,\pi) &= \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi^2(\mathcal{O}_i^b,\mathcal{O}_i^c)}\Big[\widehat{\Phi}_i^*(\Theta)\Big]^{\otimes 2},\\
\widehat{\Sigma}_{0,n}^*(\Theta,\pi) &= \frac{1}{n}\sum_{i=1}^n \Big[\widehat{\Phi}_i^*(\Theta)\Big]^{\otimes 2},
\end{aligned}$$

avec

$$\widehat{\Phi}_i^*(\Theta) = \frac{\sum_{k=1}^n \delta_k \Phi_k(\Theta) K_h(\mathcal{O}_k^b = \mathcal{O}_i^b, \mathcal{O}_k^c - \mathcal{O}_i^c)}{\sum_{\ell=1}^n \delta_\ell K_h(\mathcal{O}_\ell^b = \mathcal{O}_i^b, \mathcal{O}_\ell^c - \mathcal{O}_i^c)}.$$

La preuve de la consistance de $\widehat{\Theta}_{Ws}$ repose sur le fait que $\frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\widehat{\pi}(\mathcal{O}_i)}\Phi_i(\Theta)$ converge vers une limite qui est nulle en Θ_0. Ainsi, nous appliquons le théorème de la fonction inverse de Foutz (1977). Les conditions du théorème Foutz (1977) sont données sous formes de lemmes techniques que nous prouvons.

Lemme 3.1 *Supposons que les hypothèses* (***B4***) *et* (***B6***) *à* (***B9***) *sont vérifiées. Alors,* $\frac{1}{\sqrt{n}}U_{Ws,n}(\Theta,\widehat{\pi})$ *converge en probabilité vers* $\mathbf{0}$.

Preuve du Lemme 3.1

On a

$$\frac{1}{\sqrt{n}}U_{Ws,n}(\Theta,\widehat{\pi}) = \Big(\frac{1}{\sqrt{n}}U_{Ws,n}(\Theta,\widehat{\pi}) - \frac{1}{\sqrt{n}}U_{W,n}(\Theta,\pi)\Big) + \frac{1}{\sqrt{n}}U_{W,n}(\Theta,\pi).$$

Ensuite, nous montrons que $U_{Ws,n}(\Theta,\widehat{\pi}) - U_{Ws,n}(\Theta,\pi)$ converge en probabilité vers $\mathbf{0}$ lorsque n tend vers l'infini. On a

$$\begin{aligned}
&\frac{1}{\sqrt{n}}U_{Ws,n}(\theta,\widehat{\pi}) - \frac{1}{\sqrt{n}}U_{W,n}(\theta,\pi)\\
=\;& \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{\widehat{\pi}(\mathcal{O}_i^b,\mathcal{O}_i^c)} - \frac{1}{\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)}\right]\delta_i\Phi_i(\theta),\\
=\;& -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\widehat{\pi}(\mathcal{O}_i^b,\mathcal{O}_i^c)-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b,\mathcal{O}_i^c)} + O_{\mathbb{P}}\Big(\big(\widehat{\pi}(\mathcal{O}_i^b,\mathcal{O}_i^c)-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\big)^2\Big)\right]\delta_i\Phi_i(\theta),\\
=\;& -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\widehat{\pi}(\mathcal{O}_i^b,\mathcal{O}_i^c)-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b,\mathcal{O}_i^c)} + O_{\mathbb{P}}(\eta_n^2)\right]\delta_i\Phi_i(\theta),\\
=\;& -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\frac{\sum_{j=1}^{n}\delta_j K_h(\mathcal{O}_j^b=\mathcal{O}_i^b,\mathcal{O}_j^c-\mathcal{O}_i^c)}{\sum_{j=1}^{n}K_h(\mathcal{O}_j^b=\mathcal{O}_i^b,\mathcal{O}_j^c-\mathcal{O}_i^c)}-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b,\mathcal{O}_i^c)} + O_{\mathbb{P}}(\eta_n^2)\right]\delta_i\Phi_i(\theta),\\
=\;& -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{\big[\delta_j-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\big]K_h(\mathcal{O}_j^d=\mathcal{O}_i^d,\mathcal{O}_j^c-\mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b,\mathcal{O}_i^c)h^d\widehat{f}(\mathcal{O}^d=\mathcal{O}_i^d,\mathcal{O}^c-\mathcal{O}_i^c)} + O_{\mathbb{P}}(\eta_n^2)\right]\delta_i\Phi_i(\theta),\\
=\;& -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{\big[\delta_j-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\big]\big[\delta_i-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\big]K_h(\mathcal{O}_j^b=\mathcal{O}_i^b,\mathcal{O}_j^c-\mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b,\mathcal{O}_i^c)h^d\widehat{f}(\mathcal{O}^b=\mathcal{O}_i^b,\mathcal{O}^c-\mathcal{O}_i^c)}\right]\Phi_i(\theta)\\
&-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\big[\delta_j-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\big]K_h(\mathcal{O}_j^b=\mathcal{O}_i^b,\mathcal{O}_j^c-\mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b,\mathcal{O}_i^c)h^d\widehat{f}(\mathcal{O}^b=\mathcal{O}_i^b,\mathcal{O}^c-\mathcal{O}_i^c)}\right]\Phi_i(\theta) + \mathbf{O}_{\mathbb{P}}(\eta_n).
\end{aligned}$$

Par la loi faible de grands nombres, nous avons

$$\frac{1}{n}\sum_{j=1}^{n}\left[\frac{\big[\delta_j-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\big]K_h(\mathcal{O}_j^b=\mathcal{O}_i^b,\mathcal{O}_j^c-\mathcal{O}_i^c)}{\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)h^d\widehat{f}(\mathcal{O}^b=\mathcal{O}_i^b,\mathcal{O}^c-\mathcal{O}_i^c)}\right]$$

converge en probabilité vers $\mathbf{0}$.

De plus, en utilisant le théorème de Slutsky sous l'hypothèse **B4**, on a

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{\left[\delta_i-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\right]\left[\delta_j-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\right]K_h(\mathcal{O}_j^b=\mathcal{O}_i^b,\mathcal{O}_j^c-\mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b,\mathcal{O}_i^c)h^d\widehat{f}(\mathcal{O}^b=\mathcal{O}_i^b,\mathcal{O}^c-\mathcal{O}_i^c)}\right]\left[\Phi_i(\Theta)\right]$$

$$+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{\left[\delta_j-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\right]K_h(\mathcal{O}_j^b=\mathcal{O}_i^b,\mathcal{O}_j^c-\mathcal{O}_i^c)}{\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)h^d\widehat{f}(\mathcal{O}^b=\mathcal{O}_i^b,\mathcal{O}^c-\mathcal{O}_i^c)}\right]\left[\Phi_i(\Theta)\right]\xrightarrow{\mathbb{P}}\mathbf{0}.$$

Par conséquent, il suit que $\frac{1}{\sqrt{n}}U_{Ws,n}(\Theta,\widehat{\pi})-\frac{1}{\sqrt{n}}U_{W,n}(\Theta,\pi)$ converge en probabilité vers $\mathbf{0}$ lorsque n tend vers l'infini. Aussi, par la loi faible des grands nombres on a $\frac{1}{\sqrt{n}}U_{W,n}(\Theta,\pi)$ converge en probabilité vers $\mathbf{0}$. Donc, en utilisant le théorème de Slutsky, on obtient la convergence en probabilité de $\frac{1}{\sqrt{n}}U_{Ws,n}(\Theta,\widehat{\pi})$ vers $\mathbf{0}$.

Lemme 3.2 *Supposons que les hypothèses* (***B1***)*,* (***B4***) *et* (***B6***) *à* (***B10***) *sont satisfaites. Alors,* $\frac{1}{\sqrt{n}}\mathcal{H}_{Ws,n}(\Theta,\widehat{\pi})$ *converge uniformément en probabilité vers* $\Sigma_1(\Theta,\pi)$ *dans un voisinage ouvert de* Θ_0*, où* $\mathcal{H}_{Ws,n}(\Theta,\pi)=\frac{\partial U_{Ws,n}(\Theta,\pi)}{\partial\Theta^\top}$.

Preuve du Lemme 3.2

On montre que

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\mathcal{H}_{Ws,n}(\Theta,\widehat{\pi})-\frac{1}{\sqrt{n}}\mathcal{H}_{W,n}(\Theta,\pi)\\
=\;&\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{\widehat{\pi}(\mathcal{O}_i^b,\mathcal{O}_i^c)}-\frac{1}{\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)}\right]\delta_i\left[\frac{\partial\Phi_i(\theta)}{\partial\Theta^\top}\right],\\
=\;&-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{\left[\delta_j-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\right]K_h(\mathcal{O}_j^d=\mathcal{O}_i^d,\mathcal{O}_j^c-\mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b,\mathcal{O}_i^c)h^d\widehat{f}(\mathcal{O}^d=\mathcal{O}_i^d,\mathcal{O}^c-\mathcal{O}_i^c)}+O_{\mathbb{P}}(\eta_n^2)\right]\delta_i\left[\frac{\partial\Phi_i(\theta)}{\partial\Theta^\top}\right],\\
=\;&-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{\left[\delta_j-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\right]\left[\delta_i-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\right]K_h(\mathcal{O}_j^b=\mathcal{O}_i^b,\mathcal{O}_j^c-\mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b,\mathcal{O}_i^c)h^d\widehat{f}(\mathcal{O}^b=\mathcal{O}_i^b,\mathcal{O}^c-\mathcal{O}_i^c)}\right]\left[\frac{\partial\Phi_i(\theta)}{\partial\Theta^\top}\right]\\
&-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\left[\delta_j-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\right]K_h(\mathcal{O}_j^b=\mathcal{O}_i^b,\mathcal{O}_j^c-\mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b,\mathcal{O}_i^c)h^d\widehat{f}(\mathcal{O}^b=\mathcal{O}_i^b,\mathcal{O}^c-\mathcal{O}_i^c)}\right]\left[\frac{\partial\Phi_i(\theta)}{\partial\Theta^\top}\right]+\mathbf{O}_{\mathbb{P}}(\eta_n).
\end{aligned}$$

De plus, en utilisant les mêmes arguments que dans la preuve du Lemme 3.1, nous montrons que $\frac{1}{\sqrt{n}}\mathcal{H}_{Ws,n}(\Theta,\widehat{\pi})-\frac{1}{\sqrt{n}}\mathcal{H}_{W,n}(\Theta,\pi)$ converge en probabilité vers la matrice nulle. Ensuite, par la loi faible des grands nombres, $\frac{1}{\sqrt{n}}\mathcal{H}_{Ws,n}(\Theta,\pi)$ converge en probabilité vers $\Sigma_1(\Theta,\pi)$. D'autre part, en utilisant le théorème de Slutsky, nous avons $\frac{1}{\sqrt{n}}\mathcal{H}_{W,n}(\Theta,\widehat{\pi})$ converge en probabilité vers $\Sigma_1(\Theta,\pi)$. De plus, sous les hypothèses (**B1**) et (**B10**), nous avons la convergence uniforme en probabilité de $\frac{1}{\sqrt{n}}\mathcal{H}_{Ws,n}(\Theta,\widehat{\pi})$ vers $\Sigma_1(\Theta,\pi)$ dans un voisinage ouvert de Θ_0.

Par conséquent, en utilisant le théorème de la fonction inverse de Foutz (1977), il existe une unique solution consistante de l'équation d'estimation $U_{Ws,n}(\Theta, \widehat{\pi}) = \mathbf{0}$ dans un voisinage de Θ_0. Il s'ensuit que $\widehat{\Theta}_{Ws}$ est un estimateur consistant de Θ.

Maintenant, nous établissons la distribution asymptotique de $\sqrt{n}(\widehat{\Theta}_{Ws,n} - \Theta_0)$. D'abord nous prouvons le Lemme 3.3 suivant.

Lemme 3.3 *Supposons que les hypothèses* (***B6***) *à* (***B10***) *sont satisfaites. Alors,* $U_{Ws,n}(\Theta, \widehat{\pi}) - U_{Ws,n}(\Theta, \pi) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{\delta_i - \pi(\mathcal{O}_i)}{\pi(\mathcal{O}_i)} \right] \Phi_i^*(\Theta) + \mathbf{O}_{\mathbb{P}}(\eta_n)$.

Preuve du Lemme 3.3

On a

$$
\begin{aligned}
& U_{Ws,n}(\theta, \widehat{\pi}) - U_{W,n}(\theta, \pi) \\
= & -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{\widehat{\pi}(\mathcal{O}_i^b, \mathcal{O}_i^c) - \pi(\mathcal{O}_i^b, \mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b, \mathcal{O}_i^c)} + O_{\mathbb{P}}\Big(\big(\widehat{\pi}(\mathcal{O}_i^b, \mathcal{O}_i^c) - \pi(\mathcal{O}_i^b, \mathcal{O}_i^c) \big)^2 \Big) \right] \delta_i \Phi_i(\theta), \\
= & -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{\widehat{\pi}(\mathcal{O}_i^b, \mathcal{O}_i^c) - \pi(\mathcal{O}_i^b, \mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b, \mathcal{O}_i^c)} + \mathbf{O}_{\mathbb{P}}(\eta_n^2) \right] \delta_i \Phi_i(\theta),
\end{aligned}
$$

$$
\begin{aligned}
= & -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{\frac{\sum_{j=1}^{n} \delta_j K_h(\mathcal{O}_j^b = \mathcal{O}_i^b, \mathcal{O}_j^c - \mathcal{O}_i^c)}{\sum_{j=1}^{n} K_h(\mathcal{O}_j^b = \mathcal{O}_i^b, \mathcal{O}_j^c - \mathcal{O}_i^c)} - \pi(\mathcal{O}_i^b, \mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b, \mathcal{O}_i^c)} + O_{\mathbb{P}}(\eta_n^2) \right] \delta_i \Phi_i(\theta), \\
= & -\frac{1}{\sqrt{n^3}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\frac{\big[\delta_j - \pi(\mathcal{O}_i^b, \mathcal{O}_i^c) \big] K_h(\mathcal{O}_j^b = \mathcal{O}_i^b, \mathcal{O}_j^c - \mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b, \mathcal{O}_i^c) h^d \widehat{f}(\mathcal{O}^b = \mathcal{O}_i^b, \mathcal{O}^c - \mathcal{O}_i^c)} + O_{\mathbb{P}}(\eta_n^2) \right] \delta_i \Phi_i(\theta), \\
= & -\frac{1}{\sqrt{n^3}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\frac{\big[\delta_j - \pi(\mathcal{O}_i^b, \mathcal{O}_i^c) \big] \big[\delta_i - \pi(\mathcal{O}_i^b, \mathcal{O}_i^c) \big] K_h(\mathcal{O}_j^b = \mathcal{O}_i^b, \mathcal{O}_j^c - \mathcal{O}_i^c) \Phi_i(\theta)}{\pi^2(\mathcal{O}_i^b, \mathcal{O}_i^c) h^d \widehat{f}(\mathcal{O}^b = \mathcal{O}_i^b, \mathcal{O}^c - \mathcal{O}_i^c)} \right] \\
& -\frac{1}{\sqrt{n^3}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\frac{\pi(\mathcal{O}_i^b, \mathcal{O}_i^c) \big[\delta_j - \pi(\mathcal{O}_i^b, \mathcal{O}_i^c) \big] K_h(\mathcal{O}_j^b = \mathcal{O}_i^b, \mathcal{O}_j^c - \mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b, \mathcal{O}_i^c) h^d \widehat{f}(\mathcal{O}^b = \mathcal{O}_i^b, \mathcal{O}^c - \mathcal{O}_i^c)} \right] \Phi_i(\theta) + \mathbf{O}_{\mathbb{P}}(\eta_n), \\
= & -\mathbb{T}_{1n} - \mathbb{T}_{2n} + \mathbf{O}_{\mathbb{P}}(\eta_n),
\end{aligned}
$$

où

$$
\begin{aligned}
\mathbb{T}_{1n} &= \frac{1}{\sqrt{n^3}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\frac{\Big[\delta_j - \pi(\mathcal{O}_j^d, \mathcal{O}_j^c) \Big] \Big[\delta_i - \pi(\mathcal{O}_i^b, \mathcal{O}_i^c) \Big] K_h(\mathcal{O}_j^b = \mathcal{O}_i^b, \mathcal{O}_j^c - \mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b, \mathcal{O}_i^c) h^d \widehat{f}(\mathcal{O}^b = \mathcal{O}_i^b, \mathcal{O}^c - \mathcal{O}_i^c)} \right] \Phi_i(\Theta), \\
\mathbb{T}_{2n} &= \frac{1}{\sqrt{n^3}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\frac{\pi(\mathcal{O}_i^b, \mathcal{O}_i^c) \big[\delta_j - \pi_i(\mathcal{O}_i^b, \mathcal{O}_i^c) \big] K_h(\mathcal{O}_j^b = \mathcal{O}_i^b, \mathcal{O}_j^c - \mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b, \mathcal{O}_i^c) h^d \widehat{f}(\mathcal{O}^b = \mathcal{O}_i^b, \mathcal{O}^c - \mathcal{O}_i^c)} \right] \Phi_i(\theta).
\end{aligned}
$$

D'abord, on montre que $\mathbb{T}_{1n} = \mathbf{O}_{\mathbb{P}}(\eta_n)$. Pour le prouver, nous montrons que $\mathbb{E}(\mathbb{T}_{1n}) = \mathbf{O}(\eta_n)$ et $\mathbf{var}(\mathbb{T}_{1n}) = \mathbf{O}(\eta_n^2)$. Dans la suite, la notation $\mathbf{O}(a_n)$ désigne un vecteur colonne ou une matrice dont les composantes sont uniformément $O(a_n)$.
Posons

$$\tau_{ij} = \left[\frac{\Big[\delta_j - \pi(\mathcal{O}_i^b, \mathcal{O}_i^c)\Big]\Big[\delta_i - \pi(\mathcal{O}_i^b, \mathcal{O}_i^c)\Big] K_h(\mathcal{O}_j^b = \mathcal{O}_i^b, \mathcal{O}_j^c - \mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b, \mathcal{O}_i^c) h^d \widehat{f}(\mathcal{O}^b = \mathcal{O}_i^b, \mathcal{O}^c - \mathcal{O}_i^c)}\right]\Phi_i(\Theta).$$

On remarque que

$$\mathbb{E}(\tau_{ij}) = \mathbb{E}\Big[\mathbb{E}\Big(\tau_{ij}\Big|\mathcal{O}_j^b = \mathcal{O}_i^b, \mathcal{O}_j^c - \mathcal{O}_i^c\Big)\Big] = \begin{cases} \mathbf{0} & \text{si } \ i \neq j \\ \mathbb{E}\left[\frac{\left[1-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\right]}{\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)}\Phi_i(\Theta)\right] & \text{si } \ i = j, \end{cases}$$

ensuite

$$\begin{aligned} \mathbb{E}\left[\frac{\left[1-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\right]}{\pi^2(\mathcal{O}_i^b,\mathcal{O}_i^c)}\Phi_i(\Theta)\right] &= \mathbb{E}\left(\mathbb{E}\left[\frac{\left[1-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\right]}{\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)}\Phi_i(\Theta)\Big|\mathcal{O}_i^b,\mathcal{O}_i^c\right]\right), \\ &= \mathbb{E}\left[\frac{\left[1-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\right]}{\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)}\Phi_i^*(\Theta)\right]. \end{aligned}$$

Donc, nous avons

$$\begin{aligned} \mathbb{E}(\mathbb{T}_{1n}) &= \frac{1}{\sqrt{n^3}}\sum_{j=1}^{n}\sum_{i=1}^{n}\mathbb{E}(\tau_{ij}) \\ &= \frac{1}{\sqrt{n^3}}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\left[1-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\right]}{\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)}\Phi_i^*(\Theta)\right], \\ &= \frac{1}{\sqrt{n}}\mathbb{E}\left[\frac{\left[1-\pi(\mathcal{O}_1^b,\mathcal{O}_1^c)\right]}{\pi(\mathcal{O}_1^b,\mathcal{O}_1^c)}\Phi_1^*(\Theta)\right], \\ \mathbb{E}(\mathbb{T}_{1n}) &= \mathbf{O}(\eta_n). \end{aligned}$$

De plus, on a

$$\begin{aligned} \mathbf{cov}(\tau_{ij}, \tau_{k\ell}) = \mathbb{E}\Bigg(&\left[\frac{\Big[\delta_i - \pi(\mathcal{O}_i^b, \mathcal{O}_i^c)\Big]\Big[\delta_j - \pi(\mathcal{O}_i^b, \mathcal{O}_i^c)\Big] K_h(\mathcal{O}_j^b = \mathcal{O}_i^b, \mathcal{O}_j^c - \mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b, \mathcal{O}_i^c) h^d \widehat{f}(\mathcal{O}^b = \mathcal{O}_i^b, \mathcal{O}^c - \mathcal{O}_i^c)}\right]\Phi_i(\Theta) \\ &\times\left[\frac{\Big[\delta_k - \pi(\mathcal{O}_k^b, \mathcal{O}_k^c)\Big]\Big[\delta_\ell - \pi(\mathcal{O}_k^b, \mathcal{O}_k^c)\Big] K_h(\mathcal{O}_\ell^b = \mathcal{O}_k^b, \mathcal{O}_\ell^c - \mathcal{O}_k^c)}{\pi^2(\mathcal{O}_k^b, \mathcal{O}_k^c) h^d \widehat{f}(\mathcal{O}^b = \mathcal{O}_k^b, \mathcal{O}^c - \mathcal{O}_k^c)}\right]\Phi_k^\top(\Theta)\Bigg). \end{aligned}$$

Alors, nous obtenons

$$\mathrm{cov}(\tau_{ij},\tau_{k\ell}) \;=\; \begin{cases} \mathbf{0} & \text{si } \; k\neq i,j \text{ et } \; \ell\neq i,j, \\ \mathbf{0} & \text{si } \; k\neq i,j \text{ et } \; \ell= i \text{ ou } j, \\ \mathbb{E}\left[\frac{\left[1-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\right]^2}{\pi^2(\mathcal{O}_i^b,\mathcal{O}_i^c)}\left[\Phi_i^*(\Theta)\right]^{\otimes 2}\right] & \text{si } \; k= j \text{ et } \ell= i. \end{cases}$$

En remarquant que chaque terme des sommes doubles dans l'espérance ci-dessous résulte d'un terme d'ordre $\mathbf{O}\Big(\frac{1}{nh^{2d}}\Big)$, on obtient

$$\begin{aligned}
\mathbb{E}\Big[\mathbb{T}_{1n}^{\otimes 2}\Big] &= \mathbb{E}\left[\frac{1}{n(nh^d)^2}\left(\sum_{i\neq i',j=j'}+\sum_{i=i',j=j'}+\sum_{i=j,i'=j',j\neq j'}\right)\tau_{ij}\tau_{i'j'}^\top\right], \\
&= \mathbb{E}\left[\frac{1}{n(nh^d)^2}\sum_{i\neq i',j=j'}\tau_{ij}\tau_{i'j'}^\top\right]+\mathbf{O}\Big(\frac{1}{nh^{2d}}\Big) \\
&= \mathbb{E}\left[\frac{1}{n(nh^d)^2}\sum_{i\neq i',i\neq j,i'\neq j}\tau_{ij}\tau_{i'j'}^\top\right]+\mathbf{O}\Big(\frac{1}{h^{2d}}\Big), \\
&= \mathbb{E}\left[\frac{1}{n(nh^d)^2}\sum_{i=1}^n\frac{\Big[\delta_i-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\Big]^2\Big[\Phi_i(\Theta)\Big]^{\otimes 2}}{\pi^4(\mathcal{O}_i^b,\mathcal{O}_i^c)\widehat{f}^2(\mathcal{O}^b=\mathcal{O}_i^b,\mathcal{O}^c-\mathcal{O}_i^c)}\sum_{j=1}^n\big(\delta_j-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\big)\times\right. \\
&\quad \left. K_h(\mathcal{O}_j^b=\mathcal{O}_i^b,\mathcal{O}_j^c-\mathcal{O}_i^c)\sum_{j'=1}^n\big(\delta_{j'}-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\big)K_h(\mathcal{O}_{j'}^b=\mathcal{O}_i^b,\mathcal{O}_{j'}^c-\mathcal{O}_i^c)\right]+\mathbf{O}\Big(\frac{1}{nh^{2d}}\Big), \\
&= \mathbb{E}\left[\frac{1}{(nh^d)^2}\sum_{i=1}^n\frac{\Big[\delta_1-\pi(\mathcal{O}_1^b,\mathcal{O}_1^c)\Big]^2\Big[\Phi_1(\Theta)\Big]^{\otimes 2}}{\pi^4(\mathcal{O}_1^b,\mathcal{O}_1^c)\widehat{f}^2(\mathcal{O}^b=\mathcal{O}_1^b,\mathcal{O}^c-\mathcal{O}_1^c)}\sum_{j=1}^n\big(\delta_j-\pi(\mathcal{O}_1^b,\mathcal{O}_1^c)\big)\times\right. \\
&\quad \left. K_h(\mathcal{O}_j^b=\mathcal{O}_1^b,\mathcal{O}_j^c-\mathcal{O}_1^c)\sum_{j'=1}^n\big(\delta_{j'}-\pi(\mathcal{O}_1^b,\mathcal{O}_1^c)\big)K_h(\mathcal{O}_{j'}^b=\mathcal{O}_1^b,\mathcal{O}_{j'}^c-\mathcal{O}_1^c)\right]+\mathbf{O}\Big(\frac{1}{nh^{2d}}\Big), \\
&= \mathbf{O}(\eta_n^2).
\end{aligned}$$

Par conséquent, $\mathbb{T}_{1n}=\mathbf{O}_{\mathbb{P}}(\eta_n)$.

Par ailleurs, nous avons

$$\begin{aligned}
\mathbb{T}_{2n} &= \frac{1}{\sqrt{n^3}}\sum_{j=1}^n\sum_{i=1}^n\left[\frac{\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\big[\delta_k-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)\big]K_h(\mathcal{O}_j^b=\mathcal{O}_i^b,\mathcal{O}_j^c-\mathcal{O}_i^c)}{\pi^2(\mathcal{O}_i^b,\mathcal{O}_i^c)h^d\widehat{f}(\mathcal{O}=\mathcal{O}_i^b,\mathcal{O}^c-\mathcal{O}_i^c)}\right]\Phi_i(\Theta), \\
&= \frac{1}{\sqrt{n}}\sum_{j=1}^n\left[\frac{\big[\delta_j-\pi(\mathcal{O}_j^b,\mathcal{O}_j^c)\big]}{\pi(\mathcal{O}_j^b,\mathcal{O}_j^c)}\frac{\frac{1}{n}\sum_{i=1}^n K_h(\mathcal{O}_i^b=\mathcal{O}_j^b,\mathcal{O}_i^c-\mathcal{O}_j^c)\Phi_i(\theta)}{h^d\widehat{f}(\mathcal{O}^b=\mathcal{O}_j^b,\mathcal{O}^c-\mathcal{O}_j^c)}\right],
\end{aligned}$$

$$\begin{aligned}\mathbb{T}_{2n} &= \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\frac{\delta_k-\pi(\mathcal{O}_k^b,\mathcal{O}_k^c)}{\pi(\mathcal{O}_k^b,\mathcal{O}_k^c)}\left[\Phi_k^*(\Theta)+\mathbf{O}_{\mathbb{P}}(\eta_n^2)\right],\\ &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\delta_j-\pi(\mathcal{O}_j^b,\mathcal{O}_j^c)}{\pi(\mathcal{O}_j^b,\mathcal{O}_j^c)}\Phi_j^*(\Theta)+\mathbf{O}_{\mathbb{P}}(\eta_n).\end{aligned}$$

Ce qui implique que

$$U_{Ws,n}(\Theta,\widehat{\pi})-U_{W,n}(\Theta,\pi)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\frac{\delta_i-\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)}{\pi(\mathcal{O}_i^b,\mathcal{O}_i^c)}\Big]\Phi_i^*(\Theta)+\mathbf{O}_{\mathbb{P}}(\eta_n).$$

Ce qui achève la preuve du Lemme 3.3.

Par un développement de Taylor de $U_{Ws,n}(\widehat{\Theta}_{Ws},\widehat{\pi})$ en $(\Theta_0,\widehat{\pi})$, on a

$$\mathbf{0}=U_{Ws,n}(\widehat{\Theta}_{Ws},\widehat{\pi})=U_{Ws,n}(\Theta_0,\widehat{\pi})+\left[\frac{1}{\sqrt{n}}\frac{\partial U_{Ws,n}(\Theta_0,\widehat{\pi})}{\partial\Theta^\top}\right]\sqrt{n}(\widehat{\Theta}_{Ws}-\Theta_0)+\mathbf{O}_{\mathbb{P}}(\eta_n).$$

De plus,

$$\begin{aligned}\sqrt{n}(\widehat{\Theta}_{Ws}-\Theta_0) &= \left[\frac{1}{\sqrt{n}}\frac{\partial U_{Ws,n}(\Theta_0,\widehat{\pi})}{\partial\Theta^\top}\right]^{-1}U_{Ws,n}(\Theta_0,\widehat{\pi})+\mathbf{O}_{\mathbb{P}}(\eta_n)\\ &= \left[\frac{1}{\sqrt{n}}\frac{\partial U_{Ws,n}(\Theta_0,\widehat{\pi})}{\partial\Theta^\top}\right]^{-1}\Big[U_{W,n}(\Theta_0,\pi)+\Big(U_{Ws,n}(\Theta_0,\widehat{\pi})-U_{W,n}(\Theta_0,\pi)\Big)\Big]+\mathbf{O}_{\mathbb{P}}(\eta_n).\end{aligned}$$

Nous avons

$$\begin{aligned}\mathbf{var}\Big[U_{Ws,n}(\Theta,\widehat{\pi})] &= \mathbf{var}\Big[U_{W,n}(\Theta,\pi)+(U_{Ws,n}(\Theta,\widehat{\pi})-U_{W,n}(\Theta,\pi))\Big]\\ &= \mathbf{var}\Big[U_{W,n}(\Theta,\pi)\Big]+\mathbf{var}\Big[U_{Ws,n}(\Theta,\widehat{\pi})-U_{W,n}(\Theta,\pi)\Big]+\\ &\quad 2\mathbf{cov}\Big[U_{W,n}(\Theta,\pi),(U_{Ws,n}(\Theta,\widehat{\pi})-U_{W,n}(\Theta,\pi))\Big].\end{aligned}$$

D'autre part, par les hypothèses **(B4)**, **(B6)**, **(B8)** et **(B9)**, nous avons

$$\mathbf{var}\Big[U_{W,n}(\Theta,\pi)\Big]=\mathbb{E}\left[\frac{1}{\pi(\mathcal{O}_1^b,\mathcal{O}_1^c)}\big[\Phi_1(\Theta)\big]^{\otimes 2}\right].$$

Ensuite, on montre

$$\begin{aligned}&\mathbf{cov}\Big[U_{W,n}(\Theta,\pi),\big(U_{Ws,n}(\Theta,\widehat{\pi})-U_{W,n}(\Theta,\pi)\big)\Big]\\ =\ &-\mathbb{E}\left[\frac{\delta_1\big(\delta_1-\pi(\mathcal{O}_1^b,\mathcal{O}_1^c)\big)}{\pi^2(\mathcal{O}_1^b,\mathcal{O}_1^c)}\Big[\Phi_1(\Theta)\Big]^{\otimes 2}\right]+\mathbf{O}(\eta_n),\end{aligned}$$

$$
\begin{aligned}
&= -\mathbb{E}\left(\mathbb{E}\left[\frac{\delta_1\big(\delta_1-\pi(\mathcal{O}_1^b,\mathcal{O}_1^c)\big)}{\pi^2(\mathcal{O}_1^b,\mathcal{O}_1^c)}\Big[\Phi_1^*(\Theta)\Big]^{\otimes 2}\Big|\mathcal{O}_1^b,\mathcal{O}_1^c\right]\right)+\mathbf{O}(\eta_n),\\
&= -\mathbb{E}\left[\frac{1-\pi(\mathcal{O}_1^b,\mathcal{O}_1^c)}{\pi(\mathcal{O}_1^b,\mathcal{O}_1^c)}\Big[\Phi_1^*(\Theta)\Big]^{\otimes 2}\right]+\mathbf{O}(\eta_n),
\end{aligned}
$$

Enfin, on a

$$
\begin{aligned}
\mathbf{var}\Big[U_{Ws,n}(\Theta,\widehat{\pi})-U_{W,n}(\Theta,\pi)\Big] &= \mathbb{E}\left[\frac{\Big(\delta_1-\pi(\mathcal{O}_1^b,\mathcal{O}_1^c)\Big)^2}{\pi^2(\mathcal{O}_1^b,\mathcal{O}_1^c)}\Big[\Phi_1(\Theta)\Big]^{\otimes 2}\right]+\mathbf{O}(\eta_n)\\
&= \mathbb{E}\left(\mathbb{E}\left[\frac{\Big(\delta_1-\pi(\mathcal{O}_1^b,\mathcal{O}_1^c)\Big)^2}{\pi^2(\mathcal{O}_1^b,\mathcal{O}_1^c)}\Big[\Phi_1^*(\Theta)\Big]^{\otimes 2}\Big|\mathcal{O}_1^b,\mathcal{O}_1^c\right]\right)+\mathbf{O}(\eta_n)\\
&= \mathbb{E}\left[\frac{1-\pi(\mathcal{O}_1^b,\mathcal{O}_1^c)}{\pi(\mathcal{O}_1^b,\mathcal{O}_1^c)}\Big[\Phi_1^*(\Theta)\Big]^{\otimes 2}\right]+\mathbf{O}(\eta_n).
\end{aligned}
$$

Il s'ensuit que

$$
\begin{aligned}
\mathbf{var}\Big[U_{Ws,n}(\Theta,\widehat{\pi})\Big] &= \mathbb{E}\left[\frac{\big[\Phi_1(\Theta)\big]^{\otimes 2}}{\pi(\mathcal{O}_1^b,\mathcal{O}_1^c)}\right]-\mathbb{E}\left[\frac{\big[\Phi_1^*(\Theta)\big]^{\otimes 2}}{\pi(\mathcal{O}_1^b,\mathcal{O}_1^c)}\right]+\mathbb{E}\Big[\big[\Phi_1^*(\Theta)\big]^{\otimes 2}\Big]+\mathbf{O}(\eta_n)\\
&= \Sigma_2(\Theta_0,\pi)-\big[\Sigma_2^*(\Theta_0,\pi)-\Sigma_0^*(\Theta_0,\pi)\big]+\mathbf{O}(\eta_n).
\end{aligned}
$$

Par conséquent, en utilisant le Théorème Central Limite, nous avons $U_{Ws,n}(\Theta,\widehat{\pi})\xrightarrow{d}\mathcal{N}\Big(\mathbf{0},\Sigma_2(\Theta_0,\pi)-\big[\Sigma_2^*(\Theta_0,\pi)-\Sigma_0^*(\Theta_0,\pi)\big]\Big)$. De plus, comme $\frac{1}{\sqrt{n}}\mathcal{H}_{Ws,n}(\Theta_0,\widehat{\pi})$ converge en probabilité vers $\Sigma_1(\Theta_0)$. Ainsi, en utilisant le théorème de Slutsky, on a $\sqrt{n}(\widehat{\Theta}_{Ws}-\Theta_0)$ converge en distribution vers une loi normale multivariée centrée de matrice de variance-covariance Σ_{Ws}, où

$$
\Sigma_{Ws}=\Big[\Sigma_1(\Theta_0)^{-1}\Big]\Big(\Sigma_2(\Theta_0,\pi)-\big[\Sigma_2^*(\Theta_0,\pi)-\Sigma_0^*(\Theta_0,\pi)\big]\Big)\Big[\Sigma_0^{-1}(\Theta_0)\Big]^{\top}.
$$

Ce qui achève la preuve du Théorème 3.2.

3.4 Méthodes d'imputation multiple non-paramétrique

L'imputation multiple (MI) est une méthode simple mais puissante pour traiter les données manquantes. L'objectif de l'imputation multiple est de générer des valeurs possibles pour les valeurs manquantes et de créer ainsi plusieurs ensembles de données " complets ". Dans ce qui suit, nous proposons deux méthodes d'imputation multiple non paramétrique pour estimer les paramètres du modèle de régression ZIBP lorsque certaines covariables sont partiellement observées. Les méthodes MI

présentées s'inspirent des travaux de Wang et Chen (2009).
Soit la distribution conditionnelle empirique

$$\widehat{F}(z|\mathcal{O}_i) = \frac{\sum_{k=1}^{n} \delta_k K_h(\mathcal{O}_k^b = \mathcal{O}_i^b, \mathcal{O}_k^c - \mathcal{O}_i^c)}{\sum_{\ell=1}^{n} \delta_\ell K_h(\mathcal{O}_\ell^b = \mathcal{O}_i^b, \mathcal{O}_\ell^c - \mathcal{O}_i^c)} I(\mathcal{Z}_k \leq z), \tag{3.8}$$

où $I(.)$ désigne la fonction indicatrice qui est définie comme suit: $I(\mathcal{Z}_k \leq z) = 1$ si toutes les composantes de $\mathcal{Z}_k$ sont inférieures ou égales aux composantes correspondantes de z et $I(\mathcal{Z}_k \leq z) = 0$ sinon.

3.4.1 Methode 1

La première méthode MI est une méthode d'estimation du type de Rubin (2004). On la note MI1. Considérons $\mathcal{O} = (Y_1, Y_2, \mathbf{X}^{(obs),\top}, \mathbf{W}^{(obs),\top})^\top$, le vecteur des variables qui sont toujours observées sur chaque individu. On suppose que $\mathcal{O}$ est un vecteur contenant un mélange de variables discrètes, catégorielles et continues. Comme pour la section précédente, on pose $\mathcal{O} = \left(\mathcal{O}^b, \mathcal{O}^c\right)$, où $\mathcal{O}^c \in \mathbb{R}^d$ est un vecteur de variables aléatoires continues de dimension d et $\mathcal{O}^b$ est un vecteur de variables aléatoires discrètes et catégorielles de dimension b. Nous rappelons que les variables qui contiennent au moins une données manquantes $\mathcal{Z} := \left\{\mathbf{X}^{(mis)}, \mathbf{W}^{(mis)}\right\}$ sont MAR. Étant donné un nombre M de replications, la procédure d'imputation de la méthode MI1 peut être résumée comme suit :
`Étape 1` : Lorsque $\delta_i = 0$, nous proposons d'imputer un $\mathcal{Z}_i$ avec un $\widetilde{\mathcal{Z}}_{i,v}$, qui est généré aléatoirement à partir de la distribution conditionnelle estimée $\widehat{F}(z|\mathcal{O}_i)$, pour $i = 1, \ldots, n$, $v = 1 \ldots, M$. Ainsi, nous estimons la valeur inconnue de la covariable $\mathcal{Z}_i$ par

$$\widetilde{\mathcal{Z}}_{i,v} = \frac{\sum\limits_{k=1}^{n} \delta_k \mathcal{Z}_k K_h(\mathcal{O}_k^b = \mathcal{O}_i^b, \mathcal{O}_k^c - \mathcal{O}_i^c)}{\sum\limits_{\ell=1}^{n} \delta_\ell K_h(\mathcal{O}_\ell^b = \mathcal{O}_i^b, \mathcal{O}_\ell^c - \mathcal{O}_i^c)}, \quad v = 1 \ldots, M.$$

`Étape 2` : Sur la base des M ensembles de données imputées, le paramètre du modèle sous-jacent est ensuite estimé séparément en utilisant chaque ensemble de données imputées.
Soit $\widehat{\Theta}_v$ la solution des équations d'estimation

$$\widetilde{U}_v(\Theta) = \sum_{i=1}^{n} \widetilde{U}_{i,v}(\Theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big[\delta_i \Phi_i(\Theta) + (1 - \delta_i)\widetilde{\Phi}_{i,v}(\Theta)\Big] = 0, \quad v = 1, \ldots, M, \tag{3.9}$$

où $\widetilde{\Phi}_{i,v}(\Theta)$ est le score obtenu à partir du v-ième ensemble de données imputées pour $i = 1, \ldots, n$ et $v = 1, \ldots, M$.

`Étape 3` : l'estimation par imputation multiple de Θ est donnée par

$$\widehat{\Theta}_{mi1}^{(M)} = \frac{1}{M}\sum_{v=1}^{M}\widehat{\Theta}_v.$$

Selon Rubin (1976), $\widehat{\text{var}}(\widehat{\Theta}_{mi1}^{(M)})$ est un estimateur de $\text{var}(\widehat{\Theta}_{mi1}^{(M)})$, où

$$\begin{aligned}\widehat{\text{var}}(\widehat{\Theta}_{mi1}^{(M)}) &= \frac{1}{M}\sum_{v=1}^{M}\left\{\left(-\frac{\partial\widetilde{U}_v(\widehat{\Theta}_v)}{\partial\Theta^\top}\right)\mathcal{H}_v(\widehat{\Theta}_v)\left(-\frac{\partial\widetilde{U}_v(\widehat{\Theta}_v)}{\partial\Theta^\top}\right)^\top\right\} \\ &\quad +\left(1+\frac{1}{M}\right)\frac{\sum_{v=1}^{M}\left(\widehat{\Theta}_v-\hat{\Theta}_{mi1}\right)^{\otimes 2}}{M-1},\end{aligned} \tag{3.10}$$

avec

$$\mathcal{H}_v(\Theta) = \sum_{i=1}^{n}\left[\widetilde{U}_{i,v}(\Theta)\right]\left[\widetilde{U}_{i,v}(\Theta)\right]^\top.$$

3.4.2 Méthode 2

La deuxième méthode MI est inspiré de Fay (1996), nous la notons MI2. La procédure de cette méthode d'imputation multiple se déroule aussi en trois étapes. Pour l'`Étape 2` de cette procédure MI, nous définissons $\overline{\Phi}_i(\Theta) = \frac{1}{M}\sum_{v=1}^{M}\widetilde{\Phi}_{i,v}(\Theta)$ et nous utilisons l'expression $\widetilde{U}_v(\Theta)$ dans l'équation (3.9) pour proposer la fonction d'estimation suivante :

$$U_{mi2}(\Theta) = \frac{1}{M}\sum_{v=1}^{M}\widetilde{U}_v(\Theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\delta_i\Phi_i(\Theta)+(1-\delta_i)\overline{\Phi}_i(\Theta)\right]. \tag{3.11}$$

Soit $\widehat{\Theta}_{mi2}^{(M)}$ solution de l'équation (3.11). $\widehat{\Theta}_{mi2}^{(M)}$ est l'estimateur de Θ obtenu par la deuxième méthode MI (MI2). Selon Rubin (2004), un estimateur $\widehat{\text{var}}\left[U_{mi2}(\Theta)\right]$ est donné par :

$$\begin{aligned}\widehat{\text{var}}\left[U_{mi2}^{(M)}(\Theta)\right] &= \frac{1}{M}\sum_{v=1}^{M}\widehat{\text{var}}\left(\widetilde{U}_v(\Theta)\right)+\left(1+\frac{1}{M}\right)\frac{\sum_{v=1}^{M}\left(\widetilde{U}_v(\Theta)-U_{mi2}(\Theta)\right)^{\otimes 2}}{M-1}, \\ &= \frac{1}{M}\sum_{v=1}^{M}\sum_{i=1}^{n}\left[\widetilde{U}_{i,v}(\Theta)\right]^{\otimes 2}+\left(1+\frac{1}{M}\right)\frac{\sum_{v=1}^{M}\left(\widetilde{U}_v(\Theta)-U_{mi2}(\Theta)\right)^{\otimes 2}}{M-1}, \\ &= \frac{1}{M}\sum_{v=1}^{M}\sum_{i=1}^{n}\left[\widetilde{U}_{i,v}(\Theta)\right]^{\otimes 2}+\left(1+\frac{1}{M}\right)\frac{\sum_{v=1}^{M}\left(\widetilde{U}_v(\Theta)\right)^{\otimes 2}}{M-1}.\end{aligned}$$

Le passage de la première ligne à la deuxième est justifié par le fait que les $\widetilde{U}_{i,v}(\Theta)$ sont des vecteurs aléatoires indépendants et identiquement distribués et on utilise $U_{mi2}(\widehat{\Theta}_{mi2}^{(M)}) = \mathbf{0}$ pour la deuxième à la troisième.

Soit $\mathcal{H}_{mi2}(\Theta)$ le gradient de $-U_{mi2}(\Theta)$. Par un développement de Taylor, on obtient

$$\mathbf{0} = U_{mi2}(\widehat{\Theta}_{mi2}^{(M)}) = U_{mi2}(\Theta) - \mathcal{H}_{mi2}(\Theta)(\widehat{\Theta}_{mi2}^{(M)} - \Theta) + \mathbf{o}_{\mathbb{P}}(1).$$

D'où, $\text{var}(\widehat{\Theta}_{mi2})$ peut être estimé par

$$\begin{aligned}
\widehat{\text{var}}(\widehat{\Theta}_{mi2}^{(M)}) &= \Big[\mathcal{H}_{mi2}^{-1}(\Theta)\Big]\widehat{\text{var}}\Big[U_{mi2}^{(M)}(\Theta)\Big]\Big[\mathcal{H}_{mi2}^{-1}(\Theta)\Big]^{\top} && (3.12)\\
&= \Big[\mathcal{H}_{mi2}^{-1}(\Theta)\Big]\Bigg(\frac{1}{M}\sum_{v=1}^{M}\sum_{i=1}^{n}\Big[\widetilde{U}_{i,v}(\Theta)\Big]^{\otimes 2}\\
&\quad + \Big(1+\frac{1}{M}\Big)\frac{\sum_{v=1}^{M}\Big(\widetilde{U}_{v}(\Theta)\Big)^{\otimes 2}}{M-1}\Bigg)\Big[\mathcal{H}_{mi2}^{-1}(\Theta)\Big]^{\top}. && (3.13)
\end{aligned}$$

Les expressions des estimateurs de $\text{var}(\widehat{\Theta}_{mi1}^{(M)})$ et $\text{var}(\widehat{\Theta}_{mi1}^{(M)})$ dans 3.10 et 3.12 ont été proposées par Lee *et al.* (2016). Motivés par le fait que les estimateurs de variances des estimateurs MI $\text{var}(\widehat{\Theta}_{mi1}^{(M)})$ et $\text{var}(\widehat{\Theta}_{mi1}^{(M)})$ basés sur l'idée d'estimation de la variance proposée par Rubin (2004) fonctionnent mal dans le cas du modèle de régression ZIP avec covariables MAR, nous proposons une formule modifiée pour estimer les variances des estimateurs MI basés sur la variance asymptotique de l'estimateur IPW semi-paramétrique proposé par Lukusa *et al.* (2016).

Théorème 3.3 *Supposons les hypothèses* (***B4***), (***B6***) *à* (***B10***) *sont vérifiées. Alors* $\sqrt{n}(\widehat{\Theta}_{mi1}^{(M)} - \Theta_0)$ *et* $\sqrt{n}(\widehat{\Theta}_{mi2}^{(M)} - \Theta_0)$ *sont distribués asymptotiquement suivant une normale multivariée centrée et de matrice de covariance* Σ^**, où*

$$\Sigma^* = \Big[\Sigma_1^{-1}(\Theta,\pi)\Big]\Delta(\Theta,\pi)\Big[\Sigma_1^{-1}(\Theta,\pi)\Big]^{\top},$$

avec

$$\Sigma_1^{-1}(\Theta,\pi) = \mathbb{E}\Bigg[-\frac{1}{\sqrt{n}}\frac{\partial U_{Ws,s}(\Theta,\pi)}{\partial\Theta^{\top}}\Bigg]$$

et

$$\Delta(\Theta,\pi) = \mathbb{E}\Bigg[\bigg(\frac{\delta_1}{\pi(\mathcal{O}_1)}\Phi_1(\Theta) + \Big(1-\frac{\delta_1}{\pi(\mathcal{O}_1)}\Big)\Phi_1^*(\Theta)\bigg)^{\otimes 2}\Bigg].$$

Remarque 3.3 *Un estimateur consistant de la matrice de variance-covariance* Σ^* *est donné par :*

$$\frac{1}{n}\Big[\Sigma_{1,n}(\Theta,\hat{\pi})^{-1}\Big]\widehat{\Delta}(\Theta,\hat{\pi})\Big[\Sigma_{1,n}(\Theta,\hat{\pi})^{-1}\Big]^{\top},$$

où

$$\begin{aligned}\widehat{\Delta}(\Theta,\hat{\pi}) \;:=\; & \frac{1}{n}\sum_{i=1}^{n}\Bigg[\left(\frac{\delta_i}{\hat{\pi}(\mathcal{O}_i)}\Phi_i(\Theta)+\Big[1-\frac{\delta_i}{\hat{\pi}(\mathcal{O}_i)}\Big]\widehat{\Phi}_i^*(\Theta)\right)^{\otimes 2}\\ & +\frac{1}{M^2}(1-\delta_i)\big[\tilde{\Phi}_{i,v}(\Theta)-\widehat{\Phi}_i^*(\Theta)\big]^{\otimes 2}\Bigg]\end{aligned}$$

et

$$\hat{\Phi}_i^*(\Theta)=\frac{\sum_{k=1}^{n}\delta_k\Phi_k(\Theta)K_h(\mathcal{O}_k^b=\mathcal{O}_i^b,\mathcal{O}_k^c-\mathcal{O}_i^c)}{\sum_{j=1}^{n}\delta_j K_h(\mathcal{O}_j^b=\mathcal{O}_i^b,\mathcal{O}_j^c-\mathcal{O}_i^c)}.$$

Preuve du Théorème 3.3

La preuve du Théorème 3.3 est la conséquence des deux lemmes suivants que nous énonçons et établissons dans la suite de la section.

Lemme 3.4 *Supposons les hypothèses* (***B4***)*,* (***B6***) *à* (***B10***) *sont vérifiées. Alors* $U_{mi2}(\Theta)$ *converge en distribution vers une normale multivariée centrée et de matrice de variance-covariance* $\Delta(\Theta,\pi)$ *lorsque* n*,* $M\to\infty$*.*

Rappelons que $\Phi_i^*(\Theta)=\mathbb{E}\big[\Phi_i(\Theta)|\mathcal{O}_i^b,\mathcal{O}_i^c\big]$ et $\overline{\Phi}_i(\Theta)=\frac{1}{M}\sum_{v=1}^{M}\widetilde{\Phi}_{i,v}(\Theta)$, $i=1,...,n$.
On a

$$\begin{aligned}U_{mi2}(\Theta) \;=\; & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\delta_i\Phi_i(\Theta)+(1-\delta_i)\Phi_i^*(\Theta)\Big)\\ & +\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)\Big(\overline{\Phi}_i(\Theta)-\mathbb{E}_{\widehat{F}}(\widetilde{\Phi}_{i,1}(\Theta)\Big)\\ & +\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)\Big(\mathbb{E}_{\widehat{F}}\big(\widetilde{\Phi}_{i,1}(\Theta)\big)-\Phi_i^*(\Theta)\Big).\end{aligned}\tag{3.14}$$

D'abord, on peut remarquer que le second terme de la décomposition de $U_{mi2}(\Theta)$ dans (3.14) nous permet d'écrire :

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)\Big(\overline{\Phi}_i(\Theta)-\mathbb{E}_{\widehat{F}}(\widetilde{\Phi}_{i,1}(\Theta)\Big)=\mathbf{O}_{\mathbb{P}}(1).$$

En effet, étant donné $\mathcal{O}_i$, $\widetilde{\Phi}_{i,v}(\Theta)$, $v = 1, \ldots, M$ sont indépendants et identiquement distribuées selon $\widehat{F}(z|\mathcal{O}_i)$. Ainsi, on a

$$\overline{\Phi}_i(\Theta) - \mathbb{E}_{\widehat{F}}(\widetilde{\Phi}_{i,v}(\Theta) = \sum_{v=1}^{M}\Big[\overline{\Phi}_i(\Theta) - \mathbb{E}_{\widehat{F}}\Big(\widetilde{\Phi}_{i,1}(\Theta)\big|\mathcal{O}_i\Big)\Big] = \mathbf{O}_{\mathbb{P}}(M^{-1/2}).$$

Il s'ensuit que

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)\Big(\overline{\Phi}_i(\Theta) - \mathbb{E}_{\widehat{F}}(\widetilde{\Phi}_{i,1}(\Theta)\Big) = \mathbf{O}_{\mathbb{P}}(1).$$

Ensuite, le troisième terme de la décomposition de $U_{mi2}(\Theta)$ dans (3.14) peut s'écrire comme suit :

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)\Big(\mathbb{E}_{\widehat{F}}\big(\widetilde{\Phi}_{i,1}(\Theta)\big) - \Phi_i^*(\Theta)\Big)\\
=\;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)\left\{\sum_{k=1}^{n}\frac{\delta_k\Phi_k(\Theta)K_h(\mathcal{O}_k^b=\mathcal{O}_i^b,\mathcal{O}_k^c-\mathcal{O}_i^c)}{\sum_{j=1}^{n}\delta_jK_h(\mathcal{O}_j^b=\mathcal{O}_i^b,\mathcal{O}_j^c-\mathcal{O}_i^c)} - \Phi_i^*\sum_{k=1}^{n}\frac{\delta_kK_h(\mathcal{O}_k^b=\mathcal{O}_i^b,\mathcal{O}_k^c-\mathcal{O}_i^c)}{\sum_{j=1}^{n}\delta_jK_h(\mathcal{O}_j^b=\mathcal{O}_i^b,\mathcal{O}_j^c-\mathcal{O}_i^c)}\right\},\\
=\;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)\left\{\sum_{k=1}^{n}\frac{\delta_k(\Phi_k(\Theta)-\Phi_k^*(\Theta))K_h(\mathcal{O}_k^b=\mathcal{O}_i^b,\mathcal{O}_k^c-\mathcal{O}_i^c)}{\sum_{j=1}^{n}\delta_jK_h(\mathcal{O}_j^b=\mathcal{O}_i^b,\mathcal{O}_j^c-\mathcal{O}_i^c)}\right\},\\
=\;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)\left\{\sum_{k=1}^{n}\frac{\delta_k(\Phi_k(\Theta)-\Phi_k^*(\Theta))K_h(\mathcal{O}_k^b=\mathcal{O}_i^b,\mathcal{O}_k^c-\mathcal{O}_i^c)}{\sum_{\ell=1}^{n}K_h(\mathcal{O}_\ell^b=\mathcal{O}_i^b,\mathcal{O}_\ell^c-\mathcal{O}_i^c)}\right\}\\
&\times\left\{\frac{\sum_{\ell=1}^{n}K_h(\mathcal{O}_\ell^b=\mathcal{O}_i^b,\mathcal{O}_\ell^c-\mathcal{O}_i^c)}{\sum_{j=1}^{n}K_h(\mathcal{O}_j^b=\mathcal{O}_i^b,\mathcal{O}_j^c-\mathcal{O}_i^c)}\right\},\\
=\;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1-\delta_i}{\hat{\pi}(\mathcal{O}_i)}\left\{\sum_{k=1}^{n}\frac{\delta_k\big(\Phi_k(\Theta)-\Phi_k^*(\Theta)\big)K_h(\mathcal{O}_k^b=\mathcal{O}_i^b,\mathcal{O}_k^c-\mathcal{O}_i^c)}{\sum_{\ell=1}^{n}K_h(\mathcal{O}_\ell^b=\mathcal{O}_i^b,\mathcal{O}_\ell^c-\mathcal{O}_i^c)}\right\},\\
=\;& \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\delta_k\big(\Phi_k(\Theta)-\Phi_k^*(\Theta)\big)\left\{\sum_{i=1}^{n}\frac{1-\delta_i}{\hat{\pi}(\mathcal{O}_i)}\left[\frac{K_h(\mathcal{O}_k^b=\mathcal{O}_i^b,\mathcal{O}_k^c-\mathcal{O}_i^c)}{\sum_{\ell=1}^{n}K_h(\mathcal{O}_\ell^b=\mathcal{O}_i^b,\mathcal{O}_\ell^c-\mathcal{O}_i^c)}\right]\right\},\\
=\;& \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\delta_k\big(\Phi_k(\Theta)-\Phi_k^*(\Theta)\big)\left\{\frac{1-\hat{\pi}(\mathcal{O}_k)}{\hat{\pi}(\mathcal{O}_k)}\right\},\\
=\;& \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\delta_k\left(\frac{1-\pi(\mathcal{O}_k)}{\pi(\mathcal{O}_k)}\right)\Big(\Phi_k(\Theta)-\Phi_k^*(\Theta)\Big) + \mathbf{o}_{\mathbb{P}}(1). \qquad (3.15)
\end{aligned}$$

En utilisant (3.14) et (3.15), on peut écrire

$$U_{mi2}(\Theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\delta_i}{\pi(\mathcal{O}_i)}\Phi_i(\Theta) + \left(1 - \frac{\delta_i}{\pi(\mathcal{O}_i)}\right)\Phi_i^*(\Theta)\right] + \mathbf{O}_{\mathbb{P}}(M^{-1/2}) + \mathbf{o}_{\mathbb{P}}(1). \quad (3.16)$$

Enfin, comme le premier terme dans (3.16) est la somme de vecteurs aléatoires indépendants et identiquement distribués, on utilise le Théorème Central Limite multivarié pour montrer que $U_{mi2}(\Theta)$ converge en distribution vers une distribution normale centrée et de matrice de variance-covariance $\Delta(\Theta, \pi)$ lorsque n, $M \to \infty$.
Ce qui achève la preuve du Lemme 3.4.

Lemme 3.5 *Supposons que les hypothèses* (***B4***) *et* (***B6***) *sont vérifiées. Alors on a*

$$\mathbb{E}_{\widehat{F}}\Big[\widetilde{U}_v(\Theta)|\mathcal{O}\Big] = U_{Ws}(\Theta, \widehat{\pi})$$

et

$$\mathbb{E}_{\widehat{F}}\left[\frac{\partial\widetilde{U}_v(\Theta)}{\partial\Theta^\top}\middle|\mathcal{O}\right] = \frac{\partial U_{Ws}(\Theta, \widehat{\pi})}{\partial\Theta^\top}.$$

Preuve du Lemme 3.5

Notons que $\Phi_i(\Theta)$ est l'expression du score lorsque toutes les covariables de l'individu i sont observées et $\widetilde{\Phi}_{i,v}(\Theta)$ celle lorsqu'au moins l'une des valeurs des covariables est générée. On a

$$\mathbb{E}_{\widehat{F}}\Big[\big(\Phi_{i,v}(\Theta)|\mathcal{O}_i\big)\Big] = \sum_{k=1}^{n}\frac{\delta_k K(\mathcal{O}_k^b = \mathcal{O}_i^b, \mathcal{O}_k^c - \mathcal{O}_i^c)\Phi_k(\Theta)}{\sum_{j=1}^{n}\delta_j K(\mathcal{O}_j^b = \mathcal{O}_i^b, \mathcal{O}_j^c - \mathcal{O}_i^c)}, \quad v = 1, 2, \ldots, M.$$

Par conséquent, on a

$$\begin{aligned}
\mathbb{E}_{\widehat{F}}\big(\widetilde{U}_v(\Theta)|\mathcal{O}\big) &= \sum_{i=1}^{n}\Big[\delta_i\mathbb{E}_{\widehat{F}}\big(\Phi_i(\Theta)|\mathcal{O}_i\big)\Big) + (1-\delta_i)\mathbb{E}_{\widehat{F}}\big(\Phi_{i,v}(\Theta)|\mathcal{O}_i\big)\Big], \\
&= \sum_{i=1}^{n}\delta_i\mathbb{E}_{\widehat{F}}\Big[\Phi_i(\Theta)|\mathcal{O}_i\Big] + (1-\delta_i)\sum_{k=1}^{n}\frac{\delta_k K(\mathcal{O}_k^b = \mathcal{O}_i^b, \mathcal{O}_k^c - \mathcal{O}_i^c)\Phi_k(\Theta)}{\sum_{j=1}^{n}\delta_j K(\mathcal{O}_j^b = \mathcal{O}_i^b, \mathcal{O}_j^c - \mathcal{O}_i^c)}, \\
&= \sum_{i=1}^{n}\delta_i\Phi_i(\Theta) + \sum_{k=1}^{n}\delta_k\Phi_k(\Theta)\Big(\frac{1}{\widehat{\pi}(\mathcal{O}_k^b, \mathcal{O}_k^c)} - 1\Big), \\
&= \sum_{i=1}^{n}\frac{\delta_i}{\widehat{\pi}(\mathcal{O}_i^b, \mathcal{O}_i^c)}\Phi_i(\Theta), \\
&= U_{Ws,n}(\Theta, \widehat{\pi}).
\end{aligned}$$

En utilisant la même demarche, nous montrons que $\mathbb{E}_{\widehat{F}}\left[\frac{\partial\widetilde{U}_v(\Theta)}{\partial\Theta^\top}\middle|\mathcal{O}\right] = \frac{\partial U_{Ws,n}(\Theta,\widehat{\pi})}{\partial\Theta^\top}$.

Par conséquent, on a

$$\mathbb{E}\Big[\frac{\partial \widetilde{U}_v(\Theta)}{\partial \Theta^\top}\Big] = \mathbb{E}\Big[\frac{\partial U_{Ws,n}(\Theta, \widehat{\pi})}{\partial \Theta^\top}\Big]. \tag{3.17}$$

Donc par un développement de Taylor de $\widetilde{U}_v(\widehat{\Theta}_v)$ en Θ, on a

$$\mathbf{0} = \widetilde{U}_v(\widehat{\Theta}_v) \;\; = \;\; \widetilde{U}_v(\Theta) + \Big[\frac{\partial \widetilde{U}_v(\Theta)}{\partial \Theta^\top}\Big]\sqrt{n}(\widehat{\Theta}_v - \Theta).$$

Comme $U_{mi2}(\Theta) = \frac{1}{M}\sum_{v=1}^{M}\widetilde{U}_v(\Theta)$ et en utilisant (3.17), on obtient

$$\mathbf{0} = U_{mi2}(\widehat{\Theta}_{mi2}^{(M)}) = U_{mi2}(\Theta) - \Sigma_1(\Theta, \pi)\sqrt{n}(\widehat{\Theta}_{mi2}^{(M)} - \Theta) + \mathbf{O}_{\mathbb{P}}(M^{-1/2}) + \mathbf{O}_{\mathbb{P}}(1).$$

Par conséquent

$$\sqrt{n}(\widehat{\Theta}_{mi2}^{(M)} - \Theta) = \Sigma_1^{-1}(\Theta, \pi)U_{mi2}(\Theta) + \mathbf{O}_{\mathbb{P}}(M^{-1/2}) + \mathbf{O}_{\mathbb{P}}(1).$$

Finalement, $\sqrt{n}(\widehat{\Theta}_{mi2}^{(M)} - \Theta_0)$ est distribué asymptotiquement comme une normale multivariée centrée de matrice de covariance Σ^*.

Théorème 3.4 *Supposons que les hypothèses* (***B1***), (***B4***), (***B6***) *à* (***B10***) *sont vérifiées.*
Alors $\sqrt{n}(\widehat{\Theta}_{mi1}^{(M)} - \Theta_{mi2}^{(M)})$ *et* $\sqrt{n}(\widehat{\Theta}_{mi2}^{(M)} - \Theta_{Ws})$ *convergent en probabilité vers* ***0*** *lorsque* $M,\ n \to \infty$.

Nous montrons d'abord l'équivalence asymptotique entre les estimateurs de MI1 et MI2. Par un développement de Taylor de $\widetilde{U}_v(\widehat{\Theta}_v)$ en Θ et par des calculs, nous obtenons

$$\sqrt{n}(\widehat{\Theta}_v - \Theta) = \Sigma_1^{-1}(\Theta, \pi)n^{-1/2}\widetilde{U}_v(\Theta) + \mathbf{O}_{\mathbb{P}}(1).$$

De plus

$$\widehat{\Theta}_{mi1} = \frac{1}{M}\sum_{v=1}^{M}\widehat{\Theta}_v.$$

Ce qui entraîne que

$$\sqrt{n}(\widehat{\Theta}_{mi1} - \Theta) = \Big[\Sigma_1(\Theta, \pi)\Big]^{-1} n^{-1/2}\Big(\frac{1}{M}\sum_{v=1}^{M}\widetilde{U}_v(\Theta)\Big) + \mathbf{O}_{\mathbb{P}}(1). \tag{3.18}$$

En procédant de manière similaire et en utilisant le fait que $\overline{\Phi}_i = \frac{1}{M}\sum_{v=1}^{M}\widetilde{\Phi}_{i,v}$, nous

montrons que

$$\sqrt{n}(\widehat{\Theta}_{mi2}^{(M)} - \Theta_0) = \Big[\Sigma_1(\Theta,\pi)\Big]^{-1} n^{-1/2}\Big(\frac{1}{M}\sum_{v=1}^{M}\widetilde{U}_v(\Theta)\Big) + \mathbf{O}_{\mathbb{P}}(1). \tag{3.19}$$

À partir des expressions (3.18) et (3.19), nous avons $\sqrt{n}(\widehat{\Theta}_{mi2} - \widehat{\Theta}_{mi1}) = \mathbf{O}_{\mathbb{P}}(1)$. Donc $\sqrt{n}(\widehat{\Theta}_{mi2} - \widehat{\Theta}_{mi1})$ converge en probabilité vers $\mathbf{0}$.

Montrons maintenant que les estimateurs IPW semi-paramétriques et MI2 sont asymptotiquement équivalents.
Nous avons

$$\begin{aligned} U_{mi2}(\Theta) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\delta_i\Phi_i(\Theta) + (1-\delta_i)\mathbb{E}_{\widehat{F}}(\widetilde{\Phi}_{i,1}(\Theta)|\mathcal{O}_i^b,\mathcal{O}_i^c)\Big] \\ &\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[(1-\delta_i)\Big(\overline{\Phi}_i(\Theta) - \mathbb{E}_{\widehat{F}}(\widetilde{\Phi}_{i,1}(\Theta)|\mathcal{O}_i^b,\mathcal{O}_i^c)\Big)\Big], \end{aligned}$$

$$\begin{aligned} U_{mi2}(\Theta) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_i\Phi_i(\Theta) + \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\delta_k\Phi_k\Big[\frac{1}{\pi(\mathcal{O}_k^b,\mathcal{O}_k^c)} - 1\Big] \\ &\quad + (1-\delta_i)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\overline{\Phi}_i(\Theta) - \mathbb{E}_{\widehat{F}}(\widetilde{\Phi}_{i,1}(\Theta)|\mathcal{O}_i^b,\mathcal{O}_i^c)\Big)\Big], \\ &= U_{Ws,n}(\Theta,\widehat{\pi}) + (1-\delta_i)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\overline{\Phi}_i(\Theta) - \mathbb{E}_{\widehat{F}}\Big(\widetilde{\Phi}_{i,1}(\Theta)|\mathcal{O}_i^b,\mathcal{O}_i^c\Big)\Big] \end{aligned}$$

En outre, on a

$$\begin{aligned} &\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)\Big[\overline{\Phi}_i(\Theta) - \mathbb{E}_{\widehat{F}}\big(\widetilde{\Phi}_{i,1}(\Theta)|\mathcal{O}_i^b,\mathcal{O}_i^c\big)\Big)\Big] = \\ &\quad \frac{1}{\sqrt{M}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\frac{1}{\sqrt{M}}\sum_{v=1}^{M}(1-\delta_i)\Big(\widetilde{\Phi}_{i,v}(\Theta) - \mathbb{E}_{\widehat{F}}\big(\widetilde{\Phi}_{i,1}(\Theta)|\mathcal{O}_i^b,\mathcal{O}_i^c\big)\Big)\Big]. \end{aligned}$$

De plus,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\frac{1}{\sqrt{M}}\sum_{v=1}^{M}(1-\delta_i)\Big(\widetilde{\Phi}_{i,v}(\Theta) - \mathbb{E}_{\widehat{F}}\big(\widetilde{\Phi}_{i,1}(\Theta)|\mathcal{O}_i^b,\mathcal{O}_i^c\big)\Big)\Big] = \mathbf{O}_{\mathbb{P}}(1).$$

Par conséquent,

$$U_{mi2}(\Theta) - U_{Ws,n}(\Theta,\widehat{\pi}) = \mathbf{O}_{\mathbb{P}}(1). \tag{3.20}$$

Ensuite par un développement de Taylor de $U_{mi2}(\Theta_{mi2}^{(M)})$ en Θ et $U_{Ws}(\Theta,\widehat{\pi})$ en Θ

respectivement, on a

$$\mathbf{0} = U_{mi2}(\widehat{\Theta}_{mi2}^{(M)}) = U_{mi2}(\Theta) - \Sigma_1(\Theta,\pi)\sqrt{n}(\widehat{\Theta}_{mi2}^{(M)} - \Theta) + \mathbf{O}_{\mathbb{P}}(1) \tag{3.21}$$

et

$$\mathbf{0} = U_{Ws}(\widehat{\Theta}_{Ws},\widehat{\pi}) = U_{Ws}(\Theta,\widehat{\pi}) - \Sigma_1(\Theta_0,\pi)\sqrt{n}(\widehat{\Theta}_{Ws} - \Theta) + \mathbf{O}_{\mathbb{P}}(1). \tag{3.22}$$

En utilisant (3.20)-(3.22), on obtient

$$\begin{aligned}\sqrt{n}(\widehat{\Theta}_{mi2}^{(M)} - \widehat{\Theta}_{Ws}) &= \Sigma_1(\Theta,\pi)\sqrt{n}\Big[U_{mi2}(\Theta) - U_{Ws}(\Theta,\widehat{\pi})\Big] + \mathbf{O}_{\mathbb{P}}(1),\\ &= \mathbf{O}_{\mathbb{P}}(M^{-1/2}) + \mathbf{O}_{\mathbb{P}}(1).\end{aligned}$$

Par conséquent, $\sqrt{n}(\widehat{\Theta}_{mi2}^{(M)} - \widehat{\Theta}_{Ws})$ converge en probabilité vers $\mathbf{0}$ lorsque $M,\ n \to \infty$. Les méthodes d'imputation multiple MI de type 1 et 2 sont asymptotiquement équivalentes. Elles fournissent toutes deux des estimateurs consistants et asymptotiquement normaux. Toutefois, la deuxième procédure MI2 présente l'avantage de ne résoudre les équations d'estimation qu'une seule fois au lieu de M. En outre, il convient de noter que les approches MI peuvent être aussi utilisées lorsque les structures des covariables manquantes sont plus complexes et non monotones.

3.5 Résultats numériques

Dans cette section, nous présentons une étude de simulations numériques pour évaluer dans différents scénarios le comportement en échantillon fini des estimateurs suivants :

- $\widehat{\Theta}_{F,n}$: l'estimateur du maximum de vraisemblance qui est obtenu lorsqu'il n'y a pas de covariables manquantes. Cet estimateur est utilisé lorsque les "données sont complètes" (en anglais "full data") on le nomme FD-estimateur
- $\widehat{\Theta}_{W,n}$: l'estimateur IPW paramétrique
- $\widehat{\Theta}_{Ws}$: l'estimateur IPW semi-paramétrique
- $\widehat{\Theta}_{mi1}$: l'estimateur de la première méthode d'imputation multiple (MI1).
- $\widehat{\Theta}_{mi2}$: l'estimateur de la deuxième méthode d'imputation multiple (MI2).

3.5.1 Simulation des données

Nous simulons les données selon le modèle ZIBP (3.1)-(3.2)-(3.3) défini par ce qui suit :

$$\begin{cases}\text{logit}(\pi_i) = \gamma^\top \mathbf{W}_i,\\ \log(\lambda_{1i}) = \beta_1^\top \mathbf{X}_i,\ \ \log(\lambda_{2i}) = \beta_2^\top \mathbf{X}_i,\ \ \textbf{et}\ \ \log(\mu) = \alpha,\end{cases}$$

avec $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3})^\top$ et $\mathbf{W}_i = (W_{i1}, W_{i2}, W_{i3}, W_{i4})$ où $X_{i1} = W_{i1} = 1$, et $X_{i2}, \ldots, X_{i3}$, W_{i2} et W_{i4} sont générées indépendamment à partir des distributions de Bernoulli $\mathcal{B}(1, 0.5)$, normale $\mathcal{N}(0.9; 2.3)$, normale $\mathcal{N}(0.3; 1.3)$, et uniforme $\mathcal{U}(1.2, 1.9)$, respectivement. Sans perte de généralité, on prend $X_{i3} = W_{i3}$. Les paramètres de régression β_1, β_2 et α sont choisis comme suit:
$\beta_1 = (-0.3, 0.8, 0.2)^\top$, $\beta_2 = (0.8, 0.7, -0.4)^\top$ et $\alpha = 1.3$.
Nous prenons successivement deux valeurs du paramètre de régression γ :

- $\gamma = (-0.8, -0.5, -1, 0.2)^\top$ pour avoir 30% d'inflation de zéros.
- $\gamma = (-0.4, 0.3, -1, 0.53)^\top$ pour avoir 45% d'inflation de zéros.

En utilisant ces valeurs, les proportions moyennes d'inflation de zéros dans les ensembles de données simulées sont de 30% (respectivement, 45%).

Sous l'hypothèse MAR, la probabilité de sélection $\pi(\mathcal{O}_i)$ est d'abord estimée paramétriquement par $\pi(\omega, \mathcal{O}_i) = \mathbb{P}(\delta_i = 1|\mathcal{O}_i) = \text{logit}^{-1}(\omega^\top \mathcal{O}_i)$ où $\mathcal{O}_i := (1, Y_1, Y_2, X_2, W_4)$. Dans ce cas, le paramètre de régression ω est choisi de tel sorte que des proportions moyennes de données manquantes dans les échantillons simulés soit successivement égales à 25% et 40%. Ensuite, nous avons estimé la probabilité de sélection $\pi(\mathcal{O}_i)$ non paramétriquement par $\widehat{\pi}(o) = \sum_{k=1}^{n} \frac{\delta_k K_h(\mathcal{O}_k^b = o^b, \mathcal{O}_k^c - o^c)}{\sum_{j=1}^{n} K_h(\mathcal{O}_j^b = o^b, \mathcal{O}_j^c - o^c)}$ où K_h est un noyau multiplicatif (les noyaux discrets associés de Dirac pour les variables discrètes et le noyau Gaussien pour la variable continue W_4) et $h = 1.06 sd_{W4} n^{-1/5}$. Puisque $X_3 = W_3$, les covariables sont manquantes à la fois dans la probabilité de mélange logit(ε_i) et dans les expressions des paramètres de Poisson $\log(\lambda_{1i}) = \beta_1^\top \mathbf{X}_i$ et $\log(\lambda_{2i}) = \beta_2^\top \mathbf{X}_i$. Nous considérons les tailles d'échantillon suivantes : $n = 500$ et $= 1000$. Les estimations des méthodes d'imputation multiple MI1 et MI2 sont obtenues avec $M = 30$ imputations (d'après nos expériences numériques, ce nombre d'imputation est suffisant pour garantir la stabilité des estimations). Notons que les simulations sont réalisées à l'aide du logiciel statistique `R` (R Core Team, 2018). Nous utilisons le package `maxLik` (Henningsen et Toomet (2011)) pour résoudre les équations d'estimation (3.7) via un algorithme de Newton-Raphson.

3.5.2 Résultats des simulations

Pour chaque configuration des paramètres du plan de simulations [`taille de l'échantillon` × `proportion d'inflation de zéros` × `proportion de données manquantes`], nous simulons $N = 500$ échantillons. Pour chaque estimateur cité dans la section précédente, nous calculons le biais moyen, l'écart-type empirique (SD), l'erreur quadratique moyenne (RMSE) sur les N échantillons simulés. À des fins de comparaison, nous fournissons également les résultats qui seraient obtenus s'il n'y

avait pas de covariables manquantes. Dans les tableaux 3.1 à 3.5, nous fournissons les résultats des méthodes d'estimation pour chaque configuration des paramètres du plan de simulations avec $n = 500$. Les tableaux 3.6-3.8 fournissent les résultats pour chaque configuration des paramètres du plan de simulations avec $n = 1000$.

D'après les tableaux 3.1 et 3.8, nous observons de manière générale que le biais de nos estimations sont faibles pour une taille d'échantillon relative modérée. Puis, les biais, les SDs et les RMSEs décroissent lorsque la taille de l'échantillon augmente. En outre, comme on pouvait s'y attendre, pour une taille d'échantillon n et une proportion de données manquantes fixes (lorsque des données sont manquantes), on observe que les estimateurs des paramètres $\beta_{1,j}$s, $\beta_{2,k}$s et α_ns (respectivement, γ_is) sont plus performants lorsque la proportion d'inflation de zéros diminue (respectivement, augmente). D'autre part, on peut également observer que pour une taille d'échantillon n et une proportion d'inflation de zéros toutes deux fixées, nous observons que les estimateurs proposés sont plus performants lorsque la proportion de données manquantes diminue. Notons que l'estimateur FD est évidemment plus performant que les estimateurs proposés, mais l'estimateur FD n'est pas approprié lorsque des données manquantes sont présentes. L'estimateur FD est utilisé comme référence pour les comparaisons. Globalement, les résultats numériques obtenus montrent la bonne performance des méthodes étudiées pour l'estimation du modèle de régression ZIBP lorsque des covariables sont manquantes.

Les performances des méthodes MI se rapprochent de celles de la méthode IPW semi-paramétrique en termes de biais, SD et RMSE. De plus, les résultats numériques confirment l'équivalence asymptotique des estimateurs de type MI. Cependant, nous recommandons d'utiliser le deuxième type de MI (MI2) pour réduire la charge de calcul. En effet, l'estimateur MI2 est obtenu à partir d'une seule résolution des équations d'estimation au lieu de M dans la méthode MI de type 1. Ainsi, dans la section suivante, nous considérons comme méthode MI uniquement la méthode MI de type 2. En outre, l'estimateur IPW semi-paramétrique et les estimateurs de type MI 1 et 2 sont asymptotiquement équivalents, ce qui est conforme aux résultats théoriques.

Comme nous pouvons le remarquer, lorsque la taille de l'échantillon est importante, les performances des méthodes d'estimation proposées sont très proches de celles de la méthode d'estimation EMV (FD-estimateur) qui sert de référence pour les comparaisons.

Spécifiquement dans les résultats des simulations présentées, nous constatons que les estimations IPW-paramétrique semblent meilleures que celles des méthodes

non-paramétriques. Cependant, lorsque les probabilités de sélection sont mal spécifiées, la méthode paramétrique donne de mauvais résultats, voir Diallo *et al.* (2019) pour plus de détails.

Pour évaluer la qualité de l'approximation gaussienne énoncée dans les Théorèmes (3.2), (3.3) et (3.4) nous fournissons des graphiques Q-Q plots normaux des estimations $(\widehat{\gamma}_{i,n} - \gamma_i)/\text{s.e.}(\widehat{\gamma}_{i,n})$, $j = 1, \ldots, 4$, $(\widehat{\beta}_{1,j,n} - \beta_{1,j})/\text{s. e.}(\widehat{\beta}_{1,j,n})$, $j = 1, \ldots, 3$, $(\widehat{\beta}_{2,k,n} - \beta_{2,k})/\text{s.e.}\ (\widehat{\beta}_{2,k,n})$, $k = 1, \ldots, 3$ et $(\widehat{\alpha}_n - \alpha_l)/\text{s.e.}(\widehat{\alpha}_n)$. Nous fournissons ces graphiques pour $n = 1000$ avec 45% de proportion moyenne d'inflation de zéros dans les échantillons et 25% de données manquantes. Les figures 3.1 à 3.4 fournissent des graphiques Q-Q plot pour $\left(\widehat{\gamma}_{1,n}, \ldots, \widehat{\gamma}_{4,n}\right)$, $\left(\widehat{\beta}_{1,1,n}, \ldots, \widehat{\beta}_{1,3,n}\right)$, $\left(\widehat{\beta}_{2,1,n}, \ldots, \widehat{\beta}_{2,3,n}\right)$, $\widehat{\alpha}_n$, respectivement qui sont obtenus par la méthode IPW semi-paramétrique. De même, les figures 3.5 à 3.8 fournissent des graphiques Q-Q plot pour $\left(\widehat{\gamma}_{1,n}, \ldots, \widehat{\gamma}_{4,n}\right)$, $\left(\widehat{\beta}_{1,1,n}, \ldots, \widehat{\beta}_{1,3,n}\right)$, $\left(\widehat{\beta}_{2,1,n}, \ldots, \widehat{\beta}_{2,3,n}\right)$, $\widehat{\alpha}_n$, respectivement qui sont obtenus par la méthode MI de type 2 pour $n = 1000$, 40% de données manquantes et 45% de d'inflation de zéros. Les graphiques des autres scénarios simulés sont similaires et ne sont donc pas présentés. D'après ces figures, il apparaît que l'approximation gaussienne des distributions étudiées théoriquement sont satisfaites, même lorsque la proportion de données manquantes et la proportion d'inflation de zéros sont importantes.

		$\widehat{\gamma}_n$				$\widehat{\beta}_{1,n}$			$\widehat{\beta}_{2,n}$			$\widehat{\alpha}_n$
		$\hat{\gamma}_{1,n}$	$\hat{\gamma}_{2,n}$	$\hat{\gamma}_{3,n}$	$\hat{\gamma}_{4,n}$	$\hat{\beta}_{1,1,n}$	$\hat{\beta}_{1,2,n}$	$\hat{\beta}_{1,3,n}$	$\hat{\beta}_{2,1,n}$	$\hat{\beta}_{2,2,n}$	$\hat{\beta}_{2,3,n}$	$\hat{\alpha}_n$
FD	biais	0.0283	0.0006	-0.0188	-0.0221	-0.0233	0.0057	0.0036	-0.0074	0.0015	-0.0033	0.0016
	SD	0.9823	0.1039	0.0995	0.6277	0.1844	0.1548	0.0271	0.1373	0.1429	0.0309	0.0479
	RMSE	0.9817	0.1038	0.1012	0.6275	0.1857	0.1548	0.0273	0.1373	0.1428	0.0311	0.0479
IPW_P	biais	-0.0150	0.0048	-0.0258	-0.0020	-0.0258	0.0070	0.0034	-0.0073	0.0019	-0.0039	0.0020
	SD	1.2533	0.1402	0.1263	0.7977	0.2137	0.1705	0.0337	0.1405	0.1468	0.0337	0.0555
	RMSE	1.2521	0.1401	0.1288	0.7970	0.2150	0.1705	0.0339	0.1406	0.1467	0.0339	0.0555
IPW_S	biais	-0.1922	0.0041	-0.0187	0.1143	-0.0179	-0.0173	-0.0021	-0.0355	0.0301	-0.0022	0.0243
	SD	1.3094	0.1370	0.1233	0.8313	0.2090	0.1773	0.0317	0.1416	0.1496	0.0333	0.0517
	RMSE	1.3221	0.1369	0.1246	0.8383	0.2096	0.1780	0.0318	0.1458	0.1524	0.0333	0.0571
MI1	biais	-0.2249	0.0151	-0.3778	0.0046	-0.3593	0.0510	0.1048	-0.0930	-0.0250	0.0273	0.0370
	SD	1.0331	0.1049	0.1701	0.6492	0.2839	0.1700	0.0649	0.1472	0.1522	0.0523	0.0529
	RMSE	1.0563	0.1059	0.4143	0.6486	0.4577	0.1773	0.1232	0.1740	0.1541	0.0589	0.0645
MI2	biais	-0.2236	0.0152	-0.3765	0.0045	-0.3590	0.0509	0.1046	-0.0931	-0.0251	0.0272	0.0372
	SD	1.0322	0.1048	0.1698	0.6487	0.2838	0.1700	0.0649	0.1472	0.1522	0.0523	0.0528
	RMSE	1.0551	0.1058	0.4129	0.6480	0.4575	0.1773	0.1231	0.1740	0.1541	0.0589	0.0646

Tableau 3.1 – Résultats de la simulation pour $N = 500$ replications, taille des échantillons $n = 500$. Proportion moyenne d'inflation de zéros 30%. Proportion moyenne de données manquantes est égale à 25%. SD: écart-type empirique. RMSE: racine carrée de l'erreur quadratique moyenne.

		$\widehat{\gamma}_n$				$\widehat{\beta}_{1,n}$			$\widehat{\beta}_{2,n}$			$\widehat{\alpha}_n$
		$\hat{\gamma}_{1,n}$	$\hat{\gamma}_{2,n}$	$\hat{\gamma}_{3,n}$	$\hat{\gamma}_{4,n}$	$\hat{\beta}_{1,1,n}$	$\hat{\beta}_{1,2,n}$	$\hat{\beta}_{1,3,n}$	$\hat{\beta}_{2,1,n}$	$\hat{\beta}_{2,2,n}$	$\hat{\beta}_{2,3,n}$	$\hat{\alpha}_n$
FD	biais	0.0283	0.0006	-0.0188	-0.0221	-0.0233	0.0057	0.0036	-0.0074	0.0015	-0.0033	0.0016
	SD	0.9823	0.1039	0.0995	0.6277	0.1844	0.1548	0.0271	0.1373	0.1429	0.0309	0.0479
	RMSE	0.9817	0.1038	0.1012	0.6275	0.1857	0.1548	0.0273	0.1373	0.1428	0.0311	0.0479
IPW_P	biais	0.1339	-0.0125	-0.1341	-0.0472	-0.0878	0.0438	0.0124	-0.1614	0.1087	0.0155	-0.0080
	SD	1.9952	0.2082	0.2167	1.2738	0.3272	0.2793	0.0458	0.7522	0.7698	0.0840	0.0910
	RMSE	1.9977	0.2084	0.2547	1.2734	0.3385	0.2825	0.0474	0.7686	0.7767	0.0853	0.0913
IPW_S	biais	0.6476	-0.0069	-0.3326	-0.0727	-0.1454	0.0088	0.0227	-0.3617	0.0830	0.0572	-0.0447
	SD	1.5903	0.1644	0.1748	1.0119	0.2710	0.2384	0.0370	0.3358	0.3424	0.0657	0.0674
	RMSE	1.7156	0.1644	0.3756	1.0135	0.3073	0.2383	0.0434	0.4933	0.3520	0.0870	0.0808
MI1	biais	-0.6072	-0.2329	-1.4586	0.0741	-0.6278	0.6708	0.2663	0.8971	0.4881	-1.2574	1.4391
	SD	0.9974	0.1059	0.2051	0.6387	0.3292	0.1914	0.0936	0.4646	0.4284	0.7660	0.0521
	RMSE	0.9964	0.1071	0.5023	0.6386	0.4643	0.2307	0.1146	0.4742	0.4775	1.1492	0.1486
MI2	biais	-0.0238	0.0171	-0.4551	-0.0147	-0.3294	-0.1320	0.0678	0.1068	-0.2214	-0.8387	0.1397
	SD	1.0058	0.1054	0.2007	0.6417	0.3281	0.1897	0.0927	0.4225	0.4133	0.7452	0.0519
	RMSE	1.0051	0.1067	0.4973	0.6412	0.4647	0.2310	0.1148	0.4354	0.4685	1.1214	0.1490

Tableau 3.2 – Résultats de la simulation pour $N = 500$ replications, taille des échantillons $n = 500$. Proportion moyenne d'inflation de zéros 30%. Proportion moyenne de données manquantes est égale à 40%. SD: écart-type empirique. RMSE: racine carrée de l'erreur quadratique moyenne.

		$\widehat{\gamma}_n$				$\widehat{\beta}_{1,n}$			$\widehat{\beta}_{2,n}$			$\widehat{\alpha}_n$
		$\hat{\gamma}_{1,n}$	$\hat{\gamma}_{2,n}$	$\hat{\gamma}_{3,n}$	$\hat{\gamma}_{4,n}$	$\hat{\beta}_{1,1,n}$	$\hat{\beta}_{1,2,n}$	$\hat{\beta}_{1,3,n}$	$\hat{\beta}_{2,1,n}$	$\hat{\beta}_{2,2,n}$	$\hat{\beta}_{2,3,n}$	$\hat{\alpha}_n$
FD	biais	-0.0613	0.0081	-0.0185	0.0509	-0.0233	0.0088	0.0028	-0.0166	0.0089	-0.003	0.0018
	SD	0.9121	0.0987	0.0933	0.5860	0.1988	0.1634	0.0301	0.1760	0.1827	0.0399	0.0506
	RMSE	0.9132	0.0989	0.0951	0.5876	0.2000	0.1635	0.0302	0.1766	0.1827	0.0399	0.0505
IPW_P	biais	-0.0684	0.0092	-0.0269	0.0524	-0.0227	0.0052	0.0023	-0.0144	0.0053	-0.004	0.0025
	SD	1.0811	0.1201	0.1106	0.6938	0.2256	0.1769	0.0369	0.1811	0.1883	0.0426	0.0579
	RMSE	1.0822	0.1204	0.1137	0.6950	0.2265	0.1768	0.0370	0.1814	0.1882	0.0427	0.0579
IPW_S	biais	-0.2002	0.0089	-0.0189	0.1395	-0.0089	-0.0234	-0.0034	-0.0417	0.0304	-0.003	0.0242
	SD	1.1174	0.1179	0.1087	0.7145	0.2185	0.1786	0.0346	0.1822	0.1902	0.0421	0.0538
	RMSE	1.1341	0.1181	0.1102	0.7273	0.2185	0.1800	0.0348	0.1867	0.1924	0.0421	0.0590
MI1	biais	-0.1295	-0.0093	-0.2589	0.0557	-0.3401	0.0483	0.1009	-0.1419	-0.0198	0.0665	0.0177
	SD	0.9802	0.1020	0.1286	0.6265	0.3000	0.1782	0.0693	0.1988	0.1976	0.0778	0.0614
	RMSE	0.9877	0.1024	0.2891	0.6284	0.4533	0.1845	0.1224	0.2441	0.1984	0.1022	0.0638
MI2	biais	-0.1287	-0.0094	-0.2583	0.0553	-0.3398	0.0481	0.1008	-0.1420	-0.0200	0.0665	0.0180
	SD	0.9796	0.1020	0.1285	0.6262	0.2999	0.1782	0.0693	0.1989	0.1977	0.0779	0.0614
	RMSE	0.9871	0.1023	0.2884	0.6280	0.453	0.1844	0.1222	0.2442	0.1985	0.1023	0.0639

Tableau 3.3 – Résultats de la simulation pour $N = 500$ replications, taille des échantillons $n = 500$. Proportion moyenne d'inflation de zéros 45%. Proportion moyenne de données manquantes est égale à 25%. SD: écart-type empirique. RMSE: racine carrée de l'erreur quadratique moyenne.

		$\widehat{\gamma}_n$				$\widehat{\beta}_{1,n}$			$\widehat{\beta}_{2,n}$			$\widehat{\alpha}_n$
		$\hat{\gamma}_{1,n}$	$\hat{\gamma}_{2,n}$	$\hat{\gamma}_{3,n}$	$\hat{\gamma}_{4,n}$	$\hat{\beta}_{1,1,n}$	$\hat{\beta}_{1,2,n}$	$\hat{\beta}_{1,3,n}$	$\hat{\beta}_{2,1,n}$	$\hat{\beta}_{2,2,n}$	$\hat{\beta}_{2,3,n}$	$\hat{\alpha}_n$
FD	biais	-0.0613	0.0081	-0.0185	0.0509	-0.0233	0.0088	0.0028	-0.0166	0.0089	-0.0030	0.0018
	SD	0.9121	0.0987	0.0933	0.5860	0.1988	0.1634	0.0301	0.1760	0.1827	0.0399	0.0506
	RMSE	0.9132	0.0989	0.0951	0.5876	0.2000	0.1635	0.0302	0.1766	0.1827	0.0399	0.0505
IPW_P	biais	-0.0599	0.0143	-0.0367	0.0452	-0.0254	0.0122	0.0029	-0.0174	0.0116	-0.0041	0.0004
	SD	1.4153	0.1540	0.1466	0.9070	0.2210	0.1799	0.0340	0.1894	0.1963	0.0431	0.0565
	RMSE	1.4152	0.1545	0.1510	0.9072	0.2222	0.1802	0.0341	0.1900	0.1965	0.0433	0.0564
IPW_S	biais	-0.1577	0.0121	-0.0402	0.1289	-0.0251	0.0424	-0.0015	-0.0420	0.0451	0.0001	0.0177
	SD	1.4766	0.1490	0.1453	0.9441	0.2185	0.1799	0.0331	0.1900	0.1987	0.0421	0.0549
	RMSE	1.4836	0.1494	0.1506	0.9520	0.2198	0.1847	0.0331	0.1944	0.2035	0.0420	0.0576
MI1	biais	-0.5654	0.2899	-1.6997	0.5830	-0.4588	0.8101	0.2439	0.7250	0.7051	-0.3923	1.3187
	SD	1.0742	0.1091	0.2196	0.6777	0.2419	0.1703	0.0478	0.1940	0.1909	0.0635	0.0536
	RMSE	1.0736	0.5508	0.7333	0.8317	0.2891	0.1704	0.0649	0.2078	0.1907	0.0639	0.0567
MI2	biais	-0.1644	-0.0103	-0.6989	0.0526	-0.1589	0.0101	0.0440	-0.0751	0.0051	0.0078	0.0188
	SD	1.0733	0.1090	0.2193	0.6772	0.2419	0.1703	0.0478	0.194	0.1909	0.0635	0.0536
	RMSE	1.0847	0.1094	0.7324	0.6786	0.2892	0.1704	0.0649	0.2078	0.1908	0.0640	0.0568

Tableau 3.4 – Résultats de la simulation pour $N = 500$ replications, taille des échantillons $n = 500$. Proportion moyenne d'inflation de zéros 45%. Proportion moyenne de données manquantes est égale à 40%. SD: écart-type empirique. RMSE: racine carrée de l'erreur quadratique moyenne.

		$\widehat{\gamma}_n$				$\widehat{\beta}_{1,n}$			$\widehat{\beta}_{2,n}$			$\widehat{\alpha}_n$
		$\hat{\gamma}_{1,n}$	$\hat{\gamma}_{2,n}$	$\hat{\gamma}_{3,n}$	$\hat{\gamma}_{4,n}$	$\hat{\beta}_{1,1,n}$	$\hat{\beta}_{1,2,n}$	$\hat{\beta}_{1,3,n}$	$\hat{\beta}_{2,1,n}$	$\hat{\beta}_{2,2,n}$	$\hat{\beta}_{2,3,n}$	$\hat{\alpha}_n$
FD	biais	0.0159	-0.0020	-0.0101	-0.0123	-0.0118	0.0098	0.0006	-0.0057	0.0030	0.0000	-0.0017
	SD	0.7082	0.0719	0.0650	0.4558	0.1303	0.1042	0.0187	0.0927	0.0946	0.0230	0.0324
	RMSE	0.7077	0.0719	0.0657	0.4555	0.1307	0.1046	0.0186	0.0928	0.0946	0.0230	0.0324
IPW_P	biais	0.0034	-0.0050	-0.0145	-0.0062	-0.0191	0.0154	0.0010	-0.0080	0.0048	-0.0004	-0.0010
	SD	0.8927	0.0949	0.0868	0.5760	0.1514	0.1184	0.0233	0.0979	0.0992	0.0248	0.0361
	RMSE	0.8918	0.0950	0.0880	0.5755	0.1525	0.1192	0.0233	0.0981	0.0992	0.0247	0.0361
IPW_S	biais	-0.2020	-0.0047	-0.0109	0.1256	0.0075	-0.0064	-0.0057	-0.0311	0.0347	0.0024	0.0100
	SD	0.9234	0.0946	0.0863	0.5950	0.1448	0.1173	0.0210	0.0955	0.0983	0.0244	0.0339
	RMSE	0.9443	0.0946	0.0869	0.6076	0.1449	0.1173	0.0218	0.1004	0.1041	0.0244	0.0353
MI1	biais	-0.1839	0.0093	-0.3625	0.0051	-0.2678	0.0364	0.0742	-0.0679	-0.0138	0.0067	0.0342
	SD	0.7568	0.0752	0.1183	0.4890	0.2010	0.1127	0.0450	0.0984	0.0979	0.0281	0.0355
	RMSE	0.7781	0.0757	0.3813	0.4885	0.3347	0.1183	0.0867	0.1195	0.0988	0.0288	0.0493
MI2	biais	-0.1817	0.0092	-0.3619	0.0039	-0.2667	0.0364	0.0739	-0.0675	-0.014	0.0068	0.0339
	SD	0.7546	0.0751	0.1182	0.4873	0.2011	0.1126	0.0448	0.0984	0.0980	0.0281	0.0357
	RMSE	0.7754	0.0756	0.3806	0.4868	0.3338	0.1182	0.0864	0.1192	0.0989	0.0289	0.0492

Tableau 3.5 – Résultats de la simulation pour $N = 500$ replications, taille des échantillons $n = 1000$. Proportion moyenne d'inflation de zéros 30%. Proportion moyenne de données manquantes est égale à 25%. SD: écart-type empirique. RMSE: racine carrée de l'erreur quadratique moyenne.

		$\widehat{\gamma}_n$				$\widehat{\beta}_{1,n}$			$\widehat{\beta}_{2,n}$			$\widehat{\alpha}_n$
		$\hat{\gamma}_{1,n}$	$\hat{\gamma}_{2,n}$	$\hat{\gamma}_{3,n}$	$\hat{\gamma}_{4,n}$	$\hat{\beta}_{1,1,n}$	$\hat{\beta}_{1,2,n}$	$\hat{\beta}_{1,3,n}$	$\hat{\beta}_{2,1,n}$	$\hat{\beta}_{2,2,n}$	$\hat{\beta}_{2,3,n}$	$\hat{\alpha}_n$
FD	biais	0.0159	-0.0020	-0.0101	-0.0123	-0.0118	0.0098	0.0006	-0.0057	0.0030	0.0000	-0.0017
	SD	0.7082	0.0719	0.065	0.4558	0.1303	0.1042	0.0187	0.0927	0.0946	0.0230	0.0324
	RMSE	0.7077	0.0719	0.0657	0.4555	0.1307	0.1046	0.0186	0.0928	0.0946	0.0230	0.0324
IPW_P	biais	0.0638	-0.012	-0.1001	-0.0074	-0.0508	0.0307	0.0069	-0.0894	0.0525	0.0193	-0.0065
	SD	1.5251	0.1519	0.172	0.9703	0.2125	0.1764	0.0300	0.2686	0.2790	0.0623	0.0670
	RMSE											
IPW_S	biais	0.5974	-0.0082	-0.2883	-0.119	-0.1184	0.0209	0.0174	-0.2828	0.0604	0.0623	-0.0442
	SD	1.092	0.1084	0.1216	0.6944	0.1913	0.1654	0.025	0.2229	0.2295	0.048	0.0478
	RMSE	1.2438	0.1086	0.3129	0.7038	0.2248	0.1665	0.0304	0.3599	0.2371	0.0786	0.0651
MI1	biais	0.0512	0.0054	-0.4562	0.0000	-0.2024	-0.1296	0.0293	0.3030	-0.2621	-0.7041	0.1245
	SD	0.7525	0.0776	0.1298	0.4801	0.2733	0.1350	0.0768	0.2592	0.2485	0.3669	0.0386
	RMSE	0.7535	0.0777	0.4743	0.4796	0.3399	0.1870	0.0822	0.3986	0.3610	0.7938	0.1303
MI2	biais	0.0511	0.0055	-0.4556	0.0000	-0.2043	-0.1291	0.0298	0.3031	-0.262	-0.699	0.1246
	SD	0.7522	0.0776	0.1296	0.4799	0.2736	0.1347	0.0769	0.2570	0.2459	0.3619	0.0387
	RMSE	0.7532	0.0777	0.4736	0.4795	0.3413	0.1865	0.0824	0.3972	0.3591	0.787	0.1304

Tableau 3.6 – Résultats de la simulation pour $N = 500$ replications, taille des échantillons $n = 1000$. Proportion moyenne d'inflation de zéros 30%. Proportion moyenne de données manquantes est égale à 40%. SD: écart-type empirique. RMSE: racine carrée de l'erreur quadratique moyenne.

		$\widehat{\gamma}_n$				$\widehat{\beta}_{1,n}$			$\widehat{\beta}_{2,n}$			$\widehat{\alpha}_n$
		$\hat{\gamma}_{1,n}$	$\hat{\gamma}_{2,n}$	$\hat{\gamma}_{3,n}$	$\hat{\gamma}_{4,n}$	$\hat{\beta}_{1,1,n}$	$\hat{\beta}_{1,2,n}$	$\hat{\beta}_{1,3,n}$	$\hat{\beta}_{2,1,n}$	$\hat{\beta}_{2,2,n}$	$\hat{\beta}_{2,3,n}$	$\hat{\alpha}_n$
FD	biais	-0.0566	0.0046	-0.0071	0.0362	-0.0122	0.0087	0.0008	-0.0034	0.0004	-0.0008	-0.0013
	SD	0.6338	0.0667	0.0625	0.4049	0.1438	0.1146	0.0208	0.1182	0.1221	0.0281	0.0361
	RMSE	0.6356	0.0668	0.0628	0.4061	0.1442	0.1148	0.0208	0.1182	0.1220	0.0281	0.0361
IPW_P	biais	-0.0748	0.0057	-0.0110	0.0465	-0.0179	0.0131	0.0011	-0.0048	0.0016	-0.0007	-0.0013
	SD	0.7712	0.0805	0.0748	0.4934	0.1629	0.1239	0.0252	0.1230	0.1253	0.0308	0.0434
	RMSE	0.7741	0.0807	0.0756	0.4951	0.1638	0.1245	0.0252	0.1230	0.1252	0.0308	0.0434
IPW_S	biais	-0.2319	0.0056	-0.0069	0.1474	0.0119	-0.0144	-0.0059	-0.0269	0.0269	0.0002	0.0114
	SD	0.8021	0.0804	0.0745	0.5122	0.1554	0.1233	0.0232	0.1217	0.1263	0.0300	0.0379
	RMSE	0.8342	0.0805	0.0748	0.5325	0.1557	0.1240	0.0239	0.1245	0.1290	0.0300	0.0396
MI1	biais	-0.0910	-0.0102	-0.2583	0.0464	-0.2569	0.0328	0.0707	-0.0820	-0.0207	0.0250	0.0248
	SD	0.6709	0.0684	0.0858	0.4261	0.2154	0.1215	0.0483	0.1297	0.1273	0.0395	0.0390
	RMSE	0.6764	0.0691	0.2721	0.4282	0.3351	0.1258	0.0855	0.1533	0.1288	0.0467	0.0462
MI2	biais	-0.0906	-0.0102	-0.2580	0.0462	-0.2567	0.0327	0.0706	-0.0820	-0.0207	0.0250	0.0249
	SD	0.6707	0.0684	0.0857	0.4260	0.2153	0.1215	0.0483	0.1297	0.1273	0.0395	0.0390
	RMSE	0.6762	0.0691	0.2719	0.4281	0.3349	0.1257	0.0855	0.1533	0.1288	0.0467	0.0462

Tableau 3.7 – Résultats de la simulation pour $N = 500$ replications, taille des échantillons $n = 1000$. Proportion moyenne d'inflation de zéros 45%. Proportion moyenne de données manquantes est égale à 25%. SD: écart-type empirique. RMSE: racine carrée de l'erreur quadratique moyenne.

		$\widehat{\gamma}_n$				$\widehat{\beta}_{1,n}$			$\widehat{\beta}_{2,n}$			$\widehat{\alpha}_n$
		$\hat{\gamma}_{1,n}$	$\hat{\gamma}_{2,n}$	$\hat{\gamma}_{3,n}$	$\hat{\gamma}_{4,n}$	$\hat{\beta}_{1,1,n}$	$\hat{\beta}_{1,2,n}$	$\hat{\beta}_{1,3,n}$	$\hat{\beta}_{2,1,n}$	$\hat{\beta}_{2,2,n}$	$\hat{\beta}_{2,3,n}$	$\hat{\alpha}_n$
FD	biais	-0.0566	0.0046	-0.0071	0.0362	-0.0122	0.0087	0.0008	-0.0034	0.0004	-0.0008	-0.0013
	SD	0.6338	0.0667	0.0625	0.4049	0.1438	0.1146	0.0208	0.1182	0.1221	0.0281	0.0361
	RMSE	0.6356	0.0668	0.0628	0.4061	0.1442	0.1148	0.0208	0.1182	0.1220	0.0281	0.0361
IPW_P	biais	-0.0921	0.0101	-0.0192	0.0578	-0.0159	0.0138	0.0007	-0.0053	0.0023	-0.0004	-0.0010
	SD	0.9189	0.1070	0.1022	0.5846	0.1585	0.1228	0.0225	0.1261	0.1291	0.0312	0.0388
	RMSE	0.9226	0.1073	0.1038	0.5869	0.1591	0.1234	0.0225	0.1261	0.1290	0.0312	0.0388
IPW_S	biais	-0.2430	0.0110	-0.0196	0.1636	-0.0179	0.0529	-0.0042	-0.0290	0.0403	0.0048	0.0049
	SD	0.9679	0.1044	0.1012	0.6143	0.1547	0.1229	0.0221	0.1263	0.1314	0.0307	0.0386
	RMSE	0.9970	0.1048	0.1029	0.6352	0.1556	0.1337	0.0225	0.1294	0.1373	0.0310	0.0389
MI1	biais	-0.2249	0.0151	-0.3778	0.0046	-0.3593	0.0510	0.1048	-0.0930	-0.0250	0.0273	0.0370
	SD	0.7243	0.0760	0.1460	0.4098	0.1690	0.1159	0.0275	0.1271	0.1270	0.0345	0.0389
	RMSE	0.7350	0.0782	0.6963	0.4635	0.2074	0.1164	0.0349	0.1292	0.1278	0.0387	0.0420
MI2	biais	-0.1286	-0.0124	-0.6805	0.0535	-0.1214	0.0092	0.0304	-0.0202	-0.0067	-0.0148	0.0180
	SD	0.7241	0.0762	0.1458	0.4598	0.1694	0.1156	0.0331	0.1263	0.1253	0.0363	0.0376
	RMSE	0.7347	0.0772	0.6959	0.4624	0.2083	0.1159	0.0449	0.1277	0.1253	0.0392	0.0417

Tableau 3.8 – Résultats de la simulation pour $N = 500$ replications, taille des échantillons $n = 1000$. Proportion moyenne d'inflation de zéros 45%. Proportion moyenne de données manquantes est égale à 40%. SD: écart-type empirique. RMSE: racine carrée de l'erreur quadratique moyenne.

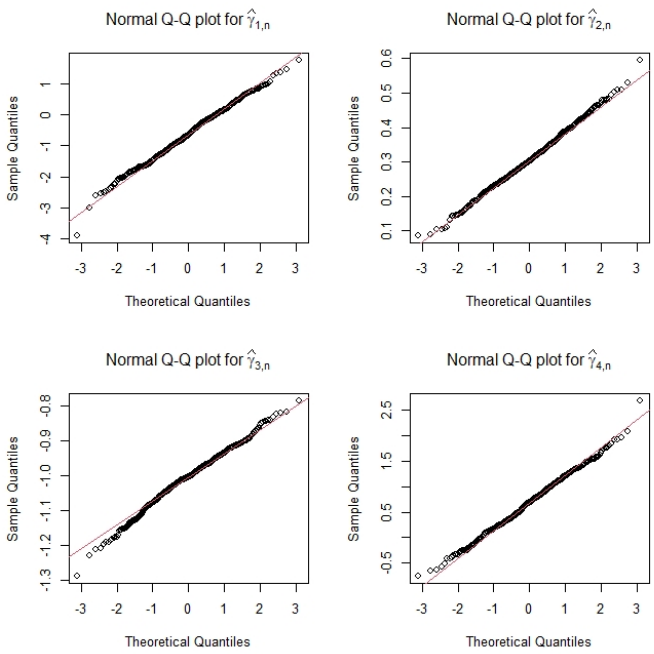

Figure 3.1 – QQ-plot normaux pour $\widehat{\gamma}_{1,n}, \ldots, \widehat{\gamma}_{4,n}$ avec $n = 1000$, 45% d'inflation de zéros et 25% de données manquantes.

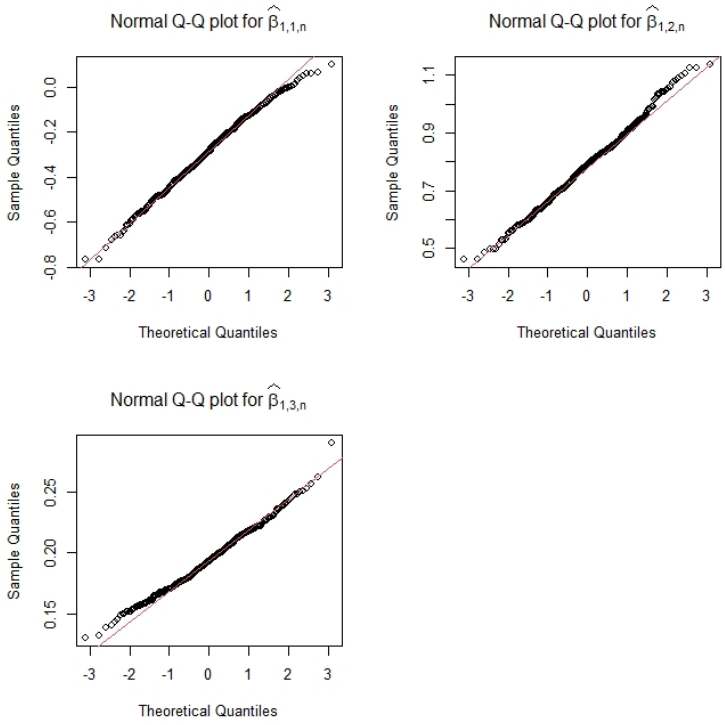

Figure 3.2 – QQ-plot normaux pour $\widehat{\beta}_{1,1,n}, \ldots, \widehat{\beta}_{1,3,n}$ avec $n = 1000$, 45% d'inflation de zéros et 25% de données manquantes.

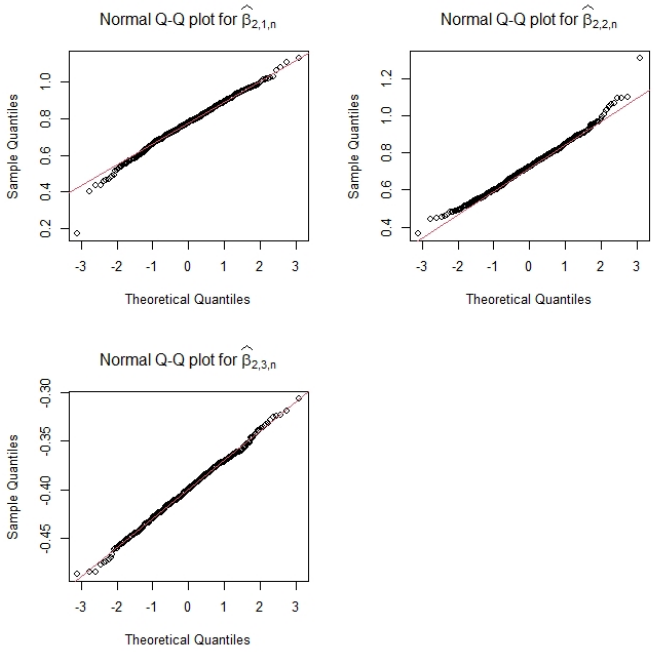

Figure 3.3 – QQ-plot normaux pour $\widehat{\beta}_{2,1,n}, \ldots, \widehat{\beta}_{2,3,n}$ avec $n = 1000$, 45% d'inflation de zéros et 25% de données manquantes.

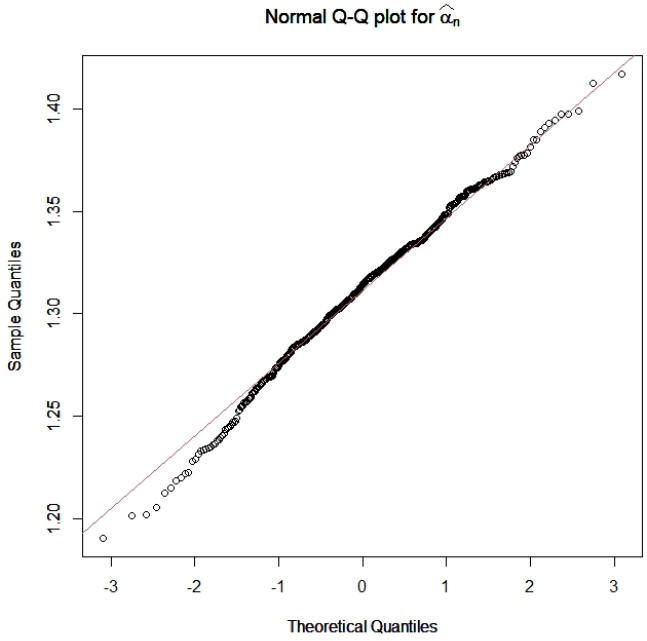

Figure 3.4 – QQ-plot normal pour $\widehat{\alpha}_n$ avec $n = 1000$, 45% d'inflation de zéros et 25% de données manquantes.

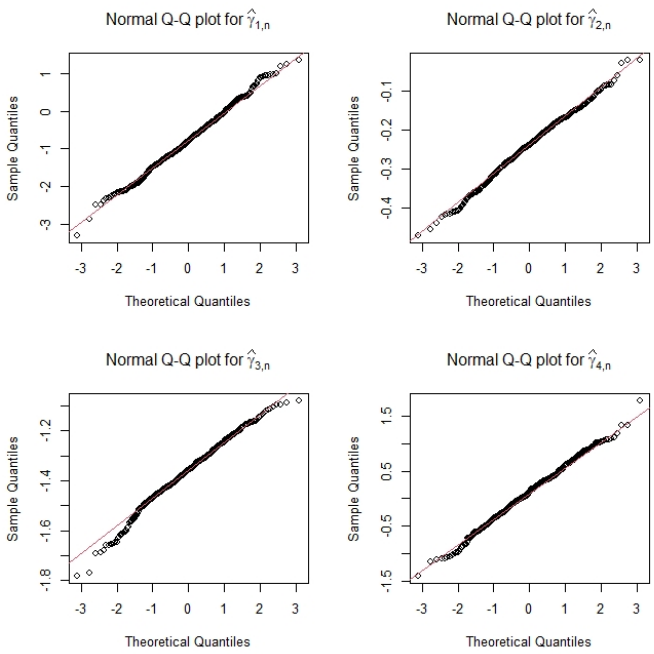

Figure 3.5 – QQ-plot normaux pour $\widehat{\gamma}_{1,n}, \ldots, \widehat{\gamma}_{4,n}$ avec $n = 1000$, 45% d'inflation de zéros et 40% de données manquantes.

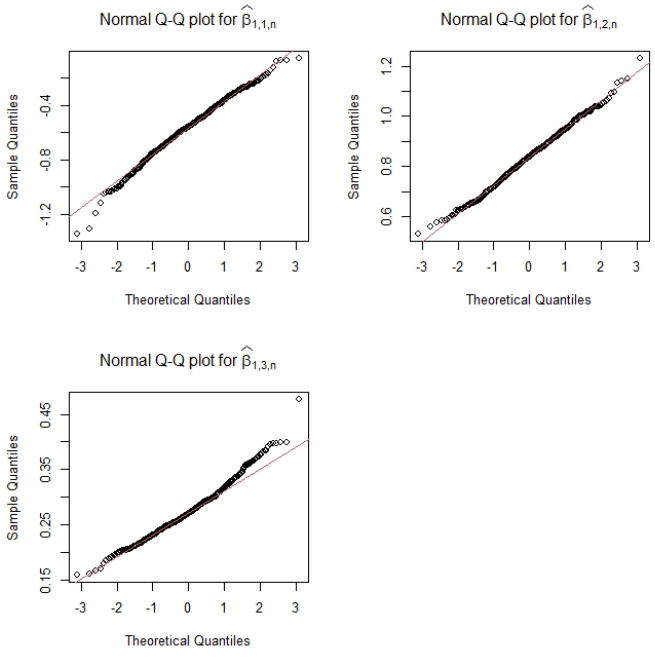

Figure 3.6 – QQ-plot normaux pour $\widehat{\beta}_{1,1,n}, \ldots, \widehat{\beta}_{1,3,n}$ avec $n = 1000$, 45% d'inflation de zéros et 40% de données manquantes.

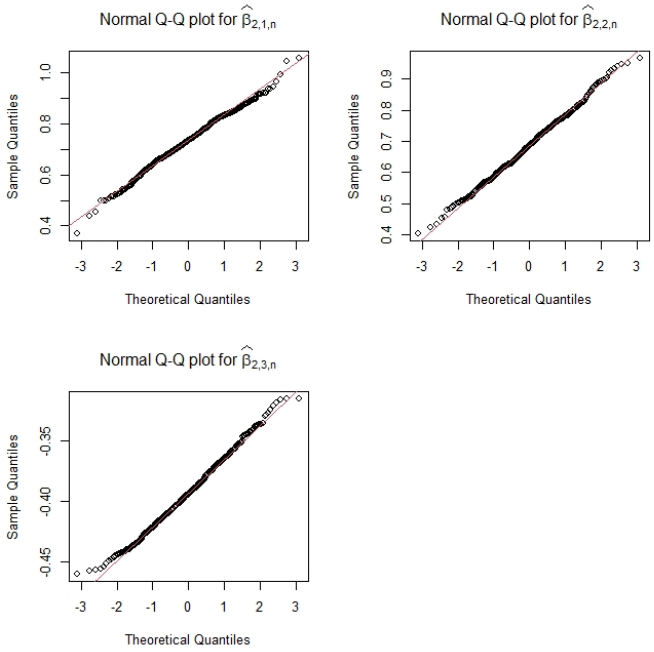

Figure 3.7 – QQ-plot normaux pour $\widehat{\beta}_{2,1,n}, \ldots, \widehat{\beta}_{2,3,n}$ avec $n = 1000$, 45% d'inflation de zéros et 40% de données manquantes.

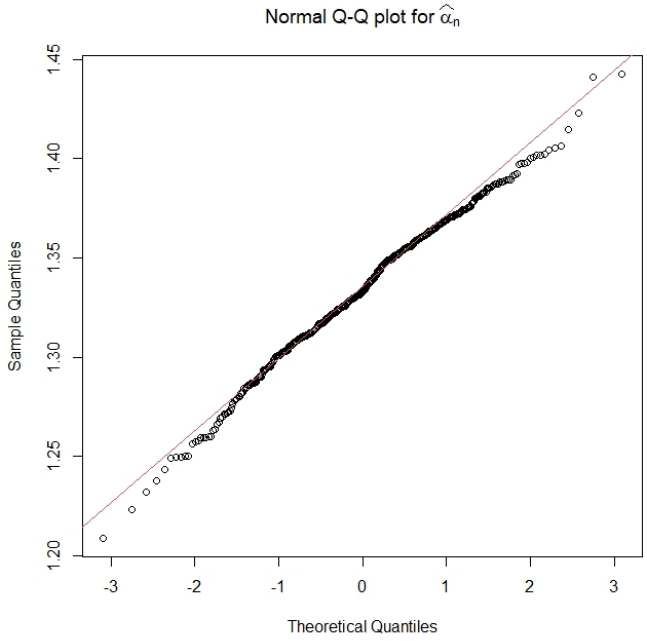

Figure 3.8 – QQ-plot normal pour $\widehat{\alpha}_n$ avec $n = 1000$, 45% d'inflation de zéros et 40% de données manquantes.

3.6 Application sur des données réelles

Dans cette section, nous présentons une application pratique des méthodologies proposées sur un jeu de données issu d'une étude de santé publique américaine. Ces données proviennent de l'enquête nationale sur les dépenses médicales (NMES), menée en 1987 et 1988 aux États-Unis. Elles fournissent un tableau détaillé de la manière dont les américains (âgés de 66 ans et plus) utilisent et paient les services de santé. Certaines mesures d'utilisation des services de santé ont été rapportées dans cette étude, comme le nombre de visites à un professionnel de la santé non médecin dans un cabinet médical et le nombre de visites à un médecin dans un cabinet médical. Des informations sur l'état de santé des patients sont également communiquées, ainsi que des variables économiques et socio-démographiques. Une description détaillée de ces données se trouve dans Deb et Trivedi (1997).

Dans ces données, nous considérons conjointement comme mesures d'utilisation des services de santé : le nombre de consultations avec un non-médecin en cabinet (`ofnd`) et le nombre de consultations avec un non-médecin en ambulatoire (`opnd`). La proportion d'observations de $(0;0)$ pour le couple de variables (`ofnd`; `opnd`) est de 59,03%. Ce chiffre indique qu'une proportion importante de patients ont renoncé aux consultations des professionnels de santé non médecin. En outre, les tests effectués avec la fonction `cor.test` du package `stats` du logiciel `R` sur les variables `ofnd` et `opnd` confirme que les variables `ofnd` et `opnd` sont corrélées. Par conséquent, nous suggérons d'analyser conjointement les observations de `ofnd` et `opnd` à l'aide d'un modèle ZIBP. Nous utilisons certaines covariables qui ont été enregistrées sur chaque individu de l'échantillon. (i) variables socio-économiques : l'âge (en années, divisé par 10, noté `age`), la durée de la scolarité (nombre d'années d'éducation, noté `school`), le revenu familial (en dix mille dollars, noté `income`); (ii) le nombre de maladies chroniques (cancer, diabète, arthrite, . . .dénoté par `chronic`), variable indiquant l'état de santé; (iii) `medicaid`, une variable binaire qui indique si la personne est couverte par medicaid ou non (medicaid est une assurance maladie américaine pour les individus ayant des revenus et des ressources limités). Nous la codons comme 1 si la personne est couverte et 0 sinon. Nous ajustons un modèle de régression ZIBP incorporant toutes ces covariables dans (3.2)-(3.3) pour chaque individu.
Afin d'illustrer les performances des méthodes proposées, nous simulons respectivement 10% (moyenne) et 30% (forte) proportions de données manquantes dans la variable `age`. Ceci nous permet de voir les effets des valeurs manquantes sur la qualité des estimations. Pour la méthode MI, le nombre d'imputations est $M = 30$. Les résultats de l'analyse des données sont présentés dans les Tableaux 3.9 à 3.10.

Des Tableaux 3.9 à 3.10, nous observons que les estimations des paramètres par les méthodes IPW paramétrique, IPW semi-paramétrique et MI de type 2 sont proches de celles obtenues sur le jeu de données complètes. De plus, les estimations obtenues par les méthodes proposées ont le même signe que celles prises comme référence. En outre, les standards errors des estimations évoluent très faiblement dans le même sens que les proportions de données manquantes. Des P-values des tests de Wald et des signes des estimations, on peut affirmer que les interprétations des résultats ne sont pas modifiées malgré l'augmentation des proportions de données manquantes. Nous pouvons donc conclure que nos méthodes proposées sont assez robustes pour fournir de bons résultats même en présence de grandes proportions de données manquantes.

Sur la base des résultats des méthodes d'estimation, nous pouvons affirmer que les facteurs qui influent sur la décision de ne jamais avoir recours à un non-médecin sont l'état de santé, le niveau d'étude, l'âge et le statut au regard de Medicaid. Plus précisément, la probabilité de ne jamais recourir à des professionnels de santé non-médecin diminue lorsque l'état de santé se dégrade. Cela peut s'expliquer par le fait qu'autant l'état de santé d'un patient se dégrade, autant il privilégie les consultations de spécialistes. Ensuite, la probabilité pour qu'un patient renonce à faire recours à des professionnels de santé non-médecin diminue avec le niveau d'étude. En effet, un niveau d'étude peu élevé s'accompagnerait d'un déficit d'information sur les possibilités de soins existantes, favorisant ainsi le renoncement aux soins. En revanche, les patients plus instruits sont mieux informés des services de soins médicaux. Par ailleurs, la probabilité de renoncer systématiquement à consulter un non-médecin augmente avec l'âge. Cela peut s'expliquer par la difficulté de mobilité associé au vieillissement. Ainsi, les patients âgés auront tendance à privilégier les consultations qu'ils considèrent plus nécessaires (consultation d'un médecin). De plus, au vu de la dégradation de l'état de santé avec le vieillissement, les patients dont l'état de santé décline sont susceptibles de favoriser les visites chez un médecin. Enfin, on peut souligner que les bénéficiaires de Medicaid sont plus susceptibles de renoncer à des consultations non médicales. Une explication est que les patients bénéficiant d'une couverture Medicaid tendent à limiter leurs consultations à celles qui sont nécessaires, c'est-à-dire aux seules visites chez le médecin (rappelons que Medicaid est une assurance maladie pour les personnes à faible revenu).

		FD-estimateur			IPW paramétrique			IPW semi-paramétrique			MI de type 2		
Variable		$\widehat{\Theta}_{F,n}$	S.E.	P-value	$\widehat{\Theta}_{Wp}$	S.E.	P-value	$\widehat{\Theta}_{Ws}$	S.E.	P-value	$\widehat{\Theta}_{MI2}$	S.E.	P-value
intercept	$\hat{\gamma}_1$	0.24384	0.42232	0.56368	0.26412	0.49772	0.59565	0.31556	0.51489	0.53996	0.37795	0.42903	0.37835
chronic	$\hat{\gamma}_2$	-0.17020	0.02345	<0.0001	-0.16656	0.02348	<0.0001	-0.18984	0.02386	<0.0001	-0.16845	0.02341	<0.0001
age	$\hat{\gamma}_3$	0.12749	0.05103	0.01248	0.10863	0.05804	0.06127	0.10390	0.06032	0.08498	0.10939	0.05322	0.03983
medicaid	$\hat{\gamma}_4$	0.31424	0.12613	0.01272	0.39519	0.12942	0.00226	0.36026	0.12825	0.00497	0.31811	0.11812	0.00707
income	$\hat{\gamma}_5$	-0.00262	0.01116	0.81419	-0.00257	0.01114	0.81719	0.01116	0.010758	0.29956	-0.00292	0.01116	0.79355
school	$\hat{\gamma}_6$	-0.08608	0.00940	<0.0001	-0.08342	0.00954	<0.0001	-0.08130	0.00957	<0.0001	-0.08659	0.00931	<0.0001
intercept	$\hat{\beta}_{1,1}$	1.53838	0.18061	< 0.0001	1.45939	0.17195	<0.0001	1.45650	0.18866	< 0.0001	1.49807	0.18472	< 0.0001
chronic	$\hat{\beta}_{1,2}$	0.01272	0.00925	0.16938	0.01359	0.00922	0.14063	0.02056	0.00925	0.02627	0.01198	0.00925	0.19552
age	$\hat{\beta}_{1,3}$	-0.11291	0.02178	<0.0001	-0.11037	0.02085	<0.0001	-0.10963	0.02253	<0.0001	-0.10678	0.02226	< 0.0001
medicaid	$\hat{\beta}_{1,4}$	0.22973	0.04929	<0.0001	0.27317	0.04932	<0.0001	0.28121	0.04983	<0.0001	0.22609	0.04918	<0.0001
income	$\hat{\beta}_{1,5}$	-0.01613	0.00453	0.00036	-0.01571	0.00450	0.00049	-0.02055	0.00459	<0.0001	-0.01605	0.00452	0.00039
school	$\hat{\beta}_{1,6}$	0.03504	0.00395	<0.0001	0.03608	0.00390	<0.0001	0.03522	0.00396	<0.0001	0.03505	0.00395	<0.0001
intercept	$\hat{\beta}_{2,1}$	4.22773	0.34159	<0.0001	4.13615	0.29429	<0.0001	4.11682	0.33702	<0.0001	4.31679	0.33570	< 0.0001
chronic	$\hat{\beta}_{2,2}$	0.13239	0.01543	<0.0001	0.13540	0.01537	<0.0001	0.14328	0.01544	<0.0001	0.13189	0.01545	<0.0001
age	$\hat{\beta}_{2,3}$	-0.46741	0.04213	<0.0001	-0.46555	0.03706	<0.0001	-0.46377	0.04202	<0.0001	-0.47739	0.04120	<0.0001
medicaid	$\hat{\beta}_4$	0.13673	0.07771	0.07849	0.19329	0.07609	0.01108	0.20508	0.07627	0.00717	0.12687	0.07786	0.10320
income	$\hat{\beta}_{2,5}$	-0.01200	0.00948	0.20546	-0.01102	0.00941	0.24156	-0.01619	0.00957	0.09063	-0.01211	0.00949	0.20183
school	$\hat{\beta}_{2,6}$	-0.10068	0.00666	<0.0001	-0.09988	0.00649	<0.0001	-0.10021	0.00659	<0.0001	-0.10118	0.00665	<0.0001

Tableau 3.9 – Résultats du modèle de régression ZIBP avec 10% données manquantes.

		FD-estimateur			IPW paramétrique			IPW semi-paramétrique			MI de type 2		
Variable		$\widehat{\Theta}_{F,n}$	S.E.	P-value	$\widehat{\Theta}_{Wp}$	S.E.	P-value	$\widehat{\Theta}_{Ws}$	S.E.	P-value	$\widehat{\Theta}_{MI2}$	S.E.	P-value
intercept	$\hat{\gamma}_1$	0.24384	0.42232	0.56368	0.31050	0.39107	0.42720	0.45839	0.36437	0.20839	0.07242	0.80095	0.92795
chronic	$\hat{\gamma}_2$	-0.17020	0.02345	<0.0001	-0.20733	0.02296	<0.0001	-0.29337	0.02627	<0.0001	-0.16710	0.02347	<0.0001
age	$\hat{\gamma}_3$	0.12749	0.05103	0.01248	0.10274	0.04579	0.02485	0.09173	0.04611	0.04667	0.15149	0.09366	0.10579
medicaid	$\hat{\gamma}_4$	0.31424	0.12613	0.01272	0.52252	0.12486	<0.0001	0.44195	0.12912	0.00062	0.31547	0.13127	0.01625
income	$\hat{\gamma}_5$	-0.00262	0.01116	0.81419	0.003238	0.01111	0.77086	0.02280	0.01023	0.02590	-0.00272	0.01117	0.80730
school	$\hat{\gamma}_6$	-0.08608	0.00940	<0.0001	-0.09317	0.00936	<0.0001	-0.08639	0.00917	<0.0001	-0.08669	0.01036	<0.0001
intercept	$\hat{\beta}_{1,1}$	1.53838	0.18061	< 0.0001	1.50947	0.17816	<0.0001	1.50576	0.17989	<0.0001	0.74288	0.15454	< 0.0001
chronic	$\hat{\beta}_{1,2}$	0.01272	0.00925	0.16938	-0.00122	0.00908	0.89248	0.06822	0.00948	<0.0001	0.00868	0.00924	0.34754
age	$\hat{\beta}_{1,3}$	-0.11291	0.02178	<0.0001	-0.12868	0.02172	<0.0001	-0.12855	0.02159	<0.0001	-0.00588	0.01771	0.73967
medicaid	$\hat{\beta}_{1,4}$	0.22973	0.04929	<0.0001	0.36889	0.04802	<0.0001	0.37310	0.05029	<0.0001	0.22223	0.04934	<0.0001
income	$\hat{\beta}_{1,5}$	-0.01613	0.00453	0.00036	-0.01212	0.00439	0.00579	-0.03606	0.00479	<0.0001	-0.01517	0.00449	0.00072
school	$\hat{\beta}_{1,6}$	0.03504	0.00395	<0.0001	0.03561	0.00393	<0.0001	0.03181	0.00395	<0.0001	0.03701	0.00391	<0.0001
intercept	$\hat{\beta}_{2,1}$	4.22773	0.34159	<0.0001	4.39974	0.33624	<0.0001	4.17739	0.33833	<0.0001	1.72831	0.22151	<0.0001
chronic	$\hat{\beta}_{2,2}$	0.13239	0.01543	<0.0001	0.10547	0.01526	<0.0001	0.17726	0.01604	<0.0001	0.11340	0.01546	<0.0001
age	$\hat{\beta}_{2,3}$	-0.46741	0.04213	<0.0001	-0.48008	0.04162	<0.0001	-0.46219	0.04280	<0.0001	-0.13618	0.02370	<0.0001
medicaid	$\hat{\beta}_4$	0.13673	0.07771	0.07849	0.12727	0.08012	0.11217	0.20221	0.07960	0.01108	0.14498	0.07812	0.06346
income	$\hat{\beta}_{2,5}$	-0.01200	0.00948	0.20546	-0.01008	0.00941	0.28404	-0.03434	0.00989	0.00052	-0.00979	0.00929	0.29191
school	$\hat{\beta}_{2,6}$	-0.10068	0.00666	<0.0001	-0.10383	0.00668	<0.0001	-0.10556	0.00664	<0.0001	-0.09136	0.00657	<0.0001

Tableau 3.10 – Résultats du modèle de régression ZIBP avec 30% données manquantes.

3.7 Conclusion

Nous avons proposé deux méthodes d'estimation IPW et deux méthodes de type MI pour estimer les paramètres d'un modèle de régression ZIBP lorsque des covariables sont MAR. L'analyse théorique et les résultats des simulations concluent que les estimateurs IPW paramétrique et semi-paramétrique puis les méthodes de type MI proposés sont efficaces. De plus, il a été aussi prouvé que les deux estimateurs de type MI sont asymptotiquement équivalents à l'estimateur IPW semi-paramétrique. Outre les preuves analytiques de l'équivalence asymptotique, nous avons fourni des preuves numériques par le biais d'une étude de simulation approfondie. Enfin, une application des méthodes proposées sur un jeu de données issu d'une étude de santé publique américaine confirme la robustesse des méthodes proposées en présence de données manquantes.

Conclusion et perspectives

Dans ce mémoire de thèse, nous nous sommes intéressés au problème de l'inférence statistique dans des modèles de Poisson bivarié à inflation de zéros. Ce travail de recherche s'articule autour de deux contributions.

Dans la première contribution, nous nous sommes particulièrement intéressé au modèle de régression de Poisson bivarié à inflation de zéros en étudiant d'abord, les propriétés asymptotiques de son estimateur du maximum de vraisemblance théoriquement. En outre, nous avons mené une étude de simulations sur plusieurs échantillons de tailles finies pour évaluer les performances de l'estimateur proposé. Ces résultats ont confirmé les bonnes propriétés de l'estimateur du maximum de vraisemblance dans le modèle ZIBP. Pour finir, une application du modèle sur un jeu de données recensant l'utilisation des services médicaux de plusieurs milliers de patients aux USA a été réalisé.

L'estimateur du maximum de vraisemblance, ne peut pas être utilisé lorsque des données qui interviennent dans la régression sont manquantes sur des covariables. Nous nous sommes donc intéressés au problème de l'inférence statistique dans le modèle ZIBP en présence de données manquantes sur les covariables dans notre seconde contribution. Dans ce contexte, nous avons proposé quatre méthodes pour estimer les paramètres du modèle de régression ZIBP avec des covariables manquantes au hasard. En supposant que la probabilité de sélection est inconnue et estimée de manière paramétrique et non-paramétrique (par un estimateur à noyau). Nous avons proposé des méthodes de pondération par l'inverse des probabilités de sélection (IPW) et d'imputation multiple (MI) pour estimer les paramètres du modèle de régression ZIBP avec des covariables manquantes au hasard. Les propriétés asymptotiques des estimateurs proposés ont été étudiées théoriquement. Puis, nous avons réalisé des

simulations pour évaluer les performances des méthodes proposées. Ces résultats numériques sont cohérents avec nos résultats théoriques. Enfin, une application des méthodes proposées sur un jeu de données issu d'une étude de santé publique américaine confirme la robustesse des méthodes proposées en présence de données manquantes.

Au terme de ces travaux, plusieurs thématiques de recherche sont à envisager. On pourra évaluer les propriétés d'autres estimateurs des paramètres du modèle ZIBP (par exemple, robust expectation-solution (RES)) plus robuste que l'estimateur du maximum de vraisemblance. Ensuite, proposer une étude du modèle ZIBP avec des effets aléatoires pour prendre en compte la corrélation qui peut exister entre les réponses observées chez les différents patients (due, par exemple, à des effets géographiques ou des clusters familiaux). La construction des estimateurs des paramètres de ce nouveau modèle et l'étude de leurs propriétés (théoriquement et/ou au moyen d'études de simulation) peuvent être abordées. Pour prendre en compte la présence de données répétées, nous pourrons introduire une dimension longitudinale dans le modèle ZIBP. On pourra proposer une méthode d'estimation adaptée à ce contexte qu'on évaluera théoriquement et/ou numériquement. Un test d'ajustement au modèle de régression de Poisson bivarié à inflation de zéros (ZIBP) avec des données manquantes peut aussi être envisagé. L'ensemble de ces travaux peut être étendu aux données de comptage de dimension supérieure à deux. Toutes ces questions pourront être abordées dans les travaux futurs.

Bibliographie

AGRESTI, A. (2015). Foundations of Linear and Generalized Linear Models. page 472.

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

ALI, E. (2021). *Modèles de régression marginaux pour des données de comptage à excès de zéros*. Thèse de doctorat, IRMAR-INSA de Rennes & LERSTAD-UGB de Saint-Louis.

ALI, E. (2022). A simulation-based study of zip regression with various zero-inflated submodels. *Communications in Statistics - Simulation and Computation*, 0(0):1–16.

ALI, E., DIOP, A. et DUPUY, J.-F. (2020). A constrained marginal zero-inflated binomial regression model. *Communications in Statistics - Theory and Methods*, 0(0):1–30.

ALMUHAYFITH, F. E., ABDULHAMID, A. A. et OMAIR, M. A. (2016). On bivariate poisson regression models. *Journal of King Saud University - Science*, 28(2):178–189.

BERMÚDEZ, L. (2009). A priori ratemaking using bivariate poisson regression models. *Insurance: Mathematics and Economics*, 44:135–141.

CARROLL, R., RUPPERT, D., STEFANSKI, L. et CRAINICEANU, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

CLAYTON, D., SPIEGELHALTER, D., DUNN, G. et PICKLES, A. (1998). Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society Series B*, 60(1):71–87.

CREEMERS, A., AERTS, M., HENS, N. et MOLENBERGHS, G. (2012). A nonparametric approach to weighted estimating equations for regression analysis with missing covariates. *Computational Statistics & Data Analysis*, 56:100–113.

CZADO, C., ERHARDT, V., MIN, A. et WAGNER, S. (2007). Zero-inflated generalized poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. *Statistical Modelling*, 7(2):125–153.

CZADO, C. et MIN, A. (2005). Consistency and asymptotic normality of the maximum likelihood estimator in a zero-inflated generalized Poisson regression. page 29.

DEB, P. et TRIVEDI, P. (1997). Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics*, 12:313–36.

DEMPSTER, A. P., LAIRD, N. M. et RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

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

DIALLO, A. O., DIOP, A. et DUPUY, J.-F. (2017). Asymptotic properties of the maximum-likelihood estimator in zero-inflated binomial regression. *Communications in Statistics - Theory and Methods*, 46(20):9930–9948.

DIALLO, A. O., DIOP, A. et DUPUY, J.-F. (2018). Analysis of multinomial counts with joint zero-inflation, with an application to health economics. *Journal of Statistical Planning and Inference*, 194:85–105.

DIALLO, A. O., DIOP, A. et DUPUY, J.-F. (2019). Estimation in zero-inflated binomial regression with missing covariates. *Statistics*, 53:1–27.

DIOP, A., DIOP, A. et DUPUY, J.-F. (2011). Maximum likelihood estimation in the logistic regression model with a cure fraction. *Electronic Journal of Statistics*, 5(none):460 – 483.

DOBSON, A. J. et BARNETT, A. G. (2018). *An Introduction to Generalized Linear Models, Fourth Edition*. Chapman and Hall/CRC.

DUPUY, J.-F. (2018). *Statistical Methods for Overdispersed Count Data*, volume 4 de *Biostatistique et sciences de la santé*.

FAHRMEIR, L. et KAUFMANN, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1):342–368.

FAMOYE, F. et SINGH, K. (2006). Zero-inflated generalized poisson regression model with an application to domestic violence data. *Journal of Data Science*, 4:117–130.

FAROUGHI, P. et ISMAIL, N. (2017). Bivariate zero-inflated negative binomial regression model with applications. *Journal of Statistical Computation and Simulation*, 87(3):457–477.

FAY, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91(434):490–498.

FENG, X. C. (2021). A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *Journal of Statistical Distributions and Applications*, 8:1–19.

FOUTZ, R. V. (1977). On the Unique Consistent Solution to the Likelihood Equations. *Journal of the American Statistical Association*, 72(357):147–148.

GARAY, A. M., HASHIMOTO, E. M., ORTEGA, E. M. et LACHOS, V. H. (2011). On estimation and influence diagnostics for zero-inflated negative binomial regression models. *Computational Statistics & Data Analysis*, 55(3):1304–1318.

GURMU, S. et ELDER, J. (2000). Generalized bivariate count data regression models. *Economics Letters*, 68(1):31–36.

HALL, D. B. (2000). Zero-inflated poisson and binomial regression with random effects: a case study. *Biometrics*, 56(4).

HALL, D. B. et BERENHAUT, K. S. (2002). Score tests for heterogeneity and overdispersion in zero-inflated poisson and binomial regression models. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 30(3):415–430.

HALL, D. B. et SHEN, J. (2010). Robust estimation for zero-inflated poisson regression. *Scandinavian Journal of Statistics*, 37(2):237–252.

HE, X., XUE, H. et SHI, N.-Z. (2010). Sieve maximum likelihood estimation for doubly semiparametric zero-inflated poisson models. *Journal of Multivariate Analysis*, 101:2026–2038.

HEILBRON, D. C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, 36:531–547.

HENNINGSEN, A. et TOOMET, O. (2011). Maxlik: A package for maximum likelihood estimation in r. *Computational Statistics*, 26:443–458.

HÉRAUD, Bousquet, V. (2012). *Traitement des données manquantes en épidémiologie : application de l'imputation multiple à des données de surveillance et d'enquêtes*. Theses, Université Paris Sud - Paris XI.

HILBE, J. M. (2007). *Negative Binomial Regression*. Cambridge University Press.

HOLGATE, P. (1964). Estimation for the bivariate Poisson distribution. *Biometrika*, 51(1-2):241–287.

HORVITZ, D. G. et THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

JENNRICH, R. I. et SAMPSON, P. F. (1976). Newton-raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics*, 18(1):11–17.

JOHNSON, N. et KOTZ, S. (1969). *Discrete Distributions*. Numéro vol. 1 de Discrete Distributions. Houghton Mifflin.

KARLIS, D. et NTZOUFRAS, I. (2003). Analysis of sports data by using bivariate Poisson models. *J Royal Statistical Soc D*, 52(3):381–393.

KEMP, C. D. et KEMP, A. W. (1988). Rapid estimation for discrete distributions. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 37(3):243–255.

LAM, K., XUE, H. et CHEUNG, Y. (2007). Semiparametric analysis of zero-inflated count data. *Biometrics*, 62:996–1003.

LAMBERT, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, 34(1):1.

LANGE, K. (2004). Computational statistics and optimization theory at ucla. *The American Statistician*, 58(1):9–11.

LEE, s.-M., HUNG, P. K. et LI, C.-S. (2021). Validation likelihood estimation method for a zero-inflated bernoulli regression model with missing covariates. *Journal of Statistical Planning and Inference*, 214.

LEE, S.-M., HWANG, W.-H. et TAPSOBA, J. d. D. (2016). Estimation in closed capture-recapture models when covariates are missing at random. *Biometrics*, 72(4):1294–1304.

LEE, S.-M., LUKUSA, T. M. et LI, C.-S. (2020). Estimation of a zero-inflated Poisson regression model with missing covariates via nonparametric multiple imputation methods. 35.

LI, C.-S. (2011). A lack-of-fit test for parametric zero-inflated poisson models. *Journal of Statistical Computation and Simulation*, 81(9):1081–1098.

LI, C.-S., LU, J.-C., PARK, J., KIM, K., BRINKLEY, P. A. et PETERSON, J. P. (1999). Multivariate Zero-Inflated Poisson Models and Their Applications. *Technometrics*, 41(1):29–38.

LIM, H. K., LI, W. K. et YU, P. L. (2014). Zero-inflated poisson regression mixture model. *Computational Statistics & Data Analysis*, 71:151–158.

LITTLE, R. et RUBIN, D. (1987). *Statistical Analysis With Missing Data*. Wiley Series in Probability and Statistics. Wiley.

LITTLE, R. J. et RUBIN, D. B. (2002). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

LITTLE, R. J. A. (1992). Regression with missing x's: A review. *Journal of the American Statistical Association*, 87(420):1227–1237.

LUKUSA, M. T. et PHOA, F. K. H. (2020). A note on the weighting-type estimations of the zero-inflated Poisson regression model with missing data in covariates. *Statistics & Probability Letters*, 158.

LUKUSA, T. M., LEE, S.-M. et LI, C.-S. (2016). Semiparametric estimation of a zero-inflated Poisson regression model with missing covariates. *Metrika*, 79(4):457–483.

MCCULLAGH, P. et NELDER, J. (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

MOGHIMBEIGI, A., ESHRAGHIAN, M. R., MOHAMMAD, K. et MCARDLE, B. (2008). Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. *Journal of Applied Statistics*, 35(10):1193–1202.

MONOD, A. (2014). Random effects modeling and the zero-inflated poisson distribution. *Communications in Statistics - Theory and Methods*, 43(4):664–680.

MULLAHY, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365.

MULLAHY, J. (1997). Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics*, 12(3):337–350.

MWALILI, S., LESAFFRE, E. et DECLERCK, D. (2008). The zero-inflated negative binomial regression model with correction for misclassification: An example in caries research. *Statistical methods in medical research*, 17:123–39.

NELDER, J. A. et WEDDERBURN, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

NEWEY, W. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, 59:1161–67.

NGUYEN, V. et DUPUY, J.-F. (2019). Asymptotic results in censored zero-inflated poisson regression. *Communications in Statistics - Theory and Methods*, 50:1–21.

NTZOUFRAS, I. et KARLIS, D. (2005). Bivariate poisson and diagonal inflated bivariate poisson regression models in r. *Journal of Statistical Software*, 14.

PREISSER, J., STAMM, J., LONG, D. L. et KINCADE, M. (2012). Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. *Caries research*, 46:413–23.

REILLY, M. et PEPE, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82(2):299–314.

RIDOUT, M., DEMÉTRIO, C. et HINLE, J. (1998). Models for count data with many zeros. international biomeric conference. *Cape Town*, 13:1–13.

RIDOUT, M., HINDE, J. et DEMÉTRIO, C. (2001). A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, 57:219–23.

ROBINS, J. M., ROTNITZKY, A. et ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

ROSEN, O., JIANG, W. et TANNER, M. A. (2000). Mixtures of marginal models. *Biometrika*, 87(2):391–404.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization.

RUBIN, D. B. (2002). Multiple imputations in sample surveys: a phenomenological bayesian approach to nonresponse.

RUBIN, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. Wiley Classics Library. Wiley.

RUBIN, D. B. et SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81(394):366–374.

SARI, N. (2009). Physical inactivity and its impact on healthcare utilization. *Health economics*, 18:885–901.

SARMA, S. et SIMPSON, W. (2006). A microeconometric analysis of canadian health care utilization. *Health economics*, 15:219–39.

SCHAFER, J. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

SCHAFER, J. et GRAHAM, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7:147–177.

SCHWARZ, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464.

SEAMAN, S. et WHITE, I. (2013). Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res.*, 22(3):278–295.

TANNER, M. A. et WONG, W. H. (1987). An application of imputation to an estimation problem in grouped lifetime analysis. *Technometrics*, 29(1):23–32.

TSIATIS, A. A. (2006). *Semiparametric theory and missing data*. Springer series in statistics. Springer, New York.

van BUUREN, S. (2012). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Press, Boca Raton, FL.

VIEIRA, A. M. C., HINDE, J. P. et DEMETRIO, C. G. B. (2000). Zero-inflated proportion data models applied to a biological control assay. *Journal of Applied Statistics*, 27(3):373–389.

WANG, B., GU, H. et QIN, P. (2021). Consistency and asymptotic normality of the maximum likelihood estimator in gaglm.

WANG, C. Y., WANG, S., ZHAO, L.-P. et OU, S.-T. (1997). Weighted semiparametric estimation in regression analysis with missing covariate data. *Journal of the American Statistical Association*, 92(438):512–525.

WANG, D. et CHEN, S. X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics*, 37(1):490 – 517.

WANG, K., LEE, A. H., YAU, K. K. et CARRIVICK, P. J. (2003). A bivariate zero-inflated poisson regression model to analyze occupational injuries. *Accident Analysis & Prevention*, 35(4):625–629.

WANG, P. (2003). A bivariate zero-inflated negative binomial regression model for count data with excess zeros. *Economics Letters*, 78(3):373–378.

WANG, S. et WANG, C.-Y. (2001). A note on kernel assisted estimators in missing covariate regression. *Statistics & Probability Letters*, 55:439–449.

YANG, M., DAS, K. et MAJUMDAR, A. (2016). Analysis of bivariate zero inflated count data with missing responses. *Journal of Multivariate Analysis*, 148:73–82.

ZHAO, L. et LIPSITZ, S. (1992). Designs and analysis of two-stage studies. *Stat Med.*, 11(6):769–782.

4

Communications écrites et orales

Travaux et publications

- KOUAKOU KONAN JEAN GEOFFROY, HILI OUAGNINA, DUPUY JEAN-FRANÇOIS. Estimation in the zero-inflated bivariate Poisson model, with an application to health-care utilization data. *Afrika Statistika 16*, 2 (2021), 2767 – 2788. URL https://doi.org/10.16929/as/2021.2767.183.
- KOUAKOU KONAN JEAN GEOFFROY, HILI OUAGNINA, DUPUY JEAN-FRANÇOIS. Estimation of zero-inflated bivariate Poisson regression with missing covariates. *Communications in Statistics - Theory and Methods*, en révision.
- KOUAKOU, KONAN JEAN GEOFFROY, HILI OUAGNINA, DUPUY JEAN-FRANÇOIS. Goodness-of-fit tests for a zero-inflated bivariate Poisson model with missing covariates, *en élaboration*.

Séminaires et Conférences

- KOUAKOU KONAN JEAN GEOFFROY, HILI OUAGNINA, DUPUY JEAN-FRANÇOIS. Estimation in the zero-inflated bivariate Poisson model, with an application to health-care utilization data. *3rd LmB Conference on Multivariate Statistical Models : Count and Semi-Continuous*. Besançon (France) du 06 au 08 Juillet 2022.
- Depuis Juillet 2018, Communications annuelles aux doctoriales de l'INP-HB.
- Séminaire Interne au sein de l'UMRI (Unité Mixte de Recherche et d'Innovation) MNTI (Mathématiques et Nouvelles Technologies de l'Information) de l'EDP-INPHB.