Modèles de régression pour données de comptage zéro-inflatées en présence de censure. Applications en économie de la santé

Van-Trinh Nguyen

UNIVERSITE BRETAGNE LOIRE MATHSTIC

INSA RENNES

THESE DE DOCTORAT DE

L'INSA RENNES
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Mathématiques et leurs Interactions*

Par

Van-Trinh NGUYEN

Modèles de régression pour données de comptage zéro-inflatées en présence de censure. Applications en économie de la santé.

Thèse présentée et soutenue à l'Amphi Bonnin de l'INSA Rennes, le 16 Mars 2020
Unité de recherche : IRMAR-UMR CNRS 6625
Thèse N° : 20ISAR 03 / D20 - 03

Rapporteurs avant soutenance :

Sophie DABO-NIANG PR, Université de Lille
Anne-Françoise YAO PR, Université Clermont Auvergne

Composition du Jury :

Sophie DABO-NIANG PR, Université de Lille
Jean-François DUPUY PR, INSA de Rennes
Valérie GARES MCF, INSA de Rennes
Christophe RAULT PR, Université d'Orléans
Anne-Françoise YAO PR, Université Clermont Auvergne

Directeur de thèse
Jean-François DUPUY PR, INSA de Rennes

Intitulé de la thèse :

Modèles de régression pour données de comptage zéro-inflatées en présence de censure. Applications en économie de la santé.

Van-Trinh NGUYEN

En partenariat avec :

Acknowledgements

"I must really admit that someone's special behavior is always in my mind and never to forget. On Monday morning in Hanoi summer, the 20/7/2015, at Vietnam Institute for Advanced Study in Mathematics (VIASM), a man who wore **the black T-shirt, blue jean and a small tattoo** on hand with the warm voice giving the lecture, he came close to me and helped to handle a `R code` **while kneeling on the ground**. Those moment and action were so different from my country and made a very good impression in my mind. It was reason why I decided to study in Ph.D. level with him. That person is Jean-François".

This dissertation would not have been completed without the help, encouragement and support from many people who all deserve my sincerest gratitude and appreciation.

In fact, I am really sure that no words may express my gratitude to my doctoral advisor Professor Jean-François DUPUY who has been close to me since the first time I came to French so far. He always shares with me his valuable advice, discussions, and support me during the darkest moments of my Ph.D. I will never forget, mostly the time when he was working abroad, he has usually had emailed me with many suggestions about our problems and discussed them in details. To be honest, during my first year, I always felt nervous before every meeting with him, however throughout the year, I found it more and more interesting challenging thanks to which I have learned and matured a lot. I am, always have been, and will always be, proud to be his student.

I greatly appreciate the Professors Sophie Dabo-Niang and Anne-Françoise Yao, the two reviewers of this manuscript, who have been so kind to spend time to read, evaluate my work and give many valuable comments. Those are indeed beneficial useful for my improvement in future research. It is my pleasure for giving my thanks to Professor Christophe Rault and assistant Professor Valérie Gares who accepted to be jury members in my thesis.

I would like to thank to Martine Fixot, Patricia Soufflet, Aurore Gouin and Claire Durand for all their help and support in administrative procedures. A special mention goes to Mr. director in science of laboratory Professor Marc Briane, who agreed to help the complement scholarship for several months during my working time.

I am grateful to all members in Department of Mathematics at INSA Rennes with their friendly and respective behavior. Thank you to Prof. Olivier Ley, Prof. Mounir Haddou, MCF Mohamed Camar-Eddine with their interesting talks and care.

I am truly thankful to my mathematical friend (also be younger brother) Manh-Khang Dao who shared and discussed valuable knowledge in fundamental mathematics, free talks about anythings in this world. A special thanks to my close-and-young friends Minh-Phuong, Thanh-Huan with many funny stories, many encouragements and many minor parties are made ourselves. At the moment that I came to Rennes, take my first friends into account Tuan-Anh Nguyen, Viet-Phuong Nguyen, Thi-Tuyen Nguyen and Ngoc-Minh Hoa with their initial-very-useful help.

Besides, I have also been lucky to meet and make friends with many nice people in Rennes. Thank you, my French, Senegal, Morocco, Ton go, Tunisia, Iranian, Algerian, Italian friends

who share experience in Ph.D. life and friendly talks: Tangi Migot, Emilie Joannopoulos, Audrey Poterie, Sammuel Corre, Alpha-Oumar Diallo, Othman Touijer, Essoham Ali, Alexandre Rantin Mazarin, Mériadec, Maryam, El Hassene, Lorenza, especially, Bilel Bousselmi, the closest-laboratory friend with lots of interesting stories, everyday conversations and mathematical discussions. To be honest, thank you all the Vietnamese friends at the big home Mirabeau residence: The Anh, Dieu Thao, Lan Anh, Thi Diep, Dinh Anh, Hoai Thuong, Duong-Trang, Luan-HongAnh, Tho-Hao with much physical support and what we done together, they made me sometimes I forget that I am living in France.

Last but not least, I would like to express all my gratitude to my family for constant encouragement and unconditional love, my parents, my brother, my younger sisters, all my friends in Vietnam as well. Especially, I dedicate this thesis to my lovely wife **Bich-Ngoc Luu**, who has been growing our two children up alone during past three years as well as has been handled a mountain of works that I have been absent home. This is also the gift for two-my-great-lovely children **Ngoc-Mai Nguyen** and **Tuan-Dung Nguyen**.

INSA-Rennes, March 2020

Van-Trinh NGUYEN

Abstract

Count data with excess of zeros commonly occur in various areas, such as ecology, epidemiology, insurance, or health economics. Zero-inflated regression models provide a useful tool to analyse such data and have given rise to an abundant literature. A further problem arises when these counts are right-censored, which also often occurs in observational studies. Right-censoring refers to the situation where one only observes a lower bound on the count of interest. So far, this setting has however attracted much less attention than uncensored zero-inflated models. This dissertation aims at investigating - both theoretically and numerically - the statistical inference in zero-inflated count regression models with right-censored data.

In the first contribution of this thesis, we investigate the theoretical and numerical properties of the maximum likelihood estimator (MLE) in the zero-inflated Poisson (ZIP) regression model with randomly right-censored counts. We establish rigorously the consistency and asymptotic normality of the MLE in this setting. A thorough simulation study is also conducted to assess finite-sample behavior of the MLE. Finally, we describe an application to a real dataset from health economy.

The ZIP model assumes that the observed overdispersion is entirely caused by zero-inflation. When additional overdispersion is present, useful alternatives to ZIP are given by the zero-inflated generalized Poisson (ZIGP) and zero-inflated negative binomial (ZINB) models. In a second contribution, we investigate the properties of the MLE in ZIGP and ZINB regression models, when the count response is subject to right-censoring. Simulations are used to examine performance (bias, mean square error, coverage probabilities and standard error calculations) of the estimates. Then, we consider the issue of variable selection. A simple, efficient and easy-to-implement methodology for variable selection is proposed. It is applicable even when the number of predictors is very large, and it yields interpretable and sound results. The proposed methods are again applied to a dataset in the field of health economy.

In a third original contribution, we investigate the statistical inference in the marginal zero-inflated Poisson (MZIP) regression model, when the count response is randomly right-censored. Contrary to the ZIP model, the marginal MZIP model directly describes the effect of the covariates on the overall population mean. For this reason, it provides an appealing alternative for interpreting covariate effects at the population level. We establish the asymptotic properties (consistency and asymptotic normality) of the MLE in the censored MZIP model and we conduct a simulation study to assess finite-sample properties. An application to a dataset in the field of healthcare demand is described.

Key words: Asymptotic properties, count data, excess of zeros, marginal model, simulations, health care utilization.

Résumé

L'excès de zéros - ou inflation de zéros - dans les données de comptage est une situation qui survient dans de nombreux domaines, tels que l'écologie, l'épidémiologie, l'assurance ou l'économie de la santé. Les modèles de régression à inflation de zéro fournissent un outil puissant pour analyser ce type de données, et ils ont à ce jour suscité une abondante littérature. Un problème supplémentaire est celui de la censure de la variable réponse. La censure à droite, en particulier, correspond à la situation où l'on observe seulement une borne inférieure sur le comptage considéré. L'analyse statistique de données de comptage en présence d'inflation de zéros et de censure n'a jusqu'à présent suscité que peu de travaux. Ce sujet constitue le thème général de ce travail de thèse.

Dans une première contribution, nous étudions les propriétés théoriques et numériques de l'estimateur du maximum de vraisemblance (EMV) dans le modèle de régression de Poisson à inflation de zéros (modèle ZIP). Nous prouvons la consistance et la normalité asymptotique de l'EMV et nous réalisons une étude de simulation approfondie, afin d'évaluer ses propriétés numériques dans des échantillons de taille finie. Nous décrivons enfin une application du modèle ZIP censuré sur des données réelles issues du domaine de l'économie de la santé.

Le modèle ZIP suppose que toute la surdispersion des données peut être expliquée par un excès de zéros. Lorsqu'une surdispersion supplémentaire est présente (due, par exemple, à une hétérogénéité inobservée des individus), on peut utiliser des modèles alternatifs, tels que les modèles de régression de Poisson généralisé à inflation de zéros (modèle ZIGP) et binomial négatif à inflation de zéros (modèle ZINB). Dans une deuxième contribution de ce travail, nous nous intéressons aux propriétés de l'EMV dans les modèles ZIGP et ZINB en présence d'une variable réponse censurée. Nous étudions ces propriétés au moyen de simulations. Puis nous nous intéressons à la question de la sélection de variables dans ces modèles et proposons une procédure de sélection simple à mettre en oeuvre. Nous illustrons cette méthodologie sur des données réelles issues du domaine de l'économie de la santé.

Dans une troisième contribution, nous nous intéressons à l'inférence statistique dans le modèle de régression ZIP marginal (modèle MZIP), en présence d'une variable réponse censurée. Contrairement au modèle ZIP, le modèle MZIP quantifie l'effet des variables explicatives directement sur la moyenne de la loi marginale du comptage (et non sur la moyenne de la loi du comptage pour les individus "susceptibles"). Nous établissons les propriétés asymptotiques de l'EMV dans ce cadre et réalisons une étude de simulations. Enfin, nous illustrons le modèle sur un jeu de données réelles.

Mots clés: Propriétés asymptotiques, données de comptage, excès de zéros, modèle marginal, simulations, applications en économie de la santé.

Notations

$\mathbb{P}(A)$	:	Probability of the event A
$\mathbb{E}(X)$	:	Mathematical expectation of a random variable X
$\text{var}(X)$	:	Variance of a random variable X
$\text{cov}(X, Y)$	:	Covariance of the two random variables X and Y
$X_n \xrightarrow{\mathbb{P}} X$	:	The sequence of random variables X_n converges in probability to X
$X_n \xrightarrow{\mathcal{D}} X$	:	The sequence of random variables X_n converges in distribution to X
$X_n \xrightarrow{a.s.} X$	:	The sequence of random variables X_n converges almost surely to X
$\mathbb{N}$	:	Set of positive integers
$\mathbb{N}^*$	:	Set of positive natural numbers
$\mathbb{R}$	:	Set of real numbers
$X^\top$	:	Transpose of X
i.i.d.	:	independent and identically distributed
GLM	:	Generalized linear models
ZI	:	Zero-inflated
ZIP	:	Zero-inflated Poisson
ZIGP	:	Zero-inflated generalized Poisson
ZINB	:	Zero-inflated negative binomial
MZIP	:	Marginal zero-inflated Poisson
MLE	:	Maximum likelihood estimator

List of Figures

3.1 Normal Q-Q plots for $\hat{\beta}_{1,n}, \ldots, \hat{\beta}_{6,n}$ with $n = 500, c = 0.4$ and a proportion of zero-inflation equal to 0.4. 43
3.2 Normal Q-Q plots for $\hat{\gamma}_{1,n}, \ldots, \hat{\gamma}_{5,n}$ with $n = 500, c = 0.4$ and a proportion of zero-inflation equal to 0.4. 44
3.3 Histogram of the normalized estimates $(\widehat{\beta}_{j,n} - \beta_j)/\text{s.e.}(\widehat{\beta}_{j,n})$, $j = 1, \ldots, 6$ with $n = 500, c = 0.4$ and a proportion of zero-inflation equal to 0.4. 45
3.4 Histogram of the normalized estimates $(\widehat{\gamma}_{j,n} - \gamma_j)/\text{s.e.}(\widehat{\gamma}_{j,n})$, $j = 1, \ldots, 5$ with $n = 500, c = 0.4$ and a proportion of zero-inflation equal to 0.4. 46

4.1 Histograms of the normalized estimates $(\widehat{\beta}_{j,n} - \beta_j)/\text{s.e.}(\widehat{\beta}_{j,n})$, $j = 1, \ldots, 6$ in censored ZIGP model (30% censoring). 63
4.2 Histograms of the normalized estimates $(\widehat{\gamma}_{j,n} - \gamma_j)/\text{s.e.}(\widehat{\gamma}_{j,n})$, $j = 1, \ldots, 5$ and $(\widehat{\varphi}_n - \varphi)/\text{s.e.}(\widehat{\varphi}_n)$ in censored ZIGP model (30% censoring). 64
4.3 Histograms of the normalized estimates $(\widehat{\beta}_{j,n} - \beta_j)/\text{s.e.}(\widehat{\beta}_{j,n})$, $j = 1, \ldots, 6$ in censored ZINB model (30% censoring). 65
4.4 Histograms of the normalized estimates $(\widehat{\gamma}_{j,n} - \gamma_j)/\text{s.e.}(\widehat{\gamma}_{j,n})$, $j = 1, \ldots, 5$ and $(\widehat{\alpha}_n - \alpha)/\text{s.e.}(\widehat{\alpha}_n)$ in censored ZINB model (30% censoring). 66
4.5 Number of doctor office visits. 67

5.1 Histograms of the normalized estimates $(\widehat{\beta}_{j,n} - \beta_j)/\text{s.e.}(\widehat{\beta}_{j,n})$, $j = 1, \ldots, 6$ in censored MZIP model (n=500, ZI=40%, c=40% censoring). 100
5.2 Histograms of the normalized estimates $(\widehat{\gamma}_{j,n} - \gamma_j)/\text{s.e.}(\widehat{\gamma}_{j,n})$, $j = 1, \ldots, 5$ and $(\widehat{\varphi}_n - \varphi)/\text{s.e.}(\widehat{\varphi}_n)$ in censored MZIP model (n=500, ZI=40%, c=40% censoring). 101
5.3 Normal Q-Q plots for $\hat{\beta}_{1,n}, \ldots, \hat{\beta}_{6,n}$ with $n = 500, c = 0.4$ and a proportion of zero-inflation equal to 0.4. 102
5.4 Normal Q-Q plots for $\hat{\gamma}_{1,n}, \ldots, \hat{\gamma}_{5,n}$ with $n = 500, c = 0.4$ and a proportion of zero-inflation equal to 0.4. 103
5.5 Number of doctor office visits. 104

List of Tables

1.1 Quelques examples de familles distributions exponentielles 2
1.2 Quelques fonctions de lien classiques 3

3.1 Health-care data analysis: estimates (standard errors) and significance codes: *** significant at the 0.1% level, ** significant at the 1% level, * significant at the 5% level, . significant at the 10% level. 35
3.2 Simulation results ($n = 500$, ZI proportion = 20%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals. 37
3.3 Simulation results ($n = 500$, ZI proportion = 40%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals. 38
3.4 Simulation results ($n = 1000$, ZI proportion = 20%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals. 39
3.5 Simulation results ($n = 1000$, ZI proportion = 40%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals. 40
3.6 Simulation results ($n = 2500$, ZI proportion = 20%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals. 41
3.7 Simulation results ($n = 2500$, ZI proportion = 40%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals. 42

4.1 Simulation results for ZIGP model. SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals. 59
4.2 Simulation results for ZINB model. SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals. 60
4.3 Summary of final censored ZIP, ZIGP and ZINB models († although not significant, `age40` remains in the model because of a significant interaction). 61

4.4 Model comparison using Vuong test: Vuong statistic, p-value and test decision (i.e., the best model according to Vuong test). 62

5.1 Simulation results ($n = 500$, ZI proportion = 20%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals. 83
5.2 Simulation results ($n = 1000$, ZI proportion = 20%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals. 84
5.3 Simulation results ($n = 2000$, ZI proportion = 20%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals. 85
5.4 Simulation results ($n = 500$, ZI proportion = 40%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals. 86
5.5 Simulation results ($n = 1000$, ZI proportion = 40%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals. 87
5.6 Simulation results ($n = 2000$, ZI proportion = 40%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals. 88
5.7 Health-care data analysis: estimates (standard errors) and p-value. 90

General Introduction

Zero-inflated regression models are widely used to accommodate count data with excess zeros. Classical zero-inflated models include the zero-inflated Poisson (ZIP) model (Lambert, 1992), zero-inflated binomial (ZIB) model (Hall, 2000; Diallo et al., 2017a), zero-inflated generalized Poisson (ZIGP) model (Famoye and Singh, 2003, 2006). Since their introduction, these models have been extended to accommodate non-linear covariate effects, missing data, cluster effect... and have been applied in a variety of settings, such as ecology, epidemiology, quantitative analysis of international relations, insurance, health economy... In this dissertation, we consider zero-inflated regression models when the count response of interest is prone to right-censoring. This problem has attracted little attention so far.

This work develops in several directions, with theoretical, methodological, numerical and applied contributions. The applications described here are concerned with real datasets in health economy. More precisely, we consider several datasets investigating healthcare demand. A first dataset arises from the National Medical Expenditure Survey (see Deb and K. Trivedi, 1997), a large study which was conducted in USA in 1987-1988 to assess the demand for medical care by elderlies. Data are available in the `R` package `AER` (see Christian Kleiber, 2008) under the name "NMES1998" and contain observations on 4,406 individuals aged 66 and over, all of whom are covered by Medicare, a federal health insurance program. Another dataset, also dealing with the demand for medical care, comes from the German Socioeconomic Panel, a large panel study which was carried out on 1812 West German men aged 25-65 years, and was reported by Jochmann (2009).

The manuscript is organized as follows. In chapter 1, we provide some background on generalized linear regression models, which are the framework for count data modeling. Some background on zero-inflated regression models is provided in Chapter 2. In the next three chapters, we describe the original contributions of this work. One first objective is to investigate, both theoretically and numerically, the properties of the maximum likelihood estimator in several zero-inflated regression models for right-censored count data with excess zeros. This is described in Chapter 3 for the zero-inflated Poisson (ZIP) model and in Chapter 5 for the marginal zero-inflated Poisson (MZIP) model. These mathematical results are supplemented by comprehensive simulation results, which aim at assessing the finite-sample properties of the maximum likelihood estimate. In Chapter 4, we conduct a methodological work, carrying out simulations to investigate properties of the maximum likelihood estimate in ZIGP and ZINB models in the case of right-censored count data. We also propose a procedure for variable selection, elaborating on the stepwise selection approach. Perspectives are provided within each chapter, along with the applications.

Contents

Acknowledgements i

Abstract iii

Résumé v

Notations vii

List of Figures ix

List of Tables xi

General introduction xiii

1 Modèles linéaires généralisés 1
- 1.1 Introduction . . . 1
- 1.2 Théorie des modèles linéaires généralisés . . . 1
  - 1.2.1 Spécification d'un modèle linéaire généralisé . . . 1
  - 1.2.2 Espérance et variance des modèles linéaires généralisés . . . 3
  - 1.2.3 Estimation . . . 4
  - 1.2.4 Propriétés asymptotiques et inférence . . . 6
  - 1.2.5 Algorithme de Newton-Raphson . . . 8
  - 1.2.6 Estimation du paramètre de dispersion . . . 9
- 1.3 Choix de modèles . . . 10

2 Modèles de régression à inflation de zéros 13
- 2.1 Introduction . . . 13
  - 2.1.1 Modèles à inflation de zéros . . . 13
  - 2.1.2 Modèles de régression à inflation de zéros . . . 15
- 2.2 Modèle de régression ZIP . . . 16
  - 2.2.1 Définition . . . 16
  - 2.2.2 Estimation dans le modèle ZIP . . . 17
- 2.3 Modèle de régression ZINB . . . 17
  - 2.3.1 Définition . . . 17
  - 2.3.2 Estimation . . . 18
- 2.4 Modèle de régression ZIGP . . . 19
  - 2.4.1 Définition . . . 19
  - 2.4.2 Estimation . . . 21

3 Zero-inflated Poisson regression model with right censored data 23
- 3.1 Introduction . . . 24
- 3.2 The censored ZIP regression model . . . 25
  - 3.2.1 Maximum likelihood estimation . . . 25

3.2.2 Some additional notations . . . 27
3.3 Asymptotic results . . . 27
3.4 Simulation study . . . 31
3.4.1 Simulation design . . . 31
3.4.2 Results . . . 32
3.5 An application in health economics: demand for physician service . . . 33
3.6 Discussion . . . 34
3.7 Appendix 1: Technical Lemma . . . 43
3.8 Appendix 2: Technical Calculations . . . 47

4 Censored zero-inflated generalized Poisson and negative binomial regression models: a simulation-based study 49
4.1 Introduction . . . 50
4.2 Censored ZIGP and ZINB models . . . 52
4.2.1 Maximum likelihood estimation in censored ZIGP regression . . . 52
4.2.2 Maximum likelihood estimation in censored ZINB regression . . . 53
4.3 A simulation study . . . 54
4.3.1 Simulation scenario . . . 54
4.3.2 Results . . . 55
4.4 Real data application . . . 56
4.5 Discussion . . . 58
4.6 Appendix 1: R code for fitting the censored ZINB model . . . 62
4.7 Appendix 2: Vuong test . . . 62
4.8 Appendix 3: Technical calculations . . . 64

5 Marginalized zero-inflated Poisson regression with right-censored counts 75
5.1 Introduction . . . 76
5.2 The censored MZIP regression model . . . 77
5.2.1 Loglikelihood estimation in MZIP regression model . . . 77
5.2.2 Loglikelihood estimation in MZIP regression model with right-censored . . . 78
5.2.3 Some further notations . . . 79
5.2.4 Regularity conditions and asymptotic results . . . 80
5.3 Simulation study . . . 81
5.3.1 Simulation design . . . 81
5.3.2 Results . . . 82
5.4 Real data application . . . 89
5.5 Conclusion . . . 89
5.6 Appendix 1: Technical Lemmas . . . 91
5.7 Appendix 2: Technical calculations . . . 96

Bibliography 105

1 Modèles linéaires généralisés

Contents

1.1 Introduction **1**

1.2 Théorie des modèles linéaires généralisés **1**

1.2.1 Spécification d'un modèle linéaire généralisé 1

1.2.2 Espérance et variance des modèles linéaires généralisés 3

1.2.3 Estimation 4

1.2.4 Propriétés asymptotiques et inférence 6

1.2.5 Algorithme de Newton-Raphson 8

1.2.6 Estimation du paramètre de dispersion 9

1.3 Choix de modèles **10**

1.1 Introduction

Les modèles linéaires généralisés sont une classe de modèles statistiques de régression permettant de traiter les problèmes où la loi de la variable réponse (ou variable à expliquer) n'est plus nécessairement gaussienne, comme dans le cas du modèle linéaire ou de l'ANOVA. La variable réponse peut ainsi être discrète. Cette classe de modèles, formulée par Nelder and Wedderburn (1972) et popularisée par McCullagh and Nelder (1989) et quelques autres (Dobson and Barnett, 2018; Dunn and Smyth, 2018) généralise donc le modèle linéaire classique, qui est caractérisé par deux hypothèses fortes : la linéarité de la relation entre le prédicteur linéaire et la variable réponse, et la normalité des erreurs.
Ce chapitre propose une introduction aux modèles linéaires généralisés : formulation, estimation, inférence, choix de modèles.

1.2 Théorie des modèles linéaires généralisés

1.2.1 Spécification d'un modèle linéaire généralisé

Les modèles linéaires généralisés sont une extension du modèle de régression linéaire standard, qui permet de s'affranchir des hypothèses de linéarité de la relation entre le prédicteur linéaire et la variable réponse et de normalité des erreurs. Pour définir un modèle linéaire généralisé, il faut se donner trois éléments que nous décrivons ci-dessous : une composante aléatoire (la distribution de la variable réponse), un prédicteur linéaire (ou composante systématique) et une fonction de lien.

Composante aléatoire

On suppose que l'échantillon statistique est constitué de n variables aléatoires $Y = (Y_1, Y_2, \ldots, Y_n)^\top$ indépendantes et admettant des distributions issues d'une structure exponentielle (voir McCullagh and Nelder (1989)). Pour une mesure dominante adaptée (mesure de *Lebesgue* pour une loi continue, mesure discrète - combinaison de masses de Dirac - pour une loi discrète), la famille de leurs densités par rapport à cette mesure s'écrit sous la forme :

$$f_{Y_i}(y_i, \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - \kappa(\theta_i)}{a(\phi)} + b(y_i, \phi)\right\}, \quad i = 1, \ldots, n \tag{1.2.1}$$

où $\theta_i \in \mathbb{R}$ est appelé *paramètre naturel* (ou *paramètre de position*, *paramètre canonique*) et $\phi \in \mathbb{R}_+^*$ est un paramètre de *dispersion*, ou de *nuisance*. Les fonctions a, b et κ sont spécifiques à chaque type de distribution.
Nous décrivons ci-dessous quelques familles exponentielles, dont nous explicitons le paramètre canonique et le paramètre de dispersion. La fonction b n'intervenant pas dans la suite, nous ne la mentionnons pas ici, et pour simplifier la lecture du tableau nous omettons l'indice i.

Distribution	Paramètre canonique	Paramètre de dispersion
$\mathcal{P}(\mu)$	$\log(\mu)$	1
$\mathcal{B}(n, \mu)$	$\log\left(\frac{\mu}{n-\mu}\right)$	1
$\mathcal{N}(\mu, \sigma^2)$	μ	σ^2
$\Gamma(\mu, \nu)$	$-\frac{1}{\mu}$	$\frac{1}{\nu}$
$\mathcal{NB}(\mu, \kappa)$	$\log\left(\frac{\kappa\mu}{1+\kappa\mu}\right)$	1

Table 1.1: Quelques examples de familles distributions exponentielles

Prédicteur linéaire

Nous notons $\mathbf{X}_i$ le vecteur (colonne) de dimension p des variables explicatives (ou covariables, ou prédicteurs) observées sur l'individu i, $i = 1, \ldots, n$. Nous notons X_{ij} ses composantes, de sorte que $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})^\top$. Les X_{ij} peuvent être quantitatives ou qualitatives. Nous notons enfin $\mathbf{X}$ la matrice, dite matrice design ou matrice du plan d'expérience, de dimension $p \times n$ contenant l'ensemble des variables explicatives observées sur l'ensemble des individus. Le vecteur $\mathbf{X}_i$ est donc le i-ème vecteur colonne de la matrice $\mathbf{X}$.
Nous notons également $\beta = (\beta_1, \ldots, \beta_p)^\top$ le vecteur de dimension p contenant les paramètres de régression inconnus.
On appelle prédicteur linéaire (pour le i-ème individu) la combinaison linéaire suivantes des variables explicatives :

$$\eta_i = \sum_{j=1}^{p} \beta_j X_{ij} = \beta^\top \mathbf{X}_i, \quad i = 1, \ldots, n$$

Cette combinaison linéaire peut inclure des transformations des variables explicatives initiales (e.g., $\log X_{ij}$,) ou des interactions (e.g., $X_{ij} \times X_{ik}$, $j \neq k$; $j, k = 1, \ldots, p$). L'ensemble des n prédicteurs linéaires peut être écrit sous la forme matricielle suivante :

$$\boldsymbol{\eta} = \beta^\top \mathbf{X}$$

avec $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top$. Le prédicteur linéaire est parfois appelé *composante systématique* du modèle, car Y_i est une variable aléatoire alors que X_{ij} est considéré comme fixe.

Fonction de lien

La troisième composante d'un modèle linéaire généralisé permet de spécifier la relation entre la variable à expliquer Y_i et le prédicteur linéaire $\beta^\top \mathbf{X}_i$. Notons $\mu_i = \mathbb{E}(Y_i|\mathbf{X}_i)$ l'espérance conditionnelle de la variable réponse sachant les variables explicatives. On pose :

$$g(\mu_i) = \beta^\top \mathbf{X}_i, \quad i = 1, \ldots, n,$$

où $g(\cdot)$, appelée fonction de lien, désigne une fonction monotone et dérivable. On choisit souvent pour $g(\cdot)$ la fonction de lien canonique, soit $g = (\partial\kappa/\partial\theta_i)^{-1}$. Alors $\theta_i = g(\mu_i) = \beta^\top \mathbf{X}_i$. Dans le tableau suivant, nous indiquons les fonctions de liens canoniques associées à quelques lois classiques.

Distribution	Poisson $\mathcal{P}(\mu)$	Binomial $\mathcal{B}(n,\mu)$	Negative Binomial $\mathcal{NB}(\mu,\kappa)$	Normal $\mathcal{N}(\mu,\sigma^2)$	Gamma $\Gamma(\kappa,\nu)$
$g(x)$	$\log(x)$	$\log(\frac{x}{1-x})$	$\log\left(\frac{\kappa x}{1+\kappa x}\right)$	x	$-\frac{1}{x}$

Table 1.2: Quelques fonctions de lien classiques

Notons que dans le modèle linéaire gaussien, la fonction de lien canonique n'apparaît pas explicitement car elle est égale à l'identité.

1.2.2 Espérance et variance des modèles linéaires généralisés

L'espérance et la variance d'une famille exponentielle admettent des expressions remarquables, dans le cas des familles exponentielles. Plus précisément, elles peuvent s'exprimer en fonction des fonctions a, b et κ.

Supposons que la variable Y_i a une densité de la forme (1.2.1) et notons g la fonction de lien, telle que $g(\mu_i) = \beta^\top \mathbf{X}_i$. Notons $\ell_{n,i} = \log f(y_i, \phi_i, \phi) = (y_i\theta_i - \kappa(\theta_i))/a(\phi) + b(y_i, \phi)$ la contribution de la i-ème observation à la log-vraisemblance et $\ell_n = \sum_{i=1}^n \ell_{n,i}$. On a alors, pour tout $i \in \{1, \ldots, n\}$:

$$\frac{\partial \ell_{n,i}}{\partial \theta_i} = \frac{y_i - \dot{\kappa}(\theta_i)}{a(\phi)} \quad \text{et} \quad \frac{\partial^2 \ell_{n,i}}{\partial \theta_i^2} = -\frac{\ddot{\kappa}(\theta_i)}{a(\phi)},$$

où $\dot{\kappa}$ et $\ddot{\kappa}$ désignent respectivement les dérivées première et seconde de κ par rapport à θ_i. En utilisant le fait que

$$\mathbb{E}\left(\frac{\partial \ell_n}{\partial \theta}\right) = 0 \quad \text{et} \quad \mathbb{E}\left(\frac{\partial^2 \ell_n}{\partial \theta^2}\right) = -\mathbb{E}\left(\left(\frac{\partial \ell_n}{\partial \theta}\right)^2\right),$$

nous obtenons :

$$\mathbb{E}(Y_i|\mathbf{X}_i) = \dot{\kappa}(\theta_i) \quad \text{et} \quad \mathrm{Var}(Y_i|\mathbf{X}_i) = \ddot{\kappa}(\theta_i)a(\phi).$$

L'expression de la variance conditionnelle aux variables explicatives, $\mathrm{Var}(Y_i|\mathbf{X}_i)$, justifie l'appellation de paramètre de dispersion pour ϕ lorsque a est la fonction identité. Notons également que l'on a une relation directe entre l'espérance μ_i de Y_i et sa variance :

$$\mathrm{var}(Y_i) = a(\phi)\ddot{\kappa}(\dot{\kappa}^{-1}\mu_i) = \phi\ddot{\kappa}(\dot{\kappa}^{-1}\mu_i).$$

1.2.3 Estimation

Dans cette section, nous décrivons la méthode d'estimation par maximum de vraisemblance dans les modèles linéaires généralisés et renvoyons le lecteur intéressé à Dupuy (2018) pour plus de détails.

Équations de vraisemblance

L'estimation des paramètres d'un modèle linéaire généralisé repose sur la méthode du maximum de vraisemblance. Supposons que l'on dispose d'un échantillon $Z_1, \dots, Z_n$ d'observations indépendantes, de densités

$$f(y_i; \theta_i, \phi) = \exp\left(\frac{\theta_i y_i - \kappa(\theta_i)}{a(\phi)} + b(y_i, \phi)\right), \quad i = 1, \dots, n.$$

Pour chaque individu i, on dispose également d'un vecteur $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{ip})^\top$ de variables explicatives (qui peuvent être quantitatives ou qualitatives).
On rappelle les notations suivantes : $\mu_i = \mathbb{E}(Y_i|\mathbf{X}_i)$ désigne l'espérance de Y_i sachant les covariables et $g(\mu_i) = \beta^\top \mathbf{X}_i$ est la fonction de lien (qui n'est pas nécessairement canonique). Dans la suite, nous noterons également $\eta_i = \beta^\top \mathbf{X}_i$.
Calculons la vraisemblance du paramètre (β, ϕ) au vu de l'échantillon indépendant $(Z_1, \mathbf{X}_1), \dots, (Z_n, \mathbf{X}_n)$:

$$L_n(\beta, \phi) = \prod_{i=1}^{n} \exp\left(\frac{\theta_i Y_i - \kappa(\theta_i)}{a(\phi)} + b(y_i, \phi)\right).$$

On en déduit la log-vraisemblance $\ell_n(\beta, \phi) = \ln L_n(\beta, \phi)$:

$$\ell_n(\beta, \phi) = \sum_{i=1}^{n} \left\{\frac{\theta_i Y_i - \kappa(\theta_i)}{a(\phi)} + b(y_i, \phi)\right\} := \sum_{i=1}^{n} \ell_{n,i}(\beta, \phi).$$

L'estimateur du maximum de vraisemblance $\hat{\beta}_n$ de β est obtenu en résolvant l'équation (ou plutôt ici, le système de p équations) :

$$\frac{\partial}{\partial \beta}\ell_n(\beta, \phi)\bigg|_{\beta=\hat{\beta}_n} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta}\ell_{n,i}(\beta, \phi)\bigg|_{\beta=\hat{\beta}_n} = 0.$$

Pour tout $j = 1, \dots, p$, calculons :

$$\frac{\partial \ell_{n,i}}{\partial \beta_j} = \frac{\partial \ell_{n,i}}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}.$$

Remarque. On adopte ici la notation dite "de Leibniz" pour la dérivée d'une composée de fonctions : si $y = f(z)$ et $z = g(x)$, la dérivée $\partial f(g(x))/x = \dot{f}(g(x))\dot{g}(x)$ sera notée $\frac{\partial y}{\partial x} = \frac{\partial y}{\partial z}\frac{\partial z}{\partial x}$. □

On a :

$$\frac{\partial \ell_{n,i}}{\partial \theta_i} = \frac{Y_i - \dot{\kappa}(\theta_i)}{a(\phi)} = \frac{Y_i - \mu_i}{a(\phi)},$$
$$\frac{\partial \mu_i}{\partial \theta_i} = V(\mu_i) = \frac{\mathrm{var}(Y_i)}{a(\phi)},$$
$$\frac{\partial \eta_i}{\partial \beta_j} = X_{ij}.$$

Le terme $\partial \mu_i / \partial \eta_i$, quant à lui, dépend de la fonction de lien $g(\mu_i) = \eta_i$ choisie. On obtient :

$$\frac{\partial}{\partial \beta_j} \ell_n(\beta, \phi) = \sum_{i=1}^{n} X_{ij} \frac{(Y_i - \mu_i)}{V(\mu_i) a(\phi)} \frac{\partial \mu_i}{\partial \eta_i}, \quad j = 1, \dots, p. \tag{1.2.2}$$

On en déduit des équations d'estimation (encore appelées équations du score ou équations de vraisemblance) pour les paramètres β_j :

$$\sum_{i=1}^{n} X_{ij} \frac{(Y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1, \dots, p. \tag{1.2.3}$$

Il s'agit d'équations non-linéaires en β, qui n'admettent pas de solution analytique. Des algorithmes itératifs (e.g., algorithme de Newton-Raphson ou de Fisher-scoring) seront utilisés pour approcher l'estimateur du maximum de vraisemblance (voir section 1.2.5).

Remarque. Les équations (1.2.3) dépendent de β au travers des μ_i et η_i, $i = 1, \dots, n$ (nous n'avons pas fait apparaître explicitement cette dépendance, afin de conserver des notations simples). En revanche, ces équations ne dépendent pas de ϕ, l'estimateur du maximum de vraisemblance de β n'en dépend donc pas non plus. ϕ peut être estimé dans un second temps (voir section 1.2.6). □

Cas d'une fonction de lien canonique

Supposons maintenant que l'on prenne pour g la fonction de lien canonique. Alors $\theta_i = g(\mu_i) = \beta^\top \mathbf{X}_i = \eta_i$ et les équations de vraisemblance se simplifient. On a :

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

et les équations (1.2.3) deviennent

$$\sum_{i=1}^{n} X_{ij} (Y_i - \mu_i) = 0, \quad j = 1, \dots, p. \tag{1.2.4}$$

On peut exprimer ces p équations sous forme matricielle. Formons la matrice du plan d'expérience :

$$\mathbb{X} = \begin{pmatrix} 1 & X_{12} & \dots & X_{1p} \\ 1 & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n2} & \dots & X_{np} \end{pmatrix},$$

dont la i-ème ligne ($i = 1, \ldots, n$) contient les variables explicatives du i-ème individu et la j-ième colonne ($j = 1, \ldots, p$) contient les valeurs de la j-ième variable sur les n individus. Dans la suite, nous supposerons que $\mathbb{X}$ est de rang p.
Notons également $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ et $\mu = (\mu_1, \ldots, \mu_n)^\top$. Alors on peut écrire le système des p équations (1.2.4) sous la forme

$$\mathbb{X}^\top(\mathbf{Y} - \mu) = 0. \tag{1.2.5}$$

1.2.4 Propriétés asymptotiques et inférence

L'inférence statistique dans les modèles linéaires généralisés repose sur les propriétés asymptotiques (consistance, normalité asymptotique) de l'estimateur du maximum de vraisemblance. Dans le cas du modèle logistique, ces propriétés ont été établies par Gourieroux and Monfort (1981). Elles ont été démontrées par Fahrmeir and Kaufmann (1985) pour un modèle linéaire généralisé quelconque (avec fonction de lien canonique ou quelconque). On pourra trouver une synthèse de ces théorèmes dans Antoniadis et al. (1992).
Le théorème suivant (voir Antoniadis et al. (1992)) énonce des conditions suffisantes pour la consistance et la normalité asymptotique de l'estimateur du maximum de vraisemblance $\hat{\beta}_n$ dans un modèle linéaire généralisé avec une fonction de lien canonique. Pour simplifier les notations, nous supposons que le paramètre de dispersion est connu et désignons la log-vraisemblance par $\ell_n(\beta)$. Nous notons β le vrai paramètre du modèle et $\mathcal{I}_n(\beta) = -\partial^2\ell_n(\beta)/\partial\beta\partial\beta^\top$.

Théorème 1.2.1 *Dans un modèle linéaire généralisé, on suppose que : i) les variables explicatives $X_{i1}, X_{i2}, \ldots, X_{ip}$ sont bornées et que ii) la plus petite valeur propre de la matrice $\mathbb{X}^\top\mathbb{X}$ tend vers l'infini quand n tend vers l'infini. Alors la suite $(\hat{\beta}_n)$ des estimateurs du maximum de vraisemblance converge en probabilité vers β et $\mathcal{I}_n(\hat{\beta}_n)^{\frac{1}{2}}(\hat{\beta}_n - \beta)$ converge en loi vers le vecteur gaussien $\mathcal{N}(0, I_p)$.*

Remarque. Dans l'asymptotique du modèle linéaire, une condition pour la consistance de l'estimateur des moindres carrés est que $(\mathbb{X}^\top\mathbb{X})^{-1}$ converge vers 0 quand n tend vers l'infini. On peut interpréter cette condition comme le fait que l'information apportée par les variables explicatives augmente indéfiniment avec n. L'hypothèse ii) ci-dessus a la même signification dans les modèles linéaires généralisés. □

Après avoir ajusté un modèle, on cherche souvent à calculer des intervalles ou des régions de confiance pour ses paramètres et à tester la significativité d'un régresseur ou d'un groupe de régresseurs. Nous décrivons ci-dessous quelques régions de confiance et tests d'hypothèses parmi les plus utilisés en pratique.

Intervalles et régions de confiance

On déduit du théorème précédent que la loi de $\hat{\beta}_n$ peut être approchée, pour n grand, par le vecteur gaussien $\mathcal{N}(\beta, \mathcal{I}_n(\hat{\beta}_n)^{-1})$. Si on note $\hat{\sigma}_j^2$ le j-ième terme diagonal de $\mathcal{I}_n(\hat{\beta}_n)^{-1}$ et $\hat{\beta}_{n,j}$ la j-ième composante de $\hat{\beta}_n$ (pour $j = 1, \ldots, p$), la loi de $\hat{\beta}_{n,j}$ peut être approchée par la loi normale $\mathcal{N}(\beta_j, \hat{\sigma}_j^2)$.
On en déduit un intervalle de confiance au niveau asymptotique $(1 - \alpha)$ pour le paramètre β_j :

$$\left[\hat{\beta}_{n,j} - u_{1-\alpha/2}\hat{\sigma}_j \,;\, \hat{\beta}_{n,j} + u_{1-\alpha/2}\hat{\sigma}_j\right],$$

où $u_{1-\alpha/2}$ désigne le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$, défini par $\mathbb{P}(\mathcal{N}(0, 1) \leqslant u_{1-\alpha/2}) = 1 - \alpha/2$ (avec $\alpha \in]0, 1[$).

Remarque. Nous avons déjà pour le modèle linéaire, la quantité $\hat{\sigma}_j$ (estimation de l'écart-type - ici asymptotique - de l'estimateur $\hat{\beta}_{n,j}$) est généralement appelée "standard error". □

On déduit également du théorème précédent que $(\hat{\beta}_n - \beta)^\top \mathcal{I}_n(\hat{\beta}_n)(\hat{\beta}_n - \beta)$ converge en loi vers un χ^2_p. On peut alors construire une région de confiance au niveau asymptotique $(1-\alpha)$ pour le paramètre $\beta = (\beta_1, \ldots, \beta_p)^\top$. Il s'agit de l'ensemble des valeurs β pour lesquelles :

$$(\hat{\beta}_n - \beta)^\top \mathcal{I}_n(\hat{\beta}_n)(\hat{\beta}_n - \beta) \leqslant c_p(1-\alpha),$$

où $c_p(1-\alpha)$ désigne le quantile d'ordre $1-\alpha$ de la loi χ^2_p.

Test sur une composante de β (test de Wald)

Supposons que l'on veuille tester la non-significativité de la j-ième variable explicative dans le prédicteur linéaire $\beta^\top \mathbf{X}_i$, soit l'hypothèse $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$. Sous H_0, la statistique de Wald $\hat{\beta}_{n,j}/\hat{\sigma}_j$ converge en loi vers la loi $\mathcal{N}(0,1)$. On peut alors prendre pour région de rejet de H_0, au niveau asymptotique α :

$$\mathcal{R}_\alpha = \left\{ \left| \frac{\hat{\beta}_{n,j}}{\hat{\sigma}_j} \right| \geqslant u_{1-\alpha/2} \right\}.$$

Test du rapport de vraisemblance

Supposons 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). Le problème de test s'écrit :

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

Le test le plus utilisé en pratique est le test du rapport de vraisemblance. Intuitivement, ce test consiste à comparer les vraisemblances sous H_0 et sous H_1 et à accepter H_0 si elles sont "proches". La vraisemblance sous H_0 vaut $L_n(\hat{\beta}_{n,H_0})$, où $\hat{\beta}_{n,H_0}$ désigne l'estimateur du maximum de vraisemblance obtenu sous la contrainte $\beta_1 = \ldots = \beta_q = 0$ (i.e., en retirant du modèle les q premières variables explicatives). Si H_0 est vraie, $L_n(\hat{\beta}_{n,H_0})$ devrait être "proche" de la vraisemblance maximale $L_n(\hat{\beta}_n)$ (bien sûr, le raisonnement reste valable si on considère la log-vraisemblance). On forme alors la statistique de test

$$D_n = 2 \ln \left(\frac{L_n(\hat{\beta}_n)}{L_n(\hat{\beta}_{n,H_0})} \right) = 2(\ell_n(\hat{\beta}_n) - \ell_n(\hat{\beta}_{n,H_0})). \tag{1.2.6}$$

On montre que si H_0 est vraie, D_n converge en loi vers un χ^2_q lorsque n tend vers l'infini. On en déduit une région de rejet de H_0, au niveau asymptotique α :

$$\mathcal{R}^D_\alpha = \{ D_n \geqslant c_q(1-\alpha) \},$$

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

Remarque. On peut également utiliser un test de Wald pour tester H_0 contre H_1. Notons $\hat{\beta}_{n,1:q}$ le vecteur composé des q premières composantes de $\hat{\beta}_n$ et $M_q(\hat{\beta}_n)$ la matrice de dimension $q \times q$ formée des q premières lignes et colonnes de $(\mathcal{I}_n(\hat{\beta}_n))^{-1}$. Sous H_0, la statistique de Wald

$$W_n = \hat{\beta}^\top_{n,1:q} M_q(\hat{\beta}_n)^{-1} \hat{\beta}_{n,1:q}$$

suit asymptotiquement un χ^2_q. On en déduit une région de rejet de H_0 au niveau asymptotique α : $\mathcal{R}^W_\alpha = \{W_n \geqslant c_q(1-\alpha)\}$. Un troisième test asymptotique, appelé test du score, peut également être utilisé. Il est décrit dans Antoniadis et al. (1992). □

Déviance

Considérons un modèle linéaire généralisé $\mathcal{M}$ défini par $g(\mu_i) = \beta^\top \mathbf{X}_i$. On appelle déviance l'écart entre les log-vraisemblances obtenues sous ce modèle et sous un modèle dans lequel on suppose seulement $\mu_i = \mathbb{E}(Y_i|\mathbf{X}_i)$, $i = 1, \ldots, n$. Dans ce dernier modèle, il y a autant de paramètres μ_i que d'individus i. On l'appelle "modèle saturé".
Notons $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$ les estimateurs du maximum de vraisemblance des μ_i dans le modèle saturé et $\hat{\beta}_n$ l'estimateur du maximum de vraisemblance de β dans le modèle $\mathcal{M}$. On définit la déviance de $\mathcal{M}$ comme :

$$\text{dév}(\mathcal{M}) = 2(\ell_n(\hat{\mu}) - \ell_n(\hat{\beta}_n)).$$

On peut interpréter cette quantité comme une mesure de la qualité d'ajustement aux données du modèle $\mathcal{M}$ (en comparaison du modèle saturé, qui s'ajuste le mieux aux données. Notons toutefois que le modèle saturé n'est pas intéressant en pratique : il possède autant de paramètres qu'il y a d'individus dans l'échantillon et n'apporte aucune information sur l'influence des variables explicatives sur la réponse).

1.2.5 Algorithme de Newton-Raphson

Sauf cas particulier, l'estimateur du maximum de vraisemblance de β dans les modèles linéaires généralisés n'admet pas d'expression explicite. On peut l'approcher numériquement à l'aide d'algorithmes itératifs, tel que l'algorithme de Newton-Raphson. Nous décrivons ci-dessous le principe de cet algorithme.

On cherche à résoudre l'équation de vraisemblance

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

(on rappelle que la solution de cette équation ne dépend pas de ϕ, aussi, dans la suite, nous notons $\ell_n(\beta)$ plutôt que $\ell_n(\beta, \phi)$). On approche la fonction $\partial \ell_n(\beta)/\partial\beta$ par son développement de Taylor à l'ordre 1. Si $\beta^{(i)}$ désigne l'approximation de $\hat{\beta}_n$ obtenue à la i-ème itération de l'algorithme, on cherche $\beta^{(i+1)}$ tel que :

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

On obtient :

$$\beta^{(i+1)} = \beta^{(i)} - \left(\frac{\partial^2 \ell_n(\beta^{(i)})}{\partial \beta \partial \beta^\top}\right)^{-1} \frac{\partial \ell_n(\beta^{(i)})}{\partial \beta}. \tag{1.2.7}$$

L'expression de $\partial \ell_n(\beta)/\partial \beta_j$ est donnée par la formule (1.2.2). Nous calculons $\partial^2 \ell_n(\beta)/\partial\beta\partial\beta^\top$ dans une annexe technique à la fin de cette section.

Partant d'une valeur initiale $\beta^{(0)}$, on itère la formule (1.2.7) jusqu'à ce qu'un critère de convergence soit satisfait (par exemple, la norme $\|\beta^{(i+1)} - \beta^{(i)}\|$ de la différence entre deux approximations successives devient plus petite qu'un seuil $\varepsilon > 0$ fixé).

Cas d'une fonction de lien canonique

Lorsque la fonction de lien est canonique, l'équation (1.2.7) se simplifie. En effet, on a :

$$\frac{\partial \ell_n(\beta)}{\partial \beta} = \frac{1}{a(\phi)}\mathbb{X}^\top(\mathbf{Y} - \mu(\beta)) \quad \text{et} \quad \frac{\partial^2 \ell_n(\beta)}{\partial \beta \partial \beta^\top} = -\frac{1}{a(\phi)}\mathbb{X}^\top W(\beta)\mathbb{X}, \tag{1.2.8}$$

où $\mu(\beta) = (\mu_1(\beta), \ldots, \mu_n(\beta))^\top$ et où $W(\beta)$ désigne la matrice diagonale dont le j-ième terme diagonal vaut $V(\mu_j(\beta))$, $j = 1, \ldots, n$. L'équation (1.2.7) devient donc :

$$\begin{aligned}
\beta^{(i+1)} &= \beta^{(i)} + (\mathbb{X}^\top W(\beta^{(i)})\mathbb{X})^{-1}\mathbb{X}^\top(\mathbf{Y} - \mu(\beta^{(i)})), \\
&= (\mathbb{X}^\top W(\beta^{(i)})\mathbb{X})^{-1}\left[\mathbb{X}^\top W(\beta^{(i)})\mathbb{X}\beta^{(i)} + \mathbb{X}^\top(\mathbf{Y} - \mu(\beta^{(i)}))\right], \\
&= (\mathbb{X}^\top W(\beta^{(i)})\mathbb{X})^{-1}\mathbb{X}^\top W(\beta^{(i)})\left[\mathbb{X}\beta^{(i)} + W(\beta^{(i)})^{-1}(\mathbf{Y} - \mu(\beta^{(i)}))\right], \\
&= (\mathbb{X}^\top W(\beta^{(i)})\mathbb{X})^{-1}\mathbb{X}^\top W(\beta^{(i)})\mathbf{U}(\beta^{(i)}),
\end{aligned}$$

où $\mathbf{U}(\beta^{(i)}) := \mathbb{X}\beta^{(i)} + W(\beta^{(i)})^{-1}(\mathbf{Y} - \mu(\beta^{(i)}))$. Ainsi, à la $(i+1)$-ième itération de l'algorithme, $\beta^{(i+1)}$ prend la forme

$$\beta^{(i+1)} \equiv (\mathbb{X}^\top W\mathbb{X})^{-1}\mathbb{X}^\top W\mathbf{U} \tag{1.2.9}$$

d'un estimateur des moindres carrés pondéré par W (le vecteur $\mathbf{U}$ joue le rôle de la variable réponse et $\mathbb{X}$ est la matrice du plan d'expérience). Notons que W et $\mathbf{U}$ sont remis à jour à chaque itération de l'algorithme car ils dépendent de $\beta^{(i)}$.
L'expression (1.2.9) explique le nom d'algorithme des moindres carrés pondérés itératifs parfois donné à cet algorithme.

Remarque. Dans le modèle linéaire gaussien, $\partial\mu_j/\partial\theta_j = 1$ pour tout $j = 1, \ldots, n$, donc $\mathbf{W}$ est la matrice identité d'ordre n. De plus, $\mu(\beta) = \mathbb{X}\beta$ donc $\mathbf{U}(\beta) = \mathbb{X}\beta + \mathbf{Y} - \mathbb{X}\beta = \mathbf{Y}$, qui ne dépend pas de l'itération i. On retrouve bien la solution explicite $(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbf{Y}$ de l'équation de vraisemblance. □

Sous `R`, les modèles linéaires généralisés sont implémentés dans la fonction `glm` du package `stats`.
La fonction `glm` utilise une version modifiée de l'algorithme de Newton-Raphson (on l'appelle algorithme de Fisher-scoring, il est souvent plus performant que l'algorithme de Newton-Raphson) dans laquelle $\mathcal{I}_n(\beta) = -\partial^2\ell_n(\beta)/\partial\beta\partial\beta^\top$ est remplacée par $\mathbb{E}[\mathcal{I}_n(\beta)]$ dans (1.2.7). Dans la littérature anglaise, $\mathcal{I}_n(\beta)$ est souvent appelée "observed information matrix" et $\mathbb{E}[\mathcal{I}_n(\beta)]$ "expected information matrix".

Remarque. D'après (1.2.8), on note que dans le cas d'une fonction de lien canonique, les matrices $\mathcal{I}_n(\beta)$ et $\mathbb{E}[\mathcal{I}_n(\beta)]$ coïncident. Il en est donc de même pour les algorithmes de Newton-Raphson et Fisher-scoring. □

1.2.6 Estimation du paramètre de dispersion

Lorsqu'il n'est pas connu, le paramètre de dispersion ϕ doit être estimé. On rappelle que $\text{var}(Y) = V(\mu)a(\phi)$, où $V(\mu)$ est la fonction variance. Considérons le cas où $a(\phi) = \phi$. On a alors

$$\phi = \frac{\text{var}(Y)}{V(\mu)},$$

avec $g(\mu) = \beta^\top \mathbf{X}$. Il existe plusieurs estimateurs de ϕ. L'un des plus utilisés est le χ^2 de Pearson, estimateur de type moment donné par

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

où $\hat{\mu}_i = g^{-1}(\hat{\beta}_n^\top \mathbf{X}_i)$.

Annexe technique

On rappelle la formule (1.2.2) :

$$\begin{aligned} \frac{\partial \ell_n(\beta)}{\partial \beta_j} &= \sum_{i=1}^{n} X_{ij} \frac{(Y_i - \mu_i)}{V(\mu_i) a(\phi)} \frac{\partial \mu_i}{\partial \eta_i}, \\ &= \sum_{i=1}^{n} s_{i,j}(\beta, \phi), \quad j = 1, \ldots, p. \end{aligned}$$

Calculons maintenant

$$\frac{\partial^2 \ell_n(\beta)}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^{n} \frac{\partial s_{i,j}(\beta, \phi)}{\partial \beta_k}, \quad j, k = 1, \ldots, p.$$

On a :

$$\frac{\partial s_{i,j}}{\partial \beta_k} = \frac{\partial s_{i,j}}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_k},$$

avec

$$\frac{\partial s_{i,j}}{\partial \mu_i} = X_{ij} \frac{\partial}{\mu_i} \left[\frac{(Y_i - \mu_i)}{V(\mu_i) a(\phi)} \frac{\partial \mu_i}{\partial \eta_i} \right] \quad \text{et} \quad \frac{\partial \eta_i}{\partial \beta_k} = X_{ik}.$$

D'où

$$\frac{\partial^2 \ell_n(\beta)}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^{n} X_{ij} \frac{\partial}{\mu_i} \left[\frac{(Y_i - \mu_i)}{V(\mu_i) a(\phi)} \frac{\partial \mu_i}{\partial \eta_i} \right] \frac{\partial \mu_i}{\partial \eta_i} X_{ik}.$$

Notons que dans le cas d'une fonction de lien canonique, on a $\frac{\partial \mu_i}{\partial \eta_i} = V(\mu_i)$ et donc :

$$\begin{aligned} \frac{\partial^2 \ell_n(\beta)}{\partial \beta_j \partial \beta_k} &= \sum_{i=1}^{n} X_{ij} \frac{\partial}{\mu_i} \left[\frac{(Y_i - \mu_i)}{a(\phi)} \right] V(\mu_i) X_{ik}, \\ &= -\frac{1}{a(\phi)} \sum_{i=1}^{n} X_{ij} V(\mu_i) X_{ik}. \end{aligned}$$

On retrouve bien l'expression de $\partial^2 \ell_n(\beta)/\partial\beta\partial\beta^\top$ donnée par (1.2.8).

1.3 Choix de modèles

Les critères de choix de modèle tels que l'AIC (Akaike, 1974) et le BIC (Schwarz, 1978), ou le test de Vuong (Vuong, 1989), sont souvent utilisés pour comparer entre eux des modèles qui ne sont pas emboîtés.

- Critère d'information d'Akaike (AIC) : pour un modèle à p paramètres, l'AIC est défini par :

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

- Critère d'information bayésien (BIC) : pour un modèle à p paramètres estimé sur n observations, le BIC est défini par :

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

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

- Test de Vuong : le principe du test est le suivant. Soit $f_0(\cdot|\cdot)$ la vraie densité conditionnelle de Y sachant $\mathbf{X}$, et $f(\cdot|\cdot,\hat{\theta})$ la densité conditionnelle estimée, où $\hat{\theta}$ est une estimation de θ. La divergence de Kullback-Leibler entre $f_0(\cdot|\cdot)$ et $f(\cdot|\cdot,\hat{\theta})$ est définie par $\mathbb{E}_0[\log f_0(Y|\mathbf{X}) - \log f(Y|\mathbf{X},\hat{\theta})]$, où $\mathbb{E}_0$ désigne l'espérance sous le vrai modèle.
  Lorsque deux modèles sont en compétition, on peut choisir celui qui présente la plus petite divergence, car il est plus proche du vrai modèle. Par exemple, si le modèle 1 est plus proche du vrai modèle, on a :

$$\mathbb{E}_0[\log f_0(Y|\mathbf{X}) - \log f(Y|\mathbf{X},\hat{\theta}^{(1)})] < \mathbb{E}_0[\log f_0(Y|\mathbf{X}) - \log f(Y|\mathbf{X},\hat{\theta}^{(2)})],$$

  où $\hat{\theta}^{(1)}$ et $\hat{\theta}^{(2)}$ sont des estimations de θ (par exemple, des estimations du maximum de vraisemblance) dans les modèles 1 et 2 respectivement. De manière équivalente,

$$\mathbb{E}_0\left[\log \frac{f(Y|\mathbf{X},\hat{\theta}^{(1)})}{f(Y|\mathbf{X},\hat{\theta}^{(2)})}\right] > 0.$$

  Notons $u_i = \log \frac{f(Y_i|\mathbf{X}_i,\hat{\theta}^{(1)})}{f(Y_i|\mathbf{X}_i,\hat{\theta}^{(2)})}$, $i = 1,\ldots,n$. La statistique de test de Vuong est définie comme suit :

$$\mathcal{Z} = \sqrt{n}\frac{n^{-1}\sum_{i=1}^n u_i}{\sqrt{n^{-1}\sum_{i=1}^n (u_i - \bar{u}_n)^2}}.$$

  Sous l'hypothèse nulle H_0 selon laquelle les modèles 1 et 2 sont également proches du vrai modèle, $\mathcal{Z}$ est distribué asymptotiquement comme une variable normale standard. Ainsi, une règle de décision au niveau asymptotique α rejette H_0 si $|\mathcal{Z}| > z_{1-\frac{\alpha}{2}}$, où $z_{1-\frac{\alpha}{2}}$ est le $(1-\frac{\alpha}{2})$ -quantile de la distribution normale standard. Si $\mathcal{Z} > z_{1-\frac{\alpha}{2}}$ (respectivement $\mathcal{Z} < -z_{1-\frac{\alpha}{2}}$), le test choisit le modèle 1 (respectivement modèle 2).

2 Modèles de régression à inflation de zéros

Contents

2.1 Introduction 13
2.1.1 Modèles à inflation de zéros 13
2.1.2 Modèles de régression à inflation de zéros 15
2.2 Modèle de régression ZIP 16
2.2.1 Définition 16
2.2.2 Estimation dans le modèle ZIP 17
2.3 Modèle de régression ZINB 17
2.3.1 Définition 17
2.3.2 Estimation 18
2.4 Modèle de régression ZIGP 19
2.4.1 Définition 19
2.4.2 Estimation 21

2.1 Introduction

Dans ce chapitre, nous nous intéressons à l'inflation de zéros. Ce phénomène, que nous définissons plus précisément dans la suite du chapitre, intervient lorsque l'on observe un nombre "excessif" de zéros 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 zéro et un modèle classique de comptage (typiquement, un modèle de Poisson, ou Poisson généralisé, ou binomial...).

2.1.1 Modèles à inflation de zéros

Le phénomène d'inflation de zéros s'observe, par exemple, dans les études économiques portant sur la consommation et le renoncement aux soins médicaux. Dans ce contexte, des exemples d'une variable réponse sujette à une inflation de zéros sont : le nombre de fois où une personne a consulté un médecin dans un intervalle de temps donné (voir, par exemple, les articles Yen et al. (2001); Sarma and Simpson (2006); Sari (2009); Pizer and Prentice (2011); Staub and Winkelmann (2013); Diallo et al. (2017a)), le nombre de prescriptions médicales reçues par les membres d'un même foyer sur une période de temps donnée (voir Street et al. (1999)), le nombre d'arrêts de travail suivant un premier arrêt consécutif à un accident du travail (voir Campolieti (2002)).

Par "inflation de zéros", on entend la situation où le nombre de zéros observés dans un échantillon de données de comptage est supérieur à celui attendu sous un modèle de comptage "classique" (tel que le modèle de Poisson, ou le modèle binomial). L'une des approches les plus utilisées pour

rendre compte de ce type de données consiste à supposer que la loi de probabilité de la variable de comptage (que nous notons Y par la suite) est un mélange d'une loi dégénérée en 0 (i.e., qui prend la valeur 0 avec probabilité 1) et d'un modèle de comptage.

Supposons, pour illustrer notre propos, que la loi du comptage considéré soit une loi de Poisson $\mathcal{P}(\lambda)$. On note ainsi la loi de Z :

$$Y \sim \omega\delta_0 + (1-\omega)\mathcal{P}(\lambda). \tag{2.1.1}$$

Dans cette expression, ω représente la probabilité pour que Z soit systématiquement égal à 0 (dans la suite, nous l'appellerons "probabilité d'inflation de zéros") et δ_0 désigne la loi dégénérée en 0. La relation (2.1.1) peut encore s'interpréter comme suit :

$$Y \sim \begin{cases} 0 & \text{avec probabilité } \omega, \\ \mathcal{P}(\lambda) & \text{avec probabilité } 1-\omega. \end{cases} \tag{2.1.2}$$

L'événement $\{Y = 0\}$ recouvre donc deux situations distinctes, correspondant aux deux types de zéros mentionnés précédemment. Dans la première situation, qui survient avec probabilité ω, l'événement $\{Y = 0\}$ est certain. Dans la deuxième, qui survient avec probabilité $1-\omega$, le zéro observé est la réalisation d'une expérience aléatoire, ici modélisée par une loi de Poisson.
Notons S la variable qui vaudrait 1 lorsque le zéro observé est structurel (première situation) et 0 lorsqu'il est circonstanciel, ou "aléatoire" (deuxième situation). Cette variable, qui nous renseigne sur le type de zéro observé, n'est pas observable. Il s'agit d'une construction théorique qui permet de calculer la loi de probabilité de Y. La loi de S est telle que $\mathbb{P}(S = 1) = \omega$ et $\mathbb{P}(S = 0) = 1-\omega$. On peut réexprimer (2.1.2) en disant que :

- conditionnellement à l'événement $\{S = 1\}$, Y vaut 0 avec probabilité 1,
- conditionnellement à l'événement $\{S = 0\}$, Y vaut y (avec $y = 0, 1, \ldots$) avec probabilité $e^{-\lambda}\frac{\lambda^y}{y!}$. En particulier, sachant que S vaut 0, Y prend la valeur 0 avec probabilité $e^{-\lambda}$.

Ainsi, d'après la formule des probabilités totales, nous avons :

$$\begin{aligned} \mathbb{P}(Y=0) &= \mathbb{P}(Y=0|S=1)\mathbb{P}(S=1) + \mathbb{P}(Y=0|S=0)\mathbb{P}(S=0), \\ &= 1\cdot\omega + e^{-\lambda}\cdot(1-\omega), \\ &= \omega + (1-\omega)e^{-\lambda}. \end{aligned}$$

De la même manière, si $z = 1, 2, \ldots$, on montre :

$$\begin{aligned} \mathbb{P}(Y=y) &= \mathbb{P}(Y=y|S=1)\mathbb{P}(S=1) + \mathbb{P}(Y=y|S=0)\mathbb{P}(S=0), \\ &= 0\cdot\omega + \frac{e^{-\lambda}\lambda^y}{y!}\cdot(1-\omega), \\ &= (1-\omega)\frac{e^{-\lambda}\lambda^y}{y!}. \end{aligned}$$

La loi de probabilité de Y peut ainsi être résumée par :

$$\mathbb{P}(Y=y) = \begin{cases} \omega + (1-\omega)e^{-\lambda} & y = 0 \\ (1-\omega)\frac{e^{-\lambda}\lambda^y}{y!} & y = 1, 2, \ldots \end{cases} \tag{2.1.3}$$

Le modèle (2.1.1) est appelé modèle de Poisson à inflation de zéros, plus connu sous l'acronyme ZIP (pour "zero-inflated Poisson") (dans la suite, nous noterons ce modèle ZIP(λ, ω)). Lorsque

ω vaut 0, le modèle ZIP se ramène à la loi de Poisson de paramètre $\lambda > 0$. On observe facilement que $\mathbb{P}(Y = 0) = e^{-\lambda} + \omega(1 - e^{-\lambda})$ est supérieure ou égale à la probabilité $e^{-\lambda}$ qu'une loi de Poisson de paramètre λ prenne la valeur 0. Dans le modèle (2.1.1), cette probabilité se trouve augmentée du terme $\omega(1 - e^{-\lambda})$.

A partir de (2.1.3), on calcule facilement l'espérance et la variance de la variable aléatoire $Y \sim \omega\delta_0 + (1-\omega)\mathcal{P}(\lambda)$. On a :

$$\begin{aligned}\mathbb{E}(Y) &= \sum_{y=0}^{+\infty} y\mathbb{P}(Y = y),\\ &= (1-\omega)\sum_{y=0}^{+\infty} y\frac{e^{-\lambda}\lambda^y}{y!},\\ &= (1-\omega)\lambda,\end{aligned}$$

car $\mathbb{E}(\mathcal{P}(\lambda)) = \sum_{y=0}^{+\infty} y\frac{e^{-\lambda}\lambda^y}{y!} = \lambda$, et

$$\begin{aligned}\text{var}(Y) &= \sum_{y=0}^{+\infty}(1-\omega)y^2\frac{e^{-\lambda}\lambda^y}{y!} - \left[(1-\omega)\lambda\right]^2,\\ &= (1-\omega)\left(\lambda+\lambda^2\right) - \left[(1-\omega)\lambda\right]^2,\\ &= (1+\omega\lambda)(1-\omega)\lambda,\\ &= (1+\omega\lambda)\mathbb{E}(Y).\end{aligned}$$

Lorsque la probabilité ω d'inflation de zéros est strictement positive, on vérifie que $\text{var}(Y) > \mathbb{E}(Y)$ et donc qu'une inflation de zéros entraîne la sur-dispersion de la distribution de Y (rappelons que la distribution de Poisson est équidispersée, i.e. $\text{var}(Y) = \mathbb{E}(Y)$). □

2.1.2 Modèles de régression à inflation de zéros

Les premières mentions de modèles à inflation de zéros dans la littérature statistique remontent aux années 1960 (le lecteur intéressé pourra se reporter à Johnson et al. (2005) pour des références bibliographiques). Un modèle ZIP(λ, ω) dans lequel le paramètre λ dépend de variables explicatives (ou régresseurs, ou covariables) est proposé dans Mullahy (1986). Dans Lambert (1992), des variables explicatives agissent également sur la probabilité ω d'inflation de zéros. Les modèles proposés dans Mullahy (1986); Lambert (1992) sont donc des modèles *de régression* de Poisson à inflation de zéros. Dans Hall (2000), l'auteur introduit un modèle de régression binomial à inflation de zéros.

Depuis, d'autres modèles de régression à inflation de zéros ont été proposés, et adaptés à des situations de plus en plus complexes (réponse multivariée, données répétées, données manquantes ou censurées...). Sans prétendre à l'exhaustivité, citons le modèle de régression de Poisson généralisé à inflation de zéros Famoye and Singh (2003); Gupta et al. (2004); Famoye and Singh (2006); Czado and Min (2005); Czado et al. (2007), les modèles de régression à inflation de zéros et effets aléatoires Hall (2000); Hall and Zhang (2004); Xiang et al. (2007); Xie et al. (2009); Zhu et al. (2017), le modèle de régression semiparamétrique de Poisson à inflation de zéros Lam et al. (2006); He et al. (2010); Feng and Zhu (2011), le modèle EIB (pour "endpoint-inflated binomial") introduit par Deng and Zhang (2015) pour modéliser un excès de zéros et de la valeur m dans un échantillon de la loi binomiale $\mathcal{B}(m, p)$ (voir également Tian et al. (2015); Dupuy

(2017)). Des modèles de régression à inflation de zéros ont été développés dans des situations de données manquantes Chen and Fu (2011); Lukusa et al. (2016); Diallo et al. (2019) ou censurées Prasad (2009). Un modèle à inflation de zéros pour une variable réponse bivariée est proposé dans Wang (2003). Un modèle de régression multinomial à inflation de zéros est proposé dans Diallo et al. (2018).
Plusieurs ouvrages très complets traitent en détail des modèles de régression ZIP et ZINB (voir Cameron and Trivedi (2013); Winkelmann (2013)). Dans les sections suivantes, nous décrivons brièvement les modèles de régression de Poisson et binomial à inflation de zéros.

2.2 Modèle de régression ZIP

2.2.1 Définition

On observe une variable de comptage Y sur un échantillon de n individus. On note Y_i l'observation de Y sur l'individu i, $i = 1, \ldots, n$. On obtient un modèle de régression de Poisson à inflation de zéros pour Y_i en autorisant, la probabilité ω et l'intensité λ à dépendre de l'individu i au travers de variables explicatives (ou *covariables*). Le modèle est alors défini par :

$$\mathbb{P}(Y_i = y_i) = \begin{cases} \omega_i + (1-\omega_i)e^{-\lambda_i} & \text{si } y_i = 0 \\ (1-\omega_i)\frac{e^{-\lambda_i}\lambda_i^y}{y!} & \text{si } y_i = 1, 2, \ldots \end{cases} \tag{2.2.1}$$

où $\omega_i = \omega_i(\gamma)$ et $\lambda_i = \lambda_i(\beta)$ sont fonctions, respectivement, des vecteurs de covariables $\mathbf{W}_i = (W_{i1}, \ldots, W_{iq})^\top$ et $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\top$ (on pose $X_{i1} = W_{i1} = 1$). Les composantes de ces vecteurs peuvent être qualitatives ou quantitatives. La probabilité ω_i est généralement décrite par une régression logistique :

$$\begin{aligned} & \text{logit}(\omega_i) = \gamma^\top \mathbf{W}_i = \gamma_1 + \gamma_2 W_{i2} + \ldots + \gamma_q W_{iq}, \\ \Longleftrightarrow \quad & \omega_i = \frac{\exp\left(\gamma^\top \mathbf{W}_i\right)}{1 + \exp\left(\gamma^\top \mathbf{W}_i\right)} \in (0, 1), \end{aligned}$$

et l'intensité λ_i est généralement modélisée par :

$$\begin{aligned} & \log(\lambda_i) = \beta^\top \mathbf{X}_i = \beta_1 + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}, \\ \Longleftrightarrow \quad & \lambda_i = \exp\left(\beta^\top \mathbf{X}_i\right), \end{aligned}$$

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

$$\forall i = 1, \ldots, n, \quad \begin{cases} Y_i \sim \omega_i \delta_0 + (1-\omega_i)\mathcal{P}(\lambda_i), \\ \text{logit}(\omega_i) = \gamma^\top \mathbf{W}_i, \\ \log(\lambda_i) = \beta^\top \mathbf{X}_i. \end{cases} \tag{2.2.2}$$

On le note $Y_i \sim \text{ZIP}(\lambda_i, \omega_i)$. Conditionnellement à $\mathbf{X}_i$ et $\mathbf{W}_i$, l'espérance et la variance de Y_i sont données par :

$$\mathbb{E}(Y_i|\mathbf{X}_i, \mathbf{W}_i) = (1-\omega_i)\lambda_i = \frac{\exp\left(\beta^\top \mathbf{X}_i\right)}{1 + \exp\left(\gamma^\top \mathbf{W}_i\right)},$$

et

$$\text{var}(Y_i|\mathbf{X}_i, \mathbf{W}_i) = (1 + \omega_i\lambda_i)(1-\omega_i)\lambda_i = (1 + \omega_i\lambda_i)\mathbb{E}(Y_i|\mathbf{X}_i, \mathbf{W}_i).$$

La distribution conditionnelle de Y_i est sur-dispersée. En effet, $(1+\omega_i\lambda_i) > 1$ donc $\text{var}(Y_i|\mathbf{X}_i, \mathbf{W}_i) > \mathbb{E}(Y_i|\mathbf{X}_i, \mathbf{W}_i)$.

2.2.2 Estimation dans le modèle ZIP

Supposons que l'on dispose d'un échantillon $(Y_1, \mathbf{X}_1, \mathbf{W}_1)$, ..., $(Y_n, \mathbf{X}_n, \mathbf{W}_n)$ d'observations indépendantes du modèle (2.2.2). Le paramètre $\psi = (\beta^\top, \gamma^\top)^\top$ du modèle peut être estimé par la méthode du maximum de vraisemblance. La vraisemblance se calcule comme suit :

$$L_n(\psi) = \prod_{i=1}^{n} \left(\omega_i + (1-\omega_i)e^{-\lambda_i}\right)^{1_{\{Y_i=0\}}} \cdot \left((1-\omega_i)e^{-\lambda_i}\frac{\lambda_i^{Y_i}}{Y_i!}\right)^{1_{\{Y_i>0\}}}.$$

On en déduit la log-vraisemblance $\ell_n(\psi) = \log L_n(\psi)$:

$$\begin{aligned}\ell_n(\psi) = &\sum_{i=1}^{n} 1_{\{Y_i=0\}} \log\left(\omega_i + (1-\omega_i)e^{-\lambda_i}\right) \\ &+ \sum_{i=1}^{n} 1_{\{Y_i>0\}} \left(\log(1-\omega_i) - \lambda_i + Y_i \log\lambda_i - \log(Y_i!)\right).\end{aligned}$$

et finalement :

$$\begin{aligned}\ell_n(\psi) \quad = \quad &\sum_{i=1}^{n} 1_{\{Y_i=0\}} \log\left(\exp\left(\gamma^\top \mathbf{W}_i\right) + \exp\left(-\exp(\beta^\top \mathbf{X}_i)\right)\right) \\ &+ \sum_{i=1}^{n} 1_{\{Y_i>0\}} \left(Y_i \beta^\top \mathbf{X}_i - \exp(\beta^\top \mathbf{X}_i) - \log(Y_i!)\right) \\ &- \sum_{i=1}^{n} \log\left(1 + \exp\left(\gamma^\top \mathbf{W}_i\right)\right).\end{aligned} \tag{2.2.3}$$

L'estimateur du maximum de vraisemblance $\hat{\psi}_n$ de ψ est obtenu en résolvant l'équation d'estimation suivante (appelée équation du score) :

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

Il s'agit en fait ici d'un système de $p+q$ équations, obtenues en dérivant $\ell_n(\psi)$ par rapport à chacun des γ_j et β_j. Cet estimateur n'admet pas d'expression explicite. Il peut être approché par des algorithmes d'optimisation non-linéaire (tels que Newton-Raphson...). Dans Lambert (1992); Hall (2000), les auteurs proposent d'utiliser l'algorithme EM (pour "Expectation-Maximisation") pour maximiser la log-vraisemblance (2.2.3).

2.3 Modèle de régression ZINB

2.3.1 Définition

Soit $\tilde{Y}_i$, $(i = 1, 2, \ldots, n)$ une variable aléatoire discrète suivant une loi binomiale négative à deux paramètres α, μ_i, soit : $\tilde{Y}_i \sim \mathcal{NB}(\alpha, \mu_i)$. Sa distribution s'exprime comme suit[1] :

$$\mathbb{P}(\tilde{Y}_i = y_i) = \frac{\Gamma(\alpha^{-1} + y_i)}{\Gamma(\alpha^{-1})y_i!}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i}, \quad y_i = 0, 1, \ldots. \tag{2.3.1}$$

[1] On rappelle que la fonction gamma est la function définie par $\Gamma : x \mapsto \int_0^\infty e^{-t}t^{x-1}dt$

Avec les mêmes notations que ci-dessus, une variable de comptage Y_i, $i = 1, 2, \ldots, n$ suit un modèle de régression binomial négatif à inflation de zéros si sa distribution est donnée par :

$$\mathbb{P}(Y_i = y_i) = \begin{cases} \omega_i + (1-\omega_i)\left(\dfrac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} & \text{si } y_i = 0 \\ (1-\omega_i)\dfrac{\Gamma(\alpha^{-1}+y_i)}{\Gamma(\alpha^{-1})y_i!}\left(\dfrac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\dfrac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i} & \text{si } y_i = 1, 2, \ldots \end{cases} \tag{2.3.2}$$

où $\omega_i = \omega_i(\gamma)$ et $\mu_i = \mu_i(\beta)$ sont fonctions, respectivement, des vecteurs de covariables $\mathbf{W}_i = (W_{i1}, \ldots, W_{iq})^\top$ et $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\top$ (on pose $X_{i1} = W_{i1} = 1$). Les composantes de ces vecteurs peuvent être qualitatives ou quantitatives. La probabilité ω_i est généralement décrite par une régression logistique :

$$\begin{aligned} & \text{logit}(\omega_i) = \gamma^\top \mathbf{W}_i = \gamma_1 + \gamma_2 W_{i2} + \ldots + \gamma_q W_{iq}, \\ \Longleftrightarrow \quad & \omega_i = \frac{\exp\left(\gamma^\top \mathbf{W}_i\right)}{1 + \exp\left(\gamma^\top \mathbf{W}_i\right)} \in (0, 1), \end{aligned}$$

et μ_i est généralement modélisée par :

$$\begin{aligned} & \log(\mu_i) = \beta^\top \mathbf{X}_i = \beta_1 + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}, \\ \Longleftrightarrow \quad & \mu_i = \exp\left(\beta^\top \mathbf{X}_i\right), \end{aligned}$$

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

$$\forall i = 1, \ldots, n, \quad \begin{cases} Y_i \sim \omega_i \delta_0 + (1-\omega_i)\mathcal{NB}(\alpha, \mu_i), \\ \text{logit}(\omega_i) = \gamma^\top \mathbf{W}_i, \\ \log(\mu_i) = \beta^\top \mathbf{X}_i. \end{cases} \tag{2.3.3}$$

On le note $Y_i \sim \text{ZINB}(\alpha, \mu_i, \omega_i)$. Conditionnellement à $\mathbf{X}_i$ et $\mathbf{W}_i$, l'espérance et la variance de Y_i sont données par :

$$\mathbb{E}(Y_i | \mathbf{X}_i, \mathbf{W}_i) = (1-\omega_i)\mu_i = \frac{\exp\left(\beta^\top \mathbf{X}_i\right)}{1 + \exp\left(\gamma^\top \mathbf{W}_i\right)},$$

et

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

La distribution conditionnelle de Y_i est sur-dispersée. En effet, $(1 + (\alpha + \omega_i)\mu_i) > 1$ donc $\text{var}(Y_i | \mathbf{X}_i, \mathbf{W}_i) > \mathbb{E}(Y_i | \mathbf{X}_i, \mathbf{W}_i)$.

2.3.2 Estimation

Supposons que l'on dispose d'un échantillon $(Y_1, \mathbf{X}_1, \mathbf{W}_1)$, $\ldots$, $(Y_n, \mathbf{X}_n, \mathbf{W}_n)$ d'observations indépendantes du modèle (2.3.3). Le paramètre $\psi = (\beta^\top, \gamma^\top)^\top$ du modèle peut être estimé par

la méthode du maximum de vraisemblance. La vraisemblance se calcule comme suit :

$$\begin{aligned} L_n(\psi) &= \prod_{i=1}^{n} \left\{ \left(\omega_i + (1-\omega_i) \left(\frac{1}{1+\alpha\mu_i} \right)^{\frac{1}{\alpha}} \right)^{1_{\{Y_i=0\}}} \right. \\ &\qquad \left. \times \left((1-\omega_i) \frac{\Gamma(\alpha^{-1}+Y_i)}{\Gamma(\alpha^{-1})Y_i!} \left(\frac{1}{1+\alpha\mu_i} \right)^{\frac{1}{\alpha}} \left(\frac{\mu_i}{\alpha^{-1}+\mu_i} \right)^{Y_i} \right)^{1_{\{Y_i>0\}}} \right\}, \\ &= \prod_{i=1}^{n} \left\{ \frac{1}{1+e^{\gamma^\top \mathbf{W}_i}} \left(e^{\gamma^\top \mathbf{W}_i} + \left(\frac{1}{1+\alpha e^{\beta_i^\top \mathbf{X}_i}} \right)^{\frac{1}{\alpha}} \right)^{1_{\{Y_i=0\}}} \right. , \\ &\qquad \left. \times \left(\frac{1}{1+e^{\gamma^\top \mathbf{W}_i}} \frac{\Gamma(\alpha^{-1}+Y_i)}{\Gamma(\alpha^{-1})Y_i!} \left(\frac{1}{1+\alpha e^{\beta^\top \mathbf{X}_i}} \right)^{\frac{1}{\alpha}} \left(\frac{e^{\beta^\top \mathbf{X}_i}}{\alpha^{-1}+e^{\beta^\top \mathbf{X}_i}} \right)^{Y_i} \right)^{1_{\{Y_i>0\}}} \right\}. \end{aligned}$$

On en déduit la log-vraisemblance $\ell_n(\psi) = \log L_n(\psi)$:

$$\begin{aligned} \ell_n(\psi) &= \sum_{i=1}^{n} 1_{\{Y_i=0\}} \left[\log \left(e^{\gamma^\top \mathbf{W}_i} + (1+\alpha e^{\beta^\top \mathbf{X}_i})^{-\frac{1}{\alpha}} \right) - \log(1+e^{\gamma^\top \mathbf{W}_i}) \right] \\ &\quad + \sum_{i=1}^{n} 1_{\{Y_i>0\}} \left[-\frac{1}{\alpha} \log(1+\alpha e^{\beta^\top \mathbf{X}_i}) + Y_i(\beta^\top \mathbf{X}_i - \log(\alpha^{-1}+e^{\beta^\top \mathbf{X}_i})) - \log(1+e^{\gamma^\top \mathbf{W}_i}) \right. \\ &\qquad \left. + \log\Gamma(\alpha^{-1}+Y_i) - \log\Gamma(\alpha^{-1}) - \log(Y_i!) \right], \end{aligned}$$

qui se simplifie en :

$$\begin{aligned} \ell_n(\psi) &= \sum_{i=1}^{n} 1_{\{Y_i=0\}} \log \left(e^{\gamma^\top \mathbf{W}_i} + (1+\alpha e^{\beta^\top \mathbf{X}_i})^{-\frac{1}{\alpha}} \right) \\ &\quad + \sum_{i=1}^{n} 1_{\{Y_i>0\}} \left[Y_i \beta^\top \mathbf{X}_i - Y_i \log(\alpha^{-1}+e^{\beta^\top \mathbf{X}_i}) - \frac{1}{\alpha} \log(1+\alpha e^{\beta^\top \mathbf{X}_i}) \right. \\ &\qquad \left. + \log\Gamma(\alpha^{-1}+Y_i) - \log\Gamma(\alpha^{-1}) - \log(Y_i!) \right] \\ &\quad - \sum_{i=1}^{n} \log(1+e^{\gamma^\top \mathbf{W}_i}). \end{aligned} \tag{2.3.4}$$

L'estimateur du maximum de vraisemblance $\hat{\psi}_n$ de ψ est obtenu en résolvant l'équation du score suivante :

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

2.4 Modèle de régression ZIGP

2.4.1 Définition

Soit $\tilde{Y}_i$, $(i = 1, 2, \ldots, n)$ une variable aléatoire discrète suivant une loi de Poisson généralisée à deux paramètres λ_i et φ (on note $\tilde{Y}_i \sim \mathcal{GP}(\lambda_i, \varphi)$). Sa fonction de probabilité s'écrit :

$$\mathbb{P}(\tilde{Y}_i = y_i) = \begin{cases} \frac{\lambda_i(\lambda_i+(\varphi-1)y_i)^{y_i-1}}{y_i!} \varphi^{-y_i} e^{-\frac{\lambda_i+(\varphi-1)y_i}{\varphi}} & \text{si} \quad y_i = 0, 1, \ldots, \\ 0 & \text{si} \quad y_i > m \text{ quand } \varphi < 1 \end{cases} \tag{2.4.1}$$

avec :

- $\lambda_i > 0$
- $\varphi > \max\left\{\frac{1}{2}, 1 - \frac{\lambda_i}{m}\right\}$, où m est le plus grand nombre entier non négatif tel que $\lambda_i + m(\varphi - 1) > 0$ quand $\varphi < 1$.

On montre :

$$\begin{aligned}\mathbb{E}(\tilde{Y}_i) &= \lambda_i,\\ \text{var}(\tilde{Y}_i) &= \varphi^2\lambda_i.\end{aligned}$$

Le cas $\varphi = 1 (> 1, < 1)$ correspond à l'équidispersion (sur-dispersion, sous-dispersion). De plus, si $\varphi = 1$, la distribution devient la distribution de Poisson standard.

On définit un modèle de Poisson généralisé à inflation de zéros en posant :

$$\mathbb{P}(Y_i = y_i) = \begin{cases} \omega_i + (1-\omega_i)\mathbb{P}(\tilde{Y}_i = y_i) & \text{si} \quad y_i = 0,\\ (1-\omega_i)\mathbb{P}(\tilde{Y}_i = y_i) & \text{si} \quad y_i = 1, 2, \ldots. \end{cases} \tag{2.4.2}$$

où $\mathbb{P}(\tilde{Y}_i = y_i)$ est donnée par (2.4.1). Si des variables explicatives sont disponibles, on construit un modèle de régression de Poisson généralisé à inflation de zéros (modèle ZIGP) en posant :

$$\begin{aligned} & \text{logit}(\omega_i) = \gamma^\top \mathbf{W}_i = \gamma_1 + \gamma_2 W_{i2} + \ldots + \gamma_q W_{iq},\\ \Longleftrightarrow \quad & \omega_i = \frac{\exp\left(\gamma^\top \mathbf{W}_i\right)}{1 + \exp\left(\gamma^\top \mathbf{W}_i\right)} \in (0,1), \end{aligned}$$

et :

$$\begin{aligned} & \log(\mu_i) = \beta^\top \mathbf{X}_i = \beta_1 + \beta_2 X_{i2} + \ldots + \beta_p X_{ip},\\ \Longleftrightarrow \quad & \mu_i = \exp\left(\beta^\top \mathbf{X}_i\right), \end{aligned}$$

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

$$\forall i = 1, \ldots, n, \quad \begin{cases} Y_i \sim \omega_i \delta_0 + (1-\omega_i)\mathcal{GP}(\lambda_i, \varphi, \omega_i),\\ \text{logit}(\omega_i) = \gamma^\top \mathbf{W}_i,\\ \log(\mu_i) = \beta^\top \mathbf{X}_i. \end{cases} \tag{2.4.3}$$

On le note $Y_i \sim \text{ZIGP}(\lambda_i, \varphi, \omega_i)$. Conditionnellement à $\mathbf{X}_i$ et $\mathbf{W}_i$, l'espérance et la variance de Y_i sont données par :

$$\mathbb{E}(Y_i|\mathbf{X}_i, \mathbf{W}_i) = (1-\omega_i)\mu_i = \frac{\exp\left(\beta^\top \mathbf{X}_i\right)}{1 + \exp\left(\gamma^\top \mathbf{W}_i\right)},$$

et

$$\text{var}(Y_i|\mathbf{X}_i, \mathbf{W}_i) = (\varphi^2 + \omega_i\mu_i)(1-\omega_i)\mu_i = (\varphi^2 + \omega_i\mu_i)\mathbb{E}(Y_i|\mathbf{X}_i, \mathbf{W}_i).$$

2.4.2 Estimation

Supposons que l'on dispose d'un échantillon $(Y_1, \mathbf{X}_1, \mathbf{W}_1)$, ..., $(Y_n, \mathbf{X}_n, \mathbf{W}_n)$ d'observations indépendantes du modèle (2.2.2). Le paramètre $\psi = (\beta^\top, \gamma^\top)^\top$ du modèle peut être estimé par la méthode du maximum de vraisemblance. La vraisemblance se calcule comme suit :

$$\begin{aligned} L_n(\psi) &= \prod_{i=1}^{n} \left\{ \left(\omega_i + (1-\omega_i) e^{-\frac{\lambda_i}{\varphi}} \right)^{1_{\{Y_i=0\}}} \right. \\ &\qquad \left. \times \left((1-\omega_i) \frac{\lambda_i (\lambda_i + (\varphi-1)Y_i)^{Y_i-1}}{Y_i!} \varphi^{-Y_i} e^{-\frac{\lambda_i + (\varphi-1)Y_i}{\varphi}} \right)^{1_{\{Y_i>0\}}} \right\}, \end{aligned}$$

On en déduit la log-vraisemblance $\ell_n(\psi) = \log L_n(\psi)$:

$$\begin{aligned} \ell_n(\psi) &= \sum_{i=1}^{n} 1_{\{Y_i=0\}} \log \left(\omega_i + (1-\omega_i) e^{-\frac{\lambda_i}{\varphi}} \right) \\ &\quad + \sum_{i=1}^{n} 1_{\{Y_i>0\}} \log \left((1-\omega_i) \frac{\lambda_i (\lambda_i + (\varphi-1)Y_i)^{Y_i-1}}{Y_i!} \varphi^{-Y_i} e^{-\frac{\lambda_i + (\varphi-1)Y_i}{\varphi}} \right), \\ &= \sum_{i=1}^{n} 1_{\{Y_i=0\}} \left[\log \left(e^{\gamma^\top \mathbf{W}_i} + e^{-\frac{e^{\beta^\top \mathbf{X}_i}}{\varphi}} \right) - \log(1 + e^{\gamma^\top \mathbf{W}_i}) \right] \\ &\quad + \sum_{i=1}^{n} 1_{\{Y_i>0\}} \left[\beta^\top \mathbf{X}_i + (Y_i - 1) \log \left(e^{\beta^\top \mathbf{X}_i} + (\varphi-1)Y_i \right) - \log(1 + e^{\gamma^\top \mathbf{W}_i}) \right. \\ &\qquad \left. - \frac{e^{\beta^\top \mathbf{X}_i} + (\varphi-1)Y_i}{\varphi} - Y_i \log \varphi - \log(Y_i!) \right], \end{aligned}$$

et finalement :

$$\begin{aligned} \ell_n(\psi &= \sum_{i=1}^{n} 1_{\{Y_i=0\}} \log \left(e^{\gamma^\top \mathbf{W}_i} + e^{-\frac{e^{\beta^\top \mathbf{X}_i}}{\varphi}} \right) \\ &\quad + \sum_{i=1}^{n} 1_{\{Y_i>0\}} \left[\beta^\top \mathbf{X}_i + (Y_i - 1) \log \left(e^{\beta^\top \mathbf{X}_i} + (\varphi-1)Y_i \right) \right. \\ &\qquad \left. - \frac{e^{\beta^\top \mathbf{X}_i} + (\varphi-1)Y_i}{\varphi} - Y_i \log \varphi - \log(Y_i!) \right] - \sum_{i=1}^{n} \log(1 + e^{\gamma^\top \mathbf{W}_i}). \end{aligned}$$

L'estimateur du maximum de vraisemblance $\hat{\psi}_n$ de ψ est obtenu en résolvant l'équation du score suivante :

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

3 Zero-inflated Poisson regression model with right censored data

Contents

3.1 Introduction **24**
3.2 The censored ZIP regression model **25**
3.2.1 Maximum likelihood estimation 25
3.2.2 Some additional notations 27
3.3 Asymptotic results **27**
3.4 Simulation study **31**
3.4.1 Simulation design 31
3.4.2 Results 32
3.5 An application in health economics: demand for physician service . **33**
3.6 Discussion **34**
3.7 Appendix 1: Technical Lemma **43**
3.8 Appendix 2: Technical Calculations **47**

Ce chapitre fait l'objet d'un article intitulé "Asymptotic results in censored zero-inflated Poisson regression" (auteurs: Van-Trinh Nguyen et Jean-François Dupuy), à paraître dans "Communications in Statistics - Theory and Methods" (DOI: 10.1080/03610926.2019.1676442).

Abstract
This paper investigates properties of the maximum likelihood estimator (MLE) in the zero-inflated Poisson regression model with randomly right-censored counts. Consistency and asymptotic normality of the MLE are rigorously established. A thorough simulation study is conducted to assess the finite-sample behavior of the MLE.

Keywords: Count data, excess of zeros, large-sample properties, simulations

Abstract in french

Dans ce chapitre, nous nous intéressons au modèle de régression de Poisson à inflation de zéro (modèle ZIP) dans un contexte de données censurées. Plus précisément, nous considérons la question de l'estimation et de l'inférence statistique dans le modèle de régression ZIP en présence de données de comptage censurées aléatoirement à droite. Nous établissons rigoureusement les propriétés asymptotiques (consistance, normalité asymptotique, estimation convergente de la variance asymptotique) de l'estimateur du maximum de vraisemblance des paramètres du modèle, dans cette situation de censure. Ce premier travail répond à une question théorique ouverte depuis la publication de Saffari et Adnan (2011) qui les premiers, avaient proposé d'utiliser le

modèle de régression ZIP dans un cadre censuré. Nos arguments utilisent des schémas de preuves dus à Czado et Min (2005), qui ont établi les propriétés asymptotiques de l'estimateur du maximum de vraisemblance dans le modèle ZIP non censuré, et à Fahrmeir et Kaufmann (1985), qui ont démontré la consistance et la normalité asymptotique de l'estimateur du maximum de vraisemblance dans les modèles linéaires généralisés, mais nous en proposons des simplifications, pour rendre les preuves plus directes.

Nous réalisons ensuite une étude de simulation approfondie, afin d'évaluer les propriétés à distance finie de l'estimateur du maximum de vraisemblance, en fonction de divers paramètres : taille d'échantillon, proportion d'inflation de zéro, proportion de censure. Les différents indicateurs numériques et représentations graphiques que nous obtenons confirment la qualité de l'estimateur proposé.

Enfin, nous décrivons une application de la méthodologie proposée à un jeu de données réelles issues du domaine de l'économie de la santé. Nous travaillons à partir d'une base de données recueillie en Allemagne et renseignant la consommation de soins de 1812 hommes, pour lesquels sont également disponibles, sous formes de variables explicatives, de nombreuses informations relatives à leur état de santé et à leur situation socio-économique. Nous identifions les déterminants du recours aux soins de ces hommes, ainsi que les variables qui influencent le nombre moyen de consultations médicales réalisées par ces hommes.

3.1 Introduction

Overdispersion is a major issue of count data analysis. Two main causes of overdispersion are excess variation between counts and zero-inflation. The first case is generally addressed using generalized Poisson models or negative binomial models, while zero-inflated models provide a useful approach when overdispersion is caused by an excess of zeros. We refer the interested reader to Hilbe (2011) and Cameron and Trivedi (2013) for a detailed treatment of overdispersion and negative binomial models. A detailed account of zero-inflation is given by Dupuy (2018). In this article, we focus on zero-inflated regression models, which mix a degenerate distribution with point mass of one at zero with a standard count regression model.

The zero-inflated Poisson (ZIP) regression model was first proposed by Lambert (1992) and was further developed to accommodate random effects (Hall, 2000; Min and Agresti, 2005; Monod, 2014), non-linear covariate effects (Lam et al., 2006; He et al., 2010), longitudinal zero-inflated counts Feng and Zhu (2011), among others. A zero-inflated negative binomial (ZINB) regression model was proposed by Ridout et al. (2001), see also Moghimbeigi et al. (2008) and Mwalili et al. (2008). When counts have an upper bound, ZIP and ZINB regression models are no longer appropriate and Hall (2000) introduced the zero-inflated binomial (ZIB) model, see also Hall and Berenhaut (2002), Diop et al. (2011, 2016) and Diallo et al. (2017a). Deng and Zhang (2015) proposed a zero-one inflated binomial regression model for bounded count data with both extra zeros and extra right-endpoints, see also Tian et al. (2015) and Dupuy (2017). Diallo et al. (2018) proposed a zero-inflated regression model for multinomial counts with joint zero-inflation.

Zero-inflated models have mainly been developed for complete data. However, although censoring is essentially associated to lifetime data analysis, count data can also be censored, the most common type being right-censoring (which occurs when it is only known that the true count is higher than the observed one). For example, consider a healthcare utilization study where patients report their number of visits to a doctor during a given period. If one possible answer is, say, "15 visits or more", all visit counts greater than 15 are right-censored at 15. Ignoring censoring may yield biased estimates and thus, incorrect inferences. The literature on

count data analysis already contains several approaches for handling right-censored counts in Poisson and generalized Poisson regressions (Terza, 1985; Caudill and Franklin G, 1995; Famoye and Wang, 2004; Xie and Wei, 2007; M. Mahmoud and M. Alderiny, 2010) and in finite mixtures of Poisson regressions Karlis et al. (2016). Much less work has been done in the case of right-censored count data with excess of zeros. Fu et al. (2018) investigate maximum likelihood estimation (MLE) in the univariate ZIP model with grouped and right-censored counts (this data structure arises when counts are reported as categories, such as "none", "1 to 2 times", "3 to 5 times", "6 or more times", rather than as exact numbers). Authors establish the consistency, asymptotic normality and asymptotic efficiency of the MLE.
Using simulations, Saffari and Adnan (2011) investigate MLE in a right-censored ZIP regression model. But this study is rather restricted since the simulation model involves few predictors and fixed censoring only. Moreover, no results are given about the finite-sample distribution of the MLE or about its theoretical properties (consistency, asymptotic normality, variance estimation). The aim of this paper is to fill this gap. We develop the asymptotic theory in ZIP regression with randomly right-censored data and we conduct a thorough simulation study to investigate the finite-sample behavior of the MLE.

The paper is organized as follows. In Section 3.2, we briefly review the censored zero-inflated Poisson regression model and we introduce some useful notations. In Section 3.3, we propose rigorous proofs of consistency and asymptotic normality of the MLE. In Section 3.4, we undertake a thorough simulation study to assess the finite-sample behaviour of the MLE. In the simulation study reported by Saffari and Adnan (2011), censoring was fixed at a constant value and the model had only one predictor for zero inflation and two predictors for Poisson mean. In our study, we consider random censoring, a larger number of predictors and we investigate the finite-sample distribution of the MLE. An application is illustrated in healthcare demand in Section 3.5. A discussion and some perspectives are provided in Section 3.6.

3.2 The censored ZIP regression model

In this section, we briefly recall the definition of the ZIP model and we describe maximum likelihood estimation when the count response is randomly right censored.

3.2.1 Maximum likelihood estimation

Let Y_i denote the count of some event for an individual i $(i = 1, \ldots, n)$ and $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})^\top$ and $\mathbf{Z}_i = (1, Z_{i2}, \ldots, Z_{iq})^\top$ be respectively p and q-dimensional vectors of covariates for this individual. Both categorical and continuous covariates are allowed. Moreover, $\mathbf{X}_i$ and $\mathbf{Z}_i$ may share some common terms or be distinct. A ZIP model specifies the distribution of Y_i as the mixture

$$Y_i \sim \begin{cases} 0 & \text{with probability } \omega_i, \\ \mathcal{P}(\lambda_i) & \text{with probability } 1 - \omega_i, \end{cases} \tag{3.2.1}$$

where $\mathcal{P}(\lambda_i)$ denotes the Poisson model with parameter $\lambda_i > 0$ and $0 \leqslant \omega_i \leqslant 1$ is some probability. Obviously, model (3.2.1) reduces to a standard Poisson model if $\omega_i = 0$. In ZIP regression, the mixing probability ω_i and parameter λ_i are usually modeled by logistic and log-linear models respectively, that is:

$$\text{logit}(\omega_i(\gamma)) = \gamma^\top \mathbf{Z}_i, \tag{3.2.2}$$

and

$$\log(\lambda_i(\beta)) = \beta^\top \mathbf{X}_i, \tag{3.2.3}$$

where $\beta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}^q$ are unknown regression parameters and $\top$ denotes the transpose operator.

Assume that we observe n independent vectors $(Y_1, \mathbf{X}_1, \mathbf{Z}_1), \ldots, (Y_n, \mathbf{X}_n, \mathbf{Z}_n)$ from the model (3.2.1)-(3.2.2)-(3.2.3), all defined on the probability space $(\Omega, \mathcal{C}, \mathbb{P})$. Based on these observations, the log-likelihood of (β, γ) can be written as:

$$\sum_{i=1}^{n} \left\{ 1_{\{Y_i=0\}} \log\left(e^{\gamma^\top \mathbf{Z}_i} + e^{-\exp(\beta^\top \mathbf{X}_i)}\right) + 1_{\{Y_i>0\}} \left(Y_i \beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} - \log(Y_i!)\right) - \log\left(1 + e^{\gamma^\top \mathbf{Z}_i}\right) \right\}. \tag{3.2.4}$$

The maximum likelihood estimator of (β, γ) is obtained by maximizing this function. It is consistent and asymptotically normally distributed (see Czado et al., 2007).

Assume now that the count response Y_i can be randomly right-censored. That is, for some individuals, we only observe a lower bound on Y_i. This can be modeled by introducing a censoring random variable C_i and by defining the observation for the i-th individual as the vector $(Y_i^*, \delta_i, \mathbf{X}_i, \mathbf{Z}_i)$, where $Y_i^* = \min(Y_i, C_i)$ and $\delta_i = 1_{\{Y_i<C_i\}}$ (if $Y_i = C_i$, we let $Y_i^* = C_i$ and $\delta_i = 0$). Thus, the censoring value can be specific to each observation. Let $J_i = 1_{\{Y_i^*=0\}}$. Based on observations $(Y_i^*, \delta_i, \mathbf{X}_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$, the likelihood of $\psi := (\beta^\top, \gamma^\top)^\top$ now writes as (see Saffari and Adnan, 2011):

$$\begin{aligned}
L_n(\psi) &= \prod_{i=1}^{n} \mathbb{P}(Y_i = Y_i^* | \mathbf{X}_i, \mathbf{Z}_i)^{\delta_i} \mathbb{P}(Y_i \geqslant Y_i^* | \mathbf{X}_i, \mathbf{Z}_i)^{1-\delta_i}, \\
&= \prod_{i=1}^{n} \left(\mathbb{P}(Y_i = Y_i^* | \mathbf{X}_i, \mathbf{Z}_i)^{1-J_i} \mathbb{P}(Y_i = 0 | \mathbf{X}_i, \mathbf{Z}_i)^{J_i}\right)^{\delta_i} \mathbb{P}(Y_i \geqslant Y_i^* | \mathbf{X}_i, \mathbf{Z}_i)^{(1-\delta_i)(1-J_i)}, \\
&= \prod_{i=1}^{n} \left(\left(e^{-\lambda_i} \frac{\lambda_i^{Y_i^*}}{Y_i^*!} (1-\omega_i) \right)^{1-J_i} \left(\omega_i + (1-\omega_i) e^{-\lambda_i}\right)^{J_i} \right)^{\delta_i} \\
&\qquad \times \left(1 - \sum_{k=0}^{Y_i^*-1} e^{-\lambda_i} \frac{\lambda_i^k}{k!} (1-\omega_i) - \omega_i \right)^{(1-\delta_i)(1-J_i)},
\end{aligned}$$

from which we easily obtain the loglikelihood $\ell_n(\psi) = \log L_n(\psi)$. If ω_i and λ_i are given by (3.2.2) and (3.2.3), straightforward algebra yields:

$$\begin{aligned}
\ell_n(\psi) = \sum_{i=1}^{n} \Bigg\{ \delta_i \left[J_i \log\left(e^{\gamma^\top \mathbf{Z}_i} + e^{-\exp(\beta^\top \mathbf{X}_i)}\right) + (1-J_i)\left(Y_i^* \beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} - \log(Y_i^*!)\right) \right] \\
+ (1-\delta_i)(1-J_i) \log\left(1 - \sum_{k=0}^{Y_i^*-1} \frac{e^{-\exp(\beta^\top \mathbf{X}_i) + k\beta^\top \mathbf{X}_i}}{k!} \right) - \log\left(1 + e^{\gamma^\top \mathbf{Z}_i}\right) \Bigg\}.
\end{aligned}$$

Note that $\ell_n(\psi)$ reduces to (3.2.4) when there is no censoring (that is, when $\delta_i = 1$ for all $i = 1, \ldots, n$).

The maximum likelihood estimator $\hat{\psi}_n := (\hat{\beta}_n^\top, \hat{\gamma}_n^\top)^\top$ of ψ solves the k-dimensional score equation

$$\frac{\partial \ell_n(\psi)}{\partial \psi} = 0, \tag{3.2.5}$$

where $k = p+q$. In the next section, we establish existence, consistency and asymptotic normality of $\hat{\psi}_n$. First, we need to introduce some further notations.

3.2.2 Some additional notations

In what follows, we note $k_i(\gamma) = e^{\gamma^\top \mathbf{z}_i}$ and $L_i(\beta) = e^{-\exp(\beta^\top \mathbf{x}_i)}$, $i = 1, \ldots, n$. Let also $S_{\lambda_i(\beta)}(u) = \mathbb{P}(\mathcal{P}(\lambda_i(\beta)) \geqslant u)$, $u = 0, 1, \ldots$ denote the survival function of $\mathcal{P}(\lambda_i(\beta))$ distribution. We have:

$$\frac{\partial \ell_n(\psi)}{\partial \beta_\ell} = \sum_{i=1}^{n} X_{i\ell} \left(-\delta_i J_i \frac{\lambda_i(\beta) L_i(\beta)}{k_i(\gamma) + L_i(\beta)} + \delta_i (1 - J_i) \left(Y_i^* - \lambda_i(\beta) \right) \right. \tag{3.2.6}$$
$$\left. -(1 - \delta_i)(1 - J_i) \sum_{k=0}^{Y_i^* - 1} \frac{L_i(\beta) \lambda_i^k(\beta) (k - \lambda_i(\beta))}{k!\, S_{\lambda_i(\beta)}(Y_i^*)} \right), \quad \ell = 1, \ldots, p,$$

and

$$\frac{\partial \ell_n(\psi)}{\partial \gamma_\ell} = \sum_{i=1}^{n} Z_{i\ell} \left(\frac{\delta_i J_i k_i(\gamma)}{k_i(\gamma) + L_i(\beta)} - \frac{k_i(\gamma)}{k_i(\gamma) + 1} \right), \quad \ell = 1, \ldots, q. \tag{3.2.7}$$

Let

$$u_i(\psi) = \frac{\lambda_i(\beta) L_i(\beta)}{(k_i(\gamma) + L_i(\beta))^2} \left[k_i(\gamma) + L_i(\beta) - \lambda_i(\beta) k_i(\gamma) \right], \quad i = 1, \ldots, n,$$

and for $Y_i^* \geqslant 1$, let

$$v_i(\psi) = \sum_{k=0}^{Y_i^* - 1} \frac{L_i(\beta) \lambda_i^k(\beta)}{k!\, S^2_{\lambda_i(\beta)}(Y_i^*)} \left\{ S_{\lambda_i(\beta)}(Y_i^*) \left((\lambda_i(\beta) - k)^2 - \lambda_i(\beta) \right) \right.$$
$$\left. -\lambda_i(\beta)(k - \lambda_i(\beta)) \mathbb{P}(\mathcal{P}(\lambda_i(\beta)) = Y_i^* - 1) \right\}, \quad i = 1, \ldots, n.$$

Then, some tedious albeit not difficult algebra shows that

$$\frac{\partial^2 \ell_n(\psi)}{\partial \beta_\ell \partial \beta_m} = \sum_{i=1}^{n} X_{i\ell} X_{im} \left\{ -\delta_i J_i u_i(\psi) - \delta_i (1 - J_i) \lambda_i(\beta) - (1 - \delta_i)(1 - J_i) v_i(\psi) \right\}, \quad \ell, m = 1, \ldots, p$$
$$\frac{\partial^2 \ell_n(\psi)}{\partial \beta_\ell \partial \gamma_m} = \sum_{i=1}^{n} X_{i\ell} Z_{im} \frac{\delta_i J_i k_i(\gamma) \lambda_i(\beta) L_i(\beta)}{(k_i(\gamma) + L_i(\beta))^2}, \quad \ell = 1, \ldots, p \text{ and } m = 1, \ldots, q$$
$$\frac{\partial^2 \ell_n(\psi)}{\partial \gamma_\ell \partial \gamma_m} = \sum_{i=1}^{n} Z_{i\ell} Z_{im} k_i(\gamma) \left(\frac{\delta_i J_i L_i(\beta)}{(k_i(\gamma) + L_i(\beta))^2} - \frac{1}{(k_i(\gamma) + 1)^2} \right), \quad \ell, m = 1, \ldots, q.$$

We note $S_n(\psi) = \partial \ell_n(\psi)/\partial \psi$, $H_n(\psi) = -\partial^2 \ell_n(\psi)/\partial \psi \partial \psi^\top$, $F_n(\psi) = \mathbb{E}(H_n(\psi))$ and I_k the identity matrix of order k. $H_n(\psi)$ is assumed positive definite.

3.3 Asymptotic results

In this section, we establish consistency and asymptotic normality of $\hat{\psi}_n$. In what follows, the space $\mathbb{R}^k$ of k-dimensional vectors is provided with the Euclidean norm $\|\cdot\|_2$ and the space of $(k \times k)$ real matrices is provided with the norm $|||A|||_2 := \sup_{\|x\|_2 = 1} \|Ax\|_2$ (for notations simplicity, we use $\|\cdot\|$ for both norms). Recall that for a symmetric real $(k \times k)$-matrix A with eigenvalues $\lambda_1, \ldots, \lambda_k$, $\|A\| = \max_i |\lambda_i|$ (from now on, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ will denote the smallest and largest eigenvalues of A respectively).

We first state some regularity conditions:

C1 Covariates are bounded, that is, there exist compact sets $\mathcal{X} \subset \mathbb{R}^p$ and $\mathcal{Z} \subset \mathbb{R}^q$ such that $\mathbf{X}_i \in \mathcal{X}$ and $\mathbf{Z}_i \in \mathcal{Z}$ for every $i = 1, 2, \ldots$

C2 The true parameter value $\psi_0 = (\beta_0^\top, \gamma_0^\top)^\top$ lies in the interior of some known compact and convex set $\mathcal{C} = \mathcal{B} \times \mathcal{G} \subset \mathbb{R}^k$ (where $\mathcal{B} \subset \mathbb{R}^p$ and $\mathcal{G} \subset \mathbb{R}^q$ are the parameter spaces of β and γ respectively).

C3 There exists a positive constant c_1 such that $n/\lambda_{\min}(F_n(\psi_0)) \leqslant c_1$ for every $n = 1, 2, \ldots$

C4 Censoring random variables $C_i, i = 1, 2, \ldots$ are strictly positive and bounded by some constant $M < \infty$.

Conditions C1-C3 are classical in generalized linear regression and zero-inflated regression models (see Fahrmeir and Kaufmann, 1985; Czado et al., 2007). Condition C4 is required in the censored setting. In the particular case of fixed censoring, this condition is obviously satisfied.

For each $n = 1, 2, \ldots$ and $\varepsilon > 0$, define the neighbourhood $N_n(\varepsilon) = \{\psi \in \mathcal{C} : (\psi - \psi_0)^\top F_n(\psi - \psi_0) \leqslant \varepsilon^2\}$ of ψ_0, where F_n is a short notation for $F_n(\psi_0)$. Our first result states that the solution of (3.2.5) exists, lies in the neighbourhood $N_n(\varepsilon)$ of ψ_0 when n is sufficiently large and is consistent for ψ_0.

Theorem 3.3.1 (Existence and consistency)**.** *Assume conditions C1-C4 hold. Then the probability that $\hat{\psi}_n$ exists and lies in $N_n(\varepsilon)$ for some ε tends to 1 as $n \to \infty$. Furthermore, $\hat{\psi}_n$ converges in probability to ψ_0 as $n \to \infty$.*

Proof of Theorem 3.3.1. Our proof uses some arguments employed by Fahrmeir and Kaufmann (1985) for proving the asymptotics in generalized linear models. But we also rely on different arguments in several parts of our demonstrations, leading to more direct proofs. A technical lemma is proved in an Appendix.

Asymptotic existence of $\hat{\psi}_n$. We show that for every $\eta > 0$, there exists $\varepsilon > 0$ and $n_1 \in \mathbb{N}$ such that

$$\mathbb{P}\left(\ell_n(\psi) - \ell_n(\psi_0) < 0 \text{ for all } \psi \in \partial N_n(\varepsilon)\right) \geqslant 1 - \eta, \qquad \text{for } n \geqslant n_1, \tag{3.3.1}$$

where $\partial N_n(\varepsilon)$ is the boundary $\{\psi \in \mathcal{C} : (\psi - \psi_0)^\top F_n(\psi - \psi_0) = \varepsilon^2\}$ of $N_n(\varepsilon)$. This will imply the existence of a local maximum of ℓ_n in $N_n(\varepsilon)$. Positive-definiteness of H_n and convexity of $\mathcal{C}$ will ensure that this maximum is global and unique.

In fact, equivalently to (3.3.1), we show that for every $\eta > 0$, there exists $\varepsilon > 0$ and $n_1 \in \mathbb{N}$ such that

$$\mathbb{P}\left(\ell_n(\psi) - \ell_n(\psi_0) \geqslant 0 \text{ for some } \psi \in \partial N_n(\varepsilon)\right) \leqslant \eta,$$

for $n \geqslant n_1$. To see this, we use Taylor's expansion to write

$$\begin{aligned} \ell_n(\psi) - \ell_n(\psi_0) &= (\psi - \psi_0)^\top S_n(\psi_0) - \frac{1}{2}(\psi - \psi_0)^\top H_n(\tilde{\psi})(\psi - \psi_0), \\ &:= (\psi - \psi_0)^\top S_n(\psi_0) - Q_n(\psi), \end{aligned}$$

where $\tilde{\psi} = a\psi + (1-a)\psi_0$ (for some $0 \leqslant a \leqslant 1$) lies between ψ and ψ_0. Let $0 < c < \frac{1}{2}$ and write f.s. for "for some". Then we have:

$$\begin{aligned} &\mathbb{P}\left(\ell_n(\psi) - \ell_n(\psi_0) \geqslant 0, \text{ f.s. } \psi \in \partial N_n(\varepsilon)\right) \\ &\quad = \mathbb{P}\left((\psi - \psi_0)^\top S_n(\psi_0) \geqslant Q_n(\psi) \text{ and } Q_n(\psi) > c\varepsilon^2, \text{ f.s. } \psi \in \partial N_n(\varepsilon)\right) \\ &\qquad\qquad + \mathbb{P}\left((\psi - \psi_0)^\top S_n(\psi_0) \geqslant Q_n(\psi) \text{ and } Q_n(\psi) \leqslant c\varepsilon^2, \text{ f.s. } \psi \in \partial N_n(\varepsilon)\right), \\ &\quad \leqslant \mathbb{P}(A) + \mathbb{P}(B), \end{aligned}$$

where A and B denote events $A = \{(\psi-\psi_0)^\top S_n(\psi_0) > c\varepsilon^2, \text{ f.s. } \psi \in \partial N_n(\varepsilon)\}$ and $B = \{Q_n(\psi) \leqslant c\varepsilon^2, \text{ f.s. } \psi \in \partial N_n(\varepsilon)\}$ respectively. Let $u_n(\psi) = \frac{1}{\varepsilon} F_n^{\frac{1}{2}}(\psi - \psi_0)$. Then

$$\begin{aligned} A &= \{u_n(\psi)^\top F_n^{-\frac{1}{2}} S_n(\psi_0) > c\varepsilon, \text{ f.s. } \psi \in \partial N_n(\varepsilon)\}, \\ &\subseteq \{\sup_{\psi \in \partial N_n(\varepsilon)} |u_n(\psi)^\top F_n^{-\frac{1}{2}} S_n(\psi_0)| > c\varepsilon\}, \\ &\subseteq \{\sup_{\|u_n(\psi)\|=1} |u_n(\psi)^\top F_n^{-\frac{1}{2}} S_n(\psi_0)| > c\varepsilon\}, \\ &= \{\|F_n^{-\frac{1}{2}} S_n(\psi_0)\| > c\varepsilon\}. \end{aligned}$$

where the second to third line comes from the fact that $\psi \in \partial N_n(\varepsilon)$ implies $\|u_n(\psi)\| = 1$. It follows that $\mathbb{P}(A) \leqslant \mathbb{P}(\|F_n^{-\frac{1}{2}} S_n(\psi_0)\| > c\varepsilon)$. By Theorem 1.5 of George A. F. Seber (2003), $\mathbb{E}\|F_n^{-\frac{1}{2}} S_n(\psi_0)\|^2 = k$ and Chebyshev's inequality implies

$$\mathbb{P}(A) \leqslant \frac{k}{c^2\varepsilon^2}.$$

Finally, letting $\varepsilon = \sqrt{\frac{2k}{\eta c^2}}$ implies that $\mathbb{P}(A) \leqslant \eta/2$. Now,

$$\begin{aligned} B &= \left\{\frac{1}{2}(\psi - \psi_0)^\top H_n(\tilde{\psi})(\psi - \psi_0) \leqslant c\varepsilon^2, \text{ f.s. } \psi \in \partial N_n(\varepsilon)\right\}, \\ &= \left\{\frac{1}{2} u_n(\psi)^\top F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}} u_n(\psi) \leqslant c, \text{ f.s. } \psi \in \partial N_n(\varepsilon)\right\}, \\ &\subseteq \left\{\frac{1}{2}\lambda_{\min}\left(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}}\right) u_n(\psi)^\top u_n(\psi) \leqslant c, \text{ f.s. } \psi \in \partial N_n(\varepsilon)\right\}, \\ &= \left\{\frac{1}{2}\lambda_{\min}\left(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}}\right) \leqslant c, \text{ f.s. } \psi \in \partial N_n(\varepsilon)\right\}. \end{aligned}$$

Thus, $\mathbb{P}(B) \leqslant \mathbb{P}(\text{there exists } \psi \in \partial N_n(\varepsilon) \text{ such that } \lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}}) \leqslant 2c)$. By Lemma 3.7.1 in Appendix, $F_n^{-\frac{1}{2}} H_n(\psi) F_n^{-\frac{1}{2}}$ converges in probability to I_k uniformly in $\psi \in N_n(\varepsilon)$, as $n \to \infty$. Thus, by Maller (2003), $\lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\psi) F_n^{-\frac{1}{2}})$ converges in probability to 1 uniformly in $\psi \in N_n(\varepsilon)$, as $n \to \infty$.
If $\tilde{\psi} = a\psi + (1-a)\psi_0$ for some $0 \leqslant a \leqslant 1$ and $\psi \in N_n(\varepsilon)$, then

$$\begin{aligned} \|F_n^{\frac{1}{2}}(\tilde{\psi} - \psi_0)\| &= \|F_n^{\frac{1}{2}}(a\psi + (1-a)\psi_0 - \psi_0)\|, \\ &= a\|F_n^{\frac{1}{2}}(\psi - \psi_0)\|, \\ &\leqslant \|F_n^{\frac{1}{2}}(\psi - \psi_0)\|, \\ &\leqslant \varepsilon, \end{aligned}$$

and thus $\tilde{\psi} \in N_n(\varepsilon)$. It follows that $\lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}})$ converges in probability to 1 as $n \to \infty$, since $|\lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}}) - 1| \leqslant \sup_{\psi \in N_n(\varepsilon)} |\lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\psi) F_n^{-\frac{1}{2}}) - 1|$. Therefore, for n sufficiently large (say, $n \geqslant n_1$), $\mathbb{P}(\text{there exists } \psi \in \partial N_n(\varepsilon) \text{ such that } \lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}}) \leqslant 2c) \leqslant \eta/2$, since $2c < 1$. This implies that $\mathbb{P}(B) \leqslant \eta/2$. Finally,

$$\mathbb{P}\left(\ell_n(\psi) - \ell_n(\psi_0) \geqslant 0, \text{ f.s. } \psi \in \partial N_n(\varepsilon)\right) \leqslant \mathbb{P}\left(A\right) + \mathbb{P}\left(B\right) \leqslant \eta,$$

which proves (3.3.1) and in turn, the existence of a unique global maximum of ℓ_n on $N_n(\varepsilon)$, which coincides with $\hat{\psi}_n$.

Consistency of $\hat{\psi}_n$. We have:

$$\begin{aligned}\lambda_{\min}(F_n)\|\hat{\psi}_n-\psi_0\|^2 &= (\hat{\psi}_n-\psi_0)^\top\lambda_{\min}(F_n)I_k(\hat{\psi}_n-\psi_0),\\ &\leqslant (\hat{\psi}_n-\psi_0)^\top F_n(\hat{\psi}_n-\psi_0),\\ &= \|F_n^{\frac{1}{2}}(\hat{\psi}_n-\psi_0)\|^2,\\ &\leqslant \varepsilon^2,\end{aligned}$$

with probability tending to 1 as $n\to\infty$, by i). By condition C3, $\lambda_{\min}(F_n)$ tends to ∞ as $n\to\infty$. Therefore $\|\hat{\psi}_n-\psi_0\|$ converges to 0 with probability tending to 1 as $n\to\infty$, which concludes the proof. □

Our second result is:

Theorem 3.3.2 (Asymptotic normality)**.** *Assume conditions C1-C4 hold. Then $F_n^{\frac{1}{2}}(\hat{\psi}_n-\psi_0)$ converges in distribution to the Gaussian vector $\mathcal{N}(0,I_k)$, as $n\to\infty$.*

Proof of Theorem 3.3.2. Czado and Min (2005) prove asymptotic normality of the MLE in the uncensored zero-inflated generalized Poisson model by using a central limit theorem with Lyapunov condition. Here, we rely on the weaker Lindeberg condition, which yields a much shorter proof.

We first prove asymptotic normality of the normalized score vector $F_n^{-\frac{1}{2}}S_n$, where S_n is a short notation for $S_n(\psi_0)$. Let u be any vector in $\mathbb{R}^k$. We show that $u^\top F_n^{-\frac{1}{2}}S_n$ converges in distribution to $\mathcal{N}(0,u^\top u)$ (without loss of generality, we set $\|u\|=1$). From (3.2.6) and (3.2.7), we remark that S_n can be written as a sum $S_n=\sum_{i=1}^n S_{n,i}$ of independent k-dimensional random vectors $S_{n,i}=(S_{n,i,1},\ldots,S_{n,i,k})^\top$. It is not difficult to see that under conditions C1, C2 and C4, components of $S_{n,i}$ are bounded by some finite positive constant c_2 that is, $|S_{n,i,\ell}|<c_2$, $\ell=1,\ldots,k$. Therefore, $\|S_{n,i}\|^2<c_3:=kc_2^2$.

Let

$$u^\top F_n^{-\frac{1}{2}}S_n=u^\top F_n^{-\frac{1}{2}}\sum_{i=1}^n S_{n,i}:=\sum_{i=1}^n S_{n,i}^*.$$

Then $\mathbb{E}(S_{n,i}^*)=0$ and $\mathrm{var}(\sum_{i=1}^n S_{n,i}^*)=1$. We now verify Lindeberg condition, namely:

$$\text{for every } \varepsilon>0,\ \sum_{i=1}^n\mathbb{E}\left(S_{n,i}^{*2}1_{\{|S_{n,i}^*|>\varepsilon\}}\right)\to 0 \text{ as } n\to\infty.$$

Let $\varepsilon>0$. We have:

$$\begin{aligned}\sum_{i=1}^n\mathbb{E}\left(S_{n,i}^{*2}1_{\{|S_{n,i}^*|>\varepsilon\}}\right) &\leqslant \sum_{i=1}^n\mathbb{E}\left(\|u\|^2\|F_n^{-\frac{1}{2}}\|^2\|S_{n,i}\|^2 1_{\{|S_{n,i}^*|>\varepsilon\}}\right),\\ &\leqslant \frac{c_1c_3}{n}\sum_{i=1}^n\mathbb{E}(1_{\{|S_{n,i}^*|>\varepsilon\}}),\end{aligned}$$

by condition C3. Now, $\{|S_{n,i}^*|>\varepsilon\}$ implies that $\{\lambda_{\min}(F_n)<c_3/\varepsilon^2\}$, therefore, $1_{\{|S_{n,i}^*|>\varepsilon\}}\leqslant 1_{\{\lambda_{\min}(F_n)<c_3/\varepsilon^2\}}$ and thus,

$$\sum_{i=1}^n\mathbb{E}\left(S_{n,i}^{*2}1_{\{|S_{n,i}^*|>\varepsilon\}}\right)\leqslant\frac{c_1c_3}{n}\sum_{i=1}^n 1_{\{\lambda_{\min}(F_n)<c_3/\varepsilon^2\}}=c_1c_3 1_{\{\lambda_{\min}(F_n)<c_3/\varepsilon^2\}}.$$

Under C3, $\lambda_{\min}(F_n) \to \infty$ as $n \to \infty$. Therefore, $\sum_{i=1}^{n} \mathbb{E}(S_{n,i}^{*2} 1_{\{|S_{n,i}^{*}|>\varepsilon\}}) \to 0$ as $n \to \infty$. It follows that for every $u \in \mathbb{R}^k$, $u^\top F_n^{-\frac{1}{2}} S_n$ converges in distribution to $\mathcal{N}(0,1)$ and by Cramer-Wold device, $F_n^{-\frac{1}{2}} S_n$ converges in distribution to $\mathcal{N}(0, I_k)$.

Weak convergence of $F_n^{\frac{1}{2}}(\hat{\psi}_n - \psi_0)$ is now obtained as usual, by expanding $S_n := S_n(\psi_0)$ about $\hat{\psi}_n$. The rest of the proof is similar to proof of Theorem 3 of Fahrmeir and Kaufmann (1985) and is thus omitted. □

In order to construct asymptotic confidence intervals and tests of hypothesis for the components of ψ, one needs to estimate F_n (by $H_n(\hat{\psi}_n)$ for example). For example, the standard error s.e.$(\hat{\beta}_{j,n})$ of the j-th component $\hat{\beta}_{j,n}$ of $\hat{\beta}_n$ can be obtained by taking the square root of the j-th diagonal term of $(H_n(\hat{\psi}_n))^{-1}$.

3.4 Simulation study

In this section, we assess finite-samples properties of the MLE under various scenarios obtained by varying the censoring and zero-inflation proportions and the sample size. As already mentioned, Saffari and Adnan (2011) report a simulation study investigating the behaviour of the MLE in the censored ZIP model. However, this study is rather restricted: the simulating model has only one predictor in the sub-model for zero inflation and two predictors in the Poisson mean. Moreover, Saffari and Adnan (2011) consider fixed censoring only and no results are given on the finite-sample distribution of the MLE nor on coverage probabilities of confidence intervals for model parameters.

3.4.1 Simulation design

We simulate the data according to the ZIP model (3.2.1)-(3.2.2)-(3.2.3) defined by:

$$\log(\lambda_i(\beta)) = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6},$$

and

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

where $X_{i1} = Z_{i1} = 1$ and the $X_{i2}, \ldots, X_{i6}, Z_{i4}, Z_{i5}$ are independently drawn from normal $\mathcal{N}(0,1)$, Bernoulli $\mathcal{B}(0.3)$, normal $\mathcal{N}(1, 2.25)$, exponential $\mathcal{E}(1)$, uniform $\mathcal{U}(2,5)$, normal $\mathcal{N}(-1,1)$ and Bernoulli $\mathcal{B}(0.5)$ distributions respectively. Linear predictors in $\log(\lambda_i(\beta))$ and $\text{logit}(\omega_i(\gamma))$ are allowed to share common terms by letting $Z_{i2} = X_{i2}$ and $Z_{i3} = X_{i3}$. We consider the following sample sizes: $n = 500, 1000, 2500$. The regression parameter β is chosen as $\beta = (0.7, 0.1, 0.4, 0.85, -0.5, 0)^\top$. The regression parameter γ is chosen as:

- case 1: $\gamma = (-0.9, -0.65, -0.2, 0.65, 0)^\top$,
- case 2: $\gamma = (0.25, -0.7, -0.2, 0.65, 0)^\top$.

Using these values, in case 1 (respectively case 2), the average percentage of zero-inflation in the simulated data sets is 20% (respectively 40%). Also, in case 1 (respectively case 2), the average percentage of zero is 34% (ranging from 30% to 37%) (respectively 50%, ranging from 47% to 53%). Random censoring values are simulated from a zero-truncated Poisson model with parameter μ, where μ is chosen to yield various average censoring proportions c in the simulated samples, namely $c = 0.1, 0.2, 0.4$. For purpose of comparison, we also provide results that would

be obtained if there were no censoring (that is, when $c = 0$) since these results will constitute a benchmark for assessing performance of the MLE when censoring is present.

For each combination of the simulation design parameters (sample size, proportions of censoring and zero-inflation), we simulate $N = 1000$ samples and we calculate the MLE $\hat{\psi}_n$. Simulations are carried out using the statistical software `R`. To solve the likelihood equation, we use the Newton-Raphson algorithm, implemented in the package `maxLik` (Henningsen and Toomet, 2011). We obtain starting values by estimating a ZIP model without taking censoring into account (this step is carried out using the function `zeroinfl` of the `R` package `pscl`, see Zeileis et al. (2008) and Jackman (2017)).

3.4.2 Results

For each configuration [`sample size` × `censoring proportion` × `zero-inflation proportion`] of the simulation parameters, we calculate the average bias and average relative bias (expressed as a percentage) of the estimates $\hat{\beta}_{j,n}$ and $\hat{\gamma}_{k,n}$ over the N simulated samples. For example, the relative bias of $\hat{\beta}_{j,n}$ is obtained as

$$\frac{1}{N}\sum_{t=1}^{N}\frac{\hat{\beta}_{j,n}^{(t)}-\beta_j}{\beta_j}\times 100,$$

where $\hat{\beta}_{j,n}^{(t)}$ denotes the MLE of β_j in the t-th simulated sample. We also obtain the average standard error (SE), empirical standard deviation (SD) and root mean square error (RMSE) for each $\hat{\beta}_{j,n}$ $(j = 1, \ldots, 6)$ and $\hat{\gamma}_{k,n}$ $(k = 1, \ldots, 5)$. SE is calculated as the average of the standard errors across the N simulated samples. For example, for $\hat{\beta}_{j,n}$, SE is calculated as $\frac{1}{N}\sum_{t=1}^{N}$ s.e.$(\hat{\beta}_{j,n}^{(t)})$, while SD (resp. RMSE) is the square root of the empirical variance (resp. the root mean square error) of $(\hat{\beta}_{j,n}^{(1)}, \ldots, \hat{\beta}_{j,n}^{(N)})$.

Finally, we provide the empirical coverage probability (CP) and average length of 95%-level confidence intervals for the β_j and γ_k. Results are given in Table 3.2 (case 1, $n = 500$), Table 3.3 (case 2, $n = 500$), Table 3.4 (case 1, $n = 1000$), Table 3.5 (case 2, $n = 1000$), Table 3.6 (case 1, $n = 2500$) and Table 3.7 (case 2, $n = 2500$) (note that the relative bias cannot be calculated for $\beta_6 = \gamma_5 = 0$, which is indicated by "-" in Tables 3.2-3.7).

We observe that the accuracy of MLEs of both β_j and γ_k decreases as sample size decreases. Accuracy of β_js estimates also decreases as censoring increases (note that the relative bias stays moderate though, even when censoring is high). On the contrary, estimates of the γ_k are rather insensitive to censoring, which can be explained as follows. The role of the zero-inflation submodel is to "separate" random and non-random zeros. Therefore, this model should not be affected by censoring since zero cannot be a censored observation, by condition C4. In fact, censoring acts indirectly on the MLEs of the γ_k (their variability slightly increases with censoring, as a consequence of the increasing variability of the MLEs of the β_j). For both β_j and γ_k, empirical coverage probabilities are close to the nominal confidence level in every case. For a given censoring proportion, we observe that MLEs of the β_j (respectively γ_k) perform better when the zero-inflation proportion decreases (respectively increases), which may be explained as follows. As zero-inflation decreases, there is more information in the data about the count sub-model and thus, one may expect MLEs of the β_j to perform better. On the other hand, as zero-inflation increases, there is more information in the sample about the zero-inflation sub-model, which allows a better estimation of the γ_k.

Finally, in order to assess quality of the Gaussian approximation stated in Theorem 3.3.2, we obtain normal Q-Q plots of the estimates and histograms of the normalized estimates $(\widehat{\beta}_{j,n}-\beta_j)/$

s.e.$(\widehat{\beta}_{j,n})$, $j = 1, \ldots, 6$ and $(\widehat{\gamma}_{k,n} - \gamma_k)/\text{s.e.}(\widehat{\gamma}_{k,n})$, $j = 1, \ldots, 5$. We provide these graphs for $n = 500$ with a proportion of zero-inflation equal to 0.4 and 40% of censoring (Figures 3.1 and 3.2 provide normal Q-Q plots for the $(\hat{\beta}_{1,n}, \ldots, \hat{\beta}_{6,n})$ and $(\hat{\gamma}_{1,n}, \ldots, \hat{\gamma}_{5,n})$ respectively; Figures 3.3 and 3.4 provide histograms of the normalized $(\hat{\beta}_{1,n}, \ldots, \hat{\beta}_{6,n})$ and $(\hat{\gamma}_{1,n}, \ldots, \hat{\gamma}_{5,n})$ respectively). Plots for the other (and more favorable) simulated scenarios yield similar observations and are thus not given. From these figures, it appears that the Gaussian approximation of the distribution of the MLE is reasonably satisfied, even when the sample size is moderate and the proportions of zero-inflation and censoring are as high as 0.4.

Sample size computation is an important topic in regression and several empirical rules have been suggested (e.g., events-per-variable criterion, one-in-ten rule). These rules should be regarded with caution here. Indeed, a ZIP model is a mixture model and several factors, interacting with one another (such as the respective proportions of each class in the mixture and number of covariates acting on each linear predictor), are susceptible to influence sample size calculation. The presence of censoring constitutes an additional source of complexity. For the model considered in this study, we observed that a minimum sample size of 200 is necessary to avoid convergence problems (results are provided in a document available at `http://dupuy.perso.math.cnrs.fr/research/Comm_in_Stats_supp.pdf` ; they yield similar observations as for larger sample sizes). Sample size computation in (censored) ZIP regression falls beyond the scope of this paper but constitutes a stimulating topic for future research.

3.5 An application in health economics: demand for physician service

We illustrate maximum likelihood estimation in a censored ZIP regression model on a dataset coming from the National Medical Expenditure Survey 1987-1988 (a large cross-sectional study carried out to assess the demand for medical care in USA, see Deb and K. Trivedi, 1997). Data are available in the `R` package `AER` (Christian Kleiber, 2008) under the name "NMES1988". They contain observations on 4,406 individuals aged 66 and over, all of whom are covered by Medicare (a federal health insurance program). The response variable is the number of visits to a physician in an office setting (denoted by `ofp` in what follows). Explanatory variables include demographics: gender (1 for female, 0 for male), age (in years, divided by 10), socio-economic variables: marital status (1 if married, 0 otherwise), educational level (number of years of education, denoted by `school`), family income (in ten-thousands of dollars), two binary variables indicating whether individual is covered by Medicaid (a US health insurance for individuals with low resources) and by a supplemental private insurance (both are coded as 1 if yes and 0 otherwise), various measures of health status: number of chronic conditions (cancer, arthritis, diabetes... denoted by `chronic`) and a variable indicating self-perceived health level (poor, average, excellent), which we recode as `health1` (1 if health is perceived as poor, 0 otherwise) and `health2` (1 health is perceived as excellent, 0 otherwise).

In the initial data, the response `ofp` is uncensored. But since NMES1988 has become a benchmark dataset for evaluating zero-inflated models, we choose to use it and to censor `ofp` artificially. We use a zero-truncated Poisson distribution to generate censoring values. Average sample

censoring proportions of `ofp` are successively set to 0.2 and 0.4. The fitted model is:

$$\begin{cases} \log(\lambda_i(\beta)) = \beta_1 + \beta_2\,\texttt{gender}_i + \beta_3\,\texttt{age}_i + \beta_4\,\texttt{marital\_status}_i + \beta_5\,\texttt{school}_i + \beta_6\,\texttt{income}_i \\ \qquad\qquad +\beta_7\,\texttt{medicaid}_i + \beta_8\,\texttt{insurance}_i + \beta_9\,\texttt{chronic}_i + \beta_{10}\,\texttt{health1}_i + \beta_{11}\,\texttt{health2}_i, \\ \\ \text{logit}(\omega_i(\gamma)) = \gamma_1 + \gamma_2\,\texttt{gender}_i + \gamma_3\,\texttt{age}_i + \gamma_4\,\texttt{marital\_status}_i + \gamma_5\,\texttt{school}_i + \gamma_6\,\texttt{income}_i \\ \qquad\qquad +\gamma_7\,\texttt{medicaid}_i + \gamma_8\,\texttt{insurance}_i + \gamma_9\,\texttt{chronic}_i + \gamma_{10}\,\texttt{health1}_i + \gamma_{11}\,\texttt{health2}_i. \end{cases}$$

Maximum likelihood estimates and standard errors for all model parameters are obtained in both censored and non-censored cases. Results are summarized in Table 4.4.

The no-censoring case was already investigated using various models (e.g., Cameron and Trivedi, 2013; Friendly, 2015; Diallo et al., 2018). Consistent with these studies, we find that both number of chronic conditions and self-perceived health are important determinants of `ofp` utilization. An increase in the number of chronic diseases increases both the probability of visiting a doctor and the average number of consultations. We find that measures of self-perceived health do not affect the decision of visiting a doctor (the effect of health status on this decision is entirely captured by `chronic`) but affect the average number of visits (this number increases as health perception degradates). Individuals with higher educational level are less susceptible to waive `ofp`-type health-care. They also seek care more often. At the same time, income is non-significant in both models for zero-inflation and visits frequency (in Deb and K. Trivedi (1997), authors formulate the hypothesis that the overall generosity of Medicare make `ofp` utilization insensitive to changes in the income). Also consistent with previous analysis, in the censored case, Medicaid and supplementary private insurance coverage are significant determinants of both decision of visiting a doctor and number of visits (as expected, covered individuals are less susceptible to waive `ofp`-type consultations, they also seek care more often).

As already observed in our simulations, parameters estimates and standard errors in $\omega_i(\gamma)$ are only slightly affected by censoring. Therefore, Wald significance tests agree whatever the censoring fraction. Parameters estimates and standard errors in the Poisson model part are more sensitive to censoring. Wald significance tests agree for all regressors except age, gender and marital status, where both increasing bias and standard errors resulting from censoring yield misleading conclusions.

Overall, MLEs in censored ZIP regression seem to be rather robust to censoring. In particular, estimation and variable selection in the zero-inflation model part are not affected by censoring (even when the censoring fraction is large).

3.6 Discussion

In this paper, we consider the randomly right-censored zero-inflated Poisson regression model. We establish rigorously the consistency and asymptotic normality of the MLE in this model. Our numerical study confirms the good performance of the MLE and suggests that estimation and variable selection in the zero-inflation sub-model part are not affected by censoring.
Now, several issues deserve attention. Random right-censoring is one of many possible censoring schemes. In practice, count data may also be left-censored or interval-censored. Estimation and inference in ZIP regression in these settings is currently an open question. Investigating estimation in more general zero-inflated models (such as the semi-parametric ZIP model for example) with censoring is also desirable. These issues constitute topics for our future work.

	no censoring		censoring = 20%		censoring = 40%	
variable	estimate (s.e.)	signif.	estimate (s.e.)	signif.	estimate (s.e.)	signif.
			Poisson model coefficients			
intercept	1.7596 (0.0878)	***	1.3491 (0.2057)	***	1.1464 (0.1246)	***
gender	0.0066 (0.0144)		0.0487 (0.0165)	**	0.0377 (0.0201)	.
age	-0.0593 (0.0107)	***	-0.0148 (0.0248)		-0.0102 (0.0152)	
marital status	-0.0796 (0.0148)	***	-0.0269 (0.0140)	.	-0.0072 (0.0209)	
school	0.0209 (0.0020)	***	0.0130 (0.0015)	***	0.0144 (0.0027)	***
income	-0.0014 (0.0023)		-0.0008 (0.0026)		-0.0038 (0.0032)	
medicaid	0.2256 (0.0254)	***	0.1799 (0.0297)	***	0.1681 (0.0369)	***
insurance	0.1892 (0.0200)	***	0.1387 (0.0228)	***	0.1638 (0.0274)	***
chronic	0.1198 (0.0046)	***	0.1150 (0.0056)	***	0.1191 (0.0070)	***
health1	0.3081 (0.0175)	***	0.2285 (0.0212)	***	0.1971 (0.0272)	***
health2	-0.3224 (0.0312)	***	-0.2532 (0.0338)	***	-0.2018 (0.0383)	***
			Zero-inflation model coefficients			
intercept	1.9786 (0.5917)	***	1.9597 (0.6870)	**	1.9709 (0.7531)	**
gender	-0.4721 (0.0974)	***	-0.4716 (0.0997)	***	-0.4840 (0.1020)	***
age	-0.1829 (0.0741)	*	-0.1818 (0.0862)	*	-0.1876 (0.0947)	*
marital status	-0.2843 (0.1033)	**	-0.2862 (0.1060)	**	-0.2921 (0.1087)	**
school	-0.0607 (0.0128)	***	-0.0614 (0.0131)	***	-0.0611 (0.0136)	***
income	-0.0104 (0.0188)		-0.0104 (0.0193)		-0.0120 (0.0201)	
medicaid	-0.4921 (0.1713)	**	-0.4928 (0.1718)	**	-0.4875 (0.1767)	**
insurance	-0.8227 (0.1102)	***	-0.8299 (0.1099)	***	-0.8272 (0.1145)	***
chronic	-0.5414 (0.0458)	***	-0.5393 (0.0464)	***	-0.5360 (0.0479)	***
health1	0.0124 (0.1615)		0.0212 (0.1605)		0.0380 (0.1669)	
health2	0.2447 (0.1505)		0.2398 (0.1594)		0.2410 (0.1575)	

Table 3.1: Health-care data analysis: estimates (standard errors) and significance codes: *** significant at the 0.1% level, ** significant at the 1% level, * significant at the 5% level, . significant at the 10% level.

Acknowledgements

Authors are grateful to the referees and associate editor for their comments and suggestions on an earlier version of this paper. Authors acknowledge financial support from the Ministry of Education and Training of the Socialist Republic of Vietnam and the French Embassy in Vietnam and logistical support from Campus France (French national agency for the promotion of higher education, international student services, and international mobility).

average proportion of censoring		$\widehat{\beta}_n$						$\widehat{\gamma}_n$				
		$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$	$\widehat{\gamma}_{5,n}$
0	bias	0.0029	-0.0013	0.0008	0.0005	-0.0011	-0.0014	-0.0241	-0.0191	0.0060	0.0180	-0.0077
	rel. bias	0.4189	-1.3293	0.2094	0.0598	0.2221	-	2.6778	2.9395	-3.0042	2.7756	-
	SD	0.0921	0.0196	0.0376	0.0136	0.0277	0.0215	0.2707	0.1615	0.3426	0.1586	0.3005
	SE	0.0878	0.0189	0.0374	0.0132	0.0276	0.0211	0.2571	0.1613	0.3258	0.1609	0.2973
	RMSE	0.1273	0.0273	0.0530	0.0189	0.0391	0.0301	0.3740	0.2290	0.4727	0.2266	0.4227
	CP	0.9500	0.9430	0.9480	0.9380	0.9510	0.9540	0.9400	0.9440	0.9400	0.9520	0.9540
	ℓ	0.3433	0.0740	0.1463	0.0512	0.1076	0.0824	1.0048	0.6298	1.2729	0.6285	1.1633
0.1	bias	-0.0049	-0.0009	0.0025	0.0041	-0.0015	-0.0009	-0.0278	-0.0198	0.0069	0.0192	-0.0073
	rel. bias	-0.7045	-0.9439	0.6349	0.4850	0.2904	-	3.0882	3.0520	-3.4358	2.9562	-
	SD	0.1263	0.0273	0.0548	0.0266	0.0347	0.0297	0.2725	0.1621	0.3448	0.1587	0.3014
	SE	0.1207	0.0266	0.0549	0.0273	0.0352	0.0296	0.2585	0.1619	0.3272	0.1613	0.2979
	RMSE	0.1747	0.0381	0.0776	0.0384	0.0494	0.0419	0.3765	0.2300	0.4753	0.2271	0.4237
	CP	0.9430	0.9390	0.9440	0.9590	0.9490	0.9530	0.9400	0.9490	0.9380	0.9540	0.9530
	ℓ	0.4725	0.1042	0.2149	0.1070	0.1377	0.1157	1.0101	0.6323	1.2781	0.6299	1.1654
0.2	bias	-0.0098	-0.0008	0.0024	0.0069	-0.0031	0.0002	-0.0292	-0.0203	0.0069	0.0198	-0.0077
	rel. bias	-1.4064	-0.8357	0.5971	0.8076	0.6201	-	3.2453	3.1206	-3.4564	3.0470	-
	SD	0.1500	0.0330	0.0697	0.0349	0.0419	0.0361	0.2731	0.1631	0.3475	0.1587	0.3013
	SE	0.1439	0.0327	0.0683	0.0355	0.0413	0.0361	0.2589	0.1624	0.3281	0.1614	0.2981
	RMSE	0.2081	0.0464	0.0976	0.0502	0.0589	0.0511	0.3773	0.2310	0.4778	0.2272	0.4238
	CP	0.9440	0.9420	0.9540	0.9460	0.9570	0.9520	0.9400	0.9500	0.9380	0.9540	0.9540
	ℓ	0.5635	0.1279	0.2676	0.1390	0.1615	0.1415	1.0118	0.6341	1.2815	0.6303	1.1661
0.4	bias	0.0016	0.0005	0.0050	0.0143	-0.0053	-0.0032	-0.0310	-0.0201	0.0062	0.0205	-0.0071
	rel. bias	0.2347	0.4716	1.2491	1.6786	1.0591	-	3.4448	3.0875	-3.0980	3.1517	-
	SD	0.2018	0.0503	0.0997	0.0534	0.0570	0.0522	0.2741	0.1652	0.3503	0.1593	0.3020
	SE	0.2063	0.0491	0.1042	0.0535	0.0567	0.0537	0.2601	0.1638	0.3312	0.1617	0.2984
	RMSE	0.2885	0.0703	0.1443	0.0769	0.0805	0.0749	0.3790	0.2335	0.4820	0.2278	0.4245
	CP	0.9520	0.9540	0.9610	0.9430	0.9540	0.9490	0.9450	0.9460	0.9370	0.9530	0.9500
	ℓ	0.8074	0.1922	0.4079	0.2092	0.2216	0.2101	1.0163	0.6395	1.2935	0.6312	1.1673

Table 3.2: Simulation results (n = 500, ZI proportion = 20%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals.

average proportion of censoring		$\widehat{\beta}_n$						$\widehat{\gamma}_n$				
		$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$	$\widehat{\gamma}_{5,n}$
0	bias	0.0021	-0.0006	0.0014	0.0001	0.0009	-0.0012	0.0051	-0.0166	-0.0089	0.0118	-0.0033
	rel. bias	0.3032	-0.5554	0.3613	0.0162	-0.1714	-	2.0303	2.3779	4.4416	1.8185	-
	SD	0.1045	0.0231	0.0442	0.0156	0.0334	0.0246	0.2234	0.1358	0.2535	0.1306	0.2442
	SE	0.1038	0.0226	0.0437	0.0158	0.0323	0.0247	0.2155	0.1315	0.2546	0.1294	0.2349
	RMSE	0.1473	0.0323	0.0621	0.0222	0.0465	0.0349	0.3103	0.1897	0.3593	0.1841	0.3388
	CP	0.9530	0.9500	0.9480	0.9530	0.9390	0.9520	0.9460	0.9530	0.9540	0.9460	0.9450
	ℓ	0.4056	0.0883	0.1708	0.0614	0.1260	0.0967	0.8438	0.5144	0.9973	0.5061	0.9205
0.1	bias	-0.0039	-0.0002	0.0014	0.0049	-0.0012	-0.0008	0.0030	-0.0172	-0.0094	0.0125	-0.0034
	rel. bias	-0.5502	-0.1754	0.3451	0.5726	0.2389	-	1.1946	2.4539	4.6880	1.9284	-
	SD	0.1496	0.0340	0.0699	0.0348	0.0458	0.0370	0.2237	0.1361	0.2540	0.1308	0.2443
	SE	0.1506	0.0341	0.0689	0.0357	0.0438	0.0370	0.2164	0.1320	0.2556	0.1296	0.2352
	RMSE	0.2123	0.0482	0.0981	0.0501	0.0634	0.0523	0.3112	0.1904	0.3604	0.1845	0.3391
	CP	0.9440	0.9490	0.9430	0.9560	0.9440	0.9440	0.9470	0.9540	0.9520	0.9470	0.9470
	ℓ	0.5894	0.1334	0.2698	0.1399	0.1709	0.1450	0.8475	0.5165	1.0012	0.5069	0.9216
0.2	bias	-0.0068	0.0016	0.0071	0.0095	-0.0041	-0.0004	0.0015	-0.0169	-0.0079	0.0130	-0.0036
	rel. bias	-0.9652	1.6303	1.7772	1.1197	0.8294	-	0.5822	2.4105	3.9740	1.9938	-
	SD	0.1915	0.0452	0.0887	0.0485	0.0556	0.0484	0.2240	0.1378	0.2549	0.1309	0.2443
	SE	0.1906	0.0448	0.0915	0.0489	0.0542	0.0483	0.2170	0.1326	0.2565	0.1297	0.2354
	RMSE	0.2702	0.0636	0.1276	0.0695	0.0777	0.0683	0.3118	0.1919	0.3616	0.1847	0.3392
	CP	0.9470	0.9490	0.9600	0.9530	0.9460	0.9480	0.9430	0.9550	0.9550	0.9460	0.9480
	ℓ	0.7457	0.1750	0.3579	0.1912	0.2117	0.1889	0.8496	0.5186	1.0047	0.5074	0.9222
0.4	bias	-0.0110	0.0078	0.0254	0.0305	-0.0154	0.0015	-0.0028	-0.0160	-0.0059	0.0147	-0.0039
	rel. bias	-1.5772	7.8277	6.3481	3.5830	3.0722	-	-1.1201	2.2840	2.9409	2.2563	-
	SD	0.3577	0.0937	0.1880	0.0926	0.0944	0.0944	0.2264	0.1408	0.2632	0.1316	0.2452
	SE	0.3550	0.0887	0.1850	0.0912	0.0924	0.0932	0.2191	0.1361	0.2627	0.1302	0.2359
	RMSE	0.5040	0.1293	0.2649	0.1335	0.1330	0.1327	0.3149	0.1964	0.3719	0.1857	0.3402
	CP	0.9520	0.9410	0.9490	0.9430	0.9480	0.9450	0.9490	0.9490	0.9530	0.9510	0.9470
	ℓ	1.3864	0.3456	0.7212	0.3555	0.3601	0.3643	0.8578	0.5321	1.0287	0.5092	0.9241

Table 3.3: Simulation results (n = 500, ZI proportion = 40%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals.

average proportion of censoring		$\widehat{\beta}_n$						$\widehat{\gamma}_n$				
		$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$	$\widehat{\gamma}_{5,n}$
0	bias	0.0008	0.0005	0.0006	0.0001	-0.0001	-0.0002	-0.0135	-0.0059	-0.0003	0.0032	0.0041
	rel. bias	0.1106	0.4539	0.1535	0.0141	0.0298	-	1.4961	0.9068	0.1458	0.4936	-
	SD	0.0605	0.0134	0.0253	0.0090	0.0198	0.0145	0.1719	0.1124	0.2299	0.1127	0.2031
	SE	0.0612	0.0131	0.0260	0.0090	0.0192	0.0147	0.1786	0.1116	0.2257	0.1118	0.2065
	RMSE	0.0860	0.0187	0.0362	0.0128	0.0276	0.0206	0.2482	0.1584	0.3221	0.1587	0.2896
	CP	0.9590	0.9470	0.9550	0.9540	0.9400	0.9540	0.9650	0.9500	0.9480	0.9510	0.9510
	ℓ	0.2394	0.0512	0.1018	0.0352	0.0752	0.0575	0.6994	0.4365	0.8834	0.4376	0.8090
0.1	bias	-0.0006	0.0005	0.0021	0.0012	-0.0003	-0.0004	-0.0149	-0.0063	0.0004	0.0036	0.0043
	rel. bias	-0.0901	0.4610	0.5329	0.1424	0.0544	-	1.6521	0.9691	-0.2008	0.5582	-
	SD	0.0854	0.0189	0.0374	0.0191	0.0245	0.0207	0.1721	0.1128	0.2295	0.1129	0.2032
	SE	0.0846	0.0186	0.0384	0.0191	0.0246	0.0207	0.1794	0.1119	0.2263	0.1120	0.2067
	RMSE	0.1202	0.0266	0.0536	0.0270	0.0347	0.0293	0.2490	0.1589	0.3222	0.1590	0.2898
	CP	0.9450	0.9420	0.9560	0.9550	0.9470	0.9570	0.9670	0.9510	0.9530	0.9490	0.9500
	ℓ	0.3315	0.0730	0.1504	0.0750	0.0965	0.0812	0.7024	0.4379	0.8859	0.4382	0.8096
0.2	bias	-0.0007	0.0006	0.0045	0.0025	-0.0016	-0.0006	-0.0158	-0.0064	0.0014	0.0038	0.0041
	rel. bias	-0.0965	0.5551	1.1235	0.2900	0.3190	-	1.7530	0.9864	-0.6932	0.5904	-
	SD	0.1013	0.0227	0.0473	0.0248	0.0286	0.0254	0.1718	0.1132	0.2292	0.1131	0.2032
	SE	0.1007	0.0229	0.0478	0.0248	0.0289	0.0253	0.1797	0.1122	0.2268	0.1120	0.2068
	RMSE	0.1429	0.0322	0.0673	0.0351	0.0406	0.0358	0.2490	0.1595	0.3224	0.1592	0.2899
	CP	0.9490	0.9570	0.9490	0.9520	0.9570	0.9440	0.9680	0.9500	0.9530	0.9470	0.9510
	ℓ	0.3947	0.0895	0.1871	0.0971	0.1131	0.0991	0.7035	0.4389	0.8878	0.4383	0.8099
0.4	bias	0.0004	0.0014	0.0074	0.0067	-0.0042	-0.0007	-0.0171	-0.0063	0.0020	0.0041	0.0041
	rel. bias	0.0544	1.4001	1.8431	0.7844	0.8493	-	1.8998	0.9736	-0.9913	0.6250	-
	SD	0.1442	0.0339	0.0721	0.0370	0.0398	0.0371	0.1724	0.1144	0.2307	0.1132	0.2034
	SE	0.1444	0.0342	0.0726	0.0372	0.0396	0.0375	0.1804	0.1131	0.2286	0.1121	0.2068
	RMSE	0.2040	.0482	0.1026	0.0529	0.0563	0.0528	0.2501	0.1609	0.3247	0.1594	0.2901
	CP	0.9580	0.9600	0.9570	0.9610	0.9560	0.9560	0.9690	0.9490	0.9540	0.9490	0.9490
	ℓ	0.5655	0.1338	0.2845	0.1458	0.1549	0.1469	0.7062	0.4424	0.8947	0.4387	0.8102

Table 3.4: Simulation results (n = 1000, ZI proportion = 20%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals.

average proportion of censoring		$\widehat{\beta}_n$						$\widehat{\gamma}_n$				
		$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$	$\widehat{\gamma}_{5,n}$
0	bias	-0.0011	0.0002	-0.0002	0.0003	0.0000	-0.0001	0.0087	-0.0069	-0.0010	0.0116	-0.0016
	rel. bias	-0.1638	0.2100	-0.0606	0.0405	0.0012	-	3.4745	0.9848	0.5148	1.7888	-
	SD	0.0702	0.0159	0.0310	0.0106	0.0226	0.0167	0.1518	0.0938	0.1808	0.0900	0.1625
	SE	0.0712	0.0155	0.0301	0.0106	0.0224	0.0170	0.1514	0.0916	0.1781	0.0906	0.1647
	RMSE	0.1000	0.0222	0.0432	0.0150	0.0318	0.0238	0.2145	0.1312	0.2538	0.1282	0.2313
	CP	0.9530	0.9400	0.9480	0.9460	0.9510	0.9540	0.9590	0.9480	0.9550	0.9550	0.9500
	ℓ	0.2785	0.0607	0.1177	0.0412	0.0874	0.0667	0.5932	0.3586	0.6980	0.3547	0.6456
0.1	bias	-0.0031	0.0011	-0.0001	0.0028	-0.0025	-0.0001	0.0072	-0.0071	-0.0011	0.0122	-0.0015
	rel. bias	-0.4410	1.1269	-0.0181	0.3284	0.4976	-	2.8914	1.0083	0.5361	1.8699	-
	SD	0.1063	0.0245	0.0476	0.0254	0.0317	0.0258	0.1528	0.0942	0.1811	0.0904	0.1626
	SE	0.1054	0.0237	0.0480	0.0250	0.0306	0.0259	0.1521	0.0919	0.1788	0.0907	0.1649
	RMSE	0.1496	0.0341	0.0676	0.0357	0.0441	0.0366	0.2156	0.1318	0.2544	0.1286	0.2315
	CP	0.9520	0.9470	0.9520	0.9540	0.9430	0.9510	0.9580	0.9500	0.9540	0.9560	0.9500
	ℓ	0.4128	0.0927	0.1881	0.0980	0.1198	0.1015	0.5957	0.3599	0.7004	0.3552	0.6461
0.2	bias	-0.0041	0.0009	0.0034	0.0050	-0.0038	-0.0001	0.0065	-0.0073	-0.0002	0.0124	-0.0018
	rel. bias	-0.5830	0.8778	0.8493	0.5876	0.7598	-	2.5978	1.0495	0.0864	1.9000	-
	SD	0.1364	0.0318	0.0616	0.0336	0.0389	0.0338	0.1530	0.0946	0.1822	0.0904	0.1626
	SE	0.1329	0.0310	0.0636	0.0341	0.0379	0.0337	0.1524	0.0922	0.1793	0.0908	0.1649
	RMSE	0.1905	0.0444	0.0886	0.0481	0.0544	0.0477	0.2159	0.1323	0.2556	0.1287	0.2316
	CP	0.9480	0.9440	0.9560	0.9580	0.9460	0.9550	0.9610	0.9490	0.9510	0.9580	0.9490
	ℓ	0.5205	0.1212	0.2493	0.1334	0.1482	0.1319	0.5969	0.3611	0.7027	0.3554	0.6462
0.4	bias	0.0012	0.0026	0.0103	0.0139	-0.0106	-0.0003	0.0048	-0.0075	0.0005	0.0129	-0.0018
	rel. bias	0.1656	2.6138	2.5854	1.6307	2.1250	-	1.9197	1.0707	-0.2254	1.9869	-
	SD	0.2476	0.0638	0.1291	0.0613	0.0660	0.0646	0.1544	0.0976	0.1857	0.0905	0.1631
	SE	0.2439	0.0603	0.1266	0.0625	0.0637	0.0641	0.1537	0.0943	0.1832	0.0909	0.1651
	RMSE	0.3475	0.0878	0.1810	0.0886	0.0923	0.0910	0.2178	0.1359	0.2608	0.1289	0.2320
	CP	0.9390	0.9350	0.9480	0.9530	0.9390	0.9470	0.9560	0.9480	0.9500	0.9570	0.9490
	ℓ	0.9544	0.2358	0.4951	0.2443	0.2488	0.2508	0.6021	0.3692	0.7178	0.3561	0.6468

Table 3.5: Simulation results (n = 1000, ZI proportion = 40%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals.

average proportion of censoring		$\widehat{\beta}_n$						$\widehat{\gamma}_n$				
		$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$	$\widehat{\gamma}_{5,n}$
0	bias	-0.0007	0.0001	0.0007	0.0000	0.0000	0.0000	-0.0034	-0.0050	0.0011	0.0032	0.0005
	rel. bias	-0.0961	0.0764	0.1761	-0.0012	-0.0030	-	0.3731	0.7663	-0.5449	0.4927	-
	SD	0.0386	0.0081	0.0165	0.0055	0.0118	0.0093	0.1116	0.0687	0.1454	0.0688	0.1311
	SE	0.0381	0.0082	0.0162	0.0055	0.0120	0.0092	0.1120	0.0698	0.1415	0.0700	0.1296
	RMSE	0.0542	0.0115	0.0232	0.0078	0.0168	0.0130	0.1581	0.0981	0.2028	0.0982	0.1843
	CP	0.9490	0.9430	0.9440	0.9460	0.9540	0.9520	0.9550	0.9510	0.9440	0.9560	0.9490
	ℓ	0.1493	0.0320	0.0636	0.0216	0.0469	0.0359	0.4389	0.2735	0.5544	0.2742	0.5080
0.1	bias	-0.0020	0.0002	0.0010	0.0003	0.0002	0.0002	-0.0037	-0.0049	0.0012	0.0033	0.0006
	rel. bias	-0.2881	0.2330	0.2396	0.0314	-0.0352	-	0.4082	0.7615	-0.5971	0.5113	-
	SD	0.0518	0.0116	0.0242	0.0120	0.0158	0.0130	0.1120	0.0689	0.1457	0.0688	0.1312
	SE	0.0532	0.0117	0.0241	0.0120	0.0155	0.0130	0.1125	0.0700	0.1418	0.0700	0.1297
	RMSE	0.0743	0.0165	0.0342	0.0170	0.0221	0.0184	0.1587	0.0983	0.2033	0.0982	0.1844
	CP	0.9570	0.9510	0.9470	0.9530	0.9390	0.9580	0.9570	0.9520	0.9470	0.9550	0.9500
	ℓ	0.2086	0.0459	0.0946	0.0472	0.0606	0.0511	0.4406	0.2741	0.5557	0.2744	0.5081
0.2	bias	-0.0028	0.0003	0.0012	0.0010	-0.0002	0.0002	-0.0041	-0.0050	0.0012	0.0035	0.0006
	rel. bias	-0.3938	0.3476	0.3096	0.1209	0.0307	-	0.4547	0.7720	-0.5901	0.5350	-
	SD	0.0625	0.0142	0.0286	0.0157	0.0183	0.0162	0.1120	0.0689	0.1459	0.0689	0.1312
	SE	0.0633	0.0143	0.0300	0.0156	0.0181	0.0159	0.1126	0.0701	0.1422	0.0701	0.1297
	RMSE	0.0890	0.0202	0.0415	0.0221	0.0257	0.0227	0.1588	0.0984	0.2037	0.0983	0.1845
	CP	0.9520	0.9550	0.9540	0.9510	0.9460	0.9440	0.9610	0.9520	0.9480	0.9560	0.9510
	ℓ	0.2482	0.0562	0.1175	0.0610	0.0710	0.0623	0.4412	0.2747	0.5570	0.2745	0.5081
0.4	bias	-0.0035	0.0005	0.0036	0.0025	-0.0010	0.0004	-0.0048	-0.0051	0.0021	0.0035	0.0005
	rel. bias	-0.4957	0.4867	0.8944	0.2962	0.2082	-	0.5302	0.7835	-1.0522	0.5447	-
	SD	0.0914	0.0212	0.0447	0.0231	0.0249	0.0241	0.1123	0.0697	0.1473	0.0689	0.1313
	SE	0.0905	0.0214	0.0455	0.0234	0.0248	0.0235	0.1130	0.0707	0.1432	0.0701	0.1297
	RMSE	0.1286	0.0301	0.0639	0.0330	0.0351	0.0337	0.1594	0.0993	0.2054	0.0983	0.1845
	CP	0.9440	0.9530	0.9630	0.9480	0.9570	0.9440	0.9610	0.9510	0.9470	0.9560	0.9490
	ℓ	0.3548	0.0838	0.1784	0.0916	0.0972	0.0922	0.4428	0.2768	0.5611	0.2747	0.5082

Table 3.6: Simulation results (n = 2500, ZI proportion = 20%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals.

average proportion of censoring		$\widehat{\beta}_n$						$\widehat{\gamma}_n$				
		$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$	$\widehat{\gamma}_{5,n}$
0	bias	-0.0025	0.0002	-0.0006	0.0004	0.0003	0.0003	0.0033	-0.0049	0.0031	0.0050	-0.0019
	rel. bias	-0.3542	0.2225	-0.1422	0.0482	-0.0621	-	1.3244	0.7032	-1.5480	0.7756	-
	SD	0.0449	0.0098	0.0188	0.0066	0.0140	0.0110	0.0964	0.0558	0.1125	0.0555	0.1024
	SE	0.0444	0.0096	0.0187	0.0065	0.0139	0.0106	0.0952	0.0576	0.1121	0.0569	0.1037
	RMSE	0.0631	0.0137	0.0266	0.0092	0.0197	0.0152	0.1355	0.0803	0.1588	0.0797	0.1457
	CP	0.9440	0.9430	0.9450	0.9500	0.9540	0.9360	0.9480	0.9580	0.9500	0.9620	0.9630
	ℓ	0.1738	0.0376	0.0733	0.0253	0.0545	0.0415	0.3732	0.2256	0.4395	0.2230	0.4064
0.1	bias	-0.0085	0.0010	0.0011	0.0024	-0.0003	0.0011	0.0021	-0.0049	0.0037	0.0054	-0.0018
	rel. bias	-1.2165	1.0182	0.2811	0.2875	0.0544	-	0.8532	0.7070	-1.8262	0.8257	-
	SD	0.0666	0.0145	0.0299	0.0159	0.0195	0.0166	0.0970	0.0559	0.1128	0.0556	0.1025
	SE	0.0661	0.0148	0.0302	0.0157	0.0191	0.0162	0.0956	0.0578	0.1125	0.0570	0.1037
	RMSE	0.0942	0.0208	0.0425	0.0225	0.0273	0.0232	0.1362	0.0805	0.1593	0.0798	0.1458
	CP	0.9480	0.9430	0.9480	0.9450	0.9410	0.9440	0.9510	0.9580	0.9530	0.9600	0.9630
	ℓ	0.2592	0.0581	0.1182	0.0617	0.0749	0.0637	0.3747	0.2264	0.4408	0.2232	0.4066
0.2	bias	-0.0070	0.0015	0.0023	0.0031	-0.0008	0.0007	0.0019	-0.0048	0.0039	0.0054	-0.0019
	rel. bias	-1.0047	1.5341	0.5652	0.3633	0.1684	-	0.7405	0.6910	-1.9621	0.8347	
	SD	0.0839	0.0192	0.0395	0.0219	0.0239	0.0212	0.0973	0.0562	0.1135	0.0557	0.1025
	SE	0.0833	0.0194	0.0399	0.0214	0.0236	0.0211	0.0958	0.0580	0.1128	0.0570	0.1038
	RMSE	0.1184	0.0273	0.0562	0.0307	0.0336	0.0299	0.1365	0.0808	0.1600	0.0798	0.1458
	CP	0.9420	0.9490	0.9520	0.9370	0.9570	0.9370	0.9490	0.9600	0.9480	0.9600	0.9620
	ℓ	0.3264	0.0759	0.1564	0.0837	0.0924	0.0827	0.3755	0.2271	0.4422	0.2233	0.4067
0.4	bias	-0.0077	0.0006	0.0033	0.0076	-0.0027	0.0013	0.0014	-0.0054	0.0034	0.0056	-0.0017
	rel. bias	-1.0942	0.6360	0.8313	0.8989	0.5363	-	0.5751	0.7733	-1.7103	0.8606	-
	SD	0.1592	0.0354	0.0767	0.0405	0.0418	0.0415	0.0977	0.0571	0.1138	0.0557	0.1025
	SE	0.1520	0.0375	0.0786	0.0388	0.0394	0.0399	0.0966	0.0592	0.1151	0.0571	0.1038
	RMSE	0.2202	0.0515	0.1099	0.0566	0.0575	0.0576	0.1373	0.0824	0.1619	0.0799	0.1459
	CP	0.9410	0.9600	0.9450	0.9430	0.9400	0.9370	0.9540	0.9580	0.9530	0.9620	0.9610
	ℓ	0.5953	0.1467	0.3080	0.1519	0.1544	0.1564	0.3785	0.2319	0.4512	0.2236	0.4068

Table 3.7: Simulation results (n = 2500, ZI proportion = 40%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals.

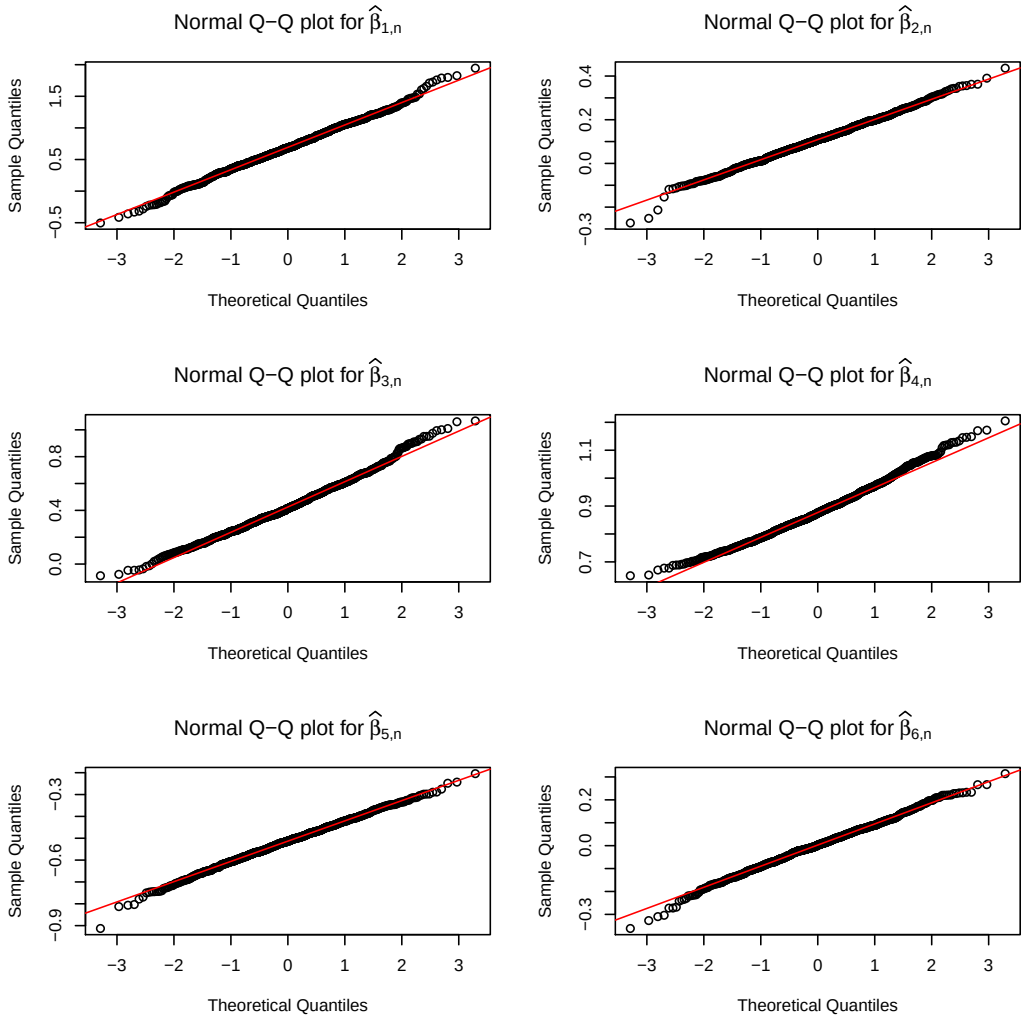

Figure 3.1: Normal Q-Q plots for $\hat{\beta}_{1,n}, \ldots, \hat{\beta}_{6,n}$ with $n = 500, c = 0.4$ and a proportion of zero-inflation equal to 0.4.

3.7 Appendix 1: Technical Lemma

Lemma 3.7.1. *Assume conditions C1-C4 hold. Then* $\sup_{\psi \in N_n(\varepsilon)} \|F_n^{-\frac{1}{2}} H_n(\psi) F_n^{-\frac{1}{2}} - I_k\|$ *converges in probability to* 0 *as* $n \to \infty$.

Proof of Lemma 3.7.1. We have

$$\begin{aligned} \|F_n^{-\frac{1}{2}} H_n(\psi) F_n^{-\frac{1}{2}} - I_k\| &= \|F_n^{-\frac{1}{2}} (H_n(\psi) - F_n) F_n^{-\frac{1}{2}}\|, \\ &\leqslant \frac{1}{\lambda_{\min}(F_n)} \|H_n(\psi) - F_n\|, \\ &\leqslant c_1 \left\| \frac{1}{n} (H_n(\psi) - \mathbb{E}(H_n(\psi))) \right\| + c_1 \left\| \frac{1}{n} (\mathbb{E}(H_n(\psi)) - F_n) \right\|, \end{aligned}$$

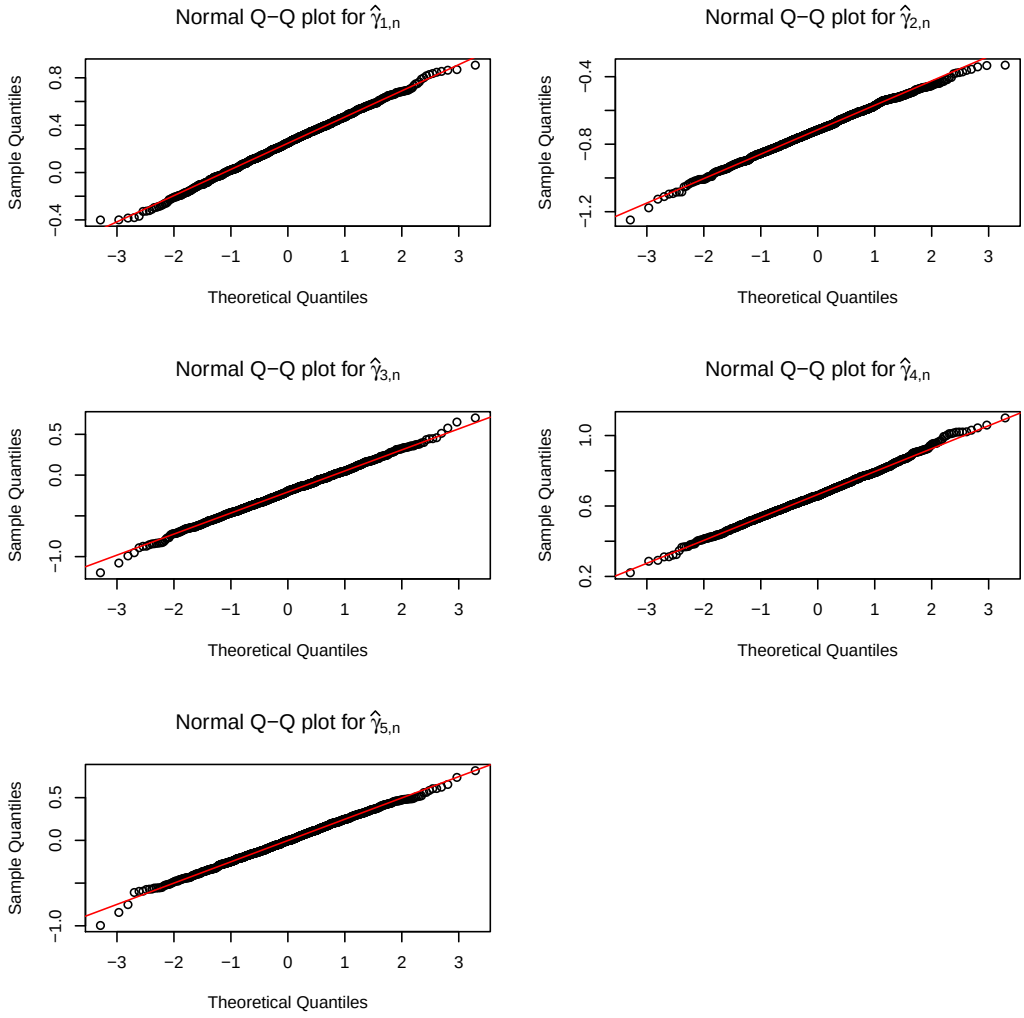

Figure 3.2: Normal Q-Q plots for $\hat{\gamma}_{1,n}, \ldots, \hat{\gamma}_{5,n}$ with $n = 500, c = 0.4$ and a proportion of zero-inflation equal to 0.4.

by C3. Thus, the lemma is proved if we can show that both terms in the right-hand side of the last inequality converge to 0 uniformly in $\psi \in N_n(\varepsilon)$ as $n \to \infty$. For illustration purposes, we show that $\sup_{\psi \in N_n(\varepsilon)} \|\frac{1}{n}(H_n(\psi) - \mathbb{E}(H_n(\psi)))\|$ converges in probability to 0 as $n \to \infty$. For this, it is sufficient to show that $\sup_{\psi \in N_n(\varepsilon)} |\frac{1}{n}(H_{n,(\ell,m)}(\psi) - \mathbb{E}(H_{n,(\ell,m)}(\psi)))|$ converges to 0 for every (ℓ, m), with $\ell, m = 1, \ldots, k$, where $H_{n,(\ell,m)}$ denotes the (ℓ, m)-element of H_n.
We illustrate the method for $\ell, m \in \{1, \ldots, p\}$. In this case, $H_{n,(\ell,m)}(\psi)$ coincides with the element $-\partial^2 \ell_n(\psi)/\partial\beta_\ell \partial\beta_m$ (other cases can be treated similarly). Using notations defined in section 3.2.2,

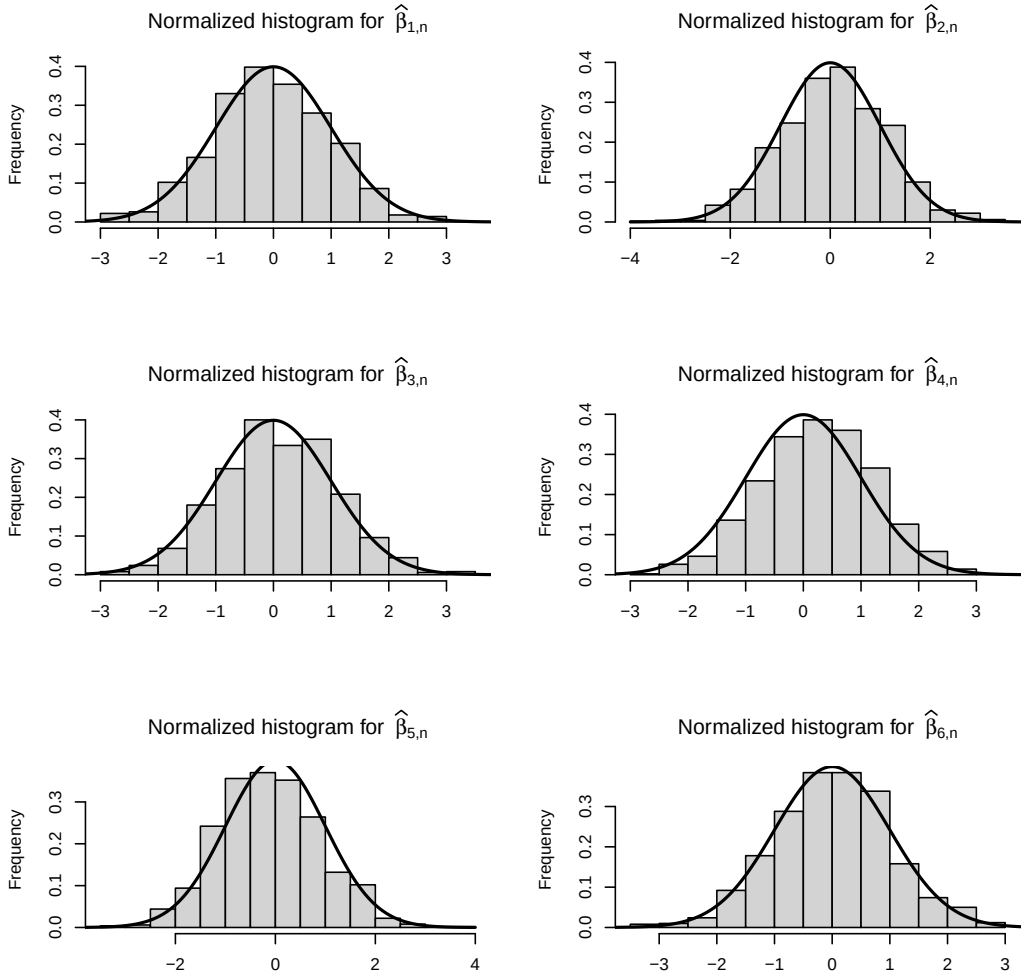

Figure 3.3: Histogram of the normalized estimates $(\widehat{\beta}_{j,n} - \beta_j)/\text{s.e.}(\widehat{\beta}_{j,n})$, $j = 1, \ldots, 6$ with $n = 500, c = 0.4$ and a proportion of zero-inflation equal to 0.4.

we have:

$$\begin{aligned}
&\left|\frac{1}{n}(H_{n,(\ell,m)}(\psi) - \mathbb{E}(H_{n,(\ell,m)}(\psi)))\right| \\
&\quad\leqslant \left|\frac{1}{n}\sum_{i=1}^{n}\left\{X_{i\ell}X_{im}\delta_i J_i u_i(\psi) - \mathbb{E}\left[X_{i\ell}X_{im}\delta_i J_i u_i(\psi)\right]\right\}\right| \\
&\qquad + \left|\frac{1}{n}\sum_{i=1}^{n}\left\{X_{i\ell}X_{im}\delta_i(1-J_i)\lambda_i(\beta) - \mathbb{E}\left[X_{i\ell}X_{im}\delta_i(1-J_i)\lambda_i(\beta)\right]\right\}\right| \\
&\qquad\quad + \left|\frac{1}{n}\sum_{i=1}^{n}\left\{X_{i\ell}X_{im}(1-\delta_i)(1-J_i)v_i(\psi) - \mathbb{E}\left[X_{i\ell}X_{im}(1-\delta_i)(1-J_i)v_i(\psi)\right]\right\}\right|.
\end{aligned}$$

We prove that $\sup_{\psi\in N_n(\varepsilon)} |\frac{1}{n}\sum_{i=1}^{n}\{X_{i\ell}X_{im}\delta_i J_i u_i(\psi) - \mathbb{E}\left[X_{i\ell}X_{im}\delta_i J_i u_i(\psi)\right]\}|$ converges in probability to 0 as $n \to \infty$ (the other two terms can be treated similarly). For this, we prove that the

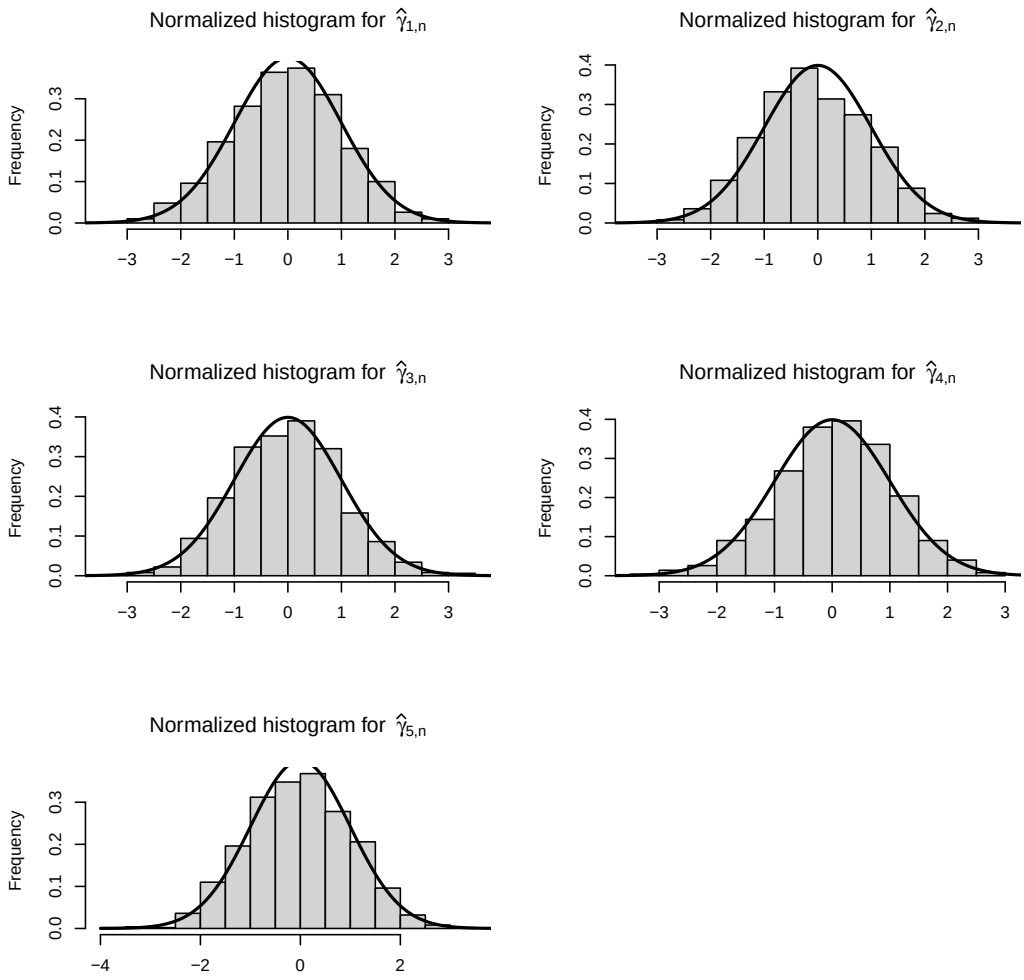

Figure 3.4: Histogram of the normalized estimates $(\widehat{\gamma}_{j,n} - \gamma_j)/\text{s.e.}(\widehat{\gamma}_{j,n})$, $j = 1, \ldots, 5$ with $n = 500, c = 0.4$ and a proportion of zero-inflation equal to 0.4.

class $\{X_{i\ell}X_{im}\delta_i J_i u_i(\psi) : \psi \in \mathcal{C}\}$ is Donsker (and thus Glivenko-Cantelli, which will ensure the required uniform (over ψ) convergence). We refer to Aad W. van der Vaart (1996) for definitions and properties of this classes.
The class $\{X_{i\ell}X_{im}\delta_i J_i\}$ is obviously Donsker. Under C1 and C2, classes $\{\beta^\top \mathbf{X}_i : \beta \in \mathcal{B}\}$ and $\{\gamma^\top \mathbf{Z}_i : \gamma \in \mathcal{G}\}$ are Donsker. The exponential function is Lipschitz on compact sets therefore classes $\{e^{\beta^\top \mathbf{X}_i} : \beta \in \mathcal{B}\}$, $\{e^{-\exp(\beta^\top \mathbf{X}_i)} : \beta \in \mathcal{B}\}$ and $\{e^{\gamma^\top \mathbf{Z}_i} : \gamma \in \mathcal{G}\}$ are also Donsker. Now, products and sums of bounded Donsker classes are Donsker (Aad W. van der Vaart, 1996), therefore, the class $\{X_{i\ell}X_{im}\delta_i J_i u_i(\psi) : \psi \in \mathcal{C}\}$ is Donsker.
Hence,

$$\sup_{\psi \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \{X_{i\ell}X_{im}\delta_i J_i u_i(\psi) - \mathbb{E}\left[X_{i\ell}X_{im}\delta_i J_i u_i(\psi)\right]\} \right|$$

converges in probability to 0 as $n \to \infty$. Since $N_n(\varepsilon) \subset \mathcal{C}$, so that

$$\sup_{\psi \in N_n(\varepsilon)} |\frac{1}{n} \sum_{i=1}^{n} \{X_{i\ell} X_{im} \delta_i J_i u_i(\psi) - \mathbb{E}\left[X_{i\ell} X_{im} \delta_i J_i u_i(\psi)\right]\}|$$

also converges to 0 as $n \to \infty$, which concludes the proof. □

3.8 Appendix 2: Technical Calculations

In this Appendix, we give some calculation details. We have:

$$\begin{aligned}
&\frac{\partial}{\partial \beta_m} \mathcal{S}_{\mathcal{P}(\lambda_i(\beta))}(Y_i^*) = X_{im} \lambda_i(\beta) \mathbb{P}(\mathcal{P}(\lambda_i(\beta)) = Y_i^* - 1), \\
&\frac{\partial}{\partial \beta_\ell} \left(\frac{\lambda_i(\beta) L_i(\beta)}{k_i(\gamma) + L_i(\beta)} \right) = X_{i\ell} \frac{\lambda_i(\beta) L_i(\beta)}{(k_i(\gamma) + L_i(\beta))^2} \left[(1 - \lambda_i(\beta))(k_i(\gamma) + L_i(\beta)) + \lambda_i L_i(\beta) \right], \\
&\frac{\partial}{\partial \beta_\ell} \left(L_i(\beta) \lambda_i(\beta)^k (k - \lambda_i(\beta)) \right) = X_{i\ell} L_i(\beta) \lambda_i(\beta)^k \left((\lambda_i(\beta) - k)^2 - \lambda_i(\beta) \right),
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial}{\partial \beta_\ell} \left(\frac{L_i(\beta) \lambda_i(\beta)^k (k - \lambda_i(\beta))}{\mathcal{S}_{\mathcal{P}(\lambda_i(\beta))}(Y_i^*)} \right) = X_{i\ell} \frac{L_i(\beta) \lambda_i(\beta)^k}{\mathcal{S}_{\mathcal{P}(\lambda_i(\beta))}(Y_i^*)^2} \Big(&[(\lambda_i(\beta) - k)^2 - \lambda_i(\beta)] \mathcal{S}_{\mathcal{P}(\lambda_i(\beta))}(Y_i^*) \\
&- \lambda_i(\beta)(k - \lambda_i(\beta)) \mathbb{P}(\mathcal{P}(\lambda_i(\beta)) = Y_i^* - 1) \Big).
\end{aligned}$$

To see this, note that:

$$\begin{aligned}
\frac{\partial}{\partial \beta_m} \mathcal{S}_{\mathcal{P}(\lambda_i(\beta))}(Y_i^*) &= \frac{\partial}{\partial \beta_m} \left(1 - \sum_{k=0}^{Y_i^* - 1} \frac{\lambda_i(\beta)^k L_i(\beta)}{k!} \right) \\
&= - \sum_{k=0}^{Y_i^* - 1} \frac{1}{k!} \left(k \lambda_i(\beta)^{k-1} \lambda_i(\beta) X_{im} L_i(\beta) - \lambda_i(\beta)^k L_i(\beta) \lambda_i(\beta) X_{im} \right) \\
&= - \sum_{k=0}^{Y_i^* - 1} \frac{X_{im} L_i(\beta) \lambda_i^k(\beta)}{k!} (k - \lambda_i(\beta)) \\
&= -X_{im} \left(\sum_{k=1}^{Y_i^* - 1} \frac{L_i(\beta) \lambda_i^k(\beta)}{(k-1)!} - \sum_{k=0}^{Y_i^* - 1} \frac{L_i(\beta) \lambda_i^{k+1}(\beta)}{k!} \right) \\
&= X_{im} \frac{L_i(\beta) \lambda_i^{Y_i^*}(\beta)}{(Y_i^* - 1)!} \\
&= X_{im} \lambda_i(\beta) \mathbb{P}(\mathcal{P}(\lambda_i(\beta)) = Y_i^* - 1).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{\partial}{\partial\beta_\ell}\left(\frac{\lambda_i(\beta)L_i(\beta)}{k_i(\gamma)+L_i(\beta)}\right) &= \frac{1}{(k_i(\gamma)+L_i(\beta))^2}\left[(\lambda_i'(\beta)L_i(\beta)+\lambda_i(\beta)L_i'(\beta))(k_i(\gamma)+L_i(\beta))\right.\\
&\qquad\left.-\lambda_i(\beta)L_i(\beta)L_i'(\beta)\right],\\
&= \frac{1}{(k_i(\gamma)+L_i(\beta))^2}\left[X_{i\ell}\lambda_i(\beta)L_i(\beta)-\lambda_i^2(\beta)X_{i\ell}L_i(\beta)(k_i+L_i(\beta))\right.\\
&\qquad\left.+X_{i\ell}\lambda_i(\beta)^2L_i(\beta)^2\right],\\
&= X_{i\ell}\frac{1}{(k_i(\gamma)+L_i(\beta))^2}\left[(1-\lambda_i(\beta))(k_i(\gamma)+L_i(\beta))+\lambda_i(\beta)L_i(\beta)\right]\\
&\qquad\times\lambda_i(\beta)L_i(\beta),\\
&= X_{i\ell}\frac{\lambda_i(\beta)L_i(\beta)}{(k_i(\gamma)+L_i(\beta))^2}\left[(1-\lambda_i(\beta))(k_i(\gamma)+L_i(\beta))+\lambda_iL_i(\beta)\right],
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial}{\partial\beta_\ell}\Big(L_i(\beta)\lambda_i(\beta)^k(k-\lambda_i(\beta))\Big) &= L_i(\beta)'(\lambda_i(\beta)^kk-\lambda_i(\beta)^{k+1})+L_i(\beta)(k^2X_{i\ell}\lambda_i(\beta)^k\\
&\qquad-(k+1)X_{i\ell}\lambda_i(\beta)^{k+1}),\\
&= -X_{i\ell}L_i(\beta)\lambda_i(\beta)^k\left(k\lambda_i(\beta)-\lambda_i(\beta)^2-k^2+(k+1)\lambda_i(\beta)\right),\\
&= X_{i\ell}L_i(\beta)\lambda_i(\beta)^k\left(\lambda_i(\beta)^2+k^2-\lambda_i(\beta)(2k+1)\right),\\
&= X_{i\ell}L_i(\beta)\lambda_i(\beta)^k\left((\lambda_i(\beta)-k)^2-\lambda_i(\beta)\right),
\end{aligned}$$

Finally,

$$\begin{aligned}
\frac{\partial}{\partial\beta_\ell}\left(\frac{L_i(\beta)\lambda_i(\beta)^k(k-\lambda_i(\beta))}{\mathcal{S}_{\mathcal{P}(\lambda_i(\beta))}(Y_i^*)}\right) &= \left[\Big(L_i(\beta)\lambda_i(\beta)^k(k-\lambda_i(\beta))\Big)'\mathcal{S}_{\mathcal{P}(\lambda_i(\beta))}(Y_i^*)\right.\\
&\qquad\left.-L_i(\beta)\lambda_i(\beta)^k(k-\lambda_i(\beta))\mathcal{S}'_{\mathcal{P}(\lambda_i(\beta))}(Y_i^*)\right]\Big/\mathcal{S}_{\mathcal{P}(\lambda_i(\beta))}(Y_i^*)^2,\\
&= \frac{1}{\mathcal{S}_{\mathcal{P}(\lambda_i(\beta))}(Y_i^*)^2}\Big(X_{i\ell}L_i(\beta)\lambda_i(\beta)^k((\lambda_i(\beta)-k)^2-\lambda_i(\beta))\mathcal{S}_{\mathcal{P}(\lambda_i(\beta))}(Y_i^*)\\
&\qquad-L_i(\beta)\lambda_i(\beta)^{k+1}(k-\lambda_i(\beta))X_{i\ell}\mathbb{P}(\mathcal{P}(\lambda_i(\beta))=Y_i^*-1)\Big),\\
&= X_{i\ell}\frac{L_i(\beta)\lambda_i(\beta)^k}{\mathcal{S}_{\mathcal{P}(\lambda_i(\beta))}(Y_i^*)^2}\Big(\Big[(\lambda_i(\beta)-k)^2-\lambda_i(\beta)\Big]\mathcal{S}_{\mathcal{P}(\lambda_i(\beta))}(Y_i^*)\\
&\qquad-\lambda_i(\beta)(k-\lambda_i(\beta))\mathbb{P}(\mathcal{P}(\lambda_i(\beta))=Y_i^*-1)\Big).
\end{aligned}$$

□

4 Censored zero-inflated generalized Poisson and negative binomial regression models: a simulation-based study

Contents

4.1 Introduction **50**
4.2 Censored ZIGP and ZINB models **52**
4.2.1 Maximum likelihood estimation in censored ZIGP regression 52
4.2.2 Maximum likelihood estimation in censored ZINB regression 53
4.3 A simulation study **54**
4.3.1 Simulation scenario 54
4.3.2 Results 55
4.4 Real data application **56**
4.5 Discussion **58**
4.6 Appendix 1: R code for fitting the censored ZINB model **62**
4.7 Appendix 2: Vuong test **62**
4.8 Appendix 3: Technical calculations **64**

Ce chapitre fait l'objet d'un article intitulé "Zero-inflated regression models for right-censored counts, with an application to health utilization" (auteurs: Van-Trinh Nguyen et Jean-François Dupuy), publié dans la revue "Biostatistics and Health Sciences" (Vol. 1, No. 1, 1-19, 2019 ; DOI: 10.21494/ISTE.OP.2019.0393).

Abstract
Zero-inflated models for censored and overdispersed count data have received little attention so far, except for the zero-inflated Poisson (ZIP) model which assumes that overdispersion is entirely caused by zero-inflation. When additional overdispersion is present, useful alternatives to ZIP are given by the zero-inflated generalized Poisson (ZIGP) and zero-inflated negative binomial (ZINB) models. This paper investigates properties of the maximum likelihood estimator (MLE) in ZIGP and ZINB regression models when the count response is subject to right-censoring. Simulations are used to examine performance (bias, mean square error, coverage probabilities and standard error calculations) of the MLE. Results suggest that maximum likelihood yields accurate inference. A simple, efficient and easy-to-implement methodology for variable selection is also proposed. It is applicable even when the number of predictors is very large and yields interpretable and sound results. The proposed methods are applied to a dataset of healthcare demand.

Key words: excess of zeros, maximum likelihood, simulations

Abstract in french

Dans ce chapitre, nous nous intéressons à l'estimation dans le modèle de régression de Poisson généralisé à inflation de zéro (modèle ZIGP) et dans le modèle de régression négatif binomial à inflation de zéro (modèle ZINB), en présence de données de comptage censurées aléatoirement à droite, mais cette fois-ci, du point de vue de la sélection de modèles. Le problème de la sélection de modèles (i.e., des variables explicatives pertinentes) dans les modèles de régression zéro-inflatés a été abordé, dans la littérature, par des méthodes d'estimation pénalisée (pénalité lasso, ridge, elastic-net par exemple) mais outre le fait que ces méthodes requièrent le paramétrage fin de plusieurs paramètres (poids de la pénalité par exemple), elles peuvent également être prises en défaut lorsque le nombre de prédicteurs possibles est trop grand (ce qui arrive lorsque l'on souhaite tester l'effet de toutes les interactions d'ordre 2 existant entre les prédicteurs, par exemple).

Nous proposons donc un algorithme de sélection de variables en trois étapes, qui consiste à ajuster, tout d'abord, un modèle de régression logistique à l'indicatrice d'égalité à 0 des observations (ce modèle est alors conçu comme une approximation du modèle d'excès de zéros), afin de sélectionner les variables explicatives et interactions pertinentes - la sélection repose ici sur une méthode pas-à-pas ascendante et le critère BIC. Dans un second temps, nous proposons d'ajuster un modèle de comptage censuré (Poisson généralisé ou négatif binomial, par exemple) aux données de comptage, sans tenir compte de l'excès de zéros - et de sélectionner des variables explicatives en utilisant les mêmes outils que dans la première étape. Enfin, dans une troisième étape, nous proposons d'ajuster un modèle censuré de régression à inflation de zéros en utilisant les variables et interactions identifiées aux deux premières étapes, et de mener une sélection de variables séquentielle, basée sur le test de Wald et le critère BIC.

Nous évaluons la pertinence de l'algorithme proposé, en l'appliquant à un jeu de données réelles issues du domaine de l'économie de la santé. Il s'agit d'une base de données recueillie en Allemagne et renseignant la consommation de soins de 1812 hommes, pour lesquels sont également disponibles, sous formes de variables explicatives, de nombreuses informations relatives à leur état de santé et à leur situation socio-économique. Nous identifions les déterminants du recours aux soins de ces hommes, ainsi que les variables qui influencent leur nombre moyen de consultations médicales.

4.1 Introduction

Healthcare utilization refers to the measure of a population's use of available healthcare services. It is often reported as the number of healthcare services (e.g., hospital resources, physician resources) used over a period of time. Count-valued outcomes arising from healthcare utilization studies can be modeled using discrete distributions, such as Poisson or negative binomial. However, healthcare utilization data often contain large numbers of zeros, i.e. there is a large number of non-users of the corresponding healthcare service over the study period. When there are more zeros than expected under a standard count model, the data are said to be zero-inflated, which is a particular cause of zero-inflation.

Various models have been developed to address zero-inflation, such as zero-inflated (ZI) models which mix a degenerate distribution at zero with a standard count model. If predictors are present (e.g., age, income, health satisfaction...), ZI models can be extended to the regression setting by modeling zero-inflation and count sub-distributions as functions of the predictors. For example, zero-inflated Poisson (ZIP) regression model was proposed by Lambert (1992), and further developed to accommodate random effects (Hall, 2000; Min and Agresti, 2005;

Monod, 2014), non-linear covariate effects (Lam et al., 2006; He et al., 2010; Lu and Li, 2015), longitudinal zero-inflated counts Feng and Zhu (2011). The ZIP model assumes that overdispersion in the data is entirely caused by an excess of zeros. When some additional overdispersion is present, useful alternatives to ZIP are the zero-inflated negative binomial (ZINB) model (Ridout et al., 2001; Moghimbeigi et al., 2008; Mwalili et al., 2008) and zero-inflated generalized Poisson (ZIGP) model (Czado and Min, 2005; Czado et al., 2007), which both contain an additional overdispersion parameter.

Count data can also be affected by censoring, the most common type being right-censoring (which occurs when it is only known that the true count is higher than the observed one). For example, consider a healthcare utilization study where patients report their number of visits to a doctor during a given period. If one possible answer is, say, "15 visits or more", all visit counts greater than 15 are right-censored at 15. Ignoring censoring yields biased estimates and incorrect inference.

Count data analysis with censoring has been investigated by several authors, including cases of Poisson and generalized Poisson regressions (Terza, 1985; Caudill and Franklin G, 1995; Famoye and Wang, 2004; Xie and Wei, 2007; M. Mahmoud and M. Alderiny, 2010), zero-truncated Poisson regression (Yeh et al., 2012) and finite mixtures of Poisson regressions (Karlis et al., 2016). In contrast, much less work has been done for censored counts with zero-inflation. Saffari and Adnan (2011) and Nguyen and Dupuy (2019b) investigate ZIP regression with right-censored data. Adnan et al. (2012); Saffari et al. (2013) address estimation in right-censored hurdle negative binomial and hurdle generalized Poisson regression models. But to date, applicability of ZIGP and ZINB regression models to censored data has not been evaluated. Our aim is to fill this gap. We conduct simulations to explore properties of the maximum likelihood estimator in right-censored ZIGP and ZINB models. We also investigate the question of variable selection in these models.

Variable selection is a crucial issue in regression modeling. When many potential risk factors are available (which is usually the case in healthcare utilization studies), it is important to identify the predictors (and eventual interactions) which have a significant impact on the response, as parsimonious models offer easier interpretation and more accurate estimates. Several authors addressed variable selection in uncensored ZIP and ZINB models. For example, Czado et al. (2007) use sequential elimination (based either on hypothesis testing or information criteria) to select significant predictors in an application dealing with patent outsourcing. Buu et al. (2011), Wang et al. (2014, 2015), Zeng et al. (2014) and Chatterjee et al. (2018) investigate penalized maximum likelihood estimation. This approach, however, requires specific computing algorithms and elaborated strategies for tuning parameter selection, which can discourage its use. Moreover, from our experience, penalized estimation in zero-inflated models can fail to converge when the number of predictors is too large (the problem may even arise with a moderate number of risk factors, if all second-order interactions are included in the model). Stepwise regression can avoid this problem (although the method also has its own disadvantages). Furthermore, in practice, stepwise regression often selects similar subsets of predictors as penalized methods (see for example Wang et al., 2014, 2015). Variable selection for right-censored zero-inflated counts has not been addressed. We discuss this issue here, with the objective of providing a simple methodology that can be applied with existing softwares.

This paper is organized as follows. In Section 3.2, we briefly review the ZIGP and ZINB models and we describe maximum likelihood estimation (MLE) with right-censored counts. In Section 4.3, we conduct a detailed simulation study to assess the performance of the MLE. Section 4.4 describes an application to a dataset of healthcare demand. We present a simple, efficient and easy-to-implement methodology for selecting predictors and interactions in both

zero-inflation and counts submodels. This approach is demonstrated on the healthcare demand data. Discussion and concluding remarks are presented in Section 4.5

4.2 Censored ZIGP and ZINB models

4.2.1 Maximum likelihood estimation in censored ZIGP regression

Let Z_i denote the count of some event (such as the number of doctor visits) for an individual i $(i = 1, \dots, n)$ and $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{ip})^\top$ and $\mathbf{W}_i = (W_{i1}, W_{i2}, \dots, W_{iq})^\top$ be respectively p and q-dimensional vectors of risk factors for this individual. Both categorical and continuous variables are allowed. Moreover, $\mathbf{X}_i$ and $\mathbf{W}_i$ may share some common terms or be distinct. To include intercepts, we set $X_{i1} = 1$ and $W_{i1} = 1$.

A zero-inflated generalized Poisson model (Czado and Min, 2005; Czado et al., 2007) for Z_i is defined as

$$Z_i \sim \begin{cases} 0 & \text{with probability } \omega_i, \\ \mathcal{GP}(\lambda_i, \varphi) & \text{with probability } 1 - \omega_i, \end{cases} \tag{4.2.1}$$

where $0 \leqslant \omega_i \leqslant 1$ is the probability of zero-inflation and $\mathcal{GP}(\lambda_i, \varphi)$ is the generalized Poisson distribution with parameters $\lambda_i > 0$ and φ (Consul and Famoye, 1992). Both under- and overdispersion are allowed, depending on whether $\varphi < 1$ or $\varphi > 1$. However, in case of underdispersion, the support of $\mathcal{GP}(\lambda_i, \varphi)$ depends on λ_i and φ, which makes them difficult to estimate. For this reason, the generalized Poisson is usually considered for modelling overdispersed data, which is also the most common case in practice. We also restrict to this case here and assume that $\varphi > 1$.

The probability density function of the ZIGP model is given by

$$\mathbb{P}(Z_i = z) = \begin{cases} \omega_i + (1 - \omega_i) e^{-\frac{\lambda_i}{\varphi}} & \text{for } z = 0, \\ (1 - \omega_i) \frac{\lambda_i (\lambda_i + (\varphi - 1) z)^{z-1} \varphi^{-z}}{z!} e^{-\frac{(\lambda_i + (\varphi - 1) z)}{\varphi}} & \text{for } z = 1, 2, \dots \end{cases} \tag{4.2.2}$$

From this, it is straightforward to see that the mean and variance of Z_i are given by $\mathbb{E}(Z_i) = (1 - \omega_i)\lambda_i$ and $\operatorname{var}(Z_i) = \mathbb{E}(Z_i)(\varphi^2 + \lambda_i \omega_i)$ respectively, where φ is called overdispersion parameter. Therefore, the ZIGP model can accommodate two different sources of overdispersion, namely zero-inflation and heterogeneity between individuals. The ZIGP model reduces to the usual ZIP when $\varphi = 1$. We refer the reader to Czado et al. (2007) for an application of ZIGP model to uncensored counts.

When risk factors are available, the mixing probability ω_i is usually modeled by a logistic regression: $\operatorname{logit}(\omega_i(\gamma)) = \gamma^\top \mathbf{W}_i$ and λ_i is classically modeled as $\lambda_i(\beta) = \exp(\beta^\top \mathbf{X}_i)$. Vectors $\beta = (\beta_1, \dots, \beta_p)^\top \in \mathbb{R}^p$ and $\gamma = (\gamma_1, \dots, \gamma_q)^\top \in \mathbb{R}^q$ are unknown regression parameters.

Assume now that the count response Z_i can be right-censored. That is, for some individuals, we only observe a lower bound on Z_i. This can be modeled by introducing a positive censoring value C_i and defining the count data for the i-th individual as the pair (Z_i^*, δ_i), where $Z_i^* = \min(Z_i, C_i)$ and $\delta_i = 1_{\{Z_i < C_i\}}$ (if $Z_i = C_i$, we let $Z_i^* = C_i$ and $\delta_i = 0$). The censoring value can either be the same for all individuals (fixed threshold) or be specific to each observation. Let $J_i = 1_{\{Z_i^* = 0\}}$ and $\bar{J}_i = 1 - J_i$. Let also $\bar{\delta}_i = 1 - \delta_i$.

Suppose that we observe n independent vectors $(Z_i^*, \delta_i, \mathbf{X}_i, \mathbf{W}_i)$, $i = 1, \dots, n$. Let $\psi :=$

$(\beta^\top, \gamma^\top, \varphi)^\top$ denote the set of all unknown parameters. Then, the likelihood of ψ is:

$$\begin{aligned} L_n(\psi) &= \prod_{i=1}^n \mathbb{P}(Z_i = Z_i^*|\mathbf{X}_i, \mathbf{W}_i)^{\delta_i} \mathbb{P}(Z_i \geqslant Z_i^*|\mathbf{X}_i, \mathbf{W}_i)^{\bar\delta_i}, \\ &= \prod_{i=1}^n \left(\mathbb{P}(Z_i = Z_i^*|\mathbf{X}_i, \mathbf{W}_i)^{\bar J_i} \mathbb{P}(Z_i = 0|\mathbf{X}_i, \mathbf{W}_i)^{J_i}\right)^{\delta_i} \mathbb{P}(Z_i \geqslant Z_i^*|\mathbf{X}_i, \mathbf{W}_i)^{\bar\delta_i \bar J_i}, \end{aligned}$$

with $\mathbb{P}(Z_i \geqslant Z_i^*|\mathbf{X}_i, \mathbf{W}_i) = 1 - \sum_{k=0}^{Z_i^*-1} \mathbb{P}(Z_i = k|\mathbf{X}_i, \mathbf{W}_i)$. Suppose that ω_i and λ_i are given as above and let $S_{\mathcal{GP}(\lambda_i,\varphi)}$ denote the survival function of the generalized Poisson $\mathcal{GP}(\lambda_i, \varphi)$ distribution, that is, $S_{\mathcal{GP}(\lambda_i,\varphi)}(z) = \mathbb{P}(\mathcal{GP}(\lambda_i, \varphi) \geqslant z)$. Using (4.2.2) and some algebra, the loglikelihood $\ell_n(\psi) = \log L_n(\psi)$ can be written as:

$$\begin{aligned} \ell_n(\psi) &= \sum_{i=1}^n \delta_i \left[J_i \log\left(e^{\gamma^\top \mathbf{W}_i} + e^{-\frac{\exp(\beta^\top \mathbf{X}_i)}{\varphi}}\right) + \bar J_i \left\{ \beta^\top \mathbf{X}_i + (Z_i^* - 1)\log\left(e^{\beta^\top \mathbf{X}_i} + (\varphi - 1)Z_i^*\right) \right.\right. \\ &\qquad \left.\left. - Z_i^* \log\varphi - \frac{1}{\varphi}\left(e^{\beta^\top \mathbf{X}_i} + (\varphi - 1)Z_i^*\right) - \log(Z_i^*!) \right\} \right] \\ &\quad - \sum_{i=1}^n \log\left(1 + e^{\gamma^\top \mathbf{W}_i}\right) + \sum_{i=1}^n \bar\delta_i \bar J_i \log S_{\mathcal{GP}(\lambda_i,\varphi)}(Z_i^*), \end{aligned} \tag{4.2.3}$$

with

$$S_{\mathcal{GP}(\lambda_i,\varphi)}(Z_i^*) = 1 - \sum_{z=0}^{Z_i^*-1} e^{\beta^\top \mathbf{X}_i} (e^{\beta^\top \mathbf{X}_i} + (\varphi - 1)z)^{z-1} \varphi^{-z} e^{-\frac{(\exp(\beta^\top \mathbf{X}_i) + (\varphi-1)z)}{\varphi}} \frac{1}{z!}.$$

If $\delta_i = 1$ for every $i = 1, \ldots, n$, (4.2.3) reduces to the loglikelihood given by Czado and Min (2005) in the uncensored ZIGP model. If $\varphi = 1$, (4.2.3) reduces to the loglikelihood given by Nguyen and Dupuy (2019b) in the censored ZIP model.

The MLE $\hat\psi_n := (\hat\beta_n^\top, \hat\gamma_n^\top, \hat\varphi_n)^\top$ is obtained by solving the score equation $\partial\ell_n(\psi)/\partial\psi = 0$, which can be achieved by nonlinear optimization. In this paper, all estimates are obtained using the `R` function `maxLik` (Henningsen and Toomet, 2011), which implements Newton-type algorithms. A sample code is provided in the Appendix. The function also provides the Hessian matrix of ℓ_n, which is needed for variance estimation of the MLE. Precisely, we estimate the variance-covariance matrix of $\hat\psi_n$ by $\hat\Sigma_n = [-\partial^2\ell_n(\hat\psi_n)/\partial\psi\partial\psi^\top]^{-1}$. Standard errors of parameter estimates are obtained as the square roots of the diagonal terms of $\hat\Sigma_n$.

A rigorous assessment of asymptotic properties of $\hat\psi_n$ is likely to be challenging, in light of complicacy of the calculations in the censored ZIP model (Nguyen and Dupuy, 2019b). In that paper, it is shown that the MLE in the censored ZIP model, which is a particular case of censored ZIGP, is consistent and asymptotically normal. Such properties can be expected in the ZIGP model also. However, leaving aside the distributional theory, we propose to investigate these properties by means of simulations.

4.2.2 Maximum likelihood estimation in censored ZINB regression

The zero-inflated negative binomial model can be defined similarly as the ZIGP model, by replacing the generalized Poisson distribution in (4.2.1) by a negative binomial distribution. The probability density function of ZINB model is given by

$$\mathbb{P}(Z_i = z) = \begin{cases} \omega_i + (1 - \omega_i)\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}} & \text{for } z = 0, \\ (1 - \omega_i)\frac{\Gamma(z+\alpha^{-1})}{\Gamma(\alpha^{-1})z!}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^z \left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}} & \text{for } z = 1, 2, \ldots \end{cases} \tag{4.2.4}$$

where $0 \leqslant \omega_i \leqslant 1$, $\mu_i \geqslant 0$ and α is a positive overdispersion parameter. The mean and variance of Z_i are $(1-\omega_i)\mu_i$ and $(1-\omega_i)(\mu_i+\alpha\mu_i^2+\omega_i\mu_i^2)$ respectively. From this, we note that the ZINB model also allows two sources of overdispersion, one coming from zero-inflation and the other from heterogeneity.

When risk factors are available, ω_i is usually modeled as $\text{logit}(\omega_i(\gamma)) = \gamma^\top \mathbf{W}_i$ and μ_i is taken as $\mu_i(\beta) = \exp(\beta^\top \mathbf{X}_i)$, where $\beta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}^q$ are unknown parameters. If counts Z_i are right-censored and if we observe n independent vectors $(Z_i^*, \delta_i, \mathbf{X}_i, \mathbf{W}_i)$ (with same notations as above), the loglikelihood of $\theta := (\beta^\top, \gamma^\top, \alpha)^\top$ can be calculated as in the previous section and is given by:

$$\begin{aligned}\ell_n(\theta) &= \sum_{i=1}^n \delta_i \left[J_i \log\left(e^{\gamma^\top \mathbf{W}_i} + \frac{1}{(1+\alpha e^{\beta^\top \mathbf{X}_i})^{\alpha^{-1}}} \right) + \bar{J}_i \left\{ Z_i^* \beta^\top \mathbf{X}_i + Z_i^* \log\alpha \right.\right. \\ &\qquad \left.\left. -(Z_i^* + \alpha^{-1}) \log\left(1+\alpha e^{\beta^\top \mathbf{X}_i}\right) + \log\Gamma(Z_i^* + \alpha^{-1}) - \log\Gamma(\alpha^{-1}) - \log(Z_i^*!) \right\} \right] \\ &\quad - \sum_{i=1}^n \log\left(1 + e^{\gamma^\top \mathbf{W}_i}\right) + \sum_{i=1}^n \bar{\delta}_i \bar{J}_i \log S_{\mathcal{NB}(\mu_i,\alpha)}(Z_i^*), \end{aligned} \tag{4.2.5}$$

where

$$S_{\mathcal{NB}(\mu_i,\alpha)}(Z_i^*) = 1 - \sum_{z=0}^{Z_i^*-1} \frac{\Gamma(z+\alpha^{-1})}{\Gamma(\alpha^{-1})z!} \left(\frac{\alpha e^{\beta^\top \mathbf{X}_i}}{1+\alpha e^{\beta^\top \mathbf{X}_i}} \right)^z \left(\frac{1}{1+\alpha e^{\beta^\top \mathbf{X}_i}} \right)^{\alpha^{-1}}.$$

The MLE $\hat{\theta}_n := (\hat{\beta}_n^\top, \hat{\gamma}_n^\top, \hat{\alpha}_n)^\top$ is obtained by solving the score equation $\partial\ell_n(\psi)/\partial\theta = 0$, which again requires numerical optimization. Properties of this MLE are investigated by simulations in the next section. As for the ZIGP model, we obtain standard errors as $\sqrt{\text{diag}(\hat{\Sigma}_n)}$, where $\hat{\Sigma}_n = [-\partial^2\ell_n(\hat{\theta}_n)/\partial\theta\partial\theta^\top]^{-1}$.

4.3 A simulation study

In this section, we investigate properties of the MLE in censored ZIGP and ZINB models.

4.3.1 Simulation scenario

First, we simulate data from the ZIGP model (4.2.1), with:

$$\log(\lambda_i(\beta)) = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6},$$

and

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

where $X_{i1} = W_{i1} = 1$ and the $X_{i2}, \ldots, X_{i6}, W_{i4}, W_{i5}$ are independently drawn from normal $\mathcal{N}(0,1)$, Bernoulli $\mathcal{B}(0.3)$, normal $\mathcal{N}(1, 2.25)$, exponential $\mathcal{E}(1)$, uniform $\mathcal{U}(2,5)$, normal $\mathcal{N}(-1,1)$ and Bernoulli $\mathcal{B}(0.5)$ distributions respectively. Linear predictors in $\log(\lambda_i(\beta))$ and $\text{logit}(\omega_i(\gamma))$ are allowed to share two common terms, namely $W_{i2} = X_{i2}$ and $W_{i3} = X_{i3}$. Regression parameters β and γ are taken as $\beta = (0.7, 0.1, 0.4, 0.85, -0.5, 0)^\top$ and $\gamma = (-0.9, -0.65, -0.2, 0.65, 0)^\top$. The proportion of zero-inflated data in the simulated sample is approximately equal to 0.2. The overdispersion parameter φ is taken as 2, which ensures some further overdispersion.

Censoring values C_i are simulated from a zero-truncated Poisson model with parameter μ, where μ is chosen to yield various average proportions of censored counts in the simulated data (here 0.15 and 0.3). For purpose of comparison, we also provide results that would be obtained if there were no censoring (these results will constitute a benchmark for assessing performance of the MLE when censoring is present).

The MLE of β, γ and φ are obtained by solving the score equation described in Section 4.2.1. Numerical optimization is carried out using the function `maxLik` (Henningsen and Toomet, 2011) of `R` (a free software environment for statistical computing, Team, 2018). We need to provide initial estimates to `maxLik`. We propose to obtain initial values for β and γ by fitting an uncensored ZIP model to the data, using the `R` function `zeroinfl` from package `pscl` (Jackman, 2017). For φ, note that if Z follows the ZIGP model (4.2.1), we have $\mathbb{E}(Z) = (1-\omega)\lambda$ and $\text{var}(Z) = \mathbb{E}(Z)(\varphi^2 + \lambda\omega)$, therefore,

$$\varphi = \left(\frac{\text{var}(Z)}{\mathbb{E}(Z)} - \mathbb{E}(Z)\frac{\omega}{1-\omega} \right)^{1/2}.$$

A reasonable starting value for φ can be obtained by estimating $\mathbb{E}(Z)$ and $\text{var}(Z)$ by the empirical mean and variance of the Z_i, $i = 1, \ldots, n$ (denoted by $\bar{Z}_n$ and S_n^2 respectively) and ω by the proportion $\hat{\omega} = n^{-1}\sum_{i=1}^{n} 1_{\{Z_i=0\}}$ of observations equal to 0 (note that $\hat{\omega}$ is not an estimate of the probability of zero-inflation, since some observed zeros may arise from the generalized Poisson distribution ; however, our simulations suggest that this rough approximation is sufficient to ensure a reasonable initial value for φ). Thus, we consider the following initial estimate for φ:

$$\hat{\varphi}_n^{init} = \left(\frac{S_n^2}{\bar{Z}_n} - \bar{Z}_n\frac{\hat{\omega}}{1-\hat{\omega}} \right)^{1/2}.$$

The simulation was performed 1000 times and several summary measures are obtained. Specifically, for a sample size of $n = 1000$, Table 4.1 presents the average bias, average relative bias (expressed as a percentage), average standard error, empirical standard deviation, root mean square error and corresponding empirical coverage probability for each parameter in the model (we consider 95% Wald-type confidence intervals). We also report the average length of these intervals.

Simulation design for the censored ZINB model is similar. We simulate 1000 samples from model (4.2.4) with $\log(\mu_i(\beta)) = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6}$, $\text{logit}(\omega_i(\gamma)) = \gamma_1 W_{i1} + \gamma_2 W_{i2} + \gamma_3 W_{i3} + \gamma_4 W_{i4}, + \gamma_5 W_{i5}$ (we use the same values as above for β and γ) and $\alpha = 0.5$. With these values, the average proportion of zero-inflated data in the simulated samples is 0.2. Numerical optimization is implemented via `maxLik`. Starting values for all model parameters are obtained by fitting an uncensored ZINB model to the data, with `zeroinfl`. Table 4.2 provides the same summary measures as for ZIGP model.

4.3.2 Results

From Table 4.1 and Table 4.2, we note that the MLE has generally low bias. Model-based standard errors and empirical standard deviations are close to each other for all parameters, suggesting that $\hat{\Sigma}_n$ is an adequate estimate of estimates variance.

For every censoring fraction, Wald-type confidence intervals based on model standard errors have coverage probabilities near the nominal confidence level (their average length increases with censoring, though, since standard errors increase with censoring). This correct coverage confirms that the model-based variance $\hat{\Sigma}_n$ is an adequate estimate of MLEs variance, in both

censored ZIGP and ZINB models. Unreported simulations show that as expected, bias, standard errors and average length of the confidence intervals decrease with increasing sample size, for all parameters, and that the MLE of β, φ and α (respectively γ) perform better when the proportion of zero-inflated counts decreases (respectively increases).
Wald-type confidence intervals are based on approximate normality of parameters estimates. To assess the finite-sample distribution of the MLE, we plot histograms of the normalized estimates $(\widehat{\beta}_{j,n}-\beta_j)/\text{s.e.}(\widehat{\beta}_{j,n})$, $j=1,\ldots,6$, $(\widehat{\gamma}_{k,n}-\gamma_k)/\text{s.e.}(\widehat{\gamma}_{k,n})$, $k=1,\ldots,5$, $(\widehat{\varphi}_n-\varphi)/\text{s.e.}(\widehat{\varphi}_n)$ and $(\widehat{\alpha}_n-\alpha)/\text{s.e.}(\widehat{\alpha}_n)$, where "s.e." denotes model-based standard error of the corresponding parameter. Graphs are provided for a censoring fraction equal to 0.3 (plots for 0.15 yield similar observations and are thus omitted). Histograms for ZIGP (respectively ZINB) model are given by Figures 4.1 and 4.2 (respectively Figures 4.3 and 4.4). On these graphs, the black curve represents the density function of the standard normal distribution. These graphs indicate that the distribution of the MLE can be reasonably approximated by a normal distribution, for every parameter.

Overall, these results suggest that maximum likelihood estimation yields adequate inference on both regression and overdispersion parameters in ZIGP and ZINB models, when censoring is present.

4.4 Real data application

In this section, we illustrate the censored ZIGP and ZINB models on a real data set from the German Socioeconomic Panel (a survey aimed at investigating healthcare utilization by German households). We also describe a simple and efficient methodology for selecting predictors and interactions in zero-inflation and counts components. Finally, we compare the fitted models using Vuong's test (A brief reminder of Vuong test is given in Appendix 2).

The dataset considered here contains the number of doctor office visits (the response variable) for 1812 West German men aged 25-65 years, during the last three months of 1994. Several risk factors are available, including age, socio-economic variables: marital status (1 if married, 0 otherwise), educational level (number of years of schooling), household monthly net income (in German marks/1000) and composition (coded as 1 if children under 16 live in the household, 0 otherwise), two binary variables indicating whether individual is covered by a public health insurance and by a supplemental private insurance (both are coded as 1 if yes and 0 otherwise), employment characteristics (coded as `self`: 1 if self employed, 0 otherwise ; `civil`: 1 if civil servant ; `bluec`: 1 if blue collar employee ; `employed`: 1 if employed), various measures of health status: health satisfaction (`health`, coded as 0 if low to 10 if high), handicap status (`handicap`: 1 if handicapped, 0 otherwise) and degree of handicap in percentage points (`hdegree`). Following Jochmann (2009), who first described these data, we study a more complex effect of age by considering linear spline variables `age30`, `age35`,..., `age60` (where `ageXX` is 1 if age $\geqslant$ XX and 0 otherwise). Therefore, a total of 20 candidate predictors are available. Jochmann (2009) also suggests to consider interactions between health satisfaction and age variables (i.e., `age30`$\times$`health`, `age35`$\times$`health`,...). There is no reason, however, to limit ourselves to these interactions and one may wish to assess all possible second-order interactions (except for meaningless ones, such as interactions between `ageXX` variables).

In Figure 4.5, we plot the number of doctor office visits, censored at 15 visits for illustrative purpose. The plot strongly suggests that data are zero-inflated (41.2% of the observed counts are equal to 0). Thus, we fit the following three models: i) a censored ZIGP model, ii) a censored ZIP model (obtained by letting $\varphi=1$ in (4.2.2)) and iii) a censored ZINB model, with all risk factors and second-order interactions, which results in a very large number of possible predictors. Several authors recently addressed variable selection in high-dimensional uncensored ZIP and

ZINB models via penalized maximum likelihood, and various penalty functions are implemented in the `R` package `mpath` (Wang, 2019). Thus, in a first approximation, we tried to fit penalized ZIP and ZINB models to the healthcare demand data, using all risk factors and interactions and ignoring censoring. None of the methods implemented in `mpath` converged. Therefore, we propose an alternative methodology for model fitting and variable selection in censored ZIGP and ZINB regressions:

1. First, we determine appropriate predictors for zero-inflation modelling.
   We fit a logistic regression model to the indicators $1_{\{Z_i=0\}}$, $i = 1, \ldots, n$, considered as the response variable. Note that this is not a model for zero-inflation since some of the 0 may arise from the count distribution. However, we may expect that this rough procedure will still identify a relevant subset of predictors, that will be used in a second step in the logistic model for ω_i. Given the very large number of potential predictors, we use stepwise logistic regression, starting from a model with no variables (null model). The largest possible model contain all risk factors and interactions. At each step, we use Bayesian information criterion (BIC) to select variables. Based on this strategy, we select the following predictors: `age50` and `health`.

2. In the second step, we select a preliminary set of predictors for modelling the count component of the considered zero-inflated model (ZIP, ZIGP or ZINB).
   The strategy is the same as above. For example, we use stepwise Poisson regression to select risk factors and interactions that will be used in the count component of the censored ZIP and ZIGP models. Again, variable selection is based on BIC. Starting from the null model, the chosen predictors are `age40`, `age50`, `handicap`, `hdegree`, `health`, `civil`, `self`, `health`×`hdegree`, `civil`×`age40`, `self`×`age40`.
   We use the same strategy to select a preliminary set of predictors for the count component of ZINB model. The chosen variables are `health`, `age50`, `self` and `civil`.

3. In the third step, we estimate the censored ZIP, ZIGP and ZINB models defined by

$$\text{logit}(\omega_i) = \gamma_1 + \gamma_2 \times \texttt{age50} + \gamma_3 \times \texttt{health}$$

   and

   - for ZIP and ZIGP models:

$$\begin{aligned}\lambda_i = \exp(&\beta_1 + \beta_2\,\texttt{age40} + \beta_3\,\texttt{age50} + \beta_4\,\texttt{handicap} + \beta_5\,\texttt{hdegree} + \beta_6\,\texttt{health}\\ &+\beta_7\,\texttt{civil} + \beta_8\,\texttt{self} + \beta_9\,\texttt{health}\times\texttt{hdegree} + \beta_{10}\,\texttt{civil}\times\texttt{age40}\\ &+\beta_{11}\,\texttt{self}\times\texttt{age40})\end{aligned}$$

   - for ZINB model:

$$\mu_i = \exp(\beta_1 + \beta_2\,\texttt{health} + \beta_3\,\texttt{age50} + \beta_4\,\texttt{self} + \beta_5\,\texttt{civil}).$$

   Then we use sequential elimination to obtain the final models. At each step, we remove the less significant predictor, based on Wald test at level 0.01 (if removal decreases the BIC).

Parameter estimates, standard errors and p-values of the corresponding Wald tests are given in Table 4.3. The final models are not nested, thus they are compared using Vuong test (Vuong, 1989). Results are given in Table 4.4.

We now discuss the results of our analysis. First, we observe that the decision of not seeking care is driven by age and health satisfaction. Men aged 50 years and over are less likely to waive doctor visits and the probability of renouncing doctor visits increases with health satisfaction, which is a natural finding. Then, we observe that adding a dispersion parameter has a strong beneficial impact on model fit: comparing censored ZIP and censored ZIGP (respectively ZIP and ZINB) models, Vuong statistic is -9.30 (respectively -9.59) with p-value less than 10^{-19} (respectively 10^{-21}). There is also a large difference between BIC values of final models (7843 for ZIP against 7031 for ZIGP and 7011 for ZINB), which again clearly indicates superiority of ZIGP and ZINB models over ZIP.
Except for `handicap`, ZIGP and ZINB models select the same risk factors in their count component. Both models indicate higher healthcare utilization by older men (aged 50 or more) and by those having low health satisfaction. Both models also suggest that self-employed (respectively civil servants) have lower healthcare demand than not self-employed (respectively not civil servants). In the German health insurance system, self-employed and civil servants can choose to remain uninsured. The lack of financial compensation may thus explain the fact that these individuals are less likely to visit a doctor. Vuong statistic for comparing ZIGP and ZINB models is -0.85 with p-value 0.40, which suggests that there is no statistically significant difference between the two models. Rather, it is interesting to consider their results jointly. These results confirm the presence of additional overdispersion that is not accounted for by a ZIP model, and give strong evidence of the impact of a few risk factors on healthcare demand.

4.5 Discussion

In this paper, we investigate MLEs properties in ZIGP and ZINB regression models with right-censored counts. Our simulations suggest that the MLE performs well and that reliable statistical inference on model parameters can be based on the normal approximation of MLEs distribution and on approximation of MLEs variance by Fisher information matrix derived from the censored likelihood.
Variable selection in zero-inflated models is a challenging issue. We observed that variable selection techniques based on penalized maximum likelihood can fail in the uncensored case when the number of possible predictors is too large. Moreover, penalized techniques are currently not available for censored ZI models. Therefore, we propose a simple and efficient strategy for variable selection. This strategy can be implemented using existing softwares.

Our results allow to extend the scope of ZI models to censored data. Now, several issues still deserve attention. For example, random right-censoring is only one of many possible censoring types. In practice, count data may also be left-censored or interval-censored. For now, statistical inference in ZI models in these contexts is an open question. Another question of interest relates to longitudinal data. Here, we are concerned with cross-sectional data but panel data often arise in applications. Extending the current work to the longitudinal setting is therefore of interest and constitutes a topic for our future work.

Acknowledgements

Authors are grateful to the referees and associate editor for their comments and suggestions on an earlier version of this paper. Authors acknowledge financial support from the Ministry of Education and Training of the Socialist Republic of Vietnam and the French Embassy in Vietnam and logistical support from Campus France (French national agency for the promotion of higher education, international student services, and international mobility).

average proportion of censoring		$\widehat{\beta}_n$						$\widehat{\gamma}_n$					$\widehat{\varphi}_n$
		$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$	$\widehat{\gamma}_{5,n}$	
0	bias	-0.0114	-0.0002	-0.0001	0.0021	-0.0018	0.0013	-0.0018	-0.0050	-0.0145	0.0104	-0.0019	-0.0257
	rel. bias	-1.6245	-0.2287	-0.0361	0.2461	0.3685	-	0.2036	0.7627	7.2651	1.5946	-	-1.2875
	SD	0.1199	0.0256	0.0496	0.0188	0.0365	0.0273	0.2125	0.1320	0.2768	0.1339	0.2377	0.0889
	SE	0.1162	0.0251	0.0497	0.0171	0.0352	0.0277	0.2104	0.1311	0.2654	0.1306	0.2399	0.0746
	RMSE	0.1673	0.0358	0.0702	0.0255	0.0508	0.0389	0.2990	0.1861	0.3836	0.1873	0.3376	0.1188
	CP	0.9470	0.9430	0.9470	0.9390	0.9440	0.9530	0.9530	0.9460	0.9530	0.9440	0.9520	0.9140
	ℓ	0.4544	0.0980	0.1946	0.0665	0.1377	0.1085	0.8225	0.5122	1.0362	0.5099	0.9382	0.2916
0.15	bias	-0.0107	-0.0010	0.0001	0.0039	-0.0035	0.0012	-0.0121	-0.0099	-0.0144	0.0144	-0.0014	-0.0123
	rel. bias	-1.5314	-1.0238	0.0172	0.4557	0.6930	-	1.3493	1.5292	7.1805	2.2170	-	-0.6169
	SD	0.1554	0.0365	0.0717	0.0321	0.0447	0.0378	0.2114	0.1324	0.2826	0.1325	0.2383	0.1034
	SE	0.1561	0.0352	0.0719	0.0321	0.0436	0.0381	0.2126	0.1329	0.2670	0.1311	0.2394	0.1024
	RMSE	0.2205	0.0507	0.1015	0.0455	0.0626	0.0537	0.3000	0.1878	0.3890	0.1869	0.3377	0.1460
	CP	0.9540	0.9410	0.9480	0.9460	0.9410	0.9490	0.9570	0.9460	0.9530	0.9510	0.9520	0.9420
	ℓ	0.6114	0.1376	0.2817	0.1259	0.1707	0.1494	0.8315	0.5194	1.0439	0.5123	0.9370	0.3995
0.30	bias	-0.0085	-0.0018	-0.0007	0.0059	-0.0029	0.0005	-0.0108	-0.0108	-0.0188	0.0149	-0.0009	-0.0100
	rel. bias	-1.2103	-1.7514	-0.1696	0.6893	0.5822	-	1.1946	1.6578	9.4151	2.2988	-	-0.4993
	SD	0.1936	0.0451	0.0956	0.0427	0.0510	0.0477	0.2141	0.1356	0.2875	0.1338	0.2372	0.1581
	SE	0.1908	0.0442	0.0915	0.0410	0.0508	0.0471	0.2141	0.1349	0.2717	0.1315	0.2395	0.1546
	RMSE	0.2719	0.0632	0.1323	0.0594	0.0720	0.0670	0.3029	0.1915	0.3959	0.1881	0.3370	0.2213
	CP	0.9480	0.9490	0.9370	0.9450	0.9450	0.9470	0.9570	0.9450	0.9530	0.9510	0.9520	0.9360
	ℓ	0.7465	0.1730	0.3583	0.1603	0.1987	0.1843	0.8373	0.5271	1.0617	0.5134	0.9372	0.5975

Table 4.1: Simulation results for ZIGP model. SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals.

average proportion of censoring		$\widehat{\beta}_n$						$\widehat{\gamma}_n$					$\widehat{\alpha}_n$
		$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$	$\widehat{\gamma}_{5,n}$	
0	bias	0.0089	-0.0009	0.0006	-0.0005	-0.0013	-0.0028	-0.0195	-0.0089	-0.0011	0.0080	0.0024	-0.0090
	rel. bias	1.2685	-0.8718	0.1434	-0.0560	0.2532	-	2.1721	1.3657	0.5461	1.2306	-	-1.7993
	SD	0.1472	0.0344	0.0746	0.0266	0.0404	0.0371	0.2084	0.1314	0.2540	0.1300	0.2208	0.0486
	SE	0.1522	0.0354	0.0726	0.0255	0.0411	0.0388	0.2061	0.1306	0.2580	0.1293	0.2325	0.0474
	RMSE	0.2119	0.0493	0.1041	0.0368	0.0576	0.0537	0.2937	0.1854	0.3620	0.1835	0.3206	0.0685
	CP	0.9480	0.9480	0.9480	0.9450	0.9530	0.9600	0.9530	0.9510	0.9600	0.9540	0.9640	0.9300
	ℓ	0.5960	0.1384	0.2843	0.1000	0.1609	0.1519	0.8052	0.5098	1.0089	0.5047	0.9097	0.1849
0.15	bias	0.0102	-0.0005	-0.0005	0.0015	-0.0024	-0.0034	-0.0208	-0.0093	-0.0024	0.0087	0.0025	-0.0084
	rel. bias	1.4617	-0.5322	-0.1156	0.1818	0.4867	-	2.3146	1.4248	1.1870	1.3347	-	-1.6895
	SD	0.1660	0.0398	0.0879	0.0388	0.0462	0.0427	0.2100	0.1320	0.2557	0.1305	0.2213	0.0626
	SE	0.1735	0.0413	0.0864	0.0374	0.0462	0.0448	0.2090	0.1321	0.2604	0.1303	0.2330	0.0630
	RMSE	0.2403	0.0573	0.1232	0.0539	0.0653	0.0620	0.2970	0.1869	0.3649	0.1845	0.3213	0.0892
	CP	0.9500	0.9490	0.9440	0.9400	0.9470	0.9540	0.9570	0.9580	0.9640	0.9500	0.9630	0.9360
	ℓ	0.6794	0.1616	0.3382	0.1465	0.1806	0.1753	0.8164	0.5151	1.0180	0.5082	0.9112	0.2454
0.30	bias	0.0048	-0.0022	0.0029	0.0030	-0.0025	-0.0022	-0.0234	-0.0117	-0.0005	0.0100	0.0017	-0.0067
	rel. bias	0.6814	-2.2118	0.7200	0.3526	0.4913	-	2.5974	1.8045	0.2466	1.5393	-	-1.3323
	SD	0.2000	0.0478	0.1084	0.0484	0.0545	0.0524	0.2144	0.1353	0.2593	0.1312	0.2215	0.0833
	SE	0.2047	0.0500	0.1061	0.0496	0.0539	0.0535	0.2126	0.1345	0.2643	0.1317	0.2335	0.0865
	RMSE	0.2862	0.0692	0.1517	0.0693	0.0766	0.0749	0.3027	0.1911	0.3702	0.1861	0.3218	0.1202
	CP	0.9610	0.9610	0.9400	0.9590	0.9490	0.9500	0.9560	0.9590	0.9680	0.9510	0.9640	0.9350
	ℓ	0.8010	0.1957	0.4150	0.1939	0.2107	0.2092	0.8296	0.5243	1.0325	0.5133	0.9132	0.3358

Table 4.2: Simulation results for ZINB model. SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals.

	ZIP			ZIGP			ZINB		
parameter	estimate	std. error	p-value	estimate	std. error	p-value	estimate	std. error	p-value
				Zero-inflation submodel					
intercept	-2.408137	0.216428	$<$ 2e-16	-2.61000	0.32792	1.73e-15	-2.98345	0.35760	$<$ 2e-16
health	0.300900	0.028496	$<$ 2e-16	0.24136	0.04360	3.11e-08	0.30594	0.04356	2.16e-12
age50	-0.550811	0.122265	6.64e-06	-0.60107	0.22778	0.00832	-0.65075	0.20204	0.001278
				Count submodel					
intercept	2.286474	0.047607	$<$ 2e-16	2.25776	0.08888	$<$ 2e-16	2.47741	0.09967	$<$ 2e-16
health	-0.140050	0.006565	$<$ 2e-16	-0.16361	0.01346	$<$ 2e-16	-0.19345	0.01439	$<$ 2e-16
age40	-0.083821	0.045467	0.065248†						
age50	0.234934	0.043893	8.68e-08	0.19988	0.07002	0.00431	0.26571	0.06734	7.94e-05
handicap	0.436929	0.084145	2.07e-07	0.23111	0.07416	0.00183			
self	-0.233767	0.068345	0.000625	-0.31783	0.11808	0.00711	-0.36536	0.11685	0.001768
civil	-0.553545	0.118488	2.99e-06	-0.27089	0.09936	0.00640	-0.38925	0.10194	0.000134
hdegree	-0.005052	0.001423	0.000386						
civil:age40	0.377772	0.135995	0.005472						
φ	—	—	—	1.98527	0.07430	$<$ 2e-16	—	—	—
α	—	—	—	—	—	—	0.68102	0.06884	$<$ 2e-16
AIC		7777.276			6976.789			6962.151	
BIC		7843.302			7031.811			7011.671	

Table 4.3: Summary of final censored ZIP, ZIGP and ZINB models († although not significant, `age40` remains in the model because of a significant interaction).

	ZIP vs ZIGP	ZIP vs ZINB	ZIGP vs ZINB
Vuong test	-9.30	-9.59	-0.85
p-value	$< 10^{-19}$	$< 10^{-21}$	0.40
decision	ZIGP	ZINB	equal fit

Table 4.4: Model comparison using Vuong test: Vuong statistic, p-value and test decision (i.e., the best model according to Vuong test).

4.6 Appendix 1: `R` code for fitting the censored ZINB model

The code below fits the censored ZINB model to a data set simulated as in Section 4.3. In this code, `b`, `g` and `a` represent β, γ and α respectively. Functions `dnbinom` and `pnbinom` from the `R` package `stats` calculate the density and distribution function of the negative binomial distribution (note that these functions use a slightly different parameterization for the overdispersion parameter).

Before running this code, the user needs to specify the design matrices `X` and `W` (where each row corresponds to a risk factor and the first rows are made of 1).

The following code builds the censored ZINB loglikelihood:

```
loglikfunZINB=function(param){
b=param[1:p]
g=param[(p+1):(p+q)]
a=param[p+q+1]
sum(delta*J*log(exp(t(g)%*%W)+(1+a*exp(t(b)%*%X))^(-1/a))+delta*(1-J)*
log(dnbinom(z,size=1/a,mu=exp(t(b)%*%X))))+sum((1-J)*(1-delta)
*log(1-pnbinom(z-1,size=1/a,mu=exp(t(b)%*%X)))-log(1+exp(t(g)%*%W)))
}
```

The code below determines the initial estimates of $\beta_1, \ldots, \beta_6, \gamma_1, \ldots, \gamma_5$ and α and calculates the MLE (intercepts are estimated by default by `zeroinfl`, thus it is not useful to specify X_1 and W_1 in the model formula):

```
ZINB=zeroinfl(z~X2+X3+X4+X5+X6|W2+W3+W4+W5,dist="negbin")
ZINBcensored=maxLik(logLik=loglikfunZINB,start=c(unlist(ZINB$coeff),1/ZINB$theta))
```

Estimates, standard errors and several other summaries can be obtained using the `R` function `summary`.

4.7 Appendix 2: Vuong test

The principle of the test is as follows. Let $f_0(\cdot|\cdot)$ be the true conditional density of Z given $(\mathbf{X}, \mathbf{W})$ and $f(\cdot|\cdot, \hat{\theta})$ be the estimated conditional density, where $\hat{\theta}$ is an estimate of θ (such as the MLE). Kullback-Leibler divergence between $f_0(\cdot|\cdot)$ and $f(\cdot|\cdot, \hat{\theta})$ is defined as $\mathbb{E}_0[\log f_0(Z|\mathbf{X}, \mathbf{W}) - \log f(Z|\mathbf{X}, \mathbf{W}, \hat{\theta})]$, where $\mathbb{E}_0$ denotes expectation under the true model.
If two competing models are present, one may choose the one with smallest divergence, since it is closer to the true model. For example, if model 1 is closer to the true model, we have:

$$\mathbb{E}_0[\log f_0(Z|\mathbf{X}, \mathbf{W}) - \log f(Z|\mathbf{X}, \mathbf{W}, \hat{\theta}^{(1)})] < \mathbb{E}_0[\log f_0(Z|\mathbf{X}, \mathbf{W}) - \log f(Z|\mathbf{X}, \mathbf{W}, \hat{\theta}^{(2)})],$$

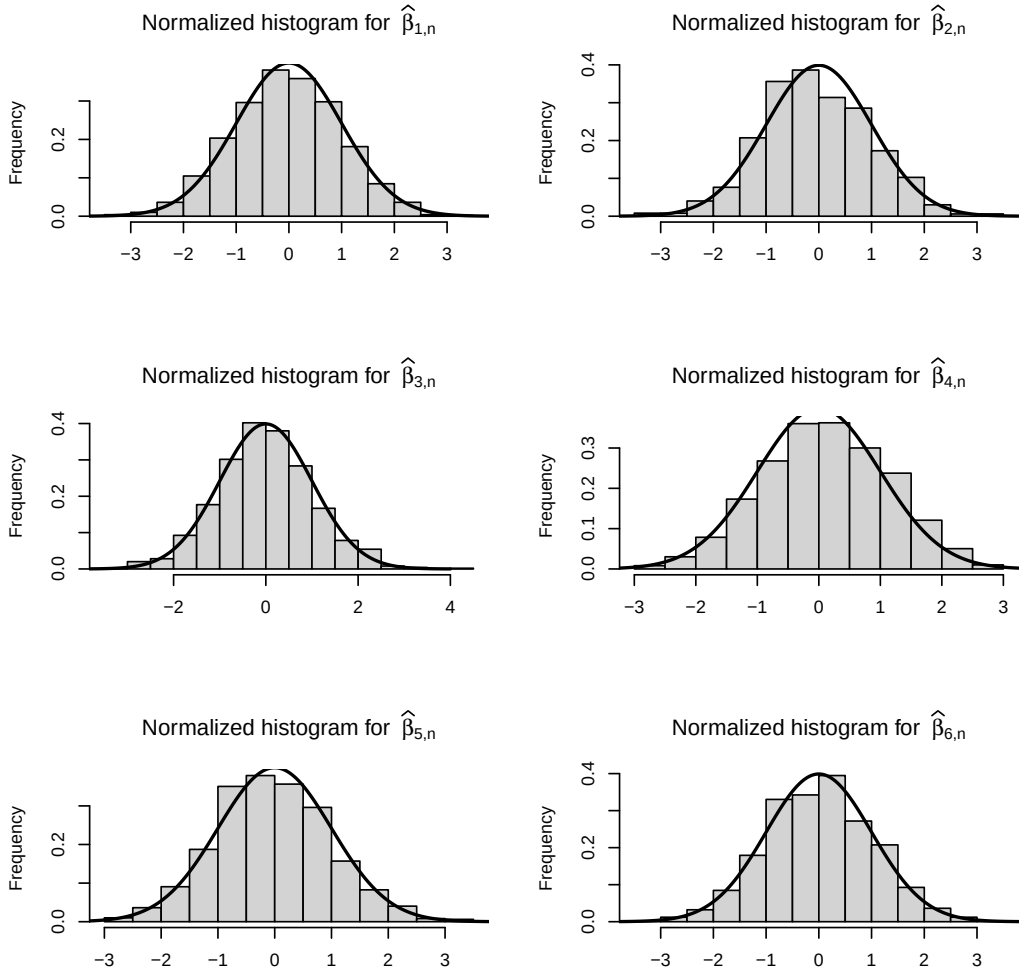

Figure 4.1: Histograms of the normalized estimates $(\widehat{\beta}_{j,n}-\beta_j)/\text{s.e.}(\widehat{\beta}_{j,n})$, $j = 1, \ldots, 6$ in censored ZIGP model (30% censoring).

where $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$ are the MLE in models 1 and 2 respectively. Equivalently,

$$\mathbb{E}_0\left[\log\frac{f(Z|\mathbf{X},\mathbf{W},\hat{\theta}^{(1)})}{f(Z|\mathbf{X},\mathbf{W},\hat{\theta}^{(2)})}\right] > 0.$$

Let $u_i = \log\frac{f(Z_i|\mathbf{X}_i,\mathbf{W}_i,\hat{\theta}^{(1)})}{f(Z_i|\mathbf{X}_i,\mathbf{W}_i,\hat{\theta}^{(2)})}$, $i = 1, \ldots, n$. Vuong test statistic is defined as

$$\mathcal{Z} = \sqrt{n}\frac{n^{-1}\sum_{i=1}^n u_i}{\sqrt{n^{-1}\sum_{i=1}^n (u_i-\bar{u}_n)^2}}.$$

Under the null hypothesis H_0 that models 1 and 2 are equally close to the true model, $\mathcal{Z}$ is asymptotically distributed as a standard normal variable. Thus, a decision rule at the asymptotic level α rejects H_0 if $|\mathcal{Z}| > z_{1-\frac{\alpha}{2}}$, where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution. If $\mathcal{Z} > z_{1-\frac{\alpha}{2}}$ (respectively $\mathcal{Z} < -z_{1-\frac{\alpha}{2}}$), the test chooses model 1 (respectively model 2).

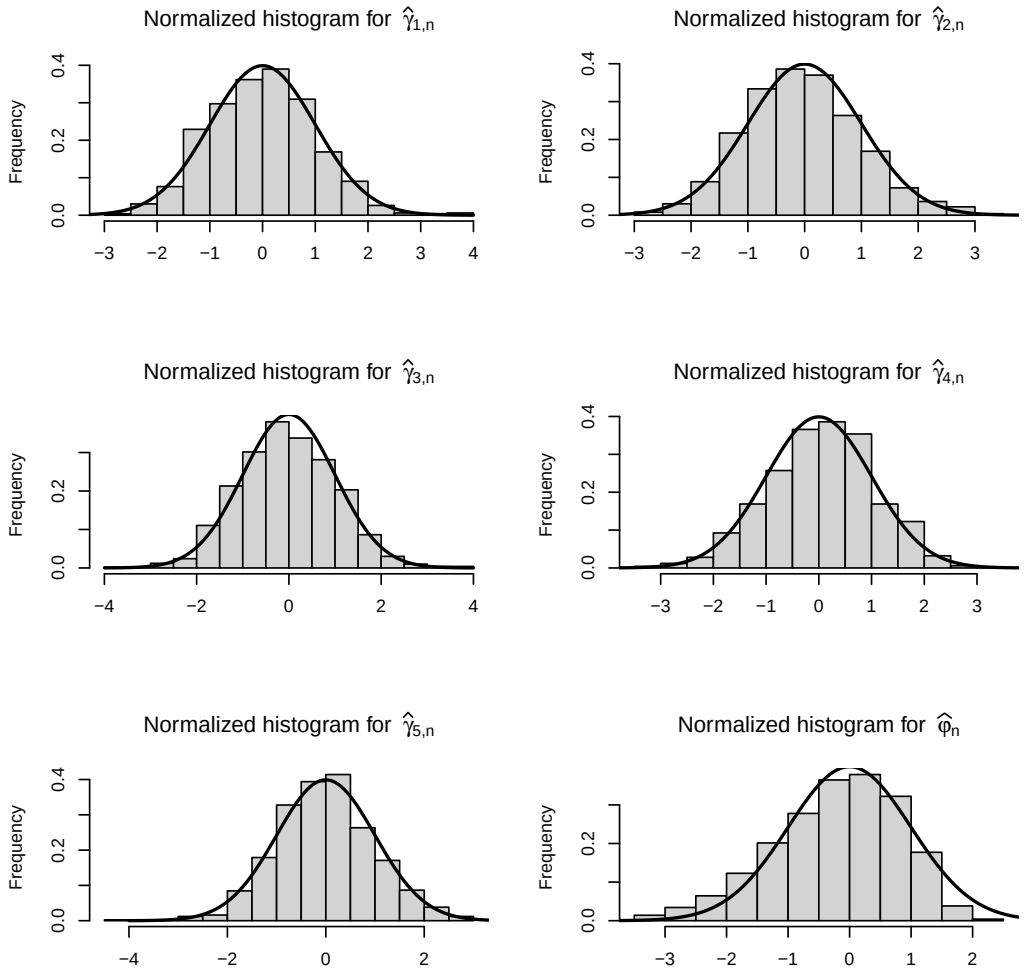

Figure 4.2: Histograms of the normalized estimates $(\widehat{\gamma}_{j,n} - \gamma_j)/\text{s.e.}(\widehat{\gamma}_{j,n})$, $j = 1, \ldots, 5$ and $(\widehat{\varphi}_n - \varphi)/\text{s.e.}(\widehat{\varphi}_n)$ in censored ZIGP model (30% censoring).

4.8 Appendix 3: Technical calculations

Zero-Inflated Generalized Poisson (ZIGP)

For the sake of convenience, we omit the β, θ, ψ in the $\lambda_i(\beta)$, $k_i(\gamma)$, $L_i(\psi)$ to become λ_i, k_i, L_i. By letting $Q_i := Q_i(\psi) = e^{\beta^\top X_i} + (\varphi - 1)k$, the density function (2.4.1) is

$$\begin{aligned}
\mathbb{P}(\tilde{Y}_i = k) &= \frac{\lambda_i(\beta)\left(\lambda_i(\beta) + (\varphi-1)k\right)^{k-1}}{k!}\varphi^{-k}e^{-\frac{\lambda_i(\beta)+(\varphi-1)k}{\varphi}} \\
&= \frac{1}{k!}\lambda_i Q_i^{k-1}\varphi^{-k}e^{-\frac{Q_i}{\varphi}}
\end{aligned}$$

for $k = 0, 1, \ldots, Y_i^* - 1$. We have some first calculations.

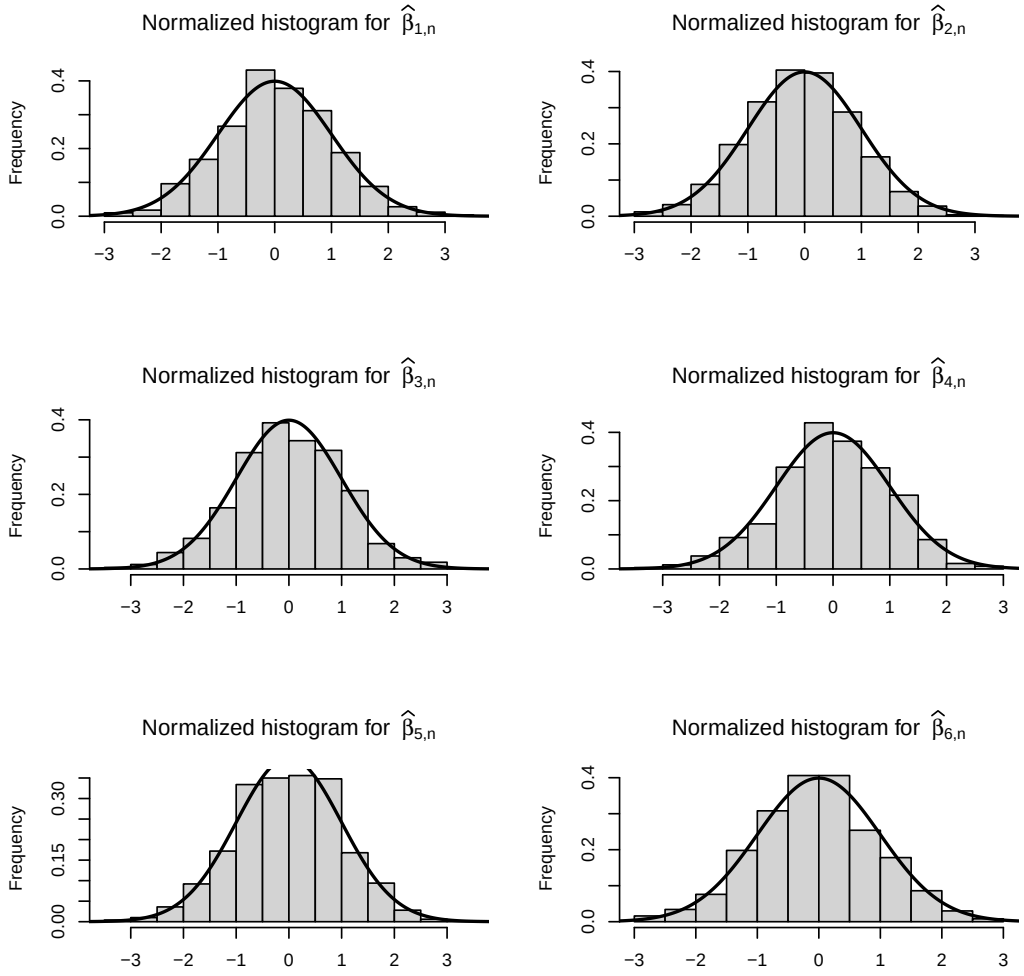

Figure 4.3: Histograms of the normalized estimates $(\widehat{\beta}_{j,n}-\beta_j)/\text{s.e.}(\widehat{\beta}_{j,n})$, $j=1,\ldots,6$ in censored ZINB model (30% censoring).

Lemma 4.8.1. *The partial derivatives in the generalized Poisson distribution are as follows:*

$$\begin{aligned}\frac{\partial}{\partial\beta_m}\mathbb{P}(\tilde{Y}_i=k) &= X_{im}\mathbb{P}(\tilde{Y}_i=k)\left(1+\frac{(k-1)\,\lambda_i}{Q_i}-\frac{\lambda_i}{\varphi}\right), \quad m=1,2,\ldots,p,\\ \frac{\partial}{\partial\varphi}\mathbb{P}(\tilde{Y}_i=k) &= \mathbb{P}(\tilde{Y}_i=k)\left(\frac{k(k-1)}{Q_i}-\frac{k}{\varphi}+\frac{\lambda_i-k}{\varphi^2}\right).\end{aligned}$$

Proof of the Lemma 4.8.1. We have:

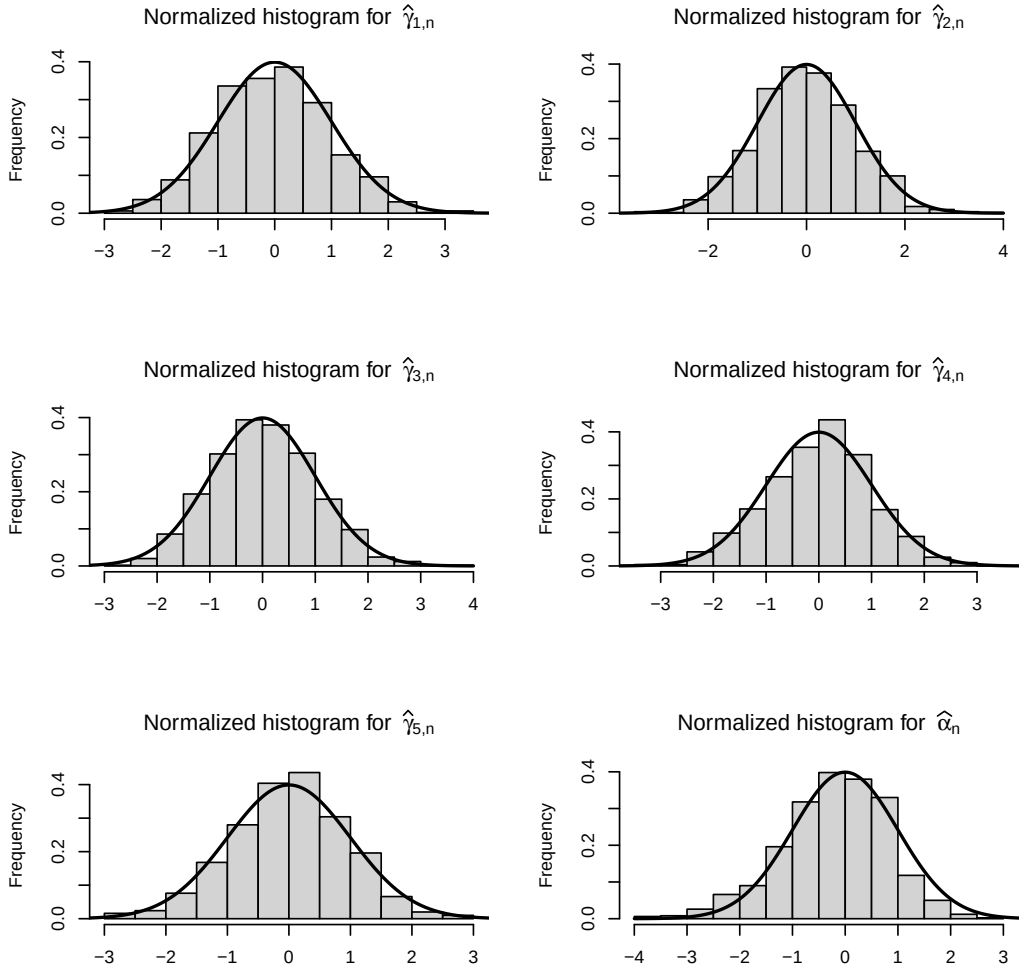

Figure 4.4: Histograms of the normalized estimates $(\widehat{\gamma}_{j,n} - \gamma_j)/\text{s.e.}(\widehat{\gamma}_{j,n})$, $j = 1, \ldots, 5$ and $(\widehat{\alpha}_n - \alpha)/\text{s.e.}(\widehat{\alpha}_n)$ in censored ZINB model (30% censoring).

$$
\begin{aligned}
\frac{\partial}{\partial \beta_m}\mathbb{P}(\tilde{Y}_i = k) &= \frac{\partial}{\partial \beta_m}\left(\frac{1}{k!}\lambda_i Q_i^{k-1}\varphi^{-k}e^{-\frac{Q_i}{\varphi}}\right), \\
&= \frac{\varphi^{-k}}{k!}\left(\lambda_i' Q_i^{k-1}e^{-\frac{Q_i}{\varphi}} + \lambda_i {Q_i^{k-1}}' e^{-\frac{Q_i}{\varphi}} + \lambda_i' Q_i^{k-1}e^{-\frac{{Q_i}'}{\varphi}}\right), \\
&= \frac{\varphi^{-k}}{k!}\left(X_{im}\lambda_i Q_i^{k-1}e^{-\frac{Q_i}{\varphi}} + \lambda_i\,(k-1)\,X_{im}Q_i^{k-2}\lambda_i e^{-\frac{Q_i}{\varphi}}\right. \\
&\qquad \left. + \lambda_i Q_i^{k-1}X_{im}e^{-\frac{Q_i}{\varphi}}\left(-\frac{\lambda_i}{\varphi}\right)\right),
\end{aligned}
$$

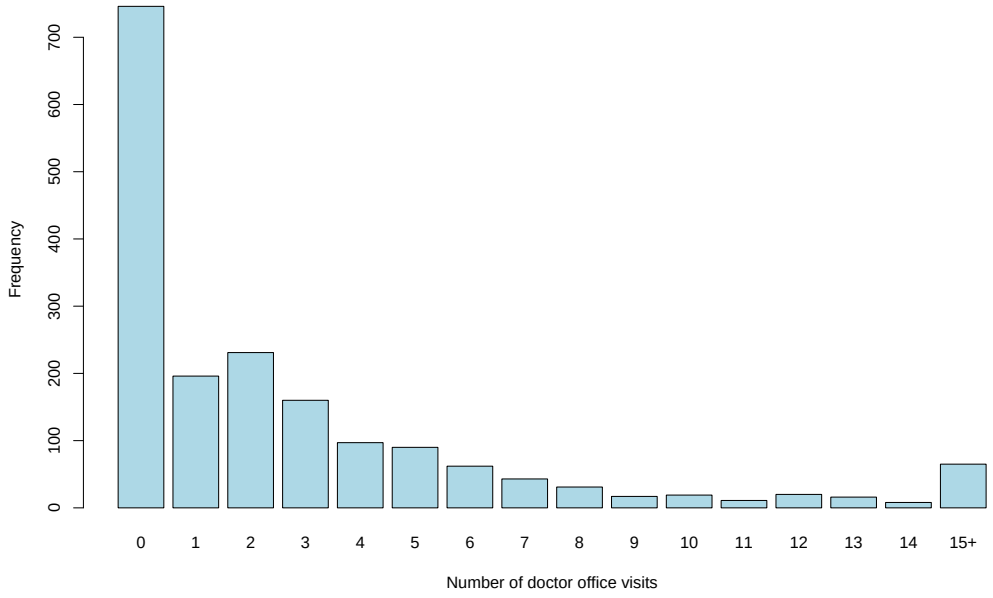

Figure 4.5: Number of doctor office visits.

$$
\begin{aligned}
&= X_{im}\frac{\lambda_i Q_i^{k-1}\varphi^{-k}e^{-\frac{Q_i}{\varphi}}}{k!}\left(1+\frac{(k-1)\lambda_i}{Q_i}-\frac{\lambda_i}{\varphi}\right),\\
&= X_{im}\mathbb{P}(\tilde{Y}_i=k)\left(1+\frac{(k-1)\,\lambda_i}{Q_i}-\frac{\lambda_i}{\varphi}\right),
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\frac{\partial}{\partial\varphi}\mathbb{P}(\tilde{Y}_i=k) &= \frac{\partial}{\partial\varphi}\left(\frac{1}{k!}\lambda_i Q_i^{k-1}\varphi^{-k}e^{-\frac{Q_i}{\varphi}}\right),\\
&= \frac{\lambda_i}{k!}\left(Q_i^{k-1}\varphi^{-k'}e^{-\frac{Q_i}{\varphi}}+Q_i^{k-1}\varphi^{-k'}e^{-\frac{Q_i}{\varphi}}+Q_i^{k-1}\varphi^{-k}e^{-\frac{Q_i}{\varphi}'}\right),\\
&= \frac{\lambda_i}{k!}\left((k-1)Q_i^{k-2}k\varphi^{-k}e^{-\frac{Q_i}{\varphi}}+Q_i^{k-1}(-k)\frac{\varphi^{-k}}{\varphi}e^{-\frac{Q_i}{\varphi}}+Q_i^{k-1}\varphi^{-k}e^{-\frac{Q_i}{\varphi}}\frac{\lambda_i-k}{\varphi^2}\right),\\
&= \frac{\lambda_i Q_i^{k-1}\varphi^{-k}e^{-\frac{Q_i}{\varphi}}}{k!}\left(\frac{k(k-1)}{Q_i}-\frac{k}{\varphi}+\frac{\lambda_i-k}{\varphi^2}\right),\\
&= \mathbb{P}(\tilde{Y}_i=k)\left(\frac{k(k-1)}{Q_i}-\frac{k}{\varphi}+\frac{\lambda_i-k}{\varphi^2}\right).
\end{aligned}
$$

□

Noting the score vector of log-likelihood function by the expression

$$\boldsymbol{s}_n(\psi) = \frac{\partial}{\partial\psi}\ell_n(\psi) = \left(\underbrace{s_{n,1}(\psi),\ldots,s_{n,p}(\psi)}_{\partial/\partial\beta_m}, \underbrace{s_{n,p+1}(\psi),\ldots,s_{n,p+q}(\psi)}_{\partial/\partial\gamma_\ell}, \underbrace{s_{n,p+q+1}(\psi)}_{\partial/\partial\varphi}\right)^\top. \quad (4.8.1)$$

for every $m = 1,\ldots,p$; $\ell = 1,\ldots,q$. Then from Lemma (4.8.1), we have:

Theorem 4.8.2 (**The score vector**).

$$\begin{aligned}
s_{n,m}(\psi) &= \sum_{i=1}^{n} X_{im}\left\{-\delta_i J_i \frac{\frac{1}{\varphi}\lambda_i L_i}{k_i+L_i} + \delta_i(1-J_i)\left(1+\frac{(Y_i^*-1)\lambda_i}{\lambda_i+(\varphi-1)Y_i^*}-\frac{\lambda_i}{\varphi}\right)\right.\\
&\quad\left.-(1-\delta_i)(1-J_i)\sum_{k=0}^{Y_i^*-1}\frac{\mathbb{P}\left(\tilde{Y}_i=k\right)}{\mathcal{S}_{\mathcal{GP}(\lambda_\rangle,\varphi)}(Y_i^*)}\left(1+\frac{(k-1)\lambda_i}{Q_i}-\frac{\lambda_i}{\varphi}\right)\right\},\\
s_{n,p+\ell}(\psi) &= \sum_{i=1}^{n} W_{i\ell}\left\{\delta_i J_i\frac{k_i}{k_i+L_i}-\frac{k_i}{1+k_i}\right\},\\
s_{n,p+q+1}(\psi) &= \sum_{i=1}^{n}\left\{\delta_i J_i\frac{\frac{1}{\varphi^2}L_i\lambda_i}{k_i+L_i}+\delta_i(1-J_i)\left(\frac{Y_i^*(Y_i^*-1)}{\lambda_i+(\varphi-1)Y_i^*}-\frac{Y_i^*}{\varphi}+\frac{\lambda_i-Y_i^*}{\varphi^2}\right)\right.\\
&\quad\left.-(1-\delta_i)(1-J_i)\sum_{k=0}^{Y_i^*-1}\frac{\mathbb{P}\left(\tilde{Y}_i=k\right)}{\mathcal{S}_{\mathcal{GP}(\lambda_\rangle,\varphi)}(Y_i^*)}\left(\frac{k(k-1)}{Q_i}-\frac{k}{\varphi}+\frac{\lambda_i-k}{\varphi^2}\right)\right\}.
\end{aligned}$$

where the survival function for generalized Poisson distribution is:

$$S_{\mathcal{GP}(\lambda_\rangle,\varphi)}(u) := \mathbb{P}\left(\tilde{Y}_i \geqslant u\right) = 1-\sum_{k=0}^{u-1}\mathbb{P}\left(\tilde{Y}_i=k\right),$$

The following two lemmas are useful for calculating the second derivatives of the loglikelihood.

Lemma 4.8.3. *With the setting of λ_i, L_i, k_i, Q_i as above, we have:*

$$\begin{aligned}
&\frac{\partial}{\partial\beta_m}\left(\frac{L_i\lambda_i}{k_i+L_i}\right) = X_{im}\frac{\lambda_i L_i}{(k_i+L_i)^2}\left[\left(1-\frac{\lambda_i}{\varphi}\right)(k_i+L_i)+\frac{L_i\lambda_i}{\varphi}\right],\\
&\frac{\partial}{\partial\varphi}\left(\frac{\frac{1}{\varphi}\lambda_i L_i}{k_i+L_i}\right) = \frac{\lambda_i L_i}{\varphi^3(k_i+L_i)^2}\left[(\lambda_i-\varphi)(k_i+L_i)-L_i\lambda_i\right],\\
&\frac{\partial}{\partial\gamma_\ell}\left(\frac{L_i\lambda_i}{k_i+L_i}\right) = W_{i\ell}\frac{-k_i\lambda_i L_i}{(k_i+L_i)^2},\\
&\frac{\partial}{\partial\beta_m}\left(\frac{\lambda_i}{\lambda_i+(\varphi-1)Y_i^*}\right) = X_{im}\frac{\lambda_i(\varphi-1)Y_i^*}{(\lambda_i+(\varphi-1)Y_i^*)^2},\\
&\frac{\partial}{\partial\varphi}\left(\frac{\lambda_i}{\lambda_i+(\varphi-1)Y_i^*}\right) = -\frac{\lambda_i Y_i^*}{(\lambda_i+(\varphi-1)Y_i^*)^2}
\end{aligned}$$

Proof of the Lemma 4.8.3. We prove the two first equalities. Calculations for the others are similar and are therefore omitted. We have:

$$\begin{aligned}
\frac{\partial}{\partial\beta_m}\left(\frac{L_i\lambda_i}{k_i+L_i}\right) &= \frac{\left(L_i'\lambda_i+L_i\lambda_i'\right)(k_i+L_i)-L_i'L_i\lambda_i}{(k_i+L_i)^2} \\
&= \frac{\left(-X_{im}\frac{L_i\lambda_i}{\varphi}\lambda_i+L_iX_{im}\lambda_i\right)(k_i+L_i)-X_{im}\frac{-L_i\lambda_i}{\varphi}L_i\lambda_i}{(k_i+L_i)^2} \\
&= X_{im}\frac{\lambda_iL_i}{(k_i+L_i)^2}\left[\left(1-\frac{\lambda_i}{\varphi}\right)(k_i+L_i)+\frac{L_i\lambda_i}{\varphi}\right]
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial}{\partial\varphi}\left(\frac{\frac{1}{\varphi}\lambda_iL_i}{k_i+L_i}\right) &= \lambda_i\frac{\left(-\frac{1}{\varphi^2}L_i+\frac{L_i\lambda_i}{\varphi^3}\right)(k_i+L_i)-\frac{1}{\varphi}L_i\frac{L_i\lambda_i}{\varphi^2}}{(k_i+L_i)^2} \\
&= \frac{\lambda_i\frac{L_i}{\varphi^3}}{(k_i+L_i)^2}\left[(\lambda_i-\varphi)(k_i+L_i)-L_i\lambda_i\right] \\
&= \frac{\lambda_iL_i}{\varphi^3(k_i+L_i)^2}\left[(\lambda_i-\varphi)(k_i+L_i)-L_i\lambda_i\right].
\end{aligned}$$

□

Lemma 4.8.4. *From the Lemmas 4.8.1, 4.8.3, it is easy to calculate the partial derivatives of the survival function with respect to β_m and φ:*

$$\frac{\partial}{\partial\beta_m}\left(S_{\mathcal{GP}(\lambda_\rangle,\varphi)}(Y_i^*)\right) = -\sum_{k=0}^{Y_i^*-1}\frac{\partial}{\partial\beta_m}\left(\mathbb{P}(\tilde{Y}_i=k)\right) = -X_{im}\sum_{k=0}^{Y_i^*-1}\mathbb{P}\left(\tilde{Y}_i=k\right)\left(1+\frac{(k-1)\lambda_i}{Q_i}-\frac{\lambda_i}{\varphi}\right),$$

$$\frac{\partial}{\partial\varphi}\left(S_{\mathcal{GP}(\lambda_\rangle,\varphi)}(Y_i^*)\right) = -\sum_{k=0}^{Y_i^*-1}\frac{\partial}{\partial\varphi}\left(\mathbb{P}(\tilde{Y}_i=k)\right) = \sum_{k=0}^{Y_i^*-1}\mathbb{P}\left(\tilde{Y}_i=k\right)\left(\frac{k(k-1)}{Q_i}-\frac{k}{\varphi}+\frac{\lambda_i-k}{\varphi^2}\right).$$

Finally, these preliminary calculations allow us to obtain the second derivatives, as:

$$\begin{aligned}
\frac{\partial^2 \ell_n(\psi)}{\partial\beta_m\partial\beta_\ell} &= \sum_{i=1}^{n} X_{im}X_{i\ell}\left\{-\delta_i J_i \frac{\frac{1}{\varphi}\lambda_i L_i}{(k_i+L_i)^2}\left[\left(1-\frac{\lambda_i}{\varphi}\right)(k_i+L_i)+\frac{L_i\lambda_i}{\varphi}\right]\right. \\
&\quad + \delta_i(1-J_i)\left[(Y_i^*-1)\,Y_i^*\frac{\lambda_i(\varphi-1)}{(\lambda_i+(\varphi-1)Y_i^*)^2}-\frac{\lambda_i}{\varphi}\right] \\
&\quad \left.-(1-\delta_i)(1-J_i)\sum_{k=0}^{Y_i^*-1}\frac{\mathbb{P}\left(\tilde{Y}_i=k\right)}{S^2_{\mathcal{GP}(\lambda_\rangle,\varphi)}(Y_i^*)}M_i(\psi)\right\}, \quad \forall m,\ell=1,\ldots,p, \\
\frac{\partial^2 \ell_n(\psi)}{\partial\beta_m\partial\gamma_\ell} &= \sum_{i=1}^{n} X_{im}W_{i\ell}\delta_i J_i\frac{\frac{1}{\varphi}k_i\lambda_i L_i}{(k_i+L_i)^2}, \quad \forall m=1,\ldots,p;\ \ell=1,\ldots,q, \\
\frac{\partial^2 \ell_n(\psi)}{\partial\beta_m\partial\varphi} &= \sum_{i=1}^{n} X_{im}\left\{-\delta_i J_i\frac{\lambda_i L_i}{\varphi^3(k_i+L_i)^2}\left[(\lambda_i-\varphi)(k_i+L_i)-Li\lambda_i\right]\right. \\
&\quad + \delta_i(1-J_i)\left[-\frac{\lambda_i Y_i^*(Y_i^*-1)}{(\lambda_i+(\varphi-1)Y_i^*)^2}+\frac{\lambda_i}{\varphi}\right] \\
&\quad \left.-(1-\delta_i)(1-J_i)\sum_{i=1}^{Y_i^*-1}\frac{\mathbb{P}\left(\tilde{Y}_i=k\right)}{S^2_{\mathcal{GP}(\lambda_\rangle,\varphi)}(Y_i^*)}N_i(\psi)\right\}, \quad \forall m=1,\ldots,p, \\
\frac{\partial^2 \ell_n(\psi)}{\partial\gamma_\ell\partial\gamma_m} &= \sum_{i=1}^{n} W_{i\ell}W_{im}\left\{\delta_i J_i\frac{k_i L_i}{(k_i+L_i)^2}-\frac{k_i}{(1+k_i)^2}\right\}, \quad \forall \ell,m=1,\ldots,q, \\
\frac{\partial^2 \ell_n(\psi)}{\partial\gamma_\ell\partial\varphi} &= -\sum_{i=1}^{n} W_{i\ell}\delta_i J_i\left\{\frac{\frac{1}{\varphi^2}k_i L_i\lambda_i}{(k_i+L_i)^2}\right\}, \quad \forall \ell=1,\ldots,q, \\
\frac{\partial^2 \ell_n(\psi)}{\partial\varphi^2} &= \sum_{i=1}^{n}\left\{\delta_i J_i\frac{1}{\varphi^4}L_i\frac{(\lambda_i-2\varphi L_i)(k_i+L_i)-L_i\lambda_i}{(k_i+L_i)^2}\right. \\
&\quad + \delta_i(1-J_i)\left(-\frac{Y_i^{*2}(Y_i^*-1)}{(\lambda_i+(\varphi-1)Y_i^*)^2}+\frac{Y_i^*}{\varphi^2}-\frac{2(\lambda_i-Y_i^*)}{\varphi^3}\right) \\
&\quad \left.-(1-\delta_i)(1-J_i)\sum_{k=0}^{Y_i^*-1}\frac{\mathbb{P}\left(\tilde{Y}_i=k\right)}{S^2_{\mathcal{GP}(\lambda_\rangle,\varphi)}(Y_i^*)}R_i(\psi)\right\},
\end{aligned}$$

where

$$\begin{aligned}
M_i(\psi) &= \left(\left(1+\frac{(k-1)\lambda_i}{Q_i}-\frac{\lambda_i}{\varphi}\right)^2+\frac{\lambda_i(\varphi-1)k(k-1)}{Q_i^2}-\frac{\lambda_i}{\varphi}\right)S_{\mathcal{GP}(\lambda_\rangle,\varphi)}(Y_i^*) \\
&\quad +\left(1+\frac{(k-1)\lambda_i}{Q_i}-\frac{\lambda_i}{\varphi}\right)\sum_{j=0}^{Y_i^*-1}\mathbb{P}\left(\tilde{Y}_i=j\right)\left(1+\frac{(j-1)\lambda_i}{Q_i}-\frac{\lambda_i}{\varphi}\right),
\end{aligned}$$

$$
\begin{aligned}
N_i(\psi) &= \left(\left(\frac{k(k-1)}{Q_i} - \frac{k}{\varphi} + \frac{\lambda_i - k}{\varphi}\right)\left(1 + \frac{(k-1)\lambda_i}{Q_i} - \frac{\lambda_i}{\varphi}\right) - \frac{k(k-1)\lambda_i}{Q_i^2} + \frac{\lambda_i}{\varphi^2}\right) S_{\mathcal{GP}(\lambda_\rangle,\varphi)}(Y_i^*) \\
&\quad - \left(1 + \frac{(k-1)\lambda_i}{Q_i} - \frac{\lambda_i}{\varphi}\right) \sum_{j=0}^{Y_i^*-1} \mathbb{P}\left(\tilde{Y}_i = j\right)\left(\frac{j(j-1)}{Q_i} - \frac{j}{\varphi} + \frac{\lambda_i - j}{\varphi^2}\right), \\
R_i(\psi) &= \left(\left(\frac{k(k-1)}{Q_i} - \frac{k}{\varphi} + \frac{\lambda_i - k}{\varphi^2}\right)^2 - \frac{k^2(k-1)}{Q_i} + \frac{k}{\varphi^2} - \frac{2\,(\lambda_i - k)}{\varphi^3}\right) S_{\mathcal{GP}(\lambda_\rangle,\varphi)}(Y_i^*) \\
&\quad - \left(\frac{k(k-1)}{Q_i} - \frac{k}{\varphi} + \frac{\lambda_i - k}{\varphi^2}\right) \sum_{j=0}^{Y_i^*-1} \mathbb{P}(\tilde{Y}_i = j)\left(\frac{j(j-1)}{Q_i} - \frac{j}{\varphi} + \frac{\lambda_i - j}{\varphi^2}\right).
\end{aligned}
$$

Zero-Inflated negative binomial (ZINB)

For simplicity of exposition, we let $\mu_i := \mu_i(\beta)$. Then, we have the result:

Lemma 4.8.5. *Assume that $i = 1, \ldots, n$, a random variable $\tilde{Y}_i \sim \mathcal{NB}(\mu_i, \alpha)$ draws from the negative binomial of the form (2.3.1)*

$$
\mathbb{P}(\tilde{Y}_i = k|\mu_i, \alpha) = \frac{\Gamma(k + \alpha^{-1})}{k!\Gamma(\alpha^{-1})}\left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^k \left(\frac{1}{1 + \alpha\mu_i}\right)^{\frac{1}{\alpha}}, \quad k = 0, 1, \ldots,
$$

We then have:

$$
\begin{aligned}
\frac{\partial}{\partial\beta_m}\mathbb{P}(\tilde{Y}_i = k|\mu_i, \alpha) &= X_{im}\mathbb{P}(\tilde{Y}_i = k|\mu_i, \alpha)\left(\frac{k - \mu_i}{1 + \alpha\mu_i}\right), \\
\frac{\partial}{\partial\alpha}\mathbb{P}(\tilde{Y}_i = k|\mu_i, \alpha) &= \mathbb{P}(\tilde{Y}_i = k|\mu_i, \alpha)\left(\frac{1}{\alpha^2}\left(\log(1 + \alpha\mu_i) - \sum_{j=0}^{k-1}\frac{1}{\frac{1}{\alpha} + j}\right) + \frac{k - \mu_i}{\alpha(1 + \alpha\mu_i)}\right).
\end{aligned}
$$

Proof of the Lemma 4.8.5. Indeed, we have:

$$
\begin{aligned}
\frac{\partial}{\partial\beta_m}\mathbb{P}(\tilde{Y}_i = k|\mu_i, \alpha) &= \frac{\Gamma(k + \alpha^{-1})}{k!\Gamma(\alpha^{-1})}\left\{\left[\left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^k\right]'\left(\frac{1}{1 + \alpha\mu_i}\right)^{\frac{1}{\alpha}} + \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^k\left[\left(\frac{1}{1 + \alpha\mu_i}\right)^{\frac{1}{\alpha}}\right]'\right\}, \\
&= \frac{\Gamma(k + \alpha^{-1})}{k!\Gamma(\alpha^{-1})}\left\{k\left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{k-1}\frac{\alpha\mu_i X_{im}}{(1 + \alpha\mu_i)^2}\left(\frac{1}{1 + \alpha\mu_i}\right)^{\frac{1}{\alpha}}\right. \\
&\qquad \left. - \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^k\frac{1}{\alpha}\left(\frac{1}{1 + \alpha\mu_i}\right)^{\frac{1}{\alpha}-1}\frac{1}{(1 + \alpha\mu_i)^2}\alpha\mu_i X_{im}\right\}, \\
&= X_{im}\frac{\Gamma(k + \alpha^{-1})}{k!\Gamma(\alpha^{-1})}\left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^k\left(\frac{1}{1 + \alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{k}{1 + \alpha\mu_i} - \frac{\mu_i}{1 + \alpha\mu_i}\right), \\
&= X_{im}\mathbb{P}(\tilde{Y}_i = k|\mu_i, \alpha)\left(\frac{k - \mu_i}{1 + \alpha\mu_i}\right),
\end{aligned}
$$

note that $\forall k \in \mathbb{N}^*$, and $\alpha^{-1} > 0$, we have:

$$
\frac{\Gamma(\alpha^{-1} + k)}{\Gamma(\alpha^{-1})} = \prod_{j=0}^{k-1}(\alpha^{-1} + j) \longrightarrow \left[\frac{\Gamma(\alpha^{-1} + k)}{\Gamma(\alpha^{-1})}\right]' = -\frac{1}{\alpha^2}\frac{\Gamma(\alpha^{-1} + k)}{\Gamma(\alpha^{-1})}\sum_{j=0}^{k-1}\frac{1}{\alpha^{-1} + j},
$$

therefore

$$\begin{aligned}
\frac{\partial}{\partial\alpha}\mathbb{P}(\tilde{Y}_i=k|\mu_i,\alpha) &= \frac{1}{k!}\left\{\left[\frac{\Gamma(k+\alpha^{-1})}{\Gamma(\alpha^{-1})}\right]'\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^k\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\right.\\
&\quad+\frac{\Gamma(k+\alpha^{-1})}{\Gamma(\alpha^{-1})}\left[\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^k\right]'\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\\
&\quad\left.+\frac{\Gamma(k+\alpha^{-1})}{\Gamma(\alpha^{-1})}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^k\left[\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\right]'\right\},\\
&= \frac{1}{k!}\left\{-\frac{1}{\alpha^2}\sum_{j=0}^{k-1}\frac{1}{\alpha^{-1}+j}\cdot\frac{\Gamma(k+\alpha^{-1})}{\Gamma(\alpha^{-1})}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^k\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\right.\\
&\quad+\frac{\Gamma(k+\alpha^{-1})}{\Gamma(\alpha^{-1})}k\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{k-1}\frac{\mu_i}{(1+\alpha\mu_i)^2}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\\
&\quad\left.+\frac{\Gamma(k+\alpha^{-1})}{\Gamma(\alpha^{-1})}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^k\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left[\frac{\log(1+\alpha\mu_i)}{\alpha^2}-\frac{\mu_i}{\alpha(1+\alpha\mu_i)}\right]\right\},\\
&= \mathbb{P}(\tilde{Y}_i=k|\mu_i,\alpha)\left(\frac{1}{\alpha^2}\left(\log(1+\alpha\mu_i)-\sum_{j=0}^{k-1}\frac{1}{\alpha^{-1}+j}\right)+\frac{k-\mu_i}{\alpha(1+\alpha\mu_i)}\right).
\end{aligned}$$

□

Setting $k_i=e^{\gamma^\top\mathbf{w}_i}$, $L_i=1+e^{\beta^\top\mathbf{x}_i}$, and the survival function $\mathcal{S}_{\mathcal{NB}(\mu_i,\alpha)}(Y_i^*)=\mathbb{P}(\mathcal{NB}(\mu_i,\alpha)\geqslant Y_i^*)=1-\sum_{k=0}^{Y_i^*-1}\mathbb{P}(\mathcal{NB}(\mu_i,\alpha)=k)$, the loglikelihood function turns out

$$\begin{aligned}
\ell_n(\psi) &= \sum_{i=1}^n\left[\delta_i J_i\log\left(k_i+L_i^{-\frac{1}{\alpha}}\right)-\log(1+k_i)\right]\\
&\quad+\sum_{i=1}^n\delta_i(1-J_i)\left[Y_i^*\beta^\top\mathbf{X}_i+Y_i^*\log\alpha-(Y_i^*+\frac{1}{\alpha})\log L_i\right.\\
&\qquad\left.+\log\Gamma(Y_i^*+\frac{1}{\alpha})-\log\Gamma(\frac{1}{\alpha})-\log(Y_i^*!)\right]\\
&\quad+\sum_{i=1}^n(1-\delta_i)(1-J_i)\log\mathcal{S}_{\mathcal{NB}(\mu_i,\alpha)}(Y_i^*),
\end{aligned}$$

From the Lemma 4.8.5, we have the explicit derivative expression of the survival function of the negative binomial regression model as follows

Lemma 4.8.6 (**The derivative of the survival function**).

$$\begin{aligned}
\frac{\partial}{\partial\beta_m}\mathcal{S}_{\mathcal{NB}(\mu_i,\alpha)}(Y_i^*) &= -\sum_{k=1}^{Y_i^*-1}X_{im}\mathbb{P}(\tilde{Y}_i=k)\frac{k-\mu_i}{1+\alpha\mu_i},\\
\frac{\partial}{\partial\alpha}\mathcal{S}_{\mathcal{NB}(\mu_i,\alpha)}(Y_i^*) &= -\sum_{k=1}^{Y_i^*-1}\mathbb{P}(\tilde{Y}_i=k)\left[\frac{1}{\alpha^2}\left(\log(1+\alpha\mu_i)-\sum_{j=0}^{k-1}\frac{1}{\alpha^{-1}+j}\right)+\frac{k-\mu_i}{\alpha(1+\alpha\mu_i)}\right].
\end{aligned}$$

Noting the score vector of log-likelihood function by the expression

$$\boldsymbol{s}_n(\psi) = \frac{\partial}{\partial \psi}\ell_n(\psi) = \left(\underbrace{s_{n,1}(\psi), \ldots, s_{n,p}(\psi)}_{\partial/\partial\beta_m}, \underbrace{s_{n,p+1}(\psi), \ldots, s_{n,p+q}(\psi)}_{\partial/\partial\gamma_\ell}, \underbrace{s_{n,p+q+1}(\psi)}_{\partial/\partial\alpha} \right)^\top . \quad (4.8.2)$$

for every $m = 1, \ldots, p$; $\ell = 1, \ldots, q$. Then from the two previous Lemma 4.8.5, and 4.8.6, with some straightforward calculating, we have the first derivative:

Theorem 4.8.7 (**The score vector**). *With the setting as above, the derivatives of the loglikelihood with respect to each corresponding element are:*

$$\begin{aligned}
\frac{\partial \ell_n(\psi)}{\partial \beta_m} &= \sum_{i=1}^{k} X_{im} \left\{ \delta_i J_i \frac{-L_i^{-\frac{1}{\alpha}-1}\mu_i}{k_i + L_i^{-\frac{1}{\alpha}}} + \delta_i(1-J_i)\left(Y_i^* - (Y_i^* + \frac{1}{\alpha})\frac{\alpha\mu_i}{L_i}\right) \right. \\
&\quad \left. -(1-\delta_i)(1-J_i)\sum_{k=0}^{Y_i^*-1} \frac{\mathbb{P}(\tilde{Y}_i = k)}{\mathcal{S}_{\mathcal{NB}(\mu_i,\alpha)}(Y_i^*)} \frac{k-\mu_i}{1+\alpha\mu_i} \right\}, \\
\frac{\partial \ell_n(\psi)}{\partial \gamma_\ell} &= \sum_{i=1}^{k} W_{i\ell} \left\{ \delta_i J_i \frac{k_i}{k_i + L_i^{-\frac{1}{\alpha}}} - \frac{k_i}{1+k_i} \right\}, \\
\frac{\partial \ell_n(\psi)}{\partial \alpha} &= \sum_{i=1}^{k} \left\{ -\delta_i J_i \frac{L_i^{-\frac{1}{\alpha}}}{\alpha(k_i + L_i^{-\frac{1}{\alpha}})} \left(\frac{\log L_i}{\alpha^2} - \frac{\mu_i}{\alpha L_i} \right) \right. \\
&\quad + \delta_i(1-J_i)\left(\frac{Y_i^*}{\alpha} + \frac{1}{\alpha^2}\log L_i - (Y_i^* + \frac{1}{\alpha})\mu_i - \frac{1}{\alpha^2}\sum_{k=0}^{Y_i^*-1}(\frac{1}{\alpha}+k) \right) \\
&\quad \left. -(1-\delta_i)(1-J_i)\sum_{k=0}^{Y_i^*-1} \frac{\mathbb{P}(\tilde{Y}_i = k)}{\mathcal{S}_{\mathcal{NB}(\mu_i,\alpha)}(Y_i^*)} \left[\frac{1}{\alpha^2}\left(\log L_i - \sum_{j=0}^{k-1} \frac{1}{\frac{1}{\alpha}+j} \right) + \frac{k-\mu_i}{\alpha L_i} \right] \right\}.
\end{aligned}$$

Next, we state the second derivative expressions without proof.

Theorem 4.8.8 (**The second derivatives**).

$$\begin{aligned}
\frac{\partial^2 \ell_n(\psi)}{\partial \beta_m \partial \beta_\ell} &= \sum_{i=1}^{n} X_{im} X_{i\ell} \left\{ \delta_i J_i \frac{\mu_i L^{-\frac{1}{\alpha}-2}\left[((\alpha+1)\mu_i - L_i)\left(k_i + L_i^{-\frac{1}{\alpha}}\right) - \mu_i L_i^{-\frac{1}{\alpha}} \right]}{\left(k_i + L_i^{-\frac{1}{\alpha}}\right)^2} \right. \\
&\quad -\delta_i(1-J_i)\left(Y_i^* + \frac{1}{\alpha}\right)\frac{\alpha\mu_i}{L_i} \\
&\quad -(1-\delta_i)(1-J_i)\sum_{k=0}^{Y_i^*-1}\left[\mathbb{P}(\tilde{Y}_i = k)\frac{(k-\mu_i)^2 - \mu_i(1+\alpha k)}{L_i^2}\mathcal{S}_{\mathcal{NB}(\mu_i,\alpha)}(Y_i^*) \right. \\
&\quad \left. \left. + \left(\sum_{k=0}^{Y_i^*-1} \mathbb{P}(\tilde{Y}_i = k)\frac{k-\mu_i}{L_i} \right)^2 \right] \Big/ \mathcal{S}_{\mathcal{NB}(\mu_i,\alpha)}(Y_i^*)^2 \right\}, \quad \forall m, \ell = 1, \ldots, p,
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 \ell_n(\psi)}{\partial \beta_m \partial \gamma_\ell} &= \sum_{i=1}^{n} X_{im} W_{i\ell} \left\{ \delta_i J_i \frac{L_i^{-\frac{1}{\alpha}-1} \mu_i k_i}{\left(k_i + L_i^{-\frac{1}{\alpha}}\right)^2} \right\}, \quad \forall m = 1, \ldots, p, \forall \ell = 1, \ldots, q, \\
\frac{\partial^2 \ell_n(\psi)}{\partial \beta_m \partial \alpha} &= \sum_{i=1}^{n} X_{im} \left\{ -\delta_i J_i \frac{\mu_i L_i^{-\frac{1}{\alpha}-1} \left(\frac{k_i \log L_i}{\alpha^2} - \left(\frac{1}{\alpha}+1\right) \frac{k_i \mu_i}{L_i} - \mu_i L_i^{-\frac{1}{\alpha}-1} \right)}{\left(k_i + L_i^{-\frac{1}{\alpha}}\right)^2} \right. \\
&\quad -\delta_i (1-J_i) \left(-\frac{\mu_i}{\alpha L_i} + \left(Y_i^* + \frac{1}{\alpha}\right) \frac{(L_i - \alpha\mu_i)\mu_i}{L_i^2} \right) \\
&\quad \left. -(1-\delta_i)(1-J_i) K_i(\psi) \right\}, \quad \forall m = 1, \ldots, p, \\
\frac{\partial^2 \ell_n(\psi)}{\partial \gamma_\ell \partial \gamma_m} &= \sum_{i=1}^{n} W_{i\ell} W_{im} \left\{ \delta_i J_i \frac{L_i^{-\frac{1}{\alpha}}}{\left(k_i + L_i^{-\frac{1}{\alpha}}\right)^2} - \frac{k_i}{(1+k_i)^2} \right\}, \quad \forall m, \ell = 1, \ldots, q, \\
\frac{\partial^2 \ell_n(\psi)}{\partial \gamma_\ell \partial \alpha} &= -\sum_{i=1}^{n} W_{i\ell} \left\{ \delta_i J_i \frac{k_i}{\left(k_i + L_i^{-\frac{1}{\alpha}}\right)^2} L_i^{-\frac{1}{\alpha}} \left(\frac{1}{\alpha^2} \log L_i - \frac{\mu_i}{\alpha L_i} \right) \right\}, \quad \forall \ell = 1, \ldots, q, \\
\frac{\partial^2 \ell_n(\psi)}{\partial \alpha^2} &= -\sum_{i=1}^{n} \left\{ \delta_i J_i E_i(\psi) + \delta_i (1-J_i) F_i(\psi) - (1-\delta_i)(1-J_i) H_i(\psi) \right\},
\end{aligned}
$$

where

$$
\begin{aligned}
E_i(\psi) &= \frac{k_i L_i^{-\frac{1}{\alpha}}}{\left(k_i + L_i^{-\frac{1}{\alpha}}\right)^2} \left(\frac{\log L_i}{\alpha^2} - \frac{\mu_i}{\alpha L_i} \right)^2 + \frac{L_i^{-\frac{1}{\alpha}}}{k_i + L_i^{-\frac{1}{\alpha}}} \left(\frac{2\mu_i}{\alpha^2 L_i} - \frac{2\log L_i}{\alpha^3} + \frac{\mu_i}{\alpha L_i^2} \right), \\
F_i(\psi) &= \frac{\mu_i - Y_i^*}{\alpha^2} - \frac{2}{\alpha^3} \left(\log L_i - \sum_{k=0}^{Y_i^*-1} \frac{1}{\alpha^{-1}+k} \right) + \frac{1}{\alpha^2} \left(\frac{\mu_i}{L_i} - \sum_{k=0}^{Y_i^*-1} \frac{1}{(1+\alpha k)^2} \right),
\end{aligned}
$$

and

$$
\begin{aligned}
H_i(\psi) &= \left[\left(\sum_{k=0}^{Y_i^*-1} \mathbb{P}(\tilde{Y}_i = k) G_i(\psi) \right) \mathcal{S}_{\mathcal{NB}(\mu_i,\alpha)}(Y_i^*) \right. \\
&\quad \left. + \left(\sum_{k=0}^{Y_i^*-1} \mathbb{P}(\tilde{Y}_i = k) \left(\frac{1}{\alpha^2} \left(\log L_i - \sum_{j=0}^{k-1} \frac{1}{\alpha^{-1}+j} \right) + \frac{k-\mu_i}{\alpha L_i} \right) \right)^2 \right] \Big/ \mathcal{S}_{\mathcal{NB}(\mu_i,\alpha)}(Y_i^*)^2, \\
G_i(\psi) &= \left[\frac{1}{\alpha^2} \left(\ln L_i - \sum_{j=0}^{k-1} \frac{1}{\alpha^{-1}+j} \right) + \frac{k-\mu_i}{\alpha L_i} \right]^2 + \left[\frac{\mu_i}{\alpha^2 L_i} - \frac{2\ln L_i}{\alpha^3} + \sum_{j=0}^{k-1} \frac{2\alpha+1}{\alpha^2(1+\alpha)^2} - \frac{(k-\mu_i)\mu_i}{L_i^2} \right].
\end{aligned}
$$

5 Marginalized zero-inflated Poisson regression with right-censored counts

Contents

5.1 Introduction **76**
5.2 The censored MZIP regression model **77**
5.2.1 Loglikelihood estimation in MZIP regression model 77
5.2.2 Loglikelihood estimation in MZIP regression model with right-censored 78
5.2.3 Some further notations 79
5.2.4 Regularity conditions and asymptotic results 80
5.3 Simulation study **81**
5.3.1 Simulation design 81
5.3.2 Results 82
5.4 Real data application **89**
5.5 Conclusion **89**
5.6 Appendix 1: Technical Lemmas **91**
5.7 Appendix 2: Technical calculations **96**

Abstract
The marginalized zero-inflated Poisson regression model is often used to analyze the covariate effects on the overall mean response. This article extends the model to randomly right-censored count data. In this model, maximum likelihood estimators are constructed and their properties are rigorously established. A simulation study is performed to assess finite-sample behavior in various scenarios. An application in health care utilization is also illustrated.

Keyword: Count data, excess of zeros, marginalized model, large-sample properties, simulations

Abstract in french

Dans ce chapitre, nous nous intéressons à l'inférence statistique dans le modèle de Poisson zéro-inflaté marginal (modèle MZIP), en présence d'une censure aléatoire à droite portant sur les données de comptage. Dans le modèle MZIP, contrairement au modèle ZIP classique des précédents chapitres, c'est l'espérance mathématique de la loi marginale du comptage qui est modélisée en fonction des variables explicatives, et non l'espérance de la loi du comptage dans la population à risque. Cette distinction permet d'interpréter l'influence des variables explicatives directement sur la population entière, ce qui est parfois plus pertinent au regard des questions que se pose le statisticien. Nous construisons des estimateurs du maximum de vraisemblance de l'ensemble des paramètres du modèle MZIP, puis nous en établissons rigoureusement les

propriétés asymptotiques (consistance, normalité asymptotique, estimation convergente de la variance asymptotique). De nouveau, nous réalisons une étude de simulation approfondie, afin d'évaluer les propriétés à distance finie de l'estimateur du maximum de vraisemblance, en fonction de divers paramètres : taille d'échantillon, proportion d'inflation de zéro, proportion de censure. Les indicateurs numériques et représentations graphiques que nous obtenons confirment la qualité de l'estimateur proposé. Une application en soci-économie de la santé est enfin décrite.

5.1 Introduction

Zero-inflated regression models which view data as being generated from a mixture of a degenerated distribution with mass point of one at zeros and a standard count regression model have become a popular power statistical tool to analyze count data with excess of zeros.
A large number of well-known applications are conducted in a variety of disciplines, such as ecology, epidemiology, health care utilization, industry, assurance and so on. Zero-inflated Poisson regression model was proposed by Lambert (1992) and further developed by Böhning (2000), Lim et al. (2014), Monod (2014), Fu et al. (2018), recently by Nguyen and Dupuy (2019b), and references therein. Recent variants of ZIP regression include random-effects ZIP models (Hall, 2000; Min and Agresti, 2005) and semiparametric ZIP models (Lam et al., 2006; Feng and Zhu, 2011). Zero-inflated negative binomial (ZINB) regression model was proposed by Ridout et al. (2001), see also Moghimbeigi et al. (2008) and Mwalili et al. (2008). When counts have an upper bound, ZIP and ZINB regression models are no longer appropriate. Hall (2000) thus introduced the zero-inflated binomial (ZIB) model, see also Hall and Berenhaut (2002), Diop et al. (2011), Diop et al. (2016) and Diallo et al. (2017a). Statistical modeling of bounded count data containing both extra zeros and extra right-endpoints has recently appealing attention. Deng and Zhang (2015) proposed a zero-one inflated binomial regression model for such data, see also Tian et al. (2015) and Dupuy (2017). In Diallo et al. (2018), authors propose a zero-inflated regression model for multinomial counts with joint zero-inflation. To answer for the question of evaluating the covariate effects on the marginalized mean of outcome variable, several marginalized regression models are mentioned, such as Long et al. (2014), Long et al. (2015), Todem et al. (2016), see also Martin and Hall (2017), Benecha (2018), Inan et al. (2018). But in the real life, data may occurs censored and those models are no longer adequate. A popular situation is right-censored, this means that only a lower bound of the count of interest is observed, or on the other words, when we know that the true count value is higher than the observed one. Ignoring censoring can yield biased estimates and incorrect statistical inference. Terza (1985) considers estimation in Poisson regression when the response is right-censored at a constant threshold. Famoye and Wang (2004) and Karlis et al. (2016) investigate randomly right-censoring in generalized Poisson regression model and in mixtures of Poisson regressions, respectively. Estimation in zero-inflated Poisson regression model with censoring is handled by Saffari and Adnan (2011), Nguyen and Dupuy (2019b). In our knowledge, there has been no paper researched in theory about the covariate effects on overall population mean when data is right-censored and that is reason why our work is conducted to fulfill this gap.
In the marginalized zero-inflated Poisson (MZIP) regression model with overall exposure effects proposed by Long et al. (2014), authors only established the model without proof of properties of MLEs, a simulation study is illustrated with the two predictors for zero-inflated part and themselves for overall exposure mean. In Martin and Hall (2017), authors showed the existence of MLEs by using EM algorithm through three models, say, MZIP, MZIB and MZINB. The work of Todem et al. (2016) presented the method to obtain the MLEs by describing the model

in two ways: derived marginal models and direct marginal models. In addition, a comparison between them is also conducted.
The remainder of this paper is organized as follows. In section 5.2, we give a brief definition of the likelihood estimation in MZIP model and we describe maximum likelihood estimation when the count response is randomly right-censored. In section 5.3, we present results of comprehensive simulation study to assess the finite-sample behaviour of the MLE, in our study, a large number of predictors are carried out and the finite-sample distribution of MLE is also investigated. Section 5.4 reports an application in health care utilization. A conclusion and perspective are provided in Section 5.5. All of the proofs are postponed to an Appendix.

5.2 The censored MZIP regression model

In this section, we consider the Marginalized ZIP model and we describe maximum likelihood estimation when the count response is randomly right-censored.

5.2.1 Loglikelihood estimation in MZIP regression model

Let $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})^\top$ and $\mathbf{W}_i = (1, W_{i2}, \ldots, W_{iq})^\top$ be respectively p and q-dimensional vectors of covariates affecting to overall mean and to zero-inflated part with $i = 1, \ldots, n$, $\top$ denotes the transpose operator. Set $\beta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}^q$ are unknown regression parameters. The MZIP model assumes that the count response variable Z_i which directly models the marginalized mean of the mixture distribution (see Long et al., 2014; Martin and Hall, 2017) by describing as follows:

$$\text{logit}(\omega_i) \quad = \quad \gamma^\top \mathbf{W}_i \tag{5.2.1}$$

and

$$\log(\nu_i) \quad = \quad \beta^\top \mathbf{X}_i \tag{5.2.2}$$

where $\nu_i = \mathbb{E}(Z_i|\mathbf{X}_i) > 0$ is the marginalized mean, and $0 \leqslant \omega_i \leqslant 1$ is certain probability, $i = 1, \ldots, n$. The Poisson process mean λ_i is redefined as the jointly function of two parameters (γ, β) by letting $\nu_i = (1 - \omega_i)\lambda_i$, thus

$$\lambda_i = (1 + e^{\gamma^\top \mathbf{W}_i})\, e^{\beta^\top \mathbf{X}_i} \tag{5.2.3}$$

A MZIP model specifies the distribution of Z_i as the mixture

$$Z_i \sim \begin{cases} 0 & \text{with probability } \omega_i, \\ \mathcal{P}(\lambda_i) & \text{with probability } 1 - \omega_i, \end{cases} \tag{5.2.4}$$

Considering n independent vectors $(Z_1, \mathbf{X}_1, \mathbf{W}_1), \ldots, (Z_n, \mathbf{X}_n, \mathbf{W}_n)$ from the model (5.2.1)-(5.2.2)-(5.2.3)-(5.2.4), all of them are defined on a probability space $(\Omega, \mathcal{C}, \mathbb{P})$. Based on these observations, the log-likelihood of MZIP can be expressed as:

$$\sum_{i=1}^{n} \left\{ J_i \log\left(e^{\gamma^\top \mathbf{W}_i} + e^{-e^{\beta^\top \mathbf{X}_i}\left(1+e^{\gamma^\top \mathbf{W}_i}\right)} \right) + (Z_i - 1 - J_i Z_i) \log\left(1 + e^{\gamma^\top \mathbf{W}_i}\right) \right.$$
$$\left. + (1 - J_i)\left[Z_i \beta_i^\top \mathbf{X_i} - e^{\beta^\top \mathbf{X}_i}\left(1 + e^{\gamma^\top \mathbf{W}_i}\right) - \log(Z_i!) \right] \right\},$$

where $J_i = 1_{\{Z_i=0\}}$. Now, because of $J_i Z_i = 0$ for all $i = 1, \ldots, n$, the above function is written as follows:

$$\sum_{i=1}^{n} \left\{ J_i \log\left(e^{\gamma^\top \mathbf{W}_i} + e^{-e^{\beta^\top \mathbf{X}_i}\left(1+e^{\gamma^\top \mathbf{W}_i}\right)} \right) + (Z_i - 1) \log\left(1 + e^{\gamma^\top \mathbf{W}_i}\right) \right.$$
$$\left. + (1 - J_i) \left[Z_i \beta_i^\top \mathbf{X_i} - e^{\beta^\top \mathbf{X}_i} \left(1 + e^{\gamma^\top \mathbf{W}_i}\right) - \log(Z_i!) \right] \right\}. \qquad (5.2.5)$$

The maximum likelihood estimator of (β, γ) can be obtained by maximizing this function.

5.2.2 Loglikelihood estimation in MZIP regression model with right-censored

Assume Z_i can be randomly right-censored at the lower bound by a censoring random variable C_i. This means, true value for a certain individual is actual greater than C_i and we do not know that unobserved value. Letting $Z_i^* = \min(Z_i, C_i)$, $\delta_i = 1_{\{Z_i < C_i\}}$ and $J_i = 1_{\{Z_i^*=0\}}$. We define the observation for the i-individual as the vector $(Z_i^*, \delta_i, \mathbf{X}_i, \mathbf{W}_i)$ (if $Z_i = C_i$, we refer to $Z_i^* = C_i$), the censoring value thus can be specified to each observation. The likelihood of $\psi := (\beta^\top, \gamma^\top)^\top$ based on observations $(Z_i^*, \delta_i, \mathbf{X}_i, \mathbf{W}_i)$, $i = 1, \ldots, n$ is now calculated as:

$$\begin{aligned}
L_n(\psi) &= \prod_{i=1}^{n} \mathbb{P}(Z_i = Z_i^* | \mathbf{X}_i, \mathbf{W}_i)^{\delta_i} \mathbb{P}(Z_i \geqslant Z_i^* | \mathbf{X}_i, \mathbf{W}_i)^{1-\delta_i}, \\
&= \prod_{i=1}^{n} \left(\mathbb{P}(Z_i = Z_i^* | \mathbf{X}_i, \mathbf{W}_i)^{1-J_i} \mathbb{P}(Z_i = 0 | \mathbf{X}_i, \mathbf{W}_i)^{J_i} \right)^{\delta_i} \mathbb{P}(Z_i \geqslant Z_i^* | \mathbf{X}_i, \mathbf{W}_i)^{(1-\delta_i)(1-J_i)} \\
&\qquad \times \left((1 - \omega_i) \left(1 - \sum_{k=0}^{Z_i^*-1} \frac{e^{-\lambda_i} \lambda_i^k}{k!} \right) \right)^{(1-\delta_i)(1-J_i)}, \\
&= \prod_{i=1}^{n} \left(\frac{e^{\gamma^\top \mathbf{W}_i}}{1 + e^{\gamma^\top \mathbf{W}_i}} + \frac{1}{1 + e^{\gamma^\top \mathbf{W}_i}} e^{-e^{\beta^\top \mathbf{X}_i}\left(1+e^{\gamma^\top \mathbf{W}_i}\right)} \right)^{\delta_i J_i} \\
&\qquad \times \left(\frac{1}{1 + e^{\gamma^\top \mathbf{W}_i}} \frac{e^{-e^{\beta^\top \mathbf{X}_i}\left(1+e^{\gamma^\top \mathbf{W}_i}\right)} e^{Z_i^* \beta^\top \mathbf{X}_i} \left(1 + e^{\gamma^\top \mathbf{W}_i}\right)^{Z_i^*}}{Z_i^*!} \right)^{\delta_i (1-J_i)} \\
&\qquad \times \left(\frac{1}{1 + e^{\gamma^\top \mathbf{W}_i}} \left(1 - \sum_{k=0}^{Z_i^*-1} \frac{e^{-\lambda_i} \lambda_i^k}{k!} \right) \right)^{(1-\delta_i)(1-J_i)},
\end{aligned}$$

from which, we readily obtain the loglikelihood $\ell_n(\psi) = \log L_n(\psi)$. If ω_i and λ_i are given by (5.2.1) and (5.2.3) and note that $J_i = 1_{\{Z_i^*=0\}}$ so $J_i Z_i^* = 0$, $\forall i = 1, \ldots, n$, straightforward algebra yields:

$$\begin{aligned}
\ell_n(\psi) &= \sum_{i=1}^{n} \left\{ \delta_i J_i \log\left(e^{\gamma^\top \mathbf{W}_i} + e^{-e^{\beta^\top \mathbf{X}_i}\left(1+e^{\gamma^\top \mathbf{W}_i}\right)} \right) + (\delta_i Z_i^* - 1) \log\left(1 + e^{\gamma^\top \mathbf{W}_i}\right) \right. \\
&\quad + \delta_i (1 - J_i) \left(Z_i^* \beta^\top \mathbf{X}_i - e^{\beta^\top \mathbf{X}_i} \left(1 + e^{\gamma^\top \mathbf{W}_i}\right) - \log(Z_i^*!) \right) \\
&\quad \left. + (1 - \delta_i)(1 - J_i) \log\left(1 - \sum_{k=0}^{Z_i^*-1} \frac{e^{-e^{\beta^\top \mathbf{X}_i}\left(1+e^{\gamma^\top \mathbf{W}_i}\right)} e^{k \beta^\top \mathbf{X}_i} \left(1 + e^{\gamma^\top \mathbf{W}_i}\right)^k}{k!} \right) \right\}.
\end{aligned}$$

Clearly that $\ell_n(\psi)$ reduces to (5.2.5) when there is no censoring.

The maximum likelihood estimator $\hat{\psi}_n := (\hat{\beta}_n^\top, \hat{\gamma}_n^\top)^\top$ of ψ can be achieved by solving k-dimensional score equation

$$\frac{\partial \ell_n(\psi)}{\partial \psi} = 0, \tag{5.2.6}$$

where $k = p + q$.

Solution of this non-linear equation is relatively straightforward by using a standard mathematical software. In our simulation study and real-data analysis, we use `R` package `optimx` of Nash and Varadhan (2011); Nash (2014), which provides efficient computational tools for solving equation such as (5.2.6).

Next, we need to introduce some further notations and a few modelling assumptions before stating asymptotic properties of $\hat{\psi}_n$.

5.2.3 Some further notations

In what follows, we denote $k_i(\gamma) = e^{\gamma^\top \mathbf{W}_i}$, $\nu_i(\beta) = e^{\beta^\top \mathbf{X}_i}$ $\lambda_i(\psi) = e^{\beta^\top \mathbf{X}_i}\left(1 + e^{\gamma^\top \mathbf{W}_i}\right)$, $L_i(\psi) = e^{-e^{\beta^\top \mathbf{X}_i}\left(1+e^{\gamma^\top \mathbf{W}_i}\right)}$. Let also $S_{\lambda_i(\beta)} = \mathbb{P}(\mathcal{P}(\lambda_i(\beta)) \geqslant u)$, $u = 0, 1, \ldots$ denote the survival function of $\mathcal{P}(\lambda_i(\beta))$ distribution. First, we have:

$$\begin{aligned}
\frac{\partial \ell_n(\psi)}{\partial \beta_j} &= \sum_{i=1}^{n} X_{ij} \left\{ -\delta_i J_i \frac{L_i(\psi)\lambda_i(\psi)}{k_i(\gamma) + L_i(\psi)} + \delta_i (1 - J_i) \left(Z_i^* - \lambda_i(\psi)\right) \right. \\
&\quad \left. -(1-\delta_i)(1-J_i) \sum_{k=0}^{Z_i^*-1} \frac{L_i(\psi)\lambda_i(\psi)^k \left(k - \lambda_i(\psi)\right)}{k! \mathcal{S}_{\lambda_i(\psi)}\left(Z_i^*\right)} \right\}, \; j = 1, \ldots, p, \qquad (5.2.7) \\
\frac{\partial \ell_n(\psi)}{\partial \gamma_\ell} &= \sum_{i=1}^{n} W_{i\ell} \left\{ \delta_i J_i \frac{k_i(\gamma) - L_i(\psi)k_i(\gamma)\nu_i(\beta)}{k_i(\gamma) + L_i(\psi)} + \left(\delta_i Z_i^* - 1\right) \frac{k_i(\gamma)}{1 + k_i(\gamma)} - \delta_i (1 - J_i)\nu_i(\beta)k_i(\gamma) \right. \\
&\quad \left. -(1-\delta_i)(1-J_i) \sum_{k=0}^{Z_i^*-1} \frac{L_i(\psi)\nu_i(\beta)\lambda_i(\psi)^{k-1}k_i(\gamma)\left(k - \lambda_i(\psi)\right)}{k! \mathcal{S}_{\lambda_i(\psi)}\left(Z_i^*\right)} \right\}, \ell = 1, \ldots, q. \qquad (5.2.8)
\end{aligned}$$

For all $i = 1, \ldots, n$ and for $Z_i^* \geqslant 1$, let

$$\begin{aligned}
u_i(\psi) &= \frac{L_i(\psi)\lambda_i(\psi)\left(L_i(\psi) + k_i(\gamma) - \lambda_i(\psi)k_i(\gamma)\right)}{\left(k_i(\gamma) + L_i(\psi)\right)^2}, \\
v_i(\psi) &= \sum_{k=0}^{Z_i^*-1} \frac{L_i(\psi)\lambda_i(\psi)^k}{k! \mathcal{S}_{\lambda_i(\psi)}\left(Z_i^*\right)^2} \\
&\quad \times \left[\left(\left(k - \lambda_i(\psi)\right)^2 - \lambda_i(\psi) \right) \mathcal{S}_{\lambda_i(\psi)}\left(Z_i^*\right) - \left(k - \lambda_i(\psi)\right)\lambda_i(\psi)\mathbb{P}\left(\mathcal{P}\left(\lambda_i(\psi)\right) = Z_i^* - 1\right) \right] \\
s_i(\psi) &= \frac{L_i(\psi)k_i(\gamma)\left(\nu_i(\beta)k_i(\gamma) + \nu_i(\beta)L_i(\psi) - \nu_i(\beta)k_i(\gamma)\lambda_i(\psi) - \lambda_i(\psi)\right)}{\left(k_i(\gamma) + L_i(\psi)\right)^2}
\end{aligned}$$

$$
\begin{aligned}
t_i(\psi) &= \sum_{k=0}^{Z_i^*-1} \frac{L_i(\psi)\nu_i(\beta)k_i(\gamma)\lambda_i(\psi)^{k-2}}{k!\mathcal{S}_{\lambda_i(\psi)}\left(Z_i^*\right)^2} \\
&\quad \times\Big[\Big(\left(k-\lambda_i(\psi)\right)^2-\lambda_i(\psi)\Big)\,\mathcal{S}_{\lambda_i(\psi)}\left(Z_i^*\right)-\lambda_i(\psi)(k-\lambda_i(\psi))\mathbb{P}\left(\mathcal{P}\left(\lambda_i(\psi)\right)=Z_i^*-1\right)\Big] \\
f_i(\psi) &= \frac{k_i(\gamma)\Big[\left(1-L_i(\psi)\nu_i(\beta)-L_i(\psi)k_i(\gamma)\nu_i(\beta)^2\right)\left(k_i(\gamma)+L_i(\psi)\right)-k_i(\gamma)\left(1-L_i(\psi)\nu_i(\beta)\right)^2\Big]}{\left(k_i(\gamma)+L_i(\psi)\right)^2} \\
g_i(\psi) &= \sum_{k=0}^{Z_i^*-1} \frac{L_i(\psi)\nu_i(\beta)k_i(\gamma)\lambda_i(\psi)^{k-2}}{k!\mathcal{S}_{\lambda_i(\psi)}\left(Z_i^*\right)^2} \\
&\quad \times\Big[\left(k-\lambda_i(\psi)\right)\left(\lambda_i(\psi)+(k-1)\,\nu_i(\beta)k_i(\gamma)-\lambda_i(\psi)k_i(\gamma)\nu_i(\beta)\right)-\lambda_i(\psi)k_i(\gamma)\nu_i(\beta) \\
&\qquad -\lambda_i(\psi)k_i(\gamma)\nu_i(\beta)\left(k-\lambda_i(\psi)\right)\mathbb{P}\left(\mathcal{P}\left(\lambda_i(\psi)\right)=Z_i^*-1\right)\Big]
\end{aligned}
$$

Then, some tedious albeit not difficult algebra shows that $\forall j,m=1,\ldots,p$ and $\forall \ell,s=1,\ldots,q$, we have:

$$
\begin{aligned}
\frac{\partial^2\ell_n(\psi)}{\partial\beta_j\partial\beta_m} &= \sum_{i=1}^n X_{ij}X_{im}\left\{-\delta_i J_i u_i(\psi)-\delta_i(1-J_i)\lambda_i(\psi)-(1-\delta_i)(1-J_i)v_i(\psi)\right\}, \\
\frac{\partial^2\ell_n(\psi)}{\partial\beta_j\partial\gamma_\ell} &= \sum_{i=1}^n X_{ij}W_{i\ell}\left\{-\delta_i J_i s_i(\psi)-\delta_i(1-J_i)\nu_i(\beta)k_i(\gamma)-(1-\delta)(1-J_i)t_i(\psi)\right\}, \\
\frac{\partial^2\ell_n(\psi)}{\partial\gamma_\ell\partial\gamma_s} &= \sum_{i=1}^n W_{i\ell}W_{is}\left\{\delta_i J_i g_i(\psi)+(\delta_i Z_i^*-1)\,\frac{k_i(\gamma)}{\left(1+k_i(\gamma)\right)^2}-\delta_i(1-J)\nu_i(\beta)k_i(\gamma)\right. \\
&\qquad \left.-(1-\delta_i)(1-J_i)f_i(\psi)\right\}.
\end{aligned}
$$

5.2.4 Regularity conditions and asymptotic results

In this section, we present consistency and asymptotic normality of $\hat{\psi}_n$. In what follows, the space $\mathbb{R}^k$ of k-dimensional vectors is provided with the Euclidean norm $\|\cdot\|_2$ and the space of $(k\times k)$ real matrices is provided with the norm $|||A|||_2 := \sup_{\|x\|_2=1}\|Ax\|_2$ (for the sake of notational simplicity, we use $\|\cdot\|$ for both norms). Recall that for a symmetric real $(k\times k)$-matrix A with eigenvalues $\lambda_1,\ldots,\lambda_k$, $\|A\|=\max_i|\lambda_i|$ (here and in the sequel, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ will denote the smallest and largest eigenvalues of symmetric matrix A respectively).

We first state some regularity conditions:

C1 Covariates are bounded, that is, there exist compact sets $\mathcal{X}\subset\mathbb{R}^p$ and $\mathcal{W}\subset\mathbb{R}^q$ such that $\mathbf{X}_i\in\mathcal{X}$ and $\mathbf{W}_i\in\mathcal{W}$ for every $i=1,2,\ldots$

C2 The true parameter value $\psi_0=(\beta_0^\top,\gamma_0^\top)^\top$ lies in the interior of some known compact and convex set $\mathcal{C}=\mathcal{B}\times\mathcal{G}\subset\mathbb{R}^k$ (where $\mathcal{B}\subset\mathbb{R}^p$ and $\mathcal{G}\subset\mathbb{R}^q$ are the parameter spaces of β and γ respectively).

C3 There exists a positive constant κ such that $n/\lambda_{\min}(F_n(\psi_0))\leqslant\kappa$ for every $n=1,2,\ldots$

C4 Censoring random variables $C_i, i = 1, 2, \ldots$ are strictly positive and bounded by some constant $M < \infty$ (for example, M can be the end of the study period, at which every individual still under study is censored).

Conditions C1-C3 are classical in generalized linear regression and zero-inflated regression models (see Fahrmeir and Kaufmann, 1985; Czado et al., 2007; Nguyen and Dupuy, 2019b). Condition C4 is required in the censored setting.

Next, we state asymptotic properties of the estimator $\hat{\psi}_n$. Their proofs are outlined in Appendix section.

For each $n = 1, 2, \ldots$ and $\varepsilon > 0$, define the neighbourhood $N_n(\varepsilon) = \{\psi \in \mathcal{C} : (\psi - \psi_0)^\top F_n(\psi - \psi_0) \leqslant \varepsilon^2\}$ of ψ_0, where F_n is a short notation for $F_n(\psi_0)$. Our first result states that the solution of (5.2.6) exists, lies in the neighbourhood $N_n(\varepsilon)$ of ψ_0 when n is sufficiently large and is consistent for ψ_0.

Theorem 5.2.1 (**Existence and consistency**). *Assume conditions C1-C4 hold. Then the probability that $\hat{\psi}_n$ exists and lies in $N_n(\varepsilon)$ for some ε tends to 1 as $n \to \infty$. Furthermore, $\hat{\psi}_n$ converges in probability to ψ_0 as $n \to \infty$.*

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

Theorem 5.2.2 (**Asymptotic normality**). *Assume the conditions C1-C4 hold. Then $F_n^{\frac{1}{2}}(\hat{\psi}_n - \psi_0)$ converges in distribution to the Gauss vector $\mathcal{N}(0, I_k)$, as $n \to \infty$.*

where I_k is identity matrix of order k.

5.3 Simulation study

In this section, we investigate the finite-samples properties for the MLE under various scenarios which obtained by varying the censoring and zero-inflation proportions and the sample size.

5.3.1 Simulation design

First, we simulate the data according to the MZIP model (5.2.1)-(5.2.2)-(5.2.3)-(5.2.4) as follows:

$$\begin{aligned} \log(\nu_i(\beta)) &= \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6}, \\ \text{logit}(\omega_i(\gamma)) &= \gamma_1 W_{i1} + \gamma_2 W_{i2} + \gamma_3 W_{i3} + \gamma_4 W_{i4}, + \gamma_5 W_{i5}, \end{aligned}$$

and

$$\lambda_i = (1 + e^{\gamma^\top \mathbf{W}_i}) e^{\beta^\top \mathbf{X}_i},$$

where $X_{i1} = W_{i1} = 1$ and the $X_{i2}, \ldots, X_{i6}, W_{i4}, W_{i5}$ are independently drawn from normal $\mathcal{N}(0, 1)$, Bernoulli $\mathcal{B}(0.3)$, normal $\mathcal{N}(1, 2.25)$, exponential $\mathcal{E}(1)$, uniform $\mathcal{U}(2, 5)$, normal $\mathcal{N}(-1, 1)$ and Bernoulli $\mathcal{B}(0.5)$ distributions respectively. Linear predictors in $\log(\nu_i(\beta))$ and $\text{logit}(\omega_i(\gamma))$ are allowed to share common terms by letting $W_{i2} = X_{i2}$ and $W_{i3} = X_{i3}$. We consider the following sample sizes: $n = 500, 1000, 2000$. The regression parameter β is chosen as $\beta = (0.7, 0.1, 0.4, 0.85, -0.5, 0)^\top$. The regression parameter γ is chosen as:

- case 1: $\gamma = (-0.9, -0.65, -0.2, 0.65, 0)^\top$,
- case 2: $\gamma = (0.25, -0.7, -0.2, 0.65, 0)^\top$.

With these chosen values, in case 1 (respectively case 2), the average percentage of zero-inflation in the simulated data sets is 20% (respectively 40%). Censoring values are simulated from a zero-truncated Poisson model with parameter μ, where μ is chosen to yield various average censoring proportions c in the simulated samples, namely $c = 10\%, 20\%, 40\%$. For purpose of comparison, we also provide results that would be obtained if there were no censoring (that is, when $c = 0\%$) since these results will constitute a benchmark for assessing performance of the MLE when censoring is present.

For each combination of the simulation design parameters (sample size, proportions of censoring and zero-inflation), we simulate $N = 1000$ samples and we calculate the MLE $\hat{\psi}_n$. Simulations are carried out using the statistical software `R`. In order to solve the likelihood equation, we use the package `optimx` (Nash, 2014) which implements various Newton-Raphson algorithms. We obtain starting values by estimating a MZIP model without taking censoring into account.

5.3.2 Results

For each configuration [`sample size` × `censoring proportion` × `zero-inflation proportion`] of the simulation parameters, we calculate the average bias and average relative bias (expressed as a percentage) of the estimates $\hat{\beta}_{j,n}$ and $\hat{\gamma}_{k,n}$ over the N simulated samples. For example, the relative bias of $\hat{\beta}_{j,n}$ is obtained as

$$\frac{1}{N}\sum_{t=1}^{N}\frac{\hat{\beta}_{j,n}^{(t)}-\beta_j}{\beta_j}\times 100,$$

where $\hat{\beta}_{j,n}^{(t)}$ denotes the MLE of β_j in the t-th simulated sample. We also obtain the average standard error (SE), empirical standard deviation (SD) and root mean square error (RMSE) for each $\hat{\beta}_{j,n}$ $(j = 1, \ldots, 6)$ and $\hat{\gamma}_{k,n}$ $(k = 1, \ldots, 5)$. Finally, we provide the empirical coverage probability (CP) and average length of 95%-level confidence intervals for the β_j and γ_k. Results are given in Table 5.1 (case 1, $n = 500$), Table 5.4 (case 2, $n = 500$), Table 5.2 (case 1, $n = 1000$), Table 5.5 (case 2, $n = 1000$), Table 5.3 (case 1, $n = 2000$) and Table 5.6 (case 2, $n = 2000$).

From these results, we observe, as expected, that accuracy of MLEs of both β_j and γ_k decreases as sample size decreases. Accuracy of β_js estimates also decreases as censoring increases (note that the relative bias stays moderate though, even when censoring is high). On the contrary, estimates of the γ_k are rather insensitive to censoring, which can be explained by the fact that censoring does not affect zero counts. For both β_j and γ_k, empirical coverage probabilities are close to the nominal confidence level in every case. As may also be expected, for a given censoring proportion, we observe that MLEs of the β_j (respectively γ_k) perform better when the zero-inflation proportion decreases (respectively increases).

Finally, in order to assess quality of the Gaussian approximation stated in Theorem 5.2.2, we obtain normal Q-Q plots of the estimates and histograms of the normalized estimates $(\widehat{\beta}_{j,n} - \beta_j)/\text{standard error}(\widehat{\beta}_{j,n})$, $j = 1, \ldots, 6$ and $(\widehat{\gamma}_{k,n} - \gamma_k)/\text{standard error}(\widehat{\gamma}_{k,n})$, $j = 1, \ldots, 5$. We provide these graphs for $n = 500$ with a proportion of zero-inflation equal to 40% and 40% of censoring (see figures 5.1 to 5.4). Plots for the other (and more favorable) simulated scenarios yield similar observations and are thus not given. From these figures, it appears that the Gaussian approximation of the distribution of the MLE is reasonably satisfied, even when the sample size is moderate and the proportions of zero-inflation and censoring are as high as 40%.

average proportion of censoring		$\widehat{\beta}_n$						$\widehat{\gamma}_n$				
		$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$	$\widehat{\gamma}_{5,n}$
0	bias	0.0005	0.0008	-0.0026	0.0002	-0.0003	-0.0005	-0.0115	-0.0051	0.0081	0.0051	-0.0023
	rel. bias	0.0647	0.7783	-0.6404	0.0187	0.0592	-	1.2748	0.7839	-4.0539	0.7835	-
	SD	0.0868	0.0252	0.0497	0.0119	0.0243	0.0196	0.1546	0.1011	0.2126	0.0845	0.1458
	SE	0.0841	0.0250	0.0497	0.0124	0.0251	0.0194	0.1542	0.0987	0.2131	0.0832	0.1379
	RMSE	0.1208	0.0355	0.0704	0.0172	0.0349	0.0276	0.2186	0.1414	0.3011	0.1187	0.2006
	CP	0.9480	0.9470	0.9510	0.9560	0.9520	0.9430	0.9510	0.9350	0.9480	0.9520	0.9330
	ℓ	0.3291	0.0975	0.1944	0.0482	0.0980	0.0758	0.6025	0.3837	0.8314	0.3227	0.5381
0.1	bias	0.0003	0.0004	-0.0038	0.0022	-0.0023	-0.0005	-0.0159	-0.0084	0.0128	0.0111	-0.0023
	rel. bias	0.0484	0.3679	-0.9508	0.2579	0.4636	-	1.7710	1.2887	-6.4220	1.7046	-
	SD	0.1134	0.0310	0.0629	0.0244	0.0309	0.0275	0.1795	0.1213	0.2489	0.1038	0.1835
	SE	0.1118	0.0311	0.0637	0.0250	0.0320	0.0270	0.1793	0.1192	0.2542	0.1035	0.1806
	RMSE	0.1592	0.0439	0.0896	0.0350	0.0445	0.0385	0.2542	0.1703	0.3559	0.1470	0.2574
	CP	0.9370	0.9440	0.9510	0.9560	0.9540	0.9490	0.9530	0.9420	0.9520	0.9500	0.9490
	ℓ	0.4378	0.1217	0.2492	0.0980	0.1251	0.1057	0.7007	0.4652	0.9938	0.4029	0.7059
0.2	bias	0.0023	0.0006	-0.0021	0.0029	-0.0032	-0.0008	-0.0213	-0.0127	0.0117	0.0125	0.0027
	rel. bias	0.3278	0.6132	-0.5131	0.3356	0.6499	-	2.3717	1.9477	-5.8457	1.9181	-
	SD	0.1360	0.0360	0.0734	0.0314	0.0361	0.0339	0.1947	0.1318	0.2663	0.1201	0.2123
	SE	0.1324	0.0357	0.0738	0.0327	0.0376	0.0330	0.1953	0.1299	0.2733	0.1162	0.2068
	RMSE	0.1898	0.0507	0.1040	0.0454	0.0522	0.0473	0.2766	0.1854	0.3816	0.1676	0.2963
	CP	0.9420	0.9480	0.9620	0.9570	0.9570	0.9510	0.9570	0.9420	0.9620	0.9490	0.9440
	ℓ	0.5183	0.1397	0.2887	0.1281	0.1469	0.1292	0.7629	0.5071	1.0682	0.4527	0.8086
0.4	bias	0.0077	0.0005	-0.0020	0.0044	-0.0042	-0.0015	-0.0265	-0.0167	0.0086	0.0100	-0.0011
	rel. bias	1.1023	0.4880	-0.4904	0.5206	0.8476	-	2.9415	2.5730	-4.3083	1.5378	-
	SD	0.1921	0.0479	0.1022	0.0467	0.0510	0.0501	0.2227	0.1444	0.2928	0.1396	0.2577
	SE	0.1890	0.0478	0.1004	0.0499	0.0519	0.0492	0.2236	0.1448	0.2985	0.1379	0.2521
	RMSE	0.2696	0.0676	0.1433	0.0684	0.0728	0.0702	0.3166	0.2051	0.4181	0.1964	0.3604
	CP	0.9470	0.9440	0.9470	0.9610	0.9601	0.9420	0.9520	0.9530	0.9540	0.9460	0.9600
	ℓ	0.7400	0.1871	0.3932	0.1952	0.2028	0.1925	0.8732	0.5656	1.1668	0.5376	0.9860

Table 5.1: Simulation results (n = 500, ZI proportion = 20%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals.

average proportion of censoring		$\widehat{\beta}_n$						$\widehat{\gamma}_n$				
		$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$	$\widehat{\gamma}_{5,n}$
0	bias	0.0006	-0.0002	0.0007	0.0000	-0.0002	-0.0005	-0.0028	-0.0008	-0.0046	0.0008	0.0030
	rel. bias	0.0825	-0.1970	0.1676	-0.0010	0.0388	-	0.3144	0.1229	2.2814	0.1220	-
	SD	0.0583	0.0171	0.0340	0.0085	0.0167	0.0133	0.1023	0.0682	0.1434	0.0572	0.0930
	SE	0.0583	0.0173	0.0345	0.0085	0.0174	0.0134	0.1066	0.0673	0.1459	0.0562	0.0940
	RMSE	0.0824	0.0243	0.0484	0.0120	0.0241	0.0189	0.1477	0.0959	0.2046	0.0802	0.1323
	CP	0.9510	0.9520	0.9540	0.9550	0.9580	0.9540	0.9570	0.9370	0.9480	0.9460	0.9540
	ℓ	0.2283	0.0676	0.1350	0.0331	0.0682	0.0525	0.4170	0.2626	0.5706	0.2190	0.3677
0.1	bias	-0.0019	-0.0006	-0.0006	0.0010	-0.0011	0.0002	-0.0059	-0.0035	-0.0006	0.0042	0.0035
	rel. bias	-0.2663	-0.6178	-0.1596	0.1160	0.2101	-	0.6523	0.5453	0.2833	0.6535	-
	SD	0.0777	0.0212	0.0436	0.0175	0.0217	0.0190	0.1218	0.0833	0.1759	0.0707	0.1211
	SE	0.0781	0.0217	0.0446	0.0175	0.0223	0.0188	0.1246	0.0821	0.1766	0.0710	0.1247
	RMSE	0.1101	0.0304	0.0623	0.0248	0.0312	0.0267	0.1743	0.1170	0.2491	0.1003	0.1739
	CP	0.9460	0.9600	0.9590	0.9450	0.9610	0.9520	0.9590	0.9450	0.9450	0.9550	0.9550
	ℓ	0.3058	0.0851	0.1746	0.0687	0.0875	0.0738	0.4877	0.3212	0.6911	0.2774	0.4883
0.2	bias	0.0005	-0.0003	-0.0003	0.0009	-0.0014	-0.0003	-0.0052	-0.0057	-0.0020	0.0056	0.0025
	rel. bias	0.0656	-0.2962	-0.0785	0.1081	0.2766	-	0.5758	0.8731	1.0165	0.8632	-
	SD	0.0905	0.0246	0.0506	0.0227	0.0260	0.0230	0.1301	0.0905	0.1952	0.0785	0.1364
	SE	0.0926	0.0250	0.0516	0.0229	0.0263	0.0231	0.1355	0.0895	0.1900	0.0799	0.1429
	RMSE	0.1294	0.0350	0.0723	0.0323	0.0370	0.0326	0.1879	0.1274	0.2723	0.1121	0.1976
	CP	0.9580	0.9500	0.9550	0.9540	0.9550	0.9480	0.9580	0.9470	0.9460	0.9550	0.9570
	ℓ	0.3627	0.0977	0.2023	0.0897	0.1029	0.0904	0.5304	0.3502	0.7438	0.3122	0.5596
0.4	bias	-0.0046	-0.0003	0.0033	0.0010	-0.0007	0.0010	-0.0081	-0.0071	-0.0019	0.0057	0.0017
	rel. bias	-0.6520	-0.3274	0.8314	0.1163	0.1369	-	0.9043	1.0869	0.9672	0.8708	-
	SD	0.1320	0.0334	0.0671	0.0334	0.0346	0.0347	0.1506	0.1013	0.2069	0.0948	0.1723
	SE	0.1321	0.0334	0.0704	0.0348	0.0362	0.0343	0.1551	0.0999	0.2076	0.0953	0.1745
	RMSE	0.1868	0.0473	0.0973	0.0483	0.0501	0.0488	0.2163	0.1424	0.2930	0.1345	0.2452
	CP	0.9450	0.9420	0.9630	0.9570	0.9690	0.9460	0.9610	0.9470	0.9510	0.9500	0.9460
	ℓ	0.5175	0.1308	0.2758	0.1364	0.1418	0.1346	0.6070	0.3909	0.8126	0.3726	0.6835

Table 5.2: Simulation results (n = 1000, ZI proportion = 20%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals.

average proportion of censoring		$\widehat{\beta}_n$						$\widehat{\gamma}_n$				
		$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$	$\widehat{\gamma}_{5,n}$
0	bias	0.0011	-0.0002	0.0011	0.0001	-0.0004	-0.0004	-0.0027	-0.0012	-0.0072	0.0036	0.0024
	rel. bias	0.1602	-0.2263	0.2665	0.0091	0.0855	-	0.3047	0.1860	3.6012	0.5607	-
	SD	0.0417	0.0124	0.0238	0.0058	0.0118	0.0094	0.0749	0.0480	0.1050	0.0411	0.0650
	SE	0.0406	0.0120	0.0240	0.0058	0.0121	0.0093	0.0743	0.0462	0.1009	0.0386	0.0648
	RMSE	0.0582	0.0173	0.0338	0.0082	0.0169	0.0132	0.1055	0.0666	0.1458	0.0565	0.0918
	CP	0.9390	0.9430	0.9500	0.9570	0.9620	0.9460	0.9550	0.9400	0.9430	0.9360	0.9540
	ℓ	0.1590	0.0470	0.0940	0.0227	0.0476	0.0364	0.2909	0.1805	0.3951	0.1507	0.2539
0.1	bias	-0.0001	-0.0001	0.0009	-0.0007	-0.0012	0.0005	-0.0048	-0.0028	-0.0103	0.0034	0.0036
	rel. bias	-0.0160	-0.1024	0.2369	-0.0801	0.2466	-	0.5370	0.4348	5.1459	0.5279	-
	SD	0.0557	0.0153	0.0321	0.0126	0.0156	0.0133	0.0872	0.0588	0.1297	0.0521	0.0863
	SE	0.0548	0.0153	0.0313	0.0123	0.0157	0.0132	0.0873	0.0573	0.1238	0.0495	0.0871
	RMSE	0.0781	0.0216	0.0448	0.0176	0.0222	0.0188	0.1235	0.0821	0.1796	0.0719	0.1227
	CP	0.9530	0.9510	0.9400	0.9390	0.9540	0.9500	0.9450	0.9430	0.9390	0.9490	0.9470
	ℓ	0.2149	0.0598	0.1227	0.0482	0.0615	0.0518	0.3419	0.2243	0.4850	0.1937	0.3414
0.2	bias	0.0000	0.0000	0.0007	-0.0011	-0.0011	0.0007	-0.0057	-0.0028	-0.0096	0.0016	0.0014
	rel. bias	0.0020	-0.0455	0.1871	-0.1350	0.2161	-	0.6306	0.4283	4.7799	0.2500	-
	SD	0.0652	0.0170	0.0370	0.0157	0.0183	0.0163	0.0966	0.0629	0.1394	0.0585	0.1002
	SE	0.0651	0.0176	0.0364	0.0161	0.0185	0.0162	0.0950	0.0625	0.1333	0.0558	0.1001
	RMSE	0.0921	0.0244	0.0519	0.0225	0.0260	0.0230	0.1355	0.0886	0.1931	0.0808	0.1416
	CP	0.9530	0.9640	0.9440	0.9550	0.9530	0.9490	0.9500	0.9530	0.9410	0.9380	0.9510
	ℓ	0.2549	0.0689	0.1425	0.0631	0.0724	0.0635	0.3721	0.2446	0.5223	0.2182	0.3921
0.4	bias	-0.0022	0.0003	-0.0006	-0.0018	-0.0008	0.0016	-0.0064	-0.0031	-0.0087	0.0024	0.0010
	rel. bias	-0.3174	0.2770	-0.1411	-0.2074	0.1594	-	0.7128	0.4797	4.3513	0.3627	-
	SD	0.0936	0.0239	0.0482	0.0224	0.0256	0.0246	0.1094	0.0691	0.1506	0.0676	0.1247
	SE	0.0926	0.0235	0.0494	0.0245	0.0254	0.0241	0.1086	0.0698	0.1457	0.0665	0.1222
	RMSE	0.1316	0.0335	0.0690	0.0332	0.0361	0.0344	0.1543	0.0983	0.2097	0.0948	0.1746
	CP	0.9430	0.9460	0.9630	0.9720	0.9470	0.9460	0.9460	0.9590	0.9400	0.9490	0.9480
	ℓ	0.3630	0.0919	0.1937	0.0959	0.0995	0.0943	0.4254	0.2734	0.5707	0.2603	0.4789

Table 5.3: Simulation results (n = 2000, ZI proportion = 20%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals.

average proportion of censoring		$\widehat{\beta}_n$						$\widehat{\gamma}_n$				
		$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$	$\widehat{\gamma}_{5,n}$
0	bias	-0.0039	0.0011	-0.0046	0.0007	-0.0010	0.0001	-0.0017	-0.0038	0.0095	0.0030	0.0016
	rel. bias	-0.5560	1.1167	-1.1573	0.0875	0.2000	-	-0.6740	0.5360	-4.7423	0.4555	-
	SD	0.0939	0.0330	0.0643	0.0136	0.0271	0.0201	0.1063	0.0720	0.1546	0.0530	0.0772
	SE	0.0928	0.0328	0.0650	0.0132	0.0258	0.0200	0.1063	0.0718	0.1521	0.0502	0.0788
	RMSE	0.1320	0.0465	0.0915	0.0190	0.0374	0.0284	0.1503	0.1017	0.2170	0.0730	0.1103
	CP	0.9450	0.9410	0.9480	0.9420	0.9490	0.9470	0.9490	0.9520	0.9420	0.9440	0.9560
	ℓ	0.3627	0.1279	0.2538	0.0514	0.1005	0.0783	0.4162	0.2795	0.5943	0.1948	0.3079
0.1	bias	-0.0024	0.0016	-0.0058	-0.0005	-0.0018	0.0006	0.0021	-0.0072	0.0042	0.0044	-0.0024
	rel. bias	-0.3392	1.5631	-1.4438	-0.0572	0.3622	-	0.8593	1.0300	-2.1098	0.6694	-
	SD	0.1300	0.0416	0.0828	0.0288 0	.0352	0.0305	0.1332	0.0941	0.1923	0.0742	0.1188
	SE	0.1271	0.0410	0.0831	0.0291	0.0352	0.0300	0.1297	0.0931	0.1926	0.0715	0.1174
	RMSE	0.1818	0.0584	0.1174	0.0410	0.0498	0.0428	0.1858	0.1326	0.2722	0.1031	0.1670
	CP	0.9480	0.9570	0.9480	0.9530	0.9470	0.9510	0.9430	0.9500	0.9520	0.9440	0.9450
	ℓ	0.4975	0.1602	0.3250	0.1140	0.1374	0.1175	0.5077	0.3643	0.7545	0.2788	0.4596
0.2	bias	-0.0040	0.0020	-0.0095	-0.0025	0.0003	0.0016	0.0059	-0.0064	-0.0008	0.0029	-0.0096
	rel. bias	-0.5746	1.9845	-2.3643	-0.2982	-0.0557	-	2.3598	0.9093	0.3994	0.4409	-
	SD	0.1615	0.0484	0.0966	0.0401	0.0445	0.0397	0.1517	0.1045	0.2120	0.0887	0.1461
	SE	0.1585	0.0478	0.0975	0.0409	0.0440	0.0394	0.1473	0.1040	0.2113	0.0857	0.1448
	RMSE	0.2262	0.0680	0.1376	0.0573	0.0626	0.0559	0.2115	0.1475	0.2992	0.1233	0.2059
	CP	0.9490	0.9420	0.9530	0.9480	0.9420	0.9470	0.9360	0.9530	0.9490	0.9380	0.9430
	ℓ	0.6205	0.1869	0.3817	0.1600	0.1718	0.1540	0.5768	0.4069	0.8276	0.3344	0.5671
0.4	bias	-0.0132	0.0009	-0.0075	-0.0093	0.0052	0.0049	0.0090	-0.0088	-0.0065	0.0042	-0.0093
	rel. bias	-1.8839	0.8937	-1.8767	-1.0969	-1.0404	-	3.6120	1.2524	3.2268	0.6493	-
	SD	0.2664	0.0756	0.1559	0.0659	0.0703	0.0702	0.1849	0.1169	0.2309	0.1117	0.2062
	SE	0.2833	0.0740	0.1522	0.0755	0.0742	0.0743	0.1855	0.1191	0.2349	0.1126	0.2003
	RMSE	0.3890	0.1058	0.2180	0.1006	0.1023	0.1023	0.2620	0.1670	0.3294	0.1586	0.2876
	CP	0.9740	0.9400	0.9420	0.9660	0.9640	0.9700	0.9490	0.9580	0.9540	0.9520	0.9480
	ℓ	1.1072	0.2891	0.5949	0.2950	0.2894	0.2906	0.7263	0.4660	0.9202	0.4400	0.7845

Table 5.4: Simulation results (n = 500, ZI proportion = 40%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals.

average proportion of censoring		$\widehat{\beta}_n$						$\widehat{\gamma}_n$				
		$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$	$\widehat{\gamma}_{5,n}$
0	bias	-0.0023	0.0002	-0.0004	-0.0001	-0.0010	0.0003	0.0002	-0.0014	-0.0007	0.0007	0.0016
	rel. bias	-0.3344	0.1925	-0.0958	-0.0113	0.2064	-	0.0970	0.2044	0.3449	0.1106	-
	SD	0.0634	0.0208	0.0443	0.0086	0.0177	0.0132	0.0720	0.0461	0.1006	0.0343	0.0523
	SE	0.0636	0.0224	0.0445	0.0088	0.0176	0.0136	0.0735	0.0479	0.1034	0.0334	0.0531
	RMSE	0.0898	0.0306	0.0628	0.0123	0.0250	0.0190	0.1029	0.0665	0.1442	0.0479	0.0745
	CP	0.9570	0.9620	0.9510	0.9540	0.9450	0.9570	0.9460	0.9640	0.9530	0.9460	0.9560
	ℓ	0.2490	0.0877	0.1740	0.0343	0.0689	0.0534	0.2881	0.1869	0.4045	0.1302	0.2078
0.1	bias	0.0000	-0.0001	-0.0015	-0.0023	-0.0005	0.0004	0.0001	0.0006	-0.0021	-0.0005	0.0005
	rel. bias	-0.0020	-0.1409	-0.3789	-0.2662	0.1009	-	0.0317	-0.0820	1.0543	-0.0843	-
	SD	0.0853	0.0273	0.0569	0.0203	0.0248	0.0198	0.0896	0.0626	0.1345	0.0506	0.0804
	SE	0.0885	0.0285	0.0580	0.0203	0.0244	0.0208	0.0905	0.0644	0.1342	0.0491	0.0811
	RMSE	0.1229	0.0395	0.0813	0.0288	0.0347	0.0287	0.1273	0.0898	0.1900	0.0705	0.1142
	CP	0.9570	0.9600	0.9450	0.9480	0.9450	0.9620	0.9520	0.9580	0.9560	0.9420	0.9520
	ℓ	0.3466	0.1117	0.2271	0.0795	0.0953	0.0816	0.3544	0.2520	0.5258	0.1921	0.3179
0.2	bias	0.0007	0.0002	-0.0025	-0.0047	0.0004	0.0005	0.0001	0.0004	-0.0025	-0.0013	-0.0001
	rel. bias	0.0988	0.2384	-0.6136	-0.5510	-0.0884	-	0.0251	-0.0569	1.2592	-0.1993	-
	SD	0.1089	0.0324	0.0667	0.0276	0.0301	0.0265	0.1009	0.0729	0.1473	0.0597	0.0969
	SE	0.1106	0.0334	0.0683	0.0285	0.0305	0.0274	0.1030	0.0723	0.1474	0.0593	0.1005
	RMSE	0.1552	0.0466	0.0955	0.0400	0.0428	0.0381	0.1442	0.1027	0.2084	0.0841	0.1396
	CP	0.9470	0.9560	0.9600	0.9510	0.9560	0.9530	0.9600	0.9530	0.9470	0.9520	0.9600
	ℓ	0.4331	0.1308	0.2674	0.1118	0.1194	0.1073	0.4035	0.2832	0.5777	0.2319	0.3938
0.4	bias	-0.0011	0.0011	-0.0036	-0.0125	0.0031	0.0021	0.0003	0.0004	-0.0046	-0.0028	0.0001
	rel. bias	-0.1595	1.0633	-0.8947	-1.4695	-0.6155	-	0.1126	-0.0521	2.3191	-0.4366	-
	SD	0.1897	0.0485	0.1014	0.0445	0.0482	0.0498	0.1251	0.0814	0.1680	0.0792	0.1314
	SE	0.1961	0.0510	0.1055	0.0527	0.0512	0.0514	0.1293	0.0828	0.1639	0.0782	0.1393
	RMSE	0.2728	0.0704	0.1463	0.0700	0.0703	0.0716	0.1798	0.1161	0.2347	0.1113	0.1914
	CP	0.9560	0.9680	0.9630	0.9540	0.9660	0.9550	0.9590	0.9540	0.9460	0.9510	0.9590
	ℓ	0.7677	0.1996	0.4132	0.2061	0.2001	0.2012	0.5064	0.3243	0.6422	0.3059	0.5458

Table 5.5: Simulation results (n = 1000, ZI proportion = 40%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals.

average proportion of censoring		$\widehat{\beta}_n$						$\widehat{\gamma}_n$				
		$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\beta}_{6,n}$	$\widehat{\gamma}_{1,n}$	$\widehat{\gamma}_{2,n}$	$\widehat{\gamma}_{3,n}$	$\widehat{\gamma}_{4,n}$	$\widehat{\gamma}_{5,n}$
0	bias	-0.0013	0.0004	0.0000	-0.0004	0.0002	0.0002	0.0022	-0.0001	-0.0015	-0.0002	0.0002
	rel. bias	0.1892	0.3625	0.0096	-0.0449	-0.0318	-	0.8732	0.0196	0.7471	-0.0241	
	SD	0.0423	0.0158	0.0313	0.0059	0.0126	0.0094	0.0526	0.0330	0.0728	0.0236	0.0367
	SE	0.0443	0.0157	0.0310	0.0060	0.0123	0.0094	0.0514	0.0329	0.0715	0.0230	0.0365
	RMSE	0.0612	0.0223	0.0440	0.0084	0.0176	0.0133	0.0735	0.0466	0.1020	0.0329	0.0518
	CP	0.9640	0.9550	0.9480	0.9540	0.9470	0.9540	0.9460	0.9510	0.9560	0.9330	0.9540
	ℓ	0.1734	0.0615	0.1214	0.0235	0.0481	0.0370	0.2013	0.1285	0.2799	0.0897	0.1431
0.1	bias	-0.0016	0.0009	-0.0009	-0.0032	0.0013	0.0011	0.0011	0.0002	-0.0031	-0.0033	-0.0002
	rel. bias	-0.2331	0.9048	-0.2283	-0.3774	-0.2621	-	0.4289	-0.0327	1.5533	-0.5063	-
	SD	0.0578	0.0200	0.0408	0.0140	0.0176	0.0138	0.0644	0.0451	0.0956	0.0349	0.0590
	SE	0.0621	0.0201	0.0408	0.0143	0.0171	0.0146	0.0635	0.0450	0.0942	0.0341	0.0567
	RMSE	0.0849	0.0283	0.0577	0.0202	0.0245	0.0201	0.0905	0.0637	0.1342	0.0489	0.0819
	CP	0.9690	0.9550	0.9460	0.9430	0.9370	0.9640	0.9420	0.9470	0.9480	0.9510	0.9320
	ℓ	0.2434	0.0786	0.1600	0.0559	0.0669	0.0572	0.2489	0.1763	0.3694	0.1336	0.2224
0.2	bias	-0.0018	0.0012	-0.0028	-0.0061	0.0031	0.0015	-0.0002	0.0016	-0.0039	-0.0046	0.0016
	rel. bias	-0.2623	1.2099	-0.7059	-0.7203	-0.6149	-	-0.0975	-0.2294	1.9321	-0.7083	-
	SD	0.0753	0.0226	0.0490	0.0195	0.0211	0.0190	0.0720	0.0508	0.1062	0.0408	0.0716
	SE	0.0775	0.0235	0.0480	0.0200	0.0214	0.0192	0.0723	0.0506	0.1037	0.0413	0.0704
	RMSE	0.1080	0.0326	0.0687	0.0286	0.0302	0.0270	0.1020	0.0717	0.1485	0.0582	0.1004
	CP	0.9550	0.9610	0.9450	0.9450	0.9470	0.9580	0.9550	0.9440	0.9460	0.9500	0.9460
	ℓ	0.3038	0.0919	0.1881	0.0784	0.0837	0.0752	0.2833	0.1983	0.4066	0.1618	0.2759
0.4	bias	-0.0024	0.0035	-0.0037	-0.0115	0.0050	0.0019	-0.0028	0.0011	-0.0044	-0.0071	0.0017
	rel. bias	-0.3424	3.4750	-0.9343	-1.3472	-1.0046	-	-1.1180	-0.1627	2.1784	-1.0940	-
	SD	0.1329	0.0363	0.0772	0.0324	0.0340	0.0351	0.0910	0.0589	0.1165	0.0525	0.0998
	SE	0.1371	0.0357	0.0740	0.0369	0.0357	0.0359	0.0907	0.0581	0.1155	0.0547	0.0978
	RMSE	0.1909	0.0510	0.1070	0.0504	0.0496	0.0503	0.1285	0.0827	0.1640	0.0761	0.1398
	CP	0.9600	0.9500	0.9350	0.9610	0.9550	0.9550	0.9410	0.9440	0.9470	0.9530	0.9460
	ℓ	0.5370	0.1399	0.2900	0.1445	0.1399	0.1408	0.3555	0.2278	0.4526	0.2142	0.3835

Table 5.6: Simulation results (n = 2000, ZI proportion = 40%). SD: empirical standard deviation. SE: average standard error. RMSE: empirical root mean square error. CP: empirical coverage probability of 95%-level confidence intervals. ℓ : average length of the confidence intervals.

5.4 Real data application

In this section, we illustrate the censored MZIP model on a real data set from the German Socioeconomic Panel (a survey aimed at investigating healthcare utilization by German households). Here, we use the predictors and their interactions affect on zero-inflation and count components, that they are thoroughly considered in details in our previous work, see Nguyen and Dupuy (2019a).

The dataset considered here contains the number of doctor office visits (the response variable) for 1812 West German men aged 25-65 years, during the last three months of 1994. Several risk factors are available, including age, socio-economic variables: marital status (1 if married, 0 otherwise), educational level (number of years of schooling), household monthly net income (in German marks/1000) and composition (coded as 1 if children under 16 live in the household, 0 otherwise), two binary variables indicating whether individual is covered by a public health insurance and by a supplemental private insurance (both are coded as 1 if yes and 0 otherwise), employment characteristics (coded as `self`: 1 if self employed, 0 otherwise ; `civil`: 1 if civil servant ; `bluec`: 1 if blue collar employee ; `employed`: 1 if employed), various measures of health status: health satisfaction (`health`, coded as 0 if low to 10 if high), handicap status (`handicap`: 1 if handicapped, 0 otherwise) and degree of handicap in percentage points (`hdegree`). Following Jochmann (2009), who first described these data, we study a more complex effect of age by considering linear spline variables `age30`, `age35`,..., `age60` (where `ageXX` is 1 if age $\geqslant$ XX and 0 otherwise). Therefore, a total of 20 candidate predictors are available. Jochmann (2009) also suggests to consider interactions between health satisfaction and age variables (i.e., `age30`$\times$`health`, `age35`$\times$`health`,...). There is no reason, however, to limit ourselves to these interactions and one may wish to assess all possible second-order interactions (except for meaningless ones, such as interactions between `ageXX` variables).

In Figure 5.5, we plot the number of doctor office visits, censored at 15 visits for illustrative purpose. The plot strongly suggests that data are zero-inflated (41.2% of the observed counts are equal to 0). Thus, we fit the MZIP model with all risk factors and second-order interactions, which results in a very large number of possible predictors. Several authors recently addressed variable selection in high-dimensional uncensored ZIP and ZINB models via penalized maximum likelihood, and various penalty functions are implemented in the `R` package `mpath` (Wang, 2019).

In the result table, see Table 5.7, we consider the data in three scenarios (no censoring, censoring proportion is 10% and 20%) and we report estimates, their standard errors (in parenthesis) as well as corresponding p-value for each one. We observe that all of covariates affecting to the both marginalized sub-model and zero-inflation sub-model are high statistical significance (p-value < 0.05) except only a single coefficient `age40` (in bold letters) of non zero-inflation part. But we still retain because it combine with `civil` constituting interaction covariate `civil:age40`. Moreover, when the censored percentages increase, the standard errors also increase in both sub-models (but only slightly increasing in zero-inflation part, this is explained as section 5.3).

5.5 Conclusion

In this paper, we introduce a new regression model for evaluating the covariate effects on overall response when occurring right-censoring. Maximum likelihood estimation is conducted to investigate distribution properties under various of scenarios throughout our numerical study. Moreover, in analysis of health care utilization, the proposed model conveys a reasonable explanation and interpretation. In addition, the model also provides the plausible way to decide

	no censoring		censoring = 10%		censoring = 21%	
parameter	estimate (s.e.)	p-value.	estimate (s.e.)	p-value.	estimate (s.e.)	p-value.
			Marginalized model coefficients			
intercept	2.4281 (0.0539)	< 2e-16	1.8957 (0.0665)	< 2e-16	1.7043(0.0783)	<2e-16
health	-0.2373 (0.0074)	< 2e-16	-0.1825 (0.0091)	< 2e-16	-0.1746(0.0107)	<2e-16
age50	0.3113 (0.0487)	1.62e-10	0.3253 (0.0569)	1.06e-08	0.2612(0.0651)	6.08e-05
handicap	0.3593 (0.0729)	8.37e-07	0.3076 (0.0996)	0.00202	0.2888(0.1253)	0.02116
self	-0.2609 (0.0661)	8.00e-05	-0.2252 (0.0757)	0.00291	-0.2471(0.0881)	0.00505
civil	-0.6002 (0.1186)	4.16e-07	-0.4924 (0.1247)	7.86e-05	-0.4302(0.1318)	0.00110
hdegree	-0.0043 (0.0012)	0.000421	-0.0034 (0.0017)	0.04233	-0.0028(0.0021)	0.17682
age40	0.0141 (0.0423)	**0.739926**	-0.0280 (0.0502)	**0.57637**	-0.0180(0.0597)	**0.76293**
civil : age40	0.4125 (0.1342)	0.002115	0.3947 (0.1454)	0.00664	0.3613(0.1606)	0.02442
			Zero-inflation model coefficients			
intercept	-2.1040 (0.1882)	< 2e-16	-2.1466 (0.1962)	< 2e-16	-2.2992 (0.2147)	< 2e-16
health	0.2560 (0.0241)	< 2e-16	0.2603 (0.0250)	< 2e-16	0.2728 (0.0274)	< 2e-16
age50	-0.5252 (0.1139)	3.99e-06	-0.5831 (0.1218)	1.68e-06	-0.6075 (0.1350)	6.79e-06

Table 5.7: Health-care data analysis: estimates (standard errors) and p-value.

whether or not how a covariate affect on mean response variables.

In many contexts, marginalized zero-inflated model prove more convenient and useful. It provides not only a direct interpreted way of marginalized mean effects but also produce more accurate estimations, as shown in our simulation study, like coverage and error rates.

While in this paper we have concentrated on models for randomly right-censoring and independent data, some more extensions should be deserved attention, for instance, the randomly left-censored data or interval censored, longitudinal/clustered data. Estimation and inference in MZIP regression in these setting still may be an open question. Investigating estimation in more general marginalized zero-inflated models (e.g., marginalized zero-inflated generalized Poisson regression model, marginalized zero-inflated negative binomial regression model...) with randomly censoring are also desirable. All these issues constitute topic for our future work.

5.6 Appendix 1: Technical Lemmas

We note $S_n(\psi) = \partial \ell_n(\psi)/\partial\psi$, $H_n(\psi) = -\partial^2 \ell_n(\psi)/\partial\psi\partial\psi^\top$, $F_n(\psi) = \mathbb{E}(H_n(\psi))$ in which $H_n(\psi)$ is assumed positive definite. We first prove an important technical lemma.

Lemma 5.6.1. *Assume conditions $C1 - C4$ hold. Then $\sup_{\psi\in N_n(\varepsilon)} \|F_n^{-\frac{1}{2}} H_n(\psi) F_n^{-\frac{1}{2}} - I_k\|$ converges in probability to 0 as $n \to \infty$.*

Proof of the Lemma 5.6.1. We have almost surely

$$\begin{aligned}
\|F_n^{-\frac{1}{2}} H_n(\psi) F_n^{-\frac{1}{2}} - I_k\| &= \|F_n^{-\frac{1}{2}} (H_n(\psi) - F_n) F_n^{-\frac{1}{2}}\|, \\
&\leqslant \frac{1}{\lambda_{\min}(F_n)} \|H_n(\psi) - F_n\|, \\
&\leqslant \kappa \left\| \frac{1}{n}(H_n(\psi) - \mathbb{E}(H_n(\psi))) \right\| + \kappa \left\| \frac{1}{n}(\mathbb{E}(H_n(\psi)) - F_n) \right\|,
\end{aligned}$$

since $C3$. Therefore, to show the lemma, we need to prove convergence of both terms in the right hand side in probability to 0 by taking the supremum over $\psi \in N_n(\varepsilon)$ as $n \to \infty$. We thus prove for the first term, it is sufficient to show that all elements of matrix $\sup_{\psi\in N_n(\varepsilon)} \left\|\frac{1}{n}(H_n(\psi) - \mathbb{E}(H_n(\psi)))\right\|$ converge in probability to zero, i.e.,

$$\sup_{\psi\in N_n(\varepsilon)} \left| \frac{1}{n}(h_{n(\ell,m)}(\psi) - \mathbb{E}(h_{n,(\ell,m)}(\psi))) \right| \xrightarrow{\mathbb{P}} 0 \text{ as } n \to \infty,$$

where $h_{n(\ell,m)}(\psi)$ denote $(\ell, m)-$ element of matrix $H_n(\psi)$ for every $\ell, m = 1, \ldots, k$. Without loss of generality, we illustrate for $\ell, m = 1, \ldots, p$, so in this case, $h_{n(\ell,m)}(\psi)$ is exact of the element $-\partial^2 \ell_n(\psi)/\partial\beta_\ell \partial\beta_m$. We have

$$\begin{aligned}
&\left| \frac{1}{n}(h_{n,(\ell,m)}(\psi) - \mathbb{E}(h_{n,(\ell,m)}(\psi))) \right| \\
&\quad \leqslant \left| \frac{1}{n} \sum_{i=1}^{n} \{X_{i\ell} X_{im} \delta_i J_i u_i(\psi) - \mathbb{E}\left[X_{i\ell} X_{im} \delta_i J_i u_i(\psi)\right]\} \right| \\
&\quad + \left| \frac{1}{n} \sum_{i=1}^{n} \{X_{i\ell} X_{im} \delta_i (1 - J_i) \lambda_i(\beta) - \mathbb{E}\left[X_{i\ell} X_{im} \delta_i (1 - J_i) \lambda_i(\beta)\right]\} \right| \\
&\quad + \left| \frac{1}{n} \sum_{i=1}^{n} \{X_{i\ell} X_{im} (1 - \delta_i)(1 - J_i) v_i(\psi) - \mathbb{E}\left[X_{i\ell} X_{im} (1 - \delta_i)(1 - J_i) v_i(\psi)\right]\} \right|,
\end{aligned}$$

Since the similar structure of the three terms on the right hand side of inequality above, we only prove for the $\sup_{\psi\in N_n(\varepsilon)}\left|\frac{1}{n}\sum_{i=1}^{n}\{X_{i\ell}X_{im}\delta_i J_i u_i(\psi)-\mathbb{E}\left[X_{i\ell}X_{im}\delta_i J_i u_i(\psi)\right]\}\right|$ converges in probability to 0 as $n\to\infty$. We thus indicate consecutively the class $\{X_{i\ell}X_{im}\delta_i J_i u_i(\psi):\psi\in\mathcal{C}\}$ is Donsker, and then so is the $\{X_{i\ell}X_{im}\delta_i J_i u_i(\psi):\psi\in N_n(\varepsilon)\}$.
Indeed, the class $\{X_{i\ell}X_{im}\delta_i J_i\}$ is obviously Donsker because it includes only one class with only one function, the function here is the identity function therefore it evidently satisfies a condition of square integrable functions, so the assertion holds by the conclusion of van de Vaart and Wellner (see Aad W. van der Vaart, 1996, pg. 83). In addition, due to assumption that the boundary of covariates $\mathbf{X}_i$, $\mathbf{W}_i$ as well as the compact and convex set of $\mathcal{C}$ leading to Donsker class for $\left\{\beta^\top\mathbf{X}_i:\beta\in\mathcal{B}\right\}$ and $\{\gamma^\top\mathbf{W}_i:\gamma\in\mathcal{G}\}$. Furthermore, derivative of an exponential function is always bounded on compact sets, therefore it satisfies Lipchitz condition, deducing the classes $\{e^{\beta^\top\mathbf{X}_i}:\beta\in\mathcal{B}\}$, $\{e^{-\exp(\beta^\top\mathbf{X}_i)}:\beta\in\mathcal{B}\}$ and $\{e^{-\exp(\gamma^\top\mathbf{W}_i)}:\gamma\in\mathcal{G}\}$ are also Donsker. On the other hand, products and sums of bounded Donsker classes are Donsker, thus

$$u_i(\psi)=\left(L_i(\psi)\lambda_i(\psi)\left(L_i(\psi)+k_i(\gamma)-\lambda_i(\psi)k_i(\gamma)\right)\right)/\left(k_i(\gamma)+L_i(\psi)\right)^2$$

is Donsker class, see Aad W. van der Vaart (1996), so is $\{X_{i\ell}X_{im}u_i(\psi):\psi\in\mathcal{C}\}$. Hence, by Glivenko-Cantelli

$$\sup_{\psi\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}\{X_{i\ell}X_{im}\delta_i J_i u_i(\psi)-\mathbb{E}\left[X_{i\ell}X_{im}\delta_i J_i u_i(\psi)\right]\}\right|$$

converges in probability to 0 as $n\to\infty$. Since $N_n(\varepsilon)\subset\mathcal{C}$ so that

$$\sup_{\psi\in N_n(\varepsilon)}\left|\frac{1}{n}\sum_{i=1}^{n}\{X_{i\ell}X_{im}\delta_i J_i u_i(\psi)-\mathbb{E}\left[X_{i\ell}X_{im}\delta_i J_i u_i(\psi)\right]\}\right|$$

also converges to 0 as $n\to\infty$, which completes the proof. □

Proof of the Theorem 5.2.1. Our proof is followed along the light of Fahrmeir and Kaufmann (1985), Czado and Min (2005), Czado et al. (2007) in no censored case, and recently of Nguyen and Dupuy (2019b).
We first prove for the asymptotic existence of MLEs $\hat{\psi}_n$ on a boundary $\partial N_n(\varepsilon)$ of a certain neighborhood of a true value ψ_0, say $\partial N_n(\varepsilon)=\{\psi\in\mathcal{C}:(\psi-\psi_0)^\top F_n(\psi-\psi_0)=\varepsilon^2\}$. Because of the convex and compact properties of set $\mathcal{C}$ combining with positive definiteness of $H_n(\psi)$ leading to the global and unique maximum of loglikelihood function $\ell_n(\psi)$. We first shall point out that for every $\xi>0$, there exits $\varepsilon>0$ and $n^*\in\mathbb{N}$ such that for $n\geqslant n^*$

$$\mathbb{P}\left(\ell_n(\psi)-\ell_n(\psi_0)<0\text{ for all }\psi\in\partial N_n(\varepsilon)\right)\geqslant 1-\xi,$$

or equivalently, for $n\geqslant n^*$

$$\mathbb{P}\left(\ell_n(\psi)-\ell_n(\psi_0)\geqslant 0\text{ for some }\psi\in\partial N_n(\varepsilon)\right)\leqslant\xi, \tag{5.6.1}$$

Indeed, using the second order Taylor's expansion, we have

$$\begin{aligned}\ell_n(\psi)-\ell_n(\psi_0) &= (\psi-\psi_0)^\top S_n(\psi_0)-\frac{1}{2}(\psi-\psi_0)^\top H_n(\tilde{\psi})(\psi-\psi_0),\\ &:= (\psi-\psi_0)^\top S_n(\psi_0)-R_n(\psi),\end{aligned}$$

where $\tilde{\psi} = \psi_0 + (\psi - \psi_0)t$ lies on the closed interval $[\psi_0, \psi]$ for certain $t \in [0,1]$. Let $0 < c < \frac{1}{2}$, we have

$$\begin{aligned}
&\mathbb{P}(\ell_n(\psi) - \ell_n(\psi_0) \geqslant 0, \text{ for some } \psi \in \partial N_n(\varepsilon)) \\
&\qquad = \mathbb{P}\left((\psi - \psi_0)^\top S_n(\psi_0) \geqslant R_n(\psi) \text{ and } R_n(\psi) > c\varepsilon^2\right) \\
&\qquad\quad + \mathbb{P}\left((\psi - \psi_0)^\top S_n(\psi_0) \geqslant R_n(\psi) \text{ and } R_n(\psi) \leqslant c\varepsilon^2\right), \\
&\qquad \leqslant \mathbb{P}(M) + \mathbb{P}(N),
\end{aligned}$$

where $M = \{(\psi - \psi_0)^\top S_n(\psi_0) > c\varepsilon^2 \text{ for some } \psi \in \partial N_n(\varepsilon)\}$ and $N = \{R_n(\psi) \leqslant c\varepsilon^2 \text{ for some } \psi \in \partial N_n(\varepsilon)\}$. Let $u_n(\psi) = \frac{1}{2} F_n^{\frac{1}{2}}(\psi - \psi_0)$, we derive that if $\psi \in \partial N_n(\psi)$ then $\|u_n(\psi)\| = 1$, we thus have:

$$\begin{aligned}
M &= \left\{ u_n(\psi)^\top F_n^{-\frac{1}{2}} S_n(\psi_0) > c\varepsilon, \text{ for some } \psi \in \partial N_n(\varepsilon) \right\}, \\
&\subseteq \left\{ \sup_{\psi \in N_n(\varepsilon)} \left| u_n(\psi)^\top F_n^{-\frac{1}{2}} S_n(\psi_0) \right| > c\varepsilon \right\}, \\
&\subseteq \left\{ \sup_{\|u_n(\psi)\|=1} \left| u_n(\psi)^\top F_n^{-\frac{1}{2}} S_n(\psi_0) \right| > c\varepsilon \right\}, \\
&= \left\{ \|F_n^{-\frac{1}{2}} S_n(\psi_0)\| > c\varepsilon \right\}.
\end{aligned}$$

From Chebyshev's inequality, the above leads to that $\mathbb{P}(M) \leqslant \mathbb{P}(\|F_n^{-\frac{1}{2}} S_n(\psi_0)\| > c\varepsilon) \leqslant \mathbb{E}\|F_n^{-\frac{1}{2}} S_n(\psi_0)\|^2/(c\varepsilon)^2$. Moreover, since $\mathbb{E}(S_n(\psi_0)) = 0$ and $\mathrm{Var}(S_n(\psi_0)) = F_n$ so by Theorem 1.5 of George A. F. Seber (2003) $\mathbb{E}\|F_n^{-\frac{1}{2}} S_n(\psi_0)\|^2 = tr(F_n^{-\top} F_n) = k$. It follows that

$$\mathbb{P}(M) \leqslant \frac{k}{(c\varepsilon)^2}$$

By letting $\varepsilon = \sqrt{2k/\xi c^2}$ deduce that $\mathbb{P}(M) \leqslant \xi/2$. Turn to the remaining part, we have:

$$\begin{aligned}
N &= \left\{ \frac{1}{2}(\psi - \psi_0)^\top H_n(\tilde{\psi})(\psi - \psi_0) \leqslant c\varepsilon^2, \text{ for some } \psi \in \partial N_n(\varepsilon) \right\}, \\
&= \left\{ \frac{1}{2} u_n(\psi)^\top F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}} u_n(\psi) \leqslant c, \text{ for some } \psi \in \partial N_n(\varepsilon) \right\}, \\
&\subseteq \left\{ \frac{1}{2} \lambda_{\min}\left(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}} \right) u_n(\psi)^\top u_n(\psi) \leqslant c, \text{ for some } \psi \in \partial N_n(\varepsilon) \right\}, \\
&= \left\{ \frac{1}{2} \lambda_{\min}\left(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}} \right) \leqslant c, \text{ for some } \psi \in \partial N_n(\varepsilon) \right\}.
\end{aligned}$$

where the second line to third line comes from in fact that $u_n(\psi)^\top \lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}}) u_n(\psi) \leqslant u_n(\psi)^\top F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}} u_n(\psi)$. Therefore,

$$\mathbb{P}(N) \leqslant \mathbb{P}(\text{there exists } \psi \in N_n(\varepsilon) \text{ such that } \lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}}) \leqslant 2c).$$

From the technique Lemma 5.6.1, $F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}}$ converges in probability to I_k uniformly in $\psi \in N_n(\varepsilon)$, as $n \to \infty$. This implies that $\lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}}$ converges in probability to 1

uniformly in $\psi \in N_n(\varepsilon)$, as $n \to \infty$ (see Maller, 2003, pg. 52).
Furthermore, for some certain $0 \leqslant t \leqslant 1$ and $\psi \in N_n(\varepsilon)$, it is not difficult to have that if $\tilde{\psi} = t\psi + (1-t)\psi_0$ then $\tilde{\psi} \in N_n(\varepsilon)$. On the other hand, since $|\lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}}) - 1| \leqslant \sup_{\psi \in N_n(\varepsilon)} |\lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\psi) F_n^{-\frac{1}{2}}) - 1|$, it follows that $\lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}})$ converges in probability to 1 as $n \to \infty$. Thus, for every $\xi > 0$, for n sufficiently large (say, $n \geqslant n^*$), the assertion $\mathbb{P}(N) \leqslant \mathbb{P}(\text{there exists } \psi \in N_n(\varepsilon) \text{ such that } \lambda_{\min}(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}) F_n^{-\frac{1}{2}}) \leqslant 2c) \leqslant \xi/2$ holds, since $2c < 1$. Finally,

$$\mathbb{P}(\ell_n(\psi) - \ell_n(\psi_0) \geqslant 0, \text{ for some } \psi \in N_n(\varepsilon)) \leqslant \mathbb{P}(M) + \mathbb{P}(N) \leqslant \xi,$$

that completely establishes the proof of existence of unique global maximum of $\ell_n(\psi)$ on $N_n(\varepsilon)$. For the consistency of MLE $\hat{\psi}_n$. Since F_n is symmetric, there exists a orthogonal matrix P such that $P^{-1} F_n P = \text{diag}(\lambda_1, \ldots, \lambda_k)$ where $\lambda_1, \ldots, \lambda_k$ be all eigenvalues of matrix F_n (by principal axis theorem). We have:

$$\begin{aligned}
\lambda_{\min}(F_n)\|\hat{\psi}_n - \psi_0\|^2 &= (\hat{\psi}_n - \psi_0)^\top \lambda_{\min}(F_n) I_k (\hat{\psi}_n - \psi_0), \\
&= (\hat{\psi}_n - \psi_0)^\top P \lambda_{\min}(F_n) I_k P^\top (\hat{\psi}_n - \psi_0), \\
&\leqslant (P^\top(\hat{\psi}_n - \psi_0))^\top \text{diag}(\lambda_1, \ldots, \lambda_k)\, P^\top (\hat{\psi}_n - \psi_0), \\
&= (\hat{\psi}_n - \psi_0)^\top P \text{diag}(\lambda_1, \ldots, \lambda_k)\, P^{-1} (\hat{\psi}_n - \psi_0), \\
&= (\hat{\psi}_n - \psi_0)^\top F_n (\hat{\psi}_n - \psi_0), \\
&= \|F_n^{\frac{1}{2}} (\hat{\psi}_n - \psi_0)\|^2, \\
&\leqslant \varepsilon^2
\end{aligned}$$

with probability tending to 1 as $n \to \infty$. By assumption C3, $\lambda_{\min}(F_n) \to \infty$ as $n \to \infty$. It follows that $\|\hat{\psi}_n - \psi_0\| \to 0$ with probability tending to 1 as $n \to \infty$, which establishes the proof. □

Proof of the Theorem 5.2.2 We first prove asymptotic normality of the normalized score vector $F_n^{-\frac{1}{2}} S_n$, where S_n is a short notation for $S_n(\psi_0)$. Let a be any vector in $\mathbb{R}^k$. We shall point out that $a^\top F^{-\frac{1}{2}} S_n$ converges in distribution to $\mathcal{N}(0, I_k)$ (here, without loss of generality, assume that $\|a\| = 1$). It is easy to express S_n as a sum of the form $S_n = \sum_{i=1}^n S_{n,i}$, where $S_{n,i}$ is the $k-$ dimensional random vector, $S_{n,i} = (S_{n,i,1}, \ldots, S_{n,i,k})^\top$. Since the conditions C1, C2 and C4, it is not difficult to see that, components of $S_{n,i}$ are bounded by some finite positive constant c_1, thus $|S_{n,i,\ell}| < c_1$. It follows that, $\|S_{n,i}\|^2 = \sum_{\ell=1}^k S_{n,i,\ell}^2$, $\ell = 1, \ldots, k$, thus is bounded from above, say $c_2 < \infty$, $\|S_{n,i}\|^2 \leqslant c_2$.

Consider

$$a^\top F_n^{\frac{1}{2}} \sum_{i=1}^n S_{n,i} := \sum_{i=1}^n S_{n,i}^*,$$

It is clear that $\mathbb{E}(S_{n,i}^*) = 0$ and $\text{Var}(\sum_{i=1}^n S_{i,n}^*) = 1$, so to show the convergence above, we sufficiently need to point out the validity of Lindeberg condition of sequence $(S_{n,i}^*)$ as follows:

$$\text{for every } \varepsilon > 0, \sum_{i=1}^n \mathbb{E}\left(S_{n,i}^{*2} 1_{\{|S_{n,i}^*| > \varepsilon\}}\right) \to 0 \text{ as } n \to \infty.$$

From some basis algebraic properties, for instant, $\|A^{-1}\| = 1/\lambda_{\min}(A)$, and if A is positive definite

then $\|A^{1/2}\|^2 = \|A\|$. Let $\varepsilon > 0$, and by condition C3, we have:

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{E}\left(S_{n,i}^{*2} 1_{\{|S_{n,i}^*|>\varepsilon\}}\right) &\leqslant \mathbb{E}\left(\|a\|^2 \|F_n^{-\frac{1}{2}}\|^2 \|S_{n,i}\|^2 1_{\{|S_{n,i}^*|>\varepsilon\}}\right), \\ &\leqslant \frac{\kappa c_2}{n} \sum_{i=1}^{n} \mathbb{E}(1_{\{|S_{n,i}^*|>\varepsilon\}}),\end{aligned}$$

Moreover, $\varepsilon^2 \leqslant \|a^\top\|^2 \|F_n^{-\frac{1}{2}}\|^2 \|\sum_{i=1}^n S_{n,i}\|^2 \leqslant \|F_n^{-1}\| c_2$ thus $\{|S_{n,i}^*| > \varepsilon\}$ replies $\{\lambda_{\min}(F_n) < c_2/\varepsilon^2\}$, so, $1_{\{|S_{n,i}^*|>\varepsilon\}} \leqslant 1_{\{\lambda_{\min}(F_n)<c_2/\varepsilon^2\}}$ From condition C3, $\lambda_{\min}(F_n) \to \infty$ as $n \to \infty$, leading to $1_{\{\lambda_{\min}(F_n)<c_2/\varepsilon^2\}} = 0$. Therefore, $\sum_{i=1}^n \mathbb{E}(S_{n,i}^{*2} 1_{\{|S_{n,i}^*|>\varepsilon\}}) = 0$ as $n \to \infty$. It follows that for every $a \in \mathbb{R}^k$, $a^\top F_n^{-\frac{1}{2}} S_n$ converges in distribution to $\mathcal{N}(0,1)$, consequently, $F_n^{-\frac{1}{2}} S_n$ converges in distribution to $\mathcal{N}(0, I_k)$ by Cramer-Wold device Theorem.

Now, relying upon the asymptotic normality of the normalized score vector above, we shall point out that it also holds for the normalized maximum likelihood estimates $\hat{\psi}_n$. Indeed, by expanding S_n around $\hat{\psi}_n$ and from the mean value theorem for vector valued function, we have

$$\begin{aligned}S_n(\hat{\psi}_n) - S_n &= \int_0^1 S_n'\left(\psi_0 + (\hat{\psi}_n - \psi_0)t\right) dt\, (\hat{\psi}_n - \psi_0), \\ &= \int_0^1 H_n(\tilde{\psi}_n) dt.(\psi_0 - \hat{\psi}_n),\end{aligned}$$

where $0 \leqslant t \leqslant 1$ and $\tilde{\psi}_n = \psi_0 + (\hat{\psi}_n - \psi_0)t$. Note that, $S_n(\hat{\psi}_n) = 0$ and by multiplying both sides by $F_n^{-\frac{1}{2}}$, we thus obtain

$$\begin{aligned}F_n^{-\frac{1}{2}} S_n &= \int_0^1 F_n^{-\frac{1}{2}} H_n(\tilde{\psi}_n) F_n^{-\frac{1}{2}} dt\, F_n^{\frac{1}{2}} (\hat{\psi}_n - \psi_0), \\ &= \left(\int_0^1 F_n^{-\frac{1}{2}} H_n(\tilde{\psi}_n) F_n^{-\frac{1}{2}} dt - I_k\right) F_n^{\frac{1}{2}} (\hat{\psi}_n - \psi_0) + F_n^{\frac{1}{2}} (\hat{\psi}_n - \psi_0), \\ &= \int_0^1 \left(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}_n) F_n^{-\frac{1}{2}} - I_k\right) dt\, F_n^{\frac{1}{2}} (\hat{\psi}_n - \psi_0) + F_n^{\frac{1}{2}} (\hat{\psi}_n - \psi_0),\end{aligned}$$

Suppose that $\hat{\psi}_n$ belongs to $N_n(\varepsilon)$, so does $\tilde{\psi}_n$. Furthermore,

$$F_n^{-\frac{1}{2}} H_n(\tilde{\psi}_n) F_n^{-\frac{1}{2}} - I_k \leqslant \|F_n^{-\frac{1}{2}} H_n(\tilde{\psi}_n) F_n^{-\frac{1}{2}} - I_k\| \leqslant \sup_{\psi \in N_n(\varepsilon)} \|F_n^{-\frac{1}{2}} H_n(\psi) F_n^{-\frac{1}{2}} - I_k\|,$$

and since

$$\int_0^1 \left(\sup_{\psi \in N_n(\varepsilon)} \|F_n^{-\frac{1}{2}} H_n(\psi) F_n^{-\frac{1}{2}} - I_k\|\right) dt = \sup_{\psi \in N_n(\varepsilon)} \|F_n^{-\frac{1}{2}} H_n(\psi) F_n^{-\frac{1}{2}} dt - I_k\|$$

by Lemma (5.6.1) yields

$$\int_0^1 \left(\sup_{\psi \in N_n(\varepsilon)} \|F_n^{-\frac{1}{2}} H_n(\psi) F_n^{-\frac{1}{2}} - I_k\|\right) dt \to 0 \text{ in probability, as } n \to \infty,$$

We thus have

$$\int_0^1 \left(F_n^{-\frac{1}{2}} H_n(\tilde{\psi}_n) F_n^{-\frac{1}{2}} - I_k\right) dt \to 0 \text{ in probability, as } n \to \infty$$

Finally, combining the asymptotic normality of the normalized score vector and Slutzky theorem, getting the deserved result

$$F_n^{\frac{1}{2}}(\hat{\psi}_n - \psi_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I_k) \text{ as } n \to \infty.$$

□

5.7 Appendix 2: Technical calculations

For the purposes of simplicity, we write some notations as: $k_i(\gamma) = e^{\gamma^\top \mathbf{W}_i}$, $\nu_i(\beta) = e^{\beta^\top \mathbf{X}_i}$, $\lambda_i(\psi) = e^{\beta^\top \mathbf{X}_i}\left(1 + e^{\gamma^\top \mathbf{W}_i}\right)$, $L_i(\psi) = e^{-e^{\beta^\top \mathbf{X}_i}\left(1+e^{\gamma^\top \mathbf{W}_i}\right)}$, the loglikelihood function above becomes

$$\begin{aligned}
\ell_n(\psi) &= \sum_{i=1}^{n} \{\delta_i J_i \log\left(k_i(\gamma) + L_i(\psi)\right) + (\delta_i Y_i^* - 1)\log\left(1 + k_i(\gamma)\right) \\
&\quad + \delta_i\left(1 - J_i\right)\left(Y_i^* \beta^\top \mathbf{X}_i - \lambda_i(\psi) - \log\left(Y_i^*!\right)\right) \\
&\quad + (1-\delta_i)(1-J_i)\log\left(1 - \sum_{k=0}^{Y_i^*-1} \frac{L_i(\psi)\lambda_i(\psi)^k}{k!}\right)\Bigg\}.
\end{aligned}$$

Lemma 5.7.1 (The derivative of survival function). *Let* $\mathcal{S}_{\lambda_i(\psi)}(Y_i^*) = \mathbb{P}\left(\mathcal{P}\left(\lambda_i(\psi)\right) \geqslant Y_i^*\right) = 1 - \sum_{k=0}^{Y_i^*-1} \frac{L_i(\psi)\lambda_i(\psi)^k}{k!}$, *then*

$$\begin{aligned}
\frac{\partial}{\partial \beta_j}\mathcal{S}_{\lambda_i(\psi)}(Y_i^*) &= \lambda_i(\psi)\mathbb{P}\left(\mathcal{P}\left(\lambda_i(\psi)\right) = Y_i^* - 1\right), \quad \forall j = 1, \ldots, p, \qquad (5.7.1)\\
\frac{\partial}{\partial \gamma_\ell}\mathcal{S}_{\lambda_i(\psi)}(Y_i^*) &= \nu_i(\psi)k_i(\psi)\mathbb{P}\left(\mathcal{P}\left(\lambda_i(\psi)\right) = Y_i^* - 1\right) \quad \forall \ell = 1, \ldots, q.
\end{aligned}$$

Proof of the Lemma 5.7.1. We will indicate for the (5.7.1), the rest can be treatment similarly. We have:

$$\begin{aligned}
\frac{\partial}{\partial \beta_j}\mathcal{S}_{\lambda_i(\psi)}(Y_i^*) &= -\sum_{k=0}^{Y_i^*-1} \frac{1}{k!}\left(L_i'(\psi)\lambda_i(\psi)^k + L_i(\psi)\lambda_i'(\psi)\right), \\
&= -\sum_{k=0}^{Y_i^*-1} \frac{1}{k!}\left(-L_i(\psi)\lambda_i(\psi)\lambda_i(\psi)^k + L_i(\psi)k\lambda_i(\psi)^{k-1}\lambda_i(\psi)\right), \\
&= \sum_{k=0}^{Z_i^*-1} \frac{L_i(\psi)\lambda_i(\psi)^{k+1}}{k!} - \sum_{k=0}^{Y_i^*-1} \frac{kL_i(\psi)\lambda_i(\psi)^k}{k!}, \\
&= \sum_{k=0}^{Y_i^*-1} \frac{L_i(\psi)\lambda_i(\psi)^{k+1}}{k!} - \sum_{k=0}^{Y_i^*-2} \frac{L_i(\psi)\lambda_i(\psi)^{k+1}}{k!}, \\
&= \lambda_i(\psi)\mathbb{P}\left(\mathcal{P}\left(\lambda_i(\psi)\right) = Z_i^* - 1\right), \quad \forall j = 1, \ldots, p.
\end{aligned}$$

□

Theorem 5.7.2 (The first derivatives).

$$\begin{aligned}
\frac{\partial \ell_n(\psi)}{\partial \beta_j} &= \sum_{i=1}^{n} X_{ij} \left\{ -\delta_i J_i \frac{L_i(\psi)\lambda_i(\psi)}{k_i(\gamma)+L_i(\psi)} + \delta_i(1-J_i)\left(Y_i^* - \lambda_i(\psi)\right) \right. \\
&\qquad \left. -(1-\delta_i)(1-J_i) \sum_{k=0}^{Y_i^*-1} \frac{L_i(\psi)\lambda_i(\psi)^k \left(k-\lambda_i(\psi)\right)}{k!\mathcal{S}_{\lambda_i(\psi)}\left(Y_i^*\right)} \right\}, \\
\frac{\partial \ell_n(\psi)}{\partial \gamma_\ell} &= \sum_{i=1}^{n} W_{i\ell} \left\{ \delta_i J_i \frac{k_i(\gamma) - L_i(\psi)k_i(\gamma)\nu_i(\beta)}{k_i(\gamma)+L_i(\psi)} + \left(\delta_i Y_i^* - 1\right)\frac{k_i(\gamma)}{1+k_i(\gamma)} - \delta_i(1-J_i)\nu_i(\beta)k_i(\gamma) \right. \\
&\qquad \left. -(1-\delta_i)(1-J_i) \sum_{k=0}^{Y_i^*-1} \frac{L_i(\psi)\nu_i(\beta)\lambda_i(\psi)^{k-1}k_i(\gamma)\left(k-\lambda_i(\psi)\right)}{k!\mathcal{S}_{\lambda_i(\psi)}\left(Y_i^*\right)} \right\}.
\end{aligned}$$

Proof of the Theorem 5.7.2 Because of their similarity, we therefore only show for the first expression, the rest one can be treatment analogously. In deed, by taking the partial derivative with respect to β_j, $j = 1, \ldots, p$, we have:

$$\begin{aligned}
\frac{\partial \ell_n(\psi)}{\partial \beta_j} &= \sum_{i=1}^{n} \left\{ \delta_i J_i \frac{e^{-e^{\beta^\top \mathbf{x}_i}(1+e^{\gamma^\top \mathbf{w}_i})}(1+e^{\gamma^\top \mathbf{w}_i})e^{\beta^\top \mathbf{x}_i}(-X_{ij})}{e^{\gamma^\top \mathbf{w}_i} + e^{-e^{\beta^\top \mathbf{x}_i}(1+e^{\gamma^\top \mathbf{w}_i})}} \right. \\
&\quad +\delta_i(1-J_i)\left(Y_i^* X_{ij} - e^{\beta^\top \mathbf{x}_i}(1+e^{\gamma^\top \mathbf{w}_i})X_{ij}\right) \\
&\quad \left. -(1-\delta_i)(1-J_i)\frac{\sum_{k=0}^{Y_i^*-1} \frac{(1+e^{\gamma^\top \mathbf{w}_i})^k}{k!} e^{-e^{\beta^\top \mathbf{x}_i}(1+\gamma^\top \mathbf{w}_i)} X_{ij} e^{k\beta^\top \mathbf{x}_i}\left(-(1+e^{\gamma^\top \mathbf{w}_i})e^{\beta^\top \mathbf{x}_i}+k\right)}{\mathcal{S}_{\lambda_i}(\psi)(Y_i^*)} \right\}, \\
&= \sum_{i=1}^{n} \left\{ \delta_i J_i \frac{e^{-e^{\beta^\top \mathbf{x}_i}(1+e^{\gamma^\top \mathbf{w}_i})}(1+e^{\gamma^\top \mathbf{w}_i})e^{\beta^\top \mathbf{x}_i}(-X_{ij})}{e^{\gamma^\top \mathbf{w}_i} + e^{-e^{\beta^\top \mathbf{x}_i}(1+e^{\gamma^\top \mathbf{w}_i})}} \right. \\
&\quad +\delta_i(1-J_i)\left(Y_i^* X_{ij} - e^{\beta^\top \mathbf{x}_i}(1+e^{\gamma^\top \mathbf{w}_i})X_{ij}\right) \\
&\quad \left. -(1-\delta_i)(1-J_i)\sum_{k=0}^{Y_i^*-1} \frac{X_{ij} e^{-e^{\beta^\top \mathbf{x}_i}(1+\gamma^\top \mathbf{w}_i)} e^{k\beta^\top \mathbf{x}_i}(1+e^{\gamma^\top \mathbf{w}_i})^k \left(-(1+e^{\gamma^\top \mathbf{w}_i})e^{\beta^\top \mathbf{x}_i}+k\right)}{k!\mathcal{S}_{\lambda_i}(\psi)(Y_i^*)} \right\}, \\
&= \sum_{i=1}^{n} X_{ij} \left\{ \delta_i J_i \frac{-L_i(\psi)\lambda_i(\psi)}{k_i(\gamma)+L_i(\psi)} + \delta_i(1-J_i)\left[Y_i^* - \lambda_i(\psi)\right] \right. \\
&\quad \left. -(1-\delta_i)(1-J_i)\sum_{k=0}^{Y_i^*-1} \frac{L_i(\psi)\lambda_i(\psi)^k\left[k-\lambda_i(\psi)\right]}{k!\mathcal{S}_{\lambda_i}(\psi)(Y_i^*)} \right\}.
\end{aligned}$$

□

Now, we shall implement for calculating the second derivatives. Rather, we omit the parameters ψ, β, γ in the expressions $L_i(\psi), \lambda_i(\psi), k_i(\gamma), \nu_i(\beta)$ to become the simpler form $L_i, \lambda_i, k_i, \nu_i$, but before stating, $\forall i = 1, \ldots, n$, we denote

$$
\begin{aligned}
u_i(\psi) &= \frac{L_i\lambda_i\left(L_i+k_i-\lambda_i k_i\right)}{\left(k_i+L_i\right)^2},\\
v_i(\psi) &= \sum_{k=0}^{Y_i^*-1}\frac{L_i\lambda_i^k}{k!\mathcal{S}_{\lambda_i}\left(Y_i^*\right)^2}\\
&\quad\times\Big[\left((k-\lambda_i)^2-\lambda_i\right)\mathcal{S}_{\lambda_i}\left(Y_i^*\right)-\left(k-\lambda_i\right)\lambda_i\mathbb{P}\left(\mathcal{P}\left(\lambda_i\right)=Y_i^*-1\right)\Big],\\
s_i(\psi) &= \frac{L_i k_i\left(\nu_i k_i+\nu_i L_i-\nu_i k_i\lambda_i-\lambda_i\right)}{\left(k_i+L_i\right)^2},\\
t_i(\psi) &= \sum_{k=0}^{Y_i^*-1}\frac{L_i\nu_i k_i\lambda_i^{k-2}}{k!\mathcal{S}_{\lambda_i}\left(Y_i^*\right)^2}\\
&\quad\times\Big[\left((k-\lambda_i)^2-\lambda_i\right)\mathcal{S}_{\lambda_i}\left(Y_i^*\right)-\lambda_i(k-\lambda_i)\mathbb{P}\left(\mathcal{P}\left(\lambda_i\right)=Y_i^*-1\right)\Big],
\end{aligned}
$$

$$
\begin{aligned}
f_i(\psi) &= \frac{k_i\left[\left(1-L_i\nu_i-L_i k_i\nu_i^2\right)\left(k_i+L_i\right)-k_i\left(1-L_i\nu_i\right)^2\right]}{\left(k_i+L_i\right)^2},\\
g_i(\psi) &= \sum_{k=0}^{Y_i^*-1}\frac{L_i\nu_i k_i\lambda_i^{k-2}}{k!\mathcal{S}_{\lambda_i}\left(Y_i^*\right)^2}\times\Big[\left(k-\lambda_i\right)\left(\lambda_i+\left(k-1\right)\nu_i k_i-\lambda_i k_i\nu_i\right)-\lambda_i k_i\nu_i\\
&\qquad -\lambda_i k_i\nu_i\left(k-\lambda_i\right)\mathbb{P}\left(\mathcal{P}\left(\lambda_i\right)=Y_i^*-1\right)\Big],
\end{aligned}
$$

Theorem 5.7.3 (**The second derivatives**).

$$
\begin{aligned}
\frac{\partial^2\ell_n(\psi)}{\partial\beta_j\partial\beta_m} &= \sum_{i=1}^{n}X_{ij}X_{im}\left\{-\delta_i J_i u_i(\psi)-\delta_i(1-J_i)\lambda_i(\psi)-(1-\delta_i)(1-J_i)v_i(\psi)\right\},\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\forall j,m=1,\ldots,p,\\
\frac{\partial^2\ell_n(\psi)}{\partial\beta_j\partial\gamma_\ell} &= \sum_{i=1}^{n}X_{ij}W_{i\ell}\left\{-\delta_i J_i s_i(\psi)-\delta_i(1-J_i)\nu_i(\beta)k_i(\gamma)-(1-\delta)(1-J_i)t_i(\psi)\right\},\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\forall j=1,\ldots,p,\forall\ell=1,\ldots,q,\\
\frac{\partial^2\ell_n(\psi)}{\partial\gamma_\ell\partial\gamma_s} &= \sum_{i=1}^{n}W_{i\ell}W_{is}\left\{\delta_i J_i g_i(\psi)+\left(\delta_i Y_i^*-1\right)\frac{k_i(\gamma)}{\left(1+k_i(\gamma)\right)^2}-\delta_i(1-J)\nu_i(\beta)k_i(\gamma)\right.\\
&\qquad\left.-(1-\delta_i)(1-J_i)f_i(\psi)\right\},\quad\forall\ell,s=1,\ldots,q.
\end{aligned}
$$

Proof of the Theorem 5.7.3. We have:

$$
\begin{aligned}
\frac{\partial^2\ell_n(\psi)}{\partial\beta_j\partial\beta_m} &= \sum_{i=1}^{n}X_{ij}X_{im}\left\{-\delta_i J_i\frac{\left(L_i'\lambda_i+L_i\lambda_i'\right)\left(k_i+L_i\right)-\left(k_i'+L_i'\right)\left(L_i\lambda_i\right)}{\left(k_i+L_i\right)^2}-\delta_i(1-J_i)\lambda_i'\right.\\
&\quad\left.-(1-\delta_i)(1-J_i)\sum_{k=0}^{Y_i^*-1}\frac{[L_i'(k\lambda_i^k-\lambda_i^{k+1})+(k\lambda_i^k-\lambda_i^{k+1})'L_i]\mathcal{S}_{\lambda_i}(Y_i^*)-\mathcal{S}_{\lambda_i}'(Y_i^*)L_i\lambda_i^k(k-\lambda_i)}{k!\mathcal{S}_{\lambda_i}(Y_i^*)^2}\right\},\\
&= \sum_{i=1}^{n}X_{ij}X_{im}\left\{-\delta_i J_i\frac{L_i\lambda_i\left[L_i\lambda_i+(1-\lambda_i)(k_i+L_i)\right]}{\left(k_i+L_i\right)^2}-\delta_i(1-J_i)\lambda_i\right.\\
&\quad\left.-(1-\delta_i)(1-J_i)\sum_{k=0}^{Y_i^*-1}\frac{L_i\lambda_i^k[-\lambda_i(k-\lambda_i)+k^2-(k+1)\lambda_i]\mathcal{S}_{\lambda_i}(Y_i^*)-\mathcal{S}_{\lambda_i}'(Y_i^*)L_i\lambda_i^k(k-\lambda_i)}{k!\mathcal{S}_{\lambda_i}(Y_i^*)^2}\right\},
\end{aligned}
$$

finally, combining the derivative of survival function 5.7.1, the abbreviations of $u_i(\psi)$, $v_i(\psi)$, letting common factors and rearrangement the order, leads to

$$\frac{\partial^2 \ell_n(\psi)}{\partial\beta_j \partial\beta_m} = \sum_{i=1}^{n} X_{ij}X_{im}\left\{-\delta_i J_i u_i(\psi) - \delta_i(1-J_i)\lambda_i(\psi) - (1-\delta_i)(1-J_i)v_i(\psi)\right\}.$$

□

Acknowledgements

Authors acknowledge financial support from the Ministry of Education and Training of the Socialist Republic of Vietnam and the French Embassy in Vietnam and logistical support from Campus France (French national agency for the promotion of higher education, international student services, and international mobility).

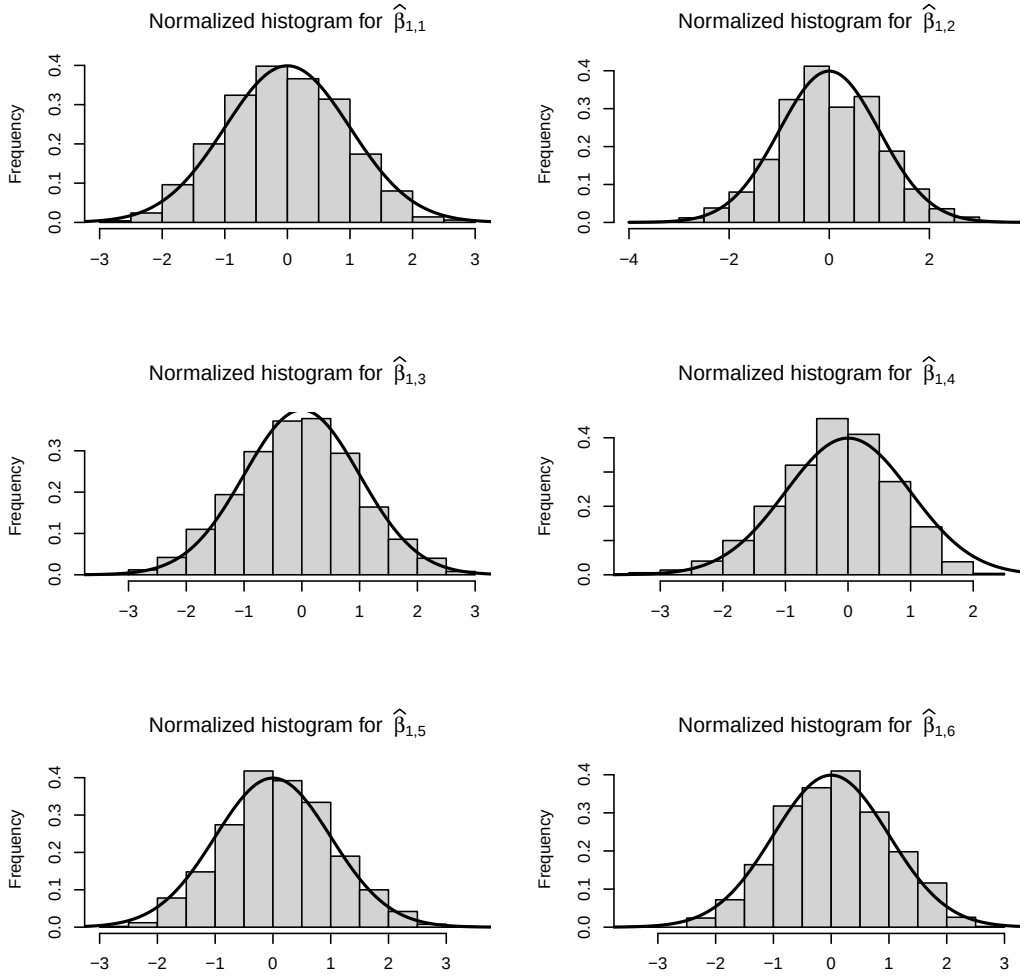

Figure 5.1: Histograms of the normalized estimates $(\widehat{\beta}_{j,n}-\beta_j)/\text{s.e.}(\widehat{\beta}_{j,n})$, $j = 1, \ldots, 6$ in censored MZIP model (n=500, ZI=40%, c=40% censoring).

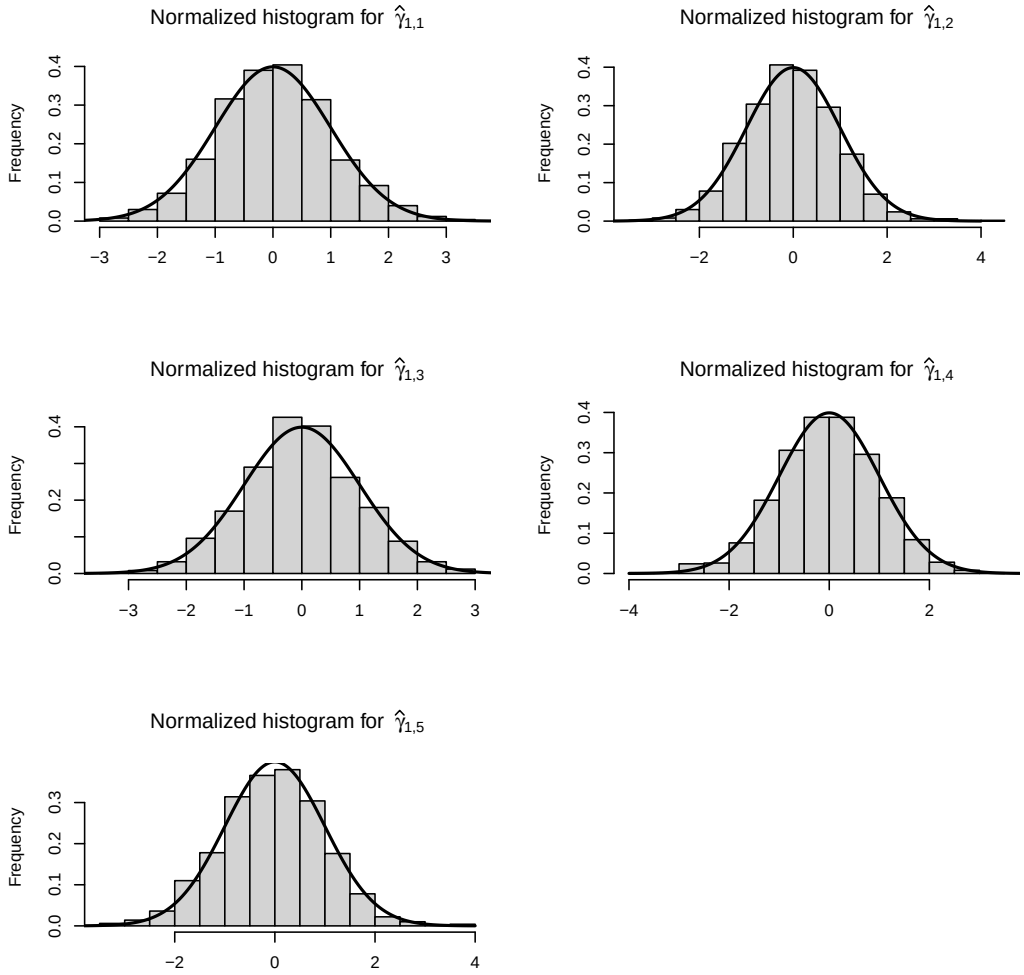

Figure 5.2: Histograms of the normalized estimates $(\widehat{\gamma}_{j,n} - \gamma_j)/\text{s.e.}(\widehat{\gamma}_{j,n})$, $j = 1, \ldots, 5$ and $(\widehat{\varphi}_n - \varphi)/\text{s.e.}(\widehat{\varphi}_n)$ in censored MZIP model (n=500, ZI=40%, c=40% censoring).

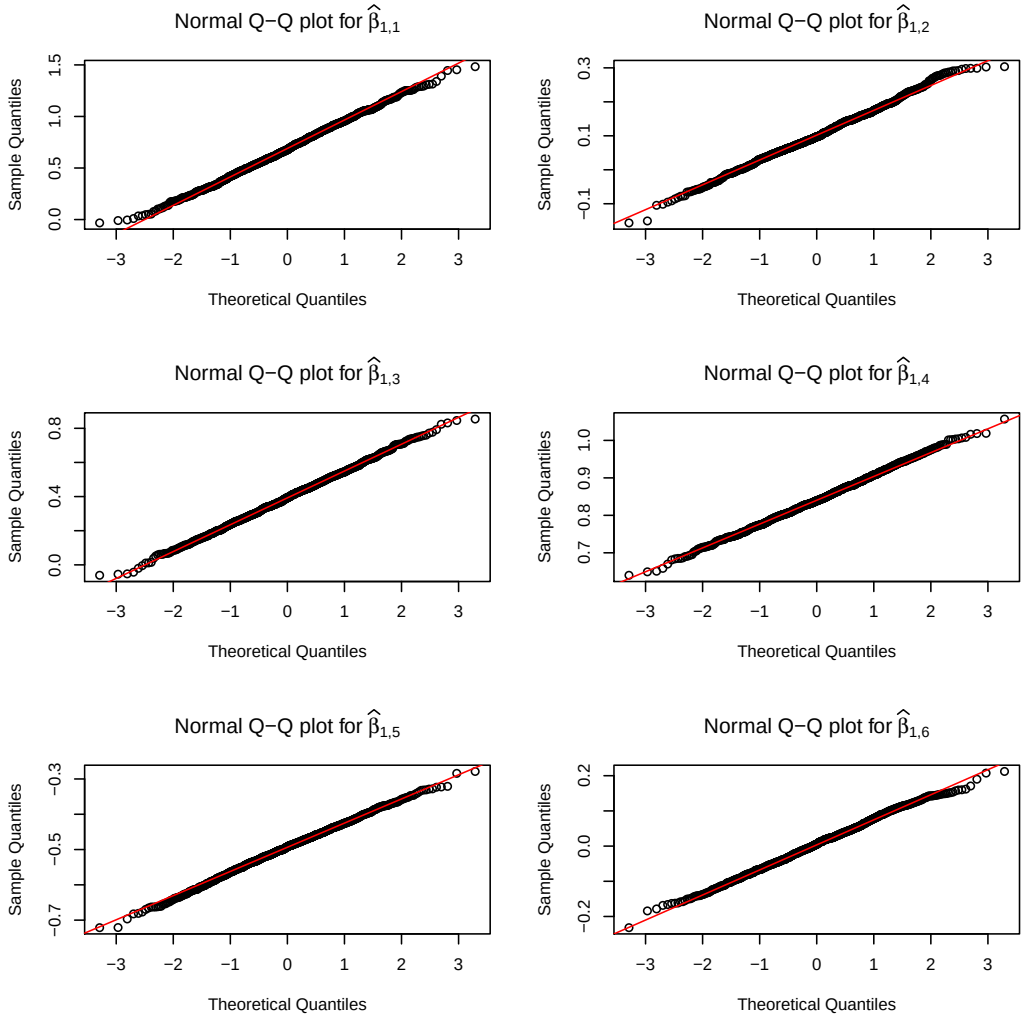

Figure 5.3: Normal Q-Q plots for $\hat{\beta}_{1,n}, \ldots, \hat{\beta}_{6,n}$ with $n = 500, c = 0.4$ and a proportion of zero-inflation equal to 0.4.

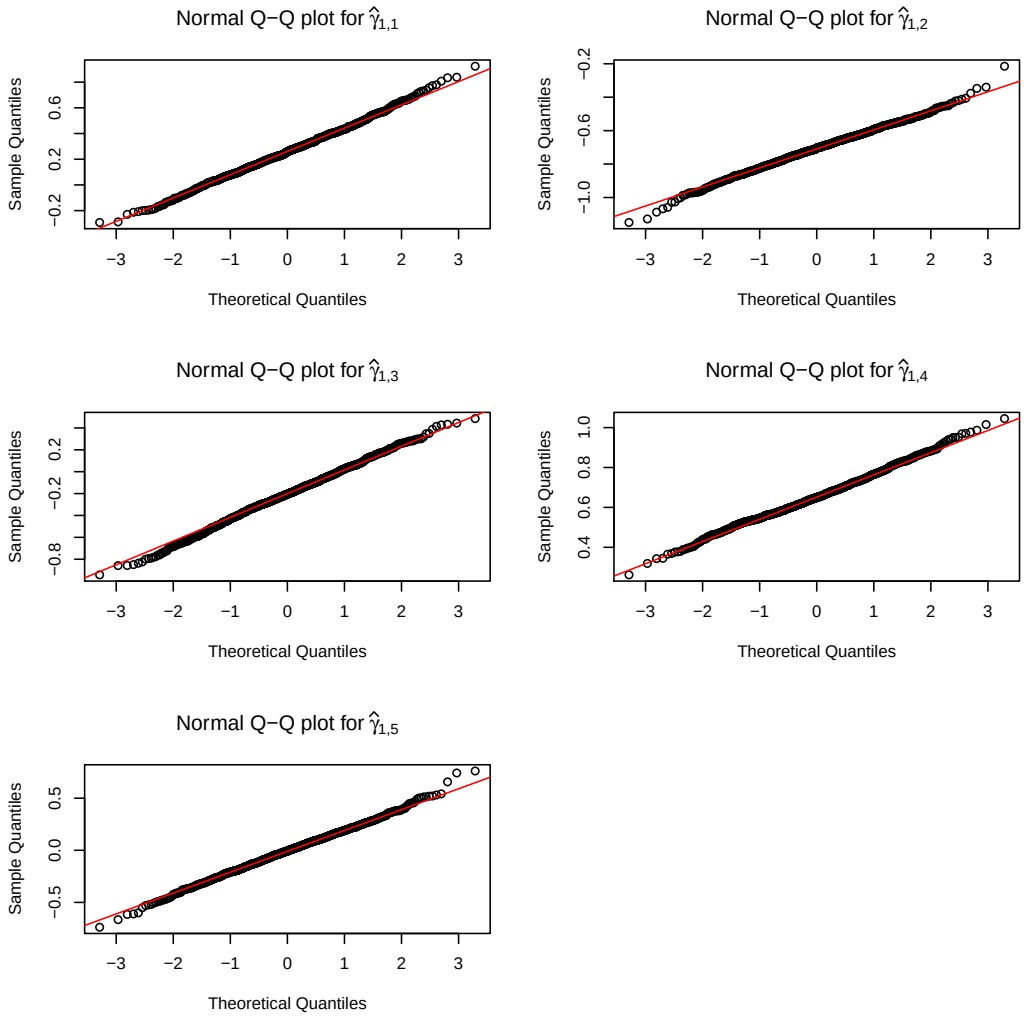

Figure 5.4: Normal Q-Q plots for $\hat{\gamma}_{1,n}, \ldots, \hat{\gamma}_{5,n}$ with $n = 500, c = 0.4$ and a proportion of zero-inflation equal to 0.4.

.

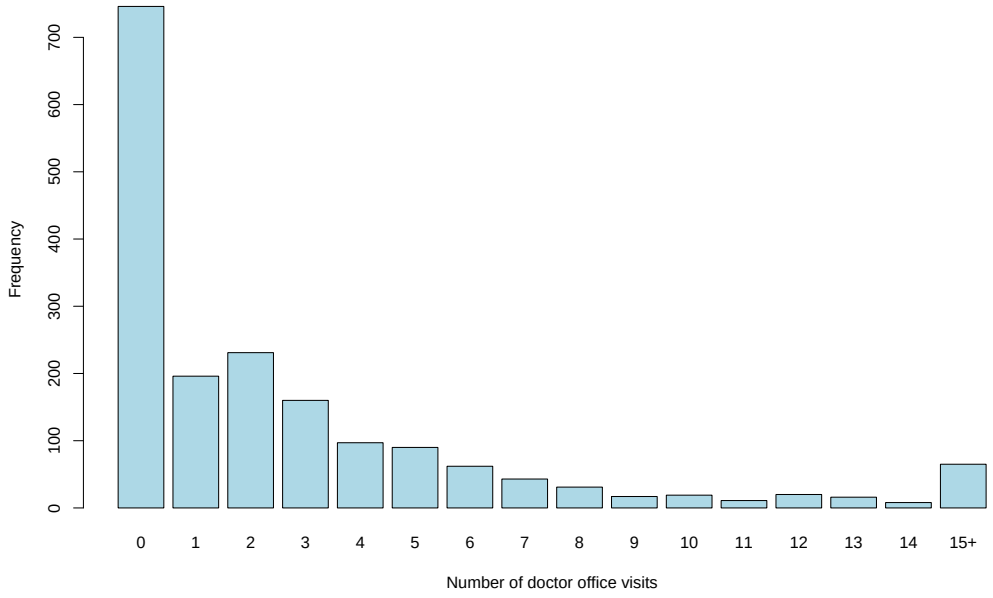

Figure 5.5: Number of doctor office visits.

Bibliography

Jon A. Wellner (auth.) Aad W. van der Vaart. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag New York, 1 edition, 1996. ISBN 978-1-4757-2547-6,978-1-4757-2545-2. URL https://doi.org/10.1007/978-1-4757-2545-2.

Robiah Adnan, Seyed Ehsan Saffari, and William Greene. Hurdle negative binomial regression model with right censored count data. *Journal of Statistics and Operations Research Transactions*, 36:181–194, 12 2012.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974. ISSN 0018-9286. doi: 10.1109/TAC.1974.1100705.

A. Antoniadis, J. Berruyer, and R. Carmona. *Régression non linéaire et applications*. Collection "Economie et statistiques avancées.": Série Ecole nationale de la statistique et de l'administration économique et Centre d'études des programmes économiques. Economica, 1992.

John S.; Divaris Kimon; Herring Amy H.; Das Kalyan Benecha, Habtamu K.; Preisser. Marginalized zero-inflated poisson models with missing covariates. *Biometrical Journal*, 2018. ISSN 0323-3847,1521-4036. doi: 10.1002/bimj.201600249. URL /scimag/10.1002%2Fbimj.201600249.

Anne Buu, Norman J Johnson, Runze Li, and Xianming Tan. New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Statistics in medicine*, 30:2326–40, 08 2011. doi: 10.1002/sim.4268.

Ekkehart Dietz; Dankmar Böhning. On estimation of the poisson parameter in zero-modified poisson models. *Computational Statistics & Data Analysis*, 34:441–459, 2000. ISSN 0167-9473. doi: 10.1016/s0167-9473(99)00111-5. URL /scimag/10.1016%2Fs0167-9473%2899%2900111-5.

A. Colin Cameron and Pravin K. Trivedi. *Regression analysis of count data*, volume 53 of *Econometric Society Monographs*. Cambridge University Press, Cambridge, second edition, 2013. ISBN 978-1-107-66727-3. doi: 10.1017/CBO9781139013567. URL https://doi.org/10.1017/CBO9781139013567.

Michele Campolieti. The recurrence of occupational injuries: estimates from a zero inflated count model. *Applied Economics Letters*, 9(9):595–600, 2002.

Steven Caudill and Mixon Franklin G, Jr. Modeling household fertility decisions: Estimation and testing of censored regression models for count data. *Empirical Economics*, 20:183–96, 02 1995. doi: 10.1007/BF01205434.

Saptarshi Chatterjee, Shrabanti Chowdhury, Himel Mallick, Prithish Banerjee, and Broti Garai. Group regularization for zero-inflated negative binomial regression models with an application

to healthcare demand in germany. *Statistics in Medicine*, 37:3012–3026, 09 2018. doi: 10.1002/sim.7804.

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

Achim Zeileis (auth.) Christian Kleiber. *Applied Econometrics with R*. Use R. Springer-Verlag New York, 1 edition, 2008. ISBN 0387773169,9780387773162. doi: 10.1007/978-0-387-77318-6. URL https://www.springer.com/la/book/9780387773162.

P.C. Consul and Felix Famoye. Generalized poisson regression model. *Communications in Statistics - Theory and Methods*, 21(1):89–109, 1992. doi: 10.1080/03610929208830766. URL https://doi.org/10.1080/03610929208830766.

Claudia Czado and Aleksey Min. Consistency and asymptotic normality of the maximum likelihood estimator in a zero-inflated generalized poisson regression, 2005. URL http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-1792-8.

Claudia Czado, Vinzenz Erhardt, Aleksey Min, and Stefan Wagner. Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. *Stat. Model.*, 7(2):125–153, 2007. ISSN 1471-082X. doi: 10.1177/1471082X0700700202. URL https://doi.org/10.1177/1471082X0700700202.

Partha Deb and Pravin K. Trivedi. Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics*, 12:313–36, 05 1997. doi: 10.1002/(SICI)1099-1255(199705)12:33.0.CO;2-G.

Dianliang Deng and Yu Zhang. Score tests for both extra zeros and extra ones in binomial mixed regression models. *Comm. Statist. Theory Methods*, 44(14):2881–2897, 2015. ISSN 0361-0926. doi: 10.1080/03610926.2013.809118. URL https://doi.org/10.1080/03610926.2013.809118.

Alpha Oumar Diallo, Aliou Diop, and Jean-François Dupuy. Asymptotic properties of the maximum-likelihood estimator in zero-inflated binomial regression. *Communications in Statistics - Theory and Methods*, 46(20):9930–9948, 2017a.

Alpha Oumar Diallo, Aliou Diop, and Jean-François Dupuy. Analysis of multinomial counts with joint zero-inflation, with an application to health economics. *J. Statist. Plann. Inference*, 194:85–105, 2018. ISSN 0378-3758. doi: 10.1016/j.jspi.2017.09.005. URL https://doi.org/10.1016/j.jspi.2017.09.005.

Alpha Oumar Diallo, Aliou Diop, and Jean-François Dupuy. Estimation in zero-inflated binomial regression with missing covariates. *Statistics*, 53(4):839–865, 2019. doi: 10.1080/02331888.2019.1619741. URL https://doi.org/10.1080/02331888.2019.1619741.

Aba Diop, Aliou Diop, and Jean-François Dupuy. Maximum likelihood estimation in the logistic regression model with a cure fraction. *Electron. J. Stat.*, 5:460–483, 2011. ISSN 1935-7524. doi: 10.1214/11-EJS616. URL https://doi.org/10.1214/11-EJS616.

Aba Diop, Aliou Diop, and Jean-François Dupuy. Simulation-based inference in a zero-inflated Bernoulli regression model. *Comm. Statist. Simulation Comput.*, 45(10):3597–3614, 2016. ISSN 0361-0918. doi: 10.1080/03610918.2014.950743. URL https://doi.org/10.1080/03610918.2014.950743.

Annette J. Dobson and Adrian G. Barnett. *An introduction to generalized linear models*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL, fourth edition, 2018. ISBN 978-1-138-74151-5; 978-1-138-74168-3. For the third edition see [MR2459739].

Peter K. Dunn and Gordon K. Smyth. *Generalized linear models with examples in R*. Springer Texts in Statistics. Springer, New York, 2018. ISBN 978-1-4419-0117-0; 978-1-4419-0118-7. doi: 10.1007/978-1-4419-0118-7. URL https://doi.org/10.1007/978-1-4419-0118-7.

Jean-François Dupuy. Inference in a generalized endpoint-inflated binomial regression model. *Statistics*, 51(4):888–903, 2017. ISSN 0233-1888. doi: 10.1080/02331888.2017.1316724. URL https://doi.org/10.1080/02331888.2017.1316724.

Jean François Dupuy. *Méthodes statistiques pour l'analyse de données de comptage surdispersées*, volume 4 of *Biostastique et sciences de la santé*. ISTE Press, France, first edition, 2018. ISBN 978-1-78405-522-6,978-1-784056-522-5.

Ludwig Fahrmeir and Heinz Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.*, 13(1):342–368, 03 1985. doi: 10.1214/aos/1176346597. URL https://doi.org/10.1214/aos/1176346597.

F. Famoye and K. P. Singh. On inflated generalized poisson regression models. *Advances and applications in statistics*, 3(2):145–158, 2003.

F. Famoye and K. P. Singh. Zero-inflated generalized poisson model with an application to domestic violence data. *Journal of Data Science*, 4:117–130, 2006.

Felix Famoye and Weiren Wang. Censored generalized Poisson regression model. *Comput. Statist. Data Anal.*, 46(3):547–560, 2004. ISSN 0167-9473. doi: 10.1016/j.csda.2003.08.007. URL https://doi.org/10.1016/j.csda.2003.08.007.

Jiarui Feng and Zhongyi Zhu. Semiparametric analysis of longitudinal zero-inflated count data. *J. Multivariate Anal.*, 102(1):61–72, 2011. ISSN 0047-259X. doi: 10.1016/j.jmva.2010.08.001. URL https://doi.org/10.1016/j.jmva.2010.08.001.

David Friendly, Michael; Meyer. *Discrete Data Analysis with R : Visualization and Modeling Techniques for Categorical and Count Data*. Chapman and Hall/CRC Texts in Statistical Science Ser. Chapman and Hall/CRC, 1st edition, 2015. ISBN 9781498725859,1498725856. URL https://www.crcpress.com/Discrete-Data-Analysis-with-R-Visualization-and-Modeling-Techniques-for/Friendly-Meyer/p/book/9781498725835.

Qiang Fu, Xin Guo, and Kenneth C. Land. A poisson-multinomial mixture approach to grouped and right-censored counts. *Communications in Statistics - Theory and Methods*, 47(2):427–447, 2018. doi: 10.1080/03610926.2017.1303736. URL https://doi.org/10.1080/03610926.2017.1303736.

Alan J. Lee(auth.) George A. F. Seber. *Linear Regression Analysis, Second Edition*. Wiley series in probability and statistics. Wiley-Interscience, 2 edition, 2003. ISBN 9780471415404,9780471722199. URL https://www.wiley.com/en-us/Linear+Regression+Analysis%2C+2nd+Edition-p-9780471415404.

Christian Gourieroux and Alain Monfort. Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. *Journal of Econometrics*, 17(1):83–97, 1981.

P. L. Gupta, R. C. Gupta, and R. C. Tripathi. Score test for zero inflated generalized poisson regression model. *Communications in Statistics - Theory and Methods*, 33:47–64, 2004.

Daniel B. Hall. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, 56(4):1030–1039, 2000. ISSN 0006-341X. doi: 10.1111/j.0006-341X.2000.01030.x. URL https://doi.org/10.1111/j.0006-341X.2000.01030.x.

Daniel B. Hall and Kenneth S. Berenhaut. Score tests for heterogeneity and overdispersion in zero-inflated Poisson and binomial regression models. *Canad. J. Statist.*, 30(3):415–430, 2002. ISSN 0319-5724. doi: 10.2307/3316145. URL https://doi.org/10.2307/3316145.

Daniel B. Hall and Zhengang Zhang. Marginal models for zero inflated clustered data. *Statistical Modelling*, 4:161–180, 2004.

Xuming He, Hongqi Xue, and Ning-Zhong Shi. Sieve maximum likelihood estimation for doubly semiparametric zero-inflated poisson models. *Journal of Multivariate Analysis*, 101(9):2026 – 2038, 2010. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2010.05.003. URL http://www.sciencedirect.com/science/article/pii/S0047259X10001107.

Arne Henningsen and Ott Toomet. maxlik: A package for maximum likelihood estimation in r. *Computational Statistics*, 26(3):443–458, Sep 2011. ISSN 1613-9658. doi: 10.1007/s00180-010-0217-1. URL https://doi.org/10.1007/s00180-010-0217-1.

Joseph M. Hilbe. *Negative Binomial Regression*. Cambridge University Press, 2nd edition, 2011. ISBN 978-0-52119-815-8.

Gul Inan, John Preisser, and Kalyan Das. A score test for testing a marginalized zero-inflated poisson regression model against a marginalized zero-inflated negative binomial regression model. *Journal of Agricultural, Biological and Environmental Statistics*, 23(1):113–128, Mar 2018. ISSN 1537-2693. doi: 10.1007/s13253-017-0314-5. URL https://doi.org/10.1007/s13253-017-0314-5.

Simon Jackman. *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory*. United States Studies Centre, University of Sydney, Sydney, New South Wales, Australia, 2017. URL https://github.com/atahk/pscl/. R package version 1.5.2.

Markus Jochmann. What belongs where? variable selection for zero-inflated count models with an application to the demand for health care. *Computational Statistics*, 28, 01 2009. doi: 10.1007/s00180-012-0388-z.

Norman Lloyd Johnson, Adrienne W. Kemp, and Samuel Kotz. *Univariate discrete distributions*. John Wiley & Sons, New York, 2005.

Dimitris Karlis, Purushottam Papatla, and Sudipt Roy. Finite mixtures of censored Poisson regression models. *Stat. Neerl.*, 70(2):100–122, 2016. ISSN 0039-0402. doi: 10.1111/stan.12079. URL https://doi.org/10.1111/stan.12079.

K. F. Lam, Hongqi Xue, and Yin Bun Cheung. Semiparametric analysis of zero-inflated count data. *Biometrics*, 62(4):996–1003, 2006. ISSN 0006-341X. doi: 10.1111/j.1541-0420.2006.00575.x. URL https://doi.org/10.1111/j.1541-0420.2006.00575.x.

Diane Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992. doi: 10.1080/00401706.1992.10485228. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1992.10485228.

Hwa Kyung Lim, Wai Keung Li, and Philip L. H. Yu. Zero-inflated Poisson regression mixture model. *Comput. Statist. Data Anal.*, 71:151–158, 2014. ISSN 0167-9473. doi: 10.1016/j.csda.2013.06.021. URL https://doi.org/10.1016/j.csda.2013.06.021.

D Leann Long, John Preisser, Amy Herring, and Carol Golin. A marginalized zero-inflated poisson regression model with overall exposure effects. *Statistics in medicine*, 33, 12 2014. doi: 10.1002/sim.6293.

D Leann Long, John Preisser, Amy Herring, and Carol Golin. A marginalized zero-inflated poisson regression model with random effects. *Journal of the Royal Statistical Society Series C Applied Statistics*, 04 2015. doi: 10.1111/rssc.12104.

Minggen Lu and Chin-Shang Li. Spline-based semiparametric estimation of a zero-inflated poisson regression single-index model. *Annals of the Institute of Statistical Mathematics*, 7 2015. ISSN 0020-3157. doi: 10.1007/s10463-015-0527-8.

T. M. Lukusa, S.-M. Lee, and C.-S. Li. Semiparametric estimation of a zero-inflated Poisson regression model with missing covariates. *Metrika*, 79(4):457–483, 2016.

Marie M. Mahmoud and Mahmoud M. Alderiny. On estimating parameters of censored generalized poisson regression model. *Applied Mathematical Sciences (Ruse)*, 4, 01 2010.

R. A. Maller. Asymptotics of regressions with stationary and nonstationary residuals. *Stochastic Process. Appl.*, 105(1):33–67, 2003. ISSN 0304-4149. doi: 10.1016/S0304-4149(02)00263-6. URL https://doi.org/10.1016/S0304-4149(02)00263-6.

Jacob Martin and Daniel B. Hall. Marginal zero-inflated regression models for count data. *J. Appl. Stat.*, 44(10):1807–1826, 2017. ISSN 0266-4763. doi: 10.1080/02664763.2016.1225018. URL https://doi.org/10.1080/02664763.2016.1225018.

P. McCullagh and J. A. Nelder. *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1989. ISBN 0-412-31760-5. doi: 10.1007/978-1-4899-3242-6. URL https://doi.org/10.1007/978-1-4899-3242-6. Second edition [of MR0727836].

Yongyi Min and Alan Agresti. Random effect models for repeated measures of zero-inflated count data. *Stat. Model.*, 5(1):1–19, 2005. ISSN 1471-082X. doi: 10.1191/1471082X05st084oa. URL https://doi.org/10.1191/1471082X05st084oa.

Abbas Moghimbeigi, Mohammed Reza Eshraghian, Kazem Mohammad, and Brian McArdle. Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. *J. Appl. Stat.*, 35(9-10):1193–1202, 2008. ISSN 0266-4763. doi: 10.1080/02664760802273203. URL https://doi.org/10.1080/02664760802273203.

Anthea Monod. Random effects modeling and the zero-inflated Poisson distribution. *Comm. Statist. Theory Methods*, 43(4):664–680, 2014. ISSN 0361-0926. doi: 10.1080/03610926.2013.814782. URL https://doi.org/10.1080/03610926.2013.814782.

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

Samuel M. Mwalili, Emmanuel Lesaffre, and Dominique Declerck. The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. *Stat. Methods Med. Res.*, 17(2):123–139, 2008. ISSN 0962-2802. doi: 10.1177/0962280206071840. URL https://doi.org/10.1177/0962280206071840.

John C. Nash. On best practice optimization methods in R. *Journal of Statistical Software*, 60 (2):1–14, 2014. URL http://www.jstatsoft.org/v60/i02/.

John C. Nash and Ravi Varadhan. Unifying optimization algorithms to aid software system users: optimx for R. *Journal of Statistical Software*, 43(9):1–14, 2011. URL http://www.jstatsoft.org/v43/i09/.

JA Nelder and R.W.M. Wedderburn. Generalized linear models. *J. R. Stat. Soc. Ser. A*, 19: 92–100, 01 1972. doi: 10.1007/978-1-4612-4380-9_39.

V.-T Nguyen and J.-F Dupuy. Modèles de régression à inflation de zéro et données censurées - application au recours aux soins de santé. *Biostatistiques et sciences de la santé*, 1(Numéro 1), 2019a. ISSN 2632-8291. doi: 10.21494/ISTE.OP.2019.0393. URL https://www.openscience.fr/Modeles-de-regression-a-inflation-de-zero-et-donnees-censurees-application-au.

Van Trinh Nguyen and Jean-François Dupuy. Asymptotic results in censored zero-inflated poisson regression. *Communications in Statistics - Theory and Methods*, 0(0):1–21, 2019b. doi: 10.1080/03610926.2019.1676442. URL https://doi.org/10.1080/03610926.2019.1676442.

Steven D. Pizer and Julia C. Prentice. Time is money: Outpatient waiting times and health insurance choices of elderly veterans in the united states. *Journal of Health Economics*, 30 (4):626–636, 2011.

Jonathan P. Prasad. *Zero-inflated censored regression models: an application with episode of care data*. http://scholarsarchive.byu.edu/etd/2226, 2009.

Martin Ridout, John Hinde, and Clarice G. B. Demétrio. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, 57 (1):219–223, 2001. ISSN 0006-341X. doi: 10.1111/j.0006-341X.2001.00219.x. URL https://doi.org/10.1111/j.0006-341X.2001.00219.x.

Seyed Ehsan Saffari and Robiah Adnan. Zero-inflated Poisson regression models with right censored count data. *Matematika (Johor Bahru)*, 27(1):21–29, 2011. ISSN 0127-8274.

Seyed Ehsan Saffari, Robiah Adnan, and William Greene. Investigating the impact of excess zeros on hurdle-generalized poisson regression model with right censored count data. *Statistica Neerlandica*, 67(1):67–80, 2013. doi: 10.1111/j.1467-9574.2012.00532.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.2012.00532.x.

Nazmi Sari. Physical inactivity and its impact on healthcare utilization. *Health Economics*, 18 (8):885–901, 2009.

Sisira Sarma and Wayne Simpson. A microeconometric analysis of canadian health care utilization. *Health Economics*, 15(3):219–239, 2006.

Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 03 1978. doi: 10.1214/aos/1176344136. URL https://doi.org/10.1214/aos/1176344136.

Kevin E. Staub and Rainer Winkelmann. Consistent estimation of zero-inflated count models. *Health Economics*, 22(6):673–686, 2013.

Andrew Street, Andrew Jones, and Aya Furuta. Cost-sharing and pharmaceutical utilisation and expenditure in russia. *Journal of Health Economics*, 18(4):459–472, 1999.

R Core Team. R: A language and environment for statistical computing. 2018. doi: RFoundationforStatisticalComputingVienna,Austria,https://www.R-project.org/.

Joseph V. Terza. A tobit-type estimator for the censored poisson regression model. *Economics Letters*, 18:0–365, 1985. ISSN 0165-1765. doi: 10.1016/0165-1765(85)90053-9. URL /scimag/10.1016%2F0165-1765%2885%2990053-9.

Guo-Liang Tian, Huijuan Ma, Yong Zhou, and Dianliang Deng. Generalized endpoint-inflated binomial model. *Comput. Statist. Data Anal.*, 89:97–114, 2015. ISSN 0167-9473. doi: 10.1016/j.csda.2015.03.009. URL https://doi.org/10.1016/j.csda.2015.03.009.

David Todem, KyungMann Kim, and Wei-Wen Hsu. Marginal mean models for zero-inflated count data. *Biometrics*, 72(3):986–994, 2016. ISSN 0006-341X. doi: 10.1111/biom.12492. URL https://doi.org/10.1111/biom.12492.

Quang Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57:307–33, 02 1989. doi: 10.2307/1912557.

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

Z. Wang. mpath: Regularized linear models, r package version 0.3-7. 2019. doi: https://CRAN.R-project.org/package=mpath.

Zhu Wang, Shuangge Ma, Ching-Yun Wang, Michael Zappitelli, Prasad Devarajan, and Chirag Parikh. Em for regularized zero inflated regression models with applications to postoperative morbidity after cardiac surgery in children. *Statistics in Medicine*, 33, 12 2014. doi: 10.1002/sim.6314.

Zhu Wang, Shuangge Ma, and Ching-Yun Wang. Variable selection for zero-inflated and overdispersed data with application to health care demand in germany. *Biometrical Journal*, 57, 06 2015. doi: 10.1002/bimj.201400143.

R. Winkelmann. *Econometric Analysis of Count Data*. Springer Berlin Heidelberg, 2013.

Liming Xiang, Andy H. Lee, Kelvin K. W. Yau, and Geoffrey J. McLachlan. A score test for overdispersion in zero-inflated poisson mixed regression model. *Statistics in Medicine*, 26(7):1608–1622, 2007.

Feng-Chang Xie and Bo-Cheng Wei. Diagnostics analysis in censored generalized poisson regression model. *Journal of Statistical Computation and Simulation*, 77(8):695–708, 2007. doi: 10.1080/10629360600581316. URL https://doi.org/10.1080/10629360600581316.

Feng-Chang Xie, Bo-Cheng Wei, and Jin-Guan Lin. Score tests for zero-inflated generalized poisson mixed regression models. *Computational Statistics & Data Analysis*, 53(9):3478–3489, 2009.

Hung-Wen Yeh, Byron Gajewski, Purna Mukhopadhyay, and Fariba Behbod. The zero-truncated poisson with right censoring: An application to translational breast cancer research. *Statistics in biopharmaceutical research*, 4:252–263, 08 2012. doi: 10.1080/19466315.2011.636279.

Steven T. Yen, Chao-Hsiun Tang, and Shew-Jiuan B. Su. Demand for traditional medicine in taiwan: a mixed gaussian Poisson model approach. *Health Economics*, 10(3):221–232, 2001.

Achim Zeileis, Christian Kleiber, and Simon Jackman. Regression models for count data in r. *Journal of Statistical Software, Articles*, 27(8):1–25, 2008. ISSN 1548-7660. doi: 10.18637/jss.v027.i08. URL https://www.jstatsoft.org/v027/i08.

Ping Zeng, Yongyue Wei, Yang Zhao, Jin Liu, Liya Liu, Ruyang Zhang, Jianwei Gou, Shuiping Huang, and Feng Chen. Variable selection approach for zero-inflated count data via adaptive lasso. *Journal of Applied Statistics*, 41, 04 2014. doi: 10.1080/02664763.2013.858672.

Huirong Zhu, Sheng Luo, and Stacia M DeSantis. Zero-inflated count models for longitudinal measurements with heterogeneous random effects. *Statistical Methods in Medical Research*, 26(4):1774–1786, 2017.

INSA de RENNES
Service des Formations

AVIS DU JURY SUR LA REPRODUCTION DE LA THESE SOUTENUE

Titre de la thèse:
Modèles de régression pour données de comptage zéro-inflatées en présence de censure. Applications en économie de la santé

Nom Prénom de l'auteur : NGUYEN VAN TRINH

Membres du jury :
- Madame GARES Valérie
- Madame DABO-NIANG Sophie
- Monsieur DUPUY Jean-François
- Madame YAO Anne-Françoise
- Monsieur RAULT Christophe

Président du jury : Monsieur Christophe RAULT

Date de la soutenance : 16 Mars 2020

Reproduction de la these soutenue

☒ Thèse pouvant être reproduite en l'état
☐ Thèse pouvant être reproduite après corrections suggérées

Fait à Rennes, le 16 Mars 2020

Signature du président de jury

Le Directeur,

M'hamed DRISSI

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

INSA RENNES

Titre : Modèles de régression pour données de comptage zéro-inflatées en présence de censure. Applications en économie de la santé.

Mot clés : Propriétés asymptotiques, données de comptage, excès de zéros, modèle marginal, simulations, applications en économie de la santé.

Résumé : L'excès de zéros dans les données de comptage est une situation qui survient dans de nombreux domaines (épidémiologie, économie de la santé...). . Les modèles de régression à inflation de zéro fournissent un outil puissant pour analyser ce type de données. Un problème supplémentaire est celui de la censure de la variable réponse. La censure à droite, en particulier, correspond à la situation où l'on observe seulement une borne inférieure sur le comptage considéré. L'analyse statistique de données de comptage en présence d'inflation de zéros et de censure constitue le thème général de ce travail de thèse. Dans une première contribution, nous étudions les propriétés théoriques et numériques de l'estimateur du maximum de vraisemblance dans le modèle de régression de Poisson à inflation de zéros. Nous décrivons également une application de ce modèle sur des données issues du domaine de l'économie de la santé. Le modèle de Poisson à inflation de zéros suppose que toute la surdispersion des données peut être expliquée par un excès de zéros. Lorsqu'une surdispersion supplémentaire est présente, on peut utiliser des modèles alternatifs, tels que les modèles de régression de Poisson généralisé à inflation de zéros (modèle ZIGP) et binomial négatif à inflation de zéros (modèle ZINB). Dans une deuxième contribution, nous nous intéressons aux propriétés de l'estimateur du maximum de vraisemblance dans les modèles ZIGP et ZINB en présence d'une variable réponse censurée. Nous étudions ces propriétés au moyen de simulations. Puis nous proposons une procédure de sélection simple à mettre en oeuvre. Nous illustrons cette méthodologie sur des données issues du domaine de l'économie de la santé. Dans une troisième contribution, nous nous intéressons à l'inférence statistique dans le modèle de régression de Poisson à inflation de zéros marginal (modèle MZIP), en présence d'une variable réponse censurée. Le modèle MZIP quantifie l'effet des variables explicatives directement sur la moyenne de la loi marginale du comptage (et non sur la moyenne de la loi du comptage pour les individus "susceptibles"). Nous établissons les propriétés asymptotiques de l'estimateur du maximum de vraisemblance dans ce cadre et réalisons une étude de simulations. Enfin, nous illustrons le modèle sur un jeu de données.

Title: Zero-inflated count regression models with censoring. Applications in health economy.

Keywords: Asymptotic properties, count data, excess of zeros, marginal model, simulations, health care utilization.

Abstract: Count data with excess of zeros commonly occur in various areas, such as epidemiology and health economics. Zero-inflated regression models provide a useful tool to analyse such data. A further problem arises when these counts are right-censored, which also often occurs in observational studies. Right-censoring refers to the situation where one only observes a lower bound on the count of interest. This dissertation aims at investigating the statistical inference in zero-inflated count regression models with right-censored data. First, we investigate the properties of the maximum likelihood estimator (MLE) in the zero-inflated Poisson (ZIP) regression model with randomly right-censored counts. We establish the asymptotics of the MLE in this setting. A thorough simulation study is also conducted to assess finite-sample behavior of the MLE. Finally, we describe an application to a dataset from health economy. Overdispersion is entirely caused by zero-inflation. When additional overdispersion is present, useful alternatives to ZIP are given by the zero-inflated generalized Poisson (ZIGP) and zero-inflated negative binomial (ZINB) models. In a second contribution, we investigate the properties of the MLE in ZIGP and ZINB regression models, when the count response is subject to right-censoring. Then, we propose a simple methodology for variable selection. The proposed methods are illustrated on a dataset in the field of health economy. Finally, we investigate statistical inference in the marginal zero-inflated Poisson (MZIP) model, when the count response is randomly right-censored. The MZIP modelt provides an appealing alternative for interpreting covariate effects at the population level. We establish asymptotic of the MLE in the censored MZIP model and conduct simulations to assess finite-sample properties. Finally, we describe an application to a real dataset from health economy.