ASYMPTOTIC PROPERTIES IN THE PROBIT-ZERO-INFLATED BINOMIAL REGRESSION MODEL

LO Fatimata, BA Demba Bocar[1], and DIOP Aba

ABSTRACT. Zero-inflated regression models have had wide application recently and have provenuseful in modeling data with many zeros. Zero-inflated Binomial (ZIB) regression model is an extension of the ordinary binomial distribution that takes into account the excess of zeros. In comparing the probit model to the logistic model, many authors believe that there is little theoretical justification in choosing one formulation over the other in most circumstances involving binary responses. The logit model is considered to be computationally simpler but it is based on a more restrictive assumption of error independence, although many other generalizations have dealt with that assumption as well. By contrast, the probit model assumes that random errors have a multivariate normal distribution. This assumption makes the probit model attractive because the normal distribution provides a good approximation to many other distributions. In this paper, we develop a maximum likelihood estimation procedure for the parameters of a zero-inflated Binomial regression model with probit link function for both component of the model. We establish the existency, consistency and asymptotic normality of the proposed estimator.

1. INTRODUCTION

Zero-inflated regression models mix a degenerate distribution with point mass of one at zero with a simple regression model based on a standard distribution. Zero-infation can happen both with discrete data and continuous data

---

(see [24], [20] and [17]). Here, we concentrate on the discrete case, specialy on Zero-inflated Binomial (ZIB) regression model. [20] considered Zero-Inflated Poisson (ZIP) models and also extended Lambert's methodology to an upper bounded count situation, thereby obtaining a ZIB model. He also incorporated random intercepts into these models and used them to model insect counts in a horticulture experiment. The ZIB regression model assumes the data are from two states: a true zero state where the response always equals zero and a non zero state where the response follows a Binomial distribution. Suppose we have the response vector $y = (y_1, ..., y_n)'$ where y_i is the observed value of a random variable Y_i and $n \in \mathbb{N}^*$ is the sample size. In the ZIB regression model we assume

$$(1.1) \qquad Y_i \sim \begin{cases} 0 & \text{with probability } p_i \\ \mathcal{B}(n_i, \pi_i) & \text{with probability } 1 - p_i, \end{cases}$$

where y_i is the number of successes out of n_i trials and π_i is the probability of success for the subject i. Again, the parameters $p = (p_1, \ldots, p_n)'$ and $\pi_i = (\pi_1, \ldots, \pi_n)'$ are modeled through *logit* link generalized linear models as:

$$(1.2) \qquad \begin{aligned} \log\left[\frac{p_i}{1-p_i}\right] &= \theta_1 + \theta_2 \mathbf{z}_{i2} + \ldots + \theta_p \mathbf{z}_{iq} := \theta' \mathbf{z}_i, \\ \log\left[\frac{\pi_i}{1-\pi_i}\right] &= \beta_1 + \beta_2 \mathbf{x}_{i2} + \ldots + \beta_p \mathbf{x}_{ip} := \beta' \mathbf{x}_i. \end{aligned}$$

Here, β and θ are regression parameters and $\mathbf{X}$ and $\mathbf{Z}$ are corresponding design matrices. [13] studied the asymptotics of the model 1.1 by considering the logit link function for the two components of the model (binomial and zero inflation). They established the consistency and asymptotic normality of the maximum likelihood estimator. [33] applied the ZIB model estimated by a quasi-likelihood method to a biological control assay. This model was initially built to model overdispersion generated by individual variability in the probability of success.

For other types of discrete outcomes, such as binary, multinomial or ordinal, various single value inflated models were developed, including: zero-inflated Bernoulli model ([16]), zero-inflated binomial model ([20] and [33]), zero-inflated ordered probit model ([21]), baseline or zero inflated multinomial logit model ([2], [14]), and middle category inflated ordered model ([3]). Similar extension has been made to incorporate inflation other than zero for multinomial or ordinal outcomes ([32]). The use of noncanonical link functions

is not prohibited by the fact that they are more computationally complex. [10] argued that, in some applications, the overall fit of the model as measured by the p-value of the goodness-of-fit statistics can be improved significantly by the use of a non-canonical link.

We use the probit link function for both binomial and excess of zero components for the ZIB regression model 1.1. The aim is to evaluate the misspecification problem by mistakenly choosing the appropriate link function. We do not introduce the probit model as a rival to the logistic model, but rather as an alternative, in particular in the case of excess of zero in the sample. Experience shows that in most situations the two approaches produce similar results although some differences exist. This similarity is not necessarily sustained when multivariate responses are used. Further research is needed to investigate the advantages or disadvantages in using one model over the other in data mining applications.

The use of the probit regression model dates back to [4]. Many other authors have used the probit model in other applications with success; for example, [31] compared the two models for estimating the strength of gear teeth. [30] used probit link function for data mining applications. [19] applied probit analysis in toxicological experiments. The probit model has also found popularity in economics. [9] provides a survey of the early origins of the model.

The rest of this paper is organized as follows. In Section 2, we describe the problem of ZIB regression model with probit link function, we propose an estimation method adapted to this setting. In Section 3 we establish the asymptotic propreties of the proposed estimator under some regularity conditions. A discussion and some perspectives are given in Section 4.

2. ZIB REGRESSION MODEL WITH PROBIT LINK FUNCTION

2.1. **The model set-up and estimation procedure.** Let $(Y_1, \mathbf{X}_1, \mathbf{W}_1), \ldots,$ $(Y_n, \mathbf{X}_n, \mathbf{W}_n)$ be independent and identically distributed copies of the random vector $(Y, \mathbf{X}, \mathbf{W})$ defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. For every individual $i = 1, \ldots, n$, Y_i is the number of successes out of n_i trials and π_i is the probability of success for the subject i. Let $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})^\top$ and

$\mathbf{W}_i = (1, Z_{i2}, \ldots, W_{iq})'$ be random vectors of predictors or covariates (both categorical and continuous predictors are allowed). We shall assume in the following that the $\mathbf{X}_i$'s are related to the zero inflation, while the $\mathbf{W}_i$'s are related to the binomial component. $\mathbf{X}_i$ and $\mathbf{W}_i$ are allowed to share some components. The ZIB regression model with *probit* link function is given by

$$(2.1) \qquad Y_i \sim \begin{cases} 0 & \text{with probability } p_i \\ \mathcal{B}(n_i, \pi_i) & \text{with probability } 1 - p_i \end{cases},$$

where the parameters $p = (p_1, \ldots, p_n)'$ and $\pi_i = (\pi_1, \ldots, \pi_n)'$ are modeled through *probit* link generalized linear models as:

$$(2.2) \qquad \begin{aligned} probit(p_i) &= F(\beta_1 + \beta_2 \mathbf{x}_{i2} + \ldots + \beta_p \mathbf{x}_{iq}) := F(\beta' \mathbf{x}_i), \\ probit(\pi_i) &= F(\mu_1 + \mu_2 \mathbf{w}_{i2} + \ldots + \mu_p \mathbf{w}_{ip}) := F(\mu' \mathbf{w}_i), \end{aligned}$$

where F is the cumulative distribution function of the standard normal distribution.

The model (2.1) can be rewritten as follows

$$(2.3) \qquad \mathbb{P}(Y = y_i) = \begin{cases} \pi_i + (1 - \pi_i)(1 - p_i)^{n_i} & \text{if } y_i = 0 \\ (1 - \pi_i)C_{n_i}^{y_i} p_i^{y_i}(1 - p_i)^{n_i - y_i} & \text{if } y_i \in \{1, \ldots, n\} \end{cases}.$$

Let $\theta := (\beta', \mu')'$ denote the unknown k-dimensional ($k = p + q$) parameter in the conditional distribution of Y given $\mathbf{X}_i$ and $\mathbf{W}_i$. Now, the likelihood for θ from the independent sample $(Y_i, \mathbf{X}_i, \mathbf{W}_i)$ $(i = 1, \ldots, n)$ is as follows:

$$(2.4) \qquad \begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} \mathbb{P}(Y_i = k_i | \mathbf{X}_i = \mathbf{x}_i, \mathbf{W}_i = \mathbf{w}_i) \\ &= \prod_{i=1}^{n} (F(\mu' \mathbf{w}_i) + (1 - F(\mu' \mathbf{w}_i))(1 - F(\beta' \mathbf{x}_i))^{n_i})^{k_i} \\ &\quad \times ((1 - F(\mu' \mathbf{w}_i))C_{n_i}^{Y_i} F(\beta' \mathbf{x}_i)(1 - F(\beta' \mathbf{x}_i))^{n_i - Y_i})^{1 - k_i}, \end{aligned}$$

where $k_i = \mathbf{1}_{\{Y_i = 0\}}$.

We define the maximum likelihood estimator $\widehat{\theta}_n := (\widehat{\beta}'_n, \widehat{\mu}'_n)'$ of θ as the solution (if it exists) of the k-dimensional score equation

$$(2.5) \qquad \dot{l}_n(\theta) = \frac{\partial l_n(\theta)}{\partial \theta} = 0,$$

where $l_n(\theta) := \log L_n(\theta)$ is the log-likelihood function. In the following, we shall be interested in the asymptotic properties of the maximum likelihood estimator

$\widehat{\theta}_n$. We will however obtain consistency and asymptotic normality results for the whole $\widehat{\theta}_n$. Before proceeding, we need to set some further notations.

2.2. **Some further notations.** Define first the $(p \times n)$ and $(q \times n)$ matrices

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

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

$$
\mathbb{Z} = \begin{bmatrix} \mathbb{X} & 0_{pn} \\ 0_{qn} & \mathbb{W} \end{bmatrix},
$$

where 0_{ab} denotes the $(a \times b)$ matrix whose components are all equal to zero (for any positive integer values a, b). Let $k_i(\theta) = F(\mu'\mathbf{w}_i)(h_i(\beta))^{-n_i+1} + (1 - F(\mu'\mathbf{w}_i)h_i(\beta)$, $\forall i = 1,\dots,n$ where $h_i(\beta) = 1 - F(\beta'\mathbf{x}_i)$. Let also $C(\theta) = (C_j(\theta))_{1 \le j \le 2n}$ be the 2n-dimensional column vector defined as

$$
C(\theta) = (A_1(\theta),\dots,A_n(\theta); B_1(\theta),\dots,B_n(\theta))', \quad i = 1,\dots,n,
$$

where

$$
A_i(\theta) = \frac{k_i\mathbf{w}_i f(\mu'\mathbf{w}_i)h_i(\beta)^{-n_i+1}F(\beta'\mathbf{x}_i)^{n_i}}{k_i(\theta)} - (1 - k_i)\frac{\mathbf{w}_i f(\mu'\mathbf{w}_i)}{1 - F(\mu'\mathbf{w}_i)},
$$

$$
B_i(\theta) = \frac{-k_i n_i \mathbf{x}_i f(\beta'\mathbf{x}_i)}{F(\mu'\mathbf{w}_i)k_i(\theta)} + (1 - k_i)\frac{\mathbf{x}_i f(\beta'\mathbf{x}_i)[F_i - n_i F(\beta'\mathbf{x}_i)]}{F(\beta'\mathbf{x}_i)h_i(\beta)},
$$

where $f(.)$ is the density function associate to $F(.)$.

Then, simple algebra shows that the score equation can be rewritten as

$$
\dot{l}_n(\theta) = ZC(\theta) = 0.
$$

If $H = (H_{ij})_{1 \le i \le a, 1 \le j \le b}$ denotes some $(a \times b)$ matrix, we will denote by $H_{\bullet j}$ it's j-th column $(j = 1,\dots,b)$ that is, $H_{\bullet j} = (H_{1j},\dots,H_{aj})'$. Then, it will be useful to rewrite the score vector as

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

We shall further note $\ddot{l}_n(\theta)$ the $(k \times k)$ matrix of second derivatives of $l_n(\theta)$, that is $\ddot{l}_n(\theta) = \frac{\partial^2 l_n(\theta)}{\partial\theta\partial\theta'}$. Let $\mathbf{D}(\theta) = (\mathbf{D}_{ij}(\theta))_{1 \leq i,j \leq 2n}$ be the $(2n \times 2n)$ block matrix defined as

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

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

$\mathbf{D}_{1,ii}(\theta)$

$$= \frac{k_i(\theta)w_i^2 h_i(\beta)^{-n_i+1}F(\beta'\mathbf{x}_i)[f'(\mu'\mathbf{w}_i)k_i(\theta) - f^{(2)}(\beta\mathbf{x}_i)(h_i(\beta)^{-n_i+1} - h_i(\beta)]}{k_i^2(\theta)}$$

$$- \mathbf{w}_i^2(1 - k_i(\theta))\frac{[f'(\mu\mathbf{w}_i)(1 - F(\mu'\mathbf{w}_i)) + f^{(2)}(\mu'\mathbf{w}_i)]}{(1 - F(\mu'\mathbf{w}_i))^2}$$

$\mathbf{D}_{2,ii}(\theta)$

$$= \frac{-k_i(\theta)n_i\mathbf{x}_i}{F(\mu'\mathbf{w}_i}[x_i f(\beta\mathbf{x}_i)k_i(\theta) - (n_i + 1)x_i^2 f^{(2)}(\beta'\mathbf{x}_i)(F(\mu'\mathbf{w}_i)h_i(\beta)^{n_i} - (1 - F(\mu'\mathbf{w}_i)))]$$

$$+ (1 - k_i(\theta))\frac{[\mathbf{x}_i^2 F(\beta'\mathbf{x}_i)h_i(\beta)(\mathbf{z}_i f'(\beta'\mathbf{x}_i) - n_i f'(\beta'\mathbf{x}_i)F(\beta'\mathbf{x}_i)}{(F(\beta'\mathbf{x}_i)h_i(\beta))^2} - (1 - k_i(\theta))$$

$$\cdot \frac{n_i f^{(2)}(\beta'\mathbf{x}_i)) + \mathbf{x}_i f(\beta'\mathbf{x}_i)(\mathbf{z}_i - n_i F(\beta'\mathbf{x}_i)(x_i f(\beta'\mathbf{x}_i)h_i(\beta) - x_i f(\beta'\mathbf{x}_i)F(\beta'\mathbf{x}_i))}{(F(\beta'\mathbf{x}_i)h_i(\beta))^2}$$

$\mathbf{D}_{3,ii}(\theta)$

$$= \frac{-k_i\mathbf{w}_i f(\mu'\mathbf{w}_i)k_i(\theta)[(-n_i + 1)\mathbf{x}_i h_i(\beta)^{-n_i}F(\beta'\mathbf{x}_i) + \mathbf{x}_i f(\beta'\mathbf{x}_i)h_i(\beta)^{-n_i+1})}{k_i^2(\theta)}$$

$$- \frac{k_i'(\theta)h_i(\beta)^{-n_i+1}F(\beta'\mathbf{x}_i)]}{k_i^2(\theta)}.$$

Then, some algebra shows that $\ddot{l}_n$ can be expressed as

$$\ddot{l}_n(\theta) = -\mathbb{Z}\mathbb{D}(\theta)\mathbb{Z}'.$$

Note that the size of $C(\theta), \mathbb{Z}$ and $\mathbf{D}(\theta)$ depends on n. However, in order to simplify notations, n will not be used as a lower index for these vector and matrices. In the next section, we turn to the asymptotic theory for the proposed estimator.

3. ASYMPTOTIC THEORY

We first state some regularity conditions that will be needed to study the asymptotic theory.

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

C2. Let $(\beta_0', \mu_0')'$ denote the true parameter value. β_0 and μ_0 lie in the interior of known compact sets $\mathcal{G} \subset \mathbb{R}^p$ and $\mathcal{H} \subset \mathbb{R}^q$ respectively.

C3. The Hessian matrix $\ddot{l}_n(\theta)$ is negative definite and of full rank, for every $n = 1, 2, \ldots$ Let λ_n and Λ_n be respectively the smallest and largest eigenvalues of $\mathbb{Z}\mathbf{D}(\theta_0)\mathbb{Z}'$. There exists a finite positive constant c_2 such that $\Lambda_n/\lambda_n < c_2$ for every $n = 1, 2, \ldots$

C4. Let A denote a $k \times k$ matrix such that $\|A\| = \max_{1 \leq i \leq k} |\lambda_i|$.

C5. The function f is continuous and twice differentiable.

Now we establish rigorously the existence, consistency and asymptotic normality of the maximum likelihood estimator $\widehat{\theta}_n$ in model (2.1). We prove the following results:

Theorem 3.1 (Existence and consistency)**.** *Under the conditions **C1**-**C5**, The maximum likelihood estimator $\widehat{\theta}_n$ exists almost surely as $n \to \infty$ and converges almost surely to θ_0, if and only if λ_n tends to infinity as $n \to \infty$.*

Proof. The following lemma essentially provides an intermediate technical result (see [13] for its proof). □

Lemma 3.1. *Let $\phi_n(\theta) : \mathbb{R}^k \longrightarrow \mathbb{R}^k$ be defined as*

$$\phi_n(\theta) = \theta + (\mathbb{Z}\mathbf{D}(\theta_0)\mathbb{Z}')^{-1}\dot{l}_n(\theta).$$

Then there exists an open ball $B(\theta_0, r) = \{\theta \in \mathbb{R}^k; \|\theta - \theta_0\| < r\}$ (with $r > 0$) such that ϕ_n satisfies the Lipschitz condition on $B(\theta_0, r)$ that is,

$$\|\phi_n(\theta) - \phi_n(\tilde{\theta})\| \leq c\|\theta - \tilde{\theta}\| \text{ for all } \theta, \tilde{\theta} \in B(\theta_0, r),\ 0 < c < 1.$$

Proof. Let γ_n be defined as $\gamma_n(\theta) = \theta - \phi_n(\theta) = -(\mathbb{Z}\mathbf{D}(\theta)\mathbb{Z}')\dot{l}_n(\theta)$. Then $\gamma_n(\theta_0)$ converges almost surely to 0 as $n \to \infty$. Note that

$$\gamma_n(\theta_0) = \left(\frac{1}{n}\ddot{l}_n(\theta_0)\right)^{-1} \times \left(\frac{1}{n}\dot{l}_n(\theta_0)\right).$$

The condition C3 implies that, $(\frac{1}{n}\ddot{l}_n(\theta_0))^{-1}$ converges to a matrix $\sum$. Then,

$$\frac{1}{n}\dot{l}_n(\theta_0) = \frac{1}{n}\mathbb{Z}C(\theta_0) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_{i1}A_i(\theta_0) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n X_{ip}A_i(\theta_0) \\ \frac{1}{n}\sum_{i=1}^n W_{i1}B_i(\theta_0) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n W_{iq}B_i(\theta_0) \end{pmatrix}$$

converges almost surely to 0 as $n \to \infty$. Indeed: for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$ we have

$$\mathbb{E}\left[X_{ij}A_i(\theta_0)\right] = \mathbb{E}\left[\mathbb{E}(X_{ij}A_i(\theta_0)|\mathbf{X}_i, \mathbf{W}_i)\right] = \mathbb{E}\left[X_{ij}\mathbb{E}(A_i(\theta_0)|\mathbf{X}_i, \mathbf{W}_i)\right],$$

where

$$\mathbb{E}(A_i(\theta_0)|\mathbf{X}_i, \mathbf{W}_i)$$
$$= \mathbb{E}[k_i \times \frac{\mathbf{w}_i f(\mu'\mathbf{w}_i)h_i(\beta)^{-n_i}F(\beta'\mathbf{x}_i)}{k_i(\theta)} - (1-k_i)\frac{\mathbf{w}_i f(\mu'\mathbf{w}_i)}{1-F(\mu'\mathbf{w}_i)}|\mathbf{X}_i, \mathbf{W}_i]$$
$$= \frac{w_i f(w_i\mu)h_i(\beta)^{-n_i}F(x_i\beta)}{k_i(\theta)}\mathbb{E}[k_i|\mathbf{X}_i, \mathbf{W}_i] - \mathbb{E}[(1-k_i)|\mathbf{X}_i, \mathbf{W}_i]\frac{w_i f(w_i\mu)}{1-F(w_i\mu)}|\mathbf{X}_i, \mathbf{W}_i].$$

We have

$$\begin{aligned} E(J_i|\mathbf{X}_i, \mathbf{W}_i) &= \mathbb{P}(Y_i = 0|\mathbf{X}_i, \mathbf{W}_i) \\ &= p_i + (1-\pi_i)^{m_i}(1-p_i) \\ &= F(\beta'\mathbf{x}_i) + (1-F(\mu'\mathbf{w}_i)^{-n_i}(1-F(\beta'\mathbf{x}_i) \\ &= F(\beta'\mathbf{x}_i) + (1-F(\mu'\mathbf{w}_i)^{-n_i}h_i(\beta), \end{aligned}$$

with

$$\mathbb{E}[(1-J_i)\mathbb{Z}_i|\mathbf{X}_i, \mathbf{W}_i] = n_i(1-p_i)\pi_i = n_i F(\mu'\mathbf{w}_i)h_i(\beta).$$

Then, we obtain $\mathbb{E}(A_i(\theta_0)|\mathbf{X}_i, \mathbf{W}_i) = 0$. Similarly, we show that $\mathbb{E}(B_i(\theta_0)|\mathbf{X}_i, \mathbf{W}_i) = 0$. Moreover, we have

$$
\begin{aligned}
&\mathbb{V}ar(W_{iq}B_i(\theta_0)) \\
&= \mathbb{E}(\mathbb{V}ar(W_{iq}B_i(\theta_0))|\mathbf{X}_i, \mathbf{W}_i)] + \mathbb{V}ar[\mathbb{E}(W_{iq}B_i(\theta_0))|\mathbf{X}_i, \mathbf{W}_i] \\
&= \mathbb{E}_{iq}^2[\mathbb{V}ar(B_i(\theta_0)|\mathbf{X}_i, \mathbf{W}_i)] \\
&= \mathbb{E}_{iq}^2(\frac{n_i\mathbf{x}_i f(\beta'\mathbf{x}_i)}{F(\mu'\mathbf{w}_i)k_i(\theta)} + \frac{\mathbf{x}_i f'(\beta'\mathbf{x}_i)\mathbf{z}_i - n_i F(\beta'\mathbf{x}_i))}{F(\beta'\mathbf{x}_i)h_i(\beta)})]^2 \mathbb{V}ar(k_i|\mathbf{X}_i, \mathbf{W}_i)] \\
&\leq \mathbb{E}_{ik}^2(\frac{n_i\mathbf{x}_i f(\beta'\mathbf{x}_i)}{F(\mu'\mathbf{w}_i)k_i(\theta)} + \frac{\mathbf{x}_i f'(\beta'\mathbf{x}_i)\mathbf{z}_i - n_i F(\beta'\mathbf{x}_i))}{F(\beta'\mathbf{x}_i)h_i(\beta)})^2 = c.
\end{aligned}
$$

The conditions C1, C2 and C3 implie that $c < \infty$, then

$$
\sum_{i=1}^n \frac{\mathbb{V}ar(A_i(\theta_0))}{i^2} \leq c \sum_{i=1}^n \frac{1}{i^2} \leq \infty.
$$

Kolmogorov's strong law of large numbers ensures that, as $n \to \infty$,

$$
\frac{1}{n}\sum_{i=1}^n \{X_{ij}A_i(\theta_0) - E(X_{ij}A_i(\theta_0))\} = \frac{1}{n}\sum_{i=1}^n A_i(\theta_0)X_{ij} \to 0
$$

Finaly, as $n \to \infty$ $\frac{1}{n}\dot{l}_n(\theta_0)$ and $\gamma_n(\theta_0)$ converge almost surely to 0. By using similar arguments we proove that, $\frac{1}{n}\sum_{i=1}^n W_{iq}B_i(\theta_0)$ converges almost surely to 0 as $n \to \infty$.

Let $\epsilon > 0$, the almost surely convergence of $\gamma_n(\theta_0)$ implies then for almost every $\omega \in \Omega$, there exists an integer value $n(\epsilon, \omega)$ such that fo any $n \geq n(\epsilon, \omega)$, $\|\gamma_n(\theta_0)\| \leq \epsilon$ or equivalently, $0 \in B(\gamma_n(\theta_0), \epsilon)$. In particulary, let $\epsilon = (1-c)s$ with $0 < c < 1$ such as in Lemma 3.1. Since γ_n satisfies the Lipschitz condition, the Lemma 3.1 ensures that there exists an element of $B(\theta_0, s)$ (let denote this element by $\widehat{\theta}_n$) such that $\gamma_n(\widehat{\theta}_n) = 0$, that is

$$
(\mathbb{Z}\mathbf{D}(\theta_0)\mathbb{Z}')^{-1} \times \dot{l}_n(\widehat{\theta}_n) = 0.
$$

The condition C3 implies that $\dot{l}_n(\widehat{\theta}_n) = 0$ and that $\widehat{\theta}_n$ is the unique maximizer of l_n. To summarize, we have shown that for almost every $\omega \in \Omega$ and for every $s > 0$, there exists an integer value $n(s, \omega)$ such that if $n \geq n(s, \omega)$ then the maximum likelihood estimator $\widehat{\theta}_n$ exists, and $\|\widehat{\theta}_n - \theta_0\| \leq s$ (that is, $\widehat{\theta}_n$ converges almost surely to θ_0).

This concludes the proof. $\square$

We now turn to the convergence in distribution of the proposed estimator, which is stated by the following theorem:

Theorem 3.2 (Asymptotic normality)**.** *Assume that the conditions **C1**-**C5** hold and that $\widehat{\theta}_n$ converges almost surely to θ_0. Let $\widehat{\sum}_n = \mathbb{Z}\mathbf{D}(\widehat{\theta}_n)\mathbb{Z}'$ and I_k denote the identity matrix of order k. Then $\widehat{\Sigma}_n^{\frac{1}{2}}(\widehat{\theta}_n - \theta_0)$ converges in distribution to the Gaussian vector $\mathcal{N}(0, I_k)$.*

Proof. A Taylor expansion of the score function is as

$$0 = \dot{l}_n(\widehat{\theta}_n) = \dot{l}_n(\theta_0) + \ddot{l}_n(\tilde{\theta}_n)(\widehat{\theta}_n - \theta_0)$$

where $\tilde{\theta}_n$ lies between $\widehat{\theta}_n$ and θ_0, and thus $\dot{l}_n(\theta_0) = -\ddot{l}_n(\tilde{\theta}_n)(\widehat{\theta}_n - \theta_0)$. Let $\tilde{\Sigma}_n = -\ddot{l}_n(\tilde{\theta}_n) = \mathbb{Z}\mathbf{D}(\tilde{\theta}_n)\mathbb{Z}'$ and $\Sigma_{n,0} = \mathbb{Z}\mathbf{D}(\theta_0)\mathbb{Z}'$. Now

$$(3.1) \qquad \widehat{\Sigma}_n^{\frac{1}{2}}(\widehat{\theta}_n - \theta_0) = \left[\widehat{\Sigma}_n^{\frac{1}{2}}\tilde{\Sigma}_n^{\frac{1}{2}}\right]\left[\tilde{\Sigma}_n^{\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}}\right]\Sigma_{n,0}^{\frac{1}{2}}\left[\tilde{\Sigma}_n(\widehat{\theta}_n - \theta_0)\right].$$

The two terms in brackets in (3.1) converge almost surely to I_k. To see this, we show for example that $\|\tilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}} - I_k\| \longrightarrow 0$ almost surely as $n \to \infty$. First, note that

$$\|\tilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}} - I_k\| = \|\tilde{\Sigma}_n^{-\frac{1}{2}}\left(\Sigma_{n,0}^{\frac{1}{2}} - \tilde{\Sigma}_n^{\frac{1}{2}}\right)\|$$

$$(3.2) \qquad\qquad \leq \|\tilde{\Sigma}_n^{\frac{1}{2}}\|\|\Sigma_{n,0}^{\frac{1}{2}} - \tilde{\Sigma}_n^{\frac{1}{2}}\|$$

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

and

$$\Lambda_n^{-1}\|\Sigma_{n,0} - \tilde{\Sigma}_n\| = \Lambda_n^{-1}\|\mathbb{Z}(\mathbf{D}(\theta_0) - \mathbf{D}(\tilde{\theta}_n))\mathbb{Z}'\|.$$

Note also that $\widehat{\theta}_n$ converges almost surely to θ_0 (that is, for every $w \in \tilde{\Omega}$, where $\tilde{\Omega} \subset \Omega$ and $\mathbb{P}(\tilde{\Omega}) = 1$). Let $w \in \tilde{\Omega}$. By the same arguments as in the proof of Lemma 3.1, for every $\epsilon > 0$, there exists a positive $n(\epsilon, w) \in \mathbb{N}$ such that if $n \geq n(\epsilon, w)$, then $\Lambda_n^{-1}\|\mathbb{Z}(\mathbf{D}(\theta_0) - \mathbf{D}(\tilde{\theta}_n))\mathbb{Z}'\| < \epsilon$. Hence $\Lambda_n^{-1}\|\mathbb{Z}(\mathbf{D}(\theta_0) - \mathbf{D}(\tilde{\theta}_n))\mathbb{Z}'\|$ converges almost surely to 0. By continuity of the map $x \to x^{\frac{1}{2}}$, $\|\Lambda_n^{-\frac{1}{2}}\left(\Sigma_{n,0}^{\frac{1}{2}} - \tilde{\Sigma}_n^{\frac{1}{2}}\right)\|$ converges also almost surely to 0. Moreover, for n sufficiently large, there exists a positive constant $A < \infty$, such that almost surely, $\|\tilde{\Sigma}_n^{-\frac{1}{2}}\| \leq A\frac{\Lambda_n^{\frac{1}{2}}}{\lambda_n^{1/2}} < Ac_2^{\frac{1}{2}}$. It follows from (3.2) and the condition C3 that $\|(\Sigma_{n,0}^{\frac{1}{2}}\tilde{\Sigma}_n^{-\frac{1}{2}}) - I_k\|$ converges almost

surely to 0. It remains for us to show that $\Sigma_{n,0}^{-\frac{1}{2}}(\tilde{\Sigma}_n(\widehat{\theta}_n - \theta_0))$ converges in distribution to $\mathcal{N}(0, I_k)$. Note that $\Sigma_{n,0}^{-\frac{1}{2}}(\tilde{\Sigma}_n(\widehat{\theta}_n - \theta_0)) = \Sigma_{n,0}^{-\frac{1}{2}} \sum_{j=1}^{2n} \mathbb{Z}_{\bullet j} C_j(\theta_0)$. This convergence holds if we can check the following conditions:

i) $\max_{1 \leq j \leq 2n} \mathbb{Z}'_{\bullet j}(\mathbb{Z}\mathbb{Z}')^{-1}\mathbb{Z}_{\bullet j} \to 0$ if $n \to \infty$. Condition i) follows by noting that

$$0 \leq \max_{1 \leq j \leq 2n} \mathbb{Z}'_{\bullet j}(\mathbb{Z}\mathbb{Z}')^{-1}\mathbb{Z}_{\bullet j} \leq \max_{1 \leq j \leq 2n} \|\mathbb{Z}_{\bullet j}\|^2 \|(\mathbb{Z}\mathbb{Z}')^{-1}\| = \max_{1 \leq j \leq 2n} \frac{\|\mathbb{Z}_{\bullet j}\|^2}{\tilde{\lambda}_n}.$$

and that $\|\mathbb{Z}_{\bullet j}\|^2$ is bounded above, by C1 and C2. Moreover, $\frac{1}{\tilde{\lambda}_n} \to 0$ as $n \to \infty$.

ii) $\sup_{1 \leq j \leq 2n} \mathbb{E}\left[C_j^2(\theta_0)\mathbf{1}_{\{|C_j(\theta_0)|>c\}}\right] \to 0$ as $n \to \infty$. Condition ii) follows by noting that the components $C_j(\theta_0)$ of C are bounded under C1, C2 and C3. Finally, for every $i = 1, \ldots, 2n$, $\mathbb{E}(C_j^2(\theta_0)) = Var(C_j(\theta_0))$ since $\mathbb{E}(C_j(\theta_0)) = 0$. If $u \in \{n+1, \ldots, 2n\}$, $C_j(\theta_0) = B_h(\theta_0)$ with $u = j - n$ then

$$\mathbb{V}ar(C_j(\theta_0))$$
$$= \mathbb{V}ar(B_h(\theta_0))$$
$$= \mathbb{E}(\mathbb{V}ar(B_h(\theta_0))|\mathbf{X}_u, \mathbf{W}_u)) + \mathbb{V}ar(\mathbb{E}(B_h(\theta_0)|\mathbf{X}_u, \mathbf{W}_u))$$
$$= \mathbb{E}(\mathbb{V}ar(B_h(\theta_0))|\mathbf{X}_u, \mathbf{W}_u).$$

Finally,

$$Var(B_h(\theta_0)|\mathbf{X}_h, \mathbf{W}_h)$$
$$= \left(\frac{-k_i n_i \mathbf{x}_i f(\beta'\mathbf{x}_i)}{F(\mu'\mathbf{w}_i)k_i(\theta)} + (1-k_i)\frac{\mathbf{x}_i f(\beta'\mathbf{x}_i)[\mathbf{z}_i - n_i F(\beta'\mathbf{x}_i)]}{F(\beta'\mathbf{x}_i)h_i}(\beta)\right)^2$$
$$\times Var(J_u|\mathbf{X}_u, \mathbf{W}_u)$$
$$= \left(\frac{-k_i n_i \mathbf{x}_i f(\beta'\mathbf{x}_i)}{F(\mu'\mathbf{w}_i)k_i(\theta)} + (1-k_i)\frac{\mathbf{x}_i f(\beta'\mathbf{x}_i)[\mathbf{z}_i - n_i F(\beta'\mathbf{x}_i)]}{F(\beta'\mathbf{x}_i)h_i(\beta)}\right)^2$$
$$\times \mathbb{P}(Y_u = 0|\mathbf{X}_u, \mathbf{W}_u)(1 - \mathbb{P}(Y_u = 0|\mathbf{X}_u, \mathbf{W}_u))$$
$$= \left(\frac{-k_i n_i \mathbf{x}_i f(\beta'\mathbf{x}_i)}{F(\mu'\mathbf{w}_i)k_i(\theta)} + (1-k_i)\frac{\mathbf{x}_i f(\beta'\mathbf{x}_i)[\mathbf{z}_i - n_i F(\beta'\mathbf{x}_i)]}{F(\beta'\mathbf{x}_i)h_i(\beta)}\right)^2$$
$$\times (p_u + (1-\pi_u)^{m_u}(1-p_j))(1-p_u) \times (1 - (1-\pi_u)^{m_u}).$$

Therefore $\mathbb{V}ar(B_u(\theta_0)|\mathbf{X}_u, \mathbf{W}_u) > 0$ and for all $j \in \{1, \ldots, n\}$ under conditions C1, C2 and C3, $\mathbb{V}ar(C_j(\theta_0)) > 0$.

iii) $\inf_{1 \leq j \leq 2n} \mathbb{E}(C_j^2(\theta_0)) > 0$. Condition iii) follows by using similar arguments, that is $\mathbb{V}ar(C_j(\theta_0)) > 0$ for all $j = 1, \ldots, 2n$.

To summarize, we have proved that $\Sigma_{n,0}^{-\frac{1}{2}}(\tilde{\Sigma}_n(\widehat{\theta}_n - \theta_0))$ converges in distribution to $\mathcal{N}(0, I_k)$. This result, combined with Slutsky's theorem and equation (3.1), implies that $\widehat{\Sigma}_n^{\frac{1}{2}}(\widehat{\theta}_n - \theta_0)$ converges in distribution to $\mathcal{N}(0, I_k)$.

This concludes the proof.   □

4. Discussion and perspectives

In this paper, we have proposed a procedure of maximum likelihood estimation in the Zero-inflated Binomial regression model by using probit link function for both coponents of the model (1.1). We establish the asymptotic properties (existence, consistency and asymptotic normality of the proposed maximum likelihood estimator).

Several open problems still deserve attention. First, some simulations to study the proposed of the estimator in finite sample and application on real data. It is sometimes of interest to make inference about the probability of event $\mathbb{P}(Y = y_i | \mathbf{W})$ across the whole range of the predictors $\mathbf{W}$. The calculation of simultaneous confidence bands for the probabilities $\{p(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ thus constitutes another issue of (both methodological and practical) interest. Another issue of interest deals with the inference in the ZIB regression model with probit link function, in a high-dimensional setting. We have established the theoretical properties of our estimator in a low-dimensional setting that is, when a small number of potential predictors are involved (this problem arises, for example, in genetic studies where high-dimensional data are generated using microarray technologies).

References

[1] A. AGRESTI: *Categorical Data Analysis*, Wiley, 2002, New York.

[2] B.E. BAGOZZI: *The baseline-inflated multinomial logit model for international relations research*, Conflict Management and Peace Science, **33**(2) (2016), 174-197.

[3] B.E. BAGOZZI: *A mixture model for middle category inflation in ordered survey responses*, Political Analysis, **20**(3) (2016), 369-386.

[4] C.L. BLISS: *Methods of probits*, Science, **79** (1934), 38-39.

[5] J. BOGOUSSLAVSKY, M.G. BOUSSER, J.L. MAS: *Les accidents vasculaires cérébraux*, Radiologie ed Doinville, (1993), 145.

[6] R. CALABRESE, S. OSMETTI: *Modelling SME Loan Defaults asRare Events: an Application to Credit Defaults*, Journal of AppliedStatistics, **40**(6) (2013), 1172-1188.

[7] D. COLLET: *Modelling Binary Data, 2nd ed.*, Boca Raton, FL: Chapman & Hall/CRC, 2003.

[8] G. CORDEIRO, M. DE CASTRO: *A new family of generalized distributions*, Journal of Statistical Computation and Simulation **81** (2011), 883–898.

[9] J.S. CRAMER: *Logit models from economics and other fields*, Cambridge, MA: Cambridge University Press, 2003.

[10] C. CZADO, A. MUNK: *Noncanonical links in generalized linear models: When is the effort justified?*, Journal of Statistical Planning and Inference, **87** (2000), 317-345.

[11] C. CZADO, T.J. SANTNER: *The effect of link misspecification on binary regression inference*, Journal of Statistical Planning and Inference **33** (1992), 213–231.

[12] G.M. CORDEIRO, P. MCCULLAGH: *Bias reduction in generalized linear models*, Journal of the Royal Statistical Society, Ser. B **53** (1991), 629–643.

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

[14] A.O. DIALLO, A. DIOP, J.F. DUPUY: *Analysis of multinomial counts with joint zero-inflation, with an application to health economics*, Journal of Statistical Planning and Inference, **194** (2017), 85-105.

[15] A. DIOP, A. DIOP, J.F. DUPUY: *Maximum likelihood estimation in the logistic regression model with a cure fraction*, Electronic Journal of Statistics, **5** (2011), 460-483.

[16] A. DIOP, J.F. DUPUY: *Simulation-based inference in a zero-inflated bernoulli regression model*, Communications in Statistics - Simulation and Computation, **45**(10) (2016), 3597-3614.

[17] F. FAMOYE, K.P. SINGH: *Zero-inflated generalized Poisson regression model with an application to domestic violence data*, Journal of Data Science, **4** (2004), 117-130.

[18] H.B. FANG, G. LI, J. SUN: *Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model*, Scandinavian Journal of Statistics **32** (0025), 59–75.

[19] D.J. FINNEY: *Probit analysis*, Cambridge, MA: Cambridge University Press, 1971.

[20] D.B. HALL: *Zero-inflated Poisson and binomial regression with random effects: a case study*, Biometrics, **56** (2000), 1030-1039.

[21] M.N. HARRIS, X. ZHAO: *A zero-inflated ordered probit model, with an application to modelling tobacco consumption*, Journal of Econometrics, **141**(2) (2007), 1073-1099.

[22] J.M. HILBE: *Logistic regression models*, Chapman & Hall, Boca Raton, 200(.

[23] D.A. HSIEH, C.F. MANSKI, D. MCFADDEN: *Estimation of Response Probabilities from Augmented Retrospective Observations*, Journal of the American Statistical Association, **80**(391) (1985), 651-662.

[24] D. LAMBERT: *Zero-inflated Poisson regression, with an application to defects in manufacturing*, Technometrics, **34** (1992), 1-14.

[25] J. HAUSMAN, J. ABREVAYA: *Scott-Morton F., Misclassification of the dependent variable in a discrete-response setting*, Journal of Econometrics, **87**(2) (1998), 239-269.

[26] D.W. HOSMER, S. LEMESHOW: *Applied logistic regression*, Wiley, New York, 2000.

[27] G. KING, L. ZENG: *Logistic Regression in Rare Events Data*, Political Analysis, **9** (2001), 137-163.

[28] C.F. MANSKI, S.R. LERMAN: *The Estimation of Choice Probabilities from Choice based Samples*, Econometrica **45**(8) (1977), 1977-88.

[29] P. MCCULLAGH, J.A. NELDER: *Generalized Linear Model*, Chapman Hall, New York, 1989.

[30] M. RAZZAGHI: *The Probit Link Function in Generalized Linear Models for Data Mining Applications*, Journal of Modern Applied Statistical Methods, **12**(1) (2013), 164-169.

[31] A.A. SHARIFF, A. ZAHARIM, K. SOPIAN: *The comparison of logit and probit regression analyses in estimating the strength of gear teeth*, European Journal of Scientific Research, **4** (2009), 548-553.

[32] J. SWEENEY, J. HASLETT, A.C. PARNELL: *The zero inflated binomial distribution with applications*, 2014.

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

LO FATOUMATA
DEPARTMENT OF MATHEMATICS
UNIVERSITY IBA DER THIAM OF THIES
THIES,
SENEGAL.
Email address: `fatymalo@hotmail.com`

BA DEMBA BOCAR
DEPARTMENT OF MATHEMATICS
UNIVERSITY IBA DER THIAM OF THIES
THIES,
SENEGAL.
Email address: `dbba@univ-thies.sn`

DIOP ABA
EQUIPE DE RECHERCHE EN STATISTIQUE ET MODÉLES ALÉATOIRES (ERESMA)
UNIVERSITY ALIOUNE DIOP OF BAMBEY
BAMBEY,
SENEGAL.
Email address: `aba.diop@uadb.edu.sn`