# Simultaneous confidence bands in a zero-inflated regression model for binary data

Aba Diop, Aliou Diop, Jean-François Dupuy

# Simultaneous confidence bands in a zero-inflated regression model for binary data

Aba DIOP

*MIA, Université de La Rochelle, France and LERSTAD, Université Gaston Berger, Saint Louis, Sénégal.*
*Email: aba.diop@univ-lr.fr*

Aliou DIOP

*LERSTAD, Université Gaston Berger, Saint Louis, Sénégal.*
*Email: aliou.diop@ugb.edu.sn*

Jean-François DUPUY†

*IRMAR-Institut National des Sciences Appliquées de Rennes, France.*
*Email: Jean-Francois.Dupuy@insa-rennes.fr*

**Abstract**. The logistic regression model has become a standard tool to investigate the relationship between a binary outcome and a set of potential predictors. When analyzing binary data, it often arises however that the observed proportion of zeros is greater than expected under the postulated logistic model. Zero-inflated binomial (ZIB) models have been developed to fit binary data that contain too many zeros. Maximum likelihood estimators in these models have been proposed, and their asymptotic properties recently established. In this paper, we use these asymptotic properties to construct simultaneous confidence bands for the probability of a positive outcome in a ZIB regression model. Simultaneous confidence bands are especially attractive since they allow inference to be made over the whole regressor space. We construct two types of confidence bands, based on: i) the Scheffé method for the linear regression model ii) Monte Carlo simulations to approximate the distribution of the supremum of a Gaussian field indexed by the regressor. The finite-samples properties of these two types of bands are investigated and compared in a simulation study.

**Keywords**: Logistic regression model, mixture model, simultaneous inference, simulations

## 1. Introduction

The logistic regression model has become a standard tool to describe the relationship between a binary response $Y$ and a set of potential predictors $\mathbf{X}$. In the medical setting for example, the response may represent the infection status with respect to some disease. If $Y_i$ denotes the infection status for the $i$-th individual in a sample of size $n$ ($Y_i = 1$ if the individual is infected, and $Y_i = 0$ otherwise) and $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})^\top$ is the corresponding predictor (or covariate), logistic regression models the conditional probability $\mathbb{P}(Y_i = 1|\mathbf{X}_i)$ of infection as

$$\log\left(\frac{\mathbb{P}(Y_i = 1|\mathbf{X}_i)}{1 - \mathbb{P}(Y_i = 1|\mathbf{X}_i)}\right) = \beta^\top \mathbf{X}_i, \tag{1}$$

†Corresponding author

where $\beta = (\beta_1, \ldots, \beta_p)^{\intercal} \in \mathbb{R}^p$ is an unknown regression parameter to be estimated. Estimation and testing procedures in the model (1) are well established (see for example Hosmer and Lemeshow (2000) and Hilbe (2009)). These procedures are usually based on the maximum likelihood estimator of $\beta$, which was shown to be consistent and approximately normally distributed in large samples (Gouriéroux and Monfort, 1981). This estimator and the related statistical inference are available in all standard statistical softwares and can easily be implemented by non-statisticians.

However, when analyzing binary datasets, it often arises that the observed proportion of zeros is greater than expected under the postulated logistic model. In the medical setting for example, this situation usually occurs as a result of immunity. An individual $i$ is said to be immune (or "cured", as opposed to "at-risk") when he cannot experience the outcome of interest. In this case, $Y_i = 0$ but this zero cannot be considered as generated from the model (1). For this reason, Ridout *et al.* (1998) make the distinction between structural zeros (which are inevitable, being due to individual immunity for example) and random zeros (which arise by chance, under the model (1) for example). One problem arising in this setting is that it is usually unknown who are the at-risk and the cured subjects (unless the outcome of interest has been observed). Consider for example the occurrence of infection from some disease to be the outcome of interest. Then, if a subject is uninfected, the investigator does usually not know whether this subject is immune to the infection, or at-risk albeit still uninfected (at-risk subjects are also called susceptibles). Based on such data, the statistical inference in the model (1) is no more straightforward. This problem can be considered within the general framework of zero-inflated models.

Generally speaking, zero-inflation occurs in the analysis of counts when the data generating process results into too many zeros. Failure to account for these extra zeros is known to result in biased parameter estimates and inferences. Motivated by various applications in public health, epidemiology, sociology, engineering, agriculture, ..., a variety of zero-inflated regression models have recently been proposed and applied, such as the zero-inflated Poisson (ZIP) model (see, among others, Lambert (1992), Dietz and Böhning (2000), Lam *et al.* (2006), Xiang *et al.* (2007), Li (2011)), and the zero-inflated binomial (ZIB) model (Hall (2000), Diop *et al.* (2011)) which we shall concentrate on in this paper. Various other models and numerous references can be found in Famoye and Singh (2006), Lee *et al.* (2006), Kelley and Anderson (2008), and Moghimbeigi *et al.* (2009). Zero-inflated data are commonly modeled by mixtures. Precisely, in a zero-inflated model, the count response variable is assumed to be distributed as a mixture of a count distribution and a distribution with point mass of one at zero. Diop *et al.* (2011) recently established the consistency and asymptotic normality of the maximum likelihood estimator in a ZIB model with logit links for both the response variable and the zero-inflation component.

One objective of logistic regression is to make inference on probabilities of the form $\mathbf{P}(Y = 1|\mathbf{X} = \mathbf{x})$ that is, on the probabilities of a positive outcome at $\mathbf{x}$. Based on the asymptotic results for the maximum likelihood estimator of $\beta$, one can easily construct asymptotic confidence regions for vectors of the form $(\mathbf{P}(Y = 1|\mathbf{X} = \mathbf{x}_1), \ldots, \mathbf{P}(Y = 1|\mathbf{X} = \mathbf{x}_k))$. But such pointwise confidence regions are not adequate for making inference about the response function $\mathbf{x} \mapsto \mathbf{P}(Y = 1|\mathbf{X} = \mathbf{x})$ across the whole range of $\mathbf{X}$. Simultaneous confidence bands provide the correct tool to use for that purpose. Generally speaking, simultaneous confidence bands can be used to bound unknown functions just as usual confidence intervals or regions bound unknown finite-dimensional parameters. In particular, simultaneous confidence bands provide a useful tool to quantify the plausible range of an unknown regression function and therefore, they have been widely investigated in linear

and polynomial regression, in survival regression (see Fleming and Harrington (1991) for example), and in generalized linear regression (see Sun *et al.* (2000) for example). We refer to Liu (2011) for a detailed account of simultaneous inference in regression models, with a particular emphasis on linear regression. Zero-inflated regression models are now widely applied but despite this, it seems that simultaneous confidence bands have never been investigated in this class of models. Thus in the present paper, we consider the construction of simultaneous bands for the probability of a positive outcome (viewed as a function of $\mathbf{x}$) in a zero-inflated Bernoulli regression model. As far as we know, this constitutes the first attempt to construct simultaneous confidence bands in a zero-inflated model.

The rest of the paper is organized as follows. Section 2 describes the ZIB model of interest and recalls the properties of the maximum likelihood estimator in this model. In the section 3, we construct two types of simultaneous confidence bands for the probability of a positive outcome in the ZIB model. For the first type, we follow the Scheffé method for the linear regression model. Then we propose an alternative confidence band, whose construction relies on Monte Carlo simulations to approximate the distribution of the supremum of a Gaussian field indexed by the regressor. The finite-samples properties of these two bands are evaluated and compared in a simulation study reported in section 4. A discussion and some perspectives conclude the paper (section 5).

## 2. Model and estimation

In this section, we provide a brief review of the ZIB model and we recall from Diop *et al.* (2011) the properties of the maximum likelihood estimator of the regression parameter in this model. These properties will be useful for constructing the simultaneous confidence bands in section 3.

Let $\mathcal{O}_i = (Y_i, S_i, \mathbf{X}_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$ be independent copies of the random vector $\mathcal{O} = (Y, S, \mathbf{X}, \mathbf{Z})$, where for every $i$, $Y_i$ and $S_i$ are binary variables indicating respectively whether the event of interest has occurred on the $i$-th individual ($Y_i = 1$) or not ($Y_i = 0$), and whether the individual is susceptible to the event ($S_i = 1$) or not ($S_i = 0$). If $Y_i = 0$, the value of $S_i$ is unknown. $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})^\top$ and $\mathbf{Z}_i = (1, Z_{i2}, \ldots, Z_{iq})^\top$ are covariate vectors respectively related to the event risk and to the susceptibility status. $\mathbf{X}_i$ and $\mathbf{Z}_i$ may contain quantitative and qualitative components, and may share some components (note that here and in the sequel, all vectors are column vectors and $\top$ denotes the transpose). A zero-inflated Bernoulli regression model (Hall (2000), Diop *et al.* (2011)) for the $\mathcal{O}_i$ ($i = 1, \ldots, n$) can be defined by the following equations for the event probability:

$$\begin{cases} \log\left(\frac{\mathbb{P}(Y=1|\mathbf{X}_i,S_i)}{1-\mathbb{P}(Y=1|\mathbf{X}_i,S_i)}\right) = \beta^\top \mathbf{X}_i & \text{if } \{S_i = 1\} \\ \mathbb{P}(Y = 0|\mathbf{X}_i, S_i) = 1 & \text{if } \{S_i = 0\} \end{cases} \tag{2}$$

and by the following model for the susceptibility status:

$$\log\left(\frac{\mathbb{P}(S = 1|\mathbf{Z}_i)}{1 - \mathbb{P}(S = 1|\mathbf{Z}_i)}\right) = \theta^\top \mathbf{Z}_i. \tag{3}$$

In this model, $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is an unknown regression parameter of interest ($\beta$ measures the association between the potential predictors $\mathbf{X}_i$ and the risk of event for a susceptible individual), and $\theta = (\theta_1, \ldots, \theta_q)^\top \in \mathbb{R}^q$ is an unknown nuisance parameter. Let

$\psi := (\beta^\top, \theta^\top)^\top$. The log-likelihood for $\psi$ from the sample $\mathcal{O}_1, \ldots, \mathcal{O}_n$ (where $S_i$ is unknown when $Y_i = 0$) is $l_n(\psi) = \sum_{i=1}^n l(\psi; \mathcal{O}_i)$, where $l(\psi; \mathcal{O}_i) =$

$$Y_i(\beta^\top \mathbf{X}_i + \theta^\top \mathbf{Z}_i) + (1 - Y_i)\log(1 + e^{\beta^\top \mathbf{X}_i} + e^{\theta^\top \mathbf{Z}_i}) - \log\left(1 + e^{\beta^\top \mathbf{X}_i}\right) - \log\left(1 + e^{\theta^\top \mathbf{Z}_i}\right)$$

The maximum likelihood estimator (MLE) $\widehat{\psi}_n := (\widehat{\beta}_n^\top, \widehat{\theta}_n^\top)^\top$ of $\psi$ is defined as the solution of the score equation $\partial l_n(\psi)/\partial\psi = 0$ which can be solved, for example, using the `optim` function of the software `R`. Diop *et al.* (2011) derived the asymptotic properties of $\widehat{\psi}_n$ under the following regularity conditions (we refer to Diop *et al.* (2011) for a discussion of these conditions):

**A1** The covariates are bounded (in the sequel, we will denote by $\mathcal{X}$ the space for $\mathbf{X}$). For every $i = 1, 2, \ldots$, $j = 2, \ldots, p$, $k = 2, \ldots, q$, $\text{var}[X_{ij}] > 0$ and $\text{var}[Z_{ik}] > 0$. For every $i = 1, 2, \ldots$, the $X_{ij}$ $(j = 1, \ldots, p)$ are linearly independent, and the $Z_{ik}$ $(k = 1, \ldots, q)$ are linearly independent.

**A2** There exists a continuous covariate $V$ which is in $\mathbf{X}$ but not in $\mathbf{Z}$ that is, if $\beta_V$ and $\theta_V$ denote the coefficients of $V$ in the linear predictors (2) and (3) respectively, then $\beta_V \neq 0$ and $\theta_V = 0$.

**A3** The parameters $\beta$ and $\theta$ lie in the interior of known compact sets $\mathcal{B} \subset \mathbb{R}^p$ and $\mathcal{G} \subset \mathbb{R}^q$ respectively.

**A4** The matrix $\ddot{l}_n(\psi) = \partial^2 l_n(\psi)/\partial\psi\partial\psi^\top$ is negative definite and of full rank, for every $n = 1, 2, \ldots$ Letting $\lambda_n$ and $\Lambda_n$ be the smallest and largest eigenvalues of $\ddot{l}_n(\psi_0)$ respectively, there exists a finite positive constant $c$ such that $\Lambda_n/\lambda_n < c$ for every $n = 1, 2, \ldots$

We need some further notations before stating the asymptotic properties of the MLE $\widehat{\beta}_n$ of the parameter of interest $\beta$. Let $\mathcal{I}_\psi = -\mathbb{E}[\partial^2 l(\psi; \mathcal{O})/\partial\psi\partial\psi^\top]$, $\widehat{\mathcal{I}}_{\widehat{\psi}_n} = -n^{-1}\ddot{l}_n(\widehat{\psi}_n)$, and $M$ be the $(p \times (p + q))$ block-matrix $[I_p, 0_{p,q}]$, where $I_p$ is the identity matrix of order $p$ and $0_{p,q}$ is the $(p \times q)$ matrix whose components are all equal to 0. Let also $\Sigma_\beta = M\mathcal{I}_\psi^{-1}M^\top$ and $\widehat{\Sigma}_{\beta,n} = M\widehat{\mathcal{I}}_{\widehat{\psi}_n}^{-1}M^\top$. Finally, $\xrightarrow{d}$ and $\xrightarrow{p}$ will denote the convergence in distribution and in probability respectively. Under the conditions above, Diop *et al.* (2011) prove the following theorem:

THEOREM 1 (DIOP *et al.* (2011)). *As* $n \to$, $\sqrt{n}(\widehat{\beta}_n - \beta) \xrightarrow{d} N(0, \Sigma_\beta)$ *and* $\widehat{\Sigma}_{\beta,n} \xrightarrow{p} \Sigma_\beta$.

One issue of particular interest in the model (2)-(3) is to estimate the probability $p(\mathbf{x}) := \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}, S = 1)$ of event for a susceptible individual with covariate value $\mathbf{x}$. It is straightforward to estimate $p(\mathbf{x})$ by $\exp(\widehat{\beta}_n^\top \mathbf{x})/(1 + \exp(\widehat{\beta}_n^\top \mathbf{x}))$, which is a consistent and asymptotically normal estimator of $p(\mathbf{x})$, for every $\mathbf{x}$. This result can be used to compute confidence intervals for the probabilities $p(\mathbf{x})$ (or confidence regions for sets of probabilities $(p(\mathbf{x}_1), \ldots, p(\mathbf{x}_k))$ for some set of covariate values $(\mathbf{x}_1, \ldots, \mathbf{x}_k)$).
One further step, which shall give a deeper insight into the true relationship between $\mathbf{x}$ and the event probability, is to obtain simultaneous confidence bands for the whole function $p(\cdot)$. This is the topic of the next section, where we construct two different types of simultaneous confidence bands for $p(\cdot)$ (note that one may also be interested in constructing confidence bands for the probability function $\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$. Obviously, the techniques that we will develop in the next section could be applied to this problem as well).

## 3. Simultaneous confidence bands

As mentioned above, pointwise confidence intervals are not adequate for making inference about the response function $p(\mathbf{x})$ across the whole range of $\mathbf{x}$. In this section, we investigate two methods for constructing simultaneous confidence bands for the whole $\{p(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$.

### *3.1. Method 1*

Relying on arguments similar to that for deriving Scheffé's band in a linear regression model (see Liu (2011)), we construct a first type of simultaneous confidence band for $\{p(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ in the ZIB regression model (2)-(3). Bands of the Scheffé type in the usual logistic regression model (1) have been investigated by Brand *et al.* (1973) and Hauck (1983). See also Li *et al.* (2010), who recently constructed Scheffé-type bands in a semiparametric logistic regression model.

We need first some notations. Let $\widehat{\sigma}_n^2(\mathbf{x}) = \mathbf{x}^\top \widehat{\Sigma}_{\beta,n} \mathbf{x}$, and $\chi^2_{p,1-\alpha}$ be the quantile of order $(1-\alpha)$ of the $\chi^2_p$ distribution. Then the following holds:

THEOREM 2. *Let $\alpha \in (0,1)$ and assume that the conditions A1-A4 hold. Then*

$$\lim_{n\to\infty} \mathbb{P}\left( \beta^\top \mathbf{x} \in \widehat{\beta}_n^\top \mathbf{x} \pm \widehat{\sigma}_n(\mathbf{x}) \sqrt{\frac{\chi^2_{p,1-\alpha}}{n}},\ \forall\ \mathbf{x} \in \mathcal{X} \right) \geq 1-\alpha.$$

**Proof of Theorem 2.** From Theorem 1, $n(\widehat{\beta}_n - \beta)^\top \widehat{\Sigma}^{-1}_{\beta,n} (\widehat{\beta}_n - \beta) \xrightarrow{d} \chi^2_p$ as $n \to \infty$. Then

$$\lim_{n\to\infty} \mathbb{P}\left( n(\widehat{\beta}_n - \beta)^\top \widehat{\Sigma}^{-1}_{\beta,n} (\widehat{\beta}_n - \beta) \leq \chi^2_{p,1-\alpha} \right) = 1-\alpha. \tag{4}$$

Now, using Cauchy-Schwarz inequality (Rao, 1973), we get that

$$\sup_{\mathbf{x}\in\mathcal{X}} \frac{\left(\sqrt{n}(\widehat{\beta}_n - \beta)^\top \mathbf{x}\right)^2}{\mathbf{x}^\top \widehat{\Sigma}_{\beta,n} \mathbf{x}} \leq n(\widehat{\beta}_n - \beta)^\top \widehat{\Sigma}^{-1}_{\beta,n} (\widehat{\beta}_n - \beta) \tag{5}$$

and thus

$$\mathbb{P}\left( n(\widehat{\beta}_n - \beta)^\top \widehat{\Sigma}^{-1}_{\beta,n} (\widehat{\beta}_n - \beta) \leq \chi^2_{p,1-\alpha} \right) \leq \mathbb{P}\left( \frac{\left(\sqrt{n}(\widehat{\beta}_n - \beta)^\top \mathbf{x}\right)^2}{\widehat{\sigma}_n^2(\mathbf{x})} \leq \chi^2_{p,1-\alpha},\ \forall\ \mathbf{x} \in \mathcal{X} \right).$$

It follows from (4) that

$$\begin{aligned} 1-\alpha &\leq \lim_{n\to\infty} \mathbb{P}\left( \frac{\left(\sqrt{n}(\widehat{\beta}_n - \beta)^\top \mathbf{x}\right)^2}{\widehat{\sigma}_n^2(\mathbf{x})} \leq \chi^2_{p,1-\alpha},\ \forall\ \mathbf{x} \in \mathcal{X} \right) \\ &\leq \lim_{n\to\infty} \mathbb{P}\left( \beta^\top \mathbf{x} \in \widehat{\beta}_n^\top \mathbf{x} \pm \widehat{\sigma}_n(\mathbf{x}) \sqrt{\frac{\chi^2_{p,1-\alpha}}{n}},\ \forall\ \mathbf{x} \in \mathcal{X} \right). \end{aligned} \tag{6}$$

□

From this, we can deduce a simultaneous confidence band of asymptotic confidence level greater than or equal to $(1-\alpha)$ for $\{p(\mathbf{x}) = \exp(\beta^\top \mathbf{x})/(1+\exp(\beta^\top \mathbf{x})), \forall \mathbf{x} \in \mathcal{X}\}$, as

$$\{[\widehat{l}^{(1)}_{p(\mathbf{x}),n}, \widehat{u}^{(1)}_{p(\mathbf{x}),n}], \mathbf{x} \in \mathcal{X}\}, \tag{7}$$

where

$$\widehat{l}^{(1)}_{p(\mathbf{x}),n} = \frac{\exp\left(\widehat{\beta}_n^\top \mathbf{x} - \widehat{\sigma}_n(\mathbf{x})\sqrt{\frac{\chi^2_{p,1-\alpha}}{n}}\right)}{1+\exp\left(\widehat{\beta}_n^\top \mathbf{x} - \widehat{\sigma}_n(\mathbf{x})\sqrt{\frac{\chi^2_{p,1-\alpha}}{n}}\right)} \text{ and } \widehat{u}^{(1)}_{p(\mathbf{x}),n} = \frac{\exp\left(\widehat{\beta}_n^\top \mathbf{x} + \widehat{\sigma}_n(\mathbf{x})\sqrt{\frac{\chi^2_{p,1-\alpha}}{n}}\right)}{1+\exp\left(\widehat{\beta}_n^\top \mathbf{x} + \widehat{\sigma}_n(\mathbf{x})\sqrt{\frac{\chi^2_{p,1-\alpha}}{n}}\right)},$$

**Remark.** Scheffé-type simultaneous confidence bands are conservative when the covariate $\mathbf{X}$ is restricted to a subset $\mathcal{X} \subset \mathbb{R}^p$. When $\mathbf{X}$ is unbounded, the $\leq$ sign in (5) is an equality, implying that the $\leq$ sign in (6) is also an equality. In this case, the asymptotic confidence level of the band (7) is $1-\alpha$. When $\mathbf{X}$ is bounded, Piegorsch and Casella (1988) provided a method to compute bands of the Scheffé-type with asymptotic level $1-\alpha$. However, we will observe in our simulation study that even in the case of conservative bands, the empirical coverage probability of Scheffé-type bands is lower than the desired $1-\alpha$. This may be due to the fact that the distribution of $\sup_{\mathbf{x}\in\mathcal{X}}(\sqrt{n}(\widehat{\beta}_n - \beta)^\top \mathbf{x})^2/\mathbf{x}^\top \widehat{\Sigma}_{\beta,n}\mathbf{x}$ is not well-approximated by a $\chi^2$ distribution. The approximation made in (5) is thus too rough, and the resulting confidence band appears to be narrower than it should be. Thus, we do not pursue with Piegorsch and Casella (1988)'s correction for constructing $(1-\alpha)$-level confidence bands in the zero-inflated regression model with bounded covariates. Rather, an alternative and more precise approach is suggested in the next section.

### *3.2. Method 2*

In this section, we propose an alternative method for constructing a simultaneous confidence band for $\{p(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ in the ZIB model (2)-(3). This approach relies on the weak convergence of the normalized process $W_n(\mathbf{x}) := \sqrt{n}(\widehat{\beta}_n - \beta)^\top \mathbf{x}/\widehat{\sigma}_n(\mathbf{x})$ indexed by $\mathbf{x} \in \mathcal{X}$, and uses Monte Carlo simulations (we refer to Li *et al.* (2010) for a related approach in a semiparametric logistic model). We first describe the asymptotic results needed for our construction.

#### *3.2.1. Preliminary results*

Let $C(\mathcal{X})$ be the space of real-valued continuous functions defined on $\mathcal{X}$, provided with the uniform norm. Let also $\sigma^2(\mathbf{x}) := \mathbf{x}^\top \Sigma_\beta \mathbf{x}$.

THEOREM 3. *Assume that the conditions A1-A4 hold. Then as $n \to \infty$, $W_n$ converges weakly to $W$ in $C(\mathcal{X})$, where $W$ is a Gaussian process with mean 0 and covariance function $\rho(\mathbf{x},\mathbf{y}) = \mathbf{x}^\top \Sigma_\beta \mathbf{y}/\sigma(\mathbf{x})\sigma(\mathbf{y})$.*

**Proof of Theorem 3.** In a first step, we show that as $n \to \infty$,

$$\sup_{\mathbf{x}\in\mathcal{X}} \left|\widehat{\sigma}^2_n(\mathbf{x}) - \sigma^2(\mathbf{x})\right| \xrightarrow{p} 0.$$

Note that

$$\begin{aligned}
\sup_{\mathbf{x}\in\mathcal{X}} \left|\widehat{\sigma}_n^2(\mathbf{x}) - \sigma^2(\mathbf{x})\right| &= \sup_{\mathbf{x}\in\mathcal{X}} \left|\mathbf{x}^\top \left(\widehat{\Sigma}_{\beta,n} - \Sigma_\beta\right)\mathbf{x}\right| \\
&\leq \sup_{\mathbf{x}\in\mathcal{X}} \left\{ \|\mathbf{x}\| \left\| \left(\widehat{\Sigma}_{\beta,n} - \Sigma_\beta\right)\mathbf{x} \right\| \right\} \\
&= \sup_{\mathbf{x}\in\mathcal{X},\, \|\mathbf{x}\|\neq 0} \left\{ \|\mathbf{x}\|^2 \frac{\left\| \left(\widehat{\Sigma}_{\beta,n} - \Sigma_\beta\right)\mathbf{x} \right\|}{\|\mathbf{x}\|} \right\} \\
&\leq \sup_{\mathbf{x}\in\mathcal{X},\, \|\mathbf{x}\|\neq 0} \|\mathbf{x}\|^2 . \sup_{\|\mathbf{x}\|\neq 0} \frac{\left\| \left(\widehat{\Sigma}_{\beta,n} - \Sigma_\beta\right)\mathbf{x} \right\|}{\|\mathbf{x}\|} \\
&= \sup_{\mathbf{x}\in\mathcal{X},\, \|\mathbf{x}\|\neq 0} \|\mathbf{x}\|^2 . \left\| \widehat{\Sigma}_{\beta,n} - \Sigma_\beta \right\|
\end{aligned}$$

where the first to second line follows from the Cauchy-Schwarz inequality. Now, $\mathbf{x}$ is bounded (by the assumption **A1**), and using the convergence of $\widehat{\Sigma}_{\beta,n}$ to $\Sigma_\beta$ and the continuity of the norm, it follows that the last line converges to 0 in probability as $n \to \infty$. Thus $\sup_{\mathbf{x}\in\mathcal{X}} \left|\widehat{\sigma}_n^2(\mathbf{x}) - \sigma^2(\mathbf{x})\right| \xrightarrow{p} 0$ as $n \to \infty$.
Now, the function $\phi : \mathbb{R}^p \to C(\mathcal{X})$ defined by $\phi(\mathbf{x})(\mathbf{y}) = \mathbf{x}^\top\mathbf{y}$ is obviously continuous and thus, $\phi(\sqrt{n}(\widehat{\beta}_n - \beta)) \xrightarrow{d} \phi(N(0, \Sigma_\beta))$ in $C(\mathcal{X})$ that is, $\{\sqrt{n}(\widehat{\beta}_n - \beta)^\top\mathbf{x}; \mathbf{x} \in \mathcal{X}\}$ converges weakly to $\{N(0, \Sigma_\beta)^\top\mathbf{x}; \mathbf{x} \in \mathcal{X}\}$ in $C(\mathcal{X})$. We have proved above that $\widehat{\sigma}_n^2(\cdot)$ converges uniformly in probability to $\sigma^2(\cdot)$ and thus, by Slutsky's theorem, $\{W_n(\mathbf{x}); \mathbf{x} \in \mathcal{X}\}$ converges weakly to $\{W(\mathbf{x}); \mathbf{x} \in \mathcal{X}\}$ in $C(\mathcal{X})$, where $W(\mathbf{x}) := N(0, \Sigma_\beta)^\top\mathbf{x}/\sigma(\mathbf{x})$. Finally, it is straightfirward to check that

$$\rho(\mathbf{x}, \mathbf{y}) := \mathrm{cov}(W(\mathbf{x}), W(\mathbf{y})) = \frac{\mathbf{x}^\top\Sigma_\beta\mathbf{y}}{\sigma(\mathbf{x})\sigma(\mathbf{y})}.$$

□

*3.2.2. Application to simultaneous confidence bands*

It follows from the Theorem 3 and the continuous mapping theorem (Billingsley, 1968) that

$$\sup_{\mathbf{x}\in\mathcal{X}} |W_n(\mathbf{x})| \xrightarrow{d} \sup_{\mathbf{x}\in\mathcal{X}} |W(\mathbf{x})| \text{ as } n \to \infty.$$

Then, if we knew the quantile $c_{1-\alpha}$ of order $(1-\alpha)$ of $\sup_{\mathbf{x}\in\mathcal{X}} |W(\mathbf{x})|$, we could construct an approximate simultaneous confidence band for $\{p(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$. One would simply write that

$$\begin{aligned}
1-\alpha &= \mathbb{P}\left( \sup_{\mathbf{x}\in\mathcal{X}} |W(\mathbf{x})| \leq c_{1-\alpha} \right) \\
&= \lim_{n\to\infty} \mathbb{P}\left( \sup_{\mathbf{x}\in\mathcal{X}} |W_n(\mathbf{x})| \leq c_{1-\alpha} \right) \\
&= \lim_{n\to\infty} \mathbb{P}\left( \left| \frac{\sqrt{n}(\widehat{\beta}_n - \beta)^\top\mathbf{x}}{\widehat{\sigma}_n(\mathbf{x})} \right| \leq c_{1-\alpha}, \forall \mathbf{x} \in \mathcal{X} \right) \\
&= \lim_{n\to\infty} \mathbb{P}\left( \beta^\top\mathbf{x} \in \widehat{\beta}_n^\top\mathbf{x} \pm \widehat{\sigma}_n(\mathbf{x}) \frac{c_{1-\alpha}}{\sqrt{n}}, \forall \mathbf{x} \in \mathcal{X} \right)
\end{aligned}$$

and deduce a simultaneous confidence band of asymptotic level $(1-\alpha)$ for $\{\beta^\top \mathbf{x}, \mathbf{x} \in \mathcal{X}\}$ as:

$$\left[\widehat{\beta}_n^\top \mathbf{x} - \widehat{\sigma}_n(\mathbf{x})\frac{c_{1-\alpha}}{\sqrt{n}}, \widehat{\beta}_n^\top \mathbf{x} + \widehat{\sigma}_n(\mathbf{x})\frac{c_{1-\alpha}}{\sqrt{n}}\right]$$

from which a band for $\{p(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ can easily be derived. $c_{1-\alpha}$ being unknown, we propose to estimate it using a Monte Carlo approach. The basic idea is to simulate a large number of replicates of $U_n := \sup_{\mathbf{x}\in\mathcal{X}} |W_n(\mathbf{x})|$, and to use their empirical quantile of order $(1-\alpha)$ to approximate $c_{1-\alpha}$. Precisely, the proposed algorithm is as follows:

**1.** *Draw $B$ bootstrap samples $\{(Y_i^{(b)}, \mathbf{X}_i^{(b)}, \mathbf{Z}_i^{(b)}), i = 1, \ldots, n\}$ $(b = 1, \ldots, B)$ from the original data sample, and for each bootstrap sample, estimate $\psi$ by its MLE $\widehat{\psi}_n^{(b)}$ in the model (2)-(3). Then, calculate the empirical covariance matrix $\hat{\Sigma}_{boot} = cov(\widehat{\beta}_n^{(b)}; b = 1, \ldots, B)$.*

**2.** *Draw $L$ independent $p$-vectors $\kappa_l \sim N(0, \hat{\Sigma}_{boot})$, $l = 1, \ldots, L$.*

**3.** *Calculate*

$$U_n^{(l)} = \max_{\mathbf{x}_i} \left| \frac{\mathbf{x}_i^\top \kappa_l}{\left(\mathbf{x}_i^\top \hat{\Sigma}_{boot} \mathbf{x}_i\right)^{\frac{1}{2}}} \right| \text{ for } l = 1, \ldots, L.$$

*where the maximum is taken over a grid of the covariable space $\mathcal{X}$.*

**4.** *Approximate $c_{1-\alpha}$ by the empirical quantile $\widehat{c_{1-\alpha}}$ of order $(1-\alpha)$ of $(U_n^{(1)}, \ldots, U_n^{(L)})$.*

The first step of the algorithm should result in a more robust estimate of the variance of $\widehat{\beta}_n$ than the estimate obtained from the sole original sample. Now, the empirical distribution of the $U_n^{(1)}, \ldots, U_n^{(L)}$ provides an estimate of the distribution of $\sup_{\mathbf{x}\in\mathcal{X}} |W_n(\mathbf{x})|$, which in turn approximates the distribution of $\sup_{\mathbf{x}\in\mathcal{X}} |W(\mathbf{x})|$. Thus, we may expect $\widehat{c_{1-\alpha}}$ to approximate the $(1-\alpha)$-quantile $c_{1-\alpha}$ of $\sup_{\mathbf{x}\in\mathcal{X}} |W(\mathbf{x})|$. We can now construct an approximate simultaneous confidence band of asymptotic level $(1-\alpha)$ for $\{\beta^\top \mathbf{x}, \mathbf{x} \in \mathcal{X}\}$, as:

$$\left[\widehat{\beta}_n^\top \mathbf{x} - \widehat{\sigma}_n(\mathbf{x})\frac{\widehat{c_{1-\alpha}}}{\sqrt{n}}, \widehat{\beta}_n^\top \mathbf{x} + \widehat{\sigma}_n(\mathbf{x})\frac{\widehat{c_{1-\alpha}}}{\sqrt{n}}\right].$$

From this, we deduce an approximate simultaneous confidence band of asymptotic confidence level $(1-\alpha)$ for $\{p(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$, as:

$$\{[\widehat{l}^{(2)}_{p(\mathbf{x}),n}, \widehat{u}^{(2)}_{p(\mathbf{x}),n}], \mathbf{x} \in \mathcal{X}\} \tag{8}$$

where

$$\widehat{l}^{(2)}_{p(\mathbf{x}),n} = \frac{\exp\left(\widehat{\beta}_n^\top \mathbf{x} - \widehat{\sigma}_n(\mathbf{x})\frac{\widehat{c_{1-\alpha}}}{\sqrt{n}}\right)}{1 + \exp\left(\widehat{\beta}_n^\top \mathbf{x} - \widehat{\sigma}_n(\mathbf{x})\frac{\widehat{c_{1-\alpha}}}{\sqrt{n}}\right)} \quad \text{and} \quad \widehat{u}^{(2)}_{p(\mathbf{x}),n} = \frac{\exp\left(\widehat{\beta}_n^\top \mathbf{x} + \widehat{\sigma}_n(\mathbf{x})\frac{\widehat{c_{1-\alpha}}}{\sqrt{n}}\right)}{1 + \exp\left(\widehat{\beta}_n^\top \mathbf{x} + \widehat{\sigma}_n(\mathbf{x})\frac{\widehat{c_{1-\alpha}}}{\sqrt{n}}\right)}.$$

## 4. Simulation study

The objective of this simulation study is to evaluate and compare the finite-sample performance of the proposed simultaneous confidence bands (7) and (8) in the zero-inflated Bernoulli regression model (2)-(3). Precisely, we aim at evaluating the influence of various simulation parameters (sample size, proportion of immunes, proportion of infected among the susceptibles) on the performance of the proposed bands. We first describe the simulation design that was adopted for that purpose.

The simulation setting is as follows. We consider the following models for the infection status:

$$\begin{cases} \log\left(\frac{\mathbb{P}(Y=1|\mathbf{X}_i,S_i)}{1-\mathbb{P}(Y=1|\mathbf{X}_i,S_i)}\right) = \beta_1 + \beta_2 X_{i2} + \beta_3 Z_{i2} & \text{if } S_i = 1 \\ \mathbb{P}(Y=1|\mathbf{X}_i,S_i) = 0 & \text{if } S_i = 0 \end{cases} \tag{9}$$

and the immunity status:

$$\log\left(\frac{\mathbb{P}(S=1|\mathbf{Z}_i)}{1-\mathbb{P}(S=1|\mathbf{Z}_i)}\right) = \theta_1 + \theta_2 Z_{i2}, \tag{10}$$

where $X_{i2} \sim N(0,1)$ and $Z_{i2} \sim \mathcal{U}[0,1]$. The models for infection and immunity are allowed to share one covariate. An i.i.d. sample of size $n$ of the vector $\mathcal{O}$ is generated from this model, and for each individual $i$ we get a realization $\mathcal{O}_i = (y_i, s_i, \mathbf{x}_i, \mathbf{z}_i)$, where $s_i$ is considered as unknown if $y_i = 0$. The MLE $\widehat{\beta}_n$ of $\beta = (\beta_1, \beta_2, \beta_3)^\top$ is obtained from this incomplete dataset by using the procedure described in Section 2. As a by-product of the method, an estimate of the nuisance parameter $\theta = (\theta_1, \theta_2)^\top$ is also obtained, but will not be reported here.

The properties of the proposed simultaneous confidence bands (7) and (8) are evaluated for several sample sizes ($n = 500, 1000, 1500$) and values of the percentage of immunes in the sample (25%, 50%). We also consider different values for the proportion of infected individuals among the susceptibles (30% and 70%), and several confidence levels for the bands ($1 - \alpha = 0.90, 0.95, 0.99$). The desired proportions of immunes and infected are obtained by choosing appropriate values for $\beta$ and $\theta$. The following values are considered for $\beta$:

- model $\mathcal{M}_1$: $\beta = (-1, 1.5, -0.4)^\top$. Using these values, approximately 30% of the susceptibles are infected.
- model $\mathcal{M}_2$: $\beta = (0.5, 1, -1)^\top$. Approximately 70% of the susceptibles are infected.

For each configuration `confidence level` $\times$ `sample size` $\times$ `percentage of immunes` $\times$ `percentage of infected among susceptibles` of the design parameters, $N = 1000$ samples are obtained. For each of these $N$ samples, we compute the simultaneous confidence bands (7) and (8). Then, for each type of band, we obtain the empirical coverage probability, defined as the proportion of the $N$ bands that contain the whole function $p(\mathbf{x})$. We also evaluate the precision of the bands (7) and (8) in terms of their widths. Their respective mean width (averaged over a grid of values of $\mathbf{x}$ and over the $N$ simulations) are computed. Similarly, the averaged median width and averaged first and third quartiles of the widths are obtained for each of the bands (7) and (8). All the results are reported in the tables 1 (for the model $\mathcal{M}_1$) and 2 (for the model $\mathcal{M}_2$) below. There, "Cov" indicates the empirical coverage probability. The column "Width" reports the averaged values of: the mean band width ($+$), the median band width ($\star$), the first ($\dagger$) and third ($\mp$) quartiles of the band widths.

**Table 1. Coverage probabilities and widths of the confidence bands (7) and (8) in the model $\mathcal{M}_1$, based on 1000 replicates.**

| | | Method 1 (band (7)) | | | | Method 2 (band (8)) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 25% | | 50% | | 25% | | 50% | |
| $1-\alpha$ | n | Cov | Width | Cov | Width | Cov | Width | Cov | Width |
| 0.99 | 500 | 0.946 | $0.217^{+}$ | 0.879 | $0.325^{+}$ | 0.999 | $0.734^{+}$ | 0.998 | $0.885^{+}$ |
| | | | $0.082^{\dagger}$ | | $0.138^{\dagger}$ | | $0.627^{\dagger}$ | | $0.827^{\dagger}$ |
| | | | $0.174^{\star}$ | | $0.279^{\star}$ | | $0.830^{\star}$ | | $0.954^{\star}$ |
| | | | $0.338^{\mp}$ | | $0.495^{\mp}$ | | $0.898^{\mp}$ | | $0.983^{\mp}$ |
| | 1000 | 0.931 | $0.134^{+}$ | 0.847 | $0.180^{+}$ | 0.994 | $0.593^{+}$ | 0.994 | $0.596^{+}$ |
| | | | $0.031^{\dagger}$ | | $0.053^{\dagger}$ | | $0.282^{\dagger}$ | | $0.325^{\dagger}$ |
| | | | $0.087^{\star}$ | | $0.131^{\star}$ | | $0.775^{\star}$ | | $0.727^{\star}$ |
| | | | $0.220^{+}$ | | $0.292^{+}$ | | $0.877^{+}$ | | $0.856^{+}$ |
| | 1500 | 0.837 | $0.102^{+}$ | 0.865 | $0.143^{+}$ | 0.996 | $0.588^{+}$ | 0.991 | $0.533^{+}$ |
| | | | $0.018^{\dagger}$ | | $0.031^{\dagger}$ | | $0.332^{\dagger}$ | | $0.196^{\dagger}$ |
| | | | $0.062^{\star}$ | | $0.092^{\star}$ | | $0.736^{\star}$ | | $0.690^{\star}$ |
| | | | $0.172^{+}$ | | $0.239^{+}$ | | $0.827^{+}$ | | $0.823^{+}$ |
| 0.95 | 500 | 0.854 | $0.169^{+}$ | 0.797 | $0.232^{+}$ | 0.989 | $0.667^{+}$ | 0.997 | $0.797^{+}$ |
| | | | $0.045^{\dagger}$ | | $0.082^{\dagger}$ | | $0.526^{\dagger}$ | | $0.684^{\dagger}$ |
| | | | $0.117^{\star}$ | | $0.178^{\star}$ | | $0.761^{\star}$ | | $0.908^{\star}$ |
| | | | $0.274^{+}$ | | $0.365^{+}$ | | $0.852^{+}$ | | $0.966^{+}$ |
| | 1000 | 0.791 | $0.108^{+}$ | 0.636 | $0.139^{+}$ | 0.989 | $0.530^{+}$ | 0.985 | $0.538^{+}$ |
| | | | $0.019^{\dagger}$ | | $0.031^{\dagger}$ | | $0.187^{\dagger}$ | | $0.272^{\dagger}$ |
| | | | $0.063^{\star}$ | | $0.089^{\star}$ | | $0.698^{\star}$ | | $0.622^{\star}$ |
| | | | $0.181^{\mp}$ | | $0.230^{\mp}$ | | $0.829^{\mp}$ | | $0.784^{\mp}$ |
| | 1500 | 0.836 | $0.081^{+}$ | 0.731 | $0.106^{+}$ | 0.988 | $0.518^{+}$ | 0.986 | $0.448^{+}$ |
| | | | $0.012^{\dagger}$ | | $0.020^{\dagger}$ | | $0.216^{\dagger}$ | | $0.135^{\dagger}$ |
| | | | $0.046^{\star}$ | | $0.064^{\star}$ | | $0.684^{\star}$ | | $0.547^{\star}$ |
| | | | $0.137^{\mp}$ | | $0.177^{\mp}$ | | $0.784^{\mp}$ | | $0.722^{\mp}$ |
| 0.90 | 500 | 0.846 | $0.142^{+}$ | 0.716 | $0.197^{+}$ | 0.989 | $0.607^{+}$ | 0.984 | $0.757^{+}$ |
| | | | $0.035^{\dagger}$ | | $0.059^{\dagger}$ | | $0.457^{\dagger}$ | | $0.614^{\dagger}$ |
| | | | $0.093^{\star}$ | | $0.141^{\star}$ | | $0.694^{\star}$ | | $0.875^{\star}$ |
| | | | $0.230^{\mp}$ | | $0.316^{\mp}$ | | $0.793^{\mp}$ | | $0.946^{\mp}$ |
| | 1000 | 0.788 | $0.092^{+}$ | 0.744 | $0.121^{+}$ | 0.990 | $0.485^{+}$ | 0.977 | $0.469^{+}$ |
| | | | $0.015^{\dagger}$ | | $0.025^{\dagger}$ | | $0.160^{\dagger}$ | | $0.207^{\dagger}$ |
| | | | $0.053^{\star}$ | | $0.075^{\star}$ | | $0.624^{\star}$ | | $0.547^{\star}$ |
| | | | $0.154^{+}$ | | $0.199^{+}$ | | $0.775^{+}$ | | $0.718^{+}$ |
| | 1500 | 0.800 | $0.073^{+}$ | 0.607 | $0.093^{+}$ | 0.983 | $0.462^{+}$ | 0.966 | $0.410^{+}$ |
| | | | $0.010^{\dagger}$ | | $0.016^{\dagger}$ | | $0.127^{\dagger}$ | | $0.111^{\dagger}$ |
| | | | $0.039^{\star}$ | | $0.054^{\star}$ | | $0.624^{\star}$ | | $0.479^{\star}$ |
| | | | $0.123^{+}$ | | $0.156^{+}$ | | $0.741^{+}$ | | $0.670^{+}$ |

Note: Cov: coverage probability. Width: $^{+}$: mean, $^{\dagger}$: $1^{\text{st}}$ quartile, $^{\star}$: median, $^{\mp}$: $3^{\text{rd}}$ quartile. 25% and 50% refer to the percentage of immunes in the samples.

**Table 2. Coverage probabilities and widths of the confidence bands (7) and (8) in the model $\mathcal{M}_2$, based on 1000 replicates.**

| | | Method 1 (band (7)) | | | | Method 2 (band (8)) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 25% | | 50% | | 25% | | 50% | |
| $1-\alpha$ | n | Cov | Width | Cov | Width | Cov | Width | Cov | Width |
| 0.99 | 500 | 0.831 | $0.362^{+}$ | 0.690 | $0.479^{+}$ | 0.992 | $0.692^{+}$ | 0.996 | $0.830^{+}$ |
| | | | $0.261^{\dagger}$ | | $0.339^{\dagger}$ | | $0.474^{\dagger}$ | | $0.696^{\dagger}$ |
| | | | $0.343^{\star}$ | | $0.435^{\star}$ | | $0.816^{\star}$ | | $0.927^{\star}$ |
| | | | $0.468^{\mp}$ | | $0.636^{\mp}$ | | $0.889^{\mp}$ | | $0.974^{\mp}$ |
| | 1000 | 0.775 | $0.219^{+}$ | 0.685 | $0.301^{+}$ | 0.982 | $0.526^{+}$ | 0.995 | $0.664^{+}$ |
| | | | $0.139^{\dagger}$ | | $0.201^{\dagger}$ | | $0.282^{\dagger}$ | | $0.391^{\dagger}$ |
| | | | $0.214^{\star}$ | | $0.276^{\star}$ | | $0.635^{\star}$ | | $0.812^{\star}$ |
| | | | $0.292^{+}$ | | $0.405^{+}$ | | $0.729^{+}$ | | $0.884^{+}$ |
| | 1500 | 0.735 | $0.171^{+}$ | 0.679 | $0.226^{+}$ | 0.983 | $0.350^{+}$ | 0.987 | $0.552^{+}$ |
| | | | $0.101^{\dagger}$ | | $0.143^{\dagger}$ | | $0.180^{\dagger}$ | | $0.273^{\dagger}$ |
| | | | $0.169^{\star}$ | | $0.213^{\star}$ | | $0.368^{\star}$ | | $0.677^{\star}$ |
| | | | $0.232^{+}$ | | $0.306^{+}$ | | $0.507^{+}$ | | $0.774^{+}$ |
| 0.95 | 500 | 0.734 | $0.274^{+}$ | 0.574 | $0.362^{+}$ | 0.994 | $0.598^{+}$ | 0.986 | $0.767^{+}$ |
| | | | $0.179^{\dagger}$ | | $0.231^{\dagger}$ | | $0.366^{\dagger}$ | | $0.599^{\dagger}$ |
| | | | $0.259^{\star}$ | | $0.317^{\star}$ | | $0.712^{\star}$ | | $0.871^{\star}$ |
| | | | $0.366^{+}$ | | $0.501^{+}$ | | $0.800^{+}$ | | $0.934^{+}$ |
| | 1000 | 0.662 | $0.172^{+}$ | 0.607 | $0.226^{+}$ | 0.973 | $0.443^{+}$ | 0.984 | $0.581^{+}$ |
| | | | $0.100^{\dagger}$ | | $0.136^{\dagger}$ | | $0.205^{\dagger}$ | | $0.301^{\dagger}$ |
| | | | $0.167^{\star}$ | | $0.207^{\star}$ | | $0.528^{\star}$ | | $0.714^{\star}$ |
| | | | $0.235^{\mp}$ | | $0.312^{\mp}$ | | $0.639^{\mp}$ | | $0.805^{\mp}$ |
| | 1500 | 0.638 | $0.136^{+}$ | 0.618 | $0.174^{+}$ | 0.964 | $0.287^{+}$ | 0.975 | $0.465^{+}$ |
| | | | $0.073^{\dagger}$ | | $0.098^{\dagger}$ | | $0.135^{\dagger}$ | | $0.205^{\dagger}$ |
| | | | $0.133^{\star}$ | | $0.163^{\star}$ | | $0.289^{\star}$ | | $0.552^{\star}$ |
| | | | $0.189^{\mp}$ | | $0.243^{\mp}$ | | $0.427^{\mp}$ | | $0.675^{\mp}$ |
| 0.90 | 500 | 0.631 | $0.239^{+}$ | 0.533 | $0.311^{+}$ | 0.983 | $0.543^{+}$ | 0.986 | $0.728^{+}$ |
| | | | $0.150^{\dagger}$ | | $0.189^{\dagger}$ | | $0.307^{\dagger}$ | | $0.539^{\dagger}$ |
| | | | $0.222^{\star}$ | | $0.266^{\star}$ | | $0.644^{\star}$ | | $0.835^{\star}$ |
| | | | $0.323^{\mp}$ | | $0.437^{\mp}$ | | $0.746^{\mp}$ | | $0.911^{\mp}$ |
| | 1000 | 0.588 | $0.151^{+}$ | 0.511 | $0.198^{+}$ | 0.950 | $0.403^{+}$ | 0.977 | $0.524^{+}$ |
| | | | $0.083^{\dagger}$ | | $0.113^{\dagger}$ | | $0.172^{\dagger}$ | | $0.251^{\dagger}$ |
| | | | $0.145^{\star}$ | | $0.179^{\star}$ | | $0.472^{\star}$ | | $0.632^{\star}$ |
| | | | $0.209^{+}$ | | $0.277^{+}$ | | $0.589^{+}$ | | $0.745^{+}$ |
| | 1500 | 0.570 | $0.120^{+}$ | 0.517 | $0.153^{+}$ | 0.941 | $0.255^{+}$ | 0.960 | $0.429^{+}$ |
| | | | $0.062^{\dagger}$ | | $0.082^{\dagger}$ | | $0.109^{\dagger}$ | | $0.174^{\dagger}$ |
| | | | $0.117^{\star}$ | | $0.141^{\star}$ | | $0.251^{\star}$ | | $0.503^{\star}$ |
| | | | $0.168^{+}$ | | $0.215^{+}$ | | $0.385^{+}$ | | $0.637^{+}$ |

From these tables, we note that the empirical coverage probabilities of the Scheffé-type bands are always below the nominal confidence level. As mentioned above, this may be due to the fact that the distribution of $\sup_{\mathbf{x}\in\mathcal{X}}(\sqrt{n}(\widehat{\beta}_n - \beta)^\top \mathbf{x})^2/\mathbf{x}^\top \widehat{\Sigma}_{\beta,n}\mathbf{x}$ is badly approximated by a $\chi^2$ distribution. The coverage probabilities of the alternative simulation-based confidence bands are much higher (they are almost always above the nominal level), which results in bands having somewhat large widths when the nominal confidence level is very high and/or the proportion of immunes is high (50%). This conservative feature was also observed by Li *et al.* (2010) in the semiparametric logistic regression model. Note also that when the sample size increases, the coverage probabilities tend to decrease, which is due to the fact that the width of the bands decreases. This decrease in the coverage probabilities is marked for the Scheffé-type band but almost negligible for the simulation-based confidence band. Based on these remarks, we recommend to use the simulation-based approach to construct simultaneous confidence bands in the zero-inflated regression model (2)-(3).

## 5. Conclusions and perspectives

In this paper, we have shown that simultaneous confidence bands for the regression function $p(\mathbf{x})$ can be constructed in the zero-inflated Bernoulli regression model. Simultaneous bands have been investigated in a variety of regression models but as far as we know, the present work constitutes the first attempt to construct and evaluate such bands in a zero-inflated regression model. We have first adapted the Scheffé-type confidence band to the ZIB model (2)-(3). Our simulations suggest that this type of bands enjoys poor coverage probabilities in finite samples. Thus, we have proposed an alternative construction which relies on Monte Carlo simulations to approximate the distribution of the supremum of a Gaussian field indexed by the regressors. This second type of band seems much more satisfactory.
Now, several further issues about simultaneous confidence bands in the ZIB regression model (2)-(3) deserve attention. First, one may wish to use the proposed bands to construct confidence regions for the "effective dose" (namely, the values of the covariate $\mathbf{X}$ that will produce a given response probability). One may also wish to apply the proposed confidence bands to compare two zero-inflated Bernoulli models. Finally, another stimulating topic for future research deals with the construction of simultaneous confidence bands in the model (2)-(3) in a high-dimensional setting. Several recent articles (Huang *et al.* (2008), Meier *et al.* (2008)) have considered estimation in the logistic model (without zero-inflation) when the predictor dimension is much larger than the sample size (this problem arises for example in genetic studies where high-dimensional data are generated using microarray technologies). Extending the construction of simultaneous confidence bands for the model (2)-(3) to a high-dimensional setting constitutes a further non-trivial topic of interest.

## Acknowledgements

This research was supported by AIRES-Sud (AIRES-Sud is a program from the French Ministry of Foreign and European Affairs, implemented by the "Institut de Recherche pour le Développement", IRD-DSF), by the "Service de Coopération et d'Action Culturelle" of the French Embassy in Senegal, and by Edulink (program 9-ACP-RPR-118♯18). The authors also acknowledge grants from the "Ministère de la Recherche Scientifique" of Senegal.

## References

Billingsley, P., 1968. Logistic regression models. Convergence of Probability Measures. Wiley: New York.

Brand, R.J., Pinnock, D.E. and Jackson, K.L., 1973. Large sample confidence bands for the logistic response curve and its inverse. American Statistician **27**, 157–160.

Dietz, E. and Böhning, D., 2000. On estimation of the Poisson parameter in zero-modified Poisson models. Computational Statistics & Data Analysis **34**, 441-459.

Diop, A., Diop, A. and Dupuy, J.-F., 2011. Maximum likelihood estimation in the logistic regression model with a cure fraction. Electronic Journal of Statistics **5**, 460–483.

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

Fleming, T.R. and Harrington, D.P., 1991. Counting processes and survival analysis. Wiley: New York.

Gouriéroux, C. and Monfort, A., 1981. Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. Journal of Econometrics **17**, 83–97.

Hall, D.B., 2000. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics **56**, 1030–1039.

Hauck, W.W., 1983. A note on confidence bands for the logistic response curve. American Statistician **37**, 158–160.

Hilbe, J.M., 2009. Logistic regression models. Chapman & Hall: Boca Raton.

Hosmer, D.W. and Lemeshow, S., 2000. Applied logistic regression. Wiley: New York.

Huang, J., Ma, S. and Zhang, C.H., 2008. The iterated lasso for high-dimensional logistic regression. Technical report No. 392, The University of Iowa.

Kelley, M.E. and Anderson, S.J., 2008. Zero inflation in ordinal data: incorporating susceptibility to response through the use of a mixture model. Statistics in Medicine **27**, 3674–3688.

Lam, K.F., Xue, H. and Cheung, Y.B., 2006. Semiparametric analysis of zero-inflated count data. Biometrics **62**, 996–1003.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics **34**, 1–14.

Lee, A.H., Wang, K., Scott, J.A., Yau, K.K.W. and McLachlan, G.J., 2006. Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. Statistical Methods in Medical Research **15**, 47-61.

Li, C.S., 2011. A lack-of-fit test for parametric zero-inflated Poisson models. Journal of Statistical Computation and Simulation **81**, 1081–1098.

Li, J., Zhang, C., Doksum, K.A. and Nordheim, E.V., 2010. Simultaneous confidence intervals for semiparametric logistic regression and confidence regions for the multi-dimensional effective dose. Statistica Sinica **20**, 637–659.

Liu, W., 2011. Simultaneous inference in regression. Chapman & Hall: Boca Raton.

Meier, L., van de Geer, S. and Bühlmann, P., 2008. The group Lasso for logistic regression. Journal of the Royal Statistical Society. Series B **70**, 53–71.

Moghimbeigi, A., Eshraghian, M.R., Mohammad, K. and McArdle, B., 2009. A score test for zero-inflation in multilevel count data. Computational Statistics & Data Analysis **53**, 1239–1248.

Piegorsch, W.W. and Casella, G., 1988. Confidence bands for logistic regression with restricted predictor variables. Biometrics **44**, 739–750.

Ridout, M., Demétrio, C.G.B. and Hinde, J., 1998. Models for counts data with many zeros. Proceedings of the XIXth International Biometric Conference, Cape Town, Invited Papers, 179–192.

Rao, C.R., 1973. Linear statistical inference and its applications, 2nd Edition. Wiley.

Sun, J., Loader, C. and McCormick, W.P., 2000. Confidence bands in generalized linear models. The Annals of Statistics **28**, 429–460.

Xiang, L., Lee, A.H., Yau, K.K.W. and McLachlan, G.J., 2007. A score test for overdispersion in zero-inflated Poisson mixed regression model. Statistics in Medicine **26**, 1608–1622.