Marginal effects for zero-inflated semi-continuous data

Agnondji GNON SIYA[1], Essoham ALI[2,3*], Aliou DIOP[1†]

[1] Gaston Berger University, LERSTAD, Saint-Louis, Senegal.
[2]Institut de Mathématiques Appliquées, UCO, 49000, Angers, France.
[3]Univ Bretagne Sud, CNRS UMR 6205, LMBA, Vannes, France.

*Corresponding author(s). E-mail(s): essoham.ali@univ-ubs.fr ;
Contributing authors: siya.agnondji@ugb.edu.sn ; aliou.diop@ugb.edu.sn;
†These authors contributed equally to this work.

Abstract

Zero-inflated semi-continuous data (ZISCD) are common in various fields such as medicine, economics, and social sciences, where the outcome distribution consists of a point mass at zero and a positive continuous distribution. This work uses marginalized regression models (MZIG and MZILN) to analyze semi-continuous data with zero inflation. These models allow for a direct interpretation of covariate effects on the marginal mean, accounting for distinct mechanisms generating structural zeros and continuous positive outcomes. The approach is validated through simulations and the analysis of real data, demonstrating a significant improvement in the interpretation and accuracy of estimates. {This paper introduces two novel marginalized regression models, MZIG and MZILN, that address key limitations in existing methods by enabling a clear interpretation of covariate effects on the marginal mean, while offering robust statistical properties validated through simulations and real-world applications}. This work fills a significant gap in the literature by offering a comprehensive approach to the estimation and interpretation of the MZIG and MZILN models.

Keywords: Zero-inflated ; Gamma model, Log-Normal model; Semi-continuous data; Marginalized regression; Asymptotic properties; Parameter estimation.

1 Introduction

Statistics are a fundamental pillar of applied sciences, with practical applications across all fields, from exact sciences to everyday life. Their importance lies in the analysis of data sets, which are essential for drawing inferences and reaching accurate conclusions. Among various types of data, semi-continuous data sets (SCDS) hold a special place. These data sets combine a point mass at zero and a positively skewed distribution, representing non-negative values frequently encountered in various sectors.

SCDS are utilized in diverse fields such as economics, medicine, engineering, and social sciences. For instance, [4] proposed methods to analyze meteorological data where the absence of rainfall is coded as zero, while precipitation is represented by positive values. Similarly, [3] compared models for analyzing healthcare needs, where some individuals have no claims (zero value), while others have claims with positive values. Other studies, such as those by [18] on hospitalization costs or [8] on physical activity levels among the elderly, also highlight the relevance of SCDS. Finally, [6] analyzed outcomes of emergency medical care for out-of-hospital cardiac arrest, where the majority of cases result in death (zero value). These examples demonstrate that SCDS, often characterized by a high proportion of zeros, are prevalent and are sometimes referred to as zero-inflated (ZI) data.

Zero-inflated semi-continuous data (ZISCD) present a complex bimodal structure, with structural zeros and continuous positive outcomes. These data are common in fields such as medicine, social sciences, and agriculture. For example, in agriculture, zeros may indicate a lack of production due to factors like drought, while positive outcomes reflect productive yields. Analyzing this dual structure requires advanced statistical approaches to ensure robust inferences.

Zero-inflated models, such as the zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB), are commonly used to model excess zeros in count data ([7]; [15]). Extensions of these models to semi-continuous data have led to models like the zero-inflated Gamma (ZIG) and zero-inflated Log-normal (ZILN) [12]. While these models effectively capture the bimodal nature of ZISCD, they do not always provide a clear interpretation of the effects of covariates on the marginal mean, which remains a significant challenge for applied researchers.

Marginal models play a central role in the analysis of zero-inflated data, allowing covariate effects to be directly incorporated into the distributions. Among notable contributions, the marginal zero-inflated Poisson regression model (MZIP) proposed by [9] serves as an important reference, with numerous extensions developed subsequently, including those by [10], [11], [2], and [17]. These works have significantly enriched the theoretical and practical framework of marginal zero-inflated models for count data.

Recently [16] proposed a marginalized two-part model (MTP) for semi-continuous data, enabling direct modeling of the marginal effects of covariates on the combined mean of the two parts. This model accounts for both the point mass at zero and the positively skewed distribution. However, while this approach represents a significant advancement, it has limitations, particularly the lack of a theoretical framework for analyzing the asymptotic properties of the maximum likelihood estimators (MLE). These properties are crucial for validating statistical inferences, especially when evaluating the effects of covariates on the marginal mean.

To address these limitations, this article introduces two new marginalized models tailored for SCDS: the marginalized zero-inflated Gamma (MZIG) model and the marginalized zero-inflated Log-normal (MZILN) model. These innovative models allow for a direct interpretation of covariate effects on the marginal mean while incorporating distinct mechanisms for structural zeros and continuous positive outcomes. By employing maximum likelihood estimation, these models ensure asymptotic robustness and enhanced precision in parameter estimation.

The contribution of this article is twofold: first, to rigorously study the asymptotic properties of the estimated parameters, and second, to demonstrate the practical utility of the proposed models through simulations and applications to real-world data. These models provide a comprehensive and interpretable solution for analyzing complex data where zero inflation and continuous distributions coexist.

These studies highlight the need for robust statistical approaches to analyze complex data while providing a clear interpretation of the effects of covariates. In this context, we review existing models to lay the groundwork for the methodological proposals in this article.

The organization of this study is straightforward and is structured as follows: Section 2 reviews the ZIG and ZILN models, their characteristics, and the motivations for a marginalized extension. Section 3 provides detailed descriptions of the MZIG and MZILN models, including their structures, estimation methods, and asymptotic properties. Sections 4 and 5 present numerical illustrations and practical applications to real data, respectively. Finally, Section 6 offers a discussion of the results and outlines perspectives for future research. Technical proofs are postponed to an appendix.

2 Literature review

In this section, we only review the most important formulas of the ZIG and ZILN models, we describe maximum likelihood estimation in this models.

2.1 ZIG and ZILN models for structural zeros

The Zero-Inflated Gamma (ZIG) and Zero-Inflated Log-Normal (ZILN) models are two-component frameworks designed for datasets with an excess of zeros. The first component models membership in a group where the responses are structural zeros. The second component captures the other group, where positive responses follow a Gamma or Log-Normal distribution. Let Y_i denote the random variable representing a zero-inflated count response. Based on [12], the distribution of Y_i is given by:

$$P(Y_i = y_i) = \begin{cases} 1 - \phi_i + \phi_i f(0 \mid \mu_i, \nu) & \text{if } y_i = 0, \\ \phi_i f(y_i \mid \mu_i, \nu) & \text{if } y_i > 0. \end{cases}$$

Here, $f(y_i \mid \mu_i, \nu)$ is the probability density function for the count distribution, which is Gamma in the ZIG model and Log-Normal in the ZILN model. The parameter ϕ_i represents the probability of belonging to a group where only structural zeros are observed. Zero observations, therefore, arise from both structural zeros in this group and random zeros in the other group, with group membership being unknown.

This two-component structure enables the ZIG and ZILN models to account for zero-inflated while capturing the variability in positive values. Such models are particularly suited to datasets with a significant proportion of zeros, where the positive values follow a continuous distribution like Gamma or Log-Normal.

In both models, the parameter ϕ_i, which governs excess zeros, is typically modeled using a logit link, while the parameter μ_i, characterizing the mean of the Gamma or Log-Normal distribution, is modeled using a log link, and ν is a dispersion parameter. Specifically:

$$\text{logit}(\phi_i) = \boldsymbol{\beta}^\top \mathbf{X}_i \quad \text{and} \quad \log(\mu_i) = \boldsymbol{\gamma}^\top \mathbf{W}_i, \tag{1}$$

where $\mathbf{X}_i = (1, X_{i1}, \ldots, X_{ip})^\top$ and $\mathbf{W}_i = (1, W_{i1}, \ldots, W_{iq})^\top$ are vectors of covariates. The parameter vectors $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^\top$ contain the regression coefficients, where β_1 and γ_1 are the intercepts.

2.2 Definition and effects of covariates in marginalized models: motivation for flexibility

In the context of the Marginalized Zero-Inflated Gamma (MZIG) and Marginalized Zero-Inflated Log-Normal (MZILN) models, let

$$q_i := \mathbb{E}(Y_i \mid \mathbf{X}_i)$$

represent the marginal mean of the count response, which includes both structural or random zeros and positive values. These models are based on a mixture structure: the first component accounts for the excess zeros, while the second models positive responses with either a Gamma or Log-Normal distribution. Unlike traditional models, such as Zero-Inflated Poisson (ZIP) or Zero-Inflated Negative Binomial (ZINB) models, which linearly model the logarithm of the mean of count data, the MZIG and MZILN models directly link the logarithm of the marginal mean, $\log(q_i)$, to a linear predictor. Specifically:

$$\text{logit}(\phi_i) = \boldsymbol{\beta}^\top \mathbf{X}_i \quad \text{and} \quad \log(q_i) = \boldsymbol{\gamma}^\top \mathbf{X}_i, \tag{2}$$

where q_i corresponds to the mean of the overall mixture model, and ϕ_i represents the probability of excess zeros. This formulation allows for a more direct and intuitive interpretation of how covariates influence the overall distribution of the outcome, accounting for both the zero-inflation and the continuous distribution of positive values.

A key motivation behind the development of these marginalized models is to provide better interpretability of the covariate effects on the marginal mean of the data. Traditional models, like ZIG and ZILN, struggle to fully capture datasets where zero-inflated and continuous outcomes arise from different underlying processes. In many practical applications, researchers are often more interested in understanding the overall effect of covariates on the distribution of the outcome, including both the zero-inflated and continuous parts of the data. Our proposed models allow for direct interpretation of these effects on the marginal mean, offering a more comprehensive understanding compared to traditional approaches. This is particularly valuable when the mechanisms driving zero-inflation and continuous outcomes are distinct.

The MZIG and MZILN models are flexible, providing various options for the link functions that relate covariates to both the zero-inflated probabilities and the continuous outcome distribution. By separately modeling these two components, our approach more accurately captures the underlying complexity of the data. The MZIG model extends the zero-inflated Gamma distribution, while the MZILN model

extends the zero-inflated Log-Normal distribution. Both models offer considerable flexibility, making them suitable for a wide range of applications where zero-inflation and continuous data coexist.

Moreover, the use of Maximum Likelihood Estimation (MLE) to estimate the parameters of both models provides a robust statistical framework. MLE is asymptotically efficient, ensuring that estimators are consistent and asymptotically normal, which is crucial for inference and hypothesis testing. This estimation method is particularly advantageous in practical applications, as it enables researchers to obtain reliable standard errors for the estimated parameters, improving the accuracy and interpretability of the results.

In summary, the MZIG and MZILN models provide a powerful framework for handling datasets with significant zero-inflation and continuous outcomes. These models are especially well-suited for applications in healthcare, economics, and social sciences, where both excess zeros and continuous values are commonly observed. By explicitly modeling both components of the data, the MZIG and MZILN models offer a more accurate, interpretable, and flexible approach to understanding the data generation processes in complex datasets.

In this section, we introduced the MZIG and MZILN models, which allow for a direct interpretation of the effects of covariates on the marginal mean q_i and the probability of excess zeros ϕ_i. An interesting aspect of these models is their flexibility in the choice of link functions, which helps better capture the underlying complexity of the data.

A simple but often effective approach is to use a single covariate to model both ϕ_i and q_i. This method simplifies the analysis while exploring the overall effect of this covariate on the two components of the model. It offers a coherent and intuitive interpretation of the impact of a key factor on the probability of excess zeros ϕ_i and on the marginal mean q_i of the response. This approach is particularly useful when the covariate in question plays a central role in the phenomenon being studied. However, it implicitly assumes that the interactions or effects of other covariates are negligible or irrelevant, which constitutes a strong assumption in the model.

Remark 1. *Using a single covariate to model both ϕ_i and q_i in the MZIG and MZILN models simplifies the analysis while exploring the overall effect of this covariate on the two components of the model. This approach provides a coherent and intuitive interpretation of the impact of a key factor on the probability of excess zeros ϕ_i and the marginal mean q_i of the response. It is particularly useful when the covariate under study plays a central role in the phenomenon being analyzed. However, it implicitly assumes that the interactions or effects of other covariates are negligible or irrelevant, which constitutes a strong modeling assumption.*

However, these classical models have limitations, particularly in interpreting the marginal effects of covariates. To address these challenges, we propose the MZIG and MZILN models, which are described in detail in the following section.

3 Proposed models and estimation

In this section, we present and discuss the proposed models for analyzing the data, including their structural definitions and estimation procedures.

3.1 MZIG model formulation and estimation

To account for the effects of covariates on the original data scale in an interpretable way, we define the Marginalized Zero-Inflated Gamma (MZIG) model as follows:

$$q_i = \mu_i \phi_i, \quad \text{where} \quad \mu_i = \frac{q_i}{\phi_i}.$$

Here, μ_i represents the conditional mean of the positive component. The marginal probability of an observation y_i in the MZIG model is given by:

$$P(Y_i = y_i) = \begin{cases} 1 - \phi_i & \text{if } y_i = 0, \\ \phi_i \cdot \frac{1}{\Gamma(\nu)} y_i^{\nu-1} \left(\frac{\nu\phi_i}{q_i}\right)^{\nu} \exp\left(-\frac{\nu\phi_i y_i}{q_i}\right) & \text{if } y_i > 0, \end{cases} \tag{3}$$

where ϕ_i is the probability of observing a zero (the zero-inflated component), $\Gamma(\nu)$ is the Gamma function with parameter ν (the shape of the Gamma distribution), q_i is the marginal mean of the model, which is influenced by the covariates. The MZIG model extends the family of zero-inflated models by directly linking covariates to both the probability of zero-inflated and the marginal mean on the original scale, providing enhanced interpretability and flexibility compared to traditional Gamma or hurdle models.

To better understand the practical application of the MZIG model, consider the example of household energy consumption. A large proportion of households may have zero consumption (structurally or randomly), while those that consume energy follow a Gamma distribution. In this context, ϕ_i represents the probability of having zero energy consumption, capturing the zero-inflated component, while q_i determines the marginal mean of positive energy consumption. The MZIG model effectively captures these two dynamics and provides a direct interpretation of how factors such as household size or the presence of energy-efficient appliances influence average consumption.

The parameters ϕ_i and q_i are related to the covariates via the relations given in equation (2). Specifically, ϕ_i is determined by a logit link, while $\log(q_i)$ is related to the covariates by a linear predictor, allowing for a flexible capture of the effect of covariates on both the zero-inflation and the continuous component of positive values.

Let $\mathbf{\Phi} := (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\nu})^\top$ denote the set of all unknown parameters. Then, the likelihood function of $\mathbf{\Phi}$ based on observations $(Y_i, X_i), i = 1, \cdots, n$ is calculated as:

$$L_n(\mathbf{\Phi}) = \left[\prod_{Y_i=0} \frac{1}{1+e^{\boldsymbol{\beta}^\top \mathbf{X}_i}} \prod_{Y_i>0} \frac{e^{\boldsymbol{\beta}^\top \mathbf{X}_i}}{1+e^{\boldsymbol{\beta}^\top \mathbf{X}_i}}\right] \times \left[\prod_{Y_i>0} \frac{1}{\Gamma(\nu)} Y_i^{\nu-1} \left(\nu \frac{e^{\boldsymbol{\beta}^\top \mathbf{X}_i}}{e^{\boldsymbol{\gamma}^\top \mathbf{X}_i}(1+e^{\boldsymbol{\beta}^\top \mathbf{X}_i})}\right)^{\nu} \exp\left(-\nu \frac{Y_i e^{\boldsymbol{\beta}^\top \mathbf{X}_i}}{e^{\boldsymbol{\gamma}^\top \mathbf{X}_i}(1+e^{\boldsymbol{\beta}^\top \mathbf{X}_i})}\right)\right].$$

Using (3) and some algebra, the loglikelihood $\ell_n(\mathbf{\Phi}) = \log L_n(\mathbf{\Phi})$ can be written as :

$$\begin{aligned}\ell_n(\mathbf{\Phi}) = &\sum_{i:Y_i=0} -\log(1+e^{\boldsymbol{\beta}^\top \mathbf{X}_i}) + \sum_{i:Y_i>0} \left(\boldsymbol{\beta}^\top \mathbf{X}_i - \log(1+e^{\boldsymbol{\beta}^\top \mathbf{X}_i})\right) \\ &+ \sum_{i:Y_i>0} \left[-\log(\Gamma(\nu)) + (\nu-1)\log(Y_i) + \nu \log\left(\frac{\nu e^{\boldsymbol{\beta}^\top \mathbf{X}_i}}{e^{\boldsymbol{\gamma}^\top \mathbf{X}_i}(1+e^{\boldsymbol{\beta}^\top \mathbf{X}_i})}\right) - \nu \frac{Y_i e^{\boldsymbol{\beta}^\top \mathbf{X}_i}}{e^{\boldsymbol{\gamma}^\top \mathbf{X}_i}(1+e^{\boldsymbol{\beta}^\top \mathbf{X}_i})}\right].\end{aligned} \tag{4}$$

From this log-likelihood, the Estimating Score Function (ESF) can be expressed as:

$$U_{F,n}(\mathbf{\Phi}) = \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\mathbf{\Phi})}{\partial \mathbf{\Phi}} = \frac{1}{\sqrt{n}} \begin{pmatrix} \frac{\partial \ell_n(\mathbf{\Phi})}{\partial \boldsymbol{\beta}} \\ \frac{\partial \ell_n(\mathbf{\Phi})}{\partial \boldsymbol{\gamma}} \\ \frac{\partial \ell_n(\mathbf{\Phi})}{\partial \nu} \end{pmatrix} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_i(\mathbf{\Phi}),$$

where

$$S_i(\mathbf{\Phi}) = \frac{\partial \ell_i(\mathbf{\Phi})}{\partial \mathbf{\Phi}} = \left(S_{i1}^\top(\mathbf{\Phi}), S_{i2}^\top(\mathbf{\Phi}), S_{i3}^\top(\mathbf{\Phi})\right)^\top.$$

Specifically, the components of $S_i(\mathbf{\Phi})$ are given by:

$$S_{i1}(\mathbf{\Phi}) = \frac{\partial \ell_i(\mathbf{\Phi})}{\partial \boldsymbol{\beta}}, \quad S_{i2}(\mathbf{\Phi}) = \frac{\partial \ell_i(\mathbf{\Phi})}{\partial \boldsymbol{\gamma}}, \quad S_{i3}(\mathbf{\Phi}) = \frac{\partial \ell_i(\mathbf{\Phi})}{\partial \nu}, \quad i = 1, \ldots, n.$$

The detailed expressions of $S_i(\mathbf{\Phi})$ are as follows:

$$S_{i1}(\mathbf{\Phi}) = \sum_{i:Y_i=0} -\frac{e^{\boldsymbol{\beta}^\top}}{1+e^{\boldsymbol{\beta}^\top \mathbf{X}_i}} + \sum_{i:Y_i>0} \left[\frac{1+\nu}{1+e^{\boldsymbol{\beta}^\top \mathbf{X}_i}} - \frac{\nu Y_i e^{\boldsymbol{\beta}^\top \mathbf{X}_i}}{e^{\boldsymbol{\gamma}^\top \mathbf{X}_i}\left(1+e^{\boldsymbol{\beta}^\top \mathbf{X}_i}\right)^2}\right], \tag{5}$$

$$S_{i2}(\mathbf{\Phi}) = \sum_{i:Y_i>0} \left[-\nu + \frac{\nu Y_i e^{\boldsymbol{\beta}^\top \mathbf{X}_i}}{e^{\boldsymbol{\gamma}^\top \mathbf{X}_i} \left(1 + e^{\boldsymbol{\beta}^\top \mathbf{X}_i}\right)} \right], \tag{6}$$

$$S_{i3}(\mathbf{\Phi}) = \sum_{i:Y_i>0} \left[-\psi(\nu) + \log(Y_i) + \log\left(\frac{e^{\boldsymbol{\gamma}^\top \mathbf{X}_i}(1 + e^{\boldsymbol{\beta}^\top \mathbf{X}_i})}{\nu e^{\boldsymbol{\beta}^\top \mathbf{X}_i}} \right) + 1 - \frac{Y_i e^{\boldsymbol{\beta}^\top \mathbf{X}_i}}{e^{\boldsymbol{\gamma}^\top \mathbf{X}_i}(1 + e^{\boldsymbol{\beta}^\top \mathbf{X}_i})} \right], \tag{7}$$

with $\psi(\nu)$ is the digamma function.

Notably, the explicit forms of $S_{i1}(\mathbf{\Phi})$ in (5) and $S_{i2}(\mathbf{\Phi})$ in (6) are instrumental in deriving the asymptotic properties of the estimator $\mathbf{\Phi}$.
Furthermore, we observe that $\mathbb{E}[U_{F,n}(\mathbf{\Phi})] = 0$, meaning that $U_{F,n}(\mathbf{\Phi})$ is an unbiased ESF. Consequently, the maximum likelihood estimator (MLE) $\widehat{\mathbf{\Phi}}$ of $\mathbf{\Phi}$ can be obtained by solving $U_{F,n}(\mathbf{\Phi}) = 0$.

3.1.1 Assumptions and main results for the MZIG Model

To establish the consistency and asymptotic normality of the estimator $\widehat{\mathbf{\Phi}}$, we begin by verifying the identifiability of the model parameters and introducing the necessary regularity conditions.
The following lemma demonstrates that the model parameters are identifiable, a fundamental requirement for proving the consistency and asymptotic normality of the maximum likelihood estimator. Without this condition, the primary conclusions of Theorem 1 would not be valid.

Lemma 1. *Assume that Assumptions (A1)-(A3), as stated in Appendix, are satisfied. Then the parameters $\mathbf{\Phi} = (\beta^\top, \gamma^\top, \nu)^\top$ are identifiable in the MZIG model under the specified link functions for ϕ_i and q_i, provided that the covariates X_i have sufficient variability and do not induce multicollinearity.*
The proof is given in Appendix

Discussion of identifiability: Identifiability is a critical prerequisite for statistical inference. This lemma ensures that under the assumed model structure, the true parameter vector $\mathbf{\Phi}$ is uniquely determined by the distribution of the observed data. Without identifiability, the asymptotic properties of the estimators would not hold.

Next, we assume the following regularity conditions.

(H1) Let $a^{\otimes 2} = aa^\top$ represent the outer product of any column vector a; the matrix $\mathbb{E}\left[(S_1(\mathbf{\Phi}))^{\otimes 2}\right]$ is assumed to be positive definite in the neighborhood of the true value $\mathbf{\Phi}$.

(H2) In a neighborhood of $\mathbf{\Phi}$, the first and second derivatives of $U_{F,n}(\mathbf{\Phi})$ with respect to $\mathbf{\Phi}$ are uniformly bounded above by a function of (Y, X), whose expectations exist.

Discussion of the assumptions: All these conditions are standard in the context of maximum likelihood estimation. Furthermore, the hypothesis (H1) imposes the positive definiteness of the Fisher information matrix around $\mathbf{\Phi}$, and (H2) guarantees the regularity of the estimation processes and the existence of the moments necessary for convergence.

To simplify notations in our theorems, we denote $\xrightarrow{p}$ and $\xrightarrow{d}$ as convergence in probability and in distribution, respectively. With these assumptions, we can state the following main results.

Theorem 1. *Assume that regularity conditions (H1)-(H2) hold. Then, the maximum likelihood estimator satisfies: $\widehat{\mathbf{\Phi}}_F \xrightarrow{p} \mathbf{\Phi}$ as $n \to \infty$ and $\sqrt{n}\left(\widehat{\mathbf{\Phi}}_F - \mathbf{\Phi}\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_F)$. The covariance matrix Σ_F is given by the expression: $\Sigma_F = J_F^{-1}(\mathbf{\Phi}) I_F(\mathbf{\Phi}) [J_F^{-1}(\mathbf{\Phi})]^\top$, where $I_F(\mathbf{\Phi}) = \mathbb{E}\left[S_1(\mathbf{\Phi})^{\otimes 2}\right]$ represents the expected outer product of the score function $S_1(\mathbf{\Phi})$.*

When the Fisher information matrix satisfies $J_F(\mathbf{\Phi}) = I_F(\mathbf{\Phi})$, meaning the variance of the score function equals the Fisher information matrix, the covariance matrix simplifies to: $\Sigma_F = J_F^{-1}(\mathbf{\Phi})$. This simplification arises because the additional terms $I_F(\mathbf{\Phi})$ and $[J_F^{-1}(\mathbf{\Phi})]^\top$ in the general expression are equal to the Fisher information itself, leading to Σ_F being directly given by the inverse of $J_F(\mathbf{\Phi})$, the Fisher information matrix. This result highlights the fundamental role of the Fisher information in the asymptotic efficiency of maximum likelihood estimators.
The proof is given in Appendix

3.2 MZILN model formulation and estimation

The Marginalized Zero-Inflated Log-Normal (MZILN) model extends the MZIG framework by replacing the Gamma distribution with a Log-Normal distribution. The marginal probability density function of the MZILN model is defined as:

$$P(Y_i = y_i) = \begin{cases} 1 - \phi_i & \text{if } y_i = 0, \\ \phi_i \cdot \frac{1}{y_i \sigma \sqrt{2\pi}} \exp\left(- \frac{\left(\log\left(\frac{y_i \phi_i}{q_i} \right) + \frac{\sigma^2}{2} \right)^2}{2\sigma^2} \right) & \text{if } y_i > 0. \end{cases} \tag{8}$$

Here, $0 \leq \phi_i \leq 1$ and $q_i \geq 0$. Similar to the MZIG model, the MZILN model accounts for overdispersion through two mechanisms: zero-inflation and heterogeneity in the continuous component. To illustrate the practical relevance of the MZILN model, consider the context of medical expenditures. In this domain, individuals with no expenditures constitute a significant proportion (zeros), and the positive expenditures are often highly skewed. The MZILN model leverages the Log-Normal distribution to account for this asymmetry and provides a better interpretation of covariate effects such as age or access to insurance. This example highlights the model's ability to handle complex data structures commonly observed in real-life applications. The parameters ϕ_i and q_i play a fundamental role in defining the marginal density of the MZILN model, each having a specific impact, but also a combined effect, on the shape of the distribution. The parameter ϕ_i determines the level of zero-inflation, i.e., the probability of observing a value of $Y_i = 0$. When it is high, it increases this probability, thus reducing the probability of observing positive values. In contrast, a low ϕ_i decreases this inflation and favors non-zero values. The parameter q_i, on the other hand, adjusts the location of the density for positive values of Y_i, acting as a scaling factor for the logarithmic transformation. A higher q_i shifts the density toward larger values of Y_i, while a lower q_i concentrates the density around smaller values of Y_i. The combined effect of the two parameters is crucial: ϕ_i controls the probability of observing a zero, while q_i modulates the dispersion of positive values, thus creating a subtle interaction between zero-inflation and the shape of the distribution of positive values. A change in either of these parameters can therefore affect the overall shape of the marginal density, modifying both the probability of observing zeros and the concentration of positive values.

The MZILN model defines the linear predictors as:

$$\text{logit}(\phi_i) = \boldsymbol{\alpha}^\top \mathbf{X}_i, \quad \text{and} \quad q_i = \exp(\boldsymbol{\delta}^\top \mathbf{X}_i).$$

Here, $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ are p-dimensional vectors of regression coefficients. Notably, no distinction is made between covariates influencing the marginal mean q_i and those affecting the susceptibility probability ϕ_i; all covariates are included in both submodels using the common notation $\mathbf{X}_i$. To identify covariates significantly influencing each process, Wald tests are employed. Given n independent observations $(Y_i, \mathbf{X}_i)$, the log-likelihood function of the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\delta}^\top, \sigma)^\top$ is given by:

$$\begin{aligned} \ell_n(\boldsymbol{\theta}) = & - \sum_{i:y_i=0} \log\left(1 + e^{\boldsymbol{\alpha}^\top \mathbf{X}_i}\right) + \sum_{i:y_i>0} \left(\boldsymbol{\alpha}^\top \mathbf{X}_i - \log\left(1 + e^{\boldsymbol{\alpha}^\top \mathbf{X}_i}\right) \right) \\ & + \sum_{i:y_i>0} \left(- \log(y_i \sigma \sqrt{2\pi}) - \frac{\left(\log(y_i) - \boldsymbol{\delta}^\top \mathbf{X}_i + \boldsymbol{\alpha}^\top \mathbf{X}_i - \log(1 + e^{\boldsymbol{\alpha}^\top \mathbf{X}_i}) + \frac{\sigma^2}{2} \right)^2}{2\sigma^2} \right). \end{aligned} \tag{9}$$

The log-likelihood equation (9) represents the probability of observing the data as a function of the parameters $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\delta}^\top, \sigma)^\top$. The derivatives of this log-likelihood with respect to the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and σ form the score equations, which are used to obtain parameter estimates via maximum likelihood. These derivatives reflect the impact of each parameter on the log-likelihood and enable the optimization of the function to obtain estimates for $\boldsymbol{\theta}$.

Following to $\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta})$, where $\ell_n(\boldsymbol{\theta})$ is given in (9), the ESF is offered as follows:

$$U_{S,n}(\boldsymbol{\theta}) = \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{\sqrt{n}} \begin{pmatrix} \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}} \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\delta}} \\ \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\sigma}} \end{pmatrix} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_i(\boldsymbol{\theta}). \tag{10}$$

$S_i(\boldsymbol{\theta}) = \partial \ell_i(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} = \left(S_{i1}^{\top}(\boldsymbol{\theta}), S_{i2}^{\top}(\boldsymbol{\theta}), S_{i3}^{\top}(\boldsymbol{\theta})\right)^{\top}$, $S_{i1}(\boldsymbol{\theta}) = \frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}}$, $S_{i2}(\boldsymbol{\theta}) = \frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\delta}}$, $S_{i3}(\boldsymbol{\theta}) = \frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\sigma}}$ $i = 1, \ldots, n$.

Then, some tedious albeit not difficult algebra shows that:

$$S_{i1}(\boldsymbol{\theta}) = -\sum_{i:y_i=0} \frac{e^{\boldsymbol{\alpha}^{\top}\mathbf{X}_i}}{1+e^{\boldsymbol{\alpha}^{\top}\mathbf{X}_i}} + \sum_{i:y_i>0} \frac{1}{1+e^{\boldsymbol{\alpha}^{\top}\mathbf{X}_i}} \left(1 - \frac{\log(y_i) - \boldsymbol{\delta}^{\top}\mathbf{X}_i + \boldsymbol{\alpha}^{\top}\mathbf{X}_i - \log(1+e^{\boldsymbol{\alpha}^{\top}\mathbf{X}_i}) + \frac{\sigma^2}{2}}{\sigma^2}\right),$$

$$S_{i2}(\boldsymbol{\theta}) = \sum_{i:y_i>0} \frac{1}{\sigma^2} \left(\log(y_i) - \boldsymbol{\delta}^{\top}\mathbf{X}_i + \boldsymbol{\alpha}^{\top}\mathbf{X}_i - \log(1+e^{\boldsymbol{\alpha}^{\top}\mathbf{X}_i}) + \frac{\sigma^2}{2}\right),$$

$$S_{i3}(\boldsymbol{\theta}) = \sum_{i:y_i>0} \left(-\frac{1}{\sigma} + \frac{\left(\log(y_i) - \boldsymbol{\delta}^{\top}\mathbf{X}_i + \boldsymbol{\alpha}^{\top}\mathbf{X}_i - \log(1+e^{\boldsymbol{\alpha}^{\top}\mathbf{X}_i}) + \frac{\sigma^2}{2}\right)^2}{\sigma^3}\right).$$

Remark 2. *The score equations presented above are critical in characterizing the behavior of the MZILN model and serve as the foundation for parameter estimation via maximum likelihood. However, their nonlinear and interdependent structure poses significant challenges in computation. Specifically, the interplay between the zero-inflation component ($\boldsymbol{\alpha}$), the location of the positive values ($\boldsymbol{\delta}$), and the dispersion parameter (σ) requires careful initialization and numerical optimization techniques to ensure convergence to a global or near-global maximum of the likelihood function. Additionally, the sensitivity of the log-likelihood derivatives to extreme values or boundary cases (e.g., highly skewed data or excessive zero-inflation) underscores the need for robust methods in practical implementation. These challenges highlight the importance of understanding the theoretical properties of the score function, such as consistency and asymptotic normality, which will be addressed in subsequent sections.*

3.2.1 Assumptions and main results for the MZILN Model

Finally, the regularity conditions outlined for the MZIG model also apply to the MZILN case, ensuring consistency and asymptotic normality of the parameter estimates.

(K1) We assume that $\mathbb{E}\left[S_1(\boldsymbol{\theta})\right]^{\otimes 2}$ is positive definite in a neighborhood of the true $\boldsymbol{\theta}$. Here, for any column vector a, we define $a^{\otimes 2} = aa^T$.

(K2) In a neighborhood of $\boldsymbol{\theta}$, the first and second derivatives of the score $U_{S,n}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ must be uniformly bounded by a function of (Y, X), whose expectations exist and are finite.

Theorem 2. *Assume that regularity conditions (K1)-(K2) hold. It follows that $\widehat{\boldsymbol{\theta}}_S \xrightarrow{p} \boldsymbol{\theta}$ as $n \to \infty$ and $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}\right) \xrightarrow{d} \mathcal{N}(0, \Delta_S)$ with $\Delta_S = J_S^{-1}(\boldsymbol{\theta}) I_S(\boldsymbol{\theta}) [J_S^{-1}(\boldsymbol{\theta})]^{\top}$, where $I_S = \mathbb{E}\left[S_1(\boldsymbol{\theta})^{\otimes 2}\right]$.*

Proof of Theorem 2

The proof of Theorem 2 is standard and similar to that of Theorem 1; we omit it. However, it is important to note that the asymptotic properties of the MZILN model differ in certain key aspects due to the use of the Log-Normal distribution. To assess the validity and performance of the MZIG and MZILN models in different data scenarios, we conducted a series of detailed simulations, as explained below.

4 Simulation Studies

This section outlines a series of simulation studies conducted to assess the performance of the proposed marginalized regression model in handling semi-continuous, zero-inflated count data. All simulations use a Monte Carlo sample size of $M = 5000$ with various sample sizes: $n = 200, 500, 1000, 2000$.

4.1 Simulation Experiment 1: MZIG model

In this study, data were generated using the Marginal Zero-Inflated Gamma (MZIG) model. The probability mass function of the MZIG model is given in (3). The parameters ϕ_i and q_i are modeled as functions of the explanatory variables $X_i = (X_{i1}, X_{i2}, X_{i3})$, with the following relationships: $\text{logit}(\phi_i(\beta)) = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$ and $\log(q_i(\gamma)) = \gamma_1 X_{i1} + \gamma_2 X_{i2} + \gamma_3 X_{i3}$. The regression parameters β and γ are specified to represent distinct levels of zero inflation:

- **case 1:** $\beta = (0.8, 0.9, 0.95)^\top$, $\gamma = (0.2, 0.4, 0.6)^\top$, and $\nu = 0.8$,
- **case 2:** $\beta = (-1.17, -0.12, 2.5)^\top$, $\gamma = (-0.51, 0.79, -0.36)^\top$, and $\nu = 1.65$.

For these configurations, the average proportion of zeros in the simulated datasets is approximately 20% for Case 1 and 40% for Case 2. The covariates are generated as: $X_{i1} = 1, \quad X_{i2} \sim \mathcal{N}(0, 1), \quad X_{i3} \sim \mathcal{E}(1)$.

4.2 Simulation Experiment 2: MZILN model

The simulation setup for the Marginal Zero-Inflated Log-Normal (MZILN) model follows a similar structure. Data are generated from the MZILN model given in (8) and $\text{logit}(\phi_i(\alpha)) = \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i3}$ and $\log(q_i(\delta)) = \delta_1 X_{i1} + \delta_2 X_{i2} + \delta_3 X_{i3}$, where $X_{i1} = 1$, $X_{i2} \sim \mathcal{N}(0, 1)$, and $X_{i3} \sim \mathcal{U}([-1, 1])$. The regression parameters α and δ are specified for two experiments:

- **Case 1:** $\alpha = (-0.5, 0.12, 2)^\top$, $\delta = (-0.15, -0.5, 0.4)^\top$, and $\sigma = 0.4$,
- **Case 2:** $\alpha = (0.3, 1.2, -0.75)^\top$, $\delta = (-0.15, 0.5, -1.2)^\top$, and $\sigma = 1$.

For these parameter values, the average proportion of zero-inflated data in the simulated datasets is approximately 0.26 and 0.55 for case 1 and 2, respectively. The simulations for both models are conducted using the `optim` function in the statistical software `R`.

4.3 Simulation results

For each simulation experiment and each estimator, we calculate the MLE, the average bias, the root mean square error (RMSE), as well as the empirical coverage probability . The results highlight the robustness of the MZIG and MZILN models, characterized by low average biases, decreasing RMSEs, and coverage probabilities (CP) close to 95%. For the MZIG model (Table 1 and Table 2), the estimates rapidly converge to the true values, particularly for large sample sizes ($n = 1000, 2000$), where the average biases and RMSEs significantly decrease, while the CP validates the reliability of the confidence intervals. Conversely, the MZILN model (Table 3 and Table 4) demonstrates high accuracy, with average biases close to zero for $n \geq 500$, low RMSEs, and notable stability for the dispersion parameter (σ), whose biases remain below 0.005. The MZILN model excels for highly zero-inflated data (55%), whereas the MZIG model performs better for moderate zero proportions (40%), reflecting the specific strengths of the two approaches in managing variance and covariate effects. These results confirm the effectiveness of both models for analyzing semi-continuous data with significant zero inflation, offering reliable and adaptable estimates across various contexts. Figures 1 to 4 illustrate the performance of the MZIG and MZILN models through simulations. In Figure 1, the histograms of the normalized estimators for the MZIG model show a rapid convergence toward the expected values as sample sizes increase. The mean biases and RMSE decrease, confirming the robustness of the estimation for low proportions of zero-inflated data. Figure 2, focusing on the MZILN model, indicates a similar performance but excels further for higher proportions of zero-inflated data. Figures 1 and 2 illustrate the density function of the standard normal distribution, represented by the black curve. These graphs demonstrate that the distribution of the maximum likelihood estimator (MLE) closely approximates a normal distribution for all parameters.

Table 1: Case 1: ML estimates of the MZIG model for different values of n, with an average of 20% for $Y = 0$.

		$n = 200$				$n = 500$			
Parameter	**True value**	**MLE estimator**	**bias**	**RMSE**	**CP**	**MLE estimator**	**bias**	**RMSE**	**CP**
β_1	0.80	0.7948	-0.0052	0.2983	0.9366	0.7961	-0.0039	0.1869	0.9442
β_2	0.90	0.9266	0.0266	0.2274	0.9566	0.9116	0.0116	0.1421	0.9468
β_3	0.95	1.0126	0.0626	0.3504	0.9544	0.9753	0.0253	0.2065	0.9288
γ_1	0.20	0.1947	-0.0053	0.1336	0.9220	0.1986	-0.0014	0.0837	0.9480
γ_2	0.40	0.3981	-0.0019	0.0968	0.9360	0.3988	-0.0012	0.0582	0.9658
γ_3	0.60	0.5942	-0.0058	0.0894	0.9136	0.5974	-0.0026	0.0547	0.9486
ν	0.80	0.8195	0.0195	0.0850	0.9556	0.8075	0.0075	0.0506	0.9420
		$n = 1000$				$n = 2000$			
Parameter	**True value**	**MLE estimator**	**bias**	**RMSE**	**CP**	**MLE estimator**	**bias**	**RMSE**	**CP**
β_1	0.80	0.8008	0.0008	0.1287	0.9354	0.8005	0.0005	0.0893	0.9528
β_2	0.90	0.9068	0.0068	0.0969	0.9548	0.9042	0.0042	0.0693	0.9424
β_3	0.95	0.9623	0.0123	0.1426	0.9288	0.9547	0.0047	0.1005	0.9612
γ_1	0.20	0.1993	-0.0007	0.0584	0.9436	0.1997	-0.0003	0.0411	0.9364
γ_2	0.40	0.3990	-0.0010	0.0418	0.9378	0.3995	-0.0005	0.0294	0.9476
γ_3	0.60	0.5991	-0.0009	0.0379	0.9400	0.5993	-0.0007	0.0265	0.9364
ν	0.80	0.8043	0.0043	0.0348	0.9400	0.8019	0.0019	0.0247	0.9638

Table 2: Case 2: ML estimates of the MZIG model for different values of n, with an average of 40% for $Y = 0$.

		$n = 200$				$n = 500$			
Parameter	**True value**	**MLE estimator**	**bias**	**RMSE**	**CP**	**MLE estimator**	**bias**	**RMSE**	**CP**
β_1	-1.17	-1.2071	-0.0371	0.3043	0.9720	-1.1828	-0.0128	0.1855	0.9304
β_2	-0.12	-0.1208	-0.0008	0.1850	0.9548	-0.1211	-0.0011	0.1108	0.9288
β_3	2.50	2.5837	0.0837	0.4688	0.9638	2.5313	0.0313	0.2803	0.9262
γ_1	-0.51	-0.5131	-0.0031	0.1132	0.9390	-0.5101	-0.0001	0.0700	0.9378
γ_2	0.79	0.7888	-0.0012	0.0712	0.9390	0.7904	0.0004	0.0441	0.9426
γ_3	-0.36	-0.3646	-0.0046	0.0661	0.9492	-0.3615	-0.0015	0.0407	0.9288
ν	1.65	1.7099	0.0599	0.2106	0.9622	1.6726	0.0226	0.1241	0.9622
		$n = 1000$				$n = 2000$			
Parameter	**True value**	**MLE estimator**	**bias**	**RMSE**	**CP**	**MLE estimator**	**bias**	**RMSE**	**CP**
β_1	-1.17	-1.1801	-0.0101	0.1305	0.9576	-1.1750	-0.0050	0.0934	0.9442
β_2	-0.12	-0.1215	-0.0015	0.0807	0.9470	-0.1205	-0.0005	0.0570	0.9416
β_3	2.5	2.5185	0.0185	0.1938	0.9652	2.5109	0.0109	0.1379	0.9388
γ_1	-0.51	-0.5101	-0.0001	0.0490	0.9516	-0.5094	0.0006	0.0347	0.9466
γ_2	0.79	0.7892	-0.0008	0.0308	0.9636	0.7901	0.0001	0.0221	0.9590
γ_3	-0.36	-0.3607	-0.0007	0.0283	0.9586	-0.3607	-0.0007	0.0201	0.9384
ν	1.65	1.6622	0.0122	0.0867	0.9370	1.6549	0.0049	0.0614	0.9402

Table 3: Case 1: ML estimates of the MZILN model for different values of n, with an average of 26% of $Y = 0$

		$n = 200$				$n = 500$			
Parameter	**True value**	**MLE estimator**	**bias**	**RMSE**	**CP**	**MLE estimator**	**bias**	**RMSE**	**CP**
α_1	-0.50	-0.5255	-0.0255	0.3508	0.9756	-0.5075	-0.0075	0.2108	0.9400
α_2	0.12	0.1224	0.0024	0.1828	0.9536	0.1196	-0.0004	0.1132	0.9516
α_3	2.00	2.0473	0.0473	0.4093	0.9698	2.0170	0.017 0	0.2466	0.9462
δ_1	-0.15	-0.1508	-0.0008	0.1096	0.9902	-0.1491	0.0009	0.0665	0.9376
δ_2	-0.50	-0.5001	-0.0001	0.0340	0.9688	-0.4998	0.0002	0.0216	0.9318
δ_3	0.40	0.4006	0.000 6	0.1153	0.9884	0.3996	-0.0004	0.0703	0.9378
σ	0.40	0.4012	-0.0043	0.0243	0.9602	0.4035	-0.0019	0.0151	0.9392
		$n = 1000$				$n = 2000$			
Parameter	**True value**	**MLE estimator**	**bias**	**RMSE**	**CP**	**MLE estimator**	**bias**	**RMSE**	**CP**
α_1	-0.50	-0.5009	-0.0009	0.1484	0.9372	-0.5017	-0.0017	0.1061	0.9448
α_2	0.12	0.1212	0.0012	0.0780	0.9454	0.1199	-0.0001	0.0561	0.9398
α_3	2.00	2.0048	0.0048	0.1752	0.940	2.004 4	0.0044	0.1260	0.9458
δ_1	-0.15	-0.1504	-0.0004	0.0465	0.9344	-0.1499	0.0001	0.0327	0.9574
δ_2	-0.50	-0.4998	0.000 2	0.0149	0.9528	-0.5001	-0.0001	0.0106	0.9486
δ_3	0.40	0.4002	0.000 2	0.0489	0.9376	0.3996	-0.0004	0.0347	0.9568
σ	0.40	0.4045	0.0009	0.0107	0.9478	0.4050	-0.0004	0.0075	0.9490

Table 4: ML estimates of the MZILN model for different values of n, with an average of 55% for $Y = 0$.

		$n = 200$				$n = 500$			
Parameter	**True value**	**MLE estimator**	**bias**	**RMSE**	**CP**	**MLE estimator**	**bias**	**RMSE**	**CP**
α_1	0.30	0.3020	0.0020	0.3752	0.9330	0.3006	0.0006	0.2339	0.9352
α_2	1.20	1.2279	0.0279	0.2167	0.9534	1.2128	0.0128	0.1332	0.9330
α_3	-0.75	-0.7605	-0.0105	0.4194	0.9386	-0.7544	-0.0044	0.2616	0.9402
δ_1	-0.15	-0.1479	0.0021	0.2243	0.9752	-0.1512	-0.0012	0.1383	0.9590
δ_2	0.50	0.5027	0.0027	0.1233	0.9756	0.4996	-0.0004	0.0764	0.9762
δ_3	-1.20	-1.2034	-0.0034	0.2514	0.9720	-1.1974	0.0026	0.1546	0.9660
σ	1.00	0.9786	-0.0214	0.0782	0.9556	0.9922	-0.0078	0.0488	0.9590
		$n = 1000$				$n = 2000$			
Parameter	**True Value**	**MLE estimator**	**bias**	**RMSE**	**CP**	**MLE estimators**	**bias**	**RMSE**	**CP**
α_1	0.30	0.3048	0.0048	0.1633	0.9514	0.3018	0.0018	0.1157	0.9488
α_2	1.20	1.2069	0.0069	0.0941	0.9524	1.2036	0.0036	0.0662	0.9464
α_3	-0.75	-0.7552	-0.0052	0.1857	0.9504	-0.7525	-0.0025	0.1299	0.9490
δ_1	-0.15	-0.1519	-0.0019	0.0975	0.9692	-0.1516	-0.0016	0.0707	0.9420
δ_2	0.50	0.5008	0.0008	0.0546	0.9522	0.5014	0.0014	0.0374	0.9542
δ_3	-1.20	-1.1979	0.0021	0.1104	0.9620	-1.1990	0.0010	0.0805	0.9424
σ	1.00	0.9961	-0.0039	0.0339	0.9602	0.9976	-0.0024	0.0240	0.9488

5 Analysis of cowpea data

5.1 Data and analysis

In this section, we illustrate the MZIG (Marginalized Zero-Inflated Gamma) and MZILN (Marginalized Zero-Inflated Log Normal) models using real data from a 2022 survey on cowpea producers' varietal preferences in Senegal. The survey primarily targeted cowpea producers in the four regions of the groundnut basin: Kaffrine, Kaolack, Fatick, Diourbel, and Louga. We also describe a simple and efficient methodology for selecting predictors. Finally, we compare the fitted models using the Vuong test.

The analyzed dataset includes the yields of producers (semi-continuous response variable) for 303 individuals aged 16 to 78 years. Collected in southern Senegal during 2022, the data encompasses several socio-economic and contextual factors that influence yield. The explanatory variables include the adoption status of improved varieties (`adoption status`, coded as 1 = Adopter, 0 = Non-adopter), distinguishing adopters from non-adopters; the quantity of seeds used per hectare (`seeds ha`), reflecting the intensity of use, and the total area of the plot in hectares (area); geographical location (`region`) and individual characteristics such as `age` and the decision-maker's gender (`gender`, 1 = Male, 0 = Female); structural elements such as household size, literacy level, membership in a cooperative or producer association, and secondary activities (`activity`); crop treatment (`treatment`, 1 = Yes, 0 = No), seed source (`seed source`, coded as 1 = Stock, 0 = Purchase), number of plots owned, and tilling method (`tilling`, coded, 1 = Manual, 2 = Animal traction); and finally, the availability of non-agricultural income. These variables are used to explore the determinants of agricultural yield, while also providing insights for optimizing strategies to improve productivity.

The MZIG (Marginalized Zero-Inflated Gamma) and MZILN (Marginalized Zero-Inflated Log Normal) models were selected for the analysis of cowpea producers' yields due to the particular structure of the collected data, notably the presence of a high proportion of zeros (10% of the observations are equal to zero). The presence of zeros can be attributed to various factors, such as suboptimal farming practices or difficulties accessing the necessary resources for higher cowpea production.

Classical models such as the ZIG (Zero-Inflated Gamma) and ZILN (Zero-Inflated Log Normal) are commonly used to handle this type of data, as they allow for modeling both the probability of a zero (a production failure) and positive yields simultaneously, while accounting for the distribution of observed yields. However, these models do not always consider the marginal effects of explanatory variables on the probability of a zero and the continuous part of the distribution. The MZIG and MZILN models, which include additional marginalization, offer a more flexible and robust approach to estimating the effects of various explanatory variables while properly handling the excess zeros. This marginal feature allows for better isolation of the relationships between socio-economic factors, farming practices, and cowpea producers' yields, reducing the potential bias related to the neglect of zeros.

Moreover, the MZIG and MZILN models are better suited for handling asymmetric and heterogeneous yield distributions, a common occurrence in agricultural studies where yields can vary significantly from one farm to another. This is particularly relevant in the context of this study, where yields are influenced by multiple factors, such as farming practices (crop treatment, soil management methods), individual characteristics of the producers (age, gender), and the socio-economic structure of the households.

Finally, the combination of zero handling and yield distribution fitting with these models allows for a better interpretation of the results, providing clearer insights for agricultural policies aimed at improving the productivity of cowpea producers in Senegal. The ZIG , ZILN, MZIG, and MZILN models were fitted to the data collected in this study.

Several authors have recently explored variable selection in zero-inflated models using penalized maximum likelihood methods. Notably, the `R` package `mpath` (Wang, 2019) offers various penalty functions tailored for such tasks. In this study, we propose an alternative methodology to adjust MZIG and MZILN models while simultaneously identifying relevant predictors.

The process begins with an initial variable screening aimed at identifying predictors associated with zero inflation. A logistic regression model is fitted to the binary response variable $1\{Y_i = 0\}$, where $i = 1, \ldots, n$. While this does not fully encapsulate the zero-inflation mechanism since some zeros originate from the count distribution it provides a practical approximation for identifying potential predictors. This preliminary screening is crucial, as it reduces the dimensionality of the problem and highlights variables likely to influence the zero-inflation mechanism. By identifying a subset of potentially relevant predictors, the subsequent steps can focus on refining the model while minimizing computational complexity and the risk of overfitting.

Given the typically large number of predictors, stepwise logistic regression is employed, starting from a null model. Selection at each step is guided by the Bayesian Information Criterion (BIC), which is preferred over the Akaike Information Criterion (AIC) for its emphasis on parsimony.

Following this screening, a comprehensive selection of variables is performed specifically for the MZIG and MZILN models. The Wald test is applied to assess the statistical significance of each predictor. For

a given predictor, the Wald statistic is computed as:

$$W_j = \frac{\widehat{\beta}_j^2}{SE(\widehat{\beta}_j)^2},$$

where $\widehat{\beta}_j$ represents the estimated coefficient and $SE(\widehat{\beta}_j)$ is its standard error. Under the null hypothesis of no effect, W_j follows a chi-square distribution with one degree of freedom. Predictors with p-values below 0.05 are retained, while those that are not significant are excluded. This results in a simplified, interpretable model. Throughout this process, convergence diagnostics are carefully evaluated to ensure the reliability of the final models.

The adjusted models, including ZIG, MZIG, ZILN, and MZILN, are then fine-tuned using the selected variables. Robust procedures are implemented to optimize model parameters and assess performance metrics. These adjustments are designed to accurately capture the underlying data structure and provide reliable estimates for both the zero-inflation and count components.

This systematic approach combines initial variable selection with rigorous statistical testing and model adjustment to enhance both the interpretability and performance of zero-inflated models. By leveraging stepwise regression and detailed adjustments, it ensures a robust and parsimonious model structure suitable for semi-continuous data with excess zeros. The proposed methodology is applied to various zero-inflated models to demonstrate its flexibility and effectiveness. In the following, we provide the specific formulations for each model, highlighting how the selected predictors influence the zero inflation and count components.

(i) For the ZIG model

$$\begin{cases} \text{logit}(\phi_i(\beta)) = \beta_1 + \beta_2 \times \text{age}_i + \beta_3 \times \text{area}_i + \beta_4 \times \text{tilling}_i + \beta_5 \times \texttt{treatment}_i + \beta_6 \times \texttt{activity}_i, \\ \nu_i = \exp(\gamma_1 + \gamma_2 \times \texttt{age}_i + \gamma_3 \times \texttt{area}_i + \gamma_4 \times \texttt{tilling}_i + \gamma_5 \times \texttt{treatment}_i + \gamma_6 \times \texttt{activity}_i) \end{cases}$$

(ii) For the MZIG model

$$\begin{cases} \text{logit}(\phi_i(\beta)) = \beta_1 + \beta_2 \times \texttt{region}_i + \beta_3 \times \texttt{age}_i + \beta_4 \times \texttt{area}_i + \beta_5 \times \texttt{tilling}_i + \beta_6 \times \texttt{treatment}_i \\ \qquad\qquad + \beta_7 \times \texttt{activity}_i, \\ q_i = \exp(\gamma_1 + \gamma_2 \times \texttt{region}_i + \gamma_3 \times \texttt{age}_i + \gamma_4 \times \texttt{area}_i + \gamma_5 \times \texttt{tilling}_i + \gamma_6 \times \texttt{treatment}_i \\ \qquad\qquad + \gamma_7 \times \texttt{activity}_i). \end{cases}$$

(iii) For the ZILN model

$$\begin{cases} \text{logit}(\phi_i(\alpha)) = \alpha_1 + \alpha_2 \times \texttt{tilling}_i + \alpha_3 \times \texttt{treatment}_i + \alpha_4 \times \texttt{activity}_i, \\ \nu_i = \exp(\delta_1 + \delta_2 \times \texttt{tilling}_i + \delta_3 \times \texttt{treatment}_i + \delta_4 \times \text{activity}_i). \end{cases}$$

(iv) For the MZILN model

$$\begin{cases} \text{logit}(\phi_i(\alpha)) = \alpha_1 + \alpha_2 \times \texttt{tilling}_i + \alpha_3 \times \text{treatment}_i + \alpha_4 \times \texttt{activity}_i + \alpha_5 \times \texttt{area}_i, \\ q_i = \exp(\delta_1 + \delta_2 \times \text{tilling}_i + \delta_3 \times \texttt{treatment}_i + \delta_4 \times \texttt{activity}_i + \delta_5 \times \texttt{area}_i). \end{cases}$$

5.2 Results of the data analysis

Parameter estimates, standard errors, and corresponding p-values for the Wald tests are detailed in Tables 5 and 6. Since the final models are not nested, they were compared using the Vuong test [20], with the results presented in Table 7. These tables collectively demonstrate that the MZIG and MZILN models outperform their non-marginalized counterparts (ZIG and ZILN) in terms of precision and interpretability of the factors affecting cowpea yields.

The MZIG model shows an improved overall fit (AIC of 3614.315 compared to 3652.137 for ZIG), driven by additional interactions and significant coefficients, highlighting that factors such as producer

age, agricultural practices (tilling and treatment), and secondary activities positively influence productivity, while larger cultivated areas have a negative effect. The MZILN model, which incorporates variables such as plot size, performs particularly well in scenarios with high zero-inflation, achieving a lower AIC (3253.532 compared to 3277.033 for ZILN). Its superiority is further confirmed by the Vuong test (statistic of 3.5750, $p = 0.0004$). Both models emphasize the critical roles of treatment, tilling, and secondary activities, offering complementary tools for analyzing cowpea yields across varying levels of zero inflation.

To further understand the practical implications of these findings, we analyze the significance and roles of specific parameters in real-world contexts.

The results obtained from the MZIG and MZILN models provide valuable insights into the underlying dynamics of semi-continuous data with excess zeros. Each parameter estimated in these models holds specific implications, particularly in applied contexts such as agriculture, health, or social sciences. The coefficients related to the probability of structural zeros (ϕ_i) reflect the influence of factors like tilling methods; for instance, manual tilling is associated with a higher probability of zero yields, indicative of less efficient agricultural practices. Similarly, coefficients linked to the marginal mean (q_i) demonstrate the direct impact of variables such as producer age or the use of agricultural treatments on mean yields, underscoring the importance of experience and access to modern farming inputs. Furthermore, the dispersion parameters (ν or σ) capture the intrinsic variability in the yields, often driven by unpredictable factors such as climatic conditions. These models enable a clear interpretation of covariate effects on both zero outcomes and positive values, serving as effective tools for strategic decision-making. For example, reducing plot sizes or promoting modern farming practices could decrease the prevalence of zero yields and boost overall productivity, thus offering robust bases for optimizing resources and shaping agricultural policies.

Table 5: Summary of statistical results: ZIG vs. MZIG models applied to Senegalese cowpea farming.

Parameter	**variables**	**ZIG**			**Proposed MZIG**		
		Estimate	**Std. Error**	**P-value**	**Estimate**	**Std. Error**	**P-value**
β_1	Intercept	-2.6319	0.7187	0.0003	2.7148	0.7177	0.0002
β_2	Region				-0.8464	0.0082	0.0001
β_3	Age	0.0174	0.0085	0.0400	0.0278	0.0430	0.0007
β_4	Area	-0.078	0.0388	0.0446	-0.0925	0.6128	0.0313
β_5	Tilling	-1.0545	0.6280	0.0319	-1.3572	0.3989	0.0268
β_6	Treatment	1.3173	0.3896	0.0007	1.0612	0.3735	0.0045
β_7	Activity	1.3122	0.4208	0.0018	1.139	0.3745	0.0024
γ_1	Intercept	5.3381	0.2217	0.0000	5.3509	0.2204	0.1213
γ_2	Region				-0.1413	0.3691	0.0152
γ_3	Age	-0.0862	0.0055	0.0120	0.0935	0.0055	0.0000
γ_4	Area	0.1151	0.0854	0.0178	0.2228	0.0912	0.0145
γ_5	Tilling	-1.0416	0.4941	0.0350	-1.0101	0.5036	0.0449
γ_6	Treatement	0.7268	0.1760	0.0000	0.5395	0.1638	0.0013
γ_7	Activity	1.0708	0.1710	0.000	0.9528	0.1609	0.0000
ν	Dispersion parameter	0.3888	0.0271	0.0000	0.3516	0.024	0.0000
Loglik			-1815.069			-1794.158	
AIC			3652.137			3614.315	
BIC			3692.988			3662.594	

Table 6: Summary of statistical results: ZILN vs. MZILN models applied to Senegalese cowpea farming.

Parameter	variables	ZILN			Proposed MZILN		
		Estimate	**Std. Error**	**P-value**	**Estimate**	**Std. Error**	**P-value**
α_1	Intercept	1.5720	0.3986	0.0001	1.9135	0.4325	0.0000
α_2	Tilling	-1.4899	0.5892	0.0115	-1.2705	0.6034	0.0352
α_3	Treatement	0.5830	0.4102	0.0255	0.5819	0.4087	0.0154
α_4	Activity	0.7988	0.4097	0.0512	0.5759	0.4259	0.0276
α_5	Area				-0.0967	0.0409	0.0181
δ_1	Intercept	1.4922	0.0445	0.0000	5.2344	0.2020	0.0000
δ_2	Tilling	-0.2821	0.1230	0.0218	-0.9492	0.3900	0.0150
δ_3	Treatement	-0.0804	0.0397	0.0427	-0.2177	0.1577	0.0149
δ_4	Activity	-0.0811	0.0422	0.0549	-0.2799	0.1691	0.0267
δ_5	Area				-0.1556	0.0330	0.0149
σ	Dispersion parameter	1.2654	0.0541	0.0000	1.2115	0.0518	0.0000
Loglik			-1629.517			-1615.766	
AIC			3277.033			3253.532	
BIC			3310.457			3294.383	

Table 7: Model comparison using Vuong test: Vuong statistic, p-value, and test decision (i.e., the best model according to Vuong test).

	Vuong statistic	**P-value**	**Decision**
ZIG vs MZIG	-4.2110	0.0001	MZIG
ZILN vs MZILN	-1.9909	0.0465	MZILN
MZIG vs MZILN	3.5750	0.0004	MZILN

6 Discussion

This article introduces two new marginalized regression models, MZIG and MZILN, designed to analyze semi-continuous data with a high proportion of structural zeros, providing a direct interpretation of covariate effects on the marginal mean while accounting for distinct mechanisms generating structural zeros and continuous outcomes. Equipped with robust asymptotic properties and validated through simulations and practical applications, these models stand out for their ability to overcome the limitations of traditional approaches. Simulations have demonstrated their effectiveness across various contexts of zero proportions and data structures, while their application to real-world data on cowpea production in Senegal confirmed their relevance, achieving performance comparable to or better than classical models. The perspectives of this study include methodological extensions to other continuous distributions to better address extreme asymmetry or over-dispersion, application to diverse fields such as finance or ecology, integration of random effects for longitudinal or hierarchical data, computational optimization of algorithms for large datasets, and empirical validation through comparative studies on varied datasets. These advancements pave the way for future research aimed at refining the analysis of complex data and enriching statistical tools for the benefit of researchers and practitioners.

Conflict of interest

The authors have no conflicts of interest to declare relevant to this article's content.

Funding

This research received no external funding.

Data Availability Statement

The data will be made available upon request from the corresponding author

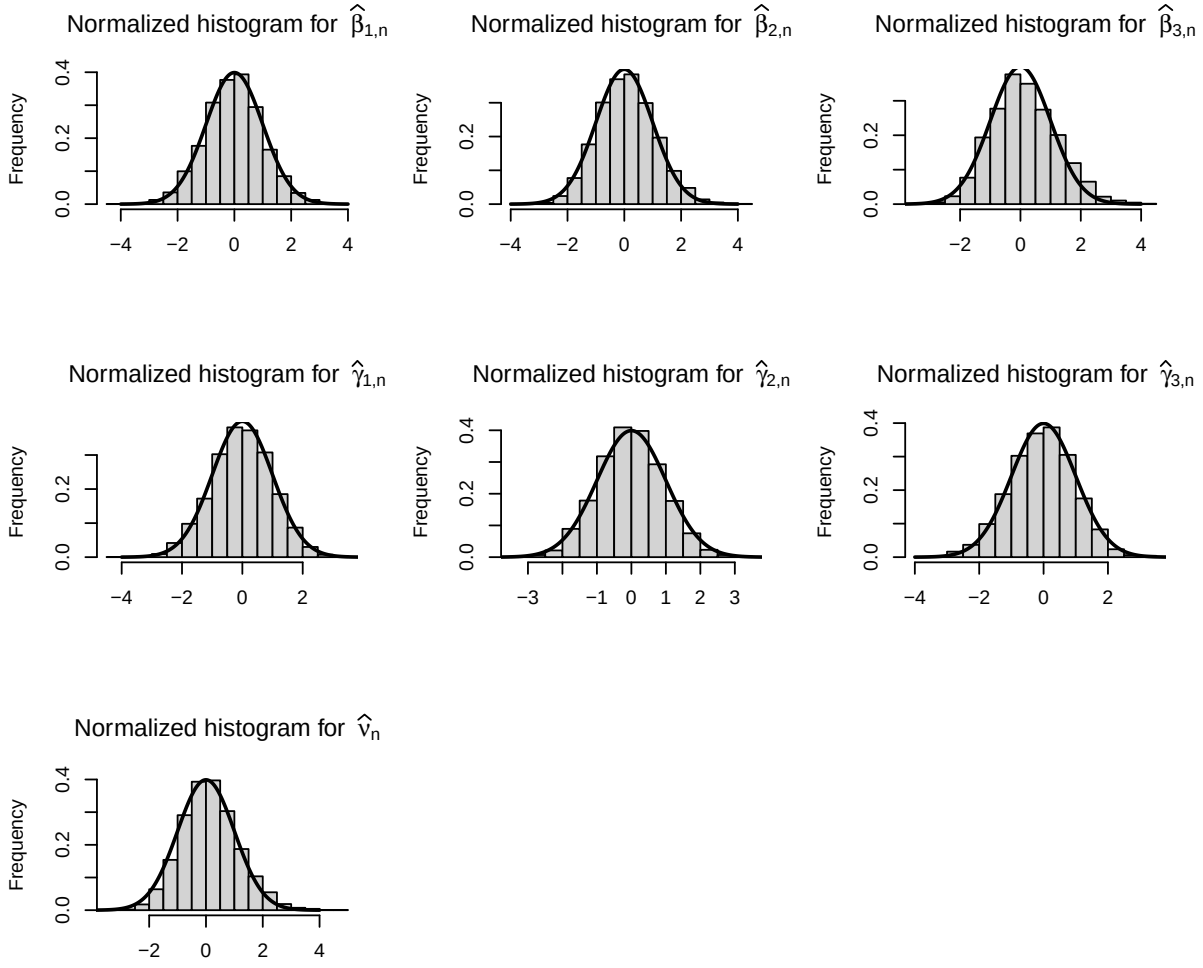

Figure 1: Histograms of normalized estimates for the MZIG model estimators with $n = 500$ and a zero-inflation fraction of 0.20.

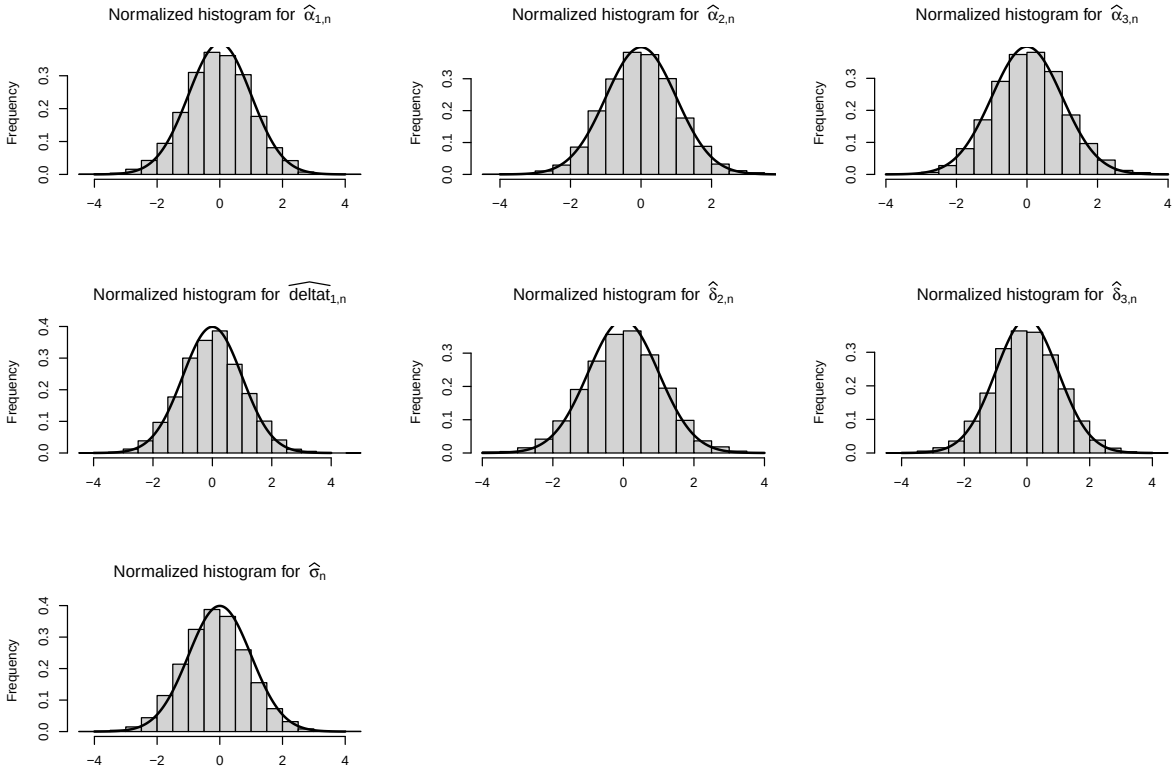

Figure 2: Histograms of normalized estimates for the MZILN model estimators with $n = 500$ and a zero-inflation fraction of 0.26.

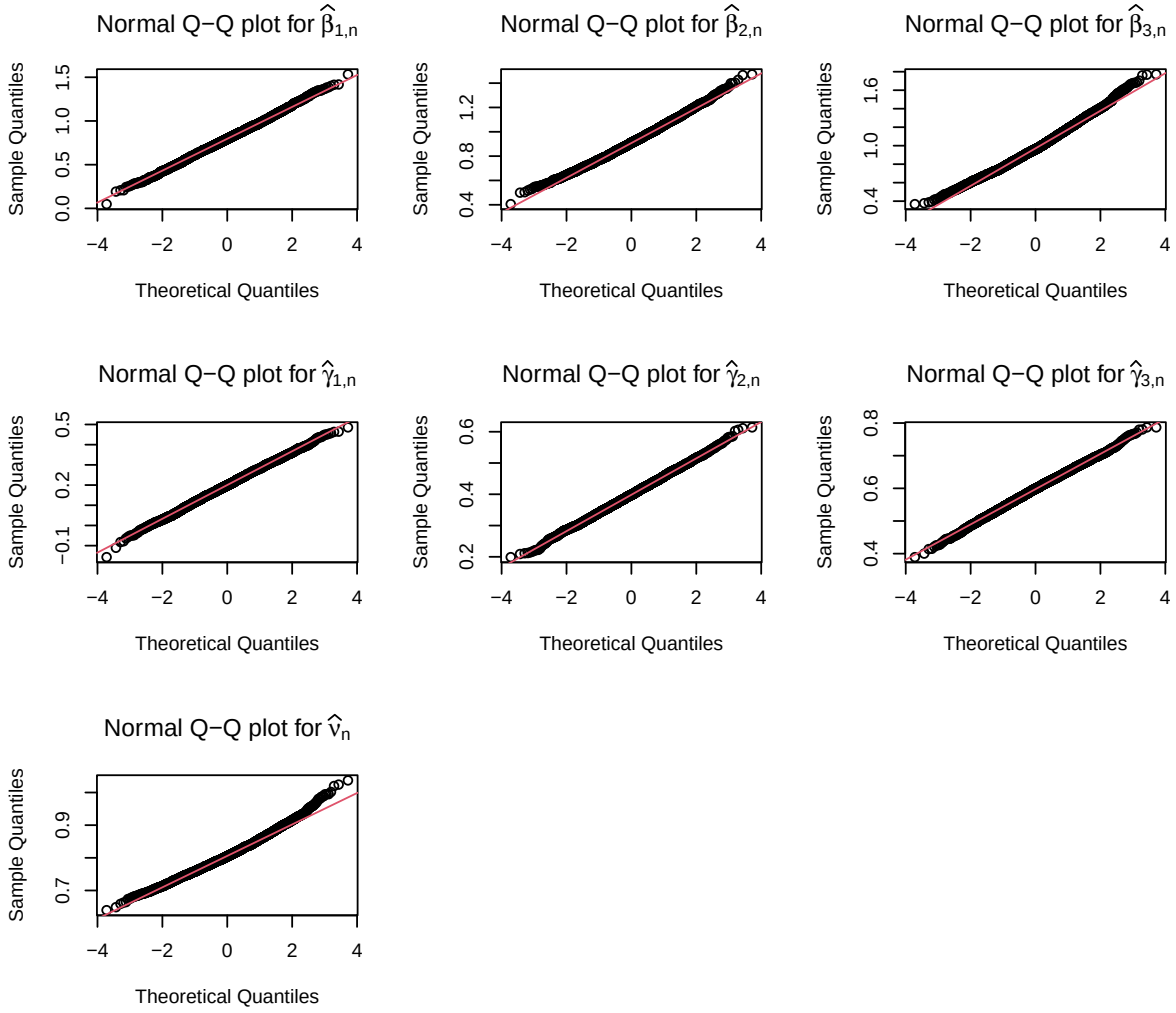

Figure 3: Normal Q-Q plots for the MZIG model estimators with $n = 500$ and a zero-inflation fraction of 0.20.

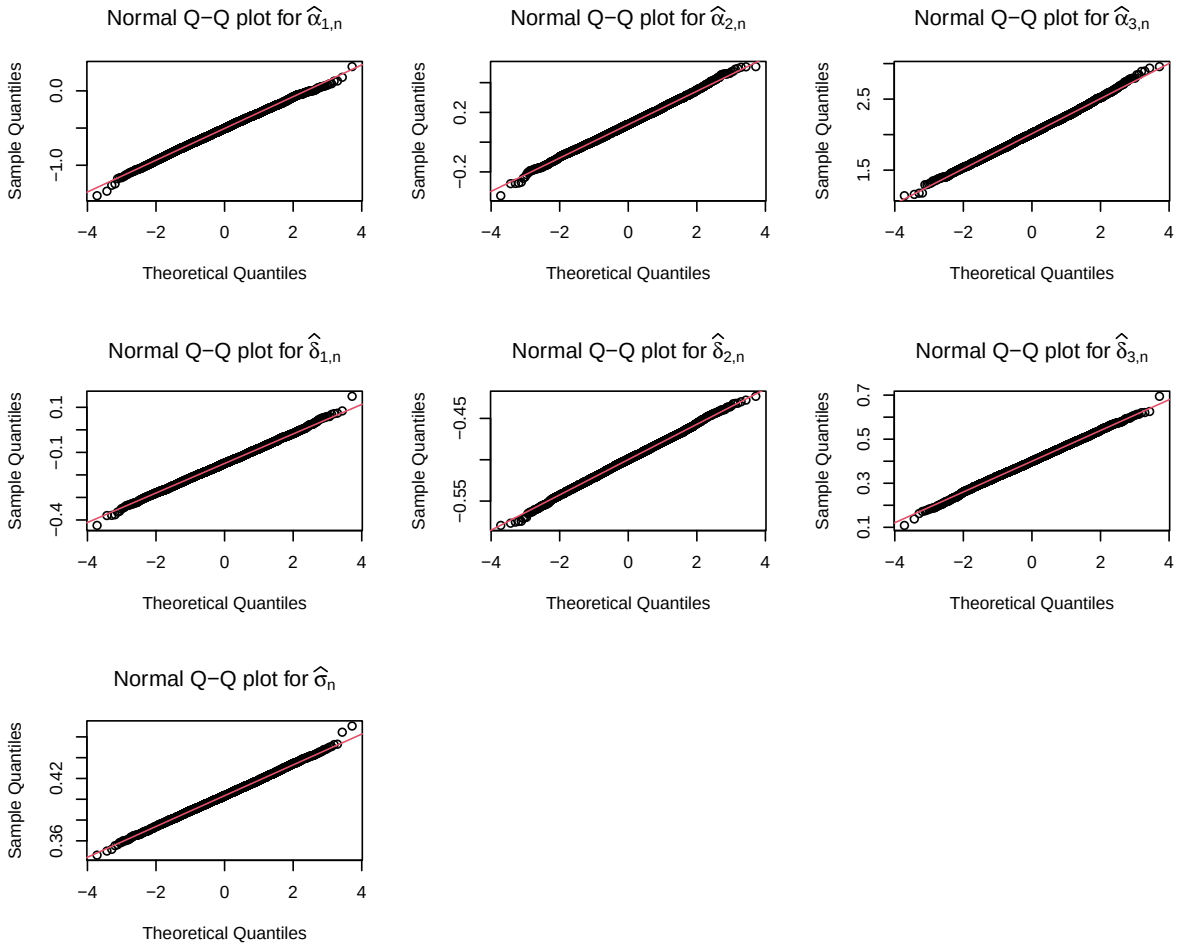

Figure 4: Normal Q-Q plots for the MZILN model estimators with $n = 500$ and a zero-inflation fraction of 0.26.

Appendix : proofs

Proof of Lemma 1.

Before proving the Lemma1, we state the following technical assumptions:

(A1) The covariates X_i have sufficient variability and avoid multicollinearity.
(A2) The link functions $g(\cdot)$ (for ϕ_i) and the log-linear mapping (for q_i) are invertible and strictly monotonic.
(A3) The model's assumptions about the distributional form (Gamma distribution for $Y_i > 0$) hold true.

The identifiability of the parameter vector $\mathbf{\Phi} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \nu)$ is established through distinct mappings of the covariates $\boldsymbol{X}_i$ to the observed outcomes. For $\boldsymbol{\beta}$, the probability of a zero outcome is given by

$$\phi_i = P(Y_i = 0) = g^{-1}(\beta^\top \boldsymbol{X}_i),$$

where $g^{-1}(\cdot)$ is the strictly monotonic inverse link function that maps probabilities back to the linear predictor $\beta^\top \boldsymbol{X}_i$. If $\boldsymbol{X}_i$ has sufficient variability (i.e., no identical rows), then $\boldsymbol{\beta}$ is uniquely determined. For $\boldsymbol{\gamma}$ and ν, when $Y_i > 0$, the outcome Y_i follows a Gamma distribution with inverse scale $q_i = \exp(\gamma^\top \boldsymbol{X}_i)$ and shape ν. The moments of the Gamma distribution provide unique values for

$$\nu = \frac{\mathbb{E}[Y_i \mid Y_i > 0]^2}{\mathrm{Var}(Y_i \mid Y_i > 0)} \quad \text{and} \quad q_i = \frac{\nu}{\mathbb{E}[Y_i \mid Y_i > 0]},$$

ensuring that $\boldsymbol{\gamma}$ is identifiable if $\boldsymbol{X}_i$ varies sufficiently.
The mappings of $\boldsymbol{X}_i$ to ϕ_i (via the logit or other link) and q_i (via the log-linear mapping) are orthogonal, preventing redundancy. As a result, $\boldsymbol{\beta}$ governs the zero-probability ϕ_i, while $\boldsymbol{\gamma}$ controls the inverse scale q_i for $Y_i > 0$. Thus, under assumptions (A1)-(A3), the parameters $\boldsymbol{\Phi} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \nu)$ are uniquely identifiable.

Proof of Theorem 1

To demonstrate that $\widehat{\boldsymbol{\Phi}}$ is a consistent estimator for $\boldsymbol{\Phi}$ and to prove its asymptotic normality, we proceed in two steps: first, we prove consistency, and second, we establish the asymptotic distribution.
To establish the consistency of the estimator $\widehat{\boldsymbol{\Phi}}$ of the true parameter $\boldsymbol{\Phi}$, we begin by defining the score function and its derivative:

$$U_{F,n}(\boldsymbol{\Phi}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_i(\boldsymbol{\Phi}),$$
$$J_{F,n}(\boldsymbol{\Phi}) = -\frac{1}{\sqrt{n}} \frac{\partial U_{F,n}(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}^\top} = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial S_i(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}^\top}.$$

Taking the expectation of $J_{F,n}(\boldsymbol{\Phi})$, and using the linearity of expectation, we obtain:

$$\mathbb{E}[J_{F,n}(\boldsymbol{\Phi})] = -\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\frac{\partial S_i(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}^\top}\right].$$

By the law of total expectation:

$$\mathbb{E}\left[\frac{\partial S_i(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}^\top}\right] = \mathbb{E}\left\{\mathbb{E}\left[\frac{\partial S_i(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}^\top}\Big| Y_i, X_i\right]\right\}.$$

Thus,

$$\mathbb{E}[J_{F,n}(\boldsymbol{\Phi})] = -\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left\{\mathbb{E}\left[\frac{\partial S_i(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}^\top}\Big| Y_i, X_i\right]\right\}.$$

Under the assumption that the $S_i(\boldsymbol{\Phi})$ are independent and identically distributed (i.i.d.), the sum simplifies to:

$$\mathbb{E}[J_{F,n}(\boldsymbol{\Phi})] = -\mathbb{E}\left\{\mathbb{E}\left[\frac{\partial S_1(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}^T}\Big| Y_1, X_1\right]\right\}.$$

By the law of total expectation:

$$\mathbb{E}\left\{\mathbb{E}\left[\frac{\partial S_1(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}^\top}\Big| Y_1, X_1\right]\right\} = \mathbb{E}\left[\frac{\partial S_1(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}^\top}\right].$$

We then deduce:

$$\mathbb{E}[J_{F,n}(\boldsymbol{\Phi})] = -\mathbb{E}\left[\frac{\partial S_1(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}^\top}\right] = J_F(\boldsymbol{\Phi}).$$

By the Weak Law of Large Numbers (WLLN), the sequence $J_{F,n}(\boldsymbol{\Phi})$ converges in probability to its expectation:

$$J_{F,n}(\boldsymbol{\Phi}) \xrightarrow{p} J_F(\boldsymbol{\Phi}).$$

Furthermore, under the regularity condition $(H2)$, this convergence is uniform in a neighborhood of the true parameter $\boldsymbol{\Phi}$.

Since the expectation of the score function is zero:

$$U_F(\boldsymbol{\Phi}) = \mathbb{E}[S_i(\boldsymbol{\Phi})] = 0,$$

there exists a unique solution to the equation $U_{F,n}(\mathbf{\Phi}) = 0$ in this neighborhood. By the inverse function theorem from [5], this solution is consistent, implying:

$$\widehat{\mathbf{\Phi}}_F \xrightarrow{p} \mathbf{\Phi} \quad \text{as } n \to \infty.$$

Consequently, $\widehat{\mathbf{\Phi}}_F$ is a consistent estimator of $\mathbf{\Phi}$.

Step 2: Asymptotic distribution of $\widehat{\mathbf{\Phi}}$. We establish here the asymptotic normality of $\widehat{\mathbf{\Phi}}$. Since $\widehat{\mathbf{\Phi}}$ is the unique solution of $U_{F,n}(\mathbf{\Phi}) = 0$, we apply a Taylor expansion of $U_{F,n}(\widehat{\mathbf{\Phi}})$ around the true value of the parameter $\mathbf{\Phi}$.

Performing a first-order expansion of the score function around $\mathbf{\Phi}$, we obtain:

$$U_{F,n}(\widehat{\mathbf{\Phi}}) = U_{F,n}(\mathbf{\Phi}) + \left(\frac{1}{\sqrt{n}} \frac{\partial U_{F,n}(\mathbf{\Phi})}{\partial \mathbf{\Phi}^\top}\right) \sqrt{n}(\widehat{\mathbf{\Phi}} - \mathbf{\Phi}) + o_p(1).$$

Using the fact that $U_{F,n}(\widehat{\mathbf{\Phi}}) = 0$, we have:

$$0 = U_{F,n}(\mathbf{\Phi}) + J_{F,n}(\mathbf{\Phi})\sqrt{n}(\widehat{\mathbf{\Phi}} - \mathbf{\Phi}) + o_p(1),$$

where $J_{F,n}(\mathbf{\Phi}) = \frac{1}{\sqrt{n}} \frac{\partial U_{F,n}(\mathbf{\Phi})}{\partial \mathbf{\Phi}^\top}$.

Rearranging the terms, we obtain:

$$\sqrt{n}(\widehat{\mathbf{\Phi}} - \mathbf{\Phi}) = -J_{F,n}^{-1}(\mathbf{\Phi})U_{F,n}(\mathbf{\Phi}) + o_p(1).$$

Under appropriate regularity conditions (H1)-(H2) and using the Central Limit Theorem (CLT), the empirical score function follows an asymptotic normal distribution: $U_{F,n}(\mathbf{\Phi}) \xrightarrow{d} \mathcal{N}(0, I_F(\mathbf{\Phi}))$, where $I_F(\mathbf{\Phi}) = \text{Var}[U_{F,n}(\mathbf{\Phi})]$.

By the Weak Law of Large Numbers (WLLN), $J_{F,n}(\mathbf{\Phi})$ converges in probability to $J_F(\mathbf{\Phi})$, where $J_F(\mathbf{\Phi})$ is the non-singular by (H3). Applying Slutsky's theorem, we get:

$$J_{F,n}^{-1}(\mathbf{\Phi})U_{F,n}(\mathbf{\Phi}) \xrightarrow{d} J_F^{-1}(\mathbf{\Phi})\mathcal{N}(0, I_F(\mathbf{\Phi})).$$

Since $I_F(\mathbf{\Phi}) = J_F(\mathbf{\Phi})$, it follows that:

$$\sqrt{n}(\widehat{\mathbf{\Phi}} - \mathbf{\Phi}) \xrightarrow{d} \mathcal{N}(0, \Sigma_F),$$

where $\Sigma_F = J_F^{-1}(\mathbf{\Phi})I_F(\mathbf{\Phi})[J_F^{-1}(\mathbf{\Phi})]^\top = J_F^{-1}(\mathbf{\Phi})$. □

References

[1] Albert, J. M., Wang, W., Nelson, S., 2014. Estimating overall exposure effects for zero-inflated regression models with application to dental caries. Statistical Methods in Medical Research 23(3), 257-278.

[2] Ali, E., Diop, A., & Dupuy, J. F. (2020). A constrained marginal zero-inflated binomial regression model. Communications in Statistics - Theory and Methods, 51(18), 6396–6422.

[3] Duan, N., Manning, W. G., Morris, C. N., & Newhouse, J. P. 1983. A comparison of alternative models for the demand for medical care. Journal of business & economic statistics, 1(2), 115-126.

[4] Feuerverger, A. (1979). On some methods of analysis for weather experiments. Biometrika, 66(3), 655-658.

[5] Foutz, R. V., 1977. On the unique consistent solution to the likelihood equations. Journal of the American Statistical Association 72, 147-148.

[6] Hallstrom, A. P. (2010). A modified Wilcoxon test for non-negative distributions with a clump of zeros. Statistics in Medicine, 29(3), 391-400.

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

[8] Lee, A. H., Zhao, Y., Yau, K. K., & Xiang, L. (2010). How to analyze longitudinal multilevel physical activity data with many zeros?. Preventive medicine, 51(6), 476-481.

[9] Long, D. L., Preisser, J. S., Herring, A. H., Golin, C. E., 2014. A marginalized zero-inflated Poisson regression model with overall exposure effects. Statistics in medicine 33(29), 5151-5165.

[10] Long, D. L., Preisser, J. S., Herring, A. H., Golin, C. E., 2015. A Marginalized Zero-inflated Poisson Regression Model with Random Effects. Journal of the Royal Statistical Society. Series C, Applied statistics 64(5), 815-830.

[11] Martin, J., Hall, D. B., 2017. Marginal zero-inflated regression models for count data. Journal of Applied Statistics 44(10), 1807-1826.

[12] Mills, E. D. (2013). Adjusting for covariates in zero-inflated gamma and zero-inflated log-normal models for semicontinuous data. PhD thesis, University of Iowa.

[13] Preisser, J. S., Das, K., Long, D. L., and Divaris, K., 2016. Marginalized zero-inflated negative binomial regression with application to dental caries. Statistics in Medicine 35(10),1722-1735.

[14] R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

[15] Ridout, M., Hinde, J., Demetrio, C. G. B., 2001. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. Biometrics 57(1), 219-223.

[16] Valerie A. Smith, John S. Preisser, Brian Neelon, Matthew L. Maciejewski, 2016. A marginalized two-part model for semicontinuous data Statistics in Medicine 33(28),4891-4903.

[17] Todem, D., Kim, K., Hsu, W. W., 2016. Marginal mean models for zero-inflated count data. Biometrics 72(3), 986-994.

[18] Tu, W., & Zhou, X. H. 1999. A Wald test comparing medical costs based on log-normal distributions with zero valued costs. Statistics in medicine, 18(20), 2749-2761.

[19] Wang, Z., 2019 mpath : Regularized Linear Models, R package version 0.3-7.

[20] Vuong, Q. H., 1989 Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica, n○ 57, 307-333.