

Article

Skew-Normal Inflated Models: Mathematical Characterization and Applications to Medical Data with Excess of Zeros and Ones

Guillermo Martínez-Flórez ¹, Roger Tovar-Falón ¹, Víctor Leiva ^{2,*} and Cecilia Castro ^{3,*}

¹ Department of Mathematics and Statistics, Universidad de Córdoba, Montería 230002, Colombia; guillermomartinez@correo.unicordoba.edu.co (G.M.-F.); rjtovar@correo.unicordoba.edu.co (R.T.-F.)

² School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso 2362807, Chile

³ Centre of Mathematics, Universidade do Minho, 4710-057 Braga, Portugal

* Correspondence: victor.leiva@pucv.cl or victorleivasanchez@gmail.com (V.L.); cecilia@math.uminho.pt (C.C.)

Abstract: The modeling of data involving proportions, confined to a unit interval, is crucial in diverse research fields. Such data, expressing part-to-whole relationships, span from the proportion of individuals affected by diseases to the allocation of resources in economic sectors and the survival rates of species in ecology. However, modeling these data and interpreting information obtained from them present challenges, particularly when there is high zero–one inflation at the extremes of the unit interval, which indicates the complete absence or full occurrence of a characteristic or event. This inflation limits traditional statistical models, which often fail to capture the underlying distribution, leading to biased or imprecise statistical inferences. To address these challenges, we propose and derive the skew-normal zero–one inflated (SNZOI) models, a novel class of asymmetric regression models specifically designed to accommodate zero–one inflation presented in the data. By integrating a continuous-discrete mixture distribution with covariates in both continuous and discrete parts, SNZOI models exhibit superior capability compared to traditional models when describing these complex data structures. The applicability and effectiveness of the proposed models are demonstrated through case studies, including the analysis of medical data. Precise modeling of inflated proportion data unveils insights representing advancements in the statistical analysis of such studies. The present investigation highlights the limitations of existing models and shows the potential of SNZOI models to provide more accurate and precise inferences in the presence of zero–one inflation.

Keywords: censoring; proportion data; regression; skew-normal model; truncation; zero–one inflation

MSC: 62J05; 62E15



Citation: Martínez-Flórez, G.; Tovar-Falón, R.; Leiva, V.; Castro, C. Skew-Normal Inflated Models: Mathematical Characterization and Applications to Medical Data with Excess of Zeros and Ones. *Mathematics* **2024**, *12*, 2486. <https://doi.org/10.3390/math12162486>

Academic Editors: MinJae Lee and Manuel Alberto M. Ferreira

Received: 20 June 2024

Revised: 20 July 2024

Accepted: 8 August 2024

Published: 12 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Analysis of phenomena represented by proportion data, restricted to values between zero and one, is crucial in various scientific fields. These data describe part-to-whole relationships and appear in diverse applications, such as disease prevalence, resource distribution in economics, species survival rates, and habitat usage in ecology [1]. Accurately modeling these data poses challenges, especially when there is high zero–one inflation, which can reflect the absence or complete manifestation of a trait or event. Such inflation creates obstacles for traditional statistical models, like the tobit model [2–4], which, despite its prevalent application in economics, medicine, and social sciences, often struggles to accurately represent the underlying distribution of proportion data with inflated extremes.

To enhance the robustness of the tobit models, recent efforts have incorporated the Student-t [5,6], power-normal [7–9], and Birnbaum–Saunders [10,11] distributions, addressing asymmetry and kurtosis in the data distribution. Similarly, the adoption of the skew-normal (SN) distribution [12–14] has provided a framework for fitting data with inherent asymmetry, demonstrating the versatility required for contemporary datasets [10,15–17].

Recent advancements include the use of SN distributions in various applications, such as quantile estimation in child weight data and classification methods for serological status [18,19]. Furthermore, issues of inference and analyses utilizing SN distributions have been extensively studied, addressing challenges in modeling near-normality data [20–23].

The beta distribution, used for unit interval data when asymmetry is apparent, necessitates beta regression when independent variables (covariates) are considered [24]. However, standard approaches often fall short when zero–one inflation is present, limiting the existing models [25–27]. Recent studies have proposed extensions to the beta distribution to handle these limitations, such as the zero–one inflated beta models, which have shown promise in dealing with highly skew distributed data [28–33].

Despite strides in modeling data with inflation and asymmetry, a gap remains in adequately addressing zero–one inflation within data of proportions. Existing models do not fully capture the unique distributional characteristics and complexities introduced by these inflations, which can lead to biased estimators and imprecise inferences [34,35].

The main objective of the present study is to propose and derive skew-normal zero–one inflated (SNZOI) models, a novel class of asymmetric regressions specifically designed to address the challenges of zero–one inflation in proportion data confined to the unit interval. The SNZOI models integrate a continuous-discrete mixture distribution with covariates, enabling them to effectively capture the complex dynamics of such data. This integration accounts for asymmetry and mixed discrete-continuous characteristics of the data distribution, providing more precise and unbiased statistical inferences.

Traditional formulations, such as the tobit and beta regressions, often struggle with zero–one inflation, leading to imprecise model fits and biased inferences. Through detailed case studies, including the analysis of disease prevalence, we demonstrate the limitations of these traditional models. In contrast, the SNZOI models exhibit superior performance, offering deeper insights and more accurate fits to the data. To achieve precise parameter estimators, we employ the maximum likelihood (ML) method supported by iterative procedures like the Newton–Raphson algorithm. The accuracy of these ML estimators is rigorously assessed using the corresponding Fisher information matrix. The present study addresses a critical gap in the field of statistical modeling by providing a robust alternative for researchers dealing with inflated data in the unit interval. Our proposal details the development, application, and comparative analysis of SNZOI models, showing their potential to enhance statistical analysis in areas where zero–one inflation is common.

The remainder of this article is structured as follows, as illustrated in Figure 1. Section 2 provides background on skew models, focusing on the SN distribution and extensions to handle inflated data. In Section 3, we present regressions for data in the unit interval with inflation, detailing the complementary log–log SN, probit SN, and logit doubly censored models, along with their information matrices. In Section 4, the new SNZOI regression model is introduced, including its formulation, parameter estimation, and its particular case based on the probit structure. In Section 5, we demonstrate the application of these models through empirical case studies on doubly censored data and data inflated at one. Section 6 states a discussion of the key findings, comparisons with previous studies, limitations, and future research directions. The article concludes with Section 7.

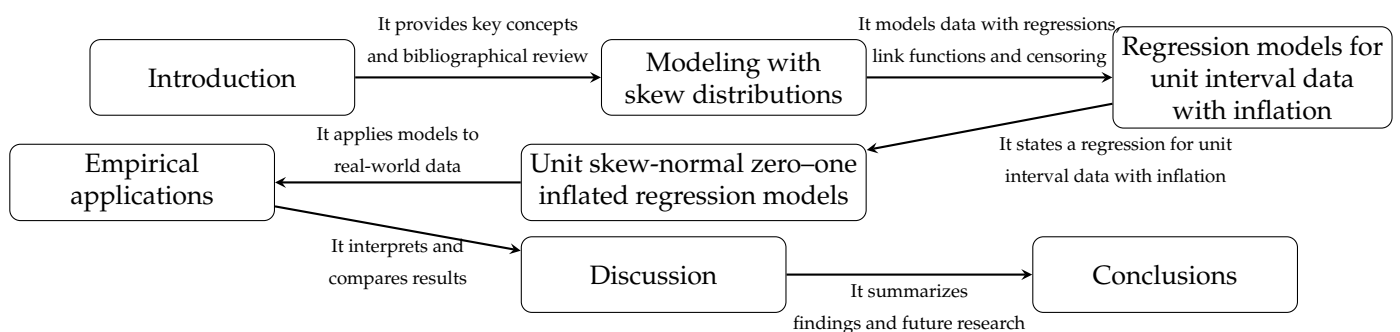


Figure 1. Flowchart of the article structure.

2. Modeling with Skew Distributions

In this section, we provide background on general concepts on statistical modeling, link functions, and skew models, with emphasis on the SN distribution. Extensions to the single and doubly censored cases are also considered. Also, in this section we indicate how the censoring thresholds are determined.

2.1. Fundamental Concepts of Regression and Link Functions

Various statistical models have been proposed to model data of rates and proportions within the closed unit interval denoted by $[0, 1]$. Let $\mathbf{x}_i = (1, x_{1i}, \dots, x_{pi})$ be the vector of p covariates for observation i , where the first element is a one to account for the intercept term, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$ be the vector of regression coefficients. The response (dependent) variable Y_i is modeled as

$$Y_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i = \mu_i + \varepsilon_i, \quad i \in \{1, \dots, n\}, \quad (1)$$

where ε_i is the error term.

The estimated regression coefficients are denoted by $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)^\top$. A common issue with some of these models is that the predicted value of the mean response, namely, $\hat{E}(Y_i) = \hat{\mu}_i = \mathbf{x}_i \hat{\boldsymbol{\beta}}$, may not always fall within the interval $[0, 1]$. To address this issue, continuous, strictly monotonic, and twice differentiable link functions are used to transform the mean response μ_i , ensuring that the predicted values fall within the unit interval.

Several link functions are commonly employed in practice, and they can be categorized as either symmetrical or asymmetrical [36]. Symmetrical link functions include the following:

- Logit link function $g(\mu) = \log(\mu / (1 - \mu))$.
- Probit link function $g(\mu) = \Phi^{-1}(\mu)$, where Φ^{-1} is the inverse of the standard normal cumulative distribution function (CDF).

These two links functions are characterized by their symmetry, that is, the response curve, which plots $g(\mu)$ against μ , is symmetric around μ , meaning they approach to the extremes at the same rate. Both the logit and probit link functions often yield similar results, except in extreme cases where predicted probabilities are close to 0 or 1. In such cases, the logit link function tends to compress extreme values more than the probit link function, leading to slightly different predictions at the boundaries of the probability scale.

In addition to symmetrical link functions, asymmetrical link functions are also utilized when the data are not symmetrically distributed within $[0, 1]$. Specifically, we have the following:

- Log-log link function $g(\mu) = \log(-\log(1 - \mu))$.
- Complementary log-log link function $g(\mu) = \log(-\log(\mu))$.

The log-log link function is suitable for modeling data where early occurrences are rare but become more frequent over time, effectively capturing the increasing hazard rate. Conversely, the complementary log-log link function is ideal for situations where the data belonging to $[0, 1]$ increase slowly from small to moderate values and then rises sharply near one. This function approaches to zero infinitely slower than others, making it particularly helpful for modeling skew distributed data with a heavy upper tail.

In summary, the rationale for selecting specific link functions is based on the characteristics of the data and the underlying assumptions of the model. For symmetrically distributed data, logit and probit link functions are often appropriate due to their symmetric nature around the mean μ . However, in cases where the response variable belonging to $[0, 1]$ has a distribution that demonstrates high skewness, particularly with a heavy tail, the complementary log-log link function provides a better fit due to its ability to handle such skewness effectively. For example, in survival analysis, where the event of interest might occur at the end of the observation period (resulting in a high proportion of censored data), the complementary log-log link function is commonly used.

Another scenario is in dose-response models, where the probability of a response might sharply increase after a certain dose level, making the complementary log–log link function more suitable. Additionally, the log–log link function may be advantageous when modeling time-to-event data, where early events are rare but become more frequent as time progresses. This function captures the increasing hazard over time, which is often seen in failure time data.

2.2. Skew-Normal Distribution and its Modeling

The probability density function (PDF) of the SN distribution is given by

$$\phi_{\text{SN}}(y; \theta) = \frac{2}{\omega} \phi\left(\frac{y - \xi}{\omega}\right) \Phi\left(\lambda \left(\frac{y - \xi}{\omega}\right)\right), \quad y \in \mathbb{R}, \quad (2)$$

where $\xi \in \mathbb{R}$ is a location parameter, $\omega \in \mathbb{R}^+$ is a scale parameter, $\lambda \in \mathbb{R}$ is an asymmetry parameter, ϕ is the standard normal PDF, and $\theta = (\xi, \omega, \lambda)$ is the parameter vector.

The notation $Y \sim \text{SN}(\xi, \omega, \lambda)$ indicates that Y follows an SN distribution with parameters ξ , ω , and λ . For $\lambda = 0$, the PDF stated in (2) reduces to the $N(\xi, \omega^2)$ PDF.

The SN distribution can be extended to regression by setting $\xi_i = x_i \beta$, where x_i is the observed covariate vector and β is the regression coefficient vector. Then, the response Y_i is modeled using the formulation stated in (1), where $\varepsilon_i \sim \text{SN}(0, \omega, \lambda)$, for $i \in \{1, \dots, n\}$.

Under the SN model, the expectation of the error term ε_i is presented as

$$E(\varepsilon_i) = \omega \sqrt{\frac{2}{\pi}} \left(\frac{\lambda}{\sqrt{1 + \lambda^2}} \right) \neq 0, \quad i \in \{1, \dots, n\},$$

which implies that $E(Y_i) \neq x_i \beta$. To correct this, an adjustment to the intercept is defined as

$$\beta_0^* = \beta_0 + \omega \sqrt{\frac{2}{\pi}} \left(\frac{\lambda}{\sqrt{1 + \lambda^2}} \right),$$

leading to $E(Y_i) = x_i \beta^*$. Thus, we redefine the coefficients as $\beta^* = (\beta_0^*, \beta_1, \dots, \beta_p)^\top$.

The log-SN (LSN) model extends the SN model to positive data and has PDF defined as

$$\phi_{\log\text{-SN}}(y; \xi, \omega, \lambda) = \frac{2}{\omega y} \phi\left(\frac{\log(y) - \xi}{\omega}\right) \Phi\left(\lambda \left(\frac{\log(y) - \xi}{\omega}\right)\right), \quad y \in \mathbb{R}^+,$$

where ξ and ω are stated in (2). For $\lambda = 0$, the LSN model is the log-normal model.

2.3. Single and Doubly Censored Data

In practical situations, the response may be censored, meaning that, for some observations, we only know that their values fall above or below a certain threshold. To handle such situations, researchers commonly use the tobit model, in which the true value of the response variable is only partially observed due to censoring.

To relax the constraint on tail probabilities inherent in the tobit specification, a two-part model was proposed in [37]. The PDF for the random variable Y_i in the two-part model, as proposed in [38], is given by

$$f(y_i) = \begin{cases} p_i, & \text{if } y_i \leq 0; \\ (1 - p_i)f(y_i), & \text{if } y_i > 0; \end{cases} \quad (3)$$

where f is the PDF of the model continuous part with positive support and $p_i = P(Y_i \leq 0) = P(Y_i = 0)$ is the probability of observation i being exactly zero due to censoring.

The model presented in (3) was generalized in [39], allowing censored (limited) responses to result from interval censoring and described by

$$f(y_i) = \begin{cases} p_i + (1 - p_i)F(c), & \text{if } y_i \leq c; \\ (1 - p_i)f(y_i), & \text{if } y_i > c; \end{cases}$$

where F is the CDF corresponding to the PDF f .

The standard tobit model is a special case of the Moulton–Halsey model stated in [39] when $p_i = 0$, indicating no inflation at zero. The Moulton–Halsey model was later extended to a two-part form in [37], where $\log(Y_i) \sim \text{SN}(x_i\beta, \omega, \lambda)$, with a probit link function defined as $p_i = \Phi(-x_i\beta)$. This extension is particularly advantageous for modeling skew distributed data with a heavy upper tail, as the probit function provides a smooth CDF-based transformation that effectively handles skewness. Furthermore, the Moulton–Halsey model, in its two-part form, accommodates limited responses resulting from interval censoring. This accommodation makes it suitable for scenarios where the response variable can take on a broader range of values, thereby improving the fit for data with complex distributions. Other probit and Heckman models can be found in [40–43].

In contrast to interval censoring, which limits responses to within specific ranges, in some cases data may be doubly censored, meaning that observations are only known to fall within a specific interval, while those outside this interval are recorded at the censoring points. Assume that Y^* follows an SN distribution with parameters ξ , ω , and λ with PDF defined in (2). Given a sample of size n , $\{Y_1^*, \dots, Y_n^*\}$ say, with observations $\{y_1^*, \dots, y_n^*\}$, only values within the interval $[c_0, c_1]$ are recorded. For $Y_i^* \leq c_0$ and $Y_i^* \geq c_1$, we record c_0 and c_1 , respectively. Therefore, for $i \in \{1, \dots, n\}$, the obtained values are defined as

$$y_i = \begin{cases} c_0, & \text{if } y_i^* \leq c_0; \\ y_i^*, & \text{if } c_0 < y_i^* < c_1; \\ c_1, & \text{if } y_i^* \geq c_1. \end{cases}$$

The sample $\{Y_1^*, \dots, Y_n^*\}$ is considered to be collected from a doubly censored skew-normal (DCSN) distribution, denoted by $Y_i^* \sim \text{DCSN}(\xi, \omega, \lambda)$, with PDF defined in (2) for $c_0 < Y_i^* < c_1$. For $Y_i^* \leq c_0$ and $Y_i^* \geq c_1$, we have

$$P(Y_i = c_0) = P(Y_i^* \leq c_0) = \Phi_{\text{SN}}(z_0), \quad P(Y_i = c_1) = P(Y_i^* \geq c_1) = 1 - \Phi_{\text{SN}}(z_1),$$

respectively, where $z_0 = (c_0 - \xi)/\omega$, $z_1 = (c_1 - \xi)/\omega$, and Φ_{SN} is the SN CDF. When $\lambda = 0$, the usual doubly censored normal tobit model follows. Considering $U = \log(Y)$, with $Y \sim \text{DCSN}(\xi, \omega, \lambda)$, the doubly censored log-skew-normal (DCLSN) model is obtained, denoted by $\text{DCLSN}(\xi, \omega, \lambda)$. Similarly, when $\lambda = 0$, we obtain the doubly censored log-normal tobit model.

2.4. Determination of Censoring Thresholds

The censoring thresholds c_0 and c_1 are determined based on prior knowledge about the data distribution or empirical observations. Typically, c_0 and c_1 are chosen to reflect cut-off points in the data. For example, c_0 might represent the minimum detectable limit of a measurement instrument, while c_1 could represent a saturation point beyond which measurements are no longer reliable.

The impact of censoring thresholds on parameter estimation may be important. If c_0 and c_1 are not appropriately chosen, they can lead to biased parameter estimators and misleading inference. To mitigate this, sensitivity analyses are often conducted to assess the robustness of the parameter estimators to different choices of c_0 and c_1 . Additionally, graphical methods, such as inspecting the distribution of the censored and uncensored observations, can provide insights into the appropriateness of the chosen thresholds.

For instance, one may plot the empirical CDF of the observed data and identify natural breakpoints or plateaus that suggest appropriate values for c_0 and c_1 . Furthermore, domain-specific knowledge can guide the selection of these thresholds, ensuring they align with practical considerations and the specific context of the study. By carefully determining and justifying the censoring thresholds, researchers may improve the reliability and validity of their parameter estimators and subsequent inferences.

3. Regression Models for Unit Interval Data with Inflation

In this section, we present three regression models for unit interval (proportion) data that account for inflation at specific values, denoted by c_0 and c_1 . The values c_0 and c_1 represent the points of inflation where the response variable can take the values zero, one, or any value between zero and one. These models are helpful in scenarios where there is a high probability of the response variable being exactly at these specific values.

3.1. Complementary Log–Log Unit Skew-Normal Regression Model

We construct the unit regression model inflated at the values c_0 and c_1 , where the link for the location parameter μ_i is the complementary log–log function $\log(\log(-\mu_i))$. We begin this construction by introducing the complementary log–log unit SN regression model truncated (SNT) at c_0 and c_1 , whose PDF is stated as

$$\phi_{\text{SNT}}(y_i; \theta) = \frac{\frac{1}{\omega} \phi_{\text{SN}}\left(\frac{y_i - \mu_i}{\omega}\right)}{\Phi_{\text{SN}}\left(\frac{c_1 - \mu_i}{\omega}\right) - \Phi_{\text{SN}}\left(\frac{c_0 - \mu_i}{\omega}\right)}, \quad c_0 < y_i < c_1, \quad i \in \{1, \dots, n\}, \quad (4)$$

where ϕ_{SN} and Φ_{SN} are defined in (2). The standardization terms, which help to normalize the data within the given bounds, are established as

$$z_{0i} = \frac{c_0 - \mu_i}{\omega}, \quad z_i = \frac{y_i - \mu_i}{\omega}, \quad z_{1i} = \frac{c_1 - \mu_i}{\omega}, \quad i \in \{1, \dots, n\}, \quad (5)$$

where μ_i is the location parameter, ω is the scale parameter, and $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})$ is a $1 \times (p + 1)$ observed covariate vector associated with the parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$. The full parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \omega, \lambda)$.

The log-likelihood function for a sample of size n , $Y_1, \dots, Y_n$ say, with observed values $y_1, \dots, y_n$ collected from the PDF ϕ_{SNT} presented in (4), is obtained by taking the natural logarithm of the joint likelihood function defined as $L(\boldsymbol{\theta}; \mathbf{y}) = \prod_{i=1}^n \phi_{\text{SNT}}(y_i; \boldsymbol{\theta})$, where now $\boldsymbol{\theta} = (\mu, \omega, \lambda)$. Taking the natural logarithm in the above expression, we obtain the log-likelihood function established as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \left(\log \left(\frac{\frac{1}{\omega} \phi_{\text{SN}}\left(\frac{y_i - \mu_i}{\omega}\right)}{\Phi_{\text{SN}}\left(\frac{c_1 - \mu_i}{\omega}\right) - \Phi_{\text{SN}}\left(\frac{c_0 - \mu_i}{\omega}\right)} \right) \right). \quad (6)$$

To simplify the logarithmic terms presented in (6), we break down the expression inside the logarithm as

$$\log \left(\frac{\frac{1}{\omega} \phi_{\text{SN}}\left(\frac{y_i - \mu_i}{\omega}\right)}{\Phi_{\text{SN}}\left(\frac{c_1 - \mu_i}{\omega}\right) - \Phi_{\text{SN}}\left(\frac{c_0 - \mu_i}{\omega}\right)} \right) = \log \left(\frac{1}{\omega} \right) + \log \left(\phi_{\text{SN}}\left(\frac{y_i - \mu_i}{\omega}\right) \right) - \log \left(\Phi_{\text{SN}}\left(\frac{c_1 - \mu_i}{\omega}\right) - \Phi_{\text{SN}}\left(\frac{c_0 - \mu_i}{\omega}\right) \right). \quad (7)$$

Note that the first term formulated in (7) simplifies to $\sum_{i=1}^n \log(1/\omega) = -n \log(\omega)$, the second term remains as $\sum_{i=1}^n \log(\phi_{\text{SN}}((y_i - \mu_i)/\omega))$, and the third term simplifies to $\sum_{i=1}^n \log(\Phi_{\text{SN}}((c_1 - \mu_i)/\omega) - \Phi_{\text{SN}}((c_0 - \mu_i)/\omega))$. Thus, the log-likelihood function (with no constant term and using the symbol $\propto$ indicating proportional to) is now given by

$$\ell(\theta) \propto -n \log(\omega) + \sum_{i=1}^n \log\left(\phi_{\text{SN}}\left(\frac{y_i - \mu_i}{\omega}\right)\right) - \sum_{i=1}^n \log\left(\Phi_{\text{SN}}\left(\frac{c_1 - \mu_i}{\omega}\right) - \Phi_{\text{SN}}\left(\frac{c_0 - \mu_i}{\omega}\right)\right). \quad (8)$$

To further simplify and generate the score equations, we introduce the following notations: $\delta_i = (1 - \mu_i) \log(1 - \mu_i)$, $w_i^\lambda = \phi_{\text{SN}}(\lambda z_i) / \Phi_{\text{SN}}(\lambda z_i)$,

$$\Delta_{jki} = \frac{z_{1i}^j \phi_{\text{SN}}^k(z_{1i}; \lambda) - z_{0i}^j \phi_{\text{SN}}^k(z_{0i}; \lambda)}{\Phi_{\text{SN}}(z_{1i}; \lambda) - \Phi_{\text{SN}}(z_{0i}; \lambda)}, \quad \Gamma_{jki} = \frac{z_{1i}^j \phi_{\text{SN}}^k(\sqrt{1 + \lambda^2} z_{1i}) - z_{0i}^j \phi_{\text{SN}}^k(\sqrt{1 + \lambda^2} z_{0i})}{\Phi_{\text{SN}}(z_{1i}; \lambda) - \Phi_{\text{SN}}(z_{0i}; \lambda)},$$

where δ_i is a term involving the mean response μ_i related to the complementary log–log link function and w_i^λ is the ratio of the SN PDF and CDF, whereas Δ_{jki} and Γ_{jki} are terms involving the standardized variables z_i , z_{1i} , and z_{0i} .

By taking the first derivatives of the function presented in (8) with respect to the parameters, we obtain the score elements. For the coefficient β_j , the element is stated as

$$\dot{\ell}(\beta_j) = -\frac{1}{\omega} \sum_{i=1}^n x_{ij} z_i \delta_i + \frac{\lambda}{\omega} \sum_{i=1}^n x_{ij} w_i^\lambda \delta_i - \frac{1}{\omega} \sum_{i=1}^n x_{ij} \Delta_{01i} \delta_i, \quad j \in \{0, 1, \dots, p\}. \quad (9)$$

For the scale parameter ω , the score function is formulated as

$$\dot{\ell}(\omega) = -\frac{n}{\omega} + \frac{1}{\omega} \sum_{i=1}^n z_i^2 - \frac{\lambda}{\omega} \sum_{i=1}^n z_i w_i^\lambda + \frac{1}{\omega} \sum_{i=1}^n \Delta_{11i}. \quad (10)$$

For the shape parameter λ , the score is defined as

$$\dot{\ell}(\lambda) = \sum_{i=1}^n w_i^\lambda z_i - \sqrt{\frac{2}{\pi}} \left(\frac{1}{1 + \lambda^2} \right) \sum_{i=1}^n \Gamma_{01i}. \quad (11)$$

The solution of the system of equations obtained by setting (9), (10), and (11) to zero yields the ML estimates of the parameters. This system is typically resolved using iterative numerical algorithms such as Newton–Raphson or quasi-Newton, which iteratively update parameter estimates to maximize the likelihood function.

The Newton–Raphson algorithm is known for its quadratic convergence properties, provided that the initial estimates are close to the true parameter values and the likelihood function is well-behaved. Specifically, the Newton–Raphson update rule is given by $\theta^{(k+1)} = \theta^{(k)} - \ddot{\ell}(\theta^{(k)})^{-1} \dot{\ell}(\theta^{(k)})$, where $\ddot{\ell}(\theta) = -\partial^2 \ell(\theta) / \partial \theta \partial \theta^\top$ is the Hessian matrix, from which the observed Fisher information matrix can be obtained, and $\dot{\ell}(\theta) = \partial \ell(\theta) / \partial \theta$ is its score function. The observed information matrix provides a local approximation to the curvature of the log-likelihood surface, which is used to iteratively update the parameter estimates. In cases where the likelihood surface is complex or the initial estimates are not close to the true values, quasi-Newton algorithms, which approximate the Hessian matrix, can provide more robust convergence. These algorithms do not require the explicit computation of the Hessian matrix at each iteration. Instead, they utilize rank-one updates to approximate the inverse Hessian matrix.

The popular quasi-Newton algorithm includes the Broyden–Fletcher–Goldfarb–Shanno (BFGS) and limited-memory BFGS (L-BFGS) approaches. The BFGS algorithm iteratively updates the parameter estimates as $\theta^{(k+1)} = \theta^{(k)} + q_k p^{(k)}$, where q_k is a step length determined by a line search and $p^{(k)} = -H_k \dot{\ell}(\theta^{(k)})$ is the search direction. Here, H_k is the approximation to the inverse Hessian matrix at iteration k updated as $H_{k+1} = (I - \rho_k s_k y_k^\top) H_k (I - \rho_k y_k s_k^\top) + \rho_k s_k s_k^\top$, where I is the identity matrix, $s_k = \theta^{(k+1)} - \theta^{(k)}$, $y_k = \dot{\ell}(\theta^{(k+1)}) - \dot{\ell}(\theta^{(k)})$, and $\rho_k = 1 / y_k^\top s_k$. The L-BFGS algorithm, a variant of the BFGS algorithm, is particularly helpful for large-scale optimization problems as it maintains only a limited number of vectors to approximate the inverse Hessian matrix, so reducing computer memory requirements.

The update rule for L-BFGS is similar to BFGS but uses a limited memory representation, as stated by $\mathbf{H}_{k+1} \approx \sum_{i=k-m+1}^k (\mathbf{I} - \rho_i \mathbf{s}_i \mathbf{y}_i^\top) \mathbf{H}_i (\mathbf{I} - \rho_i \mathbf{y}_i \mathbf{s}_i^\top) + \rho_i \mathbf{s}_i \mathbf{s}_i^\top$, where m is the number of past updates stored. This rule ensures efficiency in the optimization, making it suitable for maximizing the likelihood function in complex models [44].

3.2. Probit Unit Skew-Normal Regression Model

We now examine the unit SN model with the probit link defined as $\mu_i = \Phi(\mathbf{x}_i \boldsymbol{\beta})$. This is similar to the case of the complementary log–log link in the unit SN model. The PDF ϕ_{SNT} used to construct the likelihood function of the model is represented in (4), where z_i , z_{1i} , and z_{0i} are stated in (5), with μ_i being the location parameter, ω the scale parameter, and $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})$, a $1 \times (p+1)$ observed covariate vector associated with $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$. For parameter estimation, the score function for β_j is given by

$$\dot{\ell}(\beta_j) = \frac{1}{\omega} \sum_{i=1}^n x_{ij} z_i \phi(\mathbf{x}_i \boldsymbol{\beta}) - \frac{\lambda}{\omega} \sum_{i=1}^n x_{ij} w_i^\lambda \phi(\mathbf{x}_i \boldsymbol{\beta}) + \frac{1}{\omega} \sum_{i=1}^n x_{ij} \Delta_{01i} \phi(\mathbf{x}_i \boldsymbol{\beta}), \quad j \in \{0, 1, \dots, p\}.$$

The elements of the score equation for the parameters ω and λ are defined in (10) and (11), respectively, with the standardization terms z_i , z_{1i} , and z_{0i} stated in (5), as mentioned.

3.3. Logit Unit Skew-Normal Regression Model

Now, we explore the unit SN model with the logit link function expressed as

$$\mu_i = \frac{\exp(\mathbf{x}_i \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i \boldsymbol{\beta})}, \quad i \in \{1, \dots, n\}. \quad (12)$$

For this model, the parameters are interpreted from odds ratios between the prediction or response odds when one covariate is increased by m units (with the rest of the covariates remaining constant) and the odds without this increase. Note that this ratio is given by $\exp(m\beta_k)$, where β_k is the parameter associated with the covariate k increased by m units. As in the two link functions previously analyzed, the distribution of the response variable is $Y_i \sim \text{SN}(\mu_i, \omega, \lambda)$, for $i \in \{1, \dots, n\}$.

From the logit unit SN regression, the logit normal case is obtained with $\lambda = 0$. The parameter estimates of the SN regression model for the response in the unit interval with logit link function can be obtained using the ML method. The corresponding log-likelihood function for a sample of n observations is obtained similarly from the PDF ϕ_{SNT} defined in (4), where z_i , z_{1i} , and z_{0i} are given in (5) with μ_i being stated in (12). The elements of the score function for ω and λ are similar to those of the complementary log–log unit SN model, but with μ_i being formulated in (12), while for the parameter β_j , we have that

$$\dot{\ell}(\beta_j) = \frac{1}{\omega} \sum_{i=1}^n x_{ij} (z_i - \lambda w_i^\lambda + \Delta_{01i}) \mu_i (1 - \mu_i), \quad j \in \{0, 1, \dots, p\}.$$

As mentioned, the score equation system is reached by setting the elements of the score function to zero, obtained from the first derivative of the function with respect to the parameters $\beta_0, \beta_1, \dots, \beta_p, \omega$, and λ . Solving this system of equations provides the ML estimates. To maximize this function, iterative algorithmic methods are necessary.

3.4. Information Matrices in Skew-Normal Models

To analyze the properties and behavior of SN regression models, it is essential to examine the information matrices. These matrices provide insights into the variance and covariance of the parameter estimators, as well as their overall performance. Specifically, the observed information matrix can be approximated by the negative of the Hessian matrix, which is obtained from the second derivatives of the log-likelihood function.

To obtain the Hessian matrix elements in the complementary log–log case, we define

$$\tau_{jki} = \frac{z_{1i}^j \phi_{\text{SN}}^k(z_{1i}) - z_{0i}^j \phi_{\text{SN}}^k(z_{0i})}{\Phi_{\text{SN}}(z_{1i}; \lambda) - \Phi_{\text{SN}}(z_{0i}; \lambda)}.$$

The second derivatives of the log-likelihood function for $\beta_j \beta_{j'}$, $\beta_j \omega$, $\beta_j \lambda$, $\omega \omega$, $\omega \lambda$, and $\lambda \lambda$ are, respectively, expressed as

$$\begin{aligned} \ddot{\ell}(\beta_j \beta_{j'}) &= -\frac{1}{\omega} \sum_{i=1}^n x_{ij} x_{ij'} \delta_i (\log(1 - \mu_i) + 1) (z_i + \lambda w_i^\lambda - \Delta_{01i}) \\ &\quad - \frac{1}{\omega^2} \sum_{i=1}^n x_{ij} x_{ij'} \delta_i^2 \tau_{01i} (\phi(z_{1i}) + \phi(z_{0i})) - \frac{1}{\omega^2} \sum_{i=1}^n x_{ij} x_{ij'} \delta_i^2 (1 + \lambda^2 w_i^\lambda (\lambda z_i + w_i^\lambda) - (\Delta_{11i} + \Delta_{01i}^2)), \\ \ddot{\ell}(\beta_j \omega) &= \frac{2}{\omega^2} \sum_{i=1}^n x_{ij} \delta_i z_i - \frac{\lambda}{\omega} \sum_{i=1}^n x_{ij} \delta_i w_i^\lambda + \frac{\lambda^2}{\omega^2} \sum_{i=1}^n x_{ij} \delta_i w_i^\lambda z_i (\lambda z_i + w_i^\lambda) - \frac{1}{\omega^2} \sum_{i=1}^n x_{ij} \delta_i (\Delta_{01i} (\Delta_{11i} - 1) + (\Delta_{21i} - 2\tau_{12i})), \\ \ddot{\ell}(\beta_j \lambda) &= \frac{1}{\omega} \sum_{i=1}^n x_{ij} \delta_i w_i^\lambda - \frac{\lambda}{\omega} \sum_{i=1}^n x_{ij} \delta_i w_i^\lambda z_i (\lambda z_i + w_i^\lambda) - \sqrt{\frac{2}{\pi}} \sum_{i=1}^n x_{ij} \delta_i \left(\Gamma_{11i} + \left(\frac{1}{1 + \lambda^2} \right) \Delta_{01i} \Gamma_{01i} \right), \\ \ddot{\ell}(\omega \omega) &= -\frac{1}{\omega^2} \sum_{i=1}^n (-1 + 3z_i^2 - 2\lambda z_i w_i^\lambda + \lambda^2 z_i^2 w_i^\lambda (\lambda z_i + w_i^\lambda)) - \frac{1}{\omega^2} \sum_{i=1}^n \Delta_{11i} - \frac{1}{\omega} \sum_{i=1}^n (\Delta_{11i} - \Delta_{21i} - \Delta_{11i}^2 - 2\tau_{12i}), \\ \ddot{\ell}(\omega \lambda) &= -\frac{1}{\omega} \sum_{i=1}^n w_i^\lambda z_i (1 - \lambda z_i (\lambda z_i + w_i^\lambda)) + \sqrt{\frac{2}{\pi}} \frac{1}{\omega} \sum_{i=1}^n \left(\Gamma_{21i} + \left(\frac{1}{1 + \lambda^2} \right) \Delta_{11i} \Gamma_{01i} \right), \\ \ddot{\ell}(\lambda \lambda) &= -\sum_{i=1}^n w_i^\lambda z_i^2 (\lambda z_i + w_i^\lambda) - \sqrt{\frac{2}{\pi}} \frac{1}{(1 + \lambda^2)^2} \sum_{i=1}^n \left(\sqrt{\frac{2}{\pi}} \Gamma_{01i} - 2\lambda \right) \Gamma_{01i} + 2\sqrt{\frac{2}{\pi}} \left(\frac{\lambda}{1 + \lambda^2} \right) \sum_{i=1}^n \Gamma_{21i}. \end{aligned}$$

The elements of the Hessian matrix in the probit case are given by

$$\begin{aligned} \ddot{\ell}(\beta_j \beta_{j'}) &= -\frac{1}{\omega} \sum_{i=1}^n x_{ij} x_{ij'} \phi(\mathbf{x}_i \boldsymbol{\beta}) \left(\phi(\mathbf{x}_i \boldsymbol{\beta}) + (\mathbf{x}_i \boldsymbol{\beta}) z_i + \lambda w_i^\lambda \left(-\mathbf{x}_i \boldsymbol{\beta} + \lambda \phi(\mathbf{x}_i \boldsymbol{\beta}) (\lambda z_i + w_i^\lambda) \right) \right) \\ &\quad + \frac{1}{\omega^2} \sum_{i=1}^n x_{ij} x_{ij'} \phi(\mathbf{x}_i \boldsymbol{\beta}) \Delta_{01i} (\mathbf{x}_i \boldsymbol{\beta} + \phi(\mathbf{x}_i \boldsymbol{\beta}) \Delta_{01i}) - \frac{1}{\omega^2} \sum_{i=1}^n x_{ij} x_{ij'} \phi^2(\mathbf{x}_i \boldsymbol{\beta}) (\Delta_{11i} + \tau_{01i} (\phi(z_{1i}) + \phi(z_{0i}))), \\ \ddot{\ell}(\beta_j \omega) &= -\frac{1}{\omega^2} \sum_{i=1}^n x_{ij} \phi(\mathbf{x}_i \boldsymbol{\beta}) (2 - \lambda w_i^\lambda + \lambda^2 w_i^\lambda z_i (\lambda z_i + w_i^\lambda)) - \frac{1}{\omega^2} \sum_{i=1}^n x_{ij} \phi(\mathbf{x}_i \boldsymbol{\beta}) (\Delta_{01i} (1 - \Delta_{11i}) - (\Delta_{21i} - 2\tau_{12i})), \\ \ddot{\ell}(\beta_j \lambda) &= -\frac{1}{\omega} \sum_{i=1}^n x_{ij} \phi(\mathbf{x}_i \boldsymbol{\beta}) w_i^\lambda (1 - \lambda z_i (\lambda z_i + w_i^\lambda)) + \sqrt{\frac{2}{\pi}} \sum_{i=1}^n x_{ij} \phi(\mathbf{x}_i \boldsymbol{\beta}) \left(\Gamma_{11i} + \left(\frac{1}{1 + \lambda^2} \right) \Delta_{01i} \Gamma_{01i} \right). \end{aligned}$$

The Hessian matrix elements for the parameters ω and λ in the logit case follow the same structure as those for the complementary log–log unit SN model, with the standardization terms z_i , z_{1i} , and z_{0i} defined in (5) for $\mu_i = \Phi(\mathbf{x}_i \boldsymbol{\beta})$.

For the probit case, the elements of the Hessian matrix are similarly obtained as for the complementary log–log and logit cases and are represented as

$$\begin{aligned} \ddot{\ell}(\beta_j \beta_l) &= -\frac{1}{\omega^2} \sum_{i=1}^n x_{ij} x_{il} \mu_i^2 (1 - \mu_i)^2 \left(1 + \lambda^2 w_i^\lambda (\lambda z_i + w_i^\lambda) \right) - \frac{1}{\omega} \sum_{i=1}^n x_{ij} x_{il} \mu_i (1 - \mu_i) (1 - 2\mu_i) \left(-z_i + \lambda w_i^\lambda \right) \\ &\quad + \frac{1}{\omega^2} \sum_{i=1}^n x_{ij} x_{il} \mu_i (1 - \mu_i) \left(\omega (1 - 2\mu_i) + \mu_i (1 - \mu_i) (\Delta_{11i} - \Delta_{01i}^2 - \tau_{02i}) \right), \\ \ddot{\ell}(\beta_j \omega) &= -\frac{1}{\omega^2} \sum_{i=1}^n x_{ij} \mu_i (1 - \mu_i) \left(2z_i - \lambda w_i^\lambda + \lambda^2 z_i w_i^\lambda (\lambda z_i + w_i^\lambda) \right) - \frac{1}{\omega^2} \sum_{i=1}^n x_{ij} \mu_i (1 - \mu_i) (\Delta_{01i} (1 - \Delta_{11i}) - \Delta_{21i} - 2\tau_{12i}), \\ \ddot{\ell}(\beta_j \lambda) &= -\frac{1}{\omega} \sum_{i=1}^n x_{ij} \mu_i (1 - \mu_i) \left(w_i^\lambda (1 - \lambda z_i (\lambda z_i + w_i^\lambda)) - \left(\Gamma_{11i} + \left(\frac{1}{1 + \lambda^2} \right) \Delta_{01i} \Gamma_{01i} \right) \right). \end{aligned}$$

The Hessian matrix elements for the parameters ω and λ in the probit case are the same as the complementary log–log and logit unit SN models, with its respective link function.

The elements of the observed information matrix can be written as $\ddot{\ell}_{\theta_j\theta_l} = \sum_{i=1}^n j_{\theta_j\theta_l}$, where $j_{\theta_j\theta_l} = -\partial^2 \ell(\theta) / \partial \theta_j \partial \theta_l = -\ddot{\ell}(\theta_j\theta_l)$ and the sum is over the set of observations $i \in \{1, \dots, n\}$.

To study the conditions of the information matrix of the models under consideration, we analyze this matrix for the general case, that is, the non-inflated case at c_0 and c_1 . Thus, when $\lambda = 0$, the information matrix is stated as

$$\mathcal{I}_{\text{SN}}(\beta, \omega, \lambda) = \begin{pmatrix} \frac{1}{\omega^2} (\mathbf{M}\mathbf{x})^\top (\mathbf{M}\mathbf{x}) & \mathbf{0}_{p+1} & \frac{\sqrt{2}}{\pi\omega} \mathbf{x}^\top \mathbf{M}\mathbf{1}_n \\ \mathbf{0}_{p+1}^\top & \frac{2n}{\omega^2} & \mathbf{0}_{p+1} \\ \frac{\sqrt{2}}{\pi\omega} \mathbf{1}_n^\top \mathbf{M}\mathbf{x} & \mathbf{0}_{p+1} & \frac{2n}{\pi^2} \end{pmatrix},$$

where $\mathbf{0}_{p+1}$ is a $(p+1) \times 1$ vector of zeros, $\mathbf{1}_n$ is an $n \times 1$ vector of ones, $\mathbf{x} = (\mathbf{1}_n, \mathbf{x}')$, where $\mathbf{x}' = (x_1, \dots, x_p)$ is an $n \times p$ matrix, with $x_j = (x_{j1}, \dots, x_{jn})^\top$, for $j \in \{1, \dots, p\}$, and $\mathbf{M}$ being a diagonal matrix stated as $\mathbf{M} = \text{diag}\{\mu_1(1-\mu_1), \dots, \mu_n(1-\mu_n)\}$ for the logit link, whereas

$$\mathbf{M} = \text{diag}\{-\delta_1, \dots, -\delta_n\}, \quad \mathbf{M} = \text{diag}\{\phi(x_1\beta), \dots, \phi(x_n\beta)\},$$

for the complementary log–log and probit links, respectively. The determinant of $\mathcal{I}_{\text{SN}}$ is presented as

$$|\mathcal{I}_{\text{SN}}(\beta, \omega, \lambda)| = \frac{2n\sqrt{2}/\pi^2}{\omega^{2(p+1)}} \left| (\mathbf{M}\mathbf{x})^\top (\mathbf{M}\mathbf{x}) \right| \left| n - \mathbf{1}_n^\top \mathbf{M}\mathbf{x} (\mathbf{x}^\top \mathbf{M}\mathbf{M}\mathbf{x})^{-1} \mathbf{x}^\top \mathbf{M}\mathbf{1}_n \right|.$$

Letting $k = \mathbf{1}_n^\top \mathbf{M}\mathbf{M}\mathbf{1}_n$, $l = \mathbf{1}_n^\top \mathbf{M}\mathbf{1}_n$ and writing $\mathbf{x}$ in partitioned form as $\mathbf{x} = (\mathbf{1}_n, \mathbf{x}')$, it follows that the matrix $\mathbf{x}^\top \mathbf{M}\mathbf{M}\mathbf{x}$ can be expressed as a partitioned matrix.

We may use the existing expressions of matrix algebra to find the inverse of a partitioned matrix, since $\mathbf{M}$ is a diagonal matrix. Therefore, the rank of the matrix $\mathbf{M}\mathbf{x}$ is the same rank of $\mathbf{x}$, while $\mathbf{M}\mathbf{x}^\top \mathbf{M}\mathbf{x}$ is also full rank and then invertible. After some calculations, we obtain

$$\mathbf{1}_n^\top \mathbf{M}\mathbf{x} (\mathbf{x}^\top \mathbf{M}\mathbf{M}\mathbf{x})^{-1} \mathbf{x}^\top \mathbf{M}\mathbf{1}_n = \frac{l^2}{k} + \mathbf{1}_n^\top \mathbf{R}^\top \mathbf{V}^{-1} \mathbf{R}\mathbf{1}_n,$$

where

$$\mathbf{R} = \mathbf{x}'\mathbf{M}(k\mathbf{I} - \mathbf{I}), \quad \mathbf{V} = \mathbf{x}'\mathbf{M}(\mathbf{I} - k^{-1}\mathbf{M}\mathbf{J}\mathbf{M}),$$

with $\mathbf{I}_n$ being the $n \times n$ identity matrix and $\mathbf{J} = \mathbf{1}_n\mathbf{1}_n^\top$ the matrix of ones. Noting that $\mathbf{1}_n^\top \mathbf{R}^\top \mathbf{V}^{-1} \mathbf{R}\mathbf{1}_n > 0$ is a quadratic form, it is clear that

$$\left| n - \mathbf{1}_n^\top \mathbf{M}\mathbf{x} (\mathbf{x}^\top \mathbf{M}\mathbf{M}\mathbf{x})^{-1} \mathbf{x}^\top \mathbf{M}\mathbf{1}_n \right| \neq 0.$$

Note that, for our models, this determinant is a function of δ_i , $\mu_i(1-\mu_i)$, and $\phi(x_i\beta)$ for $i \in \{1, \dots, n\}$. In particular, for the logit model, assuming that $0 < x_i\beta < 1$, it follows that

$$\frac{16n}{e^2(1+e)^2} < \frac{l^2}{k} = \frac{(\sum_{i=1}^n \mu_i(1-\mu_i))^2}{\sum_{i=1}^n \mu_i^2(1-\mu_i)^2} < \frac{e^2(1+e)^{4n}}{16},$$

where e is the Euler or exponential number. For the complementary log–log and probit models, l^2/k is given by $(-\sum_{i=1}^n \delta_i)^2 / \sum_{i=1}^n \delta_i^2$ and $(\sum_{i=1}^n \phi(x_i\beta))^2 / \sum_{i=1}^n \phi^2(x_i\beta)$, respectively. Then, the information matrix is non-singular for each studied case of the unit SN model.

4. Unit Skew-Normal Zero-One Inflated Regression Models

In this section, we describe zero-one inflated data and explain how the ML method is used to estimate the parameters of the SNZOI model inflated at $c_0 = 0$ and $c_1 = 1$.

4.1. Formulation of the Skew-Normal Zero-One Inflated Model

With the inflations set at $c_0 = 0$ and $c_1 = 1$, the PDF for the SNZOI model is stated as

$$f(y_i) = \begin{cases} p_{0i}, & \text{if } y_i = 0; \\ \left(\frac{1 - p_{0i} - p_{1i}}{\omega} \right) \frac{\phi_{\text{SN}}((y_i - \mu_i)/\omega)}{(\Phi_{\text{SN}}((1 - \mu_i)/\omega) - \Phi_{\text{SN}}(-\mu_i/\omega))}, & \text{if } 0 < y_i < 1; \\ p_{1i}, & \text{if } y_i = 1; \end{cases}$$

with $\mu_i = g^{-1}(x_i\beta)$ for $i \in \{1, \dots, n\}$, where $p_{0i} = P(Y_i = 0)$ and $p_{1i} = P(Y_i = 1)$ are defined for the SN model. We assume that the responses inflated at $c_0 = 0$ and $c_1 = 1$ are explained by the observed covariates $x_{0i} = (1, x_{0i1}, \dots, x_{0il})$ and $x_{1i} = (1, x_{1i1}, \dots, x_{1ir})$, respectively, and data in $0 < y_i < 1$ correspond to the part of the SN model stated by a $1 \times (p + 1)$ observed covariate vector $x_i = (1, x_{i1}, \dots, x_{ip})$. Hence, the regression parameters for the SNZOI model are $\beta_0 = (\beta_{00}, \beta_{01}, \dots, \beta_{0l})^\top$, $\beta = (\beta_0, \beta_1, \dots, \beta_p)^\top$, and $\beta_1 = (\beta_{10}, \beta_{11}, \dots, \beta_{1r})^\top$.

Consider $Y_i \sim \text{SN}(\mu_i, \omega, \lambda)$, for $0 < y_i < 1$ and $i \in \{1, \dots, n\}$. Assuming a polychotomous random variable with a logit link function, we have

$$\begin{aligned} p_{0i} &= P(Y_i = 0) = \frac{\exp(x_{0i}\beta_0)}{1 + \exp(x_{0i}\beta_0) + \exp(x_{1i}\beta_1)}, \\ p_{1i} &= P(Y_i = 1) = \frac{\exp(x_{1i}\beta_1)}{1 + \exp(x_{0i}\beta_0) + \exp(x_{1i}\beta_1)}, \\ p_i &= P(0 < Y_i < 1) = 1 - p_{0i} - p_{1i} = \frac{1}{1 + \exp(x_{0i}\beta_0) + \exp(x_{1i}\beta_1)}, \end{aligned}$$

where $x_{0i} = (1, x_{0i1}, \dots, x_{0il})$ and $x_{1i} = (1, x_{1i1}, \dots, x_{1ir})$ are $1 \times (l + 1)$ and $1 \times (r + 1)$ observed covariate vectors associated with the parameter vectors β_0 and β_1 , respectively. A particular case of this model addresses zero-one response probabilities using β_0 and β_1 , and responses within the unit interval using β with the SN distribution. This particular case serves as an alternative to the beta regression model studied in [25].

4.2. Parameter Estimation in the SNZOI Model

Estimation of the parameters for the discrete part of the SNZOI model can be carried out using the ML method. Additionally, with the MNP library of the R software (version 4.3.2) [45], the estimates of the probit model can be obtained using the procedure described in [43]. For the continuous part, $\phi_{\text{SN}}(y_i; \lambda)$ namely, the score function and the elements of the Fisher information matrices were derived in the previous section for the complementary log-log, logit, and probit unit SN models. Similar to the previous case, the estimates of the parameters for the discrete and continuous parts can be obtained separately due to orthogonality. The estimation of the parameters $(\beta_0, \beta, \beta_1, \omega, \lambda)$ for the SNZOI model encompasses the parameter vectors β_0, β, β_1 , along with the parameters ω and λ .

Estimation of the SNZOI parameters can be performed using the ML method by maximizing the log-likelihood function, decomposed into parts for the discrete and continuous components of the model given by $\ell(\theta) = \ell(\beta_0, \beta_1) + \ell(\beta, \omega, \lambda)$, where

$$\begin{aligned} \ell(\beta_0, \beta_1) &\propto \sum_{i \in L} x_{0i}\beta_0 + \sum_{i \in R} x_{1i}\beta_1 - \sum_{i=1}^n \log(1 + \exp(x_{0i}\beta_0) + \exp(x_{1i}\beta_1)), \\ \ell(\beta, \omega, \lambda) &\propto \sum_{i \in S} \left(-\frac{1}{2} \left(\frac{y_i - \mu_i}{\omega} \right)^2 + \log \left(\Phi \left(\frac{\lambda(y_i - \mu_i)}{\omega} \right) \right) \right) - \sum_{i \in S} \log \left(\Phi_{\text{SN}} \left(\frac{1 - \mu_i}{\omega} \right) - \Phi_{\text{SN}} \left(-\frac{\mu_i}{\omega} \right) \right) + n_{01}(\log(2) - \log(\omega)), \end{aligned}$$

with n_{01} being the number of observations between zero and one. The set L refers to the zero-inflated (left) data with $y_i = 0$, the set R refers to the one-inflated (right) data with $y_i = 1$, and the set S includes the data $0 < y_i < 1$, for $i \in \{1, \dots, n\}$.

The score function for $\ell(\beta_0, \beta_1)$, leading to the ML estimates, is given by

$$\begin{aligned}\dot{\ell}(\beta_{0k}) &= \sum_{i \in L} x_{0ik} - \sum_{i=1}^n \frac{x_{0ik} \exp(x_{0i}\beta_0)}{1 + \exp(x_{0i}\beta_0) + \exp(x_{1i}\beta_1)}, \\ \dot{\ell}(\beta_{1m}) &= \sum_{i \in R} x_{1im} - \sum_{i=1}^n \frac{x_{1im} \exp(x_{1i}\beta_1)}{1 + \exp(x_{0i}\beta_0) + \exp(x_{1i}\beta_1)},\end{aligned}$$

for $k \in L \equiv \{1, \dots, l\}$ and $m \in R \equiv \{1, \dots, r\}$. The score function for $\ell(\beta, \omega, \lambda)$ depends on the complementary log–log, logit, or probit link function used. These scores were already obtained previously for these three models. The system of equations reached by equating the scores to zero has no closed-form solution and needs to be solved numerically. The Fisher information matrix is block diagonal, that is, it can be written as

$$\mathcal{I}(\theta) = \text{diag}\{\mathcal{I}(\beta_0, \beta_1), \mathcal{I}(\beta, \omega, \lambda)\},$$

where $\mathcal{I}(\beta_0, \beta_1)$ is the component of the information matrix for the discrete part of the model and $\mathcal{I}(\beta, \omega, \lambda)$ is its continuous part. The elements of $\mathcal{I}(\beta_0, \beta_1)$ were described in Section 4, while the elements of $\mathcal{I}(\beta, \omega, \lambda)$ for the complementary log–log, logit, and probit links were detailed in Section 3. Hence, for large n , we have that $\hat{\theta} \sim N_{p+l+r+5}(\theta, \Sigma_{\theta\theta})$, meaning that $\hat{\theta}$ is consistent and asymptotically normal distributed with $\Sigma_{\theta\theta} = \mathcal{I}(\theta)^{-1} = \text{diag}\{\mathcal{I}(\beta_0, \beta_1)^{-1}, \mathcal{I}(\beta, \omega, \lambda)^{-1}\}$.

Due to orthogonality, the parameters of the two models can be estimated separately. Estimation methods for when the likelihood function is orthogonal relative to the partition of interest have been presented in [46]. The approximated distribution $N_{p+l+r+5}(\theta, \Sigma_{\theta\theta})$ can be used for constructing confidence intervals for θ_j , presented as $\hat{\theta}_j \pm z_{1-\alpha/2}(\hat{\sigma}_{\hat{\theta}_j})^{1/2}$, where $\hat{\sigma}_{\hat{\theta}_j}$ is the j -th element in the diagonal of the matrix $\Sigma_{\hat{\theta}\hat{\theta}}$ corresponding to θ_j , and $z_{1-\alpha/2}$ is the $(1 - \alpha/2)100\%$ -th quantile of the standard normal distribution.

The zero and one inflated cases can be obtained from the SNZOI model. Taking $p_{1i} = 0$, we obtain a model with zero inflation while treating values above zero as continuous between 0 and 1. This case is referred to as left-censored because all observations at zero are treated as a single inflated point. For $p_{0i} = 0$, we obtain a model with one inflation while treating values below one as continuous between 0 and 1. This case is referred to as right-censored because all observations at one are treated as a single inflated point [25].

5. Empirical Applications

In this section, we illustrate the application of the SNZOI model and compare it with other models using real data. We demonstrate that the SNZOI model can be a valid alternative to the zero–one inflated beta regression and tobit models. Also, in this section, we indicate how the models can be selected using some statistical criteria.

5.1. Model Selection Criteria

To evaluate and compare the models, we use the Akaike information criterion (AIC) and the corrected AIC (AICc), which balance model fit and complexity [16,47]. These criteria are particularly helpful in our study as we compare multiple models with varying levels of complexity to identify the best-fitting model. The AIC is defined as

$$\text{AIC} = 2p - 2\ell, \quad (13)$$

where p is the number of parameters in the model and ℓ is the maximum value of the log-likelihood function for the model under analysis.

The AICc is a corrected version of AIC that adjusts for small sample sizes, given by

$$\text{AICc} = \text{AIC} + \frac{2p(p+1)}{n-p-1}, \quad (14)$$

where n is the sample size, providing a more accurate indicator of model quality [48]. Both AIC and AICc are obtained from the likelihood function of the observed data under the specified model. By using these criteria, we ensure a fair comparison across models, considering both their goodness of fit and their complexity.

5.2. Case Study 1: Doubly Censored Data

The data for this first application come from a clinical study previously analyzed in [49], which assessed the status and progression of periodontal disease among Gullah-speaking African-Americans with type-2 diabetes. The study used a questionnaire focused on demographical, dental, medical, and social aspects. Ethical approval was obtained, and all procedures followed the guidelines of the Helsinki Declaration, with informed consent provided by all participants. Data were collected between January 2018 and December 2019.

In this study, the clinical attachment level, a marker of periodontal disease, was measured at each of the six sites of a subject tooth. We model the dependence of the proportion of diseased sites corresponding to specific tooth types (canines, first molars, incisors, and premolars) with respect to the following covariates: gender (X_1), age (X_2), glycosylated hemoglobin (X_3), and smoking status (X_4). The dataset includes records from 290 individuals, with the response variable being the proportion of diseased tooth sites specifically for the premolars (Y). Notably, the data exhibit a high frequency of zeros ($Y = 0$) and, for some cases, ones ($Y = 1$), indicating high inflation at these points.

In Case Studies 1 and 2, we fitted several models to these data, including the new SNZOI model derived in the present article and the following models: Bernoulli one inflated (BOI), Bernoulli zero inflated (BZI), complementary log-log doubly censored log-normal (CLL-DCLN), complementary log-log doubly censored skew-normal (CLL-DCLSN), complementary log-log log-normal zero-one inflated (CLL-LNZOI), complementary log-log skew-normal zero-one inflated (CLL-SNZOI), DCLSN, logit doubly censored skew-normal (L-DCSN), logit doubly censored log-skew-normal (L-DCLSN), logit log-skew-normal zero-one inflated (L-LSNZOI), probit doubly censored log-normal (P-DCLN), probit log-normal zero-one inflated (P-LNZOI), probit log-skew-normal zero-one inflated (P-LSNZOI), and tobit.

From our analysis, we found that only covariates X_1 and X_2 are statistically significant at 5%. For the L-DCLSN model, the parameter estimates for the discrete part are reported in Table 1.

Table 2 presents the ML estimates of the model parameters for the tooth data using different models, including the detailed performance metrics AIC and AICc stated in (13) and (14) for a comprehensive comparison between the SNZOI model and other commonly employed models, such as the zero-one inflated beta and tobit models. The results demonstrate that the SNZOI model achieves competitive performance, often outperforming the other models in terms of lower AIC and AICc values.

In all cases, it can be noted that the two best models are CLL-LSNZOI and P-LSNZOI, followed by the BZI and P-LNZOI models. The CLL-LNZOI and L-LSNZOI models, as well as the doubly censored models (LN and LSN) follow, respectively. Additionally, we report that there are no significant differences statistically at 5% between the CLL-LSNZOI models with logit and probit link functions in the discrete part.

Table 1. ML estimates of the indicated parameter and model for the tooth data and their AIC and AICc.

Parameter	CLL-SNZOI	CLL-LSNZOI	CLL-DCLSN	P-LSNZOI	L-LSNZOI	L-DCLSN
β_{00}	−0.4405 (0.4174)	0.6337 (0.7408)	−3.9122 (0.8863)	0.6337 (0.7408)	0.6337 (0.7408)	0.7301 (0.8914)
β_{02}	0.0221 (0.0075)	−0.0376 (0.0135)	−0.0764 (0.0160)	−0.0376 (0.0135)	−0.0376 (0.0135)	−0.0957 (0.0161)
β_{10}	−2.2137 (1.2252)	−2.2137 (1.2252)	−2.1137 (0.5710)	−2.1123 (1.0552)	−1.2683 (1.2515)	−1.7604 (0.7648)
β_{11}	−1.5892 (0.7523)	−1.5892 (0.7523)	−0.4207 (0.1805)	−1.4316 (0.5520)	−1.1768 (0.4923)	−0.4990 (0.2381)
β_{12}	0.0509 (0.0260)	0.0509 (0.0260)	0.0274 (0.0101)	0.0533 (0.0230)	0.0517 (0.0218)	0.0291 (0.0131)
β_{20}	−2.8527 (1.9217)	−8.0316 (2.3145)	−8.0886 (1.4958)	−8.0316 (2.3145)	−8.0316 (2.3145)	−8.3703 (1.4941)
β_{22}	0.0440 (0.0274)	0.0788 (0.0358)	−0.0688 (0.0236)	0.0788 (0.0358)	0.0788 (0.0358)	0.0324 (0.0236)
ω	0.7671 (0.2731)	0.6453 (0.2279)	0.3023 (0.0545)	0.6161 (0.2362)	0.3177 (0.0477)	0.3198 (0.0667)
λ	−1.9835 (0.9456)	−1.9835 (0.9456)	−0.7232 (0.0523)	−1.7782 (0.7634)	−1.6450 (0.4274)	−0.9430 (0.1660)
AIC	310.13	309.94	325.34	308.84	313.53	326.40
AICc	312.91	312.73	328.12	311.63	316.32	329.47

Where CLL-SNZOI is complementary log–log skew-normal zero–one inflated, CLL-LSNZOI is complementary log–log log-skew-normal zero–one inflated, CLL-DCLSN is complementary log–log doubly censored skew-normal, P-LSNZOI is probit log-skew-normal zero–one inflated, L-LSNZOI is logit log-skew-normal zero–one inflated, L-DCLSN is logit doubly censored log-skew-normal, AIC is Akaike information criterion, and AICc is corrected Akaike information criterion.

We estimate the model using the MNP library, which employs a Bayesian method for this estimation. This method performs better than the ML method of the traditional case. The estimated models for the discrete part are $\hat{y}_{1i} = -2.8527 + 0.0440 \times \text{age}$ and $\hat{y}_{0-1i} = -0.4405 + 0.0221 \times \text{age}$. The CLL-LSNZOI and P-LSNZOI models can be compared against the CLL-LNZOI and P-LNZOI models using the hypotheses $H_0: \lambda = 0$ versus $H_1: \lambda \neq 0$, for each link function g , based on the test statistic $\Lambda_g = L_{\text{LN-}g}(\hat{\theta}) / L_{\text{LSN-}g}(\hat{\theta})$. After substituting the estimated values, we obtain $-2 \log(\Lambda_{\text{CLL}}) = 8.13$ and $-2 \log(\Lambda_{\text{P}}) = 6.96$, which are both greater than the critical value of 5% of the chi-square distribution, $\chi^2_{1,0.95} = 3.8414$. This leads to the conclusion that the CLL-LSNZOI and P-LSNZOI models fit the data better than the CLL-LNZOI and P-LNZOI models.

We estimate the percentage of the dataset with low response by fitting the considered model. An estimate of the percentage of observations at or below the detection limit is $100 \times 1 / (1 + \exp(1.407)) = 19.67\%$, which, compared to the observed 21.98%, indicates good agreement with the considered model.

To uncover atypical outcomes and/or model misspecifications, we examined the transformed martingale residuals as proposed in [3] and defined as

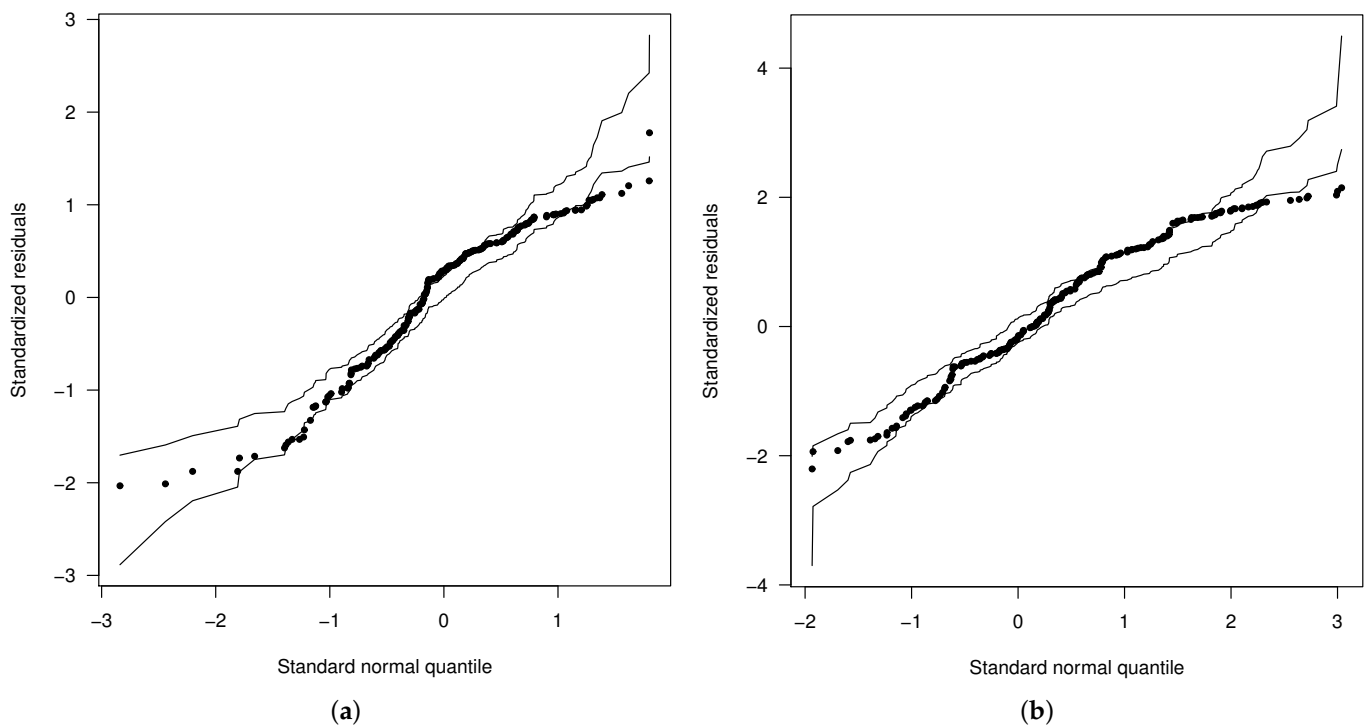
$$r_i = \text{sign}(r'_i) \sqrt{-2(r'_i + v_i \log(v_i - r'_i))}, \quad i \in \{1, \dots, n\},$$

where r'_i is the martingale residual proposed in [50], $v_i = 0$ or 1 indicates whether observation i is censored or not, respectively, and $\text{sign}(a)$ denotes the sign of a . The plots of r_i , for $i \in \{1, \dots, n\}$, with simulated envelopes, are presented in Figures 2 and 3. From these figures, it is evident that the P-LSNZOI model fits the data better than the other models. To illustrate the process followed for Case Study 1, we present Algorithm 1 and the flowchart in Figure 4.

Table 2. ML estimates of the indicated parameter and model for the tooth data and their AIC and AICc.

Parameter	BZI	P-LNZOI	CLL-DCLN	P-DCLN	CLL-LNZOI
β_{00}	0.6337 (0.7408)	0.6337 (0.7408)	0.7768 (0.7821)	1.1731 (2.1798)	0.6337 (0.7408)
β_{02}	−0.0376 (0.0135)	−0.0376 (0.0135)	−0.1791 (0.0139)	−0.2042 (0.0450)	−0.0376 (0.0135)
β_{10}	−1.3885 (0.3957)	−2.7994 (1.0573)	−3.2301 (0.5903)	−2.0330 (0.4353)	−2.8949 (1.1453)
β_{11}	−0.5366 (0.1613)	−1.1257 (0.4304)	−0.5926 (0.2087)	−0.4060 (0.4905)	−1.3134 (0.4387)
β_{12}	0.0217 (0.0068)	0.0452 (0.0186)	0.0380 (0.0106)	0.0263 (0.0091)	0.0393 (0.0194)
β_{20}	−8.0316 (2.3153)	−8.0316 (2.3145)	−8.4042 (2.8866)	−11.0378 (1.3108)	−8.0316 (2.3145)
β_{22}	0.0788 (0.0358)	0.0788 (0.0358)	−0.1049 (0.0449)	0.0611 (0.0210)	0.0788 (0.0358)
ω	0.0903 (0.0652)	0.3250 (0.0418)	0.2649 (0.1012)	0.2646 (0.0126)	0.3096 (0.0796)
AIC	311.70	313.80	324.33	323.51	316.07
AICc	314.58	316.45	326.97	326.15	318.71

Where BZI is Bernoulli zero inflated; P-LNZOI is probit log-normal zero-one inflated; CLL-DCLN is complementary log-log doubly censored log-normal; P-DCLN is probit doubly censored log-normal; CLL-LNZOI is complementary log-log log-normal zero-one inflated, AIC is Akaike information criterion, and AICc is corrected Akaike information criterion.

**Figure 2.** Plots of envelopes for r_i using (a) BZI and (b) L-DCLSN models and tooth data.

Algorithm 1 Case Study 1: Doubly censored data

- 1: **Begin**
- 2: Load the clinical study data.
- 3: Define covariates X_1 (gender), X_2 (age), X_3 (glycosylated hemoglobin), X_4 (smoking status), and response variable Y (proportion of diseased tooth sites for premolars).
- 4: Check for inflation at $Y = 0$ and $Y = 1$.
- 5: Fit the BZI, L-DCLSN, L-LSNZOI, and P-LSNZOI models.
- 6: Compare models using AIC and AICc to select the best model.
- 7: Determine statistical significance of the covariates at 5% in the best model.
- 8: Calculate and plot martingale residuals to evaluate model fit.
- 9: Identify the model to be used for predicting based on results and residual analysis.
- 10: **End**

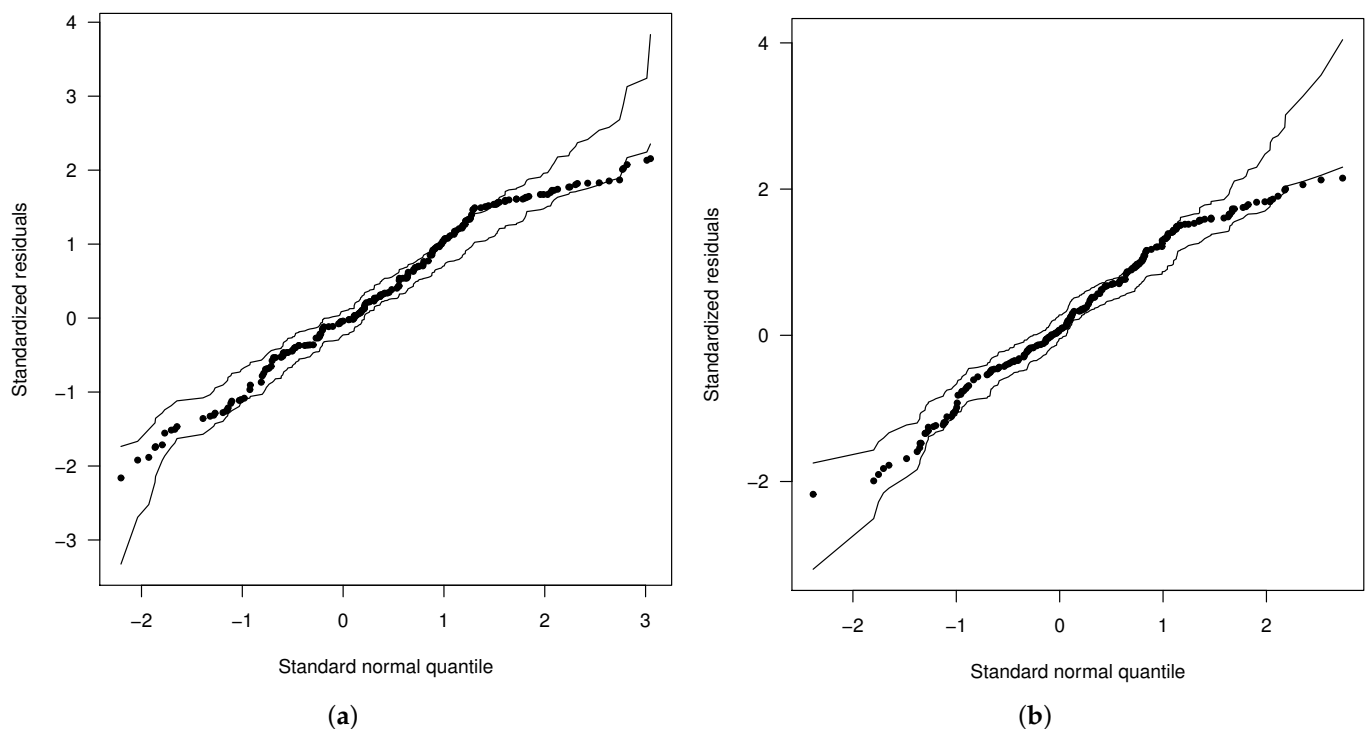


Figure 3. Plots of envelopes for r_i using (a) L-LSNZOI and (b) P-LSNZOI models and tooth data.

5.3. Case Study 2: One-Inflated Data

This case study focuses on analyzing clinical data to model the proportion of damaged incisor teeth among an individual cohort. The data are characterized by a high inflation at the value $Y = 1$, indicating full damage, which necessitates the use of specialized models to handle this inflation. We want to identify significant covariates that influence the proportion of damage on incisor teeth. To illustrate the process for Case Study 2, we present Algorithm 2 and the flowchart in Figure 4. This variable exhibits significant inflation at $Y = 1$, with 65 cases of complete damage and only one case of no damage ($Y = 0$). For this analysis, we exclude the observation where $Y = 0$, resulting in a dataset of 289 one-inflated observations with $Y = 1$.

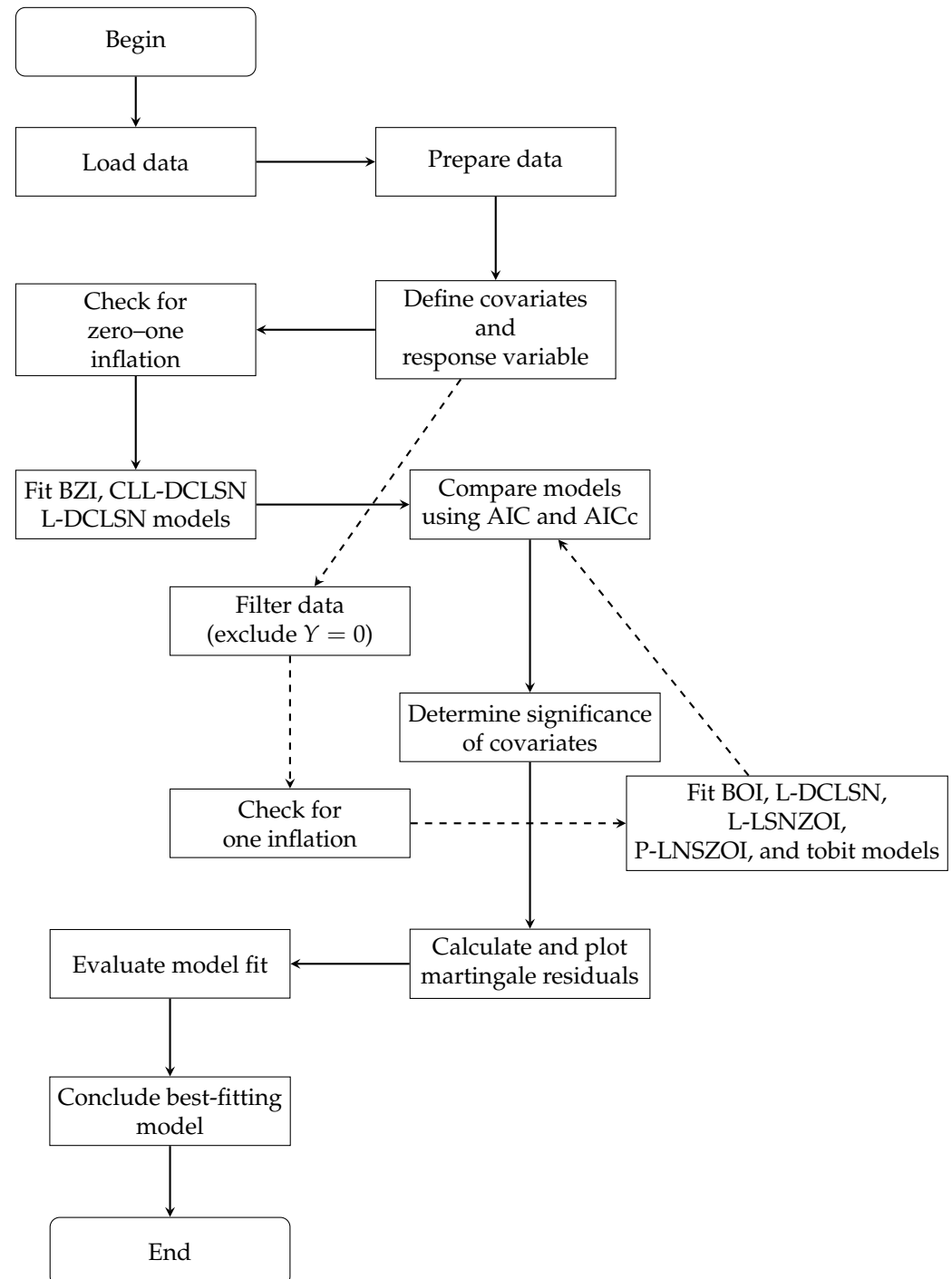


Figure 4. Overlaid flowchart for Case Studies 1 and 2.

To identify the most suitable model for these one-inflated data, we employed the L-LSNZOI and P-LSNZOI models, selecting the model with the lowest AIC and/or AICc from the combinations studied. We applied the same selection criteria to other models, including the normal tobit model (on a logarithmic scale), and the L-DCLSN and P-DCLSN models inflated at $Y = 1$. The results are reported in Table 3.

Algorithm 2 Case Study 2: One-inflated data

- 1: **Begin**
- 2: Load the clinical study data.
- 3: Define covariates X_1 (gender), X_2 (age), X_3 (glycosylated hemoglobin), X_4 (smoking status), and response variable Y (proportion of damage on incisor teeth).
- 4: Exclude observations where $Y = 0$.
- 5: Check for inflation at $Y = 1$.
- 6: Fit the BOI, L-DCLSN, L-LSNZOI, P-LSNZOI, and tobit models.
- 7: Compare models using AIC and AICc to select the best model.
- 8: Determine statistical significance of the covariates at 5% in the best model.
- 9: Calculate and plot martingale residuals to evaluate model fit.
- 10: Identify the model to be used for predicting based on results and residual analysis.
- 11: **End**

Table 3. ML estimates of the indicated parameter and model with tooth data and their AIC and AICc.

Parameter	BOI	L-LSNZOI	P-LSNZOI	L-DCLSN	Tobit
β_{10}	−0.0086 (0.3339)	−3.3573 (1.0938)	−2.8368 (1.5888)	−3.3573 (1.0938)	0.2968 (0.0897)
β_{11}	−0.2828 (0.1513)	0.1203 (0.0351)	0.1036 (0.0489)	0.1203 (0.0351)	−0.1003 (0.0393)
β_{12}	0.0182 (0.0061)				0.0104 (0.0015)
β_{20}	−6.1599 (1.0003)	−6.1599 (0.9992)	−5.7350 (1.0661)	3.6472 (0.5480)	
β_{22}	0.0853 (0.0166)	0.0853 (0.0165)	0.0756 (0.0174)	−0.0503 (0.0092)	
ω	1.3112 (0.0868)	0.3815 (0.0215)	0.3666 (0.0180)	0.3815 (0.0215)	−1.3046
λ		−7.7660 (3.1311)	−8.1582 (3.4725)	−7.7660 (3.1311)	0.0493
AIC	170.67	167.74	172.65	166.85	174.29
AICc	173.07	170.08	175.04	169.24	176.50

Where BOI is Bernoulli one inflated; L-LSNZOI is logit log-skew-normal zero-one inflated, P-LSNZOI is probit log-skew-normal zero-one inflated, L-DCLSN is logit doubly censored log-skew-normal, AIC is Akaike information criterion; and AICc is corrected Akaike information criterion.

As observed in Table 3, the L-LSNZOI and L-DCLSN models showed the lowest AIC and AICc values, indicating their superior fit to the data compared to the other models. Specifically, the L-LSNZOI model had the lowest AIC (167.74) and AICc (170.08), suggesting it as the best model. The tobit model had the highest AIC (174.29) and AICc (176.50), making it the least suitable among the models evaluated. The BOI and P-LSNZOI models also performed well but were outperformed by the L-LSNZOI model. These results show the importance of considering model selection criteria like AIC and AICc in determining the best model for one-inflated data. The significant parameters across these models also provide insights into the factors influencing the proportion of damaged incisor teeth, aiding in targeted clinical interventions.

5.4. Computational Costs of Algorithms

The most computationally intensive steps in both Algorithms 1 and 2 are the model fitting and residual analysis due to the iterative nature of parameter estimation and the subsequent evaluation of model adequacy. For example, the fitting process for the LSNZOI models involves estimating several parameters iteratively until the likelihood function converges, which can be time-consuming for large datasets. The residual analysis, especially when involving bootstrapping for simulated envelopes or generating multiple diagnostic plots, further adds to the computational load.

In our implementation, the following details highlight the computational efforts:

- **Model fitting**—The fitting process for models, such as L-LSNZOI and P-LSNZOI structures, involves maximizing the log-likelihood function through iterative numerical optimization methods. This process typically requires multiple iterations (often ranging from 50 to 200 iterations) to achieve convergence, especially in the presence of complex likelihood surfaces and multiple parameters. Each iteration involves recalculating the likelihood function and its gradient, adding to the computational time. The computational cost can be roughly estimated as $O(n \times p \times r)$, where n is the sample size, p is the number of parameters, and r is the number of iterations.
- **Residual analysis**—Calculating martingale residuals and their diagnostic plots involve extensive computations. Generating simulated envelopes through bootstrapping, for instance, necessitates multiple re-samplings of the data and re-fitting the model to each bootstrap sample, which is computationally expensive. If B bootstrap samples are used, the computational cost for this step is roughly $O(B \times n \times p \times r)$.
- **Model selection criteria**—Calculating AIC and AICc requires obtaining the log-likelihood value at the estimated parameters, which involves evaluating the likelihood function at these estimates. While this step is less computationally intensive than the fitting process, it still adds to the overall computational burden. The computational cost for this step is approximately $O(n \times p)$.

Overall, the total computational cost is a combination of these intensive steps, making the entire process demanding, especially for large datasets or complex models.

6. Discussion

In this article, we proposed a general class of skew regression models for response variables taking values in the unit interval, possibly with an excess of zeros or ones. The proposed models were developed from a continuous-discrete mixture distribution with covariates in both discrete and continuous parts. As demonstrated by applications with real data, the proposed models offer a valid alternative for describing data of rates and proportions that are inflated at zero or one.

6.1. Key Findings and Insights

Our results indicate that the SNZOI model, particularly with the logit and probit link functions (L-SNZOI and P-SNZOI), consistently outperformed other models in terms of AIC and AICc values. These models provided a better fit to the data from the clinical study on periodontal disease, where the response variable was the proportion of diseased tooth sites.

In the first application involving doubly censored data, the models effectively captured the significant effects of gender and age on the proportion of diseased premolar sites. The L-SNZOI and P-SNZOI models outperformed the other models, suggesting that the SN distribution is a more suitable assumption for this type of data. In the second application, which involved inflation at the one value, the SNZOI models again outperformed traditional beta regression and tobit models. The L-DCLSN, L-LSNZOI, and P-LSNZOI models, adjusted for right-censored data and inflation at one, demonstrated superior performance.

To further validate the model adequacy, we analyzed martingale residuals. The transformation of the martingale residuals and the generation of simulated envelopes, as proposed in [3], provided a robust method to identify model goodness of fit and outliers.

Specifically, we considered transformed martingale residuals, which help to detect discrepancies between the observed and expected values under the model. The simulated envelopes are generated to visually assess the distribution of these residuals and identify any points that fall outside the expected range, indicating potential model misfits or outliers. This comprehensive analysis of martingale residuals, along with the simulated envelopes, reinforces the robustness of our chosen models and provides a thorough evaluation of their performance, ensuring that the models accurately capture the underlying data structure.

6.2. Comparison with Previous Studies

Our findings contribute to the literature on regression models for bounded and inflated data. The superior performance of the SN models aligns with previous studies, such as [25,46], which demonstrated the advantages of using SN distributions for modeling asymmetrically distributed data with bounds. The practical implications of our study are important for clinical research where proportion data with inflation at specific values is common. The ability of the SNZOI models to handle such data accurately allows researchers to obtain more precise estimators and make better-informed decisions.

6.3. Model Limitations

Despite the promising results, there are several limitations to our study. Firstly, the complexity of the models and the need for iterative numerical methods for parameter estimation can be computationally intensive. Secondly, while the models performed well on the datasets used in this study, further validation on other types of data is necessary to confirm their generalizability.

6.4. Directions for Future Research

Future research could focus on several areas. One potential direction is the development of more efficient algorithms to reduce the computational burden of fitting these models. Additionally, exploring the application of these models to other fields, such as economics or environmental studies, could provide further validation and potentially uncover new areas of application.

While the present article focuses on the case study of periodontal disease prevalence to demonstrate the practical implications of the SNZOI model, future research could include additional case studies from diverse fields. These additional studies would provide more comprehensive evidence of the model versatility and practical utility across different domains.

As model performance is a critical aspect of our analysis, while the methods used, such as AIC, AICc, and martingale residuals, are robust for evaluating model adequacy, there is potential for further enhancement. Future work could explore additional goodness-of-fit tests that are specifically tailored to the characteristics of bounded and inflated data. Incorporating these tests could provide a more comprehensive assessment of model performance and robustness.

Another interesting direction for future research is the extension of these models to handle longitudinal data or hierarchical data structures. This would involve developing methods to account for correlations within subjects or groups, which is a common feature in many practical datasets. Additionally, investigating the robustness of these models under various misspecification scenarios can help in developing more resilient modeling strategies. Understanding how these models perform when the underlying assumptions are violated is crucial for practical applications.

7. Conclusions

The analysis of proportion data, especially when values are inflated at zero and one, poses challenges in various scientific fields. Traditional models, such as the beta and tobit regression models, often fall short in accurately capturing the complexities introduced by such data. This highlights the necessity for more advanced modeling techniques that can handle the unique distributional characteristics of zero–one inflation.

In this study, we addressed these challenges by proposing the skew-normal zero–one inflated models, particularly focusing on logit and probit link functions. These models integrate a continuous-discrete mixture distribution with covariates in both components, providing a sophisticated framework for analyzing proportion data with specific inflation points. By doing so, the skew-normal zero–one inflated models offer a robust and flexible approach for capturing asymmetrically distributed data and mixed discrete-continuous characteristics, which are common in fields such as ecology, economics, and medicine.

Our applications, including a detailed case study on periodontal disease prevalence, demonstrated that the skew-normal zero–one inflated models, particularly with the logit and probit link functions, consistently outperformed traditional models. These models showed superior performance in terms of the Akaike information criterion and its correction, offering more precise and unbiased statistical inferences. The transformation of martingale residuals and the generation of simulated envelopes further validated the robustness of our models, highlighting their ability to identify model misfits and outliers effectively. The proposed models fill a critical void in the statistical modeling and provide valuable insights and precise estimators for dealing with bounded and inflated data. The flexibility and robustness of the skew-normal zero–one inflated models make them a valid alternative for describing data of proportions that are inflated at zero or one.

Despite the promising results, several limitations remain. The complexity of the models and the need for numerical methods for parameter estimation can be computationally intensive. Additionally, while the models performed well on the datasets used in this study, further validation on other types of data is necessary to confirm their generalizability.

Future research should focus on developing more efficient algorithms to reduce the computational burden of fitting these models. Exploring the application of these models to other fields, such as economics or environmental studies, could provide further validation and potentially uncover new areas of application. Additionally, incorporating more comprehensive goodness-of-fit tests tailored to bounded and inflated data could enhance the evaluation of model performance and robustness.

In summary, the skew-normal zero–one inflated models represent an advancement in statistical modeling for proportion data with zero–one inflation. These models offer a robust and flexible framework for analyzing such data, providing deeper insights and more precise estimators. Continued research and development in this area holds promise for further advancements in statistical methodologies for bounded and inflated data.

Author Contributions: Conceptualization: G.M.-F., R.T.-F., V.L. and C.C.; data curation: G.M.-F., R.T.-F. and C.C.; formal analysis: G.M.-F., R.T.-F., V.L. and C.C.; investigation: G.M.-F., R.T.-F., V.L. and C.C.; methodology: G.M.-F., R.T.-F., V.L. and C.C.; writing—original draft: G.M.-F. and R.T.-F.; writing—review and editing: V.L. and C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by Vice-rectorate for Research of the Universidad de Córdoba, Colombia, project grant FCB-06-22 (G.M.-F. and R.T.-F.); by Vice-rectorate for Research, Creation, and Innovation (VINCI) of the Pontificia Universidad Católica de Valparaíso (PUCV), Chile, grants VINCI 039.470/2024 (regular research), VINCI 039.493/2024 (interdisciplinary associative research), VINCI 039.309/2024 (PUCV centenary), and FONDECYT 1200525 (V.L.), from the National Agency for Research and Development (ANID) of the Chilean government under the Ministry of Science, Technology, Knowledge, and Innovation; and by Portuguese funds through the CMAT—Research Centre of Mathematics of University of Minho, Portugal, within projects UIDB/00013/2020 (<https://doi.org/10.54499/UIDB/00013/2020>) and UIDP/00013/2020 (<https://doi.org/10.54499/UIDP/00013/2020>) (C.C.).

Data Availability Statement: The data and codes used in this study are available under request to the authors.

Acknowledgments: The authors would like to thank the editors and anonymous reviewers for their valuable comments and suggestions, which helped us to improve the quality of this article.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

References

1. Schaminée, J.H.; Hennekens, S.M.; Chytrý, M.; Rodwell, J.S. Vegetation-plot data and databases in Europe: An overview. *Preslia* **2009**, *81*, 173–185.
2. Tobin, J. Estimation of relationships for limited dependent variables. *Econometrica* **1958**, *26*, 24–36. [[CrossRef](#)]

3. Barros, M.; Galea, M.; González, M.; Leiva, V. Influence diagnostics in the tobit censored response model. *Stat. Methods Appl.* **2010**, *19*, 379–397. [\[CrossRef\]](#)
4. Ferreira, P.H.; Shimizu, T.K.; Suzuki, A.K.; Louzada, F. On an asymmetric extension of the tobit model based on the tilted-normal distribution. *Chil. J. Stat.* **2019**, *10*, 99–122.
5. Barros, M.; Galea, M.; Leiva, V.; Santos-Neto, M. Generalized tobit models: Diagnostics and application in econometrics. *J. Appl. Stat.* **2018**, *45*, 145–167. [\[CrossRef\]](#)
6. Arellano-Valle, R.B.; Gómez, H.W.; Quintana, F. Statistical inference for a general class of asymmetric distributions. *J. Stat. Plan. Inference* **2005**, *128*, 427–443. [\[CrossRef\]](#)
7. Gallardo, D.I.; Bourguignon, M.; Galarza, C.E.; Gómez, H.W. A parametric quantile regression model for asymmetric response variables on the real line. *Symmetry* **2022**, *14*, 1938. [\[CrossRef\]](#)
8. Gupta, R.D.; Gupta, R.C. Analyzing skewed data by power normal model. *Test* **2008**, *17*, 197–210. [\[CrossRef\]](#)
9. Pewsey, A.; Gómez, H.W.; Bolfarine, H. Developments in skew-symmetric distributions and their applications. *Symmetry* **2022**, *14*, 567.
10. Desousa, M.; Saulo, H.; Leiva, V.; Scalco, P. On a tobit-Birnbaum–Saunders model with an application to medical data. *J. Appl. Stat.* **2018**, *45*, 932–955. [\[CrossRef\]](#)
11. Sanchez, L.; Leiva, V.; Galea, M.; Saulo, H. Birnbaum–Saunders quantile regression and its diagnostics with application to economic data. *Appl. Stoch. Model. Bus. Ind.* **2021**, *37*, 53–73. [\[CrossRef\]](#)
12. Azzalini, A. A class of distributions which includes the normal ones. *Scand. J. Stat.* **1985**, *12*, 171–178.
13. Azzalini, A. Further results on a class of distributions which includes the normal ones. *Statistica* **1986**, *46*, 199–208.
14. Henze, N. A probabilistic representation of the skew-normal distribution. *Scand. J. Stat.* **1986**, *13*, 271–275.
15. Castillo, N.O.; Gómez, H.W.; Leiva, V.; Sanhueza, A. On the Fernández–Steel distribution: Inference and application. *Comput. Stat. Data Anal.* **2011**, *55*, 2951–2961. [\[CrossRef\]](#)
16. Ventura, M.; Saulo, H.; Leiva, V.; Monsueto, S. Log-symmetric regression models: Information criteria, application to movie business and industry data with economic implications. *Appl. Stoch. Model. Bus. Ind.* **2019**, *3*, 963–977. [\[CrossRef\]](#)
17. Massuia, M.B.; Garay, A.M.; Cabral, C.R.; Lachos, V.H. Bayesian analysis of censored linear regression models with scale mixtures of skew-normal distributions. *Stat. Its Interface* **2017**, *10*, 425–439. [\[CrossRef\]](#)
18. Morán-Vásquez, R.A.; Giraldo-Melo, A.D.; Mazo-Lopera, M.A. Quantile estimation using the log-skew-normal linear regression model with application to children’s weight data. *Mathematics* **2023**, *11*, 3736. [\[CrossRef\]](#)
19. Dias-Domingues, T.; Mouriño, H.; Sepúlveda, N. Classification methods for the serological status based on mixtures of skew-normal and skew-t distributions. *Mathematics* **2024**, *12*, 217. [\[CrossRef\]](#)
20. Mudholkar, G.S.; Hutson, A.D. The epsilon-skew-normal distribution for analyzing near-normal data. *J. Stat. Plan. Inference* **2000**, *83*, 291–309. [\[CrossRef\]](#)
21. Gómez, H.W.; Venegas, O.; Bolfarine, H. Skew-symmetric distributions generated by the distribution function of the normal distribution. *Environmetrics* **2007**, *18*, 395–407. [\[CrossRef\]](#)
22. Arrué, J.; Arellano-Valle, R.; Gómez, H.W.; Leiva, V. On a new type of Birnbaum–Saunders models and its inference and application to fatigue data. *J. Appl. Stat.* **2020**, *47*, 2690–2710. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Pewsey, A. Problems of inference for Azzalini’s skew-normal distribution. *J. Appl. Stat.* **2000**, *27*, 859–870. [\[CrossRef\]](#)
24. Ferrari, S.; Cribari-Neto, F. Beta regression for modelling rates and proportions. *J. Appl. Stat.* **2004**, *31*, 799–815. [\[CrossRef\]](#)
25. Ospina, R.; Ferrari, S. Inflated beta distributions. *Stat. Pap.* **2010**, *51*, 111–126. [\[CrossRef\]](#)
26. Ospina, R.; Ferrari, S. A general class of zero-or-one inflated beta regression models. *Comput. Stat. Data Anal.* **2012**, *56*, 1609–1623. [\[CrossRef\]](#)
27. Couri, L.; Ospina, R.; da Silva, G.; Leiva, V.; Figueroa-Zúñiga, J. A study on computational algorithms in the estimation of parameters for a class of beta regression models. *Mathematics* **2022**, *10*, 299. [\[CrossRef\]](#)
28. Mohammadi, Z.; Sajjadnia, Z.; Bakouch, H.S.; Sharafi, M. Zero-and-one inflated Poisson–Lindley INAR (1) process for modelling count time series with extra zeros and ones. *J. Stat. Comput. Simul.* **2022**, *92*, 2018–2040. [\[CrossRef\]](#)
29. Lee, B.S.; Haran, M. A class of models for large zero-inflated spatial data. *J. Agric. Biol. Environ. Stat.* **2024**. [\[CrossRef\]](#)
30. Figueroa-Zúñiga, J.; Niklitschek, S.; Leiva, V.; Liu, S. Modeling heavy-tailed bounded data by the trapezoidal beta distribution with applications. *REVSTAT Stat. J.* **2022**, *20*, 387–404.
31. Jornsatan, C.; Bodhisuwan, W. Zero-one inflated negative binomial-beta exponential distribution for count data with many zeros and ones. *Commun. Stat. Theory Methods* **2022**, *51*, 8517–8531. [\[CrossRef\]](#)
32. Keim, J.L.; DeWitt, P.D.; Fitzpatrick, J.J.; Jenni, N.S. Estimating plant abundance using inflated beta distributions: Applied learnings from a Lichen–Caribou ecosystem. *Ecol. Evol.* **2017**, *7*, 486–493. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Benites, L.; Maehara, R.; Lachos, V.H.; Bolfarine, H. Linear regression models using finite mixtures of skew heavy-tailed distributions. *Chil. J. Stat.* **2019**, *10*, 21–41.
34. Desousa, M.; Saulo, H.; Santos-Neto, M.; Leiva, V. On a new mixture-based regression model: Simulation and application to data with high censoring. *J. Stat. Comput. Simul.* **2020**, *90*, 2861–2877. [\[CrossRef\]](#)
35. Arellano-Valle, R.B.; Gómez, H.W.; Quintana, F. A new class of skew-normal distributions. *Commun. Stat. Theory Methods* **2004**, *33*, 1465–1480. [\[CrossRef\]](#)

36. Saulo, H.; Dasilva, A.; Leiva, V.; Sanchez, L.; de la Fuente, H. Log-symmetric quantile regression models. *Stat. Neerl.* **2022**, *76*, 124–163. [[CrossRef](#)]
37. Chai, H.S.; Bailey, K.R. Use of log-skew-normal distribution in analysis of continuous data with a discrete component at zero. *Stat. Med.* **2008**, *27*, 3643–3655. [[CrossRef](#)] [[PubMed](#)]
38. Cragg, J.G. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica* **1971**, *39*, 829–844. [[CrossRef](#)]
39. Moulton, L.H.; Halsey, N.A. A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics* **1995**, *51*, 1570–1578. [[CrossRef](#)]
40. McCulloch, R.; Rossi, P.E. An exact likelihood analysis of the multinomial probit model. *J. Econom.* **1994**, *64*, 207–240. [[CrossRef](#)]
41. Keane, M.P. A note on identification in the multinomial probit model. *J. Bus. Econ. Stat.* **1992**, *10*, 193–200. [[CrossRef](#)]
42. Heckman, J.; Sedlacek, G. Heterogeneity, aggregation, and market wage functions: An empirical model of self-selection in the labor market. *J. Political Econ.* **1985**, *93*, 1077–1125. [[CrossRef](#)]
43. Imai, K.; Van-Dyk, D. A Bayesian analysis of the multinomial probit model using marginal data augmentation. *J. Econom.* **2005**, *124*, 311–334. [[CrossRef](#)]
44. Nocedal, J.; Wright, S. *Numerical Optimization*; Springer: New York, NY, USA, 2006.
45. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022.
46. Farias, R.; Moreno-Arenas, G.; Patriota, A. Reduction of models in the presence of nuisance parameters. *Colomb. J. Stat.* **2009**, *32*, 99–121.
47. Burnham, K.P.; Anderson, D.R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*; Springer: New York, NY, USA, 2004.
48. Hurvich, C.M.; Tsai, C.L. Regression and time series model selection in small samples. *Biometrika* **1989**, *76*, 297–307. [[CrossRef](#)]
49. Galvis, D.M.; Bandyopadhyay, D.; Lachos, V.H. Augmented mixed beta regression models for periodontal proportion data. *Stat. Med.* **2014**, *33*, 3759–3771. [[CrossRef](#)]
50. Ortega, E.M.; Bolfarine, H.; Paula, G.A. Influence diagnostics in generalized log-gamma regression models. *Comput. Stat. Data Anal.* **2003**, *42*, 165–186. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.