mathematics

MDPI

Article

Mathematical Formalization and Applications to Data with Excess of Zeros and Ones of the Unit-Proportional Hazard Inflated Models

Guillermo Martínez-Flórez [1], Roger Tovar-Falón [1,*] and Héctor W. Gómez [2]

[1] Departamento de Matemáticas y Estadística, Universidad de Córdoba, Montería 230002, Colombia; guillermomartinez@correo.unicordoba.edu.co
[2] Departamento de Estadística y Ciencia de Datos, Facultad de Ciencias Básicas, Universidad de Antofagasta, Antofagasta 1240000, Chile; hector.gomez@uantof.cl
* Correspondence: rjtovar@correo.unicordoba.edu.co

Abstract: In this study, we model the rate or proportion of a specific phenomenon using a set of known covariates. To fit the regression model, which explains the phenomenon within the intervals $(0,1)$, $[0,1)$, $(0,1]$, or $[0,1]$, we employ a logit link function. This approach ensures that the model's predictions remain within the appropriate range of zero to one. In cases of inflation at zero, one, or both, the logit link function is similarly applied to model the dichotomous Bernoulli-type variable with a multinomial response. The findings demonstrate that the model yields a non-singular information matrix, ensuring valid statistical inference. This ensures the invertibility of the information matrix, allowing for hypothesis testing based on likelihood statistics regarding the parameters in the model. This is not possible with other asymmetric models, such as those derived from the skew-normal distribution, which has a singular information matrix at the boundary of the skewness parameter. Finally, empirical results show the model's effectiveness in analyzing proportion data with inflation at zero and one, proving its robustness and practicality for analyzing bounded data in various fields of research.

Keywords: unit proportional hazard distribution; censoring; proportion data; truncation; zero-one inflation

MSC: 62J05; 62E15

Citation: Martínez-Flórez, G.; Tovar-Falón, R.; Gómez, H.W. Mathematical Formalization and Applications to Data with Excess of Zeros and Ones of the Unit-Proportional Hazard Inflated Models. *Mathematics* **2024**, *12*, 3566. https://doi.org/10.3390/math12223566

Academic Editors: Domma Filippo, Francesca Condino and Manuel Alberto M. Ferreira

Received: 4 September 2024
Revised: 23 October 2024
Accepted: 13 November 2024
Published: 15 November 2024

1. Introduction

In recent years, probability distributions have seen significant advancements, particularly through the creation of new families derived from extensions or generalizations of classical distributions. These innovations aim to overcome the limitations of traditional models and provide greater flexibility to better fit the complex phenomena observed in various fields of knowledge. Examples of these distributions include those based on transformations such as the generalized beta distribution by Eugene et al. [1], the family of generalized distributions based on the Kumaraswamy distribution, referred to as Kw-distributions and introduced by Cordeiro and De Castro [2] (Kw-normal, Kw-Weibull, Kw-gamma, Kw-Gumbel, and Kw-inverse Gaussian distribution); and the beta modified Weibull distribution of Silva et al. [3]. These new distributions not only better capture data characteristics like skewness and kurtosis but also improve accuracy in modeling extreme events or phenomena with heavy tails. Furthermore, their implementation has proven useful in fields such as biomedicine, economics, and engineering, where classical models fail to adequately describe the reality of the data.

In parallel, truncated distributions have emerged as another essential tool, particularly when the data are bounded within a specific range. These distributions are modifications of classical ones, where values outside a certain interval are truncated, improving the

model's fit for data restricted by natural or experimental constraints [4]. For example, the truncated normal distribution is widely used in reliability analysis and survival studies where negative values are not possible [5,6]. Similarly, the truncated Weibull distribution has been applied in actuarial sciences to model the time to event data [7], offering greater flexibility when standard distributions fail to capture the behavior of the tail.

A method for creating new families of distributions involves using a generating distribution as a base. This method has been widely employed by various authors, including Cordeiro et al. [8,9], Zografos and Balakrishnan [10], Ristić and Balakrishnan [11], Castellares et al. [12], and Cordeiro et al. [13]. In the same context, Mahdavi and Silva [4] introduced a method for generating families of truncated distributions, producing a two-parameter extension of the base distribution. This method has been used to derive distributions such as the truncated exponential-exponential and the truncated Lomax-Exponential. These innovations in probability distributions have proven to be valuable tools in statistical analysis, providing more robust and adaptable models for complex data.

The method introduced by Mahdavi and Silva [4] can be summarized as follows:

- Definition of the Truncated Distribution: A random variable U with support in the interval (a, b), where $a \leq 0$ and $b \geq 1$, and cumulative distribution function (CDF) F is considered. The CDF of the truncated random variable U in the interval $(0, 1)$ is defined as:

$$F_{U_t}(u) = \frac{F(u) - F(0)}{F(1) - F(0)}. \tag{1}$$

- Generation of the New Family of Distributions: Using the truncated CDF, the new truncated F–G family of distributions is introduced. For each absolutely continuous G distribution (denoted as the baseline distribution), the TF–G distribution is associated. The CDF of the TF–G class of distributions is defined as:

$$G_X(x) = \frac{F(G(x)) - F(0)}{F(1) - F(0)}, \tag{2}$$

  where G is the CDF of the random variable V used to generate a new distribution.

The probability density function (PDF), $f_X(x)$, survival function, and hazard rate function are given, respectively, by:

$$f_X(x) = \frac{g(x)f(G(x))}{F(1) - F(0)}, \tag{3}$$

$$S_X(x) = \frac{F(1) - F(G(x))}{F(1) - F(0)}. \tag{4}$$

and

$$h_X(x) = \frac{g(x)f(G(x))}{F(1) - F(G(x))}, \tag{5}$$

where f and g are the PDF of the random variables U and V, respectively. The extension to the location-scale case of the model (3) is obtained from the transformation $Y = \mu + \sigma X$, where $X \sim$ TF-G, for $\mu \in \mathbb{R}$ y $\sigma \in \mathbb{R}^+$; it has PDF given by:

$$f_Y(y) = \frac{1}{\sigma} \frac{g(x)f(G(x))}{F(x_1) - F(x_0)}, \tag{6}$$

where

$$x = \frac{y - \mu}{\sigma}, \quad x_0 = \frac{a - \mu}{\sigma}, \quad x_1 = \frac{b - \mu}{\sigma}.$$

Some distributions that have been derived using the generator proposed by [4] are the truncated exponential-exponential (TEE), the truncated Lomax-Exponential by Enami [14], the truncated exponential Marshall Olkin Lomax distribution of Hadi and Al-Noor [15]

and the truncated Nadarajah-Haghighi Exponential by Al-Habib et al. [16]. The generator proposed by [4] can also be used to derive distributions useful for modeling data in the interval $(0, 1)$, such as proportions, rates, or indices.

The analysis of phenomena represented by proportion data, confined to values between zero and one, is essential across various scientific disciplines. These data elucidate part-to-whole relationships and are prevalent in numerous applications, including the prevalence of diseases, the distribution of resources in economics, the survival rates of species, and the utilization of habitats in ecology [17]. Modeling such data can be highly challenging when there is high zero-to-one inflation in proportion data. Traditional statistical models, such as the censored normal or censored log-normal models, may not be the best solution, as they often struggle to accurately characterize the underlying distribution of proportion data with inflated extremes.

Numerous authors have collaborated to develop more robust models than the censored normal and censored log-normal models for this type of data. By incorporating distributions such as the Birnbaum–Saunders [18,19], Student-t [20,21], skew-normal (SN) [22–25], and power-normal (PN) [26,27] distributions, among others, they offer a framework for analyzing data with high degrees of skewness and kurtosis compared with traditional models.

Perhaps the beta distribution is the most well-known in the statistical literature and is commonly used for fitting unit interval data. However, it has limitations when modeling unit data with zero-one inflation. Recent proposals, such as the zero-one inflated beta models, have been made to overcome this limitation and have proven to be viable alternatives for handling data with certain degrees of asymmetry [28–33]. Despite advancements in modeling data with inflation and asymmetry, there remains a gap in adequately addressing zero-one inflation in proportion data. Existing models fail to fully capture the unique distributional characteristics and complexities introduced by these inflations, leading to biased estimators and imprecise inferences [34,35].

The primary aim of this study is to introduce and develop unit-proportional hazard zero-one inflated (UPHZOI) models, a novel class of regression models specifically designed to address the challenges posed by zero-one inflation in proportional data confined to the unit interval. UPHZOI models combine a continuous-discrete mixture distribution with covariates, enabling them to effectively capture the complex dynamics of such data.

The remainder of this article is structured as follows: Section 2 provides background on the asymmetric proportional hazard model and introduces the truncated proportional hazard model. It also presents the process of parameter estimation, considering a classical approach using the maximum likelihood method. In Section 3, we introduce new regression models for unit interval data with inflation, including the model formulation, parameter estimation, and elements of the Hessian matrix. Section 4 demonstrates the application of these models through empirical case studies on doubly censored data and zero-inflated data. Section 5 presents an analysis of the major results, limitations, and future research directions. The article concludes with Section 6.

2. An Asymmetric Distribution for Skew Data

This section provides background on the proportional hazard (PH) distribution introduced by Martínez-Flórez et al. [36] for modeling data with high or low kurtosis and a wide range of skewness. Additionally, the unit-proportional hazard distribution is introduced, derived using the truncated method of [4]. The latter serves as the foundation for formulating the UPHZOI models, from which regression models for proportion data are developed.

2.1. Proportional Hazard Distribution and Its Modeling

The PDF of the PH distribution is given by

$$\phi_{\mathrm{PH}}(y;\boldsymbol{\theta}) = \alpha f\left(\frac{y-\xi}{\sigma}\right)\left\{1 - F\left(\frac{y-\xi}{\sigma}\right)\right\}^{\alpha-1}, \quad y \in \mathbb{R}, \tag{7}$$

where $\boldsymbol{\theta} = (\xi, \sigma, \alpha)$, with $\xi \in \mathbb{R}$ is a location parameter, $\sigma \in \mathbb{R}^+$ is a scale parameter, α is a positive real number and, F is an absolutely continuous distribution function with continuous density function $f = dF$. The notation $Y \sim \text{PH}(\xi, \sigma, \alpha)$ indicates that Y follows an PH distribution with parameters ξ, σ, and α.

Under the PH model, the hazard function is presented as

$$h_{\text{PH}}(y, \alpha) = \alpha h_f(y),$$

where $h_f(\cdot) = f(\cdot)/(1 - F(\cdot))$ is the hazard function regarding the density f. When the CDF F in the (7) model corresponds to the CDF of the standard normal distribution, that is, $F = \Phi$ and therefore $f = \phi$, we obtain the model denominated proportional hazard normal (PHN), whose PDF is given by

$$\phi_{\text{PHN}}(y; \boldsymbol{\theta}) = \alpha\phi\left(\frac{y-\xi}{\sigma}\right)\left\{S\left(\frac{y-\xi}{\sigma}\right)\right\}^{\alpha-1}, \; y \in \mathbb{R}, \tag{8}$$

where $S(\cdot)$ is the survival function of the standard normal PDF. This model also serves as an alternative for fitting data with much wider ranges of skewness and kurtosis than those of the normal distribution, which the latter cannot adequately capture. The CDF of the $\text{PHN}(\mu, \sigma, \alpha)$ is given by:

$$\Phi_{\text{PHN}}(y; \boldsymbol{\theta}) = 1 - \left\{S\left(\frac{y-\xi}{\sigma}\right)\right\}^{\alpha}, \; y \in \mathbb{R}. \tag{9}$$

By considering various values of α, Martínez-Flórez et al. [36] found that the range of the asymmetry and kurtosis coefficients, $\sqrt{\beta_1}$ and β_2, for the variable $Y \sim \text{PHN}(0, 1, \alpha)$ are the intervals $(-1.1578, 0.9918)$ and $(1.1513, 4.3023)$, respectively. This indicates that the PHN model is superior to both the SN and PN models in terms of asymmetry and kurtosis. Furthermore, ref. [36] demonstrate that the information matrix of the PHN distribution is non-singular. This is advantageous for statistical inference, as it allows for hypothesis testing based on likelihood ratio statistics.

2.2. Truncated Proportional Hazard Normal Distribution

Based on the TF-G distribution, we define the truncated proportional hazard normal (TPHN) distribution in the unit interval $[0, 1]$. Let $F(\cdot)$ be the CDF of the PHN distribution and $G(\cdot)$ the CDF of a continuous uniform distribution on $[0, 1]$; then, we have that the PDF of the TPHN model is

$$\phi_{\text{TPHN}}(y; \xi, \sigma, \alpha) = \frac{\frac{\alpha}{\sigma}\phi_{\text{PHN}}\left(\frac{y-\xi}{\sigma}\right)}{\left\{S\left(\frac{-\xi}{\sigma}\right)\right\}^{\alpha} - \left\{S\left(\frac{1-\xi}{\sigma}\right)\right\}^{\alpha}}, \quad 0 < y < 1, \tag{10}$$

where ϕ_{PHN} and S are defined in (8). The standardization terms, which facilitate the normalization of the data within the specified limits, are defined as

$$z = \frac{y-\xi}{\sigma}, \quad z_0 = -\frac{\xi}{\sigma}, \quad z_1 = \frac{1-\xi}{\sigma}.$$

This is denoted by $\text{TPHN}(\xi, \sigma, \alpha)$. It can be seen from (10) that the CDF, survival function, and hazard function for the TPHN distribution are given by:

$$\Phi_{\text{TPHN}}(y; \xi, \sigma, \alpha) = \frac{\{S(z_0)\}^{\alpha} - \{S(z)\}^{\alpha}}{\{S(z_0)\}^{\alpha} - \{S(z_1)\}^{\alpha}}, \tag{11}$$

$$S_{\text{TPHN}}(y; \xi, \sigma, \alpha) = \frac{\{S(z)\}^{\alpha} - \{S(z_1)\}^{\alpha}}{\{S(z_0)\}^{\alpha} - \{S(z_1)\}^{\alpha}}, \tag{12}$$

and

$$h_{\text{TPHN}}(y;\xi,\sigma,\alpha) = \frac{\alpha}{\sigma}\frac{\phi(z)\{S(z)\}^{\alpha-1}}{\{S(z)\}^{\alpha}-\{S(z_1)\}^{\alpha}} = \alpha\frac{\{S(z)\}^{\alpha}}{\{S(z)\}^{\alpha}-\{S(z_1)\}^{\alpha}}h(y), \tag{13}$$

respectively, where $h(y)$ is the hazard function of the normal distribution.

The moments of a random variable with TPHN distribution can be obtained using the expression

$$\mathbb{E}(Y^r) = \frac{\alpha\sum_{j=1}^{r}\xi^{r-j}\sigma^j\lambda^j}{\{S(z_0)\}^{\alpha}-\{S(z_1)\}^{\alpha}}, \quad r=1,2,\ldots \tag{14}$$

where

$$\lambda = \int_{S(z_1)}^{S(z_0)}\Phi^{-1}(1-u)u^{\alpha-1}du$$

being $\Phi^{-1}(\cdot)$ the inverse of the function $\Phi(\cdot)$.

2.3. Parameter Estimation in the TPHN Model

The TPHN parameters can be estimated using the maximum likelihood (ML) method by maximizing the log-likelihood function. We consider a random sample of n observations, $Y_1, Y_2, \ldots, Y_n$ from the TPHN(ξ,σ,α) distribution; the log-likelihood function of $\boldsymbol{\theta} = (\xi,\sigma,\alpha)^\top$ is obtained by taking the natural logarithm of the joint likelihood function defined as $L(\boldsymbol{\theta},\boldsymbol{y}) = \prod_{i=1}^{n}\phi_{TPHN}(y_i;\boldsymbol{\theta})$, where now $\boldsymbol{\theta} = (\xi,\sigma,\alpha)$. Taking the natural logarithm in the above expression, we obtain the log-likelihood function established as

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= n\log(\alpha) - n\log(\sigma) + \sum_{i=1}^{n}\log(\phi(z_i)) \\ &\quad +(\alpha-1)\sum_{i=1}^{n}\log(S(z_i)) - n\log(W(\xi,\sigma,\alpha)),\end{aligned} \tag{15}$$

where $z_i = \frac{y_i-\xi}{\sigma}$ and $W = W(\xi,\sigma,\alpha) = \log\left(\{S(z_0)\}^{\alpha}-\{S(z_1)\}^{\alpha}\right)$. By taking the first derivatives of the function presented in (15) with respect to the parameters, $\dot{\ell}(\boldsymbol{\theta}) = \partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, we obtain the score elements. For the location parameter ξ, the score function is formulated as

$$\dot{\ell}(\alpha) = \frac{n}{\alpha} + \sum_{i=1}^{n}\log(S(z_i)) - n\frac{\{S(z_0)\}^{\alpha}\log(S(z_0)) - \{S(z_1)\}^{\alpha}\log(S(z_1))}{W}. \tag{16}$$

For the scale parameter σ, the score function is defined as

$$\dot{\ell}(\mu) = \frac{1}{\sigma}\sum_{i=1}^{n}z_i + \frac{\alpha-1}{\sigma}\sum_{i=1}^{n}\frac{\phi(z_i)}{S(z_i)} - n\frac{\alpha}{\sigma}\frac{h(z_0)\{S(z_0)\}^{\alpha} - h(z_1)\{S(z_1)\}^{\alpha}}{W}. \tag{17}$$

For the shape parameter α, the score is formulated as

$$\dot{\ell}(\sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma}\sum_{i=1}^{n}z_i^2 + \frac{\alpha-1}{\sigma}\sum_{i=1}^{n}z_i\frac{\phi(z_i)}{S(z_i)} - n\frac{\alpha}{\sigma}\frac{z_0h(z_0)\{S(z_0)\}^{\alpha} - z_1h(z_1)\{S(z_1)\}^{\alpha}}{W}. \tag{18}$$

The maximum likelihood estimate (MLE) of the parameters is obtained by solving the system of equations formed by setting (16)–(18) equal to zero. This system is generally solved using iterative numerical methods, such as the Newton–Raphson or quasi-Newton algorithms, which iteratively refine the parameter estimates to maximize the likelihood function.

2.4. Information Matrix in TPHN Model

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. The second derivatives of the log-likelihood function for $\xi\xi, \xi\sigma, \sigma\sigma, \xi\alpha, \sigma\alpha$ an $\alpha\alpha$ are given in the Appendix A.1. To derive the information matrix, it suffices to find the expected value of the

elements of the observed information matrix. According to [36], the family of proportional hazard distributions is regular; thus, the information matrix of the PHN model is non-singular, as demonstrated in Martínez-Flórez et al. [36]. Consequently, the information matrix of the truncated distribution on $[0,1]$ is non-singular, and its covariance matrix is given by

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\xi, \sigma, \alpha) = I^{-1}(\xi, \sigma, \alpha) = (\mathbb{E}(J(\xi, \sigma, \alpha)))^{-1}.$$

It follows that, for large n, $\widehat{\boldsymbol{\theta}}$ is consistent and, furthermore, by the central limit theorem, $\widehat{\boldsymbol{\theta}}$ is asymptotically normally distributed with mean vector $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Sigma}$, i.e.,

$$\widehat{\boldsymbol{\theta}} \xrightarrow{D} N_3(\boldsymbol{\theta}, \boldsymbol{\Sigma}),$$

Details of this result can be found in [37].

In practice, since the matrix $J(\boldsymbol{\theta})$ is consistent for $I(\boldsymbol{\theta})$, we can take $\boldsymbol{\Sigma} = J^{-1}(\boldsymbol{\theta})$ as the covariance matrix of the estimator vector for the TPHN model.

2.5. Unit-Proportional Hazard Regression Model

We now introduce the unit-proportional hazard normal (UPHN) regression model to fit proportion data from the TPHN distribution by changing the location parameter ξ in (10) to the linear predictor $\xi_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, where $\boldsymbol{x}_i = (1, x_{1i}, \ldots, x_{pi})^\top$ is an observed covariate vector for the observation i, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is the regression coefficient vector. The response (dependent) variable Y_i can be modeled by

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{19}$$

where $\varepsilon_i \sim \text{TPHN}(0, \sigma, \alpha)$. It follows from the natural form that

$$Y_i \sim \text{TPHN}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma, \alpha), \quad i = 1, 2, \ldots, n.$$

Since our focus is on cases where the variable of interest lies within the unit interval $(0,1)$, issues may arise with the expected response or predicted value, which could fall outside this standard unit interval $(0,1)$, potentially resulting in negative estimates that lack interpretation and/or meaning. To avoid these issues, we change the assumption that the response variable Y is a linear function of the vector of explanatory variables $\boldsymbol{x}_i^\top = (x_1, x_2, \ldots, x_p)$ to a nonlinear transformation of this set of variables. This model will be obtained by assuming that the location parameter of y_i can be written as

$$g(\mu_i) = \xi_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad i = 1, \ldots, n, \tag{20}$$

where $g(\cdot)$ is a strictly monotonic and twice differentiable link function that maps $(0,1)$ to $\mathbb{R}$. There are several options for choosing the link function $g(\cdot)$; two commonly used for this particular case are the logit function $g(\mu_i) = \log(\mu_i/(1-\mu_i))$, and the probit function $g(\mu_i) = \Phi(\mu_i)$. These two options yield very similar results in predicted values, with some exceptions for extreme values. Because the logit and probit functions provide very similar results in terms of model fit, and unlike the probit function, the logit link function allows for simpler algebraic manipulations and obtaining expressions for the score function, elements of the information matrix and expectation calculations among others, we opt for the logit function. Thus, in this case, we write

$$\mu_i = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}, \quad i = 1, 2, \ldots, n. \tag{21}$$

For this model, the parameters are interpreted based on the odds ratio between the odds of the prediction or mean when one of the variables is increased by m units (while keeping the other explanatory variables constant) and the odds without this increase. It has been demonstrated that this odds ratio is given by $\exp(m\beta_k)$, where β_k is the parameter as-

sociated with the explanatory variable increased by m units. It follows that the distribution of the variable under study is

$$y_i \sim \text{TPHN}(\mu_i, \sigma, \alpha), \quad i = 1, 2, \ldots, n.$$

The estimates of the parameters of the UPHN regression model with a logit link function can be obtained using the ML method. The log-likelihood function for the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma, \alpha)$ given a sample of n observations is given by

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= n\log(\alpha) - n\log(\sigma) + \sum_{i=1}^{n}\log(\phi(z_i)) \\ &\quad + (\alpha-1)\sum_{i=1}^{n}\log(S(z_i)) - \sum_{i=1}^{n}\log(W_i(\mu_i, \sigma, \alpha)), \end{aligned} \tag{22}$$

where $W_i = W_i(\mu_i, \sigma, \alpha) = \log\left(\{S(z_{0i})\}^{\alpha} - \{S(z_{1i})\}^{\alpha}\right)$ with

$$z_i = \frac{y_i - \mu_i}{\sigma}, \quad z_{0i} = -\frac{\mu_i}{\sigma}, \quad z_{1i} = \frac{1-\mu_i}{\sigma}.$$

Thus, the score function, defined as the derivative of the log-likelihood function with respect to each of the parameters, is given for the vector whose components are given by:

$$\begin{aligned} \dot{\ell}(\alpha) &= \frac{n}{\alpha} + \sum_{i=1}^{n}\log(S(z_i)) - \sum_{i=1}^{n}\frac{\{S(z_{0i})\}^{\alpha}\log(S(z_{0i})) - \{S(z_{1i})\}^{\alpha}\log(S(z_{1i}))}{W_i}, \\ \dot{\ell}(\beta_j) &= \frac{1}{\sigma}\sum_{i=1}^{n}x_{ij}z_i\mu_i(1-\mu_i) + \frac{\alpha-1}{\sigma}\sum_{i=1}^{n}\frac{x_{ij}\mu_i(1-\mu_i)\phi(z_i)}{S(z_i)} \\ &\quad -\frac{\alpha}{\sigma}\sum_{i=1}^{n}\frac{x_{ij}\mu_i(1-\mu_i)\left(h(z_{0i})\{S(z_{0i})\}^{\alpha} - h(z_{1i})\{S(z_{1i})\}^{\alpha}\right)}{W_i}, \\ \dot{\ell}(\sigma) &= -\frac{n}{\sigma} + \frac{1}{\sigma}\sum_{i=1}^{n}z_i^2 + \frac{\alpha-1}{\sigma}\sum_{i=1}^{n}z_i\frac{\phi(z_i)}{S(z_i)} - \frac{\alpha}{\sigma}\sum_{i=1}^{n}\frac{z_{0i}h(z_{0i})\{S(z_{0i})\}^{\alpha} - z_{1i}h(z_{1i})\{S(z_{1i})\}^{\alpha}}{W_i}. \end{aligned}$$

Setting these expressions to zero, we get the corresponding score equations whose numerical solution leads to the MLE. The elements of the information matrix are obtained using the chain rule and are presented in Appendix A.2.

It can be seen that, for large sample sizes, we have

$$\hat{\boldsymbol{\theta}} \xrightarrow{D} N_{p+3}(\boldsymbol{\theta}, I_F(\boldsymbol{\theta})^{-1}).$$

where, "D" indicates convergence in distribution. In this way, inferences can be made about the parameters using likelihood ratio statistics.

2.6. MCMC Methods for the PHN Model

Bayesian methods can also be implemented to perform statistical inference within the PHN distribution family. Although there is limited statistical literature addressing this issue in power-normal distributions, Sarabia and Castillo [38] provides some initial ideas on how to approach it. In this section, we do not aim to propose specific Bayesian methods but rather open the door to exploring these methods within the PHN model class.

We consider the standard case of the $\text{PHN}(0, 1, \alpha) \equiv \text{PHN}(\alpha)$ model, and, similar to [38], we assume a gamma distribution for the shape parameter α. The model we consider is

$$Y \mid \alpha \sim \text{PHN}(\alpha) \tag{23}$$

$$\alpha \sim Gamma(\delta_0, \lambda_0), \tag{24}$$

where $Gamma(\delta_0, \lambda_0)$ denotes a gamma random variable with PDF proportional to $s^{\delta_0-1}e^{-\lambda_0 s}$ with δ_0 and λ_0 known. If we denote by $m(y)$ the marginal distribution of Y and by $\pi(\alpha \mid Y)$ the posterior distribution of the shape parameter α, we have that:

$$\begin{aligned} m(y) &= \int_0^\infty \alpha\phi(y)[1-\Phi(y)]^{\alpha-1}\frac{\lambda_0^{\delta_0}}{\Gamma(\delta_0)}\alpha^{\delta_0-1}e^{-\lambda_0\alpha}d\alpha \\ &= \frac{\lambda_0^{\delta_0}}{\Gamma(\delta_0)}\frac{\phi(y)}{1-\Phi(y)}\frac{\Gamma(\delta_0+1)}{\{\lambda_0-\log[1-\Phi(y)]\}^{\delta_0-1}}, \end{aligned} \tag{25}$$

from which it follows that:

$$\pi(\alpha \mid Y) = \frac{\{\lambda_0-\log[1-\Phi(y)]\}^{\delta_0+1}}{\Gamma(\delta_0+1)}\alpha^{\delta_0}e^{-(\lambda_0-\log[1-\Phi(y)])\alpha}, \tag{26}$$

which is the PDF of a random variable Gamma(δ_1, λ_1), where δ_1 and λ_1 are given by

$$\delta_1 = \delta_0 + 1, \quad \lambda_1 = \lambda_0 - \log[1-\Phi(y)]$$

Inference about the parameter α is carried out based on the posterior distribution given in (26). For the location-scale case, PHN(ξ, σ, α), prior distributions for the parameters ξ and σ that can be considered are the normal and inverse-gamma distributions, respectively.

3. UPHN Zero-One Inflated Regression Model

In this section, we present some regression models for unit interval (proportion) data that account for inflation at values zero and one or any value between zero and one.

3.1. Models for Censored Data

Cragg proposed a two-part model [39], which is a framework for fitting the mixture of a discrete and a continuous random variable. This model is represented by:

$$g(y_i) = p_i I_i + (1-p_i)f(y_i)(1-I_i),$$

where p_i is the probability that determines the relative contribution of the point mass distribution made by the discrete variable, $f(\cdot)$ is a PDF, and I_i is an indicator variable that takes values of 0 or 1. This model is optimal in cases where the model is inflated at the point mass value (for example, $y_i = a$), whose probability at $y = a$ cannot be explained by the CDF associated with the PDF $f(\cdot)$. Cragg's model can be extended to the case of a variable with double censoring or two-point mass values, for example, 0 and 1, in which case it is given by:

$$g(y_i) = p_{0i}I_{0i} + (1-p_{0i}-p_{1i})f(y_i)(1-I_{0i}-I_{1i}) + p_{1i}I_{1i},$$

where $p_{0i} = \Pr(y_i = 0)$, $p_{1i} = \Pr(y_i = 1)$, I_{0i} is the indicator variable that takes the value 1 if $y_i = 0$ and zero otherwise. Similarly, I_{1i} is the indicator variable for $y_i = 1$. In this model, the three components are determined by different stochastic processes, thus necessarily leading to a positive response from f. On the other hand, a zero or a one comes from the distribution of a point mass.

3.2. Zero-One Inflated PHN Distribution

Based on Cragg's model, we proposed the zero-one inflated PHN model as a means of

$$g(y) = \begin{cases} \rho_0, & \text{if } y = 0, \\ \frac{\alpha}{\sigma}(1-\rho_0-\rho_1)\phi(z)\{S(z)\}^{\alpha-1}, & \text{if } 0 < y < 1, \\ \rho_1, & \text{if } y = 1, \end{cases}$$

where

$$z = \frac{y-\mu}{\sigma}, \quad \rho_0 = \Pr(y=0), \quad \rho_1 = \Pr(y=1).$$

From this model, cases of inflation only at zero follow by taking $\rho_1 = 0$ or inflation only at one by taking $\rho_0 = 0$.

The CDF is represented by:

$$G(y) = \begin{cases} \rho_0, & \text{if } y \leq 0, \\ \rho_0 + (1-\rho_0-\rho_1)\left[\{S(z_0)\}^{\alpha} - \{S(z)\}^{\alpha}\right], & \text{if } 0 < y < 1, \\ 1, & \text{if } y \geq 1. \end{cases}$$

The most interesting case in this new model is when covariates are used to explain the response both in the censored part (0 and 1) and in the uncensored part (the continuous part in $(0,1)$). Thus, for the discrete part, it is assumed that the responses at zero and one can be explained by the covariate vectors $\boldsymbol{x}_{(0)i} = (1, x_{0i1}, \ldots, x_{0iq})^\top$ and $\boldsymbol{x}_{(1)i} = (1, x_{1i1}, \ldots, x_{1ir})^\top$ respectively. Then, to determine the probabilities ρ_0 and ρ_1, a logistic model with a polytomous response can be constructed such that:

$$\rho_{0i} = \Pr(y_i = 0) = \frac{\exp\left(\boldsymbol{x}_{(0)i}^\top \boldsymbol{\beta}_{(0)}\right)}{1 + \exp\left(\boldsymbol{x}_{(0)i}^\top \boldsymbol{\beta}_{(0)}\right) + \exp\left(\boldsymbol{x}_{(1)i}^\top \boldsymbol{\beta}_{(1)}\right)}, \tag{27}$$

$$\rho_{1i} = \Pr(y_i = 1) = \frac{\exp\left(\boldsymbol{x}_{(1)i}^\top \boldsymbol{\beta}_{(1)}\right)}{1 + \exp\left(\boldsymbol{x}_{(0)i}^\top \boldsymbol{\beta}_{(0)}\right) + \exp\left(\boldsymbol{x}_{(1)i}^\top \boldsymbol{\beta}_{(1)}\right)}, \tag{28}$$

$$\rho_{01i} = 1 - \rho_{0i} - \rho_{1i} = \Pr(y_i \in (0,1)) = \frac{1}{1 + \exp\left(\boldsymbol{x}_{(0)i}^\top \boldsymbol{\beta}_{(0)}\right) + \exp\left(\boldsymbol{x}_{(1)i}^\top \boldsymbol{\beta}_{(1)}\right)}, \tag{29}$$

where $\boldsymbol{\beta}_{(0)} = (\beta_{00}, \beta_{01}, \ldots, \beta_{0q})^\top$ y $\boldsymbol{\beta}_{(1)} = (\beta_{10}, \beta_{11}, \ldots, \beta_{1r})^\top$ are vectors of unknown parameters associated respectively with the covariate vectors $\boldsymbol{x}_{(0)}$ and $\boldsymbol{x}_{(1)}$.

Similarly, for the continuous component of the model, a unit model $\text{PHN}(\mu_i, \sigma, \alpha)$ is still assumed with a logit link function in the mean response, i.e., $\log(\mu_i/(1-\mu_i)) = x_i^\top \boldsymbol{\beta}$, where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is a vector of covariates with associated coefficient vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^\top$. For this model, it is easy to verify that the log-likelihood function for the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}_{(0)}^\top, \boldsymbol{\beta}_{(1)}^\top, \boldsymbol{\beta}^\top, \sigma, \alpha)^\top$ given $\boldsymbol{X}_{(0)}$, $\boldsymbol{X}$, $\boldsymbol{X}_{(1)}$ and $\boldsymbol{Y}$ can be written in the form:

$$\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\beta}_{(0)}, \boldsymbol{\beta}_{(1)}) + \ell(\boldsymbol{\beta}, \sigma, \alpha),$$

where

$$\ell(\boldsymbol{\beta}_{(0)}, \boldsymbol{\beta}_{(1)}) = \sum_{0} \boldsymbol{x}_{(0)i}\boldsymbol{\beta}_{(0)} + \sum_{1} \boldsymbol{x}_{(1)i}\boldsymbol{\beta}_{(1)} - \sum_{i=1}^{n} \log\left[1 + \exp\left(\boldsymbol{x}_{(0)i}^\top \boldsymbol{\beta}_{(0)}\right) + \exp\left(\boldsymbol{x}_{(1)i}^\top \boldsymbol{\beta}_{(1)}\right)\right].$$

and

$$\ell(\boldsymbol{\beta}, \sigma, \alpha) = \sum_{y_i \in (0,1)} (\log(\alpha) - \log(\sigma) + \log(\phi(z_i)) + (\alpha - 1)\log(S(z_i))).$$

Given these characteristics, the MLEs of the model parameters can be obtained separately for each component of the log-likelihood function. The score function is derived by differentiating each component of the log-likelihood function. It can be shown that the Fisher information matrix can be written as a block diagonal matrix in the form:

$$I(\boldsymbol{\theta}) = \text{Diag}\Big\{ I(\boldsymbol{\beta}_{(0)}, \boldsymbol{\beta}_{(1)}), I(\boldsymbol{\beta}, \sigma, \alpha), \Big\}$$

where $I(\boldsymbol{\beta}_{(0)}, \boldsymbol{\beta}_{(1)})$ corresponds to the information matrix of the discrete part. The elements of the observed information matrix for the discrete part are given in the Appendix A.3.

The respective Fisher information matrix is obtained by calculating the expectation of the elements of the observed information matrix. Furthermore, since the inverse of a block diagonal matrix is the block diagonal matrix of the respective inverses, it follows that the variance-covariance matrix is given by:

$$\boldsymbol{\Sigma} = \text{Diag}\{I^{-1}_{\left(\boldsymbol{\beta}_{(0)},\boldsymbol{\beta}_{(1)}\right)}, I^{-1}_{(\boldsymbol{\beta},\sigma,\alpha)}\}.$$

Here, for large sample sizes it follows that for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\beta}_{(0)}, \boldsymbol{\beta}_{(1)}, \sigma, \alpha)^\top$

$$\hat{\boldsymbol{\theta}} \xrightarrow{D} N_{p+q+r+3}(\boldsymbol{\theta}, I_F(\boldsymbol{\theta})^{-1}).$$

Confidence intervals for θ_r with of confidence coefficient $\omega = 100(1-\psi)\%$ can be obtained as $\hat{\theta}_r \mp z_{1-\omega/2}\sqrt{\hat{\sigma}(\hat{\theta}_r)}$. By talking $\rho_{1i} = 0$, the zero-inflated model is followed and, making $\rho_{0i} = 0$, the zero-inflated model is obtained.

3.3. The Zero-One Inflated UPHN Model

Similarly to how the zero-one inflated PHN model was constructed, a zero and/or one-inflated UPHN distribution can be proposed, which is given by:

$$f(y_i) = \begin{cases} \rho_0, & \text{if } y = 0, \\ \frac{\alpha}{\sigma}(1-\rho_0-\rho_1)\dfrac{\phi(z)\{S(z)\}^{\alpha-1}}{\{S(z_0)\}^{\alpha} - \{S(z_1)\}^{\alpha}}, & \text{if } 0 < y < 1, \\ \rho_1, & \text{if } y = 1. \end{cases}$$

where z, $\rho_0 = \Pr(y = 0)$ and $\rho_1 = \Pr(y = 1)$ are defined as in the zero-one inflated PHN model.

The CDF of this distribution is represented by

$$F(y) = \begin{cases} \rho_0, & \text{if } y \leq 0, \\ \rho_0 + (1-\rho_0-\rho_1)\dfrac{\{S(z_0)\}^{\alpha} - \{S(z)\}^{\alpha}}{\{S(z_0)\}^{\alpha} - \{S(z_1)\}^{\alpha}}, & \text{if } 0 < y < 1, \\ 1, & \text{if } y \geq 1. \end{cases}$$

For the case of covariates in the model, $\boldsymbol{x}_{(0)i} = (1, x_{0i1}, \ldots, x_{0iq})^\top$ and $\boldsymbol{x}_{(1)i} = (1, x_{1i1}, \ldots, x_{1ir})^\top$ for the zero- and one-inflated part, with associated coefficient vector $\boldsymbol{\beta}_{(0)} = (\beta_{00}, \beta_{01}, \ldots, \beta_{0q})^\top$ and $\boldsymbol{\beta}_{(1)} = (\beta_{10}, \beta_{11}, \ldots, \beta_{1r})^\top$. For the continuous component of the model, we connect the response variable with the linear predictor using the logit link function. As before, we choose this link function because, in addition to ensuring that the predictions model is within the $(0, 1)$ interval, the logit function allows for more explicit expressions of the score function elements and the information matrix compared to the probit function, which depends on the integral of the cumulative distribution function of the standard normal distribution. In this way, we assume relationship $\log(\mu_i/(1-\mu_i)) = x_i^\top \boldsymbol{\beta}$, where $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ is a vector of covariates with vector of coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^\top$.

The proposal again is to use a polytomous logistic model to explain the probabilities ρ_{0i} and ρ_{1i}. As in the case of the inflated PHN model, we have that the log-likelihood function is given by

$$\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\beta}_{(0)}, \boldsymbol{\beta}_{(1)}) + \ell(\boldsymbol{\beta}, \sigma, \alpha),$$

where $\ell(\boldsymbol{\beta}_{(0)}, \boldsymbol{\beta}_{(1)})$ is the same as the inflated PHN model, while

$$\begin{aligned}\ell(\theta;\mathbf{y}) &= n_{01}\log(\alpha) - n_{01}\log(\sigma) + \sum_{y_i\in(0,1)}\log(\phi(z_i)) + (\alpha-1)\sum_{y_i\in(0,1)}\log(S(z_i)) \\ &\quad - \sum_{y_i\in(0,1)}\log(W_i(\mu_i,\sigma,\alpha)),\end{aligned}$$

where z_i, $W_i = W_i(\mu_i,\sigma,\alpha)$, z_{0i} and z_{1i} are as defined in (22).

The score function is obtained by differentiating each component of the log-likelihood function and the Fisher information matrix can be written as a diagonal block matrix in the form:

$$I(\boldsymbol{\theta}) = \text{Diag}\Big\{I(\boldsymbol{\beta}_{(0)},\boldsymbol{\beta}_{(1)}), I(\boldsymbol{\beta},\sigma,\alpha)\Big\}.$$

The elements of the matrix $I(\boldsymbol{\beta}_{(0)},\boldsymbol{\beta}_{(1)})$ are like those given in the inflated PHN model, while the elements of the matrix $I(\boldsymbol{\beta},\sigma,\alpha)$ are like those given in the information matrix of the UPHN regression model.

3.4. Generalized Two-Part PHN Model

Cragg's two-part model [39] encounters the issue that some censored points may be values at the boundary of the censoring limit. This is particularly problematic for a distribution $f(\cdot)$ within the unit interval $[0,1]$, where a zero or one could either be a realization from the point mass distribution or a partial observation of $f(\cdot)$ having a critical value that is not precisely known but is close to $(0,T_1)$ or $(T_2,1)$ for small values of the pre-specified constants T_1 and T_2. In practice, the values T_1 and T_2 are, in some cases, defined as those for which the instruments cannot record measurements below or above, respectively, and, consequently, are treated as censoring values. In other cases, these observational limits are defined for ethical or practical reasons. For example, in clinical studies, it may be unethical to continue observing a patient under certain conditions, or the costs of prolonged observation may become prohibitive.

To address this issue in the two-part model, Moulton and Halsey [40] propose a new approach to adjust the mixture of continuous and discrete random variables. This approach allows for the possibility that some limiting responses result from an interval censoring of $f(\cdot)$. The model proposed by Moulton and Halsey (1995) for left censoring at point a is given by: $g(y_i) = [p_i + (1-p_i)F(T)]I_i + (1-p_i)f(y_i)(1-I_i)$, where F is the CDF associated to f and, T It is a pre-established constant within the interval (a,T) where some limiting responses are considered censored. Similarly to how we generalized Cragg's model, Moulton and Halsey's model can also be generalized for left and right censoring or two boundary inflation points within the definition interval of the pdf $f(\cdot)$. In our case, for the unit PHN distribution within the interval $[0,1]$, this generalization of Moulton and Halsey's model is given by:

$$\begin{aligned}g(y_i) &= \left(p_{0i} + (1-p_{0i}-p_{1i})(1-\{S(z_{0i})\}^{\alpha})\right)I_{0i} + \frac{\alpha(1-p_{0i}-p_{1i})}{\sigma}\phi(z_i)\{S(z_i)\}^{\alpha-1}I_{(0,1)i} \\ &\quad + \left(p_{1i} + (1-p_{0i}-p_{1i})\{S(z_{1i})\}^{\alpha}\right)I_{1i}.\end{aligned}$$

It can be observed that this distribution is a model with double censoring (at zero and one) and, therefore, allows for the fit of datasets with inflation at zero and one. This represents an alternative to the double-censored Tobit model, where the CDF of the normal distribution does not efficiently fit the probability of the point mass where double censoring occurs, i.e., the probability of the inflation points.

Extending this model to the case of covariates in each part of the model, we again assume that $\boldsymbol{x}_{(0)i} = (1, x_{0i1}, \ldots, x_{0iq})^{\top}$ and $\boldsymbol{x}_{(1)i} = (1, x_{1i1}, \ldots, x_{1ir})^{\top}$ are sets of auxiliary covariates for the discrete part at zero and one, respectively; and a set of covariates $x_i = (1, x_{i1}, \ldots, x_{ip})'$ for the continuous part in the interval $(0,1)$. Then, denoting by ρ_0 the proportion of observations below zero, $y_i = 0$ (lower detection limit), and by ρ_1 the proportion of observations above one, $y_i = 1$ (upper detection limit), the extension of the Moulton and Halsey model to the double-censored PHN case can be expressed through the PDF given by

$$g(y_i) = \begin{cases} \rho_{0i} + (1-\rho_{0i}-\rho_{1i})(1-\{S(z_{0i})\}^{\alpha}), & \text{if } y_i \leq 0, \\ \frac{\alpha}{\sigma}(1-\rho_{0i}-\rho_{1i})\phi(z_i)\{S(z_i)\}^{\alpha-1}, & \text{if } 0 < y_i < 1, \\ \rho_{1i} + (1-\rho_{0i}-\rho_{1i})\{S(z_{1i})\}^{\alpha}, & \text{if } y_i \geq 1, \end{cases}$$

where ρ_{0i} and ρ_{1i} are the probability masses at points zero and one, while z_{0i}, z_{1i}, z_i are as defined above; $\log(\mu_i/(1-\mu_i)) = \boldsymbol{x}_i^{\top}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the set of coefficients associated with the covariate vector $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^{\top}$.

The CDF of this model is represented by

$$G(y_i) = \begin{cases} \rho_{0i} + (1-\rho_{0i}-\rho_{1i})(1-\{S(z_{0i})\}^{\alpha}), & \text{if } y_i \leq 0, \\ \rho_{0i} + (1-\rho_{0i}-\rho_{1i})\left[1-\{S(z_i)\}^{\alpha}\right], & \text{if } 0 < y_i < 1, \\ 1, & \text{if } y_i \geq 1. \end{cases}$$

To model the responses at the point masses $y_i = 0$ and $y_i = 1$, a multinomial logistic model with a logit link function is used again, where $\boldsymbol{\beta}_{(0)}^{\top}, \boldsymbol{\beta}_{(1)}^{\top}$ are the vectors of coefficients associated with the sets of covariates $\boldsymbol{x}_{(0)i} = (1, x_{0i1}, \ldots, x_{0iq})^{\top}$ and $\boldsymbol{x}_{(1)i} = (1, x_{1i1}, \ldots, x_{1ir})^{\top}$.

The log-likelihood function for parameter vector estimation $\boldsymbol{\theta} = (\boldsymbol{\beta}_{(0)}^{\top}, \boldsymbol{\beta}_{(1)}^{\top}, \boldsymbol{\beta}^{\top}, \sigma, \alpha)^{\top}$ conditionally on $\boldsymbol{X}_{(0)}$, $\boldsymbol{X}$, $\boldsymbol{X}_{(1)}$, is given by:

$$\begin{aligned} \ell(\boldsymbol{\theta}) \;=\; & \sum_{0} \log\left[\exp(\boldsymbol{x}_{(0)i}^{\top}\boldsymbol{\beta}_{(0)}) + 1 - \{S(z_{0i})\}^{\alpha}\right] + \sum_{1} \log\left[\exp(\boldsymbol{x}_{(1)i}^{\top}\boldsymbol{\beta}_{(1)}) + \{S(z_{1i})\}^{\alpha}\right] \\ & + \sum_{i\in(0,1)} \left(\log(\alpha) - \log(\sigma) + \log(\phi(z_i)) + (\alpha-1)\log(S(z_i))\right) \\ & - \sum_{i=1}^{n} \log\left[1 + \exp\left(\boldsymbol{x}_{(0)i}^{\top}\boldsymbol{\beta}_{(0)}\right) + \exp\left(\boldsymbol{x}_{(1)i}^{\top}\boldsymbol{\beta}_{(1)}\right)\right]. \end{aligned} \tag{30}$$

The score equations are obtained by performing the first derivatives with respect to the model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}_{(0)}^{\top}, \boldsymbol{\beta}_{(1)}^{\top}, \boldsymbol{\beta}^{\top}, \sigma, \alpha)^{\top}$ while the information matrix is obtained by proceeding as in the models studied previously. Models with inflation only at zero or only at one can be studied by taking $\rho_0 = 0$ or $\rho_1 = 0$, respectively.

4. Empirical Applications

In this section, we illustrate the application of the proposed models and compare it with other models using real data. We show that the proposed model can be a valid alternative to some existing regression models in the statistical literature.

4.1. Application 1: Case Study on Students' Dropout Data

Student dropout is a major problem many Latin American countries face. In some universities in Colombia, this phenomenon can lead to more than 50% of students who enroll in a university program abandoning their higher education studies. This phenomenon has its greatest impact in the first four semesters of undergraduate studies, which is why it is important to determine the main causes leading to this abandonment of higher education.

This application refers to student dropout in the Faculty of Veterinary Medicine and Zootechnics (MVZ, by its acronym in Spanish) at the University of Córdoba, Colombia. The analyzed information corresponds to a sample of students who dropped out during one of the first four semesters (early dropout) of the programs in the MVZ Faculty at the University of Córdoba. The data correspond to variables from the SPADIES System of the Ministry of National Education (MEN by its acronym in Spanish) and the university itself.

The response variable y corresponds to the proportion of subjects passed up to the point of dropout. The explanatory variables considered were: x_1 = Saber 11 test score (exams taken at the end of secondary education); x_2 = age at the time of taking the Saber 11 test; x_3 = variable indicating whether the student received financial support

(taking values 1 = yes, 0 = no); x_4 = mother's educational level (categorized as 1 if professional and, 0 otherwise); x_5 = number of siblings; x_6 = socioeconomic status of the student (categorized as 1 if from strata 1, 2, or 3, referred to as low and 0, otherwise); and x_7 = student's gender (categorized as 1 if male and 0 otherwise).

The zero-one inflated model, PHN, UPHN, and Doubly-Censored PHN (DCPHN) were fitted since some students drop out in the first semester without passing any subjects, and others drop out in the first four semesters even after passing all enrolled subjects.

The results obtained with the models studied in this article show that in all models, the significant variables for $0 < y < 1$ were the Saber 11 test score (x_1), age at the time of taking the Saber 11 test (x_2), and number of siblings (x_5). Similarly, the censored part at zero ($y = 0$) is not explained by any variable in any of the three models, while the censored part at one ($y = 1$) showed significance in variables such as age at the time of taking the Saber 11 test (directly related to the age of university entry) and number of siblings.

Table 1 shows the results of the best-fitted model for each of the considered models. To determine which model presents better performance, we used the AIC criteria [41] and the corrected AIC (AICc) [42]. These criteria are defined as:

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

where p is the number of parameters of the model in question.

The MLEs, with standard errors in parentheses, are given in Table 1. According to the AIC and AICc criteria, the model that best fits the student dropout data is the UPHN, followed by the DCPHN model.

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

Estimador	PHN	UPHN	DCPHN
$\hat{\beta}_{00}$	−2.1624	−2.1624	−2.4371
	(0.2071)	(0.2071)	(0.3025)
$\hat{\beta}_{10}$	2.9392	0.9771	1.3859
	(0.0144)	(0.0223)	(0.6003)
$\hat{\beta}_{11}$	0.0142	0.0273	0.0208
	(0.0092)	(0.0041)	(0.0096)
$\hat{\beta}_{12}$	−0.3281	−0.2844	−0.2687
	(0.0125)	(0.0175)	(0.0905)
$\hat{\beta}_{15}$	0.1295	0.2129	0.1847
	(0.0146)	(0.0205)	(0.0910)
$\hat{\beta}_{20}$	14.5124	14.5124	16.0286
	(7.9470)	(7.9470)	(13.6058)
$\hat{\beta}_{21}$	0.0208	0.0208	
	(0.0127)	(0.0127)	
$\hat{\beta}_{22}$	−1.2230	−1.2230	−1.2024
	(0.5150)	(0.5150)	(0.8650)
$\hat{\beta}_{25}$	0.4998	0.4998	
	(0.2509)	(0.2509)	
$\hat{\sigma}$	0.1064	0.1160	0.1238
	(0.0104)	(0.0057)	(0.0598)
$\hat{\alpha}$	0.1538	0.1427	0.1721
	(0.0364)	(0.0197)	(0.1933)
AIC	195.0036	182.4216	183.6414
AICc	198.4687	185.6696	186.6646

Where PHN is proportional hazard normal, UPHN is truncated proportional hazard normal, DCPHN is doubly censored proportional hazard normal, AIC is Akaike information criterion, and AICc is corrected Akaike information criterion.

To identify outliers and/or model misspecification, we examined the transformation of the martingale residual, rMT_i, as proposed by Barros et al. [43]. These residuals are defined by

$$rMT_i = \text{sgn}(rM_i)\sqrt{-2[rM_i + \delta_i \log(\delta_i - rM_i)]}; \quad i = 1, 2, 3, \cdots, n,$$

where $rM_i = \delta_i + \log(S(e_i, \hat{\theta}))$ is the martingal residual proposed by Ortega et al. [44], where $\delta_i = 0, 1$ indicates whether the ith observation is censored or not, respectively, $\text{sgn}(rM_i)$ denotes the sign of rM_i and $S(e_i; \hat{\theta})$ represents the survival function evaluated at e_i, where $\hat{\theta}$ are the MLE for θ.

The plots of rMT_i with confidence envelope graphs generated for the PHN, UPHN, and DCPHN models, shown in Figures 1 and 2, indicate that the fitted regression models PHN, UPHN, and DCPHN, with a logit link function, exhibit a good fit.

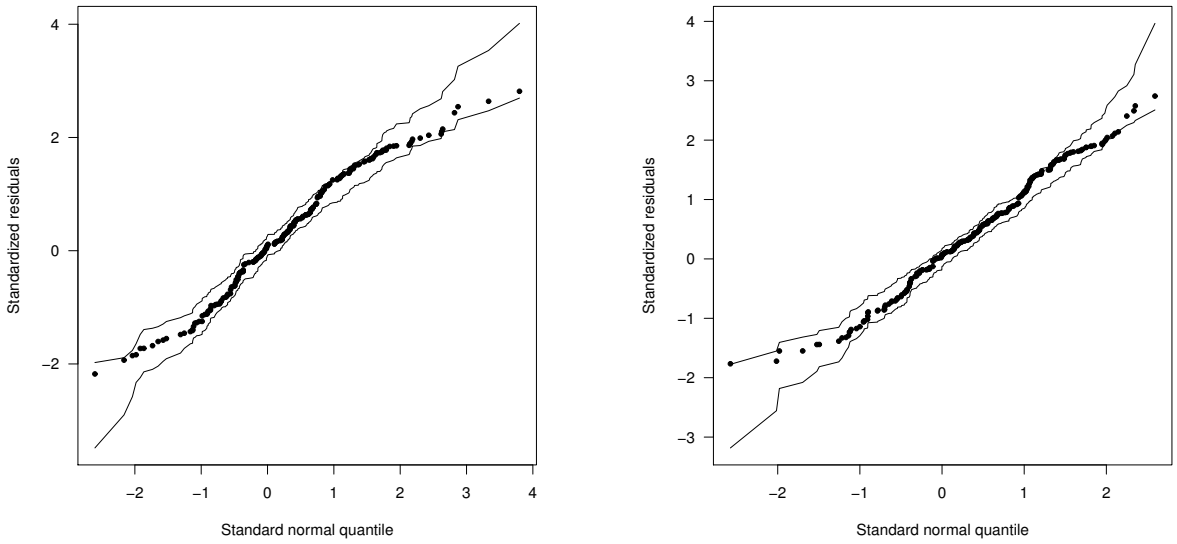

Figure 1. Plots of envelopes for rMT_i using: (**Left**) PHN and (**Right**) UPHN models and dropout data.

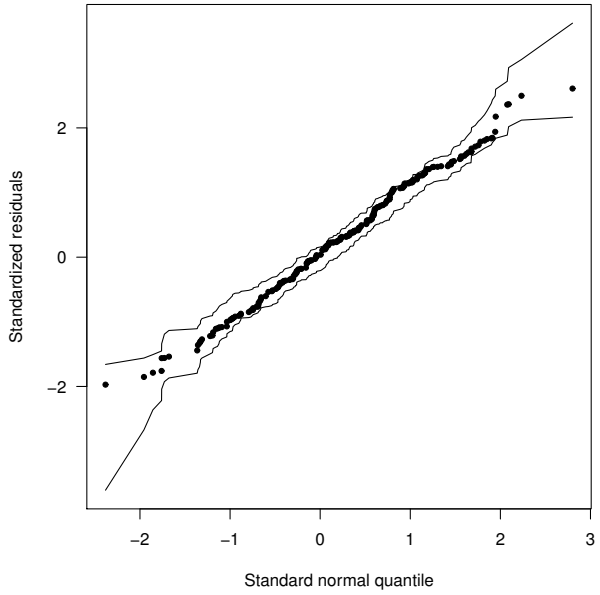

Figure 2. Plots of envelopes for rMT_i using DCPHN model and dropout data.

4.2. Application 2: Case Study on Periodontal Disease Data

The data motivating this second application come from a clinical study in which the clinical attachment level (CAL), a key marker of periodontal disease (PD), was measured at six sites on each tooth of a subject. The primary statistical question is to estimate functions that model the relationship between the "proportion of diseased sites associated with a specific tooth type (incisors, canines, premolars, and first molars)" and the covariates described below. The full dataset was previously analyzed by Galvis et al. [45] and includes information from 290 individuals. The response variable in this study is the proportion of diseased sites for the premolars (denoted as Y), with auxiliary covariates being gender (X_1), age (X_2), glycosylated hemoglobin (X_3), and smoking status (X_4).

The dataset exhibits significant inflation at $Y = 0$, but for certain subjects, we also observe $Y = 1$. To account for this, we applied the beta zero-one inflated (BIZU), truncated log-normal zero-one inflated (LNIZU), doubly censored proportional hazard normal (DCPHN), and the UPHN inflated zero-one (UPHNIZU) regression models. Our analysis revealed that only the covariates X_1 and X_2 were statistically significant. For the DCPHN model, only X_2 was significant for both the discrete outcomes.

We used several information criteria to compare the various models, including AIC and the AIC_C. We also used the Bayesian Information Criterion (BIC) and the Hannan–Quinn Information Criterion, defined as follows:

$$BIC = -2\ell(\theta) + p\log(n), \quad HQC = -2\ell(\theta) + 2p\log(\log(n)),$$

where p is the number of parameters of the model in question.

The MLEs, with standard errors in parentheses, are given in Table 2.

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

Estimador	BIZU	LNIZU	DCPHN	UPHNIZU
β_{00}	0.6337	0.6337	−7.2205	0.6337
	(0.7408)	(0.7408)	(0.8854)	(0.7408)
β_{02}	−0.0376	−0.0376	−0.0935	−0.0376
	(0.0135)	(0.0135)	(0.0161)	(0.0135)
β_{10}	−1.3885	−2.8949	−2.4039	−5.2246
	(0.3957)	(1.1453)	(0.6809)	(3.1908)
β_{11}	−0.5366	−1.3134	−0.5517	−2.9349
	(0.1613)	(0.4387)	(0.2420)	(1.4567)
β_{12}	0.0217	0.0393	0.0363	0.1325
	(0.0068)	(0.0194)	(0.0123)	(0.0735)
β_{20}	−8.0316	−8.0316	−12.7261	−8.0316
	(2.3153)	(2.3145)	(1.4938)	(2.3153)
β_{22}	0.0788	0.0788	−0.0487	0.0788
	(0.0358)	(0.0358)	(0.0236)	(0.0358)
σ	0.0903	0.3096	0.3060	0.6011
	(0.0652)	(0.0796)	(0.0305)	(0.1354)
α			1.5871	2.8429
			(0.0974)	(0.7634)
AIC	311.7097	316.0700	325.2363	308.0793
AICC	314.3525	318.7128	328.3095	310.8678
BIC	341.0687	345.4290	358.2652	341.1082
HQC	323.4723	327.8326	338.4693	321.3123

In Figures 3–6, it can be observed that the best fits correspond to the BIZU and UPHNIZU models. Additionally, note that in three of the criteria, the UPHNIZU model performs better than the BIZU model, while for the fourth criterion (BIC), no significant

differences are found between the two models. It is important to consider that the BIZU model has one less parameter, which further supports the superior fit of the UPHNIZU model. This allows us to conclude that the UPHNIZU model is a promising new alternative for modeling responses within the unit interval $[0, 1]$ with zero-one inflation.

We also generated standardized residual plots to identify the presence of outliers when fitting the UPHNIZU model. Additionally, we present the cumulative distribution function (CDF) plot of the UPHN model (Figure 5). From these, the model shows a good fit, and no outliers are detected. In addition, envelope plots were obtained for the fitted models BIZU, LNIZU, and DCLPHN, which are presented in Figures 3 and 4. These plots demonstrate that the BIZU and LNIZU models exhibit a better fit than the DCPHN model.

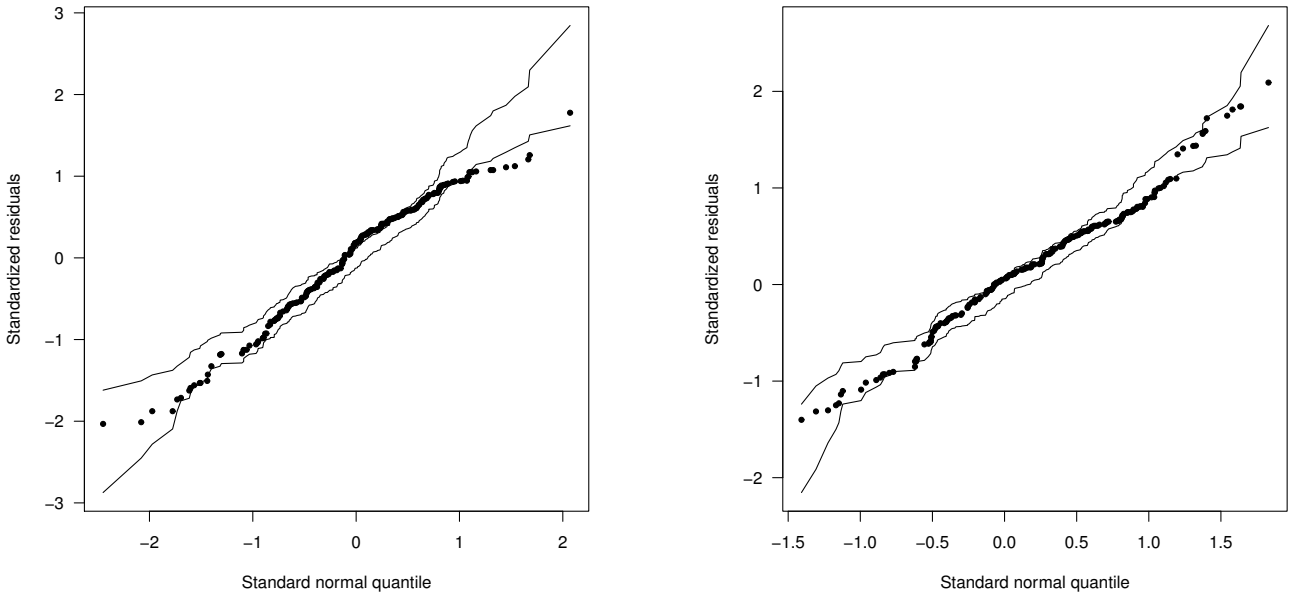

Figure 3. Plots of envelopes for rMT_i using: (**Left**) BIZU and (**Right**) LNIZU models and periodontal data.

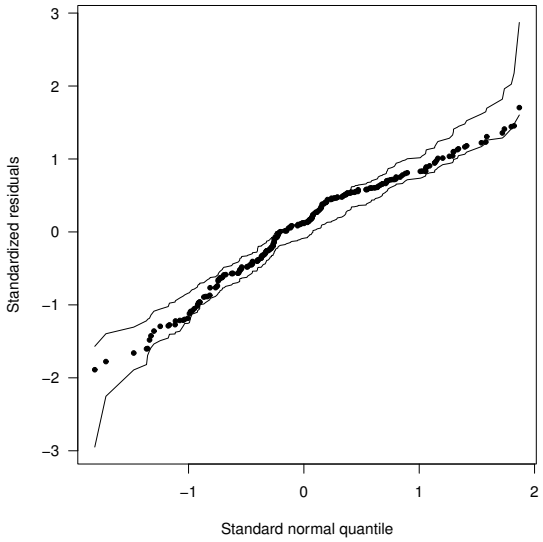

Figure 4. Plots of envelopes for rMT_i using DCPHN model and periodontal data.

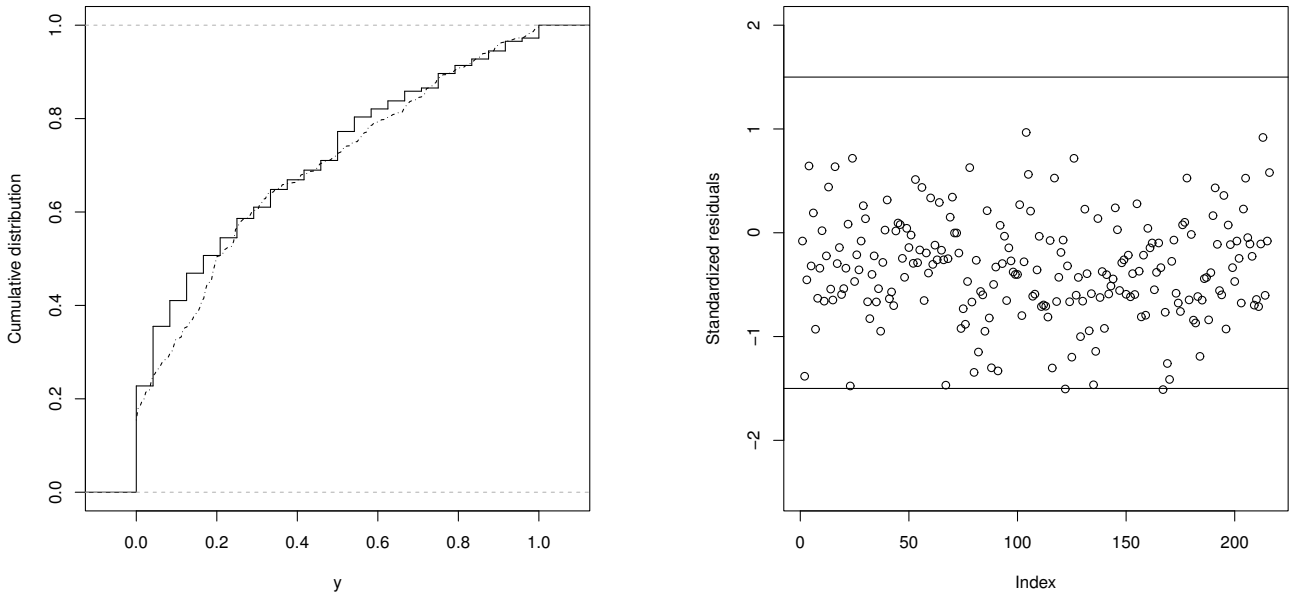

Figure 5. (**Left**) Empirical CDF of the residuals of the UPHNIZU model (solid line) and fitted CDF (dashed line). (**Right**) Plots of the standardized residuals of the UPHNIZU model, periodontal data.

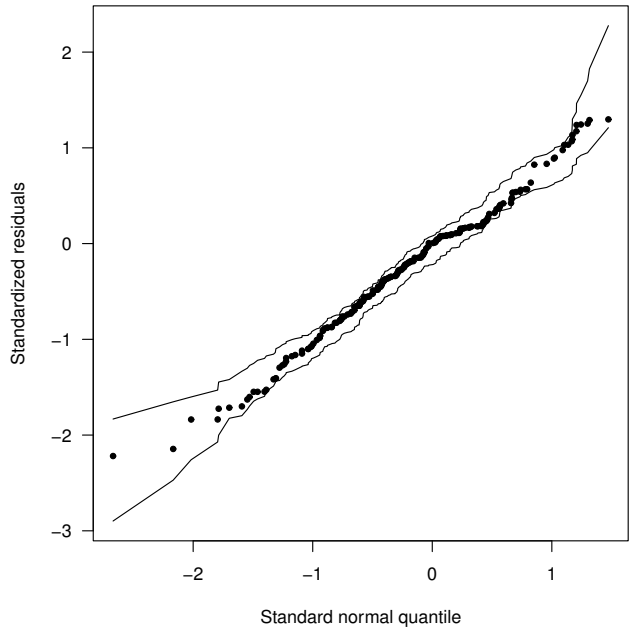

Figure 6. Plots of envelopes for rMT_i using UPHNIZU model and periodontal data.

5. Discussion

In this article, we introduced a broad class of skew regression models designed for response variables that lie within the unit interval, which may exhibit an excess of zeros or ones. These models were derived from a continuous-discrete mixture distribution that incorporates covariates in both its discrete and continuous components. As evidenced by applications using real data, the models we propose serve as a viable alternative for modeling rates and proportions that are inflated at either zero or one.

5.1. Major Results and Implications

Our findings demonstrate that the UPHNIZU model consistently surpassed other models in terms of AIC, AICc, BIC, and HQC values. These models delivered a superior fit for the data obtained from the case study on students' dropout data and the clinical study on periodontal disease, where the response variable was the proportion of diseased tooth sites.

Our findings also demonstrate that UPHNIZU models generate a non-singular information matrix, allowing valid statistical inferences and outperforming other asymmetric models like those derived from the skew-normal distribution or the beta distribution. Empirical results show the models' effectiveness in analyzing proportional data with zero and one inflation, highlighting their robustness and practicality in various research fields such as biomedicine, economics, and engineering. Additionally, they present parameter estimation methods using maximum likelihood and discuss applications in student dropout studies and periodontal disease. UPHNIZU models are a promising alternative for analyzing bounded data with extreme inflation, providing a robust and flexible tool to capture the complex characteristics of such data. The research also emphasizes the importance of innovations in probability distributions and their application in modeling complex phenomena, offering an advanced solution for the challenges of modeling proportional data with zero and one inflation.

5.2. Model Limitations

Although the results are encouraging, our study has several limitations. First, the models' complexity and reliance on iterative numerical methods for parameter estimation can lead to high computational demands. Second, while the models showed strong performance with the datasets utilized in this research, additional validation on different types of data is required to ensure their applicability in broader contexts.

5.3. Prospects for Further Investigation

Future research may explore several avenues, including the creation of more efficient algorithms to lessen the computational demands of fitting these models. Furthermore, applying these models in fields like economics or environmental studies could offer additional validation and reveal new applications.

Given the importance of model performance in our analysis, while the methods employed—such as AIC, AICc, BIC, HQC, and martingale residuals—are effective for evaluating model adequacy, there is room for improvement. Future research could investigate additional goodness-of-fit tests specifically designed for bounded and inflated data, which could offer a more thorough evaluation of model performance and robustness. Additionally, exploring Bayesian inference methods for unit interval data with inflation could provide valuable insights and enhance the analytical framework.

An intriguing avenue for future research involves adapting these models to accommodate longitudinal or hierarchical data structures. This would require methods to manage correlations within subjects or groups, often present in practical datasets. Additionally, examining the robustness of these models in various misspecification scenarios could lead to more resilient modeling strategies.

6. Conclusions

Analyzing proportion data, particularly when values are inflated at zero and one, presents significant challenges across various scientific disciplines. Conventional models, such as beta and Tobit regression models, frequently fail to accurately capture the complexities associated with such data. This underscores the need for more sophisticated modeling techniques capable of addressing the unique distributional characteristics of zero-one inflation.

This work tackled these challenges by introducing the proportional hazard normal zero-one inflated models. These models incorporate a continuous-discrete mixture distribution with covariates in both components, offering an advanced framework for analyzing

proportion data with specific inflation points. Consequently, the proportional hazard normal zero-one inflated models provide a robust and flexible method for capturing asymmetrically distributed data and mixed discrete-continuous characteristics, prevalent in fields such as medicine, sociology, humanities, and economics.

Our applications, which pertain to two case studies on student dropout and periodontal data, demonstrated that the proportional hazard normal zero-one inflated models with the logit link function are an excellent alternative to traditional models. The transformation of martingale residuals and the generation of simulated envelopes further validated the robustness of our models, underscoring their effectiveness in identifying model misfits and outliers. The proposed models address a critical gap in statistical modeling, providing valuable insights and reliable estimators for handling bounded and inflated data. The flexibility and robustness of the proportional hazard normal zero-one inflated models make them a viable alternative for describing proportion data that are inflated at zero or one.

In conclusion, the proportional hazard-normal zero-one inflated models signify a significant advancement in statistical modeling techniques for proportion data exhibiting zero-one inflation. These models provide a robust and adaptable framework for analyzing such data, yielding deeper insights and more reliable estimators.

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

Funding: This research was supported by Vice-rectorate for Research of the Universidad de Córdoba, Colombia, project grant FCB-06-22: "Estudio de la deserción en los programas de pregrado de la Universidad de Córdoba usando diferentes metodologías estadísticas" (G.M.-F. and R.T.-F).

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

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.

Appendix A. Elements of the Observed Information Matrix

Appendix A.1. Truncated Proportional Hazard Normal Model

$$
\begin{aligned}
\ddot{\ell}(\xi\xi) &= \frac{n}{\sigma^2} + n\frac{\alpha-1}{\sigma^2}\left[\overline{h^2} - \overline{zh}\right] + n\frac{\alpha}{\sigma^2}\frac{\left(h(z_0)\{S(z_0)\}^\alpha - h(z_1)\{S(z_1)\}^\alpha\right)^2}{W^2} \\
&\quad -n\frac{\alpha}{\sigma^2}\frac{z_0h(z_0)\{S(z_0)\}^\alpha - z_1h(z_1)\{S(z_1)\}^\alpha + (\alpha-1)\left(h^2(z_0)\{S(z_0)\}^\alpha - h^2(z_1)\{S(z_1)\}^\alpha\right)}{W}, \\
\ddot{\ell}(\xi\sigma) &= \frac{2n}{\sigma^2}\overline{z} + n\frac{\alpha-1}{\sigma^2}\left[-\overline{zh^2} - \overline{z^2h} + \overline{h}\right] + n\frac{\alpha}{\sigma^2}\frac{h(z_0)\{S(z_0)\}^\alpha - h(z_1)\{S(z_1)\}^\alpha}{W} \\
&\quad +n\frac{\alpha}{\sigma^2}\frac{\left(h(z_0)\{S(z_0)\}^\alpha - h(z_1)\{S(z_1)\}^\alpha\right)\left(z_0h(z_0)\{S(z_0)\}^\alpha - z_1h(z_1)\{S(z_1)\}^\alpha\right)}{W^2} \\
&\quad -n\frac{\alpha}{\sigma^2}\frac{z_0^2h(z_0)\{S(z_0)\}^\alpha - z_1^2h(z_1)\{S(z_1)\}^\alpha + (\alpha-1)\left(z_0h^2(z_0)\{S(z_0)\}^\alpha - z_1h^2(z_1)\{S(z_1)\}^\alpha\right)}{W}, \\
\ddot{\ell}(\sigma\sigma) &= -\frac{n}{\sigma^2} + \frac{3n}{\sigma^2}\overline{z^2} + n\frac{\alpha-1}{\sigma^2}\left[2\overline{zh} + \overline{z^2h^2} - \overline{z^3h}\right] + n\frac{\alpha}{\sigma^2}\frac{\left(z_0h(z_0)\{S(z_0)\}^\alpha - z_1h(z_1)\{S(z_1)\}^\alpha\right)^2}{W^2} \\
&\quad +n\frac{\alpha}{\sigma^2}\frac{-z_0h(z_0)\{S(z_0)\}^\alpha\left(-2+z_0^2+(\alpha-1)z_0h(z_0)\right) + z_1h(z_1)\{S(z_1)\}^\alpha\left(-2+z_1^2+(\alpha-1)z_1h(z_1)\right)}{W},
\end{aligned}
$$

$$
\begin{aligned}
\ddot{\ell}(\xi\alpha) &= -\frac{n}{\sigma}\overline{h} + n\frac{\alpha}{\sigma}\frac{\left(h(z_0)\{S(z_0)\}^{\alpha} - h(z_1)\{S(z_1)\}^{\alpha}\right)\left(\{S(z_0)\}^{\alpha}\log(S(z_0)) - \{S(z_1)\}^{\alpha}\log(S(z_1))\right)}{W^2} \\
&\quad -\frac{n}{\sigma}\frac{\left(h(z_0)\{S(z_0)\}^{\alpha} - h(z_1)\{S(z_1)\}^{\alpha}\right) + \alpha\left(h(z_0)\{S(z_0)\}^{\alpha}\log(S(z_0)) - h(z_1)\{S(z_1)\}^{\alpha}\log(S(z_1))\right)}{W}, \\
\ddot{\ell}(\sigma\alpha) &= -\frac{n}{\sigma}\overline{zh} + n\frac{\alpha}{\sigma}\frac{\left(z_0h(z_0)\{S(z_0)\}^{\alpha} - z_1h(z_1)\{S(z_1)\}^{\alpha}\right)\left(\{S(z_0)\}^{\alpha}\log(S(z_0)) - \{S(z_1)\}^{\alpha}\log(S(z_1))\right)}{W^2} \\
&\quad -\frac{n}{\sigma}\frac{\left(z_0h(z_0)\{S(z_0)\}^{\alpha} - z_1h(z_1)\{S(z_1)\}^{\alpha}\right) + \alpha\left(z_0h(z_0)\{S(z_0)\}^{\alpha}\log(S(z_0)) - z_1h(z_1)\{S(z_1)\}^{\alpha}\log(S(z_1))\right)}{W}, \\
\ddot{\ell}(\alpha\alpha) &= \frac{n}{\alpha} - n\frac{\left(\{S(z_0)\}^{\alpha}\log^2(S(z_0)) - \{S(z_1)\}^{\alpha}\log^2(S(z_1))\right)}{W} \\
&\quad +n\frac{\left(\{S(z_0)\}^{\alpha}\log(S(z_0)) - \{S(z_1)\}^{\alpha}\log(S(z_1))\right)^2}{W^2},
\end{aligned}
$$

where $h(z_i) = \frac{\phi(z_i)}{S(z_i)}$, $\overline{h} = \frac{1}{n}\sum_{i=1}^{n} h(z_i)$, $\overline{h^2} = \frac{1}{n}\sum_{i=1}^{n} h^2(z_i)$, $\overline{zh} = \frac{1}{n}\sum_{i=1}^{n} z_i h(z_i), \ldots, \overline{z^2h^2} = \frac{1}{n}\sum_{i=1}^{n} z_i^2 h^2(z_i)$, $\overline{z^3h} = \frac{1}{n}\sum_{i=1}^{n} z_i^3 h(z_i)$,

Appendix A.2. Unit-Proportional Hazard Normal Regression Model

$$
\begin{aligned}
\ddot{\ell}(\beta_j\beta_k) &= \frac{n}{\sigma^2} + \frac{\alpha-1}{\sigma^2}\sum_{i=1}^{n} x_{ij}x_{ik}\mu_i^2(1-\mu_i)^2\left[h^2(z_i) - z_ih(z_i)\right] \\
&\quad +\frac{\alpha}{\sigma^2}\sum_{i=1}^{n} x_{ij}x_{ik}\mu_i^2(1-\mu_i)^2\frac{\left(h(z_{0i})\{S(z_{0i})\}^{\alpha} - h(z_{1i})\{S(z_{1i})\}^{\alpha}\right)^2}{W_i^2} \\
&\quad -\frac{\alpha}{\sigma^2}\sum_{i=1}^{n} x_{ij}x_{ik}\mu_i^2(1-\mu_i)^2\frac{z_{0i}h(z_{0i})\{S(z_{0i})\}^{\alpha} - z_{1i}h(z_{1i})\{S(z_{1i})\}^{\alpha}}{W_i} \\
&\quad -\frac{\alpha(\alpha-1)}{\sigma^2}\sum_{i=1}^{n} x_{ij}x_{ik}\mu_i^2(1-\mu_i)^2\frac{\left(h^2(z_{0i})\{S(z_{0i})\}^{\alpha} - h^2(z_{1i})\{S(z_{1i})\}^{\alpha}\right)}{W_i} \\
&\quad -\frac{1}{\sigma}\sum_{i=1}^{n} x_{ij}x_{ik}\mu_i(1-\mu_i)(1-2\mu_i)z_i - \frac{\alpha-1}{\sigma}\sum_{i=1}^{n} x_{ij}x_{ik}\mu_i(1-\mu_i)(1-2\mu_i)h(z_i) \\
&\quad +\frac{\alpha}{\sigma}\sum_{i=1}^{n} x_{ij}x_{ik}\mu_i(1-\mu_i)(1-2\mu_i)\frac{h(z_{0i})\{S(z_{0i})\}^{\alpha} - h(z_{1i})\{S(z_{1i})\}^{\alpha}}{W_i}, \\
\ddot{\ell}(\beta_j\sigma) &= \frac{2}{\sigma^2}\sum_{i=1}^{n} x_{ij}\mu_i(1-\mu_i)z_i + \frac{\alpha-1}{\sigma^2}\sum_{i=1}^{n} x_{ij}\mu_i(1-\mu_i)\left[-z_ih^2(z_i) - z_i^2h(z_i) + h(z_i)\right] \\
&\quad +\frac{\alpha}{\sigma^2}\sum_{i=1}^{n} x_{ij}\mu_i(1-\mu_i)\frac{h(z_{0i})\{S(z_{0i})\}^{\alpha} - h(z_{1i})\{S(z_{1i})\}^{\alpha}}{W_i} \\
&\quad +\frac{\alpha}{\sigma^2}\sum_{i=1}^{n} x_{ij}\mu_i(1-\mu_i)\frac{\left(h(z_{0i})\{S(z_{0i})\}^{\alpha} - h(z_{1i})\{S(z_{1i})\}^{\alpha}\right)\left(z_{0i}h(z_{0i})\{S(z_{0i})\}^{\alpha} - z_{1i}h(z_{1i})\{S(z_{1i})\}^{\alpha}\right)}{W_i^2} \\
&\quad -\frac{\alpha}{\sigma^2}\sum_{i=1}^{n} x_{ij}\mu_i(1-\mu_i)\frac{z_{0i}^2h(z_{0i})\{S(z_{0i})\}^{\alpha} - z_{1i}^2h(z_{1i})\{S(z_{1i})\}^{\alpha}}{W_i} \\
&\quad -\frac{\alpha(\alpha-1)}{\sigma^2}\sum_{i=1}^{n} x_{ij}\mu_i(1-\mu_i)\frac{\left(z_{0i}h^2(z_{0i})\{S(z_{0i})\}^{\alpha} - z_{1i}h^2(z_{1i})\{S(z_{1i})\}^{\alpha}\right)}{W_i},
\end{aligned}
$$

$$
\begin{aligned}
\ddot{\ell}(\sigma\sigma) &= -\frac{n}{\sigma^2}+\frac{3}{\sigma^2}\sum_{i=1}^{n}z_i^2+\frac{\alpha-1}{\sigma^2}\sum_{i=1}^{n}\Big[2z_ih(z_i)+z_i^2h^2(z_i)-z_i^3h(z_i)\Big]\\
&\quad+\frac{\alpha}{\sigma^2}\sum_{i=1}^{n}\frac{\left(z_{0i}h(z_{0i})\{S(z_{0i})\}^{\alpha}-z_{1i}h(z_{1i})\{S(z_{1i})\}^{\alpha}\right)^2}{W_i^2}\\
&\quad-\frac{\alpha}{\sigma^2}\sum_{i=1}^{n}\frac{z_{0i}h(z_{0i})\{S(z_{0i})\}^{\alpha}\left(-2+z_{0i}^2+(\alpha-1)z_{0i}h(z_{0i})\right)}{W_i}\\
&\quad+\frac{\alpha}{\sigma^2}\sum_{i=1}^{n}\frac{+z_{1i}h(z_{1i})\{S(z_{1i})\}^{\alpha}\left(-2+z_{1i}^2+(\alpha-1)z_{1i}h(z_{1i})\right)}{W_i},\\
\ddot{\ell}(\beta_j\alpha) &= -\frac{n}{\sigma}\sum_{i=1}^{n}x_{ij}\mu_i(1-\mu_i)h(z_i)-\frac{1}{\sigma}\sum_{i=1}^{n}x_{ij}\mu_i(1-\mu_i)\frac{h(z_{0i})\{S(z_{0i})\}^{\alpha}-h(z_{1i})\{S(z_{1i})\}^{\alpha}}{W_i}\\
&\quad+\frac{\alpha}{\sigma}\sum_{i=1}^{n}x_{ij}\mu_i(1-\mu_i)\frac{h(z_{0i})\{S(z_{0i})\}^{\alpha}\left(\{S(z_{0i})\}^{\alpha}\log(S(z_{0i}))-\{S(z_{1i})\}^{\alpha}\log(S(z_{1i}))\right)}{W_i^2}\\
&\quad-\frac{\alpha}{\sigma}\sum_{i=1}^{n}x_{ij}\mu_i(1-\mu_i)\frac{h(z_{1i})\{S(z_{1i})\}^{\alpha}\left(\{S(z_{0i})\}^{\alpha}\log(S(z_{0i}))-\{S(z_{1i})\}^{\alpha}\log(S(z_{1i}))\right)}{W_i^2}\\
&\quad-\frac{\alpha}{\sigma}\sum_{i=1}^{n}x_{ij}\mu_i(1-\mu_i)\frac{h(z_{0i})\{S(z_{0i})\}^{\alpha}\log(S(z_{0i}))-h(z_{1i})\{S(z_{1i})\}^{\alpha}\log(S(z_{1i}))}{W_i},\\
\ddot{\ell}(\sigma\alpha) &= -\frac{1}{\sigma}\sum_{i=1}^{n}z_ih(z_i)+\frac{\alpha}{\sigma}\sum_{i=1}^{n}\frac{z_{0i}h(z_{0i})\{S(z_{0i})\}^{\alpha}\left(\{S(z_{0i})\}^{\alpha}\log(S(z_{0i}))-\{S(z_{1i})\}^{\alpha}\log(S(z_{1i}))\right)}{W_i^2}\\
&\quad-\frac{\alpha}{\sigma}\sum_{i=1}^{n}\frac{z_{1i}h(z_{1i})\{S(z_{1i})\}\left(\{S(z_{0i})\}^{\alpha}\log(S(z_{0i}))-\{S(z_{1i})\}^{\alpha}\log(S(z_{1i}))\right)}{W_i^2}\\
&\quad-\frac{1}{\sigma}\sum_{i=1}^{n}\frac{z_{0i}h(z_{0i})\{S(z_{0i})\}^{\alpha}-z_{1i}h(z_{1i})\{S(z_{1i})\}^{\alpha}}{W_i}\\
&\quad-\frac{\alpha}{\sigma}\sum_{i=1}^{n}\frac{z_{0i}h(z_{0i})\{S(z_{0i})\}^{\alpha}\log(S(z_{0i}))-z_{1i}h(z_{1i})\{S(z_{1i})\}^{\alpha}\log(S(z_{1i}))}{W_i},\\
\ddot{\ell}(\alpha\alpha) &= \frac{n}{\alpha}-\sum_{i=1}^{n}\frac{\{S(z_{0i})\}^{\alpha}\log^2(S(z_{0i}))-\{S(z_{1i})\}^{\alpha}\log^2(S(z_{1i}))}{W_i}\\
&\quad+\sum_{i=1}^{n}\frac{\left(\{S(z_{0i})\}^{\alpha}\log(S(z_{0i}))-\{S(z_{1i})\}^{\alpha}\log(S(z_{1i}))\right)^2}{W_i^2},
\end{aligned}
$$

where $z_i=\frac{y_i-\mu_i}{\sigma}$, $z_{0i}=-\frac{\mu_i}{\sigma}$, $z_{1i}=\frac{1-\mu_i}{\sigma}$ and $W_i=W_i(\mu_i,\sigma,\alpha)=\log\left(\{S(z_{0i})\}^{\alpha}-\{S(z_{1i})\}^{\alpha}\right)$.

Appendix A.3. UPHN Regression Model Inflated at Zero and/or One

For the discrete part, the elements of the observed information matrix are given by:

$$
\begin{aligned}
\ddot{\ell}(\beta_{(0)r}\beta_{(0)r'}) &= \sum_{i=1}^{n}\frac{x_{(0)ip}x_{(0)ip'}\exp(\boldsymbol{x}_{(0)i}^{\top}\boldsymbol{\beta}_{(0)})\left[1+\exp(\boldsymbol{x}_{(1)i}^{\top}\boldsymbol{\beta}_{(1)})\right]}{\left(1+\exp(\boldsymbol{x}_{(0)i}^{\top}\boldsymbol{\beta}_{(0)})+\exp(\boldsymbol{x}_{(1)i}^{\top}\boldsymbol{\beta}_{(1)})\right)^2},\\
\ddot{\ell}(\beta_{(1)q}\beta_{(0)r}) &= -\sum_{i=1}^{n}\frac{x_{(0)ip}x_{(1)iq}\exp(\boldsymbol{x}_{(0)i}^{\top}\boldsymbol{\beta}_{(0)})\exp(\boldsymbol{x}_{(1)i}^{\top}\boldsymbol{\beta}_{(1)})}{\left(1+\exp(\boldsymbol{x}_{(0)i}^{\top}\boldsymbol{\beta}_{(0)})+\exp(\boldsymbol{x}_{(1)i}^{\top}\boldsymbol{\beta}_{(1)})\right)^2},\\
\ddot{\ell}(\beta_{(1)q}\beta_{(1)q'}) &= \sum_{i=1}^{n}\frac{x_{(1)iq}x_{(1)iq'}\exp(\boldsymbol{x}_{(1)i}^{\top}\boldsymbol{\beta}_{(1)})\left[1+\exp(\boldsymbol{x}_{(0)i}^{\top}\boldsymbol{\beta}_{(0)})\right]}{\left(1+\exp(\boldsymbol{x}_{(0)i}^{\top}\boldsymbol{\beta}_{(0)})+\exp(\boldsymbol{x}_{(1)i}^{\top}\boldsymbol{\beta}_{(1)})\right)^2},
\end{aligned}
$$

while the elements for the continuous part are given by:

$$\begin{aligned}
\ddot{\ell}(\beta_j\beta_k) &= \frac{n_{01}}{\sigma^2} + \frac{\alpha-1}{\sigma^2}\sum_{y_i\in(0,1)} x_{ij}x_{ik}\mu_i^2(1-\mu_i)^2\Big[h^2(z_i) - z_i h(z_i)\Big] \\
&\quad -\frac{1}{\sigma}\sum_{y_i\in(0,1)} x_{ij}x_{ik}\mu_i(1-\mu_i)(1-2\mu_i)z_i - \frac{\alpha-1}{\sigma}\sum_{y_i\in(0,1)} x_{ij}x_{ik}\mu_i(1-\mu_i)(1-2\mu_i)h(z_i), \\
\ddot{\ell}(\beta_j\sigma) &= \frac{2}{\sigma^2}\sum_{y_i\in(0,1)} x_{ij}\mu_i(1-\mu_i)z_i + \frac{\alpha-1}{\sigma^2}\sum_{y_i\in(0,1)} x_{ij}\mu_i(1-\mu_i)\Big[-z_i h^2(z_i) - z_i^2 h(z_i) + h(z_i)\Big], \\
\ddot{\ell}(\sigma\sigma) &= -\frac{n_{01}}{\sigma^2} + \frac{3}{\sigma^2}\sum_{y_i\in(0,1)} z_i^2 + \frac{\alpha-1}{\sigma^2}\sum_{y_i\in(0,1)}\Big[2z_i h(z_i) + z_i^2 h^2(z_i) - z_i^3 h(z_i)\Big], \\
\ddot{\ell}(\beta_j\alpha) &= -\frac{n}{\sigma}\sum_{y_i\in(0,1)} x_{ij}\mu_i(1-\mu_i)h(z_i), \\
\ddot{\ell}(\sigma\alpha) &= -\frac{1}{\sigma}\sum_{y_i\in(0,1)} z_i h(z_i), \\
\ddot{\ell}(\alpha\alpha) &= \frac{n_{01}}{\alpha},
\end{aligned}$$

where $z_i = \frac{y_i-\mu_i}{\sigma}$ and n_{01} is the number of sample elements that belong to the interval $(0,1)$.

References

1. Eugene, N.; Lee, C.; Famoye, F. Beta-normal Distribution and Its Applications. *Commun.-Stat.-Theory Methods* **2002**, *31*, 497–512. [CrossRef]
2. Cordeiro, G.M.; de Castro, M. A New Family of Generalized Distributions. *J. Stat. Comput. Simul.* **2011**, *81*, 883–898. [CrossRef]
3. Silva, G.O.; Ortega, E.M.M.; Cordeiro, G.M. The Beta Modified Weibull Distribution. *Lifetime Data Anal.* **2010**, *16*, 409–430. [CrossRef] [PubMed]
4. Mahdavi, M.; Silva, G.O. A method to expand family of continuous distributions based on truncated distributions. *J. Statist. Res.* **2016**, *13*, 231–247. [CrossRef]
5. Chen, S.; Gui, W. Estimation of unknown parameters of Truncated Normal Distribution under Adaptive Progressive Type II Censoring Scheme. *Mathematics* **2021**, *9*, 49. [CrossRef]
6. Taketomi, N.; Yamamoto, K.; Chesneau, C.; Emura, T. Parametric Distributions for Survival and Reliability Analyses, a Review and Historical Sketch. *Mathematics* **2022**, *10*, 3907. [CrossRef]
7. Kreer, M.; Kızılersü, A.; Thomas, A.W.; Egídio dos Reis, A.D. Goodness-of-fit tests and applications for left-truncated Weibull distributions to non-life insurance. *Eur. Actuar. J.* **2015**, *5*, 139–163. [CrossRef]
8. Cordeiro, G.M.; Silva, G.O.; Ortega, E.M.M. The Beta-Weibull Geometric Distribution. *Statistics* **2013**, *47*, 817–834. [CrossRef]
9. Cordeiro, G.M.; Ortega, E.M.M.; Silva, G.O. The Kumaraswamy Modified Weibull Distribution: Theory and Applications. *J. Stat. Comput. Simul.* **2014**, *84*, 1387–1411. [CrossRef]
10. Zografos, K.; Balakrishnan, N. On Families of Beta-and Generalized Gamma generated Distributions and Associated Inference. *Stat. Methodol.* **2009**, *6*, 344–362. [CrossRef]
11. Ristić, M.M.; Balakrishnan, N. The Gamma-exponentiated Exponential Distribution. *J. Stat. Comput. Simul.* **2012**, *82*, 1191–1206. [CrossRef]
12. Castellares, F.; Santos, M.A.C.; Montenegro, L.C.; Cordeiro, G.M. A Gamma-Generated Logistic Distribution: Properties and Inference. *Am. J. Math. Manag. Sci.* **2015**, *34*, 14–39.
13. Cordeiro, G.M.; Lima, M.C.S.; Cysneiros, A.H.M.A.; Pascoa, M.A.R.; Pescim, R.R.; Ortega, E.M.M. An Extended Birnbaum–Saunders Distribution: Theory, Estimation, and Applications. *Commun. Stat.-Theory Methods* **2016**, *45*, 2268–2297. [CrossRef]
14. Enami, S.H. Truncated Lomax-exponential distribution ans its fitting to financial data. *J. Mahani Math. Res.* **2023**, *12*, 201–216.
15. Hadi, H.H.; Al-Noor, N.H. Truncated exponential Marshall Olkin Lomax distribution: Properties, entropies, and applications. *AIP Conf. Proc.* **2023**, *2414*, 201–216.
16. Al-Habib, K.H.; Khaleel, M.H.; Al-Mofleh, H. Statistical Properties and Application for [0,1] Truncated Nadarajah-Haghighi Exponential Distribution. *Ibn-Haitham J. Pure Appl. Sci.* **2024**, *37*, 363–377.
17. Schaminée, J.H.; Hennekens, S.M.; Chytry, M.; Rodwell, J.S. Vegetation-plot data and databases in Europe: An overview. *Preslia* **2009**, *81*, 173–185.
18. 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]
19. 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]

20. 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]
21. 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]
22. Azzalini, A. A class of distributions which includes the normal ones. *Scand. J. Stat.* **1985**, *12*, 171–178.
23. Azzalini, A. Further results on a class of distributions which includes the normal ones. *Statistica* **1986**, *46*, 199–208.
24. Henze, N. A probabilistic representation of the skew-normal distribution. *Scand. J. Stat.* **1986**, *13*, 271–275.
25. 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]
26. Gupta, R.D.; Gupta, R.C. Analyzing skewed data by power normal model. *Test* **2008**, *17*, 197–210. [CrossRef]
27. Pewsey, A.; Gómez, H.W.; Bolfarine, H. Likelihood-based inference for power distributions. *Test* **2012**, *21*, 775–789. [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. Jornsatian, 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. Martínez-Flórez, G.; Moreno-Arenas, G.; Vergara-Cardozo, S. Properties and inference for proportional hazard Models. *Rev. Colomb. Estadística* **2013**, *36*, 95–114.
37. Lehmann, E.L.; Casella, G. *Theory of Point Estimation*, 2nd ed.; Springer: New York, NY, USA, 1998.
38. Sarabia, J.M.; Castillo, E. About a class of max-stable families with applications to income distributions. *Int. J. Stat.* **2005**, *LXIII*, 505–527.
39. Cragg, J. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica* **1971**, *39*, 829–844. [CrossRef]
40. Moulton, L.; Halsey, N.A. A mixture model with detection limits for regression analysses of antibody response to vaccine. *Biometrics* **1995**, *51*, 1570–1578. [CrossRef]
41. Akaike, H. A new look at statistical model identification. *IEEE Trans. Autom. Control.* **1974**, *AU-19*, 716–722. [CrossRef]
42. Cavanaugh, J.E. Unifying the derivations for the Akaike and corrected Akaike information criteria. *Stat. Probab. Lett.* **1997**, *33*, 201–208. [CrossRef]
43. Barros, M.; Paula, G.A.; Leiva, V. A new class of survival regression models with heavy-tailed errors: Robustness and diagnostics. *Lifetime Data Anal.* **2010**, *14*, 316–332. [CrossRef] [PubMed]
44. 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]
45. Galvis, D.M.; Bandyopadhyay, D.; Lachos, V.H. Augmented mixed beta regression models for periodontal proportion data. *Stat. Med.* **2014**, *33*, 3759–3771. [CrossRef]