


Article

Dichotomous Proportional Hazard Regression Model: A Case Study on Students' Dropout

Guillermo Martínez-Flórez ^{1,†} , Roger Tovar-Falón ^{1,†}  and Carlos Barrera-Causil ^{2,*,†} 

¹ Departamento de Matemáticas y Estadística, Universidad de Córdoba, Montería 230002, Colombia; guillermomartinez@correo.unicordoba.edu.co (G.M.-F.); rjtovar@correo.unicordoba.edu.co (R.T.-F.)

² Grupo de Investigación Davinci, Facultad de Ciencias Exactas y Aplicadas, Instituto Tecnológico Metropolitano, Medellín 050034, Colombia

* Correspondence: carlosbarrera@itm.edu.co

† These authors contributed equally to this work.

Abstract: In problems involving binary classification, researchers often encounter data suitable for modeling dichotomous responses. These scenarios include medical diagnostics, where outcomes are classified as “disease” or “no disease”, and credit scoring in finance, determining whether a loan applicant is “high risk” or “low risk”. Dichotomous response models are also useful in many other areas for estimating binary responses. The logistic regression model is one option for modeling dichotomous responses; however, other statistical models may be required to improve the quality of fits. In this paper, a new regression model is proposed for cases where the response variable is dichotomous. This novel, non-linear model is derived from the cumulative distribution function of the proportional hazard distribution, and is suitable for modeling binary responses. Statistical inference is performed using a classical approach with the maximum likelihood method for the proposed model. Additionally, it is demonstrated that the introduced model has a non-singular information matrix. The results of a simulation study, along with an application to student dropout data, show the great potential of the proposed model in practical and everyday situations.

Keywords: dichotomous response; logistic regression; maximum likelihood estimation; proportional hazard distribution

MSC: 62J12



Citation: Martínez-Flórez, G.; Tovar-Falón, R.; Barrera-Causil, C. Dichotomous Proportional Hazard Regression Model: A Case Study on Students' Dropout. *Mathematics* **2024**, *12*, 2170. <https://doi.org/10.3390/math12142170>

Academic Editor: Heng Lian

Received: 11 June 2024

Revised: 8 July 2024

Accepted: 9 July 2024

Published: 11 July 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

In recent statistical literature, new probability distributions have been introduced as extensions of other known distributions. This methodology serves as a foundation for generating new families of distributions applicable across various fields. Among other authors, this approach was utilized by Eugene et al. [1] to propose the Beta-G class of distributions. Subsequently, Silva et al. [2] introduced the modified Weibull beta distribution families and the Weibull beta geometry, as noted by Cordeiro et al. [3]. Moreover, building upon this methodology, Cordeiro and de Castro [4] defined the Kumaraswamy-G class of distributions, followed by the suggestion of the Kumaraswamy modified Weibull by Cordeiro et al. [5]. Similarly, Zografos and Balakrishnan [6] and Ristić and Balakrishnan [7] presented a new family of distributions generated by gamma random variables, leading to the development of the Gamma-Generated-Logistic distributions by Castellares et al. [8] and the Gamma-Birnbaum-Saunders distributions by Cordeiro et al. [9].

On the other hand, Martínez-Flórez et al. [10] examined the exponentiated-skew-normal distribution. Similarly, Martínez-Flórez et al. [11] proposed the family of proportional hazard distributions based on the distribution of the minimum in the sample. All of these new families of distributions have proven useful in analyzing responses of interest by adjusting both linear and nonlinear regression models. For instance, the regression model

with skew-normal distributed errors Azzalini [12] has been widely utilized. Furthermore, extensions of regression models have been recommended, assuming errors follow the exposed skew-normal distributions (Martínez-Flórez et al. [10]; Martínez-Flórez et al. [13]). Moreover, these distribution families have been extended to encompass the case of the Birnbaum distribution [14] and the Birnbaum–Saunders log-linear regression model proposed by Rieck and Nedelman [15], showcasing the extensive range of symmetric and asymmetric families available in the literature.

All the works previously presented, and many others that have not been mentioned here, are appropriate in cases where the response variable has its support in the set of real numbers or has positive support, while very few works focus on the problem of dealing with dichotomous data. In this particular case, the issue is addressed based on non-linear functions or link functions such as the logistic regression model, known in the statistical literature as the logit model, or the non-linear alternative based on the cumulative distribution function (CDF) of the normal density, called the probit model. Thus, the limited existence of proposals in the statistical literature for the analysis of dichotomous or polytomous responses through link functions used in other types of models becomes evident.

In practice, the regression model with a dichotomous response (logistic model) has been widely used in several areas of knowledge. In the educational area, for example, it can be used to predict the probability of a student dropping out based on their academic performance, age of entry, the educational level of their parents, number of siblings, etc. In the health sector, certain patient characteristics and the application of specific treatments can be analyzed using the model to understand the connection between the patient and the implemented treatment, including the probability or odds of survival. Similarly, in finance, based on characteristics such as sex, age, race, income, and educational level, the behavior of investors can be predicted. These models are also utilized for classifying individuals into certain groups according to the predicted probability of a specific event occurring.

In this article, a new regression model is proposed to address research with dichotomous response variables. This novel model can be applied to various fields, including medicine, finance, education, and the social sciences. Our proposal is grounded in the family of proportional hazard distributions, specifically utilizing an extension of the logistic distribution within this family.

The remainder of this work is organized as follows: Section 2 provides a brief description of the logistic distribution and its associated regression model. Section 3 describes the proportional hazard and proportional hazard logistic distributions, along with some of their most important properties. In Section 4, the proportional hazard logistic regression model is introduced. Additionally, the statistical inference process is performed using a classical approach, presenting the score function and the elements of the observed information matrix. Section 5 presents an application of the introduced model to student dropout data. Finally, the conclusions of the paper are presented in Section 6.

2. Logistic Distribution

A continuous random variable with a logistic distribution has a probability density function (PDF) given by

$$f_L(z) = \frac{\exp(-z)}{(1 + \exp(-z))^2} = \frac{1}{4} \operatorname{sech}^2\left(\frac{z}{2}\right), \quad z \in \mathbb{R}. \quad (1)$$

where sech denotes the hyperbolic secant function. The shape of the logistic distribution is similar to the shape of the normal density, with heavier tails and greater kurtosis than the normal distribution.

The cumulative distribution function (CDF) of a random variable with a logistic distribution is given by

$$\mathcal{F}_L(z) = \frac{\exp(z)}{1 + \exp(z)} = \frac{1}{2} + \frac{1}{2} \tanh\left(\frac{z}{2}\right),$$

while its survival and hazard functions can be written as

$$\mathcal{S}_L(z) = \frac{1}{1 + \exp(z)} = \frac{1}{2} - \frac{1}{2} \tanh\left(\frac{z}{2}\right) \quad \text{and}$$

$$h_L(z) = \frac{\exp(z)}{1 + \exp(z)},$$

where \tanh denotes the hyperbolic tangent function.

The extension of the logistic distribution to the location-scale case is achieved by using the transformation $Y = \mu + \sigma Z$ with $\mu \in \mathbb{R}$ and $\sigma > 0$. This is denoted by $Y \sim L(\mu, \sigma)$, where μ represents the location parameter and σ the scale. Since this distribution is symmetric, then $\mathbb{E}(Y) = \mu$, $\text{Var}(Y) = \frac{\pi^2}{3}\sigma^2$, the asymmetry coefficient is zero, and its excess kurtosis is equal to $\frac{6}{5}$. Finally, the p -th percentile, for $0 < p < 1$, of this distribution is given by $y_p = \mu + \sigma \log(p/(1-p))$.

Associated with the logistic distribution is the logistic regression model, which is used to explain the probability of success of a random variable with a binomial distribution when there is a set of covariates that explain this probability (see Agresti [16]). In essence, the logistic regression model is given by

$$p_i = \Pr(Y_i = 1 \mid x_1, x_2, \dots, x_p) = \frac{\exp(x_i^\top \beta)}{1 + \exp(x_i^\top \beta)}$$

where $\mathbf{x} = (1, x_1, \dots, x_p)^\top$ represents a vector of covariates, $\beta = (\beta_0, \beta_1, \dots, \beta_p)^\top$ is the vector of model coefficients (unknown values that must be estimated), and Y_i is a Bernoulli random variable with parameter p_i .

3. Hazard Proportional Distribution

In recent decades, families of asymmetric distributions have been introduced for fitting data with tails heavier or lighter than the normal distribution. As is well known, in the presence of high degrees of skewness and/or kurtosis, inferential processes based on the assumption of normality are inadequate. Similarly, while the elliptical family may provide a solution for distributions with heavy tails, it fails to address the issue of asymmetry in the data under study.

The skew-normal (SN) distribution, introduced by Azzalini [12], is defined by the PDF given as

$$\varphi(z; \lambda) = 2\phi(z)\Phi(\lambda z), \quad z \in \mathbb{R}, \quad (2)$$

where ϕ and Φ represent the PDF and CDF of the standard normal distribution, respectively, and λ is a skewness parameter. The distribution is denoted by $Z \sim SN(\lambda)$. In addition to the work of Azzalini [12], the SN distribution described in (2) has been extensively studied by Henze [17], Pewsey [18], Chiogna [19], and Gómez et al. [20], among others.

Building on the work of Lehmann [21], Martínez-Flórez et al. [11] investigated another family of asymmetric univariate distributions called the proportional hazard. The PDF of this distribution is given by

$$\varphi_F(z; \alpha) = \alpha f(z) \{1 - F(z)\}^{\alpha-1}, \quad z \in \mathbb{R}, \quad (3)$$

where α is a positive real number, and F is a continuous CDF with continuous PDF f . This distribution is denoted by $\text{PHF}(\alpha)$. The hazard function associated with the density φ_F is

$$h_{\varphi_F}(X, \alpha) = \alpha h_f(x),$$

where $h_f = f/(1 - F)$ represents the hazard function related to the density f . When $F = \Phi(\cdot)$ and $f = \phi(\cdot)$, the distribution is called proportional hazard normal, denoted by $\text{PHN}(\alpha)$. The PDF is given by

$$\varphi_{\Phi}(z; \alpha) = \alpha \phi(z) \{S(z)\}^{\alpha-1}, \quad z \in \mathbb{R}, \quad (4)$$

where $S(z)$ is the survival function associated with the PDF $\phi(\cdot)$. This model serves as an alternative to accommodate data with asymmetry and kurtosis that fall outside the ranges allowed by the normal distribution.

The CDF of the $\text{PHN}(\alpha)$ distribution is given by:

$$F_{\Phi}(z; \alpha) = 1 - \{S(z)\}^{\alpha}, \quad z \in \mathbb{R}. \quad (5)$$

By varying the α parameter, Martínez-Flórez et al. [11] found that the range of asymmetry and kurtosis coefficients, $\sqrt{\beta_1}$ and β_2 , respectively, of the variable $Z \sim \text{PHN}(\alpha)$ falls within the intervals $(-1.1578, 0.9918)$ and $(1.1513, 4.3023)$. These ranges exhibit better skewness and kurtosis properties than those of the SN distribution. Additionally, Martínez-Flórez et al. [11] demonstrated that the information matrix of the PHN distribution in the location-scale case, denoted as $\text{PHN}(\mu, \sigma, \alpha)$, is nonsingular. A particular case of the proportional hazard family is discussed below.

Proportional Hazard Logistic Distribution

The proportional hazard logistic (PHL) distribution, denoted by $\text{PHL}(\alpha)$, is defined by the PDF given as

$$\begin{aligned} \varphi_{\text{HL}}(x; \alpha) &= \alpha \frac{\exp(x)}{(1 + \exp(x))^{\alpha+1}} \\ &= \frac{\alpha}{4} \text{sech}^2\left(\frac{x}{2}\right) \left[\frac{1}{2} - \frac{1}{2} \tanh\left(\frac{x}{2}\right)\right]^{\alpha-1} \end{aligned} \quad (6)$$

Its respective CDF is given by

$$\begin{aligned} \mathcal{F}_{\text{HL}}(x; \alpha) &= 1 - \frac{1}{(1 + \exp(x))^{\alpha}} \\ &= 1 - \left[\frac{1}{2} - \frac{1}{2} \tanh\left(\frac{x}{2}\right)\right]^{\alpha} \end{aligned} \quad (7)$$

while the survival and hazard functions can be expressed as

$$S_{\text{HL}}(x; \alpha) = \frac{1}{(1 + \exp(x))^{\alpha}} = \left[\frac{1}{2} - \frac{1}{2} \tanh\left(\frac{x}{2}\right)\right]^{\alpha} \quad (8)$$

and

$$\begin{aligned} h_{\text{HL}}(x; \alpha) &= \alpha \frac{\exp(x)}{1 + \exp(x)} \\ &= \frac{\alpha}{4} \frac{\text{sech}^2\left(\frac{x}{2}\right)}{\frac{1}{2} - \frac{1}{2} \tanh\left(\frac{x}{2}\right)} = \alpha h_L(x). \end{aligned} \quad (9)$$

respectively, where $h_L(x)$ is the hazard function of the logistic distribution.

Figure 1 illustrates the behavior of the CDF and the survival function for different values of the parameter α . It is noteworthy that, for $\alpha = 1$, the CDF corresponds to that of the logistic distribution. Moreover, the hazard function of the PHL is a multiple of the hazard function of the logistic distribution. Additionally, the adjustment of the CDF of the PHL distribution is more flexible than that of the logistic distribution. Similarly, it is observed that for $\alpha = 0.75$, the survival function converges more slowly (indicating a higher probability of survival) towards zero compared to the survival function of the logistic distribution, whereas for values greater than zero, the convergence to zero is faster (indicating a lower probability of survival).

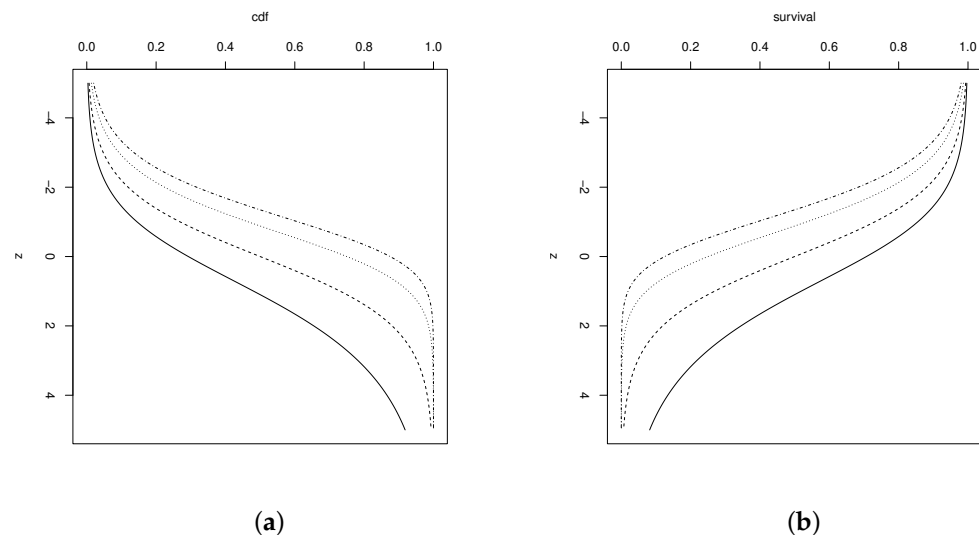


Figure 1. (a) CDF for $\alpha = 0.75$ (solid line), 1 (dotted line), 2 (dashed line), and 3 (dotted-dashed line). (b) Survival function for $\alpha = 0.75$ (solid line), 1 (dotted line), 2 (dashed line), and 3 (dotted-dashed line).

The r -th moment of the random variable $Y \sim \text{PHL}(\alpha)$ is given by:

$$\mathbb{E}(Y^r) = \int_1^\infty \frac{\log^r(u-1)}{(u-1)^2} (1-u^{-1})^{\alpha+1} du. \quad (10)$$

From Expression (10), the moments of orders 1, 2, 3, and 4 of the PHL distribution can be derived, facilitating the numerical calculation of its mean, variance, skewness, and kurtosis coefficients.

4. Proportional Hazard Logistic Regression Model

Assuming the regression model:

$$Y_i = \mathbf{X}_i^\top \boldsymbol{\beta} + \varepsilon_i = \mu_i + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (11)$$

where $\mathbf{X} = (1, x_1, \dots, x_p)^\top$ represents a set of covariates, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$ denotes a set of unknown coefficients, and $\varepsilon_i \sim \text{PHL}(0, \sigma, \alpha)$. It then follows that $Y_i \sim \text{PHL}(\mu_i, \sigma, \alpha)$, for $i = 1, 2, \dots, n$.

However, when Y is a dichotomous random variable with values zero and one, the model errors are not independent and do not satisfy the assumption of homoscedasticity. Additionally, it cannot be ensured that

$$\mathbb{E}(Y_i | x_1, \dots, x_p) = \Pr(Y_i = 1 | x_1, \dots, x_p)$$

is bounded by 0 and 1.

For this reason, it is necessary to determine a distribution function $G(\cdot)$ such that

$$\Pr(Y_i = 1 | x_1, \dots, x_p) = p_i = G(Y_i = 1 | x_1, \dots, x_p).$$

The function $G(\cdot | x_1, \dots, x_p)$ is known as a link function, and since it must ensure that the prediction lies between 0 and 1, it is commonly chosen as the distribution function of certain random variables studied in classical probability theory literature.

The link functions $G(\cdot | x_1, \dots, x_p)$ typically utilized in practice are the CDF of the logistic distribution, resulting in the logit model, and the CDF of the normal distribution, resulting in the probit model. Due to their mathematical and computational complexity, the logit model is generally preferred over the probit model in practical applications. A notable commonality between these two models is their symmetric CDF, which can be a limitation in scenarios where the probability of success for response variable Y exhibits asymmetric behavior. Moreover, both distributions have limitations in accurately modeling certain probabilities in their tails. As illustrated in Figure 1, the CDF of the proportional hazard logistic distribution displays asymmetric behavior. Additionally, the inclusion of the parameter α allows for modeling the probabilities in its tails. This parameter enhances the flexibility of the probability of success compared to the logit and probit functions, suggesting the potential for more precise adjustment of the probability of success for the variables under study.

Referring to $G(\cdot | x_1, \dots, x_p)$ as the CDF of the PHL, it follows that

$$\begin{aligned}\Pr(Y_i = 1 | x_1, \dots, x_p) &= p_i = G(Y_i = 1 | x_1, \dots, x_p) \\ &= 1 - \frac{1}{(1 + \exp(x_i^\top \beta))^\alpha} \\ &= 1 - \left[\frac{1}{2} - \frac{1}{2} \tanh\left(\frac{x_i^\top \beta}{2}\right) \right]^\alpha.\end{aligned}$$

From this, it is obtained that

$$\begin{aligned}\Pr(Y_i = 0 | x_1, \dots, x_p) &= 1 - \Pr(Y_i = 1 | x_1, \dots, x_p) \\ &= \frac{1}{(1 + \exp(x_i^\top \beta))^\alpha} \\ &= \left[\frac{1}{2} - \frac{1}{2} \tanh\left(\frac{x_i^\top \beta}{2}\right) \right]^\alpha.\end{aligned}$$

For $p_i = \Pr(Y_i = 1 | x_1, \dots, x_p)$, it follows that

$$\log\left(\frac{1 - (1 - p_i)^{1/\alpha}}{(1 - p_i)^{1/\alpha}}\right) = x_i^\top \beta, \quad i = 1, 2, \dots, n, \quad (12)$$

which will be referred to as the logit complement α -root transformation.

4.1. Properties of the PHL Regression Model

Given the structure of the probability function included in this new model, some statistics of interest are calculated for the interpretation of the parameters. Then, the odds $\text{odds}(x_1, x_2, x_3, \dots, x_p) = \text{odds}(x)$ are given by

$$\begin{aligned}\text{odds}(x_i) &= \frac{\Pr(Y_i = 1 | x_1, x_2, \dots, x_p)}{1 - \Pr(Y_i = 1 | x_1, x_2, \dots, x_p)} \\ &= \left(1 + \exp(x_i^\top \beta)\right)^\alpha - 1.\end{aligned}$$

Thus, the relative risk (RR) or odds ratio, to compare the profile of individuals i and k , is given by

$$RR(i, k) = \frac{\text{odds}(x_i)}{\text{odds}(x_k)} = \frac{(1 + \exp(x_i^\top \beta))^\alpha - 1}{(1 + \exp(x_k^\top \beta))^\alpha - 1}.$$

This expression is used when there are profiles of different individuals, or when the profiles only differ in the j th variable. Thus, to estimate the relative risk in the i th individual when the j th variable is increased by one unit, denoted as $x_j + 1$, while keeping the value of the rest of the variables constant, we have the expression

$$\frac{\text{odds}(x_1, \dots, x_{j-1}, x_j + 1, x_{j+1}, \dots, x_p)}{\text{odds}(x_1, \dots, x_{j-1}, x_j, x_{j+1}, \dots, x_p)} = \frac{(1 + \exp(\beta_j) \exp(x_i^\top \beta))^\alpha - 1}{(1 + \exp(x_i^\top \beta))^\alpha - 1}$$

This represents the odds or the number of times the risk of the event occurring increases (or decreases) when the variable x_j increases by one unit.

4.2. Maximum Likelihood Estimation

Given a random sample y_1, y_2, \dots, y_n of a random variable Y with distribution $Y_i \sim \text{Bin}(n, p_i)$, and considering a set of covariates x_1, x_2, \dots, x_p , the likelihood function is expressed as

$$\mathcal{L}_{\text{PHL}}(\beta, \alpha \mid \mathbf{X}, \mathbf{Y}) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}.$$

Then, the log-likelihood function is given by

$$\begin{aligned} \ell_{\text{PHL}}(\beta, \alpha \mid \mathbf{X}, \mathbf{Y}) &= \sum_{i=1}^n y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \\ &= \sum_{i=1}^n y_i \log\left((1 + \exp(x_i^\top \beta))^\alpha - 1\right) \\ &\quad - \alpha \sum_{i=1}^n \log(1 + \exp(x_i^\top \beta)) \end{aligned} \quad (13)$$

The score function, $U(\beta, \alpha) = (U(\beta), U(\alpha))$ with $U(\beta) = (U(\beta_0), U(\beta_1), U(\beta_2), \dots, U(\beta_p))$, which is calculated as the first derivative of the log-likelihood function concerning the parameters, is given by

$$\begin{aligned} U(\beta_j) &= \alpha \sum_{i=1}^n x_{ij} y_i \exp(x_i^\top \beta) \frac{(1 + \exp(x_i^\top \beta))^{\alpha-1}}{(1 + \exp(x_i^\top \beta))^\alpha - 1} \\ &\quad + \alpha \sum_{i=1}^n x_{ij} \frac{\exp(x_i^\top \beta)}{1 + \exp(x_i^\top \beta)} \end{aligned} \quad (14)$$

for $j = 0, 1, 2, \dots, p$, and

$$\begin{aligned} U(\alpha) &= \sum_{i=1}^n y_i \frac{(1 + \exp(x_i^\top \beta))^\alpha \log(1 + \exp(x_i^\top \beta))}{(1 + \exp(x_i^\top \beta))^\alpha - 1} \\ &\quad - \sum_{i=1}^n \log(1 + \exp(x_i^\top \beta)) \end{aligned} \quad (15)$$

The elements of the observed information matrix, $\kappa(\theta)$, defined as minus the Hessian matrix (matrix of second derivatives concerning the parameters), are given by:

$$\begin{aligned}\kappa_{\beta_j\beta_k} &= \alpha \sum_{i=1}^n x_{ij}x_{ik} \frac{\exp(x_i^\top \beta)}{(1 + \exp(x_i^\top \beta))^2} \left[1 + \frac{y_i}{p_i^2} \right. \\ &\quad \left. \left(-p_i(1 + \exp(x_i^\top \beta)) + \exp(x_i^\top \beta)(p_i + \alpha(1 - p_i)) \right) \right] \\ \kappa_{\beta_j\alpha} &= \sum_{i=1}^n x_{ij} \frac{\exp(x_i^\top \beta)}{1 + \exp(x_i^\top \beta)} \left[1 - \frac{y_i}{p_i^2} \left(\frac{p_i}{1 + \exp(x_i^\top \beta)} - \right. \right. \\ &\quad \left. \left. \alpha \log(1 + \exp(x_i^\top \beta))(1 - p_i) \right) \right] \\ \kappa_{\alpha\alpha} &= \sum_{i=1}^n \frac{1 - p_i}{p_i^2} \log^2(1 + \exp(x_i^\top \beta)).\end{aligned}$$

The elements of the information matrix, which are obtained from the expected value of the elements of the observed information matrix, $I(\theta) = \mathbb{E}(\kappa(\theta))$, are given by

$$\begin{aligned}i_{\beta_j\beta_k} &= \alpha \sum_{i=1}^n x_{ij}x_{ik} \frac{1 - p_i}{p_i} \left(\frac{\exp(x_i^\top \beta)}{1 + \exp(x_i^\top \beta)} \right)^2, \\ i_{\beta_j\alpha} &= \sum_{i=1}^n x_{ij} \left(\frac{\exp(x_i^\top \beta)}{1 + \exp(x_i^\top \beta)} \right)^2 - \\ &\quad \alpha \sum_{i=1}^n x_{ij} \frac{1 - p_i}{p_i} \frac{\exp(x_i^\top \beta)}{1 + \exp(x_i^\top \beta)} \\ &\quad \log(1 + \exp(x_i^\top \beta)) \\ i_{\alpha\alpha} &= \sum_{i=1}^n \frac{1 - p_i}{p_i^2} \log^2(1 + \exp(x_i^\top \beta)).\end{aligned}$$

When $\alpha = 1$, we obtain $p_i = \frac{\exp(x_i^\top \beta)}{1 + \exp(x_i^\top \beta)}$, and the information matrix can be written as

$$I_F(\theta) = \begin{pmatrix} \mathbf{X}^\top \mathbf{W} \mathbf{X} & \mathbf{X}^\top \mathbf{W} \\ \mathbf{M}^\top \mathbf{X} & \frac{1-p_i}{p_i^2} \log^2(1 + \exp(x_i^\top \beta)) \end{pmatrix}, \quad (16)$$

where \mathbf{W} is the diagonal matrix $\mathbf{W} = \text{diag}(p_i(1 - p_i))$, $i = 1, 2, \dots, n$, and \mathbf{M} is a vector with elements $m_i = p_i \left(p_i - \frac{1-p_i}{p_i} \log(1 + \exp(x_i^\top \beta)) \right)$.

Letting $d = \frac{1-p_i}{p_i^2} \log^2(1 + \exp(x_i^\top \beta))$, we obtain that the determinant of the information matrix is given by:

$$|I(\theta)| = d^{-p} \left| \mathbf{X}^\top \left(\mathbf{W} - \mathbf{M} \mathbf{M}^\top \right) \mathbf{X} \right| \neq 0.$$

Thus, the information matrix is non-singular, which guarantees the existence of the variance-covariance matrix of the vector of maximum likelihood estimators (MLE) $\hat{\theta}$. It can also be concluded that the variance-covariance matrix of the MLE can be written as:

$$\Sigma = I^{-1}(\hat{\theta}).$$

Therefore, for large sample sizes, we have

$$\hat{\theta} \xrightarrow{d} N_{p+2}(\theta, \Sigma),$$

meaning that the distribution of the vector of estimators is consistent and asymptotically normal, with a covariance matrix equal to the inverse of the Fisher information matrix.

Confidence intervals for coefficients θ_r of level $100(1 - \psi)\%$ can be obtained from the expression $\hat{\theta}_r \mp z_{1-\psi/2} \sqrt{\hat{\sigma}(\hat{\theta}_r)}$. Additionally, the adequacy of the proportional hazard logistic regression (PHLR) model can be evaluated through hypothesis testing:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0,$$

for at least one $j = 1, \dots, p$.

We can use the deviance function given by

$$G_p = -2(\ell(\beta_0, \alpha) - \ell(\hat{\beta}, \alpha)),$$

with distribution $G_p \sim \chi_p^2$. Similarly, two models can be compared: one complete with r variables (β_r), and another with q ($q < r$) variables (β_q) through the test statistic

$$G_{r-q} = -2(\ell(\hat{\beta}_q, \alpha) - \ell(\hat{\beta}_r, \alpha)),$$

for which we have $G_{r-q} \sim \chi_{r-q}^2$. This same statistic is useful to test the significance of the remaining $r - q$ variables that were not included in the model with q variables.

One of the strategies to validate the good fit of the logistic regression model is to analyze the proportion of correct classification that the fitted model achieves. Letting G_1 be the group of observations with $Y_i = 1$, and G_2 be the group of observations with $Y_i = 0$ then, using Bayes' Theorem, the probability of classifying an individual into group G_1 given the information of the explanatory variables x_1, x_2, \dots, x_p is given by

$$\Pr(G_1 | x) = \frac{p_1 \times \Pr(x | G_1)}{p_1 \times \Pr(x | G_1) + p_2 \times \Pr(x | G_2)}.$$

Thus, when performing the calculations for our model, we have

$$\Pr(G_1 | x) = 1 - \frac{p_2}{(p_2 - p_1) + p_1(1 + \exp(x_i^T \beta))^\alpha}.$$

Similarly, $\Pr(G_2 | x)$ is defined. In this case, the decision will be to classify the i th individual into G_1 if $\Pr(G_1 | x) > \Pr(G_2 | x)$; that is, if $\Pr(G_1 | x) > 0.5$.

To evaluate the predictive capacity of the proportional hazard logistic regression model, the overall accuracy of the model can be calculated, which is defined as the proportion of individuals that are correctly classified, as well as the sensitivity or true positive rate of the model (TPR), defined as the number of correctly classified individuals from group G_1 divided by the total number of correctly classified ones (from G_1 and G_2). Similarly, the false negative rate (FNR) of the model is defined as $(1 - TPR)$, among others.

5. Case Study: Students' Dropout Data

The data for this application consist of a sample of 413 students from the Department of Mathematics and Statistics of the University of Córdoba, which were obtained from the SPADIES System of the Ministry of National Education of Colombia (MNE). The response variable in this application takes the values $Y = 1$ (if program dropout) or $Y = 0$ (if non-dropout). The explanatory variables considered are x_1 = (cumulative general average, CGA), x_2 = character of the school (CS) of the student where they studied, taking values = 1 (if the student comes from an official school), and = 0 (if not), and x_3 = the number of periods enrolled (NPE). The logistic regression (LR) and proportional hazard logistic regression (PHLR) models were fitted. The results of the fitted models, obtained using the R Development Core Team [22] package, are given in Table 1.

Table 1. Parameter estimation of LR and PHLR models (standard errors of the estimates are given in parentheses).

Model	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\alpha}$	AIC	CAIC	BIC
LR se	7.290 (1.220)	−1.576 (0.384)	0.442 (0.423)	−0.270 (0.035)		310.4	330.5	326.5
PHLR se	1.374 (1.206)	−0.860 (0.340)	0.355 (0.293)	−0.201 (0.025)	11.93 (5.301)	295.9	325.0	316.0

The results of the model fit indicate that the variables CGA and NPE are significant, whereas the variable CS does not significantly explain the probability of university student dropout.

To compare the fitted models, we employ the Akaike Information Criterion (AIC) Akaike [23], corrected AIC (CAIC), and the Bayesian Information Criterion (BIC) by Hastie and Tibshirani [24], given by

$$AIC = -2 \times \hat{\ell}(\cdot) + 2p,$$

$$CAIC = -2 \times \hat{\ell}(\cdot) + 2p \left(1 + \frac{n+2}{n-p-2} \right)$$

and

$$BIC = -2 \times \hat{\ell}(\cdot) + p \log(n)$$

where p is the number of parameters in the model and n is the sample size. The results favor the PHLR model based on AIC, CAIC, and BIC values.

To compare the proportional hazard logistic regression (PHLR) model with the logistic regression model, we conduct the hypothesis test

$$H_0 : \alpha = 1 \quad \text{vs} \quad H_1 : \alpha \neq 1,$$

using the likelihood ratio statistic

$$\Lambda_1 = \frac{\mathcal{L}_L(\boldsymbol{\theta})}{\mathcal{L}_{\text{PHL}}(\boldsymbol{\theta}^*)},$$

where $\mathcal{L}_L(\cdot)$ and $\mathcal{L}_{\text{PHL}}(\cdot)$ represent the likelihood functions of the logistic and PHL models, respectively. Upon numerical evaluation, we obtain

$$-2 \log(\Lambda) = -2(-151.2 + 142.955) = 16.49,$$

which exceeds the value of $\chi^2_{1,95\%} = 3.84$. The PHL model exhibits the best fit compared to the logistic model.

Carrying out the hypothesis test of the significance of the explanatory variables

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0,$$

for at least one $j = 1, 2, 3$, we have

$$\begin{aligned} G_3^2 &= -2(-262.4158 + 142.9585) \\ &= 238.9147 > \chi^2_{0.05,3} = 7.8147, \end{aligned}$$

therefore, the null hypothesis is rejected. Similarly, for the hypothesis test

$$H_0 : \beta_2 = 0 \quad \text{vs} \quad H_1 : \beta_2 \neq 0,$$

it follows that

$$\begin{aligned} G_3^2 &= -2(\ell(\beta_0, \beta_1, \beta_3, \alpha) - \ell(\beta_0, \beta_1, \beta_2, \beta_3, \alpha)) \\ &= -2(-143.96 + 142.9585) = 2.003 < \chi_{0.05,1}^2 = 3.84, \end{aligned}$$

Therefore, the null hypothesis is not rejected, meaning the variable character of the school is not significant in the model. However, academic differences are observed in the classroom between students who come from official schools and those from private schools, with the latter demonstrating better preparation.

Note that in the proportional hazard logistic regression model, the case $\alpha = 1$ corresponds to the logistic distribution. However, the hypothesis test $H_0 : \alpha = 1$ vs. $H_1 : \alpha \neq 1$, which is performed using the likelihood ratio statistic, is rejected. This means that the parameter α is significantly different from one, and must be considered to explain the behavior of the data. Moreover, the AIC, CAIC, and BIC criteria are favorable to the logistic proportional hazard model when compared with the usual logistic regression model. All of the above allows us to conclude that the proportional hazard logistic regression model fits better.

So, the fitted model is given as follows:

$$P(Y = 1 \mid x_1, x_2, x_3) = \frac{(1 + e^{1.374 - 0.860x_1 + 0.355x_2 - 0.201x_3})^{11.93} - 1}{(1 + e^{1.374 - 0.860x_1 + 0.355x_2 - 0.201x_3})^{11.93} + 1}$$

Now, the sample is divided into two subsamples. The first one, called the training sample, corresponds to 70% of the total sample, and the second one is the prediction sample (30% of the sample). From this partition, the following results for the fitted PHLR model are obtained.

According to the results in Table 2, the accuracy is 77.23%, the sensitivity rate is 67.05%, and the specificity rate is 100%.

Table 2. Model predictive capacity.

Actual/Forecast	$\hat{y} = 1$	$\hat{y} = 0$	Total
$y = 1$	57	28	85
$y = 0$	0	38	38
Total	57	66	123

On the other hand, Table 3 shows the performance of the PHLR model for different values of the α parameter.

Table 3. Skewness and kurtosis of the PHLR model for different α values.

α	0.050	0.125	0.250	0.500	0.750	1.000	1.500
Skewness	0.355	0.032	0.160	0.135	0.058	0.000	−0.081
Kurtosis	1.673	2.159	2.716	2.974	2.988	3.000	3.031
α	2.500	5.000	10.000	20.000	30.000	50.000	100.000
Skewness	−0.179	−0.303	−0.410	−0.501	−0.546	−0.597	−0.655
Kurtosis	3.090	3.201	3.331	3.469	3.548	3.644	3.765

Table 3 shows the skewness and kurtosis coefficients of the proportional hazard logistic model for different α values. The results indicate that the model can fit data with both negative and positive skewness, which is an advantage over traditional logistic models. Additionally, the PHLR model can fit data with varying degrees of kurtosis, both high and low.

Diagnostic analysis is a technique to detect possible influential observations and aberrant or extraneous data. In the case of the logistic model, this technique has certain similarities with the general diagnostic analysis of regression models. However, given that the response variable only takes the values 0 and 1, a somewhat unusual situation arises. Certain difficulties may arise if there is a large number of zeros (or ones) when one expects to find few zeros or ones, which can be a sign of a lack of fit in the model. In the case of the PHLR model, the diagnostic analysis could be carried out using the Pearson residuals,

$$\tilde{r} = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}}$$

the square of which is the i th component of the Pearson chi-square statistic, the residual deviance

$$t_{D_i} = \text{sign}(\tilde{r}) \sqrt{-2 \left[y_i \log\left(\frac{y_i}{\hat{p}_i}\right) + (1 - y_i) \log\left(\frac{1 - y_i}{1 - \hat{p}_i}\right) \right]},$$

which is an adapted version of Cook's distance for the case of the logistic regression model (see Christensen [25]). When $y_i = 0$, $t_{D_i} = \text{sign}(\tilde{r}) \sqrt{-2 \log(1 - \hat{p}_i)}$, while if $y_i = 1$, $t_{D_i} = \text{sign}(\tilde{r}) \sqrt{-2 \log(\hat{p}_i)}$.

For the student dropout data, the residual deviance graph for the PHLR model is presented in Figure 2. Note that in this graph, there are no observations with high values of the residuals, which indicates that the model has a good fit. Likewise, the graph of the PHL distribution for the fitted probabilities is shown. Note that there are five values falling within the $+2.5/-2.5$ range in Figure 2b, and six values in Figure 2c, indicating that these observations are not extremely influential. Additionally, there are no observations outside the confidence bands in the envelope graphs (Figure 3b), suggesting that the PHLR model effectively handles observations that deviate slightly from the $+2/-2$ range.

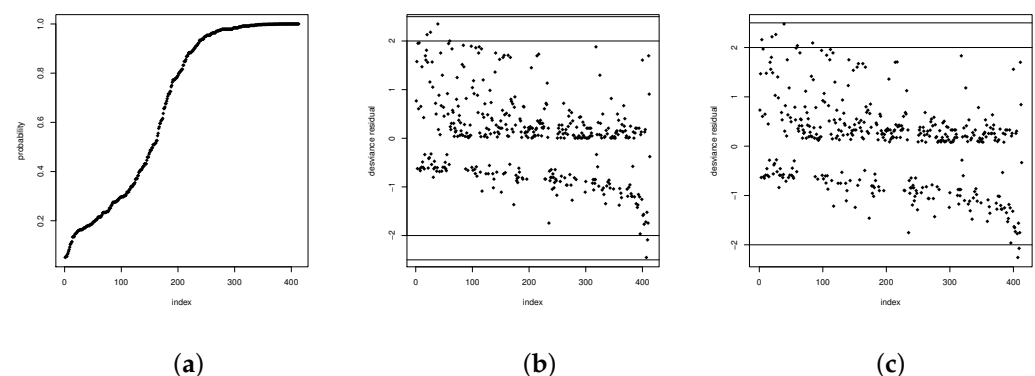


Figure 2. (a) Fitted PHLR model. (b) Residual deviance for the fitted PHLR. (c) Residual deviance for the LR model.

The rMT_i envelope graphs generated for the logistic and proportional hazard logistic models are presented in Figures 3a and 3b, respectively. It is observed that the proportional hazard logistic regression model presents a better fit than the logistic regression model.

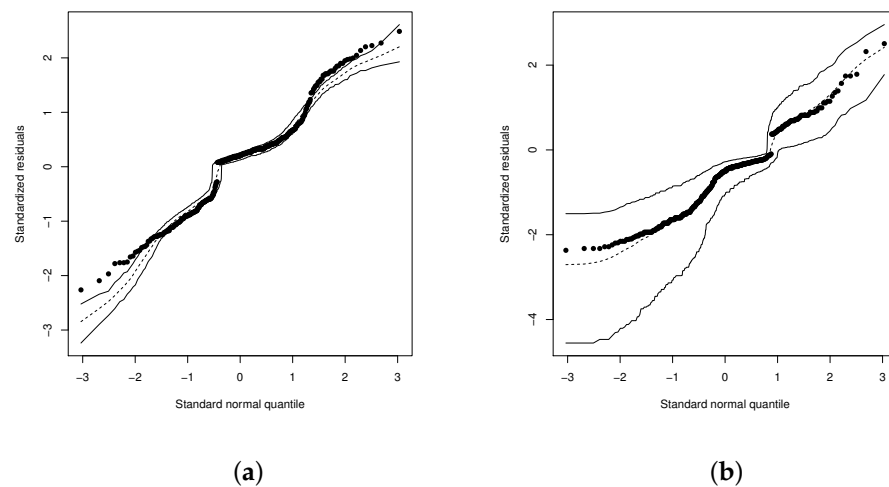


Figure 3. Envelope plots for rMT_i : (a) LR model and (b) PHLR model.

6. Conclusions

In this work, we have proposed the PHLR, a nonlinear regression model that captures complex relationships between independent variables and the response variable, particularly in the case of dichotomous data where the relationships cannot be adequately represented by a straight line. The flexibility of the PHLR model allows for a better fit to the data compared to linear models or even the logistic model.

The information matrix of the PHLR model is non-singular, ensuring that the parameters are uniquely estimable, avoiding linear dependency among them, and allowing for the proper calculation of the variance–covariance of the estimators. This guarantees the convergence of optimization and estimation algorithms, and ensures that the maximum likelihood estimators have desirable asymptotic properties, such as asymptotic normality.

In terms of information criteria such as AIC, CAIC, and BIC, the PHLR model shows a better fit than the logistic model for the analyzed student dropout data. The logistic model is revealed as a special case of the PHLR model. Additionally, the PHLR model demonstrates a good rate of correct classifications in the studied data. An alternative for the diagnostic analysis of model errors has also been proposed, offering useful tools for its implementation in educational problems or other contexts with dichotomous responses.

Author Contributions: Conceptualization, R.T.-F. and C.B.-C.; Methodology, G.M.-F. and R.T.-F.; Software, G.M.-F., R.T.-F. and C.B.-C.; Validation, G.M.-F., R.T.-F. and C.B.-C.; Formal analysis, G.M.-F., R.T.-F. and C.B.-C.; Investigation, G.M.-F., R.T.-F. and C.B.-C.; Resources, G.M.-F. and R.T.-F.; Data curation, G.M.-F., R.T.-F. and C.B.-C.; Writing—original draft, R.T.-F. and C.B.-C.; Writing—review & editing, G.M.-F., R.T.-F. and C.B.-C.; Visualization, G.M.-F., R.T.-F. and C.B.-C.; Supervision, G.M.-F., R.T.-F. and C.B.-C.; Project administration, G.M.-F. and R.T.-F.; Funding acquisition, G.M.-F. and R.T.-F. All authors have read and agreed to the published version of the manuscript.

Funding: The research of G. Martínez-Flórez and R. Tovar-Falón was supported by the project: Estudio de la deserción en los programas de pregrado de la Universidad de Córdoba usando diferentes metodologías estadísticas, FCB-06-22. Universidad de Córdoba, Colombia.

Data Availability Statement: Details about data available are given in Section 6.

Acknowledgments: Martínez-Flórez and R. Tovar-Falón acknowledge the support given by Universidad de Córdoba, Montería, Colombia. C. Barrera-Causil extends their sincere gratitude to the Instituto Tecnológico Metropolitano (ITM).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Eugene, N.; Lee, C.; Famoye, F. Beta-normal Distribution and Its Applications. *Commun. Stat.-Theory Methods* **2002**, *31*, 497–512. [CrossRef]
2. Silva, G.O.; Ortega, E.M.M.; Cordeiro, G.M. The Beta Modified Weibull Distribution. *Lifetime Data Anal.* **2010**, *16*, 409–430. [CrossRef]
3. Cordeiro, G.M.; Silva, G.O.; Ortega, E.M.M. The Beta-Weibull Geometric Distribution. *Statistics* **2013**, *47*, 817–834. [CrossRef]
4. Cordeiro, G.M.; de Castro, M. A New Family of Generalized Distributions. *J. Stat. Comput. Simul.* **2011**, *81*, 883–898. [CrossRef]
5. 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]
6. Zografos, K.; Balakrishnan, N. On Families of Beta- and Generalized Gamma generated Distributions and Associated Inference. *Stat. Methodol.* **2009**, *6*, 344–362. [CrossRef]
7. Ristić, M.M.; Balakrishnan, N. The Gamma-exponentiated Exponential Distribution. *J. Stat. Comput. Simul.* **2012**, *82*, 1191–1206. [CrossRef]
8. 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. [CrossRef]
9. 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]
10. Martínez-Flórez, G.; Bolfarine, H.; Gómez, H.W. Skew-normal alpha power model. *Statistics* **2014**, *48*, 1414–1428. [CrossRef]
11. Martínez-Flórez, G.; Moreno-Arenas, G.; Vergara-Cardozo, S. Properties and inference for proportional hazard models. *Rev. Colomb. Estadística* **2013**, *36*, 95–114.
12. Azzalini, A. A class of distributions which includes the normal ones. *Scand. J. Stat.* **1985**, *12*, 171–178.
13. Martínez-Flórez, G.; Bolfarine, H.; Gómez, H.W. Likelihood-based inference for the power regression model. *SORT-Stat. Oper. Res. Trans.* **2015**, *39*, 187–208.
14. Birnbaum, Z.W.; Saunders, S.C. A New Family of Life Distributions. *J. Appl. Probab.* **1969**, *6*, 319–327. [CrossRef]
15. Rieck, J.R.; Nedelman, J.R. A log-linear model for the Birnbaum–Saunders distribution. *Technometrics* **1991**, *33*, 51–60.
16. Agresti, A. *Categorical Data Analysis*; John Wiley & Sons Inc.: Hoboken, NJ, USA, 2002.
17. Henze, N. A probabilistic representation of the skew-normal distribution. *Scand. J. Stat.* **1986**, *13*, 271–275.
18. Pewsey, A. Problems of inference for Azzalini’s skew-normal distribution. *J. Appl. Stat.* **2000**, *27*, 859–870. [CrossRef]
19. Chiogna, M. Notes on estimation problems with scalar skew-normal distributions. *Stat. Methods Appl.* **2005**, *14*, 331–341. [CrossRef]
20. 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]
21. Lehmann, E.L. The power of rank tests. *Ann. Math. Stat.* **1953**, *24*, 23–43. [CrossRef]
22. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022. Available online: <http://www.R-project.org> (accessed on 6 March 2024).
23. Akaike, H. A new look at statistical model identification. *IEEE Trans. Autom. Contr.* **1974**, *19*, 716–722. [CrossRef]
24. Hastie, T.J.; Tibshirani, R.J. *Generalized Additive Models*, 1st ed.; Chapman and Hall/CRC: New York, NY, USA, 1990.
25. Christensen, R. *Log-Linear Models and Logistic Regression*; Springer: New York, NY, USA, 1997.

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.