Revista Colombiana de Estadística
Junio 2014, volumen 37, no. 1, pp. 181 a 196

Asymmetric Regression Models Bernoulli/Log Proportional-Hazard Distribution

Modelos de regresión asimétrico Bernoulli/distribución Log Hazard proporcional

Guillermo Martínez-Flórez[1,a], Carlos Barrera[2,b]

[1]Departamento de Matemáticas y Estadística, Universidad de Córdoba, Monteríia, Colombia

[2]Facultad de Ciencias Exactas y Aplicadas, Instituto Tecnológico Metropolitano, Medellín, Colombia

Abstract

In this paper we introduce a kind of asymmetric distribution for non-negative data called log-proportional hazard distribution (LPHF). This new distribution is used to study an asymmetrical regression model for data with limited responses (censored) through the mixture of a Bernoulli distribution with logit link and the LPHF distribution. Properties of the LPHF distribution are studied, maximum likelihood parameter estimation and information matrices are addressed. An illustration with real data shows that the model is a new alternative for studies with positive data censored.

***Key words*:** Censoring, Fisher information matrix, Maximum likelihood estimators, Proportional hazard.

Resumen

En este artículo se introduce una forma de distribución asimétrica para datos no-negativos llamada distribución log hazard proporcional (LPHF). Esta nueva distribución es usada para estudiar un modelo de regresión asimétrico para datos con respuestas limitadas (censuradas) a través de mezclas de una distribución Bernoulli con función link logit y la distribución LPHF. Propiedades de la distribución LPHF son estudiadas, se abordan las estimaciones de máxima verosimilitud de los parámetros y las matrices de información. Se presenta una ilustración con datos reales, donde se muestra que el modelo propuesto es una nueva alternativa para estudios con datos positivos censurados.

***Palabras clave*:** censura, estimadores de máxima verosimilitud, hazard proporcional, matriz de información de Fisher.

[a]Professor. E-mail: gmartinez@correo.unicordoba.edu.co
[b]Associate professor. E-mail: carlosbarrera@itm.edu.co

1. Introduction

The fundamental law of geochemistry enunciated by Ahrens (1954), "the concentration of a chemical element in a rock is distributed log-normal", is an application of the log-normal distribution. This distribution is also widely used to model different types of information, including income in the economy and lifetime distributions from materials, among others.

In many of these situations, both the kurtosis and the asymmetry of the distribution are above or below the expected for the log-normal model, reason why it is necessary to think in a more flexible model that achieves such deviation in modeling positive data.

In the case of positive data, Azzalini, dal Cappello & Kotz (2003), Mateu-Figueras & Pawlosky-Glanh (2003) and Mateu-Figueras, Pawlosky-Glanh & Barcelo-Vidal (2004) introduce the univariate distribution log-skew-normal (LSN), which contains as special case, the log-normal model.

Its density function is given by:

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

where $\xi \in \mathbb{R}$, is a location parameter, $\eta \in \mathbb{R}^+$, is a scale parameter, λ is an asymmetry parameter, $\phi(\cdot)$ is the density function of a standard normal distribution and $\Phi(\cdot)$ is the respective cumulative distribution function. Notice that if $\lambda = 0$ then the ordinary log-normal distribution follows as it is the case with the ordinary skew-normal model. Also, the information matrix is singular, thus regularity conditions are no longer satisfied. One consequence of this fact is that likelihood ratio statistics is no longer distributed according to the central chi-square distribution (Arellano-Valle & Azzalini 2008).

Moreover, in many cases the asymmetrical positive random variable in the study is limited, and in turn this is explained by a set of auxiliary covariates $X_1,\ X_2,\ ..., X_p$, thus extensions to the censured case with covariates should be addressed. The study of random variables with limited responders with covariates was presented by Tobin (1958) who studied the model popularly known as Tobit.

This model has been extensively studied in the case of normally distributed errors and is defined by considering that the observed random variable $y_i = \max\{y_i^*, 0\}$ with $y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i,\ i = 1, 2, \ldots, n$; where the error term $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$, $\mathbf{x}_i$ is a $p \times 1$ vector of known independent variables and $\boldsymbol{\beta}$ is a $p \times 1$ unknown parameter vector.

Although the Tobit model is an alternative for censoring data; in some situations the proportion of censored data cannot be well explained by the normal model since the tail of this distribution is more or less heavier than the proportion of censored data.

For instance, Moulton & Halsey (1995) show an application with 330 children in Haiti during 1987-1990 (*see* Job, Halsey, Boulos, Holt, Farrell, Albrecht, Brutus, Adrien, Andre, Chan, Kissinger, Boulos & the CiteSoleil/JHU 1991) which examines the; Immunogenecity of children before the implementation of a vaccine, here

the number of observations below the detection limit was 86 observations (26.1%), which exceeds the expected value with a normal model, whereby the proportion of censure cannot be explained by the Tobit model. These authors use asymmetric models such as log-normal model.

In this sense, other works have been published, for instance, other distributions: such as the log-skew-normal has been implemented by Chai & Bailey (2008) and recently, Martínez-Florez, Bolfarine & Gómez (2013) implemented the model log-alpha-power-normal.

The model proposed by Moulton & Halsey (1995) is a generalization of the Cragg (1971) model, that in the classical literature is known as *the two-part model,* which is an alternative to Tobit when the data rate below or above the threshold is quite different from the probability of the tail obtained with the normal model.

The probability density function of y_i under Cragg (1971) model can be expressed as

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

where p_i is the probability determining the relative contribution made by the point distribution to the overall mixture distribution, f is a density function with positive support, and $I_i = 0$ if $y_i > T$ and $I_i = 1$ if $y_i \leq T$.

Given the nature of the random variables involved in the Cragg (1971) model, different processes determine the respective components of the model.

A positive response necessarily comes from f, on the other hand, a T value comes from the point mass distribution. This model, however, does not consider the situation of a lower limit and that part of the observations may be below this lower limit.

If allowed to some limiting responses are the result of interval censored to f, we have the generalization of the two-part model exposed by Moulton & Halsey (1995). This means that an observed T value can be either a realization from the point-mass distribution or a partial observation from f with critical value not precisely known but lying somewhat in $(0, T)$ for a small pre-specified constant T. Formally,

$$g(y_i) = [p_i + (1 - p_i)F(T)]I_i + (1 - p_i)f(y_i)(1 - I_i)$$

where F is the cumulative distribution corresponding to f. If we vary the basic density f and the link function corresponding to p_i, we can generate a large family of mixed models. Models such as probit/trucated-normal, logit/lognormal, logit/log-gamma, probit/log-skew-normal and logit/log-alpha-power normal have been considered in practical applications in biology, economy, agricultural and so on (Chai & Bailey 2008, Martínez-Florez, Bolfarine & Gómez 2013). Notice that for $p_i = 0$, $i = 1, \ldots, n$, Moulton & Halsey (1995) model reduces to the Tobit model (Tobin 1958).

This is an extension of the log-normal distribution allowing for one extra parameter which will be presented in the next section.

2. Proportional Hazard Distribution

Recently, Martínez-Florez, Moreno-Arenas & Vergara-Cardozo (2013) introduced a new asymmetric model which is called proportional hazard model, this model is defined as follows:

Let F be a continuous cumulative distribution function with probability density function f, and hazard function $h = f/(1-F)$. We say that Z has a proportional hazard distribution associated with F, f and the parameter $\alpha > 0$ if its probability density function is

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

where α is a positive real number. We use the notation $Z \sim PHF(\alpha)$. The distribution function of the PHF model is given by

$$\mathbb{F}(z) = 1 - \{1-F(z)\}^{\alpha}, \ z \in \mathbb{R}.$$

This is why this type of distribution can also be regarded as an exponentiated distribution or a fractional order statistic distribution, widely studied in the literature.

If Z is a random variable from a standard $PHF(\alpha)$ distribution then the location-scale extension of Z is obtained from the transformation $X = \xi + \eta Z$, where $\xi \in R$ and $\eta \in R^+$, is a scale parameter.

In the particular case where $F = \Phi(\cdot)$, we have the family of distributions called proportional hazard normal (PHN) and denoted $PHN(\xi, \eta, \alpha)$.

In Martínez-Florez, Moreno-Arenas & Vergara-Cardozo (2013), we can see the behavior of the $PHN(0, 1, \alpha)$ density and the model hazard function for some values of the α parameter.

2.1. Log Proportional-Hazard Distribution

Let Y be a random variable with support in R^+, we say that Y follows a univariate log-proportional-hazard distribution with parameter α, if the transformed variable $X = \log(Y) \sim PHF(\alpha)$. We denote $Y \sim LPHF(\alpha)$.

Then, the pdf for the random variable Y can be written as

$$\varphi_{LF}(y;\alpha) = \frac{\alpha}{y} f(\log(y)) \left\{1 - F(\log(y))\right\}^{\alpha-1}, \quad y \in R^+$$

where F is an absolutely continuous distribution function with density function $f = dF$. This model is called standard log proportional-hazard distribution.

Let $X \sim PHF(\xi, \eta, \alpha)$, where $\xi \in \mathbb{R}$ is a location parameter and $\eta \in \mathbb{R}^+$ is a scale parameter. Hence, the transformation $X = \ln(Y)$ leads to the location-scale log proportional-hazard model, with pdf given by

$$\varphi_{LF}(y;\xi,\eta,\alpha) = \frac{\alpha}{\eta y} f\left(\frac{\log(y)-\xi}{\eta}\right) \left\{1 - F\left(\frac{\log(y)-\xi}{\eta}\right)\right\}^{\alpha-1}, \quad y \in \mathbb{R}^+$$

We use the notation $Y \sim LPHF(\xi, \eta, \alpha)$, so that $LPHN(\alpha) = LPHN(0, 1, \alpha)$.

Its cumulative distribution function can be written as

$$\mathcal{F}_F(y; \alpha) = 1 - \{1 - F(\log(y))\}^{\alpha}, \quad y \in \mathbb{R}^+. \tag{1}$$

According to (1), the inversion method can be used for generating from a random variable with distribution $LPHF(\xi, \eta, \alpha)$. That is, if $U \sim U(0, 1)$, then, random variable $Y = e^{\xi + \eta F^{-1}(1-(1-U)^{1/\alpha})}$ is distributed according to the LPHF distribution with vector of parameters $\boldsymbol{\theta} = (\xi, \eta, \alpha)'$.

In the special case where $f = \phi(\cdot)$ and $F = \Phi(\cdot)$, the density and distribution functions of the standard normal distribution, respectively, we have the standard log proportional-hazard-normal distribution.

We will denote this extension by using the notation $Y \sim LPHN(\xi, \eta, \alpha)$.

Figure 1 shows the pdf's for the LPHN distribution for α equals 0.75, 1, 2 and 3. Is clearly seen that the shape of the distribution is affected when changes the value of α. For the log-normal case, when $\alpha = 1$, the kurtosis is smaller than when $\alpha = 2$ and, similarly, for the log-skew case, when $\alpha = 3$. Furthermore, when $\alpha = 0.75$ the kurtosis for the log-normal is greater. Asymmetry is always positive and also controlled by parameter α. Hence, α controls asymmetry as well as kurtosis for the LPHN distribution.

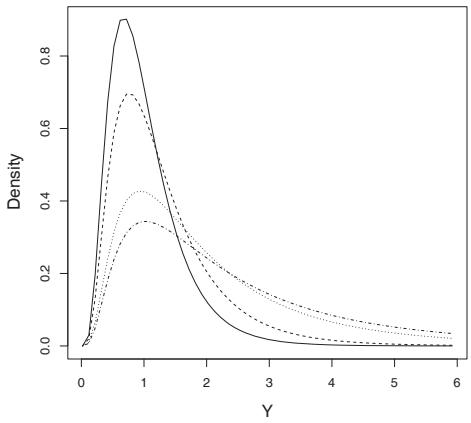

FIGURE 1: Plots of pdf $\varphi_{L\Phi}(y; 0.5, 0.75, \alpha)$, for α equals to 3 (solid line), 2 (dashed line), 1 (dotted line) and 0.75 (dashed and dotted line).

The r-th moment for the random variable $Y \sim LPHN$ is calculated numerically. Using the results of the central moments μ_r', the coefficients of variation, asymmetry and kurtosis are obtained.

Figure 2 shows the behavior of the mean and the coefficients of asymmetry and kurtosis of the LPHN model.

The survival and hazard functions for the LPHN model are, respectively, given by

$$S(t) = \{1 - \Phi(\log(t))\}^{\alpha} \quad \text{and} \quad r(t) = \frac{\alpha}{t} \frac{\phi(\log(t))}{1 - \Phi(\log(t))} = \alpha r_{LN}(t)$$

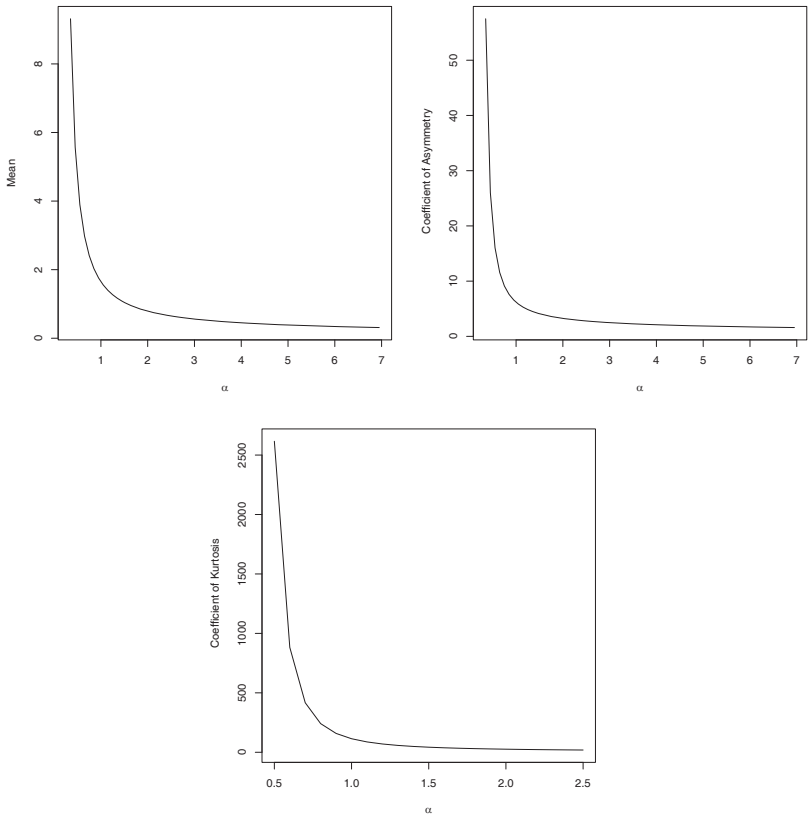

FIGURE 2: Behavior of some characteristic values of the LPHN model. (a) mean, (b) asymmetry coefficient and (c) coefficient of kurtosis.

where $r_{LN}(\cdot)$ is the hazard function of the log-normal distribution. Then, the hazard index T is proportional to the hazard index of the log-normal distribution.

2.2. Inference for Log Proportional-Hazard-Normal Model

For a random sample of size n, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$ with $Y_i \sim LPHN(\xi, \eta, \alpha)$, the log-likelihood function of $\boldsymbol{\theta} = (\xi, \eta, \alpha)'$ given $\mathbf{Y}$ is

$$\ell(\boldsymbol{\theta}; \mathbf{Y}) = n\log(\alpha) - n\log(\eta) - \sum_{i=1}^{n}\log(y) + \sum_{i=1}^{n}\log(\phi(z_i)) + (\alpha-1)\sum_{i=1}^{n}\log(1-\Phi(z_i)),$$

where $z_i = \frac{\log(y_i)-\xi}{\eta}$. The corresponding score equations are similar to the obtained in Martínez-Florez, Moreno-Arenas & Vergara-Cardozo (2013), we only need to consider the change in the log-likelihood function and obtain the MLE estimators using numerical methods.

The observed information matrix for location-scale PHN follows from minus the second derivatives of the log-likelihood function. This result is similar to that obtained by Martínez-Florez, Moreno-Arenas & Vergara-Cardozo (2013) but with minor changes due to the difference in the log-likelihood function.

2.2.1. Expected Information Matrix for the Location-Scale PHN

Considering $a_{kj} = \mathbb{E}\{z_i^k w_i^j\}$, where $w_i = \frac{\phi(z_i)}{1-\Phi(z_i)}$, the expected information matrix entries are:

$$I_{\xi\xi} = \frac{1}{\eta^2}\left[1 + (\alpha-1)(a_{02} - a_{11})\right] \qquad I_{\eta\xi} = \frac{2}{\eta^2}a_{10} + \frac{\alpha-1}{\eta^2}\left[a_{01} - a_{02} + a_{12}\right]$$

$$I_{\eta\eta} = -\frac{1}{\eta^2} + \frac{3}{\eta^2}a_{20} + \frac{\alpha-1}{\eta^2}\left[a_{22} + 2a_{11} - a_{31}\right]$$

$$I_{\alpha\xi} = -\frac{1}{\eta}a_{01} \qquad I_{\alpha\eta} = -\frac{1}{\eta}a_{11} \qquad I_{\alpha\alpha} = \frac{1}{\alpha^2}$$

The expected values of the above variables are generally calculated using numerical integration. When $\alpha = 1$, $\varphi_{L\Phi}(x;\xi,\eta,1) = \frac{1}{\eta y}\phi\left(\frac{\log(y)-\xi}{\eta}\right)$, the location-scale log-normal density function. Thus, the information matrix becomes

$$I(\boldsymbol{\theta}) = \begin{pmatrix} 1/\eta^2 & 0 & -a_{01}/\eta \\ 0 & 2/\eta^2 & -a_{11}/\eta \\ -a_{01}/\eta & -a_{11}/\eta & 1 \end{pmatrix}$$

Numerical integration shows that the determinant is $|I(\boldsymbol{\theta})| = \frac{1}{\eta^4}[2 - a_{11}^2 - 2a_{01}^2] \neq 0$, so in the case of a log-normal distribution the model's information matrix is nonsingular. The upper left 2×2 submatrix is the log-normal distribution's information matrix.

For large n and under regularity conditions we have

$$\hat{\boldsymbol{\theta}} \xrightarrow{A} N_3(\boldsymbol{\theta}, I(\boldsymbol{\theta})^{-1})$$

and the conclusion follows that $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically approaches the normal distribution with $I(\boldsymbol{\theta})^{-1}$ as covariance matrix, for large samples.

This result shows that the information matrix for the LPHN model is nonsingular and therefore the inference for large samples can be made, contrary to the log-skew-normal model, whose information matrix is singular when $\lambda = 0$, that consequently resulting likelihood ratio statistic is not distributed as a chi-square.

Note that as in the LPHN model, the information matrix of the log-skew-normal model has the same structure or shape that the location-scale skew-normal model, $SN(\xi,\eta,\lambda)$, where now $Z = (\log(y) - \xi)/\eta$. Is well known and was demonstrated by Azzalini (1985), that the information matrix of the skew-normal model is singular when its parameter of asymmetry $\lambda = 0$.

3. Asymmetric Regression Model Logit/LPHN

We now extend the LPHN model to the case of random variables with a limit of detection and the presence of covariates. Specifically, we consider the case of models with limited response and excess zeros in the response variable. Considering extensions of the generalized the two-part model of Moulton & Halsey (1995) to the situations logit/log proportional hazard-normal model, jointly with covariates at each step of the model. Initially, we develop the case of censored random variables LPHN. Thus, calling p_0 the proportion of observations at or below threshold point T, the censored model $LPHN(\xi, \eta, \alpha)$ is represented by the probability density function

$$g(y_i) = \begin{cases} p_{0i} + (1-p_{0i})\left[1 - \left\{1 - \Phi\left(\frac{\log(T)-\xi}{\eta}\right)\right\}^{\alpha}\right], & \text{if } y_i \leq T \\ (1-p_{0i})\frac{\alpha}{\eta y}\phi\left(\frac{\log(y_i)-\xi}{\eta}\right)\left\{1 - \Phi\left(\frac{\log(y_i)-\xi}{\eta}\right)\right\}^{\alpha-1}, & \text{if } y_i > T \end{cases}$$

Now we extend this model to the case of presence of covariates in limited response and when the response is not limited.

The above model can be extended to the situation where only a proportion $100p_0\%$ of censored observations comes from the censored LPHN, with the remaining $100(1-p_0)\%$ of the observations coming from the population of low responders, located below or at the point T.

Modeling this mixture as the outcome of a Bernoulli random variable D with

$$pr(D=1) = 1 - p_0$$

while for $D = 0$, $Y \leq T$ with probability one. The contribution of y_i to the likelihood conditioning on $D = 1$ when Y is assumed to follow a LPHN model can be written as

$$\left[1 - (1-p_0)\left\{1 - \Phi\left(\frac{\log(T)-\xi}{\eta}\right)\right\}^{\alpha}\right]^{I_i} \left[\frac{(1-p_0)\alpha}{\eta y_i}\phi\left(\frac{\log(y_i)-\xi}{\eta}\right)\left\{1 - \Phi\left(\frac{\log(y_i)-\xi}{\eta}\right)\right\}^{\alpha-1}\right]^{1-I_i}$$

Then, assuming that the response $y_i = T$ is explained by the set of explanatory variables $X_{11}, X_{12}, \ldots, X_{1p}$, then we model this mixture as the outcome of a Bernoulli random variable with logit link function with

$$p_{0i} = prob(y_i = T) = \frac{\exp\left(x'_{(1)i}\beta_{(1)}\right)}{1 + \exp\left(x'_{(1)i}\beta_{(1)}\right)}$$

and

$$1 - p_{0i} = \frac{1}{1 + \exp\left(x'_{(1)i}\beta_{(1)}\right)}$$

where $x_{(1)i} = (1, x_{1i1}, \ldots, x_{1ip})'$, is a covariate vector of dimension $p+1$ associated with the parameter vectors $\boldsymbol{\beta}_{(1)} = (\beta_{10}, \beta_{11}, \ldots, \beta_{1p})'$.

Taking into account the LPHN model, we have a covariate vector $x_{(2)} = (1, X_{21}, X_{22}, \ldots, X_{2r})'$ of dimension r, possibly different from $x_{(1)}$ and parameter vector $\boldsymbol{\beta}_{(2)} = (\beta_{20}, \beta_{21}, \ldots, \beta_{2r})'$, for which

$$\log(y_i) \sim PHN(x'_{(2)i}\boldsymbol{\beta}_{(2)}, \eta, \alpha), \quad y_i > T$$

where $x_{(2)i} = (1, x_{2i1}, \ldots, x_{2ir})'$.

This mixture of distributions we will call "linear logistic regression model" with proportional hazard-normal distribution and will be denoted by

$$RLLPHN(\beta_{(1)}, \beta_{(2)}, \eta, \alpha)$$

The logarithm of the likelihood function for $\boldsymbol{\theta} = (\boldsymbol{\beta}'_{(1)}, \boldsymbol{\beta}'_{(2)}, \eta, \alpha)'$ given $X_{(1)}$, $X_{(2)}$ and Y, is given by

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_i I_i \log\left[1 + \exp(x'_{(1)i}\boldsymbol{\beta}_{(1)})\left[1 - \{1 - \Phi(z_{Ti})\}^{\alpha}\right]\right] \\ &\quad - \sum_{i=1}^{n} \log\left[1 + \exp(x'_{(1)i}\boldsymbol{\beta}_{(1)})\right] \\ &+ \sum_i (1 - I_i)\left\{\log(\alpha) - \log(\eta y_i) + x'_{(1)i}\boldsymbol{\beta}_{(1)} + \log(\phi(z_i)) + (\alpha - 1)\log(1 - \Phi(z_i))\right\}\end{aligned}$$

where $z_{Ti} = \frac{\log(T) - x'_{(2)i}\boldsymbol{\beta}_{(2)}}{\eta}$ and $z_i = \frac{\log(y_i) - x'_{(2)i}\boldsymbol{\beta}_{(2)}}{\eta}$.

We denote by $\sum_0$ the sum over censored observations and $\sum_1$ the sum over noncensored observations. The score function corresponding to the log-likelihood function is given by (for $j = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, r$)

$$\begin{aligned}U(\beta_{(1)j}) &= \sum_0 \frac{x_{1ij}\exp(x'_{(1)i}\beta_{(1)})\left[1 - \{1 - \Phi(z_{T_i})\}^{\alpha}\right]}{1 + \exp(x'_{(1)i}\beta_{(1)})\left[1 - \{1 - \Phi(z_{T_i})\}^{\alpha}\right]} \\ &\quad - \sum_{i=1}^{n} \frac{x_{1ij}\exp(x'_{(1)i}\beta_{(1)})}{1 + \exp(x'_{(1)i}\beta_{(1)})} + \sum_1 x_{1ij}\end{aligned}$$

$$U(\beta_{(2)k}) = -\sum_{0} \frac{x_{2ik}\exp(x'_{(1)i}\beta_{(1)})\varphi_{L\Phi}(T, x'_{(2)i}\beta_{(2)}, \eta, \alpha)}{1+\exp(x'_{(1)i}\beta_{(1)})\left[1-\{1-\Phi(z_{T_i})\}^{\alpha}\right]}$$

$$-\frac{1}{\eta}\sum_{1} x_{2ik}\left[-z_i-(\alpha-1)\frac{\phi(z_i)}{1-\Phi(z_i)}\right]$$

$$U(\eta) = -\sum_{0} \frac{z_{T_i}\exp(x'_{(1)i}\beta_{(1)})\varphi_{L\Phi}(T, x'_{(2)i}\beta_{(2)}, \eta, \alpha)}{1+\exp(x'_{(1)i}\beta_{(1)})\left[1-\{1-\Phi(z_{T_i})\}^{\alpha}\right]}$$

$$-\frac{1}{\eta}\sum_{1}\left[1-z_i^2-(\alpha-1)z_i\frac{\phi(z_i)}{1-\Phi(z_i)}\right]$$

$$U(\alpha) = -\sum_{0} \frac{\exp(x'_{(1)i}\beta_{(1)})\{1-\Phi(z_{T_i})\}^{\alpha}\log(1-\Phi(z_{T_i}))}{1+\exp(x'_{(1)i}\beta_{(1)})\left[1-\{1-\Phi(z_{T_i})\}^{\alpha}\right]}$$

$$+\sum_{1}\left[\frac{1}{\alpha}+\log(1-\Phi(z_i))\right]$$

The system of equations obtained by equating the score to zero has no solution in closed form, and tends to be solved using iterative numerical methods.

The resulting equations require numerical procedures such as the Newton-Raphson or quasi-Newton method. These optimization algorithms can be found in the packages *maxLik* or *optimx* of the R software.

The observed information matrix is given by $J(\boldsymbol{\theta}) = -H(\boldsymbol{\theta}) = -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$, where $H(\boldsymbol{\theta})$ is the hessian matrix, which is obtained in the Appendix for the vector of parameters $\boldsymbol{\theta}$. In addition can be obtained information matrix defined as less n^{-1} times the expected value of the observed information matrix.

4. Numerical Illustration

The application of the logit/LPHN model, is carried out using the data described by Moulton & Halsey (1995) in a study of measles vaccines conducted in Haiti during 1987-1990. The detection limit was 0.1 international units (UI), or $\log(0.1) = -2.306$ in the natural log-scale. The codification for the covariates involved in the study were $X_1 = EZ$ (vaccine type; 0:Schwarz , 1:Edmonston-Zagreb); $X_2 = HI$ (vaccine dose; 0:medium, 1:high) and $X_3 = FEM$ (gender; 0:male, 1:female).

Such as Moulton & Halsey (1995), the aim in the present analysis is to study the immunogenicity differential between boys and girls using the logit/ log-proportional-hazard-normal (logit/LPHN) model.

4.1. Models

A variety of models can be adjusted given the covariates in the study. We adjust some of these models were carefully chosen from the cases studied by Moulton & Halsey (1995).

Model 1: Covariates and censored data in limited response, without censored data and covariates in the point-mass distribution located at zero;

Model 2: Censored data and covariates in limited response, without covariates in the point-mass distribution located at zero;

Model 3: Censored data, covariates in limited response and in the point-mass distribution located at zero;

Model 4: Censored data, covariates in limited response and in the point-mass distribution located at zero, a particular model.

The summary statistics we have $\overline{\log(y)} = -0.1793$, $s^2 = 1.1055$, $\sqrt{b_1} = 0.7521$ and $b_2 = 2.6286$ where the quantities $\sqrt{b_1}$ and b_2 correspond to the sample coefficients of asymmetry and kurtosis for values above 0.1. The high asymmetry degree indicated by the sample coefficient of asymmetry ($\sqrt{b_1}$) reveals that it seems worthwhile trying to fit an asymmetric model for this data set.

Moulton & Halsey (1995), and Moulton & Halsey (1996) modeled this data using the hybrids logit/log-normal (logit/LN) and logit/log-gamma (logit/LGM) models.

As a first attempt, we fitted the ordinary Tobit model with covariates (model 1), which resulted in a poor fit to the data set under study. Here $\hat{\beta}_{(2)0} = 0.565$, $\hat{\beta}_{(2)1} = 0.248$, $\hat{\beta}_{(2)2} = -0.191$ and $\hat{\beta}_{(2)3} = 0.262$, and $AIC = 1291.81$.

We adjust the mixtures logit/LN and logit/LGM, for 1-4 models, finding in both cases the model 4 presents the best fit. The estimates for these models are given in the Table 1. Note that δ is the shape parameter of the LGM model.

Table 1: Parameter estimation (standard error) and model fitting for one and two components hybrid Bernoulli/log-distributions.

		Bernoulli component			Log-distributions components		
density	AIC	INT	EZ	HI	δ	INT	FEM
LN	986.19	0.652	0.808	0.422		−0.401	0.264
		(0.220)	(0.304)	(0.288)		(0.112)	(0.155)
LGM	1022.43	0.572	0.656	0.374	−2.833	−1.179	0.053
		(0.201)	(0.261)	(0.255)	(0.510)	(0.088)	(0.056)

Hence, there is a clear indication that the conditions under which the Tobit model is adequate, are not satisfied for the measles vaccine data set.

Estimates (MLEs) for the model parameters 1-4, were obtained, and the results are shown in Tables 2.

To compare model fit, we computed the AIC criterion (Akaike 1974).

Table 2: Parameter estimation (standard error) and model fitting for one and two components hybrid logit/LPHN.

		Bernoulli component				Log-distributions components				
Model	AIC	INT	EZ	HI	FEM	INT	EZ	HI	FEM	α
(1)	1022.95					2.418	0.256	0.081	0.180	6.669
						(0.966)	(0.192)	(0.191)	(0.191)	(2.661)
(2)	992.17	1.051				-1.014	-0.162	-0.012	0.271	0.391
		(0.138)				(0.379)	(0.148)	(0.148)	(0.149)	(0.229)
(3)	979.98	0.875	1.027	0.385	-0.612	-1.514	-0.241	-0.104	0.142	0.074
		(0.258)	(0.330)	(0.273)	(0.291)	(0.480)	(0.178)	(0.153)	(0.143)	(0.152)
(4)	975.44	0.488	0.911	0.368		-1.609			0.286	0.152
		(0.203)	(0.275)	(0.262)		(0.002)			(0.060)	(0.010)

We started by fitting the censored LPHN model with covariates (Model 1). It is also fitted by the Bernoulli/LPHN model with covariates and logit link (models 2-4), for which the results are presented in the Table 2. According to the criterion AIC, the best fit clearly is presented by the hybrid logit/LPHN model.

In the case of the Bernoulli/LPHN model, we found that of all hybrid models fitted, the best is the Model 4.

In the continuous component we has that $E(Y) \neq X_{(2)}\beta_{(2)}$ since $E(\epsilon) \neq 0$. In order to have $E(Y) = X_{(2)}\beta_{(2)}$ we must correct the intercept taking $\beta^*_{(2)0} = \beta_{(0)} + E(\epsilon)$, where $\epsilon \sim LPN(0, \eta, \alpha)$, That is, the corrected estimator for the intercept of the regression model corresponding to the continuous part. Therefore, for model 4, we found that $\hat{\beta}^*_{(2)0} = -0.333$.

Here, covariates EZ and HI entered only in the Bernoulli component, and covariate FEM is the only associated with the LPHN component. Based on the Model 4, for those observations above the detection limit, the girls had $\exp(0.286) = 1.331$, and hence greater measles antibody concentration than boys.

As mentioned at the beginning of this illustration, the goal was to show that the model censored logit/LPHN was a good alternative to adjust the data set vaccine now we are going to show that this model is indeed different from the model censored logit/LN, so, we test the hypothesis

$$H_0 : \ \alpha = 1 \ \text{ versus } \ H_1 : \ \alpha \neq 1$$

Using the likelihood ratio statistics, we have that

$$-2\log(\Lambda) = -2(-511.18 + 480.72) = 60.91$$

which is greater than the 5% critical chi-square value 3.84, then we conclude that the logit/LPHN model fits the data better than the logit/LN model.

As a proof of good fit of the proposed model, we can confirm that the proportion of observations below the detection limit is 26.1% and the estimated proportion from model 2 with the hybrid model logit/LPHN is 25.90%.

Finally, in order to check the fit of the model estimates, we make the QQplot of the standardized residuals or scaled residuals of the continuous part, $e_i = (\log(y_i) -$

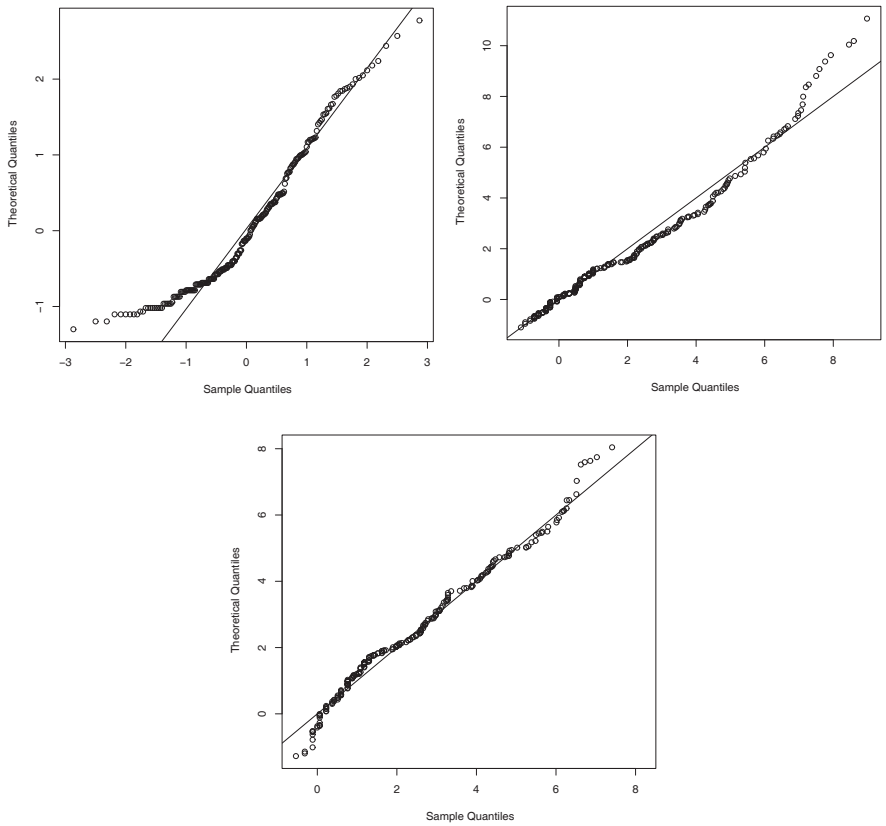

FIGURE 3: QQ-plot of the scaled residuals e_i, from the fit of Model 4. (a) log-normal (b)log-gamma and (c) log-proportional hazard-normal.

$x'_{(2)i}\hat{\beta}_{(2)})/\hat{\eta}$ based on the model and to the LN, LGM and LPHN distributions. Figure 3 presents QQplots for the scaled residuals.

Here, we can see that for vaccine data, the model LPHN fits better than the LN and LGM models, and thus, the mixed model logit/LPHN may be a new option to adjust censored data with covariates.

5. Conclusions

We proposed a new distribution that is used to study an asymmetrical regression model for data with limited responses through the mixture of a Bernoulli distribution with logit link and the LPHF distribution. Additionally, we made an illustration with real data and showed that the proposed model is an alternative for censored positive data.

[Recibido: noviembre de 2013 — Aceptado: abril de 2014]

References

Ahrens, L. H. (1954), *Quantitative Spectrochemical Analysis of Silicates*, London, Pergamon Press.

Akaike, H. (1974), 'A new look at statistical model identification', *IEEE Transaction on Automatic Control* **AU-19**, 716–722.

Arellano-Valle, R. B. & Azzalini, A. (2008), 'The centred parametrization for the multivariate skew-normal distribution', *Journal of Multivariate Analysis* **99**, 1362–1382.

Azzalini, A. (1985), 'A class of distributions which includes the normal ones', *Scandinavian Journal of Statistics* **12**, 171–178.

Azzalini, A., dal Cappello, T. & Kotz, S. (2003), 'Log-skew-normal and log-skew-t distributions as models for family income data', *Journal of Income Distribution* **11**, 12–20.

Chai, H. & Bailey, K. (2008), 'Use of log-normal distribution in analysis of continuous data with a discrete component at zero', *Statistics in Medicine* **27**, 3643–3655.

Cragg, J. (1971), 'Some statistical models for limited dependent variables with application to the demand for durable goods', *Econometrica* **39**, 829–844.

Job, J., Halsey, N., Boulos, R., Holt, E., Farrell, D., Albrecht, P., Brutus, J., Adrien, M., Andre, J., Chan, E., Kissinger, P., Boulos, C. & the CiteSoleil/JHU, P. T. (1991), 'Successful immunization of infants at 6 months of age with high dose edmonston-zagreb measles vaccine', *Pediatric Infectious Diseases Journal* **10**, 303–311.

Martínez-Florez, G., Bolfarine, H. & Gómez, H. W. (2013), 'Asymmetric regression models with limited responses with an application to antibody response to vaccine', *Biometrical Journal* **55**, 156–172.

Martínez-Florez, G., Moreno-Arenas, G. & Vergara-Cardozo, S. (2013), 'Properties and inference for proportional hazard models', *Revista Colombiana de Estadística* **36**(1), 95–114.

Mateu-Figueras, G. & Pawlosky-Glanh (2003), Una alternativa a la distribución log-normal, *in* 'Actas del XXVII Congreso Nacional de Estadística e Investigación Operativa (SEIO)', Sociedade de Estadítica e Investigación Operativa, España, pp. 1849–1858.

Mateu-Figueras, G., Pawlosky-Glanh, . & Barcelo-Vidal, C. (2004), The natural law in geochemistry: Lognormal or log skew-normal?, *in* '32th International Geological Congress', International Union of Soil Sciences, Florence, Italy, pp. 1849–1858.

Moulton, L. & Halsey, N. (1995), 'A mixture model with detection limits for regression analyses of antibody response to vaccine', *Biometrics* **51**, 1570–1578.

Moulton, L. & Halsey, N. (1996), 'A mixed Gamma model for regression analyses of quantitative assay data', *Vaccine* **14**, 1154–1158.

Tobin, J. (1958), 'Estimation of relationships for limited dependent variables', *Econometrica* **26**, 24–36.

Appendix

In this appendix we present the Hessian matrix for the logit/LPHN model. Its elements are given by

$$U(\beta_{(1)j}\beta_{(1)r}) = \sum_0 x_{1ij}x_{1ir}\left[\frac{\exp(x'_{(1)i}\beta_{(1)})[1-\{1-\Phi(z_{T_i})\}^{\alpha}]}{\{1+\exp(x'_{(1)i}\beta_{(1)})[1-\{1-\Phi(z_{T_i})\}^{\alpha}]\}^2}\right] - \sum_{i=1}^{n}\frac{x_{1ij}x_{1ir}\exp(x'_{(1)i}\beta_{(1)})}{[1+\exp(x'_{(1)i}\beta_{(1)})]^2},$$

$$U(\beta_{(2)k}\beta_{(1)j}) = \frac{-\alpha}{\eta}\sum_0 \frac{x_{2ik}x_{1ij}\phi(z_{T_i})\exp(x'_{(1)i}\beta_{(1)})\{1-\Phi(z_{T_i})\}^{\alpha-1}}{\{1+\exp(x'_{(1)i}\beta_{(1)})[1-\{1-\Phi(z_{T_i})\}^{\alpha}]\}^2}$$

$$U(\beta_{(1)j}\eta) = \frac{-\alpha}{\eta}\sum_0 \frac{x_{1ij}z_{T_i}\phi(z_{T_i})\exp(x'_{(1)i}\beta_{(1)})\{1-\Phi(z_{T_i})\}^{\alpha-1}}{\{1+\exp(x'_{(1)i}\beta_{(1)})[1-\{1-\Phi(z_{T_i})\}^{\alpha}]\}^2},$$

$$U(\beta_{(1)j}\alpha) = -\sum_0 \frac{x_{ij}\exp(x'_{(1)i}\beta_{(1)})\{1-\Phi(z_{T_i})\}^{\alpha}\log(1-\Phi(z_{T_i}))}{[1+\exp(x'_{(1)i}\beta_{(1)})[1-\{1-\Phi(z_{T_i})\}^{\alpha}]]^2},$$

$$U(\beta_{(2)k}\beta_{(2)s}) = \frac{-\alpha}{\eta^2}\sum_0 x_{2ik}x_{2is}\left\{[z_{T_i}+(\alpha-1)M_i]A_i + A_i^2\right\} + \frac{1}{\eta^2}\sum_1 x_{2ik}x_{2is}\left\{-1+(\alpha-1)z_iM_i-(\alpha-1)M_i^2\right\},$$

$$U(\beta_{(2)k}\eta) = \frac{\alpha}{\eta^2}\sum_0\left\{\left[x_{2ik}-x_{2ik}z_{T_i}^2-(\alpha-1)x_{2ik}z_{T_i}M_i\right]A_i-\alpha x_{2ik}z_{T_i}A_i^2\right\} + \frac{1}{\eta^2}\sum_1\left\{x_{2ik}\left[\frac{-2z_i}{\eta}-(1-z_i^2)(\alpha-1)M_i-z_i(\alpha-1)M_i^2\right]\right\},$$

$$U(\beta_{(2)k}\alpha) = \frac{-1}{\eta}\sum_{0}\left\{[1+\alpha\log(1-\Phi(z_{T_i}))]x_{2ik}A_i + \frac{\alpha x_{2ik}[1-\Phi(z_{T_i})]\log(1-\Phi(z_{T_i}))}{\phi(z_{T_i})}A_i^2\right\} + \frac{1}{\eta}\sum_{1}x_{2ik}M_i,$$

$$U(\eta\eta) = \frac{\alpha}{\eta^2}\sum_{0}\left\{\left[2z_{T_i}-z_{T_i}^3+(\alpha-1)z_{T_i}^2M_i\right]A_i+\alpha z_{T_i}^2A_i^2\right\} + \frac{1}{\eta^2}\sum_{1}\left\{1-z_i^2-(\alpha-1)z_iM_i\left[2-\frac{z_i^2-\phi(z_i)z_i}{\phi(z_i)}M_i\right]\right\},$$

$$U(\eta\alpha) = \frac{-1}{\eta}\sum_{0}\left\{[z_{T_i}+\alpha z_{T_i}\log(1-\Phi(z_{T_i}))]A_i + \left[\frac{\alpha z_{T_i}(1-\Phi(z_{T_i}))\log(1-\Phi(z_{T_i}))}{z_{T_i}\phi(z_{T_i})}\right]A_i^2\right\} + \frac{1}{\eta}\sum_{1}z_iM_i,$$

$$U(\alpha\alpha) = -\sum_{0}\left\{\frac{\{1-\Phi(z_{T_i})\}\log^2(1-\Phi(z_{T_i}))}{\phi(z_{T_i})}A_i + \left[\frac{\{1-\Phi(z_{T_i})\}\log(1-\Phi(z_{T_i}))}{\Phi(z_{T_i})}A_i\right]^2\right\} - \sum_{1}\frac{1}{\alpha^2}$$

where

$$A_i = \frac{\phi(z_i)\exp(x'_{(1)i}\beta_{(1)})\{1-\Phi(z_{T_i})\}^{\alpha-1}}{1+\exp(x'_{(1)i}\beta_{(1)})[1-\{1-\Phi(z_{T_i})\}^{\alpha}]} \quad \text{and} \quad M_i = \frac{\phi(z_{T_i})}{1-\Phi(z_{T_i})}.$$