Brazilian Journal of Biometrics **41** (2), 131–143 (2023)

doi:10.28951/bjb.v42i2.604

BRAZILIAN JOURNAL OF
BIOMΣTRICS
ISSN:2764-5290

ARTICLE

Bayesian binary regression using power and power reverse link functions: an application to premature birth data[1]

Rafaela Galo[1] , Robson Marcelo Rossi*[2], Diego Corrêa Alves[2] and Rosana Rosseto de Oliveira[3]

[1]Graduate Program in Biostatistics, State University of Maringá, Maringá-PR, Brazil

[2]Department of Statistics, State University of Maringá, Maringá-PR, Brazil

[3]Department of Nursing, State University of Maringá, Maringá-PR, Brazil
*Corresponding author. Email: *rmrossi@uem.br*

(Received: April 1, 2022; Revised: July 22, 2022; Accepted: August 22, 2022; Published: June, 2023)

Abstract

This study aims to determine factors associated (and quantify) with prematurity through binary regression models, considering power and reverse power link functions, with asymmetric characteristics. As criteria for the model selection, the Bayesian Deviance Information Criterion (DIC), predictive evaluation, and residual analysis. All models analyzed presented similar predictive capacity, however, the model with a reverse power logit link function, with asymmetry parameter λ=0.336 was chosen, since it presented the lowest value of DIC=3,203, residues that indicated a good fitted of the model. There was an association of prematurity with the following variables: maternal - age over 35 years (OR=1.485), with a partner (OR=0.731), and primiparous (OR=1.307); of pregnancy and childbirth - multiple pregnancy (OR=36.360), cesarean childbirth (OR=1.337) and number of prenatal consultations less than seven (OR=3.305); and newborns of white race/skin (OR=0.731) and presence of congenital malformation (OR=2.663). Considering the proposed criteria, an asymmetric link function (reverse power logit) was the most parsimonious for the model. From this, there were high chances of factors associated with the occurrence of prematurity, indicating the need for actions to minimize them.

Keywords: asymmetric link functions; premature birth; risk factors; statistical methods; unbalanced data.

1. Introduction

Preterm birth, which occurs before the 37th week of gestation, is a leading cause of neonatal morbidity and mortality (Liu; Allen; Fraser, 2008; Saigal; Doyle, 2008; Han *et al.*, 2011). Worldwide, it is estimated that the total number of premature births per year is around 15 million, and of these, 1 million children die annually from complications of this prematurity (WHO, 2012). Therefore, the assessment and identification of risk factors associated with prematurity are essential to support decision-making aimed at reducing it. Nonetheless, biological mechanisms, such as those observed in the epidemiological area, are often dispersed and unbalanced, especially when the studied outcomes are rarer. An example of this is the birth of premature babies, which, in general, are more likely to be no premature births, resulting in an imbalance in the proportion of this outcome. Still, logistic regression models (*logit* link function) are still the most used option in the epidemiological field, which can generate biased estimates, given the symmetry characteristic and little flexibility of this link function.

Studies aiming to determine which factors are associated with the birth of premature children tend to use only the binary regression model with *logit* link function (Cascaes *et al.*, 2008; Oliveira, 2015; Guimarães *et al.*, 2017), even though the event of interest must be unbalanced. Historically, binary regression models with logit link function (Logistic model) have been widely used in epidemiological investigations in different study designs, as a very important tool to model the association between a set of predictors (or explanatory) variables and a dichotomous response variable (Hosmer *et al.*, 2013).

As binary regression models are one of the classes that make up Generalized Linear Models (Nelder and Wedderburn, 1972), the *logit* link function is commonly used in these models, as it is the canonical link function of the binomial model, in addition to making estimates of its parameters easily interpreted by means of odds ratio (OR) (McCullagn and Nelder, 1989). The logit function is obtained by the inverse of the accumulated distribution function of standard logistics, which provides it with a characteristic of symmetry, that is, the probability of the response variable approaches success at the same rate at which it approaches failure. Other usual symmetrical link functions are *probit* (Gaussian) and *cauchit* (Albert and Chib, 1993). However, the assumption of models from link functions with this characteristic may not be correct in some situations, for example, when the frequency of successes and failures is unbalanced (Chen, Dey and Shao, 1999), which may lead to erroneous conclusions due to biases in parameter estimation (Czado and Santner, 1992). Thus, there are indications (Abanto-Valle *et al.*, 2015) to use asymmetric link functions, such as *cloglog* and *loglog* (Gumbel distributions of Minimum and Maximum Value, respectively).

Recently, other, more flexible, asymmetric link functions have been proposed, and called power and reverse power (Abanto-Valle *et al.*, 2014; Bazán, Romeo and Rodrigues, 2014; Bazán *et al.*, 2017). These are more flexible because they have an asymmetry parameter, which has its estimate related to the number of successes and failures in the response, which also influences the predicted probabilities. Such models are presented as a good alternative to unbalanced dichotomous data, in different fields of research (Huayanay *et al.*, 2019; Silva *et al.*, 2020).

In this context, the present study aimed to determine factors associated (and quantify such association) with the occurrence of premature births in 2017, in the city of Maringá, Paraná, through Bayesian binary regression models, considering the power and reverse power link functions, of asymmetrical characteristic (unbalanced data).

2. Materials and Methods

2.1 Data

The data used in this study were obtained from the Live Birth Information System (SINASC), which is based on the Live Birth Declaration (DNV). A total of 5,060 records of live births residing in a municipality in the south of Brazil (city of Maringá, State of Paraná), in 2017, and information on eleven possible explanatory variables (factors) related to the mother, prenatal care, and childbirth, and newborn, were used. Such records were provided by the 15th Regional Health Department of the State of Paraná, upon approval by the Permanent Research Ethics Committee (COPEP), from the State University of Maringá, under code number 406.927/2013.

The response variable, prematurity (*Y*), was obtained from the classification used by the World Health Organization (Who, 2021), that is, it is defined as “premature”, $Y=1$, the baby born alive before completing 37 weeks of gestation and $Y=0$ otherwise. From this, it can be observed an imbalance in the frequencies of the variable, that is, that premature births are less frequent (11.8%) than non-premature births (88.2%) (Table 1).

The variables selected to verify a possible association with prematurity were those obtained before the event occurred. Thus, those related to the mother (age, partner, education, and parity) were obtained; to pregnancy and childbirth (type of pregnancy, type of childbirth, prenatal consultations, and place of childbirth) and to the newborn (sex, race/skin color, and congenital malformation). Each variable was categorized into reference (lower levels) and risk, according to the criteria presented in the literature (Oliveira, 2015) (Table 1).

Table 1. Proportions according to the occurrence of prematurity in newborns, Maringá, Paraná, Brazil, 2017

Relationship	Variable	Categories	Prematurity	
			0: No	1: Yes
			%	%
Mother	Age (years)	0: [20-35)	88.7	11.3
		1: <20	90.7	9.3
		2: ≥35	85.8	14.2
	Partner	0: yes	87.8	12.2
		1: no	89.2	10.8
	Education (years)	0: ≥12	87.2	12.8
		1: <12	88.8	11.2
	Parity	0: multipara	88.9	11.1
		1: primipara	87.1	12.9
Pregnancy and childbirth	Type of pregnancy	0: only	90.2	9.8
		1: multiple	25.0	75.0
	Type of childbirth	0: vaginal	91.6	8.4
		1: cesarean	87.2	12.8
	Prenatal cosultation	0: ≥7	90.1	9.9
		1: <7	77.6	22.4
	Location of childbirth	0: hospital	88.2	11.8
		1: others	100.0	0.0
Newborn	Sex	0: female	88.2	11.8
		1: male	88.2	11.8
	Race/Skin Color	0: white	87.4	12.6
		1: no white	90.2	9.8
	Congenital malformatio	0: no	88.3	11.7
		1: yes	75.5	24.5
General			88.2	11.8

2.2 Model

In the context of Generalized Linear Models for binary data, consider $Y=(y_1, y_2, ..., y_n)^T$, the vector of independent random variables. Additionally, be $\mathbf{x}=(1, x_{i1}, x_{i2}, ..., x_{ik})$ $i=1, 2, ..., n$, the set of vectors of the values corresponding to k predictors variables, forming $X=[1,x]$, the matrix of the model's incidence. Defining $Y=1$, if the event occurred and $Y=0$, otherwise, $Y \sim Ber(p)$ (Bernoulli distribution), where $P(Y=1)=p$ (probability of occurrence of the event, premature birth). Thus, a binary regression model is specified such that $Y|\mathbf{x} \sim \mathrm{Ber}(p)$, $p=E[Y|\mathbf{x}]=F(\eta)=F(\boldsymbol{x\beta})$, or, equivalently, $\eta=\boldsymbol{x\beta}=F^{-1}(p)$, where $\boldsymbol{\beta}=(\beta_0, \beta_1, ..., \beta_k)^T$ is vector of regression coefficients, $F(.)$ is the cumulative distribution function (c.d.f.) and $F^{-1}(.)$ the link function relating p to the linear predictor η. The linear predictor determines the linear combination between the predictors variables and the regression coefficients, given by: $\eta=\beta_0+\beta_1 x_{i1}+...+\beta_k x_{ik}$, being $i=1, ..., 5{,}060$ individuals and $k=11$ regression variables initially analyzed.

2.2.1 Link functions

In the literature, the *logit*, *probit*, and *cauchit* link functions, all with symmetrical characteristics, are obtained from the Logistic, Normal, and Cauchy probability distributions, respectively; and the *cloglog*, and *loglog* link functions, of asymmetrical characteristic, obtained from the Gumbel distributions of Minimum and Maximum Value, respectively, are said to be basic (Albert and Chib, 1993; Abanto-Valle *et al.*, 2015). In addition to these usual link functions, it is intended to use here the class of link functions known as power and reverse power, as they are presented as an alternative for a better fit of the model (Bazán *et al.*, 2014).

To obtain such link functions, one must know the probability distributions from which they come. A univariate random variable T has power probability distribution, $T \sim P(\mu,\sigma^2,\lambda)$, or reverse power, $T \sim RP(\mu,\sigma^2,\lambda)$, being $\mu \in R$ location parameter, $\sigma^2>0$ scale parameter and $\lambda>0$ asymmetry parameter, if your c.d.f. are, respectively, of the form: $F_P(t \mid \mu,\sigma^2,\lambda)=\left[G\left(\frac{t-\mu}{\sigma}\right)\right]^{\lambda}$ and $F_{RP}(t \mid \mu,\sigma^2,\lambda)=1-\left[G\left(-\left(\frac{t-\mu}{\sigma}\right)\right)\right]^{\lambda}$, where $t \in R$, and G(.) the c.d.f. called the base, of any continuous probability distribution in the real numbers, being symmetric or asymmetric.

In the case where $\mu=0$ and $\sigma^2=1$, the power and reverse power distributions are presented in their standard form. The random variables W have standard power distribution, $W \sim P(\lambda)$, or standard reverse power distribution, $W \sim RP(\lambda)$, if their c.d.f. are, respectively, of the form: $F_P(w \mid \lambda)=[G(w)]^{\lambda}$ and $F_{RP}(w \mid \lambda)=1-[G(-w)]^{\lambda}$ w real number. Note that when $\lambda=1$, the power or reverse power distribution is equal to the base distribution.

In this work, those that form the usual link functions (Bazán *et al.*, 2017; Silva *et al.*, 2020), that is, the Logistics, Normal (Gaussian), Cauchy, Minimum and Maximum Value Gumbel, were also considered as base distributions for the power and reverse power distributions (Table 2).

Considering the variables of interest, binary regression models were fitted, with the link functions shown in Table 2, under the Bayesian approach, in order to choose the model with the best fit (Paulino *et al.*, 2003; Rossi, 2011). Considering that for the basics link functions, the parameters to be estimated are the regression coefficients, $\boldsymbol{\beta}=(\beta_0, \beta_1, ..., \beta_k)$, on the other hand, for the power and reverse power link functions, in addition to the previous parameters, it is also necessary to estimate λ, the parameter of asymmetry, which can be obtained from the transformation $\delta=ln(\lambda) \leftrightarrow \lambda= \exp(\delta)$, which facilitates Bayesian calculations (Bazán *et al.*, 2014).

Table 2. Basic, power and reverse power link functions for binary regression

[1]Distribution	$F(\cdot)$	Link	$F^{-1}(\cdot)$
L	$p=\frac{\exp(\eta)}{1+\exp(\eta)}$	*logit (L)*	$\eta=\log\left(\frac{p}{1-p}\right)$
N	$p=\Phi(\eta)$	*probit (P)*	$\eta=\Phi^{-1}(p)$
C	$p=0,5+\frac{\arctan(\eta)}{\pi}$	*cauchit (C)*	$\eta=\tan(\pi(p-0,5))$
GVMi	$p=1-\exp(-\exp(\eta))$	*cloglog (CLL)*	$\eta=\log(-\log(1-p))$
GVMa	$p=\exp(-\exp(-\eta))$	*loglog (LL)*	$\eta=-\log(-\log(p))$
PL	$p=\left(\frac{\exp(\eta_i)}{1+\exp(\eta)}\right)^{\lambda}$	*power logit (PL)*	$\eta=\log\left(\frac{p^{1/\lambda}}{1-p^{1/\lambda}}\right)$
PN	$p=(\Phi(\eta))^{\lambda}$	*power probit (PP)*	$\eta=\Phi^{-1}(p^{1/\lambda})$
PC	$p=\left(0,5+\frac{\arctan(\eta)}{\pi}\right)^{\lambda}$	*power cauchit (PC)*	$\eta=\tan(\pi(p^{1/\lambda}-0,5))$
PGVMi	$p=(1-\exp(-\exp(\eta)))^{\lambda}$	*power cloglog (PCLL)*	$\eta=\log(-\log(1-p^{1/\lambda}))$
PGVMa	$p=(\exp(-\exp(-\eta)))^{\lambda}$	*power loglog (PLL)*	$\eta=-\log(-\log(p^{1/\lambda}))$
RPL	$p=1-\left(\frac{\exp(-\eta)}{1+\exp(-\eta)}\right)^{\lambda}$	*reverse power logit (RPL)*	$\eta=\log\left(\frac{1-(1-p)^{1/\lambda}}{(1-p)^{1/\lambda}}\right)$
RPN	$p=1-(\Phi(-\eta))^{\lambda}$	*reverse power probit (RPP)*	$\eta=-\Phi^{-1}((1-p)^{1/\lambda})$
RPC	$p=1-\left(0,5+\frac{\arctan(-\eta)}{\pi}\right)^{\lambda}$	*reverse power cauchit (RPC)*	$\eta=-\tan(\pi((1-p)^{1/\lambda}-0,5))$
RPGVMi	$p=1-(1-\exp(-\exp(-\eta)))^{\lambda}$	*reverse power cloglog (RPCLL)*	$\eta=-\log(-\log(1-(1-p)^{1/\lambda}))$
RPGVMa	$p=1-(\exp(-\exp(\eta)))^{\lambda}$	*reverse power loglog (RPLL)*	$\eta=\log(-\log((1-p)^{1/\lambda}))$

Souce: Adapted of Bazán *et al.* (2017) and Silva *et al.*, (2020);

[1]L: Logistic; N: Normal (Gaussian); C: Cauchy; GVMi: Minimum Value Gumbel; GVMa: Maximum Value Gumbel; PL: Power Logistic; PN: Power Normal; PC: Power Cauchy; PGVMi: Power Minimum Value Gumbel; PGVMa: Power Maximum Value Gumbel;

log(.), tan(.), arctan(.) and Φ(.) represents the ln(.), tangent, inverse tangent and standard normal cumulative distribution functions, respectively.

[1]RPL: Reverse Power Logistic; RPP: Reverse Power Normal; RPC: Reverse Power Cauchy; RPGVMi: Reverse Power Minimum Value Gumbel; RPGVMa: Reverse Power Maximum Value Gumbel.

2.2.2 Bayesian estimation

Thus, priors distributions were defined as non-informative, such that $Y|\boldsymbol{\beta},\delta \sim$ Ber(p), p=F(η), $\boldsymbol{\beta} \sim$ N (0,10^{-6}) (parameterization in OpenBUGS - Bayesian analysis software of the researchers Spiegelhalter *et al.*, 1996) and $\delta \sim$ U(-2,2), since the values of λ in the interval [e^{-2},e^{2}] have a higher probability of occurrence (Bazán *et al.*, 2017; Silva *et al.*, 2020). Note that if there are models with basic link functions when $\delta=0$ ($\lambda=1$). The posterior distributions for the parameters were obtained numerically via the *R* software (R Core Team, 2022) through the *BRugs* package (Thomas *et al.*, 2006) linked to *OpenBUGS* through the interactive procedure MCMC process (Markov Chain Monte Carlo) via Gibbs Sampling algorithm (Geman and Geman, 1984).

For all models, as initial values for $\boldsymbol{\beta}$, were used frequentist (classical) estimates with *logit* link function and, to λ, a random generated value from a uniform distribution, with parameters a=-2 and b=2. In addition, 1,100,000 values were generated considering a sample discard period of 100,000 and jump of size 100 to reduce the autocorrelation generated in the MCMC process (Markov Chain Monte Carlo), forming final samples of 10,000 generated values. The convergence of the chains was verified by the Geweke (1992) and Heidelberger and Welch (1983) criteria through the *coda* package (Plummer *et al.*, 2006) in *R* software. The parameter estimates, obtained from posterior distributions, were the mean, the standard deviation and the median (used when there is an asymmetry in the posterior distribution). Furthermore, the “significance” of these parameters was verified if their respective HPD (Highest Posterior Density) intervals, with 95% credibility, did not contain a zero value.

2.2.3 Model selection, diagnostics, and predictive evaluation

Model selection, diagnosis, and predictive evaluation Some criteria were considered to carry out the selection final model fitted. As an information criterion, the Bayesian Deviance Information Criterion (DIC) (Spiegelhalter *et al.*, 2002) was used. Models with lower DIC values are indicated as more parsimonious.

In order to evaluate the quality of fit of the model to the data, it was performed the diagnosis of the residuals in addition to the verification of their predictive capacity (Giolo, 2017). Regarding the residual, it was chosen to use the normalized quantile (Dunn and Smyth, 1996) and check its suitability from the index graphs and the binomial simulated envelope (Atkinson, 1981).

For the evaluation of predictive capacity, the study population was divided into sets: training (70%) and test (30%), in order not to obtain an overfitting for the prediction model (Kuhn and Kjell, 2013). Predictive measures of sensitivity (s), specificity (e), accuracy (ACC) and the area under the ROC curve (AUC - area under ROC curve (Receiver Operating Characteristic)) - measure summarizing the overall accuracy of the predictive model (Fawcett, 2006; Borges, 2016), were calculated.

After the diagnostic stage, the best model was selected and with that, the OR’s were calculated for the associations found. Such a measure can be used in studies with a cross-sectional design, provided that its interpretation is carried out in terms of odds, and not of probability associated with risk (Castro-Costa and Ferri, 2008).

3. Results and Discussion

From the exploratory analysis, it can be verified that for the Age variable of the mother, the proportion of premature births is higher for women aged 35 years or over, when compared to the reference category and for women aged less than 20 years is less than the reference category. In the variables with only two categories: primiparous mothers, who had cesarean childbirth and with an insufficient number of prenatal consultations, in addition to newborns with congenital malformations, also present risk categories with a higher proportion of premature births, which descriptively indicates a possible association with the outcome. A variable that presents a large proportion of premature births in its risk range is formed by mothers who had multiple pregnancies. It is possible to verify a proportion of premature births three times higher than non-premature infants, indicating descriptively that the association of this variable with the outcome will be considered strong.

For the Partner and Race/Skin Color variables, the risk categories have a lower proportion of premature compared to the reference one. This result indicates descriptively that the association of these variables with the prematurity outcome has an inverse effect on the risk class. That is, individuals without partners have 73.1% (OR=0.731) of the risk of those with partners, and non-white individuals have 75% of the risk of whites (OR=0.75).

Regarding the variables: Education, Type of childbirth and sex, the proportion of premature and non-premature in the reference and risk categories present the same (or very close) pattern of premature and non-premature in the population, which indicates descriptively that there may be no association of these variables with the outcome.

The variable Age (related to the mother) was considered as a dummy variable, such that the age between 20 and 35 years was the reference class in both cases. Thus, it was defined as Age1 for less than 20 years being the intermediary risk category, and Age2 for over or equal to 35 years being the elevate risk category. The explanatory variables that were used in the proposed models were Age2, Partner, Parity, Type of pregnancy, Type of childbirth, Prenatal consults, Race/Skin Color and Congenital malformation, according to Oliveira (2015).

All the fitted models passed the Geweke, and Heidelberger and Welch tests, thus verifying the convergence in the posterior chains. It was found that from the DIC (Table 3), the models with *RPL* and *RPC* link functions are also parsimonious, besides presenting the lowest values (3,203 and 3,199, respectively) when compared to the others.

From the results of the first selection stage, it was used the criteria for evaluating the quality of the fit of the models, with link functions that stood out by the DIC (*RPL* and *RPC*), and also, for the model with logit link function, commonly used to determine associations of factors to prematurity (Cascaes *et al.*, 2008; Guimarães *et al.*, 2017).

When evaluating the accuracy, sensitivity, specificity and *AUC* measures (Table 4), it is found that all models have very close values, indicating that the predictive analysis criterion is not useful for choosing a model for the preterm birth data of the present study.

Table 3. Deviance Information Criterion (DIC) Bayesian of the considered models

Link type	Link	DIC
Basic	*logit (L)*	**3,210**
	probit (P)	3,211
	cauchit (C)	3,213
	cloglog (CLL)	3,212
	loglog (LL)	3,215
Power	*power logit (PL)*	3,213
	power probit (PP)	3,210
	power cauchit (PC)	3,211
	power cloglog (PCLL)	3,224
	power loglog (PLL)	3,215
Reverse power	*reverse power logit (RPL)*	**3,203**
	reverse power probit (RPP)	3,312
	reverse power cauchit (RPC)	**3,199**
	reverse power cloglog (RPCLL)	3,219
	reverse power loglog (RPLL)	3,212

Taking into account all the proposed criteria to verify the fitted models (Table 3), the *RPL* model was chosen. The estimates of the parameters of all the models analyzed so far (*L*, *RPL* and *RPC*), are presented in Table 5. Note that for the chosen link function *(RPL)*, all parameters were significant, since $0 \not\subset$ HPD intervals. It was decided to use the median as a point estimate for the parameters, as there is asymmetry in posterior distributions.

The diagnosis of the fit of models with *L*, *RPL* and *RPC* link functions, through the graphic analysis of the randomized quantile residuals, shows that the *RPL* model did not present excessive observations highlighted from the others, in addition to not presenting points (residuals) outside the credibility band simulated (Figure 1), indicating that the model is adequate. For the other models such behaviors were not verified.

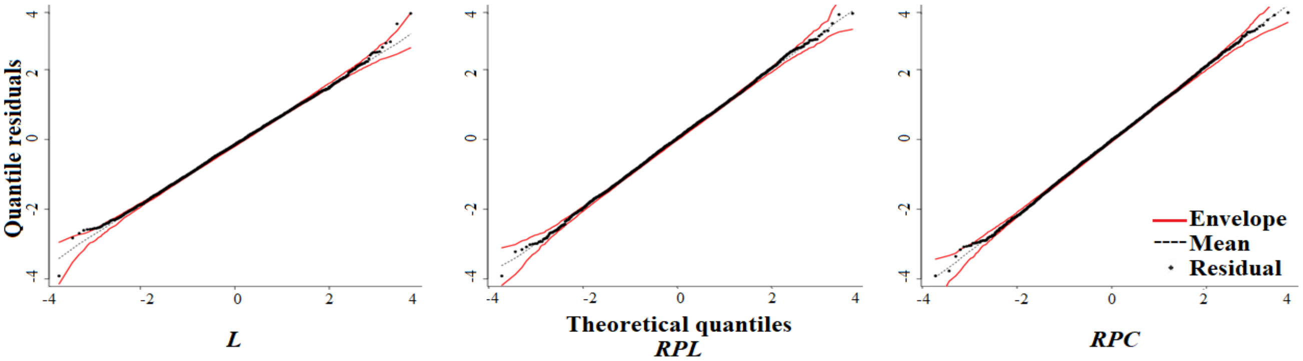

Figure 1. Normalized quantile residuals for (binomial simulated envelopes) *logit* (*L*), power reverse *logit* (*RPL*) and power reverse *cauchit* (*RPC*) regression model, respectively.

Table 4. Predictive measurements of *L*, *RPL* and *RPC* models

Link	Predicted	Observed		%			
		1	0	^{1}*ACC*	*S*	*e*	*AUC*
L	1	131	630	56.1	78.0	53.3	71.9
	0	37	720				
RPL	1	131	630	56.1	78.0	53.3	71.9
	0	37	720				
RPC	1	127	591	58.4	75.6	56.2	72.1
	0	41	759				

^{1}*ACC*: accuracy; *s*: sensitivity; *e*: specificity; *AUC*: area under the ROC curve.

Considering the *RPL* model, it can be said that the odds of prematurity was higher among those born to mothers over 35 years of age (OR=1.485) and primiparous (OR=1.307); with multiple pregnancy (OR=36.360), cesarean childbirth (OR=1.337), with insufficient number of prenatal consultations (OR=3.305) and newborns with congenital malformations (OR=2.663). In addition, the odds for prematurity was lower among those born to a mother with a partner (OR=0.731) and those with the white race/skin color self-referred by the mother (OR=0.750) (Table 5).

The present study showed that, when considering criteria for the selection of a model, it was found that the one, with asymmetric characteristic (*RPL* - has the estimate of the asymmetry parameter related to the data), was more suitable for the analysis of factors associated with the outcome, which is unbalanced, although the parameter estimate is not significant, as it includes the unit value in its high density range (λ=0.336; HPD: (0.135;2.707)) and marginally significant for the *RPC* model (λ=0.708; HPD: (0.313;1.025)). Also, when compared with the results obtained by the model with *logit* link function, which is the commonly used for this type of analysis (Cascaes *et al.*, 2008; Guimarães *et al.*, 2017), greater association measures are verified for the model chosen through the criteria of selection, that is, the asymmetric reverse power link function better captures the associations between factors and outcome.

It is also important to highlight, as the use of models with power and reverse power reverse link functions, are present in the literature for the analysis of different areas. In the application (Chen *et al.*, 1999) in which a sample of 4,000 car policyholders was used, with unbalanced amounts of successes and failures (34% and 66%, respectively), concluded that the model with the *RPC* link function would be the most parsimonious under the Bayesian perspective, to predict customers who will purchase a full insurance coverage plan for their automobile, based on sex, driving area, vehicle use, marital status, age and seniority in the company.

From an application (Lemonte and Bazán, 2018) in order to present models with power and reverse power link functions obtained from the Logistics, Normal (Gaussian), Cauchy, Laplace and *t*-Student distributions, as an alternative for modeling binary responses, used data related to coca leaf cultivation in Peru. From a sample of 1,947 coca producers, an exploratory study on the factors associated (demographic characteristics and aspects of social programs implemented in recent years in the country) to their behavior in the decision to continue or eradicate coca cultivation, was carried out. Using the frequentist (classical) approach, they opted for the model with *t*-Student reverse power link function, therefore, based on the criteria: the highest value of the likelihood function evaluated in the estimates, information criteria (AIC, BIC and HQIC), diagnostic analyzes (local influence and residues) and because it has a lower standard deviation for the estimates of the parameters, the model with this link function was chosen.

With the same distributions mentioned above, an application (Huayanay *et al.*, 2019) was presented to verify the performance of the models. From chemical attributes determined by wine connoisseurs, the objective was to verify which of these, the quality of white wine from the Minho region, in Portugal, was associated between May 2004 and February 2007. In a total of 4,898 wine samples, it was found that 21% were considered of good quality, showing the imbalance of the response variable. From certain Information Criteria, predictive analysis and the randomized quantile residual, the best fit was obtained by the model with *t*-Student power link function with 0.6 degrees of freedom.

In general, a model with a link function that has the estimate of the asymmetry parameter related to the unbalanced response, presented a better fit in different fields of study, as seen above, and in data related to prematurity. Also, the Bayesian methodology used to obtain parameter estimates was a practical option for implementing the models, either using the MCMC iterative process via the Gibbs Sampling (Geman and Geman, 1984) algorithm as here, or using other algorithms (Bazán *et al.*, 2017; Huayanay *et al.*, 2019; Silva *et al.*, 2020), in addition to highlighting the asymmetry that the posterior distribution of the parameters presents for the link function chosen.

In the study (Silva *et al.*, 2020) from a Bayesian perspective, the authors present modeling results in simulations of unbalanced dichotomous data, considering different sample sizes and conclude that, regardless of the sample size, such unbalance affects the estimation of logistic regression parameters in relation to the bias, square error and the standard deviation of the estimates. They also discuss adjustment in two distinct applications (one for linguistic data and one for educational data) considering other link functions in a regression model and conclude that power and reverse power are more suitable for the modeling.

It is common knowledge in the health area that the prevention of premature birth always needs to be improved, as complications from this type of birth are the main cause of death in children under five years of age (Who, 2021), and also a causative factor for several sequels to newborns in their childhood and even into adulthood (Oliveira, 2015; Crump *et al.*, 2011). In this sense, statistics makes its contribution, presenting more sophisticated methods, providing professionals in the area and competent bodies, a basis to better lead them to a decision-making to minimize the occurrence of this type of birth.

Table 5. The posterior estimates of the parameters and HPD and OR intervals of the L, RPL and RPC models

Parameter	*logit* (*L*)			*Reverse power logit* (*RPL*)			*Reverse power cauchit* (*RPC*)		
	Mean	Median	[3]OR	Mean	Median	OR	Mean	Median	OR
	[1](s.d.)	[2](95% HPD)		(s.d.)	(95% HPD)		(s.d.)	(95% HPD)	
Intercept	-2.770	-2.768	-	-1.851	-1.653	-	-3.290	-3.320	-
	(0.139)	(-3.038;-4.498)		(0.975)	(-3.880;-0.387)		(0.812)	(-4.873;-1.522)	
Age2	0.380	0.380	1.462	0.430	0.423	1.485	0.692	0.687	1.267
	(0.111)	(0.162;0.593)		(0.137)	(0.162;0.702)		(0.216)	(0.266;1.111)	
Partner	-3.303	-0.301	0.740	-0.334	-0.329	0.731	-0.589	-0.578	0.848
	(0.116)	(-0.532;0.076)		(0.137)	(-0.601;-0.062)		(0.228)	(-1.051;-0.150)	
Parity	0.268	0.268	1.307	0.295	0.292	1.315	0.503	0.486	1.180
	(0.099)	(0.075;0.460)		(0.120)	(0.068;0.534)		(0.190)	(0.155;0.901)	
Pregnancy type	3.343	3.338	28.163	5.583	5.088	36.360	7.485	5.017	35.077
	(0.200)	(2.951;3.741)		(2.244)	(2.574;9.769)		(12.401)	(3.663;15.793)	
Childbirth type	0.266	0.266	1.305	0.320	0.310	1.337	0.439	0.420	1.148
	(0.124)	(0.030;0.513)		(0.150)	(0.011;0.600)		(0.250)	(-0.048;0.930)	
Prenatal consultation	1.174	1.174	3.235	1.373	1.349	3.365	2.010	2.002	2.485
	(0.111)	(0.962;1.395)		(0.240)	(0.942;1.818)		(0.366)	(1.267;2.714)	
Race/Skin Color	-0.269	-0.267	0.766	-0.310	-0.303	0.750	-0.543	-0.529	0.859
	(0.110)	(-0.493;-0.060)		(0.133)	(-0.584;-0.057)		(0.233)	(-1.043;-0.137)	
Congenital malformation	0.856	0.865	2.375	1.128	1.076	2.663	1.083	1.136	1.530
	(0.370)	(0.153;1.591)		(0.570)	(0.045;2.286)		(0.650)	(-0.237;2.302)	
λ				0.707	0.336	-	0.700	0.708	-
				(0.992)	(0.135;2.707)		(0.174)	(0.313;1.025)	

[1]s.d.: standard deviation; [2]HPD: highest posterior density; [3]OR: Odds Ratio

4. Conclusions

In the proposed criteria, an asymmetric link function (reverse power *logit*) was the most parsimonious for the modeling. From this, there were high odds of factors associated with the occurrence of prematurity, indicating the need for actions to minimize them.

The use of power and reverse power link functions for binary regression models applied to unbalanced data can promote significant improvements in parameter inferences.

Acknowledgments

This work had the support of the Coordination for the Improvement of Higher Education Personnel – CAPES.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Abanto-Valle, C. A., Bazán, J. L., Smith, A. C. *State space mixed models for binary responses with skewed inverse links using JAGS*. (Instituto de Matemática da UFRJ, Departamento de métodos estatísticos), (2014).

2. Abanto-Valle, C. A, Dey, D. K., Jiang, X. Binary state space mixed models with flexible link functions: a case study on deep brain stimulation on attention reaction time. *Statistics and Its Interface* **8**, 187-194 (2015).

3. Albert, J. H., Chib, S. Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association* **88**, 669-679 (1983).

4. Atkinson, A, C. Two graphical displays for outlying and influential observations in regression. *Biometrika* **68**, 13-20 (1981).

5. Bazán, J. L., Romeo, J. S., Rodrigues, J. Bayesian skew-probit regression for binary response data. *Brazilian Journal of Probability and Statistics* **28**, 467-482 (2014).

6. Bazán, J. L., Torres-Avilés, F., Suzuki, A. K., Louzada, F. Power and reversal power links for binary regressions: An application for motor insurance policyholders. *Applied Stochastic Models in Business and Industry* **33**, 22-34 (2017).

7. Borges, L. S. R. Diagnostic Accuracy Measures in Cardiovascular Research. *International Journal of Cardiovascular Sciences* **29**, 218-22 (2016).

8. Cascaes, A. M., Gauche, H., Baramarchi, F. M., Borges, C. M., Peres, K. G. Prematuridade e fatores associados no Estado de Santa Catarina, Brasil, no ano de 2005: análise dos dados do Sistema de Informações sobre Nascidos Vivos. *Cadernos de Saúde Pública* **24**, 1024-1032 (2008).

9. Castro-Costa, E., Ferri, C. P., Measures of effect for cross-sectional studies. *Brazilian Journal of Psychiatry* **30**, 399-408 (2008).

10. Chen, M. H., Dey, D. K., Shao, Q. M. A New Skewed Link Model for Dichotomous Quantal Response Data. Journal of the American Statistical Association **94**, 1172-1186 (1999).

11. Crump, C., Sundquist, K., Sundquist, J., Winkleby, M. A. Gestational age at birth and mortality in young adulthood. *American Medical Association* **306**, 1233-1240 (2011).

12. Czado, C., Santner, T. J. The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference* **33**, 213-231 (1992).

13. Dunn, P. K., Smyth, G. K. Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics* **5**, 236-244 (1996).

14. Fawcett, T. An introduction to ROC analysis. *Pattern recognition letters* **27**, 861-874 (2006).

15. Geman, S., Geman, D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721-741 (1984).

16. Geweke, J. Evaluating the accurating of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics* **4**, 169-193 (1992).

17. Giolo, S. R. *Introdução à análise de dados categóricos com aplicações,* Blucher, (2017).

18. Guimarães, E. A. D. A., Vieira, C. S., Nunes, F. D. D., Januário, G. D. C., Oliveira, V. C. Prevalência e fatores associados à prematuridade em Divinópolis, Minas Gerais, 2008-2011: análise do Sistema de Informações sobre Nascidos Vivos. *Epidemiologia e Serviços de Saúde* **26**, 91-98 (2017).

19. Han, Z., Mulla, S., Beyene, J., Liao, G., McDonald, S. D. Maternal underweight and the risk of preterm birth and low birth weight: a systematic review and meta-analyses. *International Journal of Epidemiology* **40**, 65-101 (2011).

20. Heidelberger, P., Welch, P. D. Simulation Run Length Control in the Presence of an Initial Transient. *Institute for Operations Research and the Management Sciences (Informs)* **31**, 1109-1144 (1983).

21. Hosmer, D. W., Lemeshow, J, S., Sturdivant, R. X. *Applied logistic regression* (Wiley, 2013).

22. Huayanay, A. D. L. C., Bazán, J. L., Cancho, V. G., Dey, D. K. Performance of asymmetric links and correction methods for imbalanced data in binary regression. *Journal of Statistical Computation and Simulation* **89**(9), 1694-1714 (2019).

23. Kuhn, M., Kjell, J. *Applied predictive modeling*, Springer, (2013).

24. Lemonte, A. J., Bazán, J. L. New links for binary regression: an application to coca cultivation in Peru. *Test* **27**, 597-617 (2018).

25. Liu, S., Allen, A., Fraser, W. Fetal and infant health outcomes. In: Preterm Birth Rate. *Canadian Perinatal Health Report*, 123-132 (2008).

26. Mccullagh, P., Nelder, J. A. *Generalized linear models*. 2.ed. Boca Raton London (Chapman and Hall, 1989).

27. Nelder, J. A., Wedderburn, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society - Series A (General)* **135**(3), 370-384 (1972).

28. Oliveira, R. R. *Nascimento prematuro no Estado do Paraná e no município de Maringá*. Tese (doutorado). Maringá: Universidade Estadual de Maringá (2015).

29. Paulino, C. D. M., Turkman, M. A. A., Murteira, B. *Estatística bayesiana* (Fundação Calouste Gulbenkia, (2003).

30. Plummer, M., Best, N., Cowles, K., Vines, K. CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**, 7-11 (2006).

31. R Core Team. *R: A Language and Environment for Statistical Computing* (2022).

32. Robin, X. A., Turck, N., Hainard, A., Tiberti, N., Lisacek, F. *et al.* pROC: an open source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**(77), (2011).

33. Rossi, R. M. *Introdução aos métodos Bayesianos na análise de dados zootécnicos com uso do WinBUGS e R*, Eduem, (2011).

34. Saigal, S., Doyle, L. W. An overview of mortality and sequele of preterm birth from infancy to adulthood. *The Lancet* **371**(9608), 261-269 (2008).

35. Silva, A. N., Anyosa, S., Bazán, J. L. Modelagem bayesiana de regressão binária para dados desbalanceados usando novas ligações. *Revista Brasileira de Biometria* **38**(4), 385-417 (2020).

36. Spiegelhalter, D. J., Best, N. G., Carlin, B. P., Van Der Linde, A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society - Series B (Statistical Methodology)* **64**, 583-639 (2002).

37. Spiegelhalter, D. J., Thomas, A., Best, N., Gilks, W. *BUGS 0.5: Bayesian inference using Gibbs sampling manual (version ii)* (MRC Biostatistics Unit Archive Service, (1996).

38. Thomas, A., O'hara, B., Ligges, U., Sturtz, S. Making BUGS Open. *R News* **6**, 12-17 (2006).

39. Venables, W. N. *Modern applied statistics with S,* Springer, (2002).

40. Who. *World Health Organization*. Born too soon: the global action report on preterm (2012).

41. Who. *World Health Organization*. Preterm birth (2022).