Machine Learning for Mental Health
New Developments and Applications Using Multimodal Data

Agencia Nacional de Investigación y Desarrollo
Gobierno de Chile

UNIVERSIDAD ADOLFO IBÁÑEZ

Workshop on Machine and Statistical Learning with Applications

Binary classification and evaluation metrics using supervised machine learning models to imbalanced data

Jorge L. Bazán, Alex de la Cruz Huayanay

University of São Paulo
Interinstitutional Graduate Program in Statistics (PIPGEs) USP/UFSCar

Santiago 14/06/2023

Contact: jlbazan@icmc.usp.br and <jorgeluisbazan.weebly.com>

Content

1 Introduction

2 Power and reverse power distribution

3 Binary regression with power and reverse power link

4 Comparison of corrections vs Asymmetrical links

5 Performance of metrics of classification for imbalanced data

6 Application

7 Final remarks

Motivation

- Bazán, Romeo and Rodrigues (2014): whether the Warsaw girls was menstruating already (class 1) or not (class 0), where **class 1 =** 58.9%.

- Bazán et al. (2017): full coverage plan for vehicle insurance (full coverage plan = class 1 and not full coverage plan = class = 0), where **class 1 =** 34.7%

- Lemonte and Bazán (2018): eradicating the coca cultivation (erad = class 1 and no erad = class 0), where **class 1 =** 58%

- Huayanay et al. (2019): quality of the white *vinho branco* from Portugal (good quality = class 1 and poor quality = class 0), where **class 1 =** 21.6%

- SILVA, ANYOSA and BAZAN (2020): distinction between oral (class 1) and nasal (class 0) sounds, where **class 1 =** 70.6%. Performance level in mathematics (adequate = class 1 and no adequate = class 0), where **class 1 =** 9.8%

- Huayanay, Bazán and Deniz (2023): the presence of schizophrenia symptoms (class 1) or not (class 0), where **class 1 =** 31%.

1. INTRODUCTION

1. Introduction

- Binary classification models intend to assign an individual or observation to one of two categories or classes, based on a set of attributes.
- There are a number of methods proposed to perform classification, the most used method is logistic regression which uses a link function called logit
- In binary classification, imbalanced data result from the presence of values equal to one (or zero) in a proportion that is significantly less than the corresponding real values of zero (or one).
- In the literature there are some solutions to deal with imbalanced data: correction, asymmetrical links and sampling methods

1.1. Tactics To Combat Imbalanced in Statistics

In the presence of imbalanced data

- Fatourechi et al. (2008), Luque et al. (2019) shown that the binary regression model with the symmetric link, is unsuitable and some metrics may be inappropriate.
- Firth (1993), King and Zeng (2001) propose correction or reduction of bias to the logistic model
- Chen, Dey and Shao (1999), Wang and Dey (2011), Yin et al. (2020): Binary regression considering different assymetrical link functions.
- Nguyen, Zeno and Lars (2011), Hlosta et al. (2013), Huayanay et al. (2019): use of other metrics to assess the predictive capacity of the model.
- Bazán et al. (2017), Lemonte and Bazán (2018): Power and reversal power links for binary regressions.

1.2. Tactics To Combat Imbalanced in Machine Learning

Fernández et al. (2019) and others:

- Can You Collect More Data?
- Try Changing Your Performance Metric
- Try Resampling Your Dataset
- Try Generate Synthetic Samples
- Try Different Algorithms
- Try Penalized Models
- Try a Different Perspective
- Try Getting Creative

1.3. Objectives

- Introduce some of our binary regression models using asymmetric links proposed for when the data set is imbalanced which is an supervised machine learning models are alternative to logistic regression algorithm.
- Evaluate methods developed to deal with imbalanced data and compare them our proposed.
- Study the performance of metrics in imbalanced data using power and reverse power links.

2. POWER AND REVERSE POWER DISTRIBUTION

2. Power and reverse power distribution

According to Bazán et al. (2017)

Definition 1

A random variable Z is said to follow a power and reverse power distribution, in its standard form, when its CDF has the following form, respectively

$$F_P(z) = G\left(z\right)^{\alpha} \quad \text{and} \quad F_{RP}(z) = 1 - G\left(-z\right)^{\alpha}, \quad z \in \mathbb{R},$$

where α is a shape parameter and with $G(\cdot)$ denoting a CDF of baseline distribution with support in the real line.

We use the notation $F_l(\cdot)$ refer to the cumulative density function of power or reverse power distribution, where $l = P,\ RP$

- The probability density functions (PDF): $f_P(z) = \alpha G(z)^{\alpha-1} g(z)$ and $f_{RP}(z) = \alpha G(-z)^{\alpha-1} g(z)$, where $g(\cdot)$ is a PDF of the corresponding baseline distribution and $z \in \mathbb{R}$.
- The quantiles can be written as: $Q_P(p) = G^{-1}\left(-p^{1/\alpha}\right)$ and $Q_{RP}(p) = 1 - G^{-1}\left(-(1-p)^{1/\alpha}\right) = -Q_P(1-p)$, where p is a given probability.

2.1 Some distributions:

- Bazán, Romeo and Rodrigues (2014)
- Bazán et al. (2017),
- Chumbimune (2017),
- Lemonte and Bazán (2018)
- Huayanay et al. (2019)
- Huayanay (2019)
- Huayanay, Bazán and Russo (2023)

Distribution and notation

- Power Logistic: PL
- Reverse Power Logistic: RPL
- Power Normal: PN
- Reverse Power Normal: RPN
- Power Cauchy: PC
- Reverse Power Cauchy: RPC
- Power Reverse Gumbel: PRG
- Reverse Power Reverse Gumbel: RPRG
- Power Laplace: PLA
- Reverse Power Laplace: RPLA

Table 1: CDF, PDF and QF of P and RP link functions

Distribution	$CDF: F_{\alpha}(\eta_i)$	PDF: $f_{\alpha}(\eta_i)$	QF: $Q(p_i, \alpha)$
PL	$\left(\frac{1}{1+e^{-\eta_i}}\right)^{\alpha}$	$\alpha\left(\frac{1}{1+e^{-\eta_i}}\right)^{\alpha-1}\frac{e^{-\eta_i}}{(1+e^{-\eta_i})^2}$	$\log\left(\frac{p_i^{1/\alpha}}{1-p_i^{1/\alpha}}\right)$
RPL	$1-\left(\frac{e^{-\eta_i}}{1+e^{-\eta_i}}\right)^{\alpha}$	$\alpha\left(\frac{1}{1+e^{\eta_i}}\right)^{\alpha-1}\frac{e^{-\eta_i}}{(1+e^{-\eta_i})^2}$	$-\log\left(\frac{(1-p_i)^{1/\alpha}}{1-(1-p_i)^{1/\alpha}}\right)$
PN	$(\Phi(\eta_i))^{\alpha}$	$\alpha(\Phi(\eta_i))^{\alpha-1}\frac{1}{\sqrt{2\pi}}e^{\left\{-\frac{1}{2}\eta_i^2\right\}}$	$\Phi^{-1}\left(p_i^{1/\alpha}\right)$
RPN	$1-(\Phi(-\eta_i))^{\alpha}$	$\alpha(\Phi(-\eta_i))^{\alpha-1}\frac{1}{\sqrt{2\pi}}e^{\left\{-\frac{1}{2}\eta_i^2\right\}}$	$-\Phi^{-1}\left((1-p_i)^{1/\alpha}\right)$
PC	$\left(0.5+\frac{\arctan(\eta_i)}{\pi}\right)^{\alpha}$	$\frac{\alpha}{\pi}\left(\frac{1}{\pi}\arctan(\eta_i)+\frac{1}{2}\right)^{\alpha-1}\frac{1}{(1+\eta_i^2)}$	$\tan\left(\pi\left(p_i^{1/\alpha}-0.5\right)\right)$
RPC	$1-\left(0.5+\frac{\arctan(-\eta_i)}{\pi}\right)^{\alpha}$	$\frac{\alpha}{\pi}\left(\frac{1}{\pi}\arctan(-\eta_i)+\frac{1}{2}\right)^{\alpha-1}\frac{1}{(1+\eta_i^2)}$	$-\tan\left(\pi\left((1-p_i)^{1/\alpha}-0.5\right)\right)$
PRG	$\left(1-e^{-e^{\eta_i}}\right)^{\alpha}$	$\alpha\left(1-e^{-e^{\eta_i}}\right)^{\alpha-1}e^{-(-\eta_i+e^{\eta_i})}$	$\log\left(-\log\left(1-p_i^{1/\alpha}\right)\right)$
RPRG	$1-\left(1-e^{-e^{-\eta_i}}\right)^{\alpha}$	$\alpha\left(1-e^{-e^{-\eta_i}}\right)^{\alpha-1}e^{-(-\eta_i+e^{\eta_i})}$	$-\log\left(-\log\left(1-(1-p_i)^{1/\alpha}\right)\right)$
PLA	$\left\{\frac{1}{2}+\frac{\mathsf{sign}(\eta)}{2}\left[1-e^{-\lvert\eta\rvert}\right]\right\}^{\alpha}$	$\frac{\alpha}{2}\left\{\frac{1}{2}+\frac{\mathsf{sign}(\eta)}{2}\left[1-e^{-\lvert\eta\rvert}\right]\right\}^{\alpha-1}e^{-\lvert\eta\rvert}$	$\mathsf{sign}(p^{\frac{1}{\alpha}}-0.5)\ln\left(1-2\left\lvert p^{\frac{1}{\alpha}}-0.5\right\rvert\right)$
RPLA	$1-\left\{\frac{1}{2}-\frac{\mathsf{sign}(\eta)}{2}\left[1-e^{-\lvert\eta\rvert}\right]\right\}^{\alpha}$	$\frac{\alpha}{2}\left\{\frac{1}{2}-\frac{\mathsf{sign}(\eta)}{2}\left[1-e^{-\lvert\eta\rvert}\right]\right\}^{\alpha-1}e^{-\lvert\eta\rvert}$	$\mathsf{sign}(0.5-(1-p)^{\frac{1}{\alpha}})\ln\left(1-2\left\lvert(1-p)^{\frac{1}{\alpha}}-0.5\right\rvert\right)$

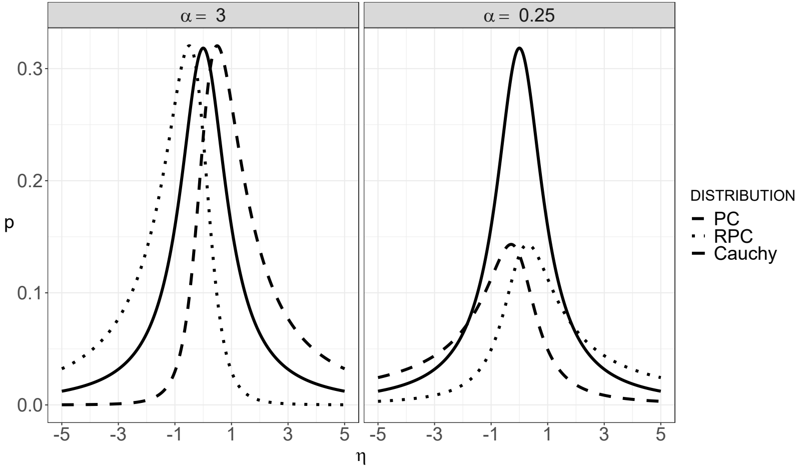

Figure 1: Density of probabilities of Cauchy, PC and RPC distribution with $\alpha = 0.5$ (left) and $\alpha = 3$ (right).

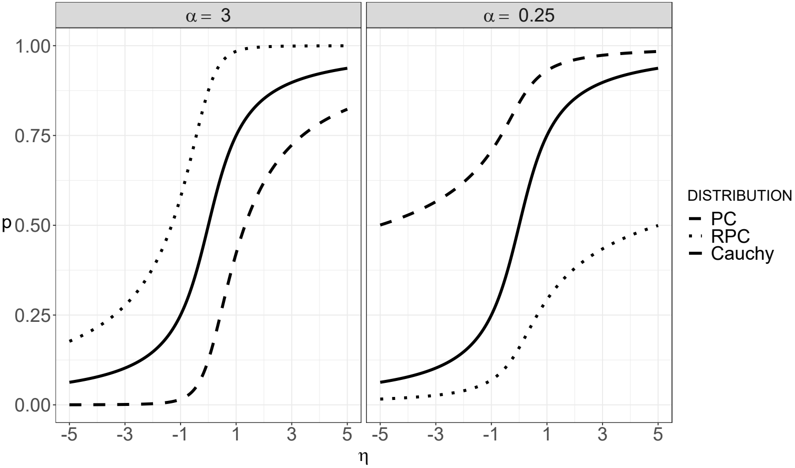

Figure 2: Curve of probabilities of Cauchy, PC and RPC distribution with $\alpha = 0.5$ (left) and $\alpha = 3$ (right).

Proposition 1

Let $U \sim$ *Uniform*$(0,1)$ *be, then* $X = Q_P(U) = G^{-1}(U^{1/\alpha})$ *follow the P distribution and* $X = Q_{RP}(U) = -G^{-1}((1-U)^{1/\alpha})$ *follow the RP distribution, where* $Q_P(U)$ *and* $Q_P(U)$ *are the values of the quantiles for* U *generated respectively by* $F_P(\cdot)$ *and* $F_{RP}(\cdot)$.

This is a direct consequence of the definition of this class of distributions and of that for continuous distributions, $F_l(X) = U$ follows a continuous uniform distribution (ROSS, 2006).

2.3. Skewness and Kurtosis

- Octile skewness coefficient : $A_O(\alpha) = \frac{(O_7 - O_4) - (O_4 - O_1)}{O_7 - O_1}$ in Brys, Hubert and Struyf (2004), where O_a denote the a^{th} octile.
- Kurtosis coefficient: $K_O = \frac{(O_7 - O_5) + (O_3 - O_1)}{O_6 - O_2}$ in Moors (1988)

In Power and Reverse Power distribution:

- $A_O(\alpha) = \frac{Q(0.875,\alpha) - 2Q(0.5,\alpha) + Q(0.125,\alpha)}{Q(0.875,\alpha) - Q(0.125,\alpha)}$
- $K_O(\alpha) = \frac{100}{1.233} \times \left[\frac{Q(0.875,\alpha) - Q(0.625,\alpha) + Q(0.375,\alpha) - Q(0.125,\alpha)}{Q(0.75,\alpha) - Q(0.25,\alpha)} - 1.233\right]$

we use the kurtosis value of the Normal (1.233) for rescale this measure. $A_O(\alpha)$ and $K_O(\alpha)$ depend of α parameter, $Q(p,\alpha)$ is the quantile function and $0 < A_O < 1$.

Table 2: Skewness $A_O(\alpha)$ for power distributions and reversal power distributions considering values between $\alpha = 0.001$ and $\alpha = 9999$.

	$A_O(\alpha)$				
Distribution	$Min.$	$Max.$	$0 < \alpha < 1$	$\alpha \geq 1$	r *
PL	-0.4248	0.1997	(-0.4248, 0.0000)	[0.0000, 0.1997)	0.6245
RPL	-0.1997	0.4248	(0.0000, 0.4248)	[-0.1997, 0.0000)	0.6245
PN	-0.1282	0.1617	(-0.1282, 0.0000)	[0.0000, 0.1617)	0.2899
RPN	-0.1617	0.1282	(0.0000, 0.1282)	[-0.1617, 0.0000)	0.2899
PC	-1.0000	0.7255	(-1.0000, 0.0000)	[0.0000, 0.7255)	1.7255
RPC	-0.7255	1.0000	(0.0000, 1.0000)	[-0.7255, 0.0000)	1.7255
PRG	-0.4219	0.1312	(-0.4219, -0.1998)	[-0.1998, 0.1312)	0.5531
RPRG	-0.1312	0.4219	(0.1998, 0.4219)	[-0.1312, 0.1996)	0.5531
PLA	-0.4248	0.2000	(-0.4248, 0.0000)	[0.1998, 0.2000)	0.2248
RPLA	-0.2000	0.4248	(0.0000, 0.4248)	[-0.2000, -0.1998)	0.2248

* $r = max. - min.$

Table 3: Kurtosis $K_O(\alpha)$ for power and reverse power distributions considering values between $\alpha = 0.001$ and $\alpha = 9999$.

Distribution	$K_O(\alpha)$				
	Min	Max	$0 < \alpha < 1$	$\alpha \geq 1$	r *
PL-RPL	3.6597	10.1144	(5.9426; 10.1144)	(3.6597; 5.9423)	6.4547
PN-RPN	-2.0454	1.9217	(-2.0454; 0.0077)	(0.0078; 1.9217)	3.9671
PC-RPC	56.5886	37527838.9177	(62.2085; 37527838.9177)	(56.5886; 73.7012)	37527782.3291
PRG-RPRG	0.2981	7.7168	(3.6582; 7.7168)	(0.2981; 0.8706)	7.4187
PLA-RPLA	3.5934	28.8118	(5.9424; 28.8118)	(3.5934; 3.6563)	25.2184

* $r = max. - min.$

3. BINARY REGRESSION WITH P AND RP LINK

3. Binary regression with P and RP link

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ be an $n \times 1$ vector of independent response variables with values 1 or 0 and $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})^\top$ be the vector of covariates. The Bayesian model of binary regression with power or reverse power link function can be given by

$$\begin{aligned} Y_i \mid \boldsymbol{\beta}, \alpha &\overset{ind.}{\sim} \text{Bernoulli}(p_i) \\ p_i &= F_l(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \\ (\boldsymbol{\beta}, \alpha)^\top &\sim \pi(\boldsymbol{\beta}, \alpha) \end{aligned} \tag{1}$$

Where $F_l(\cdot)$ is the P or RP distribution given in Table 1. As it was considered in Bazán et al. (2017) and Huayanay et al. (2019):

- $\boldsymbol{\beta} \sim N_p(\mathbf{0}, \boldsymbol{I}\sigma_\beta^2)$.
- $\delta = \log(\alpha) \sim U(-2, 2)$
- $\boldsymbol{\beta}$ and α are considered independent: $\pi(\boldsymbol{\beta}, \alpha) = \pi(\boldsymbol{\beta}) \times \pi(\alpha)$.

Considering the prior specification here, the posterior distribution of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha)^\top$, $\pi(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X})$ has the following form

$$\pi(\boldsymbol{\beta}, \alpha \mid \boldsymbol{y}, \boldsymbol{X}) \propto \prod_{i=1}^{n} \left[F_l\left(\boldsymbol{x_i}^\top \boldsymbol{\beta}\right)\right]^{y_i} \left[1 - F_l\left(\boldsymbol{x_i}^\top \boldsymbol{\beta}\right)\right]^{1-y_i} \prod_{j=1}^{p} \exp\left\{-\frac{\beta_j^2}{2\left(10^2\right)}\right\} \frac{1}{4\alpha}$$

To estimate the parameters of the models, we use a proper code using the `Stan` language through `Python` using the `Pystan` package.

4. COMPARISON OF CORRECTIONS VS ASYMMETRICAL LINKS

4. Comparison of corrections vs Asymmetrical links

A simulation study was carried out to evaluate the performance of asymmetric ligations for unbalanced data in comparison with the methods of Firth (1993) and King and Zeng (2001).
The unbalanced data are generated from the model with Cauchy power link.

$$Y_i \sim \text{Bernoulli}\,(u_i)$$

where

$$\mu_i = \left(\frac{1}{\pi} \arctan\left(\beta_1 + \beta_2 x_i\right) + \frac{1}{2} \right)^{\alpha}$$

$x \sim U(-3, 3)$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (0, 1)^\top$, $n = (500; 5000; 20000)$ e $\alpha = (1/4; 1/2; 2; 4)$
this defines 12 scenarios, each one with$M = 100$ replicas.

Note that the values of $\alpha = (1/4;\ 1/2;\ 2;\ 4)$ determines approximately the correspondent observed proportions of ones $\hat{p} = (0.20;\ 0.34;\ 0.67;\ 0.80)$. Then we fit the following method for the simulated data

- Logistic regression using glm in R.
- Logistic regression with correction KZ (LogisticKZ) from King and Zeng (2001) using Zelig in R.
- Logistic regression with correction F (LogisticF) from Firth (1993) using LogisticF in R.
- Asymmetrical links studied here using Pystan (TEAM, 2017) on the interface between Stan and Python

The performance of the methods was determined based on a measure of bias and root mean square error (RMSE) of the estimates over the replications.

Table 4: β_1 estimates for correction methods and using asymmetrical link functions by n

Bias	n	p = 0.20			p = 0.34			p = 0.67			p = 0.80		
		Estimate	Bias	RSME	Estimate	Bias	RSME	Estimate	Bias	RSME	Estimate	Bias	RSME
Logistic	500	-2.457	-2.457	2.471	-1.168	-1.168	1.178	0.966	0.966	0.971	1.765	1.765	1.771
	5000	-2.454	-2.454	2.454	-1.171	-1.171	1.172	0.934	0.934	0.934	1.748	1.748	1.748
	20000	-2.429	-2.429	2.429	-1.172	-1.172	1.172	0.936	0.936	0.936	1.731	1.731	1.731
LogisticKZ	500	-2.377	-2.377	2.383	-1.144	-1.144	1.147	0.910	0.910	0.911	1.729	1.729	1.731
	5000	-2.423	-2.423	2.424	-1.164	-1.164	1.164	0.940	0.940	0.941	1.736	1.736	1.736
	20000	-2.454	-2.454	2.454	-1.175	-1.175	1.175	0.938	0.938	0.938	1.742	1.742	1.742
LogisticF	500	-2.369	-2.369	2.376	-1.147	-1.147	1.153	0.900	0.900	0.906	1.712	1.712	1.716
	5000	-2.425	-2.425	2.425	-1.163	-1.163	1.164	0.931	0.931	0.932	1.735	1.735	1.736
	20000	-2.446	-2.446	2.446	-1.173	-1.173	1.173	0.936	0.936	0.937	1.740	1.740	1.740
PL	500	0.635	0.635	0.637	0.557	0.557	0.561	-0.540	-0.540	0.571	0.519	0.519	0.526
	5000	0.406	0.406	0.406	0.130	0.130	0.187	-0.187	-0.187	0.197	-0.289	-0.289	0.399
	20000	-0.389	-0.389	0.399	0.124	0.124	0.182	-0.174	-0.174	0.176	-0.392	-0.392	0.394
PP	500	-1.008	-1.008	1.017	-1.080	-1.080	1.236	-0.198	-0.198	0.325	0.434	0.434	0.470
	5000	0.326	0.326	0.329	0.582	0.582	0.615	-0.430	-0.430	0.514	0.084	0.084	0.209
	20000	0.501	0.501	0.501	0.852	0.852	0.525	-0.502	-0.502	0.500	-0.219	-0.219	0.190
PLaplace	500	-0.190	-0.190	1.064	0.031	0.031	0.425	-0.393	-0.393	0.660	0.363	0.363	0.729
	5000	0.150	0.150	0.185	0.048	0.048	0.103	0.657	0.657	0.657	0.545	0.545	0.640
	20000	0.148	0.148	0.162	0.042	0.042	0.057	0.416	0.416	0.417	0.477	0.477	0.477
PC	500	-0.004	-0.004	0.067	0.029	0.029	0.073	-0.295	-0.295	0.308	0.381	0.381	0.534
	5000	0.021	0.021	0.056	-0.013	-0.013	0.040	-0.022	-0.022	0.041	-0.071	-0.071	0.188
	20000	-0.001	-0.001	0.023	0.014	0.014	0.018	-0.003	-0.003	0.041	-0.044	-0.044	0.085

Table 5: β_2 estimates for correction methods and using asymmetrical link functions by n

Method	n	p = 0.20			p = 0.34			p = 0.67			p = 0.80		
		Estimate	Bias	RSME	Estimate	Bias	RSME	Estimate	Bias	RSME	Estimate	Bias	RSME
Logistic	500	1.268	0.268	0.319	1.127	0.127	0.161	0.713	-0.287	0.303	0.633	-0.367	0.375
	5000	1.254	0.254	0.257	1.094	0.094	0.100	0.703	-0.297	0.298	0.607	-0.393	0.394
	20000	1.240	0.240	0.243	1.101	0.101	0.100	0.701	-0.299	0.300	0.608	-0.392	0.392
LogisticKZ	500	1.185	0.185	0.234	1.065	0.065	0.103	0.669	-0.331	0.338	0.616	-0.384	0.394
	5000	1.229	0.229	0.235	1.093	0.093	0.099	0.714	-0.286	0.287	0.613	-0.387	0.388
	20000	1.249	0.249	0.250	1.105	0.105	0.107	0.705	-0.295	0.295	0.618	-0.382	0.382
LogisticF	500	1.180	0.180	0.209	1.054	0.054	0.090	0.665	-0.335	0.342	0.606	-0.394	0.402
	5000	1.229	0.229	0.232	1.089	0.089	0.096	0.713	-0.287	0.288	0.612	-0.388	0.389
	20000	1.244	0.244	0.244	1.103	0.103	0.104	0.705	-0.295	0.295	0.617	-0.383	0.383
PL	500	0.771	-0.229	0.230	0.812	-0.188	0.189	1.109	0.109	0.119	0.888	-0.112	0.212
	5000	0.858	-0.142	0.142	0.904	-0.096	0.110	0.923	-0.077	0.082	0.940	-0.060	0.107
	20000	0.906	-0.094	0.097	0.913	-0.087	0.087	0.951	-0.049	0.064	0.980	-0.020	0.061
PP	500	0.873	-0.127	0.165	0.738	-0.262	0.284	0.567	-0.433	0.440	0.443	-0.557	0.559
	5000	0.838	-0.562	0.162	0.429	-0.571	0.572	0.595	-0.405	0.409	0.499	-0.502	0.503
	20000	0.852	-0.148	0.149	0.390	-0.610	0.567	0.607	-0.393	0.397	0.547	-0.453	0.454
PLaplace	500	0.688	-0.313	0.429	0.667	-0.333	0.346	0.815	-0.185	0.271	0.647	-0.353	0.382
	5000	0.602	-0.398	0.398	0.613	-0.388	0.388	0.754	-0.246	0.246	0.573	-0.427	0.432
	20000	0.596	-0.404	0.404	0.607	-0.393	0.393	0.755	-0.246	0.250	0.616	-0.384	0.384
PC	500	1.222	0.222	0.234	1.102	0.102	0.108	1.384	0.384	0.388	1.333	0.333	0.347
	5000	1.049	0.049	0.062	1.009	0.009	0.016	1.037	0.037	0.064	1.024	0.024	0.101
	20000	1.010	0.010	0.014	1.009	0.009	0.008	0.998	-0.002	0.020	1.005	0.005	0.047

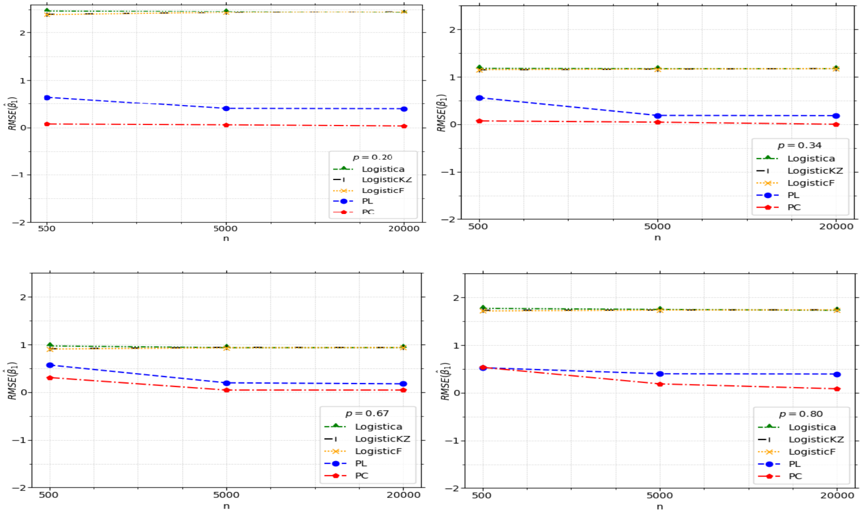

Figure 3: RMSE for β_1 with different estimation methods and sample sizes.

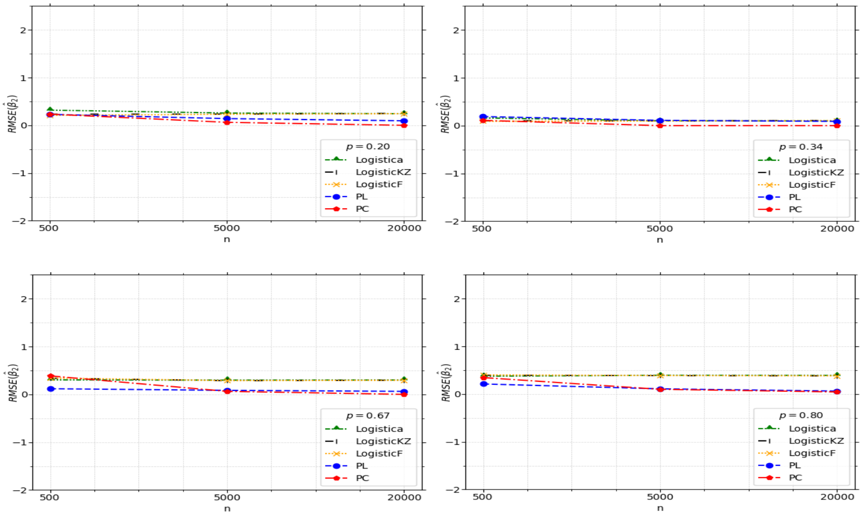

Figure 4: RMSE for β_2 with different estimation methods and sample sizes.

- The Logistic regression model presents higher Bias and higher RMSE, which means that it is not suitable for imbalanced data.
- The correction methods from Firth (1993) and King and Zeng (2001) We present better Estimates than the Logistic regression, but as the size of the sample increases, the difference is not significant in relation to the RMSE and Bias.
- The models with ligation power present the best performance, being the PC or model that outperforms the others as the size of the sample increases or RMSE is approximately zero.

5. PERFOMANCE OF METRICS OF CLASSIFICATION FOR IMBALANCED DATA

5. Performance of metrics of classification for imbalanced data

The objective is to evaluate the performance metrics for binary regression in the presence of imbalanced data

$$Y_i \sim \text{Bernoulli}\left(p_i\right), \quad p_i = \left(\frac{1}{\pi}\arctan\left(\beta_1 + \beta_2 x_i\right) + \frac{1}{2}\right)^{\alpha}$$

- $x_i \sim U(-2, 2)$
- $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (-0.5, 1.5)^\top$.
- 100 replications for different scenarios are realized:
    - 2 sample sizes $n = (5000; 10000)$
    - $\alpha = 3$ and $\alpha = 0.25$ ($p = 0.15$ and $p = 0.76$ respectively)
- Binary regression models with PC and L link were fitted for each data set.

- For each estimated model, in each replication, the confusion matrix was observed and then the metrics were computed.
- The confusion matrix was built considering an optimal value of threshold.
- To decide what is the optimal threshold to be considered, we use the correspondent threshold that produce the maximum Cohen's kappa for the true model, as suggested in Zou et al. (2016).

Table 6: Metrics in binary classification

Metric	Notation	Formula	Range
Accuracy	ACC	$\frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$	$[0;1]$
Sensitivity	TPR	$\frac{\mathrm{TP}}{\mathrm{TP+FN}}$	$[0;1]$
Specificity	TNR	$\frac{\mathrm{TN}}{\mathrm{TN+FP}}$	$[0;1]$
Critical success index	CSI	$\dfrac{TP}{TP+FP+FN}$	$[0;1]$
Sokal & Sneath index	SSI	$\dfrac{TP}{TP+2\times FP+2\times FN}$	$[0;1]$
Faith index	$FAITH$	$\dfrac{TP+0.5\times TN}{TP+FP+FN+TN}$	$[0;1]$
Pattern difference	$PDIF$	$\dfrac{4\times FP\times FN}{(TP+FP+FN+TN)^2}$	$[0;1]$
Gilbert skill score	GS	$\frac{(TP\times TN-FP\times FN)}{(FN+FP)(TP+FP+FN+TN)+(TP\times TN-FP\times FN)}$	$[0;1]$
Matthews Correlation Coefficient	MCC	$\frac{(TP\times TN-FP\times FN)}{\sqrt{(FN+FP)(TP+FN)(TN+FP)(TN+FN)}}$	$[0;1]$
G-Mean	GM	$\sqrt{TPR\times TNR}$	$[0;1]$
F_1-score	$F1$	$2\times\dfrac{TNR\times TPR}{TNR+TPR}$	$[0;1]$
Cohen's kappa	$KAPPA$	$\frac{2\times(TP\times TN-FP\times FN)}{(TP+FP)(FP+TN)+(TP+FN)(FN+TN)}$	$[0;1]$

The comparative performance of the metrics was determined based on the

- Show the distance between the curves of the models based on the metric considered. (Since that the PC is the true model we hope that the metric show a high distance between the model with PC and logistic link).
- Kolmogorov test to identify if the curves are differents: $D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$
  (Since that the PC is the true model we hope reject the test that curve of metric between the model with PC and logistic link are equals).
- Proportion of times that the metric value PC model is better: $\hat{p} = \dfrac{1}{R} \sum_{l=1}^{R} I\left\{m_1^{(r)} < m_2^{(r)}\right\}$
  (Since that the PC is the true model we hope the metric chose the true model between the model with PC and logistic link in 100% of the times).

where $m_1^{(r)}$ and $m_1^{(r)}$ are metric values for two different models, $F_{1,n}$ and $F_{2,m}$ are the empirical distribution functions of the first and second samples of size m and n respectively, for the rth replica and R is the number of replicas.

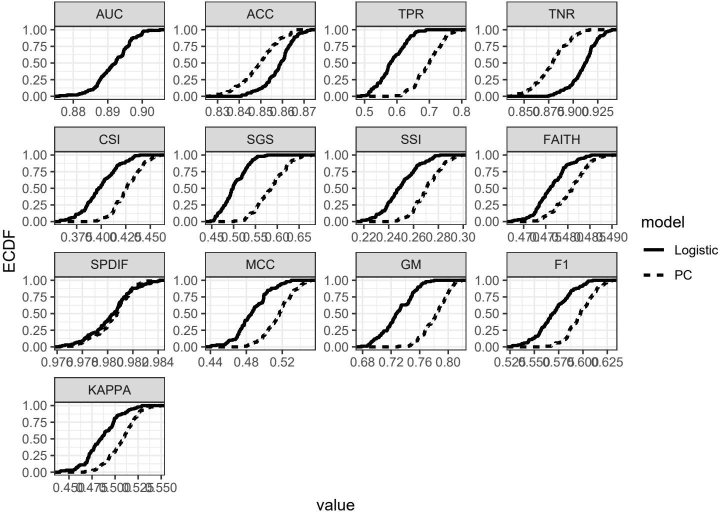

Figure 5: Empirical cumulative distribution function of metrics for logit and PC link in imbalanced data, to $\alpha = 3$ $(p = 0.15)$ and $n = 5000$.

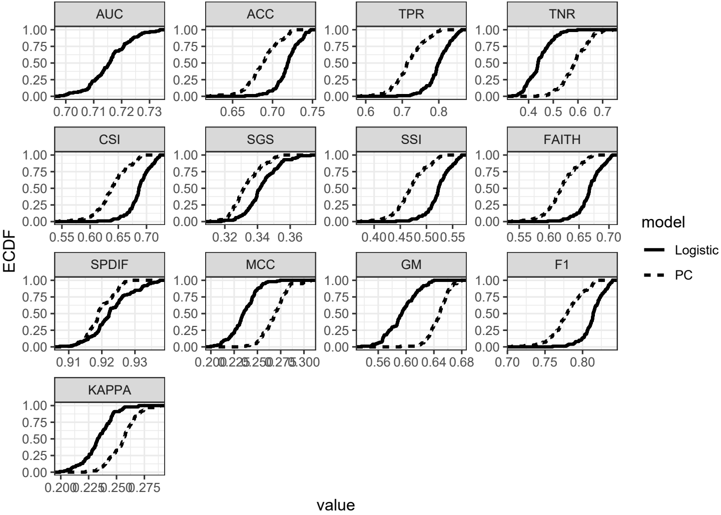

Figure 6: Empirical cumulative distribution function of metrics for logit and PC link in imbalanced data, to $\alpha = 0.25$ $(p = 0.76)$ and $n = 5000$.

Table 7: Kolmogorov test (KT) with p-value (p.val) between the metrics of the PC and logistic links for imbalanced data

Metric	$\alpha = 3(p = 0.15)$				$\alpha = 0.25(p = 0.76)$			
	$n = 5000$		$n = 10000$		$n = 5000$		$n = 10000$	
	KT	p.val	KT	p.val	KT	p.val	KT	p.val
AUC	**0.000**	**1.000**	**0.000**	**1.000**	0.000	**1.000**	**0.000**	**1.000**
ACC	0.470	0.000	0.710	0.000	0.670	0.000	0.870	0.000
TPR	0.800	0.000	0.900	0.000	0.730	0.000	0.910	0.000
TNR	0.710	0.000	0.860	0.000	0.820	0.000	0.940	0.000
CSI	0.610	0.000	0.680	0.000	0.710	0.000	0.890	0.000
SGS	0.800	0.000	0.900	0.000	0.430	0.000	0.640	0.000
SSI	0.610	0.000	0.680	0.000	0.710	0.000	0.890	0.000
FAITH	0.490	0.000	0.510	0.000	0.720	0.000	0.890	0.000
SPDIF	**0.110**	**0.581**	**0.160**	**0.155**	0.300	0.000	0.470	0.000
MCC	0.620	0.000	0.680	0.000	0.800	0.000	0.830	0.000
GM	0.790	0.000	0.910	0.000	0.860	0.000	0.950	0.000
F1	0.610	0.000	0.680	0.000	0.710	0.000	0.890	0.000
KAPPA	0.510	0.000	0.550	0.000	0.640	0.000	0.690	0.000

Table 8: Proportion of times the metric chose the correct model on imbalanced data

Metric	$\alpha = 3 (p = 0.15)$		$\alpha = 0.25 (p = 0.76)$	
	$n = 5000$	$n = 10000$	$n = 5000$	$n = 10000$
AUC	**0%**	**0%**	**0%**	**0%**
ACC	**1%**	**0%**	**0%**	**0%**
TPR	100%	100%	**0%**	**0%**
TNR	**0%**	**0%**	100%	100%
CSI	100%	100%	**0%**	**0%**
SGS	100%	100%	**0%**	**0%**
SSI	100%	100%	**0%**	**0%**
FAITH	100%	100%	**0%**	**0%**
SPDIF	**54%**	**41%**	**26%**	**19%**
MCC	100%	100%	100%	100%
GM	100%	100%	100%	100%
F1	100%	100%	**0%**	**0%**
KAPPA	100%	100%	100%	100%

When select the right model for imbalanced data

- The metrics AUC, ACC, TNR, and $SPDIF$, are not adequate to evaluate the performance of the model with low proportion of sucess.
- The metrics AUC, ACC, TPR, CSI, SGS, SSI, $FAITH$, $SPDIF$ and $F1$, are not adequate to evaluate the performance of the model with high proportion of success.
- The metrics TPR, CSI, SGS, SSI, $FAITH$, MCC, $\mathbf{GM}$, $F1$ and $KAPPA$ are adequate to evaluate the performance of the model with low proportion of success.
- The metrics TNR, MCC, $\mathbf{GM}$, and $KAPPA$ are adequate to evaluate the performance of the model with high proportion of success.

The Geometric Mean (G-Mean) or GM is a metric that measures the balance between classification performances on both the majority and minority classes.

6. APPLICATION

6. Application

The data set analyzed here is related to the Shill bidding, which is available in UCI repository (DUA; TANISKIDOU, 2017).

- Shill bidding is when a seller uses a fraudulent account to bid on their bidding and artificially increase the bidding price (ALZAHRANI; SADAOUI, 2018).
- The auctioneers can be classified in normal or suspicious behavior, this classification will be considered as a response variable
  - $Y = 0$ if the auctioneer has normal behavior
  - $Y =$ 1, if the auctioneer is suspicious.
- In addition to the response variable, the data set contains 9 attributes, but only 4 were considered, where all are numerical.

The proportion of ones is 10.7%, so the dataset is imbalanced.

The original data set contains 6331 observations and was divided into 2 subsets

- Training data set: 75% of the data was used to estimate the models.
- Test data set: 25% of the data were used to make predictions from the estimated model with the training data.
- The Training data set were used to estimate the parameters on the models and then predictions were made using Test data set data
- Since that low proportion was observed we fit the different Power links studied here with the training data set
- By considering the estimated of the training we use it with the test data set to obtain an optimal Threshold using Cohen's kappa.

Table 9: Metrics of asymmetrical power links for Shill Bidding, for Test data set

	Link					
Metric	L	PL	PN	PC	PRG	PLA
TPR	0.972	0.945	0.740	**1.000**	0.978	0.956
CSI	0.838	0.818	0.713	**0.858**	0.843	0.824
SGS	0.954	0.911	0.667	**1.000**	0.963	0.928
SSI	0.721	0.692	0.554	**0.751**	0.728	0.700
FAITH	0.545	0.542	0.525	**0.548**	0.546	0.543
MCC	0.902	0.888	0.821	**0.916**	0.905	0.892
GM	0.976	0.962	0.858	**0.989**	0.979	0.967
F1	0.912	0.900	0.832	**0.923**	0.915	0.903
KAPPA	0.900	0.886	0.814	**0.913**	0.903	0.890
Threshold	0.285	0.446	0.358	0.125	0.256	0.453

- We use the metrics indicated to low proportions following the previous study
- By considering the metrics the best model was the Binary regression using PC link
- The regression coefficients are showed on the next slide
- Note that the credible interval for the shape parameter do not include the value 1 and this is upper 1, indicating that this parameter explain the unbalancing on the data.

Table 10: Posterior parameter estimation for the binary response model with a PC link for Shill Bidding data

Variable	Parameter	Estimate	SD	95% IC	
Intercept	β_1	-13.925	3.467	(-20.994;	-7.631)
Bidder Tendency	β_2	2.045	0.597	(0.931;	3.287)
successive Outbidding	β_3	9.891	2.264	(5.366;	14.060)
Winning Ratio	β_4	5.303	1.209	(3.017;	7.707)
Auction Duration	β_5	1.108	0.414	(0.379;	1.993)
Shape parameter	α	5.9306	1.1467	(3.077;	7.339)

Table 11: Posterior parameter estimation for the binary response model with a Logistic link for Shill Bidding data

Variable	Parameter	Estimate	SD	95% IC	
Intercept	β_1	-7.014	0.500	(-8.057;	-6.070)
Bidder Tendency	β_2	0.271	0.104	(0.064;	0.471)
successive Outbidding	β_3	3.097	0.179	(2.763;	3.460)
Winning Ratio	β_4	2.603	0.314	(1.993;	3.227)
Auction Duration	β_5	0.223	0.130	(-0.020;	0.486)

7. FINAL REMARKS

7. Final remarks

- We show that asymmetry and kurtosis of Power and Reversal power distributions depend on α. They produce different possible links for binary regression.
- Our first simulation study showed that on the presence of unbalancing data, the Logistic link and corrections of this links are no appropriated (bias on the intercept). Then asymmetrical links can be a good alternative for this data separating the effect of the intercept with the effect of the curve associated with the shape of the distribution.
- Our second simulation study showed that some commonly used metrics for binary classification (AUC, ACC, and TNR) may not be the most adequate to choose the best model when the data is imbalanced. Other metrics can be recommendable depending if the we have lower or higher observed proportions of ones.
- In the application it was shown that according to the appropriate metrics for imbalanced data, a model with a power link presented better performance to describe the Shill Bidding data set.

- The properties studied in this work, show that P and RP links are good alternatives for imbalanced binary classification problems.
- Extensions for Binomial regression, Mixed regression models and Item Response Theory of some of this links had been studied recently in Alves, Bazán and Arellano-Valle (2023), Huayanay, Bazán and Deniz (2023) and Bazán et al. (2023) respectively.
- New estimation methods are necessaries for the the estimation of binary regression with asymmetrical links can be more quickly in High dimension data sets.
- Other kind of priors as the studied in Ordoñez et al. (2023) can be considered for the α parameter.
- We are conducting studies comparing the performance of asymmetrical links in binary regression with some classification algorithms as Naive Bayes, K-Nearest Neighbors, Decision Tree, Support Vector Machines and Random forest.
- We are planing studies comparing the performance of asymmetrical links in binary regression with some strategies of re-sampling and cross-validation.

Collaborations

- Adriano Suzuki: *University of São Paulo*
- Adson N. da Silva: *University of São Paulo*
- Alex de la Cruz: *University of São Paulo*
- Artur J. Lemonte: *Federal University of Rio Grande do Norte*
- Caio Azevedo: *University of Campinas*
- Cibele M Russo: *University of São Paulo*
- Dipak K. Dey: *University of Connecticut*
- Fabiano R. Coelho: *University of São Paulo*
- Francisco Louzada: *University of São Paulo*
- Francisco Torres: *University of Santiago, Chile*
- Jessica SB Alves: *University of São Paulo*
- José S. Romeo: *Massey University*
- José A. Ordoñez: *University of Campinas*
- Jossemar Rodrigues: *University of São Paulo*
- Marcos Prates: *University of Minas Gerais*
- Reinaldo Arellano-Valle: *Pontifical Catholic University of Chile*
- Sandra E. Flores: *University of São Paulo*
- Vicente Cancho: *University of São Paulo*
- Victor Lachos: *University of Connecticut*

I still don't understand this concept but everybody is invited to work to became it understandable!.

MUITO OBRIGADO!

Interested in our Program? visit <https://www.pipges.ufscar.br>

Acknowledgments

The first author was partially supported by FAPESP (2021/11720-0) and by research productivity grant from the CNP (309809/2022-30). The second author is grateful for the support from CAPES-Brazil and Inter-institutional Graduate Program in Statistics (PIPGEs) USP/UFSCar.

Reference I

- ALVES, J. S.; BAZÁN, J. L.; ARELLANO-VALLE, R. B. Flexible cloglog links for binomial regression models as an alternative for imbalanced medical data. *Biometrical Journal*, Wiley Online Library, v. 65, n. 3, p. 2100325, 2023.
- ALZAHRANI, A.; SADAOUI, S. Scraping and preprocessing commercial auction data for fraud classification. *arXiv preprint arXiv:1806.00656*, 2018.
- BAZÁN, J.; ROMEO, J.; RODRIGUES, J. Bayesian skew-probit regression for binary response data. *Brazilian Journal of Probability and Statistics*, Brazilian Statistical Association, v. 28, n. 4, p. 467–482, 2014.
- BAZÁN, J. et al. Power and reversal power links for binary regressions: An application for motor insurance policyholders. *Applied Stochastic Models in Business and Industry*, Wiley Online Library, v. 33, n. 1, p. 22–34, 2017.
- BAZÁN, J. L. et al. Revisiting the Samejima–Bolfarine–Bazán IRT models: New features and extensions. *Brazilian Journal of Probability and Statistics*, Brazilian Statistical Association, v. 37, n. 1, p. 1 – 25, 2023. Available: <https://doi.org/10.1214/22-BJPS558>.
- BRYS, G.; HUBERT, M.; STRUYF, A. A robust measure of skewness. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 13, n. 4, p. 996–1017, 2004.
- 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*, Taylor & Francis, v. 94, n. 448, p. 1172–1186, 1999.
- CHUMBIMUNE, S. A. *Regressão binária usando ligações potência e reversa de potência*. Master's Thesis (Master Thesis) — University of São Paulo, 2017.

Reference II

- DUA, D.; TANISKIDOU, E. K. *UCI Machine Learning Repository*. 2017. Available: <https://archive.ics.uci.edu/dataset/562/shill+bidding+dataset>.
- FATOURECHI, M. et al. Comparison of evaluation metrics in classification applications with imbalanced datasets. In: IEEE. *2008 seventh international conference on machine learning and applications*. [S.l.], 2008. p. 777–782.
- FERNáNDEZ, A. et al. *Learning from Imbalanced Data Sets*. [S.l.]: Springer, 2019.
- FIRTH, D. Bias reduction of maximum likelihood estimates. *Biometrika*, Oxford University Press, v. 80, n. 1, p. 27–38, 1993.
- HLOSTA, M. et al. Constrained classification of large imbalanced data by logistic regression and genetic algorithm. *International Journal of Machine Learning and Computing*, IACSIT Press, v. 3, n. 2, p. 214, 2013.
- HUAYANAY, A. De la C. *Modelos de regressão para resposta binária na presença de dados desbalanceados*. Master's Thesis (Master's Thesis) — University of São Paulo, 2019.
- HUAYANAY, A. de la C.; BAZÁN, J. L.; DENIZ, C. Longitudinal binary response models using alternative links to medical data. *Brazilian Journal of Probability and Statistics*, Brazilian Statistical Association, X, n. X, p. XXX, 2023.
- HUAYANAY, D. la C. et al. Performance of asymmetric links and correction methods for imbalanced data in binary regression. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 89, n. 9, p. 1694–1714, 2019.
- HUAYANAY, D. la C.; BAZáN, J. L.; RUSSO, C. *Performance of evaluation metrics for classification in imbalanced data*. 2023. Submitted for publication to *LACS23* June 2023.

Reference III

- KING, G.; ZENG, L. Logistic regression in rare events data. *Political analysis*, Cambridge University Press, v. 9, n. 2, p. 137–163, 2001.
- LEMONTE, A. J.; BAZÁN, J. L. New links for binary regression: an application to coca cultivation in peru. *TEST*, v. 27, n. 3, p. 597–617, Sep 2018. ISSN 1863-8260. Available: <https://doi.org/10.1007/s11749-017-0563-1>.
- LUQUE, A. et al. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, Elsevier, v. 91, p. 216–231, 2019.
- MOORS, J. A quantile alternative for kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, Wiley Online Library, v. 37, n. 1, p. 25–32, 1988.
- NGUYEN, T.-N.; ZENO, G.; LARS, S.-T. A new evaluation measure for learning from imbalanced data. In: IEEE. *The 2011 International Joint Conference on Neural Networks*. [S.l.], 2011. p. 537–542.
- ORDOÑEZ, J. A. et al. Penalized complexity priors for the skewness parameter of power links. *Canadian Journal of Statistics*, Wiley Online Library, 2023.
- ROSS, S. M. *Simulation*. [S.l.]: Academic Press, Inc., 2006.
- SILVA, A. N. da; ANYOSA, S.; BAZAN, J. L. Modelagem bayesiano de regressão binária para dados desbalanceados usando novas ligações. *Brazilian Journal of Biometrics*, v. 38, n. 4, p. 385–417, 2020.
- TEAM, S. D. *PyStan: the Python interface to Stan, Version 2.16.0.0.* 2017. Available: <http://mc-stan.org>.
- WANG, X.; DEY, D. K. Generalized extreme value regression for ordinal response data. *Environmental and ecological statistics*, Springer, v. 18, n. 4, p. 619–634, 2011.

Reference IV

- YIN, S. et al. Skewed link regression models for imbalanced binary response with applications to life insurance. *arXiv preprint arXiv:2007.15172*, 2020.
- ZOU, Q. et al. Finding the best classification threshold in imbalanced classification. *Big Data Research*, Elsevier, v. 5, p. 2–8, 2016.