Modelo de regressão para dados binários com mistura de funções de ligação

Nicholas Wagner Eugenio

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA
UFSCar-USP

NICHOLAS WAGNER EUGENIO

MODELO DE REGRESSÃO PARA DADOS BINÁRIOS COM MISTURA DE FUNÇÕES DE LIGAÇÃO

Dissertação apresentada ao Departamento de Estatística – DEs-UFSCar e ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística.

Orientador: Prof. Dr. Adriano Polpo de Campos

São Carlos
Dezembro de 2016

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA
UFSCar-USP

NICHOLAS WAGNER EUGENIO

REGRESSION MODEL WITH MIXTURE OF LINK FUNCTIONS FOR BINARY DATA

Master dissertation submitted to the Departamento de Estatística – DEs-UFSCar and to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master joint Graduate Program in Statistics.

Advisor: Prof. Dr. Adriano Polpo de Campos

São Carlos
December 2016

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Nicholas Wagner Eugenio, realizada em 08/02/2017:

Prof. Dr. Adriano Polpo de Campos
UFSCar

Profa. Dra. Aline Martines Piroutek
UNICAMP

Prof. Dr. Carlos Alberto de Bragança Pereira
USP

Este trabalho é dedicado a todos os professores pesquisadores que enxergam as dificuldades e as facilidades de seus alunos e os tornam, sob seus ensinamentos, apaixonados pelas ciências que escolheram estudar.

AGRADECIMENTOS

Agradeço, primeiramente, à minha família por todo o suporte e apoio que dão nas minhas mais (in)certas decisões: Meus pais Thelma e Valdeci, minha irmã Victoria, minha tia Ruth e todos os Wagner Eugenio que, de longe ou perto, estão comigo.

Aos membros das bancas de Qualificação e Defesa deste trabalho: Dra. Aline Martines Piroutek, Prof. Dra. Juliana Cobre, Prof. Dr. Adriano Polpo de Campos, Dr. André Rogatko e Prof. Dr. Carlos Alberto de Bragança Pereira pelos mais que produtivos questionamentos, trocas de idéias, ensinamentos, conselhos e, principalmente, reflexões.

Novamente, e com destaque, ao meu orientador Prof. Dr. Adriano Polpo de Campos pela parceria de pesquisas, aulas, conversas, orientações, ajudas, momentos de descontração e etc. Sou grato, acima de tudo, por me permitir perceber e entender que mudanças ocorrerão a todo momento, sejam elas planejadas ou inesperadas, e que cabe a cada um encará-las ou evitá-las. A resposta sempre estará dentro de nós mesmos (as vezes até transbordando).

A todos os professores do DEs-UFSCar e do ICMC-USP com quem tive a oportunidade de aprimorar meus conhecimentos em Estatística a partir de suas aulas, lições e conversas.

Não posso deixar de expressar minha imensa gratidão à Bárbara Beltrame Bettim e Karine Zanuto Mendes, minhas amadas amigas que levo no coração e têm uma grande porcentagem em todas as conquistas que tive no mundo estatístico, bem como no pessoal também. Chegamos crianças nessa São Carlos linda e saímos Mestres, que orgulho tenho de nós!

Aos meus amigos que me acolheram numa turma de Mestrado: Alan Henrique (meu parceiro de rolês), Caroline Mendes (doutoranda direta, orgulho da 015), Diego Mattozo (mato-grossense com o maior coração que já vi), Fabiano Rodrigues (merece o mundo), Gretta Rossi (ainda bem que ficou pro mestrado, ia sentir muitas saudades), Juliana Cecília (melhor parceira de bandejão, come igual a mim), Murilo Cantoni (Mumuzinho, sempre ausente, mas nunca esquecido), Natália Oliveira (sucesso é pouco demais para você), Raul Assis (Raulzão, não tem definição) e Taís Ribeiro (demorou muito pra ser minha vizinha): muito obrigado por esses anos, as madrugadas de estudo foram bem mais fáceis com vocês, as festas foram muito mais prazerosas e a despedida de Sanca muito mais dolorida. Não preciso desejar sucesso, meus amigos, isso vocês têm de sobra, desejo apenas que nos encontremos mais vezes na vida.

Aos TOs 012: Amandinha, Goiaba, Gui, Lalá, Marília, Nat e Tati que por tantas vezes me tiraram de casa para uma ida ao Mc, uma troca de idéias e risadas, para ombros-amigos, colos de mãe e aventuras em festas. Espero não precisar da profissão de vocês, mas a recomendarei

quantas vezes forem necessárias!

Também, agradeço aos meus irmãos de sangue espalhados pelo Brasil. É incrível e único o modo como estamos presentes nas vidas uns dos outros independentemente de onde estejamos. Daniel Zimand, Fabio Rotondano, Felipe Cabral, Henrique Lamonica, Igor Alves, João Paulo Ramos e Victor Larrañaga, que sorte a nossa de termos tanta afinidade.

As idas ao DEs, quase que diárias, foram caracterizadas durante esses 7 anos por uma rotina de "Bom dia, Antônio, tudo bem?" "Fala, Cholas.", dentre outros inúmeros diálogos sobre procrastinação e greves, com o Antonio Henrique Santini Ladvig; pelos "Oi, Dani!" "Oi, Cholas!", conversas sobre gatos e computadores e empréstimos de ferramentas com a Daniela Cristina Marinello; mais recentemente, e não menos saudosos, pelos "Isabel, tudo bem, cê pode me ajudar?" "Oi, Nicholas, diga!", pelas salvações nas questões burocráticas do Programa e por toda atenção com os alunos que a Maria Isabel Rinaldo Pessôa de Araujo tem. Muitíssimos obrigado a todos, vocês ficam nos bastidores e fazem o DEs funcionar.

Por fim, mas não menos importante, agradeço a todos que em algum momento fizeram parte da minha trajetória até aqui. Independentemente de modo, tempo ou intensidade toda relação interpessoal acrescentou algo em minha vida e suas lições me fizeram crescer.

Muito obrigado!

"...gente morna é aquela que faz uma coisa inacreditável:
ela economiza afeto, esquecendo que afeto e conhecimento
são duas coisas que se você guardar, você perde.
Afeto e conhecimento, se guardados, são perdidos.
Por isso, cuidado você e eu para não
sermos mornos na vida."
(Mario Sergio Cortella)

RESUMO

EUGENIO, N. W. **Modelo de regressão para dados binários com mistura de funções de ligação**. 2016. 83 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2016.

Apresenta-se um modelo de regressão para dados binários com mistura de quatro funções de ligação (*logit*, *probit*, complementar log-log e Stukel) que também são seus casos particulares. Os procedimentos de estimação frequentista são expostos e, através de estudos de simulações, mostra-se que, em relação a outros modelos, a função de ligação proposta apresenta melhor desempenho nas estimações de proporções, ao passo que para previsões é igual às demais. Sua flexibilidade em poder ser tanto uma função de ligação simétrica quanto assimétrica é corroborada pelo resultados das análises de três bancos de dados reais, bem como pelas simulações. Mostra-se ainda um caso em que, por não conseguir obter melhores resultados com as combinações de ligações, a mistura associa peso total a um de seus componentes.

Palavras-chave: Mistura, Regressão, Dados Binários, Ligação Assimétrica, Modelo Assimétrico.

ABSTRACT

EUGENIO, N. W. **Regression model with mixture of link functions for binary data**. 2016. 83 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2016.

A regression model for binary data with mixture of four link functions (*logit*, *probit*, complementary log log and Stukel) is shown and these functions are particular cases of the model. The frequentist estimation procedure is exposed and, by simulation studies, it is notable that, comparing with other models, the link function proposed presents a better performance in proportions' estimations, while for predctions they are all equal. Its flexibility in being both a symmetric or an assymmetric link function is corroborated on the real data analisys results, as the simulations. Furthermore, it is shown a case where the mixture associates total weight for a link function because it is no possible to improve the results by mixing other functions.

Keywords: Mixture, Regression, Binary Data, Assymmetric Link, Assymetric Model.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

A estimação de proporções de uma variável (resposta) com base nos valores de outras (explicativas) é um situação comum em diversas áreas da ciência e do mercado de trabalho, segundo Agresti e Finlay (2009). Sua peculiaridade é o fato de que seus resultados devem estar contidos no intervalo [0,1] e, para tanto, a relação entre tais variáveis deve ser feita através de funções que satisfaçam essa condição.

De acordo com Diniz (2015), que fez uma síntese da história das funções logística e probito apresentada por Cramer (2003) (referência principal das datas e artigo citados), o desenvolvimento da primeira se iniciou com o questionamento de Alphonse Quetelet (1975-1874) sobre valores de tamanhos populacionais implausíveis na teoria populacional malthusiana (Malthus (1872)). Ao levar a dúvida para seu pupilo Pierre-François Verhulst (1804-1849), o mesmo adicionou um termo representando a existência de um limitante ao crescimento indiscriminado e a solução do problema, no contexto de equações diferenciais, foi denominada função logística. Suas descobertas foram publicadas em três artigos: Verhulst (1838), Verhulst (1845) e Verhulst (1847).

No século seguinte Raymond Pearl (1879-1940) e Lowell J. Reed (1886-1966) depararam-se com uma curva semelhante à da Logística ao estudarem a necessidade de comida na Primeira Guerra Mundial, porém sem nomeá-la de tal forma. A nomenclatura foi retomada por Yule (1925). Ainda, funções semelhantes foram utilizadas na Química pelo professor alemão Wilhelm Ostwald em meados de 1883 para descreverem reações de processos catalíticos.

Já a invenção do modelo *probit* é creditada à Bliss (1934).Em seu artigo, estuda experimentos biológicos do tipo dose-resposta para doses fixas e respostas aleatórias que refletem a distribuição individual de níveis de tolerância. A introdução do termo *probit*, uma abreviação de *probability unit*, é apresentada em Bliss (1935) como uma “forma mais conveniente de expressar desvios da média de uma distribuição Normal", segundo Diniz (2015). Ainda, aplicações de tal modelo em áreas como economia e pesquisa de mercado apareceram nos anos 1950/60 (Farrell

(1954), Adam (1958) e Aitchison e Brown (1963)).

Com a vasta popularidade do modelo *probit*, o físico e doutor em Estatística Joseph Berkson (1899-1982), estatístico chefe da Clínica Mayo, propôs em Berkson (1944) o uso do modelo logístico em experimentos biológicos e cunhou o termo *logit*, em analogia à *probit*. Defensor da descoberta de Verhulst, alegava que estimações por Mínimos Quadrados eram bem mais eficientes que por Máxima Verossimilhança (Berkson (1980)), porém o fato do modelo *probit* possibilitar a interpretação baseada em desvios da média de uma distribuição Normal e de que na época todos os cálculos eram feitos à lápis, ou em calculadoras, fez com que adeptos do modelo de Bliss não aceitassem a proposta de Berkson muito bem, embora o mesmo tenha exposto suas razões para usa a ligação *logit* em Berkson (1951).

Somente em 1969 a eficiência das propriedades da análise logística foi reconhecida em um texto por Cox, base para Cox e Snell (1989), e ela se espalhou muito mais rápido que a outra devido às facilidades computacionais. Nelder e Weddeburn (1972), através dos Modelos Lineares Generalizados (MLG), demonstraram que a função de ligação (FL) canônica para dados seguindo distribuição Binomial é a própria função logística e a mesma se tornou padrão nos pacotes computacionais.

Por ambas citadas até agora serem funções de ligação simétricas, ou seja, a probabilidade de ocorrência do evento de interesse se aproxima dos extremos do intervalo $[0,1]$ com a mesma velocidade, há casos em que suas utilidades não produzem resultados satisfatórios. Um exemplo pode ser visto na área da saúde cujo interesse é determinar a probabilidade da presença de uma doença rara em um indivíduo.

De acordo com King e Zeng (2001) o uso de Regressão Logística em dados extremamente desbalanceados produz estimativas subestimadas para a proporção de ocorrência do evento de interesse. Desta forma, para este e casos análogos (não tão desbalanceados), o uso de funções de ligação assimétricas, como complementar log-log (assimétrica à direita) e log-log (assimétrica à esquerda), tornaram-se alternativas plausíveis. Ainda, Aranda-Ordaz (1981) apresentou uma função cujos casos particulares são a *logit* e complementar log-log. Já Prentice (1976) apresentou uma ligação baseada na função de distribuição F-Snedecor, cujos casos particulares são as *logi*, complementar log-log, log-log, *probit*, Laplace e exponencial.

Além das supracitadas, Guerrero e Johnson (1982) apresentam uma transformação baseada em Box e Cox (1964). Já a generalização de Stukel (1988) para a Regressão Logística, que lida com assimetria e curtose, é baseada na transformação de Manly (1976). Ainda, Caron (2010) apresenta uma função de ligação baseada na distribuição Weibull e que aproxima as funções *logit*, *probit* e complementar log-log.

A fim de se definir a função que melhor se adeque aos dados, faz-se necessária uma análise exploratória inicial dos mesmos em busca de peculiaridades que ajudem na escolha. Porém, são grandes as chances de se deparar com uma ou mais funções admissíveis e, ao escolher

uma em detrimento da outra, deixar de abranger, para os dados em mãos, qualidades desta que não há naquela.

Tendo como motivação para o estudo a modelagem no mercado financeiro, tem-se que na prática é quase unânime o uso da ligação *logit* nas análises de *credit scoring*, segundo Gonçalves, Gouvêa e Mantovani (2013). Indagou-se, então, o motivo para o uso quase que exclusivo desta ligação sendo que pode haver outros tão simples quanto e que apresentem, por menor que sejam, resultados iguais ou melhores de estimação e previsão. Resultados estes que, uma vez contextualizados nos valores altíssimos trabalhados nos bancos, tornam-se lucros consideráveis.

Baseada neste contexto, e com o objetivo de poder agregar as melhores características de cada FL e melhorar as estimações e previsões realizadas, é proposta nesta dissertação a mistura de funções de ligação para dados binários (MIX). Para um conjunto de covariáveis fixas, a técnica consiste em relacioná-lo à variável dependente através de uma combinação linear convexa das funções de ligação, tendo como particularidade o fato de que os parâmetros do preditor linear são os mesmos para todas elas. A idéia é apresentar uma FL tão boa quanto as utilizadas usualmente.

Uma vez que se combina diversas ligações, resultando em uma única ligação, é plausível que se tenha apenas um conjunto de parâmetros associados às covariáveis. Considera-se a função de ligação MIX como uma só, não fazendo sentido fragmentá-la em diversas funções de ligação associando a cada uma um vetor de parâmetros.

O texto está organizado da seguinte forma: No Capítulo 2 são apresentadas funções de ligação, já conhecidas e bem difundidas, para dados binários. No Capítulo 3 apresenta-se o Modelo Binomial e a ligação proposta, bem como suas propriedades e seu método de estimação. No Capítulo 4 estudos de simulações são feitos para se verificar as qualidades da ligação, entender melhor suas características e compará-la com outras FLs . No Capítulo 5 utiliza-se a MIX na modelagem de três bancos de dados reais para se corroborar seus uso e flexibilidade. Por fim, no Capítulo 6 são discutidos os resultados obtidos.

Chama-se a atenção para os fatos de que o separador decimal utilizado foi o símbolo "." (ponto), arredondamentos foram feitos na terceira casa decimal e valores estimados são acentuados pelo acento circunflexo "^". Também, nos boxplots apresentados, as linhas pontilhadas são referente às médias das observações em questão.

CAPÍTULO

2

FUNÇÕES DE LIGAÇÃO PARA RESPOSTAS DICOTÔMICAS

Diversas são as situações em que se depara com variáveis cujas possíveis respostas são apenas duas: sim ou não, sucesso ou fracasso, 1 ou 0, A ou B, dentre tantas outras maneiras de se expressar "ocorrência ou não ocorrência do evento de interesse". Essas são as chamadas variáveis binárias, dicotômicas e neste texto trabalhar-se-á com a proporção da "ocorrência do evento" sendo o objeto da modelagem.

Durante muito tempo o Modelo de Regressão Linear foi utilizado como ferramenta principal na modelagem de eventos aleatórios, independente da natureza dos dados e de suas restrições intrínsecas. Entretanto, para seu uso, algumas suposições sobre os resíduos são requeridas e, quando não satisfeitas, podem ser adotados alguns procedimentos tais como: Regressões Ridge e Robusta, transformações na variável resposta, dentre elas a mais usual para se obter e cumprir a condição de normalidade dos dados é a de Box e Cox (1964), entre outros.

A introdução dos MLG por Nelder e Weddeburn (1972) possibilitou a modelagem de variáveis respostas cujas distribuições de probabilidade não fossem gaussianas permitindo, ainda, que os parâmetros das distribuições de cada indivíduo fossem diferentes, bastando apenas que as mesmas pertencessem à Família Exponencial. Desta forma a escolha das funções de ligação ficou mais flexível.

Para a classe de modelos em estudo assume-se que cada indivíduo segue uma distribuição Bernoulli e que, conjuntamente para covariáveis iguais, seguem uma distribuição Binomial.

Antes de apresentar as formas funcionais de algumas funções de ligação faz-se necessária, para melhor fluência do texto, a introdução de notações: $\boldsymbol{Y} = \begin{bmatrix} Y_1 & \dots & Y_n \end{bmatrix}^T$ é o vetor de variáveis dependentes (variáveis de interesse) dicotômicas para os $i = 1, 2, \dots, n$ indivíduos condicionado às suas $j = 1, 2, \dots, p-1$ covariáveis (matriz de planejamento) ou variáveis preditoras $\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \boldsymbol{X_1} & \dots & \boldsymbol{X}_{p-1} \end{bmatrix}$, em que $\boldsymbol{X_j} = \begin{bmatrix} X_{j1} & \dots & X_{jn} \end{bmatrix}^T$ e $\mathbf{1}$ é um vetor de 1's de

tamanho n, conforme notação e suposições em Kutner *et al.* (2004, p. 217-218). Dado que $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \dots & \beta_{p-1} \end{bmatrix}^T$ é o vetor de p parâmetros da regressão, denota-se o preditor linear por $\boldsymbol{\eta} = \boldsymbol{X\beta}$.

Ainda, $\boldsymbol{\pi}$ é o vetor de proporções de "ocorrência do evento" e $h(\cdot) : \mathbb{R} \to [0,1]$ a função de ligação.

2.1 Algumas Funções de Ligação

Modelos de Regressão Linear pressupõe normalidade nos resíduos e têm a finalidade de estimar a esperança condicional da variável resposta, cujos valores podem variar na reta $\mathbb{R}$. A relação entre esta estatística e as covariáveis é expressa como uma equação linear:

$$E(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{X\beta} = \boldsymbol{\eta}$$

Todavia, como já dito, para dados dicotômicos as médias condicionais das proporções (e suas estimativas) devem pertencer ao intervalo $[0,1]$. Para tanto utilizam-se as funções de ligação que transformam o preditor linear e resultam em valores plausíveis para o contexto, conforme relação a seguir:

$$\begin{gathered} E(\boldsymbol{Y} \mid \boldsymbol{X}) = h(\boldsymbol{\eta}) \\ \Updownarrow \\ \boldsymbol{\pi} = h(\boldsymbol{\beta X}) \end{gathered} \tag{2.1}$$

2.1.1 Logit

A função de ligação mais utilizada é a função de densidade acumulada (FDA) da distribuição Logísitica.

Seja W $\sim$ Logística (μ, σ), com $\mu \in \mathbb{R}$ e $\sigma > 0$ parâmetros de locação e escala, respectivamente. Sua função de densidade de probabilidade (FDP) é dada por:

$$f_W(w) = \frac{exp\left(-\dfrac{w-\mu}{\sigma}\right)}{\sigma\left(1+exp\left(-\dfrac{w-\mu}{\sigma}\right)\right)^2}\, \mathbb{I}(w)_{(-\infty,+\infty)} \tag{2.2}$$

e a FDA por:

$$F_W(w) = \frac{1}{1+exp\left(-\dfrac{w-\mu}{\sigma}\right)} = \frac{exp\left(\dfrac{w-\mu}{\sigma}\right)}{1+exp\left(\dfrac{w-\mu}{\sigma}\right)} \tag{2.3}$$

Assim, adotando a forma padrão desta FDA ($\mu = 0$ e $\sigma = 1$) e substituindo a notação $F_W(w)$ pela representação da proporção π e w pelo preditor linear η obtém-se a função de ligação *logit*, conforme a Equação 2.4:

$$\boldsymbol{\pi} = \frac{\exp(\boldsymbol{\eta})}{1+\exp(\boldsymbol{\eta})} \Longleftrightarrow \boldsymbol{\eta} = \log\left(\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}\right) \tag{2.4}$$

A Tabela 1 apresenta casos peculiares para valores de $\boldsymbol{\eta}$.

Tabela 1 – Casos peculiares da função de ligação *logit*

η	π
$\to -\infty$	$\to 0$
$\to \infty$	$\to 1$
0	0.5

2.1.2 Probit

Baseada na FDA de uma distribuição Normal Padrão, $\Phi(\cdot)$, a função de ligação *probit* é apresentada na Equação 2.5:

$$\boldsymbol{\pi} = \Phi(\boldsymbol{\eta}) \Longleftrightarrow \boldsymbol{\eta} = \Phi^{-1}(\boldsymbol{\pi}) \tag{2.5}$$

Pela simetria e centralidade da distribuição em $\mu = 0$ é trivial que para $\eta = 0 \Rightarrow \pi = 0.5$.

2.1.3 Complementar log-log (CLL)

A distribuição do Valor Extremo é a base para a função de ligação complementar log-log.

Seja W $\sim$ Valor Extremo (μ, σ), com $\mu \in \mathbb{R}$ e $\sigma > 0$ parâmetros de locação e escala, respectivamente. Sua FDP é dada por:

$$f_W(w) = \frac{1}{\sigma}\exp\left(\frac{w-\mu}{\sigma}\right)\exp\left(-\exp\left(\frac{w-\mu}{\sigma}\right)\right)\mathbb{I}(w)_{(-\infty,+\infty)} \tag{2.6}$$

e a FDA por:

$$F_W(w) = 1 - \exp\left(-\exp\left(\frac{w-\mu}{\sigma}\right)\right) \tag{2.7}$$

Para a forma padrão desta FDA e, fazendo as mesmas substituições que na *logit*, obtém-se a função de ligação Complementar Log-Log (CLL), conforme Equação 2.8:

$$\boldsymbol{\pi} = 1 - \exp\left(-\exp(\boldsymbol{\eta})\right) \Longleftrightarrow \boldsymbol{\eta} = \log\left(-\log\left(1-\boldsymbol{\pi}\right)\right) \tag{2.8}$$

Para uma breve comparação, a Figura 1 apresenta os π's das funções de ligação *logit*, *probit* e complementar log-log para valores de $\eta \in [-4,4]$. Nota-se que as duas primeiras são simétricas em torno de $\pi = 0.5$ e que a outra atinge a mediana da proporção para um valor de η menor que as demais, porém é assimétrica e atinge seu ponto máximo antes delas. Ainda, a função *logit* atinge os extremos de $\boldsymbol{\pi}$ de forma mais suave que a *probit*.

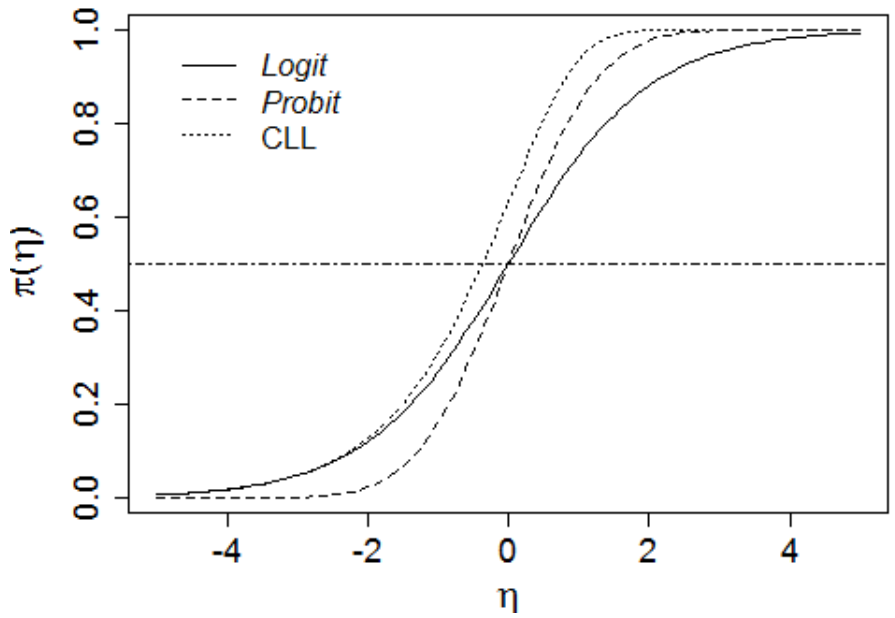


Figura 1 – Curvas das funções de ligação *logit*, *probit* e complementar log-log

2.1.4 Aranda-Ordaz (AO)

A função assimétrica apresentada por Aranda-Ordaz (1981), explicitada na Equação 2.9, é útil para casos em que há conexão com problemas de valores extremos e tem como particularidade as ligações *logit* e complementar log-log:

$$\boldsymbol{\pi} = \begin{cases} 1-[1+\lambda\exp(\boldsymbol{\eta})]^{\frac{-1}{\lambda}}, & \lambda\exp(\boldsymbol{\eta}) > -1 \\ 1, & \text{caso contrário} \end{cases} \iff \boldsymbol{\eta} = \log\left[\frac{(1-\boldsymbol{\pi})^{-\lambda}-1}{\lambda}\right],\ \lambda\exp(\boldsymbol{\eta}) > -1 \tag{2.9}$$

Se $\lambda = 1$, AO converge para a ligação *logit*, e se $\lambda \to 0$ a função tende à complementar log-log. Na Figura 2 há as curvas para três valores diferentes de λ e é possível notar que para valores de $\eta \lessapprox -2$ suas proporções π são praticamente as mesmas. A medida em que λ aumenta o crescimento da curva se dá de forma mais lenta e suave, mantendo a assimetria.

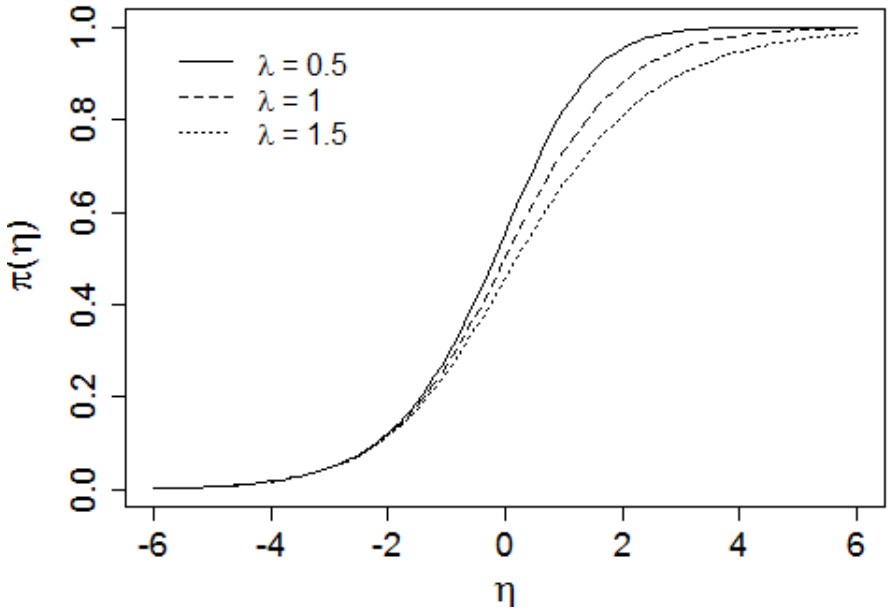


Figura 2 – Curvas da função de ligação Aranda-Ordaz para valores de $\lambda = 0.5, 1, 1.5$

2.1.5 Stukel

Proposta por Stukel (1988), a forma generalizada desta FL é descrita na Equação 2.10 e preserva a propriedade de $\eta(0) = 0.5$:

$$\boldsymbol{\pi}_{\boldsymbol{\gamma}}(\boldsymbol{\eta}) = \frac{\exp\left(g_{\boldsymbol{\gamma}}(\boldsymbol{\eta})\right)}{1+\exp\left(g_{\boldsymbol{\gamma}}(\boldsymbol{\eta})\right)} \Longleftrightarrow g_{\boldsymbol{\gamma}}(\boldsymbol{\eta}) = \log\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}, \tag{2.10}$$

em que $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ é o vetor de parâmetros que definem a forma da função de ligação modelando as caudas.

As Equação 2.11 e Equação 2.12 definem a função $g_{\boldsymbol{\gamma}}(\boldsymbol{\eta})$ não-linear e estritamente crescente.

- Se $\eta \geq 0$ (ou equivalentemente se $\pi \geq 0.5$):

$$g_{\boldsymbol{\gamma}}(\eta) = \begin{cases} \gamma_1^{-1}(\exp(\gamma_1|\eta|)-1), & \text{se } \gamma_1 > 0 \\ \eta, & \text{se } \gamma_1 = 0 \\ -\gamma_1^{-1}\log(1-\gamma_1|\eta|), & \text{se } \gamma_1 < 0 \end{cases} \tag{2.11}$$

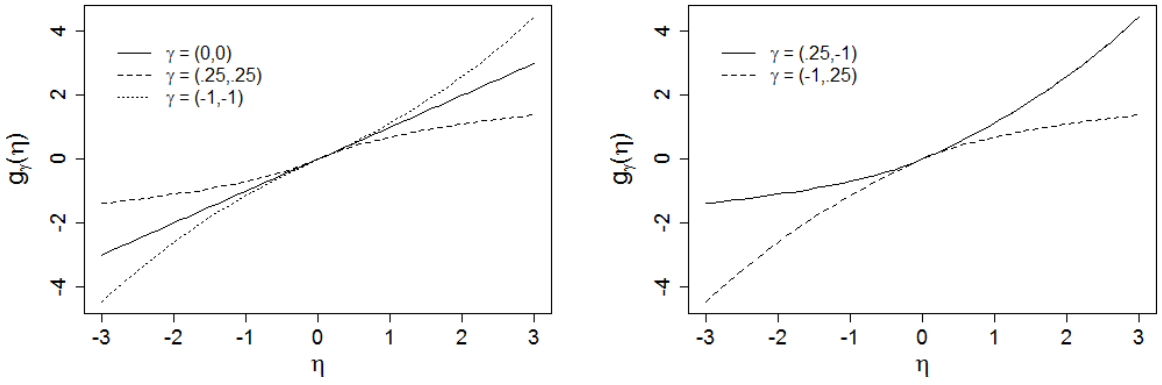
- Se $\eta \leq 0$ (ou equivalentemente se $\pi \leq 0.5$):

$$g_{\boldsymbol{\gamma}}(\eta) = \begin{cases} -\gamma_2^{-1}(\exp(\gamma_2|\eta|)-1), & \text{se } \gamma_2 > 0 \\ \eta, & \text{se } \gamma_2 = 0 \\ \gamma_2^{-1}\log(1-\gamma_2|\eta|), & \text{se } \gamma_2 < 0 \end{cases} \tag{2.12}$$

Quando $\gamma_1 = \gamma_2 = 0$ obtém-se a FL *logit*, caso contrário cada parâmetro controla uma cauda da função centrada em zero de maneiras diferentes. Se $\gamma_1 = \gamma_2$ a curva obtida é simétrica (senão é assimétrica) e para $\gamma_1 = \gamma_2 \approx 0.165$ tem-se a ligação *probit*.

Pode-se notar a simetria da função $g_\gamma(\eta)$ na Figura 3a quando os valores de γ's são iguais e, pela Figura 3b, a assimetria quando são diferentes. Nestas, ainda, ilustra-se a propriedade dita no começo da subseção. Já na Figura 4 veem-se os comportamentos das funções $\pi_\gamma(\eta)$, simétrica e assimetricamente.

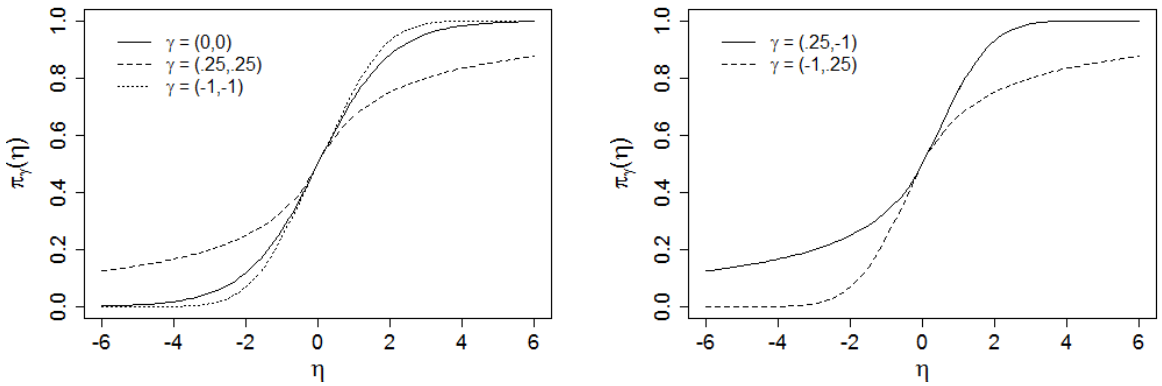
Figura 3 – Curvas das funções $g_\gamma(\eta)$ para diferentes combinações de $\boldsymbol{\gamma}$ e valores de η



(a) $\gamma_1 = \gamma_2$ (b) $\gamma_1 \neq \gamma_2$

Fonte: Elaborada pelo autor.

Figura 4 – Curvas das funções $\pi_\gamma(\eta)$ para diferentes combinações de $\boldsymbol{\gamma}$ e valores de η



(a) $\gamma_1 = \gamma_2$ (b) $\gamma_1 \neq \gamma_2$

Fonte: Elaborada pelo autor.

CAPÍTULO

3

MISTURA DE FUNÇÕES DE LIGAÇÃO

Do Capítulo 2 nota-se que as FLs são baseadas em distribuições de probabilidades, mais especificamente em suas FDAs (quando não são as próprias). Ainda, misturas de FDAs resultam em uma FDA e tal resultado pode ser expandido para as FLs devido às suas relações. Desta forma, a ligação proposta apresenta todas as características de uma FL.

Sejam $h_l(\cdot),\ l = 1,2,\ldots,k$, as distintas funções de ligação a serem combinadas na mistura e α_l suas contribuições para modelo de regressão. Desta forma propõe-se que a função de ligação $h_{MIX}(\boldsymbol{\eta})$ seja como na Equação 3.1:

$$h_{MIX}(\boldsymbol{\eta}) = \sum_{l=1}^{k-1} \alpha_l h_l(\boldsymbol{\eta}) + \left(1 - \sum_{l=1}^{k-1} \alpha_l\right) h_k(\boldsymbol{\eta}), \tag{3.1}$$

em que $\mathbf{0} \leq h_{MIX}(\boldsymbol{\eta}) \leq \mathbf{1},\ 0 \leq \alpha_l \leq 1$ e $\sum_{l=1}^{k-1} \alpha_l \leq 1$.

Reafirma-se que o preditor linear $\boldsymbol{\eta} = \boldsymbol{X\beta}$ é comum a todas funções h_l.

Por ter como objetivo ser flexível quanto às peculiaridades dos dados, no que diz respeito à (as)simetria, é mais sensato que h_{MIX} seja uma mistura de funções de ligação simétricas e assimétricas. Optou-se, portanto, por combinar as funções complementar log-log, *logit*, *probit* e Stukel (nessa ordem) e permitir aos α_l's se encarregarem naturalmente da escolha das mais adequadas.

Ainda, levando em conta que cada uma possui características próprias, é possível evidenciar, através de seus pesos, diferentes perfis comportamentais presentes nos dados tais como angulação da curva π e velocidade de aproximação de seus extremos.

Já a interpretação dos parâmetros β_0 e β_j em relação às alterações que causam na estimativa da probabilidade de sucesso é feita através de análise numérica, uma vez que, com as misturas, fica inviável encontrar uma forma analítica para a contribuição marginal de cada variável.

O vetor $\boldsymbol{\gamma}$ exerce a função de modelar as caudas da ligação Stukel.

3.1 Estimação dos Parâmetros

Considerando $\boldsymbol{Y}$ e $\boldsymbol{X}$ como apresentados no capítulo anterior, cada indivíduo Y_i segue uma distribuição Bernoulli(π_i), sendo a probabilidade de sucesso $P(Y_i = 1) = \pi_i$. As covariáveis associadas são denotadas por $\boldsymbol{X}_i$, que é o vetor linha da matriz de planejamento $\boldsymbol{X}$, isto é, $\boldsymbol{X}_i = \begin{bmatrix} 1 & X_{1i} & \ldots, & X_{(p-1)i} \end{bmatrix}$.

O vetor $\boldsymbol{\theta}$ de parâmetros a ser estimado é constituído de $k-1$ valores de $\boldsymbol{\alpha}$, dois valores de $\boldsymbol{\gamma}$ e p valores de $\boldsymbol{\beta}$, ou seja, $\boldsymbol{\theta} = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \gamma_1 & \gamma_2 & \boldsymbol{\beta} \end{bmatrix}$.

Desta forma a função de verossimilhança é definida como:

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{x}) &= \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \\ &= \pi_i^{\sum_{i=1}^{n} y_i}(1-\pi_i)^{\sum_{i=1}^{n}(1-y_i)}, \end{aligned} \tag{3.2}$$

e a log-verossimilhança:

$$\begin{aligned} \ell(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{x}) &= \log[L(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{x})] \\ &= \sum_{i=1}^{n} y_i \log(\pi_i) + \sum_{i=1}^{n}(1-y_i)\log[(1-\pi_i)], \end{aligned} \tag{3.3}$$

em que $\pi_i = h_{MIX}(\eta_i) = \sum_{l=1}^{k-1} \alpha_l h_l(\boldsymbol{X_i}^T\boldsymbol{\beta}) + \left(1 - \sum_{l=1}^{k-1}\alpha_l\right) h_k(\boldsymbol{X_i}^T\boldsymbol{\beta})$.

De forma mais detalhada $h_{MIX}(\eta_i)$ é dada por:

$$\begin{aligned} h_{MIX}(\eta_i) = \pi_{i_{MIX}} &= \alpha_{cll}\pi_{i_{cll}} + \alpha_{lgt}\pi_{i_{lgt}} + \alpha_{pbt}\pi_{i_{pbt}} + \alpha_{stu}\pi_{i_{stu}} \\ &= \alpha_{cll}\left[1 - \exp\left(-\exp(\eta_i)\right)\right] + \alpha_{lgt}\left[\frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right] \\ &+ \alpha_{pbt}\Phi(\eta_i) + \alpha_{stu}\left[\frac{\exp\left(g_{\boldsymbol{\gamma}}(\eta_i)\right)}{1+\exp\left(g_{\boldsymbol{\gamma}}(\eta_i)\right)}\right] \\ &= \alpha_{cll}\left[1 - \exp\left(-\exp(\eta_i)\right)\right] + \alpha_{lgt}\left[\frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right] \\ &+ \alpha_{pbt}\Phi(\eta_i) + (1-\alpha_{cll}-\alpha_{lgt}-\alpha_{pbt})\left[\frac{\exp\left(g_{\boldsymbol{\gamma}}(\eta_i)\right)}{1+\exp\left(g_{\boldsymbol{\gamma}}(\eta_i)\right)}\right], \end{aligned} \tag{3.4}$$

em que *cll*, *lgt*, *pbt* e *stu* são abreviações de complementar log-log, *logit*, *probit* e Stukel, respectivamente.

As estimações são feitas combinando-se duas etapas, sendo a primeira uma estimação conjunta de todos os parâmetros (substituindo $\boldsymbol{\alpha}$ por $\boldsymbol{\zeta}$, que não têm restrição, isto é, assume

valores na reta real) e a segunda estimações iterativas utilizando verossimilhança perfilada. Em ambas estimações são usadas métodos numéricos para otimização de funções, uma vez que as derivadas de interesse não foram possíveis de ser calculadas analiticamente.

3.1.1 Etapa 1

Nesta etapa utiliza-se um artifício matemático para que não seja preciso impor nenhuma restrição aos valores dos parâmetros a serem estimados, mantendo-se ainda, de forma implícita, a redundância do valor do peso da "última" função de ligação.

A manipulação feita consiste em escrever α's em função de parâmetros ζ's, de modo que $\zeta_l \in \mathbb{R}$, $0 \leq \alpha_l \leq 1$, $l = 1, \cdots, k-1$, e $\sum_{l=1}^{k-1} \alpha_l \leq 1$. Baseado no conteúdo já apresentado, escolheu-se a função Logística Multinomial para associar os valores reais ao intervalo $[0,1]$ da forma como segue:

$$\alpha_l = \frac{\exp(\zeta_l)}{1+\sum_{l=1}^{3}\exp(\zeta_l)} \quad , \quad l = 1,\ldots,3; \;\; \zeta_l \in \mathbb{R}, \quad \alpha_l \in [0,1] \text{ e } \sum_{l=1}^{k-1} \alpha_l \leq 1 \tag{3.5}$$

Assim, a função log-verossimilhança passa a ser como na Equação 3.6 e as estimativas de Máxima Verossimilhança são obtidas através da função `optim` do *software* R (R Core Team (2013)) com o método numérico *Nelder-Mead*.

$$\begin{aligned}
\ell(\zeta, \boldsymbol{\gamma}, \boldsymbol{\beta} | \boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{n} y_i \log \Bigg\{ & \frac{\exp(\zeta_1)}{1+\sum_{l=1}^{3}\exp(\zeta_l)} [\pi_{i_{cll}}] + \frac{\exp(\zeta_2)}{1+\sum_{l=1}^{3}\exp(\zeta_l)} [\pi_{i_{lgt}}] + \\
& + \frac{\exp(\zeta_3)}{1+\sum_{l=1}^{3}\exp(\zeta_l)} [\pi_{i_{pbt}}] + \left(1 - \frac{\sum_{l=1}^{3}\exp(\zeta_l)}{1+\sum_{l=1}^{3}\exp(\zeta_l)}\right) [\pi_{i_{stu}}] \Bigg\} + \\
& + \sum_{i=1}^{n} (1-y_i) \log \Bigg\{ 1 - \Bigg\{ \frac{\exp(\zeta_1)}{1+\sum_{l=1}^{3}\exp(\zeta_l)} [\pi_{i_{cll}}] + \frac{\exp(\zeta_2)}{1+\sum_{l=1}^{3}\exp(\zeta_l)} [\pi_{i_{lgt}}] + \\
& + \frac{\exp(\zeta_3)}{1+\sum_{l=1}^{3}\exp(\zeta_l)} [\pi_{i_{pbt}}] + \left(1 - \frac{\sum_{l=1}^{3}\exp(\zeta_l)}{1+\sum_{l=1}^{3}\exp(\zeta_l)}\right) [\pi_{i_{stu}}] \Bigg\} \Bigg\}.
\end{aligned} \tag{3.6}$$

Como nessa primeira etapa o objetivo é encontrar estimativas que sirvam como parâmetros fixos para o restante do processo, é tomado o cuidado de se selecionar diversos chutes iniciais, encontrar as estimativas e escolher aquela com o maior valor de log-verossimilhança. Resumidamente, tais chutes são combinações de:

(a) Peso nulo para uma função de ligação e pesos iguais para as demais (vetor $\boldsymbol{\zeta}$) ou peso 1 para uma função e nulo para as demais;

(b) Para o vetor $\boldsymbol{\beta}$ valores nulos ou os coeficientes estimados obtidos através de um Modelo Linear Geral para os mesmos dados e com funções de ligação *logit*, complementar log-log e *probit*;

(c) Chutes iguais a zero para o vetor $\boldsymbol{\gamma}$.

No total são 28 chutes iniciais diferentes para que se refine o processo.

3.1.2 Etapa 2

Com as estimativas obtidas na etapa anterior, um procedimento iterativo de Verossimilhança Perfilada é realizado. Introduzida por Fisher (1956), segundo Sprott (2008), é útil quando há interesse em somente uma parte do vetor de parâmetros (parâmetros de interesse) e não nos demais (parâmetros de perturbação).

Seguindo notação e raciocínio de Lucambio (2009), seja ϑ o vetor de parâmetros que pode ser decomposto como $\vartheta = (\psi, \varsigma)$ e que o interesse esteja em inferir sobre o vetor ψ. Em situações deste tipo se é possível construir uma função que dependa apenas dos parâmetros de interesse para se fazer inferência sobre os mesmos: são as funções de pseudoverossimilhança. Dentre elas está a Função de Verossimilhança Perfilada (FVP).

Para o vetor em questão, o logaritmo da FVP é dado por

$$\ell_P(\psi) = \max_{\varsigma} \ell(\psi, \varsigma),$$

sendo que o máximo é obtido em todo o espaço paramétrico fixando um valor de ψ.

Desta forma a maximização é obtida em $\hat{\varsigma}(\psi)$ e a log-FVP pode ser reescrita como na Equação 3.7, uma função de ψ apenas.

$$\ell_P(\psi) = \ell(\psi, \hat{\varsigma}(\psi)) \tag{3.7}$$

Em síntese, a Etapa 2 consiste em estimar cada parâmetro univariadamente (fixando os demais) avaliando sua log-verossimilhança até que haja uma convergência conjunta.

Definindo, o m-ésimo elemento do vetor $\boldsymbol{\theta}_{(5+p)\times 1}$ como $\theta_{\{m\}}, m = 1, \ldots, (5+p)$, tem-se que na primeira iteração são fixados os valores dos parâmetros obtidos na Etapa 1 (com ζ's já transformados em α's) e, isolada e sequencialmente, cada componente do vetor $\boldsymbol{\theta}$ é estimado e sua log-verossimilhança observada. Se, na estimação do parâmetro $\theta_{\{m\}}$, o valor da log-verossimilhança for maior que o da estimação do parâmetro anterior $\theta_{\{m-1\}}$, então atualiza-se $\theta_{\{m\}}$ com o valor estimado e sua log-verossimilhança fica como base de comparação para as demais neste bloco, caso contrário mantém-se os valores. Este processo é feito com o vetor $\boldsymbol{\theta}$ inteiro, que pode ser considerado um bloco, e ao final de cada iteração dos blocos verifica-se

se o limite máximo de iterações foi atingido ou se se obteve convergência (diferença de valor menor ou igual a 10^{-5} entre as log-verossimilhanças observadas nas estimações dos últimos parâmetros de cada bloco subsequente), que são os critérios de parada.

As estimativas finais dos parâmetros são aquelas do último bloco iterado.

CAPÍTULO

4

SIMULAÇÕES

4.1 Verificação da Qualidade de Estimação do Modelo

Estabelecendo uma distribuição para a covariável $\boldsymbol{X}$ e variando os valores dos parâmetros geraram-se amostras dos modelos MIX, *logit* e Stukel (FLs proposta, simétrica e assimétrica), das quais se estimou $\hat{\boldsymbol{\theta}}$, ou seus sub-conjuntos, usando diversas funções de ligação a fim de se fazerem comparações.

A implementação do algoritmo de estimação para a função MIX foi feita pelo autor e seu orientador; para as funções *logit*, *probit* e complementar log-log foi utilizada a função `glm` do pacote `stats` do *software R*; para Aranda-Ordaz e Stukel foram utilizadas as implementações feitas na dissertação de mestrado de Santos (2013).

Os procedimentos descritos a seguir são repetidos $t = 1, \ldots, n.rep$ $(n.r)$ vezes para cada função de ligação e seus resultados armazenados.

Para os três modelos do primeiro parágrafo o algoritmo de simulação consiste em, a princípio, gerar uma quantidade *n.sample* $(n.s)$ de observações da covariável e depois fixar valores para $\hat{\boldsymbol{\theta}}$ ou seus sub-conjuntos, conforme necessidade. Para cada uma das observações se calcula $\eta_{s,t} = \beta_0 + \beta_1 X_{s,t}, s = 1, \ldots, n.s$, e se utiliza tal resultado na construção de $\pi_{s,t}$, que são as suas probabilidades de sucesso associadas denominadas "proporções verdade". Com cada $\pi_{s,t}$ é gerado um valor $W_{s,t}$ da distribuição Binomial com $n.b$ ensaios: $W_{s,t} = \sum_{i=1}^{n} Y_{s,t,i}(\pi_{s,t}) \sim \text{Bin}_{s,t}(n.b, \pi_{s,t})$. Tem-se, então, um valor da soma de $n.b$ ensaios de Bernoullis para cada $\eta_{s,t}$, constituintes do vetor $\boldsymbol{W}_{t_{n.s\times 1}} = \begin{bmatrix} W_{1,t} & \ldots & W_{n.s,t} \end{bmatrix}^T$.

Com W_t estima-se o vetor de parâmetros $\hat{\boldsymbol{\theta}}_t = \begin{bmatrix} \hat{\alpha}_{1,t} & \hat{\alpha}_{2,t} & \hat{\alpha}_{3,t} & \hat{\gamma}_{1,t} & \hat{\gamma}_{2,t} & \hat{\boldsymbol{\beta}}_t \end{bmatrix}^T$ (fazendo as devidas modificações na função de log-verossimilhança) e, com ele e $\boldsymbol{X}_t$, finda-se o

processo de obtenção dos valores das proporções estimadas, como no vetor 4.1.

$$\hat{\boldsymbol{\pi}}_t = \begin{bmatrix} \hat{\pi}_{1,t} & \dots & \hat{\pi}_{n.s,t} \end{bmatrix}^T \tag{4.1}$$

Há também as "proporções observadas", descritas pelo vetor 4.2.

$$\begin{aligned} \boldsymbol{\pi}_t^o &= \begin{bmatrix} \frac{W_{1,t}}{n.b} & \dots & \frac{W_{n.s,t}}{n.b} \end{bmatrix}^T \\ &= \begin{bmatrix} \boldsymbol{\pi}_{1,t}^o & \dots & \boldsymbol{\pi}_{n.s,t}^o \end{bmatrix}^T \end{aligned} \tag{4.2}$$

Para uma melhor visualização dos componentes das iterações, as matrizes finais 4.3, 4.4, 4.5 e 4.6 são apresentadas.

$$\boldsymbol{\eta}_{n.r \times n.s} = \begin{bmatrix} \overbrace{\eta_{1,1}}^{X_{1,t}} & \overbrace{\eta_{2,1}}^{X_{2,t}} & \overbrace{\eta_{3,1}}^{X_{3,t}} & \cdots & \overbrace{\eta_{n.s,1}}^{X_{n.s,t}} \\ \eta_{1,2} & \eta_{2,2} & \eta_{3,2} & \cdots & \eta_{n.s,2} \\ \eta_{1,3} & \eta_{2,3} & \eta_{3,3} & \cdots & \eta_{n.s,3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \eta_{1,n.r} & \eta_{2,n.r} & \eta_{3,n.r} & \cdots & \eta_{n.s,n.r} \end{bmatrix} \begin{matrix} \} t = 1 \\ \} t = 2 \\ \} t = 3 \\ \\ \} t = n.r \end{matrix} = \begin{bmatrix} \boldsymbol{\eta}_1^T \\ \boldsymbol{\eta}_2^T \\ \boldsymbol{\eta}_3^T \\ \vdots \\ \boldsymbol{\eta}_{n.r}^T \end{bmatrix} \tag{4.3}$$

$$\boldsymbol{\pi}_{n.r \times n.s} = \begin{bmatrix} \overbrace{\pi_{1,1}}^{\eta_{1,t}} & \overbrace{\pi_{2,1}}^{\eta_{2,t}} & \overbrace{\pi_{3,1}}^{\eta_{3,t}} & \cdots & \overbrace{\pi_{n.s,1}}^{\eta_{n.s,t}} \\ \pi_{1,2} & \pi_{2,2} & \pi_{3,2} & \cdots & \pi_{n.s,2} \\ \pi_{1,3} & \pi_{2,3} & \pi_{3,3} & \cdots & \pi_{n.s,3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \pi_{1,n.r} & \pi_{2,n.r} & \pi_{3,n.r} & \cdots & \pi_{n.s,n.r} \end{bmatrix} \begin{matrix} \} t = 1 \\ \} t = 2 \\ \} t = 3 \\ \\ \} t = n.r \end{matrix} = \begin{bmatrix} \boldsymbol{\pi}_1^T \\ \boldsymbol{\pi}_2^T \\ \boldsymbol{\pi}_3^T \\ \vdots \\ \boldsymbol{\pi}_{n.r}^T \end{bmatrix} \tag{4.4}$$

$$\boldsymbol{W}_{n.r \times n.s} = \begin{bmatrix} \overbrace{W_{1,1}}^{\pi_{1,t}} & \overbrace{W_{2,1}}^{\pi_{2,t}} & \overbrace{W_{3,1}}^{\pi_{3,t}} & \cdots & \overbrace{W_{n.s,1}}^{\pi_{n.s,t}} \\ W_{1,2} & W_{2,2} & W_{3,2} & \cdots & W_{n.s,2} \\ W_{1,3} & W_{2,3} & W_{3,3} & \cdots & W_{n.s,3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ W_{1,n.r} & W_{2,n.r} & W_{3,n.r} & \cdots & W_{n.s,n.r} \end{bmatrix} \begin{matrix} \} t = 1 \\ \} t = 2 \\ \} t = 3 \\ \\ \} t = n.r \end{matrix} = \begin{bmatrix} \boldsymbol{W}_1^T \\ \boldsymbol{W}_2^T \\ \boldsymbol{W}_3^T \\ \vdots \\ \boldsymbol{W}_{n.r}^T \end{bmatrix} \tag{4.5}$$

$$\hat{\boldsymbol{\theta}}_{n.s\times(5+p)} = \begin{bmatrix} \hat{\alpha}_{1,1} & \hat{\alpha}_{2,1} & \hat{\alpha}_{3,1} & \hat{\gamma}_{1,1} & \hat{\gamma}_{2,1} & \hat{\boldsymbol{\beta}}_1 \\ \hat{\alpha}_{1,2} & \hat{\alpha}_{2,2} & \hat{\alpha}_{3,2} & \hat{\gamma}_{1,2} & \hat{\gamma}_{2,2} & \hat{\boldsymbol{\beta}}_2 \\ \hat{\alpha}_{1,3} & \hat{\alpha}_{2,3} & \hat{\alpha}_{3,3} & \hat{\gamma}_{1,3} & \hat{\gamma}_{2,3} & \hat{\boldsymbol{\beta}}_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{\alpha}_{1,n.r} & \hat{\alpha}_{2,n.r} & \hat{\alpha}_{3,n.r} & \hat{\gamma}_{1,n.r} & \hat{\gamma}_{2,n.r} & \hat{\boldsymbol{\beta}}_{n.r} \end{bmatrix} \begin{matrix} \} t=1 \\ \} t=2 \\ \} t=3 \\ \\ \} t=n.r \end{matrix} = \begin{bmatrix} \hat{\boldsymbol{\alpha}}_1 & \hat{\boldsymbol{\alpha}}_2 & \hat{\boldsymbol{\alpha}}_3 & \hat{\boldsymbol{\gamma}}_1 & \hat{\boldsymbol{\gamma}}_2 & \hat{\boldsymbol{\beta}} \end{bmatrix} \tag{4.6}$$

Como medidas de qualidade das estimativas dos parâmetro obtiveram-se os Erros Quadráticos Médios (EQM), Vícios Médios (Viés) e Erros Máximos Absolutos (EMA) das amostras, esperando que fossem próximos de zero, conforme explicam Lehmann e Casella (1998). O intuito é verificar se os parâmetros dos modelos são estimados corretamente quando os dados são gerados deles mesmos.

Já para as proporções estimadas, foram calculados para cada função de ligação os Erros Absolutos em Relação à Proporção EAO_t, dados pela Equação 4.7:

$$\text{EAO}_t = \begin{bmatrix} \left|\hat{\pi}_{1,t} - \pi^o_{1,t}\right| & \left|\hat{\pi}_{2,t} - \pi^o_{2,t}\right| & \cdots & \left|\hat{\pi}_{n.s,t} - \pi^o_{n.s,t}\right| \end{bmatrix}^T. \tag{4.7}$$

Para cada vetor referente à iteração t obtiveram-se a média e o máximo, resultando assim, no final do processo de simulação, em vetores de tais estatísticas.

Também calcularam-se os Erros Absolutos em Relação à Proporção Obervada Relativos à MIX (EAORM_t), que é um vetor constituído pela divisão de cada elemento do EAO_t de uma determinada função de ligação pelo seu correspondente no EAO_t da função MIX. Deles também se obtiveram as médias e os máximos.

4.1.1 Exemplo - MIX, logit e Stukel

Neste exemplo são apresentados três cenários de geração de dados e suas estimações pelas seis funções de ligação apresentadas no início deste Capítulo. Os parâmetros fixados e utilizados de acordo com a demanda de cada um foram $\boldsymbol{\alpha} = \begin{bmatrix} 0.25 & 0.25 & 0.25 \end{bmatrix}^T$, $\boldsymbol{\gamma} = \begin{bmatrix} 0.25 & -0.25 \end{bmatrix}^T$ e $\boldsymbol{\beta} = \begin{bmatrix} -10 & 0.2 \end{bmatrix}^T$.

As curvas das funções de ligação geradoras para tais combinações de parâmetros são apresentadas na Figura 5, da qual se é possível notar que para $\eta \in [-6,6]$ há elementos das Imagens no intervalo $[0,1]$, já para os outros valores de η a convergência é explícita.

Para a criação da covariável geraram-se no *software* **R** valores da V.A. $\boldsymbol{X} \sim N(\mu = 50,\ \sigma = 13)$ e, como para a simulação é interessante que se tenha proporções $\boldsymbol{\pi}$ abrangendo a maior parte possível do intervalo $[0,1]$ e não só seus limites, um procedimento que restringisse os valores de $\boldsymbol{\eta}$ foi adotado:

1. Deseja-se que $\boldsymbol{\eta} \in [-6,6]$, o que implica em

$$\begin{aligned} -6 &\lessapprox \beta_0 + \beta_1 \boldsymbol{X} \lessapprox 6 \\ -6-\beta_0 &\lessapprox \beta_1 \boldsymbol{X} \lessapprox 6-\beta_0 \\ \frac{-6-\beta_0}{\beta_1} &\lessapprox \boldsymbol{X} \lessapprox \frac{6-\beta_0}{\beta_1}. \end{aligned} \tag{4.8}$$

2. Iguala-se o extremo inferior (superior) da Equação 4.8 ao valor mínimo (máximo) da covariável decrescido (acrescido) de um $\varepsilon > 0$ obtendo-se as seguintes equações para β_0:

$$(I)\ \beta_0 \approx -6-\beta_1\{min(\boldsymbol{X})-\varepsilon\} \quad \text{e} \quad (II)\ \beta_0 \approx 6-\beta_1\{max(\boldsymbol{X})+\varepsilon\} \quad .$$

3. Por fim encontra-se β_1 em função dos extremos da covariável e o valor de β_0 fica trivial de ser determinado:

$$\beta_1 \approx \frac{-10}{\{min(\boldsymbol{X})-\varepsilon\}-\{max(\boldsymbol{X})+\varepsilon\}} \approx \frac{-10}{min(\boldsymbol{X})-max(\boldsymbol{X})-2\varepsilon}.$$

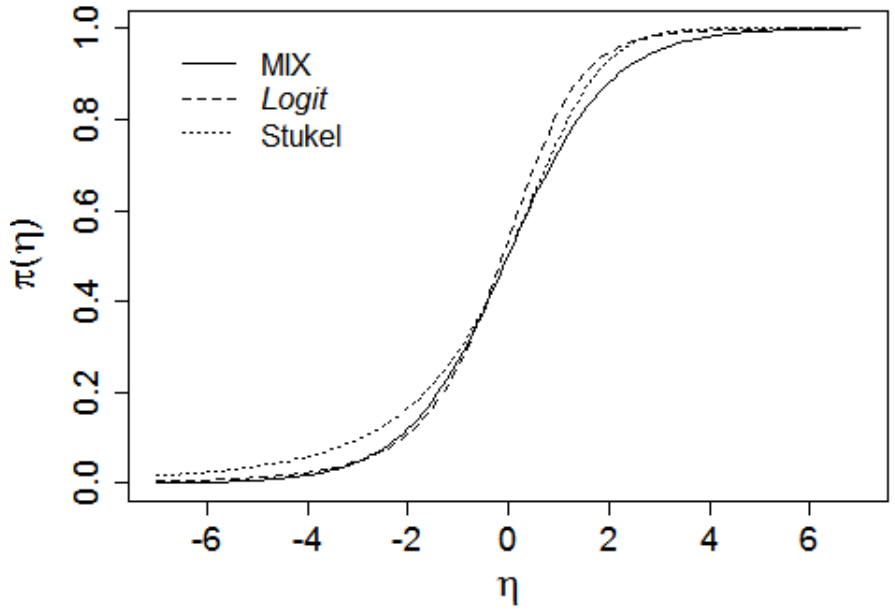


Figura 5 – Curvas das funções de ligação geradoras das amostras.

4.1.1.1 Cenário 1 - MIX

Neste cenário os dados foram gerados a partir da função de ligação MIX com $n.rep = 100$, $n.sample = (100,40,10)$ e $n.b = (100,50,20)$ totalizando 9 combinações diferentes.

As medidas de qualidade de estimações dos parâmetros estimados pela função MIX apresentadas na Tabela 2 demonstram, de um modo geral, que o vetor $\boldsymbol{\alpha}$ e o parâmetro β_1 tem seus EQMs próximos de zero, indicando uma eficiência do modelo em estimar valores próximos

dos fixados, exceto para este último parâmetro com a menor quantidade de observações na amostra (EQM $= 6.804$).

Apesar de em alguns casos as estimações terem atribuído pesos aproximados de 1 para algum componente do vetor $\boldsymbol{\alpha}$ (vide EMAs próximos de 0.7), de um modo geral α_1 teve estimações próximas de seu valor fixado para a geração, α_2 foi subestimado, de acordo com seus viéses, e α_3 superestimado (exceto em $n.s = 10$, em que ocorre o oposto).

Os valores de todas as medidas do vetor $\boldsymbol{\gamma}$ e de β_0 estão muito altos ou baixos, indicando que as estimações não resultaram em números próximos aos estabelecidos na geração. Para o primeiro, o que ocorre é que a partir de determinados valores de $\boldsymbol{\gamma}$ a log-verossimilhança perfilada fica constante e, dentre aqueles valores de parâmetro que levam aos mesmos resultados, em termos de otimização, o método numérico utilizado não distingue qual deles escolher, podendo, assim, por vezes estimar valores próximos ao fixado e por outras não. Um exemplo prático é mostrado na seção 5.1.

O fato explicado acima, contudo, não prejudica a estimação das proporções em si, uma vez que a questão de identificabilidade leva a resultados próximos aos daqueles obtidos caso as medidas de qualidade dos parâmetros fossem menores.

Já para β_0, que é um parâmetro comum à todas funções de ligação que compõe a mistura, ou seja, está presente na mesma independente dos valores de $\boldsymbol{\alpha}$, possíveis processos de estimação que não convergiram acabaram por gerar valores discrepantes de estimativas, influenciando em todas as métricas.

As Figura 6, Figura 8 e Figura 10 apresentam os gráficos de dispersão das estimativas dos parâmetros em questão para cada amostra t na combinação de $n.s = 10$ e $n.b = 20$. Ainda, gráficos de dispersão limitando os valores estimados em um intervalo pequeno em relação aos valores fixados são feitos em figuras posteriormente comentadas.

O objetivo é, apesar do abuso de generalização, compreender e explicitar melhor quais fatores levam a medidas tão altas ou baixas, nesses parâmetros.

Sem restrição de valores estimados, nota-se na Figura 6a que dois pontos são extremamente discrepantes, devido a falta de convergênca do algoritmo ou ao pseudoproblema de identificabilidade, mas o fato é que são eles que fazem as métricas de qualidade dos parâmetros atingirem valores tão estrondosos. Já quando se restringe o espaço de estimativas a um intervalo proporcionalmente próximo ao valor fixado na geração ($\gamma_1 = 0.25$), apenas em 26 amostras se encontram resultados plausíveis. Já para γ_2, mesmo com a restrição, apenas duas estimativas estão próximas ao valor fixado na geração ($\gamma_2 = -0.25$), conforme se verifica em Figura 8b, o que, posteriormente, não demonstrará ser um problema ao se analisar as estimações de proporções.

Tabela 2 – EQM, Viés e EMA para dados gerados e estimados pela função de ligação MIX com $n.s = (100, 40, 10)$, $n.b = (100, 50, 20)$, $\boldsymbol{\alpha} = (0.25, 0.25, 0.25)$, $\boldsymbol{\gamma} = (0.25, -0.25)$ e $\boldsymbol{\beta} = (-10, 0.2)$

		$n.s = 100$			$n.s = 40$			$n.s = 10$		
Par.[1]	Med. [2]	$n.b = 100$	$n.b = 50$	$n.b = 20$	$n.b = 100$	$n.b = 50$	$n.b = 20$	$n.b = 100$	$n.b = 50$	$n.b = 20$
α_1	EQM	0.040	0.029	0.031	0.047	0.032	0.045	0.058	0.049	0.059
	Viés	0.052	0.031	-0.008	0.077	-0.028	-0.030	-0.017	-0.083	-0.092
	EMA	0.489	0.538	0.585	0.545	0.503	0.564	0.692	0.463	0.724
α_2	EQM	0.040	0.043	0.046	0.042	0.050	0.055	0.046	0.058	0.059
	Viés	-0.159	-0.166	-0.194	-0.169	-0.185	-0.210	-0.180	-0.202	-0.208
	EMA	0.435	0.553	0.250	0.304	0.421	0.403	0.356	0.487	0.583
α_3	EQM	0.067	0.069	0.076	0.061	0.077	0.069	0.066	0.083	0.062
	Viés	0.132	0.117	0.092	0.050	0.107	0.004	-0.058	-0.033	-0.128
	EMA	0.550	0.592	0.603	0.629	0.710	0.711	0.640	0.723	0.724
γ_1	EQM	1.48E+08	4.86E+03	6.88E+06	8.67E+04	1.96E+03	8.18E+05	2.14E+07	1.44E+08	2.16E+08
	Viés	1298.652	14.448	335.208	42.899	15.174	117.462	-427.302	-1.57E+03	-2.06E+03
	EMA	1.21E+05	386.773	2.57E+04	2.91E+03	2.71E+02	9.00E+03	4.61E+04	1.09E+05	1.19E+05
γ_2	EQM	477.280	2.95E+18	2.46E+08	7.94E+04	1.04E+06	1.56E+07	3.57E+07	9.86E+06	5.33E+06
	Viés	7.981	-1.72E+08	1.59E+03	-12.859	-83.246	-504.061	706.505	-316.912	-38.131
	EMA	154.812	1.72E+10	1.57E+05	2.74E+03	1.01E+04	3.58E+04	5.94E+04	3.13E+04	1.96E+04
β_0	EQM	6.315	8.321	11.166	8.506	11.628	15.980	14.650	17.440	1.47E+04
	Viés	2.222	2.588	3.098	2.612	3.005	3.583	3.258	3.515	-8.223
	EMA	4.197	4.503	5.765	4.735	7.539	6.775	6.707	7.492	1.21E+03
β_1	EQM	0.003	0.003	0.004	0.003	0.005	0.006	0.006	0.007	6.804
	Viés	-0.045	-0.052	-0.062	-0.053	-0.060	-0.072	-0.066	-0.070	0.184
	EMA	0.083	0.089	0.114	0.097	0.153	0.135	0.137	0.150	26.063

A análise das estimativas de β_0 se dá de forma semelhante a do vetor anteriormente citado, porém neste caso crê-se que o problema é causado totalmente por falta de convergência do algoritmo, uma vez que pela Figura 10b nota-se que ao se limitar o intervalo de estimativas 97 delas se enquadram na restrição.

Outro ponto interessante de se observar são as relação de tais estimativas com seus devidos erros padrões (EP). Nota-se, das Figura 7a, Figura 9a e Figura 11a, que as mesmas observações que causam os excessos nas medidas de qualidade são as que possuem os maiores EP's (exceto para um ponto de γ_2), corroborando mais ainda com a suspeita de não-convergência do algoritmo nesses casos.

Por fim, supôs-se que as amostras cujas estimações de $\boldsymbol{\gamma}$ fossem discrepantes dos valores fixados não tivessem peso elevado na mistura (o que acarretaria em menores problemas de estimação de proporções). Entretanto, nota-se da Figura 12 que para $\hat{\gamma}_1$ o peso da função de ligação Stukel nas amostras em questão são, aproximadamente, 0.6 e 0.9 (há amostras com pesos maiores) e, mesmo assim, suas médias de EAOs não ultrapassam 8%.

Já para os casos peculiares de $\hat{\gamma}_2$ os pesos não são tão altos e as médias de EAOs idem. Além disso, não se pode observar uma correlação entre os componentes de $\boldsymbol{\gamma}$.

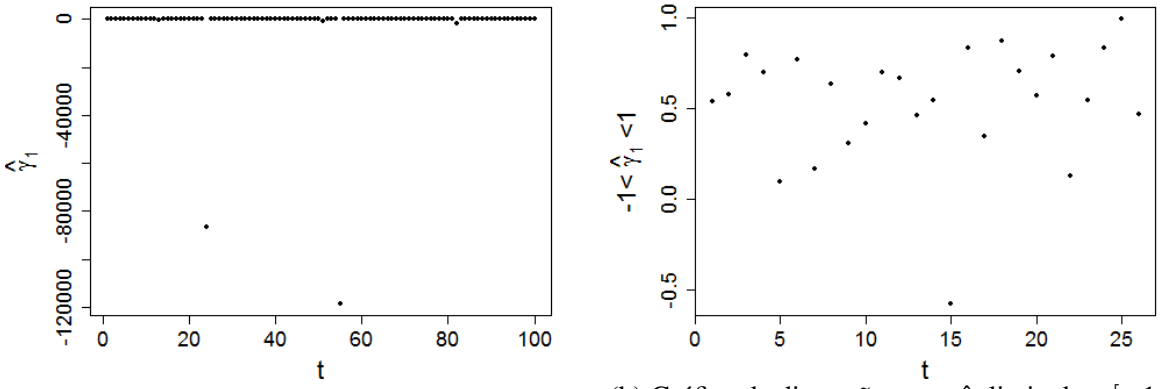
Quando se analisa a dispersão entre pesos e erros é possível notar que a maioria dos

[1] Abreviação de "Parâmetro"
[2] Abreviação de "Medida"

pesos atribuídos à ligação Stukel se encontram no intervalo $[0.2, 1]$ e que os erros das estimações da FL MIX ainda se mantém muito baixos.

Isto posto, pode-se entender que mesmo para os casos em que há estimações de valores muito altos (ou baixos) para os parâmetros que modelam as caudas da função de ligação Stukel, as estimativas de proporções da ligação MIX se mantém próximas das proporções observadas. As outras FLs que compõe a ligação proposta balanceiam as estimações marginais.

Figura 6 – Gráficos de dispersão de $\hat{\gamma}_1$ para dados gerados e estimados pela função de ligação MIX com $n.s = 10$, $n.b = 20$, $\boldsymbol{\alpha} = (0.25, 0.25, 0.25)$, $\boldsymbol{\gamma} = (0.25, -0.25)$ e $\boldsymbol{\beta} = (-10, 0.2)$

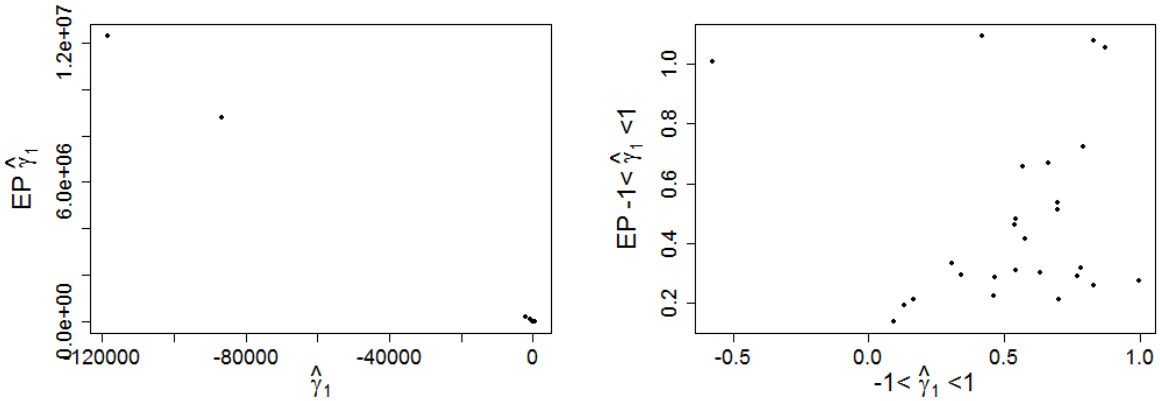


(a) Gráfico de dispersão entre $\hat{\gamma}_1$ e t

(b) Gráfico de dispersão entre $\hat{\gamma}_1$ limitado a $[-1, 1]$ e t

Fonte: Elaborada pelo autor.

Figura 7 – Gráficos de dispersão de $\hat{\gamma}_1$ e seus erros padrões (EP) para dados gerados e estimados pela função de ligação MIX com $n.s = 10$, $n.b = 20$, $\boldsymbol{\alpha} = (0.25, 0.25, 0.25)$, $\boldsymbol{\gamma} = (0.25, -0.25)$ e $\boldsymbol{\beta} = (-10, 0.2)$

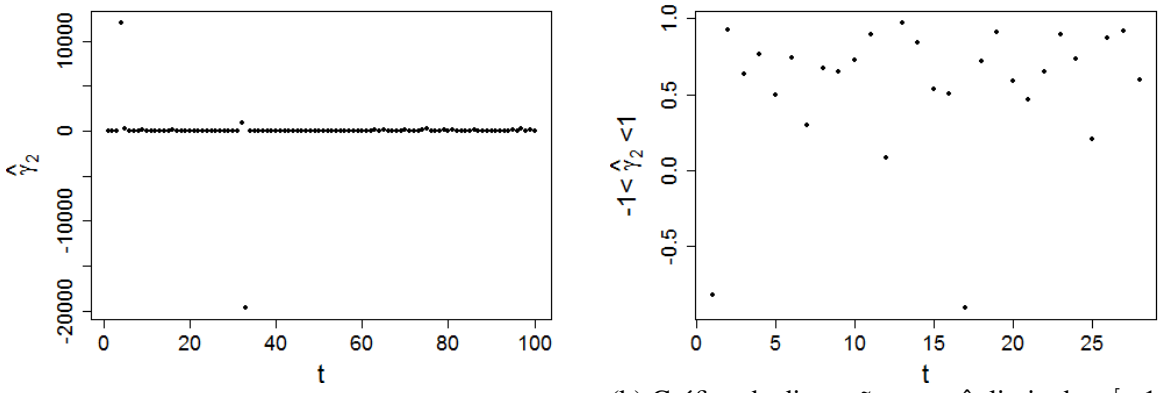


(a) Gráfico de dispersão entre $\hat{\gamma}_1$ e seus erros padrões (EP)

(b) Gráfico de dispersão entre $\hat{\gamma}_1$ limitado a $[-1, 1]$ e seus erros padrões (EP)

Fonte: Elaborada pelo autor.

Figura 8 – Gráficos de dispersão de $\hat{\gamma}_2$ para dados gerados e estimados pela função de ligação MIX com $n.s = 10$, $n.b = 20$, $\boldsymbol{\alpha} = (0.25, 0.25, 0.25)$, $\boldsymbol{\gamma} = (0.25, -0.25)$ e $\boldsymbol{\beta} = (-10, 0.2)$

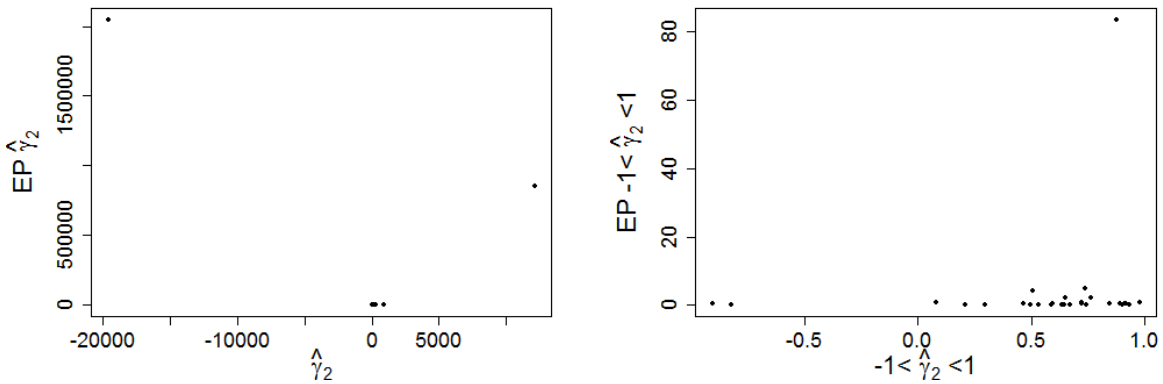


(a) Gráfico de dispersão entre $\hat{\gamma}_2$ e t

(b) Gráfico de dispersão entre $\hat{\gamma}_2$ limitado a $[-1, 1]$ e t

Fonte: Elaborada pelo autor.

Figura 9 – Gráficos de dispersão de $\hat{\gamma}_2$ e seus erros padrões (EP) para dados gerados e estimados pela função de ligação MIX com $n.s = 10$, $n.b = 20$, $\boldsymbol{\alpha} = (0.25, 0.25, 0.25)$, $\boldsymbol{\gamma} = (0.25, -0.25)$ e $\boldsymbol{\beta} = (-10, 0.2)$

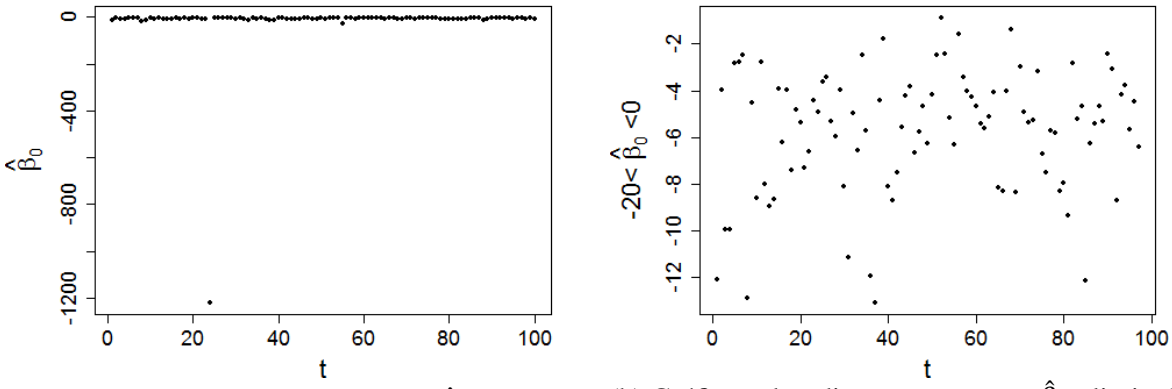


(a) Gráfico de dispersão entre $\hat{\gamma}_2$ e seus erros padrões (EP)

(b) Gráfico de dispersão entre $\hat{\gamma}_2$ limitado a $[-1, 1]$ e seus erros padrões (EP)

Fonte: Elaborada pelo autor.

Figura 10 – Gráficos de dispersão de $\hat{\beta}_0$ para dados gerados e estimados pela função de ligação MIX com $n.s = 10$, $n.b = 20$, $\boldsymbol{\alpha} = (0.25, 0.25, 0.25)$, $\boldsymbol{\gamma} = (0.25, -0.25)$ e $\boldsymbol{\beta} = (-10, 0.2)$

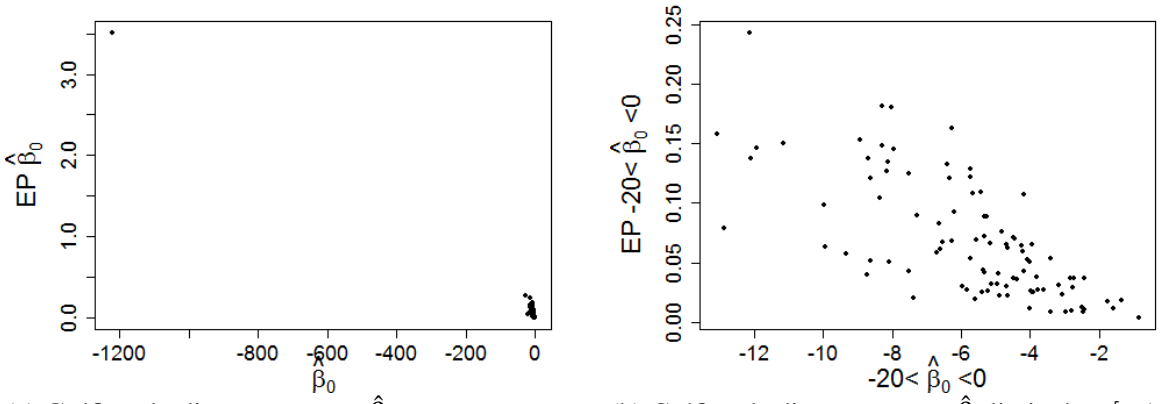


(a) Gráfico de dispersão entre $\hat{\beta}_0$ e t

(b) Gráfico de dispersão entre $\hat{\beta}_0$ limitado a $[-20, 0]$ e t

Fonte: Elaborada pelo autor.

Figura 11 – Gráficos de dispersão de $\hat{\beta}_0$ e seus erros padrões (EP) para dados gerados e estimados pela função de ligação MIX com $n.s = 10$, $n.b = 20$, $\boldsymbol{\alpha} = (0.25, 0.25, 0.25)$, $\boldsymbol{\gamma} = (0.25, -0.25)$ e $\boldsymbol{\beta} = (-10, 0.2)$



(a) Gráfico de dispersão entre $\hat{\beta}_0$ e seus erros padrões (EP)

(b) Gráfico de dispersão entre $\hat{\beta}_0$ limitado a $[-1, 1]$ e seus erros padrões (EP)

Fonte: Elaborada pelo autor.

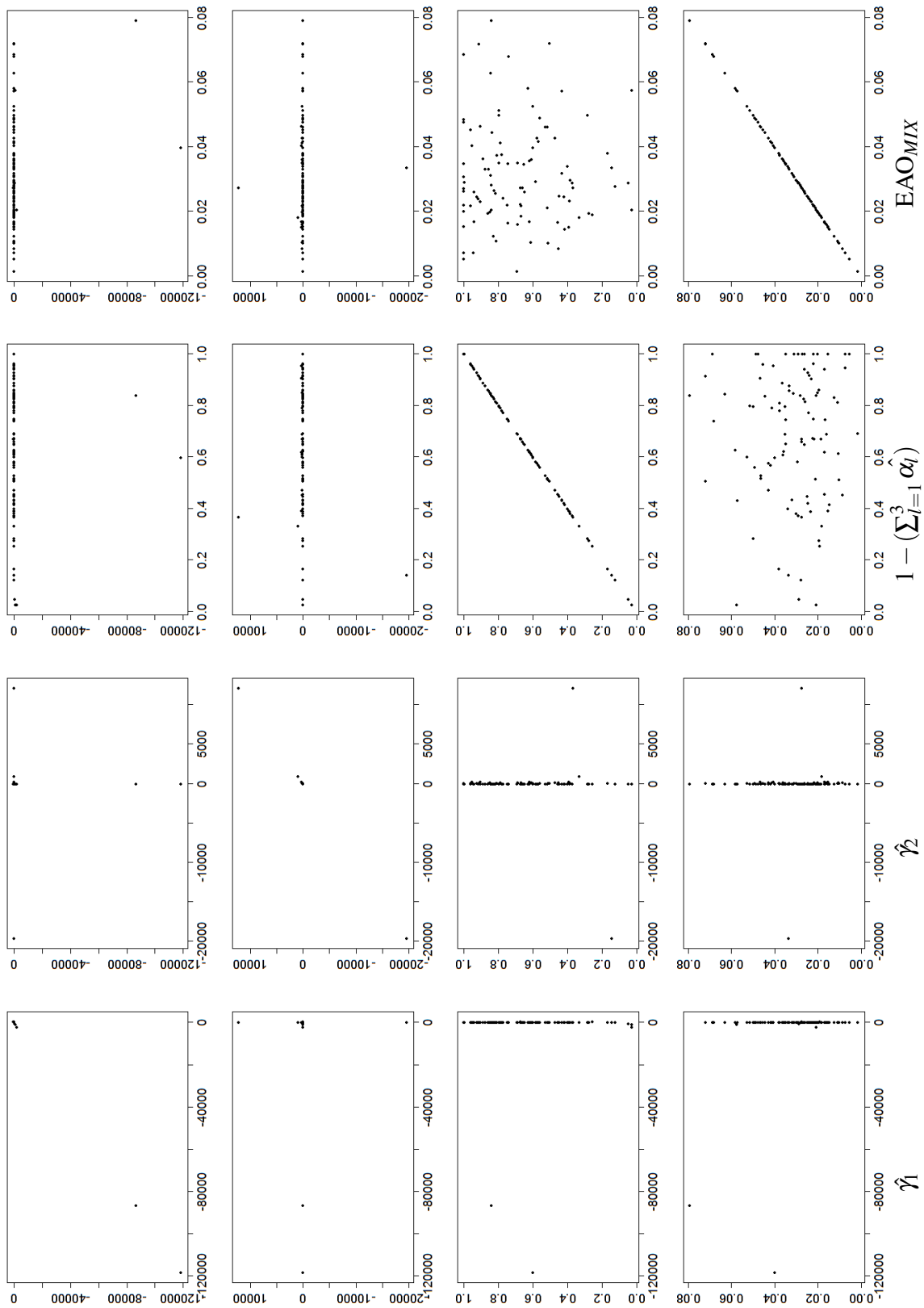


Figura 12 – Gráficos de dispersão de $\hat{\gamma}_1$, $\hat{\gamma}_2$, $1-(\sum_{l=1}^{3}\hat{\alpha}_l)$ e médias de EAO$_{MIX}$ para dados gerados e estimados pela função de ligação MIX com $n.s = 10$, $n.b = 20$, $\boldsymbol{\alpha} = (0.25, 0.25, 0.25)$, $\boldsymbol{\gamma} = (0.25, -0.25)$ e $\boldsymbol{\beta} = (-10, 0.2)$

Das métricas relativas às proporções estimadas as análises são feitas com base nas Figura 13 e Figura 14.

Analisando as distribuições das médias de erros absolutos em relação às proporções observadas a partir da Figura 13 nota-se que independentemente da combinação utilizada há quase um padrão: são, em sua maioria, simétricas e apresentam *outliers*; a ligação Stukel tem as observações mais heterogêneas e com os maiores valores de todos os quartis; a ligação MIX as mais homogêneas e com menores valores; a ligação Aranda-Ordaz demonstra um comportamento semelhante ao da função de ligação Stukel, porém um pouco deslocado para cima; e as ligações *logit*, *probit* e complementar log-log seguem, entre si e respectivamente, uma crescente nos valores de seus quartis.

Nas situações em que $n.b = (100, 50)$, a maior observação da função de ligação MIX é menor que a mediana da ligação Stukel e os maiores valores das outras distribuições são menores que o 3^{o} Quartil dela. Entretanto, para $n.b = 20$ e $n.s. = 10$, seu boxplot fica mais homogêneo e o padrão antes citado deixa de ser verdade, apesar de um de seus *outliers* errar quase 35%, mais de dez vezes o que erra a ligação MIX na simulação com o maior número de observações.

Analisando proporcionalmente a maior média de erros em relação à menor, há um aumento, desta para aquela, de aproximadamente 500 % na FL Stukel, 300% nas FLs MIX e Aranda-Ordaz e entre 200% a 250% nas outras. Isto é, além de ser a função de ligação que resulta nas maiores médias de erros é a que mais aumenta proporcionalmente essa medida na comparação dos casos extremos. Já a ligação proposta, apesar do bom desempenho das estimações, apresenta uma variação menor que esta, porém maior que as das demais funções.

Já os máximos dos erros absolutos em relação às proporções observadas têm suas distribuições apresentadas na Figura 14 e a ligação MIX continua sendo a com os menores valores de tal medida e a ligação Stukel a FL com os maiores.

Especificamente para $n.s = 10$ é possível notar que a caudas inferiores dos boxplots referentes à FL MIX estão próximos de zero indicando que, para pelos menos uma amostra, todas as estimações de proporções oscilaram infinitesimalmente em torno dos valores observados.

Fazendo uma análise marginal da mudança das quantidades de valores da covariável nota-se que, majoritariamente, aumentam-se os extremos das distribuições da medida em questão conforme o valor de $n.s$ aumenta, assim como a quantidade de pontos discrepantes também.

Para as quantidades $n.b$ definidas em cada simulação, tem-se que, quanto menor elas forem, mais próximos estarão os quartis das diferentes funções. Vale frisar que, para o caso com a menor quantidade de observações na amostra, a ligação mistura erra, no máximo, cerca de 30% referente à proporção observada, ao passo que as demais erram aproximadamente 40% ou mais. Uma diferença bastante significativa.

Mantém-se, ainda, a maior heterogeneidade para a FL Stukel, as semelhanças entre as ligações MIX e Aranda-Ordaz, a relação entre as funções de ligação *logit*, *probit* e complementar

log-log e um melhor desempenho da ligação proposta no que diz respeito aos menores máximos de erros absolutos.

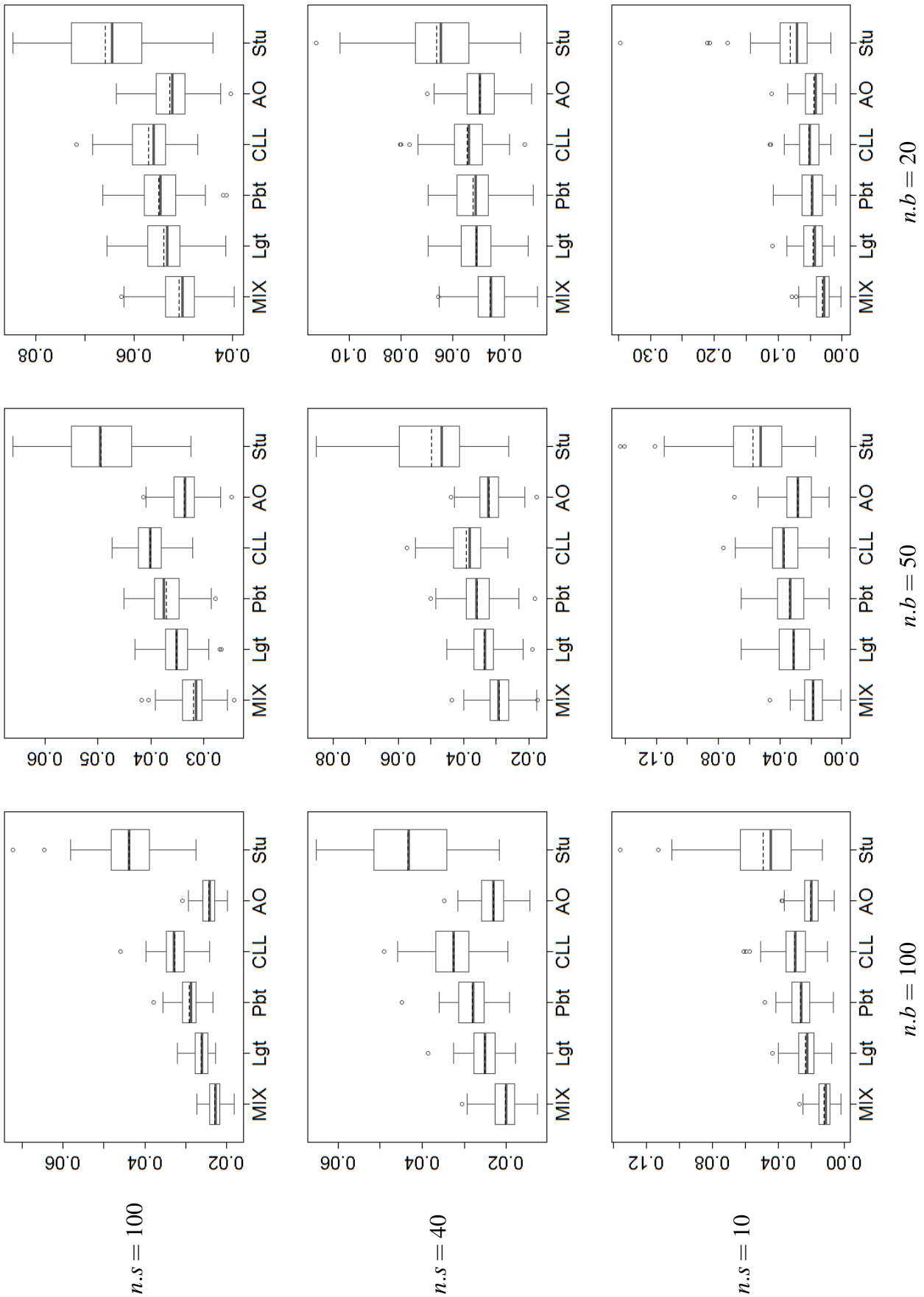


Figura 13 – Boxplot de médias dos EAO do Cenário 1

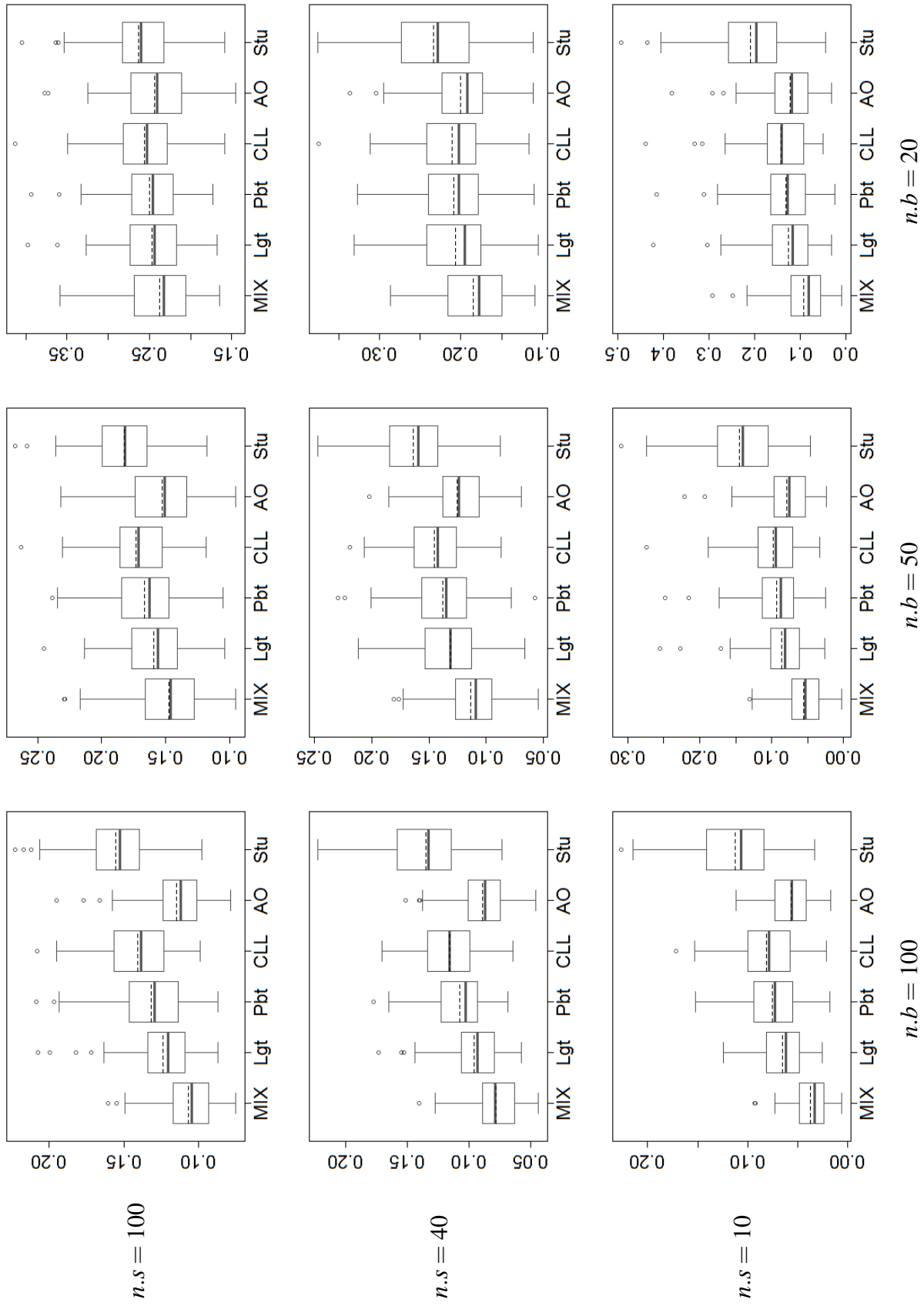


Figura 14 – Boxplot de máximos dos EAO do Cenário 1

4.1.1.2 Cenário 2 - Logit

Neste cenário os dados foram gerados a partir da função de ligação *logit* com $n.rep = 100$ fixo, $n.sample = (100, 40, 10)$ e $n.b = (100, 50, 20)$ totalizando 9 combinações diferentes.

Observa-se, das medidas apresentadas na Tabela 3, que o modelo *logit* estima valores $\hat{\beta}_1$ muito próximos ao fixado na geração das amostras. Em todas as combinações seu EQMs são praticamente 0 e seus viéses estão em torno dessa medida, mesmo com a presença de estimativas destoantes em até $1/4$ do valor de referência nas comparações de tais medidas (caso de $n.s = 40$ com $n = 20$ e $n.s = 10$ com $n.b = 50$).

Já para o outro componente do vetor $\boldsymbol{\beta}$ os viéses estão próximos de zero, indicando que as médias das estimativas nos diversos casos tendem ao valor de β_0 previamente especificado; os EQMs não estão tão próximos de 0 e aumentam conforme se diminuem $n.s$ e $n.b$, assim como os EMAs, que, no pior caso, apresentam um parâmetro quase 100% maior (8.592) que o valor usado na geração. Entretanto, para os outros valores de ensaios de Bernoulli, as medidas observadas indicam estimações plausíveis.

Tabela 3 – EQM, Viés e EMA para dados gerados e estimados pela função de ligação *logit* com $n.s = (100, 40, 10)$, $n.b = (100, 50, 20)$, $\boldsymbol{\alpha} = (0.25, 0.25, 0.25)$, $\boldsymbol{\gamma} = (0.25, -0.25)$ e $\boldsymbol{\beta} = (-10, 0.2)$

		$n.s = 100$			$n.s = 40$			$n.s = 10$		
Parâmetro	Medida	$n.b = 100$	$n.b = 50$	$n.b = 20$	$n.b = 100$	$n.b = 50$	$n.b = 20$	$n.b = 100$	$n.b = 50$	$n.b = 20$
β_0	EQM	0.035	0.081	0.201	0.119	0.219	0.607	0.350	1.034	2.960
	Viés	-0.019	0.003	-0.008	-0.010	-0.021	-0.029	0.056	-0.153	-0.187
	EMA	0.559	0.861	1.140	0.986	1.011	2.479	1.582	2.843	8.592
β_1	EQM	~ 0	~ 0	~ 0	~ 0	~ 0	~ 0	~ 0	~ 0	0.001
	Viés	~ 0	~ 0	~ 0	~ 0	~ 0	0.001	-0.001	0.003	0.004
	EMA	0.011	0.018	0.022	0.019	0.022	0.053	0.032	0.055	0.160

Partindo-se para a análise das médias de erros absolutos em relação às proporções observadas, nota-se na Figura 15 um comportamento já esperado das distribuições de tal medida para as diversas funções de ligação utilizadas: pelos fato dos dados gerados serem, em teoria, simétricos, modelos com tal características (mesmo que em casos particulares) estimam proporções mais próximas às observadas e vice-versa. Na prática, isso significa que a função de ligação complementar log-log apresenta os boxplots mais deslocados para cima, como se pode verificar nas combinações de $n.s = (100, 40)$ e $n.b = (100, 50)$ (apesar da escala do eixo das ordenadas ser muito pequeno). Há um caso em que suas observações situadas entre os 1^o e 4^o quartis são maiores que todas as observações das demais funções de ligação, porém essa situação muda para $n.s = 10$ e $n.b = 20$, em que a ligação Stukel se destoa negativamente das demais.

Deixando de lado esses padrões de piores desempenhos, de um modo geral nota-se que a diminuição marginal de $n.s$ gera um aumento na amplitude dos boxplots e explicita a diferença da função de ligação proposta para as demais quando seu terceiro quartil vai da igualdade com as medianas de *logit*, *probit* e Aranda-Ordaz para a igualdade com os primeiros quartis.

Variando de forma decrescente o valor de $n.b$ pode-se notar as medianas das FLs *logit*, *probit* e Aranda-Ordaz se igualando.

Ao se observar a Figura 16 nota-se para as distribuições dos máximos dos erros absolutos o mesmo comportamento das distribuições de médias, corroborando com as conclusões já escritas. Para $n.s = 10$ e $n.b = 100$, todas as funções apresentam boxplots com os menores valores de máximo, exceto pela ligação Stukel que sofre a influência de um *outlier*, fazendo com que seu melhor desempenho seja com $n.b = 40$.

Em suma, não existe um padrão bem definido do comportamento dos boxplots em ambas as figuras, porém a função de ligação mistura sempre apresenta os menores valores para todos os quartis e a ligação *logit*, que por pressuposição seria a que melhor modelaria os dados, por vezes se demonstra pior não só que a FL MIX como pior que a Aranda-Ordaz também, no sentido de seus máximos de EAOs serem maiores que os das demais FLs.

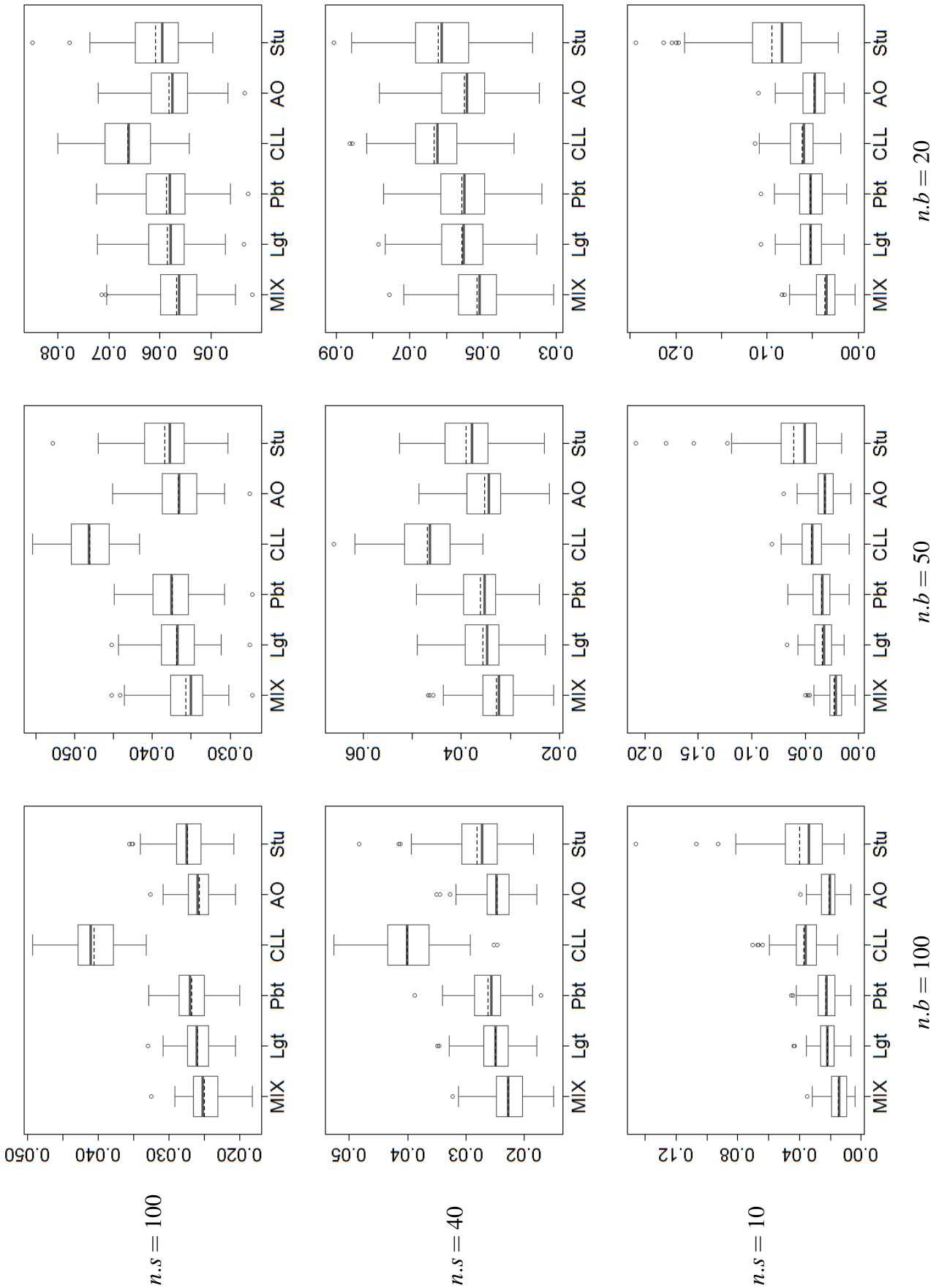


Figura 15 – Boxplot de médias dos EAO do Cenário 2

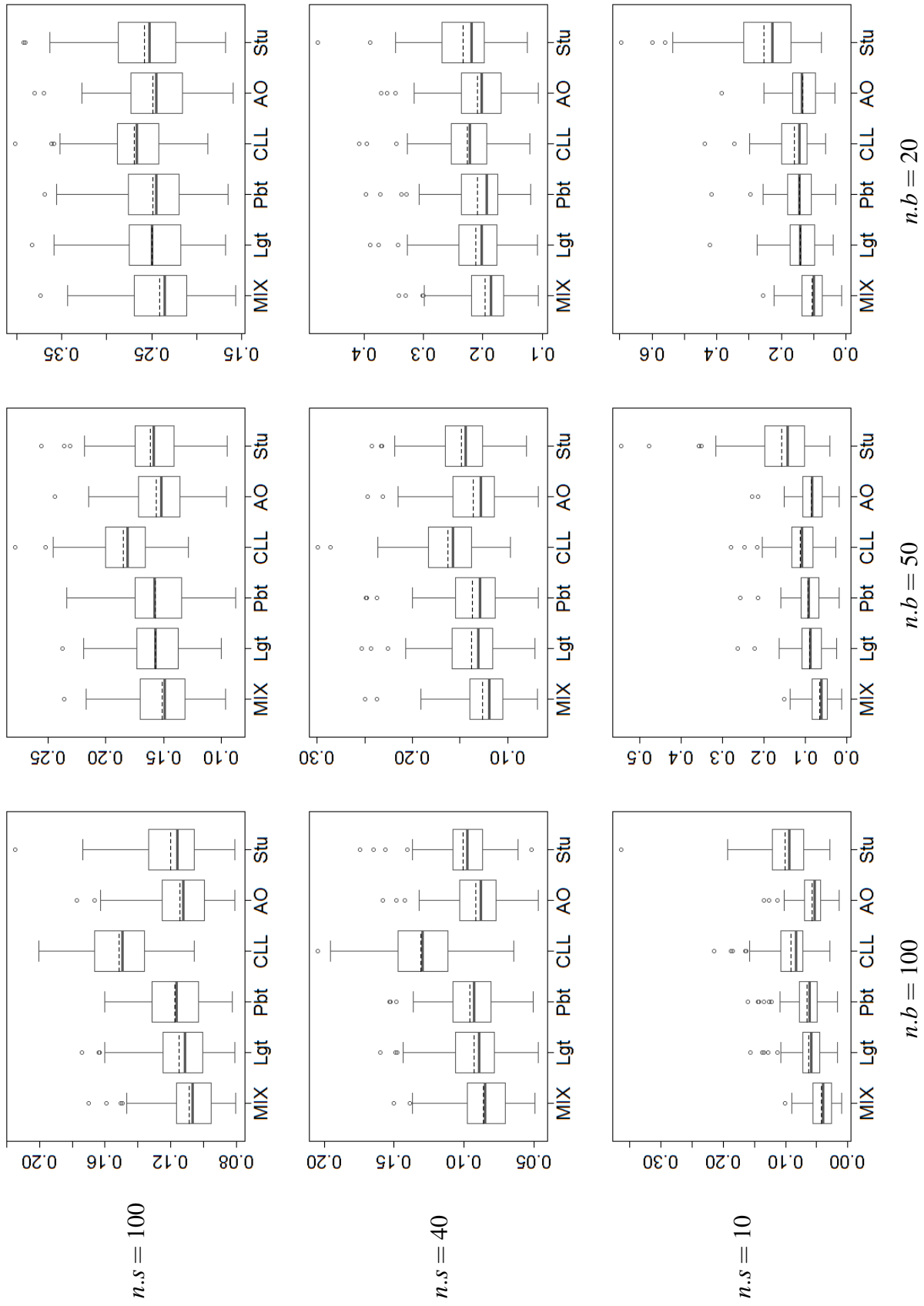


Figura 16 – Boxplot de máximos dos EAO do Cenário 2

4.1.1.3 Cenário 3 - Stukel

Neste cenário os dados foram gerados a partir da função de ligação Stukel com $n.rep = 100$ fixo, $n.sample = (100, 40, 10)$ e $n.b = (100, 50, 20)$ totalizando 9 combinações diferentes.

Começando a análise pelos parâmetros $\boldsymbol{\gamma}$ modeladores das caudas da função de ligação, observa-se da Tabela 4, que, para $n.s = (100, 40)$, os EQMs e viéses estão próximos de zero e esta última medida chega, em seu pior caso, a atingir aproximadamente 20% do valor fixado para a geração. As piores estimativas obtidas independente do valor de $n.b$, ultrapassam, em sua maioria, o dobro do parâmetro verdadeiro.

Já para $n.s = 10$, conforme se diminui $n.b$, maiores ficam os valores (estratosféricos) das medidas. No pior dos casos há um EQM de 493.343 para γ_1, quase 2000 vezes o valor do parâmetro. Viéses e EMAs também apresentam valores anormais, exceto para γ_2 com $n.b = 100$.

Dos parâmetros componentes do preditor linear, as medidas de β_1 são baixas e plausíveis de boas estimações, exceto para a combinação de $n.s = 10$ e $n.b = 20$, em que o EMA é altíssimo (28.462) e, consequentemente, seu EQM (8.604) também o é. Já para β_0, em $n.s = (100, 40)$ os viéses são baixos e os EMAs altos, chegando a atingir quase a totalidade (8.696) do valor fixado na geração. Porém, ao se analisar as medidas em $n.s = 10$ (exceto para o viés quando $n = 100$) nota-se que todos os valores são extremamente altos.

Tabela 4 – EQM, Viés e EMA para dados gerados e estimados pela função de ligação Stukel com $n.s = (100, 40, 10)$, $n.b = (100, 50, 20)$, $\boldsymbol{\alpha} = (0.25, 0.25, 0.25)$, $\boldsymbol{\gamma} = (0.25, -0.25)$ e $\boldsymbol{\beta} = (-10, 0.2)$

		$n.s = 100$			$n.s = 40$			$n.s = 10$		
Par.[1]	Med. [2]	$n.b = 100$	$n.b = 50$	$n.b = 20$	$n.b = 100$	$n.b = 50$	$n.b = 20$	$n.b = 100$	$n.b = 50$	$n.b = 20$
γ_1	EQM	0.004	0.011	0.026	0.010	0.031	0.077	2.326	37.592	493.343
	Viés	-0.020	0.008	0.024	-0.008	0.022	0.054	0.147	-0.396	-2.342
	EMA	0.160	0.311	0.538	0.313	0.626	1.177	13.579	60.672	217.553
γ_2	EQM	0.004	0.011	0.025	0.011	0.020	0.049	0.159	7.596	9.499
	Viés	-0.009	0.018	0.036	-0.002	0.007	0.008	0.032	0.290	0.553
	EMA	0.158	0.296	0.413	0.405	0.420	0.654	2.615	26.614	25.949
β_0	EQM	0.463	0.870	1.930	0.938	1.918	5.042	7.518	371.314	18660.813
	Viés	-0.196	0.071	0.190	-0.092	-0.070	-0.124	-0.170	-2.499	-19.618
	EMA	1.695	2.406	3.866	2.437	4.574	8.696	8.841	188.347	1321.383
β_1	EQM	~ 0	~ 0	0.001	~ 0	0.001	0.002	0.003	0.168	8.604
	Viés	0.004	-0.001	-0.004	0.002	0.001	0.002	0.003	0.052	0.413
	EMA	0.033	0.046	0.077	0.048	0.090	0.172	0.180	4.021	28.462

Da Figura 17, relativa às distribuições das médias de erros absolutos em relação às proporções observadas, por menos esperado que fosse, a ligação Stukel apresenta os maiores valores e as maiores amplitudes em todas as combinações, sendo seguida pelas funções de ligação simétricas *logit* e *probit*.

[1] Abreviação de "Parâmetro"
[2] Abreviação de "Medida"

As ligações assimétricas, desconsiderando a geradora dos dados, apresentam comportamentos semelhantes e, apesar da FL MIX se mostrar mais homogênea na maioria das vezes, há momentos em que ligação Aranda-Ordaz faz esse papel. Todavia, a primeira sempre tem seus quartis menores que os das demais.

Esses padrões se repetem nas médias dos erros absolutos, conforme Figura 18, exceto por alguns pontos discrepantes das FLs complementar log-log e Aranda-Ordaz em $n.s = 10$.

O comportamento da função de ligação Stukel neste Cenário é intrigante pois, apesar de ser a geradora dos dados é, também, a que apresenta as piores estimativas de proporções. Porém, tal fato (assim como as medidas elevadas da Tabela 4) pode ser decorrente de algumas estimações que não convergiram, geraram estimativas altíssimas de $\boldsymbol{\gamma}$ e β_0 e acabaram por interferir tanto nas métricas de qualidade de parâmetros quanto na construção dos boxplots. Acredita-se que este último item foi ocasionado mais pelo componente de $\boldsymbol{\beta}$ que pelo vetor de parâmetros responsáveis pela modelagem das caudas, conforme explicações já dadas sobre a convergência da log-verossimilhança perfilada.

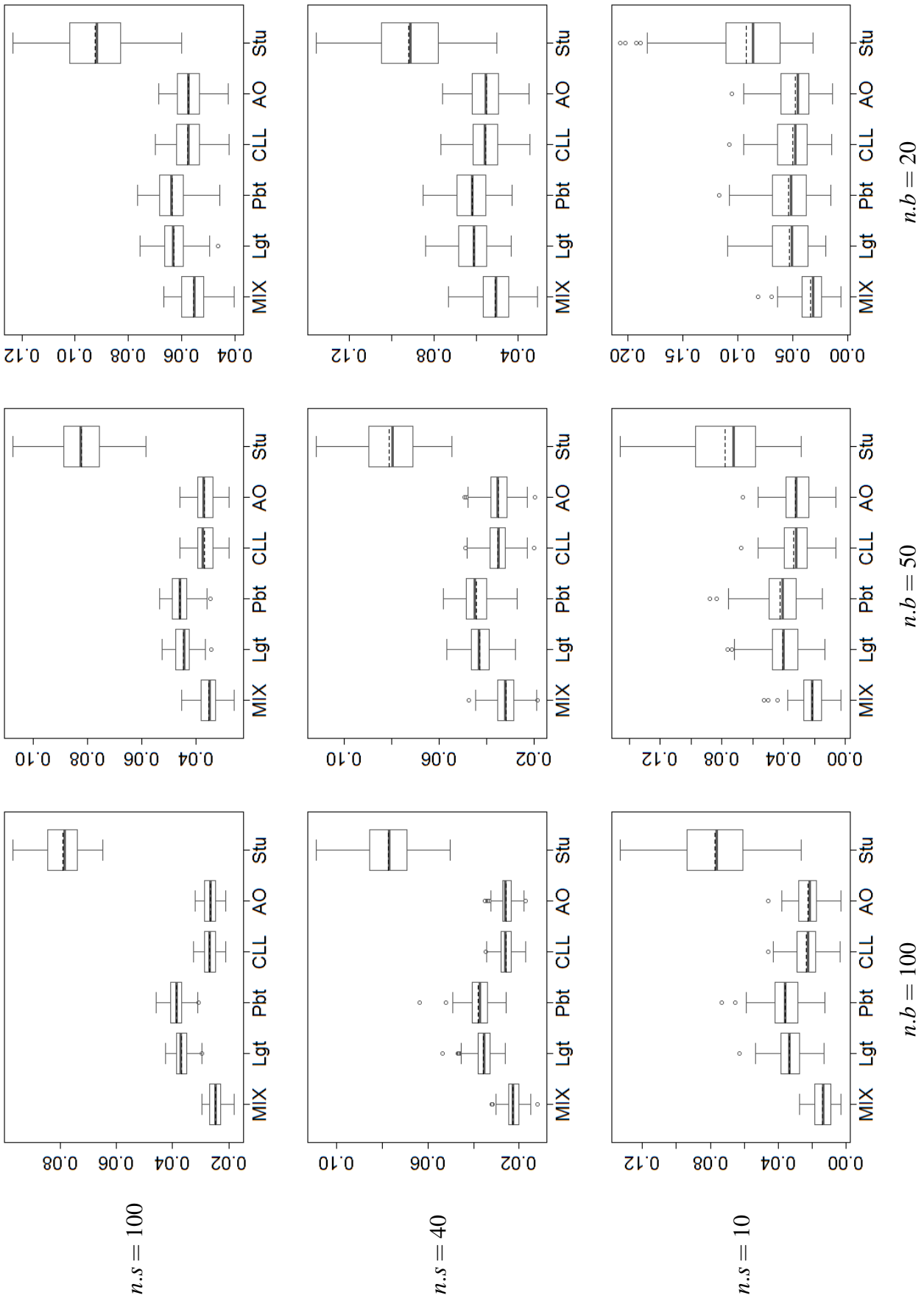


Figura 17 – Boxplot de médias dos EAO do Cenário 3

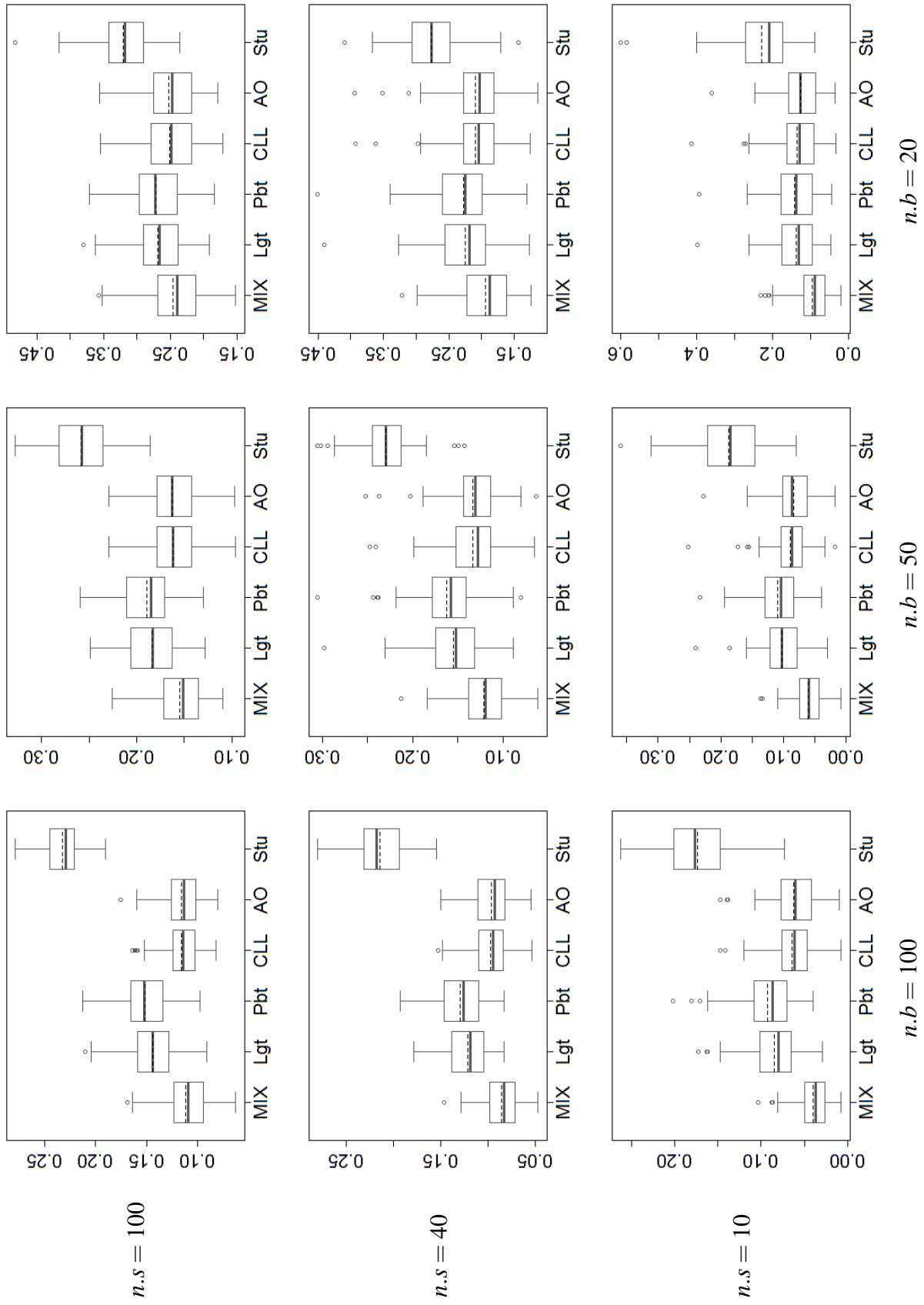


Figura 18 – Boxplot de máximos dos EAO do Cenário 3

4.2 Verificação da Qualidade de Predição

Esta sessão tem como objetivo comparar as capacidades preditivas das funções de ligação, ou seja, o quão bem elas preveem sucessos e fracassos para novas observações.

Foram encontradas dificuldades para se gerar bancos de dados tais que a relação entre as quantidades de sucessos e fracassos e suas covariáveis não fosse linear. Após diversas tentativas (não apresentadas aqui) foi constatado que, fazer previsões para observações usando as mesmas variáveis explicativas que a geraram, levam a resultados iguais para as seis funções de ligação apresentadas. Isto é, independentemente da função de ligação utilizada, as predições serão as mesmas para as observações.

A solução encontrada para a realização desta simulação foi gerar $n.s$ valores quadrivariados da distribuilção $\mathrm{N}_4 \sim (\boldsymbol{\mu}, \Sigma)$ e com elas construir a matriz 4.9 de planejamento.

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \boldsymbol{X}_1 & \boldsymbol{X}_2 & \boldsymbol{X}_3 & \boldsymbol{X}_4 & \boldsymbol{X}_1^2 & \frac{\boldsymbol{X}_3}{\boldsymbol{X}_4} & \sqrt{\boldsymbol{X}_3} & \sin(\boldsymbol{X}_1 \times \boldsymbol{X}_2) \end{bmatrix} \tag{4.9}$$

O cálculo de $\eta_{s,t}$ é feito com um vetor auxiliar de parâmetros $\boldsymbol{\beta}^{\top}$ (cujo cumprimento é igual ao número de colunas da matriz de planejamento) e os de $\pi_{s,t}$ e $W_{s,t}$ são feitos da mesma maneira que na sessão anterior, porém, agora, as observações binomiais são transformadas em $N = n.s \times n.b$ observações dicotômicas $Y_{s,t,i},\ \ i = 1, \ldots, n.b$ constituintes do vetor de variáveis respostas $\boldsymbol{Y}_{t_{N\times 1}}$.

Para uma melhor visualização e entendimento da disposição das variáveis resposta

dicotômicas ao final das *n.rep* repetições, a matriz 4.10 é apresentada.

$$\boldsymbol{Y}^{\boldsymbol{T}}_{N\times n.r} = \begin{bmatrix} \overbrace{Y_{1,1,1}}^{t=1} & \overbrace{Y_{1,2,1}}^{t=2} & \overbrace{Y_{1,3,1}}^{t=3} & \cdots & \overbrace{Y_{1,n.r,1}}^{t=n.r} \\ Y_{1,1,2} & Y_{1,2,2} & Y_{1,3,2} & \cdots & Y_{1,n.r,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{1,1,n.b} & Y_{1,2,n.b} & Y_{1,3,n.b} & \cdots & Y_{1,n.r,n.b} \\ Y_{2,1,1} & Y_{2,2,1} & Y_{2,3,1} & \cdots & Y_{2,n.r,1} \\ Y_{2,1,2} & Y_{2,2,2} & Y_{2,3,2} & \cdots & Y_{2,n.r,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{2,1,n.b} & Y_{2,2,n.b} & Y_{2,3,n.b} & \cdots & Y_{2,n.r,n.b} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ Y_{n.s,1,1} & Y_{n.s,2,1} & Y_{n.s,3,1} & \cdots & Y_{n.s,n.r,1} \\ Y_{n.s,1,2} & Y_{n.s,2,2} & Y_{n.s,3,2} & \cdots & Y_{n.s,n.r,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{n.s,1,n.b} & Y_{n.s,2,n.b} & Y_{n.s,3,n.b} & \cdots & Y_{n.s,n.r,n.b} \end{bmatrix} \begin{matrix} \left.\right\} \pi_{1,t} \\ \left.\right\} \pi_{2,t} \\ \left.\right\} \pi_{n.s,t} \end{matrix} = \begin{bmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \boldsymbol{Y}_3 & \ldots & \boldsymbol{Y}_{n.r} \end{bmatrix}. \tag{4.10}$$

Em todos os cenários de simulações foi utilizada uma adaptação da técnica de *data splitting* apresentada em Friedman, Hastie e Tibshirani (2001). Os autores sugerem a divisão da amostra em três: Treinamento (para ajuste do modelo), Validação (para estimar erro de predição) e Teste (para avaliar o erro do modelo final escolhido). Entretanto, como o objetivo nesta sessão é apenas o de se comparar os modelos e verificar seus comportamentos, não o de escolher um melhor, optou-se por dividir a amostra em 50% para Treinamento, 25% para Ponto (análogo à Validação) e 25% para Teste.

Do vetor $\boldsymbol{Y}_{t_{N\times 1}}$ sorteiam-se aleatoriamente 50% das observações (consequentemente, suas respectivas variáveis explicativas) para comporem a amostra Treinamento e indexa-se-as, e a tudo o que lhes diz respeito, pelo sobrescrito "†". As demais são divididas igual e aleatoriamente entre as amostras Ponto e Teste, indexadas por "▲" e "∗".

Com $Y^{\dagger}_{s,t,i}$ estima-se o vetor de parâmetros $\hat{\boldsymbol{\theta}}_t = [\hat{\alpha}_{1,t}, \hat{\alpha}_{2,t}, \hat{\alpha}_{3,t}, \hat{\gamma}_{1,t}, \hat{\gamma}_{2,t}, \hat{\boldsymbol{\beta}}_t]$ utilizando-se apenas as covariáveis originais da distribuição $\mathrm{N}_4 \sim (\boldsymbol{\mu}, \Sigma)$.

A exclusão das demais componentes da Matriz 4.9 foi a solução encontrada para se resolver o problema de linearidade e, de fato, poder comparar os poderes de predição dos diferentes modelos.

Com $\hat{\boldsymbol{\theta}}_t$, $\boldsymbol{X}^{\blacktriangle}_{s,t}$ e $\boldsymbol{X}^{*}_{s,t}$ estimam-se $\hat{\pi}^{\blacktriangle}_{s,t}$ e $\hat{\pi}^{*}_{s,t}$, respectivamente. Após, a curva ROC é construída com $Y^{\blacktriangle}_{s,t,i}$ e suas probabilidades de sucesso estimadas. O ponto de corte que minimiza $(1-sensiblidade)^2 + (1-especificidade)^2$, proposta de Froud e Abel (2014), é encontrado e, por fim, com ele cria-se a matriz de confusão da amostra Teste, apresentada na Tabela 5, da qual

se calculam a Acurácia (ACC), Sensibilidade (Sens.), Especificidade (Espec.), Valor Predito Positivo (VPP), Valor Predito Negativo (VPN) e o Coeficiente de Correlação de Matthews (CCM) (Matthews (1975)).

Tabela 5 – Matriz de confusão genérica

	Observado	
Predito	1	0
1	VP	FP
0	FN	VN

Os elementos da Tabela 5 são: "verdadeiros positivos" (VP) é o número de sucessos classificados corretamente, "verdadeiros negativos" (VN) é o número de fracassos classificados corretamente, "falsos positivos" (FP) é o número de fracassos classificados incorretamente e "falsos negativos" (FN) é o número de sucessos classificados incorretamente.

A explicação do significado das medidas utilizadas e seus cálculos através da matriz de confusão dão-se na listagem a seguir:

- **Acurácia** é a probabilidade de classificação correta:

$$\text{ACC} = \text{P}(\hat{Y} = 1, Y = 1) + \text{P}(\hat{Y} = 0, Y = 0).$$

$$\widehat{\text{ACC}} = \frac{\text{VP}}{\text{VP+VN+FP+FN}}.$$

- **Sensibilidade** é a probabilidade de prever uma observação como sucesso dado que ela realmente o é:

$$\text{Sens.} = \text{P}(\hat{Y} = 1 | Y = 1) = \frac{\text{P}(\hat{Y} = 1, Y = 1)}{\text{P}(Y = 1)}.$$

$$\widehat{\text{Sens.}} = \frac{VP}{VP + FN}.$$

- **Especificidade** é a probabilidade de prever uma observação como fracasso dado que ela realmente o é:

$$\text{Espec.} = \text{P}(\hat{Y} = 0 | Y = 0) = \frac{\text{P}(\hat{Y} = 0, Y = 0)}{\text{P}(Y = 0)}.$$

$$\widehat{\text{Espec.}} = \frac{VN}{VN + FP}.$$

- **Valor Predito Positivo** é a probabilidade de uma observação ser sucesso dado que ela foi prevista como:

$$\text{VPP} = \text{P}(Y = 1 | \hat{Y} = 1) = \frac{\text{P}(Y = 1, \hat{Y} = 1)}{\text{P}(\hat{Y} = 1)}.$$

$$\widehat{\text{VPP}} = \frac{VP}{VP + FP}.$$

- **Valor Predito Negativo** é a probabilidade de uma observação ser fracasso dado que ela foi prevista como:

$$\text{VPN} = \text{P}(Y = 0|\hat{Y} = 0) = \frac{\text{P}(Y = 0, \hat{Y} = 0)}{\text{P}(\hat{Y} = 0)}.$$
$$\widehat{\text{VPN}} = \frac{VN}{VN + FN}.$$

- **Coeficiente de Correlação de Matthews** mede a coerência da previsão. Varia entra $[-1, 1]$, tal que valores próximos do limite inferior indicam predição inversa e do superior predição perfeita:

$$\widehat{\text{CCM}} = \frac{(VP \times VN) - (FP \times FN)}{\sqrt{(VP + FP) \times (VP + FN) \times (VN + FP) \times (VN + FN)}}.$$

4.2.1 Exemplo - Predição

Para o estudo do comportamento das previsões das seis funções de ligação fixou-se $n.rep = 100$.

Geraram-se, para a construção da matriz de planejamento, $n.s = 1000$ observações de $\text{N}_4 \sim (\boldsymbol{\mu}, \Sigma)$, em que

$$\boldsymbol{\mu} = \begin{bmatrix} 5 & 0 & 90 & 32 \end{bmatrix}^T \quad \text{e} \quad \Sigma = \begin{bmatrix} 2 & 0.3 & 0 & 0.9 \\ 0.3 & 3 & 0 & 0 \\ 0 & 0 & 30 & -0.5 \\ 0.9 & 0 & -0,5 & 50 \end{bmatrix},$$

e para a construção de $\eta_{s,t}$ o vetor $\boldsymbol{\beta}^{\curlywedge} = \frac{1}{9}\begin{bmatrix} -160 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}^T$.

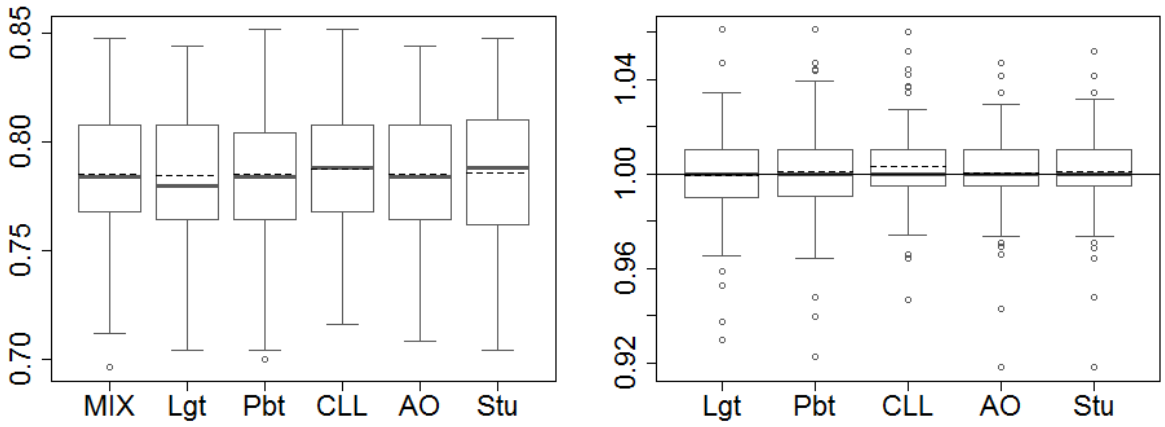
A escolha de $n = 1$ foi feita para que o banco de dados fosse o mais heterogêneo possível e não houvesse observações repetidas, o que poderia ocorrer caso as Binomiais tivessem mais ensaios de Bernoulli.

Após todo os procedimentos de estimações, construções de curvas ROC, definições de pontos de corte e previsões para as amostras Testes, obtiveram-se as distribuições para as medidas de avaliação.

Da Figura 19a nota-se que as distribuições de acurácias são, aparentemente, assimétricas à direita para todos os modelos e que os quartis deles estão bem próximos. Complementar log-log é a função de ligação com os maiores valores deles, o que é um resultado positivo para essa medida.

Já analisando a Figura 19b, constata-se que, embora as distribuições tenham bastante *outliers* e amplitudes diferentes, todas têm suas medianas iguais a 1. Tal fato indica que, independente do modelo utilizado quando comparado com a mistura, haverá um empate na quantidade de acurácias de um que são maiores que a do outro.

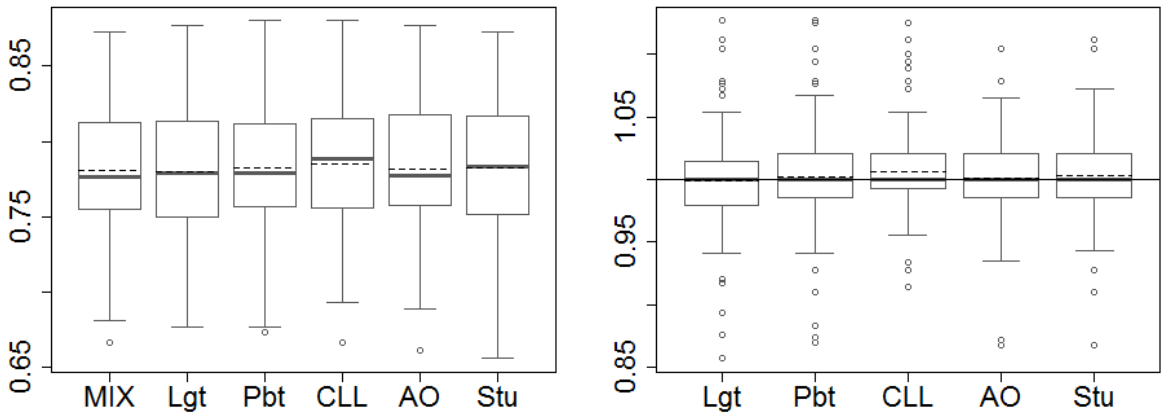
Figura 19 – Boxplots de acurácias da simulação de previsões



(a) Boxplot de acurácias (b) Boxplot de acurácias relativas à MIX

Fonte: Elaborada pelo autor.

Figura 20 – Boxplots de sensibilidades da simulação de previsões



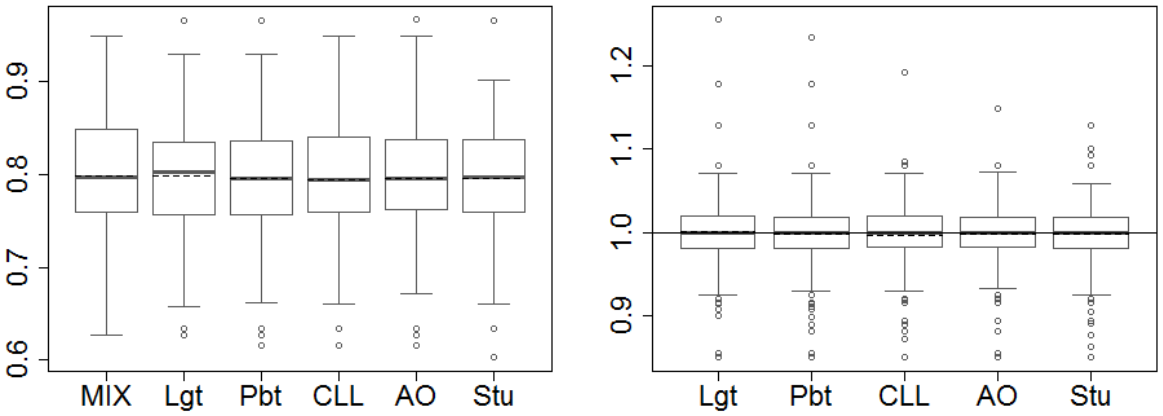
(a) Boxplot de sensibilidades (b) Boxplot de sensibilidades relativas à MIX

Fonte: Elaborada pelo autor.

Na medida "sensibilidade", a FL complementar log-log também é a que apresenta a maior mediana e, juntamente com a ligação *probit*, o maior valor de máximo. Stukel é a função de ligação mais heterogênea e com o menor valor de mínimo, entretanto sua mediana é a segunda maior, de acordo com a Figura 20a. Todavia, novamente, quando se observa os valores relativos à ligação MIX pela Figura 20b se nota que as medianas são iguais a 1.

As análises das Figura 21a, Figura 22a, Figura 23a e Figura 24a resultarão em conclusões semelhantes às já feitas e, ao se comparar ponto a ponto cada uma das medidas de cada função de ligação com as da MIX, observar-se-á que é indiferente escolher uma em detrimento da outra para se fazer previsões, ambas terão desempenhos iguais, conforme mostram as figuras Figura 21b, Figura 22b, Figura 23b e Figura 24b.

Figura 21 – Boxplots de especificidades da simulação de previsões

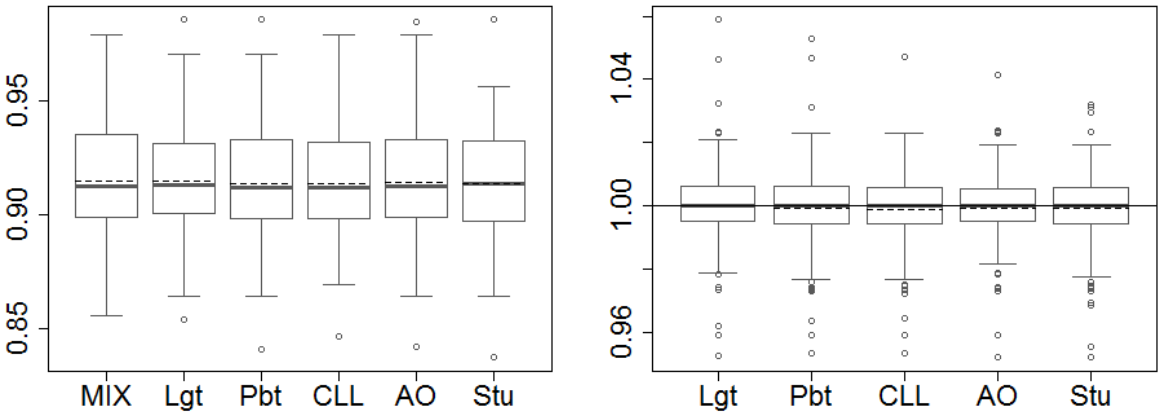


(a) Boxplot de especificidades

(b) Boxplot de especificidades relativas à MIX

Fonte: Elaborada pelo autor.

Figura 22 – Boxplots de VPPs da simulação de previsões

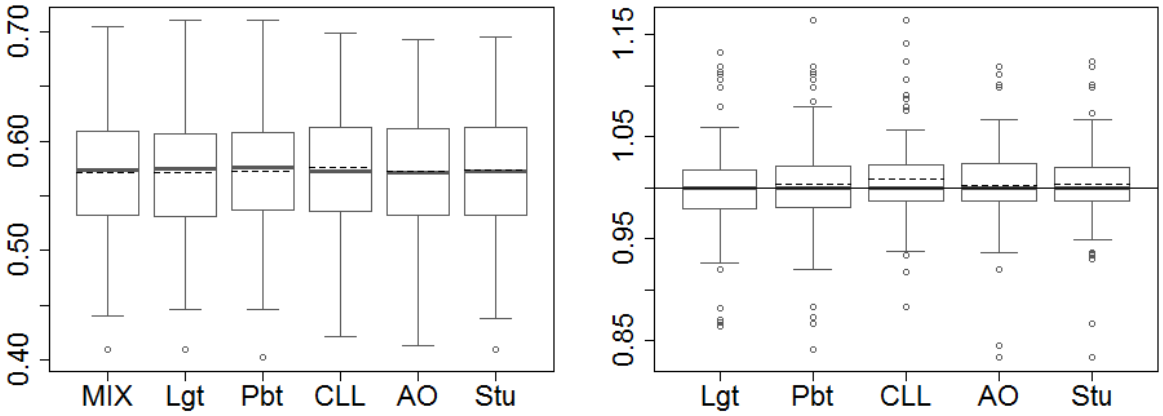


(a) Boxplot de VPPs

(b) Boxplot de VPPs relativos à MIX

Fonte: Elaborada pelo autor.

Figura 23 – Boxplots de VPNs da simulação de previsões

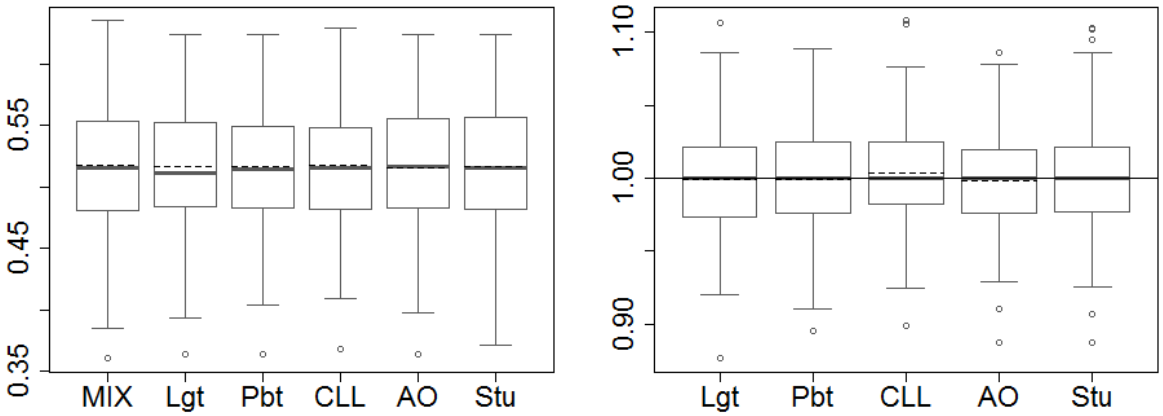


(a) Boxplot de VPNs

(b) Boxplot de VPNs relativos à MIX

Fonte: Elaborada pelo autor.

Figura 24 – Boxplots de coeficientes de correlação Matthews da simulação de previsões



(a) Boxplot de coeficientes de correlação de Matthews

(b) Boxplot de coeficientes de correlação de Matthews relativos à MIX

Fonte: Elaborada pelo autor.

CAPÍTULO

5

APLICAÇÕES A DADOS REAIS

Findados os estudos de simulações e apresentações teóricas sobre a ligação proposta, parte-se para as aplicações do mesmo a fim de se analisar seu comportamento quando os dados não são gerados computacionalmente.

Para tanto, três bancos de dados bastante difundidos na área de M.L.G. são modelados pelas seis funções de ligação e têm suas estatísticas comparadas. São eles: "Mortalidade de Besouros" (Bliss (1935)), "Garotas de Varsóvia" (Milicer e Szczotka (1966)) e "Frequência de Micronúcleos" (Balasem e Ali (1991))

Nas estimações feitas nesse Capítulo todas as estimativas foram significativas para o teste de Wald a um nível de significância de 5%.

5.1 Mortalidade de Besouros

O problema apresentado no artigo de 1935 consiste em analisar diferentes doses de dissulfeto de carbono (CS_2) e suas taxas de letalidade em besouros com o intuito de encontrar a mais eficiente. Durante 5 horas, 481 insetos foram divididos heterogeneamente em 8 grupos, sendo cada um exposto à sua respectiva dosagem. Ao fim houve 291 mortes.

O evento de interesse é a morte dos besouros e, portanto, o objetivo da análise é a estimação das proporções de óbitos para as diferentes doses de CS_2. Pelo fato da relação deles não ser bem ajustada em modelos simétricos, conforme explica o próprio autor, tais dados são interessantes para o uso da função de ligação MIX.

A Tabela 6 apresenta as estimativas dos parâmetros para as diversas funções de ligação e observa-se que a mistura proposta tem sua composição formada por 68% de FL complementar log-log e o restante de ligação Stukel, com os erro padrões (EPs) dos pesos bem próximos de zero.

Das estimativas de $\boldsymbol{\gamma}$ para a ligação MIX nota-se que o parâmetro da FL Stukel que modela a cauda à direita tem um valor muito elevado (159.235) e seu erro padrão "Não é um Número" (NéN). Isso ocorre pois a log-verossimilhança perfilada nesse ponto, embora seja ponto de máximo, é constante, conforme mostra a Figura 25. Portanto não há segunda derivada e, consequentemente, como calcular o seu erro padrão.

Ainda, a função de ligação Aranda-Ordaz convergiu para a Complementar Log-Log e as demais apresentaram resultados já vistos na literatura.

Os erros padrões do vetor $\boldsymbol{\beta}$ para as cinco funções de ligação que não a mistura apresentam erros padrões bem elevados, mas nada que interfira na significância dos parâmetros quando é realizado o teste de Wald.

Tabela 6 – Estimativas (e erros padrões) dos parâmetros de diversas funções de ligação para Mortalidade de Besouros

Função de Ligação	$\hat{\lambda}$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$
MIX	-	0.680	0	~ 0	159.235	-0.190	-37.433	20.798
	-	(0.143)	(~ 0)	(~ 0)	(NéN)	(0.579)	(0.067)	(0.037)
Logit	-	-	-	-	-	-	-60.717	34.270
	-	-	-	-	-	-	(5.181)	(2.912)
Probit	-	-	-	-	-	-	-34.935	19.728
	-	-	-	-	-	-	(2.648)	(1.487)
CLL	-	-	-	-	-	-	-39.572	22.041
	-	-	-	-	-	-	(3.240)	(1.799)
Aranda-Ordaz	~ 0	-	-	-	-	-	-39.572	22.041
	(0.074)	-	-	-	-	-	(3.240)	(1.799)
Stukel	-	-	-	-	0.164	-0.521	-69.534	39.025
	-	-	-	-	(0.092)	(0.219)	(7.123)	(3.956)

Pela Tabela 7 é possível perceber que as estimações mais próximas dos valores de proporções observados são obtidas utilizando-se as funções de ligação MIX e Stukel, o que é comprovado pela Figura 26a. Dela se verifica, ainda, que as distribuições de seus erros absolutos são mais homogêneas e menos assimétricas que das demais e que a FL MIX desvia em, no máximo, 3% do valor observado, ao passo que *logit* e *probit* chegam a mais de 10%.

A função de ligação complementar log-log (e, consequentemente a Aranda-Ordaz), tida no artigo original como a melhor para modelar tais dados, tem o valor de seu 3º Quartil de erros absolutos próximo ao do Máximo da Stukel e sua mediana maior que a dela, atualizando, portanto, as conclusões de Bliss (1935).

Por essa mesma linha de comparação o Máximo de erros absolutos da ligação MIX nem chega a atingir o 3º Quartil da Stukel, que perde seu posto de melhor modelo no que diz respeito as estimações.

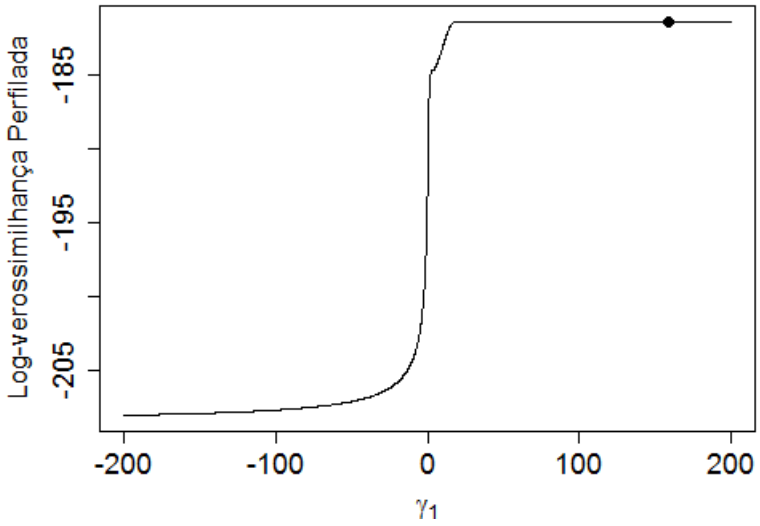


Figura 25 – Curva de γ_1 e sua log-verossimilhança perfilada para Mortalidade de Besouros

Tabela 7 – Valores observados e preditos de proporções de diversas funções de ligação para Mortalidade de Besouros

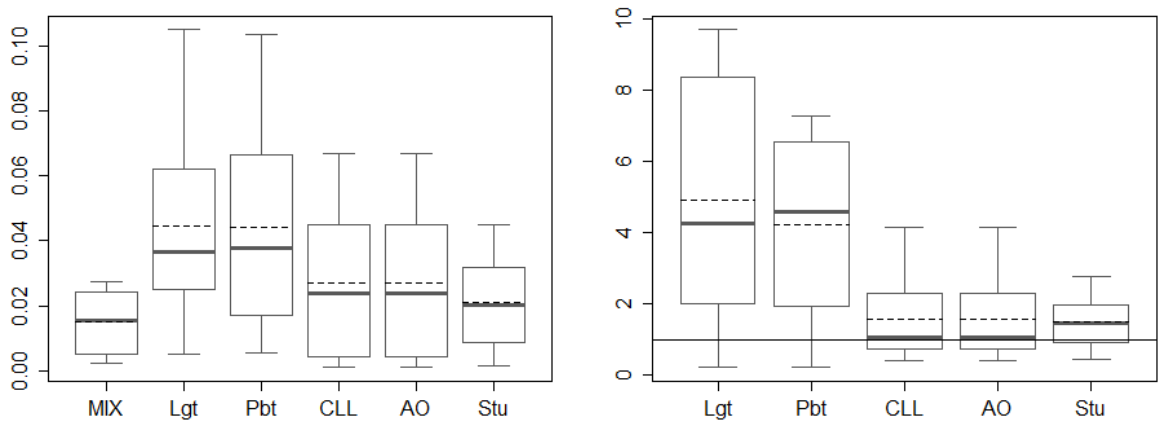
	log(Dose)							
Função de ligação	1.6907	1.7242	1.7552	1.7842	1.8113	1.8369	1.861	1.8839
Observado	0.102	0.217	0.290	0.500	0.825	0.898	0.984	1.000
MIX	0.109	0.192	0.318	0.485	0.809	0.922	0.981	0.998
Logit	0.059	0.164	0.362	0.605	0.795	0.903	0.955	0.979
Probit	0.057	0.179	0.379	0.604	0.788	0.903	0.962	0.987
CLL	0.095	0.188	0.338	0.542	0.758	0.918	0.986	0.999
Aranda-Ordaz	0.095	0.188	0.338	0.542	0.758	0.918	0.986	0.999
Stukel	0.118	0.184	0.304	0.524	0.780	0.930	0.982	0.996

Ao se comparar cada erro das demais funções de ligação com o da MIX, nota-se de suas distribuições, pela Figura 26b, que nenhuma mediana está abaixo de 1, indicando que pelo menos 50% das estimações feitas pela ligação proposta são mais próximas das proporções observadas que as estimações dos demais modelos, corroborando com as conclusões obtidas anteriormente.

Logit e *probit* são as FLs que apresentam, pelo menos para uma proporção, os menores erros relativos à MIX, assim como os maiores. Já as ligações Aranda-Ordaz e complementar log-log têm suas medianas muito próximas a 1, mas ainda acima deste valor, e seus erros são, no máximo, 4 vezes maiores que os da mistura. A função de ligação Stukel, por sua vez, apresenta os menores, mais homogêneos e simétricos erros relativos, embora aproximadamente 75% deles sejam maiores que os erros da função proposta.

Por fim, comparando-se as medidas da Tabela 8, tem-se que MIX é a função de ligação que atinge o maior valor de log-verossimilhança, satisfazendo a condição de escolha de melhor modelo para este critério, porém para AIC, a complementar log-log é tida como a melhor função

Figura 26 – Boxplot de medidas de erros para as proporções de mortes de besouros



(a) Erros absolutos

(b) Erros absolutos relativos à ligação MIX

Fonte: Elaborada pelo autor.

de ligação, seguida pela mistura.

Ainda, para o primeiro critério, os valores do modelo proposto e de Stukel são bem próximos, já para o segundo Aranda-Ordaz é quem se aproxima de MIX, o que já era esperado devido ao comportamento dos boxplots das Figura 26a e Figura 26b.

Embora para AIC o modelo proposto não seja o menor, pela análise feita há grandes evidências de que ele tenha um desempenho de estimação superior às demais funções.

Tabela 8 – Medidas de comparação de modelos para Mortalidade de Besouros

Função de Ligação	MIX	*Logit*	*Probit*	CLL	AO	Stukel
Log-verossimilhança	-181.293	-186.235	-185.679	-182.343	-182.343	-181.991
AIC	376.586	376.471	375.358	368.685	370.685	371.981

5.2 Garotas de Varsóvia

Em 1965, durante o meio de janeiro até o fim de março, 3918 mulheres varsovianas foram questionadas sobre as idades em que ocorreram suas menarcas. O objetivo do estudo era estimar a proporção de primeiras menstruações ocorridas nas diversas idades.

Como as entrevistadas poderiam responder, intencionalmente ou não, falsas datas, criaram-se intervalos de idades e neles se fizeram as contagens do evento de interesse. Para 669 delas a primeira menstruação ainda não havia chegado e o ponto médio de cada intervalo foi definido como o valor pontual da covariável em questão.

Das estatísticas apresentadas na Tabela 9 verifica-se que a mistura é composta de 32.3% de função de ligação *probit* (simétrica) e o restante de Stukel (assimétrica), com seus parâmetros

$\boldsymbol{\gamma}$ modeladores de caudas positivos. Para as demais funções de ligação o parâmetro β_0 apresenta erros padrões relativamente altos e Aranda-Ordaz e Stukel não convergem para nenhuma das outras funções de ligação.

Tabela 9 – Estimativas (e erros padrões) dos parâmetros de diversas funções de ligação para Garotas de Varsóvia

Função de Ligação	$\hat{\lambda}$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$
MIX	-	~ 0	~ 0	0.323	0.130	0.456	-14.684	1.135
	-	(~ 0)	(~ 0)	(0.117)	(0.046)	(0.082)	(0.043)	(0.003)
Logit	-	-	-	-	-	-	-21.226	1.632
	-	-	-	-	-	-	(0.770)	(0.059)
Probit	-	-	-	-	-	-	-11.819	0.908
	-	-	-	-	-	-	(0.387)	(0.030)
CLL	-	-	-	-	-	-	-12.985	0.953
	-	-	-	-	-	-	(0.426)	(0.031)
Aranda-Ordaz	1.452	-	-	-	-	-	-24.628	1.916
	(0.077)	-	-	-	-	-	(0.905)	(0.070)
Stukel	-	-	-	-	0.086	0.321	-16.999	1.315
	-	-	-	-	(0.035)	(0.054)	(0.551)	(0.043)

A Figura 27 apresenta as curvas estimadas e as proporções observadas para as seis funções de ligação. Dela é possível observar que o modelo estimado para a FL Stukel é o que mais se aproxima do para MIX, ao passo que o da complementar log-log é o que mais se distancia dela e dos pontos observados também. Da Figura 28 se verifica que o comportamento das distribuições de erros absolutos para as ligações MIX e Stukel são bem semelhantes, como esperado, apesar do fato dos 50% menores valores de erro da primeira serem mais homogêneos que o da segunda. Exceto pelos *outliers*, ambas erram no máximo 5%.

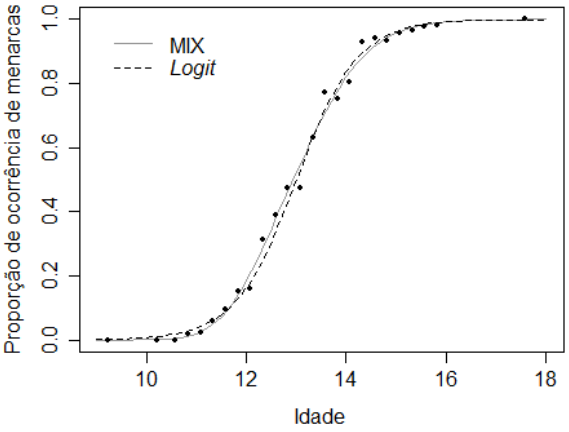
Logit, *probit* e Aranda-Ordaz são FLs que apresentam praticamente a mesma média de erros absolutos, porém os terceiros quartis das duas últimas são menores que o da primeira. Quartis esses que nem chegam a atingir a mediana de erros da complementar log-log (função de ligação com os piores resultados).

Para todas as ligações há pelo menos uma estimação cuja diferença para a proporção estimada se dá apenas a partir da terceira casa decimal, ou seja, o erro é praticamente zero.

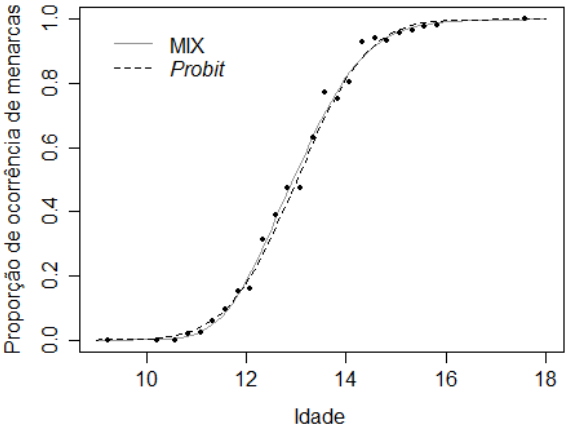
A função de ligação complementar log-log, de acordo com a Tabela 10, é a que apresenta o pior desempenho quando seus erros absolutos são comparados ao da MIX. Seu 1º Quartil já demonstra tal fato.

Exceto pela FL já citada e pela Stukel, que tem a distribuição da medida em questão mais homogênea e quase se sai melhor que a ligação proposta, as outras erram aproximadamente entre 2.5 e 4 vezes mais que a mistura nos seus 75% menores valores de erros absolutos relativos.

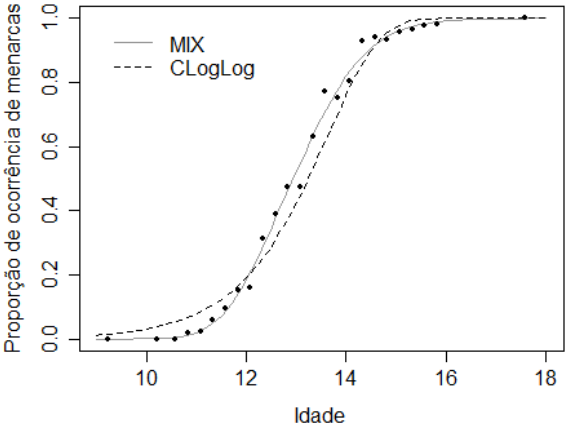
Figura 27 – Curvas estimadas e proporções de ocorrência de menarcas observadas



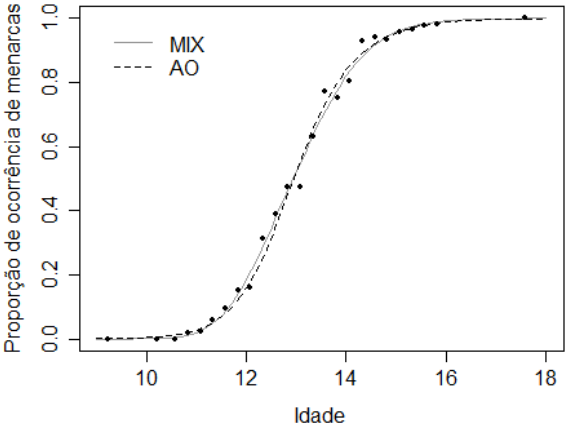
(a) Mix e *logit*



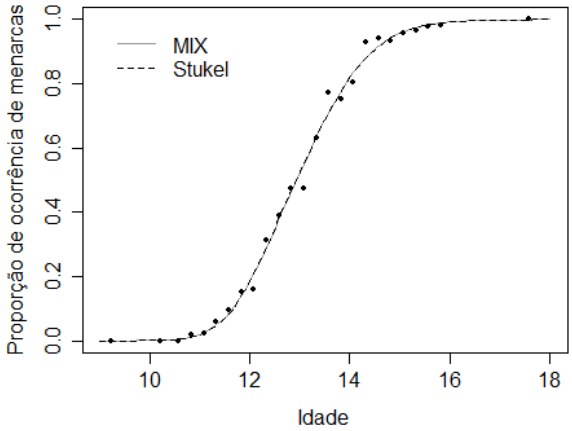
(b) Mix e *probit*



(c) Mix e CLL



(d) Mix e AO



(e) Mix e Stukel

Fonte: Elaborada pelo autor.

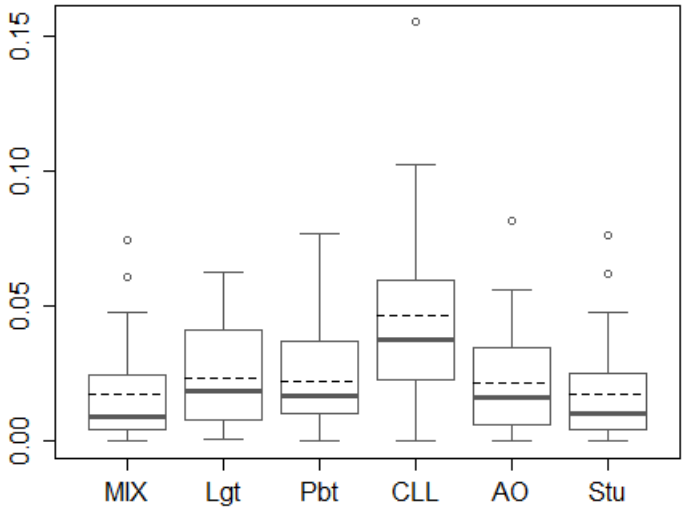


Figura 28 – Boxplot dos erros absolutos das estimações de proporções de ocorrências de menarca.

Tabela 10 – Medidas descritivas dos erros absolutos das estimações de proporções de ocorrências de menarca relativos à ligação MIX

Função de ligação	Mín.	1ºQ	Mediana	Média	3ºQ	Máx.
Logit	0.430	0.876	1.666	17.140	3.871	368.800
Probit	0.050	0.811	1.625	4.881	3.201	49.370
CLL	0.000	1.518	2.968	118.000	10.630	2679.000
Aranda-Ordaz	0.060	0.705	1.566	9.381	2.629	169.200
Stukel	0.540	0.938	1.020	0.988	1.061	1.285

Todas, contudo, têm medianas maiores que 1, indicando que MIX é a função de ligação que apresenta a maior quantidade de menores erros absolutos dentre as seis funções avaliadas para cada uma das diversas idades estudadas.

Tabela 11 – Medidas de comparação de modelos para Garotas de Varsóvia

Função de Ligação	MIX	*Logit*	*Probit*	CLL	AO	Stukel
Log-verossimilhança	-813.422	-819.652	-817.744	-865.711	-817.660	-813.594
AIC	1640.843	1643.305	1639.489	1735.422	1641.320	1635.188

Os valores de log-verossimilhança e de AIC da mistura, observados na Tabela 11, apresentam as mesmas conclusões e considerações que no exemplo anterior.

5.3 Frequência de Micronúcleos

De acordo com Flores e Yamaguchi (2009), micronúcleo é uma fragmentação de um cromossomo que gera um pequeno núcleo ao lado do original de uma célula. Decorre de uma falha durante a mitose e é usado como marcador genético para índices de toxicidade.

No estudo em questão, 5007 células de linfócitos foram divididas e expostas a 10 diferentes doses de radiação γ de Césio 137, medidas em Gy ("gray", unidade do Sistema Internacional de Unidades para dose absorvida), e se observou a frequência relativa de células com presenças de micronúcleos para cada uma das doses.

As estatísticas de estimação dos modelos estão na Tabela 12 e, relevando as diferenças nas terceiras casas decimais dos parâmetros estimados, a mistura é a própria função de ligação Stukel. A FL Aranda-Ordaz não converge para nenhum caso particular e todos os erros padrões são pequenos.

Embora as proporções estimadas pelas três funções de ligação anteriormente citadas sejam as que mais se aproximam do valor observado (de acordo com a Tabela 13), nota-se que MIX e Stukel erram mais homogeneamente que Aranda-Ordaz na Figura 29a. Apesar de ser uma função de ligação assimétrica, a complementar log-log é a que apresenta os maiores erros absolutos, ultrapassando até os 10% quase atingidos pela *logit*.

Na comparação de erros absolutos relativos à ligação MIX (Figura 29b) o boxplot da função de ligação Stukel está limitado às redondezas de 1 por motivos triviais (a ligação MIX é a própria Stukel, estão se dividindo valores iguais de erros), a ligação Aranda-Ordaz tem sua mediana acima dessa medida e as demais tem, aproximadamente, 75% dos seus erros relativos maiores que o da mistura.

Tabela 12 – Estimativas (e erros padrões) dos parâmetros de diversas funções de ligação para Frequência de Micronúcleos

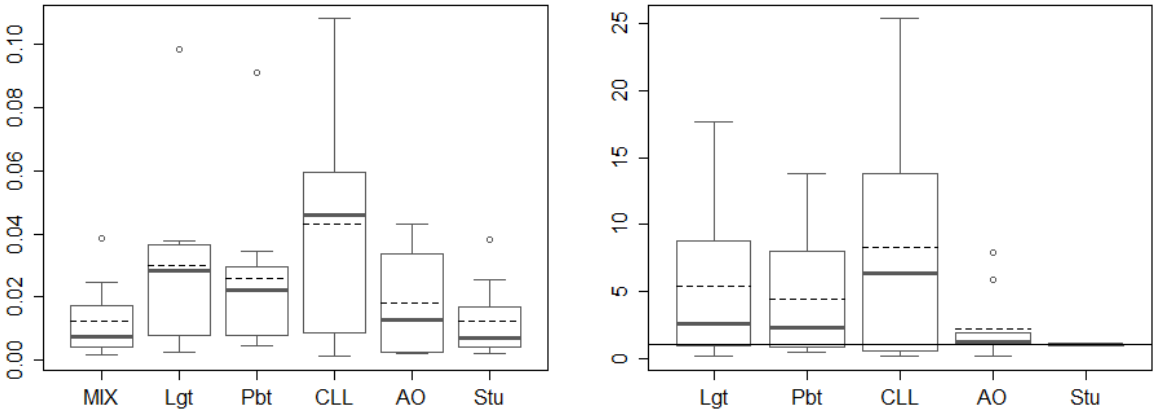
Função de Ligação	$\hat{\lambda}$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$
MIX	-	~ 0	~ 0	~ 0	0.970	1.291	-1.260	0.340
	-	(~ 0)	(~ 0)	(0.084)	(0.194)	(0.044)	(0.015)	(0.006)
Logit	-	-	-	-	-	-	-2.552	0.653
	-	-	-	-	-	-	(0.067)	(0.019)
Probit	-	-	-	-	-	-	-1.514	0.389
	-	-	-	-	-	-	(0.035)	(0.010)
CLL	-	-	-	-	-	-	-2.406	0.490
	-	-	-	-	-	-	(0.057)	(0.014)
Aranda-Ordaz	6.825	-	-	-	-	-	-2.972	1.643
	(0.237)	-	-	-	-	-	(0.107)	(0.052)
Stukel	-	-	-	-	0.934	1.266	-1.271	0.343
	-	-	-	-	(0.192)	(0.044)	(0.020)	(0.008)

Todas as análises feitas indicam que o modelo Stukel é realmente o que melhor estima as proporções de frequência de micronúcleos e, consequentemente, MIX também. Pelos valores de log-verossimilhança e AIC apresentados na Tabela 14 nota-se a coerência da conclusão com o valor de AIC: MIX convergiu pra Stukel e esta apresenta o menor AIC. Entretanto, pela quantidade maior de parâmetros, a FL proposta terá um AIC maior.

Tabela 13 – Valores observados e preditos de proporções de diversas funções de ligação para Frequência de Micronúcleos

	Dose (em Gy)									
Função de Ligação	0.05	0.1	0.25	0.5	1	2	3	4	5	6
Observado	0.038	0.046	0.058	0.100	0.138	0.322	0.392	0.521	0.652	0.759
MIX	0.044	0.048	0.060	0.084	0.146	0.297	0.431	0.527	0.635	0.763
Logit	0.074	0.077	0.084	0.097	0.130	0.223	0.356	0.515	0.671	0.797
Probit	0.068	0.070	0.078	0.094	0.130	0.231	0.364	0.516	0.666	0.793
Complementar Log-Log	0.088	0.090	0.097	0.109	0.137	0.214	0.325	0.473	0.649	0.819
Aranda-Ordaz	0.046	0.049	0.060	0.082	0.140	0.290	0.435	0.555	0.650	0.725
Stukel	0.044	0.048	0.060	0.084	0.146	0.296	0.430	0.527	0.635	0.763

Figura 29 – Boxplot de medidas de erros para as proporções de ocorrências de micronúcleos



(a) Erros absolutos

(b) Erros absolutos relativos à ligação MIX

Fonte: Elaborada pelo autor.

Tabela 14 – Medidas de comparação de modelos para Frequência de Micronúcleos

Função de Ligação	MIX	*Logit*	*Probit*	CLL	AO	Stukel
Log-verossimilhança	-2250.200	-2275.477	-2268.484	-2297.332	-2253.676	-2250.202
AIC	4514.400	4554.954	4540.969	4598.664	4513.352	4508.404

CAPÍTULO

6

CONSIDERAÇÕES FINAIS

Neste trabalho foi apresentado um modelo de regressão para dados binários com mistura de funções de ligação, seus estudos de desempenhos nas estimações de proporções e predições de observações em dados simulados e, por fim, sua aplicação a três conjuntos de dados reais. As principais considerações obtidas são apresentadas a seguir.

- No Capítulo 2 são introduzidas algumas notações importantes, é apresentado o contexto em que funções de ligação para modelos de regressão com respostas binárias são utilizados e, também, as características daquelas que foram utilizadas no estudo: *logit*, *Probit*, complementar log-log, Aranda-Ordaz e Stukel.

- No Capítulo 3 é introduzido a ligação proposta, denominado "mistura" ou "MIX", são apresentadas as interpretações de seus parâmetros e sua metodologia de estimação, que é dividida em duas etapas: a primeira para se aprimorar os chutes iniciais e a segunda para se refinar o valor das estimativas.

- No Capítulo 4 são expostos estudos de simulação para se comparar seis funções de ligação quanto aos seus desempenhos nas estimações de proporções e previsões de respostas.

  Para o primeiro estudo, a ligação MIX apresenta as melhores estimações de proporções independente do modelo gerador de dados, da quantidade de valores diferentes da covariável e dos seus respectivos números de observações dicotômicas (quando comparadas às proporções observadas). Quanto às métricas de qualidade das estimativas dos parâmetros, alguns resultados são extremamente altos ou baixos para $\boldsymbol{\gamma}$ e β_0 e discussões sobre não-convergência do algoritmo de estimação e pseudoproblemas de identificabilidade são realizadas.

  Já para o segundo estudo conclui-se que o poder preditivo das funções em questão são, em média, iguais. Assim, numa análise conjunta dos resultados, tal fato corrobora para a escolha da ligação MIX como a que apresenta o melhor desempenho na modelagem, uma

vez que não há diferença entre as funções nos resultados de previsões e para as estimações de proporções é a que se sai melhor.

- Por fim, nas análises de dados reais do Capítulo 5 se observa que o modelo proposto apresenta os melhores resultados de estimações de proporções quando comparado aos demais e seus valores de log-verossimilhança e AIC o apontam como a melhor escolha, exceto para o banco de dados de Mortalidade de Besouros. Porém, a análise gráfica contradiz esse resultado.

  Ainda, pode-se perceber a flexibilidade da função de ligação MIX no banco de dados de Frequência de Micronúcloes, em que, por não encontrar uma combinação que apresentasse melhores resultados de modelagem que a Stukel, associa peso um ao componente da mesma na mistura.

REFERÊNCIAS

ADAM, D. **Les réactions du consommateur devant le prix: contribution aux études de comportement**. [S.l.]: Sedes, 1958. v. 15. Citado na página 24.

AGRESTI, A.; FINLAY, B. **Statistical Methods for the Social Sciences**. 4. ed. [S.l.]: Pearson Prentice Hall, 2009. Citado na página 23.

AITCHISON, J.; BROWN, J. A. C. The lognormal distribution. Cambridge University Press, 1963. Citado na página 24.

ARANDA-ORDAZ, F. J. On two families of transformations to additivity for binary response data. **Biometrika**, Biometrika Trust, v. 68, n. 2, p. 357–363, 1981. Citado nas páginas 24 e 30.

BALASEM, A. N.; ALI, A. S. K. Establishment of dose-response relationships between doses of cs-137 γ-rays and frequencies of micronuclei in human peripheral blood lymphocytes. **Mutation Research/Genetic Toxicology**, Elsevier, v. 259, n. 2, p. 133–138, 1991. Citado na página 69.

BERKSON, J. Application of the logistic function to bio-assay. **Journal of the American Statistical Association**, Taylor & Francis, v. 39, n. 227, p. 357–365, 1944. Citado na página 24.

____. Why i prefer logits to probits. **Biometrics**, JSTOR, v. 7, n. 4, p. 327–339, 1951. Citado na página 24.

____. Minimum chi-square, not maximum likelihood! **The Annals of Statistics**, JSTOR, p. 457–487, 1980. Citado na página 24.

BLISS, C. I. The method of probits. **Science**, American Association for the Advancement of Science, v. 79, n. 2037, p. 38–39, 1934. Citado na página 23.

____. The calculation of the dosage-mortality curve. **Annals of Applied Biology**, Wiley Online Library, v. 22, n. 1, p. 134–167, 1935. Citado nas páginas 23, 69 e 70.

BOX, G. E.; COX, D. R. An analysis of transformations. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 211–252, 1964. Citado nas páginas 24 e 27.

CARON, R. **Regressao de dados binários: Distribuiçao Weibull**. Dissertação (Tese de Mestrado) — Departamento de Estatística, UFSCar, 2010. Citado na página 24.

COX, D. R.; SNELL, E. J. **Analysis of binary data**. [S.l.]: CRC Press, 1989. v. 32. Citado na página 24.

CRAMER, J. S. **Logit models from economics and other fields**. [S.l.]: Cambridge University Press, 2003. Citado na página 23.

DINIZ, M. A. **Modelos Bayesianos semiparamétricos para dados binários**. Tese (Tese de Doutorado) — Instituto de Matemática e Estatística, IME- USP, Agosto 2015. Citado na página 23.

FARRELL, M. J. The demand for motor-cars in the united states. **Journal of the Royal Statistical Society. Series A (General)**, JSTOR, v. 117, n. 2, p. 171–201, 1954. Citado na página 24.

FISHER, R. A. Statistical methods and scientific inference. Hafner Publishing Co., 1956. Citado na página 36.

FLORES, M.; YAMAGUCHI, M. U. Teste do micronúcleo: Uma triagem para avaliação genotóxica. **Saúde e Pesquisa**, v. 1, n. 3, p. 337–340, 2009. Citado na página 75.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. **The elements of statistical learning**. [S.l.]: Springer series in statistics Springer, Berlin, 2001. v. 1. 222 p. Citado na página 62.

FROUD, R.; ABEL, G. Using roc curves to choose minimally important change thresholds when sensitivity and specificity are valued equally: The forgotten lesson of pythagoras. theoretical considerations and an example application of change in health status. **PloS one**, Public Library of Science, v. 9, n. 12, p. e114468, 2014. Citado na página 62.

GONÇALVES, E.; GOUVÊA, M.; MANTOVANI, D. Análise de risco de crédito com o uso de regressão logística. **Revista Contemporânea de Contabilidade**, v. 10, n. 20, p. 139–160, 2013. ISSN 2175-8069. Citado na página 25.

GUERRERO, V. M.; JOHNSON, R. A. Use of the box-cox transformation with binary response models. **Biometrika**, [Oxford University Press, Biometrika Trust], v. 69, n. 2, p. 309–314, 1982. ISSN 00063444. Disponível em: <http://www.jstor.org/stable/2335404>. Citado na página 24.

KING, G.; ZENG, L. Logistic regression in rare events data. **Political analysis**, SPM-PMSAPSA, v. 9, n. 2, p. 137–163, 2001. Citado na página 24.

KUTNER, M. H.; NACHTSHEIM, C.; NETER, J. *et al.* Applied linear regression models. McGraw-Hill New York, 2004. Citado na página 28.

LEHMANN, E. L.; CASELLA, G. **Theory of point estimation**. [S.l.]: Springer, 1998. v. 31. Citado na página 41.

LUCAMBIO, F. Introdução à teoria estatística clássica e moderna. Notas de Aula, Departamento de Estatística - Universidade Federal do Paraná. 2009. Disponível em: <https://docs.ufpr.br/~lucambio/CE210/veross.pdf>. Citado na página 36.

MALTHUS, T. R. **An Essay on the Principle of Population Or a View of Its Past and Present Effects on Human Happiness, an Inquiry Into Our Prospects Respecting the Future Removal Or Mitigation of the Evils which it Occasions by Rev. TR Malthus**. [S.l.]: Reeves and Turner, 1872. Citado na página 23.

MANLY, B. Exponential data transformations. **The Statistician**, JSTOR, p. 37–42, 1976. Citado na página 24.

MATTHEWS, B. W. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. **Biochimica et Biophysica Acta (BBA)-Protein Structure**, Elsevier, v. 405, n. 2, p. 442–451, 1975. Citado na página 63.

MILICER, H.; SZCZOTKA, F. Age at menarche in warsaw girls in 1965. **Human Biology**, JSTOR, p. 199–203, 1966. Citado na página 69.

NELDER, J. A.; WEDDEBURN, R. W. M. Generalized linear models. **Journal of The Royal Statistical Society**, v. 135, n. 3, p. 370–384, 1972. Citado nas páginas 24 e 27.

PRENTICE, R. L. A generalization of the probit and logit methods for dose response curves. **Biometrics**, JSTOR, p. 761–768, 1976. Citado na página 24.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2013. Disponível em: <http://www.R-.org/>. Citado na página 35.

SANTOS, B. P. dos. **Implementação em R de modelos de regressão bináriacom ligação paramétrica**. Dissertação (Tese de Mestrado) — Instituto de Matemática e Estatística, USP, Março 2013. Citado na página 39.

SPROTT, D. A. **Statistical inference in science**. [S.l.]: Springer Science & Business Media, 2008. Citado na página 36.

STUKEL, T. A. Generalized logistic models. **Journal of the American Statistical Association**, Taylor & Francis, v. 83, n. 402, p. 426–431, 1988. Citado nas páginas 24 e 31.

VERHULST, P. F. Notice sur la loi que la population suit dans son accroissement. correspondance mathématique et physique publiée par a. **Quetelet**, v. 10, p. 113–121, 1838. Citado na página 23.

____. Recherches mathématiques sur la loi d'accroissement de la population. **Nouveaux mémoires de l'académie royale des sciences et belles-lettres de Bruxelles**, v. 18, p. 14–54, 1845. Citado na página 23.

____. Deuxième mémoire sur la loi d'accroissement de la population. **Mémoires de l'académie royale des sciences, des lettres et des beaux-arts de Belgique**, v. 20, p. 1–32, 1847. Citado na página 23.

YULE, G. U. The growth of population and the factors which control it. **Journal of the Royal Statistical Society**, [Wiley, Royal Statistical Society], v. 88, n. 1, p. 1–58, 1925. ISSN 09528385. Disponível em: <http://www.jstor.org/stable/2341575>. Citado na página 23.