UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modelos de resposta limitada para regressão e teoria de resposta ao item

Patrícia Stülp

Tese de Doutorado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Patrícia Stülp

Modelos de resposta limitada para regressão e teoria de resposta ao item

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

Área de Concentração: Estatística

Orientador: Prof. Dr. Jorge Luis Bazán Guzmán

USP – São Carlos
Outubro de 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Stülp, Patrícia
S929m Modelos de resposta limitada para regressão e teoria de resposta ao item / Patrícia Stülp; orientador Jorge Luis Bazán. -- São Carlos, 2024.
104 p.

Tese (Doutorado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Resposta limitada. 2. Proporção de tempo de resposta. 3. Modelo Hierárquico. 4. Modelo quantílico. 5. Estimação Bayesiana. I. Bazán, Jorge Luis , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Patrícia Stülp

Bounded response models for regression and item response theory

Thesis submitted to the Institute of Mathematics and Computer Science – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Doctor in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Jorge Luis Bazán Guzmán

USP – São Carlos
October 2024

UNIVERSIDADE FEDERAL DE SÃO CARLOS

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

Folha de Aprovação

Defesa de Tese de Doutorado do candidato Patrícia Stülp, realizada em 16/08/2024.

Comissão Julgadora:

Prof. Dr. Jorge Luis Bazán Guzmán (USP)

Prof. Dr. Luis Hilmar Valdivieso Serrano (PUC-Perú)

Profa. Dra. Larissa Avila Matos (UNICAMP)

Profa. Dra. Rosineide Fernando da Paz (UFC)

Prof. Dr. Caio Lucidius Naberezny Azevedo (UNICAMP)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

AGRADECIMENTOS

O presente trabalho é fruto de um processo que exigiu alguns sacrifícios, muita força de vontade e essencialmente perseverança. A sua concretização só foi possível com a ajuda de algumas pessoas e Instituições a quem desejo apresentar os meus mais sinceros agradecimentos.

Agradeço, primeiramente, a Deus, pelo dom da vida e por todas as bênçãos recebidas.

Aos meus pais, Marcia e Marino, ao meu irmão Cristiano e sua família, por se fazerem presentes na minha vida mesmo de longe, por todas as preocupações, amor e dedicação que tiveram comigo ao longo dessa trajetória. Quero um dia retribuir tudo o que já fizeram por mim.

Ao meu noivo, Roberto, pelo apoio, pela paciência, por sempre ter me encorajado a buscar a excelência e a superar meus próprios limites, por ser meu porto seguro durante todo o processo de elaboração da Tese e pelo amor demonstrado que, sem dúvidas, tornou o caminho muito mais leve.

Ao meu orientador, Prof. Dr. Jorge Bazán, por todo o conhecimento transmitido, por todas as instruções, dedicação, compreensão, paciência e pela pessoa humana e incentivadora que mostrou ser. Os meus votos de felicidades e bênçãos de Deus em sua vida.

Aos membros da banca examinadora, pelas valiosas contribuições, críticas construtivas e sugestões, que enriqueceram este trabalho. Em especial, meu agradecimento ao Prof. Dr. Luis Hilmar Valdivieso Serrano, cujo apoio como coautor do artigo foi essencial para o desenvolvimento desta pesquisa.

Aos professores do PIPGEs, por todos os ensinamentos, e à equipe técnica da USP e UFSCar, pela disponibilidade e serviços prestados.

As minhas companheiras, Jéssica e Naiara, e demais colegas do PIPGEs, agradeço pelos momentos compartilhados durante o período do doutorado, pelo incentivo, pelas alegrias e sobretudo agradeço a amizade estabelecida. Vocês são muito especiais.

Aos meus amigos, os de longe e os de perto, e àqueles que de alguma forma fazem parte da minha vida e que são essenciais para eu ser, a cada dia dessa longa jornada, uma pessoa melhor.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES).

RESUMO

STÜLP, P. **Modelos de resposta limitada para regressão e teoria de resposta ao item**. 2024. 104 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Alguns modelos de regressão de resposta limitada foram propostos recentemente na literatura para modelar taxas e proporções. Diante disso, estudamos inicialmente os modelos de regressão Beta, Simplex e L-logistic, considerando uma estimação Bayesiana dos parâmetros dos três modelos usando o algoritmo *No-U-Turn sampler* (NUTS), a partir da implementação via o pacote Stan. Um estudo comparativo de recuperação de parâmetros é implementado e um estudo de desempenho de diferentes pacotes foi desenvolvido. Finalmente foram feitas aplicações, a primeira com dados gerados e outras duas com dados reais acerca da pobreza no Peru e nos municípios do Brasil, respectivamente, e foram apresentadas comparando o desempenho dos modelos para a estimação de amostras pequenas e grandes. Além disso, propomos um novo modelo de regressão quantílica com distribuição de respostas limitadas que generaliza a distribuição L-logística. Seguindo uma abordagem bayesiana, são realizados critérios de comparação de modelos de estimação e análise de resíduos, bem como um estudo de simulação para sensibilidade da *priori* e recuperação de parâmetros. Uma aplicação da nova distribuição para modelar a vulnerabilidade à pobreza no Brasil e uma análise de regressão com dados de pobreza do Peru estão incluídas. Também são realizadas comparações com as distribuições L-Logistic e G-Logistic mostrando a grande flexibilidade do novo modelo. Outro exemplo de modelagem para respostas limitadas é o tempo em que um examinado leva para realizar um teste computadorizado, por exemplo. Neste sentido, também propomos um modelo no contexto bayesiano hierárquico, a fim de modelar o tempo de resposta (TR) seguindo uma distribuição limitada e obter um modelo conjunto em combinação com a precisão da resposta, considerando modelo Bernoulli, e para a precisão consideramos o modelo de Ogiva Normal, da Teoria de Resposta ao Item (TRI). Para modelar a proporção de tempos de resposta (PTR) propomos, inicialmente, o uso da distribuição Beta e realizamos uma abordagem Bayesiana para a estimação dos parâmetros a partir do pacote R2jags. Um rápido estudo de simulação de recuperação dos parâmetros foi desenvolvido e o modelo foi aplicado a um conjunto de dados reais de leitura computadorizada do PISA 2015.

Palavras-chave: Resposta limitada, Proporção de tempo de resposta, Modelo Hierárquico, Modelo quantílico, Estimação Bayesiana.

ABSTRACT

STÜLP, P. **Bounded response models for regression and item response theory**. 2024. 104 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Some bounded response regression models have been recently proposed in the literature to model rates and proportions. In this way, we initially studied the Beta, Simplex and L-logistic regression models considering a Bayesian estimation of the parameters of the three models using the No-U-Turn sampler (NUTS) algorithm from the implementation via the Stan package. A comparative parameter recovery study is implemented and a performance study of different packages is carried out. Finally, applications were carried out, the first with generated data and the other two with real data about poverty in Peru and in the municipalities of Brazil, respectively, and were presented comparing the performance of the models for the estimation of small and large samples. Furthermore, we propose a new quantile regression model with a bounded response distribution that generalizes the L-logistic distribution. Following a bayesian approach, estimation model comparison criteria and residual analysis are performed as well as a simulation study for prior sensitivity and parameter recovery. An application of the new distribution to model poverty vulnerability in Brazil and a regression analysis with poverty data from Peru is included. Comparison with the L-Logistic and G-Logistic distributions are also performed showing the great flexibility of the new model. Another example of modeling for bounded responses is the time it takes an examinee to perform a computerized test, for example. In this sense, we also propose a model in the hierarchical bayesian context, in order to model the response time (RT) following a bounded distribution and obtain a joint model in combination with the accuracy of the response by considering Bernoulli model and for the precision we consider the Normal Ogive model, from the Item Response Theory (IRT). To model the RT proportion, we initially propose the use of the Beta distribution and perform a Bayesian approach for parameter estimation using the R2jags package. A brief parameter recovery simulation study was developed and the model was applied to a real PISA 2015 computer readout dataset.

Keywords: Bounded response, Proportion of time response, Hierarchical model, Quantile model, Bayesian estimation.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

OK	Octile kurtosis
OS	Octile skewness
AR	Acurácia de Resposta
DIC	*Deviance Information Criterion*
EAIC	*Expected Akaike Information Criterion*
EBIC	*Expected Schwarz-Bayesian Information Criterion*
fda	função de distribuição acumulada
fdp	função densidade de probabilidade
GMRF	*Gaussian Markov Random Fields*
HQIC	*Hannan–Quinn Information Criterion*
IC	*Bayesian Predictive Information Criterion*
INLA	*Integrated Nested Laplace Approximation*
LOO	*Leave-One-Out Cross-Validation*
MCMC	*Markov Chain Monte Carlo*
NUTS	*No-U-Turn*
PCVP	Proporção de Crianças Vulneráveis à Pobreza
PISA	Programa de Avaliação Internacional de Estudantes
PTR	proporção do tempo de resposta
TG	Gamma Truncda
TN	Normal Truncada
TR	Tempo de Resposta
v.a.	variável aleatória
VGAM	*Vector Generalized Linear and Additive Models*
VGLM	*Vector Generalized Linear Model*
WAIC	*Watanabe–Akaike Information Criterion*

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Modelos com resposta limitada no intervalo $(0,1)$ têm sido estudados nos últimos tempos devido à sua ampla aplicabilidade. São muito comuns para modelar taxas, percentuais, proporções, entre outros. Para modelar este tipo de dados não é recomendado, por exemplo, utilizar regressão linear, pois seria necessária uma transformação na variável resposta para que ela pertencesse ao intervalo desejado. Segundo Ferrari e Cribari-Neto (2004), a interpretação dos parâmetros não poderia ser feita com base nos dados originais, mas sim com base na transformação, e as medidas de proporção seriam assimétricas, o que violaria a suposição de normalidade.

Neste contexto, alguns modelos são os propostos por Barndorff-Nielsen e Jørgensen (1991), Paz, Bazán e Milan (2017), Bayes, Bazán e Castro (2017), Paolino (2001) e Ferrari e Cribari-Neto (2004). No trabalho de Ferrari e Cribari-Neto (2004), por exemplo, os autores propõem um modelo de regressão cuja variável resposta segue uma distribuição Beta com parâmetro de média μ e parâmetro de precisão ϕ. O modelo relaciona esses parâmetros com covariáveis por meio de uma estrutura de regressão com função de ligação logito. Outros autores também estudaram o modelo de regressão Beta, como Kieschnick e McCullough (2003), Ospina, Cribari-Neto e Vasconcellos (2006) e Smithson e Verkuilen (2006).

Outro exemplo de modelo de regressão com resposta limitada é o modelo proposto por Barndorff-Nielsen e Jørgensen (1991), cuja variável resposta é a distribuição Simplex. Assim como no modelo Beta, López (2013) adotou uma estrutura de dependência através da função logit, para relacionar covariáveis com a média de resposta.

Ademais, Paz *et al.* (2019) apresenta um modelo de regressão no intervalo $(0,1)$, utilizando distribuição L-Logística para a variável resposta, que foi originalmente desenvolvido por Tadikamalla e Johnson (1982). O modelo proposto por Paz *et al.* (2019) considera a mediana e a dispersão como parâmetros, e a mediana da variável resposta está relacionada a um conjunto de covariáveis através da função logito. Os modelos Beta e Simplex são modelos consolidados e o

L-Logistic é um modelo novo e, por isso, estudos comparando-os não estão disponíveis.

Alguns modelos de regressão com resposta limitada são apresentados dentro do contexto de regressão quantílica como, por exemplo, o trabalho de Korkmaz, Chesneau e Korkmaz (2021), que apresenta uma proposta de modelo de regressão quantílica em termos de qualquer quantil, para variáveis resposta limitadas, através da parametrização da distribuição Normal baseada na função secante hiperbólica. Mazucheli *et al.* (2023) também derivou e aplicou um modelo de regressão quantílica na distribuição semi-normal generalizada unitária (KORKMAZ, 2020), como uma alternativa aos modelos de regressão quantílica existentes.

Também Mazucheli *et al.* (2022) realizou uma revisão de um grande conjunto de modelos paramétricos de regressão quantílica, que incluem distribuições de respostas limitadas relacionadas a Gauss no intervalo de unidades, respostas limitadas não relacionadas a Gauss neste intervalo e também distribuições para respostas positivas contínuas e o caso de uma resposta discreta. Por fim, eles desenvolveram um pacote R que permite ajustar uma variedade de distribuições e aplicaram os modelos a dois conjuntos de dados reais, biomédicos do Brasil e da COVID-19.

Já no contexto de Teoria de Resposta ao Item (TRI), um exemplo de resposta com suporte limitado é o tempo que um examinado leva para responder aos itens de uma prova. Alguns trabalhos desenvolvidos na literatura consideram o tempo no suporte $(0, \infty)$, passando a ideia de que o tempo para um examinado responder aos itens de uma prova é infinito. Por outro lado, alguns trabalhos foram desenvolvidos pensando no tempo como intervalo. Flores *et al.* (2019), por exemplo, propõe um modelo no contexto hierárquico considerando a distribuição Simplex para a proporção do tempo de resposta e um modelo TRI probit de dois parâmetros para a precisão de resposta (resposta correta ou incorreta).

1.1 Objetivos e Organização da Tese

O objetivo geral deste trabalho é apresentar, propor e detalhar modelos de regressão com resposta limitada no intervalo $(0, 1)$, dentro da perspectiva Bayesiana.

Consideramos que estimar um modelo utilizando métodos bayesianos pode ser vantajoso por vários motivos: a) possibilidade de incorporar conhecimento prévio sobre os parâmetros do modelo especificando uma distribuição a *priori* para eles considerando seu espaço paramétrico correspondente, onde informações a *priori* podem ajudar a estabilizar as estimativas dos parâmetros e melhorar a inferência, b) em vez de fornecer estimativas pontuais, a análise Bayesiana produz distribuições *posterioris* para os parâmetros, que refletem a incerteza em seus valores dados os dados e as informações a *prioris*, c) oferecer flexibilidade na especificação do modelo, permitindo o uso de modelos complexos com estruturas hierárquicas, relações não lineares e vários tipos de distribuições.

Como objetivos específicos, podemos citar:

- comparar o desempenho de diferentes algoritmos para estimar modelos de regressão de resposta limitada, bem como avaliar os modelos de regressão Beta, Simplex e L-Logistic em termos da capacidade de recuperação de parâmetros sob diferentes condições experimentais de tamanho de amostra através de um estudo de simulação;
- propor um novo modelo de regressão quantílico que generaliza o modelo de regressão L-Logistic, modelando diferentes quantis e adicionando mais flexibilidade através de um parâmetro adicional;
- apresentar uma outra proposta de modelo dentro do contexto hierárquico na Teoria de Resposta ao Item, na qual a precisão de resposta dos itens é modelada usando um modelo TRI probit de dois parâmetros e a proporção de tempo de resposta, que está compreendida no intervalo $(0, 1)$, é modelada usando a distribuição Beta, com parâmetros de posição e precisão.

Portanto, a tese está organizada da seguinte maneira:

No capítulo 2 deste trabalho apresentamos um estudo referente aos modelos de regressão Beta, Simplex e L-Logistic, que já estão bem consolidados na literatura. A partir desses modelos para resposta limitada, realizamos um estudo de simulação para alguns cenários, com o objetivo de verificar quanto à recuperação dos parâmetros e avaliar o desempenho desses modelos sob uma abordagem bayesiana, e também realizamos aplicações. A primeira aplicação realizada em um conjunto de dados gerados, a segunda em um conjunto de dados com poucas observações e a terceira realizada em um conjunto de dados com muitas observações.

O capítulo 3 apresenta uma nova proposta de modelo de regressão quantílico para variáveis respostas limitadas no intervalo $(0, 1)$, que generaliza o modelo de regressão L-Logistic. Apresentamos algumas propriedades desta nova proposta, assim como uma reparametrização em termos do quantil. Desenvolvemos um estudo de simulação de sensibilidade da *priori* e também um estudo de recuperação dos parâmetros, considerando diferentes cenários. Em termos de aplicação, consideramos uma somente utilizando a distribuiução, ou seja, sem incluir covariáveis, e outra considerando a inclusão de covariáveis de quatro maneiras diferentes.

O capítulo 4 é dedicado à apresentação do modelo conjunto Beta-Bernoulli no contexto hierárquico, que leva em consideração a proporção de tempo que um examinado responde os itens de uma prova e a precisão de resposta. Modelamos a proporção do tempo de resposta a partir de uma distribuição Beta com parâmetros de média e precisão e, separadamente, modelamos a precisão de resposta usando modelo TRI probit de dois parâmetros. Dentro do contexto hierárquico bayesiano, esse passo condiz ao primeiro nível da modelagem. No segundo nível, consideramos uma modelagem conjunta dos parâmetros de velocidade e precisão. Este trabalho foi desenvolvido com abordagem bayesiana, realizamos um pequeno estudo de simulação e

uma aplicação do modelo hierárquico Beta-Bernoulli a conjunto de dados reais de leitura computadorizada do PISA 2015.

CAPÍTULO

2

ESTIMAÇÃO BAYESIANA PARA MODELOS DE REGRESSÃO

Uma etapa importante em qualquer tipo de modelagem estatística é a estimativa dos parâmetros. Existem diversos pacotes disponíveis na biblioteca do *software* R que facilitam esse processo, como os pacotes `gamlss` (RIGBY; STASINOPOULOS, 2005; STASINOPOULOS; RIGBY *et al.*, 2007), `INLA` (RUE; MARTINO; CHOPIN, 2009), `VGAM` (YEE *et al.*, 2010), bem como `betareg` (CRIBARI-NETO; ZEILEIS, 2010), usado para estimar os parâmetros de distribuição Beta, `simplexerg` (ZHANG; QIU; SHI, 2016), usado para estimar os parâmetros de distribuição Simplex e `llbayesireg` (PAZ *et al.*, 2019), usado para estimar os parâmetros da distribuição L-Logístic, considerando a abordagem freqüentista e bayesiana.

Além dos pacotes disponíveis em R, é possível implementar os modelos em alguma linguagem de programação e estimar os parâmetros nesta programação.

Neste trabalho, realizamos o processo de estimação de parâmetros dos modelos de regressão Beta (FERRARI; CRIBARI-NETO, 2004), Simplex (BARNDORFF-NIELSEN; JØRGENSEN, 1991) e L-Logistic (PAZ *et al.*, 2019) utilizando pacotes disponíveis, além de implementar os modelos em Stan, que utiliza uma linguagem adaptada para programação em R, cuja estimação de parâmetros é realizada utilizando o pacote `rstan` (Stan Development Team, 2020), sob abordagem bayesiana. Assim, um dos objetivos deste trabalho é comparar o desempenho desses diferentes algoritmos para estimar modelos de regressão de resposta limitada, bem como avaliar os três modelos em termos de capacidade de recuperação de parâmetros sob diferentes condições experimentais de tamanho de amostra através de um estudo de simulação.

Este capítulo está estruturado da seguinte forma: na seção 2.1 apresentamos os modelos de regressão Beta, Simplex e L-Logístic, nesta ordem, para variável de resposta limitada. A seção 2.2 é dedicada à estimação Bayesiana dos parâmetros de distribuição, à apresentação dos algoritmos para comparação das estimativas em cada um dos três modelos e à apresentação

dos critérios de comparação. Na seção 2.3 realizamos um estudo de simulação para verificar a recuperação dos parâmetros em cada um dos três modelos. A seção 2.4 é dedicada à aplicação dos modelos de regressão. Finalmente na seção 2.5 apresentamos algumas conclusões sobre este trabalho.

2.1 Modelo de regressão para resposta limitada em parâmetros de localização

Nesta seção apresentamos os seguintes Modelos de Regressão limitada considerando covariáveis apenas no parâmetro localização: modelo de regressão Beta, modelo de regressão Simplex e modelo de regressão L-Logístic.

2.1.1 O modelo de regressão Beta

Na análise de regressão é mais comum modelar a média da resposta e também definir no modelo um parâmetro de precisão, ou dispersão (FERRARI; CRIBARI-NETO, 2004). Seja Y uma variável aleatória proveniente de uma distribuição Beta, cuja função densidade de probabilidade é dada por

$$f(y \mid \mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \tag{2.1}$$

com $0<\mu<1$ e $\phi>0$, $\mu = a/(a+b)$ é o parâmetro de posição (média) e $\phi = a+b$ o parâmetro de precisão. A média e a variância de Y são dadas, respectivamente, por $\mathbb{E}(Y \mid \mu,\phi) = \mu$ e $\mathbb{V}\text{ar}(Y \mid \mu,\phi) = \mu(1-\mu)/(1+\phi)$.

A Figura 1 mostra as diferentes densidades da distribuição Beta (2.1) para os diferentes valores de μ e ϕ. Nos dois gráficos apresentados variamos os valores do parâmetro de posição $\mu = (0,05, 0,25, 0,50, 0,75, 0,95)$, fixando $\phi = 5$ no primeiro gráfico e fixando $\phi = 100$ no segundo gráfico.

Na análise de regressão da distribuição Beta adotamos uma função de ligação $g_1(\cdot)$ para o parâmetro médio, estritamente monótono e duas vezes diferenciável, que está no intervalo aberto $(0,1)$. Segundo Ferrari e Cribari-Neto (2004), poderiam ser utilizadas as funções logito, probito, log-log, entre outras. Neste trabalho usamos a função de ligação logito que é escrita da seguinte forma

$$\begin{aligned} g_1(\mu_i) &= log\left(\frac{\mu_i}{1-\mu_i}\right) = X_i^T\boldsymbol{\beta} \\ \Rightarrow \quad \mu_i &= \frac{\exp(X_i^T\boldsymbol{\beta})}{1+\exp(X_i^T\boldsymbol{\beta})}, \end{aligned} \tag{2.2}$$

onde $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$, $i = 1, \cdots, n$, é um conjunto de covariáveis p e $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)$ é um vetor de parâmetros desconhecidos.

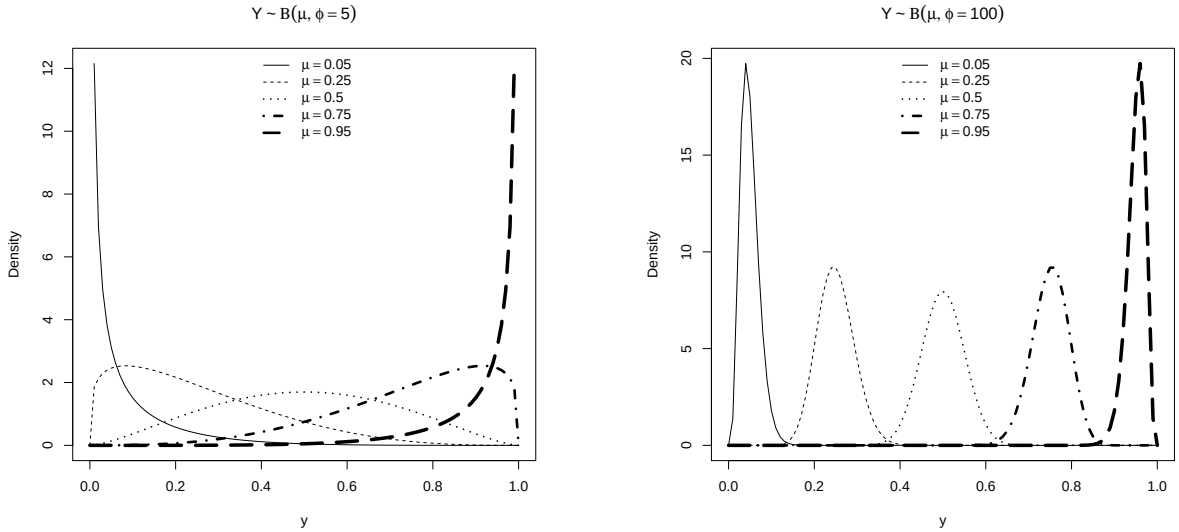


Figura 1 – Distribuição de Y para diferentes combinações dos parâmetros μ e ϕ no modelo de regressão Beta

2.1.2 O modelo de regressão Simplex

Seja Y uma variável aleatória de uma distribuição Simplex, cuja função densidade de probabilidade é dada por

$$f(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2(y(1-y))^3}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2 y(1-y)\mu^2(1-\mu)^2}\right), \tag{2.3}$$

com $0 < y < 1$, $0 < \mu < 1$, $\sigma^2 > 0$, μ o parâmetro de posição e σ^2 o parâmetro de dispersão. A média e a variância de Y são iguais a $\mathbb{E}(Y \mid \mu, \sigma^2) = \mu$ e $\mathbb{V}\text{ar}(Y \mid \mu, \sigma^2) = \mu(1-\mu) - 1/(\sqrt{2\sigma^2})\exp\left(1/(\sigma^2\mu^2(1-\mu)^2)\right)\Gamma\left(1/2; 1/(2\sigma^2\mu^2(1-\mu)^2)\right)$.

A Figura 2 mostra as diferentes densidades da distribuição Simplex (2.3) para diferentes valores de μ e σ^2. Observe que nos dois gráficos variamos os valores do parâmetro de posição $\mu = (0,1; 0,25; 0,50; 0,75)$ e fixamos o valor do parâmetro de dispersão $\sigma^2 = 1$ (primeiro gráfico) e $\sigma^2 = 9$ (segundo gráfico).

Como no modelo de regressão Beta, adotamos uma função de ligação $g_2(\cdot)$ logit para o parâmetro médio que garante que o parâmetro μ esteja no intervalo aberto $(0;1)$ (LÓPEZ, 2013). A função $g_2(\cdot)$ é estritamente monótona e duas vezes diferenciável, e é dada por

$$\begin{aligned} g_2(\mu_i) &= log\left(\frac{\mu_i}{1-\mu_i}\right) = Z_i^T \boldsymbol{\delta} \\ \Rightarrow \quad \mu_i &= \frac{\exp(Z_i^T \boldsymbol{\delta})}{1+\exp(Z_i^T \boldsymbol{\delta})}, \end{aligned} \tag{2.4}$$

com $Z_i = (z_{i1}, z_{i2}, \cdots, z_{iq})$, $i = 1, \cdots, n$, um conjunto de covariáveis q e $\boldsymbol{\delta} = (\delta_0, \cdots, \delta_q)$ um vetor de parâmetros desconhecidos.

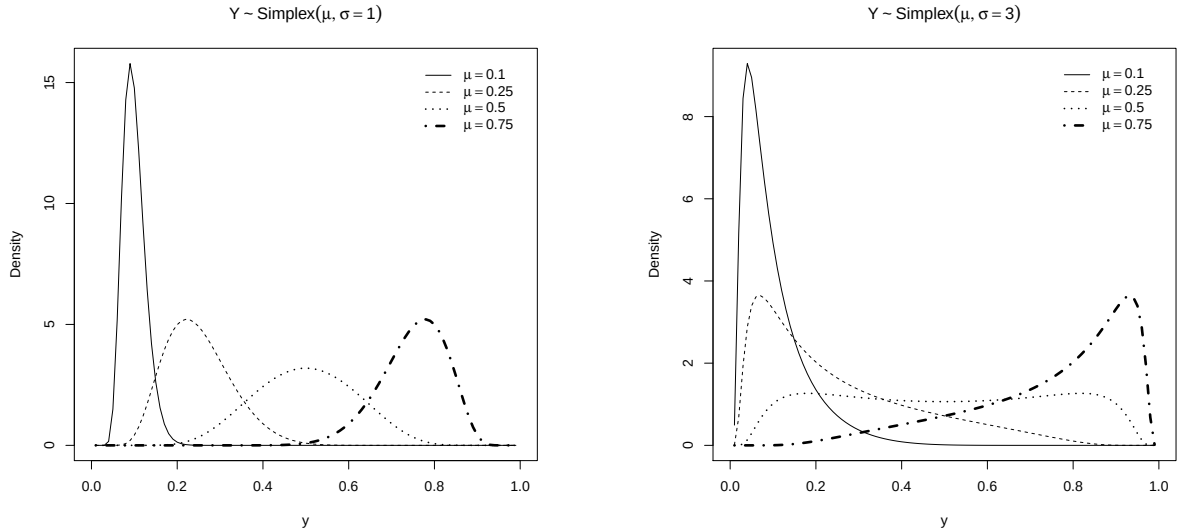


Figura 2 – Distribuição de Y para diferentes combinações dos parâmetros μ e σ^2 no modelo de regressão Simplex

2.1.3 O modelo de regressão L-Logístic

Seja Y uma variável aleatória de uma distribuição L-Logístic. Sua função de densidade de probabilidade é dada por

$$f(y \mid m,b) = \frac{b(1-m)^b m^b y^{b-1}(1-y)^{b-1}}{((1-m)^b y^b + m^b(1-y)^b)^2} \tag{2.5}$$

com $0 < y < 1$, $0 < m < 1$ e $b > 0$, onde m é a mediana e b é o parâmetro de dispersão. De acordo com Paz *et al.* (2019), a expressão para os momentos da distribuição L-Logístic é igual a $\mathbb{E}(Y^t \mid m,b) = \int_0^1 (1 + ((1-\upsilon)/\upsilon)^{1/b}(1-m)/m)^{-t} d\upsilon$ e pode ser usado para analisar alguns momentos, como $\mu_i = \mathbb{E}(Y \mid m,b)$ e $\mathbb{E}(Y^2 \mid m,b)$. A partir destes dois momentos é possível obter a esperança e a variância de Y, uma vez que não são obtidas analiticamente (PAZ *et al.*, 2019). Se quisermos obter μ_i precisamos calcular a integral $\mu_i = \int_0^1 (1 + ((1-\upsilon)/\upsilon)^{1/b}(1-m)/m)^{-1} d\upsilon$.

A Figura 3 mostra as diferentes densidades da distribuição L-Logístic (2.5) para diferentes valores de m e b.

Na análise de regressão para a distribuição L-Logístic adotamos uma função de ligação $g_3(\cdot)$ para a mediana, estritamente monótona e duas vezes diferenciável. Aqui seguimos a Paz *et al.* (2019) e optamos pela função de ligação logito, cuja justificativa é a mesma apresentada nos dois modelos anteriores (Beta e Simplex). A função $g_3(\cdot)$ é dada por

$$\begin{aligned} g_3(m_i) &= log\left(\frac{m_i}{1-m_i}\right) = T_i^T \boldsymbol{\lambda} \\ \Rightarrow \quad m_i &= \frac{\exp(T_i^T \boldsymbol{\lambda})}{1+\exp(T_i^T \boldsymbol{\lambda})}, \end{aligned} \tag{2.6}$$

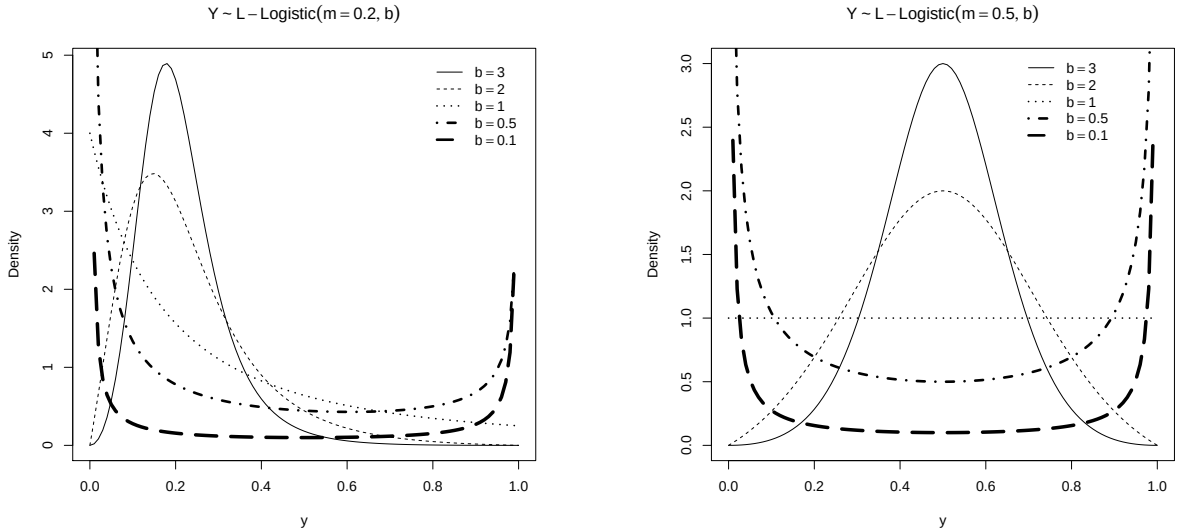


Figura 3 – Distribuição de Y para diferentes combinações dos parâmetros m e b no modelo de regressão L-Logístic

onde $T_i = (t_{i1}, t_{i2}, \cdots, t_{ik})$, $i = 1, \cdots, n$, é um conjunto de k covariáveis e $\boldsymbol{\lambda} = (\lambda_0, \cdots, \lambda_k)$ é um vetor de parâmetros desconhecidos.

2.2 Estimação

2.2.1 Estimação Bayesiana

A estimação dos parâmetros dos modelos de regressão Beta, Simplex e L-Logístic será feita utilizando a abordagem Bayesiana. Para isso, consideraremos as funções de verossimilhança escritas a partir de (2.1), (2.3) e (2.5).

Considerando $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ uma amostra independente, então

$$L(\boldsymbol{\beta}, \phi \mid \boldsymbol{Y}) = \prod_{i=1}^{n} \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y_i^{\mu\phi-1}(1-y_i)^{(1-\mu)\phi-1} \tag{2.7}$$

$$L(\boldsymbol{\delta}, \sigma^2 \mid \boldsymbol{Y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2(y_i(1-y_i))^3}} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2 y_i(1-y_i)\mu^2(1-\mu)^2}\right) \tag{2.8}$$

$$L(\boldsymbol{\lambda}, b \mid \boldsymbol{Y}) = \prod_{i=1}^{n} \frac{b(1-m)^b m^b y_i^{b-1}(1-y_i)^{b-1}}{((1-m)^b y_i^b + m^b(1-y_i)^b)^2}, \tag{2.9}$$

são as funções de verossimilhança do modelo de regressão Beta (2.7), Simplex (2.8) e L-Logístic (2.9).

Observe que para a construção da *posterior* precisamos primeiro definir as *prioris* $p(\beta, \phi)$ do modelo Beta, $p(\delta, \sigma^2)$ do modelo Simplex e $p(\lambda, b)$ do modelo L-Logístico. Como seus coeficientes de regressão são independentes dos parâmetros associados à dispersão ou à precisão,

temos

$$p(\boldsymbol{\beta},\phi) = p(\boldsymbol{\beta})p(\phi) \tag{2.10}$$
$$p(\boldsymbol{\delta},\sigma^2) = p(\boldsymbol{\delta})p(\sigma^2) \tag{2.11}$$
$$p(\boldsymbol{\lambda},b) = p(\boldsymbol{\lambda})p(b). \tag{2.12}$$

onde $p(\boldsymbol{\beta})$ é a *priori* de $\boldsymbol{\beta}$, $p(\phi)$ é a *priori* de ϕ, $p(\boldsymbol{\delta})$ é a *priori* de $\boldsymbol{\delta}$, $p(\sigma^2)$ é a *priori* de σ^2, $p(\boldsymbol{\lambda})$ é a *priori* de $\boldsymbol{\lambda}$ e $p(b)$ é a *priori* de b.

Para as *prioris* do modelo de regressão Beta (2.10) propomos $\beta_j \sim N(0,100)$ e $\phi \sim Inv-Gamma(0,01,0,01)$. Para as *prioris* do modelo de regressão Simplex (2.11) propomos $\delta_j \sim N(0,100)$ e $\sigma^2 \sim Gamma(0,01,0,01)$. Finalmente, para as *prioris* do modelo de regressão L-Logístic (2.12) propomos também $\lambda_j \sim N(0,100)$ e $b \sim Gamma(0,01,0,01)$.

Por (2.7) e (2.10), (2.8) e (2.11) e por (2.9) e (2.12) nós expressar as distribuições *posterioris*, respectivamente,

$$p(\boldsymbol{\beta},\phi \mid \boldsymbol{Y}) \propto L(\boldsymbol{\beta},\phi \mid \boldsymbol{Y})p(\boldsymbol{\beta})p(\phi) \tag{2.13}$$
$$p(\boldsymbol{\delta},\sigma^2 \mid \boldsymbol{Y}) \propto L(\boldsymbol{\delta},\sigma^2 \mid \boldsymbol{Y})p(\boldsymbol{\delta})p(\sigma^2) \tag{2.14}$$
$$p(\boldsymbol{\lambda},b \mid \boldsymbol{Y}) \propto L(\boldsymbol{\lambda},b \mid \boldsymbol{Y})p(\boldsymbol{\lambda})p(b). \tag{2.15}$$

Os três modelos de regressão foram implementados em Stan, adaptados para a linguagem de programação R através do pacote `rstan` (Stan Development Team, 2020). Rstan é a interface R do pacote Stan `C ++` e fornece inferência bayesiana completa usando o amostrador *No-U-Turn* (NUTS), uma variante do Hamiltoniano Monte Carlo (HMC).

2.2.2 Outros softwares para estimação

Utilizamos outros pacotes frequentistas e bayesianos para estimar os modelos apresentados na seção 2.1, através de pacotes que já existem na biblioteca R. Para estimar os parâmetros do modelo de regressão Beta, utilizamos os pacotes `betareg` (CRIBARI-NETO; ZEILEIS, 2010), `gamlss` (RIGBY; STASINOPOULOS, 2005; STASINOPOULOS; RIGBY *et al.*, 2007) e `INLA` (RUE; MARTINO; CHOPIN, 2009). O `betareg` ajusta modelos de regressão Beta com parâmetros de média (que vincula covariáveis à média da resposta por meio de uma função de ligação) e precisão, cujas classes associadas são projetadas para serem tão semelhantes quanto possível ao `glm` (R Core Team, 2018) padrão para ajuste de Modelos Lineares Generalizados (MLG).

O pacote `gamlss` (*Generalized additive model for location, scale and shape*) ajusta modelos cuja distribuição da variável resposta não precisa ser da família exponencial e, segundo Rigby e Stasinopoulos (2005) e Stasinopoulos, Rigby *et al.* (2007), a modelagem permite modelar não apenas a média, mas também outros parâmetros, como funções lineares e não lineares. Existem dois algoritmos básicos usados no pacote `gamlss`: `CG()` e `RS()`. O primeiro

utiliza o algoritmo de derivadas cruzadas da função de verossimilhança em relação aos parâmetros da distribuição (COLE; GREEN, 1992) e o segundo utiliza o algoritmo Rigby e Stasinopoulos (1996), Rigby e Stasinopoulos (1996) para ajuste dos modelos de média aditiva e dispersão, não utilizando as derivadas cruzadas.

O método *Integrated Nested Laplace Approximation* (INLA) é uma abordagem para inferência Bayesiana aproximada e, segundo Martino e Riebler (2014), nos últimos anos tem sido uma alternativa a outros métodos como, por exemplo, *Markov Chain Monte Carlo* (MCMC). Foi desenvolvido por Rue, Martino e Chopin (2009) e abrange uma grande família de modelos que são utilizados na prática, embora se concentre em modelos que podem ser expressos como *Gaussian Markov Random Fields* (GMRF).

No modelo de regressão Simplex, além da implementação feita em Stan, para estimar os parâmetros utilizamos os pacotes `simplexreg` (ZHANG; QIU; SHI, 2016), `gamlss` (RIGBY; STASINOPOULOS, 2005; STASINOPOULOS; RIGBY *et al.*, 2007), `VGAM` (YEE *et al.*, 2010) e `INLA` (RUE; MARTINO; CHOPIN, 2009).

O pacote `simplexreg` fornece ajuste do modelo de dispersão da distribuição Simplex e é introduzido para modelar dados proporcionais. De acordo com Zhang, Qiu e Shi (2016) ele realiza análises de regressão usando MLG, bem como equações de estimativa generalizadas baseadas na distribuição Simplex. O pacote `VGAM` é centrado no algoritmo *Iteratively Reweighted Least Squares* (IRLS) e, de acordo com Yee *et al.* (2010), ajusta modelos lineares vetoriais generalizados (do inglês, *Vector Generalized Linear Model* (VGLM)) e modelos aditivos e lineares vetoriais generalizados (do inglês *Vector Generalized Linear and Additive Models* (VGAM)), bem como modelos associados (VGLMs de baixo posto, RR-VGLMs quadráticos , VGAMs de classificação reduzida).

Por fim, também foi realizada a estimação dos parâmetros do modelo de regressão L-Logístic utilizando o método `gamlss` (RIGBY; STASINOPOULOS, 2005; STASINOPOULOS; RIGBY *et al.*, 2007) e `llbayesireg`. Este último foi desenvolvido por Paz *et al.* (2019) e é usado para estimar um modelo de regressão L-Logístic com estruturas de regressão de mediana e precisão.

Tabela 1 – Classificação dos pacotes segundo a abordagem Frequentista e Bayesiana

Modelo	Frequentista	Bayesiana
Beta	`betareg` e `gamlss`	`STAN` e `INLA`
Simplex	`simplexreg`, `gamlss` e `VGAM`	`STAN` e `INLA`
L-Logistic	`gamlss`	`STAN` e `llbayesireg`

A Tabela 1 apresenta um resumo da abordagem utilizada (frequentista ou bayesiana) para estimar os parâmetros dos três modelos em cada pacote utilizado. Podemos observar que os pacotes `betareg`, `gamlss`, `simplexreg` e `VGAM` consideram uma abordagem frequentista para estimação de parâmetros. Por outro lado, os pacotes `INLA` e `llbayesireg` e também a

implementação em STAN consideram uma abordagem bayesiana para estimação de parâmetros.

2.2.3 Critérios de comparação Bayesiana

Um passo importante na análise estatística é a comparação entre modelos. Para o cenário Bayesiano, a comparação é realizada através de alguns procedimentos de seleção propostos na literatura, como *Deviance Information Criterion* (DIC), proposto por Spiegelhalter *et al.* (2002), *Expected Akaike Information Criterion* (EAIC) e *Expected Schwarz-Bayesian Information Criterion* (EBIC), que podemos ver em Gelman *et al.* (2014), *Bayesian Predictive Information Criterion* (IC) e *Hannan–Quinn Information Criterion* (HQIC), apresentado em Anderson e Burnham (2004), *Watanabe–Akaike Information Criterion* (WAIC), presente em Gelman *et al.* (2014) e *Leave-One-Out Cross-Validation* (LOO), desenvolvida por Geisser e Eddy (1979).

O critério de seleção DIC é dado por $\text{DIC} = \bar{D} + \nu_D$, onde $\bar{D}$ é a média *posteriori* do *Deviance* e $\nu_D = \mathbb{E}(D(\theta)) - D(E(\theta))$ (θ o vetor de parâmetros) é o número efetivo de parâmetros do modelo, definido como o *deviance* esperado menos o *deviance* avaliado nas expectativas *posterioris* (SPIEGELHALTER *et al.*, 2002). Os critérios EAIC e EBIC são dados por $\text{EAIC} = \bar{D} + 2 \times \nu$ e $\text{EBIC} = \bar{D} + \nu \times log(N)$, respectivamente, em que N é o número de observações e ν é o número de parâmetros no modelo. O critério de seleção IC é calculado a partir de $\text{IC} = \bar{D} + 2 \times \nu_D$, e o critério HQIC é dado por $\text{HQIC} = \bar{D} + 2 \times \nu \times log(log(N))$. Observe que esses critérios de seleção dependem da média posterior do Deviance do modelo ($\bar{D}$).

Outros dois critérios de comparação que também são comumente utilizados, e que não dependem da média posteriori da Deviance do modelo, são *WAIC* (WATANABE; OPPER, 2010), dado por $WAIC = -2 * lppd + 2 * \nu_D$, e *LOO* (GEISSER; EDDY, 1979). Note que $lppd$ é o logarítmo da densidade preditiva pontual. Assim como WAIC, LOO utiliza o logaritmo da função de verossimilhança avaliada em simulações da distribuição *posterior* dos parâmetros.

2.3 Estudo de Recuperação de Parâmetros

Esta seção é reservada para um estudo de simulação que foi realizado a partir dos três modelos apresentados nas seções anteriores, utilizando abordagem bayesiana para a estimação dos parâmetros. Os modelos foram implementados usando o pacote rstan (Stan Development Team, 2020) e para obter convergência via MCMC, usando uma cadeia, consideramos 40000 iterações, 10 *thin* e 10000 *warmup*. Para este estudo, geramos 100 repetições de amostras de tamanho $100, 200, 500$ e 1000.

A fim de verificar a recuperação dos parâmetros, buscamos resultados quanto à média dos estimadores (Média), ao desvio padrão (DP), ao viés (Viés) e à raiz quadrada do erro quadrático médio (RMSE). O viés simulado é calculado a partir da diferença entre a média das estimativas e o verdadeiro valor do parâmetro, ou seja, é definido por $b(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$, com θ sendo o vetor

de parâmetros. O valor de RMSE e o desvio padrão das estimativas dos parâmetros do modelo são calculados pelas respectivas expressões $RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{\theta}_i - \theta)^2}{n}}$ e $\text{DP} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{\theta}_i - \mathbb{E}(\hat{\theta}))^2}{n-1}}$, no qual n é o número de amostras simuladas, θ é o valor verdadeiro do parâmetro, $\hat{\theta}_i$ é a estimativa de θ na i-ésima amostra simulada e $\mathbb{E}(\hat{\theta})$ é a média da amostra $\hat{\theta}_1, \cdots, \hat{\theta}_n$.

Para o modelo de regressão Beta, geramos uma covariável x de uma distribuição $U(-2,2)$ para predizer as respostas e fixamos os valores $\beta_0 = 0$, $\beta_1 = 1$ e $\phi = 10$ para os coeficientes de regressão, na qual $\beta_j \sim N(0,100)$ e $log(\phi) \sim N(0,10)$, para $j = 1,2$. Os resultados do estudo de simulação do modelo de regressão Beta encontram-se na Tabela 2.

Tabela 2 – Estudo de recuperação dos parâmetros do modelo de regressão Beta utilizando stan - 100 réplicas.

n	Par	Média	DP	Viés	RMSE
100	β_0	0,037	0,007	-0,037	0,001
	β_1	1,025	0,015	-0,025	0,001
	ϕ	12,152	0,243	-2,152	4,688
200	β_0	0,049	0,003	-0,049	0,002
	β_1	1,014	0,004	-0,014	0,000
	ϕ	11,040	0,114	-1,040	1,095
500	β_0	0,021	0,001	-0,021	0,000
	β_1	1,004	0,005	-0,004	0,000
	ϕ	10,787	0,085	-0,787	0,627
1000	β_0	-0,010	0,004	0,010	0,000
	β_1	0,956	0,005	0,044	0,002
	ϕ	9,570	0,097	0,430	0,195

É possível notar na Tabela 2 que o viés se aproxima de zero conforme o tamanho da amostra aumenta, com exceção de β_1, indicando que a média das estimativas dos parâmetros se aproxima do valor real do parâmetro conforme aumenta-se o tamanho da amostra. Além disso, os valores RMSE aproximam-se de zero, assim como os valores do desvio padrão. O gráfico da Figura 4 também mostra este resultado para o RMSE.

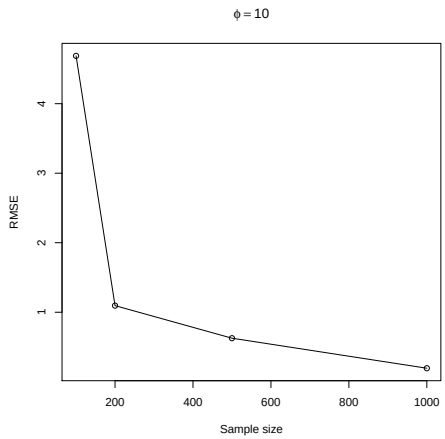


Figura 4 – RMSE de ϕ para diferentes tamanhos de amostra com stan.

A Tabela 3 apresenta os resultados do estudo de simulação do modelo de regressão Simplex. Ao considerar alguns estudos preliminares, detectamos algumas inconsistências na recuperação do parâmetro σ^2 e alguma sensibilidade na escolha da *priori*. Assim decidimos desenvolver um estudo mais detalhado com este modelo.

Tabela 3 – Estudo de recuperação dos parâmetros do modelo de regressão Simplex utilizando stan - 100 réplicas.

			$\sigma^2 = 10$				$\sigma^2 = 50$			
	n	Par	Média	DP	Viés	RMSE	Média	DP	Viés	RMSE
Priori 1	100	δ_0	0,067	0,006	-0,067	0,004	0,211	0,002	-0,211	0,045
		δ_1	0,931	0,003	0,069	0,005	0,859	0,022	0,141	0,020
		σ^2	10,483	0,066	-0,483	0,238	49,561	0,937	0,439	1,071
	200	δ_0	0,075	0,006	-0,075	0,006	0,114	0,005	-0,114	0,013
		δ_1	0,986	0,003	0,014	0,000	0,959	0,007	0,041	0,002
		σ^2	9,728	0,085	0,272	0,081	48,828	0,036	1,172	1,375
	500	δ_0	0,044	0,001	-0,044	0,002	0,121	0,014	-0,121	0,015
		δ_1	0,968	0,000	0,032	0,001	0,951	0,012	0,049	0,003
		σ^2	9,171	0,067	0,829	0,692	45,536	0,435	4,464	20,116
	1000	δ_0	0,001	0,003	-0,001	0,000	0,036	0,002	-0,036	0,001
		δ_1	1,023	0,001	-0,023	0,001	1,039	0,007	-0,039	0,002
		σ^2	9,012	0,027	0,988	0,977	44,355	0,240	5,645	31,925
			$\sigma^2 = 10$				$\sigma^2 = 50$			
	n	Par	Média	DP	Viés	RMSE	Média	DP	Viés	RMSE
Priori 2	100	δ_0	0,066	0,003	-0,066	0,004	0,214	0,000	-0,214	0,046
		δ_1	0,935	0,007	0,065	0,004	0,855	0,000	0,145	0,021
		σ^2	10,531	0,120	-0,531	0,297	49,332	0,000	0,668	0,446
	200	δ_0	0,075	0,006	-0,075	0,006	0,191	0,006	-0,191	0,037
		δ_1	0,988	0,003	0,012	0,000	1,014	0,010	-0,014	0,000
		σ^2	9,746	0,088	0,254	0,072	46,749	0,380	3,251	10,713
	500	δ_0	0,044	0,001	-0,044	0,002	0,121	0,002	-0,121	0,015
		δ_1	0,968	0,000	0,032	0,001	0,950	0,003	0,050	0,003
		σ^2	9,164	0,063	0,836	0,703	45,536	0,304	4,464	20,024
	1000	δ_0	0,001	0,003	-0,001	0,000	0,031	0,002	-0,031	0,001
		δ_1	1,024	0,001	-0,024	0,001	1,000	0,001	0,000	0,000
		σ^2	9,015	0,026	0,985	0,970	46,005	0,125	3,995	15,974
			$\sigma^2 = 10$				$\sigma^2 = 50$			
	n	Par	Média	DP	Viés	RMSE	Média	DP	Viés	RMSE
	100	δ_0	0,070	0,003	-0,070	0,005	0,216	0,003	-0,216	0,046
		δ_1	0,935	0,012	0,065	0,004	0,863	0,022	0,137	0,019
		σ^2	10,105	0,091	-0,105	0,019	47,561	0,891	2,439	6,743
		δ_0	0,074	0,007	-0,074	0,006	0,192	0,005	-0,192	0,037

	200	δ_1	0,989	0,003	0,011	0,000	1,012	0,010	-0,012	0,000
Priori 3		σ^2	9,556	0,085	0,444	0,205	45,928	0,400	4,072	16,738
		δ_0	0,045	0,001	-0,045	0,002	0,120	0,003	-0,120	0,015
	500	δ_1	0,969	0,000	0,031	0,001	0,949	0,003	0,051	0,003
		σ^2	9,098	0,065	0,902	0,817	45,219	0,305	4,781	22,953
		δ_0	-0,001	0,000	0,001	0,000	0,031	0,001	-0,031	0,001
	1000	δ_1	1,023	0,003	-0,023	0,001	1,001	0,001	-0,001	0,000
		σ^2	8,993	0,005	1,007	1,015	45,886	0,093	4,114	16,933
			$\sigma^2 = 10$				$\sigma^2 = 50$			
	n	Par	Média	DP	Viés	RMSE	Média	DP	Viés	RMSE
		δ_0	0,066	0,004	-0,066	0,004	0,214	0,003	-0,214	0,046
	100	δ_1	0,933	0,022	0,067	0,005	0,861	0,023	0,139	0,020
		σ^2	10,109	0,052	-0,109	0,015	47,809	0,861	2,191	5,543
		δ_0	0,075	0,006	-0,075	0,006	0,190	0,006	-0,190	0,036
	200	δ_1	0,987	0,003	0,013	0,000	1,013	0,010	-0,013	0,000
Priori 4		σ^2	9,556	0,085	0,444	0,204	46,070	0,404	3,930	15,604
		δ_0	0,046	0,001	-0,046	0,002	0,121	0,002	-0,121	0,015
	500	δ_1	0,968	0,000	0,032	0,001	0,948	0,003	0,052	0,003
		σ^2	9,089	0,066	0,911	0,834	45,161	0,306	4,839	23,508
		δ_0	0,000	0,003	0,000	0,000	0,030	0,001	-0,030	0,001
	1000	δ_1	1,023	0,001	-0,023	0,001	1,001	0,002	-0,001	0,000
		σ^2	8,990	0,027	1,010	1,021	45,927	0,111	4,073	16,603

Neste estudo geramos uma covariável z de uma distribuição $U(-2,2)$ para predizer as respostas e criamos dois cenários. Para o cenário 1 fixamos os seguintes valores para os coeficientes de regressão: $\delta_0 = 0$, $\delta_1 = 1$ e $\sigma^2 = 10$. Já no cenário 2, os coeficientes de regressão assumem os seguintes valores: $\delta_0 = 0$, $\delta_1 = 1$ e $\sigma^2 = 50$. Em ambos os cenários assumimos $\delta_j \sim N(0,10)$, para $j = 1,2$, e tomamos $\sigma = \frac{1}{\sqrt{\tau}}$, com quatro *prioris* distintas para τ. A *Priori* 1 corresponde à *priori* $\tau \sim Gamma(0,01,0,01)$, a *Priori* 2 corresponde à $\tau \sim Gamma(0,1,0,1)$, a *Priori* 3 corresponde à $\tau \sim Gamma(1,10e-05)$ e a *Priori* 4 corresponde à $\tau \sim Gamma(1,10e-02)$. Os resultados são apresentados na Tabela 3.

Os resultados da recuperação de parâmetros são muito bons para os parâmetros δ_0 e δ_1. Percebemos que em todos as *prioris* os resultados de viés, desvio padrão e RMSE tendem a zero, à medida que o tamanho da amostra aumenta. Por outro lado, os resultados indicam que existe uma sensibilidade do modelo na escolha da *priori* de σ^2. Para $\sigma^2 = 10$ vemos que a *priori* recupera bem o parâmetro (veja também na Figura 5, primeiro gráfico) e para $\sigma^2 = 50$ vemos que os resultados de *Priori* 2, *Priori* 3 e *Priori* 4 começam a melhorar para n suficientemente grandes (veja também Figura 5, segundo gráfico). Note-se que dentre eles, a *Priori* 2 é a mais recomendada, pois apresenta o valor de RMSE para tamanho amostral 100 muito menor que

o valor apresentado na *Priori* 3 e na *Priori* 4. Este tipo de resultados não foram observados anteriormente.

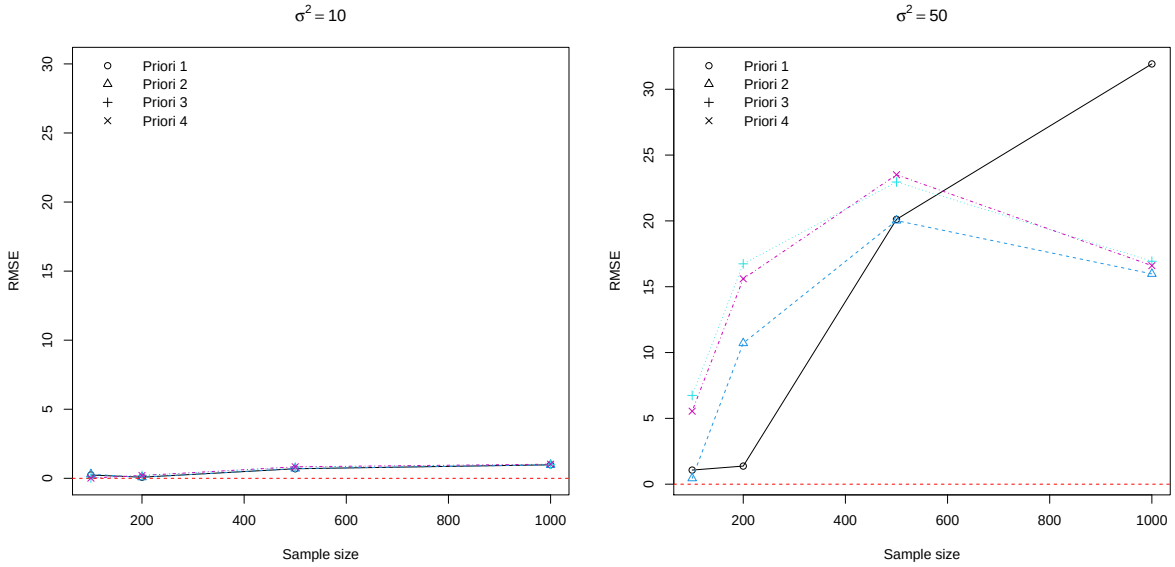


Figura 5 – RMSE de σ^2 para diferentes tamanhos de amostra com stan.

Por fim, para o modelo de regressão L-Logistic, geramos uma covariável t, também de uma distribuição $U(-2,2)$, para predizer as respostas e fixamos os valores $\lambda_0 = 0$, $\lambda_1 = 1$ e $b = 2,4$ para os coeficientes de regressão, na qual $\lambda_j \sim N(0,100)$ e $b \sim Gamma(0,01,0,01)$, para $j = 1,2$. Os resultados do estudo de simulação do modelo de regressão L-Logistic encontram-se na Tabela 4.

Tabela 4 – Estudo de recuperação dos parâmetros do modelo de regressão L-Logístic utilizando stan - 100 réplicas.

n	Par	Média	DP	Viés	RMSE
100	λ_0	0,001	0,000	-0,001	0,000
	λ_1	0,978	0,000	0,022	0,000
	b	2,462	0,000	-0,062	0,004
200	λ_0	0,007	0,000	-0,007	0,000
	λ_1	0,944	0,000	0,056	0,003
	b	2,656	0,000	-0,256	0,066
500	λ_0	-0,013	0,000	0,013	0,000
	λ_1	0,971	0,003	0,029	0,001
	b	2,459	0,000	-0,059	0,003
1000	λ_0	-0,007	0,000	0,007	0,000
	λ_1	0,973	0,001	0,027	0,001
	b	2,418	0,000	-0,018	0,000

A partir da Tabela 4 podemos perceber que os resultados de desvio padrão, viés e RMSE estão muito próximos de zero e diminuem conforme o tamanho da amostra aumenta. Além disso, os valores desses três resultados são bem próximos, para valores suficientemente grande de n.

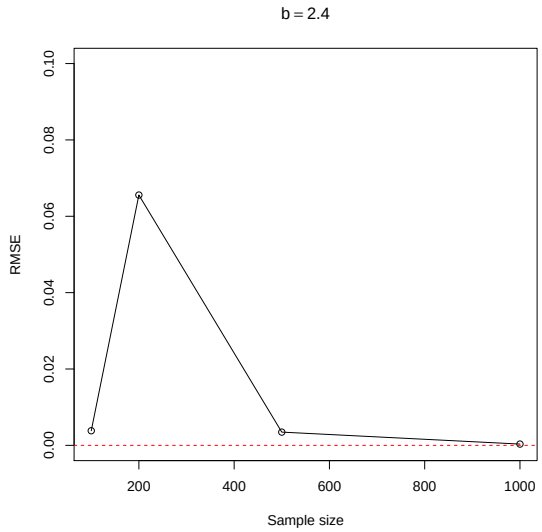


Figura 6 – RMSE de estimativas de b.

A Figura 6 mostra que a partir do tamanho de amostra 200 o valor de RMSE decresce até se aproximar de zero, como foi visto na Tabela 4.

Em resumo, considerando os resultados do estudo de simulação, podemos dizer que os modelos de regressão Beta e L-Logístic são capazes de estimar bem os parâmetros para todos os tamanhos de amostra. Por outro lado, o modelo de regressão Simplex estima bem os parâmetros para amostras maiores que 500.

2.4 Aplicações

Nesta seção, apresentamos, inicialmente, um estudo que foi realizado para verificar o desempenho dos diferentes pacotes, na qual geramos dados dos modelos de regressão Beta, Simplex e L-Logístic e fizemos o ajuste utilizando os métodos vistos em cada modelo. Na sequência, apresentamos duas aplicações realizadas com dados reais para os três modelos, cujo ajuste também foi feito utilizando métodos específicos para cada modelo.

2.4.1 Estudando o desempenho de diferentes pacotes

Para este estudo, foram gerados 1000 dados de cada distribuição proposta com coeficientes de regressão $\boldsymbol{\beta} = (\beta_0, \beta_1)$, $\boldsymbol{\delta} = (\delta_0, \delta_1)$ e $\boldsymbol{\lambda} = (\lambda_0, \lambda_1)$, associados às covariáveis x_{i1}, z_{i1} e t_{i1}, respectivamente, $i = 1, \cdots, 1000$, que se relacionam com as médias das respostas (modelos de regressão Beta e Simplex) e com a mediana (modelo de regressão L-Logistic) através das funções de ligação apresentadas em (2.2), (2.4) e (2.6), respectivamente.

Os valores dos coeficientes dos parâmetros foram fixados em $\beta_0 = 0$, $\beta_1 = 1$, $\phi = 10$, $\delta_0 = 0$, $\delta_1 = 1$, $\sigma^2 = 10$, $\lambda_0 = 0$, $\lambda_1 = 1$ e $b = 2,4$, e as covariáveis x, z, t foram geradas a partir de $U(-2,2)$. A metodologia Bayesiana foi utilizada para estimar os parâmetros dos três mode-

los e foram propostas as seguintes *prioris*: $\beta_j \sim N(0,100)$, $log(\phi) \sim N(0,10)$, $\Delta_j \sim N(0,10)$, $\tau \sim Gamma(0,1,0,1)$, $\Lambda_j \sim N(0,100)$ e $b \sim Gamma(0,01,0,01)$, $j = 0,1$. Os modelos foram implementados utilizando a linguagem R, especificamente o pacote rstan (Stan Development Team, 2020), e para obter convergência via MCMC, utilizando uma cadeia, consideramos 10000 iterações.

Com a finalidade de comparar a eficiência dos pacotes utilizados para a estimação dos parâmetros nos três modelos, nós calculamos o valor da precisão em cada pacote, que depende do valor inicial do parâmetro, da estimativa do parâmetro e do tamanho da amostra. A expressão da precisão é dada por $\sqrt{\frac{1}{k}\sum_{k=1}^{3}(\theta_k - \hat{\theta}_k)^2}$, na qual k é o número de parâmetros do modelo, θ_k é o valor inicial do k-ésimo parâmetro e $\hat{\theta}_k$ é a estimativa do k-ésimo parâmetro no referido pacote/algoritmo utilizado.

Tabela 5 – Estimativas do modelo de regressão Beta.

		Frequentista		Bayesiana	
	real	betareg	gamlss	stan	INLA
β_0	0	0,017	0,017	0,017	0,017
β_1	1	0,963	0,963	0,964	0,963
ϕ	10	9,669	10,669	9,657	9,660
Precisão:		0,193	0,387	0,199	0,198
Tempo (secs):		0,467	1,008	505,542	7,514

Tabela 6 – Estimativas do modelo de regressão Simplex.

		Frequentista			Bayesiana	
	real	simplexreg	gamlss	VGAM	stan	INLA
δ_0	0	-0,030	0,035	0,035	0,034	0,035
δ_1	1	1,049	0,995	1,123	0,994	0,994
σ^2	10	9,477	9,458	7,311	9,488	9,434
Precisão:		0,304	0,314	1,554	0,296	0,327
Tempo:		0,673	0,675	0,224	357,533	9,813

Tabela 7 – Estimativas do modelo de regressão L-Logistic.

		Frequentista	Bayesiana	
	real	gamlss	stan	llbayesireg
λ_0	0	0,006	0,006	0,006
λ_1	1	0,968	0,967	0,968
b	2,4	2,381	2,378	2,378
Precisão:		0,022	0,023	0,023
Tempo:		0,498	478,459	490,371

A Tabela 5 apresenta os resultados do modelo de regressão Beta. Note que a estimação realizada pelos pacotes é muito próxima dos valores fixados para cada parâmetro, sendo a estimativa de β_0 igual a 0,017, β_1 aproximadamente igual a 0,96 e ϕ igual a 9,66 aproximadamente, com excessão do pacote gamlss que estima ϕ em 10,66. Quanto ao valor da precisão, note que o

valor mais próximo de zero é do pacote `betareg`, seguido dos pacotes `INLA`, `stan` e `gamlss`, o que significa que as estimativas obtidas através do pacote `betareg` são as que mais se aproximam do verdadeiro valor, em comparação às estimativas obtidas pelos demais pacotes.

A Tabela 6 apresenta os resultados do modelo de regressão Simplex para cada pacote utilizado. É possível perceber que, com exceção do pacote VGAM, todos os demais pacotes recuperaram bem os valores dos parâmetros. No caso do pacote VGAM, a estimativa do parâmetro δ_1 está um pouco acima em comparação à estimativa do pacote Simplexreg e do algoritmo NUTS, por exemplo. Já a estimativa de σ^2 está muito abaixo do esperado. Essa baixa eficiência do pacote VGAM na recuperação dos parâmetros pode ser observada no valor da precisão, já que é o valor mais alto apresentado, em comparação aos demais pacotes. Por outro lado, o pacote stan apresenta a melhor eficiência.

Na Tabela 7 podemos notar que as estimativas dos parâmetros do modelo de regressão L-Logistic estão muito próximas dos valores fixados, o que indica que os pacotes utilizados recuperam bem os parâmetros. Em termos de valor da precisão, podemos considerar o pacote `gamlss` como sendo o mais eficiente na recuperação dos parâmetros do modelo, seguidos dos pacotes `llbayesireg` e `stan`, nesta ordem.

2.4.2 Desempenho dos modelos com dados de amostras pequenas

A aplicação foi realizada sobre um conjunto de dados reais sobre a pobreza no Peru. O conjunto de dados fornece informações sobre o número de pobres, extremamente pobres, não extremamente pobres e não pobres nas cidades das 25 subdivisões territoriais do país, bem como o coeficiente de variação da pobreza total, o Índice de Desenvolvimento Humano (IDH), a expectativa de vida ao nascer, alfabetização, educação, desempenho educacional e renda familiar per capita.

Uma variável dependente e uma variável independente foram consideradas nesta aplicação. O conjunto de dados tem 195 observações e estamos interessados em modelar a proporção de pessoas extremamente pobres no Peru usando o Índice de Desenvolvimento Humano (IDH). Para isso, aplicamos o conjunto de dados aos três modelos vistos anteriormente e consideramos as funções de ligação dadas em (2.2), (2.4) e (2.6), cujos preditores lineares são dados por $\eta_i = \beta_0 + \beta_1 x_{i1}$, $\eta_i = \delta_0 + \delta_1 z_{i1}$ e $\eta_i = \lambda_0 + \lambda_1 t_{i1}$, $i = 1, \cdots, 195$, respectivamente, e x_{i1}, z_{i1} e t_{i1} a variável de IDH para cada indivíduo i.

Como na aplicação anterior, estimamos os parâmetros da abordagem bayesiana usando as *prioris* $\beta_j \sim N(0,100)$, $log(\phi) \sim N(0,10)$, $\delta_j \sim N(0,10)$, $\tau \sim Gamma(0,1,0,1)$, $\lambda_j \sim N(0,100)$ e $b \sim Gamma(0,01,0,01)$, $j = 0,1$, e implementando os modelos no pacote rstan (Stan Development Team, 2020), na linguagem R, obtendo convergência via MCMC utilizando uma cadeia e 10000 iterações.

Os resultados são mostrados nas tabelas 8, 9 e 10. A Tabela 8 apresenta as estimativas

dos parâmetros β_0, β_1 e ϕ do modelo de regressão Beta e podemos observar que os valores das estimativas via `stan` estão muito próximos dos valores estimados via `betareg`, `gamlss` e `INLA`.

Os resultados apresentados na Tabela 9 referem-se às estimativas de δ_0, δ_1 e σ^2 do modelo de regressão Simplex via `stan`, `simplexreg`, `gamlss`, `VGAM` e `INLA`. Podemos ver que a estimativa de δ_0, via `stan`, é um pouco menor que as estimativas dos outros pacotes e as estimativas de δ_1 e σ^2 são um pouco maiores, em comparação com as estimativas dos demais.

Tabela 8 – Estimativas do modelo de regressão Beta.

	Frequentista		Bayesiana	
	betareg	gamlss	stan	INLA
β_0	9,741	9,741	9,718	9,708
β_1	-19,360	-19,360	-19,318	-19,296
ϕ	15,460	16,473	15,360	15,260

Tabela 9 – Estimativas do modelo de regressão Simplex.

	Frequentista			Bayesiana	
	simplexreg	gamlss	VGAM	stan	INLA
δ_0	14,667	14,665	14,746	12,892	14,479
δ_1	-28,202	-28,198	-28,325	-25,254	-27,892
σ^2	50,194	19,609	49,681	51,066	50,000

Tabela 10 – Estimativas do modelo de regressão L-Logistic.

	Frequentista	Bayesiana	
	gamlss	stan	llbayesireg
λ_0	13,048	12,789	12,768
λ_1	-25,347	-24,899	-24,865
b	2,430	2,414	2,415

A Tabela 10 apresenta as estimativas dos parâmetros λ_0, λ_1 e b do modelo de regressão L-Logístic e, assim como nos dois resultados anteriores, podemos observar que as estimativas via `stan` estão muito próximos às estimativas via `gamlss` e `llbayesireg`.

Por fim, na Tabela 11 dispomos os resultados de critério de comparação, apresentados na seção 2.2.3, dos modelos de regressão Beta, Simplex e L-Logistic.

Tabela 11 – Resultados de critério de comparação usando dados de amostras pequenas.

	BETA	SIMPLEX	LLOGISTIC
DIC	-440,748	0,203	-423,938
EAIC	-439,785	1,421	-422,630
EBIC	-433,239	7,967	-416,084
IC	-437,711	2,986	-421,245
HQIC	-437,135	4,072	-419,980
CAIC	-418,693	22,513	-401,538
WAIC	-438,562	22,698	-422,923
LOO	-438,439	17,648	-422,894

A partir dos resultados apresentados na Tabela 11, percebemos que o valor do DIC do modelo de regressão Beta é um pouco menor que o valor do DIC do modelo de regressão L-Logistic e muito menor que o DIC do modelo de regressão Simplex. Para as demais medidas, obtemos a mesma conclusão. Isso significa que o modelo de regressão Beta é o modelo que melhor se ajusta a esse conjunto de dados, nas características aqui apresentadas.

O modelo de regressão Simplex apresenta resultados muito distantes do que seria considerado plausível para comparação, um motivo pode ser devido ao tamanho de amostra pequeno. Com isso, percebemos a necessidade de um estudo mais detalhado sobre os resultados dos critérios de comparação.

2.4.3 Desempenho dos modelos com dados de amostras grandes

Na seção 2.4.2, realizamos uma aplicação com dados reais envolvendo pobreza e índice de desenvolvimento humano do Peru. Dessa aplicação percebemos uma sensibilidade do modelo de regressão Simplex quanto aos resultados a partir de um tamanho de amostra pequeno. Nesta seção, realizamos uma segunda aplicação com dados reais, cuja variáveis são a taxa de pobreza e o Índice de Desenvolvimento Humano (IDH) referentes aos municípios brasileiros.

O conjunto de dados utilizado para esta aplicação contém 5563 observações e apresenta informações dos principais indicadores das dimensões de desenvolvimento humano dos munícipos do Brasil. O conjunto de dados foi aplicado nos modelos de regressão Beta, Simplex e L-Logistic e, para isso, consideramos as funções de ligação apresentadas em (2.2), (2.4) e (2.6), cujos preditores lineares são dados por $\eta_i = \beta_0 + \beta_1 x_{i1}$, $\eta_i = \delta_0 + \delta_1 z_{i1}$ e $\eta_i = \lambda_0 + \lambda_1 t_{i1}$, $i = 1, \cdots, 5563$, respectivamente, e x_{i1}, z_{i1} e t_{i1} a variável de IDH para cada i individual.

Para as *prioris* foram utilizadas as seguintes distribuições: $\beta_j \sim N(0, 100)$, $log(\phi) \sim N(0, 10)$, $\delta_j \sim N(0, 10)$, $\tau \sim Gamma(0,1, 0,1)$, $\lambda_j \sim N(0, 100)$ e $b \sim Gamma(0,01, 0,01)$, $j = 0, 1$. Os modelos foram implementados utilizando o pacote rstan (Stan Development Team, 2020), na linguagem R, obtendo convergência via MCMC usando uma cadeia e 10000 iterações.

Os resultados encontram-se na Tabela 12, na Tabela 13 e na Tabela 14. Numa visão geral, podemos perceber que os resultados apresentados em cada modelo são muito satisfatórios. Em particular, os resultados referentes ao modelo de regressão Beta (Tabela 12) mostram que apenas o pacote `gamlss` estima ϕ um pouco acima.

Tabela 12 – Estimativas dos parâmetros do modelo de regressão Beta.

	Frequentista		Bayesiana	
	betareg	gamlss	stan	INLA
β_0	8,068	8,068	8,067	8,068
β_1	-14,452	-14,452	-14,450	-14,451
ϕ	34,192	35,204	34,175	34,150

Tabela 13 – Estimativas dos parâmetros do modelo de regressão Simplex.

	Frequentista			Bayesiana	
	simplexreg	gamlss	VGAM	stan	INLA
δ_0	8,588	8,587	8,588	8,587	8,588
δ_1	-15,353	-15,352	-15,353	-15,352	-15,353
σ^2	4,556	4,551	4,554	4,555	4,545

Tabela 14 – Estimativas dos parâmetros do modelo de regressão L-Logistic.

	Frequentista	Bayesiana	
	gamlss	stan	llbayesireg
λ_0	8,896	8,892	8,894
λ_1	-15,868	-15,862	-15,865
b	3,475	3,473	3,474

Já os resultados do modelo de regressão Simplex, apresentados na Tabela 13, indicam que o pacote `gamlss` e o algoritmo NUTS funcionam muito bem para amostras de tamanho grande, os valores estimados estão muito próximos às estimativas fornecidas pelos demais pacotes. E por fim, os resultados apresentados na Tabela 14 mostram que os pacotes utilizados na estimação dos parâmetros do modelo de regressão L-Logistic são muito eficientes.

As mesmas observações acerca da boa funcionalidade dos modelos perante uma amostra grande podem ser feitas a partir dos resultados dos critérios de comparação, mostrados na Tabela 15.

Tabela 15 – Resultado de criterios de comparação usando dados maiores.

	BETA	SIMPLEX	LLOGISTIC
DIC	-16279,65	-12155,96	-15834,12
EAIC	-16278,56	-12155,13	-15833,13
EBIC	-16272,01	-12148,58	-15826,58
IC	-16276,74	-12152,78	-15831,11
HQIC	-16275,91	-12152,48	-15830,48
CAIC	-16257,46	-12134,04	-15812,04
WAIC	-16279,50	-12148,39	-15833,97
LOO	-16279,49	-12148,29	-15833,97

Podemos ver que o valor do DIC do modelo de regressão Beta é menor que o valor do DIC do modelo de regressão Simplex e L-Logistic, assim como também as medidas de EAIC, EBIC, IC, HQIC, CAIC, WAIC e LOO. Embora os modelos de regressão Simplex e L-Logistic tenham apresentado ótima eficiência.

2.5 Comentários Finais

Inicialmente realizamos um estudo de simulação com os três modelos para verificar quanto à recuperação dos parâmetros dos pacotes e algoritmos utilizados em cada um. Percebe-

mos que no caso do modelo de regressão Simplex, por exemplo, o modelo apresentava melhores resultados para tamanho de amostra grande.

Na sequência, realizamos três aplicações: a primeira aplicação com dados simulados e na sequência duas com dados reais acerca da pobreza no Peru e nos municípios do Brasil, nesta ordem, sendo a variável explicativa o Índice de Desenvolvimento Humano.

Nas três aplicações usamos a função de ligação logit para relacionar a covariável à resposta média (modelo de regressão Beta e Simplex) e à mediana (modelo de regressão L-Logístic), e o objetivo das aplicações foi verificar a eficiência do algoritmo NUTS (rstan).

Dentro da aplicação com dados simulados, observamos a eficiência do algoritmo, que se mostrou satisfatória, assim como os demais pacotes utilizados, com excessão do pacote `VGAM` no modelo de regressão Simplex. Já na aplicação com dados reais, observamos que o algoritmo NUTS apresentou-se eficientemente em comparação aos pacotes utilizados para estimação dos parâmetros.

Quanto aos critérios de compração nas aplicações com dados reais, observamos que o modelo de regressão Simplex foi passível de comparação quando o tamanho da amostra do conjunto de dados era grande, que foi o caso da Aplicação 3, com tamanho de amostra 5563. Na Aplicação 2, cujo tamanho da amostra era igual a 195, não foi possível realizar uma comparação do modelo de regressão Simplex com os outros dois, devido aos resultados serem muito ruins.

De forma geral, percebemos que o algoritmo via Stan apresenta eficiência satisfatória, e na sua utilização podemos apontar algumas vantagens, como por exemplo, a possibilidade de implementar qualquer distribuição de interesse do pesquisador, ao contrário de alguns pacotes prontos, que trazem uma lista de distribuições disponíveis para o uso desse pacote específico. Outra observação interessante é que a implementação via Stan permite especificar o tipo de variável em estudo (inteiro, vetor, se está em um intervalo específico, etc).

CAPÍTULO

3

UM NOVO MODELO DE REGRESSÃO PARA RESPOSTAS LIMITADAS E APLICAÇÕES

Segundo Robert *et al.* (2019), o objetivo da regressão tradicional é avaliar o efeito de uma ou mais variáveis explicativas na média da variável resposta, porém em um contexto paramétrico, esta análise é inviável quando a distribuição estatística não possui uma forma simples para a média, dificultando a avaliação dos efeitos das covariáveis na resposta média (MAZUCHELI *et al.*, 2022). Uma alternativa é utilizar a regressão quantílica, proposta em Koenker e Machado (1999), cuja média é substituída por um conjunto definido de quantis, permitindo uma melhor verificação das relações entre a resposta e as covariáveis.

Neste sentido, este capítulo está estruturado da seguinte forma: A seção 3.1 apresenta a proposta da distribuição LG-Logistic e sua reparametrização em termos de quantis. A seção 3.2 apresenta o modelo de regressão quantílica LG-Logistic. A seção 3.3 é dedicada à estimação bayesiana do modelo. Na seção 3.4 realizamos um estudo de simulação para sensibilidade a priori e um estudo de parâmetros de recuperação. A seção 3.5 apresenta duas aplicações com dados reais, uma utilizando a distribuição e outra o modelo de regressão correspondente. A seção 3.6 apresenta algumas conclusões.

3.1 A distribuição LG-Logistic

Propomos nesta seção uma nova distribuição limitada que generaliza a distribuição L-Logistic.

Dizemos que uma variável aleatória (v.a.) Y segue a distribuição LG-Logistic, e será denotada por $Y \sim \text{LGL}(m, \tau, \alpha)$, se sua função densidade de probabilidade (fdp) for dada por

$$f(y \mid m, \tau, \alpha) = \alpha\tau \left[1 + \left(\frac{m(1-y)}{(1-m)y}\right)^{\tau}\right]^{-(\alpha+1)} \left(\frac{m(1-y)}{(1-m)y}\right)^{\tau} \frac{1}{y(1-y)} \quad , 0 < y < 1, \quad (3.1)$$

onde $0 < m < 1$ é um parâmetro de localização, $\tau > 0$ é um parâmetro de escala e $\alpha > 0$ é um parâmetro de forma. A função de distribuição acumulada (fda) da distribuição LG-Logistic é dada por

$$F(y \mid m, \tau, \alpha) = \left[1 + \left(\frac{m(1-y)}{(1-m)y}\right)^{\tau}\right]^{-\alpha} \tag{3.2}$$

que pode ser facilmente invertida, para uma probabilidade fixa $p \in (0,1)$, para gerar a função quantil

$$\kappa(p) = F^{-1}(p) = \frac{\left(\frac{m}{1-m}\right)}{\left(\frac{m}{1-m}\right) + \left(\frac{1-p^{\frac{1}{\alpha}}}{p^{\frac{1}{\alpha}}}\right)^{\frac{1}{\tau}}}, \tag{3.3}$$

com $0 < \kappa(p) < 1$.

Usamos a notação LG-Logistic, pois esta usa a distribuição Logistic Generalizada (ZELTERMAN, 1987) como distribuição base.

São discutidos dois casos particulares interessantes da distribuição LG-Logistic. A primeira e a já mencionada é a distribuição L-Logistic (PAZ *et al.*, 2019), que surge quando $\alpha = 1$, que será denotado por $Y \sim \mathrm{LL}(m, \tau)$. O segundo caso é quando $m = 0,5$. Neste caso, obtemos uma distribuição com dois parâmetros, que já foi apresentada em Balakrishnan e Leung (1988) (veja mais detalhes em Alkasasbeh e Raqab (2009)). Chamaremos a essa segunda de distribuição G-Logistic e usaremos a notação $Y \sim \mathrm{GL}(\tau, \alpha)$.

A Figura 7 ilustra algumas das diferentes formas que a função densidade LG-Logistic pode assumir, de acordo com valores específicos de m, τ e α. Na Figura 7 (a), fixamos os valores de $m = 0,3$ e $\alpha = 1$, escolhendo alguns valores para o parâmetro τ. Isto corresponde ao caso da distribuição L-Logistic, mais precisamente à distribuição da v.a. $Y \sim \mathrm{LL}(m = 0,3, \tau)$. Observe que τ é de fato um parâmetro de forma e quando $\tau < 1$ a curva de distribuição tem formato de banheira, caso contrário é unimodal.

Na Figura 7 (b), fixamos os valores de $m = 0,5$ e $\alpha = 1,2$, escolhendo alguns valores para o parâmetro τ. Isso corresponde ao caso da v.a. $Y \sim \mathrm{LGL}(m = 0.5, \tau, \alpha = 1.2)$. Notamos que à medida que alteramos o valor de τ, a curva da distribuição muda de formato, quanto maior o valor de τ, mais cônica se torna a curva de distribuição.

Na Figura 7 (c), fixamos os valores de $\tau = 2,4$ e $\alpha = 1,5$, escolhendo alguns valores para o parâmetro m. Podemos notar que, à medida que o valor de m aumenta, a distribuição da curva se desloca para o lado direito, confirmando que m é um parâmetro de localização.

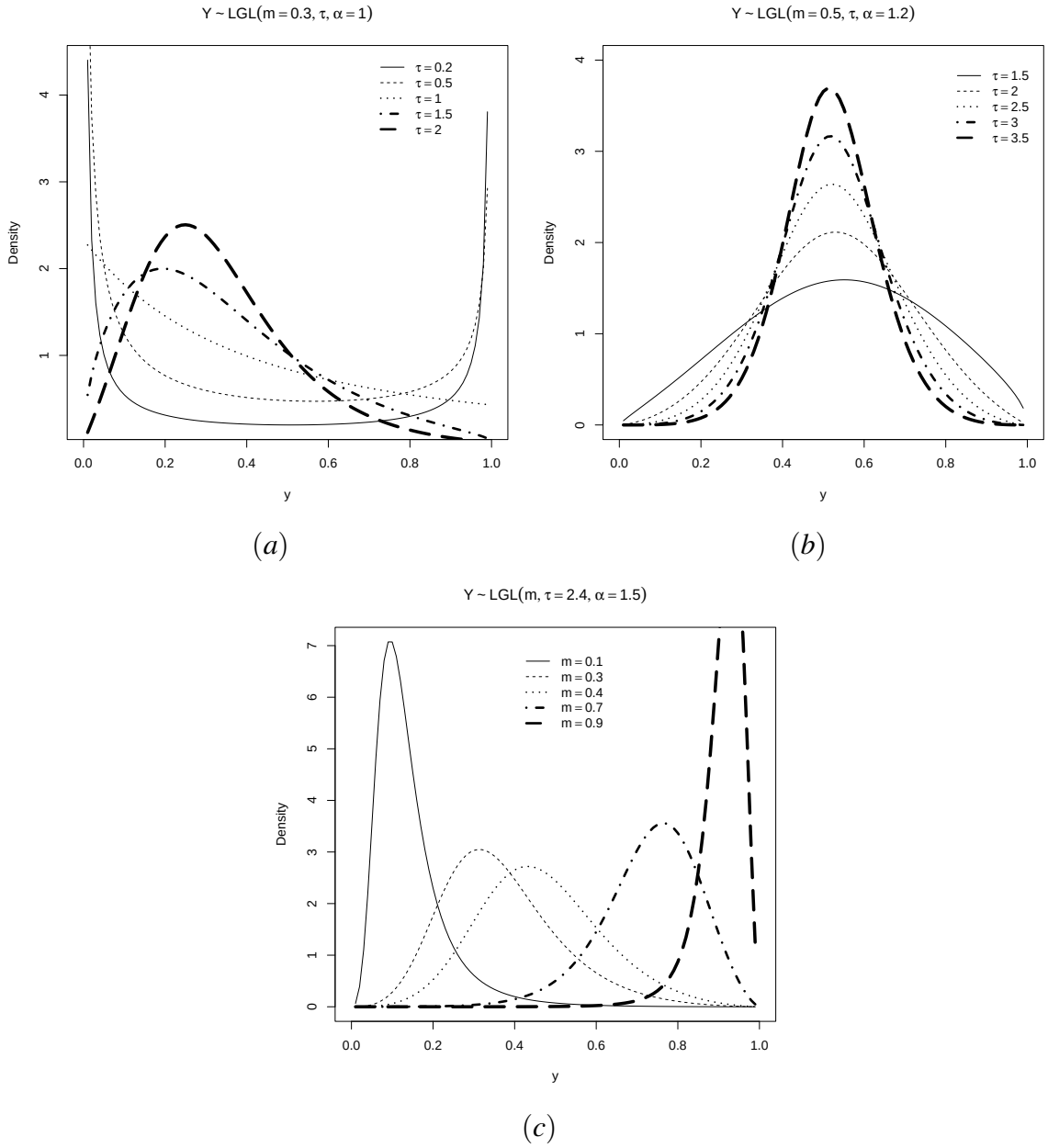


Figura 7 – Fdp de distribuição LG-Logistic para algumas escolhas dos parâmetros m, τ e α.

3.1.1 Medidas de assimetria e curtose

Esta subseção explora a assimetria e a curtose da distribuição LG-Logistic. Como esta distribuição possui uma expressão fácil para a função quantílica, escrevemos essas medidas em termos da função quantílica (3.3).

Consideramos a seguinte expressão para a assimetria (Octile skewness (OS)) da distribuição LG-Logistic, proposta por Hinkley (1975), com $p = 0,125$,

$$OS = \frac{\kappa(1-p) + \kappa(p) - 2 \times m}{\kappa(1-p) - \kappa(p)} = \frac{\kappa(0,875) + \kappa(0,125) - 2 \times \kappa(0,500)}{\kappa(0,875) - \kappa(0,125)}. \tag{3.4}$$

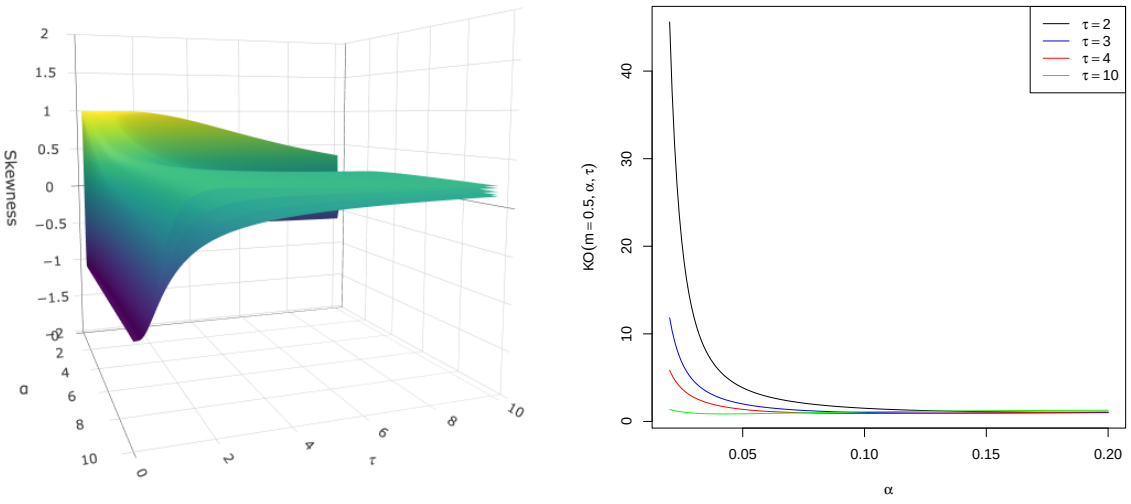
Quando OS estiver próximo de 1, a distribuição tende a ter uma assimetria extrema à direita; se estiver próximo de -1, a distribuição tende a ter uma assimetria extrema à esquerda. Quando OS é igual a 0, então a distribuição é simétrica.

Para à curtose (Octile kurtosis (OK)), usamos o seguinte índice não negativo proposto por Moors (1988)

$$OK = \frac{(\kappa(0,875) - \kappa(0,625)) + (\kappa(0,375) - \kappa(0,125))}{\kappa(0,750) - \kappa(0,250)}, \tag{3.5}$$

em que $0 < OK < \infty$. Observe que se os dois termos no numerador forem grandes (pequenos), relativamente pouca (muita) massa de probabilidade está concentrada na vizinhança de $\kappa(0,750)$ e $\kappa(0,250)$ (MOORS, 1988).

A Figura 8 apresenta o comportamento da assimetria (a) e da curtose (b) apresentadas em (3.4) e (3.5), considerando $m = 0,5$. Note que OS é uma função decrescente de m, pois podemos provar que esta pode ser escrita como a função que tem uma derivada negativa em relação a m. É por isso que podemos ver que as superfícies para diferentes m não se cruzam.



(a) (b)

Figura 8 – (a) OS considerando alguns valores α e τ, e (b) OK para diferentes valores de τ - ambos para $m = 0.5$

Em relação a OK, esta é extremamente plana, exceto quando α e τ se aproximam de 0. Neste caso estamos propondo apenas um gráfico bivariado.

Alguns valores numéricos de OS e OK são apresentados no Apêndice A.1.3, dados alguns valores de m, α e τ. Um resultado muito particular é quando $m = 0,5$ e $\alpha = 1$, o que produz uma distribuição simétrica. No entanto, esta distribuição é muito flexível e acomoda diferentes tipos de assimetrias e curtoses.

Para melhorar a interpretação, propomos na próxima seção uma nova parametrização do modelo.

3.1.2 Reparametrização da distribuição LG-Logistic

Nesta seção propomos uma parametrização conveniente da distribuição LG-Logistic em termos de seus quantis que podem ser usados para adicionar covariáveis em um modelo de regressão.

Dada uma probabilidade fixa $p \in (0,1)$ consideremos, em termos do quantil $0 < \kappa(p) < 1$, a próxima reparametrização da distribuição LG-Logistic

$$\phi = p^{\frac{1}{\alpha}} \quad \text{e} \quad \varphi = \left(\frac{1-\phi}{\phi}\right)^{\frac{1}{\tau}}. \tag{3.6}$$

Assim, fixado o valor de p, introduzimos um novo parâmetro ϕ relacionado com o parâmetro α e um novo parâmetro φ relacionado com o parâmetro τ da parametrização anterior.

Para garantir a boa definição, imporemos a restrição $\phi \in (0, 0.5)$ no novo espaço paramétrico e consequentemente $\varphi > 1$. Essa restrição implica que $p < \frac{1}{2^\alpha}$, no entanto, isso não restringe o nível de quantil, uma vez que $p \in (0,1)$ é dado. O que isso restringe é o intervalo do parâmetro de forma na parametrização inicial para $\alpha < \alpha^* = -\frac{\log(p)}{\log(2)}$. Quanto maior o valor de p (mas menor que 1), mais estreito é o intervalo de α.

Observe em (3.3), que sob esta parametrização o quantil depende de φ e m, ou seja $\kappa = \frac{1}{1-\varphi+\frac{\varphi}{m}}$ e a fdp e o fda da distribuição LG-Logistic acabam sendo, respectivamente

$$\begin{aligned} f(y|\kappa,\phi,\varphi) &= \frac{\log(p)\log\left(\frac{1-\phi}{\phi}\right)}{\log(\phi)\log(\varphi)} \left[1+\left[\left(\frac{\kappa\varphi}{1-\kappa}\right)\left(\frac{1-y}{y}\right)\right]^{\frac{\log\left(\frac{1-\phi}{\phi}\right)}{\log(\varphi)}}\right]^{-\frac{\log(p)+\log(\phi)}{\log(\phi)}} \\ &\quad \times \left[\left(\frac{\kappa\varphi}{1-\kappa}\right)\left(\frac{1-y}{y}\right)\right]^{\frac{\log\left(\frac{1-\phi}{\phi}\right)}{\log(\varphi)}} \frac{1}{y(1-y)} \end{aligned} \tag{3.7}$$

e

$$F(y \mid \kappa,\phi,\varphi) = \left[1+\left[\left(\frac{\kappa\varphi}{1-\kappa}\right)\left(\frac{1-y}{y}\right)\right]^{\frac{\log\left(\frac{1-\phi}{\phi}\right)}{\log(\varphi)}}\right]^{-\frac{\log(p)}{\log(\phi)}}. \tag{3.8}$$

Observe que esta distribuição depende agora da probabilidade fixa p usada para definir o quantil $\kappa(p)$, e então é uma distribuição quantílica paramétrica.

Considerando os dois casos particulares já discutidos da distribuição LG-Logistic, notamos que quando $\phi = p \in (0,0,5)$, in (3.8), obtemos a distribuição L-Logistic, cuja a notação

será agora $Y \sim \mathrm{LL}(\kappa, \varphi)$. Por outro lado, se $\kappa = \frac{1}{1+\varphi}$, obtemos a distribuição G-Logistic, apresentada anteriormente, cuja notação será agora dada por $Y \sim \mathrm{GL}(\phi, \varphi)$. No Apêndice A.1.2 são apresentados as fdp destes casos especiais.

Nas Figuras 9, 10 e 11 apresentamos algumas formas de (3.7) considerando $p = 0,5$.

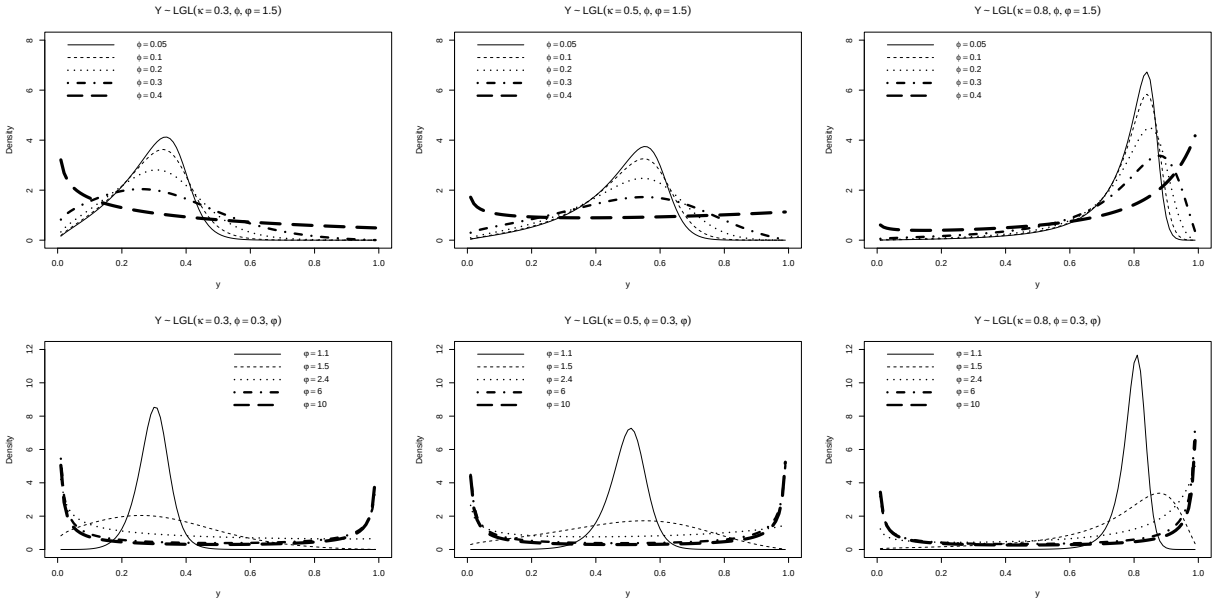


Figura 9 – Fdp da distribuição LG-Logistic para algumas escolhas dos parâmetros κ, ϕ e φ.

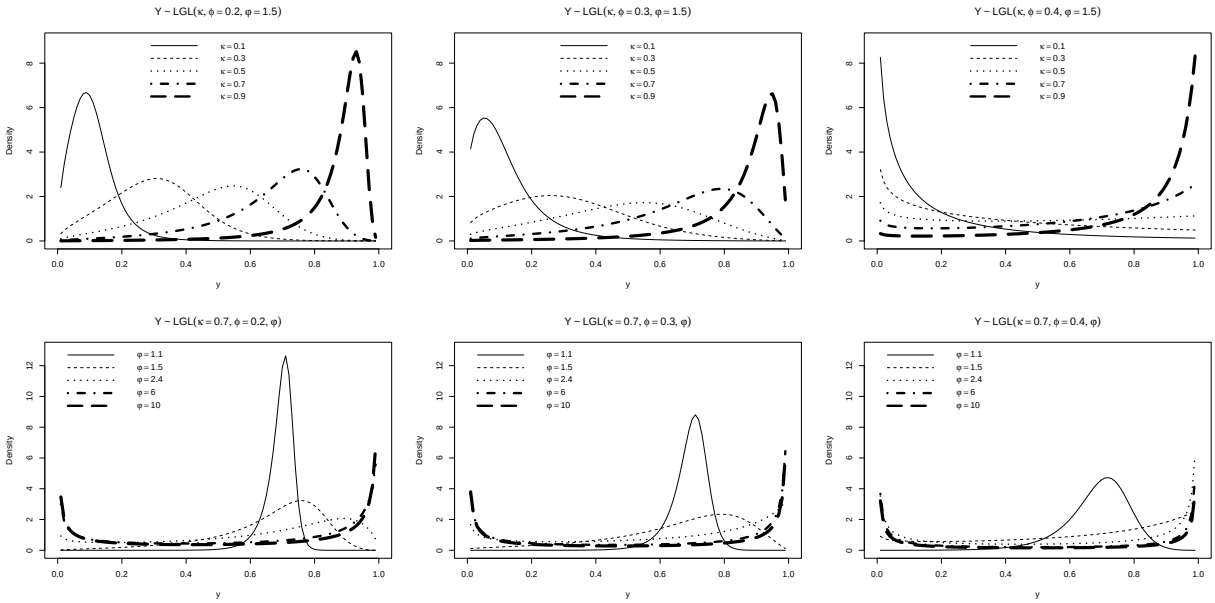


Figura 10 – Fdp da distribuição LG-Logistic para algumas escolhas dos parâmetros κ, ϕ e φ.

Observe que na Figura 9 apresentamos alguns gráficos da influência do parâmetro κ em ϕ e φ. Consideramos três valores para κ $(0,3, 0,5, 0,8)$ e, inicialmente, definimos φ como

1,5, atribuindo alguns valores a ϕ (0,05, 0,1, 0,2, 0,3 e 0,4). Podemos ver que κ influencia ϕ na localização. O mesmo acontece quando definimos ϕ como 0,3 e variamos φ (1,1, 1,5, 2,4, 6, 10), notamos que κ influencia φ na localização.

Na Figura 10, mostramos alguns gráficos da influência do parâmetro ϕ em κ e φ. Inicialmente, definimos φ como 1,5 e consideramos alguns valores para κ (0,1; 0,3; 0,5; 0,7; 0,9). Em seguida, definimos κ em 0,7 e variamos φ (1,1; 1,5; 2,4; 6; 10). Em ambos os casos usamos $\phi = 0,2, 0,3$ e 0,4. Podemos ver que ϕ interfere na escala de κ e φ.

Na Figura 11, apresentamos alguns gráficos que mostram a influência do parâmetro φ em κ e ϕ. Para isso, utilizamos alguns valores para φ (1,5, 2,4, 6). Inicialmente definimos ϕ como 0,3 e atribuímos alguns valores a κ (0,1, 0,3, 0,5, 0,7 e 0,9) e posteriormente definimos κ como 0,7 e atribuímos alguns valores a ϕ (0,1, 0,2, 0,3, 0,4, 0,5). A mudança na forma da função de distribuição é perceptível quando verificamos a influência do parâmetro φ sobre κ e ϕ.

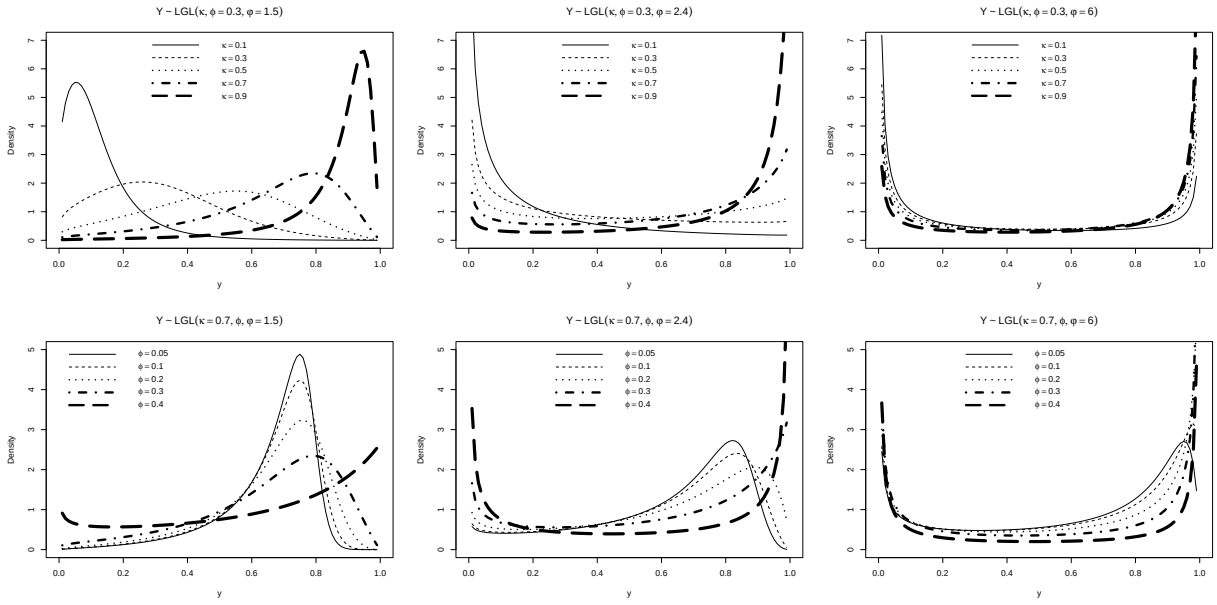


Figura 11 – Fdp da distribuição LG-Logistic para algumas escolhas dos parâmetros κ, ϕ e φ.

De agora em diante, usaremos para uma variável aleatória Y com distribuição LG-Logistic a notação $Y \sim \text{LGL}(\kappa(p), \phi, \varphi)$, sendo o parâmetro quantílico $\kappa \in (0,1)$ um parâmetro de localização, $0 < \phi < 0,5$ um parâmetro de forma, $\varphi > 1$ um parâmetro de escala e p é assumido como fixo de acordo com o percentil de interesse.

3.1.2.1 Moda

Usando a equação (3.1), determinaremos uma expressão para a moda da distribuição LG-Logistic. Tomando a derivada de (3.1) em relação a y e igualando-a a zero, ou seja $\frac{\partial}{\partial y} f(y) = 0$, temos

$$\left(\frac{1-m}{m}\right)^{\tau} = \left(\frac{1-y}{y}\right)^{\tau}\left(\frac{\alpha\tau-1+2y}{\tau+1-2y}\right). \tag{3.9}$$

Como $u = \frac{m}{1-m}$ e $\beta = \frac{1}{\tau}$, a expressão em (3.9) pode ser reescrita como segue

$$\left(\frac{1}{u}\right)^{\frac{1}{\beta}} = \left(\frac{1-y}{y}\right)^{\frac{1}{\beta}}\left(\frac{\frac{\alpha}{\beta}-1+2y}{\frac{1}{\beta}+1-2y}\right).$$

Usando a reparameterização em (3.6), temos que

$$\left(\frac{1-\kappa}{\kappa\varphi}\right)^{\frac{\log\left(\frac{1-\phi}{\phi}\right)}{\log(\varphi)}} = \left(\frac{1-y}{y}\right)^{\frac{\log\left(\frac{1-\phi}{\phi}\right)}{\log(\varphi)}}\left(\frac{\frac{\log(p)\log\left(\frac{1-\phi}{\phi}\right)}{\log(\phi)\log(\varphi)}-1+2y}{\frac{\log\left(\frac{1-\phi}{\phi}\right)}{\log(\varphi)}+1-2y}\right). \tag{3.10}$$

A moda da distribuição LG-Logistic é obtida resolvendo a equação (3.10). Desenvolvemos um código, que está disponível no Apêndice A.3.1, que foi implementado utilizando Visual Studio Code e escrito na linguagem de programação Python para determinar um valor numérico da moda, fixando valores de parâmetros.

3.1.2.2 Momentos

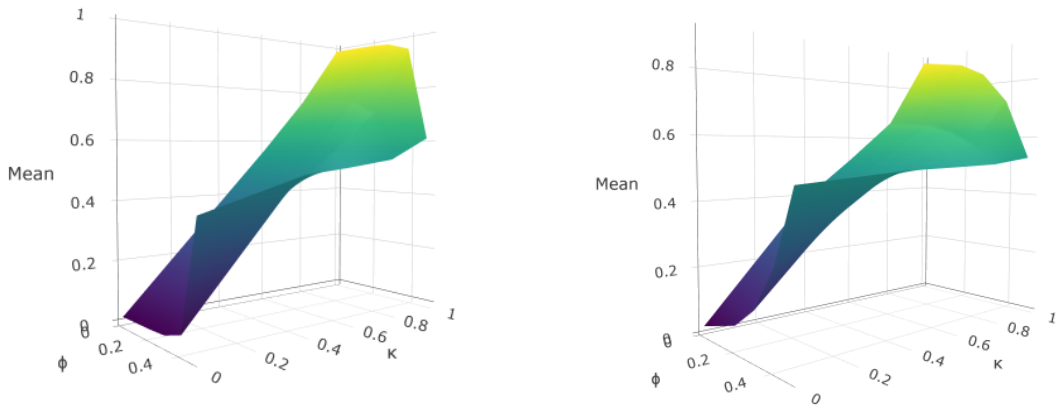
A seguinte expressão de momento da distribuição LG-Logistic é apresentada em (3.11). Mais detalhes sobre sua construção podem ser vistos no Apêndice A.1.4.

$$\mathbb{E}[Y^t] = \int_0^1 \left(1+\left(\frac{1-\kappa}{\kappa\varphi}\right)\left(\frac{1-v}{v}\right)^{\frac{\log(\varphi)}{\log\left(\frac{1-\phi}{\phi}\right)}}\right)^{-t} \frac{\log(p)}{\log(\phi)} v^{\left(\frac{\log(p)}{\log(\phi)}-1\right)} dv. \tag{3.11}$$

A integral (3.11) não possui solução analítica e requer métodos numéricos para sua avaliação. A Figura 12 e a Figura 13 apresentam o gráfico da média e da variância, respectivamente, da distribuição LG-Logística em função dos parâmetros κ e ϕ, considerando algumas escolhas do parâmetro φ.

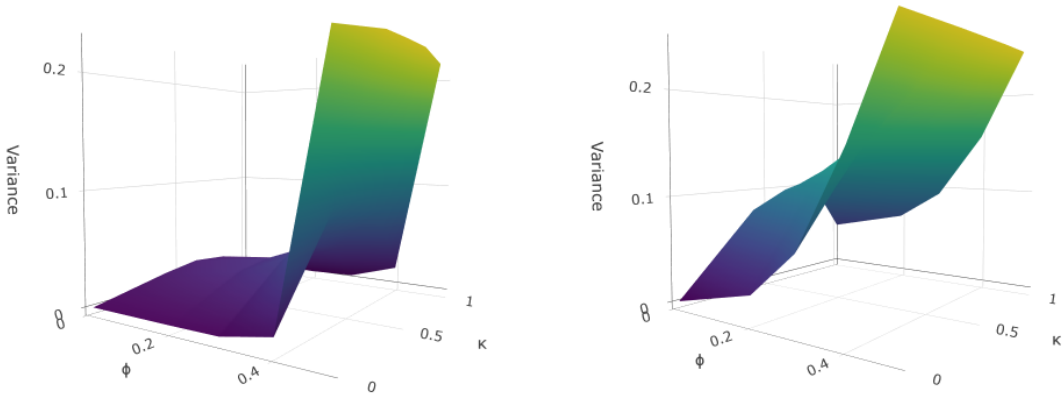
Os gráficos em (a) e (b), na Figura 12, mostram que conforme o valor de κ aumenta, o valor da média aumenta consideravelmente, haja vista $\phi \in (0, 0.3)$. Já para $\phi \in (0.3, 0.5)$, o valor da média cresce, mas mais devagar, ao passo que o valor de κ aumenta.

Nos gráficos (c) e (d) da Figura 13, podemos observar que o valor da variância aumenta à medida que ϕ cresce, para quaisquer valores de κ. Em relação aos valores de κ, no gráfico (c), a variância aumenta para $\kappa \in (0, 0.6)$ e diminui para $\kappa \in (0.6, 1.0)$. No gráfico (d), a variância aumenta quando $\kappa \in (0, 0.8)$ e diminui quando $\kappa \in (0.8, 1.0)$.



(a) (b)

Figura 12 – Média de Y ($\mathbb{E}[Y]$) da distribuição LG-Logistic para algumas escolhas de κ e ϕ, considerando $p = 0,5$, (a) $\varphi = 1,5$ e (b) $\varphi = 5,0$



(c) (d)

Figura 13 – Variância de Y ($\mathbb{V}\text{ar}[Y]$) da distribuição LG-Logistic para algumas escolhas de κ e ϕ, considerando $p = 0,5$, (c) $\varphi = 1,5$ e (d) $\varphi = 5,0$

Alguns valores numéricos de $\mathbb{E}[Y]$ e $\mathbb{V}\text{ar}[Y]$ são apresentados no Apêndice A.1.5, dados alguns valores de κ, ϕ e φ.

3.2 O modelo de regressão quantílico LG-Logistic

Nesta seção propomos o novo modelo de regressão quantílico LG-Logistic.

Seja $\boldsymbol{Y} = (Y_1, \ldots Y_n)$ um vetor $n \times 1$ de variáveis de resposta independentes com valores no intervalo $(0,1)$. Considerando a distribuição apresentada na seção 3.1.2, construímos o seguinte modelo de regressão

$$\begin{aligned} y_i &\sim \text{LGL}(\kappa_i(p), \phi, \varphi_i), \\ g_1(\kappa_i) &= X_{1i}^T \boldsymbol{\beta}, \\ g_2(\varphi_i) &= X_{2i}^T \boldsymbol{\delta_1}, \\ g_3(\phi_i) &= X_{3i}^T \boldsymbol{\delta_2}, \end{aligned} \tag{3.12}$$

onde X_{1i}^T, X_{2i}^T e X_{3i}^T, são vetores de covariáveis associados aos parâmetros de localização, escala e forma, respectivamente, $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_q)^T$, $\boldsymbol{\delta_1} = (\delta_{10}, \cdots, \delta_{1d})^T$ e $\boldsymbol{\delta_2} = (\delta_{20}, \cdots, \delta_{2d})^T$ são os vetores dos coeficientes de regressão, incluindo os interceptos β_0, δ_{10} e δ_{20} e g_1, g_2, e g_3 são funções de ligação reais estritamente monótonas e duas vezes diferenciáveis.

Para simplificar, adotaremos uma função de ligação logit para g_1, mas outras funções como probit, log-log, entre outras podem ser utilizadas (FERRARI; CRIBARI-NETO, 2004). Desta forma, temos

$$g_1(\kappa_i) = \log\left(\frac{\kappa_i}{1-\kappa_i}\right) = X_{1i}^T \boldsymbol{\beta} \;\text{ or }\; \kappa_i = \frac{\exp(X_{1i}^T \boldsymbol{\beta})}{1+\exp(X_{1i}^T \boldsymbol{\beta})}. \tag{3.13}$$

Para g_2 e g_3 adotaremos uma função de ligação *log* (BAYES; BAZÁN; CASTRO, 2017), ou seja,

$$g_2(\varphi_i) = \log(\varphi_i) = X_{2i}^T \boldsymbol{\delta_1} \;\text{ ou }\; \varphi_i = \exp(X_{2i}^T \boldsymbol{\delta_1}). \tag{3.14}$$

e

$$g_3(\phi_i) = \log(\phi_i) = X_{3i}^T \boldsymbol{\delta_1} \;\text{ ou }\; \varphi_i = \exp(X_{3i}^T \boldsymbol{\delta_2}). \tag{3.15}$$

3.3 Inferência

Apresentamos nesta seção o processo de estimação Bayesiano para a distribuição LG-Logistic e seu correspondente modelo de regressão quantílica. Para este fim, seja $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ um conjunto de observações independente da função de densidade em (3.7).

3.3.1 Estimativa bayesiana da distribuição LG-Logistic

A função de verossimilhança é dada por

$$\begin{aligned} L(\kappa,\phi,\varphi \mid \boldsymbol{Y}) &= \prod_{i=1}^{n} \frac{\log(p)\log\left(\frac{1-\phi}{\phi}\right)}{\log(\phi)\log(\varphi)} \times \\ &\times \left[1+\left[\left(\frac{\kappa\varphi}{1-\kappa}\right)\left(\frac{1-y_i}{y_i}\right)\right]^{\frac{\log\left(\frac{1-\phi}{\phi}\right)}{\log(\varphi)}}\right]^{-\frac{\log(p)+\log(\phi)}{\log(\phi)}} \times \\ &\times \left[\left(\frac{\kappa\varphi}{1-\kappa}\right)\left(\frac{1-y_i}{y_i}\right)\right]^{\frac{\log\left(\frac{1-\phi}{\phi}\right)}{\log(\varphi)}} \frac{1}{y_i(1-y_i)}. \end{aligned} \tag{3.16}$$

Para completar as especificações bayesianas do modelo, vamos assumir *prioris* independentes da seguinte forma

$$p(\kappa,\phi,\varphi) = p(\kappa)p(\phi)p(\varphi), \tag{3.17}$$

onde $p(\kappa)$ denota a *priori* para o parâmetro κ, $p(\phi)$ a *priori* para o parâmetro ϕ e $p(\varphi)$ a *priori* para o parâmetro φ. Mais precisamente, propomos que $\kappa \sim \text{Beta}(0,75, 0,75)$, que dá uma média de 0,5 e uma variância de 0,1, e $\phi \sim \text{U}(0,0,49)$, o que dá uma média de $0,2$ e uma variação de $0,02$. Por outro lado, um estudo de sensibilidade será realizado para a *priori* φ na seção 3.4.1, propondo duas distribuições de *prioris*.

De (3.16) e (3.17), temos a expressão para a distribuição *posteriori* dos parâmetros

$$p(\kappa,\phi,\varphi \mid \boldsymbol{Y}) \propto L(\kappa,\phi,\varphi \mid \boldsymbol{Y})p(\kappa)p(\phi)p(\varphi). \tag{3.18}$$

3.3.2 Estimativa bayesiana do modelo de regressão LG-Logistic

Na configuração de regressão, a função de verossimilhança é dada por

$$\begin{aligned} L(\boldsymbol{\beta},\boldsymbol{\delta_1},\boldsymbol{\delta_2} \mid \boldsymbol{Y},\boldsymbol{X}) &= \prod_{i=1}^{n} \frac{\log(p)\log\left(\frac{1-\phi_i}{\phi_i}\right)}{\log(\phi_i)\log(\varphi_i)} \times \\ &\times \left[1+\left[\left(\frac{\kappa_i\varphi_i}{1-\kappa_i}\right)\left(\frac{1-y_i}{y_i}\right)\right]^{\frac{\log\left(\frac{1-\phi_i}{\phi_i}\right)}{\log(\varphi_i)}}\right]^{-\frac{\log(p)+\log(\phi_i)}{\log(\phi_i)}} \times \\ &\times \left[\left(\frac{\kappa_i\varphi_i}{1-\kappa_i}\right)\left(\frac{1-y_i}{y_i}\right)\right]^{\frac{\log\left(\frac{1-\phi_i}{\phi_i}\right)}{\log(\varphi_i)}} \frac{1}{y_i(1-y_i)}, \end{aligned} \tag{3.19}$$

onde $\kappa_i = \frac{\exp(X_{1i}^T\boldsymbol{\beta})}{1+\exp(X_{1i}^T\boldsymbol{\beta})}$, $\varphi_i = \exp(X_{2i}^T\boldsymbol{\delta_1})$, $\phi_i = \exp(X_{3i}^T\boldsymbol{\delta_2})$ e X_{1i}, X_{2i}, e X_{3i} são as covariáveis apresentadas na seção 3.2.

Nossa especificação de *priori* será

$$p(\boldsymbol{\beta}, \boldsymbol{\delta_1}, \boldsymbol{\delta_2}) = p(\boldsymbol{\beta})p(\boldsymbol{\delta_1})p(\boldsymbol{\delta_2}), \tag{3.20}$$

onde $p(\boldsymbol{\beta})$ é a *priori* de $\boldsymbol{\beta}$, $p(\boldsymbol{\delta_1})$ é a *priori* de $\boldsymbol{\delta_1}$ e $p(\boldsymbol{\delta_2})$ é a *priori* de $\boldsymbol{\delta_2}$. Neste trabalho, propomos $\beta_j \sim \mathrm{N}(0,100)$, $j = 1,\cdots,q$, $\delta_{i1} \sim \mathrm{N}(0,100)$, $i = 1,\cdots,d_1$, $\delta_{i2} \sim \mathrm{N}(0,100)$, $i = 1,\cdots,d_2$, que são escolhas comuns para os coeficientes de regressão. Se apenas ϕ for considerado, propomos $\phi \sim \mathrm{U}(0,0.49)$, pois não temos informações prévias sobre estes parâmetros.

Por (3.19) e (3.20) podemos expressar a distribuição posterior por

$$p(\boldsymbol{\beta}, \boldsymbol{\delta_1}, \boldsymbol{\delta_2} \mid \boldsymbol{Y}) \propto L(\boldsymbol{\beta}, \boldsymbol{\delta_1}, \boldsymbol{\delta_2} \mid \boldsymbol{Y}, \boldsymbol{X})p(\boldsymbol{\beta})p(\boldsymbol{\delta_1})p(\boldsymbol{\delta_2}). \tag{3.21}$$

Todos os procedimentos de estimação serão implementados no Stan através do pacote `rstan` (Stan Development Team, 2020), e o algoritmo NUTS, uma variante do Hamiltoniano Monte Carlo (HMC). Os critérios de comparação utilizados neste capítulo são os mesmos que estão apresentados na seção 2.2.3.

3.4 Estudos de simulação

Nesta seção, apresentamos os estudos de simulação desenvolvidos para verificar a sensibilidade a *priori* dos parâmetros φ e para verificar a recuperação dos parâmetros na distribuição LG-Logistic proposta.

3.4.1 Análise de sensibilidade da priori

Segundo Paz *et al.* (2019), uma análise de sensibilidade prévia é um passo muito importante na análise Bayesiana, quando se trata do desenvolvimento de novos modelos.

Realizamos este estudo para duas escolhas de *priori* para o parâmetro φ. Como $\varphi > 1$, procuramos distribuições com suporte positivo e realizamos um truncamento atribuindo um limite inferior 1 à distribuição. Foram consideradas duas distribuições: $\varphi \sim \mathrm{N}(1,5,100)$ e $\varphi \sim \mathrm{Gamma}(1.5,1)$. No último caso, espera-se que $\mathbb{E}[\varphi] = \mathbb{V}\mathrm{ar}[\varphi] = 1,5$.

Amostras de tamanhos 20, 50 e 100 foram geradas a partir da distribuição LG-Logistic com parâmetros $\kappa = 0,7$, $\phi = 0,2$, $\varphi = 1,5$ e $p = 0,5$. Mantivemos para κ e ϕ as distribuições *prioris* apresentadas na seção 3.3.1.

O modelo foi implementado em Stan, utilizando o `rstan` (Stan Development Team, 2020) como interface, que fornece inferência bayesiana completa usando o algoritmo No-U-Turn (NUTS). Para obter convergência usamos 2 cadeias e consideramos 10000 iterações. Para comparar esses modelos, utilizamos os critérios apresentados na seção 2.2.3.

A Tabela 16 apresenta os diferentes critérios de comparação de modelos sob as duas distribuições para φ, para tamanhos de amostra $n = 20, 50$ e 100.

Tabela 16 – Critério de comparação da análise de sensibilidade da *priori* da distribuição LG-Logistic

	$n = 20$		$n = 50$		$n = 100$	
	Normal Truncada	Gamma Truncada	Normal Truncada	Gamma Truncada	Normal Truncada	Gamma Truncada
DIC	-15,040	**-15,312**	-47,112	**-47,365**	-109,659	**-109,765**
EAIC	-10,706	**-11,145**	-43,295	**-43,536**	-106,251	**-106,336**
EBIC	-7,719	**-8,158**	-37,559	**-37,800**	-98,436	**-98,521**
IC	-13,374	**-13,478**	-44,929	**-45,195**	-107,067	**-107,193**
HQIC	-10,123	**-10,562**	-41,110	**-41,351**	-103,088	**-103,173**
WAIC	-14,484	**-14,997**	-47,125	**-47,423**	-109,559	**-109,641**
LOO	-14,417	**-14,902**	-47,117	**-47,411**	-109,537	**-109,612**
Tempo (secs)	2,720	2,540	6,280	6,020	12,120	13,000

Os valores dos critérios de comparação nos dão resultados semelhantes, indicando que não há sensibilidade na escolha a priori do parâmetro φ, ou seja, encontramos evidências de que os resultados não são influenciados pela escolha da *priori*. Observe que o tempo de estimativa também é semelhante. Mesmo assim, escolhemos a distribuição Gamma Truncda (TG) como a *priori* para φ, uma vez que produz valores de critério um pouco melhores que a *priori* Normal Truncada (TN).

3.4.2 Estudo de recuperação dos parâmetros

Nesta seção realizamos um estudo de simulação para verificar a recuperação dos parâmetros na distribuição LG-Logistic. Para tanto, geramos 100 réplicas para esta distribuição em amostras de tamanho $20, 50, 100$ e 500. Usamos duas cadeias e com iterações de 10000.

Para verificar se nosso procedimento bayesiano recupera os parâmetros do modelo, consideramos a média (Média), desvio padrão (DP), viés (Viés) e raiz quadrada do erro quadrático médio (RMSE) das estimativas. O viés simulado é calculado a partir da diferença entre a média das estimativas e o valor verdadeiro do parâmetro, ou seja, é definido por $b(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$, com θ sendo o vetor de parâmetros e $\hat{\theta}$ a estimativa de θ. Os cálculos do valor RMSE e o desvio padrão das estimativas dos parâmetros do modelo são apresentados na seção 2.3.

Fixamos os valores da distribuição LG-Logistic em $\kappa = 0,7$, $\phi = 0,2$ e $\varphi = 1,5$, na qual $\kappa \sim \text{Beta}(0,75, 0,75)$, $\phi \sim \text{U}(0, 0,49)$ e φ seguem uma distribuição gama truncada, mais precisamente $\varphi \sim \text{Gamma}(1,5,1)$. Todo o desenvolvimento foi realizado para $p = 0,5$.

Tabela 17 – Comparação do estudo de recuperação da distribuição LG-Logistic considerando $p = 0,5$ - 100 réplicas.

n	Par	Real	Média	DP	Viés	RMSE
20	κ	0,7	0,674	0,040	0,026	0,048
	ϕ	0,2	0,250	0,111	-0,050	0,121
	φ	1,5	1,500	0,300	-0,001	0,300
50	κ	0,7	0,675	0,024	0,025	0,034
	ϕ	0,2	0,270	0,099	-0,070	0,121
	φ	1,5	1,396	0,198	0,104	0,224
100	κ	0,7	0,683	0,015	0,017	0,023
	ϕ	0,2	0,307	0,081	-0,109	0,135
	φ	1,5	1,291	0,128	0,209	0,245
500	κ	0,7	0,678	0,008	0,022	0,023
	ϕ	0,2	0,136	0,035	0,064	0,073
	φ	1,5	1,694	0,084	-0,194	0,212

Os resultados da Tabela 17 revelam que os parâmetros κ e ϕ são muito bem recuperados e isso melhora conforme o tamanho da amostra aumenta. Porém, não notamos exatamente o mesmo em relação aos resultados do parâmetro φ, pois o valor do RMSE diminui muito lentamente com o aumento do tamanho da amostra.

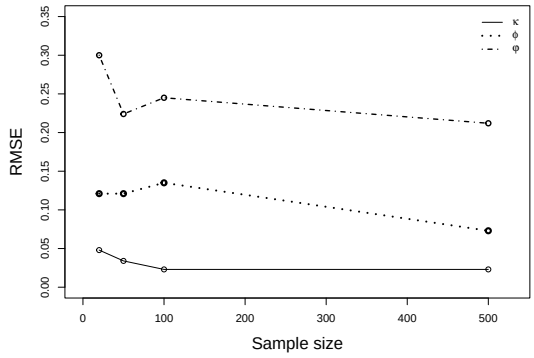


Figura 14 – Comportamento RMSE para κ, ϕ e φ nos tamanhos de amostra $20, 50, 100$ e 500 da distribuição LG-Logistic.

A Figura 14 ilustra o comportamento do RMSE na recuperação dos parâmetros κ, ϕ e φ, para tamanhos amostrais $20, 50, 100$ e 500. A observação feita anteriormente sobre φ é mais evidente, assim como os bons resultados dos parâmetros κ e ϕ.

3.5 Aplicações

Esta seção apresenta duas aplicações da distribuição LG-Logistic e sua comparação com as distribuições L-Logistic (PAZ *et al.*, 2019) e G-Logistic. Na primeira aplicação na seção

3.5.1, estimamos a distribuição da vulnerabilidade à pobreza no estado do Maranhão, Brasil, e na segunda aplicação na seção 3.5.2, usamos um conjunto de dados sobre pobreza no Peru.

3.5.1 Aplicação da distribuição LG-Logistic

Nesta subsecção, consideramos uma aplicação para modelar a Proporção de Crianças Vulneráveis à Pobreza (PCVP) (0 a 14 anos). Os dados são provenientes dos municípios do estado do Maranhão, Brasil, e foram coletados em 2010. O conjunto de dados foi elaborado usando informações disponíveis em BRASIL (2010). O estado do Maranhão está localizado na Região Nordeste do Brasil e é formado por 217 municípios.

O conjunto de dados PCVP possui 217 observações e pode ser modelado usando a distribuição LG-Logistic apresentada na seção 3.1, que chamaremos de LG-Logistic (1), ou sua reparametrização apresentada na seção 3.1.2 , que chamaremos de LG-Logistic (2). Para efeito de comparação, modelamos também a variável PCVP com as distribuições L-Logistic e G-Logistic, denominadas L-Logistic (1) e G-Logistic (1), respectivamente, bem como suas respectivas reparametrizações, denominadas L-Logistic (2) e G-Logistic (2), respectivamente.

Para a distribuição LG-Logistic (1), consideramos os seguintes *prioris*: $m \sim \mathrm{U}(0,1)$, $\alpha \sim \mathrm{Gamma}(0,001,0,001)$ e $\tau \sim \mathrm{Gamma}(0,001,0,001)$. As mesmas anteriores foram consideradas para seus casos particulares, ou seja, para a distribuição L-Logistic (1) consideramos as anteriores $m \sim \mathrm{U}(0,1)$ e $\tau \sim \mathrm{Gamma}(0,001,0,001)$ e para a distribuição G-Logistic (1) consideramos que $\alpha \sim \mathrm{Gamma}(0,001,0,001)$ e $\tau \sim \mathrm{Gamma}(0.001,0,001)$.

Para a distribuição LG-Logistic (2), consideramos as *prioris* $\kappa \sim \mathrm{Beta}(0,75,0,75)$, $\phi \sim \mathrm{U}(0,0,49)$, apresentados em seção 3.3.1, e $\varphi \sim \mathrm{Gamma}(1.5,1)$, com base nos resultados realizados na seção 3.4.1. As mesmas *priores* foram consideradas para L-Logistic (2) e G-Logistic (2), ou seja, $\kappa \sim \mathrm{Beta}(0,75,0,75)$ e $\varphi \sim \mathrm{Gamma}(1,5,1)$ e $\phi \sim \mathrm{U}(0,0,49)$ e $\varphi \sim \mathrm{Gamma}(1,5,1)$, respectivamente.

Todos os desenvolvimentos foram realizados usando 2 cadeias e 10000 iterações, considerando $p = 0,5$. A mesma aplicação foi realizada utilizando a estrutura dos modelos propostos acima, considerando outros valores para p, cujos resultados estão apresentados no apêndice A.2.1. Os resultados da aplicação dos modelos de acordo com os critérios de comparação da seção 2.2.3 são apresentados na Tabela 18.

Como visto, os critérios de comparação de modelos para LG-Logistic (1) e LG-Logistic (2), L-Logistic (1) e L-Logistic (2), G-Logistic (1) e G-Logistic (2) são semelhantes, indicando que ambas as parametrizações podem ser utilizadas. Ainda, através dos resultados apresentados na Tabela 18, podemos observar que os valores dos critérios de comparação são menores para as distribuições LG-Logistic e G-Logistic reparametrizadas, indicando que ambos os modelos se ajustam melhor à proporção de crianças vulneráveis a pobreza, em comparação com os resultados obtidos pelo ajuste da distribuição L-Logistic e também em comparação com a mesma

distribuição com a parametrização inicial.

Tabela 18 – Resultado dos critérios de comparação dos dados da pobreza do Brasil, considerando $p = 0,5$

	L-Logistic		G-Logistic		LG-Logistic	
	(1)	(2)	(1)	(2)	(1)	(2)
DIC	-254,368	-254,431	-264,220	-264,323	-264,585	**-264,619**
EAIC	-252,380	-252,398	-262,174	**-262,212**	-261,409	-261,461
EBIC	-245,620	-245,638	-255,415	**-255,452**	-251,269	-251,321
IC	-252,356	-252,463	-262,266	**-262,434**	-261,762	-261,777
HQIC	-249,649	-249,667	-259,444	**-259,481**	-257,313	-257,365
WAIC	-254,306	-254,366	-264,505	-264,596	-264,821	**-264,933**
LOO	-254,304	-254,363	-264,503	-264,594	-264,815	**-264,930**

Ao considerar os resultados do Apêndice A.2.1, percebemos que a distribuição LG-Logistic com $p = 0,10$ apresenta valores inferiores aos mostrados na Tabela 18 e poderia ser um modelo alternativo para os dados. Porém preferimos manter os resultados encontrados para $p = 0,50$ por ser mais fácil de interpretar.

Na Tabela 19, apresentamos as estimativas dos parâmetros, bem como outras informações resumidas para a distribuição LG-Logistic (2) e para seus dois casos particulares, as distribuições L-Logistic (2) e G-Logistic (2). De acordo com Gelman e Rubin (1992), o valor da estatística $\hat{R}$ é definido pela razão entre o tamanho efetivo da amostra e o tamanho total da amostra, e a razão entre o erro padrão de Monte Carlo e o desvio padrão posterior para os parâmetros estimados , que diminui para 1 como $n \to \infty$.

Tabela 19 – Resultados de informações resumidas da estimação de parâmetros considerando $p = 0,5$.

LL(κ, φ)					
Par	Média	DP*	2,5%	97,5%	$\hat{R}$**
κ	0,387	0,010	0,368	0,406	1,000
φ	1,014	0,001	1,013	1,016	1,000
GL(ϕ, φ)					
Par	Média	DP*	2,5%	97,5%	$\hat{R}$**
ϕ	0,147	0,029	0,096	0,208	1,000
φ	1,454	0,047	1,365	1,555	1,001
LGL(κ, ϕ, φ)					
Par	Média	DP*	2,5%	97,5%	$\hat{R}$**
κ	0,398	0,009	0,380	0,419	1,000
ϕ	0,222	0,062	0,108	0,352	1,000
φ	1,352	0,084	1,187	1,520	1,000

*Desvio padrão; ** Estatística $\hat{R}$

Em cada um dos modelos foram atribuídos *prioris* aos parâmetros de acordo com as especificações apresentadas anteriormente, e os resultados mostram que, em todos os casos,

$\hat{R}$ é igual ou muito próximo de 1, indicando convergência das cadeias MCMC aos anteriores utilizados.

Considerando que o G-Logistic (2) e o LG-Logistic (2) são os melhores modelos, segundo os critérios de comparação, escolhemos o modelo G-Logistic (2) por parcimônia. Então o modelo final é $\hat{y}_i \sim \text{GL}(\hat{\phi} = 0,147, \hat{\varphi} = 1,454)$.

Por fim, realizamos uma análise residual para verificar o ajuste dos modelos G-Logistic (2) e LG-Logistic (2) ao conjunto de dados. Para isso, consideramos resíduos quantílicos (DUNN; SMYTH, 1996), que tende a ser normalmente distribuído se o modelo tiver um ajuste bom. Seja $\text{F}(y, \boldsymbol{\theta})$ a função cumulativa de Y. Se F for contínua, então $\text{F}(y, \boldsymbol{\theta})$ é uniformemente distribuída no intervalo unitário.

O resíduos quantílicos pode ser definido como

$$r_{q,i} = \Phi^{-1}(\text{F}(y_i, \hat{\boldsymbol{\theta}})), \tag{3.22}$$

onde $\Phi()$ é a função cumulativa da distribuição Normal Padrão e $\hat{\boldsymbol{\theta}}$ é o vetor de parâmetros estimados da distribuição. A partir de (3.22) calculamos os resíduos quantílicos aleatórios para os dois modelos, cujo resumo é apresentado na Tabela 20.

Tabela 20 – Resultados da análise residual dos modelos G-Logistic (2) e LG-Logistic (2) para $p = 0,5$

	Min.	1^o Qu.	Mediana	Média	3^o Qu.	Max.
G-Logistic (2)	-2,362	-0,680	-0,023	-0,025	0,538	2,255
LG-Logistic (2)	-2,503	-0,668	0,042	-0,002	0,609	2,182

Observe que os resíduos quantílicos calculados para o modelo G-Logístico (2) são, em média, iguais a $-0,025$, sendo o valor mínimo igual a $-2,362$ e o valor máximo igual a $2,255$. Os resíduos quantílicos calculados para o modelo LG-Logistic (2) são inferiores aos do G-Logistic (2), sendo em média iguais a $-0,002$, com valor mínimo igual a $-2,503$ e valor máximo igual a 2.182.

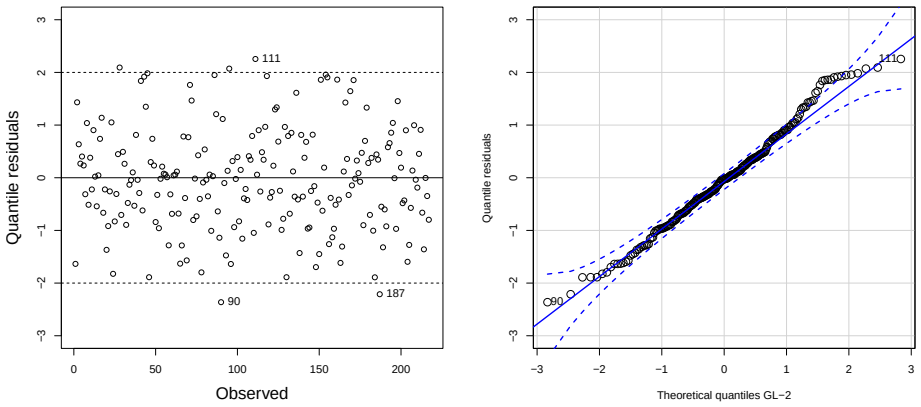


Figura 15 – Gráfico e envelope dos resíduos quantílicos do modelo G-Logístico (2), considerando $p = 0,5$.

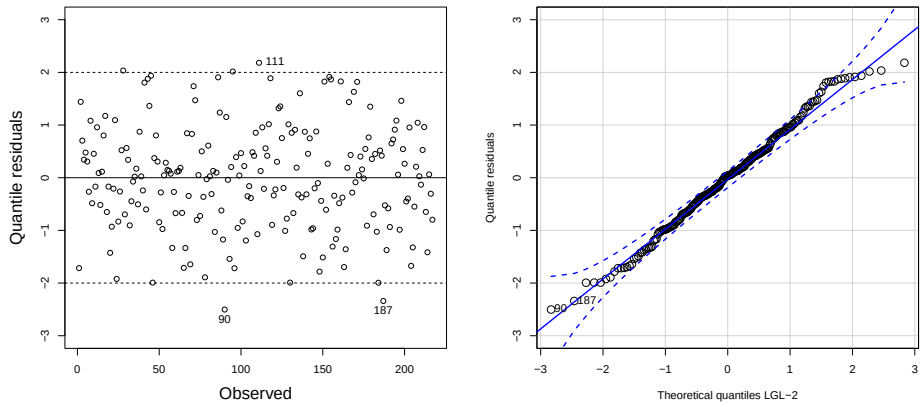


Figura 16 – Gráfico e envelope dos resíduos quantílicos do modelo LG-Logistic (2), considerando $p = 0,5$.

Verificamos graficamente a normalidade dos resíduos considerando um *QQ-plot* e seu gráfico de envelope para os resíduos quantílicos aleatórios. Os gráficos da Figura 15 referem-se ao modelo G-Logistic (2) e os gráficos da Figura 16 referem-se ao modelo LG-Logistic (2).

O envelope da Figura 15 e da Figura 16 mostra que os modelos estão se ajustando bem aos dados, com a distribuição LG-Logistic (2) tendo um ajuste melhor do que a distribuição G-Logistic (2). Além disso, é possível perceber através dos gráficos de resíduos (gráficos à esquerda) que a observação 111 apresenta resíduo quantílico positivo, enquanto as observações 90 e 187 apresentam resíduos quantílicos negativos. A observação 111 apresenta proporções de vulnerabilidade à pobreza iguais a $0,673$, enquanto as observações 90 e 187 apresentam valores iguais a $0,059$ e $0,073$, respetivamente.

3.5.2 Aplicação do modelo de regressão LG-Logistic

Nesta subseção apresentamos uma aplicação sobre a pobreza no Peru. O conjunto de dados fornece informações sobre a proporção de extrema pobreza (INEI, 2009) e o Índice de Desenvolvimento Humano (IDH) (PNUD, 2009) nas 25 subdivisões territoriais do país.

Uma variável dependente e uma variável independente foram consideradas nesta aplicação. O conjunto de dados tem 195 observações e estamos interessados em modelar a proporção de pessoas extremamente pobres no Peru usando o Índice de Desenvolvimento Humano (IDH). Aplicamos este conjunto de dados ao modelo presente na seção 3.1.2, bem como aos modelos L-Logístico (2) e G-Logístico (2).

Além disso, para cada modelo acima, propomos quatro submodelos possíveis: um modelo de regressão sem covariáveis (modelo nulo), dois modelos de regressão considerando efeitos de covariáveis em diferentes parâmetros (primeiro em κ, depois em φ) e um modelo de regressão considerando covariáveis nos dois parâmetros κ e φ (modelo completo). Detalhes desses modelos são fornecidos abaixo.

A estrutura dos quatro modelos propostos a partir do modelo de regressão L-Logistic é

$$I.\begin{cases} \text{modelo nulo :} & \text{logito}(\kappa_i)=\beta_0 \ \text{ e } \ \log(\varphi_i)=-\delta_0 \\ \text{modelo de locação :} & \text{logito}(\kappa_i)=\boldsymbol{X}_{1i}^T\boldsymbol{\beta} \ \text{ e } \ \log(\varphi_i)=-\delta_0 \\ \text{modelo de forma :} & \text{logito}(\kappa_i)=\beta_0 \ \text{ e } \ \log(\varphi_i)=\boldsymbol{X}_{1i}^T\boldsymbol{\delta} \\ \text{modelo completo :} & \text{logito}(\kappa_i)=\boldsymbol{X}_{1i}^T\boldsymbol{\beta} \ \text{ e } \ \log(\varphi_i)=\boldsymbol{X}_{2i}^T\boldsymbol{\delta} \end{cases}$$

A estrutura dos quatro modelos propostos a partir do modelo de regressão G-Logistic é

$$II.\begin{cases} \text{modelo nulo :} & \text{logito}(\phi_i)=\beta_0 \ \text{ e } \ \log(\varphi_i)=-\delta_0 \\ \text{modelo de escala :} & \text{logito}(\phi_i)=\boldsymbol{X}_{1i}^T\boldsymbol{\beta} \ \text{ e } \ \log(\varphi_i)=-\delta_0 \\ \text{modelo de forma :} & \text{logito}(\phi_i)=\beta_0 \ \text{ e } \ \log(\varphi_i)=\boldsymbol{X}_{1i}^T\boldsymbol{\delta} \\ \text{modelo completo :} & \text{logito}(\phi_i)=\boldsymbol{X}_{1i}^T\boldsymbol{\beta} \ \text{ e } \ \log(\varphi_i)=\boldsymbol{X}_{2i}^T\boldsymbol{\delta} \end{cases}$$

A estrutura dos quatro modelos propostos a partir do modelo de regressão LG-Logistic é

$$III.\begin{cases} \text{modelo nulo :} & \text{logito}(\kappa_i)=\beta_0, \ \log(\phi_i)=\lambda_0 \ \text{ e } \ \log(\varphi_i)=-\delta_0 \\ \text{modelo de locação :} & \text{logito}(\kappa_i)=\boldsymbol{X}_{1i}^T\boldsymbol{\beta}, \ \log(\phi_i)=\lambda_0 \ \text{ e } \ \log(\varphi_i)=-\delta_0 \\ \text{modelo de forma :} & \text{logito}(\kappa_i)=\beta_0, \ \log(\phi_i)=\lambda_0 \ \text{ e } \ \log(\varphi_i)=\boldsymbol{X}_{1i}^T\boldsymbol{\delta} \\ \text{modelo completo :} & \text{logito}(\kappa_i)=\boldsymbol{X}_{1i}^T\boldsymbol{\beta}, \ \log(\phi_i)=\lambda_0 \ \text{ e } \ \log(\varphi_i)=\boldsymbol{X}_{2i}^T\boldsymbol{\delta} \end{cases}$$

onde $\boldsymbol{X}_{1i}^T\boldsymbol{\beta}=\beta_0+\beta_1 X_{1i}$ e $\boldsymbol{X}_{2i}^T\boldsymbol{\delta}=\delta_0+\delta_1 X_{2i}$, $i=1,\cdots,195$, com $\boldsymbol{\beta}=(\beta_0,\beta_1)^T$, $\boldsymbol{\delta}=(\delta_0,\delta_1)^T$ e λ_0 são os coeficientes de regressão e X_{1i} e X_{2i} são as covariáveis, que neste caso representam o IDH. Observe que o último modelo de regressão proposto em *III* é apresentado na seção 3.3.2.

Para estimar os parâmetros acima, utilizamos uma abordagem bayesiana e *prioris* não informativas para os coeficientes de regressão, ou seja, $\beta_j \sim \text{N}(0,100)$, $\delta_j \sim \text{N}\,(0,100)$, para $j=1,2$ e $\lambda_0 \sim \text{N}(0,100)$. Implementamos os modelos usando o pacote `rstan` (Stan Development Team, 2020), obtendo convergência via MCMC usando 10000 iterações e 2 cadeias. O código Stan do modelo completo para esta aplicação está disponível no Apêndice A.3.2.

Todos os desenvolvimentos foram realizados para $p=0,5$. A mesma aplicação foi realizada utilizando a estrutura dos modelos propostos acima, considerando outros valores para p, cujos resultados estão apresentados no apêndice A.2.2.

Os resultados sobre os critérios de comparação dos modelos estão presentes na Tabela 21, que apresenta as três famílias de modelos descritas anteriormente: família de modelos utilizando a distribuição LG-Logistic, G-Logistic e L-Logistic, nesta ordem.

Tabela 21 – Resultados dos critérios de comparação para aplicação dos dados de pobreza do Peru na família de modelos das distribuições reparametrizadas LG-Logístic, G-Logístic e L-Logístic, considerando $p = 0,5$.

	Família de modelos usndo a distribuição LG-Logistic			
Critérios	Modelo nulo	Modelo de locação	Modelo de forma	Modelo completo
DIC	-236,710	-470,251	-244,287	-469,312
EAIC	-233,120	-465,851	-230,672	-463,689
EBIC	-223,301	-452,759	-217,580	-447,324
IC	-234,300	-466,651	-249,902	-464,935
HQIC	-229,145	-460,551	-225,371	-457,063
WAIC	-235,856	-469,397	-233,584	-468,211
LOO	-235,765	-469,376	-233,464	-468,180
	Família de modelos usndo a distribuição G-Logistic			
Critérios	Modelo nulo	Modelo de escala	Modelo de forma	Modelo completo
DIC	-230,156	-231,189	-292,283	-293,171
EAIC	-227,833	-225,960	-289,138	-288,164
EBIC	-221,287	-216,141	-279,319	-275,072
IC	-228,478	-230,418	-289,427	-290,177
HQIC	-225,183	-221,985	-285,162	-282,864
WAIC	-229,304	-229,288	-292,017	-292,979
LOO	-229,016	-228,882	-291,664	-292,630
	Família de modelos usndo a distribuição L-Logistic			
Critérios	Modelo nulo	Modelo de locação	Modelo de forma	Modelo completo
DIC	-196,164	-423,505	-259,032	-428,375
EAIC	-194,155	-422,403	-255,567	-423,798
EBIC	-187,609	-415,857	-245,748	-410,706
IC	-194,173	-420,607	-256,497	-424,951
HQIC	-191,505	-419,753	-251,592	-418,497
WAIC	-195,837	-422,249	-258,177	-425,156
LOO	-195,834	-422,245	-258,171	-425,112

Comparando os resultados, notamos que o modelo de regressão LG-Logistic apresenta o melhor ajuste a este conjunto de dados, independentemente se considerarmos o modelo sem covariáveis, com covariáveis no parâmetro localização ou covariáveis na localização e forma. Portanto, há fortes indícios de que o modelo de regressão LG-Logistic seja uma boa alternativa para dados com essas características.

Dentre esses modelos, encontramos um melhor ajuste com o modelo de locação usando a distribuição LG-Logistic e o modelo completo usando a distribuição LG-Logistic, mas por parcimônia escolhemos o primeiro. Assim, o modelo final é

$$\hat{y}_i \sim \text{LGL}(\kappa_i, \phi_i, \varphi_i)$$
$$logit(\kappa_i) = \boldsymbol{X}_{1i}^T \boldsymbol{\beta}$$
$$\log(\phi_i) = \lambda_0$$
$$\log(\varphi_i) = -\delta_0,$$

A Tabela 22 apresenta as estimativas dos parâmetros do melhor modelo, bem como o desvio padrão, percentis e estatísticas $\hat{R}$. Observe que $\hat{R}$ é igual ou muito próximo de 1 para todos os casos, indicando convergência das cadeias MCMC.

Tabela 22 – Resumo das informações das estimativas dos parâmetros do Modelo de Locação LG-Logistic, considerando $p = 0,5$.

Par	Média	DP	2,5%	97,5%	$\hat{R}$
β_0	13,123	0,670	11,728	14,406	1,001
β_1	-25,350	1,151	-27,549	-22,941	1,002
δ_0	-0,542	0,061	-0,666	-0,423	1,000
λ_0	-3,559	0,999	-5,974	-2,126	1,000

Podemos comparar os modelos entre si ao considerar os resultados disponíveis no Apêndice A.2.2. Isto é, para a família de modelos LG-Logistic, os critérios apresentam melhores valores em $p = 0,10$; os modelos utilizando a distribuição G-Logistic são melhores usando $p = 0,90$ e considerando a família de modelos L-Logistic os resultados são melhores para $p = 0,50$.

3.6 Comentários finais

Introduzimos neste trabalho a distribuição LG-Logistic, suas propriedades e sua reparametrização quantílica. Também apresentamos algumas medidas para descrever a distribuição de formas e propôr um modelo de regressão baseado na reparametrização quantílica.

Foram apresentados dois estudos de simulação: o primeiro foi uma análise de sensibilidade da *priori* e o segundo foi um estudo de recuperação de parâmetros.

Também incluímos aplicações em conjuntos de dados reais. O primeiro foi modelar a vulnerabilidade à pobreza no estado do Maranhão, Brasil, e o segundo foi uma análise do modelo de regressão para estimar a pobreza no Peru de acordo com o Índice de Desenvolvimento Humano. Em ambos os casos a distribuição LG-Logistic foi comparada com outros modelos competitivos, mostrando o primeiro um melhor desempenho na maioria dos casos.

De modo geral, os resultados deste trabalho mostram que a distribuição LG-Logistic proposta é uma alternativa excelente e flexível para modelar dados com resposta limitada no intervalo $(0,1)$.

Para pesquisas futuras, propomos estender o modelo para uma versão inflacionada que considere excessos de zeros e/ou uns. Adicionalmente, podem ser consideradas extensões do modelo proposto para modelos mistos.

Em relação às propriedades de momentos da distribuição LG-Logisitc, seria interessante estudar a possibilidade de definir uma expressão fechada para 3.11, assim como Nadarajah

e Si (2020) apresentam uma expressão de forma fechada para propriedades de momento da distribuição L-Logistic, além de algumas questões computacionais.

Outra direção para desenvolvimentos futuros é verificar o ajuste dos modelos aos dados. De uma perspectiva Bayesiana, a falta de ajuste dos dados em relação à distribuição preditiva posterior pode ser medida pela probabilidade da área da cauda, ou p-valor da quantidade do teste, e calculada usando simulações posteriores de dados replicados. Pode ser equivalente a um teste de ajuste geral sob uma abordagem de máxima verossimilhança.

Neste trabalho, utilizamos a estimativa Bayesiana devido às suas inúmeras vantagens, incluindo a incorporação de conhecimento prévio, quantificação da Incerteza e flexibilidade na especificação do modelo. Contudo, em desenvolvimentos futuros, pretendemos também explorar a estimativa de máxima verossimilhança para o modelo proposto. Esta exploração envolverá a verificação das propriedades assintóticas do modelo e a avaliação da sua eficiência computacional.

Por outro lado, explorar uma comparação com o modelo de regressão Beta quantílica (BOURGUIGNON; GALLARDO; SAULO, 2021) pode ser valiosa, visto que este modelo é inovador e promissor e considerando que uma estimativa Bayesiana é essencial para uma comparação significativa.

CAPÍTULO

4

UM NOVO MODELO CONJUNTO PARA RESPOSTA DE TEMPO E ACURÁCIA

A informação tornou-se mais acessível com o avanço da tecnologia e a realização de testes em um computador permite que, além do registro da resposta do item, e se ela está correta ou não, o tempo que um examinado gastou para responder um item da prova também seja registrado. Com isso, a necessidade de modelar e informar algo a partir deste tempo tornou-se importante como, por exemplo, informar sobre a capacidade do examinado.

Linden (2006) introduziu a modelagem de tempos de resposta (TR) dentro de uma estrutura hierárquica, na qual leva em consideração uma tendência entre velocidade ao responder determinado item e a precisão de resposta. Para isso, no primeiro nível, assumem-se duas distribuições separadas para modelar TR e respostas dos itens, cada um com um conjunto diferente de parâmetros de pessoa (velocidade e precisão, ou habilidade) e item, e leva-se em consideração que as respostas dos itens e TR são condicionalmente independentes, dados os parâmetros de velocidade e precisão. No segundo nível esses parâmetros podem ser dependentes, o que leva a uma estrutura de modelagem hierárquica, na qual a relação entre velocidade e precisão depende do nível de modelagem.

Alguns trabalhos apresentam essa estrutura hierárquica para modelagem de TR, em especial, o trabalho de Flores *et al.* (2019). Além dos autores terem proposto um modelo dentro do contexto hierárquico, propuseram um modelo de proporção de tempo, visto que o tempo de resposta em um teste não é infinito, existe um tempo limite superior para que o examinado complete todas as respostas. Neste sentido, os autores propuseram um modelo conjunto hierárquico para tempo de resposta limitado e precisão. No primeiro nível da modelagem, utilizaram a distribuição Bernoulli para modelar a precisão de resposta e a distribuição Simplex para modelar a proporção de tempo de resposta.

Como o tempo para responder um teste não é infinito, este trabalho visa apresentar uma outra proposta de modelo dentro do contexto hierárquico, na qual a precisão de resposta dos

itens é modelada usando um modelo TRI probito de dois parâmetros (assim como em Flores *et al.* (2019)) e a proporção de tempo de resposta, que está compreendida no intervalo $(0,1)$, é modelada usando a distribuição Beta (FERRARI; CRIBARI-NETO, 2004), com parâmetros de posição e precisão.

Este capítulo está estruturado da seguinte forma: na seção 4.1 apresentamos a distribuição Beta para resposta limitada no intervalo $(0,1)$ com parâmetros de posição e precisão, bem como o modelo de tempo de resposta Beta e também o modelo Normal Ogive de dois parâmetros. Na seção 4.2 apresentamos o modelo Hierárquico Beta-Bernoulli e na seção 4.3 apresentamos a abordagem Bayesiana para as estimativas dos parâmetros. Já as seções 4.4 e 4.5 são dedicadas, respectivamente, a um pequeno estudo de simulação e à aplicação do Modelo Hierárquico Beta-Bernoulli em um conjunto de dados reais do Programa de Avaliação Internacional de Estudantes (PISA). Finalmente, na seção 4.6 apresentamos algumas conclusões sobre este trabalho.

4.1 Conceitos preliminares

Esta seção é dedicada à apresentação dos modelos de Tempo de Resposta (TR) e Acurácia de Resposta (AR), de forma preliminar à apresentação da estutura do modelo hierárquico. Dessa forma, apresentamos, inicialmente, a distribuição Beta com parâmetros de posição e precisão, assim como a proposta de modelo de tempo de resposta utilizando a distribuição Beta; na sequência apresentamos o modelo Normal Ogive de dois parâmetros para modelar a acurácia de resposta.

4.1.1 Distribuição Beta

Seja Z uma variável aleatória proveniente de uma distribuição Beta, com parâmetros de média e precisão, cuja função densidade de probabilidade é dada por

$$f(z \mid \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} z^{\mu\phi-1}(1-z)^{(1-\mu)\phi-1}, \tag{4.1}$$

com $0 < z < 1$, $0 < \mu < 1$ é o parâmetro de posição (média) e $\phi > 0$ o parâmetro de precisão (FERRARI; CRIBARI-NETO, 2004). A média e a variância de Z são dadas, respectivamente, por

$$\mathbb{E}(Z \mid \mu, \phi) = \mu \quad \text{e} \quad \mathbb{V}\text{ar}(Z \mid \mu, \phi) = \frac{\mu(1-\mu)}{(1+\phi)}.$$

A Figura 17 mostra as diferentes densidades da distribuição Beta (4.1) para os diferentes valores de μ e ϕ.

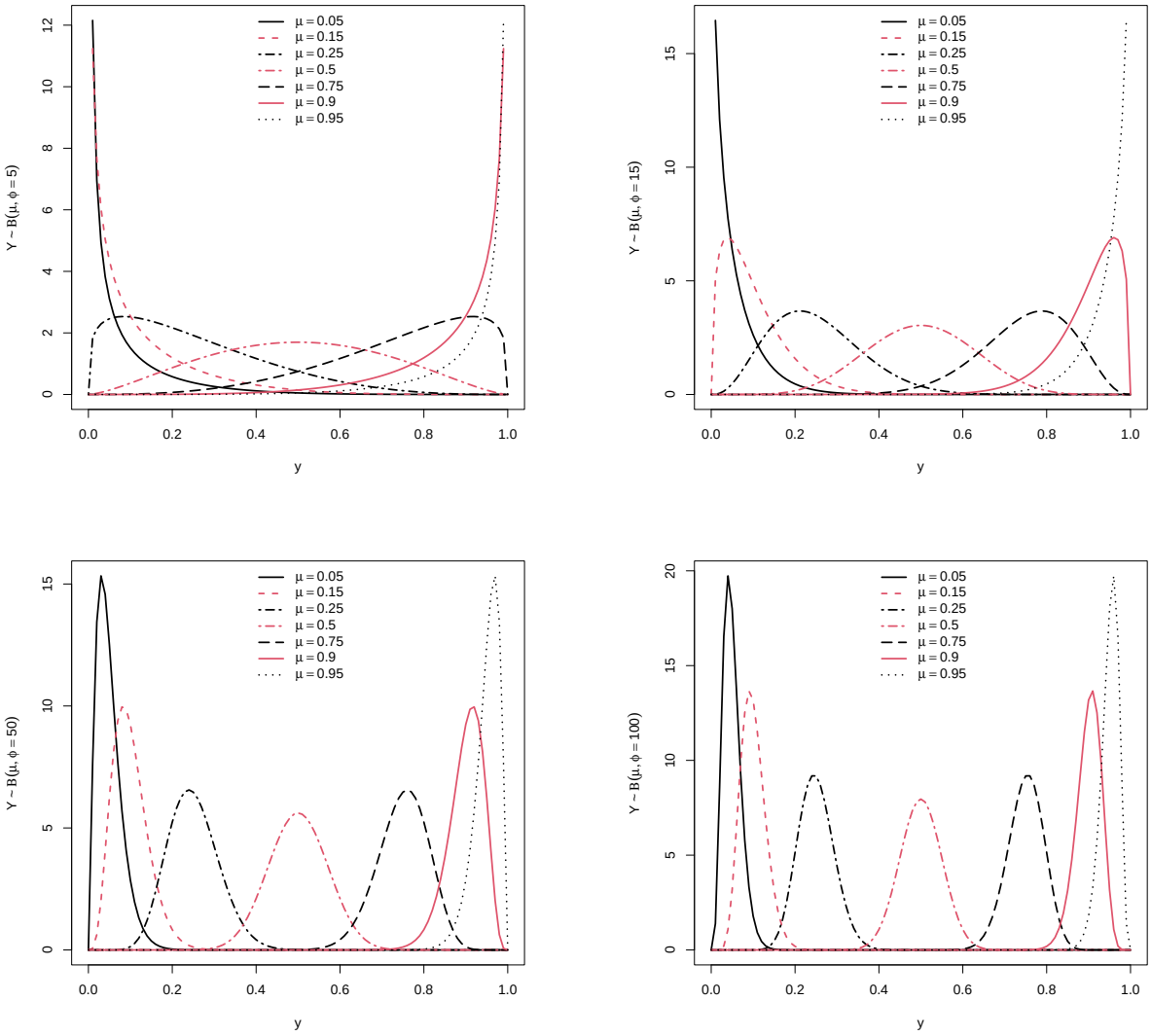


Figura 17 – Distribuição de Y para diferentes valores do parâmetro μ e (a) $\phi = 5$, (b) $\phi = 15$, (c) $\phi = 50$ e (d) $\phi = 100$ no modelo de regressão Beta.

Observe que em todos os gráficos variamos os valores do parâmetro de posição $\mu = (0,05, 0,25, 0,50, 0,75, 0,95)$ e tomamos valores diferentes para o parâmetros de precisão, ou seja, na figura (a) fixamos o valor do parâmetro de precisão em $\phi = 5$, na figura (b) fixamos em $\phi = 15$, na figura (c) fixamos em $\phi = 50$ e na figura (d) fixamos o parâmetro de precisão em $\phi = 100$. As densidades podem apresentar formas bastante diferentes dependendo dos valores tomados para os dois parâmetros, mas em todos os casos, quando $\mu = 0,50$ a distribuição apresenta-se de forma simétrica, e assimétrica quando $\mu \neq 0,50$.

4.1.2 Modelo Beta para tempo de resposta

Nesta subseção propomos usar a distribuição Beta para modelar TR. Considerando que em uma prova um examinado tem um tempo limite para responder os itens da prova de

habilidades, Flores *et al.* (2019) propuseram um modelo para resposta de tempo limitado capaz de identificar a proporção de tempo que um examinado (i) gasta para responder um item (j) da prova.

$$Z_{ij} = \frac{T_{ij} - c_j}{d_j - c_j}, \tag{4.2}$$

na qual c_j é o menor tempo que algum examinado gasta para responder a um determinado item j e d_j é o maior tempo, sendo c_j e d_j constantes definidas previamente desde os dados e assumidos como conhecidos e T_{ij} é o tempo de resposta do examinado i e um item j.

Note que Z_{ij} é uma variável aleatória que pertence ao intervalo $(0,1)$ e, portanto, pode ser modelado por (4.1).

Nós propomos um TR limitado, considerando

$$Z_{ij} \mid \tau_i, \beta_j, \alpha_j \sim Beta(\mu_{ij}, \alpha_j) \tag{4.3}$$

e

$$\text{logit}(\mu_{ij}) = \beta_j - \tau_i \tag{4.4}$$

na qual $\alpha_j > 0$ é o parâmetro de precisão que pode ser interpretado inversamente ao poder de discriminação para o item j. Ou seja, quando este parâmetro é aumentado a variância diminui, indicando que aquele item não apresenta variabilidade suficiente e então fica difícil discriminar entre os indivíduos. Além disso, como $0 < \mu_{ij} < 1$ é um parâmetro de localização, consideramos que ele é afetado por dois tipos de parâmetros: um está associado ao traço latente dos indivíduos ($\tau_i \in \mathbb{R}$) e o outro parâmetro está associado ao item ($\beta_j \in \mathbb{R}$). Este parâmetro pode ser interpretado como a dificuldade do item. Neste caso, a função logito $g(x) = \log(x/(1-x))$ é uma função de ligação, mas outra função de ligação pode ser considerada.

Neste caso, a função densidade de probabilidade da variável Z_{ij} é dada por

$$p(z_{ij} \mid \tau_i, \alpha_j, \beta_j) = \frac{\Gamma(\alpha_j)}{\Gamma\left(\frac{\alpha_j \exp(\beta_j - \tau_i)}{1+\exp(\beta_j - \tau_i)}\right)\Gamma\left(\frac{\alpha_j}{1+\exp(\beta_j - \tau_i)}\right)} z_{ij}^{\left(\frac{\alpha_j \exp(\beta_j - \tau_i)}{1+\exp(\beta_j - \tau_i)}\right)-1} (1-z_{ij})^{\left(\frac{\alpha_j}{1+\exp(\beta_j - \tau_i)}\right)-1}, \tag{4.5}$$

4.1.3 Modelo Ogiva Normal AR

A precisão das respostas y_{ij}, correta e incorreta, pode ser modelada a partir de uma distribuição de Bernoulli, cujo parâmetro é a probabilidade da resposta estar correta. Em outras palavras, a precisão das respostas é modelada usando o modelo probit TRI de dois parâmetros (a_j, b_j), ou seja

$$y_{ij} \mid \theta_i, a_j, b_j \sim Bernoulli(p_{ij})$$

isto é,

$$p(y_{ij} \mid \theta_i, a_j, b_j) = \Phi(a_j(\theta_i - b_j))^{y_{ij}} (1 - \Phi(a_j(\theta_i - b_j)))^{1-y_{ij}}, \tag{4.6}$$

onde $p_{ij} = \Phi(a_j(\theta_i - b_j))$ é a probabilidade de a resposta estar correta, $\theta_i \in \mathbb{R}$ é o parâmetro de habilidade do examinado i, $a_j > 0$ a discriminação do item j e $b_j \in \mathbb{R}$ a dificuldade do item j.

Observe que Φ corresponde à função de distribuição cumulativa da distribuição Norma Padrão e então resulta na versão Ogiva Normal do modelo TRI de dois parâmetros. Outras distribuições como logit, por exemplo, podem ver consideradas.

4.2 Um modelo para tempo de resposta limitado e acurácia de resposta

Além do tempo limitado para responder aos itens de um teste, o examinando tem a possibilidade de responder corretamente ou incorretamente um item, ou seja, a precisão da resposta é dicotômica.

Em nosso modelo, a proporção do tempo de resposta (PTR) Z_{ij} segue o modelo Beta para tempo de resposta, proposto em (4.3) e a acurácia de resposta Y_{ij} segue o modelo TRI Ogive Normal em (4.1.3) e consideramos que ambas as respostas são independentes.

Como (Z_{ij}, Y_{ij}) são modelados a partir de duas distribuições independentes para diferentes itens j e para cada examinado i, podemos escrever a distribuição conjunta condicionada à distribuição de habilidade e velocidade do examinado i ($\xi_i = (\theta_i, \tau_i)$), como segue

$$p(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\xi}, \boldsymbol{\nu}) = \prod_{i=1}^{I} \prod_{j=1}^{J} p(y_{ij} \mid \theta_i, a_j, b_j) \times p(z_{ij} \mid \tau_i, \alpha_j, \beta_j), \tag{4.7}$$

em que $p(y_{ij} \mid \theta_i, a_j, b_j)$ é definido em (4.6) e $p(z_{ij} \mid \tau_i, \alpha_j, \beta_j)$ é definido em (4.5).

Note que $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_I)$, onde $\xi_i = (\theta_i, \tau_i)$ é o vetor de parâmetros de habilidade e velocidade de cada examinado e $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_J)$ é o vetor de parâmetros do item, onde $\nu_j = (a_j, b_j, \alpha_j, \beta_j)$.

4.3 Inferência

A estimação dos parâmetros do modelo conjunto apresentado na seção 4.2 será realizada sob abordagem Bayesiana. Desta forma, temos que a função de verossimilhança conjunta é dada por

$$L(\boldsymbol{\xi}, \boldsymbol{\nu} \mid \boldsymbol{y}, \boldsymbol{z}) = \prod_{i=1}^{I} \prod_{j=1}^{J} p(y_{ij} \mid \theta_i, a_j, b_j) p(z_{ij} \mid \tau_i, \alpha_j, \beta_j). \tag{4.8}$$

Nós propomos a seguinte estrutura de *prioris*:

$$p(\boldsymbol{\xi}, \boldsymbol{\nu}) = p(\xi_i \mid \mu_\xi, \Sigma_\xi) \times p(\nu_j \mid \mu_\nu, \Sigma_\nu). \tag{4.9}$$

Para $p(\xi_i \mid \mu_\xi, \Sigma_\xi)$ podemos assumir uma distribuição *priori* Normal bivariada ou podemos propor duas *prioris* univariadas com alguma estrutura de dependência. Primeiro, assumimos

que as habilidades para o modelo AR seguem $p(\theta_i \mid \mu_\theta, \sigma_\theta^2)$ e que a velocidade para o modelo PTR segue $p(\tau_i \mid \mu_\tau, \sigma_\tau^{2c})$, com $\mu_\tau = \rho_{\theta\tau}\theta_i$ e $\sigma_\tau^2 = \sigma_\tau^{2c} + \rho_{\theta\tau}^2$, em que $\rho_{\theta\tau}$ dá a relação entre θ e τ. Já para $p(\nu_j \mid \mu_\nu, \Sigma_\nu)$ propomos uma distribuição Normal multivariada com dimensão 4, cujo vetor médio é igual a $\mu_\nu = (\mu_a, \mu_b, \mu_\theta, \mu_\tau)$ e a matriz de covariância Σ_ν é igual a

$$\Sigma_\nu = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{a\alpha} & \sigma_{a\beta} \\ \sigma_{ba} & \sigma_b^2 & \sigma_{b\alpha} & \sigma_{b\beta} \\ \sigma_{\alpha a} & \sigma_{\alpha b} & \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\beta a} & \sigma_{\beta b} & \sigma_{\beta\alpha} & \sigma_\beta^2 \end{bmatrix}.$$

Segundo Linden (2006), podemos impor algumas restrições a este tipo de modelo. Assim, a média da velocidade é fixada em $\mu_\tau = 0$, a média de precisão em $\mu_\theta = 0$ e a variância de precisão em $\sigma_\theta^2 = 1$. Adicionalmente nenhuma restrição é considerada pelos parâmetros do item.

Além disso, podemos considerar hiper-*prioris* para alguns hiperparâmetros introduzidos na especificação das *prioris* acima. Assim, a especificação *priori* para o modelo de junção PTR e AR é resumida na seguinte estrutura hierárquica:

$$\left.\begin{aligned} p(\theta_i \mid \mu_\theta, \sigma_\theta^2) &\sim N(0,1) \\ p(\tau_i \mid \mu_\tau, \sigma_\tau^{2c}) &\sim N(\rho_{\theta\tau}, \sigma_\tau^{2c} + \rho_{\theta\tau}^2) \\ \rho_{\theta\tau} &\sim N(0, 0.1) \\ \sigma_\tau^2 &\sim \text{Inverse-Gamma}(0.1, 0.1) \end{aligned}\right\} \text{Traços latentes}$$

$$\left.\begin{aligned} \mu_\nu \mid \Sigma_\nu &\sim \text{MVN}(\mu_{\nu 0}^T, \Sigma_\nu) \\ \Sigma_\nu &\sim \text{Inverse-Whishart}(I^{-1}, 4) \end{aligned}\right\} \text{Parâmetros dos itens}$$

onde I é a matriz identidade. Para mais informações sobre as escolhas deste *prioris*, ver Flores *et al.* (2019) e Linden (2006)).

Uma vez definida a estrutura dos *prioris* em nosso modelo, podemos escrever a distribuição *posteriori*, que é dada por

$$p(\boldsymbol{\xi}, \boldsymbol{\nu}, \rho_{\theta\tau}, \sigma_\tau^2, \boldsymbol{\mu}_\nu, \boldsymbol{\Sigma}_\nu \mid \boldsymbol{y}, \boldsymbol{z}) \propto p(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\xi}, \boldsymbol{\nu}) p(\boldsymbol{\xi}, \boldsymbol{\nu}) p(\rho_{\theta\tau}) p(\sigma_\tau^2) p(\mu_\nu \mid \boldsymbol{\Sigma}_\nu) p(\boldsymbol{\Sigma}_\nu) \quad (4.10)$$

O modelo aqui proposto foi implementado através do pacote `R2jags` (SU; YAJIMA, 2015) no *software* R (R Core Team, 2018). Este pacote que fornece funções *wrapper* para implementar análise Bayesiana em JAGS (*Just Another Gibbs Sampler*), usa estatísticas Rubin e Gelman Rhat para monitorar a convergência de um modelo MCMC. Para facilitar o acesso em R, usamos a função *jags*.

4.4 Estudo de simulação

Desenvolvemos um breve estudo de simulação para avaliar a recuperação dos parâmetros do modelo Beta-Bernoulli. Para isso foram considerados $j = 28$ itens e $i = 1000$ indivíduos,

além de valores fixos de a, b, α, β, θ e τ semelhante aos resultados obtidos na aplicação (ver Tabela 23), onde θ e τ foram gerados usando distribuição Normal, com média 0 e desvio padrão de 1.

Corrigindo os traços latentes e traços de velocidade, e resposta de acurácia dos parâmetros dos itens, considerando $y_{ij} \sim Bernoulli(p_{ij})$, com $p_{ij} = \Phi(a_j(\theta_i - b_j))$ e proporção de tempo de resposta, $z_{ij} \sim Beta(\mu_{ij}, \alpha_{ij})$, em que $log(\mu_{ij}) = \beta_j - \tau_i$ foram gerados.

Cada conjunto de dados gerado foi aplicado ao modelo proposto, apresentado na seção 4.2, cuja estimação de cada amostra simulada foi realizada utilizando o procedimento descrito na seção 4.3. Usamos 2 cadeias MCMC, 16000 iterações, com um descarte de 4000 iterações, e um espaçamento igual a 4. Foram consideradas 50 repetições deste procedimento.

Para verificar se nosso procedimento bayesiano recupera os parâmetros do modelo, consideramos a média (Média), desvio padrão (DP), viés (Viés) e raiz quadrada do erro quadrático médio (RMSE) das estimações, ou seja calculado por $\text{RMSE}(\Theta_l) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\Theta_l^{(r)} - \Theta_l)^2}$, na qual R é o número de réplicas na simulação, $\Theta_l^{(r)}$ é a média posterior do parâmetro Θ_j na r-ésima réplica, e Θ_j é o l-ésimo componente de $\Theta = (a, b, \alpha, \beta)^T$.

A Tabela 23 apresenta, para cada conjunto de parâmetros, a média das estatísticas entre diferentes itens para os parâmetros a, b, α e β ou entre indivíduos para os parâmetros θ e τ. Além disso, apresenta também a probabilidade de cobertura de 95% para os intervalos de credibilidade das cadeias MCMC, sendo que o valor representa a probabilidade de que o intervalo de credibilidade contenha, em pelo menos 95% das análises, o verdadeiro valor do parâmetro.

Tabela 23 – Estatísticas das estimativas dos parâmetros do modelo Hierárquico Beta-Bernoulli

Parte do modelo	Parâmetros	Real	Média da média *Posterior*	DP da média *Posterior*	Média do Viés	Média da RMSE	Média coberturas	Média $\hat{R}$
IRT	a	0,579	0,583	0,049	0,004	0,078	0,955	1,008
	b	-1,010	-0,999	0,138	0,011	0,138	0,944	1,011
	θ	0,020	0,002	0,829	-0,008	0,826	0,950	1,001
RT	α	23,892	23,979	0,971	0,087	1,501	0,936	1,003
	β	-1,978	-1,969	0,038	0,010	0,031	0,934	1,089
	τ	0,001	0,000	0,885	0,011	1,067	0,951	1,007

Podemos observar na Tabela 23 que o modelo apresenta uma boa recuperação dos parâmetros visto que os valores de RMSE são pequenos e próximos de zero, assim como a probabilidade de cobertura está em torno de 1. Além disso, a estatística $\hat{R}$ é próxima do valor 1, permitindo concluir que as cadeias MCMC estão convergindo.

4.5 Aplicações

A aplicação presente nesta seção foi realizada com dados reais computadorizados do PISA 2015 para 28 itens de leitura, que pertencem aos clusters R1 e R3. Cada cluster foi projetado para ser concluído em 30 min e, sendo assim, o tempo total gasto para responder aos dois clusters não poderia ser maior que 1h. Foram consideradas apenas as respostas completas para os clusters R1 e R3, a fim de evitar dados faltantes. Este conjunto foi utilizado em 53 países e compreende um total de 4960 tempos de respostas (TR).

Para esta aplicação utilizamos a transformação para o tempo de resposta em proporção de TR, vista na seção 4.1.2, e consideramos o modelo apresentado na seção 4.2, cujas características de estimação são apresentadas na seção 4.3. Do ponto de vista Bayesiano, ajustamos o modelo através do pacote `R2jags` (SU; YAJIMA, 2015), usando 1 cadeia de Markov, 16000 iterações, 4000 *burnin* e um espaçamento igual a 4. Os códigos estão disponíveis no Apêndice B.1.

A Tabela 24 apresenta os resultados das estimativas dos parâmetros relacionados aos item e a Tabela 25 apresenta os resultados das estimtivas dos parâmetros em relação aos traços latentes.

Tabela 24 – Resultados das estimativas dos parametros de itens do modelo TRI.

	$\hat{a}$				$\hat{b}$			
Item	Média	DP médio	P2,5	P97,5	Média	DP médio	P2,5	P97,5
1	0,494	0,036	0,428	0,566	-3,163	0,205	-3,565	-2,773
2	0,390	0,022	0,348	0,433	-0,130	0,051	-0,233	-0,033
3	0,392	0,024	0,347	0,439	-1,946	0,120	-2,175	-1,733
4	0,682	0,031	0,621	0,742	0,854	0,040	0,780	0,935
5	0,645	0,027	0,598	0,697	-0,036	0,033	-0,101	0,028
6	0,710	0,033	0,643	0,773	-2,325	0,083	-2,502	-2,173
7	0,549	0,033	0,475	0,608	-2,123	0,126	-2,428	-1,928
8	0,566	0,028	0,510	0,621	-2,052	0,089	-2,239	-1,893
9	0,584	0,028	0,529	0,639	1,017	0,054	0,908	1,120
10	0,835	0,035	0,770	0,907	-0,918	0,039	-0,995	-0,842
11	0,605	0,025	0,560	0,655	-1,145	0,053	-1,245	-1,045
12	0,935	0,052	0,841	0,104	-2,086	0,079	-2,258	-1,947
13	0,583	0,028	0,528	0,636	-1,327	0,065	-1,457	-1,208
14	0,415	0,028	0,357	0,467	-3,069	0,195	-3,513	-2,756
15	0,584	0,026	0,533	0,634	-0,642	0,044	-0,735	-0,552
16	0,713	0,027	0,659	0,765	0,087	0,032	0,025	0,151
17	0,216	0,018	0,187	0,256	0,992	0,112	0,776	1,203
18	0,548	0,027	0,496	0,603	-1,903	0,089	-2,085	-1,749
19	0,961	0,039	0,885	0,104	-1,074	0,040	-1,154	-0,999
20	0,393	0,023	0,344	0,437	-0,296	0,055	-0,400	-0,187
21	0,399	0,021	0,361	0,447	-2,318	0,117	-2,530	-2,096

22	0,392	0,021	0,350	0,437	0,018	0,047	-0,076	0,108
23	0,459	0,025	0,417	0,516	1,602	0,085	1,414	1,761
24	0,421	0,022	0,381	0,466	-0,687	0,056	-0,798	-0,577
25	0,786	0,038	0,712	0,864	-1,976	0,075	-2,122	-1,834
26	0,551	0,025	0,502	0,598	-1,340	0,064	-1,477	-1,292
27	0,688	0,028	0,634	0,741	-0,626	0,037	-0,699	-1,228
28	0,725	0,033	0,664	0,792	-1,664	0,065	-1,799	-1,549

Tabela 25 – Resultados das estimativas dos parametros de itens do modelo RT.

	$\hat{\alpha}$				$\hat{\beta}$			
Item	Média	DP médio	P2,5	P97,5	Média	DP médio	P2,5	P97,5
1	30,084	0,647	28,867	31,382	-2,119	0,009	-2,136	-2,101
2	10,041	0,203	9,647	10,462	-1,481	0,011	-1,502	-1,458
3	11,281	0,224	10,851	11,722	-1,439	0,011	-1,460	-1,418
4	22,289	0,450	21,412	23,163	-2,313	0,011	-2,334	-2,294
5	19,173	0,406	18,378	19,974	-1,890	0,010	-1,909	-1,871
6	87,089	1,766	83,591	90,749	-3,431	0,009	-3,448	-3,414
7	5,954	0,129	5,710	6,216	-2,017	0,017	-2,049	-1,902
8	14,553	0,301	13,983	15,165	-1,923	0,011	-1,946	-1,916
9	4,137	0,084	3,975	4,307	-1,351	0,016	-1,383	-1,320
10	11,831	0,238	11,362	12,288	-1,483	0,011	-1,505	-1,461
11	22,484	0,455	21,618	23,388	-1,820	0,009	-1,837	-1,803
12	45,077	0,936	43,391	47,088	-2,786	0,009	-2,804	-2,768
13	88,784	1,941	85,019	92,799	-3,099	0,008	-3,114	-3,083
14	16,356	0,321	15,727	17,002	-2,009	0,011	-2,030	-1,989
15	16,441	0,355	15,744	17,178	-2,120	0,012	-2,143	-2,098
16	11,057	0,220	10,634	11,512	-1,375	0,011	-1,397	-1,354
17	18,375	0,376	17,659	19,168	-1,984	0,010	-2,003	-1,964
18	22,384	0,467	21,514	23,287	-1,885	0,009	-1,903	-1,867
19	23,323	0,483	22,376	24,272	-1,713	0,009	-1,730	-1,696
20	13,005	0,267	12,501	13,559	-1,764	0,011	-1,786	-1,742
21	40,326	0,854	38,688	42,020	-2,117	0,008	-2,132	-2,102
22	5,848	0,113	5,629	6,071	-1,150	0,013	-1,175	-1,125
23	11,395	0,225	10,944	11,858	-1,841	0,012	-1,866	-1,817
24	38,778	0,809	37,190	40,327	-3,258	0,012	-3,281	-3,235
25	10,585	0,214	10,182	11,022	-1,648	0,012	-1,670	-1,624
26	12,088	0,234	11,632	12,550	-1,246	0,010	-1,266	-1,227
27	24,029	0,487	23,111	24,984	-2,058	0,009	-2,077	-2,039
28	32,193	0,679	30,867	33,550	-2,065	0,008	-2,082	-2,049

A Figura 18 apresenta a média posterior dos diferentes parâmetros dos itens em relação às partes TRI e TR do modelo Hierárquico Beta-Bernoulli. Note que o gráfico do modelo TRI não apresenta uma tendência entre as estimatimas como apresenta no gráfico do modelo TR, que mostra uma correlação negativa entre as estimativas.

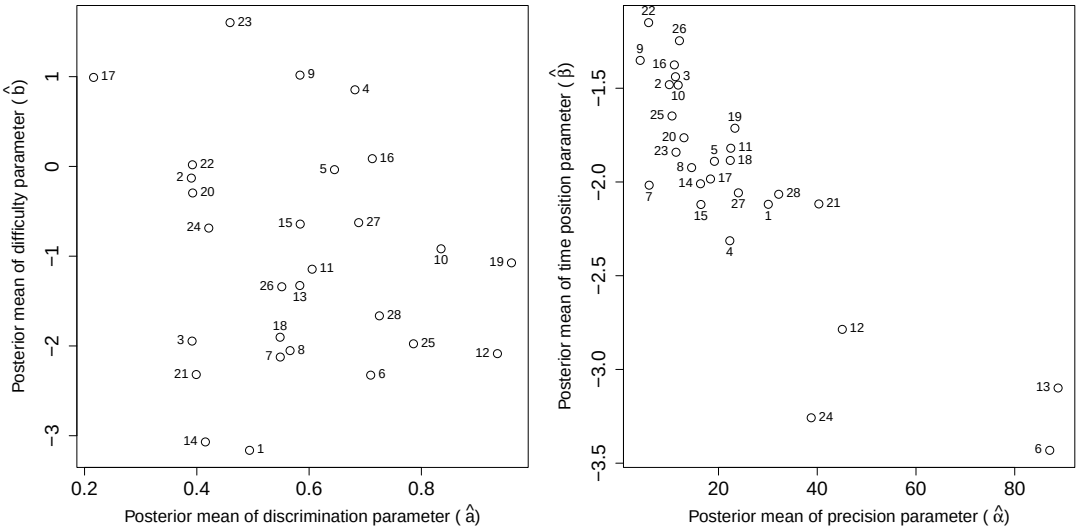


Figura 18 – Parâmetros dos 28 itens de teste do PISA 2015 (a) modelo TRI e (b) modelo TR.

Em resumo, no modelo TRI temos que a média posterior do parâmetro de discriminação ($\hat{a}$) é igual a 0,579, sendo seu valor mínimo igual a 0,216 e seu máximo igual a 0,961; já a média posterior do parâmetro de dificuldade ($\hat{b}$) é igual a -1,010, sendo o valor mínimo igual a -3,163 e o valor máximo igual a 1,602. No modelo TR, temos que a média posterior do parâmetro de precisão ($\hat{\alpha}$) é igual a 23,892, valor mínimo igual a 4,137 e valor máximo igual a 88,784. Por fim, a média posterior do parâmetro de posição ($\hat{\beta}$) é igual a -1,978, valor mínimo igual a -3,431 e valor máximo igual a -1,150.

As distribuições das estimativas dos parâmetros de habilidade θ_i e velocidade τ_i para a amostra são mostradas na Figura 19, além do gráfico de dispersão entre as estimativasa dos parâmetros θ_i e τ_i.

No gráfico das estimativas do parâmetro de habilidade é possível notar uma assimetria à esquerda, o que não é visto no gráfico das estimativas do parâmetro de velocidade. Em resumo, o valor médio da posterior do parâmetro de habilidade é igual a -0,002, sendo que o valor mínimo é igual a -3,059 e o máximo igual a 2,188. Já o valor médio da posterior do parâmetro de velocidade é igual a -0,001, sendo o valor mínimo igual a -1,165 e o valor máximo igual a 1,414.

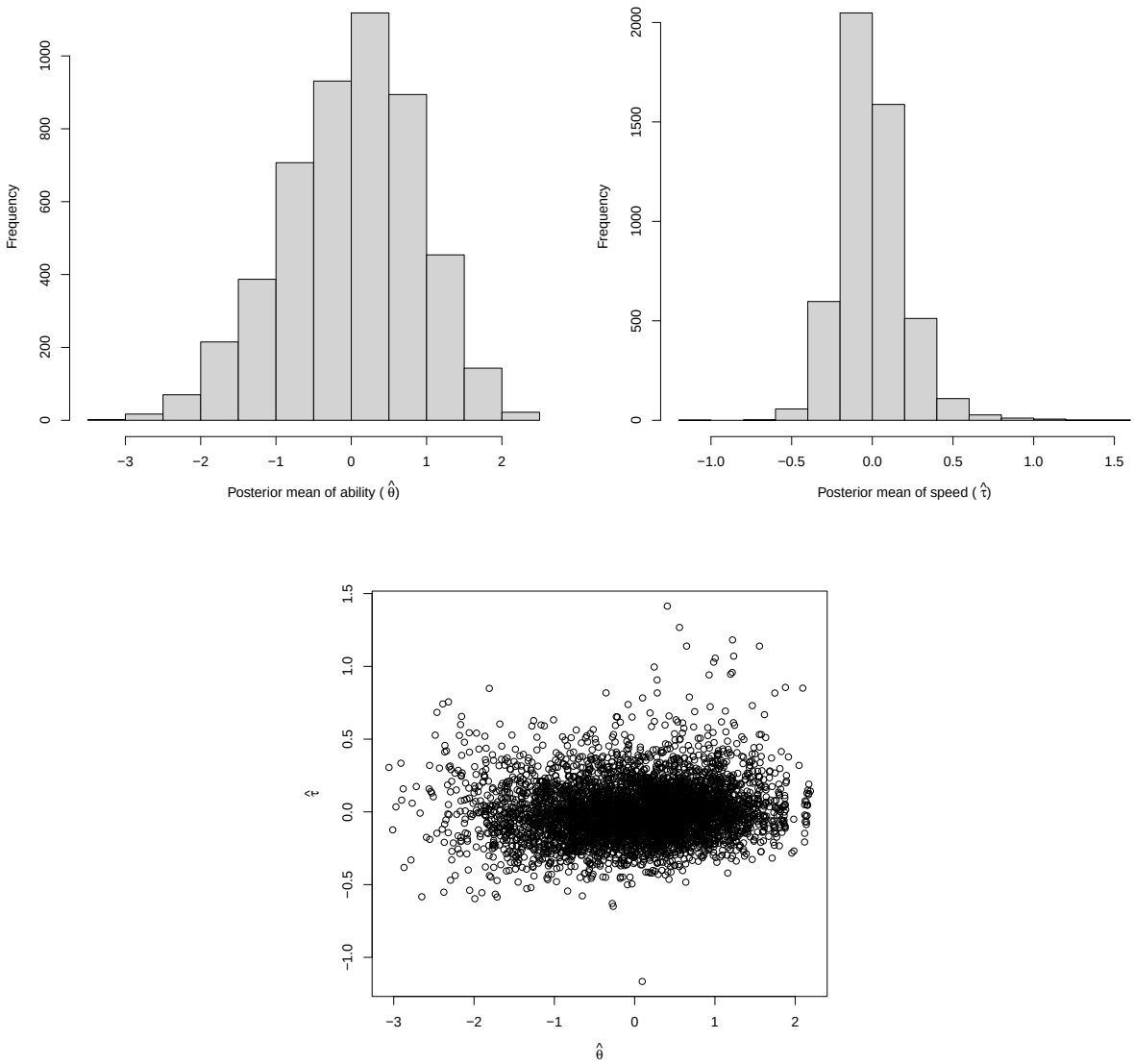


Figura 19 – Distribuições de estimativas de (a) parâmetro de habilidade e (b) parâmetro de velocidade, e gráfico de dispersão entre $\hat{\theta}$ e $\hat{\tau}$.

4.6 Comentários finais

Inicialmente realizamos um estudo teórico acerca da modelagem de tempos de respostas e percebemos a importância de corroborar com este tipo de modelagem, dado o avanço tecnológico para a realização de testes e a necessidade de mais informação a respeito da pessoa que está sendo examinada (sobre suas habilidades, por exemplo).

Reparamos também a importância de tratar o tempo dentro de um limite, pois nenhum teste é realizado até o tempo infinito, ou seja, o examinado tem um limite de tempo para responder aos itens do teste e esse fato é importante ser levado em consideração. Além do limite de tempo, percebemos que é interessante modelar a proporção de tempo de resposta, uma vez

que o examinado pode distribuir o tempo total do teste entre todos os itens.

A partir dessas observações, desenvolvemos um modelo conjunto dentro do contexto Hierárquico e realizamos um breve estudo de simulção de recuperação dos parâmetros, além de uma aplicação do modelo em um conjunto de dados reais de Programme for International Student Assessment (PISA). Embora alguns estudos mais detalhados necessitem ser feitos com esse modelo, ele já se apresenta como uma boa alternativa para ser utilizado neste contexto.

CAPÍTULO

5

COMENTÁRIOS FINAIS E DESENVOLVIMENTOS FUTUROS

Neste capítulo, são apresentados comentários finais referentes aos resultados principais e às contribuições deste trabalho. Também são descritas as produções resultantes desta pesquisa. Por fim, são feitos comentários gerais sobre possíveis desenvolvimentos futuros.

5.1 Comentários Finais

Neste trabalho estudamos os modelos de regressão considerando variável resposta limitada no intervalo $(0, 1)$. Respostas com essa característica estão muito presentes no nosso dia a dia, no entando, os modelos para este tipo de conjunto de dados precisam ser mais explorados.

No capítulo 2 estudamos três modelos de regressão para resposta limitada já existentes na literatura, na qual pudemos observar algumas peculiaridades através de um estudo de simulação que foi desenvolvido como, por exemplo a eficiência do algoritmo NUTS e uma sensibilidade no modelo de regressão Simplex com a escolha da *priori* do parâmetro de dispersão. Além disso, nas aplicações realizadas pecebemos que os modelos de regressão Beta e L-Logistic tiveram um bom ajuste tanto para amostra pequena quanto para amostra grande. Por outro lado, não foi possível comparar o desempenho do modelo de regressão Simplex aos dois outros modelos na aplicação realizada com amostra pequena, somente com amostra grande. O algoritmo NUTS apresentou muitas vantagens em comparação ao desempenho dos pacotes utilizados na aplicação.

Já no capítulo 3, propusemos um novo modelo de regressão quantílico, que apresentou resultados muito promissores. Desenvolvemos uma estimativa Bayesiana utilizando o algoritmo NUTS, além de diversos estudos e aplicações de simulação. As aplicações foram realizadas utilizando conjunto de dados de pobreza nos municípios do Brasil, na qual consideramos o modelo com base na distribuição (sem a inclusão de covariáveis) e com a inclusão de covariáveis

de quatro formas diferentes. Os resultados foram comparados a outros modelos, que são casos particulares desta proposta, e mostraram-se muito satisfatórios, indicando que o modelo proposto é uma ótima alternativa para dados relacionados à pobreza nos municípios do Brasil.

Por fim, no capítulo 4 propusemos um modelo hierárquico dentro do contexto da Teoria de Resposta ao Item. De acordo com Fox *et al.* (2007), os tempos de resposta (TR) de um teste podem ser modelados em três perspectivas diferentes: A primeira forma de modelar TR é dentro da abordagem de TRI, para as variáveis de resposta para os mesmos itens; a segunda perspectiva trata de modelar TR de forma separada, ou seja, modela TR independentemente das variáveis de resposta para os itens; a terceira forma de modelar TR é dentro de uma estrutura hierárquica. Neste capítulo desenvolvemos um breve estudo de simulação em termos da recuperação de parâmetros, assim como uma aplicação a um conjunto de dados reais. Os resultados alcançados até o presente momento são bastante animadores, embora alguns estudos ainda necessitem ser desenvolvidos.

5.2 Produções

Neste momento, as produções associadas a este projeto são:

5.2.1 Trabalhos apresentados em eventos

- Poster: "new proposed regression model for bounded responses". 67^a Reunião Anual da Sociedade Internacional de Biometria e 20^o Simpósio de Estatística Aplicada à Experimentação Agronômica, 2023. Londrina, PR, Brasil.

- Poster: "LG-Logistic median regression model: a new proposal for bounded response". 7th Latin American Conference on Statistical Computing - LACSC, 2023. Lima, Peru.

- Poster: "Desempenho dos pacotes bayesianos para a estimação de modelos limitados". 66ª Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria (RBras), 2022. Florianópolis, SC, Brasil.

- Poster: "Modelo hierárquico Beta-Bernoulli para tempo de resposta e precisão". 24º SINAPE Simpósio Nacional de Probabilidade e Estatística, 2022. Gramado, RS, Brasil.

- Apresentação assíncrona: "Bounded Regression Models using NUTS Algorithm". IX WORKSHOP ON PROBABILISTIC AND STATISTICAL METHODS, 2022. São Carlos, SP, Brasil.

5.2.2 Artigos submetidos

- STULP, P., BAZÁN, J. L, VALDIVIESO, L. H. A new quantile regression model for bounded responses with applications. Computational Statistics, Trabalho submetido em 30 de setembro de 2023.

REFERÊNCIAS

ALKASASBEH, M. R.; RAQAB, M. Z. Estimation of the generalized logistic distribution parameters: Comparative study. **Statistical Methodology**, Elsevier, v. 6, n. 3, p. 262–279, 2009. Citado na página 48.

ANDERSON, D.; BURNHAM, K. Model selection and multi-model inference. **Second. NY: Springer-Verlag**, v. 63, n. 2020, p. 10, 2004. Citado na página 34.

BALAKRISHNAN, N.; LEUNG, M. Order statistics from the type i generalized logistic distribution. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 17, n. 1, p. 25–50, 1988. Citado na página 48.

BARNDORFF-NIELSEN, O. E.; JØRGENSEN, B. Some parametric models on the simplex. **Journal of multivariate analysis**, Elsevier, v. 39, n. 1, p. 106–116, 1991. Citado nas páginas 23 e 27.

BAYES, C. L.; BAZÁN, J. L.; CASTRO, M. D. A quantile parametric mixed regression model for bounded response variables. **Statistics and its interface**, International Press of Boston, v. 10, n. 3, p. 483–493, 2017. Citado nas páginas 23 e 56.

BOURGUIGNON, M.; GALLARDO, D. I.; SAULO, H. **A parametric quantile beta regression for modeling case fatality rates of COVID-19**. 2021. Disponível em: <https://arxiv.org/abs/2110.04428>. Citado na página 68.

BRASIL, A. **Atlas do desenvolvimento humano no Brasil**. 2010. Disponível em: <http://www.atlasbrasil.org.br/consulta/planilha>. Citado na página 61.

COLE, T. J.; GREEN, P. J. Smoothing reference centile curves: the lms method and penalized likelihood. **Statistics in medicine**, Wiley Online Library, v. 11, n. 10, p. 1305–1319, 1992. Citado na página 33.

CRIBARI-NETO, F.; ZEILEIS, A. Beta regression in r. **Journal of statistical software**, v. 34, n. 1, p. 1–24, 2010. Citado nas páginas 27 e 32.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996. Citado na página 63.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of applied statistics**, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004. Citado nas páginas 23, 27, 28, 56 e 70.

FLORES, S. *et al.* A hierarchical joint model for bounded response time and response accuracy. In: SPRINGER. **The Annual Meeting of the Psychometric Society**. [S.l.], 2019. p. 95–109. Citado nas páginas 24, 69, 70, 72 e 74.

FOX, J.-P. *et al.* Modeling of responses and response times with the package cirt. **Journal of Statistical software**, v. 20, n. 7, p. 1–14, 2007. Citado na página 82.

GEISSER, S.; EDDY, W. F. A predictive approach to model selection. **Journal of the American Statistical Association**, Taylor & Francis, v. 74, n. 365, p. 153–160, 1979. Citado na página 34.

GELMAN, A. *et al.* Understanding predictive information criteria for bayesian models. **Statistics and computing**, Springer, v. 24, n. 6, p. 997–1016, 2014. Citado na página 34.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical science**, Institute of Mathematical Statistics, v. 7, n. 4, p. 457–472, 1992. Citado na página 62.

HINKLEY, D. V. On power transformations to symmetry. **Biometrika**, Oxford University Press, v. 62, n. 1, p. 101–111, 1975. Citado na página 49.

INEI. Mapa de pobreza provincial y distrital 2009. el enfoque de la pobreza monetaria. dirección técnica de demografía e indicadores sociales. **Instituto Nacional de Estadística e Informática. Lima, Perú**, 2009. Citado na página 64.

KIESCHNICK, R.; MCCULLOUGH, B. D. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. **Statistical modelling**, Sage Publications Sage CA: Thousand Oaks, CA, v. 3, n. 3, p. 193–213, 2003. Citado na página 23.

KOENKER, R.; MACHADO, J. A. Goodness of fit and related inference processes for quantile regression. **Journal of the american statistical association**, Taylor & Francis, v. 94, n. 448, p. 1296–1310, 1999. Citado na página 47.

KORKMAZ, M. Ç. The unit generalized half normal distribution: A new bounded distribution with inference and application. Univ Politehnica Bucharest, 2020. Citado na página 24.

KORKMAZ, M. Ç.; CHESNEAU, C.; KORKMAZ, Z. S. On the arcsecant hyperbolic normal distribution. properties, quantile regression modeling and applications. **Symmetry**, MDPI, v. 13, n. 1, p. 117, 2021. Citado na página 24.

LINDEN, W. J. van der. A lognormal model for response times on test items. **Journal of Educational and Behavioral Statistics**, Sage Publications Sage CA: Los Angeles, CA, v. 31, n. 2, p. 181–204, 2006. Citado nas páginas 69 e 74.

LÓPEZ, F. O. A bayesian approach to parameter estimation in simplex regression model: A comparison with beta regression. **Revista Colombiana de Estadística**, Universidad Nacional de Colombia., v. 36, n. 1, p. 1–21, 2013. Citado nas páginas 23 e 29.

MARTINO, S.; RIEBLER, A. Integrated nested laplace approximations (inla). **Wiley StatsRef: Statistics Reference Online**, Wiley Online Library, p. 1–19, 2014. Citado na página 33.

MAZUCHELI, J.; ALVES, B.; MENEZES, A. F.; LEIVA, V. An overview on parametric quantile regression models and their computational implementation with applications to biomedical problems including COVID-19 data. **Computer Methods and Programs in Biomedicine**, Elsevier, v. 221, p. 106816, 2022. Citado nas páginas 24 e 47.

MAZUCHELI, J.; KORKMAZ, M. Ç.; MENEZES, A. F.; LEIVA, V. The unit generalized half-normal quantile regression model: Formulation, estimation, diagnostics, and numerical applications. **Soft Computing**, Springer, v. 27, n. 1, p. 279–295, 2023. Citado na página 24.

MOORS, J. A quantile alternative for kurtosis. **Journal of the Royal Statistical Society: Series D (The Statistician)**, Wiley Online Library, v. 37, n. 1, p. 25–32, 1988. Citado na página 50.

NADARAJAH, S.; SI, Y. A note on the "l-logistic regression models. **Brazilian Journal of Probability and Statistics**, JSTOR, v. 34, n. 1, p. 183–187, 2020. Citado na página 68.

OSPINA, R.; CRIBARI-NETO, F.; VASCONCELLOS, K. L. Improved point and interval estimation for a beta regression model. **Computational Statistics & Data Analysis**, Elsevier, v. 51, n. 2, p. 960–981, 2006. Citado na página 23.

PAOLINO, P. Maximum likelihood estimation of models with beta-distributed dependent variables. **Political Analysis**, Cambridge University Press, v. 9, n. 4, p. 325–346, 2001. Citado na página 23.

PAZ da *et al.* L-logistic regression models: Prior sensitivity analysis, robustness to outliers and applications. **Brazilian Journal of Probability and Statistics**, Brazilian Statistical Association, v. 33, n. 3, p. 455–479, 2019. Citado nas páginas 23, 27, 30, 33, 48, 58 e 60.

PAZ, R. F. da; BAZÁN, J. L.; MILAN, L. A. Bayesian estimation for a mixture of simplex distributions with an unknown number of components: Hdi analysis in Brazil. **Journal of Applied Statistics**, Taylor & Francis, v. 44, n. 9, p. 1630–1643, 2017. Citado na página 23.

PNUD. Informe sobre desarrollo humano perú 2009: Por una densidad del estado al servicio de la gente. **Programa de las Naciones Unidas para el Desarrollo. Lima, Perú**, 2009. Citado na página 64.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2018. Disponível em: <https://www.R-project.org/>. Citado nas páginas 32 e 74.

RIGBY, R. A.; STASINOPOULOS, D. A semi-parametric additive model for variance heterogeneity. **Statistics and Computing**, Springer, v. 6, n. 1, p. 57–65, 1996. Citado na página 33.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 54, n. 3, p. 507–554, 2005. Citado nas páginas 27, 32 e 33.

RIGBY, R. A.; STASINOPOULOS, M. D. Mean and dispersion additive models. **Statistical theory and computational aspects of smoothing**, Springer, p. 215–230, 1996. Citado na página 33.

ROBERT, B. M.; BRINDHA, G.; SANTHI, B.; KANIMOZHI, G.; PRASAD, N. R. Computational models for predicting anticancer drug efficacy: A multi linear regression analysis based on molecular, cellular and clinical data of oral squamous cell carcinoma cohort. **Computer methods and programs in biomedicine**, Elsevier, v. 178, p. 105–112, 2019. Citado na página 47.

RUE, H.; MARTINO, S.; CHOPIN, N. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. **Journal of the royal statistical society: Series b (statistical methodology)**, Wiley Online Library, v. 71, n. 2, p. 319–392, 2009. Citado nas páginas 27, 32 e 33.

SMITHSON, M.; VERKUILEN, J. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. **Psychological methods**, American Psychological Association, v. 11, n. 1, p. 54, 2006. Citado na página 23.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. V. D. Bayesian measures of model complexity and fit. **Journal of the royal statistical society: Series b (statistical methodology)**, Wiley Online Library, v. 64, n. 4, p. 583–639, 2002. Citado na página 34.

Stan Development Team. **RStan: the R interface to Stan**. 2020. R package version 2.21.2. Disponível em: <http://mc-stan.org/>. Citado nas páginas 27, 32, 34, 40, 41, 43, 58 e 65.

STASINOPOULOS, D. M.; RIGBY, R. A. *et al.* Generalized additive models for location scale and shape (gamlss) in r. **Journal of Statistical Software**, v. 23, n. 7, p. 1–46, 2007. Citado nas páginas 27, 32 e 33.

SU, Y.-S.; YAJIMA, M. Package R2jags. 2015. R package version 0.03-08. Disponível em: <http://CRAN.R-project.org/package=R2jags>. Citado nas páginas 74 e 76.

TADIKAMALLA, P. R.; JOHNSON, N. L. Systems of frequency curves generated by transformations of logistic variables. **Biometrika**, Oxford University Press, v. 69, n. 2, p. 461–465, 1982. Citado na página 23.

WATANABE, S.; OPPER, M. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. **Journal of machine learning research**, v. 11, n. 12, 2010. Citado na página 34.

YEE, T. W. *et al.* The vgam package for categorical data analysis. **Journal of Statistical Software**, Citeseer, v. 32, n. 10, p. 1–34, 2010. Citado nas páginas 27 e 33.

ZELTERMAN, D. Parameter estimation in the generalized logistic distribution. **Computational Statistics & Data Analysis**, Elsevier, v. 5, n. 3, p. 177–184, 1987. Citado na página 48.

ZHANG, P.; QIU, Z.; SHI, C. simplexreg: An r package for regression analysis of proportional data using the simplex distribution. **Journal of Statistical Software**, v. 71, n. 11, 2016. Citado nas páginas 27 e 33.

APÊNDICE

A

APÊNDICE DO CAPÍTULO 3

A.1 Propriedades da distribuição LG-Logistic

A.1.1 Quantil

Calculando a função inversa de (3.2), temos

$$
\begin{aligned}
& p = F(y \mid \cdot) = \left[1+\left(\frac{m(1-y)}{(1-m)y}\right)^{\tau}\right]^{-\alpha} \\
\Rightarrow \quad & p^{\frac{1}{\alpha}} = \frac{1}{1+\left(\frac{m(1-y)}{(1-m)y}\right)^{\tau}} \Rightarrow 1+\left(\frac{m}{1-m}\right)^{\tau}\left(\frac{1-y}{y}\right)^{\tau} = \frac{1}{p^{\frac{1}{\alpha}}} \\
\Rightarrow \quad & \left(\frac{m}{1-m}\right)^{\tau}\left(\frac{1-y}{y}\right)^{\tau} = \frac{1-p^{\frac{1}{\alpha}}}{p^{\frac{1}{\alpha}}} \Rightarrow \frac{1}{y}-1 = \left(\frac{\frac{1-p^{\frac{1}{\alpha}}}{p^{\frac{1}{\alpha}}}}{\left(\frac{m}{1-m}\right)^{\tau}}\right)^{\frac{1}{\tau}} \\
\Rightarrow \quad & \frac{1}{y} = \frac{\left(\frac{1-p^{\frac{1}{\alpha}}}{p^{\frac{1}{\alpha}}}\right)^{\frac{1}{\tau}}+\left(\frac{m}{1-m}\right)}{\left(\frac{m}{1-m}\right)} \Rightarrow y = \frac{\left(\frac{m}{1-m}\right)}{\left(\frac{m}{1-m}\right)+\left(\frac{1-p^{\frac{1}{\alpha}}}{p^{\frac{1}{\alpha}}}\right)^{\frac{1}{\tau}}}.
\end{aligned}
$$

Portanto

$$
\kappa_Y(p) = F_Y^{-1}(p) = \frac{\left(\frac{m}{1-m}\right)}{\left(\frac{m}{1-m}\right)+\left(\frac{1-p^{\frac{1}{\alpha}}}{p^{\frac{1}{\alpha}}}\right)^{\frac{1}{\tau}}}.
$$

Usando a reparametrização em (3.6), temos

$$
\kappa_Y(p) = F_Y^{-1}(p) = \frac{u}{u+\varphi}.
$$

A.1.2 Casos particulares da distribuição LG-Logistic distribution

A função acumulada da distribuição LG-Logistic é dada por

$$F(y \mid m,\tau,\alpha) = \left[1+\left(\frac{m(1-y)}{(1-m)y}\right)^{\tau}\right]^{-\alpha},$$

conforme proposto em (3.2). Tomando $\alpha = 1$, temos como caso particular a distribuição L-Logistic, cuja notação é $Y \sim \mathrm{LL}(m,\tau)$, e quando $m = 0.5$, temos como caso particular a distribuição G-Logistic, cuja notação é $Y \sim \mathrm{GL}(\tau,\alpha)$. Matematicamente, as distribuições têm as seguintes funções acumuladas, respectivamente,

$$\begin{aligned} F_{\mathrm{LL}}(y \mid m,\tau) &= \left[1+\left(\frac{m(1-y)}{(1-m)y}\right)^{\tau}\right]^{-1} \\ F_{\mathrm{GL}}(y \mid \tau,\alpha) &= \left[1+\left(\frac{1-y}{y}\right)^{\tau}\right]^{-\alpha} \end{aligned} \tag{A.1}$$

Considerando a reparametrização apresentada em (3.6), para $\alpha = 1$ no primeiro caso e $m = 0,5$ no segundo, e substituindo em (A.1), temos

$$\begin{aligned} F_{\mathrm{LL}}(y \mid \kappa,\varphi) &= \left[1+\left[\left(\frac{\kappa\varphi}{1-\kappa}\right)\left(\frac{1-y}{y}\right)\right]^{\frac{\log\left(\frac{1-p}{p}\right)}{\log(\varphi)}}\right]^{-1} \\ F_{\mathrm{GL}}(y \mid \phi,\varphi) &= \left[1+\left(\frac{1-y}{y}\right)^{\frac{\log\left(\frac{1-\phi}{\phi}\right)}{\log(\varphi)}}\right]^{-\frac{\log(p)}{\log(\phi)}} \end{aligned} \tag{A.2}$$

e

$$\begin{aligned} f_{LL}(y|\kappa,\varphi) &= \frac{\log\left(\frac{1-p}{p}\right)}{\log(\varphi)}\left[1+\left[\left(\frac{\kappa\varphi}{1-\kappa}\right)\left(\frac{1-y}{y}\right)\right]^{\frac{\log\left(\frac{1-p}{p}\right)}{\log(\varphi)}}\right]^{-2} \\ &\quad \times\left[\left(\frac{\kappa\varphi}{1-\kappa}\right)\left(\frac{1-y}{y}\right)\right]^{\frac{\log\left(\frac{1-p}{p}\right)}{\log(\varphi)}}\frac{1}{y(1-y)} \\ f_{GL}(y \mid \phi,\varphi) &= \frac{\log(p)\log\left(\frac{1-\phi}{\phi}\right)}{\log(\phi)\log(\varphi)}\left[1+\left(\frac{1-y}{y}\right)^{\frac{\log\left(\frac{1-\phi}{\phi}\right)}{\log(\varphi)}}\right]^{-\frac{\log(p)+\log(\phi)}{\log(\phi)}} \\ &\quad \times\left(\frac{1-y}{y}\right)^{\frac{\log\left(\frac{1-\phi}{\phi}\right)}{\log(\varphi)}}\frac{1}{y(1-y)} \end{aligned} \tag{A.3}$$

as funções de distribuição acumuladas e densidades L-Logistic e G-Logistic, respectivamente. Em termos de notação, temos $Y \sim \mathrm{LL}(k,\varphi)$ e $Y \sim \mathrm{GL}(\phi,\varphi)$.

A.1.3 Medidas de assimetria e curtose

Na Tabela 26, nós apresentamos algumas medidas de assimetria e curtose, usando (3.4) and (3.5) em (3.3), escolhendo alguns valores de m, τ e α.

Tabela 26 – Medidas de assimetria (OS) e curtose (OK) da distribuição LG-Logistic para alguns valores escolhidos de m, τ e α.

Par*	Valores escolhidos								
m	0,2	0,2	0,2	0,5	0,5	0,5	0,8	0,8	0,8
τ	0,6	5,0	10,0	0,6	5,0	10,0	0,6	5,0	10,0
α	0,6	0,6	0,6	0,6	0,6	0,6	0,6	0,6	0,6
OS	0,823	0,054	-0,026	0,533	-0,080	-0,096	-0,059	-0,226	-0,169
OK	1,645	1,295	1,311	0,846	1,299	1,321	0,602	1,359	1,347
m	0,2	0,2	0,2	0,5	0,5	0,5	0,8	0,8	0,8
τ	0,6	5,0	10,0	0,6	5,0	10,0	0,6	5,0	10,0
α	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
OS	0,555	0,115	0,058	0,000	0,000	0,000	-0,555	-0,115	-0,058
OK	1,020	1,307	1,307	0,723	1,290	1,302	1,020	1,307	1,307
m	0,2	0,2	0,2	0,5	0,5	0,5	0,8	0,8	0,8
τ	0,6	5,0	10,0	0,6	5,0	10,0	0,6	5,0	10,0
α	5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0	5,0
OS	-0,524	0,220	0,194	-0,695	0,127	0,149	-0,744	0,053	0,108
OK	1,179	1,304	1,296	1,609	1,261	1,276	1,838	1,245	1,263
m	0,2	0,2	0,2	0,5	0,5	0,5	0,8	0,8	0,8
τ	0,6	5,0	10,0	0,6	5,0	10,0	0,6	5,0	10,0
α	10,0	10,0	10,0	10,0	10,0	10,0	10,0	10,0	10,0
OS	-0,654	0,231	0,212	-0,710	0,140	0,168	-0,725	0,074	0,131
OK	1,491	1,300	1,295	1,702	1,255	1,273	1,772	1,237	1,258

* Parâmetro

A.1.4 Momentos

Usando a expressão em (3.1) e aplicando-a à definição de momentos, temos

$$\mathbb{E}[Y^t] = \int_0^1 y^t \alpha\tau \left[1+\left(\frac{m(1-y)}{(1-m)y}\right)^{\tau}\right]^{-(\alpha+1)} \left(\frac{m(1-y)}{(1-m)y}\right)^{\tau} \frac{1}{y(1-y)} dy. \tag{A.4}$$

Considerando a transformação $Z = \tau\log\left(\frac{y(1-m)}{(1-y)m}\right) \Rightarrow dZ = \frac{\tau}{y(1-y)}dy$, temos a função base G-Logística, pois $\exp(-Z) = \left(\frac{m(1-y)}{(1-m)y}\right)^{\tau}$, e então $F(Z) = [1+\exp(-Z)]^{-\alpha}$, $\alpha > 0$.

Fazendo alguns cálculos, temos $y = \frac{1}{1+\left(\frac{1-m}{m}\right)\exp\left(-\frac{Z}{\tau}\right)}$ e, fazendo as devidas substituições em (A.4), temos

$$\mathbb{E}[Y^t] = \int_{-\infty}^{\infty} \left(\frac{1}{1+\left(\frac{1-m}{m}\right)\exp\left(-\frac{Z}{\tau}\right)}\right)^t \alpha \frac{(\exp(Z))^{\alpha}}{(1+\exp(Z))^{\alpha+1}} dZ. \tag{A.5}$$

Seja $v = \frac{\exp(Z)}{1+\exp(Z)}$, então $Z = log\left(\frac{v}{1-v}\right)$ e, consequentemente, $dZ = \frac{1}{v(1-v)}dv$.

Retornando à equação (A.5), temos que

$$\mathbb{E}[Y^t] = \int_0^1 \left(1 + \left(\frac{1-m}{m}\right)\left(\frac{1-v}{v}\right)^{\frac{1}{\tau}}\right)^{-t} \alpha v^{\alpha-1} dv. \tag{A.6}$$

Como $u = \frac{m}{1-m}$ e $\beta = \frac{1}{\tau}$, a expressão em (A.6) pode ser reescrita da seguinte forma

$$\mathbb{E}[Y^t] = \int_0^1 \left(1 + \left(\frac{1}{u}\right)\left(\frac{1-v}{v}\right)^{\beta}\right)^{-t} \alpha v^{\alpha-1} dv.$$

Usando reparametrização em (3.6), temos

$$\mathbb{E}[Y^t] = \int_0^1 \left(1 + \left(\frac{1-\kappa}{\kappa\varphi}\right)\left(\frac{1-v}{v}\right)^{\frac{\log(\varphi)}{\log\left(\frac{1-\phi}{\phi}\right)}}\right)^{-t} \frac{\log(p)}{\log(\phi)} v^{\left(\frac{\log(p)}{\log(\phi)}-1\right)} dv. \tag{A.7}$$

A.1.5 Medidas de média e variância

A Tabela 27 apresenta, para $p = 0,5$, a média e a variância da distribuição LG-Logistic com alguns valores para κ, ϕ e φ parâmetros, usando (A.7)

Tabela 27 – Média e variância da distribuição LG-Logistic para algumas escolhas de κ, ϕ e φ.

Par*	Valores escolhidos							
κ	0,4	0,8	0,4	0,8	0,4	0,8	0,4	0,8
ϕ	0,2	0,2	0,4	0,4	0,2	0,2	0,4	0,4
φ	1,5	1,5	1,5	1,5	5,0	5,0	5,0	5,0
$\mathbb{E}[Y]$	0,393	0,758	0,431	0,700	0,437	0,645	0,477	0,574
$\mathbb{V}ar[Y]$	0,026	0,022	0,088	0,079	0,123	0,125	0,194	0,192

* Parâmetro

A.2 Resultados Aplicações

A.2.1 Resultados da aplicação da distribuição LG-Logistic

A mesma aplicação realizada utilizando $p = 0,5$ foi reproduzida aqui, usando as distribuições reparametrizadas L-Logistic (2), G-Logistic (2), LG-Logistic (2), considerando $p = 0,10,0,25,0,75,0,90$. Os resultados referente aos critérios DIC, EAIC, EBIC, IC, HQIC, WAIC e LOO, utilizados para fins de comparação entre modelos, são apresentados nas tabelas 28 e 29.

Na Tabela 28, considerando $p = 0,10$ e $p = 0,25$, é possível notar que, embora os valores sejam muito próximos, a distribuição LG-Logistic apresentou melhores resultados quanto aos valores dos critérios de comparação.

Tabela 28 – Resultdos dos critérios de comparação para dados de pobreza do Brasil, considerando $p = 0,10$ e $p = 0,25$

$p = 0,10$			
	L-Logistic	G-Logistic	LG-Logistic
DIC	-254,403	-264,360	-265,542
EAIC	-252,386	-261,913	-260,934
EBIC	-245,626	-255,153	-250,795
IC	-252,420	-262,806	-264,150
HQIC	-249,655	-259,183	-256,838
WAIC	-254,324	-264,353	-264,432
LOO	-254,322	-264,350	-264,428
$p = 0,25$			
	L-Logistic	G-Logistic	LG-Logistic
DIC	-254,429	-264,303	-264,965
EAIC	-252,394	-262,118	-261,372
EBIC	-245,634	-255,358	-251,232
IC	-252,465	-262,488	-262,559
HQIC	-249,663	-259,387	-257,276
WAIC	-254,349	-264,493	-264,822
LOO	-254,346	-264,491	-264,817

Já na Tabela 29 constam apenas os resultados da aplicação usando as distribuições reparametrizadas G-Logistic e LG-Logistic. Uma vez que a distribuição L-Logistic é caso particular da distribuição LG-Logistic quando $\frac{\log(p)}{\log(\phi)} = 1$ (ou seja, $p = \phi$), na equação (8), foi possível considerar apenas $p = 0,10$ e $p = 0,25$.

Os resultados da Tabela 29 mostram que, considerando $p = 0,75$, os resultados são melhores usando a distribuição G-Logistic e considerando $p = 0,90$, a distribuição LG-Logistic apresenta melhores resultados.

Tabela 29 – Resultados dos critérios de comparação para dados de pobreza do Brasil, considerando $p = 0,75$ and $p = 0,90$

	$p = 0,75$		$p = 0,90$	
	G-Logistic	LG-Logistic	G-Logistic	LG-Logistic
DIC	-264,428	-263,324	-198,114	-232,019
EAIC	-262,034	-259,684	-195,119	-228,115
EBIC	-255,275	-249,544	-188,359	-217,976
IC	-262,822	-260,963	-197,110	-229,923
HQIC	-259,304	-255,588	-192,388	-224,019
WAIC	-264,637	-263,404	-197,930	-231,160
LOO	-264,634	-263,401	-197,927	-231,167

A.2.2 Resultado da aplicação do modelo de regressão LG-Logistic

As quatro propostas de modelos apresentadas na seção 6.2 foram utilizadas neste material suplementar, a fim de comparar os modelos utilizando $p = 0,1, 0,25, 0,75, 0,90$. Os resultados

dos critérios de comparação utilizados são apresentados, respectivamente, na Tabela 30, Tabela 31, Tabela 32 e na Tabela 33.

Tabela 30 – Resultados dos critérios de comparação da aplicação dos dados de pobreza no Peru na família de modelos das distribuições reparametrizadas L-Logistic, G-Logistic e LG-Logistic, considerando $p = 0,10$.

	Família de modelos usando a distribuição LG-Logistic			
Critérios	Modelo nulo	Modelo de locação	Modelo de forma	Modelo completo
DIC	-244,107	-474,232	-390,671	-474,634
EAIC	-233,258	-466,139	-379,255	-463,635
EBIC	-223,439	-453,047	-366,163	-447,270
IC	-248,955	-474,325	-394,088	-475,634
HQIC	-229,283	-460,838	-373,954	-457,009
WAIC	-236,317	-470,051	-383,251	-468,349
LOO	-236,295	-470,044	-383,219	-468,323
	Família de modelos usando a distribuição G-Logistic			
Critérios	Modelo nulo	Modelo de escala	Modelo de forma	Modelo completo
DIC	-233,018	-233,516	-291,231	-291,818
EAIC	-227,574	-225,828	-286,982	-285,612
EBIC	-221,028	-216,009	-277,163	-272,520
IC	-234,463	-235,205	-289,479	-290,025
HQIC	-224,923	-221,852	-283,007	-280,311
WAIC	-230,071	-230,164	-290,350	-290,946
LOO	-230,044	-230,117	-290,265	-290,871
	Família de modelos usando a distribuição L-Logistic			
Critérios	Modelo nulo	Modelo de locação	Modelo de forma	Modelo completo
DIC	-196,290	-423,627	-218,100	-426,716
EAIC	-192,178	-418,445	-213,204	-420,286
EBIC	-182,359	-405,353	-200,112	-403,921
IC	-194,402	-420,808	-214,996	-423,145
HQIC	-188,202	-413,145	-207,903	-413,660
WAIC	-195,900	-422,348	-216,155	-424,233
LOO	-195,897	-422,346	-216,136	-424,198

Na Tabela 30 nota-se que o LG-Logistic apresenta melhores resultados quando comparado aos demais modelos sem incluir covariáveis (modelo nulo), assim como também o Modelo completo usando a distribuição LG-Logistic apresenta resutados melhores em relação ao modelo completo usando a distribuição G-Logistic e L-Logistic. Já quando inclui-se covariáveis ao parâmetro de locação (modelo de locação), o modelo LG-Logistic apresenta melhores resultados;

e o mesmo resultado é visto quando inclui-se covariáveis ao parâmetro de forma (modelo de forma).

Tabela 31 – Resultados dos critérios de comparação da aplicação dos dados de pobreza no Peru na família de modelos das distribuições reparametrizadas L-Logistic, G-Logistic e LG-Logistic, considerando $p = 0,25$.

	Família de modelos usando a distribuição LG-Logistic			
Critérios	Modelo nulo	Modelo de locação	Modelo de forma	Modelo completo
DIC	-239,004	-471,763	-326,869	-470,975
EAIC	-233,275	-465,985	-315,263	-463,731
EBIC	-223,456	-452,893	-302,171	-447,366
IC	-238,734	-469,541	-330,475	-468,219
HQIC	-229,299	-460,684	-309,962	-457,105
WAIC	-236,216	-469,729	-318,780	-468,269
LOO	-236,186	-469,721	-318,694	-468,236
	Família de modelos usando a distribuição G-Logistic			
Critérios	Modelo nulo	Modelo de escala	Modelo de forma	Modelo completo
DIC	-231,551	-231,633	-292,283	-292,684
EAIC	-227,942	-226,010	-288,785	-287,299
EBIC	-221,396	-216,191	-278,966	-274,207
IC	-231,160	-231,256	-289,780	-290,070
HQIC	-225,292	-222,035	-284,810	-281,998
WAIC	-229,918	-229,808	-291,953	-292,313
LOO	-229,829	-229,634	-291,736	-292,126
	Família de modelos usando a distribuição L-Logistic			
Critérios	Modelo nulo	Modelo de locação	Modelo de forma	Modelo completo
DIC	-196,112	-423,477	-194,427	-426,590
EAIC	-192,107	-418,367	-189,746	-420,335
EBIC	-182,288	-405,275	-176,654	-403,970
IC	-194,117	-420,588	-191,108	-422,846
HQIC	-188,132	-413,065	-184,445	-413,709
WAIC	-195,712	-422,161	-193,974	-424,262
LOO	-195,708	-422,158	-193,963	-424,226

Analisando os resultados da Tabela 31, percebe-se que as conclusões são muito semelhantes às anteriores. Ou seja, o modelo LG-Logistic apresenta melhores resultdos em relação aos demais, considerando $p = 0,25$, quando comparando os modelos sem covariável (modelo nulo), covariável no parâmetro de locação (modelo de locação), de forma (modelo de forma) e também em ambos (modelo completo).

Visto que o modelo L-Logistic é um caso particular do modelo LG-Logistic quando $\frac{\log(p)}{\log(\phi)} = 1$ (ou seja, $p = \phi$), em (8), então consideramos esse modelo somente para quando $p = 0,10, 0,25$, um vez que $\phi \in (0,0,5)$. Portanto, nas tabelas 32 e 33 encontram-se os resultados dos critérios de comparação para os modelos de regressão G-Logistic e LG-Logistic.

Comparando os resultados do modelo sem covariáveis (modelo nulo) e o modelo completo (modelo completo), na Tabela 32, percebemos que o modelo LG-Logistic apresenta melhores resultados. Por outro lado, os resultados do modelo incluindo covariáveis no parâmetro de forma (modelo de forma) são melhores no modelo G-Logistic.

Tabela 32 – Resultados dos critérios de comparação da aplicação dos dados de pobreza no Peru na família de modelos das distribuições reparametrizadas L-Logistic, G-Logistic e LG-Logistic, considerando $p = 0,75$.

	Família de modelos usando a distribuição LG-Logistic			
Critérios	Modelo nulo	Modelo de locação	Modelo de forma	Modelo completo
DIC	-236,079	-470,069	-281,687	-468,764
EAIC	-233,073	-465,849	-276,918	-463,420
EBIC	-223,254	-452,757	-263,826	-447,055
IC	-233,086	-466,285	-278,455	-464,109
HQIC	-229,098	-460,548	-271,617	-456,794
WAIC	-235,771	-469,626	-281,050	-468,007
LOO	-235,659	-469,618	-280,742	-467,966
	Família de modelos usando a distribuição G-Logistic			
Critérios	Modelo nulo	Modelo de escala	Modelo de forma	Modelo completo
DIC	-229,784	-272,967	-292,349	-296,162
EAIC	-227,794	-268,862	-289,164	-289,766
EBIC	-221,248	-259,043	-279,345	-276,674
IC	-227,775	-271,073	-289,534	-294,559
HQIC	-225,143	-264,886	-285,188	-284,465
WAIC	-229,096	-273,276	-292,066	-294,540
LOO	-228,880	-273,227	-291,804	-294,256

Na Tabela 33 pode-se notar que o modelo sem covariáveis (modelo nulo) e o modelo completo (modelo completo) apresentam melhores resultados ao utilizar o modelo LG-Logistic, assim como também o modelo de forma.

Tabela 33 – Resultados dos critérios de comparação da aplicação dos dados de pobreza no Peru na família de modelos das distribuições reparametrizadas L-Logistic, G-Logistic e LG-Logistic, considerando $p = 0,90$.

	Famíia de modelos usando a distribuição LG-Logistic			
Critérios	Modelo nulo	Modelo de locação	Modelo de forma	Modelo completo
DIC	-235,804	-466,775	-292,017	-463,412
EAIC	-231,712	-461,404	-286,588	-457,178
EBIC	-221,893	-448,312	-273,496	-440,813
IC	-233,895	-464,147	-289,446	-459,645
HQIC	-227,736	-456,103	-281,288	-450,552
WAIC	-234,948	-465,400	-291,532	-461,510
LOO	-234,805	-465,383	-291,455	-461,424
	Família de modelos usando a distribuição G-Logistic			
Critérios	Modelo nulo	Modelo de escala	Modelo de forma	Modelo completo
DIC	-230,153	-294,726	-291,060	-298,523
EAIC	-227,891	-290,568	-287,454	-293,238
EBIC	-221,345	-280,749	-277,635	-280,146
IC	-228,416	-292,890	-288,666	-295,807
HQIC	-225,240	-286,592	-283,478	-287,938
WAIC	-229,513	-295,185	-290,936	-297,757
LOO	-229,289	-295,123	-290,587	-297,381

A.3 Código

A.3.1 Moda

A solução da equação (3.10) não possui forma fechada, portanto desenvolvemos um código que foi implementado utilizando Visual Studio Code e escrito na linguagem de programação Python, que determina um valor numérico para a moda assumindo alguns valores de parâmetros.

Utilizamos a função minimize_scalar para determinar o valor mínimo da função $-f(x)$, apresentada em (3.7), ou seja, o valor máximo de $f(x)$, que é a moda de $f(x)$.

```
# INITIALIZING VARIABLES
controle = 0
k        = 0
phi      = 0
varphi   = 1
p        = 0

!pip install numpy scipy
# pip show numpy
!pip install numpy
import numpy as np
```

```
# DENSITY FUNCTION f(y)
def f(y):
return ( (np.log(p) * np.log( (1 - phi) / phi) ) / ( np.log(phi) * np.log(varphi) ) ) * \
( ( 1 + ( ( (k * varphi) / (1 - k) ) * ( (1 - y) / y ) ) ** ( np.log( (1 - phi) / phi ) /  \
np.log(varphi) ) ) ** (- ( np.log(p) + np.log(phi) ) / np.log(phi) ) ) * \
( ( ( (k * varphi) / (1 - k) ) * ( (1 - y) / y ) ) ** ( np.log( (1 - phi) / phi ) / \
np.log(varphi) ) ) * ( 1 / ( y * (1 - y) ) )

# WHILE LOOP
while controle == 0:

# ================================================================================================
print("\n SET COEFFICIENTS \n")

aux = 0 #-------------------------------------------------------------------------
while aux == 0:
k = float(input("ENTER A VALUE FOR k: "))

if 0 < k < 1:
aux = 1
else:
print("\n ==> VALUE IS OUTSIDE VALUE RANGE (0, 1) ")


aux = 0 #-------------------------------------------------------------------------
while aux == 0:
phi = float(input("ENTER A VALUE FOR phi : "))

if 0 <= phi <= 0.5:
aux = 1
else:
print("\n ==> VALUE IS OUTSIDE VALUE RANGE (0, 0.5) ")

aux = 0 #-------------------------------------------------------------------------
while aux == 0:
varphi = float(input("ENTER A VALUE FOR varphi : "))

if 1 < varphi:
aux = 1
else:
print("\n ==> VALUE IS OUTSIDE VALUE RANGE, GREATER THAN 1 ")


aux = 0 #-------------------------------------------------------------------------
while aux == 0:
p = float(input("ENTER A VALUE FOR p: "))
```

```
if 0 <= p <= 1:
aux = 1
else:
print("\n ==> VALUE IS OUTSIDE VALUE RANGE [0, 1] ")


# ============================================================================================================
# Define the domain
xmin = 0.0000001
xmax = 0.9999999

# Create a grid of x points
x_values = np.linspace(xmin, xmax, 1000)  # 1000 points between xmin and xmax

from scipy.optimize import minimize_scalar

# Find the maximum of the function -f(y)
result = minimize_scalar(lambda x: -f(x), bounds=(xmin, xmax), method = 'bounded')

print(result.x)

controle = int(input("\n WANT TO CALCULATE AGAIN? (0:yes ou 1:no): "))

print("\n END OF PROGRAM, COME BACK ALWAYS. \n")
```

A.3.2 Modelo

Apresentamos o programa STAN utilizado para implementar o modelo de regressão LGL completo, que foi aplicado em 3.5.2.

```
functions{
  real LGL_lpdf(real y, real u, real b, real a){
    real lprob;
    lprob = log(a) - log(b) -
    (a+1)*log(1 + pow((u*(1-y))/y, 1/b)) + (1/b)*(log(u) + log(1-y)
    - log(y)) - log(y) - log(1-y);
    return lprob;
 }
  real loglik_LGL(real[] y, real[] u, real[] b, real a){
    real lprob[num_elements(y)];
    for(i in 1:num_elements(y)){
      lprob[i] = log(a) - log(b[i]) -
      (a+1)*log(1 + pow(((u[i])*(1-y[i]))/y[i], 1/b[i]))
```

```
      + (1/b[i])*(log(u[i]) + log(1-y[i]) - log(y[i]))
      - log(y[i]) - log(1-y[i]);
    }
  return sum(lprob);
  }
}
data {
  int<lower = 0> n;
  real<lower=0,upper=1> y[n];
  vector[n] x;
  real<lower = 0,upper = 1> q;
  real L;
}
parameters {
  real beta0;
  real beta1;
  real delta0;
  real delta1;
  real lambda0;
}
transformed parameters{
  real k[n];
  real u[n];
  real b[n];
  real<lower=L> varphi[n];
  real a;
  real<lower = 0> phi;
  phi = exp(lambda0);

  a = log(q)/log(phi);

  for (i in 1:n){
    k[i]      = inv_logit(beta0 + beta1 * x[i]);
    varphi[i] = exp(-delta0-delta1*x[i]);
    u[i]      = k[i]*varphi[i]/(1-k[i]);
    b[i]      = log(varphi[i])/log((1-phi)/phi);
  }
}
model {
```

```
  beta0 ~ normal(0,10);
  beta1 ~ normal(0,10);
  delta0 ~ normal(0,10);
  delta1 ~ normal(0,10);
  lambda0 ~ normal(0,10);

  target += loglik_LGL(y, u, b, a);
}
generated quantities{
  real dev;
  real log_lik[n];
  dev = 0;

  for (i in 1:n)
  {
    log_lik[i] = LGL_lpdf(y[i] | u[i], b[i], a);
    dev = dev + (-2)*log_lik[i];
  }
}
```

APÊNDICE

B

APÊNDICE DO CAPÍTULO 4

B.1 Código

Apresentamos o código BUGS para o modelo Beta Normal Ogive, apresentado na seção 2.4.

```
model
{
   for(j in 1:J)
{
psi[j,1:4] ~ dmnorm(mu_i[], Omega_i[,])    # a, b, alpha, beta
a[j] <- exp(psi[j,1])    # a> 0, Bernoulli
alpha[j] <- exp(psi[j,3])    # alpha > 0, Beta
}
   for(i in 1:I)
{
theta[i] ~ dnorm(0,1) # Bernoulli
cm[i] <- rho*theta[i]
tau[i] ~ dnorm(cm[i],ctau)
for(j in 1:J)
{
u[i,j] ~ dbern(p[i,j])                 # Bernoulli
m[i,j] <- a[j]*(theta[i] - psi[j,2])   # Bernoulli
p[i,j] <- phi(m[i,j])                  # Bernoulli
z[i,j] ~ dbeta(mu[i,j]*alpha[j], alpha[j]*(1 - mu[i,j]))
     logit(mu[i,j]) <- psi[j,4]-tau[i]
}  }
```

```
rho ~ dnorm(0,0.1)
ctau ~ dgamma(0.1,0.1)
   mu_i[1:4] ~ dmnorm(mm[], Omega_m[,])
   Omega_m[1:4,1:4] ~ dwish(Ri[,],4)
   Omega_i[1:4,1:4] ~ dwish(Ri[,],4)
   Sigma_i[1:4,1:4] <- inverse(Omega_i[,])
}
```


ICMC
USP
SÃO CARLOS