UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Diagnóstico e seleção de modelos com resposta binária e função de ligação assimétrica

Fabiano Rodrigues Coelho

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

ICMC USP SÃO CARLOS

Fabiano Rodrigues Coelho

Diagnóstico e seleção de modelos com resposta binária e função de ligação assimétrica

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 Doutor em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientadora: Profa. Dra. Cibele Maria Russo Novelli
Coorientador: Prof. Dr. Jorge Luis Bazán Guzmán

USP – São Carlos
Fevereiro 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)

Rodrigues Coelho, Fabiano
R672d Diagnóstico e seleção de modelos com resposta binária e função de ligação assimétrica / Fabiano Rodrigues Coelho; orientadora Cibele Maria Russo Novelli; coorientador Jorge Luis Bazán Guzmán. -- São Carlos, 2024.
83 p.

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

1. função de ligação assimétrica. 2. dados desbalanceados. 3. modelos binários mistos. 4. análise de resíduos. 5. estimação bayesiana. I. Maria Russo Novelli, Cibele , orient. II. Luis Bazán Guzmán, Jorge, coorient. III. 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

Fabiano Rodrigues Coelho

Diagnostic and models selection with binary response and asymmetric link function

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Doctorate Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Profa. Dra. Cibele Maria Russo Novelli
Co-advisor: Prof. Dr. Jorge Luis Bazán Guzmán

USP – São Carlos
February 2024

Em memória de Odonélia Loura de Amorim Coelho, cuja bondade pura permanece como inspiração em nossos corações.

AGRADECIMENTOS

Agradeço a Deus, cuja presença esteve ao meu lado em toda a jornada. Em momentos de trevas, sempre senti Sua proteção, guiando-me através dos desafios. Ao enfrentar um mar amargo, onde a escuridão era densa e os lamentos ecoavam, o Senhor envolveu-me com Sua luz, permitindo-me atravessar o oceano. Sou profundamente grato a Deus por me conceder forças para chegar até aqui.

Expresso minha mais profunda gratidão aos meus pais, Celino Rodrigues Coelho e Cicera Aparecida da Silva Coelho, e à minha irmã, Daniele da Silva Coelho. Seu apoio constante e as energias positivas têm sido o alicerce da minha vida. Em meio aos altos e baixos, vocês estiveram sempre ao meu lado, oferecendo amor incondicional, orientação e incentivo. Suas presenças são pilares de força, e não há palavras suficientes para expressar o quanto valorizo e aprecio cada um de vocês.

À Eliandra de Mello Bonotto, que esteve presente em cada etapa desta jornada, compartilhando lágrimas e sorrisos. Você é uma daquelas pessoas que, ao olharmos, pensamos: deve ser um anjo enviado por Deus. Você é incrível. Obrigado por tornar minha vida melhor.

Minha orientadora, Prof[a]. Dra. Cibele Maria Russo Novelli, agradeço por todos esses anos de parceria, aprendizado, paciência, críticas construtivas e conselhos. Agradeço pela forma respeitosa com que sempre me tratou e por toda a empatia demonstrada quando passei pelo momento mais difícil de minha vida.

Meu Coorientador, Prof. Dr. Jorge Luis Bazán Guzmán, expresso minha mais profunda admiração e gratidão. Ao longo de nossa jornada de pesquisa, tive o privilégio de aprender com sua experiência e orientação, enriquecendo minha compreensão e habilidades na área. Estou imensamente grato por ter a oportunidade de trabalhar sob sua orientação.

Aos membros das bancas de Qualificação e Defesa deste trabalho: Prof. Dra. Cibele Maria Russo Novelli, Prof. Dr. Jorge Luis Bazán Guzmán, Prof. Dr. Jony Arrais Pinto Junior, Prof. Dra. Lizbeth Naranjo Albarrán, Prof. Dr. Marcos Oliveira Prates, Prof. Dr. Cristian Marcelo Villegas Lobos, agradeço pelas ponderações, intercâmbio de pensamentos e questionamentos que promoveram um notável progresso em minha carreira e vida pessoal.

Aos docentes do Departamento de Estatística da Universidade Federal de São Carlos (UFSCar) e do Departamento de Matemática Aplicada e Estatística do ICMC - USP, os quais eu tive a honra de ser instruído e aprender muito.

Aos meus amigos do PIPGEs, pessoas maravilhosas, parceiros de coração com a qual

pude dividir os momentos bons e ruins ao longo do doutorado, desejo tudo de bom a eles. Obrigado por tudo.

O servidor Julio Cezar de Barros e aos funcionários do serviço de Pós-Graduação do ICMC-USP, por todo suporte e solicitude quando foi necessário resolver algum assunto pendente.

A Hélio Azevedo e sua esposa, por terem uma das melhores e mais lindas características de um ser humano. Tal característica é a abnegação. Vocês fazem o seu melhor para ajudar sem esperar nada em troca. Vocês moram no meu coração.

Gostaria de expressar minha profunda gratidão e admiração à Equipe de Hematologia do Hospital das Clínicas de Ribeirão Preto pelo cuidado excepcional e pelo tratamento eficaz que recebi durante minha jornada de combate ao linfoma de Hodgkin. Cada membro dessa equipe demonstrou uma dedicação incomparável à minha saúde e bem-estar, desde os médicos até os enfermeiros e demais profissionais de saúde.

Ao CEPID-CeMEAI (Centro de Pesquisa, Inovação e Difusão do Centro de Ciências Matemáticas Aplicadas à Indústria), pela utilização do cluster Euler, sem ele, este trabalho se tornaria praticamente inviável e também aos seus funcionários que sempre que necessário, não deixaram de prestar auxílio, seja pessoalmente ou por e-mail.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

A todos com quem tive a oportunidade de conviver no período, muito obrigado por tudo.

RESUMO

COELHO, F. R. **Diagnóstico e seleção de modelos com resposta binária e função de ligação assimétrica**. 2024. 83 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.

Para variáveis resposta binárias, as funções de ligação probito e logito são amplamente utilizadas. No entanto, quando os dados são desbalanceados, as abordagens tradicionais podem não ser adequadas. Neste trabalho é considerado a função de ligação skew-probito como uma possível alternativa para modelos com resposta binária. Os parâmetros são estimados por meio de uma abordagem bayesiana utilizando Monte Carlo Hamiltoniano, e a análise de resíduos é desenvolvida. Além disso, uma extensão para o caso de modelos mistos é apresentada, com a estimação dos parâmetros sendo realizada por meio de integração numérica. Como aplicação prática, analisamos dois conjuntos de dados. Em ambas as aplicações, é possível verificar, por meio de critérios de seleção de modelos, que o modelo skew-probito é mais eficiente do que as abordagens tradicionais. Computacionalmente, para o modelo com efeitos fixos, utilizamos a linguagem Stan adaptada ao software R. No caso misto, consideramos a metodologia INLA. Propostas para trabalhos futuros também são discutidas.

Palavras-chave: função de ligação assimétrica, dados desbalanceados, modelos binários mistos, análise de resíduos e estimação bayesiana.

ABSTRACT

COELHO, F. R. **Diagnostic and models selection with binary response and asymmetric link function**. 2024. 83 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.

For binary response variables, probit and logit link functions are widely used. However, when the data is imbalanced, traditional approaches may not be suitable. In this thesis, we consider the skew-probit link function as a potential alternative for models with binary response. The parameters are estimated through a Bayesian approach using Hamiltonian Monte Carlo, and residual analysis is developed. Additionally, an extension for the case of mixed models is presented, with parameter estimation performed through numerical integration. As a practical application, we analyze two datasets. In both applications, it is possible to observe, through model selection criteria, that the skew-probit regression model is more efficient than traditional approaches. Computationally, for the fixed-effects model, we use the Stan language adapted to the R software. In the mixed case, the INLA methodology is considered. Proposals for future research are also discussed.

Keywords: asymmetric link function, imbalanced data, mixed binary models, residual analysis, and Bayesian estimation.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Em alguns momentos, temos o interesse em modelar duas situações distintas: sucesso ou falha. Isso pode ser exemplificado pela avaliação final de um estudante ao concluir um determinado curso, que resulta em sua aprovação ou reprovação, ou pela presença ou ausência do vírus influenza em uma pessoa. Nessas situações, a probabilidade de sucesso está associada a atributos de interesse que explicam a variável resposta. Para realizar essa associação, fazemos uso de funções de ligação. Quando lidamos com variáveis aleatórias binárias independentes, as funções de ligação mais comuns são o probito e o logito. Para informações mais detalhadas, veja, por exemplo, Meltzer *et al.* (2011) e Hailpern e Visintainer (2003).

Uma variável aleatória binária pode ser classificada com base na sua proporção de sucesso. Denotando por p a proporção de sucesso de uma variável aleatória binária X, podemos, por exemplo, afirmar que se $0,4 < p < 0,6$, então X é considerada balanceada. Se $p > 0,6$ ou $p < 0,4$, então X é considerada desbalanceada. Quando p se aproxima de zero ou um, estamos lidando com eventos raros. Um estudo realizado por Paal (2014) compara diferentes métodos para modelar eventos raros, enquanto King e Zeng (2001) propõe uma abordagem que utiliza a função de ligação logito para eventos raros.

No entanto, alguns trabalhos, como o de Chen (2004), indicam que as funções de ligação probito e logito não são satisfatórias para dados desbalanceados. Conforme mencionado por Collett (2003), em circunstâncias apropriadas, uma função de ligação assimétrica funciona melhor do que as funções de ligação convencionais. Portanto, diversos estudos surgiram na literatura com o objetivo de propor funções de ligação que solucionem esse problema. Podemos destacar, por exemplo, os trabalhos de Chen, Dey e Shao (1999), Basu e Mukhopadhyay (2000), Chen, Dey e Shao (2001), Bazán, Bolfarine e Branco (2010) e Wang, Dey *et al.* (2010).

Com o intuito de aprimorar os estudos envolvendo variáveis aleatórias com resposta binária e proporções de resposta que se encontram distantes nas duas categorias existentes, Chen, Dey e Shao (1999) propõe uma função de ligação alternativa sob a abordagem de dados

aumentados, denominada função de ligação CDS ("Covariate-Dependent Skewness"). De acordo com os autores, esse novo modelo é computacionalmente eficiente e a introdução de variáveis latentes torna o algoritmo MCMC eficaz e de fácil implementação. Isso evidencia que, ao assumir distribuições *a priori* não informativas adequadas, obtêm-se distribuições *a posteriori* apropriadas.

Bazán *et al.* (2006) apresenta uma nova versão da função de ligação skew-probito, denominada função de ligação skew-probito BBB, utilizando a teoria de resposta ao item. Essa versão inovadora introduz um novo parâmetro relacionado à assimetria das curvas de respostas ao item e define uma nova classe de funções de ligação assimétricas que controlam a probabilidade de sucesso. A função de ligação assimétrica proposta neste artigo está vinculada à função de ligação proposta por Chen, Dey e Shao (1999). Uma característica comum dos modelos skew-probito CDS e skew-probito BBB é que o parâmetro de assimetria está associado à função de distribuição acumulada da variável aleatória skew-normal, a qual induz a função de ligação.

Bazán, Bolfarine e Branco (2010) apresenta uma versão unificada de dois tipos de funções de ligação skew-probito existentes na literatura (BBB e CDS) e estabeleceu condições de existência para os estimadores de máxima verossimilhança e distribuições *a posteriori*, mesmo quando consideramos distribuições *a priori* impróprias. Também conclui que, por meio de diferentes critérios bayesianos, que funções de ligação assimétricas são uma alternativa viável para funções de ligação convencionais em situações específicas.

Outros estudos também exploraram a classe de funções de ligação skew-probito. Por exemplo, Farias e Branco (2011) propõe um amostrador de Gibbs no contexto do modelo skew-probito apresentado por Chen, Dey e Shao (1999) e conclui que esse método é mais eficiente do que abordagens convencionais.

Albert e Chib (1995) adapta o resíduo studentizado para uma abordagem bayesiana, considerando o uso de um resíduo obtido por meio de validação cruzada. A interpretação nesse caso se assemelha a uma abordagem frequentista. Farias, Branco *et al.* (2012) propõe dois novos resíduos para o modelo skew-probito (função de ligação CDS). O primeiro é uma generalização do resíduo latente simétrico, e o segundo resíduo é baseado na representação estocástica da distribuição skew-normal, seguindo uma distribuição uniforme. Esse último resíduo é baseado no método da transformada inversa em resíduos latentes convencionais.

Recentemente, Lee e Sinha (2019) considera a função de ligação skew-probito BBB ao abordar e investigar a identificabilidade dos parâmetros. Concluiu que, ao considerarmos um modelo de regressão sem variáveis explicativas, os parâmetros não são identificáveis. Por outro lado, ao incluirmos covariáveis no modelo de regressão, ele se torna identificável se as variáveis explicativas forem contínuas. No entanto, para covariáveis binárias, o modelo permanece não identificável. Os autores propõem três penalizações para a função de verossimilhança e as comparam por meio de um estudo de simulação. Além disso, apresentam uma aplicação baseada em dados cardíacos.

Naranjo, Pérez e Martín (2019) nos traz uma abordagem bayesiana para modelos de regressão com dados binários em problemas de classificação. Funções de ligação assimétricas são consideradas. Para evitar problemas computacionais, a estimação de parâmetros foi obtida usando dados aumentados. O estudo de simulação demonstra que a abordagem proposta é mais eficiente do que a abordagem tradicional para dados com problemas de classificação.

Niekerk e Rue (2021) discute a problemática da identificabilidade dos parâmetros no modelo skew-probito. Nesse sentido, ele realiza uma reformulação do intercepto do modelo. Adicionalmente, é apresentada uma nova padronização da função de ligação assimétrica, o que facilita a interpretação dos parâmetros do modelo. Outra contribuição relevante consiste na sugestão de uma nova distribuição *a priori* para o parâmetro de assimetria. Os resultados dessa pesquisa podem ser acessados no pacote R-INLA, disponível na linguagem de programação R.

Outras possibilidades de funções de ligação assimétrica para modelos de regressão com resposta binária podem ser consideradas. Por exemplo, Alves, Bazán e Arellano-Valle (2023) apresentam novas funções de ligação cloglog flexíveis para modelos de regressão binomial, incorporando um parâmetro adicional que explica a assimetria na variável resposta binomial. A abordagem de inferência Bayesiana de Monte Carlo de cadeia de Markov é desenvolvida, e simulações demonstram o desempenho do algoritmo proposto. Uma análise de sensibilidade destaca a conveniência de uma distribuição *a priori* uniforme para todos os modelos. Duas aplicações em dados médicos (idade na menarca e infecção pulmonar) ilustram as vantagens dos modelos propostos.

Na literatura científica, encontramos exemplos, na Teoria da Resposta ao Item (TRI), de curvas características de item assimétricas. Alves e Bazán (2022) propõem novos modelos assimétricos de Teoria da Resposta ao Item que têm a Curva Característica de Item (CCI) assimétrica como sua característica principal. Um caso especial desses modelos é o modelo TRI cloglog. A estimação bayesiana dos modelos propostos é discutida, e uma aplicação em dados educacionais ilustra os benefícios da nova CCI quando comparada com outros modelos de TRI propostos na literatura.

Ordoñez *et al.* (2023) introduz uma *priori* de complexidade penalizada (*priori* PC) para o parâmetro de assimetria desta família, o que é útil para lidar com dados desbalanceados. Uma expressão geral para essa densidade é obtida e demonstramos sua utilidade para alguns casos particulares, como as funções de ligação potência probito e potência logito. Um estudo de simulação e uma aplicação em dados reais são utilizados para avaliar a eficiência das densidades introduzidas em comparação com outras distribuições *a priori*. Os resultados mostram melhoria na estimação pontual e intervalos de credibilidade para os modelos considerados ao usar o *priori* PC em comparação com outras distribuições *a priori* padrão bem conhecidas.

Quando consideramos problemas de classificação, Huayanay (2023) propõe funções de ligação alternativas ao logito, que é a mais usual. Tais distribuições sugeridas são a distribuição de potência (P) e potência inversa (RP). Novas propriedades dessas distribuições em contextos

de modelos para classificação em dados desbalanceados são exploradas. Realizam-se estudos de simulação para avaliar métricas de classificação, apresenta-se uma aplicação prática e estende-se os modelos para casos mistos em estudos longitudinais. A abordagem bayesiana foi adotada para a estimação de parâmetros, utilizando o algoritmo No-U-Turn Sampler (NUTS) em um procedimento MCMC. Adicionalmente, foram avaliadas métricas de desempenho preditivo, resíduos bayesianos e uma medida de influência bayesiana para diagnóstico de modelos. A comparação entre diferentes modelos foi realizada por meio de critérios de seleção de modelo.

1.1 Objetivo e Organização da Tese

Consideramos o modelo de regressão com resposta binária e a função de ligação skew-probito. O objetivo é desenvolver técnicas de diagnóstico e seleção de modelos para essa classe de modelos e estender o modelo para o caso em que estamos interessados em modelar a correlação entre grupos ou indivíduos. Esses tópicos são fundamentais para que os pesquisadores possam modelar seus dados da forma mais parcimoniosa possível. Por essa razão, estudos envolvendo modelos de regressão binária com funções de ligação assimétricas têm ganhado crescente popularidade entre os estatísticos.

A tese está estruturada da seguinte forma: no Capítulo 2, revisamos a distribuição skew-normal e suas principais propriedades e trazemos alguns conceitos relevantes que vão utilizados ao longo do trabalho. Em seguida, definimos o modelo skew-probito, enfatizando a importância da distribuição skew-normal que induz a função de ligação skew-probito. Destacamos ainda que, para diferentes valores do vetor de parâmetros θ, resultam em diferentes tipos de função de ligação skew-probito. Por fim, discutimos a estimação bayesiana, introduzindo uma nova abordagem que utiliza uma função de verossimilhança original, em contraste com outros artigos sobre estimação bayesiana para a classe de modelos em questão.

No Capítulo 3, realizam-se três estudos de simulação. No primeiro, investiga-se a sensibilidade das possíveis distribuições *a priori* para os parâmetros relacionados à assimetria do modelo, avaliando a recuperação desses parâmetros. Isso permite verificar se a nova abordagem proposta para a estimação dos parâmetros do modelo funciona satisfatoriamente. No segundo estudo de simulação, comparamos o algoritmo NUTS e o método INLA, utilizando o RMSE (raiz do erro quadrático médio) e o Viés, a fim de responder a questões de grande relevância, tais como qual dos dois métodos é mais eficiente computacionalmente no processo de estimação e qual função de perda deve ser considerada nesse procedimento. No terceiro estudo de simulação, comparamos o resíduo studentizado e o resíduo quantílico aleatorizado, verificando qual deles apresenta melhor desempenho na detecção de observações outliers induzidas.

No Capítulo 4, ilustramos os resultados obtidos nos capítulos anteriores por meio de uma aplicação utilizando um conjunto de dados sobre diabetes. Inicialmente, realizamos a seleção de modelos para determinar o modelo a ser adotado. Destaca-se que o modelo skew-

probito é comparado com abordagens tradicionais para modelos de resposta binária. Em seguida, identificamos observações outliers e investigamos seu impacto nas estimativas do modelo escolhido após a seleção. Adicionalmente, analisamos a sensibilidade das distribuições *a priori* para a aplicação em questão, utilizando DIC e WAIC.

No Capítulo 5, apresenta-se uma extensão do modelo skew-probito que incorpora efeitos aleatórios. Realiza-se a estimação bayesiana hierárquica e uma aplicação com o conjunto de dados de esquizofrenia de Madras é apresentada. Por fim, avalia-se o desempenho preditivo dos modelos utilizados na aplicação, empregando métricas como Acurácia, Revocação, Precisão, F1-Score e a área sob a curva ROC.

No Capítulo 6, são apresentados comentários finais, contribuições científicas e sugestões para pesquisas futuras.

CAPÍTULO

2

MODELO DE REGRESSÃO SKEW-PROBITO

Neste capítulo, é apresentado o modelo skew-probito. Inicialmente, é fornecida uma breve síntese contendo alguns resultados sobre a distribuição skew-normal. Por fim, são revisitados alguns conceitos preliminares relevantes para as análises subsequentes.

2.1 Distribuição skew - normal

Uma variável aleatória D segue uma distribuição skew-normal com parâmetro $\boldsymbol{\theta} = \left(\mu, \sigma^2, \alpha\right)^\top$, onde $\mu \in \mathbb{R}$ é o parâmetro de locação, $\sigma^2 \in \mathbb{R}^+$ é o parâmetro de escala e $\alpha \in \mathbb{R}$ é um parâmetro de assimetria. A função de densidade de probabilidade de D é definida por

$$f_{\boldsymbol{\theta}}(d) = \frac{2}{\sigma}\phi\left(\frac{d-\mu}{\sigma}\right)\Phi\left(\alpha\frac{d-\mu}{\sigma}\right)\mathbb{1}_{\mathbb{R}}(d) \tag{2.1}$$

onde $\mathbb{1}$ é uma função indicadora, ϕ é a função de densidade de probabilidade da distribuição normal padrão e Φ é a função de distribuição acumulada da distribuição normal padrão. Essa distribuição é representada como $D \sim \mathrm{SN}(\boldsymbol{\theta})$. Uma parametrização alternativa pode ser obtida substituindo α por

$$\delta = \frac{\alpha}{\sqrt{1+\alpha^2}} \in [-1,1].$$

Vamos agora definir uma distribuição de probabilidade que é fundamental na construção das principais propriedades da distribuição skew - normal.

Dizemos que uma variável aleatória X tem distribuição half-normal com média 0 e variância 1 se sua função densidade de probabilidade pode ser escrita como $g(x) = 2\phi(x), x > 0$, e é denotada por $H(0,1)$. Conforme destacado em Azzalini (1985), podemos listar algumas propriedades da distribuição skew-normal:

1. **Função Geradora de Momentos:** $M_D(t) = \Phi(\delta\sigma t)\, e^{\mu t + \frac{1}{2}\sigma t}$.

2. **Esperança:** $\mathbb{E}(D) = \mu + \sqrt{\frac{2}{\pi}}\delta\sigma$.

3. **Variância:** $Var(D) = \left(1 - \frac{2}{\pi}\delta^2\right)\sigma^2$.

4. **Relação com outras distribuições:** Se $W \sim N(0,1)$ e $V \sim HN(0,1)$ variáveis aleatórias independentes e $|\delta| < 1$, então $D = \mu + \sigma\left[\delta V + \left(1-\delta^2\right)^{\frac{1}{2}} W\right] \sim SN\left(\mu, \sigma^2, \alpha\right)$ e $D|V \sim N\left(\mu + \sigma\delta v, \left(1-\delta^2\right)\sigma^2\right)$.

5. **Índice de assimetria:** $\gamma = \left(\frac{2}{\pi}\right)^{\frac{3}{2}}\left(2 - \frac{\pi}{2}\right) sign(\delta)\frac{\delta^3}{\left(1-\frac{2}{\pi}\delta^2\right)^{\frac{3}{2}}} \in [-0{,}995; 0{,}995]$.

Quando $\alpha = 0$, a função densidade de probabilidade de D é reduzida à densidade de uma distribuição normal com média μ e variância σ^2. Tomando $\mu = 0$ e $\sigma^2 = 1$, obtemos a distribuição skew-normal padrão, denotada por $S \sim SN_A(\alpha)$, e a densidade de probabilidade é dada por:

$$f_\alpha^A(s) = 2\phi(s)\Phi(\alpha s). \tag{2.2}$$

onde o suporte de s é a reta real. A Figura 1 mostra o gráfico da função densidade de probabilidade de uma variável aleatória que segue a distribuição skew-normal padrão para diferentes valores do parâmetro de assimetria. Podemos observar que, quando $\alpha > 0$, a assimetria está à direita do gráfico, enquanto se $\alpha < 0$, a assimetria está à esquerda.

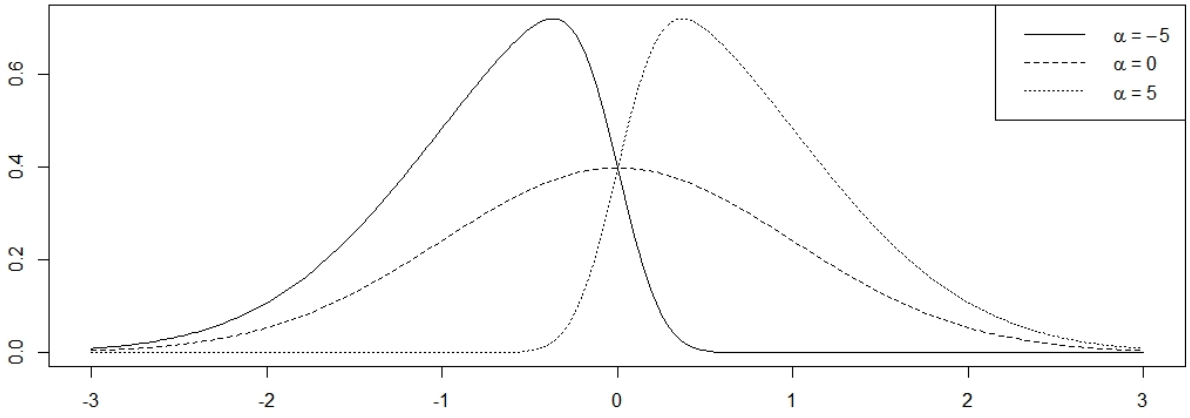


Figura 1 – Gráfico da função densidade de probabilidade da distribuição skew-normal padrão para diferentes valores do parâmetro de assimetria. Fonte: Elaborada pelo autor.

Na próxima seção, o modelo skew-probito será definido.

2.2 Modelo skew - probito

Seja $\boldsymbol{y} = (y_1, y_2, ..., y_n)^\top$, onde

$$y_i = \begin{cases} 1 \text{ com probabilidade } p_i; \\ 0 \text{ com probabilidade } 1 - p_i. \end{cases} \tag{2.3}$$

O vetor $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^\top$ é uma representação das covariáveis associadas à observação i, onde k é o número de covariáveis, e $0 < p_i < 1$ representa a probabilidade de sucesso. A matriz de dados X possui n linhas e $k+1$ colunas. Definimos $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ como o preditor linear, onde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\top$ é o vetor de coeficientes. De acordo com Bazán, Bolfarine e Branco (2010), a classe das funções de ligação skew-probito pode ser obtida considerando $F_{\boldsymbol{\theta}}^{-1}$ na especificação da probabilidade de sucesso, como sendo

$$p_i = F_{\boldsymbol{\theta}}(\eta_i), i = 1, 2, ..., n, \tag{2.4}$$

onde $F_{\boldsymbol{\theta}}(\cdot)$ é a função de distribuição acumulada da distribuição skew-normal com o vetor de parâmetros $\boldsymbol{\theta} = (\mu, \sigma^2, \alpha)$. O conjunto de equações (2.3) e (2.4) forma o modelo skew-probito, onde $\mu \in \mathbb{R}$, $\sigma^2 > 0$, e $\alpha \in \mathbb{R}$.

De acordo com Bazán, Bolfarine e Branco (2010), para valores específicos do vetor de parâmetros $\boldsymbol{\theta}$, diferentes configurações de função de ligação podem ser obtidas. Por exemplo:

- Se $\boldsymbol{\theta} = (0, 1, 0)$, então a função de ligação probito é obtida.
- Se $\boldsymbol{\theta} = (0, 1 + \alpha^2, -\alpha)$, então a função de ligação skew-probito CDS é obtida em Chen, Dey e Shao (1999).
- Se $\boldsymbol{\theta} = (0, 1, \alpha)$, então a função de ligação skew-probito BBB é obtida em Bazán *et al.* (2006).
- Se considerarmos os parâmetros μ e σ^2 para serem estimados através do conjunto de dados, obtemos uma função de ligação mais geral que pode ser chamada de função de ligação skew-probito completa.
- Se $\boldsymbol{\theta} = \left(-\frac{\sqrt{2}\delta}{\sqrt{\pi - 2\delta^2}}, \frac{\pi}{\pi - 2\delta^2}, \alpha\right)$, então obtemos a função de ligação skew - probito padrão, onde $\delta = \alpha/\sqrt{1+\alpha^2}, -1 \leq \delta \leq 1$. Observe que a escolha dos parâmetros é tal que a variável skew-normal associada tenha média 0 e variância 1.

A função de ligação padrão skew-probito é o objeto de estudo neste momento. Para aprimorar o desempenho computacional no processo de estimação, o vetor de parâmetros associado à variável aleatória assimétrica que induz a função de ligação pode ser reescrito como

$$\boldsymbol{\theta} = \left(-\frac{\sqrt{2}\alpha}{\sqrt{\pi + (\pi - 2)\,\alpha^2}}, \frac{\pi + \pi\alpha^2}{\pi + (\pi - 2)\,\alpha^2}, \alpha \right). \tag{2.5}$$

A validade da expressão (2.5) pode ser confirmada substituindo $\delta = \alpha/\sqrt{1+\alpha^2}$ na expressão original do vetor de parâmetros $\boldsymbol{\theta}$, conforme apresentado em Bazán, Bolfarine e Branco (2010).

2.3 Estimação Bayesiana

A função de verossimilhança para o modelo skew-probito é definida pela equação

$$L(\boldsymbol{\beta}, \alpha | \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{n} [F_\alpha(\eta_i)]^{y_i} [1 - F_\alpha(\eta_i)]^{1-y_i} \tag{2.6}$$

onde η_i representa o preditor linear, e $\mathbf{X}$ é a matriz de dados. Ambos os elementos são definidos acima. Neste modelo, $\boldsymbol{\beta}$ é o vetor de coeficientes associados a covariável e α é o parâmetro relacionado à assimetria do modelo.

Neste estudo, a estimação dos parâmetros é feita utilizando a função de verossimilhança original descrita em (2.6). No entanto, outra abordagem pode ser considerada, usando a função de máxima verossimilhança aumentada. Tal abordagem foi inicialmente utilizada por Chen, Dey e Shao (1999). Bazán, Bolfarine e Branco (2010) nos fala que a abordagem de dados aumentados induz um modelo hierárquico bayesiano que é muito conveniente na implementação do modelo de regressão com resposta binária e função de ligação skew-probito usando o software WinBUGS.

Vamos considerar duas parametrizações diferentes. A primeira é em relação ao parâmetro α, e a segunda em relação ao parâmetro δ.

2.3.1 Parametrização em relação a α.

Assumindo independência entre os parâmetros, podemos reescrever a distribuição *a priori* conjunta, como

$$\pi(\boldsymbol{\beta}, \alpha) = \pi(\boldsymbol{\beta})\pi(\alpha). \tag{2.7}$$

Consideramos que $\beta_j \sim N(\mu_{\beta_j}, \sigma^2_{\beta_j})$. Mais especificamente, se há pouca informação a respeito dos parâmetros β_j, consideremos $\mu_{\beta_j} = 0$ e $\sigma^2_{\beta_j} = 10000$. Sendo assim, a estrutura

hierárquica do modelo de regressão com resposta binária e função de ligação skew-probito pode ser escrita como

$$\begin{aligned} &y_i|\boldsymbol{\beta},\alpha \overset{\text{ind.}}{\sim} \text{Bernoulli}(p_i),\\ &p_i = F_\alpha(\eta_i),\\ &\eta_i = \beta_0 + \textstyle\sum_{j=1}^{k}\beta_j x_{ji},\\ &\beta_j \sim N(0,10000), \quad j = 1,\ldots,k,\\ &\alpha \sim \pi(\alpha). \end{aligned} \tag{2.8}$$

onde $i = 1,2,\ldots,n$. A função de distribuição *a posteriori*, associada com a estrutura descrita em (2.8), pode ser escrita como

$$\pi(\boldsymbol{\beta},\alpha|y,X) \propto \left(\prod_{i=1}^{n}\left[F_\alpha(\eta_i)\right]^{y_i}\left[1-F_\alpha(\eta_i)\right]^{1-y_i}\right)\left(\prod_{j=0}^{k} e^{-\frac{(\beta_j)^2}{20000}}\right)\pi(\alpha) \tag{2.9}$$

A parametrização com relação a δ pode ser discutida de maneira análoga à parametrização com relação a α.

2.3.2 Parametrização em relação a δ.

Seguindo o mesmo raciocínio utilizado na seção anterior, expressões análogas a (2.7), (2.8) e (2.9) podem ser definidas. Assumindo independência entre as distribuições *a priori*, obtemos que

$$\pi(\boldsymbol{\beta},\delta) = \pi(\boldsymbol{\beta})\pi(\delta). \tag{2.10}$$

A estrutura hierárquica deste modelo usando a parametrização em relação a δ pode ser expressa como:

$$\begin{aligned} &y_i|\boldsymbol{\beta},\delta \overset{\text{ind.}}{\sim} \text{Bernoulli}(p_i),\\ &p_i = F_\delta(\eta_i),\\ &\eta_i = \beta_0 + \textstyle\sum_{j=1}^{k}\beta_j x_{ji},\\ &\beta_j \sim N(0,10000), \quad j = 1,\ldots,k,\\ &\delta \sim \pi(\delta). \end{aligned} \tag{2.11}$$

onde $i = 1,2,\ldots,n$. A densidade da distribuição *a posteriori*, pode ser escrita como

$$\pi(\boldsymbol{\beta},\alpha|y,X) \propto \left(\prod_{i=1}^{n}\left[F_\alpha(\eta_i)\right]^{y_i}\left[1-F_\alpha(\eta_i)\right]^{1-y_i}\right)\left(\prod_{j=0}^{k} e^{-\frac{(\beta_j)^2}{20000}}\right)\pi(\alpha) \tag{2.12}$$

O espaço paramétrico para δ é o intervalo $[-1,1]$. A seguir, temos alguns conceitos importantes, que são fundamentais para entender as análises do próximos capítulos.

2.4 Conceitos Preliminares

Esta seção aborda conceitos fundamentais essenciais para a compreensão das análises estatísticas nos capítulos subsequentes. Exploramos brevemente tópicos como resíduos, métricas para avaliar estudos de simulação e métodos de seleção de modelo, estabelecendo uma base teórica crucial. Esses fundamentos desempenharão um papel vital na análise e interpretação dos resultados ao longo da tese, contribuindo para uma estrutura coesa na condução da pesquisa.

2.4.1 Métricas para avaliar estudo de simulação.

Para recuperar os verdadeiros valores dos parâmetros, são considerados o RMSE (Raiz do Erro Quadrático Médio), o viés e a probabilidade de cobertura (que representa a frequência com que o verdadeiro valor do parâmetro se encontra no intervalo de credibilidade com 95% de credibilidade). O RMSE, também conhecido como raiz do erro médio quadrático, pode ser calculado da seguinte maneira

$$RMSE\left(\hat{\nu}\right)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\nu}^{(i)}-\nu\right)^{2}} \tag{2.13}$$

e

$$vies\left(\hat{\nu}\right)=\frac{1}{n}\sum_{i=1}^{n}\hat{\nu}^{(i)}-\nu, \tag{2.14}$$

onde n é o número de réplicas e ν é o parâmetro genérico de interesse, em nosso caso, os coeficientes do vetor $\boldsymbol{\beta}$ ou parâmetros associados a assimetria do modelo.

2.4.2 Alguns critérios de seleção de modelo.

2.4.2.1 DIC(deviance information criterion)

Spiegelhalter *et al.* (2002) define o DIC (deviance information criterion) como

$$\widehat{DIC}=\overline{D\left(\boldsymbol{\beta},\alpha\right)}+p_D=2\overline{D\left(\boldsymbol{\beta},\alpha\right)}-D\left(\hat{\boldsymbol{\beta}},\hat{\alpha}\right),$$

onde p_D é o número efetivo de parâmetros, é dado por

$$p_D=\overline{D\left(\boldsymbol{\beta},\alpha\right)}-D\left(\hat{\boldsymbol{\beta}},\hat{\alpha}\right),$$

onde $D\left(\hat{\boldsymbol{\beta}}, \hat{\alpha}\right)$ é o desvio da média *a posteriori*. Os valores são obtidos através de um processo MCMC e são dados por

$$D\left(\hat{\boldsymbol{\beta}}, \hat{\alpha}\right) = D\left(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\beta}^{(k)}, \frac{1}{K}\sum_{k=1}^{K}\alpha^{(k)}\right),$$

onde K é o número de iterações num processo MCMC. O desvio bayesiano é dado por

$$D(\boldsymbol{\beta}, \alpha) = -2\log f(\mathbf{y}|\boldsymbol{\theta}) = -2\sum_{i=1}^{n}\log f(y_i|\boldsymbol{\theta}),$$

onde $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha)^{\top}$ e

$$f(y_i|\boldsymbol{\theta}) = [F_\alpha(\eta_i)]^{y_i}[1 - F_\alpha(\eta_i)]^{1-y_i}.$$

Note que $F_\alpha(\eta_i)$ é definido anteriormente neste capítulo.

2.4.2.2 WAIC (Widely Applicable Information Criterion)

De acordo com Vehtari, Gelman e Gabry (2017), o logaritmo da densidade preditiva pontual pode ser estimada como

$$lpd = \sum_{i=1}^{n}\log f(y_i|\mathbf{y}) = \sum_{i=1}^{n}\log\left[\int f(y_i|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta}\right] = \sum_{i=1}^{n}\log\left[\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}(f(y_i|\boldsymbol{\theta}))\right].$$

Para calcular *lpd*, saídas MCMC da distribuição *a posteriori* dos parâmetros em questão são usadas. Por isso,

$$\widehat{lpd} = \sum_{i=1}^{n}\log f(y_i|\mathbf{y}) = \sum_{i=1}^{n}\log\left[\frac{1}{K}\sum_{k=1}^{K} f\left(y_i|\boldsymbol{\theta}^{(k)}\right)\right],$$

onde K é o número de réplicas de um processo de MCMC. O logaritmo da densidade preditiva pontual esperada pode ser estimado por $\widehat{elpd} = \widehat{lpd} - \hat{p}$, onde $\hat{p}$ é o número efetivo de parâmetros estimados, que pode ser calculado pela expressão

$$\hat{p} = \sum_{i=1}^{n} Var_{post}\left(\log\left(f\left(y_i|\boldsymbol{\theta}^{(k)}\right)\right)\right).$$

Watanabe (2010) define o critério de seleção WAIC, através da expressão

$$\widehat{WAIC} = -2\widehat{lpd}.$$

O Deviance Information Criterion (DIC) e Widely Applicable Information Criterion (WAIC), são generalizações do Akaike information criterion (AIC) usando o desvio *a posteriori* junto com um fator de penalização. Assim, numa abordagem bayesiana, o uso do DIC e WAIC é preferível. (ver Gelman *et al.* (2013)).

2.4.3 Resíduos.

2.4.3.1 Resíduo studentizado.

Conforme destacado por Lesaffre e Lawson (2012) [p. 292], o resíduo $y_i - \mu_i$ mensura o desvio da resposta observada y_i em relação à probabilidade preditiva μ_i. Na prática, uma vez que μ_i é desconhecido, em um contexto bayesiano, sua esperança *a posteriori* é obtida, e o resíduo bayesiano ordinário pode ser definido como $y_i - \mathbb{E}(y_i|\mathbf{y})$ enquanto um resíduo padronizado pode ser calculado como

$$t_i = \frac{y_i - \mathbb{E}(y_i|\mathbf{y})}{\sqrt{Var(y_i|\mathbf{y})}}.$$

Conforme sabemos, a quantidade t_i converge em distribuição para uma distribuição normal padrão (Yan e Sedransk (2010)). Portanto, consideramos uma observação como outlier quando o resíduo studentizado padronizado t_i excede $1,96$ em módulo. No contexto do modelo skew-probito padrão, o resíduo studentizado pode ser calculado como

$$t_i^* = \frac{y_i - \widehat{p}_i}{\sqrt{\widehat{p}_i(1-\widehat{p}_i)}},$$

onde $\hat{p}_i = F_{\hat{\boldsymbol{\theta}}}(\hat{\eta}_i)$, $\hat{\eta}_i = x_i^\top \hat{\boldsymbol{\beta}}$ e

$$\hat{\boldsymbol{\theta}} = \left(-\frac{\sqrt{2}\hat{\alpha}}{\sqrt{\pi + (\pi-2)\hat{\alpha}^2}}, \frac{\pi + \pi\hat{\alpha}^2}{\pi + (\pi-2)\hat{\alpha}^2}, \hat{\alpha} \right).$$

2.4.3.2 Resíduo quantílico aleatorizado.

Vamos considerar outro tipo de resíduo, conhecido como resíduo quantílico padronizado, que foi proposto por Dunn e Smyth (1996). Esse resíduo pode ser definido por

$$r_{q,i} = \Phi^{-1}(u_i), i = 1,2,...,n.$$

No contexto do modelo skew-probito padrão, u_i é um valor aleatório obtido através de uma distribuição uniforme no intervalo

$$\left[I_{1-\widehat{p}_i}(2-y_i, y_i), I_{1-\widehat{p}_i}(1-y_i, y_i+1) \right], i = 1,2,...,n,$$

onde

$$I_x(a,b) = \frac{B(x;a,b)}{B(a,b)}$$

é a função beta regularizada, e $\Phi(.)$ é a função de distribuição acumulada de uma distribuição normal padrão. Note que

$$B(x;a,b) = \int_0^x t^{a-1}(1-t)^{b-1}\,dt$$

e

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt.$$

Segundo Dunn e Smyth (1996), o modelo de regressão apresenta ajuste satisfatório dos parâmetros quando os resíduos seguem uma distribuição aproximadamente normal padrão. Atkinson (1985) sugere a utilização do gráfico do envelope simulado para decidir se o modelo ajustado é adequado para o conjunto de dados. O resíduo quantílico aleatorizado segue uma distribuição assintoticamente normal padrão; assim, o critério para detecção de outliers é o mesmo que o resíduo studentizado descrito anteriormente.

2.4.4 Métricas de Desempenho Preditivo.

Nesta seção, apresentamos algumas métricas de desempenho preditivo e analisamos seu desempenho nos modelos propostos. Essas métricas são detalhadamente abordadas em James *et al.* (2013) e Hastie *et al.* (2009).

As métricas de desempenho preditivo são ferramentas essenciais na avaliação e validação de modelos de machine learning e estatísticos. Elas auxiliam na determinação de quão bem um modelo é capaz de realizar previsões precisas em novos dados, possibilitando que os praticantes avaliem a eficácia do modelo em relação aos objetivos específicos do problema. A seguir, apresentaremos algumas das métricas de desempenho preditivo mais comuns.

Matriz de Confusão (Confusion Matrix): A matriz de confusão é uma tabela que descreve o desempenho de um modelo de classificação em detalhes. Ela mostra o número de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos.

Tabela 1 – Matriz de confusão padrão.

	Positivo	Negativo
Positivo (Previsão)	TP	FP
Negativo (Previsão)	FN	TN

Fonte: Elaborada pelo autor.

A Tabela 1 representa a Matriz de Confusão Padrão, com quatro células que são interpretadas da seguinte forma:

- Verdadeiros Positivos (True Positives - TP): São os casos em que o modelo previu corretamente a classe positiva (verdadeiros positivos) - ou seja, acertou.
- Falsos Positivos (False Positives - FP): São os casos em que o modelo previu incorretamente a classe positiva quando na verdade era negativa (falsos positivos) - ou seja, errou.
- Falsos Negativos (False Negatives - FN): São os casos em que o modelo previu incorretamente a classe negativa quando na verdade era positiva (falsos negativos) - ou seja, errou.
- Verdadeiros Negativos (True Negatives - TN): São os casos em que o modelo previu corretamente a classe negativa (verdadeiros negativos) - ou seja, acertou.

Com base na matriz de confusão, é possível calcular diversas métricas de desempenho que proporcionam informações valiosas sobre a eficácia do seu modelo.

Acurácia (Accuracy): Mede a proporção de previsões corretas em relação ao total de previsões, sendo frequentemente utilizada em problemas de classificação binária. Entretanto, a acurácia pode ser enganosa em dados desbalanceados. Pode ser obtida por meio da matriz de confusão usando a expressão

$$\frac{TP+TN}{TP+FP+TN+FN}.$$

Precisão (Precision): Avalia a proporção de verdadeiros positivos entre todas as previsões positivas, sendo relevante quando se deseja evitar falsos positivos. Pode ser calculada a partir da matriz de confusão usando a expressão

$$\frac{TP}{TP+FP}.$$

Revocação (Recall) e Sensibilidade (Sensitivity): Medem a proporção de positivos corretamente identificados pelo modelo em relação ao total de positivos na amostra. São úteis quando é crucial evitar falsos negativos, como em testes de diagnóstico médico. Podem ser calculadas a partir da matriz de confusão usando a expressão

$$\frac{TP}{TP+FN}.$$

F1-Score: É a média harmônica entre revocação e precisão, sendo útil para equilibrar essas duas métricas, especialmente em problemas de classificação desbalanceada.

Curva ROC (Receiver Operating Characteristic Curve) e Área sob a Curva ROC (AUC-ROC): A curva ROC representa graficamente o desempenho do modelo em diversos pontos de corte, enquanto a AUC-ROC representa a área sob essa curva, oferecendo uma medida do desempenho preditivo do modelo. São particularmente valiosas em situações de classificação binária.

Essas são apenas algumas das métricas de desempenho preditivo mais comuns, e a escolha da métrica adequada depende do tipo de problema enfrentado. É fundamental selecionar as métricas mais relevantes para os objetivos do projeto e interpretá-las corretamente para tomar decisões informadas sobre o desempenho do modelo.

Nos próximos dois capítulos, exploraremos os modelos de regressão, que foram implementados por meio da linguagem de programação Stan. Essa implementação foi realizada de maneira eficiente utilizando o pacote *rstan* (Stan Development Team (2019)), que adapta a linguagem de programação R para integrar-se perfeitamente ao Stan. Esse ambiente emprega o avançado algoritmo de amostragem No-U-Turn (NUTS) para obter distribuições *a posteriori* simuladas, garantindo resultados robustos e confiáveis.

No capítulo subsequente, dedicaremos nossa discussão à configuração das distribuições *a priori*, particularmente para o parâmetro de assimetria, que será crucial nos ajustes futuros. Além disso, abordaremos outros aspectos relevantes para uma compreensão abrangente e aplicação eficaz dos modelos discutidos, proporcionando uma visão completa e coesa do tema. Para mais detalhes a respeito do algoritmo NUTTS, sugerimos a leitura do Apêndice D

CAPÍTULO

3

ESTUDOS DE SIMULAÇÃO

No capítulo anterior, apresentamos a proposta de considerar a função de verossimilhança original em conjunto com o algoritmo No-U-Turn (NUTS) para a estimação bayesiana dos parâmetros. Este capítulo, por sua vez, dedica-se a apresentar três estudos de simulação com o objetivo de avaliar a eficácia do método de estimação proposto. Esses estudos buscam verificar a capacidade do método em recuperar os verdadeiros valores dos parâmetros, além de explorar a sensibilidade da distribuição *a priori* em relação aos parâmetros associados à assimetria.

A avaliação do método proposto inclui uma comparação com o método conhecido como Integrated Nested Laplace Approximation (INLA). Por fim, investigamos uma abordagem eficaz para detectar observações outliers. Nesse contexto, realizamos uma análise comparativa de desempenho entre dois resíduos amplamente reconhecidos na detecção de outliers, aplicados ao modelo skew-probito padrão. Esses estudos proporcionam uma compreensão abrangente da consistência do método proposto, considerando diferentes cenários de simulação e comparando-o com uma abordagem estabelecida na literatura.

Ao longo deste capítulo, calculamos algumas métricas, como o RMSE e o viés, cujas expressões podem ser obtidas pelas equações (2.13) e (2.14), respectivamente.

3.1 Estudo comparativo da recuperação de parâmetros no processo de estimação, utilizando diferentes distribuições *a priori* para o parâmetro de assimetria.

Nesta seção, apresenta-se um estudo de simulação com o propósito de avaliar tanto a recuperação dos parâmetros quanto a sensibilidade das distribuições *a priori*. O objetivo deste estudo é determinar a melhor configuração de distribuições *a priori* para os parâmetros do modelo descrito nas equações (2.3) e (2.4), ao mesmo tempo que avaliamos a eficácia do método

de estimação na recuperação dos valores verdadeiros desses parâmetros.

A configuração do modelo utilizada para gerar os dados foi previamente descrita no capítulo anterior. Supomos que $\delta \sim U(-1,1)$ ou $\alpha \sim N(0,\sigma_\alpha^2)$. Três valores distintos para σ_α^2 são considerados: 1, 4, e 100. Exploramos, ainda, cenários nos quais utilizamos distribuições *a priori* informativas. Se houver conhecimento empírico sobre o parâmetro δ, é possível contemplar distribuições *a priori* informativas. Por exemplo, se $\delta < 0$, podemos supor que $\delta \sim U(-1,0)$ ou $\delta \sim U(0,1)$, caso $\delta > 0$. Outras possibilidades também podem ser consideradas, como a hipótese de δ ser próximo de 1, o que nos levaria a assumir $\delta \sim \text{Beta}(7,1)$.

Vamos considerar duas covariáveis independentes simuladas, representadas por $x_{1i} \sim U(-3,3)$ e $x_{2i} \sim Bernoulli(0,7)$, onde $i = 1,2,..,n$. Fixamos os coeficientes $\beta_0 = 0,1$, $\beta_1 = 0,5$ e $\beta_2 = -0,65$. Através dos vetores x_{1i} e x_{2i}, geramos 1000 réplicas de uma variável aleatória Bernoulli y_i por meio do modelo skew-probito padrão, considerando $\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ como o preditor linear.

Nesse contexto, exploramos oito cenários distintos que envolvem quatro tamanhos amostrais diferentes, $n = \{50, 500, 1000, 5000\}$, e dois valores distintos para o parâmetro $\delta = \{0,71, 0,98\}$. Essa abordagem nos permite investigar as variações nos resultados conforme alteramos o tamanho da amostra e o valor do parâmetro δ em oito situações distintas.

Uma variável de resposta desbalanceada, com uma baixa proporção de uns é esperada quando $\alpha = 4,5$, enquanto uma variável resposta balanceada é esperada quando $\alpha = 1$, com uma proporção de valores iguais a 1 próxima de $0,5$. Ressalta-se que o cenário oposto, representado por $\alpha = -4,5$, é omitido, visto que é análogo ao caso em que $\alpha = 4,5$.

A Tabela 2 apresenta o viés calculado para a estimativa dos parâmetros β_0, β_1, β_2, α, e δ, variando as configurações das distribuições *a priori* de α e δ. Já na Tabela 3, são exibidas as probabilidades de cobertura dos parâmetros β_0, β_1, β_2, α, e δ, considerando distintos tamanhos amostrais e configurações das distribuições *a priori*. Os intervalos com 95% de credibilidade são calculados para cada réplica.

Na Tabela 4, encontram-se os valores do RMSE (raiz do erro quadrático médio) para os parâmetros β_0, β_1, β_2, α, e δ, variando as configurações das distribuições *a priori* para α e δ. Os parâmetros são estimados no estudo de simulação com base na abordagem bayesiana e uma extensão do método Hamiltoniano Monte Carlo (HMC), conhecida como algoritmo No-U-Turn (NUTS), conforme proposto por Hoffman, Gelman *et al.* (2014). A convergência das cadeias de Markov é avaliada utilizando a estatística $\widehat{R}$, conforme sugerido por Gelman, Rubin *et al.* (1992).

Tabela 2 – Viés das estimativas dos parâmetros α, β_0, β_1, β_2 e δ para diferentes tamanhos amostrais e configurações de distribuições *a priori*.

$\alpha = 4,5$										
	$n = 50$					$n = 500$				
Prioris	α	δ	β_0	β_1	β_2	α	δ	β_0	β_1	β_2
$\delta \sim U(-1,1)$	3,459	0,978	0,494	0,116	0,092	3,449	0,575	0,363	0,017	0,043
$\alpha \sim N(0,100)$	1,180	0,981	0,475	0,279	0,269	1,907	0,657	0,453	0,076	0,045
$\alpha \sim N(0,4)$	4,598	1,215	0,586	0,187	0,198	3,108	0,384	0,304	0,025	0,011
$\alpha \sim N(0,1)$	4,526	1,038	0,606	0,089	0,023	4,081	0,639	0,422	0,018	0,050
$\delta \sim U(0,1)$	2,120	0,276	0,324	0,139	0,097	2,648	0,210	0,275	0,019	0,047
$\delta \sim \text{Beta}(7,1)$	0,803	0,041	0,102	0,203	0,146	1,141	0,043	0,100	0,015	0,018
	$n = 1000$					$n = 5000$				
$\delta \sim U(-1,1)$	2,537	0,311	0,263	0,013	0,027	0,459	0,014	0,034	0,005	0,005
$\alpha \sim N(0,100)$	0,239	0,283	0,196	0,027	0,023	0,241	0,002	0,010	0,001	0,003
$\alpha \sim N(0,4)$	2,214	0,154	0,178	0,016	0,016	0,732	0,014	0,034	0,009	0,011
$\alpha \sim N(0,1)$	3,602	0,379	0,362	0,012	0,038	1,797	0,043	0,100	0,018	0,019
$\delta \sim U(0,1)$	2,058	0,144	0,214	0,010	0,030	0,395	0,010	0,026	0,005	0,007
$\delta \sim \text{Beta}(7,1)$	0,834	0,032	0,085	0,011	0,011	0,285	0,007	0,020	0,005	0,006
$\alpha = 1$										
	$n = 50$					$n = 500$				
Prioris	α	δ	β_0	β_1	β_2	α	δ	β_0	β_1	β_2
$\delta \sim U(-1,1)$	0,301	0,745	0,083	0,095	0,106	1,019	0,737	0,130	0,005	0,001
$\alpha \sim N(0,100)$	1,995	0,782	0,019	0,289	0,307	2,580	0,921	0,312	0,139	0,120
$\alpha \sim N(0,4)$	1,260	0,931	0,137	0,155	0,190	1,083	0,762	0,143	0,069	0,066
$\alpha \sim N(0,1)$	1,064	0,777	0,130	0,068	0,072	1,018	0,724	0,128	0,004	0,003
	$n = 1000$					$n = 5000$				
$\delta \sim U(-1,1)$	0,952	0,671	0,123	0,003	0,010	0,705	0,499	0,089	0,025	0,036
$\alpha \sim N(0,100)$	1,752	0,732	0,205	0,109	0,102	0,402	0,377	0,035	0,022	0,018
$\alpha \sim N(0,4)$	0,847	0,627	0,112	0,056	0,054	0,475	0,398	0,049	0,012	0,006
$\alpha \sim N(0,1)$	0,900	0,629	**0,110**	0,000	0,000	0,652	0,463	0,083	0,013	0,023

Fonte: Elaborada pelo autor.

Tabela 3 – Probabilidade de cobertura dos intervalos com 95% de credibilidade associados aos parâmetros α, β_0, β_1, β_2 e δ para diferentes tamanhos amostrais e configurações de distribuições *a priori*.

$\alpha = 4.5$										
	$n = 50$					$n = 500$				
Prioris	α	δ	β_0	β_1	β_2	α	δ	β_0	β_1	β_2
$\delta \sim U(-1,1)$	0,186	0,186	0,723	0,855	0,827	0,576	0,576	0,728	0,989	0,948
$\alpha \sim N(0,100)$	0,518	0,518	0,474	0,532	0,511	0,726	0,726	0,712	0,766	0,923
$\alpha \sim N(0,4)$	0,110	0,110	0,791	0,878	0,852	0,590	0,590	0,834	0,987	0,954
$\alpha \sim N(0,1)$	0,000	0,000	0,744	0,926	0,908	0,000	0,000	0,509	0,996	0,945
$\delta \sim U(0,1)$	0,555	0,555	0,731	0,777	0,772	0,682	0,682	0,822	0,974	0,939
$\delta \sim Beta(7,1)$	0,684	0,684	0,655	0,663	0,655	0,929	0,929	0,925	0,946	0,928
	$n = 1000$					$n = 5000$				
$\delta \sim U(-1,1)$	0,704	0,704	0,772	0,992	0,958	0,910	0,910	0,929	0,971	0,949
$\alpha \sim N(0,100)$	0,837	0,837	0,841	0,857	0,933	0,939	0,939	0,942	0,946	0,945
$\alpha \sim N(0,4)$	0,724	0,724	0,860	0,988	0,957	0,891	0,891	0,924	0,947	0,943
$\alpha \sim N(0,1)$	0,000	0,000	0,477	0,995	0,967	0,074	0,074	0,707	0,911	0,945
$\delta \sim U(0,1)$	0,727	0,727	0,833	0,984	0,958	0,922	0,922	0,942	0,963	0,952
$\delta \sim Beta(7,1)$	0,923	0,923	0,913	0,942	0,949	0,937	0,937	0,950	0,949	0,948
$\alpha = 1$										
	$n = 50$					$n = 500$				
Prioris	α	δ	β_0	β_1	β_2	α	δ	β_0	β_1	β_2
$\delta \sim U(-1,1)$	0,936	0,936	0,886	0,901	0,885	0,992	0,992	0,976	0,997	0,977
$\alpha \sim N(0,100)$	0,625	0,625	0,608	0,586	0,592	0,668	0,668	0,681	0,710	0,915
$\alpha \sim N(0,4)$	0,970	0,970	0,937	0,908	0,905	0,989	0,989	0,981	0,980	0,963
$\alpha \sim N(0,1)$	0,994	0,994	0,924	0,939	0,931	0,999	0,999	0,981	0,997	0,974
	$n = 1000$					$n = 5000$				
$\delta \sim U(-1,1)$	0,981	0,981	0,980	0,999	0,980	0,962	0,962	0,969	1,000	0,992
$\alpha \sim N(0,100)$	0,695	0,695	0,709	0,743	0,897	0,948	0,948	0,948	0,977	0,987
$\alpha \sim N(0,4)$	0,975	0,975	0,975	0,993	0,962	0,964	0,964	0,963	0,989	0,994
$\alpha \sim N(0,1)$	0,997	0,997	0,984	0,998	0,972	0,972	0,972	0,974	1,000	0,995

Fonte: Elaborada pelo autor.

Tabela 4 – RMSE dos parâmetros α, β_0, β_1, β_2 e δ para diferentes tamanhos amostrais e configurações de distribuições *a priori*.

$\alpha = 4,5$										
	$n = 50$					$n = 500$				
Prioris	α	δ	β_0	β_1	β_2	α	δ	β_0	β_1	β_2
$\delta \sim U(-1,1)$	6,118	1,061	1,068	0,388	1,094	4,823	0,684	0,415	0,054	0,184
$\alpha \sim N(0,100)$	10,359	1,386	1,231	0,437	1,202	7,784	1,125	0,812	0,145	0,216
$\alpha \sim N(0,4)$	4,899	1,364	1,352	0,461	1,515	3,349	0,654	0,396	0,061	0,191
$\alpha \sim N(0,1)$	4,549	1,064	1,129	0,353	1,047	4,099	0,700	0,449	0,053	0,181
$\delta \sim U(0,1)$	4,705	0,310	0,916	0,444	1,027	6,278	0,245	0,334	0,057	0,182
$\delta \sim Beta(7,1)$	7,000	0,057	0,830	0,451	1,042	4,144	0,056	0,208	0,057	0,194
	$n = 1000$					$n = 5000$				
$\delta \sim U(-1,1)$	3,911	0,442	0,323	0,041	0,132	1,198	0,045	0,080	0,021	0,059
$\alpha \sim N(0,100)$	5,294	0,735	0,507	0,094	0,151	1,224	0,027	0,058	0,021	0,059
$\alpha \sim N(0,4)$	2,467	0,362	0,252	0,042	0,136	1,007	0,028	0,070	0,021	0,059
$\alpha \sim N(0,1)$	3,636	0,450	0,382	0,041	0,128	1,850	0,054	0,121	0,025	0,062
$\delta \sim U(0,1)$	4,682	0,188	0,269	0,040	0,125	1,141	0,021	0,067	0,020	0,060
$\delta \sim Beta(7,1)$	3,810	0,047	0,156	0,042	0,132	1,065	0,016	0,060	0,020	0,060
$\alpha = 1$										
	$n = 50$					$n = 500$				
Prioris	α	δ	β_0	β_1	β_2	α	δ	β_0	β_1	β_2
$\delta \sim U(-1,1)$	3,772	0,815	0,623	0,282	0,744	1,239	0,837	0,217	0,055	0,166
$\alpha \sim N(0,100)$	10,313	1,243	1,065	0,421	1,043	5,670	1,283	0,630	0,165	0,238
$\alpha \sim N(0,4)$	1,751	1,099	0,668	0,333	0,821	1,766	1,039	0,311	0,123	0,199
$\alpha \sim N(0,1)$	1,100	0,801	0,543	0,249	0,666	1,118	0,812	0,196	0,050	0,164
	$n = 1000$					$n = 5000$				
$\delta \sim U(-1,1)$	1,195	0,792	0,186	0,046	0,121	0,914	0,623	0,129	0,038	0,067
$\alpha \sim N(0,100)$	4,251	1,132	0,508	0,135	0,180	1,137	0,695	0,180	0,037	0,065
$\alpha \sim N(0,4)$	1,644	0,948	0,274	0,075	0,144	1,039	0,675	0,159	0,031	0,061
$\alpha \sim N(0,1)$	1,052	0,753	0,160	0,037	0,119	0,908	0,632	0,127	0,029	0,060

Fonte: Elaborada pelo autor.

Para determinar a melhor configuração da distribuição *a priori* para os parâmetros de assimetria, avaliam-se o RMSE mais baixo, o menor viés e a probabilidade de cobertura mais próxima a 95%. Se o RMSE diminui, o viés diminui e a probabilidade de cobertura aumenta à medida que o tamanho da amostra cresce, indicando que o processo de estimação sugerido é eficaz na recuperação dos verdadeiros valores dos parâmetros. Além disso, analisa-se em qual distribuição *a priori* os critérios de seleção de modelo apresentam os menores valores.

A eficácia da probabilidade de cobertura é maior quanto mais próxima estiver do coeficiente de confiança considerado. Por exemplo, uma probabilidade de cobertura de 0,94 é melhor do que 0,99 se o coeficiente de confiança for 0,95.

De acordo com as Tabelas 2, 3 e 4, o método de estimação demonstra um bom desempenho na recuperação dos valores reais dos parâmetros, uma vez que o RMSE diminui à medida que o tamanho da amostra aumenta, independentemente da escolha da distribuição *a priori*. O viés também diminui com o aumento do tamanho da amostra, independentemente da distribuição *a priori* considerada. A probabilidade de cobertura aumenta à medida que o tamanho da amostra aumenta e atinge estabilidade acima de $0,90$, oscilando ligeiramente acima desse nível. Nota-se que, na maioria dos casos, a probabilidade de cobertura não atinge $0,95$.

O RMSE para a estimação dos parâmetros δ e α é mais alto do que o RMSE para $\boldsymbol{\beta}$. Independentemente da escolha da distribuição *a priori* para α ou δ, o RMSE é sempre menor para δ. Para o caso desbalanceado, a distribuição *a priori* $\alpha \sim N(0,1)$ apresenta um RMSE menor para os parâmetros α, β_1 e β_2 quando o tamanho da amostra é 50. No entanto, essa distribuição *a priori* apresenta um RMSE mais elevado para o parâmetro de assimetria α, sugerindo que o método de estimação proposto não é adequado para tamanhos amostrais pequenos. Para outros tamanhos amostrais, a distribuição *a priori* $\alpha \sim N(0,4)$ é recomendada. No entanto, para tamanhos amostrais maiores, a distribuição *a priori* $\alpha \sim N(0,4)$ ainda possui um RMSE menor, mas sem ganhos significativos. O viés das estimativas de α e δ é maior do que o viés dos parâmetros $\boldsymbol{\beta}$, exceto quando distribuições *a priori* informativas são consideradas. Para tamanhos de amostra acima de 500, o viés é menor na maioria dos parâmetros quando se utiliza a distribuição *a priori* $N(0,4)$.

Para $\alpha = 4,5$, a probabilidade de cobertura é inferior a 50% em muitos dos cenários apresentados, especialmente quando se utiliza a distribuição *a priori* $\alpha \sim N(0,1)$, que apresenta uma probabilidade de cobertura de 0% para os parâmetros de assimetria em quase todos os tamanhos de amostra. A probabilidade de cobertura dos parâmetros de assimetria é maior quando se assume a distribuição *a priori* $\alpha \sim N(0,100)$. O intercepto do modelo apresenta uma maior probabilidade de cobertura quando a distribuição *a priori* $\alpha \sim N(0,4)$ é considerada. Para os outros coeficientes, a distribuição *a priori* $\alpha \sim N(0,1)$ resulta em uma melhor probabilidade de cobertura em comparação com outras opções. Além disso, se houver conhecimento empírico sobre o parâmetro δ e ele for positivo e próximo de 1, a distribuição *a priori*, $\delta \sim Beta(7,1)$ pode ser considerada.

Adicionalmente, diferentes critérios de comparação de modelos são calculados para diferentes cenários. Os critérios WAIC e DIC são calculados em cada réplica, conforme demonstrado na Tabela 5, do seguinte modo:

$$DIC = \frac{1}{1000} \sum_{i=1}^{1000} DIC_i, \tag{3.1}$$

$$WAIC = \frac{1}{1000} \sum_{i=1}^{1000} WAIC_i \tag{3.2}$$

onde DIC_i e $WAIC_i$ são os valores dos respectivos critérios de seleção de modelos calculada na i - ésima réplica.

Tabela 5 – *DIC* e *WAIC* para diferentes tamanhos amostrais e configurações de distribuições *a priori*.

$\alpha = 4,5$								
Prioris	$n = 50$		$n = 500$		$n = 1000$		$n = 5000$	
	DIC	*WAIC*	*DIC*	*WAIC*	*DIC*	*WAIC*	*DIC*	*WAIC*
$\delta \sim U(-1,1)$	50,794	51,456	485,198	485,343	967,739	967,887	4870,194	4870,254
$\alpha \sim N(0,100)$	49,664	50,081	483,194	483,479	966,547	966,786	4869,657	4869,729
$\alpha \sim N(0,4)$	51,144	51,841	485,111	485,251	967,289	967,385	4869,909	4869,944
$\alpha \sim N(0,1)$	50,786	51,527	485,820	485,910	969,437	969,492	4874,076	4874,098
$\delta \sim U(0,1)$	50,305	50,924	485,256	485,405	970,012	970,154	4871,751	4871,813
$\delta \sim Beta(7,1)$	49,806	50,350	483,943	484,117	966,522	966,675	4871,427	4871,489
$\alpha = 1$								
Prioris	$n = 50$		$n = 500$		$n = 1000$		$n = 5000$	
	DIC	*WAIC*	*DIC*	*WAIC*	*DIC*	*WAIC*	*DIC*	*WAIC*
$\delta \sim U(-1,1)$	54,654	55,328	520,799	520,926	1040,952	1041,034	5235,177	5235,196
$\alpha \sim N(0,100)$	53,659	54,140	521,471	521,796	1042,092	1042,319	5235,855	5235,886
$\alpha \sim N(0,4)$	54,875	55,594	522,063	522,212	1041,134	1041,232	5235,531	5235,555
$\alpha \sim N(0,1)$	54,943	55,655	520,800	520,894	1041,712	1041,766	5235,142	5235,159

Fonte: Elaborada pelo autor.

Se for considerada uma distribuição *a priori* normal com variância grande para α, obtém-se um valor menor de DIC e WAIC, embora a diferença em relação a uma distribuição *a priori* $\alpha \sim N(0,4)$ não seja significativa, o que não ocorre no caso balanceado ($\alpha = 1$), A seguir, os resultados apresentados nesta seção são comparados com o método INLA, o qual pode ser utilizado para estimar os parâmetros do modelo skew-probito.

3.2 Um estudo comparativo entre dois métodos de estimação

O estudo de simulação conduzido na seção 3.1 revela que a melhor configuração de distribuição *a priori* é considerar $\alpha \sim N(0,4)$. Também podemos obter as estimativas do modelo skew-probito padronizado usando o método INLA (Integrated Nested Laplace Approximation) (ver Rue, Martino e Chopin (2009)).

Nesta seção, estamos interessados em comparar a performance do pacote *rstan* (No-U-Turn Sampler algorithm) com o pacote R - INLA (Laplace integration). Quatro tamanhos de amostra foram considerados $n = (100, 500, 1000, 5000)$. O preditor linear é dado por

$$\eta_i = \beta_0 + \beta_1 x_{1i},$$

onde $x_{1i} \sim U(-3,3), i = 1,2,...,n$ e fixe $\beta_0 = -0,25$, $\beta_1 = 0,5$ e $\alpha = 4,5$. Com essas especificações, 1000 conjuntos de dados não balanceados são gerados usando o modelo skew-probito padronizado.

Depois que os conjuntos de dados são simulados usando o modelo skew-probito, o processo de estimação é realizado usando os pacotes *rstan* e R-INLA. Para comparar a eficiência dos dois métodos, são calculados o RMSE (raiz do erro quadrático médio) e o viés. As distribuições *a priori* consideradas são

- $\beta_j \sim N(0, 10000), i = 0,1$ e
- $\alpha \sim N(0,4)$.

As cadeias de Markov são amostradas com 20.000 iterações, com 10.000 iterações de burn-in e espaçamento 1, e a convergência é avaliada através da estatística $\hat{R}$. O tempo gasto no processo de estimativa dos parâmetros com cada um dos dois pacotes usados neste estudo de simulação é registrado.

A Tabela 6 mostra os resultados de RMSE e viés para diferentes tamanhos de amostra, dois pacotes de linguagem R e duas funções de perda específicas para os parâmetros e o tempo médio gasto pelo modelo padrão skew - probito em diferentes tamanhos de amostra e para dois pacotes da linguagem de programação R, que utilizam diferentes estratégias de estimação de parâmetros.

Tabela 6 – RMSE e o viés dos parâmetros α, β_0, β_1 e δ para diferentes formas de estimar o modelo skew - probito padrão e tempo médio (segundos) gasto pelo modelo skew - probito padronizado nas estimativas em diferentes abordagens e tamanhos de amostra.

$n = 100$								
Função de	*rstan*				*R-INLA*			
Perda	*Absoluta*		*Quadrática*		*Absoluta*		*Quadrática*	
	RMSE	Viés	RMSE	Viés	RMSE	Viés	RMSE	Viés
β_0	0,544	0,489	0,640	0,587	0,524	0,522	0,527	0,524
β_1	0,148	0,057	0,155	0,071	0,134	0,100	0,130	0,094
α	4,503	4,425	4,489	4,437	4,339	4,328	4,389	4,383
δ	1,089	0,953	0,980	0,959	0,864	0,734	0,880	0,840
Tempo	41,260				4,117			
$n = 500$								
Função de	*rstan*				*R-INLA*			
Perda	*Absoluta*		*Quadrática*		*Absoluta*		*Quadrática*	
	RMSE	Viés	RMSE	Viés	RMSE	Viés	RMSE	Viés
β_0	0,334	0,297	0,383	0,382	0,540	0,540	0,541	0,540
β_1	0,053	0,009	0,043	0,015	0,147	0,140	0,144	0,137
α	3,493	3,330	3,496	3,496	4,043	4,043	4,124	4,115
δ	0,628	0,413	0,660	0,660	0,533	0,366	0,610	0,519
Tempo	184,282				4,049			
$n = 1000$								
Função de	*rstan*				*R-INLA*			
Perda	*Absoluta*		*Quadrática*		*Absoluta*		*Quadrática*	
	RMSE	Viés	RMSE	Viés	RMSE	Viés	RMSE	Viés
β_0	0,243	0,201	0,300	0,255	0,546	0,545	0,546	0,546
β_1	0,037	0,004	0,036	0,005	0,160	0,156	0,157	0,153
α	2,727	2,513	2,874	2,662	3,901	3,893	3,969	3,961
δ	0,363	0,188	0,502	0,412	0,340	0,215	0,424	0,331
Tempo	355,108				7,143			
$n = 5000$								
Função de	*rstan*				*R-INLA*			
Perda	*Absoluta*		*Quadrática*		*Absoluta*		*Quadrática*	
	RMSE	Viés	RMSE	Viés	RMSE	Viés	RMSE	Viés
β_0	0,069	0,048	0,078	0,056	0,553	0,553	0,553	0,553
β_1	0,018	0,006	0,018	0,006	0,181	0,180	0,180	0,179
α	1,115	0,870	1,113	0,843	3,724	3,722	3,741	3,740
δ	0,028	0,017	0,058	0,031	0,089	0,075	0,109	0,091
Tempo	1616,309				18,799			

Fonte: Elaborada pelo autor.

A Tabela 6 demonstra que, no contexto da estimativa de parâmetros para o modelo de regressão skew-probito padrão, o uso do pacote R-INLA requer que a mediana *a posteriori* seja considerada, pois a recuperação dos parâmetros é mais eficaz quando comparada à utilização da média *a posteriori*. Para o pacote *rstan*, também é aconselhável considerar a mediana *a posteriori* durante o processo de estimação do modelo de regressão skew-probito padrão, visto que isso resulta em uma recuperação mais precisa dos verdadeiros valores dos parâmetros, em contraste com a utilização da média *a posteriori*. Além disso, os valores de RMSE e viés são menores nesse cenário.

De forma geral, observa-se que o pacote R-INLA exibe menor eficiência em relação à estimativa de parâmetros para o modelo de regressão skew-probito padrão quando o tamanho da amostra ultrapassa 500. Por outro lado, o tempo de processamento é maior ao utilizar o pacote *rstan*. O parâmetro de assimetria δ deve ser destacado, pois é melhor recuperado no processo de estimação pelo pacote R-INLA, exceto quando $n = 5000$.

Ambos os pacotes da linguagem de programação R demonstram ser eficazes na recuperação de parâmetros de assimetria, pois as estimativas de RMSE e viés diminuem à medida que o tamanho da amostra aumenta. Por outro lado, quando consideramos o método INLA, o viés e o RMSE associados aos coeficientes β_0 e β_1 aumentam à medida que o tamanho da amostra aumenta.

Na próxima seção, exploramos técnicas de detecção de outliers no contexto do modelo skew-probito padrão.

3.3 Um estudo comparativo sobre a detecção de outliers em dois resíduos.

Uma das técnicas mais úteis para a detecção de outliers é a análise de resíduos. Esta seção se dedica à comparação de dois resíduos amplamente conhecidos, anteriormente definidos no Capítulo 2: o resíduo studentizado e o resíduo quantílico.

A performance dos resíduos quantílico e studentizado na detecção de observações induzidas como outliers é investigada por meio de um estudo de simulação. O objetivo é comparar o comportamento dos resíduos mencionados acima no modelo de regressão skew-probito padrão. Simulamos 1000 conjuntos de dados de tamanho 1000 com base no modelo skew-probito padrão, com vetor de coeficientes $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$, covariável x_{1i} e parâmetro de assimetria α. O modelo simulado pode ser descrito como

$$\begin{aligned} y_i|\boldsymbol{\beta}, \alpha &\overset{\text{ind.}}{\sim} Bernoulli\,(p_i)\,, \\ p_i &= F_\alpha\,(\eta_i)\,, \\ \eta_i &= \beta_0 + \beta_1 x_{1i}, \end{aligned} \tag{3.3}$$

onde $F_{\alpha}(.)$, indica a função de distribuição acumulada de uma variável aleatória skew - normal padrão, e a covariável é gerada como $x_{1i} \sim U(-3,3)$.

Três valores diferentes para os parâmetros são assumidos para a obtenção de três situações específicas.

(a) Quando $\beta_0 = -1,5$, $\beta_1 = 0,5$ e $\alpha = -4,5$, a variável resposta y com uma proporção média de uns sendo $0,316$.

(b) Para $\beta_0 = 0,3$, $\beta_1 = 0,5$ e $\alpha = 1$, a variável resposta y é gerada com uma proporção média de uns sendo $0,553$.

(c) Considerando $\beta_0 = 1,25$, $\beta_0 = 0,5$ e $\alpha = 4,5$, a variável resposta y é gerada com uma proporção média de uns sendo $0,649$.

Depois de gerados os conjuntos de dados, são induzidas perturbações em determinadas percentagens das observações, são elas 1%, 2%, 3%, 4% e 5%. Existem três possibilidades para definir o parâmetro e 5 valores para porcentagens de pertubação de uma amostra. Isso produz um total de 15 cenários neste estudo de simulação. Três tipos de perturbações são usados no conjunto de dados.

Pertubação tipo A aplica - se ao caso em que $\beta_0 = -1,5$, $\beta_0 = 0,5$ e $\alpha = -4,5$ (caso desbalanceado, onde há uma proporção menor de uns). O procedimento consiste nos seguintes passos

- organize o vetor X em ordem crescente;
- escolha os $W = \frac{k}{100} * 1000$ menores índices de tal forma que $y = 0$, onde k é a porcentagem de observações perturbadas na amostra e
- para os W valores do segundo item, troque $y = 0$ por $y = 1$.

Pertubação tipo B aplica - se ao caso em que $\beta_0 = 1,25$, $\beta_0 = 0,5$ e $\alpha = 4,5$ (caso desbalanceado, onde há uma maior proporção de uns). O procedimento nos seguintes passos:

- organize o vetor X em ordem decrescente;
- escolha os $W = \frac{k}{100} * 1000$ maiores índices tal que $y = 1$ e
- para os W valores do segundo item, troque $y = 1$ por $y = 0$.

Pertubação tipo C aplica - se ao caso, quando $\beta_0 = 0,3$, $\beta_0 = 0,5$ e $\alpha = 1$ (caso balanceado). O procedimento consiste nos seguintes passos:

- Calcule o valor $W = \frac{k}{100} * 1000$, onde k é a porcentagem da amostra que se tem interesse em perturbar;
- Ordene o vetor X em ordem decrescente;
- Escolha os $\frac{W}{2}$ índices mais altos, tal que $y = 1$;
- Para os $\frac{W}{2}$ indivíduos selecionados no item anterior, troque $y = 1$ por $y = 0$;
- Ordene o vetor X em ordem crescente;
- fixe os $\frac{W}{2}$ menores índices, tal que $y = 0$ e
- para as $\frac{W}{2}$ observações fixadas no item anterior, troque $y = 0$ por $y = 1$.

As estimativas realizadas durante o estudo de simulação são obtidas usando a biblioteca *rstan* da linguagem de programação R (veja Stan Development Team (2019) para mais detalhes). Para cada modelo ajustado, uma cadeia de Markov com 10000 iterações e espaçamento de 1 entre as amostras coletadas foi gerada. A Tabela 7 mostra a porcentagem de observações perturbadas detectadas tanto no resíduo studentizado, como no resíduo quantílico aleatorizado para diferentes valores dos parâmetros e diferentes porcentagens de perturbação da amostra. Também o tempo computacional médio gasto para estimar o modelo. O resíduo com a maior porcentagem de observações perturbadas detectadas é considerado o mais eficiente.

Os resultados do estudo de simulação na Tabela 7 revelam que o resíduo quantílico aleatorizado é mais eficiente do que o resíduo studentizado para o modelo skew-probito, considerando diferentes cenários de perturbação (caso desbalanceado: Perturbação tipo A e B e caso balanceado: Perturbação tipo C) e diferentes porcentagens de outliers induzidos. O resíduo studentizado não detecta outliers do tipo B e possui uma porcentagem de detecção pequena para o tipo A, diminuindo à medida que a porcentagem de outliers induzidos aumenta e sendo ainda menor no caso balanceado, considerando a perturbação do tipo C. Com base nesses resultados, recomendamos fortemente o uso do resíduo quantílico aleatorizado para o modelo skew-probito. Este resíduo apresenta uma porcentagem de detecção próxima a 70% no caso balanceado e pode diminuir sua detecção à medida que a porcentagem de observações perturbadas aumenta no caso desbalanceado. Novos mecanismos de perturbação de observação podem ser considerados em futuros estudos de simulação.

Tabela 7 – Porcentagem de observações perturbadas detectadas pelo resíduo studentizado e resíduo quantílico aleatorizado para diferentes valores dos parâmetros e diferentes porcentagens de perturbação na amostra.

Perturbação tipo A, $\beta_0 = -1,5$, $\beta_1 = 0,5$ e $\alpha = -4,5$				
Porcentagem de outliers induzidos	Número de outliers induzidos	Detecção: resíduo studentizado (%)	Detecção: resíduo quantílico (%)	Tempo (segundos)
1%	10	26,83	73,17	52,57
2%	20	4,59	95,42	53,20
3%	30	7,89	89,51	56,28
4%	40	10,04	67,69	56,28
5%	50	9,85	54,15	48,39
Perturbação tipo B, $\beta_0 = 1,25$, $\beta_1 = 0,5$ e $\alpha = 4,5$				
Porcentagem de outliers induzidos	Número de outliers induzidos	Detecção: resíduo studentizado (%)	Detecção: resíduo quantílico (%)	Tempo (segundos)
1%	10	0,00	68,08	58,20
2%	20	0,00	50,08	58,51
3%	30	0,00	33,48	54,84
4%	40	0,00	25,11	54,26
5%	50	0,00	20,09	54,29
Perturbação tipo C, $\beta_0 = 0,3$, $\beta_1 = 0,5$ e $\alpha = 1$				
Porcentagem de outliers induzidos	Número de outliers induzidos	Detecção: resíduo studentizado (%)	Detecção: resíduo quantílico (%)	Tempo (segundos)
1%	10	35,09	64,91	49,06
2%	20	26,80	73,21	49,71
3%	30	27,59	72,41	53,19
4%	40	26,23	73,77	52,38
5%	50	21,86	78,13	52,58

Fonte: Elaborada pelo autor.

CAPÍTULO

4

APLICAÇÃO

Vamos ilustrar os resultados obtidos no capítulo anterior, através de uma aplicação.

4.1 Conjunto de dados e Seleção de Modelos

O conjunto de dados refere-se a membros do sexo feminino de uma população indígena que reside no estado do Arizona, nos Estados Unidos. Estudos indicam que essa população apresenta uma alta incidência de diabetes durante a gravidez. O banco de dados original pode ser baixado do pacote **mlbench** com o nome *PimaIndiansDiabetes2*. Para mais detalhes, consulte Leisch e Dimitriadou (2010) e Newman *et al.* (1998).

Neste estudo, consideramos seis variáveis explicativas e uma variável dependente. O conjunto de dados possui 768 observações. Observações com dados faltantes são excluídas, resultando em 724 indivíduos remanescentes, uma redução de apenas $5,73\%$ em relação aos dados originais. A variável dependente é o diabetes, classificada em dois níveis, onde 1 corresponde à presença de diabetes ($34,4\%$ dos casos) conforme o exame, e 0 indica ausência ($65,6\%$ dos casos).

Para fins de comparação com o modelo proposto, também ajustamos o modelo de regressão com resposta binária e função de ligação probito e logito. Além disso, exploramos o modelo skew-probito sem intercepto, uma vez que o intercepto do modelo skew-probito tem sido objeto de discussão em artigos sobre o tema, devido à sua identificabilidade em alguns tipos específicos de função de ligação skew-probito, conforme discutido, por exemplo, por Naranjo, Pérez e Martín (2019). O modelo skew-probito para esses dados é:

$$\begin{aligned} y_i &\overset{\text{ind.}}{\sim} Bernoulli(p_i), \\ p_i &= F_\alpha(\eta_i). \end{aligned} \tag{4.1}$$

onde $i = 1, 2, ..., 724$ e F_α denota a função de distribuição acumulada de uma variável aleatória

que segue uma distribuição skew-normal padrão com vetor de parâmetros

$$\theta = \left(-\frac{\sqrt{2}\alpha}{\sqrt{\pi + (\pi - 2)\,\alpha^2}}, \frac{\pi + \pi\alpha^2}{\pi + (\pi - 2)\,\alpha^2}, \alpha \right)$$

.

O preditor linear é definido pela expressão

$$\eta_i = \beta_0 + \sum_{j=1}^{6} \beta_j x_{ji}, \tag{4.2}$$

onde $i = 1, 2, ..., 724$, x_{1i} é o número de gestações ("pregnant"na base de dados), x_{2i} é a concentração de glicose no sangue ("glucose"), x_{3i} é o índice de massa corpórea do indivíduo ("mass"), x_{4i} é a pressão diastólica ("pressure"), x_{5i} é um escore que classifica a probabilidade de diabetes com base no histórico familiar ("pedigree") e x_{6i} é a idade ("age"). Neste caso, as distribuições *a priori* $\beta_j \sim N(0, 10000)$, $j = 0, 1, ..., 6$ e $\alpha \sim N(0, 4)$ são consideradas. Os modelos são implementados usando a linguagem R, especificamente o pacote *rstan* (Gelman, Lee e Guo (2015)). Para obter convergência via MCMC, usamos uma cadeia, 5.000 iterações são realizadas com 2.500 de aquecimento, sem espaçamento. Para verificar a convergência da cadeia de Markov, a estatística de Gelman-Rubin (Gelman, Rubin *et al.* (1992)) é utilizada. O código utilizado está disponível no Apêndice A.

A Tabela 8 apresenta as estimativas dos modelos propostos. A mediana *a posteriori* é considerada, e cada estimativa é acompanhada de seu respectivo intervalo com 95% de credibilidade. A convergência da cadeia de Markov pode ser assegurada pela estatística de Gelman-Rubin, pois é igual a 1 para todos os parâmetros. O DIC e WAIC também são apresentados na Tabela 8 para cada modelo. Portanto, o modelo mais adequado é aquele cuja função de ligação é skew-probito padrão com intercepto.

Conforme mencionado por Henderson *et al.* (2016), valores mais baixos de DIC implicam em melhores ajustes de modelo. Seguindo uma regra geral sugerida por Carlin e Louis (2008), a qual indica que diferenças significativas entre os valores de DIC começam em diferenças superiores a três a cinco. Portanto, há evidências de que o modelo de regressão com resposta binária e função de ligação skew-probito é o mais justificável para esses dados, dentre os modelos propostos.

A Figura 2 exibe a curva de probabilidade para os diferentes modelos apresentados na Tabela 8. A curva de probabilidade para os modelos probito e logito é simétrica em torno do ponto $\eta = 0$. A curva de probabilidade para o modelo skew-probito situa-se abaixo das curvas do modelo probito e logito e cruza a linha vertical $\eta = 0$ quando a probabilidade de sucesso é inferior a $0,5$. A curva de probabilidade do modelo skew-probito sem o intercepto quase coincide com a do modelo skew-probito com intercepto.

Tabela 8 – Estimativa dos parâmetros do modelo para dados de diabetes.

		Modelo probito	**Modelo Logito**	**Modelo skew - probito**	**Modelo skew - probito sem intercepto**
Variável	**Parâmetro**	**Estimativas**	**Estimativas**	**Estimativas**	**Estimativas**
Intercepto	β_0	-5,29 [-6,20; -4,44]	-9,08 [-10,73; -7,46]	-6,97 [-8,26;-5,76]	
Pregnant	β_1	0,07 [0,03; 0,11]	0,12 [0,05; 0,19]	0,09 [0,04; 0,15]	0,09 [0,05; 0,14]
Glucose	β_2	0,02 [0,02; 0,02]	0,04 [0,03; 0,04]	0,03 [0,02; 0,03]	0,02 [0,01; 0,02]
Mass	β_3	0,05 [0,04; 0,07]	0,09 [0,06; 0,12]	0,08 [0,05; 0,10]	0,02 [0,00; 0,04]
Pressure	β_4	-0,01 [-0,02; 0,00]	-0,01 [-0,03; 0,01]	-0,01 [-0,02; 0,01]	-0,05 [-0,06; -0,03]
Pedigree	β_5	0,48 [0,14; 0,81]	0,98 [0,37; 1,58]	0,81 [0,34; 1,29]	0,38 [-0,04; 0,81]
Age	β_6	0,01 [0,00; 0,02]	0,02 [0,00; 0,04]	0,02 [0,00; 0,03]	0,00 [-0,01; 0,02]
Parâmetro de assimetria	α			3,23 [1,49; 5,49]	3,05 [1,72; 5,04]
	DIC	686,93	687,60	679,06	834,60
	WAIC	687,38	687,88	679,59	834,95

Fonte: Elaborada pelo autor.

A Tabela 9 apresenta as estimativas dos parâmetros para os modelos completo e reduzido. Comparando estes modelos usando os critérios de seleção de modelos DIC e WAIC, a covariável pressure (pressão diastólica) não é significativo para prever se cada indivíduo tem diabetes ou não. No entanto, não há ganho significativo ao excluir essa variável. O modelo completo é escolhido como o modelo adequado, dentre os modelos ajustados.

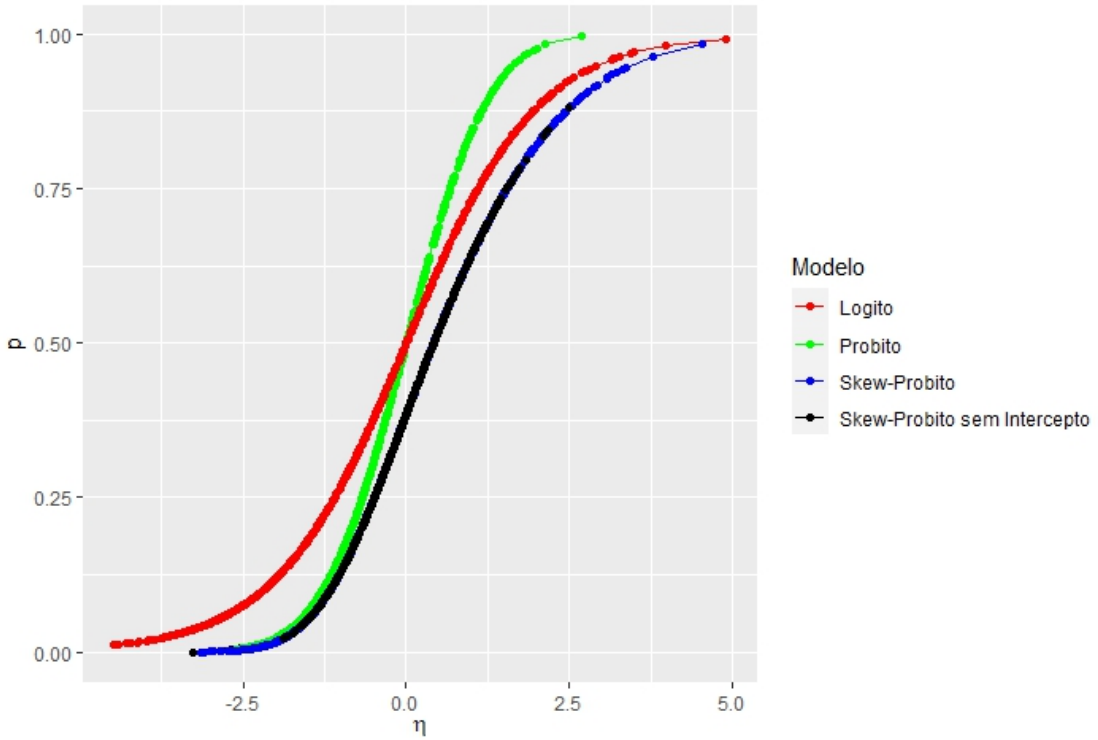


Figura 2 – Curva de probabilidade para os diferentes modelos ajustados na Tabela 8. Fonte: Elaborada pelo autor.

Tabela 9 – Comparação das estimativas do modelo completo com os modelos reduzidos.

Variável	**Parâmetro**	**Modelo completo**	**Modelo sem a variável pressure**	**Modelo sem a variável age**	**Modelo sem a variável pressure e age**
Intercepto	β_0	-6,97 [-8,26; -5,76]	-7,23 [-8,35; -6,07]	-6,78 [-8,10; -5,47]	-6,96 [-8,09; -5,90]
Pregnant	β_1	0,09 [0,04; 0,15]	0,09 [0,04; 0,14]	0,11 [0,07; 0,16]	0,11 [0,07; 0,16]
Glucose	β_2	0,03 [0,02; 0,03]	0,03 [0,02; 0,03]	0,03 [0,02; 0,04]	0,03 [0,02; 0,04]
Mass	β_3	0,08 [0,05; 0,10]	0,05 [0,07; 0,10]	0,07 [0,05; 0,10]	0,07 [0,05; 0,10]
Pressure	β_4	-0,01 [-0,02; 0,01]	-	0,00 [-0,02; 0,01]	-
Pedigree	β_5	0,81 [0,34; 1,29]	0,82 [0,36; 1,31]	0,80 [0,33; 1,29]	0,81 [0,36; 1,29]
Age	β_6	0,02 [0,00; 0,03]	0,01 [0,00; 0,03]	-	-
Parâmetro de assimetria	α	3,23 [1,49; 5,49]	3,25 [1,41; 5,79]	2,95 [1,01; 5,28]	3,02 [1,27; 5,50]
	DIC	679,06	677,74	681,30	679,61
	WAIC	679,59	678,15	681,73	679,93

Fonte: Elaborada pelo autor.

4.2 Detecção de Outliers

Com o objetivo de verificar a adequação do modelo, é realizada a análise de resíduos. O resíduo do quantílico aleatorizado é calculado com base na expressão apresentada na seção 2.4.3.2. A condição $|r| > 3$ é estabelecido como critério para detecção de outliers. Sob este critério, dois possíveis outliers são detectados, observações 79 e 478. A Figura 3 mostra o gráfico, histograma e envelope simulado do resíduo quantílico aleatorizado para dados de diabetes. Não há evidência de assimetria nos resíduos.

A Tabela 10 mostra as observações detectadas como possíveis outliers. As principais características dessas observações são o alto número de gestações (pregnancies), glicose (glucose) alta (> 100), pedigree baixo e pressão (pressure) baixa. Além disso, um dos pacientes pode ser classificado como pré-obeso e o outro como obeso grau II, mas sem diabetes.

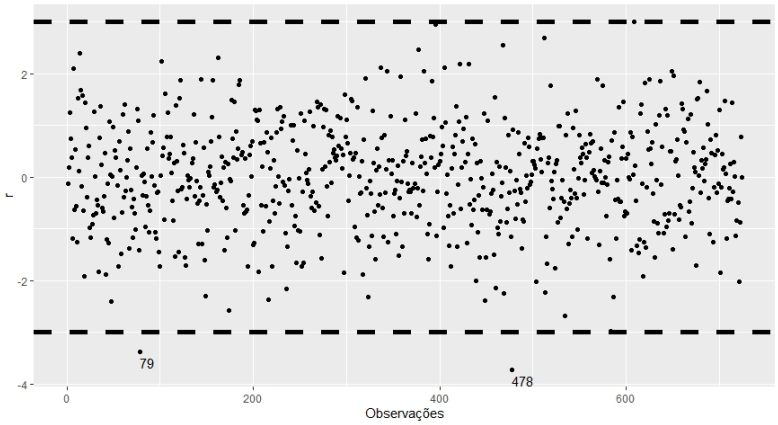
Tabela 10 – Descrição das observações detectadas como outliers para dados de diabetes.

Ponto	**Diabetes**	**Pregnant**	**Glucose**	**Mass**	**Pressure**	**Pedigree**	**Age**
79	0	13	106	36,6	72	0,178	45
478	0	8	120	25,0	78	0,409	64

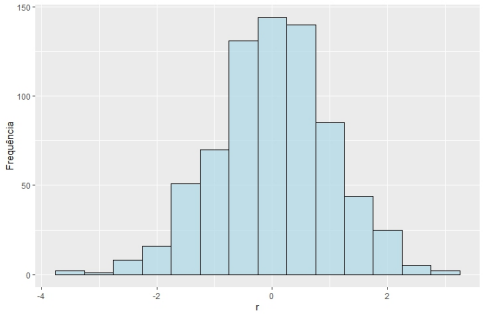
Fonte: Elaborada pelo autor.

4.3 Análise de Influência Global

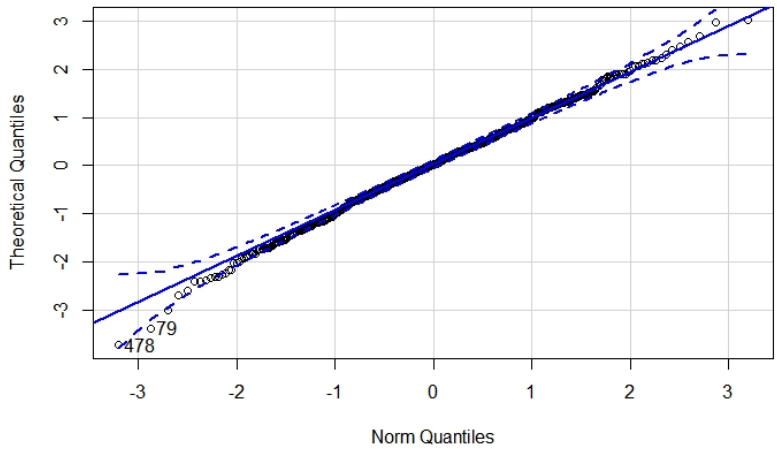
A Tabela 11 compara as estimativas do modelo skew - probito ajustado com todas as observações e e modelos sem as observações 79 e 478. Note que não há mudança inferencial quando removemos a observação detectada com sendo outlier. Os critérios de seleção de modelo DIC e WAIC também são apresentados. Os modelos estimados sem as observações 79 e 478 tem menor DIC e WAIC do que o modelo completo.



(a) Gráfico do resíduo quantílico aleatorizado.



(b) Histograma do resíduo quantílico aleatorizado.



(c) Envelope simulado do resíduo.

Figura 3 – Gráfico, histograma e envelope do resíduo quantílico aleatorizado para dados de diabetes, respectivamente. Fonte: Elaborada pelo autor.

Tabela 11 – Comparação de estimativas sob o modelo completo e modelos sem observações 79 e 478.

Variável	Parâmetros	**Modelo completo**	**Modelo sem a observação 79**	**Modelo sem a observação 478**	**Modelo sem a observação 79 e 478**
Intercepto	β_0	-6,97 [-8,26; -5,76]	-6,99 [-8,26; -5,70]	-6,96 [-8,30; -5,69]	-6,98 [-8,29; -5,78]
Pregnant	β_1	0,09 [0,04; 0,15]	0,09 [0,05; 0,14]	0,09 [0,04; 0,14]	0,09 [0,04; 0,15]
Glucose	β_2	0,03 [0,02; 0,03]	0,03 [0,02; 0,04]	0,03 [0,02; 0,04]	0,03 [0,02; 0,03]
Mass	β_3	0,08 [0,05; 0,10]	0,08 [0,05; 0,10]	0,07 [0,05; 0,10]	0,08 [0,05; 0,10]
Pressure	β_4	-0,01 [-0,02; 0,01]	-0,01 [-0,02; 0,01]	-0,01 [-0,02; 0,01]	-0,01 [-0,02; 0,01]
Pedigree	β_5	0,81 [0,34; 1,29]	0,80 [0,34; 1,29]	0,81 [0,34; 1,28]	0,82 [0,33; 1,27]
Age	β_6	0,02 [0,00; 0,03]	0,02 [0,00; 0,03]	0,02 [0,00; 0,03]	0,02 [0,00; 0,03]
Parâmetro de assimetria	α	3,23 [1,49; 5,49]	3,27 [1,41; 5,80]	3,26 [1,41; 5,60]	3,30 [1,51; 5,65]
	DIC	679,06	677,31	677,82	676,37
	WAIC	679,59	677,87	678,34	676,89

Fonte: Elaborada pelo autor.

A Tabela 12 também mostra a variação relativa absoluta em cada parâmetro estimado, que pode ser obtida através da expressão

$$ARC_{\hat{\theta}} = 100 \left| \frac{\hat{\theta} - \hat{\theta}_{(i)}}{\hat{\theta}} \right| \tag{4.3}$$

onde $\hat{\theta}$ é a estimativa de parâmetro relacionada ao modelo completo e $\hat{\theta}_{(i)}$ é uma estimativa de parâmetro em relação ao modelo sem a i-ésima observação.

Tabela 12 – Influência global para o modelo skew - probito padrão.

	Observação outlier	β_0	β_1	β_2	β_3	β_4	β_5	β_6	α
estimativas	Modelo original	-6,98	0,09	0,03	0,07	-0,01	0,82	0,02	3,21
estimativas	79	-6,98	0,09	0,03	0,08	-0,01	0,81	0,02	3,22
ARC(%)		0,00	0,00	0,00	14,28	0,00	1,22	0,00	0,31
estimativas	478	-7,00	0,09	0,03	0,07	-0,01	0,81	0,02	3,24
ARC(%)		0,28	0,00	0,00	0,00	0,00	1,22	0,00	0,93

Fonte: Elaborada pelo autor.

Observando a Tabela 12, com exceção do parâmetro β_3, que apresenta variação relativa de $14,28\%$ no caso do modelo estimado sem observação 79, todos os demais parâmetros não apresentam variação relativa significativa. Portanto, com base na Tabela 11 e na Tabela 12, podemos dizer que não há evidências de que as observações 79 e 478 sejam globalmente

influentes. O modelo final pode ser escrito como

$$\begin{aligned} y_i &\overset{\text{ind.}}{\sim} Bernoulli\,(p_i)\,, \\ p_i &= F_{(\alpha=3,21)}\,(\eta_i)\,, \\ \eta_i &= -6,98 + 0,09x_{1i} + 0,03x_{2i} + 0,07x_{3i} - 0,01x_{4i} + 0,81x_{5i} + 0,02x_{6i}. \end{aligned} \tag{4.4}$$

Ressalta-se que as variáveis explicativas pregnancy, glucose, mass, pedigree e age contribuem para o aumento da probabilidade do paciente testar positivo para diabetes. Os histogramas das saídas do MCMC e trace plot de cada um dos parâmetros no modelo final podem ser encontrados no Apêndice B. Há outras formas de identificar se uma determinada observação é influente, Prates *et al.* (2011) considera o uso de uma métrica para detectar observações influentes.

4.4 Sensibilidade da distribuição *a priori* associada ao parâmetro de assimetria.

Adicionalmente, analisamos a sensibilidade da distribuição *a priori* considerada para o parâmetro de assimetria no processo de estimação do modelo de regressão com resposta binária e função de ligação skew - probito para dados de Diabetes. O processo de estimação do modelo de regressão skew - probito para dados de diabetes foi realizado novamente para diferentes distribuições *a priori* para α ou δ, conforme descrito na seção 3.1. Estimativas e intervalo com 95% de credibilidade são dados na Tabela 13.

Tabela 13 – Modelo final com diferentes configurações de distribuições *a priori*.

Variável	Parâmetros	**Modelo com** $\delta \sim U\,(-1,1)$	**Modelo com** $\alpha \sim N\,(0,100)$	**Modelo com** $\alpha \sim N\,(0,4)$	**Modelo com** $\alpha \sim N\,(0,1)$
Intercepto	β_0	-6,91 [-8,23;-5,75]	-6,77 [-8,00;-5,68]	-6,99[-8,22; -5,73]	-6,83 [-7,96; -5,32]
Pregnant	β_1	0,09 [0,04; 0,14]	0,09 [0,04; 0,13]	0,09 [0,04; 0,15]	0,09 [0,04; 0,13]
Glucose	β_2	0,03 [0,02; 0,03]	0,03 [0,02; 0,03]	0,03 [0,02; 0,03]	0,03 [0,02; 0,03]
Mass	β_3	0,08 [0,05; 0,10]	0,07 [0,05; 0,10]	0,08 [0,05; 0,10]	0,07 [0,05; 0,10]
Pressure	β_4	-0,01 [-0,02; 0,01]	-0,01 [-0,02; 0,01]	-0,01 [-0,02; 0,01]	-0,01 [-0,02; 0,01]
Pedigree	β_5	0,78 [0,29; 1,32]	0,81 [0,34; 1,29]	0,82 [0,40; 1,29]	0,74 [0,31; 1,18]
Age	β_6	0,02 [0,00; 0,03]	0,02 [0,00; 0,03]	0,02 [0,00; 0,03]	0,01 [0,00; 0,03]
Parâmetro de assimetria	α	3,88 [1,16; 8,67]	6,50 [2,58; 12,88]	3,38 [1,48; 5,79]	1,95 [-0,25; 3,25]
	δ	0,94 [0,76; 0,99]	0,98 [0,93; 1,00]	0,95 [0,83; 0,99]	0,83 [-0,24; 0,96]
	DIC	679,52	677,34	678,20	680,80
	WAIC	680,38	678,26	678,73	681,26

Fonte: Elaborada pelo autor.

A Tabela 13 não mostra evidências de diferença significativa entre as estimativas, quando é considerado os intervalos com 95% de credibilidade para α quando o modelo skew - probito é ajustado com diferentes distribuições *a priori*. A interseção desses intervalos de credibilidade não é vazia. O mesmo pode ser dito para o parâmetro δ.

Os códigos utilizados neste capítulo para estimar o modelo skew-probito encontram-se no Apêndice A. No Apêndice B, apresentamos os gráficos das saídas MCMC. Na Figura 6, exibimos os histogramas das distribuições marginais *a posteriori* do parâmetro α e do vetor de parâmetros $\boldsymbol{\beta}$. Na Figura 7, apresentamos o trace plot para cada um dos parâmetros do modelo final. Os gráficos no Apêndice B indicam que há evidências de convergência da cadeia de Markov. No próximo capítulo, consideraremos o uso de efeitos aleatórios no modelo skew-probito.

CAPÍTULO

5

MODELO DE REGRESSÃO SKEW - PROBITO MISTO

O objetivo deste capítulo é apresentar uma extensão do modelo skew-probito previamente proposto por Bazán, Bolfarine e Branco (2010). A estimação hierárquica bayesiana é realizada, e uma aplicação com o bem conhecido conjunto de dados de Madras é apresentada.

5.1 Introdução

Em determinadas circunstâncias, a suposição de independência entre as observações é violada. Tais observações podem ser classificadas com base em critérios, que podem ser espaciais ou alguma característica intrínseca comum a um grupo de observações. Por exemplo, no contexto escolar, os alunos são agrupados em classes, que, por sua vez, são agrupadas em escolas e assim por diante. Outra situação relevante ocorre quando medidas repetidas estão disponíveis para o mesmo indivíduo em intervalos de tempo igualmente espaçados entre elas. Os dados coletados nessas situações são denominados dados longitudinais.

Para a situação descrita acima, é necessário incorporar a correlação entre indivíduos do mesmo grupo ou medidas repetidas do mesmo indivíduo. Uma possível solução é incluir efeitos aleatórios. Neste capítulo, é apresentada a extensão do modelo skew-probito, proposto por Bazán, Bolfarine e Branco (2010), que incorpora esses efeitos aleatórios.

Dados de resposta binária longitudinal são comumente analisados usando modelos de regressão logística mista. No entanto, quando os dados são desbalanceados, esse modelo pode ser inadequado. Motivados por aplicações com variáveis resposta desbalanceadas, propomos o uso da função de ligação skew-probito sob abordagem bayesiana e realizamos o processo de estimação bayesiana por meio do método INLA. Para a aplicação, utilizamos os dados conhecidos de esquizofrenia de Madras, nos quais a variável binária y indica a presença ou ausência de sintomas psiquiátricos nos meses $t = 0, \ldots, 11$ durante o primeiro ano após a hospitalização por

esquizofrenia de $i = 1, ..., 86$ pacientes. Outras variáveis, como idade e sexo, são consideradas. Por meio de métodos de comparação de modelos, é possível verificar que o modelo proposto é mais apropriado do que o modelo tradicional. Além disso, são fornecidas interpretações das estimativas baseadas na aplicação.

5.2 Motivação

O estudo longitudinal da esquizofrenia de Madras acompanhou 90 pacientes esquizofrênicos por dez meses, com o objetivo principal de caracterizar a história natural da doença (o histórico natural de uma doença descreve o curso que a doença tende a seguir em termos de seu desenvolvimento, progressão e desfecho, na ausência de intervenções médicas ou outras influências significativas). Após eliminar as observações com dados faltantes, obtivemos uma amostra de tamanho 86.

A esquizofrenia é um transtorno mental caracterizado por comportamento social incomum e incapacidade de distinguir o que é real do que não é. Entre os sintomas mais comuns estão delírios, pensamento confuso ou pouco claro, alucinações auditivas, diminuição da interação social, expressão de emoções reduzida e ausência de motivação. Pessoas com esquizofrenia geralmente apresentam outros problemas de saúde mental, como transtornos de ansiedade, depressão ou transtornos por abuso de substâncias. Os sintomas geralmente se manifestam gradualmente na idade adulta e persistem por um longo período.

A variável binária y (y_{ij}) indica a presença ou ausência de sintomas psiquiátricos durante $j = 0, \ldots, 11$ meses durante o primeiro ano após a hospitalização por esquizofrenia para $i = 1, ..., 86$ pacientes. No conjunto de dados, existem outras variáveis, tais como

- X_{1ij} (mês): $(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$ e $11)$,
- X_{2ij} (idade): $(0 = \text{idade } \geq 20, 1 = \text{idade } < 20)$ e
- X_{3ij} (gênero): $(0 = \text{masculino }, 1 = \text{feminino })$.

A Tabela 14 mostra a proporção de respostas em cada uma das variáveis no conjunto de dados. Nas Figuras 4 e 5, podem ser observados os gráficos do perfil médio para as variáveis sexo e idade, respectivamente. Não há evidências significativas de diferença no comportamento da doença nos subgrupos de idade e sexo. A hospitalização faz com que a proporção de pacientes com sintomas diminua significativamente em ambos os subgrupos de sexo e idade.

O objetivo é verificar se a taxa de declínio dos sintomas difere entre os subgrupos de gênero e idade.

Tabela 14 – Estatística descritiva das covariáveis do conjunto de dados Madras sobre esquizofrenia.

Variável	1	0
y	$30,91\%$ (Sim)	$69,09\%$ (Não)
Gênero	$47,29\%$ (Feminino)	$52,71\%$ (Masculino)
Idade	$36,11\%$ (< 20)	$63,89\%$ (≥ 20)

Fonte: Elaborada pelo autor.

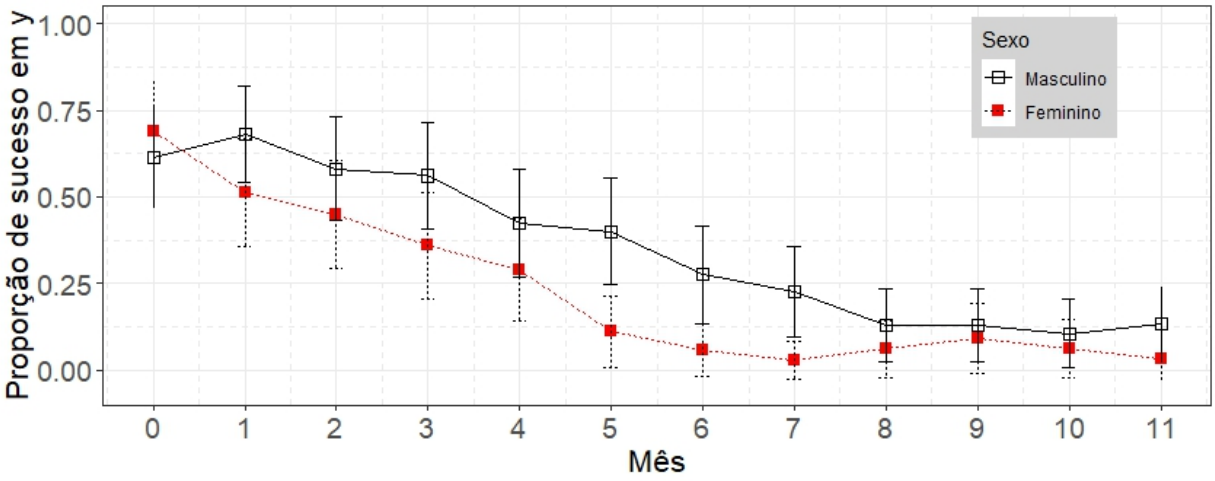


Figura 4 – Gráfico do perfil médio da variável *y* no subgrupo sexo. Fonte: Elaborada pelo autor.

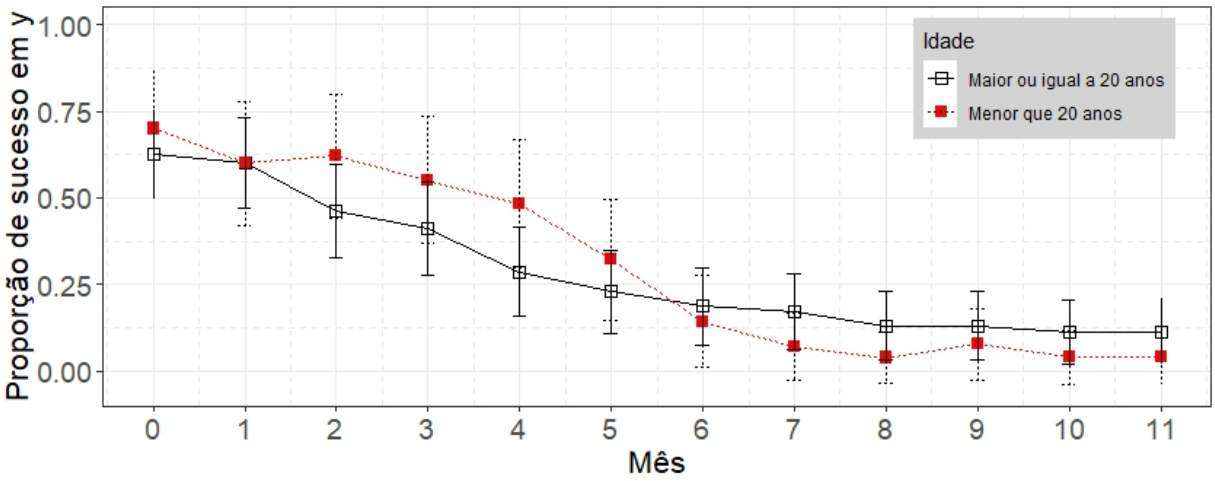


Figura 5 – Gráfico do perfil médio da variável y no subgrupo idade. Fonte: Elaborada pelo autor.

5.3 O Modelo Skew - Probito Misto

Proposto por Bazán, Bolfarine e Branco (2010), o modelo skew-probito surge da necessidade de modelar variáveis binárias desbalanceadas. Essas variáveis não são efetivamente modeladas pelos modelos de regressão usuais, como o probito e o logito.

Em alguns problemas, encontramos medidas repetidas para indivíduos ou grupos. A abordagem comum é introduzir efeitos aleatórios no modelo de regressão, incorporando consequentemente a correlação intra-individual (ou intragrupo) no modelo. Considere $i = 1, 2, \ldots, n$ (unidades de nível 2) e $j = 1, 2, \ldots, n_i$ observações repetidas (unidades de nível 1), agrupadas por sujeito. Neste caso, $\mathbf{x}_{ij}$ é o vetor $k \times 1$ de covariáveis, onde x_{ij1} pode ser igual a 1, correspondendo a um intercepto, e $\boldsymbol{\beta}$ é um vetor $k \times 1$ de coeficientes de regressão. Também denotamos $\mathbf{z}_{ij}$ como o vetor $r \times 1$ de variáveis com efeitos aleatórios (geralmente inclui-se uma coluna de uns para o intercepto aleatório) e $\mathbf{b}_i$ o vetor de efeitos aleatórios. Definimos $\eta_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i$ como o preditor linear. Propomos um modelo misto como extensão do modelo proposto por Bazán, Bolfarine e Branco (2010), que pode ser descrito pelo conjunto de equações a seguir.

$$\begin{aligned} &y_{ij}|\mathbf{b}_i \overset{\text{ind.}}{\sim} Bernoulli\left(p_{ij}\right), \\ &p_{ij} = E(y_{ij}|\mathbf{b}_i) = F_\theta(\eta_{ij}), \\ &\mathbf{b}_i \sim N_r(\mathbf{0}, \boldsymbol{\Sigma_b}). \end{aligned} \tag{5.1}$$

onde $j = 1, \ldots, n_i$, $i = 1, \ldots, n$ e $F_\theta(.)$ denota a função de distribuição acumulada da distribuição skew - normal padrão, com vetor de parâmetros $\theta = \left(\mu, \sigma^2, \alpha\right)$.

A função de verossimilhança para a classe de modelos skew - probito é dada por

$$L(\boldsymbol{\beta}, \mathbf{b}_i, \theta|\mathbf{y}, \mathbf{X}) = \prod_{i=1}^{n}\prod_{j=1}^{n_i}\left[F_\theta\left(\eta_{ij}\right)\right]^{y_{ij}}\left[1 - F_\theta\left(\eta_{ij}\right)\right]^{1-y_{ij}} \tag{5.2}$$

onde η_{ij} é o preditor linear, $\mathbf{X}$ é a matriz de dados previamente definida no capítulo anterior, e $F_\theta\left(\eta_{ij}\right)$ foi definido anteriormente. Vamos assumir independência entre as distribuições *a priori*.

O processo de estimação é realizado utilizando a linguagem de programação R, por meio do pacote R-INLA. No Apêndice C, discute-se brevemente sobre o método INLA e alguns aspectos importantes. Essa metodologia é adotada devido à falta de obtenção de convergência das cadeias de Markov simuladas com a metodologia NUTS.

5.4 Aplicação

São estimados quatro modelos com as variáveis mencionadas anteriormente, utilizando duas funções de ligação diferentes para cada modelo (skew-probito e probito) são assumidas. O

modelo 1 é um modelo probito que incorpora as covariáveis **mês**, **idade** e **sexo** e é dado pelo seguinte conjunto de equações.

$$\begin{aligned} y_i &\overset{\text{ind.}}{\sim} Bernoulli\left(p_i\right), \\ p_i &= \Phi\left(\beta_0 + \sum_{j=1}^{3} \beta_j X_{ji}\right). \end{aligned} \tag{5.3}$$

onde, $i = 1,2,...,922$ e Φ denota a função de distribuição acumulada de uma variável aleatória normal com média 0 e variância 1.

O **modelo 2** é um modelo probito com intercepto aleatório e que incorpora as covariáveis **mês**, **idade** e **gênero**, e é modelado da seguintes equações

$$\begin{aligned} y_{ij}|b_i &\overset{\text{ind.}}{\sim} Bernoulli\left(p_{ij}\right), \\ p_{ij} &= \Phi\left(\beta_0 + b_i + \sum_{k=1}^{3} \beta_k X_{kij}\right), \\ b_i &\overset{\text{i.i.d.}}{\sim} N\left(0, \sigma_b^2\right) \end{aligned} \tag{5.4}$$

onde $i = 1,2,...,86$; $j = 1,2,...,12$.

O **modelo 3** é o modelo skew - probito que incorpora as covariáveis **mês**, **idade** e **gênero**. Este modelo pelas seguintes equações

$$\begin{aligned} y_i &\overset{\text{ind.}}{\sim} Bernoulli\left(p_i\right), \\ p_i &= F_\alpha\left(\beta_0 + \sum_{j=1}^{3} \beta_j X_{ji}\right). \end{aligned} \tag{5.5}$$

onde, $i = 1,2,...,922$ e F_α denota a função de distribuição acumulada de uma variável aleatória skew - normal com média 0 e variância 1.

O **modelo 4** é um modelo skew - probito com intercepto aleatório e que incorpora as covariáveis **mês**, **idade** e **gênero**, representado pelas seguintes equações:

$$\begin{aligned} y_{ij}|b_i &\overset{\text{ind.}}{\sim} Bernoulli\left(p_{ij}\right), \\ p_{ij} &= F_\alpha\left(\beta_0 + b_i + \sum_{k=1}^{3} \beta_k X_{kij}\right), \\ b_i &\overset{\text{i.i.d.}}{\sim} N\left(0, \sigma_b^2\right) \end{aligned} \tag{5.6}$$

onde $i = 1,2,...,86$; $j = 1,2,...,12$. As suposições de distribuição *a priori* são:

- $\alpha \sim Normal(0,4)$;
- $\beta_k \sim Normal(0,1000)$, para $k = 0,1,\ldots,3$;
- $b \sim N(0,\sigma_b^2)$;
- $\frac{1}{\sigma_b^2} \sim Gamma(0{,}01;0{,}01)$.

O método de estimação (método INLA) usado para estimar os quatro modelos apresentados nessa seção, esta descrito no Apêndice C. Os modelos foram comparados utilizando os critérios de seleção de modelos DIC e WAIC. Detalhes sobre DIC e WAIC são fornecidos no Capítulo 2.

Quando observamos a Tabela 15, podemos notar que o modelo 4, utilizando a função de ligação skew - probito com intercepto aleatório, possui o menor valor de DIC e WAIC. Contudo, a diferença entre os modelos 2 e 4 não é significativa em nenhum dos critérios de seleção de modelos utilizados. Vamos trabalhar comm o modelo Skew - probito misto.

Ao analisar as estimativas e os intervalos com 95% de credibilidade do modelo skew - probito com intercepto aleatório, observamos que a variável idade não é significativa. Para verificar se ela deve ser mantida, o modelo é ajustado novamente sem a variável idade e comparado com o modelo completo.

Tabela 15 – Estimativas de parâmetros dos modelos 1, 2, 3 e 4.

	Modelo 1	**Modelo 2**	**Modelo 3**	**Modelo 4**
Parâmetros	Estimativas (I.C.)	Estimativas (I.C.)	Estimativas (I.C.)	Estimativas (I.C.)
β_0	0,57 [0,38;0,76]	0,58 [0,45;0,81]	0,71 [0,66;0,76]	0,71 [0,63;0,78]
β_1	-0,19 [-0,22;-0,16]	-0,20 [-0,23;-0,17]	-0,19 [-0,22;-0,16]	-0,20[-0,23;-0,17]
β_2	0,13 [-0,05;0,33]	0,15 [-0,06;0,38]	0,13 [-0,06;0,33]	0,16 [-0,05;0,37]
β_3	-0,47[-0,66;-0,28]	-0,47 [-0,69;-0,26]	-0,46 [-0,65;-0,27]	-0,47 [-0,67;-0,26]
σ_b^2		0,08 [0,02;0,22]		1,39 [1,27;1,56]
α			-0,01 [-0,14;0,11]	-0,01 [-0,15;0,11]
DIC	937,62	919,33	937,50	917,29
WAIC	937,77	919,20	937,75	917,45

Fonte: Elaborada pelo autor.

A Tabela 16 mostra as estimativas do modelo 4 em comparação com o modelo reduzido. O modelo completo possui os menores valores em todos os critérios de comparação utilizados. Portanto, o modelo escolhido possui as variáveis explicativas mês, ano e idade.

Tabela 16 – Estimativas do modelo completo e dos modelos reduzidos.

	Modelo 4	Modelo 4 sem a variável idade
Parâmetro	Estimativas (I.C.)	Estimativas (I.C.)
β_0	0,71 [0,63;0,78]	0,72 [0,66;0,78]
β_1	-0,20[-0,23;-0,17]	-0,19 [-0,22;-0,16]
β_2	0,16 [-0,05;0,37]	-
β_3	-0,47[-0,67;-0,26]	-0,44 [-0,62-0,25]
α	-0,01 [-0,15;0,11]	0,00 [-0,14;0,12]
σ_b^2	1,39 [1,27;1,56]	1,36 [1,27;1,49]
DIC	917,29	937,29
WAIC	917,45	937,49

Fonte: Elaborada pelo autor.

O modelo final pode ser escrito da forma:

$$\begin{aligned} &y_{ij}|b_i \overset{\text{ind.}}{\sim} Bernoulli\left(p_{ij}\right), \\ &p_{ij} = F_{(\alpha=-0{,}01)}\left(b_i + 0{,}71 - 0{,}20X_{1ij} + 0{,}16X_{2ij} - 0{,}47X_{3ij}\right), \\ &b_i \overset{\text{i.i.d.}}{\sim} N\left(0, 1{,}39\right) \end{aligned} \tag{5.7}$$

onde, $i = 1, 2, ..., 86$; $j = 1, 2, ..., 12$. Observa-se que a variável tempo influencia negativamente na probabilidade do paciente apresentar sintomas de esquizofrenia, ou seja, à medida que os meses vão passando, há evidências de que a chance dos pacientes que estão sendo submetidos ao tratamento apresentarem sintomas diminui. Da mesma forma, o gênero contribui positivamente para a probabilidade dos pacientes apresentarem sintomas de esquizofrenia, isto é, pacientes do gênero feminino têm maior probabilidade de apresentarem os sintomas. Por fim, se o paciente tem menos de 20 anos de idade, a probabilidade de ele apresentar os sintomas da doença diminui em relação aos outros pacientes.

5.5 Desempenho Preditivo

Nesta seção, iremos avaliar o desempenho preditivo dos modelos ajustados para o conjunto de dados de Madras. Para isso, utilizaremos as métricas descritas no capítulo 2. Essas métricas são detalhadamente abordadas em James *et al.* (2013) e Hastie *et al.* (2009).

A seguir, apresentaremos as métricas definidas anteriormente para os quatro modelos propostos na seção 5.4, utilizando dados de esquizofrenia de Madras. Nas Tabelas 17, 18, 19 e 20, encontram-se as matrizes de confusão correspondentes aos modelos descritos anteriormente.

Tabela 17 – Matriz de confusão para o modelo 1: Modelo Probito.

	Valor Previsto	
Valor Real	0	1
0	563	74
1	156	129

Fonte: Elaborada pelo autor.

Tabela 18 – Matriz de confusão para o modelo 2: Modelo Probito misto.

	Valor Previsto	
Valor Real	0	1
0	567	70
1	139	146

Fonte: Elaborada pelo autor.

Na Tabela 21, apresentamos os resultados de algumas métricas de desempenho preditivo para os modelos 1, 2, 3 e 4 descritos na seção 5.5. Podemos concluir que não há diferenças

Tabela 19 – Matriz de confusão para o modelo 3: Modelo Skew - Probito.

	Valor Previsto	
Valor Real	0	1
0	563	74
1	156	129

Fonte: Elaborada pelo autor.

Tabela 20 – Matriz de confusão para o modelo 4: Modelo Skew - Probito Misto.

	Valor Previsto	
Valor Real	0	1
0	566	71
1	138	147

Fonte: Elaborada pelo autor.

significativas no desempenho preditivo quando consideramos os modelos Probito Misto ou Skew - Probito Misto para dados de esquizofrenia de Madras.

Tabela 21 – Resultados de algumas métricas de desempenho preditivos associados aos modelos 1, 2, 3 e 4.

	Modelo 1	**Modelo 2**	**Modelo 3**	**Modelo 4**
Acurácia	0,75	0,77	0,75	0,77
Revocação	0,45	0,51	0,45	0,52
Precisão	0,64	0,68	0,64	0,67
F1 - score	0,53	0,58	0,53	0,58
Área sob a curva ROC	0,79	0,82	0,79	0,81

Fonte: Elaborada pelo autor.

5.6 Conclusão

Neste capítulo, propomos modelos de regressão com resposta binária e função de ligação skew-probito com efeitos aleatórios. O modelo proposto mostrou-se melhor do que os modelos usuais. O sexo e a idade não afetam a incidência de sintomas de esquizofrenia. A taxa de declínio dos sintomas não difere nos subgrupos de sexo e idade. Além disso, analisamos a capacidade preditiva dos modelos propostos utilizando métricas como acurácia, revocação, precisão, F1-score e área sob a curva ROC.

Os modelos propostos foram implementados na linguagem **INLA** sem dificuldades. Eles também podem ser aplicados a outros problemas em que o uso de efeitos aleatórios seja justificado. No próximo capítulo, discutiremos algumas propostas futuras.

CAPÍTULO

6

CONSIDERAÇÕES FINAIS

Neste capítulo, discutiremos considerações finais e propostas para trabalhos futuros.

6.1 Comentários Finais

A modelagem de regressão para respostas binárias geralmente utiliza as funções de ligação probito e logito. Contudo, estudos na literatura, como os de Agresti (2012) e Prentice (1976), sugerem que em situações de desbalanceamento de dados, funções de ligação assimétricas podem ser mais adequadas. Com base nessa motivação, consideramos a função de ligação skew-probito como uma alternativa para problemas de dados desbalanceados. Em nosso trabalho, consideramos a função ligação skew-probito padrão. A função de ligação skew-probito padrão recebe esse nome devido à variável aleatória skew-normal que induz essa função de ligação, sendo escolhida de forma que tenha média zero e variância um.

Primeiramente, consideramos o modelo de regressão skew-probito padrão sob a abordagem bayesiana. Mostramos a nova abordagem para o processo de estimação do modelo de regressão skew-probito. Utilizamos a função de verossimilhança original e um processo Monte Carlo Hamiltoniano (HMC) para obter estimativas. Computacionalmente, o processo de estimação dos parâmetros é desenvolvido por meio da linguagem de programação Stan.

Realizamos três estudos de simulação. No primeiro, investigamos o método de estimação proposto para recuperar os verdadeiros valores dos parâmetros, usando RMSE, viés e a probabilidade de cobertura como instrumentos e avaliando a sensibilidade da distribuição *a priori* do parâmetro de assimetria. Nosso objetivo é identificar a configuração mais adequada das distribuições *a priori*. No segundo estudo, comparamos nosso método de estimação com o método INLA e refletimos sobre qual função de perda é mais conveniente considerar no processo de estimação. No terceiro estudo, comparamos dois métodos de detecção de outliers: o resíduo studentizado e o resíduo quantílico aleatorizado, no contexto do modelo skew-probito.

Finalmente, ilustramos os resultados dos estudos de simulação com uma aplicação em dados reais.

O processo de estimação de parâmetros realizado usando o algoritmo HMC e a função de verossimilhança original se mostra como uma alternativa viável à abordagem de dados aumentados proposta por Bazán *et al.* (2006), pois o primeiro estudo de simulação mostrou boa recuperação dos parâmetros e sugere considerar uma distribuição *a priori* para o parâmetro α, como sendo $N(0,4)$. Também mostramos que o método de estimação proposto é mais eficaz do que a abordagem INLA, sugerindo considerar uma função de perda absoluta ao estimar os parâmetros do modelo. Para detectar outliers no modelo de regressão skew-probito, o resíduo quantílico aleatorizado teve um desempenho superior do que seu concorrente em situações que possuem dados desbalanceados.

Na aplicação, o modelo de regressão skew-probito padronizado mostrou-se mais adequado do que os modelos de resposta binária usuais, e todas as covariáveis utilizadas foram significativas para explicar a probabilidade da presença ou ausência de diabetes nos pacientes. O método de estimação proposto mostrou-se eficiente e de fácil implementação, enquanto o resíduo quantílico aleatorizado apresentou bom desempenho na detecção de possíveis outliers.

No Capítulo 5, apresentamos uma extensão do modelo skew-probito proposto por Bazán, Bolfarine e Branco (2010), que é a iclusão de efeitos aleatórios. Realizamos os ajustes utilizando o método INLA, conforme descrito por Niekerk e Rue (2021), em uma aplicação aos dados de esquizofrenia de Madras. Optamos pelo método INLA devido à falta de convergência nas cadeias de Markov. Utilizamos os critérios de seleção de modelos DIC e WAIC para comparar o modelo de regressão skew-probito misto com os modelos skew-probito, probito e probito misto. Observamos que o DIC e o WAIC favorecem o modelo skew-probito misto, embora a diferença não seja significativa.

Conforme mencionado anteriormente, ajustamos o modelo de regressão skew-probito misto com intercepto aleatório para os dados de esquizofrenia de Madras, utilizando o método INLA. O intervalo de credibilidade de 95% é apresentado na Tabela 16, e observamos a partir disso que não há evidências de que a covariável idade seja significativa. Portanto, ajustamos o modelo reduzido sem a covariável idade e o comparamos com o modelo completo por meio de critérios de seleção de modelos, constatando que há evidências de que o modelo completo é preferível ao modelo reduzido.

Além disso, avaliamos o desempenho preditivo dos modelos ajustados aos dados de esquizofrenia de Madras utilizando métricas apropriadas, concluindo que os modelos probito misto e skew-probito misto apresentam desempenho preditivo superior aos outros modelos ajustados, nos quais não há presença de efeitos aleatórios na formulação do modelo. Ressaltamos ainda que não há evidências de diferença significativa no desempenho preditivo dos modelos ajustados quando a presença de efeitos aleatórios é considerada.

6.2 Produções científicas

- Coelho, F. R.; Bazán, J. L.; Russo, C. M. (2020). "Bayesian skew-probit regression model with random effects: An application to longitudinal data". UFSCar/USP, Brazil. *8th Workshop on Probabilistic and Statistical Methods.* Pôster.
- Coelho, F. R.; Russo, C. M.; & Bazán, J. L. (2022). "On outliers detection and prior distribution sensitivity in standard skew-probit regression models". *Brazilian Journal of Probability and Statistics*, 36(3), 441-462.

6.3 Propostas Futuras

- Refazer o estudo de simulação sobre outliers utilizando uma outra forma de perturbar as observações.
- Investigar o desempenho dos critérios de seleção de modelos (DIC, WAIC, entre outros), na seleção de modelos skew - probito misto.
- Propor medidas de diagnostico influência. O ponto de partida são medidas de divergência, tais como: divergência Kullback-Leibler, divergência J - distância, divergência qui-quadrado, entre outras. Tais medidas são apresentadas em Peng e Dey (1995) e Dey e Birmiwal (1994).
- Desenvolver métodos de seleção de variáveis. Uma possibilidade é ajustar o modelo de regressão skew - probito misto usando distribuições *a priori* "spyke and slab prior distribution".Como referência sobre o assunto, temos Andersen, Winther e Hansen (2014).
- Propor um nova versão do modelo de regressão skew-probito misto utilizando diferentes distribuições de probabilidade para os efeitos aleatórios.
- Desenvolver análise de resíduo para o modelo misto.
- Investigar a eficiência de métricas de desempenho preditivo, apresentadas no Capítulo 2, através de estudo de simulação.
- Comparar o desempenho da função de ligação skew - probito com algoritmos de aprendizagem de máquina.

REFERÊNCIAS

AGRESTI, A. **Categorical data analysis**. [S.l.]: John Wiley & Sons, 2012. v. 792. Citado na página 67.

ALBERT, J.; CHIB, S. Bayesian residual analysis for binary response regression models. **Biometrika**, Oxford University Press, v. 82, n. 4, p. 747–769, 1995. Citado na página 20.

ALVES, J. S.; BAZÁN, J. L.; ARELLANO-VALLE, R. B. Flexible cloglog links for binomial regression models as an alternative for imbalanced medical data. **Biometrical Journal**, Wiley Online Library, v. 65, n. 3, p. 2100325, 2023. Citado na página 21.

ALVES, J. S. B.; BAZÁN, J. L. New flexible item response models for dichotomous responses with applications. In: SPRINGER. **The Annual Meeting of the Psychometric Society**. [S.l.], 2022. p. 311–323. Citado na página 21.

ANDERSEN, M. R.; WINTHER, O.; HANSEN, L. K. Bayesian inference for structured spike and slab priors. **Advances in Neural Information Processing Systems**, v. 27, 2014. Citado na página 69.

ATKINSON, A. Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis. **Oxford Statistical Science Series, Oxford University Press: Oxford**, 1985. Citado na página 33.

AZZALINI, A. A class of distributions which includes the normal ones. **Scandinavian journal of statistics**, JSTOR, p. 171–178, 1985. Citado na página 25.

BASU, S.; MUKHOPADHYAY, S. Bayesian analysis of binary regression using symmetric and asymmetric links. **Sankhyā: The Indian Journal of Statistics, Series B**, JSTOR, p. 372–387, 2000. Citado na página 19.

BAZÁN, J. L.; BOLFARINE, H.; BRANCO, M. D. A framework for skew-probit links in binary regression. **Communications in Statistics—Theory and Methods**, Taylor & Francis, v. 39, n. 4, p. 678–697, 2010. Citado nas páginas 19, 20, 27, 28, 59, 62 e 68.

BAZÁN, J. L.; BRANCO, M. D.; BOLFARINE, H. *et al.* A skew item response model. **Bayesian analysis**, International Society for Bayesian Analysis, v. 1, n. 4, p. 861–892, 2006. Citado nas páginas 20, 27 e 68.

BESKOS, A.; PILLAI, N.; ROBERTS, G.; SANZ-SERNA, J.-M.; STUART, A. Optimal tuning of the hybrid monte carlo algorithm. 2013. Citado na página 81.

CARLIN, B. P.; LOUIS, T. A. **Bayesian methods for data analysis**. [S.l.]: CRC press, 2008. Citado na página 52.

CHEN, M.-H. Skewed link models for categorical response data. In: **Skew-Elliptical Distributions and Their Applications**. [S.l.]: Chapman and Hall/CRC, 2004. p. 151–172. Citado na página 19.

CHEN, M.-H.; DEY, D. K.; SHAO, Q.-M. A new skewed link model for dichotomous quantal response data. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 94, n. 448, p. 1172–1186, 1999. Citado nas páginas 19, 20, 27 e 28.

____. Bayesian analysis of binary data using skewed logit models. **Calcutta Statistical Association Bulletin**, SAGE Publications Sage India: New Delhi, India, v. 51, n. 1-2, p. 11–30, 2001. Citado na página 19.

COLLETT, D. **Modeling binary data**. [S.l.], 2003. Citado na página 19.

CORTES, R. X. Estimando modelos dinâmicos utilizando o inla para campos aleatórios markovianos não gaussianos. 2014. Citado na página 80.

DEY, D. K.; BIRMIWAL, L. R. Robust bayesian analysis using divergence measures. **Statistics & Probability Letters**, Elsevier, v. 20, n. 4, p. 287–294, 1994. Citado na página 69.

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 nas páginas 32 e 33.

FARIAS, R. B.; BRANCO, M. D. Efficient algorithms for bayesian binary regression model with skew-probit link. In: **Recent Advances in Biostatistics: False Discovery Rates, Survival Analysis, and Related Topics**. [S.l.]: World Scientific, 2011. p. 143–168. Citado na página 20.

FARIAS, R. B.; BRANCO, M. D. *et al.* Latent residual analysis in binary regression with skewed link. **Brazilian Journal of Probability and Statistics**, Brazilian Statistical Association, v. 26, n. 4, p. 344–357, 2012. Citado na página 20.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; DUNSON, D. B.; VEHTARI, A.; RUBIN, D. B. **Bayesian data analysis**. [S.l.]: CRC press, 2013. Citado na página 31.

GELMAN, A.; GILKS, W. R.; ROBERTS, G. O. Weak convergence and optimal scaling of random walk metropolis algorithms. **The annals of applied probability**, Institute of Mathematical Statistics, v. 7, n. 1, p. 110–120, 1997. Citado na página 81.

GELMAN, A.; LEE, D.; GUO, J. Stan: A probabilistic programming language for bayesian inference and optimization. **Journal of Educational and Behavioral Statistics**, Sage Publications Sage CA: Los Angeles, CA, v. 40, n. 5, p. 530–543, 2015. Citado na página 52.

GELMAN, A.; RUBIN, D. B. *et al.* Inference from iterative simulation using multiple sequences. **Statistical science**, Institute of Mathematical Statistics, v. 7, n. 4, p. 457–472, 1992. Citado nas páginas 38 e 52.

HAILPERN, S. M.; VISINTAINER, P. F. Odds ratios and logistic regression: further examples of their use and interpretation. **The Stata Journal**, SAGE Publications Sage CA: Los Angeles, CA, v. 3, n. 3, p. 213–225, 2003. Citado na página 19.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H.; FRIEDMAN, J. H. **The elements of statistical learning: data mining, inference, and prediction**. [S.l.]: Springer, 2009. v. 2. Citado nas páginas 33 e 65.

HENDERSON, N. C.; LOUIS, T. A.; WANG, C.; VARADHAN, R. Bayesian analysis of heterogeneous treatment effects for patient-centered outcomes research. **Health Services and Outcomes Research Methodology**, Springer, v. 16, p. 213–233, 2016. Citado na página 52.

HOFFMAN, M. D.; GELMAN, A. *et al.* The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. **J. Mach. Learn. Res.**, v. 15, n. 1, p. 1593–1623, 2014. Citado na página 38.

HUAYANAY, A. de la C. Modelos alternativos para classificação em dados desbalanceados. Universidade Federal de São Carlos, 2023. Citado na página 21.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. *et al.* **An introduction to statistical learning**. [S.l.]: Springer, 2013. v. 112. Citado nas páginas 33 e 65.

KING, G.; ZENG, L. Logistic regression in rare events data. **Political analysis**, Cambridge University Press, v. 9, n. 2, p. 137–163, 2001. Citado na página 19.

LEE, D.; SINHA, S. Identifiability and bias reduction in the skew-probit model for a binary response. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 89, n. 9, p. 1621–1648, 2019. Citado na página 20.

LEISCH, F.; DIMITRIADOU, E. **mlbench: Machine Learning Benchmark Problems**. [S.l.], 2010. R package version 2.1-1. Citado na página 51.

LESAFFRE, E.; LAWSON, A. B. **Bayesian biostatistics**. [S.l.]: John Wiley & Sons, 2012. Citado na página 32.

MELTZER, E. B.; BARRY, W. T.; D'AMICO, T. A.; DAVIS, R. D.; LIN, S. S.; ONAITIS, M. W.; MORRISON, L. D.; SPORN, T. A.; STEELE, M. P.; NOBLE, P. W. Bayesian probit regression model for the diagnosis of pulmonary fibrosis: proof-of-principle. **BMC medical genomics**, BioMed Central, v. 4, n. 1, p. 1–13, 2011. Citado na página 19.

NARANJO, L.; PÉREZ, C. J.; MARTÍN, J. Skewed link-based regression models for misclassified binary data. **Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A. Matemáticas**, Springer, v. 113, n. 2, p. 1585–1599, 2019. Citado nas páginas 21 e 51.

NEWMAN, D.; HETTICH, S.; BLAKE, C.; MERZ, C. **UCI Repository of machine learning databases**. 1998. Disponível em: <http://www.ics.uci.edu/~mlearn/MLRepository.html>. Citado na página 51.

NIEKERK, J. V.; RUE, H. Skewed probit regression—identifiability, contraction and reformulation. **Revstat Statistical Journal**, Instituto Nacional De Estatistica, v. 19, n. 1, p. 1–22, 2021. Citado nas páginas 21 e 68.

ORDOÑEZ, J. A.; PRATES, M. O.; BAZÁN, J. L.; LACHOS, V. H. Penalized complexity priors for the skewness parameter of power links. **Canadian Journal of Statistics**, Wiley Online Library, 2023. Citado na página 21.

PAAL, B. Van der. A comparison of different methods for modelling rare events data. **PhD thesis**, Ghent University, 2014. Citado na página 19.

PAIXÃO, R. S. Método zero-variance para monte carlo hamiltoniano aplicado a modelos garch univariados e multivariados. Universidade Federal de São Carlos, 2021. Citado nas páginas 81 e 82.

PENG, F.; DEY, D. K. Bayesian analysis of outlier problems using divergence measures. **Canadian Journal of Statistics**, Wiley Online Library, v. 23, n. 2, p. 199–213, 1995. Citado na página 69.

PRATES, M. O.; DEY, D. K.; WILLIG, M. R.; YAN, J. Intervention analysis of hurricane effects on snail abundance in a tropical forest using long-term spatiotemporal data. **Journal of Agricultural, Biological, and Environmental Statistics**, Springer, v. 16, p. 142–156, 2011. Citado na página 57.

PRENTICE, R. L. A generalization of the probit and logit methods for dose response curves. **Biometrics**, JSTOR, p. 761–768, 1976. Citado na página 67.

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 45, 79 e 80.

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 30.

Stan Development Team. **RStan: the R interface to Stan**. 2019. R package version 2.19.2. Disponível em: <http://mc-stan.org/>. Citado nas páginas 34 e 49.

VEHTARI, A.; GELMAN, A.; GABRY, J. Practical bayesian model evaluation using leave-one-out cross-validation and waic. **Statistics and computing**, Springer, v. 27, n. 5, p. 1413–1432, 2017. Citado na página 31.

WANG, X.; DEY, D. K. *et al.* Generalized extreme value regression for binary response data: An application to b2b electronic payments system adoption. **The Annals of Applied Statistics**, Institute of Mathematical Statistics, v. 4, n. 4, p. 2000–2023, 2010. Citado na página 19.

WATANABE, S. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. **Journal of Machine Learning Research**, v. 11, n. Dec, p. 3571–3594, 2010. Citado na página 31.

YAN, G.; SEDRANSK, J. A note on bayesian residuals as a hierarchical model diagnostic technique. **Statistical Papers**, Springer, v. 51, p. 1–10, 2010. Citado na página 32.

APÊNDICE

A

CÓDIGOS EM *RSTAN*.

O código STAN usado para estimar o modelo na aplicação é mostrado abaixo.

```
skew.probito_cod <- '
data {
int<lower=0> n;  // numero de observac(
int<lower=0,upper=1> y[n];          // variavel resposta
int<lower=0,upper=17> pregnant[n];
int<lower=44,upper=199> glucose[n];
real<lower=18.2,upper=67.1> mass[n];
int<lower=24,upper=122> pressure[n];
real<lower=0.07,upper=2.42> pedigree[n];
int<lower=21,upper=81> age[n];
}
parameters {
vector[6] beta;
real alfa;
real beta0;
}
transformed parameters {
vector[n] prob;
vector[n] eta;
for(i in 1:n){
eta[i]=  beta0 + beta[1]*pregnant[i] + beta[2]*glucose[i] +
beta[3]*mass[i] + beta[4]*pressure[i] + beta[5]*pedigree[i] +
beta[6]*age[i];
prob[i] = skew_normal_cdf(eta[i],(-sqrt(2)*alfa)/(sqrt(pi() +
```

```
(pi() - 2)*pow(alfa,2))),(pi() +pi()*pow(alfa,2))/(pi() +
(pi() - 2)*pow(alfa,2)),alfa);
}
}
model{
for (i in 1:6){beta[i] ~ normal(0,100); }
beta0 ~ normal(0, 100);
alfa ~ normal(0, 2);
y ~ bernoulli(prob);
}
generated quantities {
real dev;
vector[n] log_lik;
dev = 0;
for (i in 1:n){
log_lik[i] = bernoulli_lpmf(y[i] | prob[i]);
dev = dev + (-2)*log_lik[i];
}
}
'
```

APÊNDICE

B

GRÁFICO DAS SAÍDAS MCMC.

A Figura 6 mostra o histograma das saídas do MCMC para cada um dos parâmetros do modelo final (equação (4.4)) e a Figura 7 mostra o trace plot das saídas MCMC para cada um dos parâmetros. Podemos ver que a convergência da cadeia de Markov é obtida para a distribuição marginal *a posteriori* dos parâmetros α e $\boldsymbol{\beta}$.

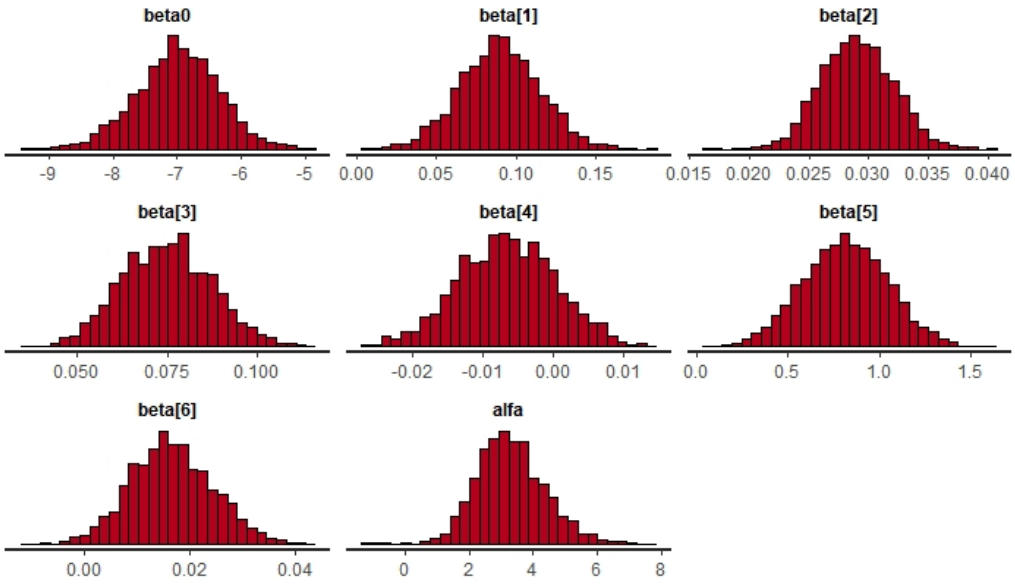


Figura 6 – Histogramas das saídas do MCMC para o modelo final. Fonte: Elaborada pelo autor.

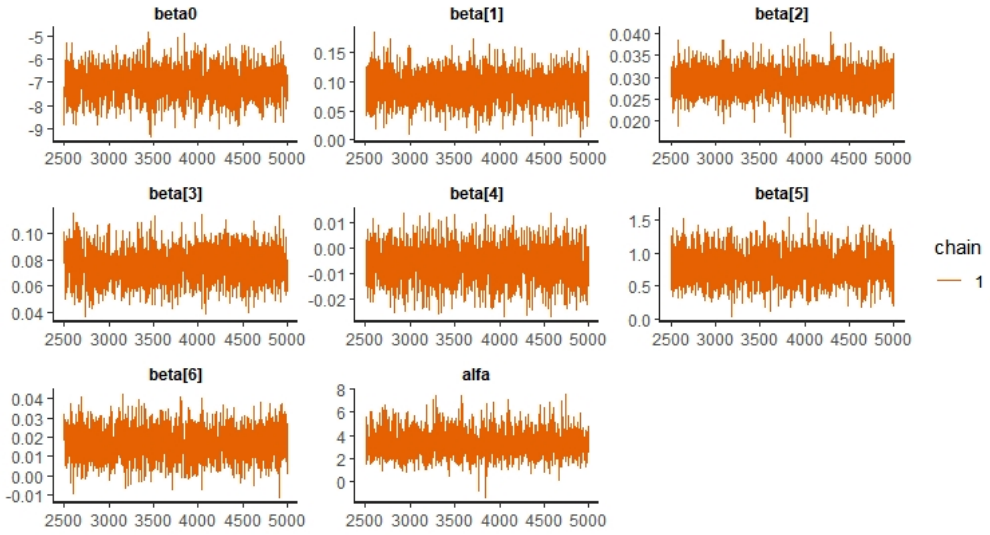


Figura 7 – Trace plot para cada um dos parâmetros no modelo final. Elaborada pelo autor.

APÊNDICE

C

O MÉTODO INLA.

O método INLA (Approximate Bayesian Inference Using Integrated Nested Laplace Approximations) é uma técnica avançada de inferência bayesiana que visa fornecer uma aproximação eficaz da distribuição *a posteriori* em modelos estatísticos complexos. Desenvolvido recentemente por Rue, Martino e Chopin (2009), o INLA combina técnicas de integração numérica e aproximação de Laplace para oferecer uma alternativa computacionalmente eficiente aos métodos tradicionais de inferência bayesiana, como o Markov Chain Monte Carlo (MCMC). Os métodos MCMC podem ser computacionalmente intensivos e requerem um tempo significativo para convergir, especialmente em modelos estatísticos complexos. O INLA é amplamente utilizado em várias áreas, incluindo estatística, epidemiologia, geociências e outras.

A principal ideia por trás do INLA é aproximar a distribuição *a posteriori* de um modelo estatístico complexo por meio de uma distribuição Gaussiana multivariada, que é mais fácil de lidar computacionalmente. O método INLA combina elementos da teoria de aproximações bayesianas e técnicas de integração numérica para obter resultados aproximados de alta qualidade em uma fração do tempo necessário para a execução de uma cadeia de Markov Monte Carlo.

O processo geral do INLA pode ser dividido em quatro etapas principais:

- **Construção do Modelo:** Começa-se com a especificação de um modelo estatístico, que inclui a escolha de uma distribuição *a priori* para os parâmetros do modelo e uma função de verossimilhança que relaciona os dados aos parâmetros.

- **Aproximação da distribuição *a posteriori*:** O INLA aproxima a distribuição *a posteriori* dos parâmetros usando uma combinação de uma distribuição Gaussiana multivariada (aproximação de Laplace) e uma técnica chamada discretização do espaço de parâmetros (aproximação de grade). Essas duas abordagens aproximam a distribuição *a posteriori* de maneira eficiente.

- **Cálculo de Estatísticas de Interesse:** Uma vez que a distribuição *a posteriori* aproximada é obtida, é possível calcular várias estatísticas de interesse, como médias, medianas, intervalos de credibilidade e muito mais.

- **Diagnóstico e Avaliação:** É importante realizar diagnósticos para verificar a qualidade da aproximação do INLA. Isso pode envolver a verificação da convergência da aproximação e a avaliação da adequação do modelo.

A eficiência do INLA em fornecer resultados de alta qualidade e sua aplicabilidade em uma ampla gama de modelos têm contribuído para sua crescente popularidade em diversas áreas científicas. Segundo Cortes (2014), a principal diferença entre o INLA e as abordagens tradicionais é que não há necessidade de simulações estocásticas de uma distribuição marginal *a posteriori*. Ele afirma que as principais vantagens da abordagem em questão são:

- Alta qualidade do ajuste;
- Agilidade computacional; e
- Método que não sofre de problemas de convergência MCMC.

Porém, o método INLA também possui algumas limitações:

- **Aproximação:** Como o próprio nome sugere, o INLA é uma técnica de aproximação. Embora seja altamente eficiente, a qualidade da aproximação pode variar dependendo do modelo e dos dados.

- **Restrições no Modelo:** Nem todos os modelos estatísticos podem ser facilmente tratados pelo INLA. Modelos muito complexos ou com características especiais podem não ser adequados para esta abordagem.

- **Requer Aprendizado:** Dominar o método INLA requer um certo aprendizado, especialmente para a escolha adequada de hiperparâmetros e a interpretação dos resultados.

O método INLA foi proposto por Rue, Martino e Chopin (2009) para uma classe específica de modelos, conhecidos como modelos gaussianos latentes, nos quais a variável de resposta Y_i, com média μ_i, está relacionada à estrutura aditiva do preditor η_i por meio da função de ligação $g(\mu_i) = \eta_i$. A estrutura do preditor linear leva em consideração a presença das covariáveis e pode ser escrita como

$$\eta_i = \alpha + \sum_{j=1}^{n_h} h^{(j)}\left(\mu_{ji}\right) + \sum_{k=1}^{\eta_\beta} \beta z_{ki}. \tag{C.1}$$

APÊNDICE

D

O ALGORITMO NUTTS.

Neste apêndice, introduzimos o algoritmo NUTS, uma das abordagens mais comuns para estimar parâmetros de modelos de regressão de forma bayesiana.

D.1 Método HMC (Hamiltonian Monte Carlo)

Os desafios decorrentes das baixas taxas de aceitação de candidatos nas técnicas de Metropolis e do desempenho limitado do algoritmo de Gibbs em problemas multidimensionais com geometrias complexas nas distribuições *a posteriori*, levaram ao desenvolvimento de uma nova técnica MCMC que utiliza a dinâmica Hamiltoniana, em homenagem ao físico irlandês William Rowan Hamilton (1805-1865).

O HMC é uma adaptação da técnica Metropolis que emprega um esquema direcionado para a geração de novos candidatos. Isso aprimora a taxa de aceitação e, consequentemente, a eficiência. Mais especificamente, o HMC utiliza o gradiente do logaritmo da distribuição *a posteriori* para orientar a cadeia de Markov em direção às regiões de maior densidade, onde a maioria das amostras é coletada. Como resultado, uma cadeia de Markov com o algoritmo HMC bem ajustado aceitará candidatos em uma taxa muito mais alta do que o algoritmo Metropolis tradicional.(Beskos *et al.* (2013), Gelman, Gilks e Roberts (1997))

Segundo Paixão (2021), o HMC é descrito de forma simplificada da seguinte maneira: o método propõe valores candidatos para $(\theta, \mathbf{r})$, gerados em dois estados, antes de serem submetidos ao passo de aceitação do método Metropolis-Hastings. No primeiro estado, um valor de $\mathbf{r}$ é simulado a partir de uma distribuição normal com média $\mathbf{0}$ e matriz de covariância M, independente de θ. No segundo estado, é simulado um sistema conjunto em $(\theta, \mathbf{r})$ que segue a

dinâmica hamiltoniana. O sistema é desenvolvido por meio das equações hamiltonianas D.1.

$$\begin{aligned} \frac{\partial \theta}{\partial t} &= \frac{H(\theta,\mathbf{r})}{\partial \mathbf{r}} = \frac{\partial K(\mathbf{r})}{\partial \mathbf{r}}, \\ \frac{\partial \mathbf{r}}{\partial t} &= -\frac{H(\theta,\mathbf{r})}{\partial \theta} = -\frac{\partial U(\theta)}{\partial \theta}. \end{aligned} \tag{D.1}$$

As equações hamiltonianas podem ser aproximadas através da discretização do tempo, usando um pequeno passo de tamanho ε através do método *leapfrog*. Este método é utilizado para resolver as equações hamiltonianas por meio da aplicação de L passos, cada um definido por

$$\begin{aligned} \mathbf{r}_i\left(t+\tfrac{\varepsilon}{2}\right) &= \mathbf{r}_i(t) - \tfrac{\varepsilon}{2}\tfrac{\partial}{\partial \theta_i}U(\theta(t)) \\ \theta_i(t+\varepsilon) &= \theta_i\left(t+\tfrac{\varepsilon}{2}\right) - \varepsilon\tfrac{\partial}{\partial r_i}K\left(\mathbf{r}\left(t+\tfrac{\varepsilon}{2}\right)\right) \\ \mathbf{r}_i(t+\varepsilon) &= \mathbf{r}_i\left(t+\tfrac{\varepsilon}{2}\right) - \tfrac{\varepsilon}{2}\tfrac{\partial}{\partial \theta_i}U(\theta(t+\varepsilon)). \end{aligned} \tag{D.2}$$

Ao final, é simulado o estado (θ^*, r^*), que corresponde ao valor do sistema no tempo fictício $L\varepsilon$. Em seguida, utiliza-se a etapa de aceitação do método *Metropolis - Hastings*, em que o estado (θ^*, ω^*) é aceito como próximo estado da cadeia de MArkov com a seguinte probabilidade:

$$\mathbb{P}(\theta,\omega;\theta^*,\omega^*) = \min\{1, \exp\{H(\theta,\omega) - H(\theta^*,\omega^*)\}\} \tag{D.3}$$

Segundo Paixão (2021), para execução do método HMC, deve-se definir a escolha de três parâmetros.

- ε, referente à probabilidade de aceitação apresentada na equação D.3. Quanto menor for seu valor, maiores serão a probabilidade de aceitação e a auto-dependência da cadeia.
- L, referente à autodependência da cadeia. Quanto maior for seu valor, menor será essa autodependência, porém maior será o custo computacional.
- M, que é a combinação do dois parâmetros recém-mencionados.

Apesar de existirem alguns algoritmos para encontrar o valor ideal para cada um desses parâmetros, não há consenso na literatura quanto ao critério de escolha.

D.2 Algoritmo NUTS

O algoritmo NUTS (*No-U-Turn Sampler*) é uma técnica avançada de amostragem usada na inferência Bayesiana. Ele é uma extensão eficiente do algoritmo de Monte Carlo Hamiltoniano (HMC), que, por sua vez, é baseado na dinâmica Hamiltoniana da física.

O algoritmo HMC é utilizado para gerar amostras de uma distribuição de probabilidade alvo, simulando o comportamento de uma partícula fictícia em um campo de força definido pelo

logaritmo negativo da probabilidade do modelo estatístico. A partícula é simulada de acordo com as equações de movimento da dinâmica Hamiltoniana. No entanto, amostrar eficientemente nessa dinâmica requer a especificação de um tamanho de passo adequado, o que pode ser desafiador em problemas complexos.

O NUTS resolve esse problema ajustando automaticamente o tamanho do passo durante a amostragem, eliminando a necessidade de ajuste manual de hiperparâmetros. Isso é realizado por meio de um algoritmo de amostragem que decide adaptativamente quanto tempo a partícula fictícia deve se mover em cada direção antes de decidir se deve ou não retornar, daí o nome "No-U-Turn Sampler".

Agora, vamos mergulhar um pouco mais nas considerações matemáticas por trás do NUTS. O algoritmo utiliza uma abordagem de dinâmica Hamiltoniana para simular a trajetória da partícula fictícia. Isso requer o cálculo do gradiente da log-verossimilhança dos dados em relação aos parâmetros ($\nabla \log p(D|\theta)$) e o gradiente do logaritmo da distribuição *a priori* dos parâmetros ($\nabla \log p(\theta)$). Esses gradientes são utilizados para atualizar a posição e a quantidade de movimento da partícula fictícia ao longo do tempo.

Durante a amostragem, o algoritmo NUTS inicia com uma partícula fictícia em um determinado ponto do espaço de parâmetros e simula seu movimento de acordo com as equações de movimento da dinâmica Hamiltoniana por um tempo aleatório. Durante esse processo, ele ajusta adaptativamente o tamanho do passo e decide se deve continuar a mover-se em uma determinada direção ou se deve parar e retornar, dependendo da condição "sem retorno".

A condição "sem retorno"é verificada continuamente enquanto a partícula está em movimento. Ela determina se a partícula está retornando a uma região previamente explorada, o que indica que a amostragem está progredindo bem e não há necessidade de continuar explorando nessa direção. Isso aumenta significativamente a eficiência da amostragem, evitando gastar tempo amostrando áreas do espaço de parâmetros que já foram exploradas adequadamente.

Em resumo, o algoritmo NUTS é uma técnica sofisticada de amostragem que combina a dinâmica Hamiltoniana com a parada adaptativa sem retorno para amostrar eficientemente distribuições de probabilidade complexas. Sua capacidade de ajustar automaticamente o tamanho do passo e decidir quando parar de explorar determinadas direções o torna uma ferramenta poderosa para a inferência Bayesiana em modelos estatísticos complexos.


ICMC
USP
SÃO CARLOS