UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Novos desenvolvimentos para dados de contagem

Naiara Caroline Aparecido dos Santos

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

ICMC USP SÃO CARLOS

Naiara Caroline Aparecido dos Santos

Novos desenvolvimentos para dados de contagem

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

Área de Concentração: Estatística

Orientador: Prof. Dr. Jorge Luis Bazán Guzmán
Coorientador: Prof. Dr. Artur José Lemonte

USP – São Carlos
Setembro 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)

A237n

Aparecido Dos Santos, Naiara Caroline
Novos desenvolvimentos para dados de contagem / Naiara Caroline Aparecido Dos Santos; orientador Jorge Luis Bazán Guzmán; coorientador Artur José Lemonte. -- São Carlos, 2024.
86 p.

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

1. Dados de contagem. 2. Análise de resíduos. 3. Estimação bayesiana. 4. Modelos Rasch. 5. Modelos mistos. I. Bazán Guzmán, Jorge Luis, orient. II. Lemonte, Artur José, 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

Naiara Caroline Aparecido dos Santos

New developments to counts data

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

Concentration Area: Statistics

Advisor: Prof. Dr. Jorge Luis Bazán Guzmán
Co-advisor: Prof. Dr. Artur José Lemonte

USP – São Carlos
September 2024

UNIVERSIDADE FEDERAL DE SÃO CARLOS

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

Folha de Aprovação

Defesa de Tese de Doutorado da candidata Naiara Caroline Aparecido dos Santos, realizada em 30/07/2024.

Comissão Julgadora:

Prof. Dr. Jorge Luis Bazán Guzmán (USP)

Prof. Dr. Juvêncio Santos Nobre (UFC)

Prof. Dr. Jorge Andrés González Burgos (PUC-Chile)

Profa. Dra. Paula Fariña (UDP)

Profa. Dra. Katiane Silva Conceição (USP)

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

Aos meus pais, Vania e João,
a minha madrinha Marlene,
sem os quais minha vida seria um ponto final.

AGRADECIMENTOS

A vida acadêmica é uma verdadeira montanha-russa, e sem o apoio incondicional da família e das grandes amizades, não seria possível enfrentá-la. Durante essa longa e desafiadora jornada, tive a sorte de contar com as melhores pessoas ao meu lado. Cada uma delas, com sua presença, carinho e incentivo, foi fundamental para que eu mantivesse a força e a determinação necessárias para seguir em frente. A cada curva, subida e descida dessa montanha-russa, o apoio inabalável que recebi fez toda a diferença, e esta conquista é, sem dúvida, compartilhada com cada um de vocês.

Primeiramente, agradeço a Deus, que me concedeu saúde, força e sabedoria para perseverar em cada etapa dessa caminhada.

À minha família, meu porto seguro. Aos meus pais, João e Vânia, sou profundamente grata pelo amor incondicional, ensinamentos e por sempre acreditarem no meu potencial. Vocês são minha maior inspiração! À minha madrinha, Marlene, que sempre esteve presente para me ajudar no que fosse preciso, apoiando-me de todas as formas. Ao meu companheiro, William, meu amor e gratidão por sua paciência, compreensão e apoio constante, por sempre estar ao meu lado em todos os momentos.

Aos meus queridos amigos Breno e Talita, que me acompanham desde a graduação. Nossa jornada acadêmica foi longa, cheia de desafios e conquistas, e nunca soltamos a mão um do outro. Passamos por muitos momentos juntos, apoiando-nos mutuamente em cada etapa, e essa parceria, que começou nos estudos, estendeu-se para a vida. Agradeço de coração pelo apoio incondicional e por estarem sempre ao meu lado.

Minha profunda gratidão às meninas do apartamento, Graziela e Izadora, com quem dividi a moradia durante essa fase. Vocês tornaram minha vida muito mais leve enquanto estive longe da família. Agradeço pelos momentos inesquecíveis que compartilhamos, pelo apoio constante e pela companhia, que fizeram toda a diferença em minha caminhada. Vocês foram um suporte essencial, e sou imensamente grata por cada memória que criamos juntas.

Aos meus amigos e colegas de jornada, Patricia, Jessica, Alex, Danilo, Isaac, Gabriel, Ana, Marina, Laila e Adriane, que compartilharam comigo esta experiência acadêmica, sou eternamente grata pelas conversas, pelas inúmeras horas de estudo, pelo apoio mútuo e por todos os momentos vividos. Sem vocês, essa trajetória teria sido muito mais árdua.

Ao meu orientador, Professor Jorge Bazán, minha mais sincera gratidão por acreditar no meu potencial e por me guiar com paciência, conhecimento e gentileza ao longo deste processo.

Seus conselhos e dedicação foram fundamentais para que este trabalho tomasse forma e, sem dúvida, meu crescimento profissional e pessoal está diretamente ligado à sua orientação.

Aos professores e colaboradores da USP e UFSCar, pelo ensino de excelência e pelas oportunidades de aprendizado que me proporcionaram. Em especial, à Professora Vera Tomazzela, por todo o carinho, incentivo e confiança. Suas palavras e orientação foram fundamentais para que eu explorasse todo o meu potencial e enxergasse novas oportunidades. À Professora Juliana Cobre, pela experiência de estágio ao seu lado, que me proporcionou um crescimento profissional imensurável. Obrigada por abrir meus olhos para uma nova perspectiva da estatística.

Aos professores membros da banca, Juvêncio Nobre, Jorge González, Paula Fariña e Katiane Conceição, sou extremamente grata por dedicarem seu tempo para avaliar este trabalho. Suas contribuições e comentários foram inestimáveis para o aprimoramento da pesquisa.

Agradeço a todos que, direta ou indiretamente, contribuíram para a realização deste trabalho. Cada gesto, palavra de encorajamento e colaboração foi essencial para que eu chegasse até aqui.

E, por fim, se minha tese fosse um modelo de machine learning, vocês, amigos e familiares, seriam as variáveis mais significativas no meu algoritmo. Obrigada por me ajudarem a evitar o overfitting da vida acadêmica!

“Nem todo mundo vai compreender
isso tudo que você é
o que não significa
que você deva se esconder
ou se calar
O mundo tem medo
de mulheres extraordinárias.”
(Ryane Leão)

“Mesmo as noites,
totalmente sem estrelas
podem anunciar a aurora
de uma grande realização.”
(Martin Luther King)

RESUMO

SANTOS, N.C.A. **Novos desenvolvimentos para dados de contagem**. 2024. 86 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.

Este trabalho investiga novos desenvolvimentos para a análise de dados de contagem, concentrando-se em duas metodologias: modelos de Teoria de Resposta ao Item (TRI) e Modelos Lineares Generalizados Mistos (MLGM). A pesquisa foca no desenvolvimento e na aplicação de métodos clássicos e bayesianos para aprimorar a análise de modelos de contagem existentes e na proposição de novos modelos. Os capítulos desta tese compreendem manuscritos desenvolvidos ao longo do doutorado. Primeiramente, é apresentado um estudo com o modelo de contagem Rasch Poisson, já presente na literatura, visando um melhor entendimento e introduzindo uma nova abordagem por meio do método de aproximações de Laplace encaixadas e integradas. São mostradas técnicas de análise de resíduos através de visualização gráfica, utilizando os resíduos quantílicos aleatorizados, e a metodologia é aplicada na área de Psicologia. exploramos modelos alternativos para respostas de contagem, que superam algumas das limitações do modelo Rasch-Poisson. Detalhamos sua formulação e métodos de estimação, sob as abordagens Clássica e Bayesiana. Além disso, demonstramos o potencial da análise de resíduos por meio de gráficos aplicados a dados de um teste de atenção. A seguir, ao considerar modelos mistos, introduzimos uma nova proposta para respostas de contagem, baseada na distribuição Bell de um parâmetro, explicitando os detalhes de sua formulação e estimação sob as abordagens Clássica e Bayesiana. Avaliamos a recuperação de parâmetros da metodologia de estimação proposta por meio de um estudo de simulação e também mostramos o potencial de seu uso em uma aplicação para dados de ataques epilépticos. Por fim, propomos um novo modelo de regressão misto, baseado na distribuição Bell-Touchard parametrizada pela média. São apresentados estudos de simulação e a metodologia é aplicada em um experimento neurofisiológico. Os diversos estudos e aplicações ao longo do texto mostram que as propostas trazem bons resultados e têm potencial de uso por pesquisadores de diversas áreas, com os códigos utilizados para a estimação dos parâmetros disponibilizados.

Palavras-chave: Dados de contagem, Análise de resíduos, Estimação bayesiana, Modelos Rasch, Modelos mistos.

ABSTRACT

SANTOS, N.C.A. **New developments to counts data**. 2024. 86 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.

This work investigates new developments in count data analysis, focusing on two methodologies: Item Response Theory (IRT) models and Generalized Linear Mixed Models (GLMM). The research focuses on the development and application of classical and Bayesian methods to improve the analysis of existing count models and to propose new models. The chapters of this thesis comprise manuscripts developed throughout the doctoral program. First, a study is presented with the Rasch Poisson count model, already existing in the literature, aiming at a better understanding and introducing a new approach through the method of nested and integrated Laplace approximations. Techniques for residual analysis are shown through graphical visualization, using randomized quantile residuals, and the methodology is applied in the field of Psychology. We explore alternative models for count responses, which overcome some of the limitations of the Rasch-Poisson model. We detail its formulation and estimation methods, under both Classical and Bayesian approaches. Additionally, we demonstrate the potential of residual analysis using graphs applied to data from an attention test. Next, when considering mixed models, we introduce a new proposal for count responses, based on the one-parameter Bell distribution, explicitly detailing its formulation and estimation under Classical and Bayesian approaches. We evaluate the parameter recovery of the proposed estimation methodology through a simulation study and also show the potential of its use in an application to epileptic seizure data. Finally, we propose a new mixed regression model based on the Bell-Touchard distribution parameterized by the mean. Simulation studies are presented, and the methodology is applied in a neurophysiological experiment. The various studies and applications throughout the text show that the proposals yield good results and have potential use by researchers in various fields, with the codes used for parameter estimation made available.

Keywords: Count data, Residual analysis, Bayesian estimation, Rasch models, Mixed models.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

A análise de dados de contagem emerge como um componente fundamental em uma diversidade de campos de aplicação, destacando-se por sua relevância na medicina, educação e psicologia, entre outros. No contexto médico, essa análise é crucial, por exemplo, para o monitoramento eficaz da frequência de interações clínicas em pesquisas longitudinais, como no caso de idosos em Minnesota (WALLER; ZELTERMAN, 1997), bem como para o acompanhamento da frequência de convulsões em indivíduos acometidos por epilepsia (THALL; VAIL, 1990). No âmbito educacional e psicológico, as técnicas de análise de dados de contagem permitem avaliar habilidades específicas, como a capacidade de distinção rápida de dígitos sob condições de tempo limitado em testes de atenção (BEYZAEE, 2017), ou na mensuração da acurácia da memória, onde os participantes são desafiados a identificar emblemas em contextos variados (JENDRYCZKO; BERKEMEYER; HOLLING, 2020).

Dados dessa natureza podem apresentar particularidades como equidispersão, onde a média e a variância são iguais, subdispersão ou superdispersão, nas quais a variância é, respectivamente, menor ou maior que a média, além da presença de excessos de zeros. Tais características são extensivamente discutidas na literatura (MOLENBERGHS; VERBEKE; DEMÉTRIO, 2007; SELLERS; SHMUELI, 2010; FORTHMANN; GÜHNE; DOEBLER, 2019; PINHEIRO *et al.*, 2019; SIDUMO; SONONO; TAKAIDZA, 2023), sublinhando a contínua evolução deste campo de estudo.

Tradicionalmente, a distribuição Poisson é o modelo mais popular utilizado para a análise de dados de contagem, caracterizada pela premissa de equidispersão, na qual a média e a variância dos dados são iguais (MCCULLAGH, 2019). No entanto, essa característica de equidispersão muitas vezes não se sustenta em aplicações práticas, limitando a aplicabilidade do modelo Poisson em contextos variados. Em resposta a essas limitações, foram desenvolvidas várias distribuições alternativas que buscam oferecer maior flexibilidade, tais como: o modelo Binomial Negativo; o modelo Poisson-Normal (HINDE, 1982); o modelo Conway–Maxwell–Poisson

(COM-Poisson) (CONWAY; MAXWELL, 1962); o modelo Bell (CASTELLARES; FERRARI; LEMONTE, 2018); e sua extensão, o modelo Bell-Touchard (CASTELLARES; LEMONTE; MORENO-ARENAS, 2020). Para situações com excesso de zeros, tem-se o modelo Poisson Inflacionada de Zeros (ZIP) e o Binomial Negativa Inflacionada de Zeros (ZINB) (RIDOUT; DEMÉTRIO; HINDE, 1998), além do Bell Inflacionada de Zeros (ZIBELL) (LEMONTE; MORENO-ARENAS; CASTELLARES, 2020), entre outros modelos.

Diversas abordagens podem ser adotadas para modelar adequadamente a relação média-variância, com destaque para a estrutura proposta por Breslow e Clayton (1993), conhecida como Modelos Lineares Generalizados Mistos (MLGM). Esses modelos combinam Modelos Lineares Generalizados (MLG) com efeitos aleatórios normais no preditor linear. No entanto, a estrutura de efeitos aleatórios pode não ser suficiente para modelar os dados de forma adequada. Nesses casos, torna-se necessário um modelo de contagem mais flexível que considere a dispersão. Portanto, diversas distribuições de probabilidade discretas, adequadas para descrever fenômenos de contagem, foram introduzidas no contexto de modelos mistos (BOOTH *et al.*, 2003; ZHANG *et al.*, 2017; MORRIS; SELLERS, 2017).

No contexto da Teoria de Resposta ao Item (TRI), aplicada principalmente em testes de desempenho educacional para quantificar acertos e erros, a inadequação dos modelos de Poisson também se torna evidente. Diante dessa limitação, foram propostos modelos alternativos ou extensões do modelo Poisson, como o modelo Rasch Poisson bidimensional de Forthmann *et al.* (2018), TRI Poisson inflacionado de zeros de Wang (2010), TRI Binomial negativa de Hung (2012), TRI COM-Poisson de Forthmann, Gühne e Doebler (2019), entre outros.

Diante das considerações apresentadas, torna-se evidente a necessidade contínua de desenvolver e implementar modelos e ferramentas estatísticas flexíveis para a modelagem de dados de contagem, assim como de estudar os critérios de comparação de modelos e a análise de resíduos. Neste contexto, o objetivo desta tese é apresentar procedimentos metodológicos necessários para a elaboração de um trabalho ímpar, envolvendo novos desenvolvimentos associados a dados de contagem, com um enfoque particular na Teoria de Resposta ao Item e nos Modelos Lineares Generalizados Mistos. O trabalho está organizado da seguinte forma:

No Capítulo 2, exploramos o modelo de contagem Rasch Poisson, analisando as metodologias existentes para a estimativa de parâmetros e introduzindo uma nova abordagem por meio do método de aproximações de Laplace encaixadas e integradas, utilizando o pacote `INLA` em R. Além disso, desenvolvemos técnicas de análise de resíduos através de visualização gráfica e aplicamos o modelo a um conjunto de dados reais de um teste de atenção, demonstrando sua utilidade prática.

No Capítulo 3, abordamos modelos alternativos para respostas de contagem que superam algumas das limitações do modelo Rasch Poisson, tais como o modelo Binomial Negativa e sua variante inflacionada de zero, e o modelo Poisson Inflacionada de Zero. Para o ajuste desses modelos, exploramos tanto a abordagem bayesiana, utilizando o pacote `INLA`, quanto o

método clássico, por meio do pacote `gamlss`. Assim como no capítulo anterior, apresentamos direcionamentos para a análise de resíduos, destacando o uso de resíduos quantílicos aleatorizados para avaliação do ajuste dos modelos. Incluímos também uma aplicação prática ao conjunto de dados reais derivados de um teste de atenção, demonstrando a aplicabilidade do modelo proposto.

No Capítulo 4, introduzimos um novo modelo de regressão misto para respostas de contagem, baseado na distribuição Bell de um parâmetro. Este modelo representa uma alternativa interessante aos modelos mistos tradicionais para dados de contagem, especialmente com superdispersão. Consideramos abordagens clássica e bayesiana para a inferência, e conduzimos estudos de simulação para avaliar o desempenho do modelo proposto, bem como dos critérios de comparação de modelos. Além disso, demonstramos sua aplicabilidade prática em um conjunto de dados reais sobre a contagem de convulsões em pacientes epilépticos.

No Capítulo 5, assim como no capítulo anterior, introduzimos um novo modelo de regressão misto para variáveis de resposta de contagem, baseado na distribuição Bell-Touchard parametrizada pela média. Este modelo surge também como uma alternativa interessante aos modelos tradicionais de efeitos mistos para dados de contagem. Para o ajuste do modelo, exploramos o método clássico utilizando o algoritmo de Rigby e Stasinopoulos (RS) com o pacote `gamlss`. Realizamos estudos de simulação em diferentes cenários para avaliar o desempenho do modelo e dos critérios de seleção. Utilizamos também um conjunto de dados reais para ilustrar sua aplicabilidade prática.

Ressalta-se que os capítulos desta tese compreendem manuscritos desenvolvidos ao longo do doutorado. As discussões e conclusões do trabalho, resumindo as contribuições da tese no estado da arte dos modelos de contagem, os trabalhos realizados durante este período e as possibilidades para pesquisas futuras são apresentados no Capítulo 6.

CAPÍTULO

2

ANÁLISE DE RESÍDUOS EM MODELOS DE CONTAGEM RASCH POISSON

O trabalho apresentado neste capítulo desenvolve e discute a análise de resíduos para avaliar o ajuste de modelos alternativos ao Rasch Poisson. O modelo mais geral proposto é o modelo TRI Binomial Negativa zero-inflacionado (ZIBN), que inclui o modelo TRI Poisson como um caso particular. Para estimar os parâmetros dos modelos, propõe-se o uso da máxima verossimilhança penalizada por meio do algoritmo de Rigby e Stasinopoulos, bem como uma abordagem bayesiana com aproximações de Laplace encaixadas e integradas. O estudo propõe o uso de resíduos quantílicos aleatorizados com o propósito de avaliar o ajuste dos itens. Para ilustrar a metodologia, é utilizado um conjunto de dados de um teste de atenção seletiva, no qual 228 respondentes da 3ª e 4ª série tiveram que riscar dois dígitos em 20 blocos, cada um contendo três linhas com dígitos e letras arranjados aleatoriamente. O resultado dessa aplicação traz achados interessantes e demonstra o potencial da análise de resíduos utilizando gráficos de violino para auxiliar no processo de diagnóstico dos modelos, identificando o melhor modelo para os dados.

O conteúdo deste capítulo está publicado em Santos e Bazán (2021).

2.1 Introdução

No contexto da avaliação escolar, os modelos de Teoria de Resposta ao Item (TRI) são um conjunto de modelos probabilísticos onde as características latentes dos indivíduos que fazem um teste e as características latentes dos itens desse teste são consideradas para explicar as respostas obtidas (BAZÁN, 2018). Os modelos TRI mais conhecidos são aqueles em que a resposta é dicotômica, por exemplo, o chamado modelo TRI de três parâmetros utilizado no Exame Nacional do Ensino Médio (ENEM), um exame padrão não obrigatório que avalia alunos do ensino médio no Brasil.

Esse modelo considera três características dos itens: dificuldade, discriminação e acerto ao acaso. Estes são denominados parâmetros dos itens e precisam ser estimados junto com as características dos indivíduos, chamados de traços latentes. Casos particulares desse modelo são os chamados modelos de dois parâmetros (apenas os parâmetros de dificuldade e discriminação são considerados) e os modelos com um parâmetro (que considera apenas o parâmetro de dificuldade) (HAMBLETON; SWAMINATHAN, 2013). Em particular, o modelo de um parâmetro, também denominado modelo Rasch, foi originalmente formulado por Rasch (1960) e, considerando sua especificação aditiva, este pode ser visto como um Modelo Linear Generalizado Misto (MLGM), uma importante classe de modelos de regressão (WANG; YUE; FARAWAY, 2018). Portanto, métodos para ajustar MLGMs e diagnosticá-los também podem ser aplicados ao modelo Rasch (DOEBLER; HOLLING, 2016).

Atualmente, é comum observar respostas de contagem nas avaliações dos alunos. Por exemplo, considere uma avaliação em que a tarefa é identificar palavras escritas corretamente em uma longa lista de palavras. Portanto, as respostas aos itens correspondem às pontuações totais (contagens) ou ao número total de erros. Nestes casos, também é necessário desenvolver modelos TRI para respostas de contagem. Um modelo com essas características foi formulado por Rasch (1960), denominado modelo de contagem Rasch Poisson (do inglês, RPC). Embora esse modelo não seja novo, ele tem sido cada vez mais utilizado em avaliações recentes e modelos mais complexos tem sido formulados usando esse modelo como base (HUNG, 2012; FORTHMANN; GÜHNE; DOEBLER, 2019, ver).

O modelo RPC pode ser ajustado usando abordagem clássica (BAGHAEI; RAVAND; NADRI, 2019) e abordagem bayesiana (MUTZ; DANIEL, 2018). Recentemente, Baghaei e Doebler (2019) mostraram que o modelo RPC pode ser ajustado usando o pacote `lme4` (BATES *et al.*, 2015) no R, considerando este como um MLGM. Neste trabalho, desenvolvemos uma abordagem bayesiana para o modelo de contagem Rasch Poisson, utilizando uma formulação análoga à apresentada por Baghaei e Doebler (2019).

A análise de resíduos é uma ferramenta importante para avaliar o ajuste de um modelo a um determinado conjunto de dados. No caso dos modelos TRI, estamos interessados na análise de resíduos para os itens do teste. Assim, uma ferramenta de visualização gráfica dos resíduos

torna-se importante para avaliar se um item pode ser considerado seguindo o modelo proposto. No caso dos modelos de contagem Rasch, os resíduos de Pearson foram estudados com base nos valores obtidos pelo método de máxima verossimilhança e estão atualmente disponíveis para MLGMs no pacote `stats` no R, por meio da função `resid` (R Core Team, 2020). Nesse contexto, propomos a utilização do resíduo quantílico aleatorizado desenvolvido por Dunn e Smyth (1996), especificamente, desenvolvemos a análise de resíduos por meio de visualização gráfica considerando o gráfico de violino proposto por Hintze e Nelson (1998).

2.2 Modelo de contagem Rasch Poisson

Considere um teste de k itens aplicado a n indivíduos, em que as respostas obtidas correspondem a contagens como número de soluções corretas ou número de erros, entre outros. O modelo de contagem Rasch Poisson assume que as respostas de contagem Y_{ij} do indivíduo i no item j são independentes e Poisson distribuídas (RASCH, 1960) com:

$$Y_{ij} \sim Poisson(\mu_{ij})$$

$$P\left(Y_{ij} = y_{ij}\right) = \frac{\mu_{ij}^{y_{ij}} e^{-\mu_{ij}}}{y_{ij}!} \tag{2.1}$$

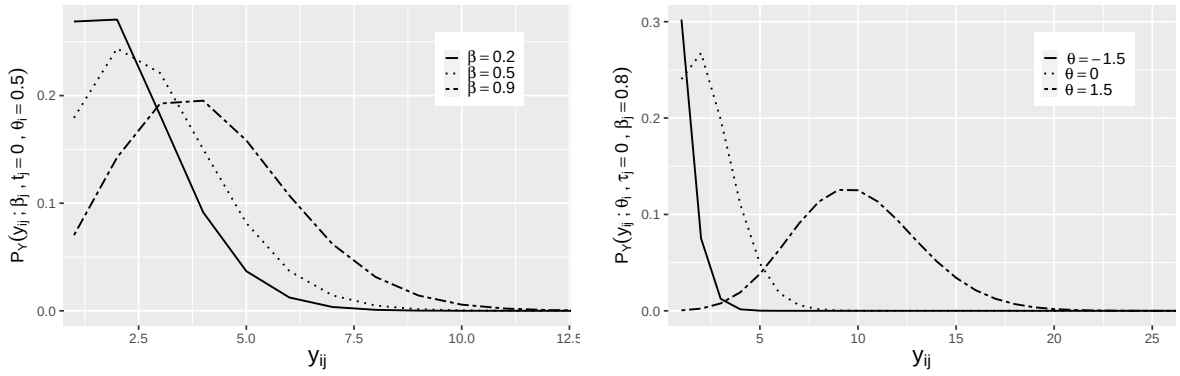
onde $E[Y_{ij}] = \mu_{ij} > 0$ é a contagem esperada do indivíduo i no item j, com $i = 1, \cdots, n$ e $j = 1, \cdots, k$. Em outras palavras, μ_{ij} é uma função do parâmetro associado ao traço latente ou habilidade do indivíduo i e a facilidade do item j. Além disso, assume-se que tem uma composição aditiva, usando a função de ligação logarítmica, expressa por:

$$\log\left(\mu_{ij}\right) = \beta_j + \theta_i + t_j \tag{2.2}$$

sendo: $\mu_{ij} = \exp\{\beta_j + \theta_i + t_j\}$ com $i = 1, \cdots, n$ e $j = 1, \cdots, k$, onde n é o número de indivíduos (tamanho da amostra), k é o número de itens, β_j é o parâmetro de facilidade do item j, θ_i é o traço latente do indivíduo i e t_j é o limite de tempo conhecido para o item j, correspondendo na expressão a uma variável *offset*, que é relevante ao modelar razões ou taxas quando os indivíduos não levam o mesmo tempo para responder a cada item, podendo ser fixado em zero no caso de não haver limite de tempo.

Na Figura 1 ilustramos a pontuação esperada para algumas combinações selecionadas dos parâmetros de facilidade do item e valores dos traços latentes, considerando um tempo ilimitado ($t_j = 0$). Em outras palavras, apresentamos as curvas da distribuição dos y_{ij} para diferentes valores de β_j e θ_i. Em particular, na Figura 1a, fixamos o valor do traço latente $\theta_i = 0.5$ e variamos os valores da facilidade do item $\beta_j = (0.2, 0.5, 0.9)$. Em contrapartida, na Figura 1b, fixamos o valor da facilidade do item $\beta_j = 0.8$ e variamos os valores do traço latente $\theta_i = (-1.5, 0, 1.5)$. Considerando a Figura 1a, observamos que dado um valor do traço latente, o valor esperado da variável resposta é maior se o item é mais fácil. De igual forma, na Figura 1b,

dado um valor da facilidade do item, o valor esperado da variável resposta é maior se o indivíduo possui uma maior habilidade.



(a) Traço latente é fixo e a facilidade do item varia. (b) Facilidade do item fixa e os traços latentes variam.

Figura 1 – Distribuição dos y_{ij} para diferentes combinações dos parâmetros do modelo RPC.

O modelo RPC assume a propriedade de independência condicional, ou seja, para o indivíduo i, as respostas y_{ij} correspondentes aos itens j são condicionalmente independentes dado os valores do traço latente do indivíduo, θ_i. Além disso, o modelo pressupõe independência entre as respostas de diferentes indivíduos. Assim, sejam $\beta = (\beta_1, \cdots, \beta_k)^\top$, $\theta = (\theta_1, \cdots, \theta_n)^\top$, $\mathbf{t} = (t_1, \cdots, t_k)^\top$, e $\mathbf{y} = (y_{ij})$ a matriz de respostas observadas, considerando as suposições do modelo e a Equação 2.1, a função de verossimilhança é dada por:

$$L(\beta, \theta \mid \mathbf{y}, \mathbf{t}) = \prod_{i=}^{n} \prod_{j=1}^{k} \frac{\mu_{ij}^{y_{ij}} e^{-\mu_{ij}}}{y_{ij}!} \qquad \text{com} \quad \mu_{ij} = e^{\beta_j + \theta_i + t_j}. \tag{2.3}$$

Existem vários métodos para estimar os parâmetros do modelo RPC (VERHELST; KAMPHUIS, 2009), tais como: máxima verossimilhança condicional (CML), máxima verossimilhança conjunta (JML), máxima verossimilhança marginal (MML), entre outros (BAGHAEI; DOEBLER, 2019). Alguns autores, como Jansen e Duijn (1992), impõem a restrição $\sum_{j=1}^{k} \beta_j t_j = 1$, que é adicionada para identificação do modelo quando usado o procedimento de estimação de máxima verossimilhança marginal. Além disso, Jansen (1994) propôs o uso de uma distribuição para traços latentes dentro de um algoritmo *Expectation-Maximization* (EM), $\theta_i \sim N\left(0, \sigma^2\right)$. Usando esta especificação e a formulação de efeitos aditivos (Equação 2.2), o modelo RPC pode ser visto como um MLGM, no qual consideramos θ_i como o efeito aleatório do indivíduo, β_j o efeito fixo associado aos itens e t_j um *offset*, ou seja, uma constante conhecida adicionada à equação de regressão.

2.3 Inferência Bayesiana e Análise de Resíduos

Para obter as estimativas dos parâmetros do modelo RPC, consideramos a abordagem bayesiana e propomos prioris para θ_i e β_j a fim de obter a distribuição a posteriori desses

parâmetros. Nesse contexto, propomos o uso do método de aproximações de Laplace encaixadas e integradas (INLA - *Integrated Nested Laplace Approximations*) desenvolvido por Rue, Martino e Chopin (2009). O método INLA é uma abordagem determinística para inferência bayesiana em uma ampla estrutura de modelos gaussianos latentes, incluindo MLGM (RUE *et al.*, 2017), em que a variável resposta Y_{ij}, com média μ_{ij}, é ligada à estrutura aditiva do preditor linear η_{ij} por meio de uma função de ligação $g(\cdot)$, tal que $g(\mu_{ij}) = \eta_{ij}$.

A estrutura aditiva do modelo RPC dada pela Equação 2.2, como já mencionado, pode ser considerada como um MLGM e, portanto, pode ser escrita usando uma estrutura hierárquica:

$$\begin{aligned} Y_{ij} \mid \theta_i, \beta_j, t_j &\sim Poisson(\mu_{ij}) \\ log(\mu_{ij}) &= \eta_{ij} \\ \eta_{ij} &= \beta_j + \theta_i + t_j \\ i = 1, \cdots, n \quad &\text{e} \quad j = 1, \cdots, k \\ \theta_i \mid \sigma_\theta^2 &\sim N\left(0, \sigma_\theta^2\right) \\ \sigma_\theta^{-2} &\sim Gama\left(1, 10^{-5}\right) \\ \beta_j &\sim N\left(0, 1000\right), \end{aligned} \tag{2.4}$$

onde $\sigma_\theta^{-2} = \tau_\theta$ é o parâmetro de precisão.

Os resíduos carregam informações importantes para verificar as suposições que fundamentam os modelos estatísticos e, portanto, desempenham um papel importante na análise de dados. O uso dos resíduos permite detectar discrepâncias de algumas observações específicas do modelo, além de fornecer uma visão geral em termos de qualidade do ajuste. Para o modelo Poisson, os resíduos de Pearson são comumente utilizados (BAGHAEI; DOEBLER, 2019) e definidos como

$$r_{ij} = \left(y_{ij} - \widehat{\mu}_{ij}\right) \widehat{\mu}_{ij}^{-1/2}, \tag{2.5}$$

onde $\widehat{\mu}_{ij} = \widehat{E[Y_{ij}]}$ é a média a posteriori de Y_{ij}, obtida usando $\widehat{\mu}_{ij} = \exp\{\widehat{\beta}_j + \widehat{\theta}_i + t_j\}$ com $\widehat{\theta}_i$ e $\widehat{\beta}_j$ sendo as médias a posteriori do traço latente e da facilidade do item, respectivamente.

Neste trabalho, propomos o uso dos resíduos quantílicos aleatorizados de Dunn e Smyth (1996), definidos por

$$q_{ij} = \Phi^{-1}\left(F(y_{ij} - 1; \hat{\mu}_{ij}) + u_i . f(y_{ij}; \hat{\mu}_{ij})\right), \tag{2.6}$$

onde $f(\cdot)$ e $F(\cdot)$ representam a função massa de probabilidade e função de distribuição acumulada da distribuição Poisson, respectivamente, e u_i é um valor da distribuição uniforme no intervalo $(0, 1)$.

Em um modelo bem especificado, espera-se que os resíduos se concentrem em torno de zero, cobrindo uniformemente uma faixa de aproximadamente -1.96 a 1.96, considerando um

nível de confiança de 95%. Para verificação do ajuste de um determinado item, apresentamos a distribuição dos resíduos dos diferentes indivíduos pra esse item usando métodos gráficos: *boxplot* e gráfico de violino.

O gráfico de violino, proposto por Hintze e Nelson (1998), combina o gráfico de *boxplot* e a estimação da densidade em um único gráfico. Em outras palavras, adiciona as informações disponíveis das estimativas de densidade às estatísticas básicas resumidas inerentes a um *boxplot*. Dessa forma, essa combinação do formato da densidade e estatísticas resumidas em único gráfico fornece uma ferramenta útil para ilustrar a adequação do modelo, detectar especificações incorretas da distribuição de erros, bem como identificar o comportamento da distribuição dos erros e potenciais itens com ajuste problemático.

2.4 Aplicação

Ilustramos a abordagem bayesiana do modelo RPC considerando os dados apresentados por Baghaei e Doebler (2019), referente a um estudo de 228 examinados para um teste de atenção seletiva proposto por Beyzaee (2017). O teste consiste em 20 blocos com um limite de tempo de 15 segundos para realização da tarefa em cada bloco, em que os participantes precisam riscar os números 2 e 7 em três linhas de dígitos e letras dispostos aleatoriamente. Um exemplo de bloco do teste é mostrado na Figura 2.

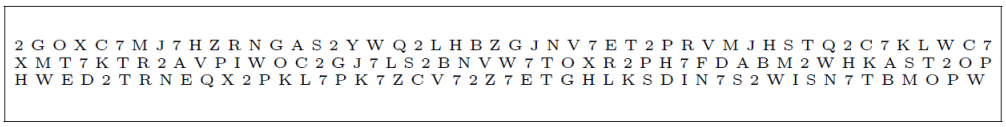


Figura 2 – Exemplo de um bloco do teste de atenção seletiva.

Fonte: Baghaei e Doebler (2019).

O conjunto de dados em análise possui as seguintes variáveis: “ID”, que corresponde ao número de identificação do aluno; “Item”, que se refere ao bloco de letras e números que os alunos devem verificar; “Hit”, representando o número total de verificações corretas em cada item por examinado; e “TL”, indicando o tempo, em segundos, utilizado para responder cada item. Este conjunto de dados pode ser solicitado diretamente aos autores do artigo conforme mencionado em Baghaei e Doebler (2019).

Conforme descrito por Baghaei e Doebler (2019), cada bloco é considerado um item, e o número total de verificações corretas de 2’s e 7’s, registrado na variável “Hit”, é modelado como a unidade de análise. Dessa forma, ajustamos o modelo RPC conforme definido na Equação 2.4, utilizando a abordagem bayesiana (Seção 2.3). A Tabela 1 apresenta as estimativas a posteriori para o parâmetro de facilidade do modelo RPC, incluindo a média (MD), desvio padrão (DP) e intervalo de credibilidade de 95% (Q2.5, Q97.5) para cada item avaliado. Podemos observar

que as médias das estimativas de facilidade (β) para os itens variam de 0.157 a 0.648. Itens com maior média de facilidade são considerados mais fáceis, enquanto aqueles com menor média são mais difíceis. Portanto, o item β_{12} é considerado o mais fácil, com uma média de 0.648, enquanto o item β_2 é o mais difícil, com uma média de 0.157. O desvio padrão das estimativas é relativamente pequeno para todos os itens, variando de 0.017 a 0.019, indicando uma consistência nas estimativas de facilidade dos itens. A precisão das estimativas é ainda reforçada pelos intervalos de credibilidade de 95%, que são estreitos para a maioria dos itens.

Tabela 1 – Estimativas: Média (MD), Desvio padrão (DP) e intervalo de credibilidade de 95% (Q2.5, Q97.5) para o parâmetro de facilidade do modelo RPC.

Parâmetro	MD	DP	Q2.5	Q97.5	Parâmetro	MD	DP	Q2.5	Q97.5
β_1	0.265	0.018	0.230	0.300	β_{11}	0.552	0.018	0.516	0.587
β_2	0.157	0.019	0.120	0.193	β_{12}	0.648	0.018	0.612	0.683
β_3	0.427	0.018	0.391	0.463	β_{13}	0.414	0.018	0.378	0.450
β_4	0.579	0.019	0.542	0.615	β_{14}	0.632	0.018	0.596	0.668
β_5	0.378	0.019	0.342	0.415	β_{15}	0.263	0.018	0.228	0.298
β_6	0.593	0.018	0.557	0.629	β_{16}	0.540	0.018	0.504	0.575
β_7	0.455	0.019	0.418	0.492	β_{17}	0.295	0.018	0.259	0.331
β_8	0.553	0.018	0.518	0.589	β_{18}	0.612	0.018	0.576	0.648
β_9	0.237	0.018	0.201	0.273	β_{19}	0.342	0.018	0.306	0.378
β_{10}	0.341	0.017	0.306	0.375	β_{20}	0.263	0.018	0.227	0.299

Em resumo, os resultados apresentados na Tabela 1 revelam uma ampla diversidade de dificuldades entre os itens do teste. Itens como β_2 e β_9 apresentam estimativas mais baixas, indicando maior dificuldade, enquanto itens como β_{12} e β_{14} são significativamente mais fáceis, conforme mostrado na Figura 3. Os resultados obtidos aqui são muito próximos aos alcançados utilizando a abordagem clássica através do pacote `lme4`.

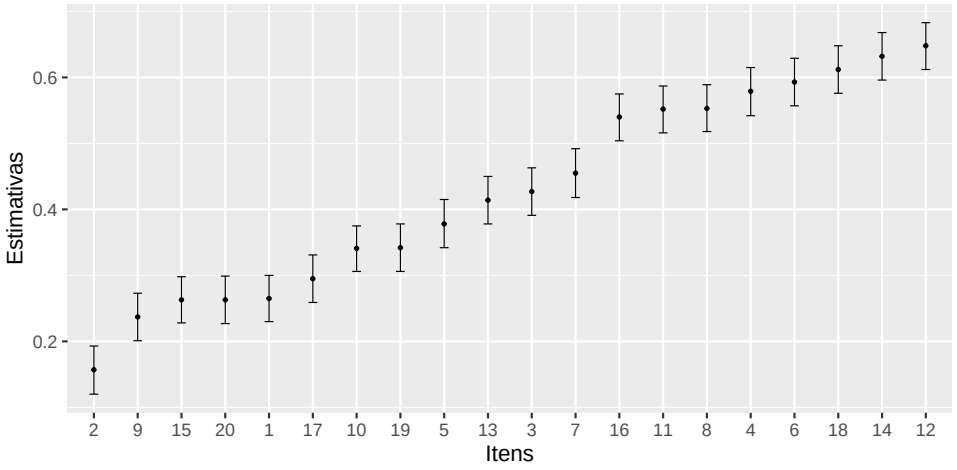


Figura 3 – Médias a posteriori e intervalos de credibilidade de 95% dos parâmetros de itens.

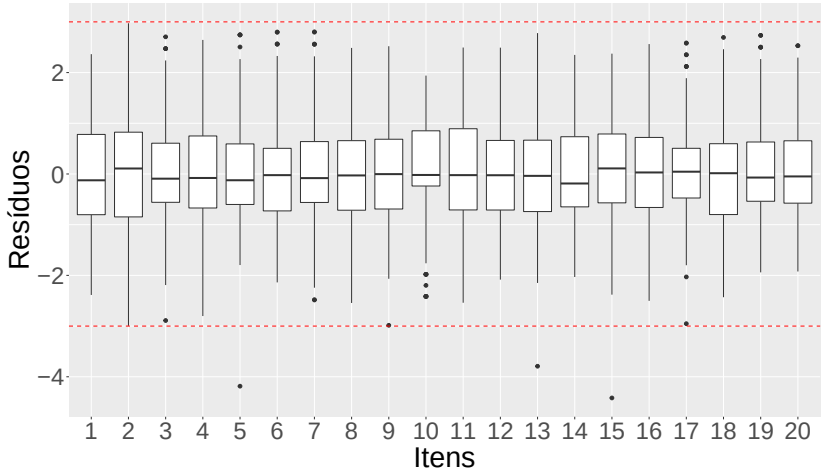
Para avaliar a qualidade do ajuste do modelo, conduzimos uma análise de resíduos utilizando os resíduos quantílicos aleatorizados e de Pearson. A Tabela 2 apresenta as medidas descritivas a posteriori de ambos os tipos de resíduos. Considerando esses resultados, podemos observar que os resíduos de Pearson possuem uma variabilidade em torno de 1, com valores mínimos e máximos variando consideravelmente. Isso indica que, embora a maioria dos itens se ajuste bem ao modelo, há exceções notáveis com grandes desvios. Em particular, os itens 5, 13, e 15 destacam-se com valores de resíduos mínimos muito baixos, denotando potenciais pontos discrepantes, conforme mostrado na Figura 4a.

Por outro lado, os resíduos quantílicos aleatorizados exibem um padrão similar, com desvio padrão próximo a 1 e uma variação significativa nos valores mínimos e máximos. Além disso, conforme mostrado na Figura 4b, um número maior de itens é identificado como problemáticos, incluindo os itens 2, 3, 5, 9, 13, 15 e 17. A presença de múltiplos *outliers*, conforme identificados pelos resíduos quantílicos, reforça a necessidade de uma análise mais detalhada desses itens específicos para entender as razões subjacentes às discrepâncias observadas.

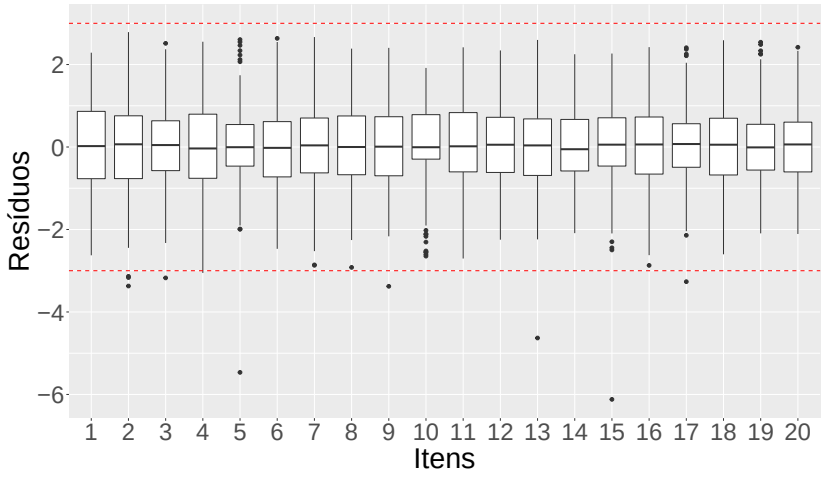
Tabela 2 – Medidas descritivas: Média (MD), Desvio padrão (DP), Mínimo (Min) e Máximo (Max) dos resíduos para o modelo RPC aplicado aos dados do teste de atenção.

Itens	Resíduos de Pearson				Resíduos Quantílicos			
	MD	DP	Min	Max	MD	DP	Min	Max
1	0.048	1.156	-2.388	2.359	0.044	1.154	-2.626	2.287
2	0.045	1.169	-2.997	2.968	0.028	1.173	**-3.371**	2.786
3	0.046	0.964	-2.892	2.701	0.049	0.963	**-3.174**	2.514
4	0.046	1.066	-2.804	2.640	0.044	1.054	**-3.053**	2.553
5	0.045	0.932	**-4.187**	2.739	0.055	0.950	**-5.464**	2.605
6	0.046	1.069	-2.141	2.794	0.045	1.062	-2.469	2.634
7	0.045	1.013	-2.485	2.796	0.048	1.011	-2.867	2.666
8	0.047	1.073	-2.546	2.482	0.038	1.066	-2.919	2.386
9	0.047	1.056	-2.986	2.516	0.047	1.046	**-3.377**	2.404
10	0.050	0.914	-2.417	1.936	0.065	0.926	-2.646	1.918
11	0.047	1.026	-2.542	2.490	0.050	1.029	-2.704	2.417
12	0.047	1.019	-2.086	2.488	0.045	1.016	-2.247	2.342
13	0.046	1.040	**-3.794**	2.775	0.044	1.036	**-4.631**	2.596
14	0.047	0.923	-2.034	2.344	0.052	0.919	-2.088	2.249
15	0.048	0.979	**-4.419**	2.369	0.045	1.016	**-6.119**	2.267
16	0.047	1.009	-2.505	2.558	0.055	1.002	-2.870	2.422
17	0.047	0.922	-2.955	2.578	0.056	0.926	**-3.265**	2.407
18	0.046	1.001	-2.433	2.689	0.044	1.001	-2.602	2.589
19	0.046	0.891	-1.943	2.730	0.062	0.871	-2.094	2.541
20	0.046	0.983	-1.926	2.526	0.049	0.977	-2.108	2.419

Nota: Os valores em negrito denotam pontos discrepantes.



(a) Resíduos de Pearson.



(b) Resíduos Quantílicos.

Figura 4 – Boxplot dos resíduos para o modelo RPC aplicado aos dados do teste de atenção.

Para esclarecer a distribuição dos *outliers* detectados usando os resíduos quantílicos aleatorizados, apresentamos a distribuição dos resíduos desses itens por meio do gráfico de violino (Figura 5). Optamos por relatar os resultados dos resíduos quantílicos, uma vez que se espera um comportamento de distribuição normal desses resíduos, ao contrário dos resíduos de Pearson, que apresentam apenas comportamento normal assintótico. Embora seja possível demonstrar esse comportamento utilizando o *QQ-plot*, escolhemos o gráfico de violino por permitir a comparação de outras características entre diferentes itens.

O gráfico de violino permite observar tanto a amplitude quanto a distribuição dos resíduos para cada item. A análise revela várias informações importantes sobre a qualidade do ajuste do modelo RPC aos dados. A forma dos gráficos de violino revela a distribuição dos resíduos. Idealmente, esperamos que os resíduos sigam uma distribuição simétrica em torno de zero para um bom ajuste do modelo. No entanto, os itens 2, 4 e 13 mostram desvios significativos da normalidade, indicando que o modelo não captura bem os dados para esses itens. Além disso,

itens como 5 e 15 exibem caudas longas e finas, sugerindo a presença de *outliers*. Esses *outliers* são pontos onde os resíduos se afastam significativamente do restante dos dados, indicando possíveis anomalias ou problemas específicos com esses itens no modelo.

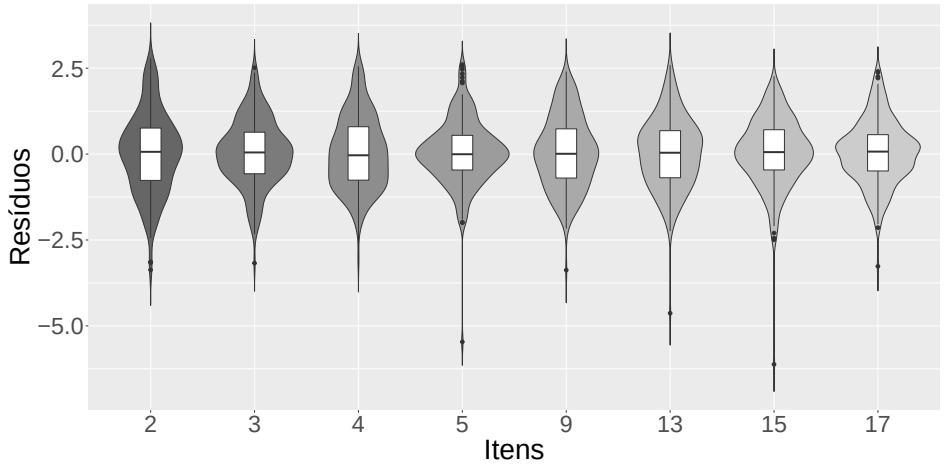


Figura 5 – Gráficos de violino dos resíduos quantílicos para os itens problemáticos.

Comparando todos os itens apresentados no gráfico, itens como 9 e 17 exibem distribuições mais simétricas e concentradas em torno de zero, sugerindo um melhor ajuste do modelo para esses itens em comparação com os demais itens. Esses resultados indicam que, embora o modelo RPC estime bem os parâmetros da maioria dos itens, há itens específicos onde o ajuste é inadequado. A análise detalhada dos resíduos é crucial para identificar esses problemas e melhorar o modelo. Em particular, itens com grande dispersão dos resíduos ou com presença de *outliers* devem ser examinados mais detalhadamente para entender as razões subjacentes às discrepâncias e ajustar o modelo de forma a capturar melhor as características dos dados. A identificação de itens problemáticos destaca a importância da análise de resíduos como uma ferramenta diagnóstica essencial em modelos estatísticos.

Em TRI, a normalidade dos traços latentes é crucial, pois a suposição de normalidade é frequentemente utilizada para modelar a habilidade dos indivíduos. A Figura 6 corrobora essa suposição, reforçando a validade do modelo RPC aplicado, demonstrando que é aproximadamente normal em torno de zero. Além disso, o hiperparâmetro associado à dispersão do parâmetro do traço latente apresenta uma média a posteriori de $\hat{\mu}_\theta = 44.310$ e um desvio padrão a posteriori de 4.582. Esses valores indicam uma pequena variância para o traço latente, o que significa que a maioria dos traços latentes dos indivíduos está concentrada em torno de um valor médio com pouca dispersão.

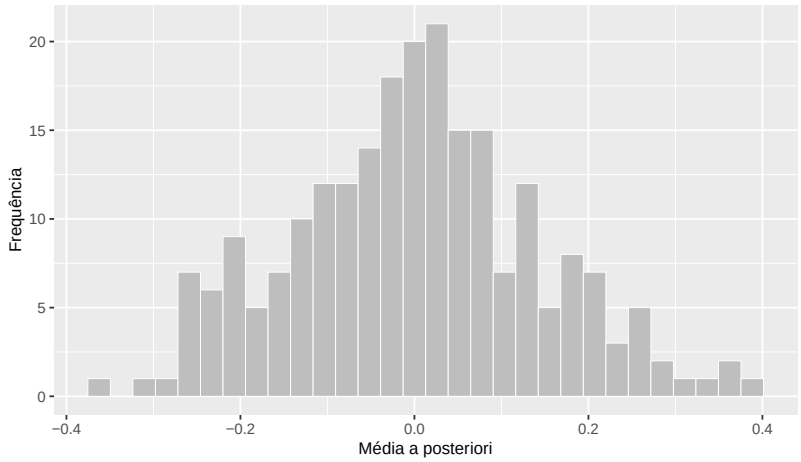


Figura 6 – Gráfico de barras das médias a posteriori do traço latente.

2.5 Comentários finais

Neste capítulo, apresentamos a abordagem bayesiana para estimar os parâmetros do modelo RPC, utilizando o método INLA. Este método é uma alternativa aos métodos bayesianos comumente usados na literatura estatística, embora seja menos frequente na literatura psicométrica. Utilizamos a especificação do modelo RPC como um modelo de regressão Poisson misto. Os resultados, embora não apresentados aqui, indicaram que nossas estimativas foram muito próximas das obtidas utilizando a função `glmer`, que aplica a abordagem dos MLGM com o método de máxima verossimilhança marginal, conforme proposto em Baghaei e Doebler (2019).

Considerando nossa formulação, extensões para modelos Rasch, propostas por Boeck e Wilson (2004), podem ser exploradas e facilmente implementadas a partir de uma abordagem bayesiana usando o método INLA. Ilustramos o método de estimativa com dados reais do teste de atenção apresentados por Baghaei e Doebler (2019). Além disso, introduzimos o uso dos resíduos quantílicos aleatorizados para avaliar o ajuste de cada item do teste. Demonstramos que os resíduos quantílicos e de Pearson fornecem informações discrepantes sobre o número de itens que não estão bem ajustados ao modelo RPC. Mostramos que os gráficos de violino são ferramentas eficazes para verificar o ajuste dos itens de teste. Com base no método proposto de análise de resíduos nos dados analisados, nossos resultados sugerem que, embora o modelo RPC forneça um ajuste geral robusto, há itens que requerem atenção especial para garantir a precisão e a validade do modelo. A identificação e análise desses *outliers* são cruciais para refinar o modelo e melhorar a qualidade do ajuste geral. Portanto, outros modelos, como os de Wang (2010), Hung (2012) e Forthmann, Gühne e Doebler (2019), podem ser estudados futuramente com este conjunto de dados. A aplicação desses modelos pode proporcionar *insights* adicionais e melhorar a precisão das inferências realizadas.

CAPÍTULO

3

ANÁLISE DE RESÍDUOS EM MODELOS DE CONTAGEM RASCH

No capítulo anterior, foi verificado que o modelo de contagem Rasch Poisson (RPC) apresenta resíduos maiores em alguns itens e, consequentemente, não é o melhor modelo para o conjunto de dados considerado. Este capítulo traz extensões do modelo RPC, considerando modelos alternativos à Poisson para a resposta observada. Inicialmente, são consideradas as contagens Binomiais Negativas (NBI) e, posteriormente, estende-se este modelo para incluir o excesso de zeros, utilizando os modelos de contagem Poisson Inflacionado de Zeros (ZIP) e Binomial Negativo Inflacionado de Zeros (ZINBI). Enquanto os modelos TRI-NBI e TRI-ZIP foram introduzidos por Wang (2010) e Hung (2012), respectivamente, o modelo TRI-ZINBI, proposto neste estudo, é uma nova contribuição para a literatura psicométrica e inclui os dois modelos anteriores como casos particulares. Para os modelos propostos, são considerados os métodos de estimação clássico e bayesiano, seguindo a formulação do modelo RPC como MLGM apresentada por Baghaei e Doebler (2019), e para avaliar o ajuste dos itens, são utilizados os resíduos quantílicos aleatorizados. A metodologia é ilustrada com dados de um teste de atenção e verifica-se que o modelo TRI-ZIP é o melhor entre os apresentados, sendo capaz de explicar adequadamente os dados.

O conteúdo deste capítulo está publicado em Santos e Bazán (2021).

3.1 Modelos de contagem Rasch

O modelo RPC, proposto por Rasch (1960) assume que as respostas de contagem Y_{ij} do indivíduo i no item j de um teste são condicionalmente independentes do traço latente θ_i e Poisson distribuídas. Além disso, assume-se que tem uma composição aditiva, usando a função de ligação logarítmica, expressa por:

$$\begin{aligned} Y_{ij} &\sim Pois(\mu_{ij}) \\ \log\left(\mu_{ij}\right) &= \beta_j + \theta_i + t_j, \end{aligned} \tag{3.1}$$

com $\theta_i \sim N\left(0, \sigma^2\right)$, $i = 1, \cdots, n$ e $j = 1, \cdots, k$, onde μ_{ij} é a contagem esperada para o indivíduo i no item j, θ_i é a habilidade do indivíduo i (traço latente), β_j é a facilidade do item j, e t_j é o limite de tempo para o item j (variável *offset*). Aqui, $Pois(\mu_{ij})$ denota a função massa de probabilidade da distribuição Poisson com parâmetro $\mu_{ij} > 0$, dada por:

$$P(y_{ij}; \mu_{ij}) = \frac{\mu_{ij}^{y_{ij}} e^{-\mu_{ij}}}{y_{ij}!} \tag{3.2}$$

onde a média e variância são dadas por $E[Y_{ij}] = \mu_{ij}$ e $Var[Y_{ij}] = \mu_{ij}$, respectivamente.

3.1.1 Distribuições de contagem alternativas

A distribuição Poisson possui a característica de equidispersão (média igual a variância) e pode não modelar adequadamente a sub ou superdispersão presente nos dados. Assim, para uma maior flexibilidade na relação entre a média e variância, as distribuições Binomial negativa, Poisson inflacionada de zeros e Binomial negativa inflacionada de zeros são propostas como alternativas para modelar dados de contagem com essas características de dispersão.

- *Distribuição Binomial negativa (NBI)*: a função massa de probabilidade da distribuição Binomial Negativa, denotada por $NBI(\mu, \phi)$, é dada por

$$P_Y(y; \mu, \phi) = \frac{\Gamma\left(y + \frac{1}{\phi}\right)}{\Gamma\left(\frac{1}{\phi}\right)\Gamma(y+1)} \left(\frac{\phi\mu}{1+\phi\mu}\right)^y \left(\frac{1}{1+\phi\mu}\right)^{1/\phi}, \tag{3.3}$$

  onde $\mu > 0$, $\phi > 0$ é o parâmetro de dispersão e $\Gamma(\cdot)$ é a função gamma. A média e a variância da NBI são $E[Y] = \mu$ e $Var[Y] = \mu + \phi\mu^2$, respectivamente.

- *Distribuição Poisson inflacionada de zeros (ZIP)*: Seja $Y = 0$ com probabilidade ω e $Y \sim Pois(\mu)$ com probabilidade $(1-\omega)$. Então dizemos que Y tem distribuição Poisson inflacionada de zeros, denotada por $ZIP(\mu,\omega)$, se sua função de probabilidade é dada por

$$P_Y(y;\mu,\omega) = \begin{cases} \omega + (1-\omega)e^{-\mu}, & y = 0 \\ (1-\omega)\dfrac{\mu^y e^{-\mu}}{y!}, & y = 1,2,\cdots \end{cases}, \tag{3.4}$$

  onde $\mu > 0$ e $0 < \omega < 1$ é a probabilidade de zeros. A média e a variância de uma variável aleatória ZIP são dadas por $E[Y] = (1-\omega)\mu$ e $Var[Y] = \mu(1-\omega)(1+\mu\omega)$, respectivamente.

- *Distribuição Binomial Negativa inflacionada de zeros (ZINBI)*: Seja $Y = 0$ com probabilidade ω e $Y \sim NBI(\mu,\phi)$ com probabilidade $(1-\omega)$. Então dizemos que Y tem distribuição Binomial negativa inflacionada de zeros, denotada por $ZINBI(\mu,\phi,\omega)$, com função de probabilidade dada por

$$P_Y(y;\mu,\phi,\omega) = \begin{cases} \omega + (1-\omega)(1+\phi\omega)^{-1/\phi}, & y = 0 \\ (1-\omega)\dfrac{\Gamma\left(y+\frac{1}{\phi}\right)}{\Gamma\left(\frac{1}{\phi}\right)\Gamma(y+1)}\left(\dfrac{\phi\mu}{1+\phi\mu}\right)^y\left(\dfrac{1}{1+\phi\mu}\right)^{1/\phi}, & y = 1,2,\cdots \end{cases}, \tag{3.5}$$

  onde $\mu > 0$, $\phi > 0$ é o parâmetro de dispersão, $0 < \omega < 1$ é a probabilidade de zeros e $\Gamma(\cdot)$ é a função gamma. A média e a variância da ZINBI são $E[Y] = (1-\omega)\mu$ e $Var[Y] = (1-\omega)[1+(\phi+\omega)\mu]$, respectivamente.

Em particular, a mais geral dessas distribuições é a ZINBI, enquanto as outras distribuições são casos particulares. Mais detalhes sobre esses modelos podem ser vistos em Hung (2012), Wang (2010), Magnus e Thissen (2017), Gonzalez (2018).

3.1.2 Modelo de contagem Rasch alternativo

Como alternativa ao modelo RPC, temos que as respostas de contagem Y_{ij} seguem uma distribuição NBI, ZIP ou ZINBI. Para essas distribuições, consideramos a composição aditiva do modelo Rasch usando a função de ligação logarítmica (Equação 3.1). A seguir, apresentamos apenas o modelo de contagem ZINBI, uma vez que os outros modelos são casos particulares desse modelo.

$$\begin{aligned} Y_{ij} &\sim ZINBI(\mu_{ij},\phi,\omega) \\ \log(\mu_{ij}) &= \beta_j + \theta_i + t_j, \end{aligned} \tag{3.6}$$

com $\theta_i \sim N\left(0,\sigma^2\right)$, $i = 1,\cdots,n$, $j = 1,\cdots,k$, onde θ_i é a habilidade do indivíduo i, β_j é a facilidade do item j, t_j é uma constante conhecida para o item j, ϕ é o parâmetro de dispersão ω é a probabilidade de zeros.

3.2 Métodos de estimação

A estimação dos parâmetros dos modelos propostos pode ser realizada considerando a formulação equivalente desses modelos como um MLGM. Em outras palavras, os modelos de contagem Rasch, por meio de sua especificação aditiva (Equação 3.1), podem ser visto como um MLGM considerando o traço latente θ_i como um efeito aleatório do indivíduo, β_j como um efeito fixo associado aos itens e t_j como uma variável *offset*, ou seja, uma constante conhecida adicionada à equação de regressão. Considerando esta formulação, propomos o uso de dois métodos de estimação. Primeiramente, sob abordagem clássica, consideramos a verossimilhança marginal penalizada (PML), usando o algoritmo de Rigby e Stasinopoulos (RS) por meio do pacote `GAMLSS` (RIGBY *et al.*, 2019). Considerando a abordagem bayesiana, utilizamos o método de aproximações de Laplace encaixadas e integradas, método INLA, por meio do pacote `INLA` no R (WANG; YUE; FARAWAY, 2018).

Para comparar os modelos alternativos ao modelo RPC, sob as abordagens consideradas, fazemos o uso de alguns critérios de comparação de modelos. Especificamente, na abordagem clássica, consideramos os critérios AIC (*Akaike information criterion*) de Akaike (1983) e SBC (*Schwartz Bayesian criterion*) de Schwarz (1978) definidos, respectivamente, por

$$AIC = GD + 2 \times df,$$

$$SBC = GD + \log(n) \times df,$$

onde df denota os graus de liberdade efetivos totais usados no modelo e $GD = -2\ell(\hat{\theta})$ é o deviance global. Rigby *et al.* (2019) comentam que, na prática, o SBC é muito mais restritivo que AIC. Sob abordagem bayesiana, consideramos os critérios DIC (*Deviance Information Criterion*) de Spiegelhalter *et al.* (2002) e WAIC (*Watanabe-Akaike information criterion*) de Watanabe e Opper (2010) definidos como

$$DIC = \bar{D} + p_D,$$

$$WAIC = -2lppd + 2p_D,$$

respectivamente, onde p_D é o número de parâmetros efetivos no modelo, $\bar{D}$ é a média a posteriori do deviance do modelo e $lppd$ é o logaritmo da densidade preditiva pontual.

Para todos os critérios mencionados, valores menores indicam melhor ajuste.

3.3 Análise de resíduos

A análise de resíduos é uma ferramenta importante para avaliar o ajuste de um modelo a um determinado conjunto de dados, onde é possível identificar possíveis *outliers*. Dentre os resíduos existentes descritos na literatura, consideramos os resíduos de Pearson, definidos como (CORDEIRO; SIMAS, 2009):

$$r_{ij} = \frac{y_{ij} - \widehat{E[Y_{ij}]}}{\sqrt{\widehat{V[Y_{ij}]}}} \tag{3.7}$$

com $r_{ij} \approx N(0,1)$, onde $\widehat{E[Y_{ij}]}$ e $\widehat{V[Y_{ij}]}$ são as estimativas da média e variância de Y_{ij}, respectivamente, considerando a distribuição de contagem adotada.

Feng, Sadeghpour e Li (2017), usando estudos de simulação, compararam os resíduos quantílicos com os resíduos de Pearson e concluíram que a distribuição dos resíduos quantílicos é melhor aproximada pela distribuição normal padrão do que os resíduos de Pearson. Além disso, os autores mostraram que os resíduos quantílicos são melhores para detectar falta de ajuste. Diante disso, consideramos os resíduos quantílicos aleatorizados propostos por Dunn e Smyth (1996), definidos por

$$q_{ij} = \Phi^{-1}(U_i) \tag{3.8}$$

com $q_{ij} \sim N(0,1)$, onde $\Phi(\cdot)$ é a função de distribuição acumulada da normal padrão e U_i é uma variável aleatória uniforme no intervalo $(a_i, b_i]$ com $a_i = F(y_i^-; \widehat{\eta})$ e $b_i = F(y_i; \widehat{\eta})$, em que $F(\cdot)$ é a função distribuição acumulada do modelo de contagem correspondente considerado e $\widehat{\eta}$ é o vetor de parâmetros estimados.

A Tabela 3 mostra como os resíduos de Pearson e quantílicos aleatorizados são calculados para os diferentes modelos de contagem Rasch, onde *dpois*, *dnbinom*, *dzip* e *dzinb* denotam a função massa de probabilidade para as distribuições Pois, NBI, ZIP e ZINBI respectivamente; e a função distribuição acumulada por *ppois*, *pnbinom*, *pzip* e *pzinb* respectivamente. Adicionalmente, desenvolvemos funções R genéricas que podem calcular os resíduos quantílicos e de Pearson para os diferentes modelos de contagem adotados. Essas funções foram implementadas com base nas saídas dos ajustes dos pacotes `gamlss` e `INLA`.

Tabela 3 – Resíduos de Pearson e quantílicos aleatorizados de diferentes modelos.

Modelos	Resíduos de Pearson	Resíduos Quantílicos
Pois	$\frac{y_i-\hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$	$\Phi^{-1}\left(ppois(y_i^-;\hat{\mu}_{ij})+u_i.dpois(y_i;\hat{\mu}_{ij})\right)$
NBI	$\frac{y_i-\hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}+\hat{\phi}\hat{\mu}_{ij}^2}}$	$\Phi^{-1}\left(pnbinom(y_i^-;\hat{\mu}_{ij},\hat{\phi})+u_i.dnbinom(y_i;\hat{\mu}_{ij},\hat{\phi})\right)$
ZIP	$\frac{y_i-(1-\hat{\omega})\hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-\hat{\omega})(1+\hat{\mu}_{ij}\hat{\omega})}}$	$\Phi^{-1}\left(pzip(y_i^-;\hat{\mu}_{ij},\hat{\omega})+u_i.dzip(y_i;\hat{\mu}_{ij},\hat{\omega})\right)$
ZINBI	$\frac{y_i-(1-\hat{\omega})\hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-\hat{\omega})\left[1+(\hat{\phi}+\hat{\omega})\hat{\mu}_{ij}\right]}}$	$\Phi^{-1}\left(pzinb(y_i^-;\hat{\mu}_{ij},\hat{\phi},\hat{\omega})+u_i.dzinb(y_i;\hat{\mu}_{ij},\hat{\phi},\hat{\omega})\right)$

No caso dos modelos TRI, estamos interessados na análise de resíduos para testar os itens do teste. Portanto, consideramos os resíduos de Pearson e quantílicos para estimar a distribuição dos resíduos para cada item j, denotados por r_{ij} e q_{ij} respectivamente. Então, para cada item j, temos um vetor com n valores dos resíduos de Pearson para os indivíduos, $r_{1j},\cdots,r_{nj}$. Da mesma forma, temos um vetor de n valores dos resíduos quantílicos para cada item j, $q_{1j},\cdots,q_{nj}$. Assim, estatísticas resumidas para esses resíduos, como média, desvio padrão, mínimo e máximo, podem ser relatadas.

Para verificar o ajuste do modelo para um determinado item, propomos o uso do gráfico de violino (HINTZE; NELSON, 1998) dos resíduos considerados. Recomendamos esses gráficos, pois eles exibem a distribuição dos resíduos juntamente com informações sobre as estatísticas resumos e o comportamento da densidade, fornecendo uma ferramenta útil para análise de resíduos. No programa R, esses gráficos podem ser obtidos usando o pacote `ggplot2` (WICKHAM, 2016).

3.4 Aplicação

Ilustramos a análise de resíduos para os modelos de contagem considerados (RPC, NBI, ZIP e ZINBI) com uma aplicação a um conjunto de dados reais. Consideramos a análise das contagens de verificações corretas obtidas pela aplicação de um teste de atenção seletiva proposto por Beyzaee (2017), que corresponde às respostas de 228 alunos a 20 itens com um limite de tempo para completar a tarefa de 15 segundos, onde a tarefa é riscar os dígitos 2 e 7 em três linhas de dígitos e letras dispostos aleatoriamente (BAGHAEI; DOEBLER, 2019).

Para ilustração, mostramos a seguir apenas a estrutura hierárquica do modelo ZINBI sob abordagem bayesiana. Como já mencionado, o modelo pode ser visto como um MLGM em que

o indivíduo é considerado como efeito aleatório, o item é considerado como efeito fixo e podem ser incluídas prioris para os parâmetros de interesse. As prioris consideradas aqui são as prioris padrão (*default*) do pacote, utilizando o método INLA no R. Assim, o modelo mais geral que pode ser proposto é dado por

$$\begin{aligned} Y_{ij} \mid \theta_i, \beta_j, t_j, \phi, \omega &\sim ZINBI(\mu_{ij}, \phi, \omega) \\ log(\mu_{ij}) &= \eta_{ij} \\ \eta_{ij} &= \beta_j + \theta_i + t_j \\ i = 1, \cdots, n \quad \text{e} &\quad j = 1, \cdots, k \\ \theta_i \mid \sigma_\theta^2 &\sim N\left(0, \sigma_\theta^2\right) \\ \sigma_\theta^{-2} &\sim Gamma\left(1, 10^{-5}\right); \\ \beta_j &\sim N(0, 1000) \\ \phi &\sim N(0, 0.2) \\ logit(\omega) &\sim N(-1, 0.2). \end{aligned} \tag{3.9}$$

onde $\sigma_\theta^{-2} = \tau_\theta$ é o parâmetro de precisão e t_j é uma variável *offset* conhecida. Na formulação hierárquica do modelo acima, θ_i é um efeito aleatório com hiper prioris para o parâmetro de precisão correspondente. Além disso, as prioris para β_j, ϕ e ω são definidas. Modelos de casos particulares como RPC, NBI e ZIP podem ser obtidos eliminando algumas linhas na especificação acima.

Considerando as abordagens clássica e bayesiana, ajustamos os modelos de contagem Rasch propostos utilizando os pacotes `gamlss` e `INLA`, respectivamente. A Tabela 4 apresenta a comparação de ajuste dos modelos, baseada nos critérios de seleção discutidos na Seção 3.2. Nessa comparação, identificamos que o modelo ZIP foi o que melhor se ajustou aos dados, considerando todos os critérios avaliados.

Tabela 4 – Critérios de comparação de modelos dos modelos de contagem Rasch considerados.

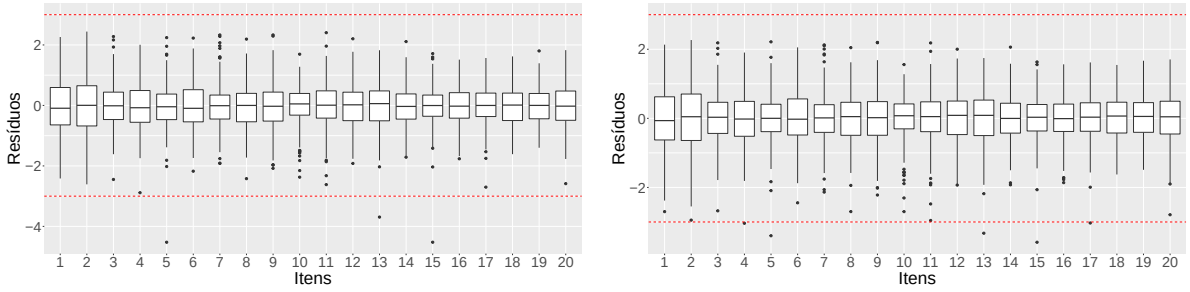
Modelos	Abordagem Clássica		Abordagem Bayesiana	
Rasch	AIC	SBC	DIC	WAIC
Pois	24639.39	26146.26	24633.84	24539.88
NBI	24641.39	26154.69	25052.45	25154.31
ZIP	24542.26	25843.72	24589.59	24493.55
ZINBI	24544.26	25852.14	24592.04	24495.15

Para verificar o ajuste do modelo ZIP aos itens, realizamos uma análise detalhada dos resíduos. Os gráficos *boxplots* apresentados na Figura 7 ilustram essa análise, com linhas adicionadas nos valores -3 e 3 para destacar os pontos discrepantes. Analisando os resultados considerando a abordagem clássica, os resíduos de Pearson na Figura 7a mostram que os itens 5, 13 e 15 apresentam pontos discrepantes significativos. Esses *outliers* indicam que o modelo não

está estimando adequadamente esses itens específicos, resultando em resíduos que se desviam significativamente do restante dos dados. Por outro lado, a análise dos resíduos quantílicos na Figura 7b revela um cenário ligeiramente diferente. Além dos itens 5, 13 e 15, o item 4 também apresenta pontos discrepantes.

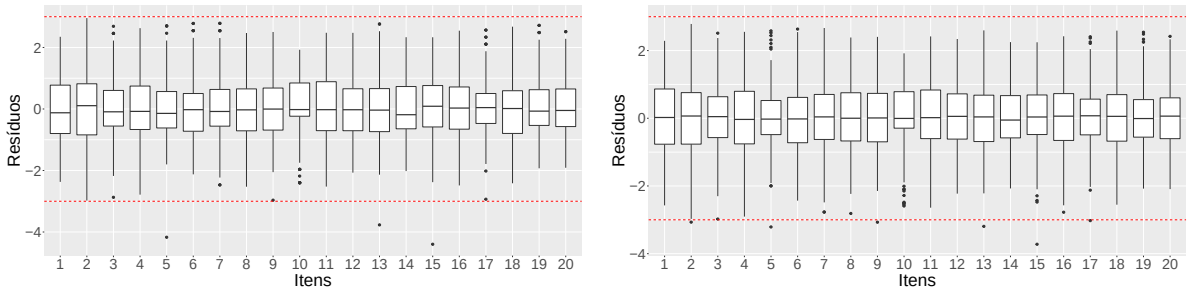
Sob a abordagem bayesiana, analisando os resíduos de Pearson (Figura 8a), identificamos que os itens 5, 13 e 15 também apresentam pontos discrepantes significativos, de maneira similar ao observado na abordagem clássica. A análise dos resíduos quantílicos, apresentada na Figura 8b, revela um cenário mais complexo. Além dos itens 5, 13 e 15, que já haviam sido identificados como discrepantes pelos resíduos de Pearson, também observamos pontos discrepantes nos itens 2, 9 e 17. Isso sugere que os resíduos quantílicos podem estar capturando outras nuances nos dados que os resíduos de Pearson não detectam, fornecendo uma visão mais completa sobre o ajuste do modelo.

Esses resultados indicam que, embora o modelo ZIP proporcione um ajuste geral robusto, há itens específicos que requerem atenção especial. Os pontos discrepantes observados tanto nos resíduos de Pearson quanto nos resíduos quantílicos destacam a necessidade de refinamento do modelo para esses itens.



(a) Resíduos de Pearson. (b) Resíduos Quantílicos.

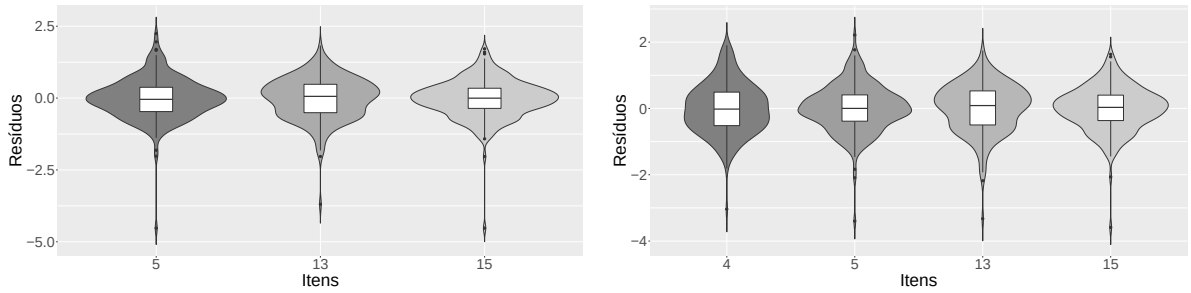
Figura 7 – Boxplot dos resíduos para o modelo Rasch ZIP sob abordagem clássica.



(a) Resíduos de Pearson. (b) Resíduos Quantílicos.

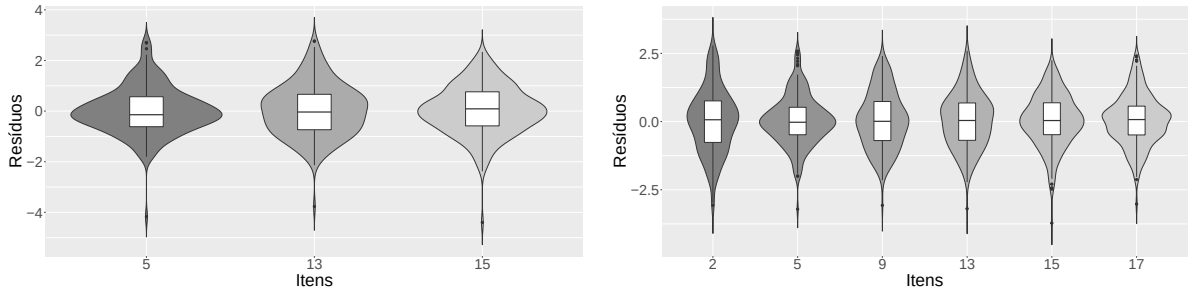
Figura 8 – Boxplot dos resíduos para o modelo Rasch ZIP sob abordagem bayesiana.

Com o intuito de esclarecer a distribuição dos resíduos nos itens identificados com discrepâncias, utilizamos gráficos de violino para ilustrar a distribuição desses resíduos sob as abordagens clássica e bayesiana, como mostrado nas Figuras 9 e 10. Esses gráficos permitem visualizar tanto a amplitude quanto a forma da distribuição dos resíduos para os itens problemáticos.



(a) Resíduos de Pearson. (b) Resíduos Quantílicos.

Figura 9 – Gráficos de violino dos resíduos para o modelo Rasch ZIP sob abordagem clássica para os itens problemáticos.



(a) Resíduos de Pearson. (b) Resíduos Quantílicos.

Figura 10 – Gráficos de violino dos resíduos para o modelo Rasch ZIP sob abordagem bayesiana para os itens problemáticos.

Na abordagem clássica, os gráficos de violino dos resíduos de Pearson (Figura 9a) indicam que os itens 5, 13 e 15 apresentam um afastamento significativo da normalidade. Idealmente, em um modelo bem ajustado, os resíduos deveriam exibir uma distribuição em forma de sino, ou seja, uma distribuição normal. No entanto, os gráficos mostram que esses itens possuem distribuições assimétricas e caudas longas, sugerindo problemas no ajuste do modelo. Os resíduos quantílicos sob a abordagem clássica (Figura 9b) revelam um cenário ligeiramente diferente. Além dos itens 5, 13 e 15, o item 4 também apresenta uma distribuição que se desvia da normalidade, indicando que mais itens possuem problemas de ajuste quando analisados com resíduos quantílicos.

Sob a abordagem bayesiana, os gráficos de violino dos resíduos de Pearson (Figura 10a) mostram que os itens 5, 13 e 15 continuam a apresentar discrepâncias significativas, semelhantes

às observadas na abordagem clássica. No entanto, os resíduos quantílicos (Figura 10b) revelam uma maior complexidade. Além dos itens 5, 13 e 15, identificamos pontos discrepantes nos itens 2, 9 e 17, sugerindo que a abordagem bayesiana pode estar capturando outras nuances nos dados que não são detectadas pelos resíduos sob abordafem clássica.

Esses resultados indicam que, embora o modelo ZIP proporcione um ajuste geral robusto, ainda há itens específicos que necessitam de maior atenção. A presença de pontos discrepantes tanto nos resíduos de Pearson quanto nos resíduos quantílicos destaca a necessidade de refinamento do modelo para esses itens. A análise de resíduos é, portanto, fundamental para diagnosticar e melhorar o ajuste do modelo, garantindo maior precisão e validade nas inferências realizadas. A maior quantidade de itens identificados como problemáticos pelos resíduos quantílicos na abordagem bayesiana sugere que esta abordagem pode ser mais conservadora, ou seja, essa característica implica que a abordagem tende a ser mais rigorosa na detecção de discrepâncias, proporcionando uma avaliação mais detalhada e precisa dos ajustes necessários no modelo.

3.5 Comentários finais

Neste capítulo, apresentamos a modelagem de modelos de contagem TRI, explorando tanto as abordagens clássica quanto a bayesiana. Entre os modelos discutidos, destacamos a introdução do modelo TRI Binomial negativa inflacionado de zeros como o mais abrangente. Utilizamos a especificação do modelo de contagem Rasch como um MLGM, com ajustes realizados pelos pacotes `gamlss` e `INLA` no software R.

Propusemos a utilização dos resíduos quantílicos aleatorizados para verificar o ajuste dos modelos, devido à sua superioridade em comparação aos resíduos de Pearson, oferecendo uma distribuição mais próxima da normalidade. Adicionalmente, apresentamos o uso de gráficos de violino como uma ferramenta gráfica eficaz para a análise da distribuição dos resíduos nos itens de um teste. Os resultados desta aplicação revelam achados interessantes e destacam o potencial da análise de resíduos, especialmente com o uso de gráficos de violino, no diagnóstico dos modelos. Esses gráficos proporcionam uma visualização detalhada da distribuição dos resíduos, facilitando a identificação de discrepâncias e aprimorando o processo de escolha do modelo mais adequado para os dados.

Os modelos discutidos mostraram-se alternativas úteis, especialmente em casos de superdispersão nos dados, como evidenciado na aplicação prática onde o modelo ZIP foi identificado como o mais adequado entre os estudados. No entanto, a análise residual indicou que, apesar de ser o melhor entre os avaliados, o modelo ZIP ainda apresenta limitações no ajuste aos dados. A identificação desses *outliers* reforça a importância de utilizar múltiplos critérios de análise de resíduos para obter uma avaliação mais abrangente do desempenho do modelo. Ajustes adicionais e possíveis modificações no modelo podem ser necessários para abordar as discrepâncias

identificadas, garantindo que o modelo final seja o mais robusto e preciso possível.

Esses achados sugerem a necessidade de novas investigações e desenvolvimentos futuros para aprimorar o ajuste dos modelos. Estudos de simulação e novas aplicações devem ser continuados, visando a elaboração de propostas que possam superar as limitações identificadas, contribuindo assim para o avanço da modelagem de dados de contagem em psicometria.

CAPÍTULO

4

MODELO DE REGRESSÃO BELL MISTO PARA DADOS DE CONTAGEM MÉDICO SUPERDISPERSOS

A distribuição Bell é uma distribuição discreta de um parâmetro que pode modelar dados de contagem com sobredispersão. Sua função de massa de probabilidade é simples, não necessitando de parâmetros adicionais para lidar com a sobredispersão, o que a torna mais parcimoniosa comparada a outras distribuições de dois ou três parâmetros. O trabalho apresentado neste capítulo se baseia em uma colaboração que propõe um novo modelo, a extensão do modelo de regressão Bell para o contexto misto, adicionando efeitos aleatórios aos efeitos fixos já existentes no preditor linear, o que é particularmente útil quando há dependência entre as observações, como em dados longitudinais ou de medidas repetidas. A estimação dos parâmetros é realizada por meio da abordagem clássica e bayesiana, utilizando o algoritmo de Rigby e Stasinopoulos (RS) no gamlss e o algoritmo No-U-Turn Sampler (NUTS) no stan. Resultados de estudos de simulação são reportados, mostrando que os modelos e métodos inferenciais propostos trazem boa recuperação dos parâmetros. A metodologia é utilizada para uma aplicação na área de saúde, sobre contagens de crises epilépticas, ilustrando sua flexibilidade e eficácia prática. A comparação com modelos tradicionais reforça a relevância do modelo proposto. Além disso, o modelo proposto neste capítulo enriquece o ferramental de alternativas para a modelagem e análise de dados de contagem.

Por mais que os desenvolvimentos não tenham relação direta com o tópico de TRI, as ideias presentes nesta parte da tese podem ser futuramente utilizadas, por exemplo, para propor um novo modelo para respostas de contagem dentro da classe dos modelos TRI, algo que, conforme ressaltado no capítulo anterior, pode ser bastante interessante no avanço do estado da arte dos modelos TRI.

O trabalho desenvolvido neste capítulo foi submetido a um periódico especializado.

4.1 Introdução

Dados de contagem ocorrem em muitos problemas práticos, como na pesquisa médica. Exemplos desse tipo de dado incluem o número de consultas que mulheres grávidas têm com um médico, as visitas ao serviço de saúde e o número de hospitalizações de pacientes, entre muitos outros. Na prática, é bem conhecido que a distribuição de Poisson de um parâmetro é a mais popular, principalmente devido à sua simplicidade. No entanto, uma desvantagem dessa distribuição é a propriedade de equidispersão (sua variância é restrita a ser igual à média). Frequentemente, nos deparamos com conjuntos de dados que exibem subdispersão ou sobredispersão, e, portanto, a distribuição de Poisson pode não ser adequada para tais casos. Essa variabilidade excessiva tem sido amplamente considerada na literatura, e para lidar com a sobredispersão, distribuições discretas com dois (CONWAY; MAXWELL, 1962; WINKELMANN, 1995) e três parâmetros (BONAT *et al.*, 2018) foram propostas como alternativas.

Recentemente, Castellares, Ferrari e Lemonte (2018) introduziram uma nova distribuição discreta baseada em uma expansão em série, chamada distribuição Bell. É uma distribuição de um parâmetro capaz de modelar dados de contagem com sobredispersão. Além disso, Castellares, Ferrari e Lemonte (2018) destacam várias propriedades interessantes, como, por exemplo, a distribuição Bell pertence à família de distribuições exponenciais de um parâmetro. A vantagem desta distribuição de um parâmetro é que ela possui uma forma muito simples para sua função de massa de probabilidade e não necessita de um parâmetro adicional para lidar com a sobredispersão, tornando a distribuição Bell mais parcimoniosa do que outras distribuições de dois e três parâmetros disponíveis na literatura para lidar com a sobredispersão.

Com base na distribuição Bell, Castellares, Ferrari e Lemonte (2018) também introduziram um novo modelo de regressão quando a variável resposta é uma contagem. Em uma configuração semelhante ao Modelo Linear Generalizado (MLG), a resposta média da variável resposta é relacionada a um preditor linear por meio de uma função de ligação. Neste artigo, estendemos o trabalho de (CASTELLARES; FERRARI; LEMONTE, 2018) introduzindo um novo modelo de regressão que adiciona efeitos aleatórios aos efeitos fixos existentes no preditor linear. Essa extensão dos MLGs é chamada de Modelos Lineares Generalizados Mistos (MLGM); veja, por exemplo, Breslow e Clayton (1993). Os MLGMs assumem que as medições Y_{ij} ($i = 1,\ldots,n$ e $j = 1,\ldots,n_i$) são independentes e pertencem à família exponencial, isto é,

$$Y_{ij} \mid \mathbf{u}_i \overset{ind}{\sim} f(y_{ij} \mid \mathbf{u}_i),$$

$$\begin{aligned} f(y_{ij} \mid \mathbf{u}_i) &= \exp\left\{[y_{ij}\eta_{ij} - a(\eta_{ij})]/\phi + c(y_{ij},\phi)\right\}, \\ g(\mu_{ij}) &= g\left(E\left[Y_{ij} \mid \mathbf{u}_i\right]\right) = \mathbf{x}_{ij}^T\beta + \mathbf{z}_{ij}^T\mathbf{u}_i, \end{aligned} \tag{4.1}$$

onde $g(\cdot)$ é a função de ligação, $\mathbf{x}_{ij}$ é um vetor associado às covariáveis, $\mathbf{z}_{ij}$ é um vetor associado aos efeitos aleatórios, β é um vetor de efeitos fixos, e $\mathbf{u}_i$ é um vetor de dimensão q de efeitos aleatórios com distribuição $N_q(\mathbf{0},\Sigma_u)$, sendo Σ_u a matriz de covariância positiva definida.

Finalmente, $a(\cdot)$ e $c(\cdot,\cdot)$ são funções conhecidas, e $\phi > 0$. Vale ressaltar que os MLGMs são adotados quando não há independência entre as observações, onde tais observações podem ser classificadas com base em alguns critérios espaciais ou alguma característica intrínseca para um grupo de observações ou agrupamentos. Além disso, essa classe de modelos também é aplicada quando medições repetidas estão disponíveis para o mesmo indivíduo ao longo do tempo (i.e., dados longitudinais) e, portanto, é necessário incorporar a correlação entre indivíduos no mesmo grupo, agrupamento ou medidas repetidas ao longo do tempo no mesmo indivíduo. Em todos esses casos, é necessário incluir efeitos aleatórios.

A inferência para MLGM pode ser realizada usando várias abordagens, incluindo abordagens clássicas baseadas em verossimilhança, e abordagens bayesianas baseadas em Cadeias de Markov Monte Carlo (MCMC), que frequentemente envolvem expressões que não têm uma forma fechada e cujas soluções geralmente são iterativas e podem ser computacionalmente intensivas. Devido ao constante avanço computacional, vários pacotes de *software* permitem o ajuste desses modelos. Considerando o *software* R (R Core Team, 2021), várias funções estão disponíveis, por exemplo: (i) sob a abordagem clássica: `lmer` no pacote `lme4` (BATES *et al.*, 2015), `glmmPQL` no pacote `MASS` (VENABLES; RIPLEY, 2002), `gamlss` no pacote `gamlss` (RIGBY *et al.*, 2019), entre outros; e (ii) sob a abordagem bayesiana: `bugs` no pacote `R2WinBUGS` (STURTZ; LIGGES; GELMAN, 2005), `inla` no pacote `INLA` (RUE; MARTINO; CHOPIN, 2009), `stan` no pacote `rstan` (Stan Development Team, 2020), entre outros. Em particular, para a implementação do modelo de regressão Bell misto proposto, consideramos o pacote `gamlss` sob a abordagem clássica, e `stan` no contexto bayesiano.

4.2 Distribuição Bell e modelo de regressão

A distribuição Bell é um modelo flexível para dados de contagem que permite sobredispersão (CASTELLARES; FERRARI; LEMONTE, 2018). Sua função massa de probabilidade tem a forma

$$P(Y = y \mid \mu) = \exp\left\{1 - e^{W_0(\mu)}\right\} \frac{W_0(\mu)^y B_y}{y!}, \tag{4.2}$$

onde $y = 0, 1, 2, \ldots$, $W_0(\cdot)$ é a função Lambert (CORLESS *et al.*, 1996), e $B_y = e^{-1}\sum_{k=0}^{\infty} k^y/k!$ são os números de Bell (BELL, 1934). Se Y tem fmp conforme em (4.2), escrevemos $Y \sim Bell(\mu)$. A média e a variância de Y são dadas, respectivamente, por $E[Y] = \mu$ e $Var[Y] = \mu[1 + W_0(\mu)]$. Note que $Var[Y]/E[Y] = 1 + W_0(\mu) > 1$, uma vez que $W_0(\mu) > 0$ para todos $\mu > 0$. Consequentemente, a distribuição Bell pode ser adequada para modelar dados de contagem com sobredispersão, embora possa não acomodar todas as possíveis formas de sobredispersão; veja Castellares, Ferrari e Lemonte (2018). Note também que a fmp da distribuição Bell pode ser expressa como

$$P(Y = y \mid \mu) = \exp\left\{\phi^{-1}\left[y\xi - a(\xi)\right] + c(y, \phi)\right\},$$

que está na forma da família exponencial com $\phi = 1$, $\xi = \log(W_0(\mu))$, $a(\xi) = e^{W_0(\mu)}$, e $c(y,\phi) = \log(B_y/y!) + 1$. Portanto, a distribuição Bell pertence à família de distribuições exponenciais, como mencionado anteriormente.

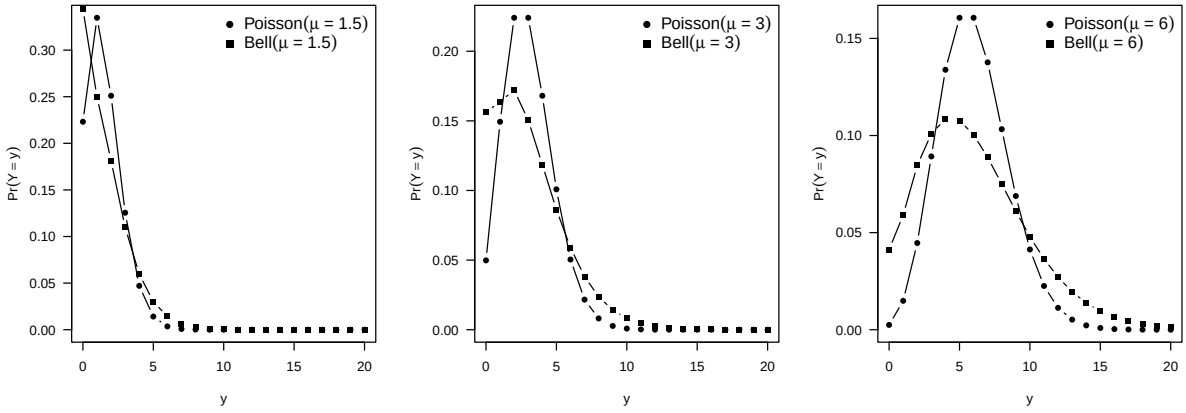


Figura 11 – Comportamento das distribuições Bell e Poisson para diferentes valores de μ.

Quanto à geração de amostras pseudo-aleatórias da distribuição $Y \sim Bell(\mu)$, recorremos a Lemonte, Moreno-Arenas e Castellares (2020), onde é demonstrado que a variável aleatória Y tem a mesma distribuição que a soma de N variáveis aleatórias independentes e identicamente distribuídas (iid) de Poisson zero truncada (ZTP) com parâmetro $W_0(\mu) > 0$, onde N é Poisson distribuída com parâmetro $e^{W_0(\mu)} - 1$. Em outras palavras, se $X_i \overset{iid}{\sim} ZTP(W_0(\mu))$ para $i = 1, \ldots, N$, com $N \sim Poisson\left(e^{W_0(\mu)} - 1\right)$ independente da sequência $X_1, \ldots, X_N$, então $Y = X_1 + \cdots + X_N \sim Bell(\mu)$. A Figura 11 ilustra algumas formas da distribuição Bell para diferentes valores de μ junto com a distribuição Poisson. A partir desta figura, é evidente que o comportamento da ditribuição muda dependendo do valor de μ. Um estudo recente sobre a diferença entre as distribuições Bell e Poisson foi fornecido por Lemonte (2022a).

Castellares, Ferrari e Lemonte (2018) estenderam a distribuição Bell para o contexto de regressão, permitindo que a média μ varie através de variáveis explicativas na forma

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \mathbf{x}_i^T \beta. \tag{4.3}$$

A função de log-verossimilhança associada é dada por

$$\ell(\beta) = \sum_{i=1}^{n} \left(1 - e^{W_0(\mu_i)} + y_i \log(W_0(\mu_i)) + \log(B_{y_i}) - \log(y_i!)\right). \tag{4.4}$$

A estimativa de máxima verossimilhança (ML), $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^T$, de $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ não possui uma expressão fechada e, portanto, deve ser obtida numericamente. No *software* R, as estimativas dos parâmetros de regressão $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ podem ser obtidas, tanto sob abordagens clássicas quanto bayesianas, usando a função `bellreg(.)` do pacote `bellreg` (DEMARQUI, 2020).

Finalmente, espera-se que o modelo de regressão Bell possa ter um desempenho melhor do que o modelo de regressão Poisson como base para um modelo de regressão misto para dados de contagem com sobredispersão, uma vez que é imediato verificar que o índice de dispersão da distribuição Bell reduz-se a

$$I_{\text{Bell}} = \frac{\text{variance} - \text{mean}}{\text{mean}} = W_0(\mu_i), \tag{4.5}$$

enquanto o índice de dispersão da distribuição de Poisson é

$$I_{\text{Poisson}} = \frac{\text{variance} - \text{mean}}{\text{mean}} = 0. \tag{4.6}$$

Assim, $I_{\text{Bell}} > 0$ para todos $\mu_i > 0$, o que significa que o modelo de regressão Bell pode lidar com sobredispersão. Por outro lado, temos que $I_{\text{Poisson}} = 0$ é constante, e assim o modelo usual de regressão de Poisson pode não ser capaz de lidar com sobredispersão.

4.3 Modelo de regressão Bell misto

Nesta seção, generalizamos o modelo de regressão Bell proposto por Castellares, Ferrari e Lemonte (2018) acomodando variabilidade extra através de uma variável não observada, o efeito aleatório, que é calculado para cada conjunto de observações dependentes. Seja $\mathbf{Y}_i = (y_{i1}, \ldots, y_{in_i})^T$ o vetor de respostas para a unidade de amostra i com n_i replicatas, onde cada componente $y_{ij} \in \{0, 1, 2, \ldots\}$. A seguinte notação é válida quando as observações dependentes são para medidas repetidas e estudos longitudinais:

$$\begin{aligned} \mathbf{Y}_{ij} \mid \mathbf{b}_i &\sim \mathit{Bell}\left(\mu_{ij}\right), \\ \mathbf{b}_i &\sim N_q\left(\mathbf{0}, \mathbf{D}\right), \\ g\left(\mu_{ij}\right) &= \mathbf{x}_{ij}^T\beta + \mathbf{z}_{ij}^T\mathbf{b}_i, \end{aligned} \tag{4.7}$$

para $j = 1, \ldots, n_i$ e $i = 1, \ldots, n$, onde $\mathbf{Y}_{ij}$ é a i-ésima amostra da unidade medida no j-ésimo ponto temporal (em um estudo longitudinal) ou no j-ésimo grupo (em um estudo de medidas repetidas), $g(\cdot)$ é uma função de ligação apropriada, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ é um vetor de coeficientes de regressão (efeitos fixos), $\mathbf{b}_i = \left(b_{i1}, \ldots, b_{iq}\right)^T$ é um vetor de efeitos aleatórios, $\mathbf{x}_{ij} = \left(1, x_{ij1}, \ldots, x_{ijp}\right)^T$ e $\mathbf{z}_{ij} = \left(z_{ij1}, \ldots, z_{ijq}\right)^T$ são vetores contendo covariáveis associadas aos efeitos fixos e aleatórios, respectivamente. Supomos que os efeitos aleatórios $\mathbf{b}_1, \ldots, \mathbf{b}_n$ são independentes e normalmente distribuídos, onde $\mathbf{D}$ é uma matriz definida positiva. A função de ligação $g(\cdot)$ relaciona os parâmetros às covariáveis, e é estritamente monótona e duas vezes diferenciável. Há várias opções possíveis para a função de ligação (MCCULLAGH, 2019). Como μ_{ij} deve ser estritamente positivo, assumimos a função de ligação logarítmica $g(\mu_{ij}) = \log(\mu_{ij})$,

mas outras funções de ligação também podem ser exploradas. A função de verossimilhança completa do modelo de regressão Bell misto é dada por

$$L(\beta, \mathbf{D}, \mathbf{b} \mid \mathbf{Y}) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \exp\left\{1 - e^{W(\mu_{ij})}\right\} \frac{W(\mu_{ij})^{y_{ij}} B_{y_{ij}}}{y_{ij}!} \, \phi_q(\mathbf{b}_i \mid \mathbf{0}, \mathbf{D}), \tag{4.8}$$

onde $\phi_q(\cdot)$ é a função densidade de probabilidade de uma distribuição normal multivariada com vetor de médias $\mathbf{0}$ e matriz de covariância $\mathbf{D}$ de dimensão q.

4.4 Inferência

Baseando-se na abordagem clássica, a inferência em MLGMs pode ser realizada usando máxima verossimilhança penalizada (PML):

$$\ell_p = \ell(\theta \mid y) - \frac{1}{2}\theta^T \mathbf{G} \theta, \tag{4.9}$$

onde $\ell(\theta|\mathbf{y})$ é a função de log-verossimilhança convencional, θ é o vetor de parâmetros do modelo (que inclui parâmetros de efeitos fixos e aleatórios), $\mathbf{y}$ é o vetor de dados observados, e $\mathbf{G}$ é uma matriz de penalidade simétrica definida positiva. A maximização de ℓ_p em relação aos parâmetros do modelo pode ser alcançada através de algoritmos de otimização, como o algoritmo de Rigby e Stasinopoulos (RS) implementado no pacote `gamlss` (RIGBY *et al.*, 2019) no *software* R. O algoritmo RS é um processo iterativo que utiliza o método de Newton-Raphson para maximizar a função de log-verossimilhança penalizada, estimando os parâmetros de efeitos fixos e aleatórios.

No contexto bayesiano, assumimos distribuições para os parâmetros, chamadas de distribuições a priori, que expressam nossa incerteza sobre os parâmetros. O modelo de regressão Bell misto bayesiano, com uma função de ligação logarítmica, é dado por

$$\begin{gathered} \mathbf{Y}_{ij} \mid \mathbf{b}_i \overset{ind}{\sim} Bell\left(\mu_{ij}\right), \quad \mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{D}), \\ (\beta, \mathbf{D}) \sim \pi(\beta, \mathbf{D}), \\ \log\left(\mu_{ij}\right) = \mathbf{x}_{ij}^T \beta + \mathbf{z}_{ij}^T \mathbf{b}_i, \end{gathered} \tag{4.10}$$

onde os parâmetros β e $\mathbf{D}$ são assumidos como independentes de modo que $\pi(\beta, \mathbf{D}) = \pi_1(\beta)\,\pi_2(\mathbf{D})$. Uma distribuição normal multivariada é proposta como a distribuição a priori para os efeitos fixos, ou seja, $\beta \sim N_p\left(\mathbf{0}, \sigma_\beta^2 \mathbf{I}_p\right)$, onde $\mathbf{I}_p$ denota uma matriz identidade $p \times p$. Para a matriz de covariância dos efeitos aleatórios, há várias possibilidades de especificações a priori disponíveis na literatura, por exemplo, a distribuição Wishart invertida pode ser adotada, ou para um componente particular de $\mathbf{D}$, σ_b^2, a distribuição gama inversa pode ser considerada.

Considerando a função de verossimilhança em (4.8), a distribuição a posteriori, denotada por $\pi(\beta, \mathbf{D}, \mathbf{b} \mid \mathbf{Y})$, pode ser expressa como

$$\begin{aligned}\pi(\beta, \mathbf{D}, \mathbf{b} \mid \mathbf{Y}) &\propto L(\beta, \mathbf{D}, \mathbf{b} \mid \mathbf{Y})\, \pi(\beta, \mathbf{D}) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{n_i} \exp\left\{1 - e^{W(\mu_{ij})}\right\} \frac{W(\mu_{ij})^{y_{ij}} B_{y_{ij}}}{y_{ij}!} \\ &\quad \times \phi_q(\mathbf{b}_i \mid \mathbf{0}, \mathbf{D})\ \phi_p\left(\beta \mid \mathbf{0}, \sigma_\beta^2 \mathbf{I}\right) \pi_2(\mathbf{D}). \end{aligned} \tag{4.11}$$

A distribuição a posteriori não possui uma expressão fechada, pois é analiticamente intratável. No entanto, uma aproximação pode ser obtida usando métodos MCMC, obtendo amostras da distribuição a posteriori. Na literatura, há vários algoritmos disponíveis para esse propósito, um dos quais é o algoritmo No-U-Turn-Sampler (NUTS) desenvolvido por Hoffman, Gelman *et al.* (2014), que é uma extensão do algoritmo de Monte Carlo Hamiltoniano (HMC). Os autores mencionam que os algoritmos HMC e NUTS são capazes de produzir resultados mais precisos em comparação com outros métodos MCMC, para vários modelos estatísticos. Os passos do algoritmo são definidos por descartar uma amostra inicial (aquecimento, que são as iterações de *burn-in* para o amostrador) e fazer saltos nos valores gerados (afinamento, que é o intervalo de afinamento usado na simulação) para que ocorra a convergência. Portanto, até que a convergência seja alcançada, o número de iterações, os valores de aquecimento e o intervalo de afinamento podem ser alterados. Este algoritmo está implementado no *software* Stan. Assim, para ajustar o modelo de regressão Bell misto proposto, consideramos o *software* R através do pacote `rstan` (Stan Development Team, 2020).

Critérios de Comparação de Modelos

Critérios de informação são medidas estatísticas para avaliação comparativa de modelos candidatos. Entre as várias medidas na literatura, consideramos o critério de informação de Akaike (AIC) proposto por Akaike (1998), o critério de informação Bayesiano (BIC) discutido por Schwarz (1978), o critério de informação de Hannan-Quinn (HQIC) desenvolvido por Hannan e Quinn (1979), o critério de informação consistente de Akaike (CAIC) proposto por Bozdogan (1987) e o critério de informação corrigido de Kullback–Leibler (KICC) discutido por Seghouane (2006), que são critérios de informação clássicos desenvolvidos com base no método ML e podem ser calculados, respectivamente, como

$$AIC = -2\log(L) + 2p,$$

$$BIC = -2\log(L) + p\log(n),$$

$$HQIC = -2\log(L) + 2p\log(\log(n)),$$

$$CAIC = -2\log(L) + p(\log(n) + 1),$$

$$KICC = -2\log(L) + ((p+1)(3n - p - 2)) + (p/(n-p)),$$

onde p é o número efetivo de parâmetros, L é a função de verossimilhança maximizada e n é o número de observações.

Sob a abordagem bayesiana, consideramos o critério de informação de desvio (DIC) proposto por Spiegelhalter *et al.* (2002), o critério de informação de Akaike esperado (EAIC) e sua versão Bayesiana (EBIC); veja, por exemplo, Gelman, Hwang e Vehtari (2014). Esses critérios são construídos com base na média posteriori do desvio e definidos como $DIC = \overline{D(\xi)} + 2p_D$, $EAIC = \overline{D(\xi)} + 2p$ e $EBIC = \overline{D(\xi)} + p\log(n)$, onde $p_D = \overline{D(\xi)} - D(\widehat{\xi})$ é chamado de número efetivo de parâmetros, com $D(\widehat{\xi})$ sendo o desvio obtido a partir das estimativas posteriori dos parâmetros. Também consideramos o critério de informação de Watanabe (WAIC) proposto por Watanabe e Opper (2010), obtido ao adicionar uma correção (p_{WAIC}) para o número efetivo de parâmetros, $WAIC = -2lppd + 2p_{WAIC}$, onde $lppd$ é o logaritmo da densidade preditiva, e para o cálculo de p_{WAIC}, neste trabalho, adotamos a versão de variância devido às suas propriedades de estabilidade, $p_{WAIC} = \sum_{i=1}^{n} Var\left[\log\left(p(y_i \mid \xi)\right)\right]$. Também consideramos o critério de informação *leave-one-out* (LOOIC) proposto por Geisser e Eddy (1979), que é um caso especial de validação cruzada no qual um ponto de dados é deixado de fora por vez e o $lppd$ é calculado com os pontos de dados restantes.

Dado um conjunto de modelos candidatos, o modelo que produz o menor valor desses critérios é o que melhor se ajusta aos dados.

4.5 Estudos de Simulação

Desempenho dos métodos de estimação

Nesta seção, realizamos um breve experimento de Monte Carlo para avaliar a capacidade de recuperação de parâmetros do modelo de regressão Bell misto proposto (introduzido na Seção 4.3) utilizando abordagens clássica e bayesiana. Para a geração de dados, utilizamos uma covariável contínua gerada a partir de uma distribuição normal padrão. Em cada conjunto de dados de Monte Carlo, as observações y_{ij} foram geradas como condicionalmente independentes dado b_i de acordo com $y_{ij} \mid b_i \sim Bell\left(\mu_{ij}\right)$ e $b_i \sim N(0, \sigma_b^2)$, onde $\log\left(\mu_{ij}\right) = \beta_0 + \beta_1 x_{ij} + b_i$, para $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, e $n_i = 3$.

Assumimos os valores verdadeiros dos parâmetros como $\beta_0 = 1.5$, $\beta_1 = 0.5$, e $\sigma_b = 0.6$, que permaneceram fixos ao longo dos experimentos de Monte Carlo, e amostras de tamanhos $n = 50$, 100, 250, e 500. Para calcular as estimativas de máxima verossimilhança (ML), usamos o algoritmo RS disponível no pacote `gamlss`. Para a inferência bayesiana, usamos o pacote `rstan` no R, considerando três cadeias MCMC. Para cada réplica, após descartar as primeiras 5000 iterações do amostrador de Gibbs, utilizamos 20000 iterações com intervalos de desbaste de 10, resultando em amostras de tamanho 1500 para cada parâmetro. As distribuições a priori foram especificadas da seguinte forma: $\beta \sim N_2(\mathbf{0}, 10^2\mathbf{I}_2)$ e $\sigma_b^{-2} \sim \text{Gamma}(1, 10^{-4})$, que são priores

comuns em análise de regressão.

Tabela 5 – Resumo do estudo de simulação Monte Carlo para o modelo Bell misto.

True	n	ML (sd)	ML Bias	ML RMSE	BE Mean (sd)	BE Mean Bias	BE Mean RMSE	CP
$\beta_0 = 1.5$	50	1.532 (0.110)	-0.032	0.115	1.514 (0.101)	-0.014	0.102	0.960
	100	1.523 (0.076)	-0.023	0.079	1.504 (0.083)	-0.004	0.083	0.920
	250	1.521 (0.050)	-0.021	0.055	1.500 (0.051)	0.000	0.051	0.950
	500	1.528 (0.032)	-0.028	0.042	1.494 (0.037)	0.006	0.037	0.950
$\beta_1 = 0.5$	50	0.501 (0.100)	-0.001	0.100	0.494 (0.110)	0.006	0.110	0.940
	100	0.486 (0.074)	0.014	0.075	0.495 (0.080)	0.005	0.080	0.930
	250	0.493 (0.042)	0.007	0.042	0.497 (0.049)	0.003	0.049	0.940
	500	0.500 (0.034)	0.000	0.034	0.492 (0.028)	0.008	0.029	0.960
$\sigma_b = 0.6$	50	0.638 (0.093)	-0.038	0.100	0.557 (0.127)	0.043	0.134	0.880
	100	0.684 (0.068)	-0.084	0.108	0.582 (0.063)	0.018	0.066	0.930
	250	0.677 (0.047)	-0.077	0.090	0.592 (0.039)	0.008	0.040	0.980
	500	0.676 (0.029)	-0.076	0.081	0.599 (0.030)	0.001	0.030	0.950

As estimativas bayesiana e ML foram avaliadas de acordo com o desvio padrão (SD), Viés $= (1/R)\sum_{r=1}^{R}(\widehat{\theta}^{(r)} - \theta)$, onde $\widehat{\theta}^{(r)}$ denota a estimativa de θ obtida na r-ésima amostra de Monte Carlo, e θ é um dos parâmetros de interesse, e a raiz do erro quadrático médio (RMSE) definido como $RMSE = \sqrt{\sum_{r=1}^{R}(\widehat{\theta}^{(r)} - \theta)^2/R}$. Um resumo do estudo de Monte Carlo é apresentado na Tabela 5. Além disso, "BE-Média" relata a média amostral das 100 estimativas obtidas para cada parâmetro sob a abordagem bayesiana. Na última coluna, CP indica a probabilidade de cobertura dos intervalos de credibilidade de 95% (calculada como a média dos limites inferior e superior entre as 100 amostras de Monte Carlo geradas).

Com base nos resultados listados na Tabela 5, observamos que tanto as estimativas de máxima verossimilhança (ML) quanto as estimativas bayesianas demonstram uma capacidade satisfatória de recuperar os valores verdadeiros dos parâmetros. À medida que o número de amostras aumenta, as estimativas dos parâmetros tornam-se mais precisas e menos variáveis. Além disso, pode-se observar que as estimativas estão relativamente próximas aos valores verdadeiros dos parâmetros, com viés próximo de zero em todos os casos, refletindo a precisão das estimativas. Também notamos que os valores de RMSE foram pequenos em todos os casos,

destacando a precisão e baixa dispersão das estimativas. Quanto à probabilidade de cobertura, que avalia a proporção de intervalos de credibilidade que englobam os valores verdadeiros dos parâmetros, observamos que a grande maioria estão muito próximos ao nível nominal fixo de 95% em todas as situações, indicando uma boa calibração dos intervalos e fornecendo uma cobertura confiável para os valores verdadeiros dos parâmetros. Finalmente, com base nas descobertas acima, os resultados obtidos no estudo de simulação de Monte Carlo demonstram que tanto as abordagens clássica quanto bayesiana são confiáveis para recuperar os parâmetros do modelo de regressão Bell misto. À medida que o tamanho da amostra aumenta, a precisão e a confiabilidade das estimativas melhoram, como esperado.

Desempenho dos critérios de comparação de modelos

Realizamos um breve estudo de simulação para avaliar se alguns dos critérios de comparação de modelos clássicos e bayesianos podem identificar corretamente o modelo subjacente verdadeiro quando outros modelos são ajustados a um conjunto de dados gerado a partir de um dos modelos ajustados. Aqui, consideramos o mesmo cenário apresentado na seção anterior. Comparamos o modelo Bell com o modelo Poisson. Consideramos os seguintes critérios de seleção de modelo: (i) sob a abordagem clássica: AIC, BIC, HQIC, CAIC e KICC; e (ii) sob a abordagem bayesiana: DIC, WAIC, LOOIC, EAIC e EBIC. Os resultados são resumidos na Tabela 6, que apresenta a porcentagem de vezes que um dos dois modelos é escolhido como o melhor modelo usando os diferentes critérios.

De acordo com a Tabela 6, observamos que os critérios de comparação de modelos clássicos (AIC, BIC e HQIC) identificam corretamente 100% das vezes o verdadeiro modelo subjacente simulado quando o outro modelo é ajustado a um conjunto de dados dado. Por outro lado, ao analisar os resultados para os critérios de comparação de modelos bayesianos, notamos que o critério LOOIC pode ser recomendado entre os critérios de comparação de modelos bayesianos.

Tabela 6 – Results (in %) of the comparison criteria.

Models	**n**	**AIC**	**BIC**	**HQIC**	**CAIC**	**KICC**	**DIC**	**WAIC**	**LOOIC**	**EAIC**	**EBIC**
Bell	50	100	100	100	100	100	79	98	99	61	61
Poisson		0	0	0	0	0	21	2	1	39	39
Bell	100	100	100	100	100	100	86	100	100	70	70
Poisson		0	0	0	0	0	14	0	0	30	30
Bell	250	100	100	100	100	100	96	100	100	88	88
Poisson		0	0	0	0	0	4	0	0	12	12
Bell	500	100	100	100	100	100	100	100	100	93	93
Poisson		0	0	0	0	0	0	0	0	7	7

4.6 Aplicação

Para ilustrar a flexibilidade prática do modelo de regressão Bell misto, consideramos o conjunto de dados de epilepsia originalmente analisado por Thall e Vail (1990), que está disponível no *software* R (R Core Team, 2021) através do pacote `MASS` (VENABLES; RIPLEY, 2002). O conjunto de dados consiste em quatro medidas repetidas de contagens de crises (cada uma tomada durante um período de duas semanas antes de uma visita clínica) para 59 pacientes epilépticos, resultando em um total de 236 observações. Assumimos que, condicional em μ_i (ou seja, condicional nos efeitos aleatórios), as contagens de crises Y_{ij} são independentes entre os sujeitos $i = 1,2,\ldots,59$ e medidas repetidas $j = 1,2,3$ e 4. Assumimos que $Y_{ij} \sim \text{Bell}(\mu_{ij})$, e

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{Trt}_i + \beta_2 \log(\text{Base}_i) + b_i, \tag{4.12}$$

onde Base_i denota as contagens de crises no período de referência para o paciente i, Trt_i indica o recebimento do medicamento anti-epiléptico Progabide em oposição a um placebo (ou seja, é uma variável indicadora para o grupo de tratamento do paciente i) e $b_i \sim N(0,\sigma_b^2)$ é o intercepto aleatório para o sujeito. Para auxiliar na convergência ao ajustar o modelo, a covariável $\log(\text{Base}_i)$ é centralizada em torno de sua média respectiva. Incluímos o modelo convencional de regressão de Poisson para fins de comparação, ou seja, também assumimos que $Y_{ij} \sim \text{Pois}(\lambda_{ij})$, onde $\log(\lambda_{ij}) = \beta_0 + \beta_1 \text{Trt}_i + \beta_2 \log(\text{Base}_i) + b_i$. Adicionalmente, também consideramos o modelo de regressão Poisson Inversa Gaussiana (IG) para fins de comparação, onde $Y_{ij} \sim \text{Pois-IG}(\lambda_{ij},\phi)$, sendo $\log(\lambda_{ij}) = \beta_0 + \beta_1 \text{Trt}_i + \beta_2 \log(\text{Base}_i) + b_i$ e $\phi > 0$; (DEAN; LAWLESS; WILLMOT, 1989).

Consideramos duas rotinas de estimação, especificamente o pacote `gamlss` e o pacote `stan`, ambos acessíveis através do *software* R para ajuste de modelos. Sob a abordagem bayesiana, consideramos 3 cadeias com 60000 iterações MCMC por cadeia, das quais as primeiras 20000 iterações foram consideradas como período de aquecimento. A seleção de modelo foi realizada com base no AIC, BIC e CAIC (abordagem clássica), e WAIC e LOOIC (abordagem bayesiana). Além disso, para inferências bayesianas, consideramos $\beta = (\beta_0,\beta_1,\beta_2)^T \sim N_3(\mathbf{0},10^2\mathbf{I}_3)$, e $\sigma_b^{-2} \sim \text{Gamma}(1,10^{-4})$.

Tabela 7 – Estimativas dos parâmetros para os modelos de regressão misto aos dados de contagem de crises.

	Poisson		**Pois-IG**		**Bell**	
	ML estimate (SE)	**p value**	**ML estimate (SE)**	**p value**	**ML estimate (SE)**	**p value**
β_0	1.9897 (0.0361)	< 0.0001	1.7695 (0.0480)	< 0.0001	2.1162 (0.0578)	< 0.0001
β_1	-0.2918 (0.0457)	< 0.0001	-0.3258 (0.0658)	< 0.0001	-0.3176 (0.0769)	< 0.0001
β_2	1.0025 (0.0287)	< 0.0001	0.9957 (0.0434)	< 0.0001	1.0057 (0.0474)	< 0.0001
ϕ			0.0831 (0.2447)			
σ_b	0.4687		0.5511		0.5834	
AIC	1290.01		1205.10		1214.23	
BIC	1434.56		1341.58		1320.88	
CAIC	1476.29		1380.98		1351.66	

A Tabela 7 lista as estimativas de máxima verossimilhança (ML) dos parâmetros para os modelos de regressão mista Bell, Poisson e Pois-IG, os erros padrão (SE) e o p-valor associado. Um *script* simples em R para obter as estimativas ML do modelo de regressão Bell misto é fornecido no Apêndice A. Note que as estimativas ML de todos os modelos de regressão mista são próximas, e os parâmetros são significativos em qualquer nível de significância usual. Adicionalmente, com base nos critérios BIC e CAIC para os modelos de regressão misto ajustados, é evidente que o novo modelo de regressão Bell misto supera o modelo tradicional de regressão Poisson misto e o modelo de regressão Pois-IG misto. Com base no critério AIC, o modelo de regressão Pois-IG misto supera os outros. Note que o modelo de regressão Bell misto é competitivo com o modelo de regressão Pois-IG misto, mesmo com menos parâmetros.

A Tabela 8 fornece as médias a posteriori e os intervalos de credibilidade de 95% para os modelos de regressão Bell misto e Poisson sob abordagem bayesiana. A convergência das cadeias MCMC foi verificada usando o critério de Gelman e Rubin (1992), ou seja, $\hat{R}$. Conforme a Tabela 8, observe que os valores de $\hat{R}$ estão próximos de um em todos os casos, indicando uma boa convergência, sugerindo que as cadeias estão amostrando adequadamente a distribuição posteriori dos parâmetros. Além disso, com base nos critérios bayesianos WAIC e LOOIC, podemos concluir que o modelo de regressão Bell misto apresenta um ajuste melhor aos dados de epilepsia do que o modelo de regressão Poisson misto tradicional.

Tabela 8 – Resumos a posteriori dos parâmetros para os modelos de regressão mistos aos dados de contagem de crises.

	Poisson			**Bell**		
	Mean	**CI 95%**	$\hat{R}$	**Mean**	**CI 95%**	$\hat{R}$
β_0	1.8019	(1.5870,2.0153)	0.9999	1.8151	(1.5965, 2.0331)	0.9998
β_1	-0.3351	((-0.6429, -0.0325)	0.9999	-0.3087	(-0.6122, -0.0132)	0.9999
β_2	1.0107	(0.8041, 1.2140)	1.0001	1.0167	(0.8172, 1.2122)	0.9999
σ_b	0.5353	(0.4245, 0.6714)	0.9998	0.4712	(0.3540, 0.6133)	1.0000
WAIC		1334.88			1235.61	
LOOIC		1341.60			1238.98	

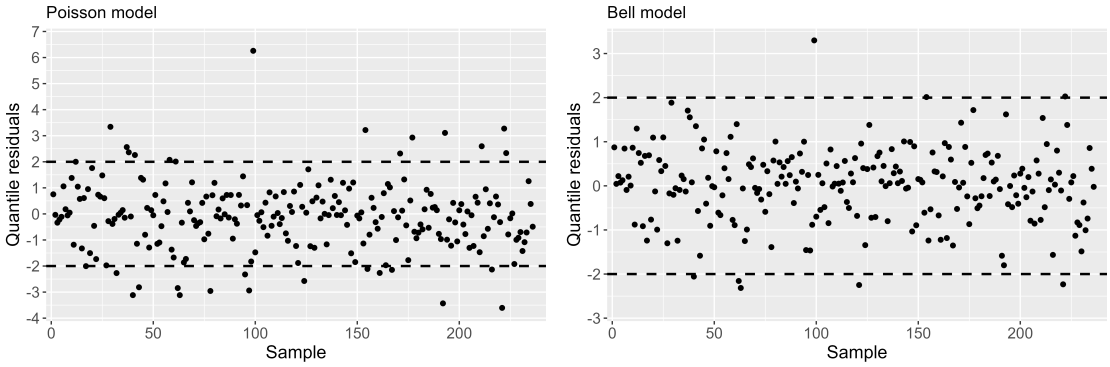


Figura 12 – Resíduo quantílico aleatorizado dos modelos de regressão Bell misto e Poisson ajustados aos dados de contagem de crises.

Finalmente, a Figura 12 exibe os resíduos quantílicos padronizados em relação ao índice dos modelos mistos Bell e Poisson. Observe que os resíduos para o modelo de regressão Bell misto parecem satisfatórios (aleatórios) e muito melhor distribuídos do que os resíduos do modelo de regressão Poisson misto, o que confirma a superioridade do modelo de regressão Bell misto proposto em relação ao conhecido modelo de regressão Poisson misto.

4.7 Considerações Finais

Neste artigo, introduzimos um novo modelo de regressão misto para lidar com variáveis de resposta de contagem. O modelo misto foi construído com base na distribuição Bell de um parâmetro recentemente proposta por Castellares, Ferrari e Lemonte (2018) como uma solução para lidar com dados superdispersos no contexto de contagem. Abordamos a inferência utilizando duas perspectivas: clássica e bayesiana. Na abordagem clássica, empregamos a estimativa de parâmetros para o modelo de regressão Bell misto usando máxima verossimilhança penalizada através do pacote `gamlss`. Por outro lado, na abordagem bayesiana, adotamos o método MCMC com o auxílio do pacote `stan`.

Os resultados de experimentos numéricos indicaram que tanto a estimativa de máxima verossimilhança quanto a abordagem bayesiana podem ser aplicadas efetivamente para estimar os parâmetros do modelo de regressão Bell misto. Além disso, com base em uma aplicação a dados reais, verificamos que o modelo de regressão Bell misto pode ser uma alternativa interessante ao modelo tradicional de regressão Poisson mista na prática e ao modelo de regressão Poisson inversa Gaussiana misto, mostrando que o modelo proposto pode ser competitivo mesmo com menos parâmetros em comparação com este modelo.

É evidente que o modelo de regressão Bell misto proposto pode ser aplicado em diferentes dados de saúde de contagem que apresentam sobredispersão. Para isso, desenvolvemos os códigos em R para estimar o modelo de regressão Bell misto como uma nova família para o pacote gamlss, o que pode ser útil para os profissionais. Em conclusão, os desenvolvimentos apresentados em relação ao modelo de regressão Bell misto destacam sua relevância e aplicabilidade potencial em cenários práticos envolvendo variáveis de resposta de contagem.

CAPÍTULO

5

MODELO DE REGRESSÃO BELL-TOUCHARD COM EFEITOS MISTOS

Quando interessados em modelar dados de contagem que apresentam sobredispersão, o uso de distribuições e modelos flexíveis se torna essencial. Este capítulo propõe um novo modelo de regressão Bell-Touchard com efeitos mistos, representando uma inovação significativa nesse contexto. Fundamentado na distribuição Bell-Touchard, uma distribuição bi-paramétrica, este modelo oferece uma alternativa eficaz para lidar com dados de contagem superdispersos. A estimação dos parâmetros do modelo é realizada por meio de métodos inferenciais clássicos, utilizando a maximização da verossimilhança penalizada, com implementação via algoritmo RS no *software* R. Estudos de simulação de Monte Carlo são conduzidos para avaliar o desempenho do modelo proposto, evidenciando sua capacidade de recuperar parâmetros com precisão em diferentes cenários e tamanhos de amostra. O modelo proposto é, por fim, utilizado em uma aplicação para respostas sobre os movimentos de *grooming* em ratos, coletados em um experimento neurofisiológico. Os métodos de diagnóstico, incluindo a análise de resíduos quantílicos aleatorizados, confirmam a adequação do modelo proposto. Os resultados mostram que o modelo de regressão Bell-Touchard misto supera modelos tradicionais de contagem, demonstrando seu potencial.

O trabalho deste capítulo será submetido a um periódico especializado.

5.1 Introdução

A distribuição mais comumente utilizada para dados de contagem é o modelo Poisson. Os modelos de regressão Poisson consideram a relação entre a resposta de contagem e os dados explicativos sob a suposição de equidispersão (ou seja, variância igual a média). No entanto, é comum que dados de contagem exibam subdispersão (variância menor que a média) ou sobredispersão (variância maior que a média). Para abordar a sobredispersão, os modelos de regressão de efeitos aleatórios Poisson são frequentemente empregados (MORRIS; SELLERS, 2022). No entanto, a estrutura de efeitos aleatórios por si só pode não ser adequada para modelar adequadamente os dados. Nesses casos, é necessário um modelo de contagem mais flexível para considerar a dispersão.

Uma abordagem proposta por Booth *et al.* (2003) consiste na inclusão de efeitos aleatórios no preditor linear, introduzindo um parâmetro adicional de dispersão através de um modelo Binomial negativa. A distribuição Binomial negativa é frequentemente utilizada para lidar com a sobredispersão. Zhang *et al.* (2017) discutem o uso de modelos mistos Binomial negativa para analisar dados de contagem de microbioma, levando em consideração tanto a sobredispersão quanto a correlação entre amostras. Além disso, pesquisadores propõem a modelagem de regressão com efeitos aleatórios fundamentada na distribuição Conway–Maxwell–Poisson (MORRIS; SELLERS, 2017; MORRIS; SELLERS; MENGER, 2017), uma alternativa para lidar com a dispersão inerente aos dados de contagem. Obviamente, outras distribuições de probabilidade discretas interessantes, que descrevem fenômenos de contagem, foram introduzidas na literatura estatística. Em particular, Castellares, Lemonte e Moreno-Arenas (2020) estudaram a família de distribuições de Bell-Touchard de dois parâmetros, com função massa de probabilidade na forma:

$$P(Y=y)=\frac{e^{\phi(1-e^{\alpha})}\alpha^{\phi}T_y(\phi)}{y!},\quad y=0,1,\cdots \tag{5.1}$$

onde $\alpha>0$, $\phi>0$ e $T_y(\phi)$ são os polinômios Touchard (TOUCHARD, 1933) dados por $T_y(\phi)=\sum_{k=0}^{y}S(y,k)\,\phi^k$, onde $S(y,k)$ são os números Stirling (COMTET, 2012).

A distribuição Bell-Touchard de dois parâmetros (abreviada como “BeTo") é um modelo flexível para dados de contagem e possui várias propriedades interessantes. O modelo BeTo foi mostrado como útil para modelar dados de contagem superdispersos. Modelagem de regressão baseada na distribuição BeTo foi originalmente proposta por Lemonte (2022b), que descreve a estimação de máxima verossimilhança do modelo de regressão Bell-Touchard com uma nova reparametrização da média, $\mu=\phi\alpha e^{\alpha}$ e portanto $\alpha=W_0(\mu/\phi)$, onde $W(\cdot)$ denota a função Lambert (CORLESS *et al.*, 1996). No *framework* de modelo misto, continuamos o estudo do modelo de regressão Bell-Touchard de Lemonte (2022b) para incluir a estrutura de efeito aleatório para modelar dados de contagem.

5.2 Distribuição Bell-Touchard e modelo de regressão

Introduzida por Lemonte (2022b), uma variável aleatória discreta Y é dita ser distribuída de acordo com a distribuição de Bell-Touchard com parâmetros $\mu > 0$ e $\phi > 0$, denotada por $Y \sim \text{BeTo}(\mu, \phi)$, se sua função massa de probabilidade é dada por

$$P(Y = y) = \exp\left\{\phi\left[1 - e^{W_0(\mu/\phi)}\right]\right\} \frac{W_0(\mu/\phi)^y T_y(\phi)}{y!}, \tag{5.2}$$

onde $W_0(\cdot)$ é a função de Lambert e $T_y(\cdot) = \sum_{k=0}^{y} S(y,k)\phi^k$ são os polinômios de Touchard (TOUCHARD, 1933), onde $S(y,k)$ são os números de Stirling do segundo tipo (COMTET, 2012). A Figura 13 ilustra algumas formas da distribuição BeTo para diferentes valores de μ e ϕ.

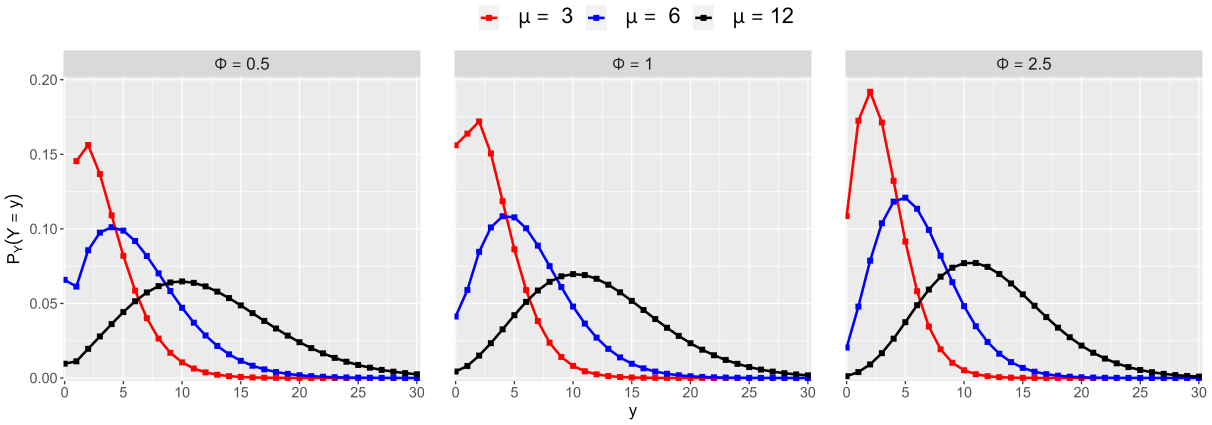


Figura 13 – Função massa de probabilidade BeTo para diferentes valores de μ e ϕ.

A média e a variância da distribuição BeTo são dadas por $E[Y] = \mu$ e $Var[Y] = \mu\,[1 + W_0(\mu/\phi)]$, respectivamente. O indice de dispersão, definido como $Var[Y]/E[Y]$, é dado por $I_Y = 1 + W_0(\mu/\phi) > 1,\ \forall\ \mu, \phi > 0$, o que implica que a distribuição BeTo pode ser adequada para modelar dados de contagem com sobredispersão (LEMONTE, 2022b). Note que a fmp da distribuição BeTo pode ser expressa como

$$\begin{aligned} P(Y = y) &= \exp\left\{y\log(W_0(\mu/\phi)) - \phi e^{W_0(\mu/\phi)} + \log\left(\frac{T_y(\phi)}{y!}\right) + \phi\right\} \\ &= \exp\{\xi M(y) - A(\mu) + C(y)\} \end{aligned} \tag{5.3}$$

que está na forma da família exponencial com $\xi = \log(W_0(\mu/\phi))$, $M(y) = y$, $A(\mu) = \phi e^{W_0(\mu/\phi)}$ e $c(y) = \log\left(T_y(\phi)/y!\right) + \phi$, para $y = 0, 1, \ldots$.

A distribuição BeTo inclui duas distribuições como casos especiais: para $\phi = 1$ reduz a distribuição Bell introduzida por Castellares, Ferrari e Lemonte (2018) e se aproxima da distribuição de Poisson conforme $\phi \to \infty$. Outro resultado da distribuição BeTo é fornecido na seguinte proposição.

Proposição 1. Seja $Y \sim \text{BeTo}(\mu, \phi)$, onde $\mu > 0$ e $\phi > 0$. Então, a variável aleatória Y tem a mesma distribuição que a soma de N variáveis aleatórias independentes e identicamente distribuídas Poisson zero-truncada com parâmetros $W_0(\mu/\phi) > 0$, com $N \sim Poisson\,(\phi(\exp\{W_0(\mu/\phi)\} - 1))$.

A Proposição 1 nos permite fornecer a seguinte caracterização: Seja $X_1, X_2, X_3, \cdots$ uma sequência de variáveis aleatórias independentes e idênticamente distribuídas, de modo que X_n tenha uma distribuição Poisson zero-truncada com parâmetro $W_0(\mu/\phi) > 0$, e seja N uma distribuição Poisson com parâmetro $\phi(\exp\{W_0(\mu/\phi)\} - 1)$ e independente da sequência $\{X_n, n \geq 1\}$. Então, a variável aleatória $Y = X_1 + X_2 + \cdots + X_N$ tem distribuição $BeTo(\mu, \phi)$, onde $\mu > 0$ e $\phi > 0$.

Lemonte (2022b) estendeu a distribuição BeTo para o contexto de regressão, permitindo que o parâmetro μ varie para cada observação i. Esta formulação de regressão relaciona μ_i às variáveis explicativas usando a ligação

$$\log \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \mathbf{x}_i^T \beta \tag{5.4}$$

cuja função log-verossimilhança associada é dada por

$$\ell(\beta, \phi) = \phi \sum_{i=1}^{n} \left[1 - e^{W_0(\mu_i/\phi)}\right] + \sum_{i=1}^{n} y_i \log\left(W_0(\mu_i/\phi)\right) + \sum_{i=1}^{n} \log\left(T_{y_i}(\phi)\right). \tag{5.5}$$

Dada esta construção, o modelo de regressão BeTo tem como caso especial o modelo de regressão Bell quando $\phi = 1$. O leitor é referenciado a (CASTELLARES; LEMONTE; MORENO-ARENAS, 2020; LEMONTE, 2022b) para uma descrição mais detalhada sobre a família de distribuições BeTo de dois parâmetros.

5.3 Modelo de regressão Bell-Touchard misto

Em conformidade com a notação apresentada por Verbeke e Molenberghs, seja y_{ij} a j-ésima medição disponível para o i-ésimo grupo (ou *clusters*), $i = 1, \ldots, n$ e $j = 1, \ldots, n_i$, onde grupo tem um sentido geral para incluir tipos específicos de dados correlacionados, tais como longitudinais, medidas repetidas, bem como dados espaciais e familiares. O modelo de regressão Bell-Touchard misto é definido, assumindo-se que, condicionalmente a um vetor q-dimensional de efeitos aleatórios $\mathbf{b}_i$, as variáveis respostas Y_{ij} são consideradas independentes e têm distribuição Bell-Touchard

$$\begin{aligned} Y_{ij} \mid \mathbf{b}_i &\sim BeTo\left(\mu_{ij}, \phi\right) \\ \log\left(\mu_{ij}\right) &= \mathbf{x}_{ij}^T \beta + \mathbf{z}_{ij}^T \mathbf{b}_i \end{aligned} \tag{5.6}$$

em que $\mathbf{x}_{ij}^T$ e $\mathbf{z}_{ij}^T$ são vetores p-dimensional e q-dimensional, respectivamente, β é um vetor de parâmetros de regressão e $\mathbf{b}_i = (b_{i1}, \ldots, b_{iq})^T$ é um vetor de efeitos aleatórios. Para ilustrar,

consideramos o caso com apenas um efeito aleatório $(q = 1)$. Neste caso, o modelo de regressão BeTo com intercepto aleatório é definido por

$$\begin{aligned} Y_{ij} \mid b_i &\sim BeTo\left(\mu_{ij}, \phi\right), \\ \log\left(\mu_{ij}\right) &= \mathbf{x}_{ij}^T \beta + b_i \\ b_i &\sim f\left(b_i \mid \theta\right) \end{aligned} \tag{5.7}$$

onde $f(b_i \mid \theta)$ é a densidade do efeito aleatório b_i caracterizado por parâmetros θ. Isto é, o resultado de contagem para o grupo i na ocorrência j segue uma distribuição Bell-Touchard condicional a um efeito aleatório específico do grupo, um conjunto de covariáveis e um vetor de parâmetros de regressão $(\beta_1, \ldots, \beta_p)$ que são comuns a todos os sujeitos. Logo, a função massa de probabilidade condicional é

$$f\left(y_{ij} \mid b_i\right) = \frac{W_0(\mu_{ij}/\phi)^{y_{ij}} T_{y_{ij}}(\phi)}{y_{ij}!} \exp\left\{\phi\left(1 - e^{W_0(\mu_{ij}/\phi)}\right)\right\}. \tag{5.8}$$

Na literatura de modelos mistos, é usual assumir que os efeitos aleatórios possuem distribuição normal: $b_i \sim N(0, \sigma_b^2)$. Ao assumir essa suposição para o nosso modelo misto BeTo, não obtemos uma forma fechada para a verossimilhança marginal e, portanto, a estimação de máxima verossimilhança demanda o uso de técnicas computacionais. A verossimilhança marginal para o modelo BeTo, com $b_i \sim N(0, \sigma_b^2)$, para o grupo i é dada por

$$\begin{aligned} L_i\left(\beta, \phi, \sigma_b^2\right) &= \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} f\left(y_{ij} \mid b_i\right) g\left(b_i\right) db_i \\ &= \int_{-\infty}^{\infty} \left[\prod_{j=1}^{n_i} \frac{W_0(\mu_{ij}/\phi)^{y_{ij}} T_{y_{ij}}(\phi)}{y_{ij}!} \exp\left\{\phi\left(1 - e^{W_0(\mu_{ij}/\phi)}\right)\right\}\right] g\left(b_i\right) db_i \end{aligned} \tag{5.9}$$

5.4 Inferência

Estimação

Inferencialmente, as estimativas de maxima verossimilhança dos parâmetros do modelo BeTo misto podem ser obtidas no R utilizando integração numérica ou otimização para maximizar a log-verossimilhança marginal (MORRIS; SELLERS, 2017). No entanto, optamos pela verossimilhança penalizada em vez da verossimilhança marginal. Os estimadores de verossimilhança penalizada são obtidos maximizando a verossimilhança penalizada e pode ser calculados numericamente pelo algoritmo Rigby e Stasinopoulos (RS) (STASINOPOULOS *et al.*, 2017). A função log-verossimilhança penalizada é dada por:

$$\ell_p = \ell(\theta|\mathbf{y}) - \frac{1}{2}\theta^T \mathbf{G} \theta, \tag{5.10}$$

onde a função log-verossimilhança $\ell(\boldsymbol{\theta}|\mathbf{y})$ é

$$\begin{aligned}\ell(\boldsymbol{\theta}|\mathbf{y}) &= \sum_{i=1}^{n}\sum_{i=1}^{n_i}\log\left[f(y_{ij}\mid\mu_{ij},\phi\right] \\ &= \sum_{i=1}^{n}\sum_{i=1}^{n_i}\log\left[\frac{W_0(\mu_{ij}/\phi)^{y_{ij}}T_{y_{ij}}(\phi)}{y_{ij}!}\right]\phi\left(1-e^{W_0(\mu_{ij}/\phi)}\right)\end{aligned} \tag{5.11}$$

$\boldsymbol{\theta}$ é o vetor de parâmetros do modelo e $\mathbf{G}$ é uma matriz de penalização simétrica e positiva definida. O algoritmo RS envolve três etapas: a iteração externa, a iteração interna e o ajuste retroativo modificado. Estas etapas são estruturadas de forma aninhada, de modo que cada iteração externa chama repetidamente a iteração interna, que, por sua vez, invoca o algoritmo de ajuste retroativo modificado repetidas vezes. A convergência é alcançada quando os três algoritmos convergem (STASINOPOULOS *et al.*, 2017). Este método está implementado no pacote `gamlss` do *software* R (R Core Team, 2021; RIGBY *et al.*, 2019).

Critérios de Comparação de Modelos

Os critérios de informação são estatísticas utilizadas para comparar modelos candidatos. Dentre as diversas medidas presentes na literatura, consideramos o critério de informação de Akaike (AIC) (AKAIKE, 1998), o critério de informação bayesiano (BIC) (SCHWARZ, 1978), o critério de informação Hannan-Quinn (HQIC) (HANNAN; QUINN, 1979), o critério consistente de informação de Akaike (CAIC) (BOZDOGAN, 1987) e o critério de informação Kullback-Leibler corrigido (KICC) (SEGHOUANE, 2006). Esses critérios baseiam-se no método da máxima verossimilhança (ML) e podem ser calculados, respectivamente, como:

$$AIC = -2\log(L) + 2p,$$

$$BIC = -2\log(L) + p\log(n),$$

$$HQIC = -2\log(L) - 2p\log(\log(n)),$$

$$CAIC = -2\log(L) + p(\log(n)+1),$$

$$KICC = -2\log(L) + ((p+1)(3n-p-2)) + (p/(n-p)),$$

onde p é o número efetivo de parâmetros, L é a função de verossimilhança maximizada e n é o número de observações. Todos esses critérios estão disponíveis no pacote `ICglmm` no R (R Core Team, 2021; SAGLAM; DUNDER, 2021).

Entre um conjunto de modelos candidatos, aquele que apresentar o menor valor dentre os critérios de informação é considerado o que melhor se ajusta aos dados.

Inferência sobre efeitos aleatórios

Para a seleção da parte aleatória em modelos mistos, conduzimos testes de hipóteses sobre os componentes de variância utilizando o Teste de Razão de Verossimilhança (TRV). As hipóteses são formuladas como $H_0 : \sigma^2 = 0$ *versus* $H_A : \sigma^2 > 0$, onde σ é o componente de variância em questão. A estatística do teste é calculada por $TRV = 2\left[\ell(\hat{\theta}_{H_A}) - \ell(\hat{\theta}_{H_0})\right]$, comparando a verossimilhança maximizada do modelo completo com a do modelo restrito, mantendo constante o número de parâmetros de efeito fixo e variando o número de parâmetros de efeito aleatório. No caso de o teste estar no limite do espaço paramétrico, a distribuição do TRV é uma mistura de qui-quadrados, especificamente: $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ (VERBEKE; MOLENBERGHS, 2003).

Diagnóstico do Modelo

Para avaliar a adequação do modelo proposto, empregamos métodos de diagnóstico centrados na análise dos resíduos quantílicos aleatorizados, conforme descrito por Dunn e Smyth (1996), expressos pela equação $q_{ij} = \Phi^{-1}(U_i)$, onde Φ denota a função de distribuição acumulada da normal padrão e U_i é uma variável uniformemente distribuída no intervalo $(a_i, b_i]$. Os termos a_i e b_i são derivados da função de distribuição acumulada do modelo de contagem ajustado, $F(\hat{\theta})$, avaliada nos limites inferiores e superiores dos dados observados, respectivamente, e $\hat{\theta}$ é o vetor de parâmetros estimados. A inspeção visual do ajuste é realizada por meio do gráfico semi-normal com envelopes simulados, como proposto por Atkinson (1987). Quando o modelo se mostra bem ajustado, espera-se que a maioria dos pontos se localize dentro do envelope simulado, implicando que aproximadamente 95% desses pontos estejam dentro das bandas de confiança.

5.5 Estudos de simulação

Nesta seção, consideramos experimentos de simulação Monte Carlo para avaliar o desempenho do modelo e do método de estimação propostos. As simulações Monte Carlo foram realizadas considerando

$$\begin{aligned} y_{ij} \mid b_i &\sim BeTo\left(\mu_{ij}, \phi\right) \\ \log(\mu_{ij}) &= 1.5x_{ij1} + 0.5x_{ij2} + b_i, \\ b_i &\sim N(0, \sigma_b^2) \end{aligned} \tag{5.12}$$

para $i = 1, \cdots, n$ e $j = 1, \cdots, n_i = 3$, onde $x_{ij1} = 1$ e os valores de x_{ij2} foram obtidos como amostras aleatórias da distribuição normal padrão. Para componente de variância do intercepto aleatório consideramos $\sigma_b = 0.6$, adotamos três diferentes valores de ϕ: 0.5, 1.0 e 2.5; e quatro tamanhos de amostra, $n = 50, 100$, 200 e 350. Simulamos 100 amostras para cada cenário e os

modelos foram ajustados utilizando o algoritmo RS no pacote `gamlss` (RIGBY *et al.*, 2019) disponível no R.

Recuperação de parâmetros

Um estudo de recuperação de parâmetros foi desenvolvido para ilustrar o desempenho das estimativas clássicas do modelo proposto. O objetivo deste estudo de simulação é mostrar o comportamento das estimativas baseada na raiz do erro quadrático médio (REQM) e na média. Os conjuntos de dados foram simulados de acordo com o modelo BeTo, como estabelecidos na Seção 5.5.

Tabela 9 – Estimativas clássicas dos parâmetros do modelo BeTo e raiz do erro quadrático médio (REQM), baseadas em 100 conjuntos de dados simulados.

	Parametros	β_0	β_1	ϕ	σ_b	β_0	β_1	ϕ	σ_b	β_0	β_1	ϕ	σ_b
	Real	1.5	0.5	0.5	0.6	1.5	0.5	1.0	0.6	1.5	0.5	2.5	0.6
$n = 50$	Média	1,272	0,669	0,256	0,749	1,181	0,571	0.886	0,777	1.568	0.334	2.462	0.321
	REQM	0,259	0,145	0,398	0,186	0,342	0,143	0,366	0,206	0.205	0.328	0.121	0.278
$n = 100$	Média	1.648	0,592	0,657	0,756	1.433	0.544	0,904	0.673	1.576	0.431	2.354	0.549
	REQM	0.158	0,134	0,177	0,152	0.204	0.097	0.146	0.115	0.106	0.205	0.157	0.123
$n = 200$	Média	1.547	0.532	0.473	0.552	1.608	0.522	1.097	0.544	1.477	0.598	2.441	0.594
	REQM	0.103	0.111	0.104	0.146	0.112	0.100	0.111	0.143	0.099	0.110	0.101	0.144
$n = 350$	Média	1.501	0.540	0.522	0.608	1.499	0.487	1.021	0.617	1.509	0.491	2.493	0.535
	REQM	0.059	0.092	0.086	0.102	0.010	0.023	0.095	0.100	0.061	0.099	0.094	0.104

A média e o REQM das estimativas de cada parâmetro foram calculados para cada cenário. Os resultados são apresentados na Tabela 9. Observa-se que o REQM diminui conforme o tamanho da amostra aumenta. Além disso, a diferença entre a média das estimativas e os valores verdadeiros dos parâmetros é pequena. Concluímos que o modelo BeTo misto, discutido na Seção 5.3, com o método de estimação detalhado na Seção 5.4, representa uma alternativa viável para ajustar dados de contagem. Em resumo, os resultados da simulação de Monte Carlo forneceram uma indicação clara de que o método de estimação utilizando o algoritmo RS pode ser usado efetivamente para estimar os parâmetros do modelo BeTo.

Seleção de modelo

Para analisar o desempenho dos critérios de seleção AIC, BIC, HQIC, CAIC e KICC, apresentados na Seção 5.4, realizamos um estudo de simulação considerando diferentes cenários apresentados anteriormente. Comparamos o modelo Bell-Touchard (BT) com o modelo Poisson (PO) e Binomial Negativa (BN) para cada amostra. Em seguida, determinamos a porcentagem de casos em que cada distribuição apresentou o menor valor dos critérios. O propósito desta análise é verificar se à medida que o tamanho da amostra aumenta, os critérios de seleção conseguem escolher o modelo correto.

Tabela 10 – Porcentagem de vezes em que o modelo correto foi selecionado, em um total de 100 réplicas para cada cenário.

n	Modelos	$\phi = 0.5$					$\phi = 1.0$					$\phi = 2.5$				
		AIC	BIC	HQIC	CAIC	KICC	AIC	BIC	HQIC	CAIC	KICC	AIC	BIC	HQIC	CAIC	KICC
50	PO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	BN	0	2	0	2	0	0	1	1	1	1	11	14	11	14	11
	BE	42	48	45	48	48	92	91	92	91	91	6	6	6	6	6
	BT	58	50	55	50	52	8	8	7	8	8	83	80	83	80	83
100	PO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	BN	0	0	0	0	0	0	0	0	0	0	4	4	4	4	4
	BE	37	44	43	44	44	51	57	57	57	57	0	0	0	0	0
	BT	63	56	57	56	56	49	43	43	43	43	96	96	96	96	96
200	PO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	BN	0	0	0	0	0	0	0	0	0	0	0	2	0	2	0
	BE	12	7	7	7	7	28	27	28	27	28	0	0	0	0	0
	BT	88	93	93	93	93	72	73	72	73	72	100	98	100	98	100
350	PO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	BN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	BE	0	0	0	0	0	8	12	8	12	8	0	0	0	0	0
	BT	100	100	100	100	100	92	88	92	88	92	100	100	100	100	100

Os resultados da simulação, conforme apresentados na Tabela 10, ilustram a eficácia dos critérios de seleção de modelos (AIC, BIC, HQIC, CAIC e KICC) em identificar o modelo mais adequado em diversos cenários com diferentes tamanhos de amostra. Observa-se que, à medida que o tamanho da amostra aumenta, a porcentagem de escolha do modelo Bell-Touchard (BT) também cresce significativamente em todos os critérios, com variações de 88% a 100% para amostras de 350 observações.

Para valores menores de ϕ (0.5), o modelo BT continua sendo predominantemente escolhido, porém com uma frequência menor em comparação com os demais cenários. No cenário em que ϕ é igual a 1, os critérios demonstram uma preferência pelo modelo BE em detrimento do modelo BT. Esta observação pode ser atribuída à relação específica entre os dois modelos: o modelo BE é um caso especial do modelo BT quando $\phi = 1$. Sob tais condições, os critérios de seleção tendem a favorecer o modelo BE, possivelmente devido à sua simplicidade e ao ajuste mais específico aos dados quando a sobredispersão é moderada. Quando ϕ é 2.5, mesmo em amostras menores ($n = 50$), o modelo BT é claramente preferido pelos critérios de seleção de modelo, com porcentagens que variam de 83% a 100% dependendo do critério. Este padrão reforça a ideia de que o modelo BT é particularmente adequado para dados que apresentam uma variabilidade maior do que a esperada, característica típica de dados com sobredispersão.

5.6 Aplicação

Para ilustrar a flexibilidade e aplicabilidade prática do modelo de regressão misto Bell-Touchard, utilizamos um conjunto de dados proveniente de um experimento neurofisiológico realizado no Laboratório Experimental de Neurofisiologia e Neuroetologia da Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Brasil, envolvendo oito ratos machos da espécie Guerra (ACHCAR; COELHO-BARROS; MARTINEZ, 2008). O objetivo do experimento era analisar o número de movimentos de *grooming* (contagem de *grooming*) em ratos após

receberem dois tratamentos: solução salina (placebo) seguido de ocitocina. Após cada aplicação de tratamento, as contagens de *grooming* foram registradas 12 vezes a cada cinco minutos, resultando em um total de 24 observações por rato.

O conjunto de dados, disponibilizado e previamente analizados por Achcar, Coelho-Barros e Martinez (2008), consiste em 192 observações da variável aleatória discreta Y_{ij}, que representa a j-ésima contagem de *grooming* do i-ésimo rato $(i = 1, \ldots, 8)$. Considerando uma função de ligação logarítmica, o preditor linear é expresso como

$$log\left(\mu_{ij}\right) = \beta_0 + \beta_1 \text{Tratamento}_i + \beta_2 \log(\text{Tempo}) + b_i \tag{5.13}$$

onde β_1 é o coeficiente que mede o efeito do tipo de `Tratamento` (0: solução salina, 1: ocitocina), β_2 mede o efeito do `Tempo` na escala logaritmica e $b_i \sim N\left(0, \sigma_b^2\right)$ é um intercepto aleatório específico do rato. A análise desses dados teve como foco determinar se a administração de ocitocina influencia a ocorrência e a persistência dos movimentos de *grooming* ao longo do tempo, considerando também efeitos aleatórios específicos a cada sujeito para ajustar as variações entre as medidas repetidas de cada rato.

Todas as análises foram executadas no *software* R (R Core Team, 2021), com a implementação do algoritmo RS através do pacote `gamlss` (RIGBY *et al.*, 2019). Comparamos o desempenho do modelo Bell-Touchard com modelos mistos convencionais, como Poisson e Binomial Negativa; e também o modelo Bell. Para avaliar a performance dos modelos, utilizamos os critérios AIC e BIC, conforme detalhado na Seção 5.4. Também aplicamos o teste da razão de verossimilhança, com um nível de significância de 5%, para testar a hipótese nula $H_0 : \sigma_b^2 = 0$, com o intuito de determinar a importância estatística da inclusão de efeitos aleatórios no modelo.

Tabela 11 – Estimativas dos parâmetros dos dados de grooming (erros padrão) e valores p associado, critérios de seleção de modelos e valores p para testes da razão de verossimilhança para a hipótese de teste $H_0 : \sigma_b = 0$ são exibidos entre colchetes ao lado da estimativa associada.

	Poisson		Bell		Neg. Binomial		Bell-Touchard	
	estimate	p value	estimate	p value	estimate	p value	estimate	p value
β_0	2.5416 (0.0627)	< 0.0001	2.5944 (0.1039)	< 0.0001	2.8838 (0.2307)	< 0.0001	2.5769 (0.1431)	< 0.0001
β_1	-1.0772 (0.0707)	< 0.0001	-1.1105 (0.1059)	< 0.0001	-1.2582 (0.1780)	< 0.0001	-1.1001 (0.1507)	< 0.0001
β_2	-0.0139 (0.0018)	< 0.0001	-0.0164 (0.0029)	< 0.0001	-0.0257 (0.0057)	< 0.0001	-0.0156 (0.0041)	< 0.0001
ϕ	–	–	–	–	1.1466 (0.1567)	< 0.0001	0.0352 (0.3781)	< 0.0001
σ_b	0.3856 [0.0000]		0.3976 [0.0000]		0.4075 [0.0000]		0.3909 [0.0000]	
AIC	1318.22		1024.26		994.30		929.95	
BIC	1346.05		1054.90		1023.87		959.34	

A Tabela 11 apresenta um resumo das estimativas dos parâmetros e seus correspondentes valores-p para os modelos Poisson, Binomial Neg., Bell e Bell-Touchard aplicados às contagens de *groomings*. Podemos observar, que para todos os modelos testados, o coeficiente para o tratamento é negativo e significativo, sugerindo que a administração de ocitocina reduz significativamente a contagem de *grooming* em comparação com a solução salina. O efeito do tempo também é negativo e significativo em todos os modelos, indicando uma tendência de diminuição do comportamento de *grooming* ao longo do tempo.

Todos os modelos indicam um efeito de agrupamento/longitudinal. Isso é evidente por meio das estimativas do componente de variância do efeito aleatório dos modelos, onde $\sigma_b > 0$. Testes de razão de verossimilhança para todos os modelos indicam que o efeito é estatisticamente significativo. Quando consideramos os critérios de informação, observamos que o modelo BeTo apresenta os menores valores tanto para o AIC (929.95) quanto para o BIC (959.34), indicando um melhor ajuste quando comparado aos demais modelos. Este resultado sugere que o modelo Bell-Touchard proporciona uma representação mais adequada das contagens de *groomings*.

Adicionalmente, realizamos uma análise dos resíduos utilizando os resíduos quantílicos aleatorizados para o modelo escolhido e determinamos o envelope para eles seguindo o método proposto na Seção 5.4. Assim, em consonância com os resultados apresentados, podemos observar na Figura 14 que o gráfico de resíduos quantílicos aleatorizados em relação às observações não mostra padrões claros, o que é indicativo de um bom ajuste do modelo, além disso, o gráfico semi-normal reforça essa interpretação, com a linha dos resíduos seguindo de perto a linha esperada para uma distribuição normal padrão, e apenas 4.7% dos pontos residuais situam-se fora do envelope simulado, o que é aceitável para um bom ajuste de modelo. Estas constatações realçam a adequação do modelo Bell-Touchard misto.

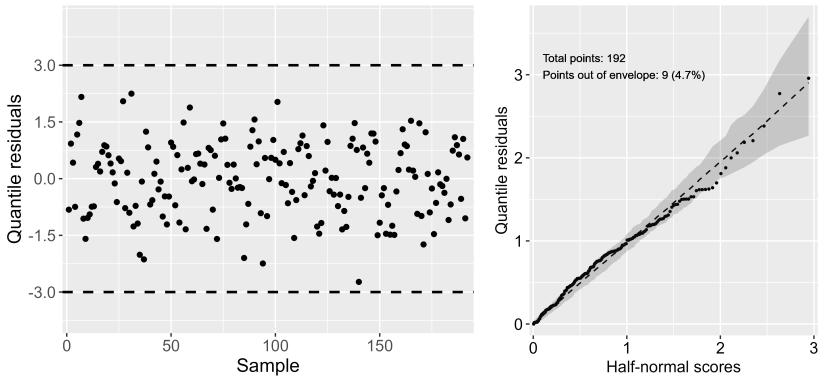


Figura 14 – Resíduo quantílico aleatorizado do modelo de regressão Bell-Touchard misto ajustados aos dados de contagem de *grooming*.

5.7 Considerações finais

Ao lidar com dados de contagem, certas estruturas inerentes a esses dados podem não ser bem representadas por distribuições padrão, levando pesquisadores a questionar a aplicabilidade de tais modelos em alguns contextos. As estruturas mais prevalentes que governam a natureza de fenômenos discretos incluem a sobredispersão e a heterogeneidade individual devido ao agrupamento ou medidas repetidas. Neste contexto, nosso objetivo foi introduzir uma alternativa para lidar com essas estruturas de dados. Nossa abordagem foi baseada em propor um novo modelo de regressão de efeitos mistos baseado na distribuição Bell-Touchard parametrizada pela média.

Neste capítulo, todos os procedimentos inferenciais foram realizados sob abordagem clássica. O algoritmo RS implementado no pacote gamlss foi empregado para maximização da verossimilhança penalizada. Estudos de simulação de Monte Carlo foram realizados, e os resultados obtidos permitiram-nos avaliar as propriedades empíricas dos estimadores e então concluir sobre a adequação da metodologia adotada para os cenários predefinidos.

O modelo proposto foi considerado para analisar um conjunto de dados reais sobre o número de movimentos de *grooming* praticados por oito ratos em um experimento neurofisiológico. A escolha do modelo de regressão Bell-Touchard misto é justificável, uma vez que a variável resposta foi identificada como superdispersa e com heterogeneidade individual devido a medidas repetidas ao longo do tempo. Ao olhar para os critérios de comparação, notamos que o modelo Bell-Touchard com intercepto aleatório superou seus concorrentes. A adequação do modelo foi avaliada utilizando os resíduos quantílicos aleatorizados por meio do gráfico semi-normal com envelopes simulados. Portanto, com base nos resultados obtidos, concluímos afirmando que o modelo proposto pode ser considerado uma excelente adição à classe de modelos de regressão de efeitos mistos, pois se mostrou ser flexível e competitivo quando se trata de modelar dados de contagem.

Para trabalhos futuros, é possível explorar o modelo proposto considerando a abordagem bayesiana, realizando estudos de sensibilidade das distibuições a priori dos parâmetros. Também há diversas possibilidades de avanços no modelo proposto. Por exemplo, pode-se considerar o cenário em que há excesso de zeros, ou ainda, flexibilizar a distribuição dos efeitos aleatórios. Além disso, no contexto de modelos de teoria de resposta ao item, a distribuição apresentada pode ser considerada como uma proposta de modelo alternativo para respostas de contagem.

CAPÍTULO

6

DISCUSSÃO E CONCLUSÕES

Este capítulo é dedicado a descrever as contribuições do presente trabalho, detalhar as produções resultantes e apresentar alguns direcionamentos de propostas futuras.

6.1 Contribuições no estado da arte

A literatura de dados de contagem ainda possui espaços para ser enriquecida com novas pesquisas, como as propostas nesta tese de Doutorado, com avanços no campo dos Modelos de Teoria de Resposta ao Item e dos Modelos Lineares Generalizados Mistos. De maneira geral, as maiores contribuições para essa área de pesquisa realizadas neste trabalho foram:

- Proposta de uma nova abordagem bayesiana para o Modelo Rasch Poisson: foi desenvolvida uma nova metodologia para a estimação de parâmetros do modelo Rasch Poisson utilizando aproximações de Laplace encaixadas e integradas com o pacote INLA em R, além de incluir técnicas de análise de resíduos através de visualização gráfica.

- Modelos TRI alternativos para respostas de contagem: a tese propôs e explorou modelos alternativos ao Rasch Poisson, como o modelo Binomial Negativo e suas variantes inflacionadas de zeros (ZIP e ZINB), através do uso de abordagens bayesianas e clássicas. A tese ainda destaca direcionamentos para a análise de resíduos, com o uso de resíduos quantílicos aleatorizados para avaliação do ajuste dos modelos propostos.

- Introdução de um novo modelo de regressão misto baseado na distribuição Bell: um novo modelo de regressão misto foi introduzido, utilizando a distribuição Bell de um parâmetro, que é particularmente eficaz para dados de contagem com superdispersão. Este modelo foi validado tanto por métodos clássicos quanto bayesianos, o qual ainda não estava presente

na literatura, e demonstrou sua utilidade prática em dados reais sobre convulsões em pacientes epilépticos.

- Proposta do modelo de regressão Bell-Touchard misto: a tese traz a proposta de um modelo misto inédito para resposta de contagem, considerando abordagem clássica. Especificamente, propõe um novo modelo de regressão Bell-Touchard, integrando efeitos mistos e oferecendo uma alternativa inovadora e eficaz para modelar dados superdispersos, o qual não estava presente na literatura nem sob abordagem clássica nem bayesiana.
- Para todos os modelos propostos, TRI e mistos, a disponibilização de códigos que possibilitam aos pesquisadores utilizarem as metodologias para seus próprios estudos futuros.

Essas contribuições destacam a importância da pesquisa em ampliar o ferramental disponível para a análise de dados de contagem, oferecendo alternativas metodológicas aplicáveis a diversas áreas do conhecimento, enriquecendo a literatura existente.

6.2 Produções

Trabalhos para eventos

- Apresentação de pôster: dos Santos N. C. A., Bazán J. L. *Bayesian approach to the Rasch Poisson model with application in Attention data* (2020). Poster Presentation at Workshop on Probabilistic and Statistical Methods, February 12-14, WPSM 2020.
- Apresentação de conferência: dos Santos N. C. A., Bazán J. L. *Residual Analysis in Single Level and Multilevel Rasch Counts Models* (2020). Oral Presentation at International Meeting of the Psychometric Society Virtual, July 13-17, IMPS 2020.
- Apresentação de seminário: dos Santos N. C. A., Bazán J. L. *Residual Analysis in Single Level and Multilevel Rasch Counts Models* (2020). Seminar Presentation at Latent Variable Models Group Virtual, September 25, 2020.
- Apresentação de conferência: dos Santos N. C. A., Bazán J. L. *A new Rasch count model* (2021). Oral Presentation at International Meeting of the Psychometric Society Virtual, July 19-23, IMPS 2021.
- Apresentação de pôster: dos Santos N. C. A., Bazán J. L. *Modelo de regressão Bell mistos para dados de contagem* (2022). Apresentação de poster no 24^o Simpósio Nacional de Probabilidade e Estatística, Gramado/RS, 31 de Julho - 05 de Agosto, SINAPE 2022.
- Apresentação de conferência: dos Santos N. C. A., Bazán J. L. *Using exams in R to develop evaluative quizzes in a Statistics I course* (2022). Oral Presentation at $11th$ International Conference on Teaching Statistics, Rosario/Argentina, September 11-16, ICOTS 2022.

- Apresentação de pôster: dos Santos N. C. A., Bazán J. L. *Modelo de regressão Bell de efeitos mistos para dados de contagem* (2022). Apresentação de pôster na 66ª Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria, Florianópolis/SC, 16-18 de Novembro, RBras 2022.

Artigos publicados ou submetidos

- dos Santos, N. C. A. and Bazán, J. L. (2021). *Residual Analysis in Rasch Poisson counts models*. Revista Brasileira de Biometria. vol 39, 206-2020. <https://doi.org/10.28951/rbb.v39i1.531>

- dos Santos N. C. A. and Bazán J. L. (2021) *Residual Analysis in Rasch Counts Models*. In: Wiberg M., Molenaar D., González J., Böckenholt U., Kim JS. (eds) Quantitative Psychology. Springer Proceedings in Mathematics & Statistics, vol 353. Springer, Cham. 285-295. <https://doi.org/10.1007/978-3-030-74772-5_26>

- dos Santos N. C. A. e Bazán J. L. (2022) *Usando exams no R para elaborar questionários avaliativos numa disciplina de Estatística I*. In: S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics. <DOI:10.52041/iase.icots11.T9A>

- dos Santos N. C. A., Bazán J. L. and Lemonte A. J. *A mixed Bell regression model for overdispersed medical count data*. (Submetido ao *Journal of Applied Statistics*)

Repositórios Github

Os repositórios do Github referenciados abaixo contêm os códigos desenvolvidos ao longo da pesquisa apresentada na tese, organizados por capítulo e finalidade. Cada repositório tem como objetivo facilitar o acesso e a replicabilidade dos estudos realizados. A seguir, detalhamos os conteúdos específicos de cada repositório:

- **Teoria de Resposta ao Item:**

  Este repositório contém os códigos referentes aos Capítulos 2 e 3 da tese. Encontram-se aqui os scripts utilizados nas análises dos modelos TRI, incluindo rotinas de estimação dos parâmetros e verificação de ajuste dos modelos, que exemplificam os conceitos abordados. (Link: https://github.com/NaiCaroline/Modelos-TRI)

- **Modelos mistos:**

  Neste repositório estão os códigos correspondentes às análises realizadas nos Capítulos 4 e 5. Dispõe-se de scripts para simulações, Testes de Razão de Verossimilhança (TRV) e os procedimentos detalhados para a análise de resíduos, abrangendo tanto os novos modelos mistos sugeridos quanto aqueles previamente estabelecidos na literatura. (Link: será disponibilizado juntamente com o artigo associado.)

- **R-exams:**

  O repositório r-exams contém códigos dos exercícios desenvolvidos ao longo dos quatro estágios de monitoria durante o doutorado. Este repositório oferece uma série de exercícios práticos e um breve tutorial introdutório sobre a utilização dos recursos disponibilizados. (Link: https://github.com/NaiCaroline/R-exams)

6.3 Possibilidades de Trabalhos Futuros

Ao longo dos capítulos, foram discutidas algumas considerações sobre as possibilidades de trabalhos futuros específicos a cada um deles. No entanto, é importante ressaltar algumas das principais direções que podem ser seguidas.

- Na continuação dos estudos sobre os modelos de contagem Rasch apresentados nos Capítulos 2 e 3, há a possibilidade de explorar a identificabilidade desses modelos. Além disso, pode-se investigar a inclusão de covariáveis na estimação dos traços latentes, utilizando estudos de simulação apropriados e análises de dados reais. Outra oportunidade consiste na aplicação de procedimentos de validação cruzada para a escolha de modelos, utilizando abordagens de aprendizado de máquina.

- Tanto o modelo Bell misto quanto o Bell-Touchard misto oferecem um amplo espaço para exploração, já que esta é uma abordagem pioneira na literatura. Diversas metodologias para a estimação dos parâmetros podem ser exploradas. No futuro, pode-se propor uma versão bayesiana para o modelo Bell-Touchard misto, detalhando sua elaboração e implementação. Ademais, podem ser desenvolvidos extensões desses modelos que considerem o excesso de zeros nos dados (LEMONTE; MORENO-ARENAS; CASTELLARES, 2020, p.ex.).

- Também podem ser propostos novos modelos TRI baseados nas distribuições Bell e Bell-Touchard, tanto sob uma abordagem clássica quanto bayesiana, visto que estes modelos ainda não foram abordados na literatura.

- Diante de todos os modelos de contagem estudados, existe a possibilidade de desenvolver um novo pacote em R que inclua os modelos propostos e trabalhados, tanto no contexto da TRI quanto de modelos mistos, abrangendo abordagens clássica e bayesiana.

REFERÊNCIAS

ACHCAR, J. A.; COELHO-BARROS, E. A.; MARTINEZ, E. Z. Statistical analysis for longitudinal counting data in the presence of a covariate considering different" frailty" models. **Brazilian Journal of Probability and Statistics**, JSTOR, p. 183–205, 2008. Citado nas páginas 71 e 72.

AKAIKE, H. Information measures and model selection. **Int Stat Inst**, v. 44, p. 277–291, 1983. Citado na página 40.

____. Information theory and an extension of the maximum likelihood principle. **Selected papers of hirotugu akaike**, Springer, p. 199–213, 1998. Citado nas páginas 55 e 68.

ATKINSON, A. C. Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis. **Clarendon Press, Oxford**, 1987. Citado na página 69.

BAGHAEI, P.; DOEBLER, P. Introduction to the rasch poisson counts model: An r tutorial. **Psychological reports**, SAGE Publications Sage CA: Los Angeles, CA, v. 122, n. 5, p. 1967–1994, 2019. Citado nas páginas 26, 28, 29, 30, 35, 37 e 42.

BAGHAEI, P.; RAVAND, H.; NADRI, M. Is the d2 test of attention rasch scalable? analysis with the rasch poisson counts model. **Perceptual and motor skills**, SAGE Publications Sage CA: Los Angeles, CA, v. 126, n. 1, p. 70–86, 2019. Citado na página 26.

BATES, D.; MÄCHLER, M.; BOLKER, B.; WALKER, S. Fitting linear mixed-effects models using lme4. **Journal of Statistical Software**, v. 67, n. 1, p. 1–48, 2015. Citado nas páginas 26 e 51.

BAZÁN, J. L. Psicometria e avaliação por testes: um marco metodológico. In: ____. [S.l.]: EdUFSCar, 2018. Citado na página 26.

BELL, E. T. Exponential polynomials. **Annals of Mathematics**, JSTOR, p. 258–277, 1934. Citado na página 51.

BEYZAEE, S. A latent variable modeling of verbal reasoning, cognitive flexibility, processing speed, sustained attention, and reading comprehension among iranian efl learners. **Unpublished master's thesis). Mashhad, Iran: Islamic Azad University**, 2017. Citado nas páginas 21, 30 e 42.

BOECK, P. D.; WILSON, M. **Explanatory item response models: A generalized linear and nonlinear approach**. [S.l.]: Springer Science & Business Media, 2004. Citado na página 35.

BONAT, W. H.; JØRGENSEN, B.; KOKONENDJI, C. C.; HINDE, J.; DEMÉTRIO, C. G. Extended poisson–tweedie: Properties and regression models for count data. **Statistical Modelling**, SAGE Publications Sage India: New Delhi, India, v. 18, n. 1, p. 24–49, 2018. Citado na página 50.

BOOTH, J. G.; CASELLA, G.; FRIEDL, H.; HOBERT, J. P. Negative binomial loglinear mixed models. **Statistical Modelling**, Sage Publications Sage CA: Thousand Oaks, CA, v. 3, n. 3, p. 179–191, 2003. Citado nas páginas 22 e 64.

BOZDOGAN, H. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. **Psychometrika**, Springer, v. 52, n. 3, p. 345–370, 1987. Citado nas páginas 55 e 68.

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. **Journal of the American statistical Association**, Taylor & Francis, v. 88, n. 421, p. 9–25, 1993. Citado nas páginas 22 e 50.

CASTELLARES, F.; FERRARI, S. L.; LEMONTE, A. J. On the bell distribution and its associated regression model for count data. **Applied Mathematical Modelling**, v. 56, p. 172 – 185, 2018. Citado nas páginas 22, 50, 51, 52, 53, 61 e 65.

CASTELLARES, F.; LEMONTE, A. J.; MORENO-ARENAS, G. On the two-parameter bell–touchard discrete distribution. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 49, n. 19, p. 4834–4852, 2020. Citado nas páginas 22, 64 e 66.

COMTET, L. **Advanced Combinatorics: The art of finite and infinite expansions**. [S.l.]: Springer Science & Business Media, 2012. Citado nas páginas 64 e 65.

CONWAY, R. W.; MAXWELL, W. L. A queuing model with state dependent service rates. **Journal of Industrial Engineering**, v. 12, n. 2, p. 132–136, 1962. Citado nas páginas 22 e 50.

CORDEIRO, G. M.; SIMAS, A. B. The distribution of pearson residuals in generalized linear models. **Computational statistics & data analysis**, Elsevier, v. 53, n. 9, p. 3397–3411, 2009. Citado na página 41.

CORLESS, R. M.; GONNET, G. H.; HARE, D. E.; JEFFREY, D. J.; KNUTH, D. E. On the lambertw function. **Advances in Computational mathematics**, Springer, v. 5, n. 1, p. 329–359, 1996. Citado nas páginas 51 e 64.

DEAN, C.; LAWLESS, J. F.; WILLMOT, G. E. A mixed poisson-inverse-gaussian regression model. **The Canadian Journal of Statistics**, v. 17, n. 2, p. 171–181, 1989. Citado na página 59.

DEMARQUI, F. **bellreg: Count Regression Models Based on the Bell Distribution**. [S.l.], 2020. R package version 0.0.1. Disponível em: <https://CRAN.R-project.org/package=bellreg>. Citado na página 52.

DOEBLER, A.; HOLLING, H. A processing speed test based on rule-based item generation: An analysis with the rasch poisson counts model. **Learning and Individual Differences**, Elsevier, v. 52, p. 121–128, 2016. Citado na página 26.

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 27, 29, 41 e 69.

FENG, C.; SADEGHPOUR, A.; LI, L. Randomized quantile residuals: an omnibus model diagnostic tool with unified reference distribution. **arXiv preprint arXiv:1708.08527**, v. 7, 2017. Citado na página 41.

FORTHMANN, B.; ÇELIK, P.; HOLLING, H.; STORME, M.; LUBART, T. Item response modeling of divergent-thinking tasks: A comparison of rasch's poisson model with a two-dimensional model extension. **The International Journal of Creativity & Problem Solving**, Korean Assn for Thinking Development, 2018. Citado na página 22.

FORTHMANN, B.; GÜHNE, D.; DOEBLER, P. Revisiting dispersion in count data item response theory models: The conway–maxwell–poisson counts model. **British Journal of Mathematical and Statistical Psychology**, Wiley Online Library, 2019. Citado nas páginas 21, 22, 26 e 35.

GEISSER, S.; EDDY, W. F. A predictive approach to model selection. **Journal of the American Statistical Association**, Taylor & Francis, v. 74, n. 365, p. 153–160, 1979. Citado na página 56.

GELMAN, A.; HWANG, J.; VEHTARI, A. Understanding predictive information criteria for bayesian models. **Statistics and computing**, Springer, v. 24, n. 6, p. 997–1016, 2014. Citado na página 56.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical science**, Institute of Mathematical Statistics, v. 7, n. 4, p. 457–472, 1992. Citado na página 60.

GONZALEZ, L. F. P. **Item response models for counting responses**. Dissertação (Mestrado) — Universidade Estadual de Campinas, SP, 2018. Citado na página 39.

HAMBLETON, R. K.; SWAMINATHAN, H. **Item response theory: Principles and applications**. [S.l.]: Springer Science & Business Media, 2013. Citado na página 26.

HANNAN, E. J.; QUINN, B. G. The determination of the order of an autoregression. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 41, n. 2, p. 190–195, 1979. Citado nas páginas 55 e 68.

HINDE, J. Compound poisson regression models. In: SPRINGER. **Glim 82: Proceedings of the international conference on generalised linear models**. [S.l.], 1982. p. 109–121. Citado na página 21.

HINTZE, J. L.; NELSON, R. D. Violin plots: a box plot-density trace synergism. **The American Statistician**, Taylor & Francis, v. 52, n. 2, p. 181–184, 1998. Citado nas páginas 27, 30 e 42.

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 55.

HUNG, L.-F. A negative binomial regression model for accuracy tests. **Applied Psychological Measurement**, Sage Publications Sage CA: Los Angeles, CA, v. 36, n. 2, p. 88–103, 2012. Citado nas páginas 22, 26, 35, 37 e 39.

JANSEN, M. G. Parameters of the latent distribution in rasch's poisson counts model. In: **Contributions to mathematical psychology, psychometrics, and methodology**. [S.l.]: Springer, 1994. p. 319–326. Citado na página 28.

JANSEN, M. G.; DUIJN, M. A. van. Extensions of rasch's multiplicative poisson model. **Psychometrika**, Springer, v. 57, n. 3, p. 405–414, 1992. Citado na página 28.

JENDRYCZKO, D.; BERKEMEYER, L.; HOLLING, H. Introducing a computerized figural memory test based on automatic item generation: An analysis with the rasch poisson counts model. **Frontiers in Psychology**, Frontiers Media SA, v. 11, p. 945, 2020. Citado na página 21.

LEMONTE, A. J. A note about model selection and hypothesis test procedure to discriminate Poisson and Bell models. **Journal of Statistical Theory and Practice**, v. 16, p. 19, 2022. Citado na página 52.

____. On the mean-parameterized bell–touchard regression model for count data. **Applied Mathematical Modelling**, Elsevier, v. 105, p. 1–16, 2022. Citado nas páginas 64, 65 e 66.

LEMONTE, A. J.; MORENO-ARENAS, G.; CASTELLARES, F. Zero-inflated bell regression models for count data. **Journal of Applied Statistics**, Taylor & Francis, v. 47, n. 2, p. 265–286, 2020. Citado nas páginas 22, 52 e 78.

MAGNUS, B. E.; THISSEN, D. Item response modeling of multivariate count data with zero inflation, maximum inflation, and heaping. **Journal of Educational and Behavioral Statistics**, SAGE Publications Sage CA: Los Angeles, CA, v. 42, n. 5, p. 531–558, 2017. Citado na página 39.

MCCULLAGH, P. **Generalized linear models**. [S.l.]: Routledge, 2019. Citado nas páginas 21 e 53.

MOLENBERGHS, G.; VERBEKE, G.; DEMÉTRIO, C. G. An extended random-effects approach to modeling repeated, overdispersed count data. **Lifetime data analysis**, Springer, v. 13, p. 513–531, 2007. Citado na página 21.

MORRIS, D.; SELLERS, K.; MENGER, A. Fitting a flexible model for longitudinal count data using the nlmixed procedure. In: **SAS Global Forum Proceedings; SAS Institute: Cary, NC, USA**. [S.l.: s.n.], 2017. Citado na página 64.

MORRIS, D. S.; SELLERS, K. F. A com-poisson mixed model with normal random effects for clustered count data. In: **Proceedings of the 61st World Statistics Congress of the International Statistical Institute, Marrakech, Morocco**. [S.l.: s.n.], 2017. p. 16–21. Citado nas páginas 22, 64 e 67.

____. A flexible mixed model for clustered count data. **Stats**, MDPI, v. 5, n. 1, p. 52–69, 2022. Citado na página 64.

MUTZ, R.; DANIEL, H.-D. The bibliometric quotient (bq), or how to measure a researcher's performance capacity: A bayesian poisson rasch model. **Journal of Informetrics**, Elsevier, v. 12, n. 4, p. 1282–1295, 2018. Citado na página 26.

PINHEIRO, H. P.; MAIA, R. P.; NETO, E. A. L.; RODRIGUES-MOTTA, M. Zero-one augmented beta and zero-inflated discrete models with heterogeneous dispersion for the analysis of student academic performance. **Statistical Methods & Applications**, Springer, v. 28, n. 4, p. 749–767, 2019. Citado na página 21.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2020. Disponível em: <https://www.R-project.org/>. Citado na página 27.

____. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2021. Disponível em: <https://www.R-project.org/>. Citado nas páginas 51, 59, 68 e 72.

RASCH, G. Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests. **Oxford, England: Nielsen & Lydiche**, 1960. Citado nas páginas 26, 27 e 38.

RIDOUT, M.; DEMÉTRIO, C. G.; HINDE, J. Models for count data with many zeros. In: INTERNATIONAL BIOMETRIC SOCIETY INVITED PAPERS CAPE TOWN, SOUTH AFRICA. **Proceedings of the XIXth international biometric conference**. [S.l.], 1998. v. 19, p. 179–192. Citado na página 22.

RIGBY, R. A.; STASINOPOULOS, M. D.; HELLER, G. Z.; BASTIANI, F. D. **Distributions for modeling location, scale, and shape: Using GAMLSS in R**. [S.l.]: CRC press, 2019. Citado nas páginas 40, 51, 54, 68, 70 e 72.

RUE, H.; MARTINO, S.; CHOPIN, N. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). **Journal of the Royal Statistical Society B**, v. 71, p. 319–392, 2009. Citado nas páginas 29 e 51.

RUE, H.; RIEBLER, A.; SØRBYE, S. H.; ILLIAN, J. B.; SIMPSON, D. P.; LINDGREN, F. K. Bayesian computing with inla: a review. **Annual Review of Statistics and Its Application**, Annual Reviews, v. 4, p. 395–421, 2017. Citado na página 29.

SAGLAM, F.; DUNDER, E. **ICglm: Information Criteria for Generalized Linear Regression**. [S.l.], 2021. R package version 0.1.0. Disponível em: <https://CRAN.R-project.org/package=ICglm>. Citado na página 68.

SANTOS, N. C. A. d.; BAZáN, J. L. Residual analysis in rasch poisson counts models. **Brazilian Journal of Biometrics**, v. 39, n. 1, p. 206–220, Mar. 2021. Disponível em: <https://biometria.ufla.br/index.php/BBJ/article/view/531>. Citado na página 25.

SANTOS, N. C. A. dos; BAZÁN, J. L. Residual analysis in rasch counts models. In: WIBERG, M.; MOLENAAR, D.; GONZÁLEZ, J.; BÖCKENHOLT, U.; KIM, J.-S. (Ed.). **Quantitative Psychology**. Cham: Springer International Publishing, 2021. p. 285–295. ISBN 978-3-030-74772-5. Citado na página 37.

SCHWARZ, G. Estimating the dimension of a model. **The annals of statistics**, JSTOR, p. 461–464, 1978. Citado nas páginas 40, 55 e 68.

SEGHOUANE, A.-K. A note on overfitting properties of kic and kicc. **Signal Processing**, Elsevier, v. 86, n. 10, p. 3055–3060, 2006. Citado nas páginas 55 e 68.

SELLERS, K. F.; SHMUELI, G. A flexible regression model for count data. **The Annals of Applied Statistics**, JSTOR, p. 943–961, 2010. Citado na página 21.

SIDUMO, B.; SONONO, E.; TAKAIDZA, I. Count regression and machine learning techniques for zero-inflated overdispersed count data: Application to ecological data. **Annals of Data Science**, Springer, p. 1–15, 2023. Citado na página 21.

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 nas páginas 40 e 56.

Stan Development Team. **RStan: the R interface to Stan**. 2020. R package version 2.21.2. Disponível em: <http://mc-stan.org/>. Citado nas páginas 51 e 55.

STASINOPOULOS, M. D.; RIGBY, R. A.; HELLER, G. Z.; VOUDOURIS, V.; BASTIANI, F. D. **Flexible regression and smoothing: using GAMLSS in R**. [S.l.]: CRC Press, 2017. Citado nas páginas 67 e 68.

STURTZ, S.; LIGGES, U.; GELMAN, A. R2winbugs: A package for running winbugs from r. **Journal of Statistical Software**, v. 12, n. 3, p. 1–16, 2005. Citado na página 51.

THALL, P. F.; VAIL, S. C. Some covariance models for longitudinal count data with overdispersion. **Biometrics**, JSTOR, p. 657–671, 1990. Citado nas páginas 21 e 59.

TOUCHARD, J. **Propriétés arithmétiques de certains nombres récurrents**. [S.l.]: Secrétariat de la société scientifique, 1933. Citado nas páginas 64 e 65.

VENABLES, W. N.; RIPLEY, B. D. **Modern Applied Statistics with S**. Fourth. New York: Springer, 2002. ISBN 0-387-95457-0. Disponível em: <http://www.stats.ox.ac.uk/pub/MASS4/>. Citado nas páginas 51 e 59.

VERBEKE, G.; MOLENBERGHS, G. The use of score tests for inference on variance components. **Biometrics**, Oxford University Press, v. 59, n. 2, p. 254–262, 2003. Citado na página 69.

VERHELST, N.; KAMPHUIS, F. A poisson-gamma model for speed tests. **Measurement and Research Department Reports**, v. 2, p. 2010–1, 2009. Citado na página 28.

WALLER, L. A.; ZELTERMAN, D. Log-linear modeling with the negative multinomial distribution. **Biometrics**, JSTOR, p. 971–982, 1997. Citado na página 21.

WANG, L. Irt–zip modeling for multivariate zero-inflated count data. **Journal of Educational and Behavioral Statistics**, SAGE Publications Sage CA: Los Angeles, CA, v. 35, n. 6, p. 671–692, 2010. Citado nas páginas 22, 35, 37 e 39.

WANG, X.; YUE, Y. R.; FARAWAY, J. J. **Bayesian regression modeling with INLA**. [S.l.]: CRC Press, 2018. Citado nas páginas 26 e 40.

WATANABE, S.; OPPER, M. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. **Journal of machine learning research**, v. 11, n. 12, 2010. Citado nas páginas 40 e 56.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. Disponível em: <https://ggplot2.tidyverse.org>. Citado na página 42.

WINKELMANN, R. Duration dependence and dispersion in count-data models. **Journal of business & economic statistics**, Taylor & Francis, v. 13, n. 4, p. 467–474, 1995. Citado na página 50.

ZHANG, X.; MALLICK, H.; TANG, Z.; ZHANG, L.; CUI, X.; BENSON, A. K.; YI, N. Negative binomial mixed models for analyzing microbiome count data. **BMC bioinformatics**, Springer, v. 18, p. 1–10, 2017. Citado nas páginas 22 e 64.

APÊNDICE

A

CÓDIGO R APLICAÇÃO MODELO BELL MISTO

Desenvolvemos os códigos em R para estimar o modelo de regressão Bell misto como uma nova família para o pacote `gamlss`. Como exemplo, apresentamos abaixo o código em R para estimar os parâmetros da regressão Bell na aplicação.

```
# Load packages
library(gamlss)
library(MASS)

# Read Bell Model
source("Bell_model_gamlss.R")

# Parameter Estimate
fit.bell = gamlss(y ~ trt + lbase + re(random = ~1|Subject),
                  data = dados, family = BELL)
summary(fit.bell)

## ******************************************************************
## Family:  c("BELL", "Bell")
##
## Call:  gamlss(formula = y ~ trt + lbase + re(random = ~1 | Subject),
##        family = BELL, data = dados, gd.tol = Inf)
##
## Fitting method: RS()
## ------------------------------------------------------------------
```

```
## Mu link function:  log
## Mu Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.11623    0.05775  36.642  < 2e-16 ***
## trtprogabide -0.31762    0.07694  -4.128 5.31e-05 ***
## lbase         1.00570    0.04745  21.194  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## ------------------------------------------------------------------
## No. of observations in the fit:  236
## Degrees of Freedom for the fit:  30.78876
##       Residual Deg. of Freedom:  205.2112
##                       at cycle:  12
## Global Deviance:     1152.651
##             AIC:     1214.228
##             SBC:     1320.875
## ******************************************************************
```


ICMC
USP
SÃO CARLOS