UNIVERSIDADE FEDERAL DO PARANÁ

JOSÉ EVANDEILTON LOPES

MODELOS DE REGRESSÃO BETA PARA DADOS DE ESCALA

CURITIBA

2023

JOSÉ EVANDEILTON LOPES

MODELOS DE REGRESSÃO BETA PARA DADOS DE ESCALA

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em estatística no Programa de Pós-Graduação em Métodos Numéricos em Engenharia (PPGMNE), Setor de Ciências Exatas, da Universidade Federal do Paraná.

Orientador: Prof. Wagner Hugo Bonat, PhD

CURITIBA

2023

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIA E TECNOLOGIA

Lopes, José Evandeilton
Modelos de regressão beta para dados de escala / José Evandeilton Lopes. – Curitiba, 2023.
1 recurso on-line : PDF.

Dissertação (Mestrado) - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-Graduação em Métodos Numéricos em Engenharia.

Orientador: Wagner Hugo Bonat

1. Modelos estatísticos. 2. Análise de regressão. 3. Escala do tipo Likert. I. Universidade Federal do Paraná. II. Programa de Pós-Graduação em Métodos Numéricos em Engenharia. III. Bonat, Wagner Hugo. IV . Título.

Bibliotecário: Leticia Priscila Azevedo de Sousa CRB-9/2029

ATA Nº371

ATA DE SESSÃO PÚBLICA DE DEFESA DE MESTRADO PARA A OBTENÇÃO DO GRAU DE MESTRE EM MÉTODOS NUMÉRICOS EM ENGENHARIA

No dia trinta de agosto de dois mil e vinte e tres às 19:00 horas, na sala Sala de reunião do DEST, Sala de reuniões do DEST, foram instaladas as atividades pertinentes ao rito de defesa de dissertação do mestrando **JOSÉ EVANDEILTON LOPES**, intitulada**: MODELOS DE REGRESSÃO BETA PARA DADOS DE ESCALA**, sob orientação do Prof. Dr. WAGNER HUGO BONAT. A Banca Examinadora, designada pelo Colegiado do Programa de Pós-Graduação MÉTODOS NUMÉRICOS EM ENGENHARIA da Universidade Federal do Paraná, foi constituída pelos seguintes Membros: WAGNER HUGO BONAT (UNIVERSIDADE FEDERAL DO PARANÁ), ANDERSON LUIZ ARA SOUZA (UNIVERSIDADE FEDERAL DO PARANÁ), SÍLVIA EMIKO SHIMAKURA (DEPARTAMENTO DE ESTATíSTICA DA UNIVERSIDADE FEDERAL DO PARANá). A presidência iniciou os ritos definidos pelo Colegiado do Programa e, após exarados os pareceres dos membros do comitê examinador e da respectiva contra argumentação, ocorreu a leitura do parecer final da banca examinadora, que decidiu pela APROVAÇÃO. Este resultado deverá ser homologado pelo Colegiado do programa, mediante o atendimento de todas as indicações e correções solicitadas pela banca dentro dos prazos regimentais definidos pelo programa. A outorga de título de mestre está condicionada ao atendimento de todos os requisitos e prazos determinados no regimento do Programa de Pós-Graduação. Nada mais havendo a tratar a presidência deu por encerrada a sessão, da qual eu, WAGNER HUGO BONAT, lavrei a presente ata, que vai assinada por mim e pelos demais membros da Comissão Examinadora.

Curitiba, 30 de Agosto de 2023.

Assinatura Eletrônica
19/09/2023 19:40:04.0
WAGNER HUGO BONAT
Presidente da Banca Examinadora

Assinatura Eletrônica
19/10/2023 10:46:13.0
ANDERSON LUIZ ARA SOUZA
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
18/09/2023 17:42:57.0
SÍLVIA EMIKO SHIMAKURA
Avaliador Externo (DEPARTAMENTO DE ESTATíSTICA DA UNIVERSIDADE FEDERAL DO PARANá)

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação MÉTODOS NUMÉRICOS EM ENGENHARIA da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **JOSÉ EVANDEILTON LOPES** intitulada: **MODELOS DE REGRESSÃO BETA PARA DADOS DE ESCALA**, sob orientação do Prof. Dr. WAGNER HUGO BONAT, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 30 de Agosto de 2023.

Assinatura Eletrônica
19/09/2023 19:40:04.0
WAGNER HUGO BONAT
Presidente da Banca Examinadora

Assinatura Eletrônica
19/10/2023 10:46:13.0
ANDERSON LUIZ ARA SOUZA
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
18/09/2023 17:42:57.0
SÍLVIA EMIKO SHIMAKURA
Avaliador Externo (DEPARTAMENTO DE ESTATíSTICA DA UNIVERSIDADE FEDERAL DO PARANá)

Dedico esta dissertação, em primeiro lugar, a Deus, fonte de toda sabedoria e conhecimento. Agradeço à minha querida esposa, Dorinha, por sua constante companhia e apoio, nos momentos de alegria e nos desafios.

AGRADECIMENTOS

Gostaria de expressar minha profunda gratidão ao meu orientador, Professor Dr. Wagner Hugo Bonat[1]. Seu empenho, atenção meticulosa e orientação contínua foram cruciais para a realização deste trabalho com excelência.

Um agradecimento especial aos ilustres membros da banca de avaliação, Drª. Silvia Emiko Shimakura[2] e Dr. Anderson Luiz Ara Souza[3]. Suas valiosas contribuições foram fundamentais para o refinamento e aprimoramento da pesquisa.

Adicionalmente, estendo meus sinceros agradecimentos ao Professor Dr. Paulo Justiniano Ribeiro Jr.[4]. Suas observações perspicazes e sugestões construtivas enriqueceram substancialmente o debate da banca.

1 http://www.leg.ufpr.br/~wagner/
2 http://www.leg.ufpr.br/~silvia/
3 http://www.leg.ufpr.br/~ara/
4 http://www.leg.ufpr.br/~paulojus/

“Aqui termina meu relato. Esta é minha conclusão: tema a Deus e obedeça a seus mandamentos, pois esse é o dever de todos. Eclesiastes 12:13, [NVI]

RESUMO

A pesquisa focou no estudo de dados em escala obtidos através de instrumentos como *Numerical Rating Scale (NRS)* e escalas *Likert*, amplamente utilizados em domínios médicos, especialmente na avaliação da dor. Três modelos estatísticos foram analisados: "betareg" (M1), "betaregesc" (M2) e, notavelmente, o "quasibeta" (M3) - que se diferencia ao focar nos primeiros e segundos momentos estatísticos, afastando-se das suposições distributivas convencionais. Estes modelos foram testados em dados mapeáveis na escala beta, típicos das ferramentas mencionadas. Para M3, abordagens baseadas em pseudo-verossimilhança foram adotadas, com referências como (BONAT; JØRGENSEN, 2016) e (BONAT et al., 2019). As simulações mostraram que, enquanto M1 e M2 tiveram precisão e robustez notáveis, o "quasibeta" (M3) apresentou características únicas, como tendências de subestimação de efeitos de covariáveis. Contudo, quando aplicados a dados reais, todos os modelos revelaram sua importância, embora M3 tenha apresentado algumas limitações no contexto avaliado. Em resumo, a pesquisa destacou a importância dos modelos de regressão beta, especialmente o M3, para interpretar dados beta. A seleção do modelo deve considerar as características específicas de cada conjunto de dados, como tamanho da amostra e contexto de aplicação.

Palavras-chaves: Regressão beta; quasi-beta; Dados de escala; Escalas NRS e Likert; Pseudo verossimilhança

ABSTRACT

The research focused on studying scale data obtained from instruments like the *Numerical Rating Scale (NRS)* and *Likert* scales, widely used in medical domains, especially in pain assessment. Three statistical models were analyzed: "betareg" (M1), "betaregesc" (M2), and notably, the "quasibeta" (M3) - which stands out by focusing on the first and second statistical moments, moving away from conventional distributional assumptions. These models were tested on data mappable to the beta scale, typical of the aforementioned tools. For M3, pseudo-likelihood-based approaches were adopted, with references like (BONAT; JØRGENSEN, 2016) and (BONAT et al., 2019). Simulations revealed that while M1 and M2 showcased notable precision and robustness, the "quasibeta" (M3) exhibited unique traits, such as tendencies to underestimate covariate effects. However, when applied to real data, all models underscored their significance, though M3 showed some limitations in the evaluated context. In summary, the research emphasized the importance of the beta regression models, especially M3, in interpreting beta data. Model selection should account for specific data set characteristics, such as sample size and application context.

Key-words: Beta regression. quasi-beta. Scale data. NRS and Likert. Pseudo likelihood

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Ao longo das últimas décadas, a estatística tem desempenhado um papel crucial na análise de dados em uma ampla gama de campos do conhecimento. Dentre as ferramentas estatísticas disponíveis, os modelos de regressão têm se destacado como os mais populares para modelar a relação entre uma variável resposta (dependente) e uma ou mais variáveis explicativas (independentes).

Modelos de regressão evoluíram ao longo do tempo, passando por diversas modificações e aprimoramentos. Uma das primeiras classes de modelos de regressão é a dos modelos lineares simples, os quais objetivam modelar a relação entre uma variável resposta Y e uma única variável explicativa X de forma linear. Esse tipo de modelo foi inicialmente proposto por Galton (1886) e, posteriormente, aprimorado por Pearson (1901). Uma das principais aplicações dos modelos lineares simples encontra-se na área da econometria, onde se busca modelar a relação entre uma variável resposta, como o preço de um produto, e uma variável explicativa, como a demanda por esse produto. Um exemplo de aplicação dos modelos lineares simples pode ser observado no estudo realizado por Shahbaz et al. (2020), que analisou a relação entre o consumo de energia e o crescimento econômico nos países da *ASEAN*. Entretanto, os modelos lineares simples apresentam algumas limitações: eles pressupõem que a relação entre as variáveis é linear e que os erros são independentes e identicamente distribuídos, com média zero e variância constante. Quando essas suposições são violadas, os resultados obtidos podem ser inválidos. Para contornar essas limitações, modelos mais avançados foram propostos, como os modelos de regressão múltipla e os modelos não lineares.

Os modelos lineares múltiplos foram desenvolvidos para abordar situações nas quais há mais de uma variável explicativa relacionada à variável resposta. Esse tipo de modelo possibilita a análise conjunta dos efeitos das variáveis explicativas sobre a variável resposta. Os modelos lineares múltiplos foram propostos inicialmente por Fisher (1922) e ganharam popularidade a partir da década de 1950, com o advento da computação. Esses modelos têm diversas aplicações em diferentes áreas do conhecimento. Por exemplo, na área da saúde, podem ser utilizados para modelar características entre múltiplas variáveis explicativas e a mortalidade em uma determinada população (CHEN et al., 2020). Outra aplicação dos modelos lineares múltiplos ocorre na análise de dados de mercado, onde se busca modelar a relação entre o preço de um produto, suas características e a demanda pelo produto (MARSCHAK; ANDREWS, 1944).

Os Modelos Lineares Generalizados (MLGs) representam uma classe de modelos que englobam tanto os modelos lineares simples quanto os modelos lineares

múltiplos, permitindo a modelagem de relações mais complexas entre a variável resposta e as variáveis explicativas. Os MLGs foram propostos por Nelder e Wedderburn (1972) como uma extensão dos modelos lineares clássicos, apresentando a chamada família exponencial, que é uma família de modelos probabilísticos com características em comum. A principal característica dos MLGs é a possibilidade de se modelar diferentes tipos de variáveis resposta, como variáveis categóricas e contínuas com distribuição assimétrica. Além disso, os MLGs permitem a modelagem de diferentes funções de ligação entre as variáveis resposta e as variáveis explicativas, o que pode ser útil em situações onde a relação entre essas variáveis é não-linear. Essa classe de modelos tem aplicação em diversas áreas do conhecimento, como na análise de dados biomédicos e epidemiológicos, exemplificado em Biggerstaff e Jackson (2008), e na modelagem de dados financeiros, como abordado por Tseng et al. (2017). Outro exemplo de aplicação dos MLGs ocorre na análise de dados de contagem, como o número de ocorrências de um determinado evento em uma população. Nesse caso, pode-se utilizar a função de ligação logarítmica (função log) para modelar a relação entre a variável resposta e as variáveis explicativas (AGRESTI, 2002).

Na evolução das técnicas estatísticas, surgiram os Modelos Lineares Generalizados com Efeitos Mistos (MLGM) e os Modelos Longitudinais (MLGL). Estes são extensões dos Modelos Lineares Generalizados (MLG) e permitem a modelagem de dados com estruturas de correlação mais complexas. Os MLGM são utilizados na análise de dados longitudinais, que consistem em observações repetidas ao longo do tempo de uma unidade experimental, como estudos clínicos e epidemiológicos. Estes modelos combinam efeitos fixos, que representam os efeitos médios das variáveis explicativas na variável resposta, e efeitos aleatórios, que representam a variabilidade não explicada pelos efeitos fixos. Os efeitos aleatórios são modelados por um ou mais termos de efeitos aleatórios, que descrevem a estrutura de correlação entre observações repetidas, como a correlação intraclasse (ICC) e a correlação autoregressiva (AR).

Os MLGL possibilitam a modelagem de dados longitudinais por meio de uma função de correlação específica, como o modelo de regressão com erros correlacionados Hedeker e Gibbons (2006), utilizado para modelar dados com correlação intraclasse. Esses modelos permitem uma análise mais completa e detalhada dos dados longitudinais, levando em consideração a estrutura de correlação entre as observações repetidas, mas requerem cuidado na escolha da estrutura de correlação e na inclusão de efeitos aleatórios para evitar problemas de sobreajuste ou subajuste. Além disso, podem ser aplicados em estudos observacionais ou experimentais com outras estruturas complexas, como medidas repetidas em diferentes locais ou com diferentes tratamentos, como no caso de um estudo que avalia o impacto de diferentes doses de um medicamento em pacientes com hipertensão arterial (MONETTE, 2010).

Outra classe importante de modelos são os Modelos Aditivos Generalizados (GAMs), que surgiram na década de 1980 como uma extensão dos Modelos Lineares Generalizados (MLGs) para lidar com relações não-lineares entre a variável resposta e as variáveis explicativas. Esses modelos permitem a formação de modelos de maneira flexível, através da combinação aditiva de funções suaves, como funções *spline* (polinômios de baixo grau conectados em pontos de transição), *wavelet* e *Fourier*. Além disso, os GAMs podem ser estendidos para modelos de efeitos mistos (GAMMs) para tratar dados longitudinais ou espaciais.

Os Modelos Aditivos Generalizados com Parâmetros de Suavização Variáveis (GAMLSS) são uma evolução que permite modelar não apenas a média, mas também a variância, assimetria e curtose da variável resposta, através da especificação de uma distribuição adequada e dos parâmetros de suavização variáveis. As aplicações desses modelos estatísticos abrangem diversas áreas, desde ecologia e biologia até economia, epidemiologia e ciências sociais. Por exemplo, um GAM pode ser utilizado para avaliar a relação não-linear entre a poluição do ar e a mortalidade por doenças cardiovasculares, controlando por outras variáveis (HASTIE; TIBSHIRANI, 1990; RIGBY; STASINOPOULOS, 2005; DOMINICI et al., 2006; WOOD, 2017).

Outra classe de modelos que tem crescido em aplicações e que não pertencem à família exponencial padrão são os chamados modelos de regressão Beta. Eles se aplicam em casos onde a variável resposta está restrita ao intervalo unitário (0,1). Por exemplo, Ferrari e Cribari-Neto (2004) apresentaram o modelo de regressão beta e Bonat et al. (2015) apresentaram aplicações de sua versão mista no contexto de máxima verossimilhança. Ainda nessa classe, outra escolha possível e pouco explorada são os modelos de regressão simplex, primeiramente introduzidos por Kieschnick e McCullough (2003b) e com aplicações na forma mista por Qiu et al. (2008). Expandindo ainda mais a área dos modelos para dados restritos ao intervalo unitário, Bonat et al. (2012) apresentaram um *framework* cobrindo, além dos modelos beta e simplex, modelos gaussianos restritos e modelo Kumaraswamy. Eles testaram os efeitos destas abordagens em dados reais com diversas funções de ligação adequadas, como `logit`, `probit`, `cauchit`, `log-log` e complemento `log-log`.

Outro *framework* visando unificação e flexibilização do processo de modelagem foi introduzido por Bonat e Jørgensen (2016), onde os autores se basearam apenas nas suposições de primeiro e segundo momentos, integrando na estrutura de variância $\phi\mu^p(1-\mu)^p$ tanto o parâmetro de dispersão ϕ quanto o de potência p, e expandiram a modelagem para variáveis respostas simples, mistas, inflacionadas e múltiplas. O *framework* dos *Multiple response variables regression models* foi implementado no pacote R `mcglm` por Bonat (2018b). Através do `mcglm`, foi possível modelar diversos tipos de dados, inclusive respostas não normais múltiplas, estruturas temporais, es-

paço temporais, tipos mistos, medidas repetidas e dados longitudinais, uma vez que flexibiliza a introdução de estruturas de variância-covariância diversas no contexto de verossimilhança e quasi-verossimilhança.

Até o momento, realizou-se uma revisão abrangente das técnicas e *frameworks* de modelos estatísticos propostos para variados tipos de dados, como respostas contínuas, contagens, dados binários, respostas mistas e múltiplas. Contudo, há uma abordagem estatística chamada **Análise de Sobrevivência**, na qual as variáveis resposta estão relacionadas com o tempo e, em geral, produzem dados ditos **censurados**. À medida que o tempo passa, alguns indivíduos podem ser afetados por eventos diversos, como falência, desistência, mudança de localidade etc., e sair do estudo de coleta de dados. Para dados de sobrevivência, outras abordagens de modelagem foram propostas para lidar com esse tipo de resposta conjugada com o tempo de falha ou censura associado aos indivíduos. Entre os modelos estudados na Análise de Sobrevivência estão: Curvas de sobrevivência de Kaplan-Meier (KAPLAN; MEIER, 1958), (KLEINBAUM; KLEIN, 2012a); Modelo de risco proporcional de Cox (COX, 1972), (BORGAN et al., 1995); e Modelos paramétricos de sobrevivência (PETERSEN, 1986) e (KLEINBAUM; KLEIN, 2012b).

No contexto dos modelos de regressão, não há muitas referências sobre dados em escala, como por exemplo, dados em escala *Likert* ou Escalas de Classificação Numérica. Neste texto, será feito um apanhado sobre esses tipos de dados como embasamento para o objeto de estudo em questão.

A análise de dados em escala, incluindo escalas *Likert*, Escalas Numéricas de Classificação (*NRS*) e outras, é uma prática comum em pesquisas nas áreas de ciências sociais, psicologia, saúde e educação. Essas escalas permitem avaliar atitudes, sensações, percepções e comportamentos dos indivíduos de maneira quantitativa, possibilitando a aplicação de técnicas estatísticas (STREINER et al., 2015).

Entre as formas mais comuns de coleta de dados em pesquisas estão as escalas de *Likert*, que consistem em declarações ou perguntas acompanhadas de uma série de respostas possíveis, geralmente variando de "concordo fortemente"a "discordo fortemente"(NORMAN, 2010). Outra abordagem é a Escala Numérica de Classificação (*NRS*), na qual os participantes atribuem uma classificação numérica a um conceito ou item, geralmente em uma escala de 0 a 10.

Ao analisar dados em escala, é crucial determinar o nível de medida adequado, que pode ser ordinal, intervalar ou de taxa. As escalas de *Likert* e *NRS* são geralmente consideradas ordinais, pois os valores representam uma ordem ou classificação, mas a diferença entre os valores não é necessariamente uniforme (JAMIESON, 2004). No entanto, alguns argumentam que, sob certas condições, as escalas de *Likert* podem ser tratadas como dados de intervalo (CARIFIO; PERLA, 2007), ampliando assim as

possibilidades para modelagem estatística.

Em relação aos dados censurados, observações conhecidas apenas por estarem abaixo de um limite de detecção são chamadas de dados censurados à esquerda, enquanto observações conhecidas apenas por estarem acima de um limite de quantificação são chamadas de dados censurados à direita. Dados conhecidos por estarem entre dois limites são chamados de dados com censura intervalar. Dados censurados à direita são comumente encontrados em dados de sobrevivência (KLEIN; MOESCHBERGER, 2003) e analisados no contexto de máxima verossimilhança, conforme apresentado por Helsel et al. (2005).

Um dado particularmente relevante na área médica é a medida da dor, também referida como escores de dor. Disciplinas médicas, como a Anestesiologia, exploram os padrões físicos, neurológicos e psicológicos relacionados ao fenômeno da dor. Devido à sua natureza intrinsecamente subjetiva, uma variedade de instrumentos foi concebida para quantificar essa sensação, incluindo as Escalas Visuais e as Escalas de Classificação Numérica (ECN) de dor. Em inglês, a ECN é conhecida como *Numerical Rating Scale* (*NRS*) e é extensivamente empregada no contexto médico (BENZON et al., 2011).

A dor, sendo um fenômeno multifacetado, possui inevitavelmente um grau de erro de medida associado, tanto ao instrumento quanto à avaliação individual, que precisa ser levado em consideração na análise. Especificamente, a *NRS*, e em particular a *NRS*-11, pode resultar em dados que sejam ordinais, discretos ou até mesmo contínuos, dependendo da escala adotada. No que tange a dados ordinais ou discretos, é viável transformá-los para uma escala contínua, intervalar ou taxa ao redor de uma região específica, gerando assim um conjunto de medidas possíveis delineadas por subintervalos definidos. Esse procedimento habilita a implementação de métodos estatísticos pertinentes, inclusive modelos voltados para taxas e proporções. Assim, a análise de dados escalonados, como as escalas *Likert* e *NRS*, exige um entendimento detalhado sobre o tipo de medida em questão e as técnicas estatísticas ideais para dados ordinais, censurados e intervalares, assegurando análises acuradas e resultados robustos.

1.1 OBJETO DE ESTUDO

Conforme discutido anteriormente, muitas abordagens e ferramentas envolvendo modelos estatísticos estão disponíveis na literatura para tratar de diversos tipos de dados. No entanto, identifica-se uma notável lacuna na divulgação de estudos, dados e modelos estatísticos relacionados a respostas obtidas por meio de escalas, tais como a *NRS* e *Likert*. Portanto, o objetivo deste trabalho é revisitar e enriquecer a literatura voltada para essa categoria específica de dados.

A fim de preencher essa lacuna, propomos uma formulação adaptada do modelo de regressão beta para dados intervalares que apresentam censura, empregando o paradigma de máxima verossimilhança. Para assegurar um mapeamento beta coeso, sugerimos transformações de escala e estratégias para lidar com a incerteza associada à medida observada. Este procedimento transita a partir da escala discreta original para uma contínua e intervalar, compatível com o suporte da distribuição beta. Para avaliar as características do modelo sugerido, serão conduzidos estudos de simulação abrangendo uma variedade de cenários.

Como aplicação prática, analisaremos um conjunto de dados referentes a escores de dor de pacientes submetidos a cirurgias no joelho, coletados por meio da *NRS-11*. Com base nesse dataset, propondremos transformações de escala e conduziremos estudos descritivos.

1.2 JUSTIFICATIVA

Na literatura médica, observa-se a recorrência na geração de dados por meio de escalas numéricas. Tais dados são frequentemente coletados em contextos clínicos variados, incluindo avaliações de dor após procedimentos cirúrgicos ou durante tratamentos de pacientes com dor crônica. No entanto, apesar da relevância desses dados para a medicina, verifica-se uma carência na literatura estatística específica para sua análise apropriada.

Essa lacuna na literatura estatística sinaliza a imperativa necessidade de desenvolver e expandir abordagens analíticas voltadas para esses conjuntos de dados. Existe uma oportunidade significativa para a implementação de modelos estatísticos avançados, particularmente o modelo de regressão beta e o modelo quasi-beta. Ambos os modelos se destacam por sua capacidade de capturar detalhadamente as probabilidades associadas a cada subintervalo de medida.

O modelo quasi-beta, por sua natureza menos complexa, pode oferecer benefícios específicos na análise de dados de escala, permitindo uma flexibilidade adicional em comparação ao modelo de regressão beta tradicional.

Além disso, a presença de conjuntos de dados reais, combinada com a possibilidade de conduzir simulações, fornece um meio eficaz para avaliar a aderência e eficácia desses modelos a esses tipos específicos de dados. Assim, este estudo não busca apenas suprir uma deficiência identificada na literatura estatística, mas visa oferecer ferramentas analíticas robustas que podem beneficiar a pesquisa e prática médica.

1.3 OBJETIVOS

Propor modelos de regressão beta e quasi-beta em um *framework* desenhado para acomodar respostas oriundas de escalas, sejam elas *NRS*, *Likert* ou outros instrumentos geradores de dados beta mapeáveis.

1.3.1 Gerais

Aplicar o **modelo de regressão beta** e **quasi-beta** utilizando o método de máxima verossimilhança, fundamentando-se nos trabalhos de Gentleman e Geyer (1994), Klein e Moeschberger (2003), Helsel et al. (2005), Bogaerts et al. (2017), Ferrari e Cribari-Neto (2004), e Bonat e Jørgensen (2016), entre outros.

1.3.2 Específicos

- Propor, formular e validar modelos de regressão beta adaptados para dados de escala utilizando o princípio da máxima verossimilhança, e testar sua eficácia em *datasets* simulados e coletados empiricamente;
- Investigar detalhadamente as escalas e escores de dor e *Likert*, identificando suas aplicações predominantes, características intrínsecas e limitações metodológicas;
- Realizar uma revisão sistemática e crítica da literatura estatística que aborda dados com censura intervalar independente do tempo, focalizando especificamente em dados gerados por instrumentos como *NRS* e *Likert*;
- Estudar e implementar transformações estatisticamente robustas para dados de escala, visando otimizar sua adequação para modelagens estatísticas avançadas;
- Delinear e explorar uma metodologia baseada na máxima verossimilhança, adaptada especificamente para tratar dados intervalares que apresentem censura;
- Desenvolver e documentar um pacote na linguagem *R*, voltado para a parametrização e ajuste de modelos beta aplicados a dados de escala.

1.4 LIMITAÇÕES E DESAFIOS

Embora a literatura médica frequentemente utilize dados de escala, especialmente em relação às escalas de dor, o acesso a dados abertos deste tipo permanece notoriamente limitado. Neste estudo, apenas um conjunto de dados referente a esse tópico está disponível, cortesia do Hospital do Trabalhador em Curitiba-PR. Para futuras validações, simulações computacionais serão conduzidas, em alinhamento com o paradigma de máxima verossimilhança proposto.

É importante observar que a dor, devido à sua natureza complexa, é especialmente desafiadora para medir com precisão. Instrumentos de medição contemporâneos enfrentam limitações que resultam em algum grau de imprecisão. Similarmente, escalas *Likert*, quando usadas para coletar opiniões, também apresentam suas próprias restrições. A seguir, detalham-se algumas das limitações inerentes à análise de dados de escala.

- Para viabilizar a medição utilizando escalas, os dados gerados frequentemente se enquadram em categorias ordinais, discretas ou discretizadas. Essa categorização induz a um nível de censura que é intrínseco ao desenho da escala;
- A presença de censura intervalar, como observada na *NRS-11*, requer abordagens estatísticas especializadas. Modelos estatísticos convencionais muitas vezes são adequados apenas para dados contínuos ou discretos, o que torna a análise mais complexa;
- A implementação de modelos de regressão não tradicionais, como o modelo de regressão beta e o modelo quasi-beta, nesse contexto representa um desafio considerável. O ajuste desses modelos requer um esforço adicional para manter a integridade dos dados, minimizando a perda de informação e evitando a introdução de viés adicional.

1.5 ORGANIZAÇÃO DO TRABALHO

Esta dissertação está estruturada em seis capítulos. O conteúdo do primeiro capítulo já foi discutido anteriormente. No segundo capítulo, realiza-se uma revisão da literatura abrangendo a medicina da dor, a escala de dor *NRS*, dados censurados, o modelo de regressão beta e a teoria de verossimilhança adaptada para dados intervalares censurados. O terceiro capítulo apresenta o conjunto de dados relacionados à cirurgia de joelhos usados neste estudo, além da estratégia adotada para a simulação de dados visando testar as propriedades da verossimilhança em cenários de dados intervalares censurados. O quarto capítulo se dedica a detalhar a estimação do modelo beta considerando a censura intervalar. No quinto capítulo, são expostos os resultados, englobando tanto análises de dados reais quanto simulações. O sexto e último capítulo oferece as conclusões e considerações finais, bem como sugestões para investigações futuras no âmbito do tema abordado.

2 FUNDAMENTAÇÃO TEÓRICA

Esse capítulo aborda aspectos envolvendo a literatura médica da dor e a escala de dor NRS. além de abordar sobre escalas Likert. Trata ainda dos métodos e modelos estatísticos comumente utilizados em dados de escala de dor. Apresenta ainda a distribuição e o modelo de regressão beta, uma revisão da teoria de máxima verossimilhança e sua adaptação da dados onde há censura intervalar.

2.1 REVISÃO DE LITERATURA

Embora na área médica haja uma ampla utilização de escalas como a *NRS-11*, a análise estatística desses tipos de dados ainda não é comum na literatura, especialmente quando se aborda modelos estatísticos avançados. Dada a natureza da marcação nas escalas, as variáveis de resposta podem não seguir uma distribuição normal ou, em alguns casos, apresentar uma inflação de zeros. Isso implica a necessidade de transformações nos dados antes da aplicação de diversas técnicas de modelagem.

Nair e Diwan (2020) discutem o ceticismo existente sobre qual estatística é ideal para analisar escores de dor. Eles enfatizam a importância do acompanhamento de um analista ou bioestatístico desde o início do estudo e a utilização de testes de aderência, como Kolmogorov-Smirnov ou Shapiro-Wilk, para identificar corretamente a distribuição dos dados. Devido às diversas formas de coleta, os escores de dor podem ser categorizados como nominais, ordinais, intervalares, razão, discretos ou contínuos. Portanto, é crucial compreender a natureza do estudo, determinar corretamente a escala dos dados e investigar as transformações aplicáveis, garantindo o uso de técnicas estatísticas adequadas. Muitas vezes, os dados de escalas de dor são tratados como ordinais, dada a sua natureza. Por exemplo, em uma escala *NRS-11*, um paciente pode selecionar o nível de dor 6. Considerando a natureza intrínseca desse instrumento de medição, é razoável supor uma incerteza em torno dessa seleção. Em relação a esse aspecto, é uma prática comum em estatística tratar dados dessa natureza através de métodos estatísticos não paramétricos. No entanto, com transformações adequadas, é possível abordar a análise tanto pelo viés não paramétrico quanto pelo paramétrico, sem perder generalidade.

Nair e Diwan (2020) fazem referência ao emprego de várias técnicas listadas na TABELA 1 no *Korean Journal of Pain*. Esta tabela apresenta alguns métodos e testes estatísticos adequados para vários tipos de dados resultantes do uso de ferramentas como escalas visuais ou numéricas, por exemplo, a *NRS-11*. No entanto, é importante salientar que a tabela não abrange todos os métodos estatísticos aplicáveis a escores

de dor, mas destaca os principais já reconhecidos e utilizados na literatura.

TABELA 1 – REVISÃO DE MÉTODOS ESTATÍSTICOS

	Escala simples		Escala intervalar ou razão	
Nível de escala	Nominal	Ordinal	Com normalidade	Sem normalidade
Dois grupos independentes	Teste Chi-quadrado; Teste Exato de Fisher	Teste da Soma de rank de Wilcoxon	Teste t	Teste da Soma de rank de Wilcoxon;Teste de Mann-Whitney
Três ou mais grupos independentes	Teste Chi-quadrado; Teste Exato de Fisher	Teste de Kruskal-Wallis	ANOVA One-Way	Teste de Kruskal-Wallis
Duas amostras correlacionadas	Teste de McNemar	Teste do Rank de sinais de Wilcoxon	Teste t pareado	Teste da Soma de rank de Wilcoxon; Teste de Mann-Whitney
Três ou mais amostras correlacionadas	Teste Q de Cochran; Regressão logística com efeito misto	Regressão logística ordinal com efeito misto; Teste de Friedman	RMANOVA	Teste de Friedman; Regressão linear com efeito misto

FONTE: Adaptada de Nair e Diwan (2020)

A literatura atual demonstra diversas abordagens estatísticas sendo empregadas na análise de dados de escalas. No entanto, o objetivo deste trabalho é ampliar o escopo de modelos estatísticos aplicáveis a esse tipo de dado. A seleção de um método estatístico é influenciada por diversos fatores, incluindo a escala dos dados, a presença ou ausência da suposição de normalidade, o número de grupos comparativos e suas relações interdependentes.

Goulet et al. (2015) destacam algumas estratégias adotadas na análise de escores de dor em veteranos de guerra afetados por distúrbios musculoesqueléticos (MSD), obtidos por meio da escala NRS no momento do diagnóstico. O autor sugere a transformação dos escores para facilitar o uso de certas abordagens estatísticas, como a regressão via Método dos Mínimos Quadrados Ordinários (OLS). Uma alternativa seria a binarização dos dados para empregar a Regressão Logística. No entanto, é importante salientar que tais transformações, apesar de serem válidas, podem levar à perda de informações significativas. Portanto, elas devem ser realizadas com a supervisão e o discernimento tanto do especialista médico quanto do analista estatístico. Além das abordagens de Regressão OLS (i) e Logística (ii), os autores também exploraram os modelos de Regressão Poisson (iii) e Binomial Negativa (iv) para contagem, bem como o Modelo de Contagem Inflacionada em Zero (v) e o *Logit* Cumulativo (vi).

Adicionalmente, com a intenção de enriquecer ainda mais a gama de ferra-

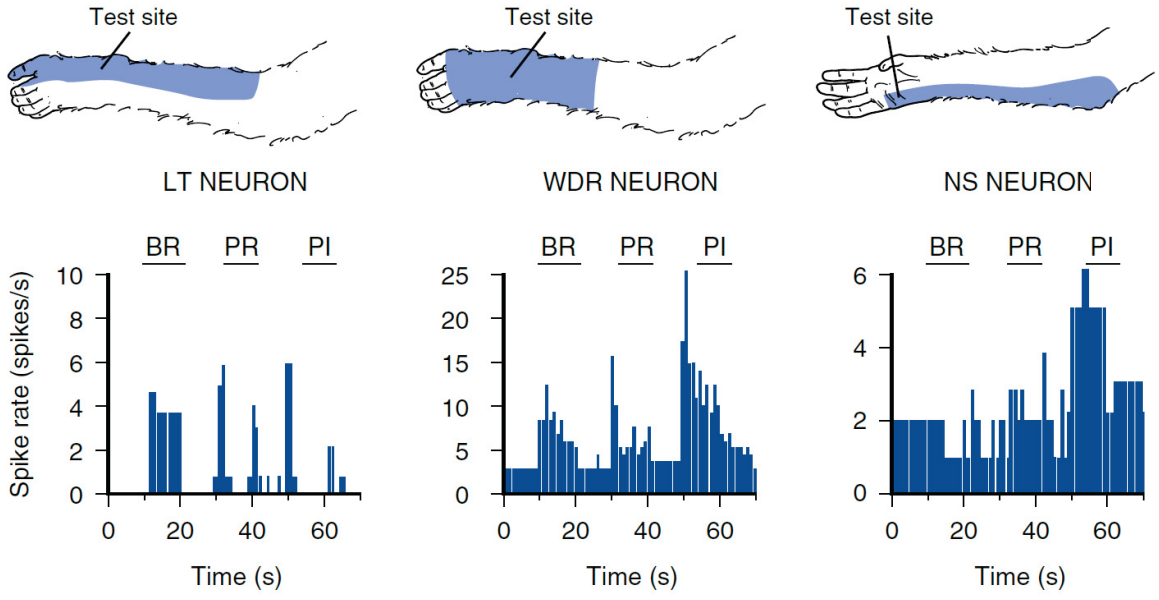
mentas disponíveis para análise de dados intervalares, Lopes (2023) introduzem o modelo de regressão beta. Este modelo é aplicado especificamente aos dados da escala *NRS-11* e é desenvolvido sob o paradigma da máxima verossimilhança.

2.2 MEDICINA DA DOR

O conceito de dor, embora frequentemente mencionado na literatura médica, como evidenciado em Raja et al. (2018), é surpreendentemente complexo e multifacetado. Mas, afinal, o que é dor? Essa questão, que à primeira vista pode parecer simples, carrega consigo uma riqueza de nuances e implicações. Por exemplo, o dicionário Michaelis oferece diversas definições para a dor, abordando tanto seus aspectos fisiológicos quanto os psicológicos. Neste trabalho, o foco será exclusivamente na compreensão da dor do ponto de vista médico, mais especificamente na mensuração da dor através de ferramentas como escalas.

Dependendo da disciplina — seja medicina ou psicologia — a definição de dor pode variar. Do ponto de vista médico, a dor é caracterizada como uma "sensação desagradável ou penosa, de intensidade variável, causada por uma anormalidade no organismo ou em parte dele, e é mediada pela estimulação de fibras nervosas que transmitem impulsos dolorosos ao cérebro; referindo-se ao sofrimento físico". Em termos mais simplificados, a dor pode ser interpretada como uma reação psicológica a algum dano tecidual em organismos vivos, detectável tanto por instrumentos específicos quanto pela expressão verbal ou sonora do indivíduo afetado.

FIGURA 1 – HISTOGRAMAS PARA FREQUÊNCIA DE RESPOSTAS DE NEURÔNIOS DO TRATO ESPINOTALÂMICO DE PRIMATAS.



FONTE: Obtido de Raja et al. (2018), p. 6.

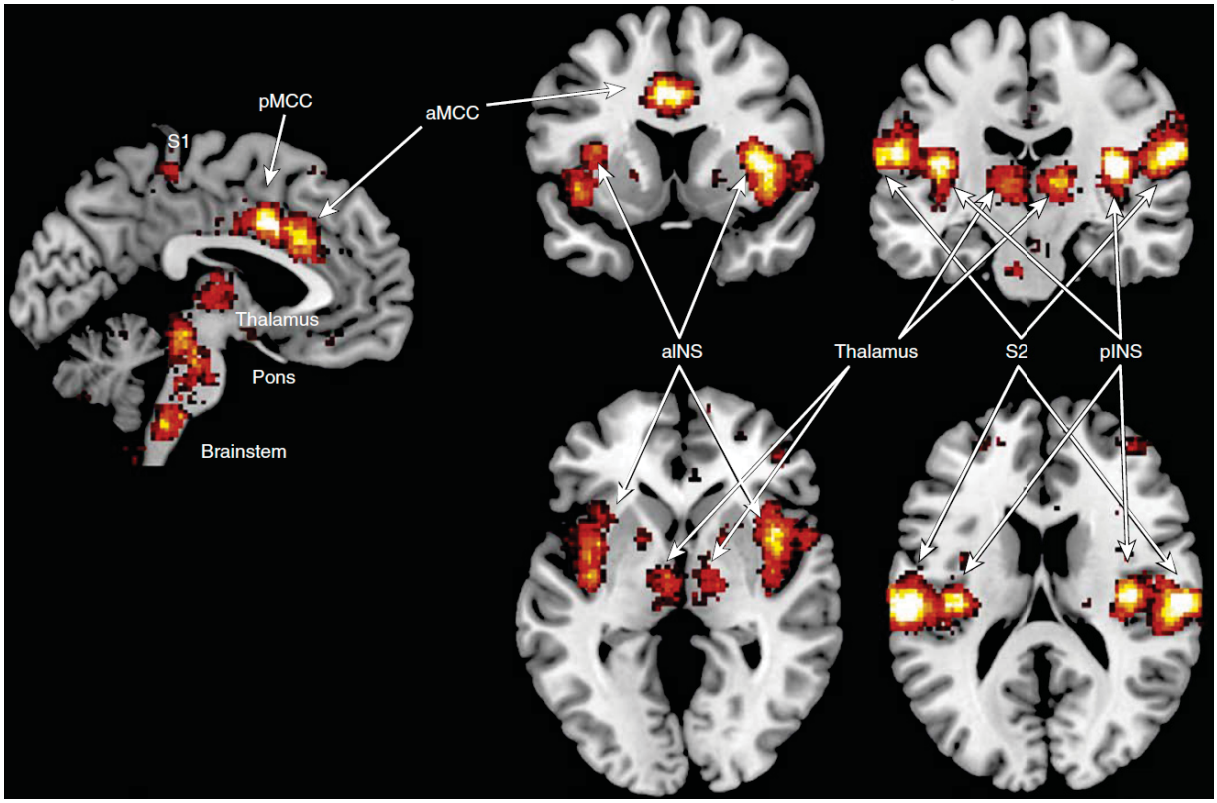
A FIGURA 1 apresenta um resumo gráfico dos resultados provenientes dos testes de resposta sensorial realizados em primatas. Os histogramas de frequência ilustram as reações dos neurônios do trato espinotalâmico de primatas, categorizados em três classes: baixo limiar (LT), ampla faixa dinâmica (WDR) e nociceptivo-específico (NS). A atividade destas células foi induzida mediante a aplicação de variados estímulos mecânicos de intensidades graduais em diferentes áreas do corpo, situadas dentro do campo receptivo de cada célula. Os momentos e os locais de cada estímulo são destacados por linhas e rótulos na parte superior dos histogramas.

Para o estímulo de escova suave (BR), utilizou-se uma escova delicada feita de pelo de camelo. Já a pressão inócua (PR) foi obtida através de um grande clipe arterial, enquanto que uma sensação nociva (PI) foi produzida por um clipe arterial menor. Observando o histograma central relativo à célula WDR, percebe-se que as reações são consistentes com a intensidade dos estímulos aplicados. Contrariamente, o neurônio NS, representado à direita, só demonstra reações significativas ao estímulo mais forte. Por outro lado, o neurônio LT, à esquerda, responde exclusivamente ao estímulo leve de escovação na pele (as respostas breves decorrentes da aplicação e remoção dos clipes arteriais se originam do contato inicial e subsequente desprendimento).

O cérebro, esse órgão magnificamente complexo, desempenha um papel central na percepção e processamento da dor. A FIGURA 2 destaca, através de vários cortes anatômicos, as áreas cerebrais ativas na sinalização de dor. Importante ressaltar que essa visualização foi obtida com base em estudos de imagem, mais especificamente uma meta-análise reversa via mapa estatístico de inferência de 420 estudos de fMRI com o termo “dor” criado em *Neurosynth.org*.

As estruturas que compõem o sistema somatossensorial, como o tronco encefálico, o diencefálico e regiões corticais, são partes fundamentais desse intrincado mecanismo. Quando se observa a neuroanatomia da dor, percebe-se que dois caminhos somatossensoriais distintos se dirigem ao tronco encefálico e ao diencéfalo. Vários axônios e colaterais axônicos que se originam dos neurônios de projeção espinhal se desprendem e terminam em diversos núcleos no tronco encefálico e no mesencéfalo. O conceito de "matriz da dor"engloba áreas como os córtices somatossensoriais primário e secundário, a ínsula, o córtex cingulado anterior, o córtex pré-frontal, e vários núcleos talâmicos. Estas regiões demonstraram consistentemente sua atividade em estudos de imagem, em contextos de dor tanto aguda quanto crônica. Essa cartografia cerebral, revelando as áreas responsáveis pela sinalização da dor, evidencia a multidimensionalidade e a intrincada natureza dos mecanismos de dor. Mas a dor não é meramente um fenômeno biológico. Aspectos psicológicos, como resiliência e experiências anteriores com a dor, influenciam profundamente sua percepção. Dada essa complexidade, a mensuração da dor torna-se um desafio.

FIGURA 2 – REGIÕES CEREBRAIS ENVOLVIDAS NA SINALIZAÇÃO DA DOR.



FONTE: De Raja et al. (2018), p. 8.

Em que pese ser uma experiência pessoal e intransferível, a dor é algo que a medicina busca avaliar e quantificar. Embora não exista um instrumento definitivo para medir a dor, ferramentas como as NRS têm mostrado ser métricas confiáveis. Na avaliação clínica da dor, escalas categóricas baseadas em NRSs são frequentemente preferidas em relação às escalas visuais analógicas. Vale ressaltar que as NRSs são aplicáveis a um vasto público, incluindo adultos, adolescentes, idosos e até mesmo crianças com deficiências cognitivas (RAJA et al., 2018).

2.3 *NUMERIC RATE SCALES (NRS)*

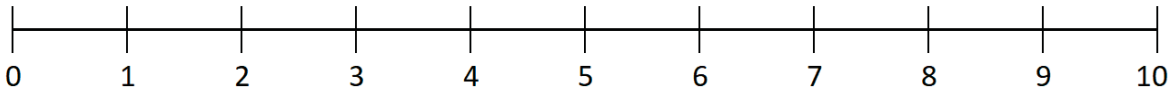
A implementação da Escala de Classificação Numérica (ECN) ou *Numeric Rate Scales (NRS)* na medicina representa um avanço significativo na avaliação da dor. A simplicidade e a objetividade dessa técnica permitiram que profissionais de saúde quantificassem um sintoma que é intrinsecamente subjetivo. Esta escala, embora aparentemente básica, fornece um *insight* valioso sobre a experiência de dor de um paciente.

Dada a variabilidade na percepção da dor entre os indivíduos, a ECN oferece uma ferramenta padronizada para medir a intensidade da dor. Ao pedir ao paciente para localizar sua dor em uma escala com intervalos fixos, os profissionais de saúde

podem avaliar a eficácia de tratamentos, o progresso da recuperação ou a gravidade da condição de um paciente.

Um aspecto crucial da *NRS* é a sua universalidade. Seja qual for o range escolhido – seja de 0 a 10, 0 a 20 ou 0 a 100 – o princípio fundamental é o mesmo: proporcionar uma métrica quantitativa para uma sensação que, até então, era difícil de descrever e quantificar. A clareza proporcionada por esta técnica beneficia não apenas os pacientes, que agora têm uma maneira de comunicar sua dor, mas também os profissionais de saúde, que podem fazer avaliações e ajustes de tratamento com base em dados mais tangíveis.

FIGURA 3 – EXEMPLO DE *NRS-11*



FONTE: Produzida pelo autor com base em Raja et al. (2018), p. 40.
NOTA: Quanto maior o número marcado na escala maior será a expressão da dor experimentada e informada pelo paciente.

A metodologia por trás da *NRS-11*, como ilustrado na FIGURA 3, é meticulosamente planejada para garantir uma compreensão clara e objetiva por parte do paciente. A simplicidade e intuitividade dessa escala foram pensadas para minimizar qualquer confusão e garantir que a marcação escolhida pelo paciente seja tão próxima quanto possível de sua experiência real.

Os marcadores, representados por números de 0 a 10, atuam como referências claras para o paciente. O intervalo de cada número representa uma graduação específica da sensação de dor, com uma transição percebida entre os números. Isto é particularmente útil porque a dor não é um fenômeno binário; ela ocorre em um espectro.

O uso da *NRS-11* e outras escalas semelhantes oferece uma abordagem quantitativa para a avaliação da dor. Mas é importante entender que, enquanto os números fornecem uma representação tangível, eles ainda são baseados em uma percepção subjetiva do paciente. A experiência de "5"em uma escala de dor pode não ser a mesma para dois pacientes diferentes.

A decisão de marcar "7", por exemplo, não só implica que a dor é bastante intensa, mas também que é percebida como mais intensa do que "6"e menos do que "8". A região de probabilidade mencionada é vital na análise estatística posterior, pois reconhece a natureza subjetiva e variável da dor.

A finalidade última da *NRS-11* é fornecer aos profissionais de saúde uma ferramenta que traduza a experiência individual e intangível da dor em dados quantificáveis

e comparáveis, permitindo uma melhor gestão da dor e um tratamento mais eficaz. A natureza subjetiva da dor torna essas ferramentas imperativas, já que elas estabelecem um padrão pelo qual as experiências dos pacientes podem ser comunicadas e compreendidas.

2.4 DADOS DE ESCALA E NÍVEIS DE MEDIDA

A análise de dados de escala, tais como as escalas *Likert* e *NRS*, é frequentemente utilizada em pesquisas nas áreas de ciências sociais, psicologia e educação. Estas escalas fornecem uma ferramenta quantitativa para avaliar atitudes, percepções e comportamentos dos indivíduos, facilitando assim a aplicação de técnicas estatísticas (STREINER et al., 2015).

Dentre as diversas ferramentas de coleta de dados, as escalas de *Likert* são amplamente adotadas em pesquisas. Elas se baseiam em afirmações ou questões e oferecem uma variedade de respostas, que vão, comumente, de "concordo fortemente"a "discordo fortemente"(NORMAN, 2010). Já a *NRS* proporciona uma abordagem distinta, na qual os participantes atribuem uma classificação numérica, comumente em uma escala de 0 a 10, a um determinado conceito ou item.

Na análise de dados dessa natureza, é imperativo estabelecer o nível de medida apropriado, seja ele nominal, ordinal, intervalar ou de razão. Comumente, as escalas de *Likert* e *NRS* são classificadas como ordinais, visto que os valores indicam uma sequência ou classificação. Contudo, a distância entre esses valores pode não ser uniforme (JAMIESON, 2004). Apesar disso, há quem defenda que, em determinadas circunstâncias, as escalas de *Likert* podem ser interpretadas como dados de intervalo (CARIFIO; PERLA, 2007).

As estatísticas descritivas são frequentemente usadas para resumir e descrever os dados coletados em escalas. Algumas estatísticas descritivas comuns incluem média, mediana, moda, amplitude, variância e desvio padrão. Para dados ordinais, é mais apropriado usar a mediana e a moda, pois a média pode ser enganosa devido à natureza não equidistante das categorias (SULLIVAN; ARTINO, 2013).

Para visualizar a distribuição dos dados e identificar possíveis tendências, padrões e outliers, gráficos como histogramas, gráficos de barras e gráficos de caixa podem ser empregados (STREINER et al., 2015). A escolha adequada de estatísticas descritivas e representações gráficas é essencial para transmitir de maneira clara e precisa as informações dos dados de escala.

No contexto de dados de escala, é habitual a realização de análises inferenciais, seja para testar hipóteses ou para comparar grupos distintos. Há uma gama de técnicas estatísticas aplicáveis a dados ordinais, englobando testes não paramétricos

e modelos de regressão ordinal (STREINER et al., 2015). Algumas das abordagens mais frequentes incluem:

- **Teste de Mann-Whitney-Wilcoxon:** Conhecido também como teste U de Mann-Whitney, serve para comparar duas amostras independentes. Funciona como uma alternativa não paramétrica ao teste t de Student para dados ordinais (WINTER; DODOU, 2010).
- **Teste de Kruskal-Wallis:** Representa uma extensão do teste de Mann-Whitney-Wilcoxon, voltado para a comparação de três ou mais conjuntos independentes. Age como uma substituta não paramétrica à ANOVA de um fator (STREINER et al., 2015).
- **Teste de Friedman:** Este teste não paramétrico é direcionado para a comparação de três ou mais conjuntos, surgindo como uma alternativa ao teste ANOVA de medidas repetidas (STREINER et al., 2015).
- **Regressão ordinal:** Consiste em uma análise de regressão direcionada à modelagem da relação entre uma variável ordinal e uma ou mais variáveis independentes. Esta técnica pressupõe uma ordem entre as categorias, mas não necessariamente uma distância uniforme entre elas (STREINER et al., 2015).

Durante a execução de análises inferenciais, é primordial assegurar que as premissas dos testes estatísticos estão sendo cumpridas e, se necessário, realizar ajustes nas análises. Por exemplo, os testes não paramétricos descritos anteriormente pressupõem que as amostras são independentes e que os dados configuram-se como ordinais (WINTER; DODOU, 2010).

Na análise de dados de escala, é essencial reconhecer certas questões recorrentes e os equívocos relacionados:

- **Tratamento de dados ordinais como dados de intervalo:** Apesar da argumentação de que, em determinados contextos, escalas de *Likert* podem ser tratadas como dados de intervalo, é prudente ser cauteloso ao empregar técnicas estatísticas que pressupõem distâncias equidistantes entre categorias (CARIFIO; PERLA, 2007).
- **Violação das premissas dos testes estatísticos:** Antes da aplicação de testes estatísticos, torna-se imperativo assegurar que as premissas intrínsecas a estes testes são satisfeitas. A título de ilustração, o teste t de *Student* assume homocedasticidade e normalidade dos dados, enquanto testes não paramétricos pressupõem amostras independentes (WINTER; DODOU, 2010).

- **Interpretação imprecisa de resultados:** Na leitura dos desfechos das análises inferenciais, destaca-se a necessidade de distinguir entre significância estatística e relevância prática. Uma descoberta com significância estatística não garante sua pertinência em uma perspectiva prática (SULLIVAN; ARTINO, 2013).
- **Realização de múltiplos testes:** A execução de diversas análises estatísticas em uma mesma coleta de dados pode ampliar o risco de erros do Tipo I (falsos positivos). A correção de Bonferroni surge como uma estratégia habitual para ajustar a taxa de erro quando múltiplos testes são realizados (STREINER et al., 2015).

2.5 DADO INTERVALAR CENSURADO

Dados com algum tipo de censura intervalar são comuns em análise de sobrevivência e as NRSs são apenas um caso em que esse efeito acontece. De maneira mais formal uma variável aleatória Y resultante de um evento pode estar sujeita a censura em algum momento e isso traz consigo algum nível de imprecisão em sua medida que não deve ser negligenciada. A TABELA 2 ilustra os tipos mais comuns de censura em

TABELA 2 – TIPOS DE CENSURA E FORMA ESPERADA PARA A CONTRIBUIÇÃO EM TERMOS DA FUNÇÃO DE VEROSSIMILHANÇA.

Observação	Intervalo $\lfloor l_i, u_i \rfloor$	Tipo	Contribuição
Sem censura (exata)	y_i	$\delta = 0$	$f(y_i)$
Censura à direita	$y_i \in (l_i, \infty)$	$\delta = 1$	$F(\infty) - F(l_i)$
Censura à esquerda	$y_i \in (0, u_i)$	$\delta = 2$	$F(u_i) - F(0)$
Censura intervalar	$y_i \in (l_i, u_i)$	$\delta = 3$	$F(u_i) - F(l_i)$

FONTE: Adaptado de Bogaerts et al. (2017), p. 5. e Colosimo e Giolo (2006), p. 255

dados. Assim, pode-se representar o intervalo $I = \lfloor l_i, u_i \rfloor$ com l_i sendo o limite inferior e u_i o limite superior do intervalo para uma medida qualquer para um dado indivíduo. No paradigma de verossimilhança, o tipo de censura define qual o tipo de função de probabilidade deve ser aplicada. Lembrando da teoria das probabilidades que

$$\begin{aligned} F_Y(y_i) &= F_Y(Y = u_i) - F_Y(Y = l_i) \\ &= \int_{-\infty}^{u_i} f_Y(Y = u_i) du_i - \int_{-\infty}^{l_i} f_Y(Y = l_i) dl_i \end{aligned} \tag{2.1}$$

se $f_Y(Y = u_i)$ é contínua e

$$\begin{aligned} F_Y(Y = y_i) &= F_Y(y_i = u_i) - F_Y(y_i = l_i) \\ &= \sum_{-\infty}^{u_i} P_Y(Y \leq u_i) - \sum_{-\infty}^{l_i} P_Y(Y \leq l_i) \end{aligned} \tag{2.2}$$

se discreta. Além disso, vale notar na 2.1 e 2.2 se $f_Y(y_i)$ é uma distribuição de probabilidade definida em um intervalo $I = \lfloor l_i, u_i \rfloor$, tem-se que $F_Y(l_i = 0) = 0$ e $F_Y(u_i = \infty) = 1$. Nessa notação a construção da função de verossimilhança é atendida para qualquer δ escolhido na TABELA 2.

É relevante ressaltar uma nuance que pode não ser imediatamente perceptível: a censura, na maioria dos contextos, relaciona-se com o tempo, determinando o desfecho da variável resposta. Em estudos voltados à análise de sobrevivência, a variável resposta, frequentemente binária, vincula-se ao tempo, sendo modelada em conjunto. Neste estudo, a abordagem é similar, com a exceção de que a censura aplicada ao escore de dor é independente do tempo. Em outras palavras, a medida obtida na escala serve como variável resposta, e a censura vincula-se ao valor máximo da escala, que corresponde ao ponto indicado pelo paciente e ao ponto imediatamente anterior. Essa abordagem estabelece um intervalo de probabilidade. Como exemplificado, se o paciente indicou 7 na escala, presume-se que a intensidade da dor esteja entre 6 e 7 na escala *NRS-11*. Ao proceder dessa maneira, alcança-se uma resposta contínua e intervalar censurada, em contraste com uma medida discreta e precisa.

2.6 MUDANÇA DE ESCALA INTERVALAR

Em análise de dados, é comum o emprego de transformação de dados para tratar alguma particularidade. Em seu livro, Stuart (1984) dedica o capítulo quatro para discutir transformações de dados. O autor lista algumas razões para o emprego de transformação, dentre elas: melhorar a interpretação, promover a simetria, estabilizar a dispersão dos dados, aprimorar a relação entre variáveis, e controlar o range dos dados. Estas transformações podem ser aplicadas tanto em variáveis resposta quanto em variáveis explicativas.

Sem perda de generalidade em estatística, a transformação de dados é a aplicação de uma função matemática determinística a cada ponto de um conjunto de dados. Assim, cada valor de dados d_i é substituído pelo valor transformado $d_i^* = f(d_i)$, onde f é uma função com uma inversa definida. As transformações são frequentemente aplicadas para que os dados atendam mais diretamente às suposições de um procedimento estatístico inferencial a ser utilizado, ou para melhorar a interpretabilidade ou visualização gráfica. A função f deve ter uma inversa para permitir a reversão à escala original, auxiliando na interpretação após a aplicação dos métodos estatísticos.

No contexto das *NRS* e com foco em modelos de regressão como o beta, cuja variável resposta está limitada ao intervalo $(0, 1)$, é possível utilizar uma transformação de range para adaptar a escala de números entre zero e dez, vinte ou cem para o intervalo $(0, 1)$. No entanto, é vital considerar os limites inferior e superior, denominado **efeito de borda**. Na distribuição beta, os valores zero e um não são inclusos, portanto, o

analista deve escolher uma transformação que minimize ou evite a perda de informação devido ao efeito de borda. Uma opção que respeita esse efeito é a transformação

$$y^* = \frac{\frac{y(n-1)}{R} + \frac{1}{2}}{n} \tag{2.3}$$

A inversa para a escala original é

$$y = \frac{R(1 - 2ny^*)}{2(1-n)} \tag{2.4}$$

Onde $R = y_{max} - y_{min}$ representa o range dos dados e n o total de observações do vetor y. É relevante mencionar que na equação 2.3, $\lim_{n\to\infty} y^* = y/R$ e que $\lim_{n\to\infty} y = y^*R$ e na equação 2.4. Além disso, para os dados de *NRS* o valor mínimo é zero e os máximos possíveis são 10, 20 ou 100. Quanto mais extenso for o conjunto de dados, menor será o impacto no efeito de borda. Como a função possui uma inversa e o total de observações é finito, o efeito de borda é abordado sem perda de informação.

Outra alternativa de transformação para a escala $(0, 1)$, que incorpora as bordas, é conhecida como transformação *min-max*. Ela é dada pela equação

$$y^* = \frac{y - y_{min}}{y_{max} - y_{min}}, \tag{2.5}$$

cuja unção inversa para recuperar os dados na escala original é representada por

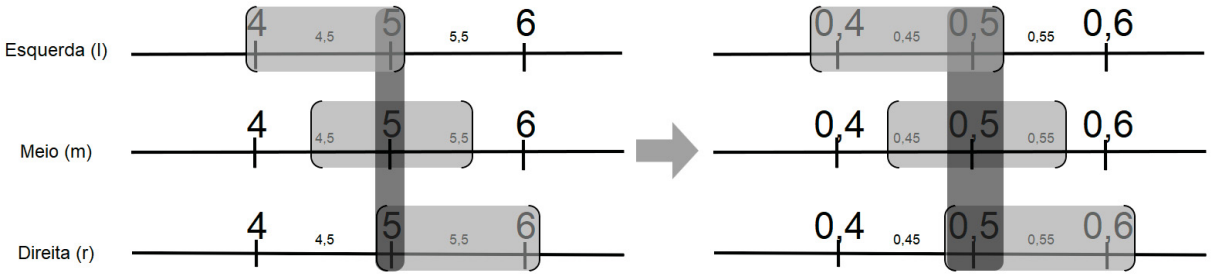
$$y = y^*(y_{max} - y_{min}) + y_{min} \tag{2.6}$$

Na formulação da 2.5 e 2.6, y_{max} e y_{min} representam, respectivamente, os valores máximos e mínimos registrados. Ao aplicar a transformação *min-max* com a intenção de adaptar os dados a modelos de regressão confinados ao intervalo unitário, é essencial realizar um ajuste adicional nos valores extremos. Isso é necessário porque, intrinsecamente, tais modelos não operam nos limites do intervalo. Considerando uma *NRS-11*, por exemplo, onde o valor máximo registrado é 10 e o mínimo é 0, na escala *min-max*, tais valores seriam traduzidos como 0.0 e 1.0, respectivamente. Dessa forma, essa transformação pode ser simplesmente interpretada como uma divisão dos valores por 10. Para abordar os valores de borda, uma estratégia é determinar um grau de precisão para a medição, tal como $\zeta = 0.0001$, e então adicionar ou subtrair esse valor dos extremos. Assim, para uma entrada $y_i = 0$, o resultado transformado seria $y_i = \zeta$. Já para $y_i = 1$, o valor transformado seria $y_i = 1 - \zeta$.

2.7 MAPEAMENTO DE INTERVALOS PARA BETA

O processo de mapeamento beta, seja partindo da escala de dor ou de outras escalas, envolve a divisão do intervalo principal em porções menores. A FIGURA 4

FIGURA 4 – TRECHO SE UMA *NRS-11* EXEMPLIFICANDO O TRATAMENTO DOS SUBINTERVALOS DE INCERTEZA



FONTE: Produzida pelo autor
NOTA: Quanto maior o número marcado na escala maior será a expressão da dor experimentada e informada pelo paciente.

ilustra o mapeamento de intervalos utilizando três diferentes abordagens para adequar à compatibilidade com o modelo de regressão beta, que é restrito ao intervalo unitário.

Conforme apresentado na FIGURA 4, no contexto de uma *NRS-11* e escolhendo arbitrariamente o ponto 5, propõem-se três abordagens para o mapeamento de intervalos beta:

- Caso 01 - Meio (m): Esta abordagem ilustra a quantificação da incerteza do instrumento, centralizando em torno do valor 5. Aqui, somamos e subtraímos meio ponto, tanto à direita quanto à esquerda do 5. Assim, define-se um subintervalo centrado no valor registrado, onde o limite inferior é $l_i = 5 - 0,5 = 4,5$ e o superior é $l_s = 5 + 0,5 = 5,5$. Portanto, $y_{[i=5]} = (l_i, l_s) = (4,5; 5,5)$.
- Caso 02 - Direita (r): Nesta abordagem, adiciona-se uma unidade ao valor 5 na direção crescente, indicando que a medida informada pelo paciente pode ter sido subestimada. Assim, o limite inferior do intervalo é $l_i = 5$ e o superior é $l_s = 5 + 1 = 6$. Portanto, $y_{[i=5]} = (l_i, l_s) = (5; 6)$.
- Caso 03 - Esquerda (l): Considera-se, nesta abordagem, uma possível superestimação do valor informado pelo paciente. Subtrai-se uma unidade da medida registrada, resultando em um limite inferior de $l_i = 5 - 1 = 4$ e um limite superior de $l_s = 5$. Assim, $y_{[i=5]} = (l_i, l_s) = (4; 5)$.

Após o tratamento dos intervalos, pode-se aplicar os resultados da seção seção 2.6 para finalizar o mapeamento dos intervalos, tornando as observações compatíveis com a distribuição beta. Em estudos de simulação e aplicação, as três abordagens serão testadas.

2.8 DISTRIBUIÇÃO BETA

A distribuição beta tem sido utilizada frequentemente como opção para modelagem de dados restritos ao intervalo unitário. Na literatura é comum encontrar a distribuição beta com seus parâmetros de forma dados por α e β, contudo neste trabalho foi adotada a notação utilizada por Johnson et al. (1995), também a utilizada por Ferrari e Cribari-Neto (2004).

Dito isso, se Y é uma variável aleatória que segue uma distribuição beta com parâmetros de forma p e q e expressa por $Y \sim \mathbf{B}(p, q)$, então sua distribuição de probabilidade é dada por:

$$f(y, p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}; y \in (0,1); p, q > 0 \tag{2.7}$$

A média e a variância de Y são dadas por:

$$\begin{aligned} \mathrm{E}[Y] &= \frac{p}{p+q} \\ \mathrm{V}[Y] &= \frac{pq}{(p+q)^2(p+q+1)} \end{aligned} \tag{2.8}$$

A Equação 2.7 representa a principal forma da distribuição beta encontrada na literatura. No entanto, com essa notação, a definição de modelos de regressão torna-se complicada, visto que a média e a variância da distribuição beta, conforme 2.8, dependem dos dois parâmetros de forma da distribuição. Diante disso, Ferrari e Cribari-Neto (2004) propuseram uma nova parametrização da distribuição beta, adequada para modelos de regressão. Esses autores perceberam que, ao fazer uma mudança de variável, dada por $\mu = \frac{p}{p+q}$ e $\phi = p+q$; $p = \mu\phi$ e $q = (1-\mu)\phi$, seria possível obter uma versão mais flexível da distribuição beta para modelagem via regressão, sem perda das propriedades da distribuição, mantendo o mesmo suporte de y. Após algumas manipulações algébricas, chega-se à seguinte forma da distribuição beta, denominada aqui de parametrização 1:

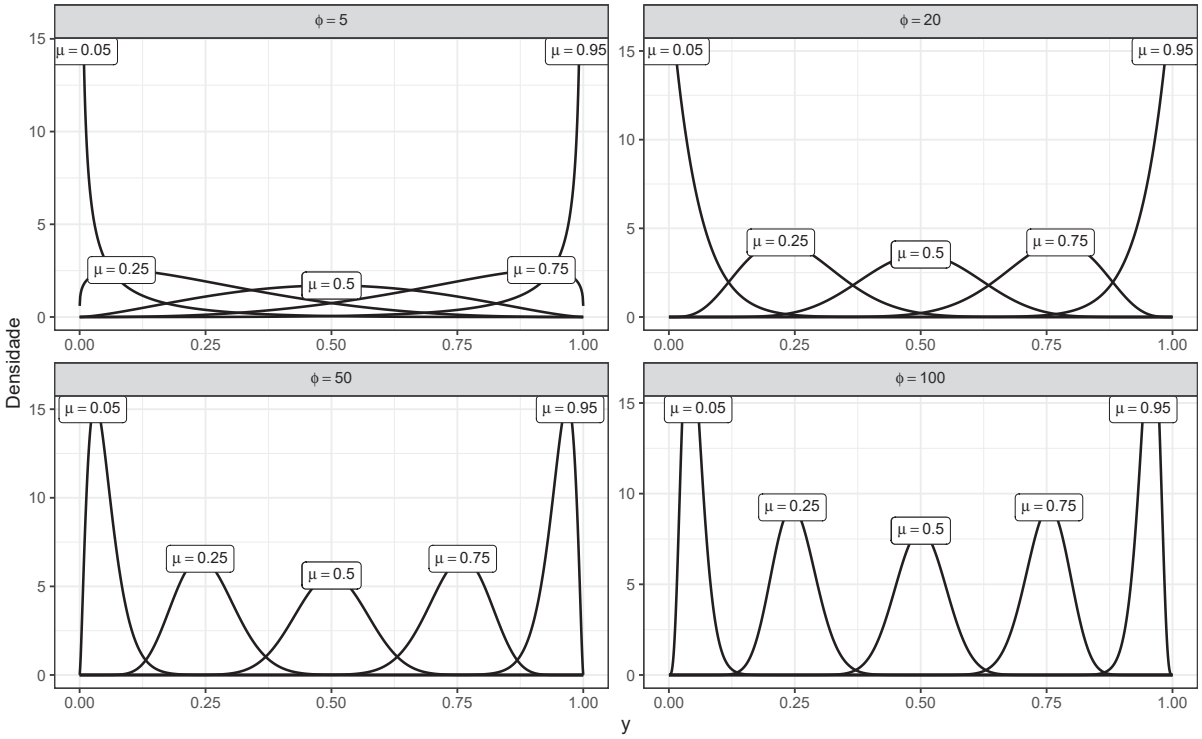
$$f(y, \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}; y, \mu \in (0,1); \phi > 0 \tag{2.9}$$

Cuja média e variância são dadas por:

$$\begin{aligned} \mathrm{E}[Y] &= \mu \\ \mathrm{V}[Y] &= \frac{V(\mu)}{1+\phi} = \frac{\mu(1-\mu)}{1+\phi}. \end{aligned} \tag{2.10}$$

Na parametrização 1 a média de Y depende exclusivamente de μ. O parâmetro ϕ atua como um parâmetro de precisão e é sempre positivo. É relevante mencionar que, quando $\lim_{\phi \to 0} \mathrm{V}[Y] = \mu(1-\mu)$, isso representa a variância de uma distribuição binomial. Além disso, quando $\lim_{\phi \to \infty} \mathrm{V}[Y] = 0$, sugere-se que, quanto maior a precisão, menor será a variância de Y.

FIGURA 5 – PERFIS DISTRIBUIÇÃO BETA PARAMETRIZAÇÃO 1 (μ, ϕ).



FONTE: O autor

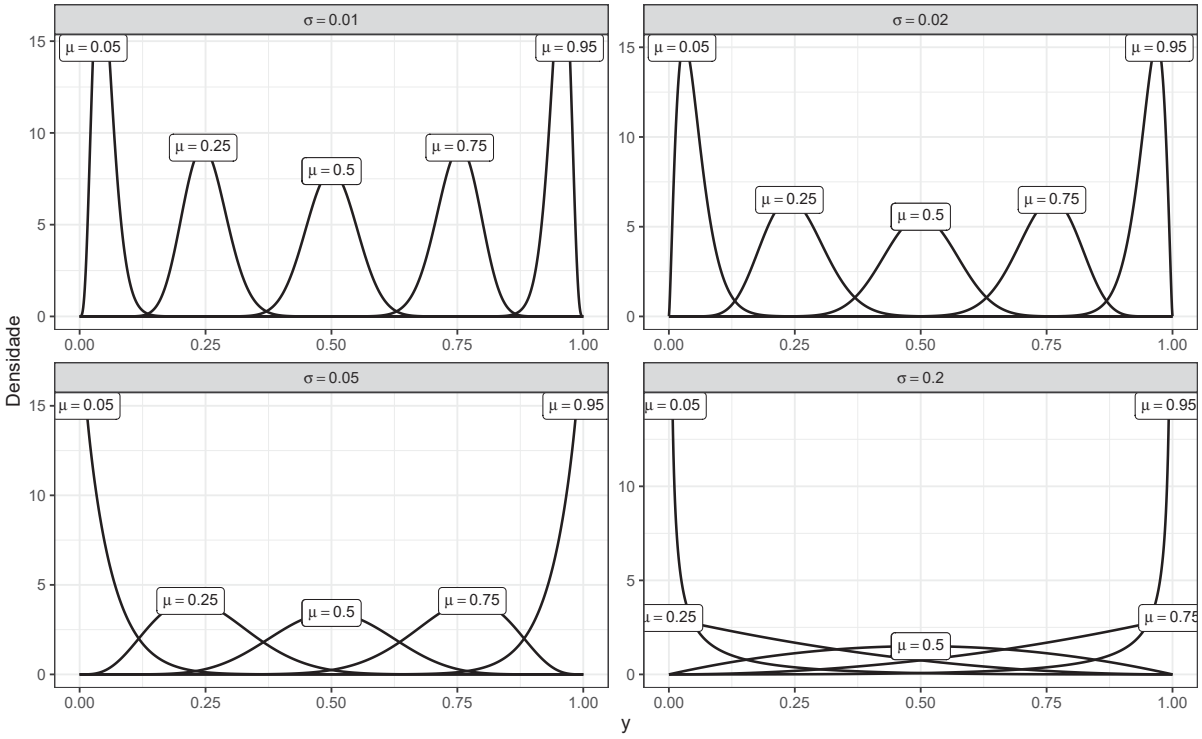
A Figura 5 mostra os perfis de cobertura das curvas originadas pela distribuição beta na parametrização 1 para variados valores de μ e ϕ. Identifica-se que valores reduzidos de ϕ resultam em dados mais dispersos e distribuídos por todo o domínio de Y. Em contrapartida, valores mais elevados de ϕ centralizam a distribuição ao redor de μ. Desta forma:

- Quando μ é reduzido e ϕ também é baixo, observa-se uma concentração abaixo da mediana de Y;
- Quando μ é elevado e ϕ é baixo, a concentração é vista acima da mediana de Y;
- Se μ é reduzido e ϕ é elevado, há uma marcante concentração abaixo da mediana de Y;
- Se μ é elevado e ϕ também é, percebe-se uma intensa concentração acima da mediana de Y;
- Se μ situa-se na mediana, o valor de ϕ determina a variabilidade de Y em torno de μ.

Consequentemente, é evidente que, sem considerar as magnitudes de μ e ϕ, a distribuição beta tem uma ampla adaptabilidade na cobertura do suporte para qualquer

valor de $Y = y_i$. Logo, essa distribuição é fortemente aconselhada para a modelagem de dados limitados ou passíveis de conversão para o intervalo unitário.

FIGURA 6 – PERFIS DISTRIBUIÇÃO BETA PARAMETRIZAÇÃO 2 (μ, σ).



FONTE: O autor

Outra reparametrização conveniente para modelos de regressão beta com dispersão fixa ou variável é apresentada por Mariano Bayer (2011) onde o autor aplica a seguinte transformação $\mu = p/(p+q)$ e $\sigma = 1/(1+\phi)$ de modo a gerar uma nova distribuição beta com a seguinte aparência:

$$f(y,\mu,\sigma) = \frac{\Gamma(\frac{1-\sigma}{\sigma})}{\Gamma(\mu\frac{1-\sigma}{\sigma})\Gamma((1-\mu)\frac{1-\sigma}{\sigma})} y^{\mu(\frac{1-\sigma}{\sigma})-1}(1-y)^{(1-\mu)(\frac{1-\sigma}{\sigma})-1}; y,\mu,\sigma \in (0,1). \quad (2.11)$$

Cuja média e variância são dadas por:

$$\begin{aligned} \mathrm{E}[Y] &= \mu \\ \mathrm{V}[Y] &= V(\mu)\sigma = \mu(1-\mu)\sigma. \end{aligned} \quad (2.12)$$

Note que na parametrização 2 a média de Y depende apenas de μ assim como na primeira reparametrização. O parâmetro σ, por sua vez faz o papel de parâmetro de dispersão e está restrito ao intervalo unitário tal como μ. Nessa reparametrização tem-se o benefício de não ter um parâmetro no denominador de uma fração, o que poderia trazer problemas numerico-computacionais no processo de estimação dos parâmetros.

Conforme perfis de comportamento de μ e σ na Figura 6 nota-se um padrão de inversão em relação de à parametrização 1, pois para valores maiores de σ há maior dispersão de Y e para valores pequenos há concentração, conforme esperado, uma vez que σ assume o efeito de dispersão e não mais de precisão como ϕ.

2.9 FUNÇÃO ACUMULADA DA DISTRIBUIÇÃO BETA

Para análise dados censurados a aplicação da distribuição acumulada é mandatória. Sendo assim, a seguir é apresentada a distribuição acumulada da beta na parametrização 2. A expressão fechada para a acumulada da beta depende de duas funções especiais: a função **beta** dada por:

$$\begin{aligned} \mathrm{B}(p,q) &= \int_0^1 t^{p-1}(1-t)^{q-1}\,dt \\ &= \frac{\Gamma(p)\,\Gamma(q)}{\Gamma(p+q)} \end{aligned} \tag{2.13}$$

e a função **beta incompleta** dada por:

$$\mathrm{B}(y;\,p,q) = \int_0^y t^{p-1}\,(1-t)^{q-1}\,dt. \tag{2.14}$$

Com isso a acumulada é expressa por:

$$\begin{aligned} F(y,p,q) &= \frac{\int_0^y t^{p-1}(1-t)^{q-1}dt}{B(p,q)} \\ &= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\int_0^y t^{p-1}(1-t)^{q-1}dt, \quad 0 \leq y \leq 1; p,q > 0. \end{aligned} \tag{2.15}$$

Sem perda de generalidade, a acumulada pode ser expressa também como a integral da distribuição de probabilidade beta a partir do limite inferior $y = 0$ até algum ponto $t = y$ do suporte de Y. Aplicando a parametrização da Equação 2.9 na Equação 2.15 e simplificando

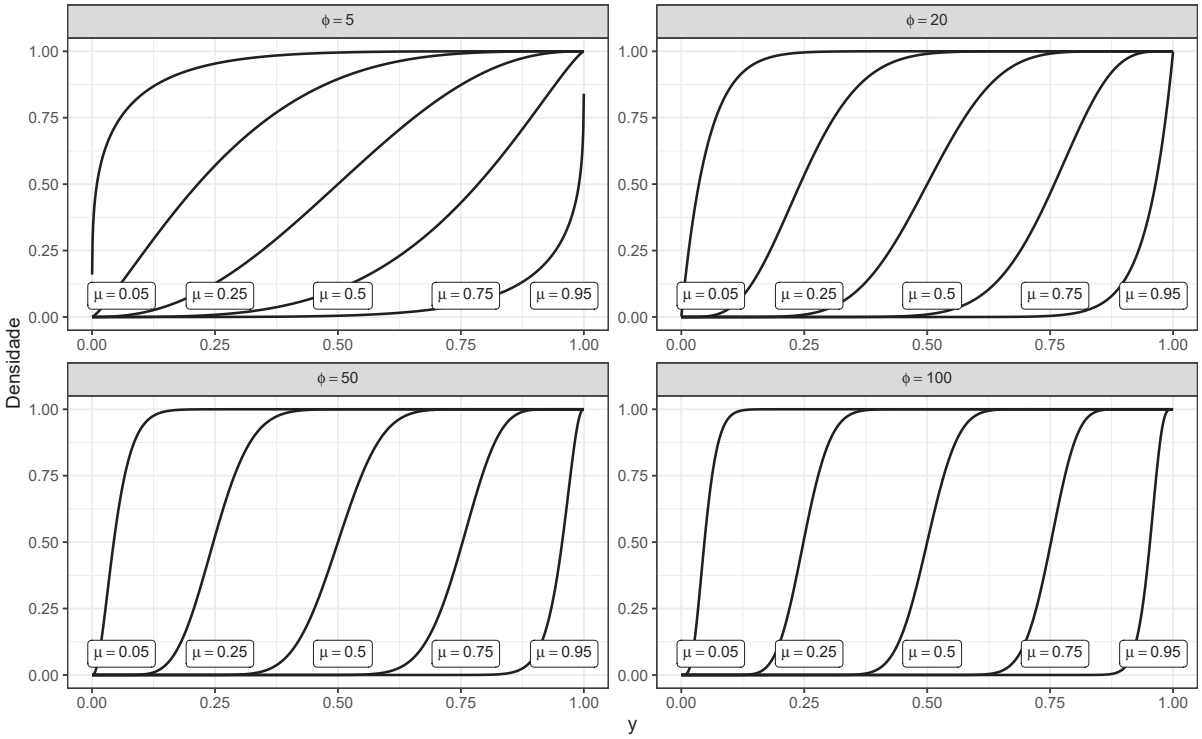
$$F(y,p,q) = \frac{\mathrm{B}(y,\mu\phi,\phi(1-\mu)) \times \Gamma(\phi)}{\Gamma(\mu\phi) \times \Gamma(\phi(1-\mu))}, \tag{2.16}$$

que depende das funções beta incompleta e gama aplicadas no intervalo desejado.

A Figura 5 mostra os perfis de cobertura das curvas originadas pela distribuição beta na parametrização 1 para variados valores de μ e ϕ. Identifica-se que valores reduzidos de ϕ resultam em dados mais dispersos e distribuídos por todo o domínio de Y. Em contrapartida, valores mais elevados de ϕ centralizam a distribuição ao redor de μ. Desta forma:

- Quando μ é reduzido e ϕ também é baixo, observa-se uma concentração abaixo da mediana de Y;

FIGURA 7 – PERFIS DISTRIBUIÇÃO BETA ACUMULADA PARAMETRIZAÇÃO 1 (μ, ϕ).



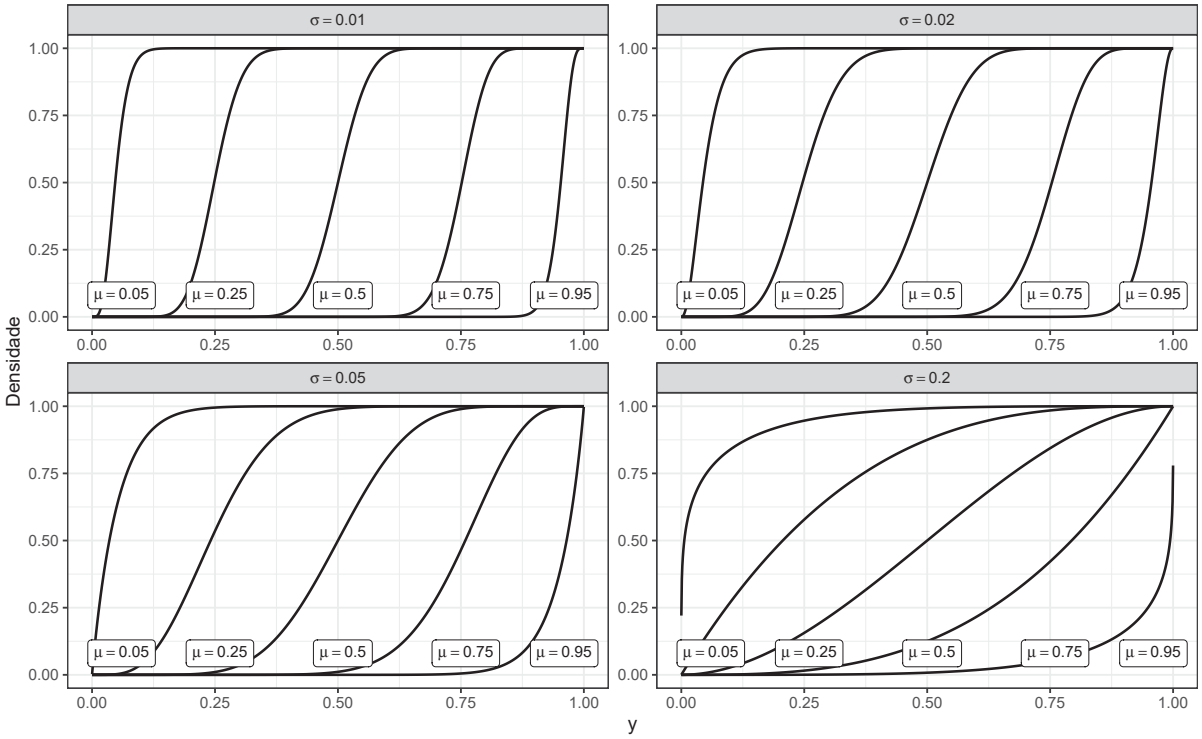
FONTE: O autor

- Quando μ é elevado e ϕ é baixo, a concentração é vista acima da mediana de Y;
- Se μ é reduzido e ϕ é elevado, há uma marcante concentração abaixo da mediana de Y;
- Se μ é elevado e ϕ também é, percebe-se uma intensa concentração acima da mediana de Y;
- Se μ situa-se na mediana, o valor de ϕ determina a variabilidade de Y em torno de μ.

Consequentemente, é evidente que, sem considerar as magnitudes de μ e ϕ, a distribuição beta tem uma ampla adaptabilidade na cobertura do suporte para qualquer valor de $Y = y_i$. Logo, essa distribuição é fortemente aconselhada para a modelagem de dados limitados ou passíveis de conversão para o intervalo unitário.

Na parametrização 2, os perfis da beta acumulada são apresentados na Figura 8. Observa-se um maior acúmulo de massa quando σ é reduzido, como previsto, uma vez que valores menores de σ resultam em valores de Y menos dispersos. Por outro lado, quando σ se aproxima de um, verifica-se um espalhamento mais acentuado, indicando assim uma maior dispersão.

FIGURA 8 – PERFIS DISTRIBUIÇÃO BETA ACUMULADA PARAMETRIZAÇÃO 2 (μ, σ).



FONTE: O autor

2.10 REGRESSÃO BETA

A classe dos modelos de regressão beta é útil para modelar respostas que se situam no intervalo (0,1), empregando uma estrutura de regressão que engloba uma função de ligação, covariáveis e parâmetros ainda não conhecidos. Muitos estudos em diversas áreas do saber têm recorrido à regressão beta para analisar a relação entre um conjunto de covariáveis e alguma porcentagem ou proporção. Entre eles, citam-se Brehm e Rahn (1993), Hancox e Poulton (2010), Kieschnick e McCullough (2003a), Smithson (2006) e Zucco et al. (2008).

Para obter informações adicionais sobre inferências com grandes amostras e análises de diagnóstico dentro dessa classe de modelos, recomenda-se consultar Espinheira e Meyer (2008a,b). Ospina e Ferrari (2006) propõem melhorias tanto em estimação pontual quanto intervalar. Simas et al. (2010) também discutem a estimação dentro dos modelos de regressão beta.

Nas próximas seções, serão detalhadas três formulações do modelo beta, as quais serão exploradas neste estudo, com foco em sua aplicação a dados em escala.

2.11 MODELO DE REGRESSÃO BETA

Seja $\mathcal{Y} = (\mathcal{Y}_1, ..., \mathcal{Y}_n)^T$ um vetor com n variáveis aleatórias independentes, em que cada $\mathcal{Y}_i, i = 1, ..., n$, tem distribuição beta com média μ_i e parâmetro de precisão desconhecido ϕ ou σ conforme parametrização adotada, então o modelo de regressão beta é definido:

$$\begin{aligned} y_i &\sim B(\mu_i, \phi) \\ g(\mu_i) = \eta_i &= \mathbf{x}_{ij}^{\mathsf{T}}\beta_j = \sum_{j=1}^{p} x_{ij}\beta_j \end{aligned} \tag{2.17}$$

ou conforme parametrização 2

$$\begin{aligned} y_i &\sim B(\mu_i, \sigma) \\ g(\mu_i) = \eta_i &= \mathbf{x}_{ij}^{\mathsf{T}}\beta_j = \sum_{j=1}^{p} x_{ij}\beta_j \end{aligned} \tag{2.18}$$

em que $\mathrm{g}(.)$ é uma função de ligação estritamente monótona e duplamente diferenciável, com domínio em $(0, 1)$ e imagem nos reais (ex. qualquer opção da TABELA 3).

2.12 MODELO DE REGRESSÃO BETA PARA DADO DE ESCALA

Sejam $\mathcal{Y} = \{(\mathcal{Y}_{11}, \mathcal{Y}_{12}), (\mathcal{Y}_{21}, \mathcal{Y}_{22}), (\mathcal{Y}_{31}, \mathcal{Y}_{32}), \ldots, (\mathcal{Y}_{n1}, \mathcal{Y}_{n2})\}^{\mathsf{T}}$ um vetor com n pares de observações de uma variável aleatória, tal que cada $y_i \in \mathcal{Y}$ contém o limite inferior $l_i = \mathcal{Y}_{i1}$ e o limite superior $l_s = \mathcal{Y}_{i2}$ de $\mathcal{Y}$, $\mathbf{x} = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip})^{\mathsf{T}}$ o conjunto de observações de p variáveis independentes em que $\beta = (\beta_1, \beta_2, \beta_3, \ldots, \beta_p)^{\mathsf{T}}$ é o vetor de parâmetros desconhecidos associado a cada x_i, $\mathbf{z} = (z_{i1}, z_{i2}, z_{i3}, \ldots, z_{ik})^{\mathsf{T}}$ o conjunto de k variáveis independentes relacionadas com o parâmetro de precisão ϕ, cujo vetor de parâmetros desconhecidos é dado por $\gamma = (\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_k)^{\mathsf{T}}$. O vetor completo de parâmetros desconhecidos a serem estimados é dado por $\theta = (\beta, \gamma)^{\mathsf{T}}$. Utilizando a EQUAÇÃO 2.9 com parâmetro de média μ_i associado a $\mathbf{x}$ e ϕ parâmetro de precisão. No contexto em que o parâmetro ϕ é fixo, isto é, não depende de efeito de qualquer variável, o modelo regressão beta intervalar toma a seguinte forma:

$$\begin{aligned} y_i &\sim B(\mu_i, \phi) \\ g(\mu_i) = \eta_i &= \mathbf{x}_{ij}^{\mathsf{T}}\beta_j = \sum_{j=1}^{p} x_{ij}\beta_j \end{aligned} \tag{2.19}$$

ou também em termos de σ.

$$\begin{aligned} y_i &\sim B(\mu_i, \sigma) \\ g(\mu_i) = \eta_i &= \mathbf{x}_{ij}^{\mathsf{T}}\beta_j = \sum_{j=1}^{p} x_{ij}\beta_j \end{aligned} \tag{2.20}$$

em que $\mathrm{g}(.)$ é uma função de ligação estritamente monótona e duplamente diferenciável, com domínio em $(0,1)$ e imagem nos reais (ex. qualquer opção da TABELA 3).

Note que, os modelos anteriores podem ser expandidos sem dificuldade para o caso em que o parâmetro de dispersão e/ou precisão, a depender da parametrização trabalhada, seja variável, isto é, seja também modelável por uma estrutura desrita por um segundo preditor linear.

Sendo assim, quando há o desejo de modelar também o efeito do parâmetro de dispersão ϕ sendo explicado por variáveis independentes z_i, o modelo de regressão beta passa ter dois preditores lineares, um associado a $\mathbf{x}$ e outro associado a $\mathbf{z}$:

$$\begin{aligned} y_i &\sim B(\mu_i, \phi_i) \\ g_1(\mu_i) &= \mathbf{x}_{ij}^{\mathsf{T}}\beta_j = \sum_{j=1}^{p} x_{ij}\beta_j \\ g_2(\phi_i) &= \mathbf{z}_{ij}^{\mathsf{T}}\gamma_j = \sum_{j=1}^{k} z_{ij}\gamma_j \end{aligned} \tag{2.21}$$

ou ainda

$$\begin{aligned} y_i &\sim B(\mu_i, \sigma_i) \\ g_1(\mu_i) &= \mathbf{x}_{ij}^{\mathsf{T}}\beta_j = \sum_{j=1}^{p} x_{ij}\beta_j \\ g_2(\sigma_i) &= \mathbf{z}_{ij}^{\mathsf{T}}\gamma_j = \sum_{j=1}^{k} z_{ij}\gamma_j \end{aligned} \tag{2.22}$$

em que $\mathrm{g}_1(.)$ é uma função de ligação estritamente monótona e duplamente diferenciável, com domínio em $(0,1)$ e imagem nos reais associadas as regressoras $\mathbf{x_{ij}}$ e $\mathrm{g}_1(.)$ é uma função de ligação estritamente monótona e duplamente diferenciável, com domínio em $(0,\infty)$ e imagem nos reais associadas as regressoras $\mathbf{z_{ij}}$.

2.13 FUNÇÕES DE LIGAÇÃO

Na literatura há diversas funções de ligação disponíveis para transformar o domínio de variação de variáveis aleatórias. Para modelo de regressão beta nosso o interesse está nas funções de ligação que fazem transformações $\mathrm{g}(.)(0,1) \Rightarrow \mathbb{R}$ e sua relação inversa, pois a variável resposta a ser modelada pertence ao intervalo $(0,1)$, mas as covariáveis podem pertencer a intervalos diferentes. As seguintes opções de funções de ligação presentes na TABELA 3 são opções que conferem maior estabilidade numérica e são fortemente utilizadas na literatura. Todas as funções de ligação possuem inversa e são duplamente diferenciáveis.

TABELA 3 – FUNÇÕES DE LIGAÇÃO TESTADAS NOS MODELOS DE REGRESSÃO.

Nome	$\mathrm{g}(\mu_i)$	$\mathrm{g}^{-1}(\eta_i)$	$\mathrm{g}'(\mu_i)$	Efeito
Logit	$\log\left(\frac{\mu_i}{1-\mu_i}\right)$	$\frac{e^{\eta_i}}{1+e^{\eta_i}}$	$\frac{1}{\mu_i(1-\mu_i)}$	μ_i
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$	$\sqrt{2\pi}\cdot\exp(\mu_i^2/2)$	μ_i
Cloglog	$\log(-\log(1-\mu_i))$	$1-e^{-e^{\eta_i}}$	$\frac{1}{(\mu_i-1)\log(1-\mu_i)}$	μ_i
Cauchit	$\tan(\pi(\mu_i-\frac{1}{2}))$	$\pi\cdot\csc^2(\pi\eta_i)$	$\pi\cdot\csc^2(\pi\cdot\mu_i)$	μ_i
Logaritmo	$\log(\mu_i)$	$\exp(\eta_i)$	$\frac{1}{\mu_i}$	$\phi_i;\sigma_i$
Raiz quadrada	$\sqrt{\mu_i}$	η_i^2	$\frac{1}{2\sqrt{\mu_i}}$	$\phi_i;\sigma_i$

FONTE: Adaptada de Andrade (2007)

A TABELA 3 contém a função de ligação, sua inversa e a primeira derivada que serão utilizadas na estimação dos parâmetros da regressão e foi construída com base em cálculos do autor e auxílio da referência de Andrade (2007). Em sua tese, o autor disserta sobre os impactos da escolha incorreta de funções e ligação em regressão beta e apresenta um estudo aprofundado sobre as funções de ligação mais conhecidas. Neste trabalho serão testadas essas funções em busca daquelas que melhor se aplicam ao modelo de regressão beta para dado censurado.

2.14 FUNÇÃO DE VEROSSIMILHANÇA

Em teoria das probabilidades e inferência estatística, a função de verossimilhança, frequentemente denominada apenas de "verossimilhança", descreve a probabilidade conjunta dos dados observados com base nos parâmetros de uma determinada distribuição de probabilidade ou modelo estatístico. Este conceito pressupõe que o processo gerador produza amostras independentes e identicamente distribuídas (iid). Deste modo, para cada valor específico de parâmetro no espaço paramétrico, a função de verossimilhança oferece uma previsão probabilística aos dados observados. A verossimilhança é definida pelo produto das distribuições e as probabilidades levadas em consideração abrangem tanto o processo de geração dos dados quanto o mecanismo gerador da amostra observada. Segundo Casella e Berger (2021), "o método de máxima verossimilhança é, indiscutivelmente, a técnica mais utilizada para derivar estimadores". Considerando-se uma amostra iid $Y_1, Y_2, ..., Y_n$ proveniente de uma distribuição $f(y, \theta_1, \theta_2, ..., \theta_k)$, a função de verossimilhança é representada por:

$$\begin{aligned} L(\theta|y) &= L(\theta_1, \theta_2, ..., \theta_k|y_1, y_2, ..., y_n) \\ &= \prod_{i=1}^{n} f(y_i|\theta_1, \theta_2, ..., \theta_k). \end{aligned} \tag{2.23}$$

A expressão apresentada pela EQUAÇÃO 2.23 sintetiza a ideia central da função de verossimilhança, sendo aplicável a qualquer parametrização em consideração, desde que aplicado a um modelo que segue uma distribuição de probabilidade. No contexto

da estimação, quando a verossimilhança alcança seu valor máximo, são obtidos os estimadores de máxima verossimilhança (EMV).

2.15 FUNÇÃO DE VEROSSIMILHANÇA COMPLETA COM CENSURA

Essa forma geral de verossimilhança expressa a ideia central da teoria de estimação de parâmetros, mas ela é flexível e pode ser adaptada para dados censurados e/ou intervalares. Por exemplo, com base nos trabalhos de Klein e Moeschberger (2003) e Helsel et al. (2005), que focaram em ajustes de modelos para dados censurados, Delignette-Muller e Dutang (2015) propuseram, no pacote `fitdistrplus` – um pacote na linguagem **R** que permite o ajuste de distribuições a dados com ou sem censura –, uma função de verossimilhança completa contemplando os três tipos de censura observados além do caso convencional. Visando estabelecer uma estrutura matemática completa, Lopes (2023) incluiu uma função indicadora $\mathbf{I}_{\zeta_i=0,1}$ que, caso ζ_i ocorra, assume o valor 1 e 0 caso contrário. Assim, a presença ou ausência de censura em determinado ponto de dado define o valor que $L(\theta)$ assume. Portanto, a forma completa da verossimilhança é definida por:

$$
\begin{aligned}
L(\theta) &= L(\theta)_{\delta=0} \times L(\theta)_{\delta=1} \times L(\theta)_{\delta=2} \times L(\theta)_{\delta=3} \\
L(\theta) &= \prod_{i=1}^{N_{\delta=0}} [f(y_i|\theta)]^{\zeta_i} \times \mathbf{I}_{(\zeta_i=0,1)} \\
&\quad \times \prod_{i=1}^{N_{\delta=1}} [F(y_i = u_i|\theta)]^{\zeta_i} \times \mathbf{I}_{(\zeta_i=0,1)} \\
&\quad \times \prod_{i=1}^{N_{\delta=2}} [1 - F(y_i = l_i|\theta)]^{\zeta_i} \times \mathbf{I}_{(\zeta_i=0,1)} \\
&\quad \times \prod_{i=1}^{N_{\delta=3}} [F(y_i = u_i|\theta) - F(y_i = l_i|\theta)]^{\zeta_i} \times \mathbf{I}_{(\zeta_i=0,1)}.
\end{aligned}
\tag{2.24}
$$

Na EQUAÇÃO 2.24, $\theta = (\beta, \phi)^{\mathsf{T}}$ é o vetor de parâmetros desconhecidos a serem estimados, $f(y_i = u_i|\theta)$ é a distribuição de probabilidade atribuída à variável aleatória Y e $F(y_i = u_i|\theta)$ é sua função de distribuição acumulada. Cada parte de $L(\theta)$ é composta por:

- $L(\theta)$: é a contribuição conjunta das quatro partes;
- $L(\theta)_{\delta=0}$: é a contribuição da parte sem censura;
- $L(\theta)_{\delta=1}$: é a contribuição da parte com censura à esquerda;
- $L(\theta)_{\delta=2}$: é a contribuição da parte com censura à direita;
- $L(\theta)_{\delta=3}$: é a contribuição da parte com censura intervalar.

Esta forma geral é aplicável a todas as parametrizações discutidas aqui envolvendo modelos de regressão beta, bem como em outras distribuições de probabilidade além da beta, sejam elas discretas ou contínuas. No caso da distribuição beta com censura intervalar, não existe solução analítica fechada para os estimadores de máxima verossimilhança. Portanto, métodos numéricos devem ser utilizados para calcular as estimativas dos parâmetros desconhecidos.

2.16 ESTIMAÇÃO

Quando adaptada ao modelo de regressão beta, a função de verossimilhança geral produz uma expressão complexa, cujos estimadores não têm uma forma algebricamente fechada. Dada essa complexidade, sugere-se tratar tudo numericamente utilizando um algoritmo adequado, como o BFGS, que está implementado no *core* da Linguagem R (R CORE TEAM, 2023a). O tratamento computacional pode obter as estimativas diretamente a partir da função de log-verossimilhança. Contudo, também é viável adquirir as derivadas de primeira e segunda ordens para a função escore e para a matriz hessiana por meio de derivação numérica. O pacote *numDeriv* oferece algoritmos para essa finalidade (GILBERT et al., 2009). O enfoque estritamente computacional pode aumentar o tempo de execução do processo de estimação devido ao custo computacional associado, mas representa uma solução prática diante das complexidades matemáticas presentes no processo.

A seguir, são explorados brevemente os algoritmos BFGS e L-BFGS-B, que serão utilizados nesta pesquisa.

2.17 MÉTODO BFGS (BROYDEN–FLETCHER–GOLDFARB–SHANNO)

O método BFGS é um algoritmo iterativo quase-Newton para otimização unconstrained (NASH, 1984). Ao invés de calcular diretamente o Hessiano $H(x)$, BFGS atualiza iterativamente uma aproximação H_k.

2.17.1 Passo a passo

1. Escolha um ponto inicial x_0 e inicialize H_0 como uma matriz identidade.
2. Para $k = 0, 1, 2, \ldots$:
    - Calcule o gradiente:
    $$g_k = \nabla f(x_k)$$
    - Determine a direção de busca p_k:
    $$p_k = -H_k g_k$$

- Faça uma busca linear para determinar α_k (NOCEDAL; WRIGHT, 2006).
- Atualize:

$$x_{k+1} = x_k + \alpha_k p_k$$

- Calcule:

$$s_k = \alpha_k p_k, \quad y_k = \nabla f(x_{k+1}) - g_k$$

- Atualize H_{k+1} usando a fórmula BFGS (BROYDEN, 1970).

3. Repita até convergência.

2.18 MÉTODO L-BFGS-B

L-BFGS-B é uma extensão do L-BFGS, que é uma variante de memória limitada do BFGS (BYRD et al., 1995). A principal diferença é que o L-BFGS-B pode lidar com restrições de caixa nas variáveis.

2.18.1 Passo a passo

Os passos são semelhantes ao BFGS, com algumas modificações:

1. Os vetores s_k e y_k das últimas m iterações são armazenados.
2. Durante o cálculo de p_k e α_k, as restrições de caixa são consideradas.

É importante notar que gerenciar restrições de caixa pode ser desafiador e muitas vezes requer métodos de projeção ou técnicas de penalização (NOCEDAL; WRIGHT, 2006).

2.19 INFERÊNCIA

A teoria de máxima verossimilhança é vasta, amplamente estudada e aplicada na chamada abordagem clássica, a mesma que suporta este trabalho. Nessa sessão abordaremos um pouco dessa teoria no contexto da teoria assintótica, pois ela está relacionada com as propriedades dos estimadores no contexto de grandes amostras. Para melhor aprofundamento recomenda-se consultar os trabalhos de Azzalini (1996), DasGupta (2008), Li e Babu (2019), Ibragimov e Has' Minskii (2013), Casella e Berger (2021).

2.19.1 Teoria assintótica

Com base na teoria assintótica dos estimadores de Máxima Verossimilhança, Casella e Berger (2021) discutem que a propriedade de consistência é fundamental em inferência, pois ela exige que o estimador de um parâmetro deve convergir para seu valor 'correto' à medida em que o tamanho da amostra se torna infinito. Assim, seja $\mathbf{W}_n = \mathbf{W}_n(X_1, X_2, ..., X_n)$ uma sequência de estimadores consistentes de θ, então existirá para cada θ um erro $\epsilon > 0$ que reduz com o aumento da amostra. Em termos matemáticos tem-se que

$$\lim_{n\to\infty} \mathbf{P}_\theta(|\mathbf{W}_n - \theta| < \epsilon = 1 \tag{2.25}$$

e de forma análoga

$$\lim_{n\to\infty} \mathbf{P}_\theta(|\mathbf{W}_n - \theta| \geq \epsilon = 0. \tag{2.26}$$

Conforme EQUAÇÃO 2.25 e EQUAÇÃO 2.26, à medida em que o tamanho da amostra aumenta, o estimador se aproximará cada vez mais do parâmetro verdadeiro com alta probabilidade. De forma análoga, a probabilidade de que o estimador se afaste do valor verdadeiro do parâmetro tende a zero.

Se um modelo de probabilidade for especificado corretamente, sob algumas condições de regularidade e com base na teoria assintótica, o vetor de parâmetros estimados $\hat{\theta}$ possuirá distribuição aproximadamente normal multivariada com media θ_0 e matriz de covariâncias

$$\hat{\mathbf{V}} = \left\{ -\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'} \right\}^{-1} \tag{2.27}$$

de dimensões $n \times m$ iguais ao total de estimativas p do modelo. Isto é, a distribuição de probabilidade de θ é dada por

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}_p(\boldsymbol{\theta_0}, \hat{\mathbf{V}}) \tag{2.28}$$

que também pode ser escrita em termos da distribuição χ^2 com p graus de liberdade, dada por

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\mathsf{T} \hat{\mathbf{V}}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim \boldsymbol{\chi}^2_p. \tag{2.29}$$

2.19.2 Intervalos de confiança

Intervalo de confiança (IC) é uma ferramenta estatística usada para estimar o valor verdadeiro de um parâmetro populacional com base em uma amostra de dados. Eles fornecem uma faixa de valores prováveis para o parâmetro populacional, juntamente com um nível de confiança. O nível de confiança é geralmente expressado

como uma porcentagem, como 95% e indica a probabilidade de que o intervalo de confiança realmente contenha o valor verdadeiro do parâmetro populacional. Quanto maior o nível de confiança, maior é a probabilidade de que o intervalo de confiança contenha o valor verdadeiro. Ainda no contexto da distribuição assintótica de θ, uma medida necessária para determinar o IC é o erro padrão que é determinado pela raiz quadrada da diagonal principal da matriz $\hat{\mathbf{V}}$ e pode ser escrito como

$$\widehat{SE}(\hat{\theta}) = \sqrt{\text{diag}(\hat{\mathbf{V}})}. \tag{2.30}$$

Com base no $\widehat{SE}$, um intervalo de confiança de Wald com $(1-\alpha)100\%$ de confiança para o vetor de estimativas $\hat{\theta}$ é calculado por

$$\left[\hat{\boldsymbol{\theta}} - \widehat{SE}(\hat{\boldsymbol{\theta}})\Phi^{-1}(1-\alpha/2); \hat{\boldsymbol{\theta}} + \widehat{SE}(\hat{\boldsymbol{\theta}})\Phi^{-1}(1-\alpha/2);\right], \tag{2.31}$$

em que Φ é acumulada da distribuição normal padrão, ou seja, com $\mu = 0$ e $\sigma = 1$. Outra forma de obter um IC aproximado para o vetor de θ é através da região obtida com base na aproximação dada pela EQUAÇÃO 2.29 denominada elipsoide. Nesse caso a região Θ_α é dada por

$$\Theta_\alpha = \left\{\boldsymbol{\theta} : (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \hat{\mathbf{V}}^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) < \boldsymbol{\chi}^2_p(1-\alpha)\right\}, \tag{2.32}$$

onde a parcela $\boldsymbol{\chi}^2_p(1-\alpha)$ denota o $(1-\alpha)100\%$ percentil de uma distribuição $\boldsymbol{\chi}^2_p$ com p graus de liberdade. Lembrando que p é o total de estimativas presentes no vetor $\hat{\theta}$.

2.19.3 Testes de hipóteses

A concepção dos testes de hipóteses não é recente. Talvez a obra mais completa nessa área seja atribuída a Lehmann e Lehmann (1986), que, em seu livro, procuraram unificar as ideias anteriormente abordadas por outros pesquisadores, como Ronald Fisher em meados de 1925, Jerzy Neyman e Egon Pearson por volta de 1929, e Wald (1939). Na teoria estatística, especialmente em modelos de regressão, os testes de hipóteses de Wald são frequentemente aplicados (WALD, 1939).

A partir das ideias exploradas por esses acadêmicos, a essência dos testes de hipóteses pode ser descrita da seguinte maneira: um teste de hipóteses é um método estatístico empregado para determinar se existe evidência robusta para corroborar ou refutar uma proposição (ou hipótese) acerca de uma população. O teste é formulado a partir de duas hipóteses: a hipótese nula $(H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0)$, que representa a afirmação em questão, e a hipótese alternativa $(H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0)$, que é o contraponto à hipótese nula. O sinal na hipótese alternativa pode variar, sendo $<, \leq$ (menor; menor ou igual) ou $>, \geq$ (maior; maior ou igual). O teste de hipóteses fundamenta-se em amostras aleatórias e emprega estatísticas para calcular a probabilidade de alcançar os resultados observados, ou resultados mais extremos, assumindo que a hipótese nula seja verdadeira.

Conforme o nível de significância definido, o teste pode optar por rejeitar ou aceitar a hipótese nula.

Com base na EQUAÇÃO 2.29, para testar a hipótese nula geral $(H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0)$ é possível utilizar a aproximação pela χ^2_p dada por $X^2 = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \hat{\mathbf{V}}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ e calcular um valor-p através da expressão

$$p_{val} = 1 - F_{\chi^2_p}(X^2), \tag{2.33}$$

em que $F_{\chi^2_p}(X^2)$ é distribuição χ^2_p acumulada com p graus de liberdade aplicada no ponto X^2. Para testar hipóteses individualmente para cada um dos k elementos do $\hat{\boldsymbol{\theta}}$, isto é um teste de hipóteses univariado, a hipótese apresentada antes pode ser adaptada para $(H_0 : \boldsymbol{\theta}_k = \boldsymbol{\theta}_{0,k}), k \in \{1, 2, ..., p\}$ e a estatística de teste será dada por

$$Z = \frac{\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_{0,k}}{\widehat{SE}(\hat{\theta}_k)}. \tag{2.34}$$

O valor-p para um teste bilateral baseado na aproximação normal é calculado por

$$\begin{aligned} p_{val} &= 2\{1 - \Phi(|-Z|)\} \\ &= 2\{\Phi(|-Z|)\}, \end{aligned} \tag{2.35}$$

em que Φ é acumulada da distribuição normal padrão com quantil igual a Z. Dada a simetria da distribuição normal, tanto faz a calda escolhida para determinar o p_{val}.

2.19.4 Teste da razão de verossimilhança

O Teste de Razão de Verossimilhança (TRV) é um método estatístico que serve para comparar duas hipóteses sobre uma distribuição de probabilidade e é baseado na razão das verossimilhanças das hipóteses, ou seja, a probabilidade dos dados observados dado que cada hipótese é verdadeira. O teste compara a razão das verossimilhanças sob a hipótese alternativa (H_1) e a hipótese nula (H_0), e essa razão é então comparada com um valor crítico para determinar se a hipótese nula deve ser rejeitada ou não. Os TRV são especialmente úteis quando a hipótese alternativa é um modelo aninhado, ou seja, é um caso especial da hipótese nula com um ou mais parâmetros adicionais. Nesses casos, os TRVs podem fornecer testes mais poderosos do que outros métodos, como testes chi-quadrado. Os TRVs também são usados na seleção de modelos (caso desse trabalho) onde o objetivo é selecionar o melhor modelo a partir de um conjunto de candidatos com base nos dados observados. Nesse caso, a razão de verossimilhança é usada para comparar o ajuste relativo de diferentes modelos, e o modelo com a maior razão de verossimilhança é selecionado. Há muitos trabalhos no campo do estudo dos TRV's. Mais detalhes podem ser consultados em Woolf (1957), Kent (1982) e por fim Satorra e Saris (1985) onde ideias de TRV robusto são exploradas.

No contexto desse trabalho, serão utilizados TRVs para testar a hipótese nula $H_0 : \boldsymbol{\theta} = \boldsymbol{\Theta}_0$ para algum subconjunto de $\boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$ do espaço paramétrico de Θ visando decidir qual o melhor entre um par de modelos. Por exemplo, $\boldsymbol{\Theta}_0$ pode ser obtido de um modelo mais simples de referência e comparado com outro modelo com mais componentes, por exemplo com mais covariáveis. Para maior estabilidade da medida calculada para o teste, em geral se trabalha com o logaritmo da verossimilhança $\ell(\hat{\theta})$, assim, se $\boldsymbol{\theta}_0$ for máximo da log-verossimilhança para o modelo mais simples e $\boldsymbol{\theta}$ for o máximo da log-verossimilhança para um modelo mais geral então a estatística do TRV é dada por

$$\Lambda_{TRV} = -2\{\ell(\hat{\boldsymbol{\theta}}_0) - \ell(\hat{\boldsymbol{\theta}})\}. \tag{2.36}$$

É possível demostrar que a estatística Λ_{TRV}, sob H_0 segue assintoticamente uma distribuição χ^2_ν com ν graus de liberdade e o valor-p do teste pode ser calculado com

$$p_{val} = 1 - F_{\chi^2_\nu}(\Lambda_{TRV}), \tag{2.37}$$

em que $F_{\chi^2_\nu}(X^2)$ é distribuição χ^2_ν acumulada com ν graus de liberdade, que no caso dos modelos de regressão equivale ao total de parâmetros do modelo menos 1 aplicada no ponto Λ_{TRV}. Os testes de Wald e TRV são assintoticamente equivalentes também para escolha de modelo, mas é preferível o uso do TRV nestes casos, especialmente quando se trata de modelos para dados com censura (COLLETT, 2015).

2.19.5 Seleção de modelos

Critério de informação é um conjunto de fórmulas matemáticas utilizadas para avaliar a qualidade de modelos estatísticos. Com o apoio deles é possível comparar diferentes modelos e selecionar aquele que melhor se ajusta aos dados. Os critérios de informação mais comuns são Critério de Informação de Akaike (*AIC*) proposto por Akaike (1974) e o Critério de Informação Bayesiano (*BIC*) praposto por Schwarz (1978). Ambas as medidas utilizam uma combinação da log-verossimilhança dos dados dado o modelo e sua complexidade para avaliar sua qualidade ou a bondade do ajuste. *AIC* e *BIC* dão pesos diferentes para a log-verossimilhança e sua complexidade, com *BIC* colocando geralmente mais ênfase na complexidade do modelo. Valores mais baixos de *AIC* e *BIC* indicam um melhor ajuste do modelo aos dados.

Tanto *AIC* quanto *BIC* são utilizados para avaliar modelos estatísticos, mas o *BIC* tem uma penalização mais forte para modelos complexos, o que o torna mais conservador do que o *AIC*. A equação do *AIC* é:

$$\mathbf{AIC} = 2k - 2\ell(\hat{\boldsymbol{\theta}}) \tag{2.38}$$

onde k é o número de parâmetros do modelo e $\ell(\hat{\boldsymbol{\theta}})$ é a log-verossimilhança dos dados dado o modelo. Já a equação do *BIC* é:

$$\mathbf{BIC} = k \log n - 2\ell(\hat{\boldsymbol{\theta}}) \tag{2.39}$$

onde k é o número de parâmetros do modelo, n é o número de observações.

Como dito, ambas as equações são utilizadas para avaliar a qualidade dos modelos estatísticos, sendo que *AIC* tem uma penalização menos forte para modelos complexos do que *BIC*, e *BIC* é mais conservador.

No caso do modelo quasi-beta, Bonat e Jørgensen (2016) trabalham com uma abordagem de estimação baseada na função de quasi-verossimilhança que gera uma função de quasi-escore. Para tornar os modelos comparáveis, no processo de seleção são calculadas estatísticas de pseudo *AIC* e pseudo *BIC* obtidos a partir da pseudo log-verossimilhança do modelo.

2.20 ANÁLISE DE RESÍDUOS

Como visto nesta revisão, a regressão beta é uma abordagem estatística fundamental para modelar variáveis resposta limitadas ao intervalo (0,1), como taxas, proporções e probabilidades. Essa técnica é especialmente útil quando a variável resposta não se ajusta adequadamente aos modelos lineares generalizados tradicionais, como a regressão logística e a regressão log-linear conforme Ferrari e Cribari-Neto (2004). Neste contexto, a análise de resíduos desempenha um papel crucial na avaliação da qualidade do ajuste e na identificação de possíveis problemas nos modelos de regressão beta. O objetivo dessa sessão é discutir algumas abordagens de análise de resíduos em regressão beta tradicional e encontrar formas de adaptá-los para o caso do modelo beta para dados com resposta transformada intervalar.

2.20.1 Resíduos de Pearson e *Deviance*

Os resíduos de Pearson e *Deviance* são frequentemente citados na literatura como ferramentas para avaliar a adequação dos modelos de regressão beta. Os resíduos de Pearson são descritos por:

$$r_P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(1-\hat{\mu}_i)/(1+\hat{\phi})}}$$

em que y_i representa a variável resposta observada, $\hat{\mu}_i$ é a média ajustada e $\hat{\phi}_i$ indica a dispersão ajustada (FERRARI; CRIBARI-NETO, 2004). Por outro lado, os resíduos de *Deviance* são determinados como:

$$r_D = \text{sign}(y_i - \hat{\mu}_i) \left\{ 2(\ell_i(y_i, \hat{\phi}) - \ell_i(\hat{\mu}_i, \hat{\phi})) \right\}^{1/2}.$$

Espinheira et al. (2008) destaca que tanto os resíduos de Pearson quanto os de *Deviance* apresentam restrições, especialmente quando as premissas de normalidade e homoscedasticidade dos resíduos não são atendidas.

2.20.2 Resíduos quantílicos aleatorizados e ajustados

No caso dos modelos beta, se faz necessário uma abordagem de análise de resíduos menos sensível a pressupostos. Nessa linha, Dunn e Smyth (1996) propuseram os chamados Resíduos Quantílicos Aleatorizados (RQA) como uma alternativa aos resíduos de Pearson e *Deviance*, especialmente em modelos de regressão generalizados. Os RQA são construídos a partir da inversa da função de distribuição acumulada condicional e adicionando-se um componente aleatório para garantir que os resíduos sigam uma distribuição normal padrão sob a hipótese nula de que o modelo objeto de estudo está correto. No caso da regressão beta, os resíduos quantílicos aleatorizados são dados por:

$$r_{RQA} = \Phi^{-1}\left\{F(y_i \mid \hat{\mu}_i, \hat{\phi}_i)\right\}$$

onde Φ^{-1} é a função quantil da distribuição normal padrão e $F(y_i \mid \hat{\mu}_i, \hat{\phi}_i)$ é a função de distribuição acumulada condicional da distribuição beta (ESPINHEIRA et al., 2008). Ainda nessa linha, Pereira (2019) introduziu uma abordagem alternativa, denominada resíduos quantílicos ajustados, que se baseia na transformação dos RQA usando a matriz de projeção H:

$$r_{RQAA} = H \cdot r_{RQA}$$

onde H é a matriz de projeção que ajusta os resíduos quantílicos aleatorizados para atender às restrições do modelo de regressão beta. Como sugerido, as suposições de normalidade e homoscedasticidade não precisam ser atendidas e ainda assim, tem-se nos RQA uma opção robusta para verificar a qualidade de ajuste de modelos beta.

2.20.3 Comparação e recomendação

Os resíduos de Pearson e *Deviance* são abordagens tradicionais para avaliar a adequação dos modelos de regressão beta. No entanto, eles têm limitações, principalmente quando as suposições de normalidade e homoscedasticidade não são satisfeitas (ESPINHEIRA et al., 2008). Nesses casos, os resíduos quantílicos aleatorizados e ajustados oferecem alternativas mais robustas e precisas para avaliar a qualidade do ajuste e identificar problemas potenciais nos modelos de regressão beta (DUNN; SMYTH, 1996; PEREIRA, 2019). Os resíduos quantílicos aleatorizados têm a vantagem de seguir uma distribuição normal padrão sob a hipótese nula de que o modelo está correto, o que é útil para a construção de gráficos de diagnóstico e testes de hipóteses (DUNN; SMYTH, 1996). Os resíduos quantílicos ajustados, por sua vez, fornecem uma abordagem alternativa que leva em conta as restrições do modelo de regressão beta

e pode ser usada para melhorar a interpretação e a análise dos resíduos (PEREIRA, 2019). Com base na revisão da literatura, recomenda-se o uso de resíduos quantílicos aleatorizados e ajustados em modelos de regressão beta para avaliar a qualidade do ajuste e identificar possíveis problemas. Essas abordagens são mais robustas e precisas do que os resíduos de Pearson e *Deviance* e fornecem informações valiosas para melhorar a qualidade e a interpretação dos modelos de regressão beta tratados nesse trabalho. Como no caso so modelo tratado nesse trabalho a variável resposta é intervalar, para adequar as abordagens de resíduos descritas aqui, adotou-se o ponto médio do intervalo de cada $y_i = (y_{il} + y_{is})/2$ sem perda de generalidade.

2.21 MODELO DE REGRESSÃO QUASI-BETA

Como uma alternativa para lidar com dados limitados ao intervalo unitário modeláveis via regressão beta, Bonat et al. (2019) apresentam uma categoria versátil de modelos de regressão fundamentados em pressupostos relacionados ao primeiro e segundo momentos. A estrutura de média é desenvolvida por intermédio de uma função de ligação e um preditor linear como de costume. Já a relação entre média e variância é expressa como $\phi\mu^p(1-\mu)^p$, onde μ, ϕ e p representam os parâmetros de média, dispersão e potência, respectivamente.

Assim, considere um conjunto de dados transversais, $(\mathcal{Y}_i, x_i)$, $i = 1, \ldots, n$, onde $\mathcal{Y}_i$ são realizações independentes e identicamente distribuídas de $\mathcal{Y}_i$ de acordo com uma distribuição não especificada, cuja expectativa e variância são dadas por:

$$\begin{aligned} \mathrm{E}(Y_i) &= \mu_i = g^{-1}(x_i^T\beta) \\ \mathrm{Var}(Y_i) &= \sigma_i = \phi\mu_i^p(1-\mu_i)^p \end{aligned} \tag{2.40}$$

onde x_i e β são vetores $(q \times 1)$ de covariáveis conhecidas e parâmetros de regressão desconhecidos, respectivamente. Além disso, g é uma função de ligação padrão, na qual adotamos a função logit para fornecer valores médios no intervalo $(0, 1)$, mas potencialmente qualquer outra função de ligação adequada poderia ser adotada. O modelo de regressão especificado é parametrizado por $\theta = (\beta^T, \lambda^T)^T$, onde $\lambda = (\phi, p)$ com o espaço de parâmetros usual e pode imitar a relação entre média e variância do modelo.

2.21.1 Estimação e Inferência

A metodologia de estimação para o modelo definido na EQUAÇÃO 2.40 se fundamenta nos trabalhos de Jørgensen e Knudsen (2004) e Bonat e Jørgensen (2016). Nestes estudos, duas técnicas são empregadas: a função de quasi-escore e a função de estimação de Pearson. A função de quasi-escore é aplicada para a estimação dos

parâmetros médios no modelo quasi-beta, sendo definida por:

$$\psi_\beta(\beta,\lambda)=\left(\sum_{i=1}^{n}\frac{\partial\mu_i}{\partial\beta_1}\sigma_i^{-1}(Y_i-\mu_i),\ldots,\sum_{i=1}^{n}\frac{\partial\mu_i}{\partial\beta_q}\sigma_i^{-1}(Y_i-\mu_i)\right)^{\mathsf{T}}.$$

A função de estimação de Pearson (*unbiased*), utilizada para a estimação dos parâmetros de dispersão, é dada por:

$$\psi_\lambda(\lambda,\beta)=\left(\sum_{i=1}^{n}\frac{\partial\ -\sigma_i^{-1}}{\partial\phi}[(Y_i-\mu_i)^2-\sigma_i],-\sum_{i=1}^{n}\frac{\partial\ \sigma_i^{-1}}{\partial p}[(Y_i-\mu_i)^2-\sigma_i]\right)^{\mathsf{T}}.$$

A solução do sistema de equações $\psi_\beta(\beta,\lambda)=0$ e $\psi_\lambda(\lambda,\beta)=0$ é obtida através de um processo iterativo, fazendo uso do algoritmo interativo *chaser* modificado proposto por Jørgensen e Knudsen (2004). Este algoritmo foi implementado no pacote mcglm Bonat (2018b), e a busca é conduzida pelo sistema de recorrência:

$$\beta^{(i+1)}=\beta^{(i)}-\mathbf{S}_\beta^{-1}\psi_\beta(\beta^{(i)},\lambda^{(i)})\lambda^{(i+1)}=\lambda^{(i)}-\alpha\mathbf{S}_\lambda^{-1}\psi_\lambda(\beta^{(i+1)},\lambda^{(i)})$$

Em relação ao processo iterativo, $\mathrm{S}^{-1}\beta$ e $\mathrm{S}^{-1}\lambda$ representam as inversas das matrizes que compõem a diagonal da matriz de sensitividade cruzada. Esta matriz é expressa por:

$$\mathbf{S}_\theta=\begin{pmatrix}\mathbf{S}_\beta & \mathbf{0}\\ \mathbf{S}_{\lambda\beta} & \mathbf{S}_\lambda\end{pmatrix}.$$

Além disso, tem-se a matriz de variabilidade, dada por:

$$\mathbf{V}_\theta=\begin{pmatrix}\mathbf{V}_\beta & \mathbf{V}_{\beta\lambda}\\ \mathbf{V}_{\lambda\beta} & \mathbf{V}_\lambda\end{pmatrix}.$$

No que tange ao processo de inferência, adotam-se variâncias empíricas sugeridas por Bonat e Jørgensen (2016). O objetivo desta escolha é contornar a necessidade de calcular os terceiro e quarto momentos, tarefas estas de alta complexidade. Especial atenção é dada à componente $\mathbf{V}_\lambda$, sobre a qual recai a matriz de informação de Godambe.

2.22 PSEUDO LOG-VEROSSIMILHANÇA GAUSSIANA

Nesse trabalho estão sendo comparadas duas abordagens distintas de modelagem. Na primeira, que abrange os modelos beta convencional e beta para dados de escala, a abordagem é distribucional, isto é, adota-se como referência a distribuição de probabilidade beta e portanto, a estimação depende de uma formulação robusta da função de verossimilhança conforme já apresentado. Já no segundo caso, tem-se um modelo sem suposição de distribuição, o quasi-beta. Nessa abordagem o processo

de modelagem assim como estimação não usam diretamente a distribuição beta, mas apenas as suposições de primeiro e segundo momentos amostrais, assim, o processo de estimação também muda.

Baseado nos trabalhos de Bonat e Jørgensen (2016) em seu pacote Bonat (2018a) a ideia de uma pseudo-log verossimilhança gaussiana foi adotada e através dela foram calculados pseudo-AIC e pseudo-BIC para os preditos de todos os modelos testados. Nessa técnica assume-se uma distribuição normal com média μ igual aos valores preditos pelo modelo e desvio padrão σ igual à variância estimada do modelo.

3 MATERIAIS E MÉTODOS

Este capítulo desempenha um papel crucial ao definir os critérios da metodologia e dos materiais utilizados nesta pesquisa, garantindo que outros pesquisadores possam reproduzir o trabalho. Especificamente, exploramos o conjunto de dados relacionados a cirurgias de joelho e introduzimos uma estratégia de simulação de dados beta com censura intervalar. O objetivo é testar e avaliar características do modelo beta neste contexto específico de aplicação.

Conforme destacado por Mertler e Reinhart (2017), uma metodologia clara e meticulosamente detalhada é essencial para garantir a validade e confiabilidade de um estudo. A importância de clareza neste contexto foi também enfatizada por Bem (2003), que abordou os desafios de procedimentos ambiguamente definidos em pesquisas psicológicas. Seguindo as práticas sugeridas por Patten (2014), detalhamos cada etapa do processo, desde a seleção do conjunto de dados até a implementação da estratégia de simulação.

3.1 DADOS

Antes de apresentar o planejamento dos dados segue uma breve descrição do que é o Ligamento Cruzado Anterior (LCA) e como, em geral é o processo de intervenção cirúrgica adotado em casos de lesionamento.

3.1.1 Ligamento Cruzado Anterior (LCA)

O Ligamento Cruzado Anterior (LCA) é um dos principais ligamentos que mantêm a estabilidade da articulação do joelho conforme FIGURA 9. Quando ele é lesionado, pode haver instabilidade e dor no joelho, o que pode afetar significativamente a capacidade de realizar atividades diárias e esportivas. A cirurgia de reparação do LCA é realizada para restaurar a estabilidade e a função do joelho. A reconstrução do Ligamento Cruzado Anterior (LCA) é uma cirurgia destinada a reparar esse ligamento lesionado. A abordagem cirúrgica adotada varia conforme as características do paciente e a natureza e gravidade da lesão. Geralmente, o procedimento envolve a remoção do LCA danificado e a sua substituição por um enxerto de tecido.

Um método comum de reconstrução utiliza um enxerto de tendão autógeno, ou seja, retirado do próprio paciente. Os tendões mais frequentemente empregados nesse procedimento são o quadríceps ou o patelar. Durante a cirurgia, o médico realiza incisões na frente e atrás do joelho para acessar o LCA lesionado. Após remover o ligamento danificado, o cirurgião prepara o enxerto e o fixa no local apropriado.

FIGURA 9 – ILUSTRAÇÃO LIGAMENTO CRUZADO ANTERIOR

FONTE: Website <https://clinicajoelhoombro.com>, acesso em 21/01/2023 às 18:42

Uma alternativa é a reconstrução do LCA usando um enxerto de tendão alógeno, que provém de um doador. Há também técnicas que empregam dispositivos como parafusos ou encaixes.

Depois da intervenção cirúrgica, geralmente é aplicada uma bandagem compressiva no joelho do paciente, além de uma tala para assegurar estabilidade e limitar movimentos que poderiam comprometer a recuperação. É comum a recomendação de fisioterapia para fortalecer a região e restaurar a mobilidade do joelho. O período de recuperação varia, podendo durar de semanas a meses, a depender da técnica adotada e do estado de saúde geral do paciente.

3.1.2 Dados de cirurgia do LCA

O conjunto de dados originou-se de uma pesquisa realizada no Hospital do Trabalhador, em Curitiba-PR, durante o período de março de 2010 a março de 2013.

A aprovação para a pesquisa e coleta de dados foi concedida pelo Comitê de Ética em Pesquisa em Seres Humanos da Secretaria de Estado da Saúde do Paraná. O desenho experimental foi estruturado da seguinte maneira: inicialmente, 220 pacientes foram selecionados de forma sequencial para integrar o estudo. Porém, após a aplicação de critérios de inclusão-exclusão, os quais são confidenciais, o número de pacientes na amostra foi reduzido para 166. Em seguida, esses pacientes foram distribuídos aleatoriamente em quatro grupos diferentes, e cada um foi submetido a variados procedimentos anestésico-cirúrgicos para reconstrução do LCA. O Grupo 1 contou com 45 pacientes, o Grupo 2 com 34, o Grupo 3 com 43 e o Grupo 4 com 44. É importante mencionar que os participantes da pesquisa abrangiam ambos os gêneros, com faixa etária entre 18 e 65 anos. No entanto, tais variáveis não foram consideradas nas análises, já que não foram fornecidas no conjunto de dados. Além

disso, informações detalhadas sobre os critérios de elegibilidade dos pacientes para cada grupo não foram disponibilizadas.

A dor experimentada por cada paciente foi mensurada utilizando a escala *NSR-11*. As anotações foram registradas em três momentos específicos: a primeira após 6 horas do procedimento cirúrgico, a segunda após 12 horas, e a última após 24 horas da cirurgia.

3.2 SIMULAÇÃO

Nesse estudo de simulação, o foco recai sobre as propriedades dos estimadores de máxima verossimilhança em modelos de regressão beta aplicados a dados de escala beta mapeáveis. O estudo emprega simulações computacionais para avaliar o desempenho desses estimadores em distintos cenários. Especificamente, o interesse é entender a influência de variáveis como tamanho da amostra, precisão do instrumento e dispersão nas estimativas dos coeficientes de regressão do modelo.

O software R, citado em R Core Team (2023b), foi empregado nas simulações. Estabelecendo cenários controlados, o processo de geração de amostras foi rigorosamente monitorado, reduzindo assim incertezas e ruídos potencialmente advindos de fatores externos ao experimento.

Dentre os modelos de regressão beta avaliados, estão o modelo convencional, o modelo para dado de escala e o quasi-beta. Uma comparação aprofundada entre esses modelos foi realizada, utilizando-se da parametrização 2, como apresentado na EQUAÇÃO 2.9.

A fim de proporcionar uma análise holística, diferentes cenários de simulação foram concebidos, abrangendo variados casos específicos. Através desta análise meticulosa, busca-se não apenas compreender a variabilidade dos coeficientes em cada modelo, mas também discernir a magnitude do impacto de diferentes variáveis nas estimativas. Tal abordagem fornecerá informações valiosas para aprimoramentos subsequentes no modelo de regressão beta, servindo como guia para futuras pesquisas no domínio.

Os seguintes cenários de simulação foram estabelecidos:

- **Modelos para dados de escala**: Três modelos foram testados: modelo beta usual (M1), modelo quasi-beta (M2) e modelo beta para dados de escala (M3).
- **Tratamento dos intervalos**: Este cenário se concentra na transformação de escala, como discutido em seção 2.7. Durante esse processo, também foi aplicada uma correção para o efeito de borda, observado nos valores 0 e 1, que não estão no suporte da distribuição beta nos modelos analisados.

- **Precisão do instrumento**: Foram examinadas três escalas: NRS-8, que é equivalente a uma escala Likert com 8 divisões; HRS-21, com 20 divisões; e NRS-101, com 100 divisões. Espera-se que ao aumentar a precisão do instrumento (ou seja, com mais divisões), o efeito do erro de medida seja atenuado em todos os modelos.
- **Efeito da dispersão e/ou precisão**: Aqui, o foco é avaliar o impacto do parâmetro de dispersão/precisão nas estimativas dos coeficientes de regressão do modelo. Foram consideradas duas abordagens de regressão: a primeira com dispersão fixa e a segunda com dispersão variável, onde variáveis explicativas são atribuídas à dispersão.
- **Estabilidade dos betas de** x_{ij}: Os valores simulados destas variáveis permaneceram constantes em todas as amostras. Isso possibilita uma análise mais detalhada da variabilidade dos coeficientes em cada modelo, bem como o efeito destes em relação ao tamanho da amostra.

3.2.1 Modelo com dispersão fixa

Através da fixação do parâmetro de dispersão é possível medir o impacto nas estimativas de máxima verossimilhança dos parâmetros β. Para tanto assumiu-se como função de ligação $\mathbf{logit}$ da TABELA 3 a seguinte estrutura com 1000 réplicas

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$
$$\mu_i = \frac{\exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\right)}$$

Com os seguintes critérios

- Para o preditor linear da média: $\beta_i : \beta_0 = 0.5, \beta_1 = 0.2, \beta_2 = -0.7$;
- Para o parâmetro de dispersão: $\sigma_i : \sigma_1 = 0.8, \sigma_2 = 0.4, \sigma_3 = 0.2, \sigma_4 = 0.08, \sigma_5 = 0.02$;
- Para a primeira variável preditora: $x_1 : X_1$ com distribuição normal com média $\mu = 0$ e variância $\sigma = 1$;
- Para a segunda variável preditora: $x_2 : X_2$ com distribiução binomial com tamanho $= 1$ e probabilidade $= 0.5$;
- Tamanhos de amostra testados: $n : 50, 100, 250, 500, 1000$;
- Cortes de escala testados: $5, 7, 10, 20, 100$, onde 10 equivale a uma *NRS-11* e 100 uma *NRS-101*.

3.3 VIÉS DAS ESTIMATIVAS

Neste estudo, foi adotado um critério de viés padronizado para avaliar o viés ($\hat{bias}$) nas estimativas dos coeficientes obtidos em cada simulação. Este método padronizado oferece uma representação gráfica mais clara das estimativas em cada cenário simulado e permite comparações diretas entre diferentes cenários, ao fixar a escala de viés. O critério para o cálculo do viés padronizado é explicado abaixo:

- **Cálculo do Viés Geral:** Para cada cenário, a média das estimativas obtidas nas 500 réplicas é calculada e subtraída do valor verdadeiro da medida. Matematicamente, $\hat{bias} = \sum_{i=0}^{500} \frac{\hat{\beta_{0i}}}{n} - \beta_0$. Por exemplo, se foi simulado $\beta_0 = 0.50$ e a média de todas as 500 estimativas obtidas foi 0.52, o bias geral da rodada para β_0 foi $\hat{bias} = \sum_{i=0}^{5} 00\frac{\hat{\beta_0 i}}{n} - \beta_0 = 0.52 - 0.50 = 0.02$.
- **Determinação do Intervalo de Erro:** A média dos erros padrão das estimativas de máxima verossimilhança é calculada. Este valor é adicionado e subtraído ao viés para criar um intervalo de erro centrado em zero, ou seja, $\hat{bias} \pm \hat{se}$.
- **Viés Padronizado:** Os resultados de $\hat{se}$ de um grupo de referência são utilizados para padronizar os outros resultados. Especificamente, os valores de $\hat{se}$ do grupo de referência são divididos pelos valores β_{0i} simulados com critérios correspondentes para outros tamanhos de amostra.
- **Aplicação aos Cenários:** Este procedimento é repetido para todos os cenários de simulação, assegurando assim uma comparação robusta e padronizada.

3.4 TAXA DE COBERTURA

A avaliação do desempenho e precisão das estimativas dos coeficientes em um modelo de regressão em variados cenários simulados é essencial em análises estatísticas. Este processo auxilia no entendimento de como as estimativas se alinham aos valores reais, refletindo a eficácia do modelo em suas projeções. Tanto na análise do viés quanto na análise de cobertura, vários elementos como o próprio viés, a variabilidade e a acurácia das estimativas são escrutinados. A análise de cobertura, especificamente, investiga a habilidade do modelo em fornecer intervalos de confiança que incluam os valores reais dos coeficientes com uma probabilidade predefinida, por exemplo, 95%. Essa investigação é fundamental para:

- Determinar a precisão das estimativas dos coeficientes e entender o grau de incerteza ligado a elas.

- Assegurar que os intervalos de confiança gerados para os coeficientes estão corretamente calibrados, ou seja, que abrangem os valores reais na frequência antecipada.
- Avaliar a influência de variáveis como tamanho da amostra, dispersão dos dados e correlação entre as variáveis explicativas nas projeções dos coeficientes.

Ao se debruçar sobre a cobertura das estimativas, é possível discernir as virtudes e vulnerabilidades do modelo de regressão, bem como identificar áreas para aprimoramento, visando elevar a precisão e confiabilidade das projeções. Adicionalmente, tal análise equipa os pesquisadores com *insights* preciosos para uma tomada de decisão informada e uma interpretação acertada em contextos empíricos.

4 RESULTADOS E DISCUSSÃO

Este capítulo resume os principais resultados obtidos nas simulações computacionais dos modelos M1, M2 e M3, bem como na aplicação em dados de cirurgia de joelhos, detalhados na seção 3.1.

4.1 RESULTADOS DAS SIMULAÇÕES

As simulações foram feitas em todos os cenários descritos anteriormente, contudo, decidiu-se manter no corpo do texto apenas os cenários envolvendo a estratégia de tratamento da escala centralizada ao meio (m) conforme FIGURA 4. Demais cenários estarão disponíveis no ANEXO 4.

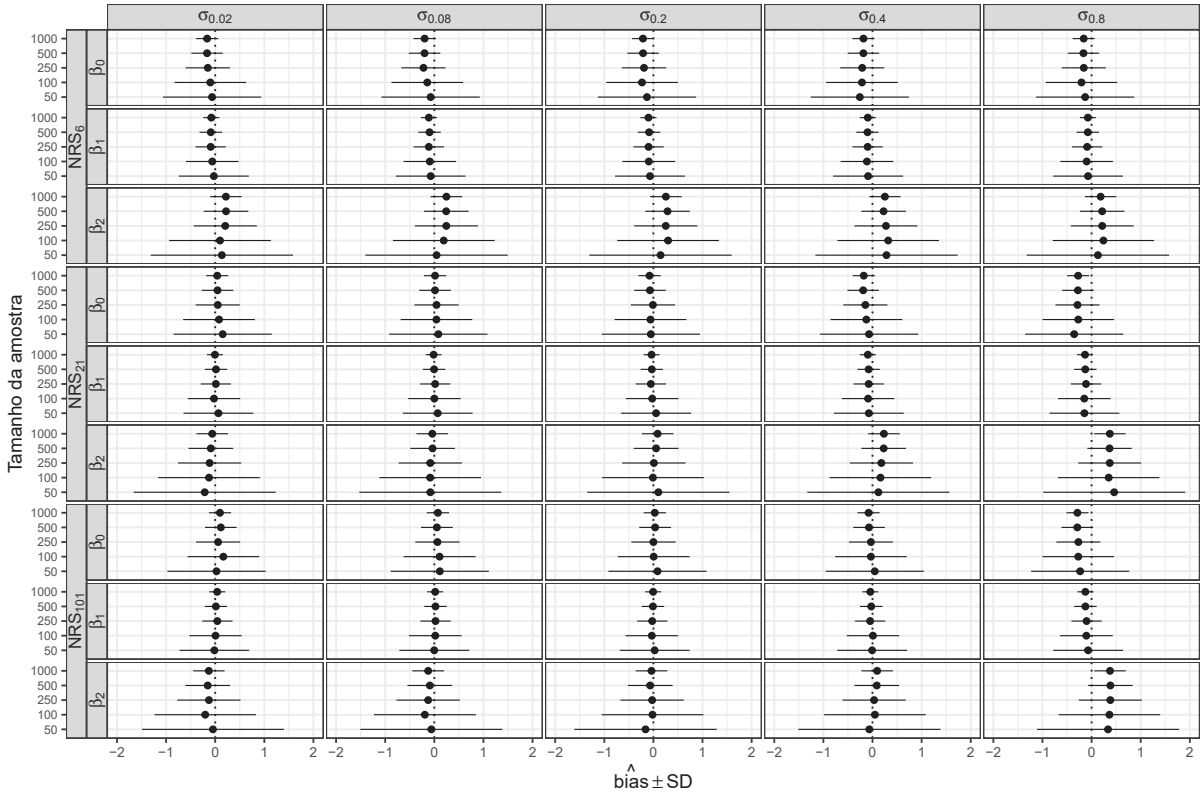
4.1.1 Análise do viés das estimativas

A FIGURA 10 revela que, em uma escala com poucas categorias — como a NRS-8, que conta com 7 intervalos — as estimativas de $\hat{\beta}_2$ (intercepto) do modelo apresentam maior variabilidade em contextos de baixa e alta dispersão. Especificamente, o viés é mais acentuado em amostras menores. Embora as estimativas dos coeficientes beta se mostrem precisas, elas exibem uma tendência crescente de viés à medida que a dispersão aumenta. Além disso, um aumento no tamanho da amostra amplifica o viés. No entanto, observa-se uma acurácia constante nas estimativas em todos os cenários analisados, mesmo quando o viés é elevado.

O parâmetro de dispersão no M1, conforme evidenciado na FIGURA 11, está diretamente associado ao número total de quebras na escala analisada. Um menor número de quebras induz a um viés ampliado nas estimativas. No entanto, à medida que o tamanho da amostra aumenta e nas situações de alta precisão, esse viés é claramente atenuado. Em cenários onde o valor simulado se aproxima da unidade, o viés se reduz de maneira significativa e a acurácia das estimativas cresce consideravelmente.

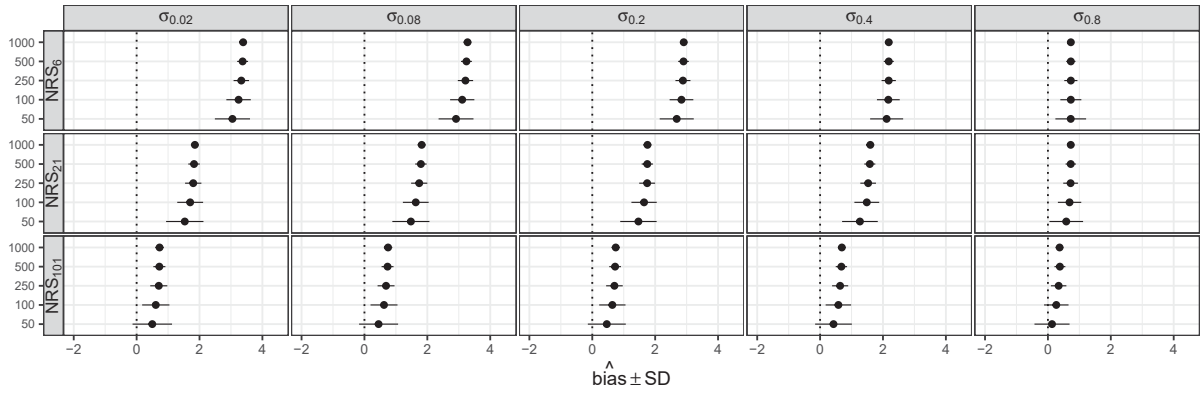
Em relação ao modelo M2, vide FIGURA 12 e FIGURA 13, há uma detalhamento das estatísticas dos coeficientes beta e do parâmetro de dispersão, respectivamente. Nesta metodologia avançada, na qual os dados são interpretados através da distribuição beta acumulada em um formato beta intervalar, identifica-se um padrão congruente ao observado no modelo M1. É evidente um viés mais expressivo no coeficiente $\hat{\beta}_2$ da inclinação do modelo, cujas causas podem estar associadas à especificidade do modelo escolhido ou a flutuações aleatórias. Além disso, percebe-se um acréscimo no viés conforme a dispersão σ se intensifica, independentemente dos cenários de cortes de escala ou dos tamanhos de amostra.

FIGURA 10 – MODELO (M1) - BETA USUAL: VIÉS DAS ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO β EM CENÁRIOS DIVERSOS.



FONTE: Produzida pelo autor

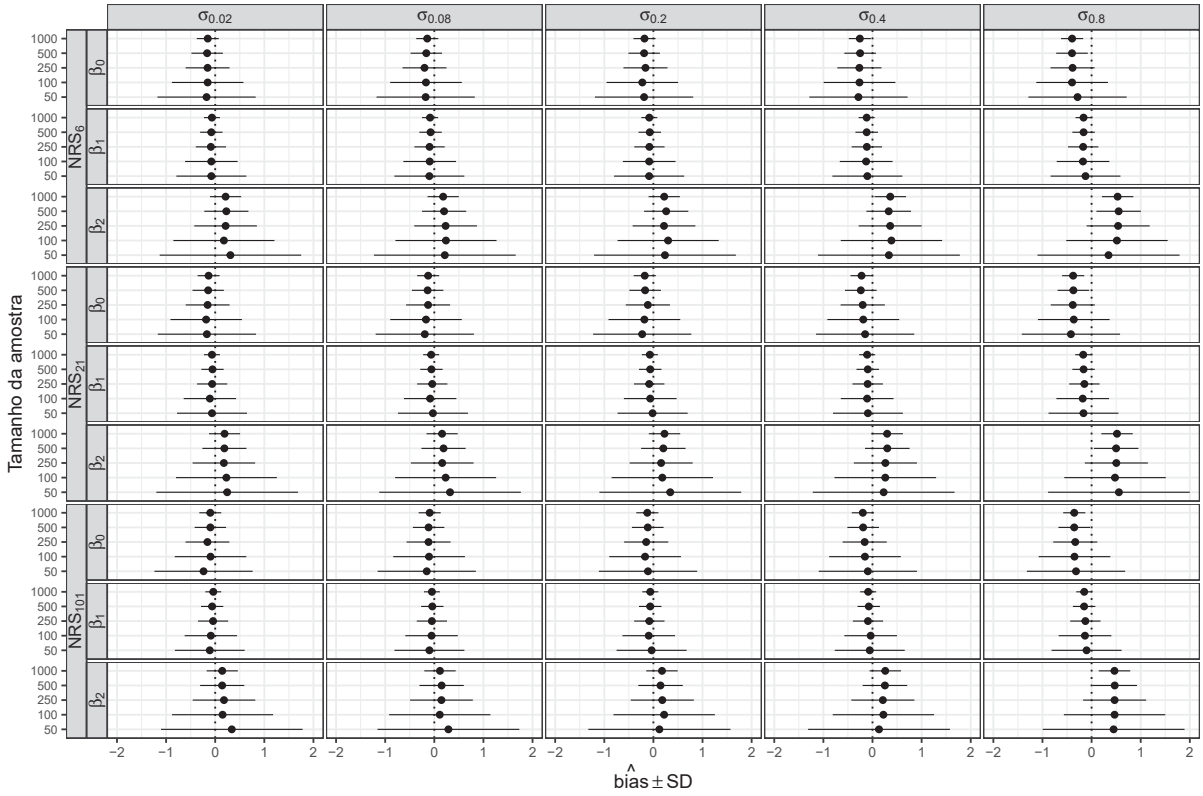
FIGURA 11 – MODELO (M1) - BETA USUAL: VIÉS DAS ESTIMATIVAS DO COEFICIENTE DE DISPERSÃO σ EM CENÁRIOS DIVERSOS.



FONTE: Produzida pelo autor

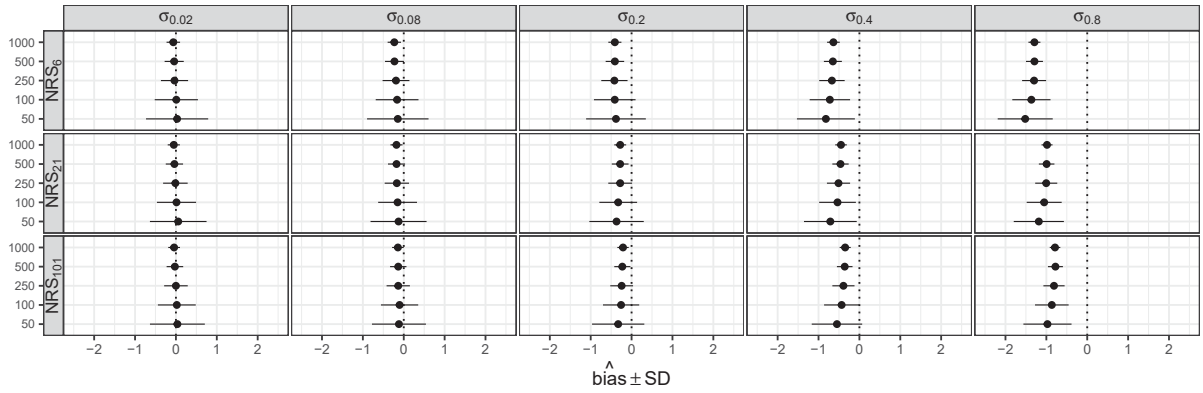
Já para o parâmetro $\hat{\sigma}$, demonstrado na FIGURA 13, registra-se uma acurácia superior e um viés atenuado em situações onde a dispersão simulada não ultrapassa $0,2$. Consequentemente, dispersões mais elevadas resultam em um viés mais pronunciado nas estimativas de σ. A influência do tamanho da amostra se mantém consistente em todos os cenários de corte das escalas em M2.

FIGURA 12 – MODELO (M2) - BETA INTERVALAR: VIÉS DAS ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO β EM CENÁRIOS DIVERSOS SIMULADOS.



FONTE: Produzida pelo autor

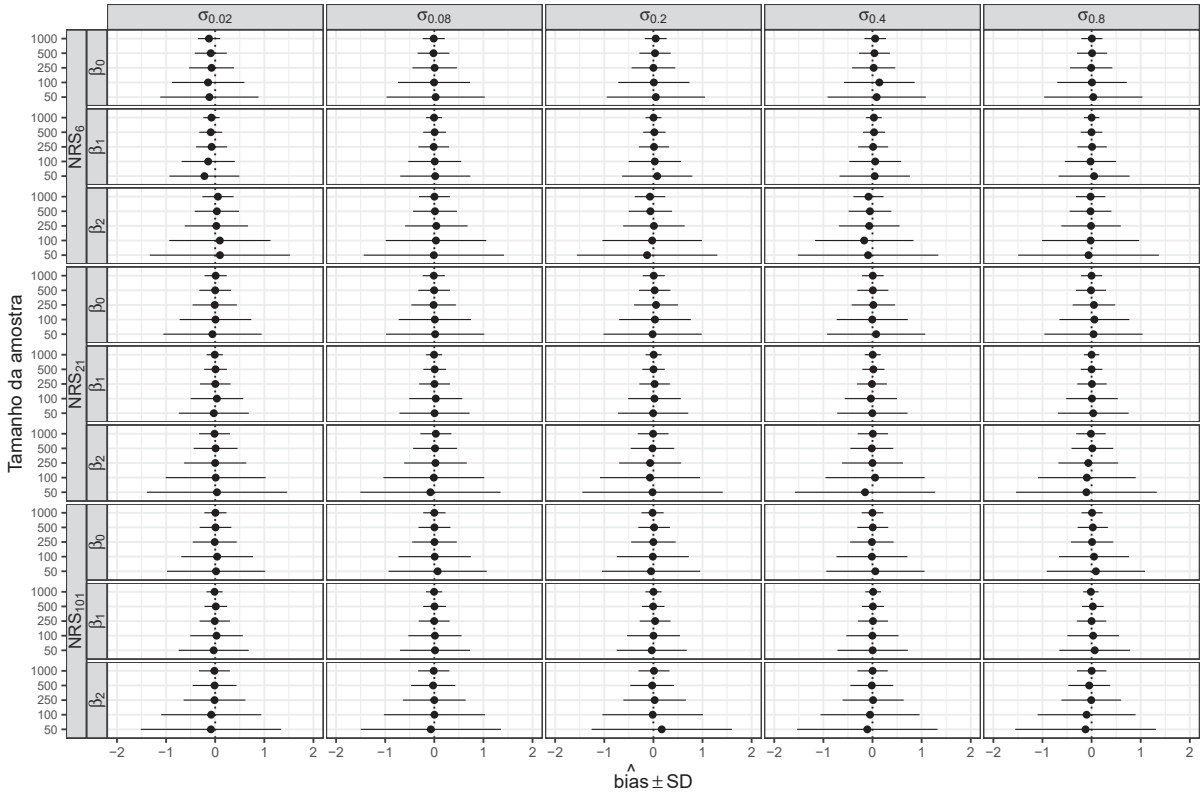
FIGURA 13 – MODELO (M2) - BETA INTERVALAR: VIÉS DAS ESTIMATIVAS DO COEFICIENTE DE DISPERSÃO σ EM CENÁRIOS DIVERSOS.



FONTE: Produzida pelo autor

Ao avaliar o modelo M3, a FIGURA 15 indica um viés mínimo em todos os cenários examinados. Independentemente do valor fixo simulado para β, do número total de quebras, do tamanho da amostra e da escala de σ, o viés dos coeficientes beta se mostra negligível, garantindo uma acurácia notável. Como esperado em todos os cenários e modelos, o intervalo de variabilidade dos coeficientes beta contrai em

FIGURA 14 – MODELO (M3) - QUASI-BETA: VIÉS DAS ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO β EM CENÁRIOS DIVERSOS SIMULADOS.



FONTE: Produzida pelo autor

amostras menores e se expande quando as amostras são maiores.

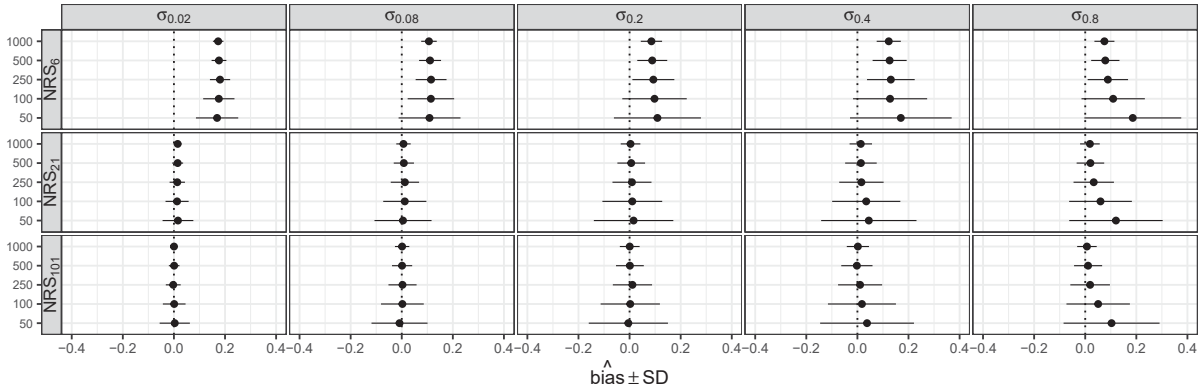
Em relação ao parâmetro σ, identifica-se um viés mais acentuado em escalas com menos de 20 quebras, independente do tamanho da amostra. No entanto, conforme o número de quebras cresce, o viés se atenua, ressurgindo apenas em valores extremos de dispersão, aproximando-se do limite superior do intervalo de σ que está confinado a $(0, 1)$, uma consequência da adoção da parametrização 2.

Em resumo, as análises de dados simulados mostraram que:

1. No modelo M1, é evidente que, em escalas com poucas categorias, o viés nas estimativas se torna mais pronunciado, sobretudo em amostras de menor tamanho. Ainda que a acurácia das estimativas persista em todos os cenários, a sensibilidade do modelo M1 ao número de intervalos na escala é inegável.

2. O modelo M2 demonstra comportamento similar ao M1. Há uma evidente ampliação do viés no coeficiente de inclinação à medida que a dispersão cresce. Apesar de garantir maior acurácia e apresentar menor viés nas estimativas de dispersão, a influência do tamanho da amostra mantém-se constante, como observado no modelo M1.

FIGURA 15 – MODELO (M3) - QUASI-BETA: VIÉS DAS ESTIMATIVAS DO COEFICIENTE DE DISPERSÃO σ EM CENÁRIOS DIVERSOS.



FONTE: Produzida pelo autor

3. O M3 destaca-se por seu mínimo viés em todos os cenários avaliados, superando os modelos M1 e M2. Sua acurácia é impecável, e a variação dos coeficientes beta contrai à medida que o tamanho da amostra decresce. O M3 ainda apresenta viés consideravelmente menor em relação ao parâmetro de dispersão, sobretudo em escalas que ultrapassam 20 quebras.

Portanto, é irrefutável que o M3 sobressai em desempenho para dados de escala quando contraposto aos modelos M1 e M2. Ele se destaca por seu baixo viés, notável acurácia e resistência robusta em face de distintos cenários e variáveis considerados.

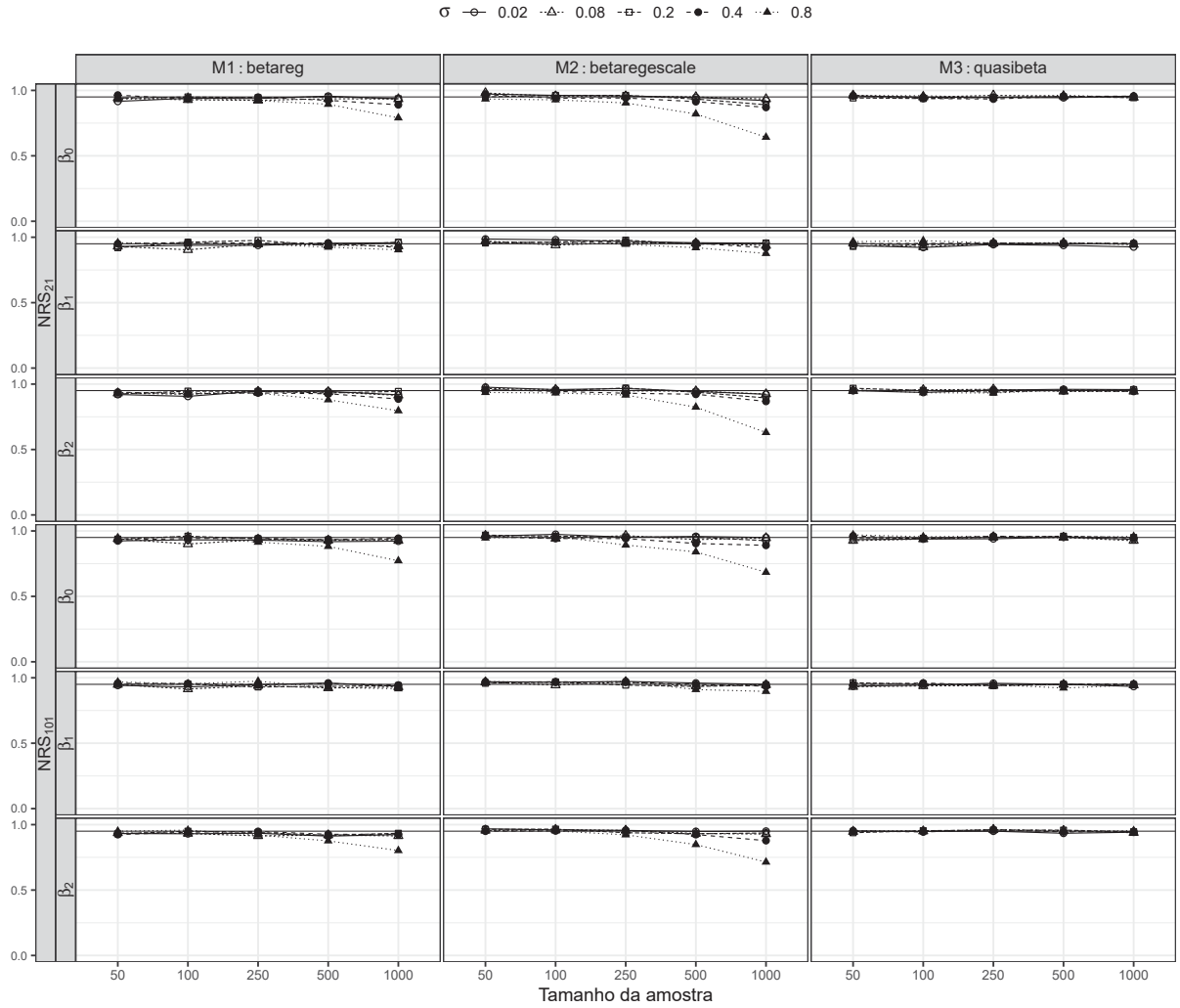
4.1.2 Análise da cobertura das estímativas

A FIGURA 16 exibe a taxa de cobertura com 95% de confiança nos cenários simulados para os coeficientes beta fixos. Intervalos de confiança de Wald foram considerados, conforme a EQUAÇÃO 2.31.

A linha preta contínua simboliza o limite de confiança de 95%, e idealmente, as estimativas deveriam orbitar este marco. Nota-se que tanto as estimativas do M1, o beta tradicional, quanto do M2, a abordagem intervalar, frequentemente exibem coberturas abaixo de 90% para todos os betas em amostras superiores a 500 em quantidade, quando o parâmetro de dispersão σ supera 0.2, independentemente da dimensão da amostra. Contudo, ambas as abordagens se mostram adequadas para amostras com tamanho inferior a 500 e dispersão limitada.

Uma justificativa para essa redução na precisão da cobertura reside na acumulação de erros de aproximação numérica. Dada a intensa complexidade computacional ao resolver integrais e derivadas das funções Γ e beta do modelo, que não têm expres-

FIGURA 16 – COBERTURA DAS ESTIMATIVAS DOS COEFICIENTES DE REGRESSÃO β EM CENÁRIOS SIMULADOS DIVERSOS.



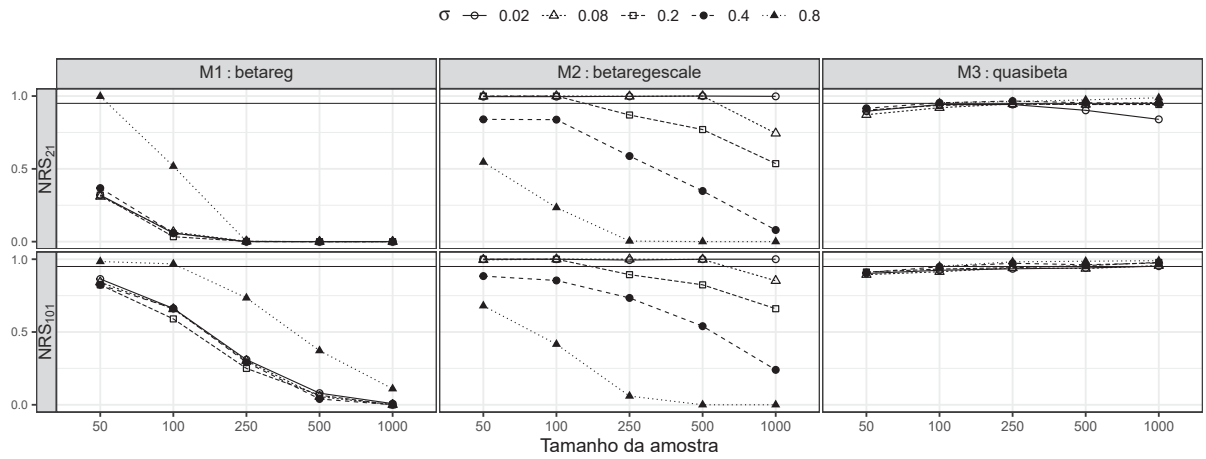
FONTE: Produzida pelo autor

são fechada, erros tendem a se acumular proporcionalmente ao aumento do tamanho da amostra. Uma segunda razão pode estar atrelada ao parâmetro de dispersão: estimativas com elevada variabilidade ou imprecisão naturalmente exercem maior influência nos coeficientes beta.

Em contraste, o M3, ou o modelo quasi-beta, demonstra consistência em todas as situações analisadas. Seja qual for o tamanho da amostra, o número de quebras ou a magnitude da dispersão, a cobertura se mantém próxima aos 95% para todos os betas simulados.

A estimação de parâmetros de dispersão ou precisão em modelos beta frequentemente confronta desafios, devido a restrições em seu espaço paramétrico, assimetrias, e outros obstáculos numérico-computacionais. Em nossa simulação, essa tendência se manteve. A FIGURA 17 ilustra que as estimativas de σ para os modelos

FIGURA 17 – COBERTURA DAS ESTIMATIVAS DO COEFICIENTE DE DISPERSÃO σ EM CENÁRIOS SIMULADOS DIVERSOS.



FONTE: Produzida pelo autor

M1 e M2 enfrentaram sérias dificuldades de cobertura, particularmente quando os valores de $\hat{\sigma}$ superavam 0.2. Esse fator pode elucidar os erros de cobertura nos betas observados anteriormente. Destaca-se que o modelo M2 proporcionou uma cobertura mais robusta que o M1 para valores de σ abaixo de 0.2. Em contrapartida, o M3 exibiu uma cobertura amplamente superior aos outros modelos em quebras acima de 8, sendo mais vulnerável apenas em contextos com um total de quebras na escala inferior a 10.

Em retrospecto, ao avaliar os modelos M1 (beta tradicional), M2 (abordagem intervalar) e M3 (quasi-beta), ficou evidente que o M3 se sobressai no que tange à cobertura das estimativas dos coeficientes do preditor de médias e do coeficiente de dispersão σ. O M3 evidenciou consistência e confiabilidade em variados contextos, assegurando uma cobertura em torno de 95% em todos os betas simulados. Os modelos M1 e M2, contudo, revelaram-se mais frágeis, sobretudo em amostras superiores a 500 e com um parâmetro de dispersão σ excedendo 0.2, conduzindo a frequentes coberturas abaixo de 90%. Estas fragilidades podem ser imputadas tanto à complexidade computacional quanto ao acúmulo de erros de aproximação nas funções Γ e beta.

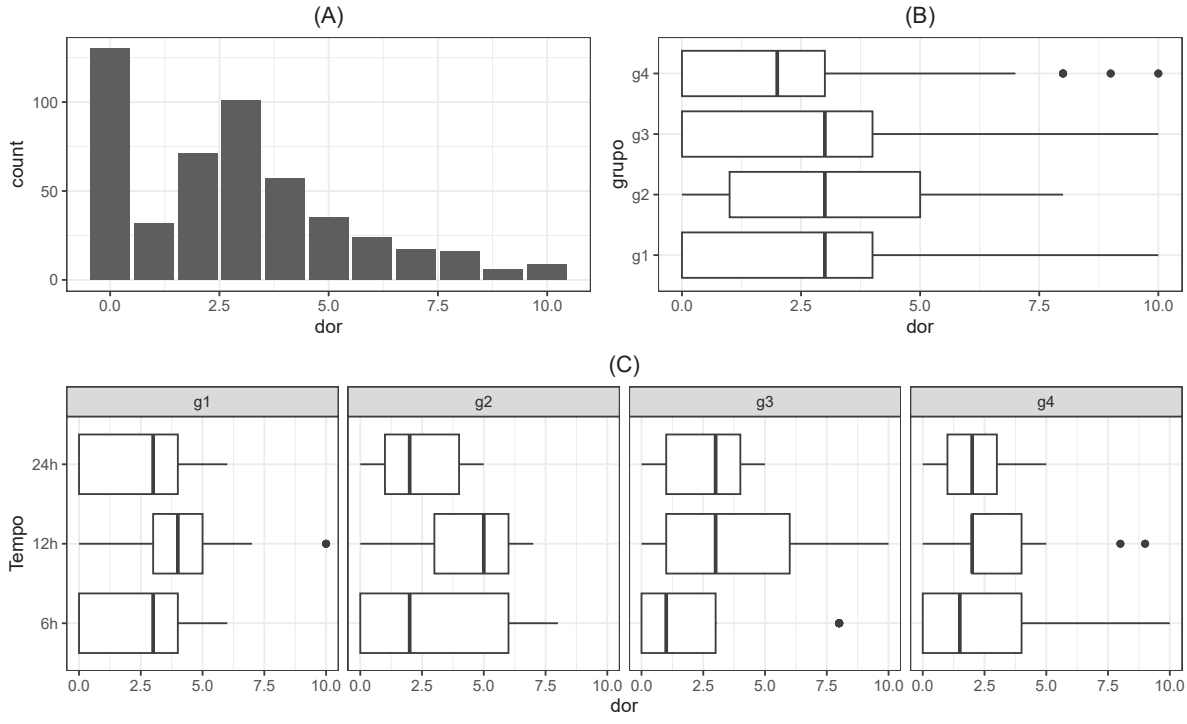
Ao analisar os três modelos de regressão - M1 (beta convencional), M2 (abordagem intervalar) e M3 (quasi-beta) - fica evidente a notável eficiência do modelo M3 em todas as simulações. Em termos simples, o quasi-beta (M3) mostrou-se mais acertado e confiável em várias situações testadas, enquanto os modelos M1 e M2 apresentaram algumas dificuldades, principalmente quando trabalhamos com amostras grandes ou com valores elevados de dispersão. Portanto, este estudo sugere que, para quem busca resultados mais precisos e confiáveis, o modelo M3 se mostra uma boa opção. Ainda no contexto das simulações os modelos M1 e M2 também possuem grande importância no contexto de dados de escala beta mapeáveis. O M2 por exemplo, é uma

proposta que se pode definir como padrão ouro, uma vez que contempla trata o dado na forma intervalar. Isso garante que não haja perda de informação por arredondamento, e mesmo sendo mais lento dada a complexidade matemática envolvida no processo de estimação, que é puramente computacional esse modelo possui grande relevância estatística e precisa ser melhor explorado no contexto de otimização e ferramental matemático.

4.2 RESULTADOS DA ANÁLISE DE DADOS DE CIRURGIA DE JOELHO

O objetivo desta análise é avaliar a performance dos três modelos e a estimação dos coeficientes de regressão em cada abordagem, considerando a presença de covariáveis reais, bem como compreender seus impactos no preditor linear e na dispersão. Adicionalmente, com a análise de dados reais, é possível inferir, a partir dos valores preditos, a estimativa de dor mais provável que seria informada para um indivíduo não incluído na amostra, mas que possua características similares àquela cujas medidas foram coletadas, especialmente em relação ao tempo e ao grupo. Esse tipo de medida pode ser de grande interesse na área médica. Uma vez que se conheça algumas características do paciente e o tipo de procedimento cirúrgico adotado é possível fazer um planejamento de cuidado visando reduzir o sofrimento no pós-cirúrgico.

FIGURA 18 – ESCORES DE DOR POR GRUPO E POR TEMPO EM HORAS PÓS CIRÚRGIA

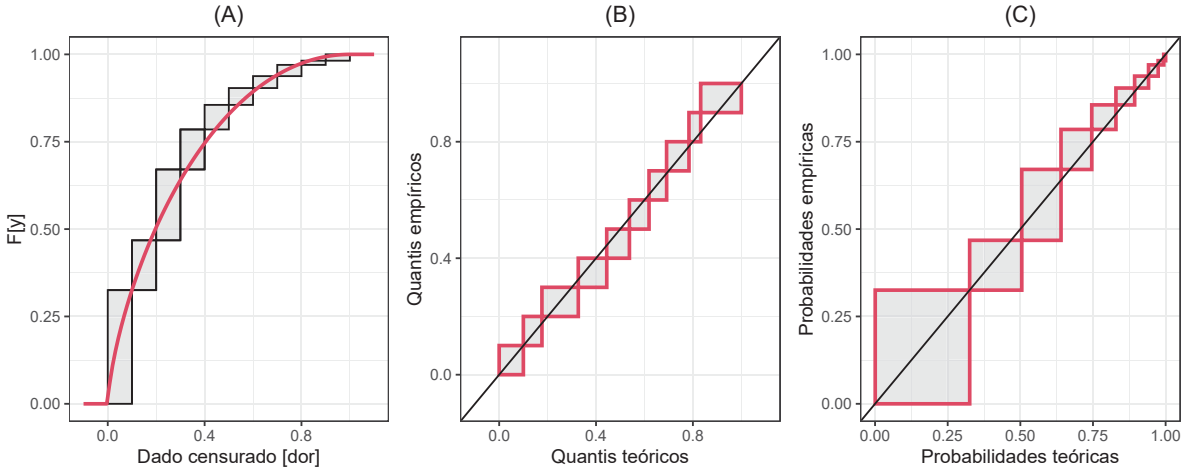


FONTE: Produzida pelo autor

Através de uma análise descritiva, pode-se ter um panorama dos dados e obter informação relevante para a modelagem. Conforme ilustrado na FIGURA 18, é evidente que a maioria dos pacientes registrou escores de dor entre 2 e 4, com uma notável prevalência no escore 3. Isso sugere uma intensidade de dor classificada entre baixa e moderada. Em relação ao aspecto temporal, há indicações claras de que, em todos os grupos, especialmente nos grupos g1 e g2, escores acima de 3 foram mais frequentes após 12 horas da cirurgia. No grupo g1, essa tendência persistiu até 24 horas após a intervenção. No entanto, em grupos como g1 e g4, foram identificados pontos atípicos, com pacientes alcançando o extremo superior da escala de dor. A investigação meticulosa desses pontos é crucial para aprimorar e entender as abordagens de manejo pós-cirúrgico.

A FIGURA 19 mostra a distribuição acumulada dos scores de dor em escala intervalar censurada já em escala $(0, 1)$ transformada pela equação 2.3 e um ajuste da distribuição beta em vermelho.

FIGURA 19 – DISTRIBUIÇÃO DOS SCORES DE DOR EM ESCALA BETA E INTERVALAR CENSURADA



FONTE: Produzida pelo autor

Em cada retângulo representado, o eixo y ilustra a probabilidade acumulada de ocorrência do escore, enquanto o eixo x evidencia o intervalo de ocorrência do escore de dor. Observa-se que em (A), conforme previamente antecipado, existe uma maior concentração de escores entre 0 e 1 $(0.0 - 0.1)$ e entre 2 e 3 $(0.2 - 0.3)$. Em qualquer intervalo de dados na escala transformada, a linha vermelha aproxima-se do centro de gravidade de cada intervalo, indicando uma boa aderência do modelo beta aos dados observados. Em contrapartida, (B) e (C) demonstram, respectivamente, as distribuições dos quantis empíricos (baseados nos dados) versus os quantis teóricos e as probabilidades empíricas em confronto com as teóricas. Considerando que o intervalo de dados transformados está na escala 0-1, a manifestação de uma linha

diagonal entre as distribuições empíricas e teóricas serve como evidência adicional, corroborando a adequação do ajuste beta aos dados censurados. seção 2.5

Para a modelagem desses dados, sugeriram-se duas estruturas de preditores lineares. A primeira estrutura refere-se a um modelo proposto contendo somente uma preditora, especificamente a variável **tempo**. Esta variável denota o momento pós-cirúrgico no qual a medida de dor foi registrada. Sua expressão matemática é:

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1\,\text{tempo}_{t2} + \beta_2\,\text{tempo}_{t3}, \tag{4.1}$$

que define o modelo M1 e outra contendo **tempo** mais **grupo** que representa o grupo de estudo considerado e que fica dada por

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1\,\text{tempo}_{t2} + \beta_2\,\text{tempo}_{t3} + \beta_3\,\text{grupo}_{g2} + \beta_4\,\text{grupo}_{g3} + \beta_5\,\text{grupo}_{g4} \tag{4.2}$$

e que define o modelo M2.

A TABELA 4 apresenta estatísticas de pseudo log-verossimilhança, incluindo o pseudo AIC e pseudo BIC para os modelos M1 e M2, avaliados para cada abordagem: **betareg**, **betaregscale** e **quasibeta**. Enquanto **betareg** e **betaregscale** são abordagens de modelos beta paramétricos, com a distribuição beta como substrato, a abordagem **quasibeta** não se baseia diretamente na distribuição beta, classificando-se como **quasi**, conforme detalhado anteriormente. Para viabilizar a comparação entre os modelos, recorreu-se a estatísticas de bondade de ajuste aproximadas baseadas na pseudo log-verossimilhança. Neste contexto, a distribuição normal é empregada na variável resposta y, com média μ e desvio padrão σ correspondendo, respectivamente, aos valores preditos e à variância estimada de cada modelo.

TABELA 4 – ESTATÍSTICAS BASEADAS NA PSEUDO LOG-VEROSSIMILHANÇA GAUSSIANA

	betareg		betaregscale		quasibeta	
Medida	M11	M12	M21	M22	M31	M32
Gl	4	7	4	7	4	7
pLogLik	-27.760	-26.069	19.011	22.420	12.490	16.200
pAIC	63.520	66.138	-30.022	-30.841	-16.980	-18.400
pBIC	80.363	95.612	-13.179	-1.367	-0.162	11.032

FONTE: O autor 2023.

Ao interpretar os resultados da TABELA 4 com base no Critério de Informação de Akaike penalizado (pAIC), observamos uma distinção clara na eficácia dos modelos betareg, betaregscale e quasibeta em termos de equilíbrio entre complexidade e ajuste aos dados. Os modelos betaregscale (M21 e M22) emergem como os mais eficientes, apesar de sua complexidade adicional em relação ao betareg. Esta eficácia é evidenciada pelos seus valores de pAIC notavelmente baixos (-30.022 para M21 e -30.841 para

M22), indicando um ajuste superior aos dados. A complexidade adicional do betaregscale, portanto, se justifica plenamente, fornecendo uma modelagem precisa e robusta dos dados. Em segundo lugar, os modelos quasibeta (M31 e M32) apresentam um equilíbrio razoável entre simplicidade e eficácia. Com valores de pAIC de -16.980 para M31 e -18.400 para M32, estes modelos são menos precisos que os betaregscale, mas ainda assim oferecem um ajuste satisfatório, especialmente considerando sua menor complexidade. O quasibeta destaca-se como uma opção viável quando a simplicidade é uma prioridade, mesmo que isso implique em uma leve redução na precisão do ajuste. Por fim, os modelos betareg (M11 e M12), apesar de sua complexidade inerente e suposições robustas baseadas na distribuição beta, não apresentam o equilíbrio mais eficiente entre ajuste e complexidade para os dados analisados. Seus valores de pAIC, 63.520 para M11 e 66.138 para M12, são superiores aos dos outros modelos, sugerindo que, apesar da complexidade, eles não proporcionam um ajuste proporcionalmente melhor.

Portanto, em termos gerais, o betaregscale se destaca como o modelo vencedor, seguido pelo quasibeta e, em último lugar, pelo betareg. Essa hierarquia reflete a capacidade de cada modelo de equilibrar eficazmente a complexidade e o ajuste aos dados em análise.

TABELA 5 – ESTIMATIVA E ERRO PADRÃO DOS COEFICIENTES DOS MODELOS TESTADOS

	betareg		betaregscale		quasibeta	
Variável	M11	M12	M21	M22	M31	M32
$\hat{\beta}_0$:Intercepto	-1.529(0.101)	-1.612(0.139)	-1.317(0.097)	-1.343(0.130)	-1.171(0.097)	-1.177(0.132)
$\hat{\beta}_1$:tempo$_{t2}$	0.820(0.133)	0.828(0.134)	0.757(0.129)	0.763(0.129)	0.614(0.131)	0.617(0.131)
$\hat{\beta}_2$:tempo$_{t3}$	0.439(0.130)	0.446(0.130)	0.301(0.129)	0.301(0.129)	0.017(0.137)	0.017(0.138)
$\hat{\beta}_3$:grupo$_{g2}$		0.161(0.158)		0.158(0.151)		0.210(0.155)
$\hat{\beta}_4$:grupo$_{g3}$		0.124(0.148)		0.054(0.143)		0.038(0.148)
$\hat{\beta}_5$:grupo$_{g4}$		0.041(0.147)		-0.095(0.143)		-0.194(0.151)
$\hat{\sigma}$	0.499(0.015)	0.498(0.015)	0.301(0.014)	0.299(0.014)	0.285(0.019)	0.284(0.020)

FONTE: O autor 2023.

Ao analisar os resultados exibidos em TABELA 4 e TABELA 5 e lembrando que o preditor linear de cada modelo foi o *logit*, pode-se interpretar os efeitos dos coeficientes em termos da razão de chances, mais conhecida como *odds ratio*. Logo, dado que a variável tempo representa medidas em três momentos distintos após a cirurgia (t1, t2 e t3), observam-se efeitos distintos nos modelos betareg, betaregscale e quasibeta. No modelo betareg, o efeito do tempo, especialmente 12 horas após a cirurgia (t2), é marcante. Para o modelo M11, a *odds ratio* para o tempo t2 é de 2.27, significando um aumento de mais do que o dobro nas chances em comparação

com o tempo base t1. Similarmente, M12 apresenta uma *odds ratio* de 2.29 para t2, reforçando a importância do tempo. As variáveis de grupo adicionadas em M12 têm *odds ratio* de 1.17 para g2, 1.13 para g3 e 1.04 para g4, indicando efeitos mais modestos. No betaregscale, observa-se ainda uma tendência similar, embora os efeitos sejam ligeiramente menores. Em M21, o *odds ratio* para t2 é de 2.13, e em M22, é de 2.14. Isso sugere que, mesmo com a complexidade adicional do modelo, o tempo após a cirurgia permanece um fator crucial. As variáveis de grupo em M22 mostram *odds ratios* de 1.17 para g2, 1.06 para g3 e 0.91 para g4. Por outro lado, o modelo quasibeta mostra um impacto menos acentuado do tempo. Em M31 e M32, as *odds ratios* para t2 são de 1.85 e 1.85, respectivamente, indicando um aumento menos pronunciado nas chances em comparação com os outros modelos. Para as variáveis de grupo em M32, as *odds ratio* são de 1.23 para g2, 1.04 para g3 e 0.82 para g4.

Em resumo, enquanto os modelos betareg e betaregscale indicam um aumento significativo nas chances devido ao tempo, principalmente 12 horas após a cirurgia, o modelo quasibeta sugere um impacto mais moderado. Estes achados destacam a importância do fator tempo na avaliação dos desfechos dos pacientes no período pós-operatório, com diferentes modelos capturando de maneira variada a influência deste fator.

TABELA 6 – ESTATÍSTICAS DO MODELO COM UMA COVARIÁVEL NAS TRÊS ABORDAGENS PROPOSTAS

Modelo	Id	Variável	Estimativa	$IC_{2.5\%}$	$IC_{97.5\%}$	EP	t-Valor	p-Valor
betareg	M11	$\hat{\beta}_0$:Intercepto	-1.5289	-1.7271	-1.3306	0.1012	-15.1148	0.0000
betareg	M11	$\hat{\beta}_1$:tempo$_{t2}$	0.8196	0.5580	1.0812	0.1335	6.1415	0.0000
betareg	M11	$\hat{\beta}_2$:tempo$_{t3}$	0.4393	0.1854	0.6932	0.1296	3.3911	0.0008
betareg	M11	$\hat{\phi}$	0.4988	0.4695	0.5281	0.0150	33.3475	0.0000
betaregscale	M21	$\hat{\beta}_0$:Intercepto	-1.3171	-1.5075	-1.1266	0.0972	-13.5549	0.0000
betaregscale	M21	$\hat{\beta}_1$:tempo$_{t2}$	0.7571	0.5041	1.0100	0.1291	5.8657	0.0000
betaregscale	M21	$\hat{\beta}_2$:tempo$_{t3}$	0.3006	0.0470	0.5542	0.1294	2.3228	0.0206
betaregscale	M21	$\hat{\phi}$	0.3012	0.2736	0.3288	0.0141	21.3879	0.0000
quasibeta	M31	$\hat{\beta}_0$:Intercepto	-1.1706	-1.3615	-0.9796	0.0974	-12.0137	0.0000
quasibeta	M31	$\hat{\beta}_1$:tempo$_{t2}$	0.6143	0.3585	0.8702	0.1305	4.7066	0.0000
quasibeta	M31	$\hat{\beta}_2$:tempo$_{t3}$	0.0166	-0.2529	0.2861	0.1375	0.1206	0.9040
quasibeta	M31	$\hat{\phi}$	0.2848	0.2479	0.3217	0.0188	15.1246	0.0000

FONTE: O autor 2023.

Dado que o efeito de grupo não foi significativo, optou-se por seguir com modelos apenas com o tempo. A TABELA 4 adiciona mais profundidade a essa análise dos modelos sem o efeito de grupo. O modelo betareg (M11) para o tempo t2 apresenta uma estimativa de 0.8196 com um p-valor extremamente baixo (0.0000) e um EP de 0.1335, reforçando a significância estatística dessa variável. De forma similar, no

modelo betaregscale (M21), a estimativa para o tempo t2 é de 0.7571, também com um p-valor de 0.0000 e um EP de 0.1291, confirmando a importância do tempo. No modelo quasibeta (M31), a influência do tempo é menos acentuada, mas ainda significativa. A estimativa para o tempo t2 é de 0.6143 com um p-valor de 0.0000 e um EP de 0.1305, indicando um efeito relevante, porém menos pronunciado do que nos modelos betareg e betaregscale.

Em conclusão, os modelos betareg e betaregscale, especialmente este último, demonstram um ajuste notável e uma significância estatística clara na modelagem do efeito do tempo após a cirurgia. Embora o modelo quasibeta também capture essa influência, ele o faz de maneira menos expressiva mesmo tendo sido o mais forte nas simulações estatísticas. Essas análises sublinham a relevância do fator tempo nos desfechos dos pacientes no período pós-operatório, sendo mais eficientemente capturado pelos modelos betaregscale e betareg.

5 CONCLUSÕES E TRABALHOS FUTUROS

Este estudo proporcionou uma análise abrangente dos modelos betareg, betaregscale e quasibeta, aplicados tanto a dados reais de dor em pacientes pós-cirurgia de joelho quanto em simulações estatísticas. As conclusões extraídas dessas análises fornecem insights valiosos sobre a aplicabilidade e robustez de cada modelo em contextos diferentes.

Na análise das simulações, o modelo quasibeta (M3) destacou-se por seu mínimo viés e alta acurácia em diversos cenários, superando os modelos betareg e betaregscale em termos de viés, acurácia e cobertura das estimativas. Na análise dos dados reais, o modelo betaregscale apresentou um ajuste superior, como indicado pelos menores valores de pAIC (-30.022 para M21 e -30.841 para M22). Este modelo demonstrou eficiência na captura da influência do tempo pós-operatório, com odds ratios significativos e baixos p-valores nas estimativas para a variável tempo. Ao relacionar as análises de dados reais e simulações, o betaregscale (M2) se mostrou mais adequado para conjuntos de dados específicos, como no caso da dor pós-cirúrgica, enquanto o quasibeta oferece uma generalização mais ampla e é menos suscetível a viés e imprecisões em diferentes cenários.

Adicionalmente valeu pontuar que modelo quasibeta (M3) é uma excelente recomendação para bases de dados grandes, típicas na era do Big Data. Sua independência da complexidade da distribuição beta o torna menos suscetível a erros numéricos e complexidades computacionais. Já O modelo betaregscale é considerado o padrão ouro para análises precisas em bases de dados menores. Seu design focado na captura completa do efeito da variável resposta, através da modelagem com uso da distribuição beta acumulada, o torna altamente eficaz nesses cenários. No entanto, é importante notar que o tempo de processamento do betaregscale é consideravelmente maior do que o do quasibeta, sendo um fator a ser considerado especialmente em grandes conjuntos de dados.

Considerando a robustez, acurácia e a aplicabilidade em cenários variados, o ranking dos modelos do mais ao menos robusto é:

1. Modelo quasibeta (M3): Demonstrou mínimo viés e alta acurácia nas simulações, sendo robusto em diversos cenários.

2. Modelo betaregscale (M2): Apresentou o melhor ajuste em dados reais, com significância estatística e eficiência na captura do efeito do tempo.

3. Modelo betareg (M1): Embora eficaz em certos aspectos, mostrou-se menos

robusto em comparação com os outros modelos.

Em conclusão, esta pesquisa revela que a escolha do modelo de regressão mais adequado depende do contexto específico da análise. O quasibeta é mais robusto e generalizável, ideal para grandes conjuntos de dados na era do Big Data, enquanto o betaregscale se destaca em situações específicas com alta precisão e eficácia, sendo mais adequado para bases de dados menores.

5.1 TRABALHOS FUTUROS

No ambiente atual, altamente orientado para dados, a eficácia e precisão na modelagem de dados beta-mapeáveis tornam-se cada vez mais cruciais. Uma área primordial de pesquisa futura é aprofundar a modelagem de dados com dispersão variável. Apesar dos avanços alcançados pelos modelos **betareg** (M1), **betaregscale** (M2) e **quasibeta** (M3), existem oportunidades significativas para desenvolver métodos que capturem melhor a heterogeneidade de variância, uma necessidade urgente dada a complexidade crescente dos conjuntos de dados.

Outro campo que merece atenção detalhada é o diagnóstico de modelos e a análise de resíduos. As ferramentas atuais, embora úteis, são insuficientes para fornecer uma compreensão completa do comportamento dos resíduos e da adequação dos modelos. Pesquisas inovadoras que ofereçam métodos mais avançados e detalhados para análise de resíduos são fundamentais para melhorar a precisão e confiabilidade das modelagens.

Além disso, à medida que nos deparamos com conjuntos de dados de tamanho e complexidade crescentes, o custo computacional torna-se uma preocupação cada vez mais premente. Há uma necessidade imperiosa de desenvolver métodos computacionalmente eficientes que possam gerenciar grandes volumes de dados sem comprometer a precisão ou a velocidade.

Finalmente, a aplicabilidade dos modelos beta-mapeáveis em diferentes campos, como medicina, ciências sociais e engenharia, oferece um vasto espaço para exploração. É vital que pesquisas futuras se concentrem em validar e adaptar esses modelos para variados contextos, ampliando assim sua utilidade e impacto.

Portanto, o campo da modelagem de dados beta-mapeáveis está repleto de oportunidades para pesquisa e desenvolvimento, prometendo avanços significativos na nossa capacidade de entender e utilizar dados de maneira eficiente e eficaz.

REFERÊNCIAS

AGRESTI, A. **Categorical Data Analysis**. [S.l.]: Wiley, 2002. Citado 1 vez na página 15.

AKAIKE, H. A new look at the statistical model identification. **Information theory and statistics**, p. 267–281, 1974. Citado 1 vez na página 49.

ANDRADE, A. C. G. d. **Efeitos da especificação incorreta da função de ligação no modelo de regressão beta**. 2007. Tese (Doutorado) – Universidade de São Paulo. Citado 1 vez na página 42.

AZZALINI, A. **Monographs on Statistics and Applied Probability**. v. 68. [S.l.]: Chapman & Hall/CRC Boca Raton, New York, 1996. Citado 1 vez na página 45.

BEM, D. J. Writing the empirical journal article. **Handbook of research methods in experimental psychology**, Blackwell, p. 3–18, 2003. Citado 1 vez na página 55.

BENZON, H.; RAJA, S. N.; FISHMAN, S. E.; LIU, S. S.; COHEN, S. P. **Essentials of pain medicine E-book**. [S.l.]: Elsevier Health Sciences, 2011. Citado 1 vez na página 18.

BIGGERSTAFF, B. J.; JACKSON, M. L. A comparison of logistic regression models with alternative link functions for predicting small-area influenza prevalence. **PLoS ONE**, Public Library of Science, v. 3, n. 12, e3853, 2008. Citado 1 vez na página 15.

BOGAERTS, K.; KOMÁREK, A.; LESAFFRE, E. **Survival analysis with interval-censored data: a practical approach with examples in R, SAS, and BUGS**. [S.l.]: Chapman e Hall/CRC, 2017. Citado 1 vezes nas páginas 20, 30.

BONAT, W. H.; PETTERLE, R. R.; HINDE, J.; DEMÉTRIO, C. G. Flexible quasi-beta regression models for continuous bounded data. **Statistical Modelling**, SAGE Publications Sage India: New Delhi, India, v. 19, n. 6, p. 617–633, 2019. Citado 3 vezes nas páginas 8, 9, 52.

BONAT, W. H. Multiple Response Variables Regression Models in R: The mcglm Package. **Journal of Statistical Software**, v. 84, n. 4, p. 1–30, 2018. DOI: 10.18637/jss.v084.i04. Citado 1 vez na página 54.

BONAT, W. H. Multiple response variables regression models in R: The mcglm package. **Journal of Statistical Software**, v. 84, p. 1–30, 2018. Citado 2 vezes nas páginas 16, 53.

BONAT, W. H.; JØRGENSEN, B. Multivariate covariance generalized linear models. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 65, n. 5, p. 649–675, 2016. Citado 8 vezes nas páginas 8, 9, 16, 20, 50, 52–54.

BONAT, W. H.; RIBEIRO JR, P. J.; ZEVIANI, W. M. Likelihood analysis for a class of beta mixed models. **Journal of Applied Statistics**, Taylor & Francis, v. 42, n. 2, p. 252–266, 2015. Citado 1 vez na página 16.

BONAT, W. H.; RIBEIRO JR, P.; ZEVIANI, W. M. Regression models with responses on the unit interval: specification, estimation and comparison. **Biometric Brazilian Journal**, v. 30, n. 4, p. 415–431, 2012. Citado 1 vez na página 16.

BORGAN, O.; GOLDSTEIN, L.; LANGHOLZ, B. Methods for the analysis of sampled cohort data in the Cox proportional hazards model. **The Annals of Statistics**, JSTOR, p. 1749–1778, 1995. Citado 1 vez na página 17.

BREHM, J.; RAHN, W. Individual-level evidence for the causes and consequences of social capital. **American Journal of Political Science**, v. 37, n. 2, p. 450–466, 1993. DOI: 10.2307/2111564. Citado 1 vez na página 39.

BROYDEN, C. G. The convergence of a class of double-rank minimization algorithms. **Journal of the Institute of Mathematics and its Applications**, Oxford University Press, v. 6, n. 1, p. 76–90, 1970. Citado 1 vez na página 45.

BYRD, R. H.; LU, P.; NOCEDAL, J.; ZHU, C. Limited memory BFGS method for large scale optimization. **Mathematical programming**, Springer, v. 45, n. 1-3, p. 503–528, 1995. Citado 1 vez na página 45.

CARIFIO, J.; PERLA, R. J. Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes. **Journal of Social Sciences**, v. 3, n. 3, p. 106–116, 2007. Citado 3 vezes nas páginas 17, 28, 29.

CASELLA, G.; BERGER, R. L. **Statistical inference**. [S.l.]: Cengage Learning, 2021. Citado 3 vezes nas páginas 42, 45, 46.

CHEN, C.-L.; HUANG, Y.-C.; LIU, Y.-W.; LIN, Y.-C.; LEE, W.-L. Multiple regression analysis for mortality prediction in medical research. **Journal of the Formosan Medical Association**, Elsevier, v. 119, n. 5, p. 1068–1076, 2020. Citado 1 vez na página 14.

COLLETT, D. **Modelling survival data in medical research**. [S.l.]: CRC press, 2015. Citado 1 vez na página 49.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. [S.l.]: Editora Blucher, 2006. Citado 0 vez na página 30.

COX, D. R. Regression models and life-tables. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972. Citado 1 vez na página 17.

DASGUPTA, A. **Asymptotic theory of statistics and probability**. [S.l.]: Springer, 2008. v. 180. Citado 1 vez na página 45.

DELIGNETTE-MULLER, M. L.; DUTANG, C. fitdistrplus: An R Package for Fitting Distributions. **Journal of Statistical Software**, v. 64, n. 4, p. 1–34, 2015. DOI: 10.18637/jss.v064.i04. Citado 1 vez na página 43.

DOMINICI, F.; PENG, R. D.; BELL, M. L.; PHAM, L.; MCDERMOTT, A.; ZEGER, S. L.; SAMET, J. M. Air pollution and hospital admissions for cardiovascular and respiratory diseases in Rome, Italy. **American Journal of Respiratory and Critical Care Medicine**, American Thoracic Society, v. 173, n. 6, p. 656–661, 2006. Citado 1 vez na página 16.

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 3 vez na página 51.

ESPINHEIRA, P. L.; FERRARI, S. L.; CRIBARI-NETO, F. On beta regression residuals. **Journal of Applied Statistics**, Taylor & Francis, v. 35, n. 4, p. 407–419, 2008. Citado 3 vez na página 51.

ESPINHEIRA, P.; MEYER, R. A beta regression model for earthquake magnitude distribution. **Journal of Applied Statistics**, v. 35, n. 5, p. 515–531, 2008. DOI: 10.1080/02664760802084900. Citado 1 vez na página 39.

_____. Bayesian inference for beta regression models with parametric mean and dispersion functions. **Computational Statistics and Data Analysis**, v. 52, n. 7, p. 3629–3640, 2008. DOI: 10.1016/j.csda.2007.12.016. Citado 1 vez na página 39.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of applied statistics**, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004. Citado 6 vezes nas páginas 16, 20, 34, 50.

FISHER, R. A. On the mathematical foundations of theoretical statistics. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, The Royal Society, v. 222, n. 594-604, p. 309–368, 1922. Citado 1 vez na página 14.

GALTON, F. Regression towards mediocrity in hereditary stature. **The Journal of the Anthropological Institute of Great Britain and Ireland**, v. 15, p. 246–263, 1886. Citado 1 vez na página 14.

GENTLEMAN, R.; GEYER, C. J. Maximum likelihood for interval censored data: Consistency and computation. **Biometrika**, Oxford University Press, v. 81, n. 3, p. 618–623, 1994. Citado 1 vez na página 20.

GILBERT, P.; VARADHAN, R.; GILBERT, M. P. Package 'numDeriv'. **differential equations**, Citeseer, v. 3, p. 203–267, 2009. Citado 1 vez na página 44.

GOULET, J.; BUTA, E.; CARROLL, C.; BRANDT, C. (124) Statistical methods for the analysis of NRS pain data. **The Journal of Pain**, Elsevier, v. 16, n. 4, s7, 2015. Citado 1 vez na página 23.

HANCOX, R. J.; POULTON, R. Watching television is associated with childhood obesity: but is it clinically important? **International Journal of Obesity**, v. 34, n. 1, p. 1–3, 2010. DOI: 10.1038/ijo.2009.260. Citado 1 vez na página 39.

HASTIE, T.; TIBSHIRANI, R. **Generalized additive models**. [S.l.]: CRC press, 1990. Citado 1 vez na página 16.

HEDEKER, D.; GIBBONS, R. D. **Longitudinal data analysis**. [S.l.]: John Wiley & Sons, 2006. Citado 1 vez na página 15.

HELSEL, D. R. et al. **Nondetects and data analysis. Statistics for censored environmental data.** [S.l.]: Wiley-Interscience, 2005. Citado 3 vezes nas páginas 18, 20, 43.

IBRAGIMOV, I. A.; HAS' MINSKII, R. Z. **Statistical estimation: asymptotic theory**. [S.l.]: Springer Science & Business Media, 2013. v. 16. Citado 1 vez na página 45.

JAMIESON, S. Likert scales: how to (ab)use them. **Medical Education**, v. 38, n. 12, p. 1217–1218, 2004. Citado 2 vezes nas páginas 17, 28.

JOHNSON, N. L.; KOTZ, S.; BALAKRISHNAN, N. **Continuous univariate distributions, volume 2**. [S.l.]: John wiley & sons, 1995. v. 289. Citado 1 vez na página 34.

JØRGENSEN, B.; KNUDSEN, S. J. Parameter orthogonality and bias adjustment for estimating functions. **Scandinavian Journal of Statistics**, Wiley Online Library, v. 31, n. 1, p. 93–114, 2004. Citado 2 vezes nas páginas 52, 53.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American statistical association**, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958. Citado 1 vez na página 17.

KENT, J. T. Robust properties of likelihood ratio tests. **Biometrika**, Oxford University Press, v. 69, n. 1, p. 19–27, 1982. Citado 1 vez na página 48.

KIESCHNICK, R.; MCCULLOUGH, B. D. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. **Statistical Modelling**, v. 3, n. 3, p. 193–213, 2003. DOI: 10.1191/1471082x03st056oa. Citado 1 vez na página 39.

KIESCHNICK, R.; MCCULLOUGH, B. D. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. **Statistical modelling**, Sage Publications Sage CA: Thousand Oaks, CA, v. 3, n. 3, p. 193–213, 2003. Citado 1 vez na página 16.

KLEIN, J. P.; MOESCHBERGER, M. L. **Survival analysis: techniques for censored and truncated data**. [S.l.]: Springer, 2003. v. 1230. Citado 3 vezes nas páginas 18, 20, 43.

KLEINBAUM, D. G.; KLEIN, M. Kaplan-Meier survival curves and the log-rank test. In: SURVIVAL analysis. [S.l.]: Springer, 2012. P. 55–96. Citado 1 vez na página 17.

_____. Parametric survival models. In: SURVIVAL analysis. [S.l.]: Springer, 2012. P. 289–361. Citado 1 vez na página 17.

LEHMANN, E. L.; LEHMANN, E. **Testing statistical hypotheses**. [S.l.]: Springer, 1986. v. 2. Citado 1 vez na página 47.

LI, B.; BABU, G. J. **A graduate course on statistical inference**. [S.l.]: Springer, 2019. Citado 1 vez na página 45.

LOPES, J. E. **MODELOS DE REGRESSÃO BETA PARA DADOS DE ESCALA**. 2023. Tese (Doutorado) – UFPR - Universidade Federal do Paraná. Citado 2 vezes nas páginas 24, 43.

MARIANO BAYER, F. **Modelagem e Inferência em Regressão Beta**. 2011. Tese (Doutorado) – UFPE - Universidade Federal de Pernambuco. Citado 1 vez na página 36.

MARSCHAK, J.; ANDREWS, W. H. Random simultaneous equations and the theory of production. **Econometrica**, JSTOR, v. 12, n. 3, p. 143–205, 1944. Citado 1 vez na página 14.

MERTLER, C. A.; REINHART, R. V. **Advanced and Multivariate Statistical Methods: Practical Application and Interpretation**. [S.l.]: Routledge, 2017. Citado 1 vez na página 55.

MONETTE, G. **Longitudinal data analysis with mixed models**. [S.l.]: York University Summer Program in Data Analysis, 2010. Citado 1 vez na página 15.

NAIR, A. S.; DIWAN, S. Pain scores and statistical analysis—the conundrum. **Ain-Shams Journal of Anesthesiology**, SpringerOpen, v. 12, n. 1, p. 1–2, 2020. Citado 2 vezes nas páginas 22, 23.

NASH, S. G. **Newton-type minimization via the Lanczos method**. [S.l.]: SIAM, 1984. Citado 1 vez na página 44.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. **Journal of the Royal Statistical Society: Series A (General)**, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972. Citado 1 vez na página 15.

NOCEDAL, J.; WRIGHT, S. J. **Numerical optimization**. [S.l.]: Springer Science & Business Media, 2006. Citado 2 vez na página 45.

NORMAN, G. Likert scales, levels of measurement and the “laws” of statistics. **Advances in Health Sciences Education**, v. 15, n. 5, p. 625–632, 2010. Citado 2 vezes nas páginas 17, 28.

OSPINA, R.; FERRARI, S. A general class of zero-or-one inflated beta regression models. **Computational Statistics and Data Analysis**, v. 50, n. 8, p. 2028–2050, 2006. DOI: 10.1016/j.csda.2005.02.007. Citado 1 vez na página 39.

PATTEN, M. L. **Proposing Empirical Research: A Guide to the Fundamentals**. [S.l.]: Routledge, 2014. Citado 1 vez na página 55.

PEARSON, K. On lines and planes of closest fit to systems of points in space. **Philosophical Magazine**, Taylor & Francis, v. 2, n. 11, p. 559–572, 1901. Citado 1 vez na página 14.

PEREIRA, G. H. On quantile residuals in beta regression. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 48, n. 1, p. 302–316, 2019. Citado 3 vezes nas páginas 51, 52.

PETERSEN, T. Fitting parametric survival models with time-dependent covariates. **Journal of the Royal Statistical Society Series C: Applied Statistics**, Oxford University Press, v. 35, n. 3, p. 281–288, 1986. Citado 1 vez na página 17.

QIU, Z.; SONG, P. X.-K.; TAN, M. Simplex mixed-effects models for longitudinal proportional data. **Scandinavian Journal of Statistics**, Wiley Online Library, v. 35, n. 4, p. 577–596, 2008. Citado 1 vez na página 16.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2023. Disponível em: https://www.R-project.org/. Citado 1 vez na página 44.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2023. Disponível em: https://www.R-project.org/. Citado 1 vez na página 57.

RAJA, S.; LIU, S. S.; FISHMAN, S.; COHEN, S. P. **Essentials of pain medicine**. [S.l.]: Elsevier/Saunders, 2018. Citado 2 vezes nas páginas 24, 26, 27.

RIGBY, R. A.; STASINOPOULOS, D. M. **Generalized additive models for location, scale and shape**. [S.l.]: CRC press, 2005. Citado 1 vez na página 16.

SATORRA, A.; SARIS, W. E. Power of the likelihood ratio test in covariance structure analysis. **Psychometrika**, Springer, v. 50, n. 1, p. 83–90, 1985. Citado 1 vez na página 48.

SCHWARZ, G. E. A Bayesian criterion for the selection of variables in linear regression. **Journal of the American Statistical Association**, Taylor & Francis, v. 73, n. 364, p. 59–66, 1978. Citado 1 vez na página 49.

SHAHBAZ, M.; TAHIR, M. I.; AHMAD, A.; AL., et. Energy consumption and economic growth in ASEAN countries: evidence from panel data analysis. **Environmental Science and Pollution Research**, Springer, v. 27, p. 32051–32066, 2020. Citado 1 vez na página 14.

SIMAS, A. B.; BARRETO-SOUZA, W.; ROCHA, A. V. Beta regression for modelling rates and proportions. **Journal of Applied Statistics**, v. 37, n. 7, p. 1297–1315, 2010. DOI: 10.1080/02664760902962426. Citado 1 vez na página 39.

SMITHSON, M. Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. **Educational and Psychological Measurement**, v. 66, n. 2, p. 218–227, 2006. DOI: 10.1177/0013164405278555. Citado 1 vez na página 39.

STREINER, D. L.; NORMAN, G. R.; CAIRNEY, J. **Health Measurement Scales: A practical guide to their development and use**. [S.l.]: Oxford University Press, 2015. Citado 8 vezes nas páginas 17, 28–30.

STUART, M. **Understanding robust and exploratory data analysis**. [S.l.]: Wiley Online Library, 1984. Citado 1 vez na página 31.

SULLIVAN, G. M.; ARTINO, A. R. Analyzing and interpreting data from Likert-type scales. **Journal of Graduate Medical Education**, v. 5, n. 4, p. 541–542, 2013. Citado 2 vezes nas páginas 28, 30.

TSENG, K.-M.; LI, Y.-C.; LEE, H.-H.; CHANG, C.-Y. A generalized linear model for analysis of panel count data with excess zeros and repeated measures. **Journal of Biopharmaceutical Statistics**, Taylor & Francis, v. 27, n. 2, p. 214–229, 2017. Citado 1 vez na página 15.

WALD, A. Contributions to the theory of statistical estimation and testing hypotheses. **The Annals of Mathematical Statistics**, JSTOR, v. 10, n. 4, p. 299–326, 1939. Citado 2 vez na página 47.

WINTER, J. C. F. de; DODOU, D. Five-Point Likert Items: t test versus Mann-Whitney-Wilcoxon. **Practical Assessment, Research & Evaluation**, v. 15, n. 1, p. 11, 2010. Citado 3 vez na página 29.

WOOD, S. N. **Generalized Additive Models: An Introduction with R**. [S.l.]: CRC press, 2017. Citado 1 vez na página 16.

WOOLF, B. The log likelihood ratio test (the G-test). **Annals of human genetics**, Wiley Online Library, v. 21, n. 4, p. 397–409, 1957. Citado 1 vez na página 48.

ZUCCO, C.; THOMASSEN, J.; SCHMITT, H. A Europeanized Politics? European Integration and Political Parties in Europe. **West European Politics**, v. 31, n. 1-2, p. 244–266, 2008. DOI: 10.1080/01402380701829691. Citado 1 vez na página 39.

ANEXO 1 – PACOTE BETAREGSCALE

O pacote **betaregscale** consiste em uma biblioteca de funções em R projetadas para ajustar modelos de regressão beta a dados derivados de escalas mapeáveis no suporte da distribuição beta. Isso inclui, por exemplo, escalas de dor e *Likert*, entre outras, de maneira que a incerteza do instrumento é avaliada de forma intervalar. Este pacote foi desenvolvido para dar suporte à pesquisa de mestrado sob o tema MODELOS DE REGRESSÃO BETA PARA DADOS DE ESCALA para o PPGMNE/UFPR. Para mais detalhes consulte o site do projeto no GitHub em https://evandeilton.github.io/betaregscale/

Este *framework* acomoda modelos com dispersão fixa ou variável, todos sob o paradigma de máxima verossimilhança. Adicionalmente, oferece funções para simulações e avaliação do desempenho dos modelos no processo de estimação, além de outras para ajuste do modelo a dados reais. O código-fonte e contribuições podem ser acessados no repositório oficial no *GitHub*. Informações detalhadas sobre instalação e uso estão disponíveis na documentação do pacote.

O **betaregscale** é voltado para a modelagem de dados com variável resposta mapeável em intervalo contínuo, i.e., $y = (y_s; y_i)$, incluindo censura à esquerda, direita ou intervalar independente do tempo. Encontra aplicação em pesquisas de opinião, avaliações de produtos, escalas de dor como NRS-11, NRS-21 e NRS-101 ou escalas *Likert*, avaliações de compostos químicos, entre outros. Utilizando a distribuição beta, ele acomoda características dos dados em uma estrutura de regressão, associando variáveis explicativas à variável resposta intervalar e permitindo preditores lineares para coeficientes relacionados à média e à dispersão, fornecendo estimativas robustas e confiáveis dos parâmetros do modelo.

PRINCIPAIS FUNCIONALIDADES

Entre as principais funcionalidades tem-se:

- Ajuste de modelos de regressão beta com dispersão fixa e variável.
- Funções para simulação de dados, permitindo a avaliação do desempenho dos modelos em diferentes cenários.
- Estatística de bondade do ajuste como AIC e BIC, por exemplo em gof().

- Funções genéricas como *coef*, *vcov*, *fitted*, *residuals*, *summary* e *print* foram implementadas para a classe **betaregscale** para facilitar o acesso às medidas do ajuste.
- Funções para ajuste e comparação de modelos com diferentes combinações de variáveis explicativas tanto para μ como ϕ.

Acesse a documentação detalhada de cada função e exemplos de uso neste site para obter informações sobre como utilizar o pacote **betaregscale** em suas análises.

ANEXO 2 – CÓDIGOS DA ANÁLISE DE DADOS

```r
## ------------------------------------------------------------------- ##
## titulo:  MODELOS DE REGRESSAO BETA PARA DADOS DE ESCALA
## autor:   Jose Evandeilton Lopes
## local:   Curitiba
## code:    Modelagem
## ------------------------------------------------------------------- ##

## Extras
require(tidyverse)
require(patchwork)
require(latex2exp)

aux_coef_mcglm <- function(fit){
  e <- summary(fit, verbose = FALSE, print = FALSE)[[1]]
  p <- e$tau;rownames(p) <- paste0("phi", 1:nrow(p))
  i <- confint(fit)
  a <- cbind(rbind(e$Regression, p), i)
  b <- data.frame(variable = rownames(a),
  estimate = a$Estimates,
  ci_lower = a$`0.025%`,
  ci_upper = a$`0.975%`,
  se = a$Std.error,
  t_value = a$`Z value`,
  p_value = a$`Pr(>|z|)`
  )
  return(b)
}

# Calcula log pseudo-verossimilhanca
plogLik_iid <- function(pred, y, dispersion, model = "beta") {
  if(model == "beta") {
    #variancia <- pred*(1-pred)/(1+dispersion)
    variancia <- pred*(1-pred)*dispersion
  }
  if(model == "simplex") {
    variancia <- mean_variance(mu = pred, phi = dispersion,
    model = "simplex")
  }
  if(model == "gaussian") {
    variancia <- sum( ((pred - y)^2)/length(y))
  }
  pll <- sum(dnorm(y, mean = pred, sd = sqrt(variancia), log = TRUE))
  return(pll)
```

```
}

pbic <- function(ll, k, n){
  return(k*log(n)-2*ll)
}

paic <- function(ll, k){
  return(2*k-2*ll)
}

fn_min_max <- function(x){
  xmin <- min(x, na.rm = TRUE)
  xmax <- max(x, na.rm = TRUE)
  o <- (x-xmin)/(xmax-xmin)
  o[round(o, 8) == 0] <- 0.00001
  o[round(o, 8) == 1] <- 0.99999
  attr(o, "min") <- xmin
  attr(o, "max") <- xmax
  return(o)
}

fn_min_max_back <- function(x){
  xmin <- attr(x, "min")
  xmax <- attr(x, "max")
  o <- x*(xmax-xmin) + xmin
  return(o)
}

## Transformacao de mu e sigma para p e q
fnp <- function(mu, phi){
  p <- mu*phi
  q <- (1-mu)*phi
  c(p, q)
}

## Transformacao de mu e lambda para p e q
fnp2 <- function(mu, phi){
  p <- mu*((1-phi)/phi)
  q <- (1-mu)*((1-phi)/phi)
  c(p, q)
}

fn_plot_fitdist <- function(Y){

  f1 <- fitdistrplus::fitdistcens(censdata = Y, distr = "beta")

  g2 <- fitdistrplus::cdfcompcens(f1,
```

```
    plotstyle = "ggplot",
    xlab = "Dado censurado [dor]",
    ylab = "F[y]",
    main = "(A)",
    Turnbull.confint = FALSE) +
    theme(strip.text.x = element_blank(), legend.position="none")

  g3 <- fitdistrplus::qqcompcens(f1,
    plotstyle = "ggplot",
    xlab = "Quantis teoricos",
    ylab = "Quantis empiricos",
    main = "(B)") +
    theme(strip.text.x = element_blank(), legend.position="none")

  g4 <- fitdistrplus::ppcompcens(f1,
    plotstyle = "ggplot",
    xlab = "Probabilidades teoricas",
    ylab = "Probabilidades empiricas",
    main = "(C)") +
    theme(strip.text.x = element_blank(), legend.position="none")
  return(g2 + g3 + g4)
}

fn_plot_fitdist1 <- function(Y, main = "(A)"){
  f1 <- fitdistrplus::fitdistcens(censdata = Y, distr = "beta")

  fitdistrplus::cdfcompcens(f1,
    plotstyle = "ggplot",
    xlab = "Dado censurado [dor]",
    ylab = "F[y]",
    main = main,
    Turnbull.confint = FALSE) +
  theme(strip.text.x = element_blank(), legend.position="none")
}

plot_phi_fixo_bias <-  function(dados, bias, lower, upper, limits = c
    (-2, 2), titulo = NULL){
  g <- dados %>%
  dplyr::as_tibble() %>%
  dplyr::mutate(phi = as.numeric(as.character(phi)),
  ncuts_ = as.numeric(as.character(ncuts))) %>%
  #dplyr::filter(type == 'm') %>%
  dplyr::mutate(beta_latex = paste0("$\\beta_{", substr(as.character(
    coefs),2,2),"}$"),
  phi_latex = paste0("$\\sigma_{", format(as.numeric(as.character(phi)),
    nsmall = 3),"}$"),
  nrs_latex = paste0("$NRS_{", as.numeric(as.character(ncuts))+1, "}$"))
```

```
    %>%
  dplyr::group_by(n, phi) %>%
  ggplot(aes(n, {{bias}}))

  g +
  geom_pointrange(aes(ymin = {{lower}}, ymax = {{upper}}), size=0.2,
   linewidth = 0.3, show.legend = FALSE) +
  labs(title = titulo) +
  ylab(expression(hat(bias) %+-% SD)) +
  xlab("Tamanho da amostra")+
  ggh4x::facet_nested(reorder(TeX(nrs_latex, output = "char"), ncuts_) ~
    reorder(TeX(phi_latex, output = "char"), phi),
  switch = "y", labeller = label_parsed)+
  theme_bw() +
  theme(axis.text.x = element_text(size = 8),
  axis.text.y = element_text(size = 7),
  strip.text.x = element_text(size = 10, margin = margin(0.1, 0, 0.1, 0,
    "cm")),
  strip.text.y = element_text(size = 10, margin = margin(0.1, 0, 0.1, 0,
    "cm")),
  panel.spacing.x = unit(0.12, "lines"),
  panel.spacing.y = unit(0.12,"lines"),
  plot.title = element_text(hjust = 0.5)) +
  coord_flip() +
  geom_hline(yintercept = 0.0, linetype="dotted")+
  scale_y_continuous(limits = limits)
  #ylab(label = expression(hat(MAPE) + SD)) +
  #xlab(label = "Tamanho da amostra")
}

plot_beta_coverage <- function(dados, coverage, yintercept = 0.95,
   limits = c(0.7, 1.0)){
  g <- dados %>%
  dplyr::mutate(phi = as.numeric(as.character(phi)),
  ncuts_ = as.numeric(as.character(ncuts))) %>%
  dplyr::mutate(beta_latex = paste0("$\\beta_{",substr(as.character(
   coefs),2,2),"}$"),
  phi_latex  = paste0("$\\sigma_{", phi,"}$"),
  nrs_latex  = paste0("$NRS_{", as.numeric(as.character(ncuts))+1, "}$")
   ) %>%
  ggplot(aes(x = n, y = {{coverage}}, group = factor(phi)))

  g <- g +
  geom_line(aes(linetype=factor(phi)), linewidth = 0.3) +
  geom_point(aes(shape=factor(phi))) +
  scale_shape_manual(values = c(1, 2, 22, 19, 17))+
  ggh4x::facet_nested(reorder(TeX(nrs_latex, output = "char"), ncuts_) +
```

```
    reorder(TeX(beta_latex, output = "char"), phi) ~ modelo,
  switch = "y", labeller = label_parsed) +
  geom_hline(yintercept = yintercept, linewidth = 0.2) +
  #guides(shape=guide_legend(title = expression(gamma)))+
  guides(shape=guide_legend(title = expression(sigma)),
  linetype=guide_legend(title = expression(sigma)), label.vjust = -2) +
  theme_bw() +
  theme(axis.text.x = element_text(size = 8),
  axis.text.y = element_text(size = 7),
  strip.text.x = element_text(size = 10, margin = margin(0.1, 0, 0.1, 0,
    "cm")),
  strip.text.y = element_text(size = 10, margin = margin(0.1, 0, 0.1, 0,
    "cm")),
  panel.spacing.x = unit(0.12, "lines"),
  panel.spacing.y = unit(0.12,"lines"),
  plot.title = element_text(hjust = 0.5),
  legend.position = "top"
  #legend.spacing.y = unit(5, 'cm')
  #legend.spacing.y = element_text(size = 30, margin = margin(10, "cm"))
  ) +
  scale_y_continuous(limits = limits, n.breaks = 4) +
  ylab(NULL) +
  xlab("Tamanho da amostra")
  return(g)
}

plot_phi_coverage <- function(dados, coverage, yintercept = 0.95, limits
    = c(0.7, 1.0)){
  g <- dados %>%
  dplyr::mutate(phi = as.numeric(as.character(phi)),
  ncuts_ = as.numeric(as.character(ncuts))) %>%
  dplyr::mutate(beta_latex = paste0("$\\beta_{",substr(as.character(
   coefs),2,2),"}$"),
  phi_latex  = paste0("$\\sigma_{", phi,"}$"),
  nrs_latex  = paste0("$NRS_{", as.numeric(as.character(ncuts))+1, "}$")
   ) %>%
  ggplot(aes(x = n, y = {{coverage}}, group = factor(phi)))

  g <- g +
  geom_line(aes(linetype=factor(phi)), linewidth = 0.3) +
  geom_point(aes(shape=factor(phi))) +
  scale_shape_manual(values = c(1, 2, 22, 19, 17))+
  ggh4x::facet_nested(reorder(TeX(nrs_latex, output = "char"), ncuts_) ~
    modelo,
  switch = "y", labeller = label_parsed) +
  geom_hline(yintercept = yintercept, linewidth = 0.2) +
  #guides(shape=guide_legend(title = expression(gamma)))+
```

```
  guides(shape=guide_legend(title = expression(sigma)),
  linetype=guide_legend(title = expression(sigma)), label.vjust = -2) +
  theme_bw() +
  theme(axis.text.x = element_text(size = 8),
  axis.text.y = element_text(size = 7),
  strip.text.x = element_text(size = 10, margin = margin(0.1, 0, 0.1, 0,
    "cm")),
  strip.text.y = element_text(size = 10, margin = margin(0.1, 0, 0.1, 0,
    "cm")),
  panel.spacing.x = unit(0.12, "lines"),
  panel.spacing.y = unit(0.12,"lines"),
  plot.title = element_text(hjust = 0.5),
  legend.position = "top"
  #legend.spacing.y = unit(5, 'cm')
  #legend.spacing.y = element_text(size = 30, margin = margin(10, "cm"))
  ) +
  scale_y_continuous(limits = limits, n.breaks = 4) +
  ylab(NULL) +
  xlab("Tamanho da amostra")
  return(g)
}

require(tidyverse)
require(betaregscale)
require(patchwork)
require(mcglm)
source("Extras2.R")

## ------------------------------------------------------------------- ##
## Dados
## ------------------------------------------------------------------- ##
da <- readr::read_csv("dados/dados_escala_dor.csv", show_col_types =
    FALSE) %>%
dplyr::mutate(
  y = fn_min_max(Dor),
  tempon = parse_number(Tempo),
  tempoh = dplyr::case_when(
    Tempo == "t1" ~ 6,
    Tempo == "t2" ~ 12,
    Tempo == "t3" ~ 24,
    TRUE ~ NA_real_),
  D1 = dplyr::case_when(
    Dor <= 0  ~ "0.00001;0.05",
    Dor <= 1  ~ "0.05;0.15",
    Dor <= 2  ~ "0.15;0.25",
    Dor <= 3  ~ "0.25;0.35",
    Dor <= 4  ~ "0.35;0.45",
```

```
    Dor <= 5  ~ "0.45;0.55",
    Dor <= 6  ~ "0.55;0.65",
    Dor <= 7  ~ "0.65;0.75",
    Dor <= 8  ~ "0.75;0.85",
    Dor <= 9  ~ "0.85;0.95",
    Dor <= 10 ~ "0.95;0.99999",
    TRUE ~ NA_character_),
  ntrial = 1) %>%
  tidyr::separate(col = D1, into = c("left","right"), sep = ";") %>%
  janitor::clean_names() %>%
  dplyr::mutate_at(c("left","right"), "as.numeric") %>%
  as.data.frame() %>%
  dplyr::filter(!(grupo == 'g1' & tempo == 't2' & dor == 10))

## ------------------------------------------------------------------- ##
## 1. M1:Beta convencional com dispersao fixa
## 2. M2:Beta intervalar com dispersao fixa
## 3. M3:Quasi beta - MCGLM
## ------------------------------------------------------------------- ##

## Parametros
f1 <- dor ~ tempo
f2 <- dor ~ tempo + grupo
controle <- list(tuning = 1.0, max_iter = 20, tol = 1e-04, verbose =
    FALSE)

link1 <- "logit"
link2 <- "identity"
ncuts <- 10

## ------------------------------------------------------------------- ##
## Estimativas
## ------------------------------------------------------------------- ##
fn_ests_dois_modelos <- function(){
  modelos <- list(
  M11 = betaregscale::betaregscale(formula = f1, dados = da, ncuts =
    ncuts,
  link = link1, link_phi = link2, type = "m", repar = "2",
  acumulada = FALSE, method = "L-BFGS-B"),

  M12 = betaregscale::betaregscale(formula = f2, dados = da, ncuts =
    ncuts,
  link = link1, link_phi = link2, type = "m", repar = "2",
  acumulada = FALSE, method = "L-BFGS-B"),

  M21 = betaregscale::betaregscale(formula = f1, dados = da, ncuts =
    ncuts,
```

```
  link = link1, link_phi = link2, type = "m", repar = "2",
  acumulada = TRUE, method = "L-BFGS-B"),

  M22 = betaregscale::betaregscale(formula = f2, dados = da, ncuts =
   ncuts,
  link = link1, link_phi = link2, type = "m", repar = "2",
  acumulada = TRUE, method = "L-BFGS-B"),

  M31 = mcglm(linear_pred = c(y ~ tempo), matrix_pred = list(mc_id(da)),
  link = c("logit"), covariance = "identity", variance = c("binomialP"),
  Ntrial = list(da$ntrial), control_algorithm = controle, data = da),

  M32 = mcglm(linear_pred = c(y ~ tempo + grupo), matrix_pred = list(mc_
   id(da)),
  link = c("logit"), covariance = "identity", variance = c("binomialP"),
  Ntrial = list(da$ntrial), control_algorithm = controle, data = da)
  )

  ESTS <- dplyr::bind_rows(
  purrr::map_df(modelos[1:4], est, .id = "modelos"),
  purrr::map_df(modelos[5:6], aux_coef_mcglm, .id = "modelos") %>%
  dplyr::mutate(variable = dplyr::case_when(variable == "phi1" ~ "phi",
   TRUE ~ variable))
  )

  BE <- ESTS %>%
  dplyr::transmute(modelos,
  variable,
  est = paste0(formatC(estimate, format = "f", digits = 3), "(",
  formatC(se, format = "f", digits = 3), ")")) |>
  tidyr::pivot_wider(names_from = "modelos", values_from = "est")

  BE <- dplyr::bind_rows(
  BE %>% dplyr::filter(variable != "phi"),
  BE %>% dplyr::filter(variable == "phi")
  )

## ------------------------------------------------------------------- ##
## pseudo log-verossimilhanca, AIC e BIC
## ------------------------------------------------------------------- ##

  p_LIK <- dplyr::bind_cols(
  M11 = plogLik_iid(pred = as.numeric(fitted(modelos$M11)),
  dispersion = modelos$M11$hatphi,
  y = modelos$M11$dados$yt,
  model = "beta"),

```

```
345   M12 = plogLik_iid(pred = as.numeric(fitted(modelos$M12)),
346   dispersion = modelos$M12$hatphi,
347   y = modelos$M12$dados$yt,
348   model = "beta"),
349
350   M21 = plogLik_iid(pred = as.numeric(fitted(modelos$M21)),
351   dispersion = modelos$M21$hatphi,
352   y = apply(modelos$M21$dados[,1:2], MARGIN = 1, mean),
353   model = "beta"),
354
355   M22 = plogLik_iid(pred = as.numeric(fitted(modelos$M22)),
356   dispersion = modelos$M22$hatphi,
357   y = apply(modelos$M22$dados[,1:2], MARGIN = 1, mean),
358   model = "beta"),
359
360   M31 = plogLik(modelos$M31, verbose = FALSE)$plogLik,
361   M32 = plogLik(modelos$M32, verbose = FALSE)$plogLik
362   )
363
364   p_AIC <-  dplyr::bind_cols(
365   M11 = paic(ll = p_LIK$M11, k = length(coef(modelos$M11))),
366   M12 = paic(ll = p_LIK$M12, k = length(coef(modelos$M12))),
367
368   M21 = paic(ll = p_LIK$M21, k = length(coef(modelos$M21))),
369   M22 = paic(ll = p_LIK$M22, k = length(coef(modelos$M22))),
370
371   M31 = as.numeric(pAIC(modelos$M31)),
372   M32 = as.numeric(pAIC(modelos$M32))
373   )
374
375   p_BIC <-  dplyr::bind_cols(
376   M11 = pbic(ll = p_LIK$M11, k = length(coef(modelos$M11)), n = nrow(da)
      ),
377   M12 = pbic(ll = p_LIK$M12, k = length(coef(modelos$M12)), n = nrow(da)
      ),
378
379   M21 = pbic(ll = p_LIK$M21, k = length(coef(modelos$M21)), n = nrow(da)
      ),
380   M22 = pbic(ll = p_LIK$M22, k = length(coef(modelos$M22)), n = nrow(da)
      ),
381
382   M31 = as.numeric(pBIC(modelos$M31)),
383   M32 = as.numeric(pBIC(modelos$M32))
384   )
385
386   pESTS <- round(data.frame(c(rbind(p_LIK, p_AIC, p_BIC))), 3)
387   rownames(pESTS) <- c("pLogLik","pAIC","pBIC")
```

```
388
389   pESTS <- rbind(Gl = rep(c(4,7), 3), pESTS)
390   tb <- data.frame(Medida = row.names(pESTS), pESTS)
391   row.names(tb) <- NULL
392
393   return(list(ESTS = ESTS, BE = BE, LL = tb, modelos = modelos))
394 }
395
396 fn_ests_um_modelo <- function(dados, link = "logit"){
397   modelos <- list(
398   M11 = betaregscale::betaregscale(formula = dor ~ tempo, dados = dados,
        ncuts = 10,
399   link = link, link_phi = "identity", type = "m", repar = "2",
400   acumulada = FALSE, method = "L-BFGS-B"),
401
402   M21 = betaregscale::betaregscale(formula = dor ~ tempo, dados = dados,
        ncuts = 10,
403   link = link, link_phi = "identity", type = "m", repar = "2",
404   acumulada = TRUE, method = "L-BFGS-B"),
405
406   M31 = mcglm(linear_pred = c(y ~ tempo), matrix_pred = list(mc_id(dados
       )),
407   link = c(link), covariance = "identity", variance = c("binomialP"),
408   Ntrial = list(da$ntrial), control_algorithm = controle, data = dados)
409   )
410
411   ESTS <- dplyr::bind_rows(
412   purrr::map_df(modelos[1:2], est, .id = "modelos"),
413   purrr::map_df(modelos[3], aux_coef_mcglm, .id = "modelos") %>%
414   dplyr::mutate(variable = dplyr::case_when(variable == "phi1" ~ "phi",
       TRUE ~ variable))
415   )
416
417   BE <- ESTS %>%
418   dplyr::transmute(modelos,
419   variable,
420   est = paste0(formatC(estimate, format = "f", digits = 3), "(",
421   formatC(se, format = "f", digits = 3), ")")) |>
422   tidyr::pivot_wider(names_from = "modelos", values_from = "est")
423
424   BE <- dplyr::bind_rows(
425   BE %>% dplyr::filter(variable != "phi"),
426   BE %>% dplyr::filter(variable == "phi")
427   )
428
429 ## ------------------------------------------------------------------ ##
430 ## pseudo log-verossimilhanca, AIC e BIC
```

```r
## ------------------------------------------------------------------ ##

  p_LIK <- dplyr::bind_cols(
  M11 = plogLik_iid(pred = as.numeric(fitted(modelos$M11)),
  dispersion = modelos$M11$hatphi,
  y = modelos$M11$dados$yt,
  model = "beta"),

  M21 = plogLik_iid(pred = as.numeric(fitted(modelos$M21)),
  dispersion = modelos$M21$hatphi,
  y = apply(modelos$M21$dados[,1:2], MARGIN = 1, mean),
  model = "beta"),

  M31 = plogLik(modelos$M31, verbose = FALSE)$plogLik
  )


  p_AIC <-  dplyr::bind_cols(
  M11 = paic(ll = p_LIK$M11, k = length(coef(modelos$M11))),

  M21 = paic(ll = p_LIK$M21, k = length(coef(modelos$M21))),

  M31 = as.numeric(pAIC(modelos$M31))
  )

  p_BIC <-  dplyr::bind_cols(
  M11 = pbic(ll = p_LIK$M11, k = length(coef(modelos$M11)), n = nrow(
    dados)),

  M21 = pbic(ll = p_LIK$M21, k = length(coef(modelos$M21)), n = nrow(
    dados)),

  M31 = as.numeric(pBIC(modelos$M31))
  )

  pESTS <- round(data.frame(c(rbind(p_LIK, p_AIC, p_BIC))), 3)
  rownames(pESTS) <- c("pLogLik","pAIC","pBIC")

  pESTS <- rbind(Gl = rep(c(4), 3), pESTS)
  tb <- data.frame(Medida = row.names(pESTS), pESTS)
  row.names(tb) <- NULL
  return(list(ESTS = ESTS, BE = BE, LL = tb, modelos = modelos))
}

A <- fn_ests_dois_modelos()

knitr::kable(A$BE, "latex", booktabs = T, align = "lcccccc") |>
```

```
kableExtra::add_header_above(c("","betareg" = 2,"betaregscale" = 2, "quasibeta" = 2)) |>
kableExtra::kable_styling(font_size = 10)

knitr::kable(A$LL, "latex", booktabs = T, align = "lccccc", digits = 3) |>
kableExtra::add_header_above(c("","betareg" = 2,"betaregscale" = 2, "quasibeta" = 2)) |>
kableExtra::kable_styling(font_size = 10)

A$ESTS %>%
dplyr::filter(modelos %in% c("M11","M21","M31")) %>%
dplyr::mutate_at(.vars = c('estimate', 'ci_lower', 'ci_upper', 'se', 't_value', 'p_value'),
.funs = round, digits = 6) %>%
dplyr::mutate(modelo = case_when(
  modelos %in% c("M11","M12") ~ "betareg",
  modelos %in% c("M21","M22") ~ "betaregscale",
  modelos %in% c("M31","M32") ~ "quasibeta")
) %>%
dplyr::select(modelo, modelos, everything()) %>%
knitr::kable("latex", booktabs = T, align = "lccccc", digits = 4) %>%
#kableExtra::add_header_above(c("","betareg" = 2,"betaregscale" = 2, "quasibeta" = 2)) |>
kableExtra::kable_styling(font_size = 10)

pseudo_ests <- A$LL %>%
tidyr::pivot_longer(cols = M11:M32, names_to = "id", values_to = "medida") %>%
dplyr::transmute(modelo = id,
modelo_desc = case_when(
  modelo %in% c("M11","M12") ~ "betareg",
  modelo %in% c("M21","M22") ~ "betaregscale",
  modelo %in% c("M31","M32") ~ "quasibeta"),
  medida,
  medida_desc = Medida
)
#readr::write_csv2(pseudo_ests, "outputs/pseudo_ests.csv")

LK <- list(logit = "logit", probit = "probit", cauchit = "cauchit", cloglog = "cloglog")
B <- lapply(LK, function(l) fn_ests_um_modelo(da, link = l))

## ------------------------------------------------------------------ ##
## Residuo
## ------------------------------------------------------------------ ##
modelos <- A$modelos
```

```
RR <- purrr::map_df(modelos, function(m){
  data.frame(
  residuo = as.numeric(resid(m)),
  ajustados = as.numeric(fitted(m))
  )
}, .id = "modelo") %>%
dplyr::filter(modelo %in% c("M11","M21","M31"))

RR %>%
ggplot(aes(x = ajustados, y = residuo)) +
geom_point() +
facet_wrap(.~modelo)
```

ANEXO 3 – DEMONSTRAÇÕES

MÉDIA E VARIÂNCIA DISTRIBUIÇÃO BETA

Prova. Sabendo que a função beta é dada por $\mathrm{B}(p,q) = \frac{\Gamma(p)\cdot\Gamma(q)}{\Gamma(p+q)}$ e também que a função gama é dada por $\Gamma(x+1) = \Gamma(x)\cdot x$, segue que

$$
\begin{aligned}
\mathrm{E}(X) &= \int_{\mathcal{X}} x \cdot f_X(x)\,\mathrm{d}x\ .\\
\mathrm{E}(X) &= \int_0^1 x \cdot \frac{\Gamma(p+q)}{\Gamma(p)\cdot\Gamma(q)}\, x^{p-1}\,(1-x)^{q-1}\,\mathrm{d}x\\
&= \frac{\Gamma(p+q)}{\Gamma(p)} \cdot \frac{\Gamma(p+1)}{\Gamma(p+1+q)} \int_0^1 \frac{\Gamma(p+1+q)}{\Gamma(p+1)\cdot\Gamma(q)}\, x^{(p+1)-1}\,(1-x)^{q-1}\,\mathrm{d}x\ .\\
&= \frac{\Gamma(p+q)}{\Gamma(p)} \cdot \frac{p\cdot\Gamma(p)}{(p+q)\cdot\Gamma(p+q)} \int_0^1 \frac{\Gamma(p+1+q)}{\Gamma(p+1)\cdot\Gamma(q)}\, x^{(p+1)-1}\,(1-x)^{q-1}\,\mathrm{d}x\\
&= \frac{p}{p+q} \int_0^1 \frac{\Gamma(p+1+q)}{\Gamma(p+1)\cdot\Gamma(q)}\, x^{(p+1)-1}\,(1-x)^{q-1}\,\mathrm{d}x\\
&= \frac{p}{p+q} \int_0^1 \mathrm{B}(x; p+1, q)\,\mathrm{d}x = \frac{p}{p+q}.
\end{aligned}
\tag{3.1}
$$

$$
\begin{aligned}
\mathrm{V}(X) &= \mathrm{E}(X^2) - \mathrm{E}(X)^2\ .\\
\mathrm{E}(X^2) &= \int_0^1 x^2 \cdot \frac{\Gamma(p+q)}{\Gamma(p)\cdot\Gamma(q)}\, x^{p-1}\,(1-x)^{q-1}\,\mathrm{d}x\\
&= \frac{\Gamma(p+q)}{\Gamma(p)} \cdot \frac{\Gamma(p+2)}{\Gamma(p+2+q)} \int_0^1 \frac{\Gamma(p+2+q)}{\Gamma(p+2)\cdot\Gamma(q)}\, x^{(p+2)-1}\,(1-x)^{q-1}\,\mathrm{d}x\ .\\
&= \frac{(p+1)\cdot p}{(p+q+1)\cdot(p+q)} \int_0^1 \frac{\Gamma(p+2+q)}{\Gamma(p+2)\cdot\Gamma(q)}\, x^{(p+2)-1}\,(1-x)^{q-1}\,\mathrm{d}x\\
&= \frac{(p+1)\cdot p}{(p+q+1)\cdot(p+q)} \int_0^1 \mathrm{B}(x; p+2, q)\,\mathrm{d}x\\
&= \frac{(p+1)\cdot p}{(p+q+1)\cdot(p+q)}\ .
\end{aligned}
\tag{3.2}
$$

$$
\begin{aligned}
\mathrm{V}(X) &= \frac{(p+1)\cdot p}{(p+q+1)\cdot(p+q)} - \left(\frac{p}{p+q}\right)^2\\
&= \frac{(p^2+p)\cdot(p+q)}{(p+q+1)\cdot(p+q)^2} - \frac{p^2\cdot(p+q+1)}{(p+q+1)\cdot(p+q)^2}\\
&= \frac{(p^3+p^2q+p^2+pq) - (p^3+p^2q+p^2)}{(p+q+1)\cdot(p+q)^2} = \frac{pq}{(p+q+1)\cdot(p+q)^2}\ .
\end{aligned}
$$

□