Modelos de regressão beta para dados de escala

José Evandeilton Lopes [1]
Wagner Hugo Bonat, [2]

Resumo

Existe uma carência na literatura relacionada a modelos de regressão aplicados a respostas intervalares censuradas ao meio, à direita e à esquerda, que sejam independentes do tempo. Essas respostas podem ser obtidas por meio de instrumentos, como as Escalas Numéricas de Classificação (NRS) ou Escalas Likert. Medidas oriundas desses tipos de instrumentos frequentemente aparecem na literatura de forma categórica nominal ou ordinal. Contudo, com algum tratamento, na maioria das situações, é possível mapear esses dados em escalas discretas, contínuas ou como taxas e proporções. Visando agregar conhecimento nessa área, propõe-se uma formulação do modelo de regressão beta conveniente para esse tipo de dado e outros tipos mapeáveis para o suporte da distribuição beta. A abordagem utilizada se baseia no paradigma de máxima verossimilhança, e é proposta uma estrutura cuja função de log-verossimilhança utiliza a distribuição acumulada da beta para lidar com a incerteza intervalar, ao invés da abordagem tradicional que utiliza a distribuição beta usual. Foram propostas transformações de escala para mapear dados de NRS-11 de forma correta para a distribuição beta, passando da escala discreta original para a contínua intervalar, bem como foram explorados três tipos de abordagens para o tratamento dos intervalos. Para verificar propriedades do modelo, bem como viés das estimativas, foi proposto um estudo de simulação abrangendo diversos cenários possíveis de aplicação, inclusive lidando com funções de ligação além da *logit*. Considerando os tratamentos: tipo de intervalo no mapeamento beta, tamanho do parâmetro ϕ e precisão da escala de medida, os resultados indicaram que a centralização do intervalo de mapeamento beta apresenta menores vieses e maior precisão nas estimativas dos coeficientes de regressão, sendo mais recomendada quando a escala possui poucas quebras (ex. NRS-11 ao invés de NRS-100). O parâmetro de dispersão fixo ϕ, que age como um parâmetro de precisão, mostrou-se sensível a variações quando seu valor aumenta, apresentando maior viés nas estimativas, mesmo em amostras grandes. O número de quebras da escala e o tamanho das amostras influenciaram na redução da variabilidade, mas não afetaram diretamente o viés do ajuste em termos gerais, exceto no caso Meio. O tipo de tratamento da incerteza em torno da medida influencia o viés de estimação do intercepto nos casos esquerdo e direito, mas o centralizado é melhor em todas as escalas.

1 Introdução

Dados com algum tipo de censura são comuns em muitas áreas da ciência, especialmente em medicina e meio ambiente. De modo geral, observações conhecidas apenas por estarem

[1]UFPR/DEST/LEG, Curitiba - evandeilton@ufpr.br
[2]UFPR/DEST/LEG, Curitiba - wbonat@ufpr.br

abaixo de um limite de detecção são dados **censurados à esquerda**. Observações conhecidas apenas por estarem acima de um limite de quantificação são dados **censurados à direita**, e dados conhecidos por estarem entre dois limites são dados com **censura intervalar**. Os limites inferiores e superiores correspondem a algum limite de detecção, quantificação ou limites de incerteza em torno da observação, que, em geral, é uma aproximação que não está livre de incerteza devido ao tipo de instrumento empregado. Dados censurados à direita são comumente encontrados com dados de sobrevivência [4]. Dados dessa natureza são analisados também no contexto de máxima verossimilhança em [3]. Um tipo especial de dado que frequentemente ocorre na área médica é dado de medida de dor, ou escores de dor. Áreas médicas, como a Anestesiologia, estudam há muito tempo os padrões físicos, neurológicos e até psicológicos associados ao fenômeno da dor e, dado o caráter de certa forma subjetivo da dor, foram desenvolvidos instrumentos diversos para aferir essa sensação desconfortável. Existem muitas abordagens e ferramentas envolvendo modelos estatísticos na literatura para diversos tipos de dados. No entanto, observa-se uma escassez na divulgação de estudos, dados e modelos de regressão para respostas intervalares com censura à direita, à esquerda e/ou intervalares independentes do tempo, como os dados obtidos por meio da Escala Numérica de Classificação (NRS). Dessa forma, o objetivo deste trabalho é revisar e expandir a literatura existente voltada para esse tipo específico de dado. Para suprir essa lacuna, propõe-se uma formulação conveniente do modelo de regressão beta para dados intervalares com presença de censura, utilizando-se do paradigma de máxima verossimilhança. Visando fazer um mapeamento beta consistente, são propostas transformações de escala e tratamento da incerteza em torno da medida observada, saindo da escala discreta original para a contínua intervalar compatível com a distribuição beta. Para verificar as propriedades do modelo proposto, são feitos estudos de simulações apropriados cobrindo cenários diversos.

1.1 Mapeamento beta, censura e efeito de borda

Uma variável aleatória Y resultante de uma medição de um fenômeno aleatório qualquer pode estar sujeita a censura em algum momento, e isso traz consigo algum nível de imprecisão em sua medida que não deve ser negligenciada.

Observação	Intervalo $\lfloor l_i, u_i \rfloor$	Tipo	Contribuição
Sem censura (exata)	$y_i = u_i$	$\delta = 0$	$f(u_i)$
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	$y_i \in (l_i, u_i)$	$\delta = 3$	$F(u_i) - F(l_i)$

Table 1: Tipos de censura e forma esperada para a contribuição em termos da função de verossimilhança.

A Tabela 1 ilustra os tipos mais comuns de censura em 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 de um dado indivíduo. No paradigma de verossimilhança, o tipo de censura define qual função de probabilidade deve ser aplicada.
A Figura 1 mostra uma abordagem proposta para o mapeamento beta, considerando um trecho onde a medida discreta é mapeada em forma de intervalo e sua região de incerteza é tratada em três possíveis situações:

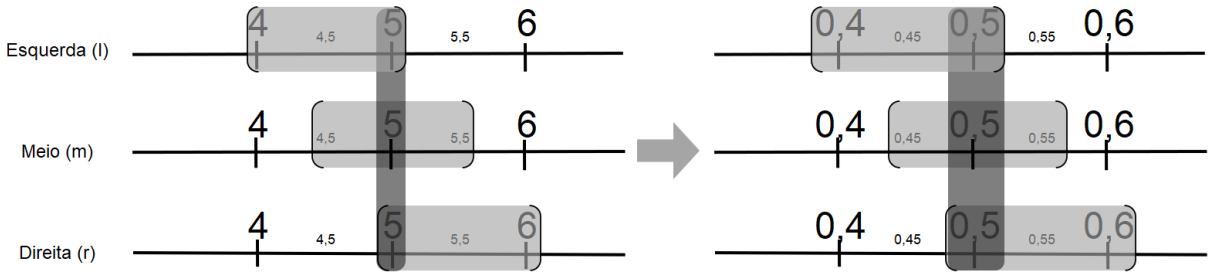

Figure 1: Exemplo de tratamento dos subintervalos de incerteza para o mapeamento beta

- Esquerda (l): nessa abordagem, considera-se que houve uma superestimação do valor informado pelo paciente. Nesse caso, é subtraída uma unidade à medida anotada. Agora o limite inferior será então $l_i = 5 - 1 = 4$ e o superior $u_i = 5$. Logo $y_{[i=5]} = (l_i, u_i) = (4; 5)$;
- Meio (m): ilustra-se a quantificação da incerteza do instrumento centralizando em torno do 5. Nessa abordagem é somado e subtraído meio ponto tanto para a direita como para a esquerda do 5. Assim, cria-se um subintervalo centralizado no valor anotado de modo que o limite inferior será $l_i = 5 - 0,5 = 4,5$ e o limite superior será $u_i = 5 + 0,5 = 5,5$. Logo $y_{[i=5]} = (l_i, u_i) = (4,5; 5,5)$;
- Direita (r): nessa abordagem é adicionado uma unidade ao 5 na direção crescente, sugerindo que houve uma subestimação da medida informada pelo paciente. Assim, o limite inferior do intervalo é $l_i = 5$ e o limite superior será $u_i = 5 + 1 = 6$. Logo $y_{[i=5]} = (l_i, u_i) = (5; 6)$;

Na análise de dados, as transformações são usadas para lidar com particularidades. Elas podem ser aplicadas às variáveis resposta e explicativas. Essas transformações visam melhorar a interpretação, simetria, dispersão, relação entre variáveis e controle de range dos dados, por exemplo [7]. Nesse trabalho, o interesse está em transformações de range para adaptar a escala de números discretos a um intervalo (0,1). Entretanto, é preciso considerar o efeito de borda, que envolve os limites inferior e superior, para evitar perda de informação. Uma opção é utilizar transformações que respeitem esse efeito ou exijam minimamente alguma penalização nos extremos da medida. As equações $y^* = \frac{\frac{y(n-1)}{R} + \frac{1}{2}}{n}$ e $y^* = \frac{y - y_{min}}{y_{max} - y_{min}}$, em que $R = y_{max} - y_{min}$ é o range dos dados e n o total de observações do vetor y, y_{max} e y_{min} são, respectivamente, o máximo e o mínimo registrados, têm sido utilizadas com frequência. Vale notar que dados transformados em escala *min-max* para trabalhar com modelos de regressão restritos ao intervalo unitário, é preciso fazer uma correção adicional no mínimo e no máximo, pois por definição, esses modelos não estão definidos nos extremos do intervalo. Note que, numa NRS-11, por exemplo, o máximo registrado é 10 e o mínimo é 0; então, na escala *min-max*, esse valor é igual a 1.0 e 0.0 no mínimo. Nesse exemplo, a transformação equivale a simplesmente dividir os valores por 10. Contudo, como resolver a questão dos extremos 0 e 1 dos limites? Aqui, como artifício técnico computacional, uma opção é determinar um nível de precisão para a medida, por exemplo, $\zeta = 0,0001$, e subtrair dos extremos. Assim, para $y_i = 0$, tem-se $y_i = \zeta - 0 = 0,0001$ e para $y_i = 1$, tem-se $y_i = 1 - \zeta = 0,9999$.

2 Modelo

A distribuição beta tem sido frequentemente utilizada como opção para modelagem de dados não gaussianos restritos ao intervalo unitário. Os autores [2] propuseram uma parametrização da beta adequada para modelos de regressão. Eles notaram 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 da beta mais flexível para modelagem via regressão, e não haveria perda de propriedades da distribuição, mantendo-se o mesmo suporte de y.

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

Cuja média e variância são dadas por $\mathrm{E}[Y] = \mu$ e $\mathrm{V}[Y] = \frac{\mu(1-\mu)}{1+\phi}$. Note que, nessa parametrização, a média de Y depende apenas de μ. O parâmetro ϕ desempenha o papel de parâmetro de precisão e é sempre positivo.
A Figura 2 mostra os perfis de cobertura das curvas geradas pela distribuição beta na parametrização dada para diversos valores de μ e ϕ. Nota-se que valores pequenos de ϕ tornam os dados mais variados e espalhados em todo o domínio de Y, enquanto valores altos de ϕ concentram a distribuição em torno de μ.

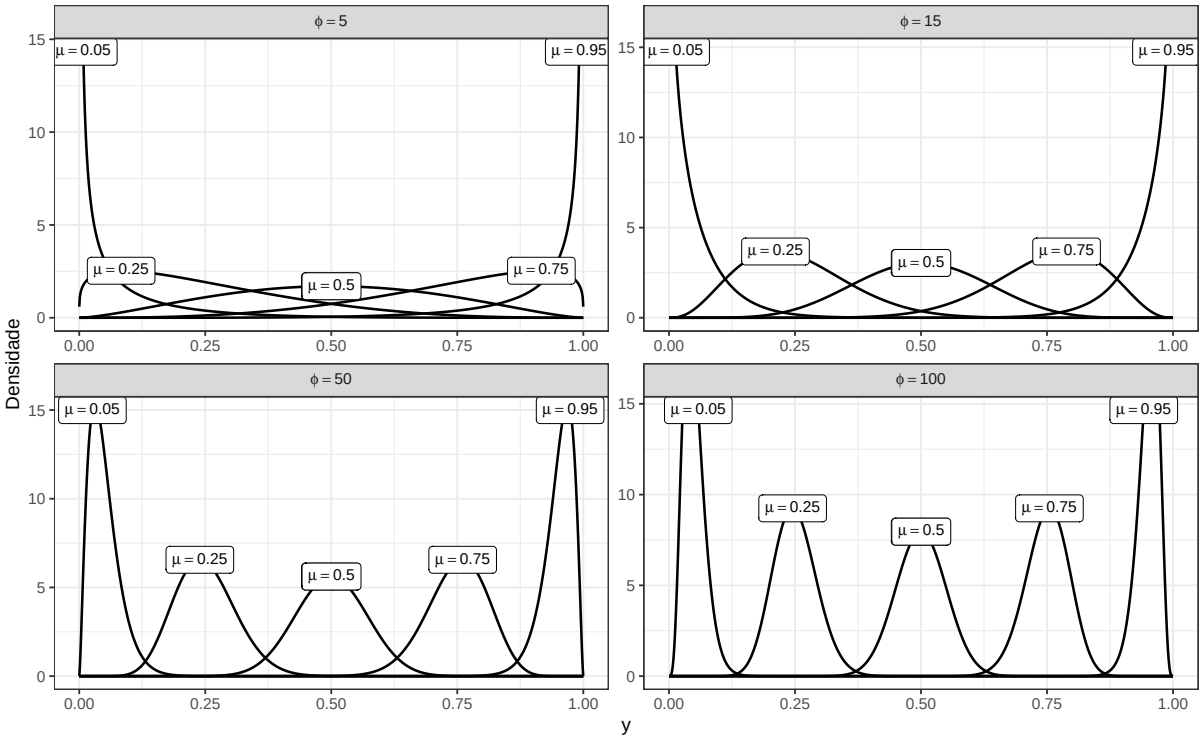

Figure 2: Perfis da distribuição beta na parametrização para modelos de regressão (μ, ϕ)

Com isso, observa-se que, independente do tamanho de μ e ϕ, a distribuição beta é totalmente flexível na cobertura do suporte para qualquer realização de $Y = y_i$. Sendo assim, trata-se de uma distribuição altamente recomendada para modelagem de dados restritos ou transformáveis ao intervalo unitário.

2.1 Modelo de regressão beta

O modelo de regressão beta pode assumir duas formas gerais: uma em que o parâmetro ϕ é fixo, ou seja, não depende de nenhuma covariável; e outra mais geral, na qual ϕ pode ser dependente do efeito de covariáveis. 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 a ter dois preditores lineares, um associado a $\mathbf{x}$ e outro associado a $\mathbf{z}$:

$$\begin{aligned} y_i &\sim B(\mu_i, \phi_i) \\ g_1(\mu_i) &= \mathbf{x}ij^{\mathsf{T}}\beta_j = \sum j = 1^p x_{ij}\beta_j \\ g_2(\phi_i) &= \mathbf{z}ij^{\mathsf{T}}\gamma_j = \sum j = 1^k z_{ij}\gamma_j \end{aligned} \tag{2}$$

Em que g1(.) e g2(.) são funções de ligação estritamente monótonas e duplamente diferenciáveis, com domínio em $(0, 1)$ e imagem nos reais, associadas às regressoras **xij** para μ e **zij** para ϕ.

2.2 Verossimilhança com censura

A forma geral da verossimilhança, que expressa a ideia geral da teoria de estimação de parâmetros, é flexível e pode ser adaptada para dados censurados. Por exemplo, com base nos trabalhos de Klein and Moeschberger [4] and Helsel et al. [3], que lidaram com ajustes de modelos para dados censurados, Delignette-Muller and Dutang [1] propuseram no pacote `fitdistrplus` em linguagem **R** uma função de verossimilhança completa contemplando os três tipos de censura observados. No entanto, essa função não permite a inclusão de preditoras. Visando montar uma estrutura matemática completa inclusive com a adição de covariáveis, Lopes [5] adicionaram uma função indicadora $\mathbf{I}\zeta_i = 0, 1$ que, caso ζ_i ocorra, recebe 1 e 0 caso contrário. Assim, a existência ou não de censura em algum ponto de dado define o valor que $L(\theta)ge$ recebe. Desse modo, a forma completa da verossimilhança fica definida por:

$$\begin{aligned} L(\theta)ge &= L(\theta)cs \times L(\theta)ce \times L(\theta)cd \times L(\theta)ci \\ L(\theta)ge &= \prod_{i=1}^{N_{cs}}[f(y_i = u_i|\theta)]^{\zeta_i} \times \mathbf{I}\zeta_i = 0, 1 \\ &+ \prod i = 1^{N_{ce}}[F(y_i = u_i|\theta)]^{\zeta_i} \times \mathbf{I}\zeta_i = 0, 1 \\ &+ \prod i = 1^{N_{cd}}[1 - F(y_i = l_i|\theta)]^{\zeta_i} \times \mathbf{I}\zeta_i = 0, 1 \\ &+ \prod i = 1^{N_{ci}}[F(y_i = u_i|\theta) - F(y_i = l_i|\theta)]^{\zeta_i} \times \mathbf{I}\zeta_i = 0, 1. \end{aligned} \tag{3}$$

Onde $\theta = (\beta, \phi)^{\mathsf{T}}$ 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. $L(\theta)ge$ é composta pelas partes: $L(\theta)cs$ é a contribuição da parte sem censura, $L(\theta)ce$ é a censura à esquerda, $L(\theta)_{cd}$ é a censura à direita e $L(\theta)_{ci}$ é para censura intervalar. Essa forma geral se aplica a outras distribuições de probabilidade além da beta, sendo elas discretas ou contínuas. No caso da beta com censura, não há

solução analítica fechada e métodos numéricos iterativos, como *quasi-Newton* ou outros, podem ser empregados para o cálculo das estimativas dos parâmetros desconhecidos.

3 Simulação

Para validar as propriedades dos estimadores de máxima verossimilhança no modelo de regressão beta para resposta transformada intervalar, recorreu-se a simulações computacionais utilizando o Software R [6]. Com base em alguns cenários desenhados adequadamente, é possível ter controle do processo gerador de amostras. Isso permite reduzir incertezas e ruídos que podem estar associados a aleatoriedades externas ao experimento. Visando abranger um bom número de casos especiais, foram criados alguns cenários de simulação conforme descritos a seguir: (1) tratamento dos intervalos: transformação de escala e correção do efeito borda, que ocorre nos valores 0 e 1 que não fazem parte do suporte da beta; (2) precisão do instrumento: para esse caso, foram simuladas três tipos de escalas: a NRS-11, que possui 10 quebras, e outras duas derivações dela com 20 e 100 quebras. É esperado que, aumentando-se a precisão do instrumento, o efeito do erro de medida seja reduzido; (3) efeito da dispersão: nesse caso, se deseja avaliar o efeito do parâmetro de dispersão nas estimativas dos coeficientes de regressão do modelo; (4) estabilidade em x_{ij} e z_{ij}: valores simulados dessas variáveis foram mantidos constantes ao longo de todas as amostras. Isso permite analisar com maior precisão a variabilidade dos coeficientes em cada modelo e também o efeito nelas quanto ao tamanho de amostra.
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 **logit** a seguinte estrutura com 500 réplicas:

$$\begin{aligned} g(\mu_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} \\ \mu_i &= \frac{\exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\right)} \end{aligned} \tag{4}$$

Com os seguintes critérios

- $\beta_i : \beta_0 = 0.3, \beta_1 = -0.6, \beta_2 = 0.3$
- $\phi_i : \phi_1 = 5, \phi_2 = 10, \phi_3 = 25, \phi_4 = 50, \phi_4 = 100$
- $x_1 : X_1 \sim \text{N(n, mu} = 1, \text{sigma} = 1)$
- $x_2 : X_2 \sim \text{Binom(n, size} = 1, \text{prob} = 0.5)$
- $n : 50, 100, 250, 500, 1000$

4 Resultados e Discussão

No modelo de regressão beta com dispersão fixa, o parâmetro ϕ é mantido constante e independente de quaisquer variáveis.

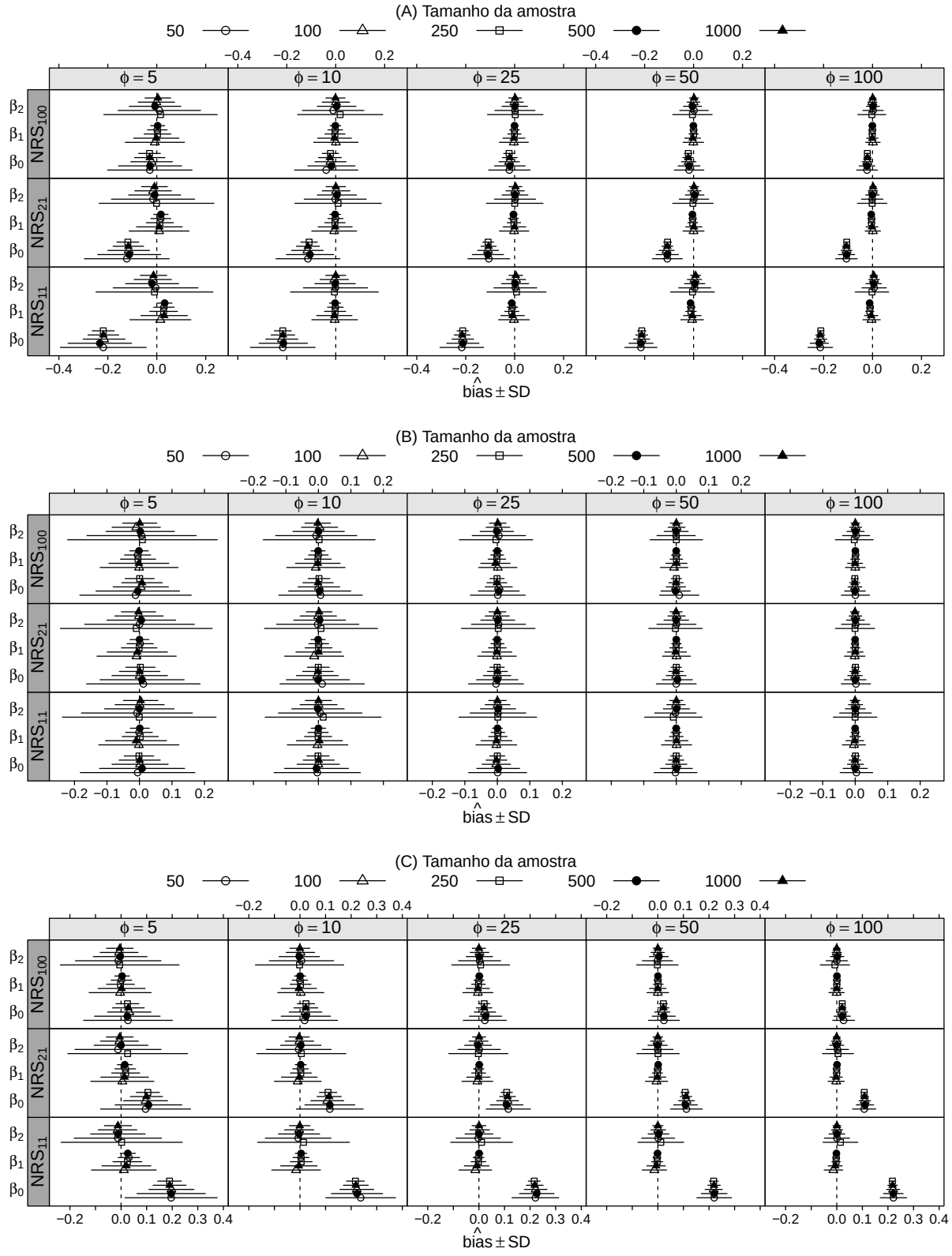

Figure 3: Viés das estimativas dos coeficientes de regressão β em cenários diversos simulados

A Figura 3 apresenta os vieses (BIAS) das estimativas dos coeficientes beta em diversos cenários simulados. Cada linha ilustra o intervalo de variação do viés em relação ao erro padrão das estimativas, levando em consideração o tamanho das amostras, representado

por formas geométricas, o valor do parâmetro ϕ, o método de tratamento do intervalo no mapeamento beta ordinal e a precisão da escala. Com base nos métodos de tratamento dos intervalos observados na Figura 1, de cima para baixo, temos respectivamente: (A) Esquerda (l), (B) Meio (m) e (C) Direita (r). Conforme estabelecido nos critérios de simulação, as variáveis explicativas X_1 e X_2 são mantidas constantes e inalteradas.

O primeiro aspecto crucial é que, ao mapear intervalos à Esquerda (l), caso (A), há um desvio significativo no coeficiente do intercepto, resultando em estimativas de $\hat{\beta}_0$ menores que o parâmetro fixo para todos os tamanhos de amostra e escalas de ϕ. Entretanto, o efeito do erro associado ao tratamento do intervalo diminui conforme o número de quebras nas escalas simuladas aumenta. Um efeito similar ocorre no caso da Direita (r), caso (C), porém com viés positivo nas estimativas. No caso do Meio (m), caso (B), há uma concordância notável entre todos os cenários, com o viés diminuindo conforme o tamanho da amostra e o parâmetro de dispersão aumentam e permanecendo invariável para diferentes quantidades de quebras.

Há evidências de que o intervalo é fundamental para o viés das estimativas dos coeficientes. Escalas com menos quebras (ex. NRS-11), apresentam maior risco de subestimação e superestimação. Nesses casos, é preferível utilizar intervalos centralizados no meio. Quando a escala possui muitas quebras (ex. NRS-101), o viés nos casos à Esquerda ou à Direita é bastante reduzido. Em resumo, para escalas com poucas quebras, a opção Meio (m) é mais recomendada, enquanto para escalas com muitas quebras, qualquer uma das três abordagens pode ser adotada.

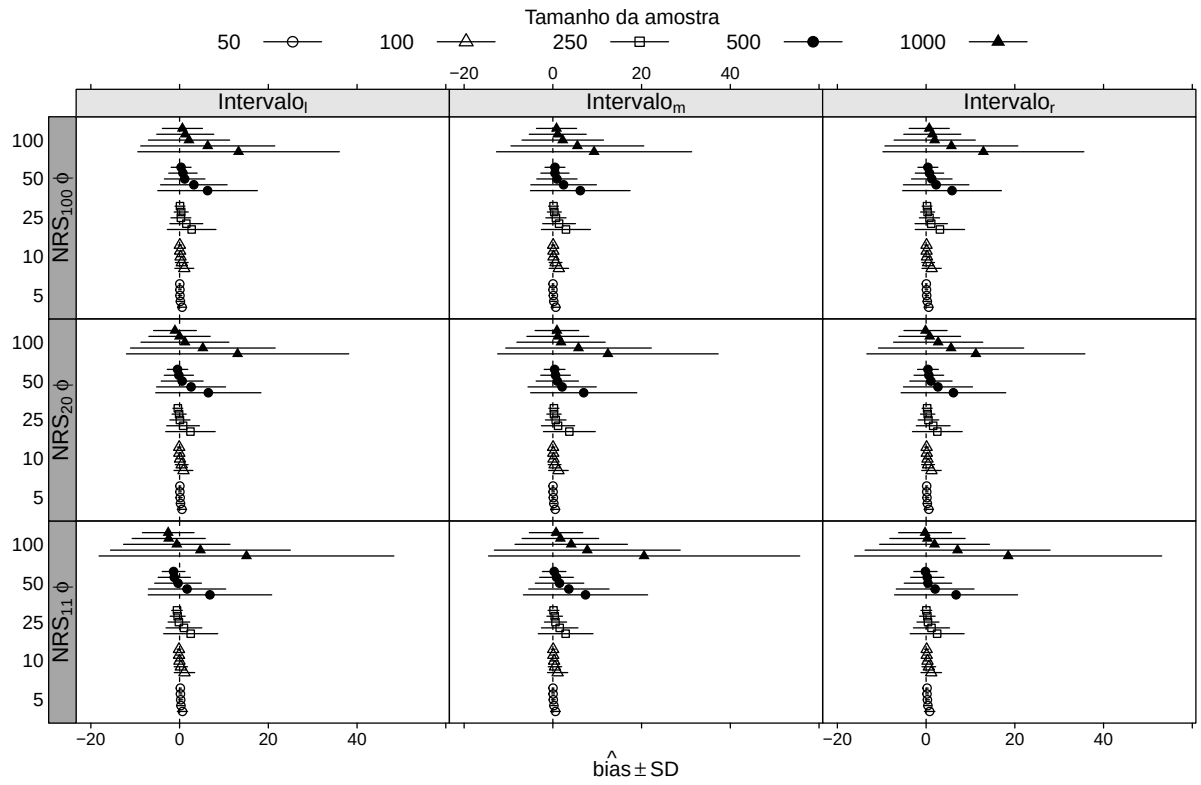

Figure 4: Viés das estimativas dos coeficientes de ϕ em cenários diversos simulados

A Figura 4 traz uma visão de como o parâmetro de dispersão fixo se comporta de acordo com o tipo de tratamento do intervalo e também quanto ao total de quebras da escala, à medida em que aumenta o tamanho das amostras simuladas. Diferentemente do observado

no β_0, o tipo de tratamento do intervalo não é afetado, uma vez que o viés observado é semelhante nas três abordagens. Em todos os cenários, tanto para o intervalo quanto para as quebras de escala, nota-se que, mesmo em amostras grandes, existe um maior viés nas estimativas de ϕ. Isso é esperado uma vez que este parâmetro é sensível quando seu valor aumenta. Outro fator importante é que, mesmo em amostras pequenas, se ϕ é pequeno ($\phi < 25$), há menos viés em suas estimativas.
Em resumo, podemos concluir com as simulações que: (1) o viés das estimativas dos coeficientes de regressão β é afetado pelo tipo de tratamento do intervalo no mapeamento beta, pelo tamanho do parâmetro ϕ e pela precisão da escala; (2) trabalhar com intervalos centralizados ao meio (caso B) é a opção mais recomendada quando a escala possui poucas quebras; (3) quando a escala possui muitas quebras (acima de 100), o viés dos casos à esquerda (caso A) ou à direita (caso C) é bastante reduzido, permitindo a escolha de qualquer uma das três abordagens; (4) o parâmetro de dispersão fixo ϕ é mais sensível a variações quando seu valor aumenta, apresentando maior viés nas estimativas, mesmo em amostras grandes. Além disso, quando ϕ é pequeno ($\phi < 25$), há menos viés em suas estimativas, mesmo em amostras pequenas.

5 Trabalhos futuros

Os resultados apresentados neste artigo fazem parte de uma pesquisa de mestrado sobre modelos para dados de escala que está sendo desenvolvida no PPGMNE/UFPR. A seguir, apresentamos uma lista dos tópicos relacionados que estão sendo investigados:

- Simular o modelo em escalas com menos de 11 quebras, no estilo Likert, com 3, 5 e 7 pontos, visando testar a capacidade de resposta do modelo e avaliar os vieses das estimativas considerando também a dispersão variável;
- Testar modelos beta usuais e modelos quesi-beta que trabalham com primeiro e segundo momento;
- Avaliar a taxa de cobertura dos betas em simulações de todos os cenários, aprofundar os estudos de bondade do ajuste, seleção de modelos e qualidade dos resíduos;
- Realizar uma aplicação em dados reais de escala de dor.

Palavras-chave: regressão beta; NRS-11; mapeamento beta; escala de dor; censura intervalar.

References

[1] Marie Laure Delignette-Muller and Christophe Dutang. "fitdistrplus: An R Package for Fitting Distributions". In: *Journal of Statistical Software* 64.4 (2015), pp. 1–34. DOI: 10.18637/jss.v064.i04.

[2] Silvia Ferrari and Francisco Cribari-Neto. "Beta regression for modelling rates and proportions". In: *Journal of applied statistics* 31.7 (2004), pp. 799–815.

[3] Dennis R Helsel et al. *Nondetects and data analysis. Statistics for censored environmental data.* Wiley-Interscience, 2005.

[4] John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Vol. 1230. Springer, 2003.

[5] José E Lopes. "Modelos de regressão beta para dados de escala". MA thesis. UFPR - Universidade Federal do Paraná, 2023.

[6] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2023. URL: https://www.R-project.org/.

[7] Michael Stuart. *Understanding robust and exploratory data analysis*. 1984.