Universidade de Brasília
Departamento de Estatística

Estudo sobre regressão beta robusta

Davi Dantas Erthal

Relatório Final apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2025

Davi Dantas Erthal

Estudo sobre regressão beta robusta

Orientadora: Profa. Terezinha Kessia de Assis Ribeiro

Relatório Final apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2025

Resumo

Modelos de regressão beta são amplamente utilizados para modelar respostas contínuas limitadas ao intervalo unitário. Exemplos incluem proporção de renda gasta com alimentação, índice de qualidade de vida, cobertura de seguro saúde, entre outros. Os parâmetros destes modelos são geralmente estimados pelo método de máxima verossimilhança, que é sensível a observações discrepantes, podendo resultar em inferências equivocadas. Este trabalho apresenta um breve estudo sobre a robustez de estimadores recentemente propostos sob modelos de regressão beta. Estas abordagens robustas, baseadas em M-estimadores, visam minimizar a influência de observações atípicas, mantendo boas propriedades estatísticas. Por fim, os diferentes métodos inferenciais robustos são avaliados e comparados por meio de uma aplicação com dados reais.

Palavras-chave: Regressão Beta, Máxima verossimilhança, Estimadores robustos, Observações discrepantes.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

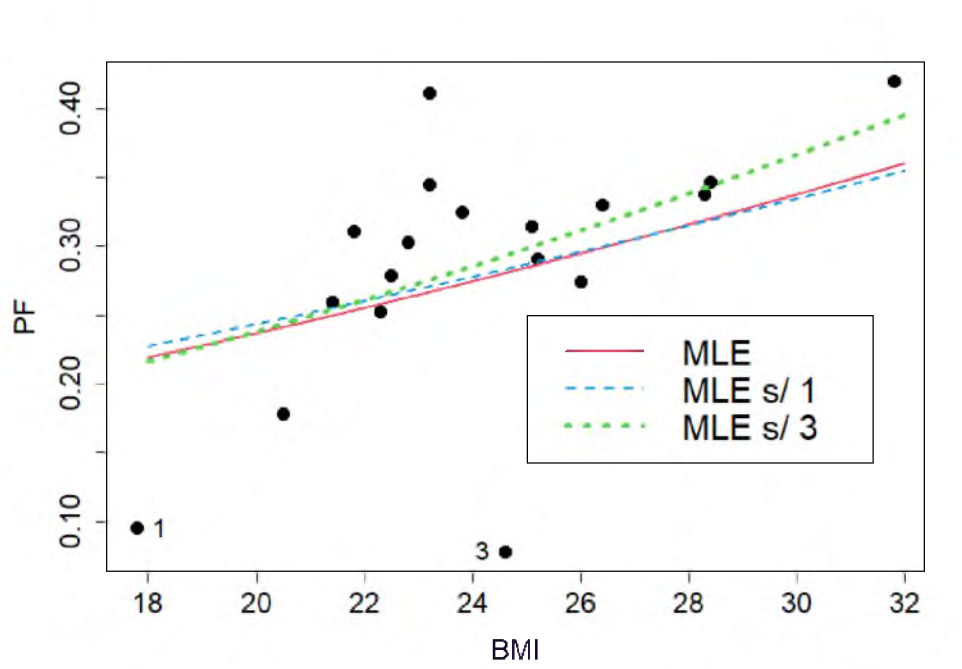
Modelos de regressão beta (FERRARI; CRIBARI-NETO, 2004; SMITHSON; VERKUILEN, 2006) são vastamente utilizados para modelar o comportamento de respostas contínuas que são limitadas ao intervalo unitário aberto. Este comportamento é explicado por meio de variáveis que são usualmente chamadas de covariáveis, explicativas ou preditoras. Exemplos de dados que podem ser modelados através desta abordagem são proporção de renda gasta com alimentação, índice de qualidade de vida de pacientes com alguma doença, cobertura de seguro saúde, fração do tempo livre gasto com atividades de lazer, taxas de desemprego e pobreza, e escore para mensurar habilidade de leitura.

O efeito que as variáveis explicativas possuem sobre a variável resposta é mensurado através dos coeficientes da regressão que são parâmetros populacionais desconhecidos que devem ser estimados. Sob modelos de regressão beta, estes parâmetros são usualmente estimados com base no método de máxima verossimilhança. Entretanto, como discutido em Ribeiro e Ferrari (2023) e Maluf, Ferrari e Queiroz (2024), este método se mostra sensível na presença de observações discrepantes. A presença de outliers pode conduzir inferências equivocadas sobre as verdadeiras relações entre as características estudadas. Uma única observação pode introduzir viés significativo nas estimativas dos coeficientes da regressão estimados via o método de máxima verossimilhança.

Como ilustração do efeito que dados discrepantes podem possuir no ajuste de um modelo de regressão beta, considere o conjunto de dados a ser melhor discutido posteriormente sobre o percentual de gordura corporal em indivíduos. A variável resposta é a porcentagem de gordura corporal (*Percent Fat*, PF), e uma variável explicativa é o índice de massa corporal (*Body Mass Index*, BMI), uma medida muito utilizada no contexto nutricional que relaciona altura e peso. Na Figura 1 tem-se o gráfico de dispersão de PF versus BMI juntamente com curvas obtidas via regressão beta ajustada utilizando o estimador de máxima verossimilhança (*Maximum Likelihood Estimator*; MLE) com todos os dados e excluindo algumas observações discrepantes. Observa-se que a retirada de um ponto discrepante impacta consideravelmente a curva ajustada aos dados, ilustrando assim a sensibilidade do MLE.

Uma solução natural para lidar com observações atípicas é trocar o procedimento inferencial. Assim, mantém-se o modelo de regressão beta como a representação do fenômeno aleatório estudado, entretanto, troca-se a forma de estimar os parâmetros. Recentemente, algumas abordagens inferenciais robustas sob modelos de regressão beta foram introduzidas. Ghosh (2019) e Ribeiro e Ferrari (2023) propuseram estimadores ro-

Figura 1: Gráfico de dispersão de PF contra BMI juntamente com as curvas ajustadas via regressão beta baseadas no MLE para os dados completos e para os dados sem o ponto discrepante



bustos que são obtidos a partir da minimização de divergências entre densidades. Ambos estimadores são casos especiais da classe de M-estimadores introduzida inicialmente por Huber (1964). Os M-estimadores são conhecidos por produzir estimadores robustos que possuem boas propriedades tais como consistência, eficiência e normalidade assintótica (MARONNA et al., 2019). Os métodos robustos sob regressão beta possuem constantes que controlam a robustez do procedimento inferencial, e contém o MLE como caso especial. Entretanto, como apontado por Ribeiro e Ferrari (2023), estes métodos possuem algumas restrições para sua obtenção. Nesse sentido, Maluf, Ferrari e Queiroz (2024) propuseram adaptações dos procedimentos inferenciais robustos de tal forma que não existam restrições para a sua obtenção.

O presente trabalho, que está dividido em 5 capítulos, visa um breve estudo sobre a recente metodologia de regressão beta robusta. No Capítulo 2, será apresentada a classe de modelos de regressão beta sob a inferência clássica baseada no método de máxima verossimilhança. No Capítulo 3, abordaremos os métodos robustos propostos por Ghosh (2019), Ribeiro e Ferrari (2023) e Maluf, Ferrari e Queiroz (2024), introduzindo estimadores alternativos capazes de reduzir a influência de valores discrepantes sobre as estimativas dos parâmetros num modelo de regressão beta. No Capítulo 4, será realizada a aplicação dos métodos discutidos a um conjunto de dados reais, permitindo a comparação entre os estimadores clássicos e robustos, bem como a análise de suas diferenças em termos de ajuste e inferência estatística. Para realizar a comparação empírica entre as aborda-

gens estudadas serão utilizadas as bibliotecas `betareg` e `robustbetareg` disponíveis no *software* R. Por fim, no Capítulo 5 são apresentadas as considerações finais deste trabalho.

2 Referencial Teórico

2.1 Distribuição Beta

A distribuição beta é uma família de distribuições de probabilidade contínuas com dois parâmetros $a > 0$ e $b > 0$. A função densidade de probabilidade de uma variável aleatória Y que segue uma distribuição beta com parâmetros a, b é dada por

$$f(y; a, b) = \frac{y^{a-1}(1-y)^{b-1}}{\beta(a,b)}, \ 0 < y < 1,$$

em que $\beta(a, b)$ é a função beta definida por

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

Considere que

$$\beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

com $\Gamma(\cdot)$ sendo a função gama definida por

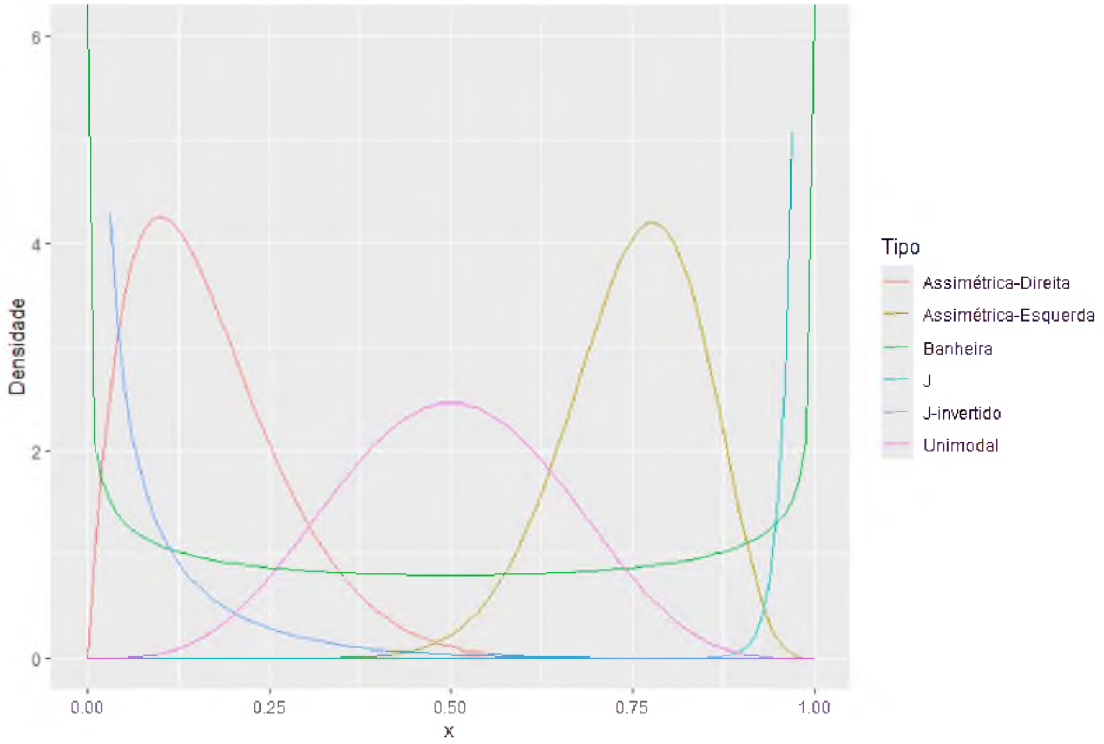
$$\Gamma(z) = \int_0^\infty u^{z-1}e^{-u}du.$$

Assim, a função densidade de probabilidade pode ser reescrita como

$$f(y; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}y^{a-1}(1-y)^{b-1}, 0 < y < 1, a, b > 0. \tag{2.1.1}$$

A flexibilidade da distribuição beta é notável, pois pode assumir uma variedade de formas dependendo dos valores fixados para os parâmetros a e b. Na Figura 2 apresentamos curvas da função de densidade da distribuição beta para alguns valores fixados dos parâmetros. Este modelo é frequentemente usado para representar uma variedade de tipos de fenômenos aleatórios em que a variável de interesse está limitada ao intervalo contínuo (0, 1). Entretanto, a distribuição beta também pode ser usada para fenômenos com valores restritos ao intervalo contínuo (c, d), com c e d sendo constantes reais. Este intervalo contínuo pode ser ajustado ao suporte exigido pela distribuição beta por meio da transformação $y' = (y - c)/(d - c)$ com $y' \in (0, 1)$.

Figura 2: Curvas para a função de densidade de probabilidade da distribuição beta para diferentes valores de (a,b).



A média e a variância da distribuição beta são expressos por

$$\mathrm{E}(Y) = \frac{a}{a+b},$$

$$\mathrm{Var}(Y) = \frac{ab}{(a+b)^2(a+b+1)}.$$

Desse modo, a e b contribuem para a média e variância de Y, entretanto não possuem uma interpretação simples. Estes resultados da distribuição beta servem como base para introdução do modelo de regressão beta. Neste tipo de regressão, a média de Y será explicada através de covariáveis.

Ferrari e Cribari-Neto (2004) propuseram uma reparametrização da distribuição beta com o objetivo de estabelecer uma estrutura de regressão para modelar a média de uma variável resposta distribuída segundo o modelo beta. Sob essa reparametrização, a média e a precisão são definidas como $\mu = \mathrm{E}(Y) = a/(a+b)$ e $\phi = a+b$, respectivamente, com $a = \mu\phi$ e $b = (1-\mu)\phi$ sendo os parâmetros originais da distribuição beta.

A expressão para a função de densidade de probabilidade da distribuição beta

reparametrizada fica dada por

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \;\; 0<y<1. \tag{2.1.2}$$

Denotamos por $Y \sim \mathcal{B}(\mu,\phi)$. Sob esta reparametrização, a média e a variância de Y ficam expressos por

$$\mathrm{E(Y)} = \frac{\mu\phi}{\mu\phi+(1-\mu)\phi} = \mu,$$

$$\mathrm{Var(Y)} = \frac{\mu(1-\mu)}{\phi+1}.$$

Observe que, mantendo a média μ constante, a variância de Y tende a reduzir quanto maior for o valor de ϕ. Ainda, valores baixos de ϕ geram valores altos para a variância de Y. Por isso, ϕ é considerado como parâmetro de precisão da distribuição beta reparametrizada.

2.2 Método da Máxima Verossimilhança

O método de máxima verossimilhança é uma das mais populares técnicas para obtenção de estimadores. Tome $Y_1,\ldots,Y_n$ tais que Y_i possui função densidade $f(y_i;\theta)$ com $\theta=(\theta_1,\ldots,\theta_k)^\top$. A função de verossimilhança para θ é definida como

$$L(\theta;y) = f(y;\theta),$$

com $f(y;\theta)$ denotando a função de densidade conjunta de $Y=(Y_1,\ldots,Y_n)^\top$. Sob independência de $Y_1,Y_2,\ldots,Y_n$, ficamos com

$$L(\theta;y) = \prod_{t=1}^{n} f(y_t;\theta),$$

em que $f(y_t;\theta)$ é a função densidade de Y_t avaliada no ponto y_t.

Para cada ponto y da amostra, seja $\hat{\theta}(y)$ um valor do parâmetro θ ao qual $L(\theta;y)$ atinge o seu máximo como uma função de θ, para y fixado. O MLE do parâmetro θ baseado na amostra Y é $\hat{\theta}(Y)$.

O MLE é o valor mais provável do parâmetro θ para qual a amostra observada foi gerada. Intuitivamente, o MLE é uma escolha razoável para estimar θ. Em geral, o MLE é um bom estimador pontual, pois possui boas propriedades tais como normalidade assintótica, eficiência assintótica e consistência. Casella e Berger (2024) apontam duas

desvantagens inerentes associadas com o problema geral de encontrar o máximo de uma função, e portanto da estimação pelo método de máxima verossimilhança. A primeira é verdadeiramente encontrar o máximo e verificar que, de fato, o máximo global foi encontrado. E a segunda, mais relacionada com este trabalho, é a sensibilidade numérica ou falta de robustez. O quão sensível é a estimativa de máxima verossimilhança à pequenas mudanças nos dados? Em alguns casos uma amostra minimamente diferente produz um MLE extremamente diferente, fazendo com que suspeitemos dessa estimativa, e do uso do método.

Para lidar com a primeira desvantagem citada, note que, se a função de verossimilhança é diferenciável em θ_i, os possíveis candidatos para o MLE são os valores de $(\theta_1, ..., \theta_k)$ que satisfazem as seguintes equações

$$\frac{\partial L(\theta; y)}{\partial \theta_j} = 0, \;\; j = 1, ..., k.$$

Assim, a solução $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)^\top$ é um possível candidato para o MLE. Além disso, o procedimento de igualar zeros com a primeira derivada nos ajudam a localizar os pontos extremos no interior do domínio da função. Na maioria dos casos, desde que seja possível, é mais fácil trabalhar com o logaritmo da função de verossimilhança $\log L(\theta; y)$ do que com a função de verossimilhança diretamente, uma vez que a função logaritmo é estritamente crescente no intervalo $(0, \infty)$.

O logaritmo da função de verossimilhança para θ é

$$\ell(\theta; y) = \sum_{t=1}^{n} \log[f(y_t; \theta)].$$

Daí, o vetor escore, que é o vetor de primeiras derivadas, é dado por

$$U(\theta) = \frac{d\ell(\theta)}{d\theta} = \sum_{t=1}^{n} U(y_t; \theta),$$

em que

$$U(y_t; \theta) = \frac{d \log f(y_t; \theta)}{d\theta}$$

Uma outra vantagem dos MLEs é que se a função $L(\theta; y)$ não pode ser maximizada analiticamente, pode-se obter $\hat{\theta}$ numericamente por métodos de otimização não linear.

Dado que o MLE foi encontrado, este pode ser suscetível ao problema de falta de robustez. Tomando que os dados são medidos com um erro (contaminação), nos pergun-

tamos como pequenas mudanças nos dados podem afetar o MLE. Em outras palavras, calculamos $\hat{\theta}$ baseado numa função de verossimilhança $L(\theta; y)$, entretanto gostaríamos de saber qual valor obteríamos se tivéssemos baseado os cálculos em uma função de verossimilhança $L(\theta; y + \epsilon)$. Intuitivamente, se a mudança ϵ é pequena, a diferença entre os dois estimadores deveria ser, também, pequena, o que não é o caso sempre para o MLE, principalmente quando a função de verossimilhança é plana nas proximidades do máximo e quando o estimador não pode ser obtido analiticamente.

O MLE possui algumas propriedades importantes para o estudo, tal como consistência, que serão comentadas mais à frente quando falarmos dos métodos robustos. Um estimador $\hat{\theta}$ é considerado consistente para θ se, à medida que o tamanho da amostra aumenta, este converge em probabilidade para o valor verdadeiro do parâmetro populacional θ. Em outras palavras, para uma amostra suficientemente grande (quando n tende ao infinito), a diferença entre o estimado e o verdadeiro valor do parâmetro tende a zero com probabilidade 1 (CASELLA; BERGER, 2024). Ainda, o MLE é Fisher-consistente, o que garante que o estimado irá atingir o verdadeiro valor de θ quando é calculado sob a distribuição populacional dos dados.

Outra característica importante do MLE é a normalidade assintótica. Sob certas condições (como amostras grandes e regularidade dos dados), o MLE segue aproximadamente uma distribuição normal quando o tamanho da amostra tende ao infinito; Especificamente, o MLE se aproxima de uma distribuição normal com média igual ao verdadeiro valor do parâmetro e variância dada pela inversa da informação de Fisher. Esta última propriedade é vantajosa pois possibilita a aplicação de métodos inferenciais clássicos, como a construção de intervalos de confiança e testes de hipóteses, mesmo que a distribuição dos dados originais não seja normal. O MLE é também assintóticamente eficiente, no sentido de atingir a menor variância possível entre os estimadores não viesados quando o tamanho amostral é grande. Essa eficiência está ligada ao Teorema de Cramér-Rao, que estabelece um limite inferior para a variância de qualquer estimador não viesado. Quando a amostra é suficientemente grande, o MLE atinge essa variância mínima, tornando-o o estimador mais preciso assintoticamente. Dessa forma, a eficiência assintótica garante que o MLE utiliza as informações dos dados de maneira ótima, o que o torna preferível em muitos cenários. Adicionalmente, o MLE também possui a propriedade de invariância. Isso significa que, se $\hat{\theta}$ é o MLE para o parâmetro θ, então qualquer função desse parâmetro, $g(\theta)$, pode ser estimada aplicando-se diretamente a função $g(\cdot)$ no estimador obtido para θ dado que $g^{-1}(\cdot)$ exista. Essa característica simplifica o processo de estimação de funções dos parâmetros, especialmente em casos onde se deseja estimar

quantidades derivadas, como a razão de duas variâncias ou a soma de médias de diferentes populações (CASELLA; BERGER, 2024).

2.3 Regressão Beta

Para definirmos os modelos de regressão beta considerados nesse trabalho, consideraremos o modelo proposto por Smithson e Verkuilen (2006) que é uma extensão do modelo previamente proposto por Ferrari e Cribari-Neto (2004). Nessa abordagem é considerada a precisão variável, sendo o caso de precisão constante um caso particular. Dessa forma a média μ e a precisão ϕ da variável resposta serão modeladas simultaneamente. Assim, definiremos as estruturas de regressão linear que serão atribuídas a média e ao parâmetro de precisão.

Dada a reparametrização da distribuição beta apresentada em (2.1.2), considere n realizações independentes de uma variável aleatória Y denotadas por $Y_1, \ldots, Y_n$ com $Y_i \sim B(\mu_i, \phi_i)$. Assim, assume-se que a média μ_t de cada observação y_t pode ser escrita como

$$g_\mu(\mu_t) = \eta_t = \sum_{i=1}^{k} x_{ti}\beta_i = X_t^\top \beta, \tag{2.3.1}$$

em que

- $\beta = (\beta_1, \beta_2, \ldots, \beta_k)^\top \in \mathbb{R}^k$ é um vetor de parâmetros desconhecidos (coeficientes da regressão) associado à média;
- $X_t = (x_{t1}, x_{t2}, \ldots, x_{tk})^\top \in \mathbb{R}^k$ é o vetor de valores conhecidos das k covariáveis para a t-ésima observação $(t = 1, 2, \ldots, n)$;
- $g_\mu(\cdot)$ é uma função de ligação contínua, estritamente monótona e duas vezes diferenciável.

Vale ressaltar que existem numerosas opções para a função de ligação $g_\mu(\cdot)$. Dentre as mais utilizadas estão (FERRARI; CRIBARI-NETO, 2004):

- função logito: $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$;
- função probit: $g(\mu) = \Phi^{-1}(\mu)$;
- função cloglog $g(\mu) = \log(-\log(1-\mu))$;

- função cauchit $g(\mu) = \tan(\pi(\mu - 0,5))$.

Adicionalmente, consideraremos uma estrutura de regressão que modela simultaneamente a precisão ϕ_t:

$$g_\phi(\phi_t) = \sum_{j=1}^{q} z_{tj}\gamma_j = Z_t^\top \gamma = \vartheta_t, \tag{2.3.2}$$

em que

- $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_q)^\top \in \mathbb{R}^q$ é um vetor de parâmetros desconhecidos (coeficientes da regressão) associado à precisão;
- $Z_t = (z_{t1}, z_{t2}, \ldots, z_{tq})^\top \in \mathbb{R}^k$ é o vetor de valores conhecidos das q covariáveis da precisão para a t-ésima observação ($t = 1, 2, \ldots, n$);
- $g_\phi(\cdot)$ é uma função de ligação contínua, estritamente monótona e duas vezes diferenciável.

Smithson e Verkuilen (2006) mostram que para a precisão também existem algumas funções de ligação apropriadas como a ligação logaritmo $g_\phi(\phi) = \log(\phi)$ e a ligação raiz-quadrada $g_\phi(\phi) = \sqrt{\phi}$.

Baseando-se na amostra, devemos estimar os vetores de parâmetros desconhecidos β e γ da regressão de acordo com (2.3.1) e (2.3.2). Para isso, pode-se utilizar o método de máxima verossimilhança já introduzido na Subseção 2.2. A função de verossimilhança para $\theta = (\beta^\top, \gamma^\top)^\top$ considerando $Y_1, \ldots, Y_n$ variáveis aleatórias independentes tal que $Y_t \sim B(\mu_t, \phi_t)$ é dada por

$$L(\theta; y) = \prod_{t=1}^{n} \left[\frac{\Gamma(\phi_t)}{\Gamma(\mu_t\phi_t)\Gamma((1-\mu_t)\phi_t)} y_t^{\mu_t\phi_t - 1}(1 - y_t)^{(1-\mu_t)\phi_t - 1} \right]$$

e, o logaritmo da função de verossimilhança para θ é

$$\ell(\theta; y) = \sum_{t=1}^{n} \ell_t(\mu_t, \phi_t). \tag{2.3.3}$$

Desenvolvendo $\ell_t(\mu_t, \phi_t)$ e utilizando propriedades logarítmicas, temos que

$$\begin{aligned} \ell_t(\mu_t, \phi_t) = {} & \log\Gamma(\phi_t) - \log\Gamma(\mu_t\phi_t) - \log\Gamma((1-\mu_t)\phi_t) \\ & + (\mu_t\phi_t - 1)\log(y_t) + [(1-\mu_t)\phi_t - 1]\log(1 - y_t), \end{aligned}$$

com $\mu_t = g_\mu^{-1}(\eta_t)$ uma função de β e X_t, de mesma forma que $\phi_t = g_\phi^{-1}(\vartheta_t)$ uma função de γ e Z_t.

Para obter o valor de θ que maximiza a expressão (2.3.3) é necessário obter os vetores escores para β e ϕ que são as derivadas parciais de $\ell_t(\mu_t, \phi_t)$ com relação aos respectivos parâmetros. As entradas do vetor escore para β com $i = 1, \ldots, k$ são

$$U_{\beta_i}(\theta) = \sum_{t=1}^{n} \phi_t(y_t^* - \mu_t^*) \frac{1}{g_\mu{}'(\mu_t)} x_{ti},$$

em que x_{ti} é o t-ésimo valor da i-ésima covariável, $g_\mu'(\mu_t)$ é a primeira derivada de $g_\mu(\cdot)$ avaliada em μ_t, $y_t^* = \log(y_t/(1-y_t))$ e $\mu_t^* = \mathrm{E}(y_t^*)$.

Sendo $\Phi = \mathrm{diag}\{\phi_1, \ldots, \phi_n\}$ uma matriz diagonal, obtemos que o vetor escore para β pode ser escrito por

$$U_\beta(\theta) = X^\top \Phi T(y^* - \mu^*). \tag{2.3.4}$$

As entradas do vetor escore para γ com $j = 1, \ldots, q$ são

$$U_{\gamma_j}(\theta) = \sum_{t=1}^{n} \alpha_t \frac{1}{g_\phi{}'(\phi_t)} z_{tj},$$

em que $\alpha_t = \mu_t(y_t^* - \mu_t^*) + \log(1-y_t) + \psi(\phi_t) - \psi((1-\mu_t)\phi_t)$, $\psi(\cdot)$ é a função digama.

Considere uma matriz $n \times q$ em que cada coluna representa os n valores observados para a j-ésima covariável, $j = 1, 2, \ldots, q$, $H = \mathrm{diag}\{1/g_\phi'(\phi_1), \ldots, 1/g_\phi'(\phi_n)\}$ e $\alpha = (\alpha_1, \ldots, \alpha_n)^\top$.

O vetor escore para γ fica dado por

$$U_\gamma(\theta) = Z^\top H \alpha. \tag{2.3.5}$$

Assim, o estimador de máxima verossimilhança para θ é $\hat{\theta} = (\hat{\beta}^\top, \hat{\gamma}^\top)^\top$, e este é obtido por meio da solução do seguinte sistema de equações:

$$U_\beta(\theta) = 0,$$

$$U_\gamma(\theta) = 0.$$

Constata-se que não é possível definir esses estimadores de modo analítico devido à ausência de uma expressão fechada. Em outras palavras, o sistema acima não é linear em β e γ. Dessa forma, é imprescindível empregar métodos de otimização não linear para

maximizar o logaritmo da função de verossimilhança com relação a θ, obtendo, assim, as estimativas numéricas para θ. Ferrari e Cribari-Neto (2004), por exemplo, mencionam o uso do algoritmo de *Newton-Raphson* e até sugerem valores iniciais para facilitar o processo de convergência.

Após a obtenção das estimativas pontuais dos parâmetros do modelo, podemos construir estimativas intervalares para os parâmetros de interesse e, posteriormente, realizar testes de hipóteses. Para isso, é essencial quantificar a variabilidade desses estimadores. Sob condições usuais de regularidade e para amostras grandes, OSPINA (2007) estabelece que $\hat{\theta} = (\hat{\beta}^\top, \hat{\gamma}^\top)^\top$ segue uma distribuição conjunta aproximadamente normal $(k+q)$-variada:

$$\hat{\theta} = \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} \sim^a N_{k+q}\left(\begin{pmatrix} \beta \\ \gamma \end{pmatrix} \; ; K^{-1} \right), \tag{2.3.6}$$

em que K é a matriz de Informação de Fisher do vetor de parâmetros θ definida por

$$K = K(\theta) = \begin{bmatrix} K_{\beta\beta} & K_{\beta\gamma} \\ K_{\gamma\beta} & K_{\gamma\gamma} \end{bmatrix},$$

em que

- $K_{\beta\beta} = X^\top \Phi W X$;
- $K_{\beta\gamma} = K_{\gamma\beta}^\top = X^\top C T H Z$;
- $K_{\gamma\gamma} = Z^\top D Z$;
- $W = \text{diag}\{w_1, \ldots, w_n\}$;
- $C = \text{diag}\{c_1, \ldots, c_n\}$;
- $D = \text{diag}\{d_1, \ldots, d_n\}$;
- $w_t = \phi_t\{\psi'(\mu_t\phi_t) + \psi'((1-\mu_t)\phi_t)\}\{d\mu_t/d\eta_t\}^2$;
- $c_t = \phi_t\{\psi'(\mu_t\phi_t)\mu_t - \psi'((1-\mu_t)\phi_t)(1-\mu_t)\}$;
- $d_t = [\psi'(\mu_t\phi_t)\mu_t^2 + \psi'((1-\mu_t)\phi_t)(1-\mu_t)^2 - \psi'(\phi_t)][g_\phi(\phi_t)]^2$.

OSPINA (2007) mostra que a inversa da Matriz de Informação de Fisher é expressa por

$$K^{-1} = K^{-1}(\theta) = \begin{bmatrix} K^{\beta\beta} & K^{\beta\gamma} \\ K^{\gamma\beta} & K^{\gamma\gamma} \end{bmatrix},$$

em que

- $K^{\beta\beta} = (X^\top \Phi W X - X^\top C T H Z (Z^\top D Z)^{-1} Z^\top H T C^\top X)^{-1}$;
- $K^{\beta\gamma} = (K^{\gamma\beta})^\top = -K^{\beta\beta} X^\top C T H Z (Z^\top D Z)^{-1}$;
- $K^{\gamma\gamma} = (Z^\top D Z)^{-1}\{I_q + (Z^\top H T C^\top X) K^{\beta\beta} X^\top C T H Z (Z^\top D Z)^{-1}\}$.

A partir do resultado acima, é possível obter de maneira assintótica a variabilidade do MLE sob regressão beta, a partir da diagonal principal da matriz K^{-1}. Conforme estabelecido em (2.3.6), para a s-ésima componente de $\hat{\theta}$, denotada como $\hat{\theta}_s$, temos que para obtenção de intervalos de confiança aproximados para os parâmetros do modelo de regressão beta, considere que

$$\frac{\left(\hat{\theta}_s - \theta_s\right)}{\sqrt{K(\theta)^{ss}}} \overset{a}{\sim} N(0,1),$$

em que $K(\theta)^{ss}$ é o (s,s)-ésimo elemento de $K(\theta)^{-1}$. Assim, considerando um nível de confiança de $100(1-\alpha)\%$, um intervalo de confiança aproximado para a s-ésima componente do vetor de parâmetros θ é

$$\left(\hat{\theta}_s - z_{1-\alpha/2}\sqrt{K(\hat{\theta})^{ss}}; \hat{\theta}_s + z_{1-\alpha/2}\sqrt{K(\hat{\theta})^{ss}}\right),$$

em que $z_{1-\alpha/2}$ denota o quantil da distribuição normal padrão tal que $P(Z < z_{1-\alpha/2}) = 1 - \alpha/2$, com $Z \sim N(0,1)$.

O modelo de regressão beta com precisão constante anteriormente definido por Ferrari e Cribari-Neto (2004) é visto como um caso particular do modelo discutido nesta seção, desde que é obtido supondo que $\phi_i = \phi, \forall i = 1, \ldots, n$

3 Métodos Robustos

Neste capítulo apresentaremos os estimadores recentemente propostos para realizar inferência dos parâmetros sob os modelos de regressão beta discutidos na Seção 2.3.

3.1 MDPDE

Ghosh (2019) propôs o estimador de mínima divergência potência entre densidades (*minimum density power divergence estimator*; MDPDE) para o modelo de regressão beta. O MDPDE é um estimador robusto baseado na minimização da versão empírica da divergência de potência de densidades (GHOSH, 2019) definida por

$$\mathcal{H}(\theta) = \sum_{i=1}^{n} \left[\mathcal{K}_{i,1+\alpha}(\theta) - \frac{1+\alpha}{\alpha} f_\theta(y_i; \mu_i, \phi_i)^\alpha \right],$$

em que $\alpha \geq 0$ é uma constante de afinação (*tuning constant*), e

$$\mathcal{K}_{i,1+\alpha}(\theta) = \frac{B((1+\alpha)(\mu_i\phi_i - 1) + 1, (1+\alpha)[(1-\mu_i)\phi_i - 1] + 1)}{B(\mu_i\phi_i(1-\mu_i)\phi_i)^{1+\alpha}}.$$

Assim, o MDPDE é o valor de θ solução do seguinte sistema de equações:

$$\sum_{i=1}^{n} [U(y_i; \theta) f_\theta(y_i; \mu_i, \phi_i)^\alpha - \mathcal{E}_{i,1-\alpha}(\theta)] = 0, \tag{3.1.1}$$

em que

- $f_\theta(\cdot; \mu_i, \phi_i)$ é a função de densidade de probabilidade da distribuição beta reparametrizada com média μ_i e precisão ϕ_i;
- $U(y_i; \theta) = \nabla_\theta \log(f_\theta(y_i; \mu_i, \phi_i))$ é o vetor escore para a i-ésima observação;
- $\mathcal{E}_{i,1-\alpha}(\theta) = E[U(Y_i; \theta) f_\theta(y_i; \mu_i, \phi_i)^\alpha]$ é uma constante de correção da equação de estimação.

O fator $f_\theta(y_i; \mu_i, \phi_i)^\alpha$ age como uma ponderação da i-ésima observação no processo de estimação. Vale observar que, se $\alpha = 0$ tem-se o estimador de máxima verossimilhança tradicional. Ainda, quando $\alpha \geq 1$ o estimador se torna bastante ineficiente, por essas razões, vamos nos restringir a $\alpha \in [0, 1)$, o que gera estimadores robustos. Se y_i é uma observação atípica, para α se aproximando de 1, $f_\theta(y_i; \mu_i, \phi_i)^\alpha$ será um valor pequeno

próximo a zero. Vale ressaltar que, segundo Ghosh (2019), as quantidades $K_{i,1+\alpha}(\theta)$ e $E_{i,1-\alpha}(\theta)$ estão bem definidas para $\mu_i\phi_i > \alpha/(1+\alpha)$ e $(1-\mu_i)\phi > \alpha/(1+\alpha)$. Logo, o MDPDE estará bem definido para todo $0 \leq \alpha < 1$ desde que $f_\theta(y_i;\mu_i,\phi_i)$ seja uma densidade beta limitada.

A matriz de covariâncias assintóticas do MDPDE é dada por

$$V_n(\theta) = \Psi_n(\theta)^{-1}\Omega_n(\theta)\Psi_n(\theta)^{-1^\top},$$

em que

$$\Psi_n(\theta) = \begin{pmatrix} X^\top\gamma_{11}^{(1+\alpha)}X & X^\top\gamma_{12}^{(1+\alpha)}Z \\ Z^\top\gamma_{12}^{(1+\alpha)}X & Z^\top\gamma_{22}^{(1+\alpha)}Z \end{pmatrix},$$

e

$$\Omega_n(\theta) = \begin{pmatrix} X^\top\left[\gamma_{11}^{(1+2\alpha)} - \gamma_1^{(1+\alpha)^2}\right]X & X^\top\left[\gamma_{12}^{(1+2\alpha)} - \gamma_1^{(1+\alpha)}\gamma_2^{(1+\alpha)}\right]Z \\ Z^\top\left[\gamma_{12}^{(1+2\alpha)} - \gamma_1^{(1+\alpha)}\gamma_2^{(1+\alpha)}\right]X & Z^\top\left[\gamma_{22}^{(1+2\alpha)} - \gamma_2^{(1+\alpha)^2}\right]Z \end{pmatrix}.$$

A matriz de covariâncias assintóticas é obtida a partir da distribuição assintótica dos M-estimadores, desde que o MDPDE pode ser visto como um M-estimador.

3.2 SMLE

Outro estimador robusto sob regressão beta foi proposto por Ribeiro e Ferrari (2023) que baseia-se na maximização da função de L_q-verossimilhança reparametrizada definida por

$$\ell_\alpha^*(\theta) = \sum_{i=1}^{n} L_{1-\alpha}(f_\theta^*(y_i;\mu_i,\phi_i)),$$

em que

- $f_\theta^*(y_i;\mu_i,\phi_i) = f_\theta(y_i;\mu_{i,(1-\alpha)^{-1}},\phi_{i,(1-\alpha)^{-1}})$;
- $\mu_{i,\alpha} = \phi_{i,\alpha}^{-1}[\alpha(\mu_i\phi_i - 1) + 1]$, $\phi_{i,\alpha} = \alpha(\phi_i - 2) + 2$;
- $\alpha \in (0,1]$ é a constante de afinação;
- $L_\alpha(u) = (u^\alpha - 1)/\alpha$, para $\alpha \in (0,1]$;
- $L_\alpha(u) = \log(u)$, para $\alpha = 0$.

De forma equivalente, o estimador proposto por Ribeiro e Ferrari (2023) é obtido como o valor de θ que soluciona o seguinte sistema de equações de estimação

$$\sum_{i=1}^{n} U^*(y_i;\theta) f_\theta^*(y_i;\mu_i,\phi_i)^\alpha = 0, \tag{3.2.1}$$

em que $U^*(y_i;\theta) = \nabla_\theta \log[f_\theta^*(y_i;\mu_i,\phi_i)]$ é o vetor escore modificado para θ correspondente à i-ésima observação dado por

$$U^*(y_i;\theta) = \left((1-\alpha)^{-1}\phi_{i,(1-\alpha)} \frac{(y_i^* - \mu_i^*)}{g_\mu^{'}(\mu_{i,(1-\alpha)})} X_i^\top, (1-\alpha)^{-1} \frac{\mu_{i,(1-\alpha)}(y_i^* - \mu_i^*) + y_i^\dagger - \mu_i^\dagger}{g_\phi^{'}(\phi_{i,(1-\alpha)})} Z_i^\top \right)^\top,$$

em que, $y_i^\dagger = \log(1-y_i)$ e $\mu_i^\dagger = E(Y_i^\dagger) = \psi((1-\mu_i)\phi_i) - \psi(\phi_i)$. Assim, de forma semelhante a (3.1.1) o fator $f_\theta^*(y_i;\mu_i,\phi_i)^\alpha$ em (3.2.1) funciona como uma ponderação no processo de estimação, e, ainda de forma semelhante, se $\alpha = 0$ tem-se o estimador de máxima verossimilhança tradicional. Ribeiro e Ferrari (2023) denominaram este estimador de estimador de máxima verossimilhança substituto (*surrogate maximum likelihood estimator*; SMLE). A matriz de covariâncias assintótica do SMLE é dada por (RIBEIRO; FERRARI, 2023)

$$V_{1-\alpha}(\theta) = J_{1-\alpha}(\theta)^{-1} K_{1-\alpha}(\theta) J_{1-\alpha}(\theta)^{-1^\top},$$

em que

$$J_{1-\alpha}(\theta) = (\alpha-1)^{-1} \begin{pmatrix} X^\top B_1 T_\mu^{*2} \Phi_q^2 V X & X^\top B_1 T_\mu^* T_\phi^* C^* Z \\ Z^\top B_1 T_\mu^* T_\phi^* C^* X & Z^\top B_1 T_\phi^{2*} T_\phi^* D^* X \end{pmatrix},$$

$$K_{1-\alpha}(\theta) = (1-\alpha)^{-2} \begin{pmatrix} X^\top B_2 T_\mu^{*2} \Phi_q^2 (V_{1+\alpha} + M_1^2)) X & X^\top B_2 T_\mu^* T_\phi^* (C_{1+\alpha}^* + M_3) Z \\ Z^\top B_2 T_\mu^* T_\phi^* (C_{1+\alpha}^* + M_3) X & Z^\top B_2 T_\phi^{2*} T_\phi^* (D_{1+\alpha}^* + M_2^2) Z \end{pmatrix},$$

com $B_1 = diag\{b_{1,1},\ldots,b_{n,1}\}$, $B_2 = diag\{b_{1,2},\ldots,b_{n,2}\}$,

$$b_{i_1} = \frac{B(\mu_i\phi_i,(1-\mu_i)\phi_i)^{1-\alpha}}{B(\mu_{i,(1-\alpha)}\phi_{i,(1-\alpha)},(1-\mu_{i,(1-\alpha)})\phi_{i,(1-\alpha)})},$$

$$b_{i,2} = \frac{B(\mu_{i,(1+\alpha)}\phi_{i,(1+\alpha)},(1-\mu_{i,(1+\alpha)})\phi_{i,(1+\alpha)})}{B(\mu_i\phi_i,(1-\mu_i)\phi_i)^{2\alpha} B(\mu_{i,(1-\alpha)}\phi_{i,(1-\alpha)}(1-\mu_{i,(1-\alpha)})\phi_{i,(1-\alpha)}}.$$

Analogamente ao MDPDE, as quantidades necessárias para realizar a inferência usando o SMLE estarão bem definidas para betas limitadas. Os estimadores SMLE e MDPDE são

semelhantes entre si, entretanto não são equivalentes, exceto quando $\alpha = 0$. A principal diferença entre estes dois estimadores está na forma em que se atinge a não viciosidade das respectivas equações de estimação.

Para ambos os estimadores citados, a escolha da constante de afinação é crucial para obtenção das estimativas dos parâmetros, e, quanto maior o valor de $\alpha \in [0, 1)$ mais robusto é o processo, porém, menos eficiente este se torna. A escolha de α será discutida na Seção 3.4. Adicionalmente, o MDPDE e o SMLE tem boas propriedades, como B-robustez, V-robustez e normalidade assintótica (GHOSH, 2019; RIBEIRO; FERRARI, 2023; MALUF; FERRARI; QUEIROZ, 2024).

3.3 Estimadores Robustos sob a Transformação Logit

Devido a limitação do SMLE e MDPDE estarem bem definidos para todo $\alpha \in [0, 1)$, desde que supõe-se uma distribuição beta limitada, Maluf, Ferrari e Queiroz (2024) propõem novas versões do MDPDE e do SMLE visando eliminar tais limitações em suas obtenções por meio da transformação *logit* que é definida por

$$y^* = \log\left(\frac{y}{1-y}\right).$$

Se $Y \sim B(\mu, \phi)$, a função de densidade de Y^* é dada por

$$h(y^*; \mu, \phi) = \frac{1}{B(\mu\phi, (1-\mu)\phi)} \frac{\mathrm{e}^{-y^*(1-\mu)\phi}}{(1+\mathrm{e}^{-y^*})^{\phi}}.$$

Nesse caso, a densidade resultante do logito de uma variável beta é conhecida como beta exponencial generalizada do segundo tipo (*exponential generalized beta of the second type*; EGB). Escrevemos $Y^* \sim \mathrm{EGB}(\mu, \phi)$.

Os estimadores robustos para os parâmetros do modelo de regressão beta usando a função de densidade da variável resposta transformada são construídos com base em Ghosh (2019) e Ribeiro e Ferrari (2023). São eles, o *logit minimum densisty power divergence estimator* (LMDPDE) e o *logit surrogate maximum likehood estimator* (LSMLE). Maluf, Ferrari e Queiroz (2024) descrevem as vantagens de trabalhar com os estimadores transformados. A ideia geral dos novos estimadores é obter estimadores do tipo MDPDE e SMLE, entretanto usando a função densidade de Y^* ao invés de usar a função densidade de Y. Ao fazer isto, estaremos estimando os mesmos parâmetros, sem quaisquer restrições para obtenção dos estimadores robustos.

Para obter o LMDPDE, Maluf, Ferrari e Queiroz (2024) propõem o estimador que minimiza a função

$$\mathcal{H}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{V}_i(y_i^*, \theta),$$

em que

$$\mathcal{V}_i(y_i^*, \theta) = \mathcal{K}_{i,1+\alpha}(\theta) - \frac{1+\alpha}{\alpha}h_\theta(y_i^*; \mu_i, \phi_i)^\alpha,$$

e

$$\mathcal{K}_{i,1+\alpha}(\theta) = \int_{-\infty}^{\infty} h_\theta(y_i^*; \mu_i, \phi_i)^{1+\alpha}dy^* = \frac{B(\mu_i\phi_i(1+\alpha), (1-\mu_i)\phi_i(1+\alpha))}{B(\mu_i\phi_i, (1-\mu_i)\phi_i)^{1+\alpha}},$$

com $0 \leq \alpha < 1$. A equação de estimação é dada por

$$\sum_{i=1}^{n}[U(y_i;\theta)h_\theta(y_i^*;\mu_i,\phi_i)^\alpha - \mathcal{E}_{i,1-\alpha}(\theta)] = 0.$$

Analogamente, o LSMLE é obtido a partir da maximização em θ da função

$$\ell_{1-\alpha}^*(\theta) = \sum_{i=1}^{n} L_{1-\alpha}(h_\theta^*(y_i^*;\mu_i,\phi_i)),$$

em que

$$h_\theta^*(y_i^*;\mu_i,\phi_i) = h_\theta(y_i^*;\mu_i,\phi_{i,(1-\alpha)^{-1}}),$$

e que soluciona a equação de estimação

$$\sum_{i=1}^{n} U^*(y_i^*;\theta)h_\theta^*(y_i^*;\mu_i,\phi_i)^\alpha = 0,$$

em que $U^*(y_i^*;\theta) = \nabla_\theta \log h_\theta^*(y_i^*;\mu_i,\phi_i)$ é o vetor escore modificado para a i-ésima observação, dado por

$$U^*(y_i^*;\theta) = \left(\phi_i\frac{(y_i^*-\mu_i^*)}{g_\mu'(\mu_i)}X_i^\top, (1-\alpha)^{-1}\frac{\mu_i(y_i^*-\mu_i^*)+(y_i^\dagger-\mu_i^\dagger)}{g_\phi'(\phi_{i,1-\alpha})}Z_i^\top\right)^\top.$$

É importante comentar que esses estimadores possuem algumas propriedades importantes, como a normalidade assintótica, B-robustez e V-robustez, Fisher-consistência (MALUF; FERRARI; QUEIROZ, 2024). Maluf, Ferrari e Queiroz (2024) explicitam as matrizes de

covariâncias assintóticas do LMDPDE e LSMLE como se segue. Considere $\hat{\theta}_\alpha = (\hat{\beta}_\alpha^\top, \hat{\gamma}_\alpha^\top)^\top$ e $\tilde{\theta}_\alpha = (\tilde{\beta}_\alpha^\top, \tilde{\gamma}_\alpha^\top)^\top$ o LMDPDE e o LSMLE, respectivamente, para $\alpha \in [0, 1)$ fixo. Dado que são M-estimadores, temos que $\hat{\theta}_\alpha \overset{a}{\sim} N(\theta, V_{1,\alpha}(\theta))$ e $\tilde{\theta}_\alpha \overset{a}{\sim} N(\theta, V_{2,\alpha}(\theta))$, em que $\overset{a}{\sim}$ denota distribuição assintótica, assim

$$V_{1,\alpha}(\theta) = \Lambda_{1,\alpha}^{-1}(\theta)\Sigma_{1,\alpha}(\theta)\Lambda_{1,\alpha}^{-1}(\theta),$$

$$V_{2,\alpha}(\theta) = \Lambda_{2,\alpha}^{-1}(\theta)\Sigma_{2,\alpha}(\theta)\Lambda_{2,\alpha}^{-1}(\theta),$$

em que

$$\Lambda_{1,\alpha}(\theta) = \begin{pmatrix} X^\top \gamma_{11}^{1+\alpha} X & X^\top \gamma_{12}^{1+\alpha} Z \\ Z^\top \gamma_{12}^{1+\alpha} X & Z^\top \gamma_{22}^{1+\alpha} Z \end{pmatrix},$$

$$\Sigma_{1,\alpha}(\theta) = \begin{pmatrix} X^\top \left[\gamma_{11}^{(1+2\alpha)} - \gamma_1^{(1+\alpha)^2}\right] X & X^\top \left[\gamma_{12}^{(1+2\alpha)} - \gamma_1^{(1+\alpha)}\gamma_2^{(1+\alpha)}\right] Z \\ Z^\top \left[\gamma_{12}^{(1+2\alpha)} - \gamma_1^{(1+\alpha)}\gamma_2^{(1+\alpha)}\right] X & Z^\top \left[\gamma_{22}^{(1+2\alpha)} - \gamma_2^{(1+\alpha)^2}\right] Z \end{pmatrix},$$

$$\Lambda_{2,\alpha}(\theta) = \begin{pmatrix} (\alpha-1) X^\top B_1 T_\mu^2 \Phi^2 V X & X^\top B_1 T_\mu T_\phi^* C Z \\ Z^\top B_1 T_\mu T_\phi^* C X & (\alpha-1)^{-1} Z^\top B_1 T_\phi^{*2} D Z \end{pmatrix},$$

e

$$\Sigma_{2,\alpha}(\theta) = (1-\alpha)^{-1} \begin{pmatrix} X^\top B_2 T_\mu^2 \Phi^2 (V_{1+\alpha}) X & (1-\alpha)^{-1} X^\top B_2 T_\mu T_\phi^* C_{1+\alpha} Z \\ (1-\alpha)^{-1} Z^\top B_2 T_\mu T_\phi^* C_{1+\alpha} X & (1-\alpha)^{-2} Z^\top B_2 T_\phi^{2*} D_{1+\alpha} Z \end{pmatrix}.$$

Quando $\alpha = 0$, estas matrizes de covariâncias assintóticas equivalem à matriz de covariâncias assintóticas do MLE.

3.4 Constante de Afinação

Nesta subseção discutiremos brevemente o processo de seleção da constante de afinação α. Essa constante desempenha um papel fundamental no equilíbrio entre a robustez e a eficiência dos estimadores em modelos de regressão beta robusta.

Ribeiro e Ferrari (2023) propuseram um algoritmo para determinar a constante de afinação para o SMLE. O objetivo é identificar o valor de α, em uma grade ordenada variando de $\alpha = 0$ até um α_{max}, mais próximo de zero, que resulte em estimativas sufici-

entemente estáveis dos parâmetros, preservando a eficiência total em dados não contaminados. Caso o algoritmo não atinja estabilidade até α_{max}, o valor $\alpha = 0$ (correspondente ao MLE) é selecionado. A escolha do valor ótimo da constante de afinação é uma etapa crítica especialmente em aplicações com dados reais. Valores grandes de α aumentam a robustez dos estimadores, mas podem comprometer sua eficiência assintótica. Os quatro estimadores robustos discutidos nesse trabalho terão, em aplicações práticas, a constante α escolhida pelo método orientado pelos dados proposto por Ribeiro e Ferrari (2023).

O procedimento proposto tem como objetivo selecionar um valor ideal para α que seja o mais próximo de 0, desde que atenda ao critério de estabilidade nas estimativas. Caso a estabilidade não seja alcançada, o algoritmo define $\alpha = 0$, resultando no estimador MLE. Segundo Ribeiro e Ferrari (2023), essa escolha pode ser necessária em situações onde as distribuições beta não são limitadas, desde que que o SMLE e o MDPDE podem não ser adequadamente definidos nesses casos, gerando estimativas instáveis.

3.5 Teste de Hipótese Robusto

Em aplicações práticas, é necessário avaliar se os coeficientes das covariáveis são estatísticamente significantes para explicar o comportamento médio da variável resposta. Como indicado por Ribeiro e Ferrari (2023), testes de hipóteses baseados no MLE que normalmente dependem da função de verossimilhança são sensíveis à observações atípicas e podem ter seu desempenho prejudicado em cenários de dados contaminados.

Ribeiro e Ferrari (2023) propuseram um teste de hipótese robusto que indica a nova estatística de teste como a estatística tipo-Wald. Considere a hipótese nula $H_0 : \theta_k = \theta_k^{(0)}$ contra a hipótese alternativa bilateral $H_1 : \theta_k \neq \theta_k^{(0)}$. A estatística tipo-Wald é definida por

$$W_{0,q} = \frac{\left(\hat{\theta}_{kq} - \theta_k^{(0)}\right)^2}{SE(\hat{\theta}_{kq})^2},$$

em que $\hat{\theta}_{kq}$ é o SMLE de θ_k e $SE(\hat{\theta}_{kq})$ é o erro padrão assintótico obtido a partir da matriz de covariâncias assintóticas. Este mesmo procedimento pode ser aplicado usando o MDPDE. Maluf, Ferrari e Queiroz (2024) expandem o uso deste teste para os outros estimadores vistos, substituindo as estimativas e erros padrão assintóticos pelos valores dos respectivos estimadores. A estatística $W_{0,q}$ possui uma distribuição assintótica qui-quadrado com um grau de liberdade ($W_{0,q} \sim \chi_1^2$) sob a hipótese nula H_0. Para realizar o teste de nulidade dos coeficientes da regressão basta especificar $\theta_k^{(0)} = 0$. Esta abordagem

será utilizada no Capítulo 4.

4 Aplicação

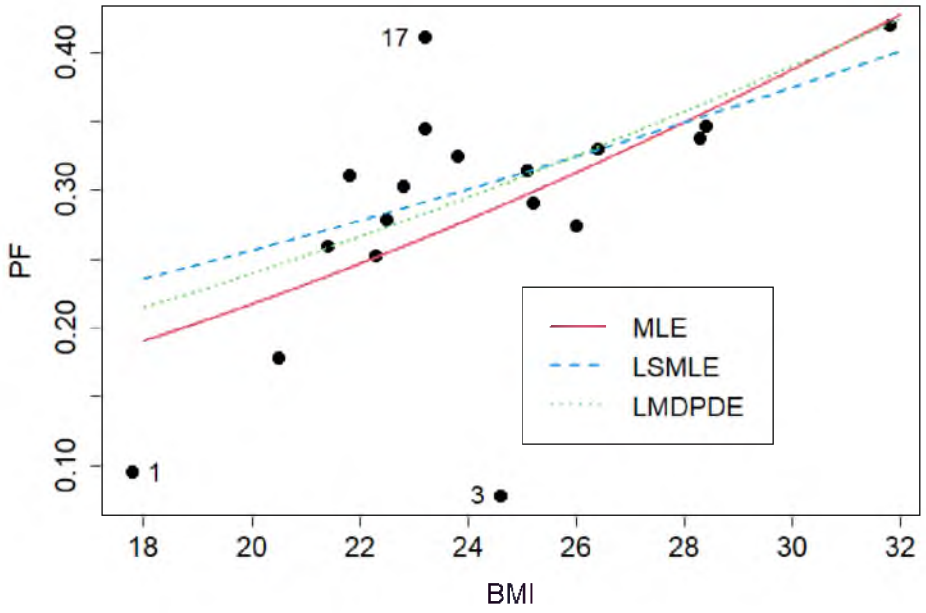
Neste capítulo ilustraremos o uso dos estimadores robustos sob regressão beta através da análise dos dados previamente apresentados no Capítulo 1. O conjunto de dados a ser analisado são os dados *humanfat*, disponível no pacote GLMsData do R. Esse conjunto contém informações sobre a composição corporal de 18 indivíduos, distribuídas em quatro variáveis. O objetivo é modelar a variável resposta, porcentagem de gordura corporal em função das variáveis explicativas, que incluem a idade em anos completos (*Age*; AGE), o índice de massa corpórea, que corresponde a, uma medida amplamente utilizada para avaliar a relação entre peso e altura, e a variável categórica gênero (*Gender*; GEN), que classifica os indivíduos como femininos (F) ou masculinos (M), sendo o gênero feminino utilizado como referência na modelagem.

O objetivo desta aplicação é comparar os diferentes métodos de estimação em termos de resultados e robustez, analisando o comportamento de cada estimador na modelagem da porcentagem de gordura corporal. A variável idade não será considerada na análise, pois sua inclusão não foi relevante na análise.

Inicialmente, considere o modelo de regressão beta com precisão constante dado por $\log(\mu_i/(1-\mu_i)) = \beta_1 + \beta_2 \text{BMI}_\text{i}$ e $\log(\phi_i) = \gamma_1$, em que μ_i e ϕ_i são a média e precisão da variável resposta PF para o i-ésimo indivíduo, $i = 1, \ldots, 18$. Ajustamos o modelo utilizando os estimadores robustos SMLE, MDPDE, LSMLE e LMDPDE para os dados completos. Denominamos este de Modelo 1. O valor ótimo selecionado para a constante α pelo estimador LSMLE e para o LMDPDE foi de $0,32$ e $0,24$ respectivamente, enquanto os outros estimadores robustos indicaram $\alpha = 0$, ou seja, indicando que não há necessidade de robustez e ajustando o modelo pelo estimador clássico não robusto MLE. É importante lembrar que as equações de estimação dos estimadores SMLE e MDPDE não são bem definidas a não ser que $\mu_i\phi_i > \alpha$ e $(1-\mu_i)\phi_i > \alpha$ (MALUF; FERRARI; QUEIROZ, 2024). Assim, esta pode ser a explicação da escolha de $\alpha = 0$ sob esses estimadores.

A Figura 3 apresenta o gráfico de dispersão de PF versus BMI com as curvas ajustadas do MLE, LSMLE e LMDPDE para os dados completos. Percebemos que as curvas geradas pelos diferente métodos de estimação se distinguem ligeiramente, de forma que o LSMLE se diferencia mais do MLE enquanto o LMDPDE se posiciona entre as outras duas estimativas. Durante a análise dos dados, identificamos visualmente que os pontos 1, 3 e 17 se destacam por estarem mais afastados da massa principal dos dados. Esses pontos podem ser possíveis outliers ou observações que apresentam características distintas em relação ao restante do conjunto. A identificação desses valores é relevante,

Figura 3: Gráfico de dispersão de PF contra BMI juntamente com as curvas ajustadas via regressão beta baseadas no MLE, LSMLE e LMDPDE para os dados completos considerando o Modelo 1.



pois podemos acompanhar o efeito dessas observações na estimativa dos parâmetros e a robustez dos métodos utilizados. Observe que a curva estimada sob o MLE fica mais próxima aos ponto atípicos.

A Tabela 1 apresenta as estimativas, erros-padrão, estatísticas-z (estimativa dividida pelo erro padrão) e valores-p assintóticos do teste tipo-Wald da nulidade dos coeficientes. Podemos notar o comportamento dos estimadores robustos analisando as variações relativas nas estimativas em relação ao MLE. Para o submodelo da média, os estimadores robustos tem estimativas maiores para o intercepto e para o coeficiente de BMI, sendo para o intercepto uma redução de 26% e 12% para o LSMLE e LMDPDE, respectivamente, e uma redução um pouco maior para o coeficiente de BMI de 33% e 14% para o LSMLE e LMDPDE, respectivamente. Ainda, no submodelo da média, os erro padrão estimados, tanto do intercepto quanto do coeficiente de BMI, apresentam uma redução ainda mais evidente, sendo de 61% para o LSMLE e 38% para o LMDPDE. No caso da precisão, há um aumento significativo nas estimativas robustas em relação ao método clăssico, que para o LSMLE é de 61%, e de 31% para o MDPDE.

No contexto deste estudo, é importante ressaltar que a pequena dimensão amostral ($n = 18$) representa uma limitação, especialmente no que se refere ăs estimativas intervalares. Embora as estimativas pontuais obtidas pelos diferentes métodos de estimação sejam interpretáveis, a confiabilidade dos intervalos de confiança pode ser afetada. Isso

ocorre porque as matrizes de covariância utilizadas na inferência são assintóticas, ou seja, sua validade teórica depende do pressuposto de amostras grandes. Com um número reduzido de observações, a aproximação assintótica pode não ser precisa, o que pode comprometer a acurácia das inferências baseadas nos intervalos de confiança.

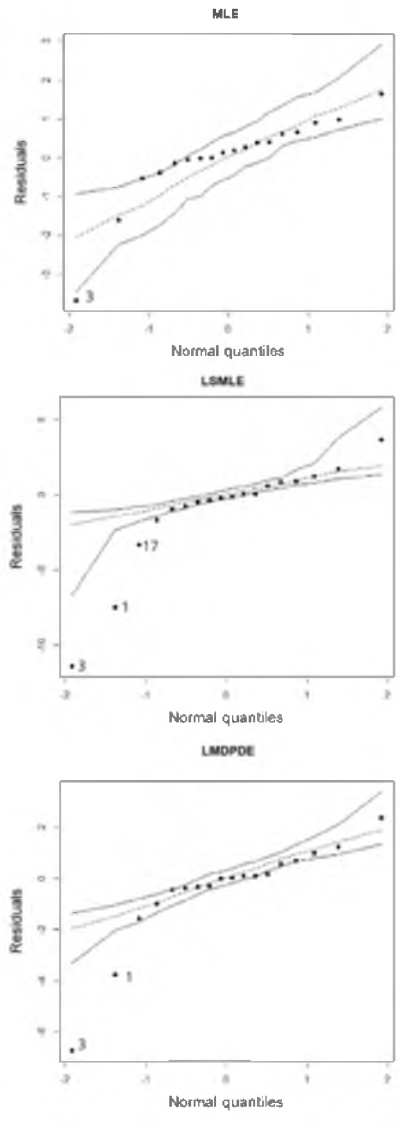
Tabela 1: Estimativas, erros padrão, estatística z e p-valor sob os modelos de regressão beta ajustados considerando o Modelo 1.

	MLE			
	Estimativa	Erro padrão	z	p-valor
média				
intercepto	−2.93245	0.75054	−3.907	< 0.001
BMI	0.0825	0.03038	2.716	0.006
precisão				
intercepto	3.327	0.3287	10.12	< 0.001
	LSMLE			
	Estimativa	Erro padrão	z	p-valor
média				
intercepto	−2.17338	0.29382	−7.397	< 0.001
BMI	0.05535	0.01195	4.633	< 0.001
precisão				
intercepto	5.3575	0.3906	13.72	< 0.001
	LMDPDE			
	Estimativa	Erro padrão	z	p-valor
média				
intercepto	−2.57342	0.46276	−5.516	< 0.001
BMI	0.07089	0.01877	3.776	< 0.001
precisão				
intercepto	4.3503	0.3507	12.41	< 0.001

Para avaliar a qualidade do ajuste sob regressão beta é comum o uso de gráficos de probabilidade normal dos resíduos com envelope simulado (RIBEIRO; FERRARI, 2023). A Figura 4 apresenta os gráficos de probabilidade normal dos resíduos com envelopes simulados sob os ajustes dos MLE, LSMLE e LMDPDE. Uma observação relevante é que o método clássico (MLE), ao buscar ajustar todos os pontos da amostra indiscriminadamente, pode mascarar a presença de discrepâncias nos dados, resultando em um ajuste global que tenta acomodar até mesmo observações atípicas. Em contraste, os métodos robustos, como o LSMLE e o LMDPDE, têm a capacidade de identificar automaticamente possíveis outliers e expô-los no gráfico. Essa característica faz com que os pontos discrepantes apresentem um ajuste visivelmente inferior nesses métodos, enquanto a massa central dos dados, que não contém valores extremos, recebe um ajuste mais pre-

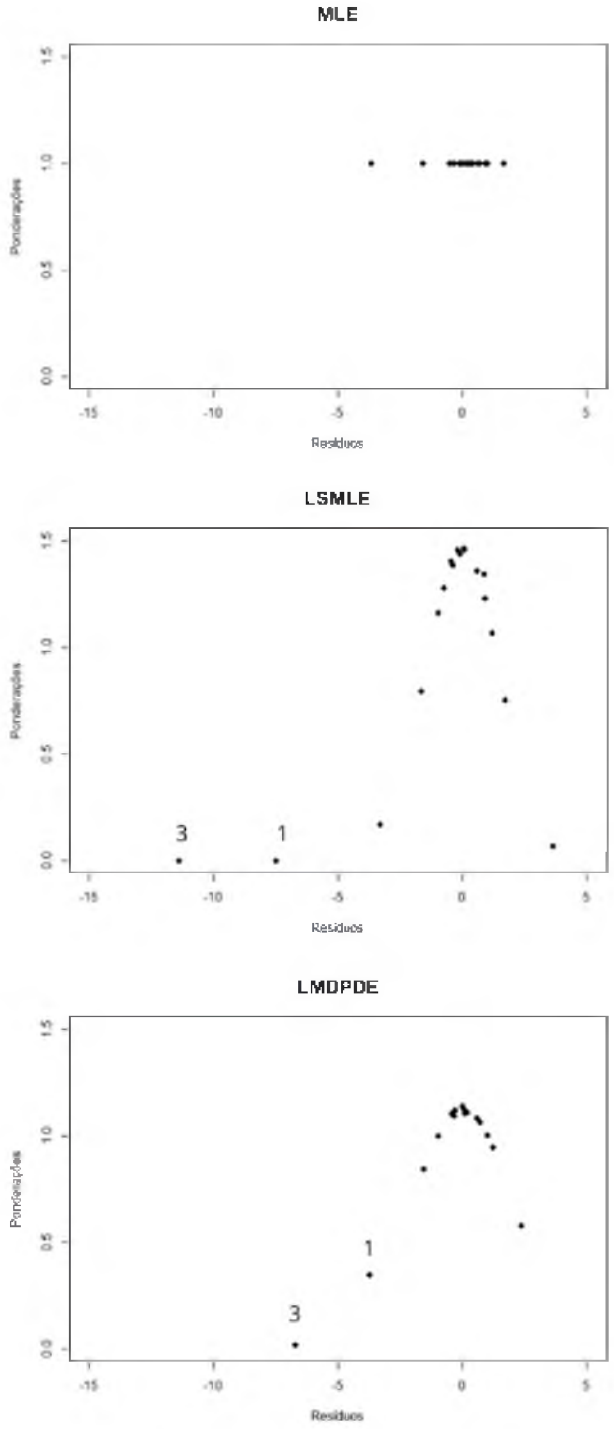
ciso. Esse comportamento é desejável no contexto da robustez estatística, pois garante que a inferência seja baseada principalmente nos dados mais representativos, reduzindo a influência de valores aberrantes na estimação dos parâmetros. A Figura 5 apresenta os gráficos dos resíduos contra as ponderações que ilustra que observações com resíduos altos tendem a receber ponderações baixas no procedimento de estimação robusto.

Figura 4: Gráficos de probabilidade normal dos resíduos com envelopes simulados sob MLE, LSMLE e LMDPDE considerando o Modelo 1.



Outra abordagem pode ser feita levando em consideração a variável gênero. Para isto, considere o modelo de regressão beta com precisão constante dado por $\log(\mu_i/(1-\mu_i)) = \beta_1 + \beta_2 \text{BMI}_\text{i} + \beta_3 \text{GEN}_\text{i}$ e $\log(\phi_i) = \gamma_1$, em que μ_i e ϕ_i são a média e precisão da variável resposta PF para o i-ésimo indivíduo, $i = 1, \ldots, 18$. Denominamos este de Modelo 2. Ajustamos o modelo utilizando os estimadores robustos SMLE, MDPDE, LSMLE e LMDPDE para os dados completos. O valor ótimo selecionado para a constante α pelo

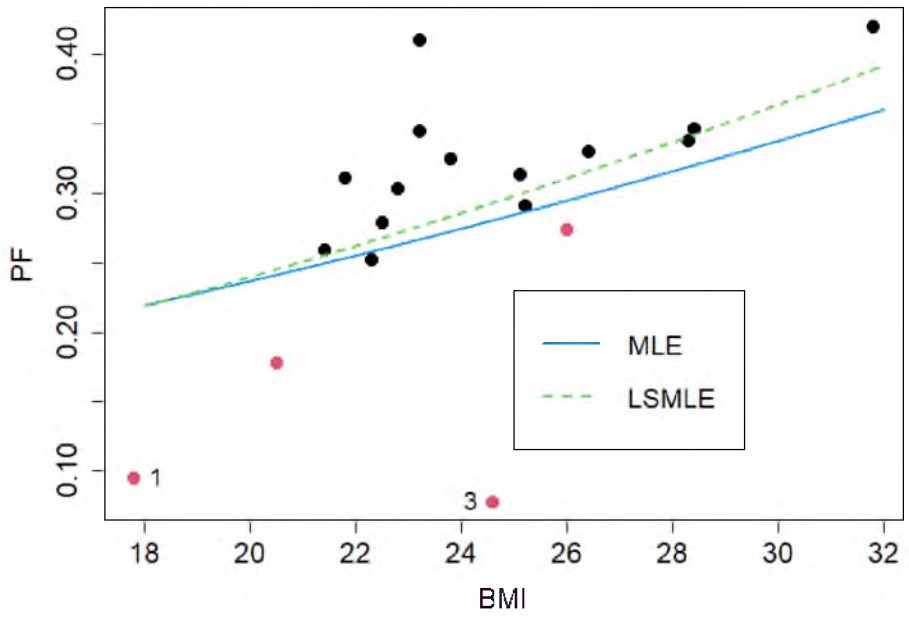
Figura 5: Gráficos dos resíduos contra as ponderações considerando o Modelo 1.



estimador LSMLE foi 0,28, enquanto que os outros estimadores, incluindo o LMDPDE, indicaram $\alpha = 0$, ou seja, indicando que não há necessidade de robustez, e ajustando o modelo pelo estimador clássico não robusto MLE. A Figura 6 apresenta o gráfico de dispersão de PF versus BMI com as curvas ajustadas do MLE e LSMLE para os dados completos, sendo que os pontos na cor preta indicam indivíduos do gênero feminino (F), que é o valor de referência, enquanto os pontos na cor vermelha indicam indivíduos do

gênero masculino (M).

Figura 6: Gráfico de dispersão de PF contra BMI com as curvas ajustadas via regressão beta baseadas no MLE e LSMLE considerando o Modelo 2.



Observa-se que os homens, em geral, apresentam uma porcentagem de gordura corporal menor do que as mulheres. Esse resultado é consistente com o conhecimento biológico, pois diferenças hormonais e de composição corporal influenciam a distribuição de gordura entre os sexos. Estudos indicam que, em média, as mulheres possuem uma maior proporção de gordura corporal devido a fatores como a função reprodutiva e a distribuição de tecido adiposo (WELLS, 2007). Esse padrão esperado reforça a importância de considerar o gênero ao modelar a porcentagem de gordura corporal em análises estatísticas.

A Tabela 2 apresenta as estimativas, erros-padrão, estatísticas-z e valores-p assintóticos do teste tipo-Wald da nulidade dos coeficientes para o modelo com a variável gênero.Ao comparar os coeficientes estimados nos dois modelos, MLE e LSMLE, observa-se variações significativas, especialmente no efeito da variável gênero. No submodelo da média, o coeficiente associado ao gênero no modelo MLE é de -0.89969, enquanto no LSMLE esse efeito é reduzido para -0.36968, representando uma diminuição percentual de aproximadamente 59%. Essa diferença indica que o modelo robusto atribui um efeito menor ao fator gênero na estimativa da porcentagem de gordura corporal em comparação com o modelo clássico. Essa redução pode ser explicada pela influência de pontos discrepantes na amostra. Como mencionado anteriormente, os pontos 1 e 3, identificados

como discrepantes, pertencem ao grupo masculino. No modelo ajustado sob o MLE, que busca ajustar todos os pontos da amostra sem distinção, esse valor atípico está ampliando artificialmente a estimativa do efeito do gênero. Já no modelo ajustado sob o LSMLE, que é robusto a outliers, há um menor impacto desse ponto extremo, resultando em uma estimativa mais conservadora para o efeito da variável gênero.

Tabela 2: Estimativas, erros padrão, estatística z e p-valor sob os modelos ajustados

	MLE			
	Estimativa	Erro padrão	z	p-valor
média				
intercepto	-1.96659	0.50290	-3.911	< 0.001
BMI	0.04974	0.02005	2.480	0.0131
GEN(M)	-0.89969	0.17810	-5.051	< 0.001
precisão				
intercepto	4.3258	0.3317	13.04	< 0.001
	LSMLE			
	Estimativa	Erro padrão	z	p-valor
média				
intercepto	-2.26517	0.26609	-8.513	< 0.001
BMI	0.05961	0.01059	5.629	< 0.001
GEN(M)	-0.36968	0.08653	-4.272	< 0.001
precisão				
intercepto	5.699	0.376	15.16	< 0.001

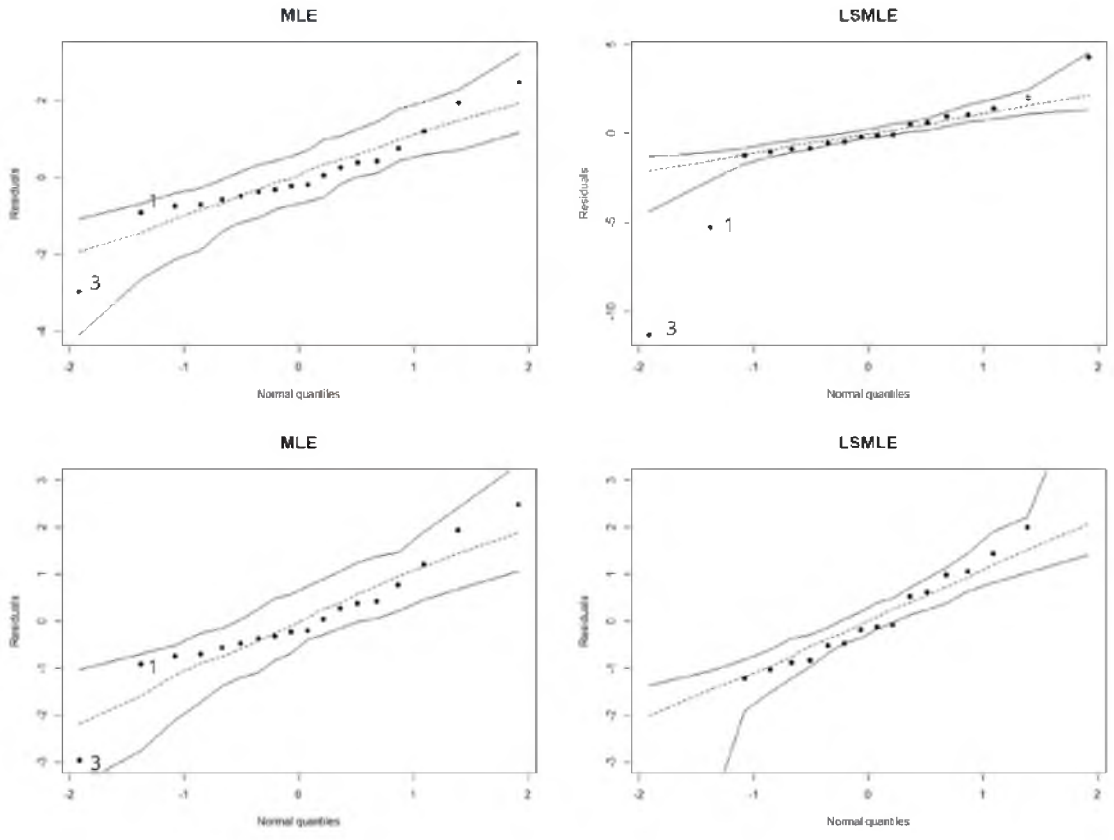
Ao comparar os erros-padrão estimados e os p-valores dos coeficientes entre os modelos ajustados sob MLE e LSMLE, algumas diferenças importantes podem ser observadas. De modo geral, os erros-padrão no submodelo da média são menores no modelo LSMLE em comparação com o modelo MLE. Essa redução sugere que o modelo robusto produz estimativas mais precisas, possivelmente por minimizar o impacto de pontos discrepantes que aumentam a variabilidade dos coeficientes no modelo clássico. Quanto aos p-valores, todos os coeficientes são altamente significativos ($p < 0.001$), com exceção do coeficiente de BMI no modelo MLE, que tem um p-valor de 0.0131. Esse valor indica que, sob um nível de significância de 5%, o coeficiente de BMI ainda seria considerado significativo, mas a diferença se torna relevante para um nível mais estrito de 1%. No modelo sob o LSMLE, o p-valor associado ao coeficiente de BMI é inferior a 0.001, reforçando que, ao reduzir a influência de pontos discrepantes, a relação entre BMI e percentual de gordura corporal se torna estatisticamente mais evidente.

Esses resultados reforçam a robustez do LSMLE, que, ao reduzir a influência de valores extremos sobre as estimativas, melhora a precisão inferencial dos coeficientes.

Nesse cenário, o objetivo principal não é ajustar bem os pontos discrepantes, mas sim garantir um melhor ajuste para a massa de dados principal. Dessa forma, mesmo que o modelo robusto apresente um "ajuste ruim" para os pontos mais afastados, isso é intencional, pois prioriza uma modelagem mais estável e representativa do comportamento geral dos dados, minimizando o impacto de observações atípicas sobre as estimativas do modelo de regressão beta.

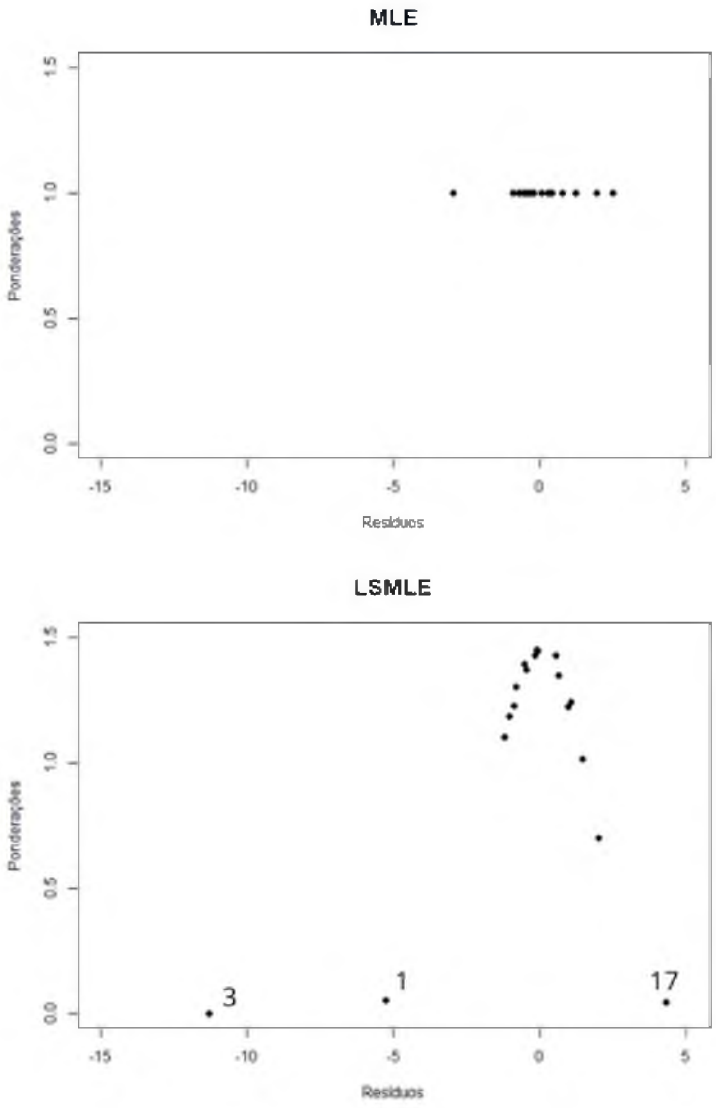
A Figura 7 exibe os gráficos de probabilidade normal dos resíduos com envelopes simulados sob os estimadores MLE e LSMLE. Os painéis da segunda linha apresentam ampliações dos painéis da primeira linha para fins de visualização. E a Figura 8 apresenta os gráficos dos resíduos contra as ponderações considerando o Modelo 2.

Figura 7: Gráficos de probabilidade normal dos resíduos com envelopes simulados sob MLE e LSMLE sob o Modelo 2.



Ao incluir a variável gênero no modelo, o comportamento dos estimadores segue o mesmo padrão observado anteriormente. O método clássico (MLE) continua ajustando os dados de forma a minimizar os resíduos, o que pode levar a um ajuste aparente melhor, mas que mascara o efeito de pontos discrepantes. Em contrapartida, o estimador robusto (LSMLE) expõe esses pontos e os trata com menor peso, permitindo que o modelo capture com mais fidelidade a estrutura predominante dos dados. Isso é particularmente relevante

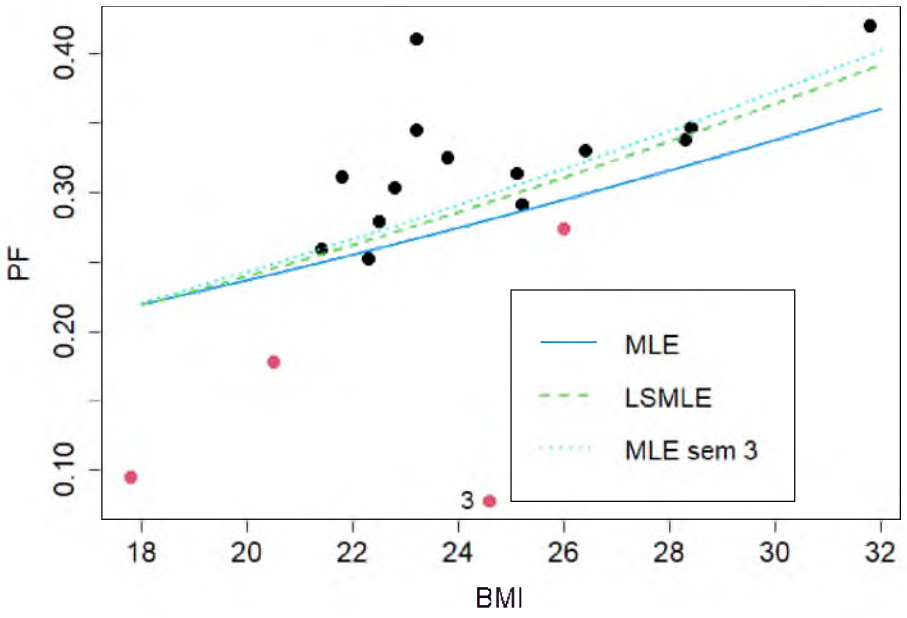
Figura 8: Gráficos dos resíduos contra as ponderações considerando o Modelo 2.



quando há variáveis categóricas, como o gênero, pois a presença de *outliers* pode distorcer a estimativa dos coeficientes. Dessa forma, o método robusto proporciona inferências mais estáveis e confiáveis sobre a relação entre as variáveis, preservando o foco na massa principal dos dados.

Para complementar a análise dos modelos de regressão beta ajustados, a Figura 9 apresenta gráfico de dispersão da porcentagem de gordura corporal (PF) em relação ao índice de massa corporal (BMI). Nesse gráfico, incluímos as curvas ajustadas via regressão beta utilizando os métodos MLE e LSMLE, tanto com os dados completos quanto com a exclusão da observação 3 no MLE. A motivação para essa abordagem vem do fato de que o LSMLE identificou a observação 3 como um ponto discrepante. Dessa forma, ao remover essa observação e refazer o ajuste pelo MLE, conseguimos avaliar o impacto dessa exclusão na curva ajustada, comparando-a com o ajuste originalmente obtido pelo LSMLE.

Figura 9: Gráfico de dispersão de PF contra BMI com as curvas ajustadas via regressão beta baseadas no MLE e LSMLE considerando o Modelo 2.



Ao examinar as curvas ajustadas, percebemos que a exclusão da observação 3 no MLE resulta em um comportamento da curva mais semelhante ao do LSMLE. Esse resultado reforça a efetividade do LSMLE em reduzir a influência de valores extremos sem precisar removê-los diretamente. Além disso, evidencia o impacto que pontos discrepantes podem ter no ajuste dos modelos tradicionais, destacando como os métodos robustos conseguem preservar melhor a tendência predominante nos dados sem serem excessivamente influenciados por valores atípicos.

5 Considerações Finais

Neste trabalho, a metodologia de regressão beta sob a perspectiva da inferência tradicional e robusta foi discutida, destacando-se a importância de métodos alternativos de estimação para reduzir o impacto de observações discrepantes. Inicialmente, revisamos a modelagem por meio de regressão beta tradicional baseada no método de máxima verossimilhança. Em seguida, apresentamos procedimentos inferenciais robustos, como o LSMLE e o LMDPDE, que, quando necessário, atribuem menor peso a valores atípicos, tornando as estimativas obtidas mais confiáveis.

Os resultados empíricos obtidos neste trabalho ilustraram que o MLE, embora amplamente utilizado, é sensível à presença de observações atípicas, o que pode comprometer a qualidade do ajuste do modelo e levar a inferências equivocadas. Em contraste, os métodos robustos demonstraram maior capacidade de identificar e mitigar o impacto de observações atípicas, proporcionando ajustes mais precisos para a massa dos dados. Essa característica é particularmente relevante em aplicações práticas, onde a presença de outliers é comum e pode distorcer as conclusões obtidas a partir de análises estatísticas.

A aplicação dos métodos robustos a um conjunto de dados reais permitiu ilustrar as vantagens das abordagens robustas em termos de identificação de outliers e avaliação da qualidade do ajuste. Os gráficos de probabilidade normal dos resíduos com envelopes simulados foram ferramentas úteis para avaliar a adequação dos modelos ajustados e identificar possíveis discrepâncias nos dados. Enquanto, o MLE tendeu a mascarar a presença de outliers e aumentar a incerteza sob o estimador, os métodos robustos destacaram essas observações, garantindo que a inferência fosse baseada principalmente nos dados mais representativos.

Em síntese, este trabalho reforça a importância da robustez estatística em modelos de regressão beta, especialmente em cenários onde há presença de outliers. Os métodos robustos apresentados oferecem uma alternativa confiável ao MLE, garantindo inferências mais robustas. Futuras pesquisas podem explorar a aplicação desses métodos em outros contextos, bem como a extensão dessas abordagens para modelos mais complexos.

Referências

CASELLA, G.; BERGER, R. *Statistical inference*. [S.l.]: CRC Press, 2024.

FERRARI, S. L. P.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004.

GHOSH, A. Robust inference under the beta regression model with application to health care studies. *Statistical methods in medical research*, SAGE Publications Sage UK: London, England, v. 28, n. 3, p. 871–888, 2019.

HUBER, P. J. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, JSTOR, v. 35, p. 73–101, 1964.

MALUF, Y. S.; FERRARI, S. L.; QUEIROZ, F. F. Robust beta regression through the logit transformation. *Metrika*, Springer, p. 1–21, 2024.

MARONNA, R. A. et al. *Robust statistics: theory and methods (with R)*. [S.l.]: John Wiley & Sons, 2019.

OSPINA, P. *Regressão beta*. Tese (Doutorado) — Instituto de Matemática e Estatística, Universidade de São Paulo, 2007.

RIBEIRO, T. K.; FERRARI, S. L. Robust estimation in beta regression via maximum l q-likelihood. *Statistical Papers*, Springer, v. 64, n. 1, p. 321–353, 2023.

SMITHSON, M.; VERKUILEN, J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, American Psychological Association, v. 11, n. 1, p. 54–71, 2006.

WELLS, J. C. K. Sexual dimorphism of body composition. *Best Pract Res Clin Endocrinol Metab*, Netherlands, v. 21, n. 3, p. 415–430, set. 2007.