UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA
MESTRADO ACADÊMICO EM MATEMÁTICA

AMANDA ALECSANDRA MOTA ROQUE RODRIGUES

MODELO DE REGRESSÃO BETA PARA DIFERENÇA DE PROPORÇÕES: TEORIA E APLICAÇÕES

MANAUS – AMAZONAS
2022

AMANDA ALECSANDRA MOTA ROQUE RODRIGUES

MODELO DE REGRESSÃO BETA PARA DIFERENÇA DE PROPORÇÕES: TEORIA E APLICAÇÕES

Dissertação apresentada ao Curso de Mestrado Acadêmico em Matemática do Programa de Pós-Graduação em Matemática do Instituto de Ciências Exatas da Universidade Federal do Amazonas, como requisito parcial à obtenção do título de mestre em Matemática. Área de Concentração: Estatística.

Orientador: Prof. Dr. Jhonnata Bezerra de Carvalho

Coorientador: Prof. Dr. Jeremias da Silva Leão

MANAUS – AMAZONAS

2022

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

R696m

Rodrigues, Amanda Alecsandra Mota Roque
Modelo de regressão beta para diferença de proporções: teoria e aplicações / Amanda Alecsandra Mota Roque Rodrigues . 2022
69 f.: il. color; 31 cm.

Orientador: Jhonnata Bezerra de Carvalho
Coorientador: Jeremias da Silva Leão
Dissertação (Mestrado em Matemática) - Universidade Federal do Amazonas.

1. Máxima verossimilhança. 2. Análise de diagnóstico. 3. Taxas e proporções. 4. Modelo de regressão. I. Carvalho, Jhonnata Bezerra de. II. Universidade Federal do Amazonas III. Título

AMANDA ALECSANDRA MOTA ROQUE RODRIGUES

MODELO DE REGRESSÃO BETA PARA DIFERENÇA DE PROPORÇÕES: TEORIA E APLICAÇÕES

Dissertação apresentada ao Curso de Mestrado Acadêmico em Matemática do Programa de Pós-Graduação em Matemática do Instituto de Ciências Exatas da Universidade Federal do Amazonas, como requisito parcial à obtenção do título de mestre em Matemática. Área de Concentração: Estatística.

Aprovada em: 20 de junho de 2022

BANCA EXAMINADORA

gov.br Documento assinado digitalmente
JHONNATA BEZERRA DE CARVALHO
Data: 17/07/2022 10:19:06-0300
Verifique em https://verificador.iti.br

Prof. Dr. Jhonnata Bezerra de Carvalho (Orientador)
Universidade Federal do Amazonas - UFAM

gov.br Documento assinado digitalmente
LAIS HELEN LOOSE
Data: 15/07/2022 16:55:11-0300
Verifique em https://verificador.iti.br

Profª. Drª Laís Helen Loose
Universidade Federal de Santa Maria - UFSM

gov.br Documento assinado digitalmente
MOIZES DA SILVA MELO
Data: 16/07/2022 14:55:32-0300
Verifique em https://verificador.iti.br

Prof. Dr. Moizés da Silva Melo
Universidade Federal do Rio Grande - FURG

Este trabalho é dedicado aos meus pais, irmãos e meu noivo por me darem todo o suporte para realizar os meus estudos e sempre me fazerem rir do meu desespero.

AGRADECIMENTOS

Quero começar com a frase clássica, mas sempre verdadeira:
Gostaria de agradecer primeiramente a Deus, por ter sido minha rocha, nos dias bons e ruins sempre esteve comigo.
Vou estender minha gratidão a meus pais, Carmen Regina e Alessessandre Roque, por me ajudarem na mudança de rotina, me darem suporte emocional durante a pandemia e, principalmente, por não me deixarem desistir e desanimar durante todo o curso.
Aos meus irmãos André e Cássio Roque por me cederem sempre que preciso seus computadores enquanto eu não tinha uma máquina para começar os meus estudos e por me fazerem rir no meio de toda a rotina bagunçada em que eu estava.
A meu futuro marido Maurício Rodrigues por ser paciente, por me incentivar e sempre dizer que eu sou inteligente e consigo rs, obrigado por me apoiar desde o primeiro dia, literalmente foi você que pegou na minha mão e me levou até a UFAM para fazer minha inscrição, obrigado por me fazer companhia nos bons e maus dias e enxugar minhas lagrimas de desespero, obrigada por tudo meu amor!
Meus queridos futuros sogros Ana Cristina e Bianey Rodrigues por me ajudarem sempre, me apoiarem nos meus sonhos e me auxiliarem em cada realização, desde a pequena até a grande, sou tão grata a vocês que não sei nem como agradecer.
Aos professores Jhonnata Bezerra e Jeremias Leão por terem acreditado no meu potencial, por me ensinarem coisas novas todos os dias, por terem paciência comigo e serem sempre compreensivos quanto as minhas dificuldades.
Ao professor Jõao Raimundo Ferreira que me incentivou muito a fazer a prova do mestrado, para alguns pode ser pouca coisa, mas o livro que me deu no início do curso fez muita diferença para mim, o fato de ter visto primeiro que eu o resultado me mostrou o quanto se importava com essa realização, obrigada por pergunta vez ou outra como eu estava e dizer que não era pra eu desistir, me chamou de mestra desde a minha inscrição na prova.
Cada pequena coisa dessas só me mostra que nós temos professores que de fato se importam com a educação, me mostra que além dos meus exemplos na família que trabalham na função de professor, eu também tenho alguns outros que se tornaram mais que meus professores, meus amigos.

Vocês todos me ensinaram algo e são inspiração para mim.
Para vocês todos meu muito obrigada!

"A cada passo precisei lutar, trabalhar duro, evoluir e sair na frente"

(Supergirl)

RESUMO

Neste trabalho é realizado um estudo referente à regressão beta, com a finalidade de modelar taxas ou proporções que variam no intervalo $(-1,1)$. A distribuição de probabilidade estudada é um caso particular da distribuição beta truncada, que assume valores no intervalo $(a,1)$, $(a<1)$, ou ainda, da fórmula geral da distribuição beta no intervalo (a,b), no qual $-\infty<a<b<\infty$. A distribuição beta no intervalo $(-1,1)$ foi denominada, no presente trabalho, como distribuição beta modular, na qual foram feitas reparametrizações dos parâmetros em função da média e precisão, e, com isso, foram atribuídas estruturas de regressão para esses novos parâmetros. Além disso, foram propostas novas funções de ligação e a estimação dos parâmetros foi feita através do método de máxima verossimilhança. Um estudo de simulação foi realizado para verificar os desempenhos dos estimadores de máxima verossimilhança considerando três diferentes tipos de funções de ligação para a média. Ademais, o modelo de regressão beta modular foi aplicado em dois conjuntos de dados reais referentes à eleição presidencial norte-americana no ano de 2016 e à diferença entre os percentuais de votação do ex-presidente Lula nas eleições de 2002 e 2006.

Palavras-chave: Máxima verossimilhança. Análise de diagnóstico. Taxas e proporções. Modelo de regressão.

ABSTRACT

In this work is related a study about a beta regression, in order to model rates or proportions that vary in the interval $(-1,1)$.The probability distribution studied is a particular case of the truncated beta distribution, which assumes values in the interval $(a,1)$, $(a<1)$, or even the general formula of the beta distribuition in the interval (a,b), where $-\infty<a<b<\infty$. The beta distribution in the interval $(-1,1)$ was called, in the present work, as modular beta distribution, in which parameterizations were made as a function of the mean and precision, and, therefore, regression structures were assigned to these new parameters. In addition, new link functions were proposed and the parameter estimation was performed using the maximum likelihood method. A simulation study was carried out to verify the performance of the maximum likelihood estimators considering three different types of link functions for the mean. Furthermore, the modular beta regression model was applied to two real data sets referring to the US presidential election in 2016 and the difference between the voting percentages of former president Lula in the 2002 and 2006 elections.

Keywords: Maximum likelihood. Diagnostic analysis. Negative rates and proportions. Regression model.

LISTA DE FIGURAS

Figura 1 – As diferentes formas da função de densidade da distribuição beta. 16
Figura 2 – Gráfico de barras para a variável resposta por estado. A cor azul (V) indica que o Donald Trump venceu e a cor rosa (P) indica que o Donald Trump perdeu. 35
Figura 3 – Histograma para a variável resposta sobreposto pela densidade da distribuição BM (em vermelho). 36
Figura 4 – Gráficos de dispersão para as variáveis no estudo. 36
Figura 5 – Análise de diagnóstico para o modelo com função de ligação probito. 40
Figura 6 – Análise de diagnóstico para o modelo com função de ligação logito. 41
Figura 7 – Análise de diagnóstico para o modelo com função de ligação complemento log-log. 42
Figura 8 – Histograma para a variável resposta sobreposto pela densidade estimada da distribuição BM. 45
Figura 9 – Gráfico de dispersão para as variáveis em estudo. 45
Figura 10 – Análise de diagnóstico para o modelo com função de ligação probito. 51
Figura 11 – Análise de diagnóstico para o modelo com função de ligação logito. 52
Figura 12 – Análise de diagnóstico para o modelo com função de ligação complemento log-log. 53
Figura 13 – Análise de diagnóstico para o modelo com função de ligação logito para o modelo de regressão BM com precisão fixa. 54

LISTA DE TABELAS

Tabela 1 – Funções inversas. 22

Tabela 2 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 1: $g_1(\boldsymbol{\mu}^*) = \log(\frac{\boldsymbol{\mu}^*}{1-\boldsymbol{\mu}^*})$, $n = 50, 100, 500$ e $\phi = 50$. 30

Tabela 3 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 2 : $g_1(\boldsymbol{\mu}^*) = \Phi^{-1}(\boldsymbol{\mu}^*)$, $n = 50, 100, 500$ e $\phi = 50$. 31

Tabela 4 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 3 : $g_1(\boldsymbol{\mu}^*) = \log(-\log(1-\boldsymbol{\mu}^*))$, $n = 50, 100, 500$ e $\phi = 50$. . . . 31

Tabela 5 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 4: $g_1(\boldsymbol{\mu}^*) = \log(\frac{\boldsymbol{\mu}^*}{1-\boldsymbol{\mu}^*})$, $g_2(\phi_i) = \log(\phi_i)$ e $n = 50, 100, 500$. 32

Tabela 6 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 5: $g_1(\boldsymbol{\mu}^*) = \Phi^{-1}(\boldsymbol{\mu}^*)$, $g_2(\phi_i) = \log(\phi_i)$ e $n = 50, 100, 500$. 33

Tabela 7 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 6: $g_1(\boldsymbol{\mu}^*) = \log(-\log(1-\boldsymbol{\mu}^*))$, $g_2(\phi_i) = \log(\phi_i)$ e $n = 50, 100, 500$. 33

Tabela 8 – Estimativas dos parâmetros para o modelo de regressão BM com dispersão variável para os dados sobre a eleição estadunidenses em 2016, ... 37

Tabela 9 – Estimativas dos parâmetros para o modelo de regressão BM com dispersão variável para os dados sobre a eleição estadunidenses em 2016, após a seleção de variáveis. 38

Tabela 10 – Ajuste dos modelos para a avaliação da diferença das taxas de votação em 2002 e 2006. 47

Tabela 11 – Ajuste dos modelos, após a seleção de variáveis, para a avaliação da diferença das taxas de votação em 2002 e 2006. 48

Tabela 12 – Ajuste do modelo de regressão BM com função de ligação logito e precisão fixa para a avaliação da diferença das taxas de votação em 2002 e 2006. . . 50

Tabela 13 – Ajuste do modelo de regressão beta com função de ligação logito e precisão fixa para a avaliação da diferença das taxas de votação em 2002 e 2006. . . 50

Tabela 14 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 4: $g_1(\boldsymbol{\mu}^*) = \log(\frac{\boldsymbol{\mu}^*}{1-\boldsymbol{\mu}^*})$ e $n = 50, 100, 500$. 64

Tabela 15 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 5 : $g_1(\boldsymbol{\mu}^*) = \Phi^{-1}(\boldsymbol{\mu}^*)$ e $n = 50, 100, 500$. 64

Tabela 16 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 6 : $g_1(\boldsymbol{\mu}^*) = \log(-\log(1-\boldsymbol{\mu}^*))$ e $n = 50, 100, 500$. 65
Tabela 17 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 7 : $g_1(\boldsymbol{\mu}^*) = \log(\frac{\boldsymbol{\mu}^*}{1-\boldsymbol{\mu}^*})$ e $n = 50, 100, 500$. 65
Tabela 18 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 8 : $g_1(\boldsymbol{\mu}^*) = \Phi^{-1}(\boldsymbol{\mu}^*)$ e $n = 50, 100, 500$. 66
Tabela 19 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 9 : $g_1(\boldsymbol{\mu}^*) = \log(-\log(1-\boldsymbol{\mu}^*))$ e $n = 50, 100, 500$. 66
Tabela 20 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 10: $g_1(\boldsymbol{\mu}^*) = \log(\frac{\boldsymbol{\mu}^*}{1-\boldsymbol{\mu}^*})$ e $n = 50, 100, 500$. 66
Tabela 21 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 11 : $g_1(\boldsymbol{\mu}^*) = \Phi^{-1}(\boldsymbol{\mu}^*)$ e $n = 50, 100, 500$. 67
Tabela 22 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 12 : $g_1(\boldsymbol{\mu}^*) = \log(-\log(1-\boldsymbol{\mu}^*))$ e $n = 50, 100, 500$. 67
Tabela 23 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 4: $g_2(\boldsymbol{\mu}^*) = \log(\frac{\boldsymbol{\mu}^*}{1-\boldsymbol{\mu}^*})$, $g(\phi_i) = \log(\phi_i)$ e $n = 50, 100, 500$. 68
Tabela 24 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 5: $g_2(\boldsymbol{\mu}^*) = \Phi^{-1}(\boldsymbol{\mu}^*)$, $g(\phi_i) = \log(\phi_i)$ e $n = 50, 100, 500$. 68
Tabela 25 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 6 : $g_2(\boldsymbol{\mu}^*) = \log(-\log(1-\boldsymbol{\mu}^*))$, $g(\phi_i) = \log(\phi_i)$ e $n = 50, 100, 500$. 69

SUMÁRIO

1	**INTRODUÇÃO**	13
2	**DISTRIBUIÇÃO BETA**	15
2.1	DEFINIÇÕES	15
2.2	DISTRIBUIÇÃO DA BETA MODULAR	16
2.2.1	**Reparametrização**	17
2.3	INFERÊNCIA SOBRE OS PARÂMETROS	18
3	**MODELOS DE REGRESSÃO**	20
3.1	MODELO DE REGRESSÃO BETA MODULAR	21
3.1.1	**Inferência para os parâmetros do modelo**	22
4	**ANÁLISE DE DIAGNÓSTICO**	26
5	**SIMULAÇÃO**	29
6	**APLICAÇÃO EM DADOS REAIS**	34
6.1	ELEIÇÕES ESTADUNIDENSES 2016	34
6.2	DIFERENÇA DAS TAXAS DE VOTAÇÃO DO CANDIDATO LULA DE 2002 E 2006	43
7	**CONSIDERAÇÕES FINAIS**	55
	REFERÊNCIAS	57
	APÊNDICE A – Cálculos	60
	APÊNDICE B – Simulação	64

1 INTRODUÇÃO

Vários estudos possuem o interesse em variáveis cujos os seus valores estão dispostos no intervalo unitário, ou seja, $(0, 1)$. Esse tipo de situação ocorre frequentemente porque se torna mais fácil ao pesquisador expressá-los como taxas ou proporções. Para modelar variáveis desse tipo, é necessário que o pesquisador conheça algumas distribuições de probabilidade presentes na literatura, para escolher a que melhor se adequa ao estudo. A distribuição beta é uma das mais utilizadas devido a facilidade de aplicação que tem em relação a outros modelos, isso se dá devido a sua flexibilidade em relação ao seu par de parâmetros.

Sabe-se que existem situações em que o objetivo é estudar o comportamento de uma certa variável em relação a outras, para compreender esse tipo de relação, pode-se utilizar modelos de regressão. Um dos mais utilizados e mais fáceis de manusear é o modelo de regressão linear normal (MRLN). Porém, cabe lembrar, que esse modelo, por mais que seja muito utilizado, não é aconselhado que se use em situações cuja a variável resposta está restrita ao intervalo unitário (SILVA, 2020).

De acordo com Turkman e Silva (2000) o modelo linear normal foi o dominante na modelação estatística por muito tempo após a sua definição, embora vários modelos não lineares ou não normais tenham sido desenvolvidos para resolver as mais diversas situações, ainda existiam aqueles aos quais não se adequavam ao modelo tradicional e aos derivados. São exemplos de aplicações de modelos consequentes de variações do tradicional, o modelo log-log complementar usado por Kannebley e Prince (2015), Ortiz, Uribe e García (2007) utilizam o modelo probit bivariado, ou ainda o modelo linear que foi a escolhido para o estudo de Garcia-Marques, Quelhas e Gomes (1997), entre outros.

Ferrari e Cribari-Neto (2004) afirmam que cada pesquisador pode transformar a variável resposta de tal forma que assumam valores reais e, dessa forma, se possa utilizar o MRLN. Porém essa abordagem traz consigo problemas que podem vir a interferir na interpretação dos parâmetros em termos da variável resposta original.

Nelder e Wedderburn (1972) definem os modelos lineares generalizados (MLG) como sendo a união dos modelos existentes, para que assim possa abranger as distribuições, desenvolvendo melhor tanto do ponto de vista teórico como o conceitual. Paula (2004) diz que essa proposta inovadora tem o objetivo básico da utilização dos MLG é dar aos pesquisadores uma gama de opções para as distribuições da variável resposta, fazendo com que esta faça parte da família exponencial de distribuições canônicas. E, por apresentar essas características passou

a ser amplamente utilizada.

Com base nisso, nota-se que existe uma necessidade de se utilizar modelos que comporte a característica numérica da variável resposta, como exemplo: uma variável resposta que assume valores no intervalo $(0, 1)$, e, assim, pode-se utilizar o modelo de regressão beta para entender o relacionamento entre as variáveis do estudo.

Neste trabalho, procura-se realizar um estudo de uma derivação nova da distribuição beta, que é chamada de distribuição beta modular, que busca modelar a diferença entre taxas ou proporções que variam entre -1 e 1. Como ilustração do modelo proposto, são utilizados dados sobre as eleições estadunidenses do ano de 2016 e os impactos que a criação do programa Bolsa Família nos resultados das eleições presidenciais em 2006 no Brasil.

No Capítulo 2, estão dispostos conceitos sobre a distribuição beta, na sua forma convencional, como a função de densidade de probabilidade, suas respectivas variância e valor esperado, e sua função de distribuição acumulada. Também é definida a distribuição beta modular, que é proposta e desenvolvida no decorrer deste trabalho, assim como uma reparametrização da distribuição beta em termos da média, representada pela letra grega μ. O Capítulo 3 , traz a definição do modelo de regressão beta modular e algumas informações sobre os parâmetros do modelo. Em seguida, no Capítulo 4 são descritas as ferramentas para a análise de diagnóstico do modelo, como o coeficiente de determinação, resíduo quantílico, critérios de qualidade de ajuste e o gráfico de probabilidade meio-normal. Além disso, no Capítulo 5 estão as simulações de Monte Carlo para avaliar os estimadores dos parâmetros e, em seguida, no Capítulo 6 são apresentadas as aplicações em dados reais. E por fim, no Capítulo 7, são descritos ao leitor as considerações finais do presente trabalho e algumas propostas de trabalhos futuros.

2 DISTRIBUIÇÃO BETA

Neste capítulo são abordados as definições e principais características da distribuição beta e da distribuição beta modular que foi desenvolvida no decorrer do presente trabalho e a transformação que à gerou.

2.1 DEFINIÇÕES

A distribuição beta é uma das mais empregadas para o estudo de variáveis cujos valores estejam entre o intervalo $(0,1)$, ou seja, que pertençam ao intervalo unitário. Suas aplicações vão além da estatística, essa distribuição, por ser muito flexível e simples de utilizar para modelar proporções, é usada nas mais diversas áreas para estudos pontuais, como na agrometeorologia, analisando dados referentes ao comportamento do vento em São Paulo (JÚNIOR *et al.*, 1995), estudos em estatística e experimentação agronômica (PESCIM, 2009), na modelagem da proporção de petróleo bruto convertido em gasolina após a destilação e fracionamento (FERRARI; CRIBARI-NETO, 2004).

Casella e Berger (2011) afirmam que a distribuição beta é uma das poucas distribuições que podem ser chamadas de comum, isso porque ela faz parte do grupo de distribuições que são utilizadas com mais frequência, além de pertencer ao conjunto das distribuições contínuas. É importante dizer que a distribuição beta está relacionada diretamente a função gama e, assim, pode-se usar as suas propriedades como ferramenta de auxílio.

Seja X uma variável aleatória contínua com distribuição beta com parâmetros $\alpha > 0$ e $\beta > 0$, denotemos por $X \sim beta(\alpha,\beta)$. Então a sua função de densidade de probabilidade (fdp) é dada por

$$f(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1, \tag{2.1}$$

em que $B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ é chamada de função beta. Como já se sabe, as distribuições gama e beta estão relacionadas, isso ocorre por meio da seguinte identidade

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Já a função de distribuição acumulada para beta é dada pela seguinte equação

$$F(x;\alpha,\beta) = \frac{B(x;\alpha,\beta)}{B(\alpha,\beta)} = \frac{\int_0^x t^{\alpha-1}(1-t)^{\beta-1}dt}{B(\alpha,\beta)} = I_x(\alpha,\beta), \tag{2.2}$$

na qual $\mathrm{F}(x;\alpha,\beta) = \mathrm{I}_x(\alpha,\beta)$ é denotada de função beta regularizada. O valor esperado e a variância são dados, respectivamente, por (CASELLA; BERGER, 2011):

$$E(X) = \frac{\alpha}{\alpha+\beta}, \tag{2.3}$$

e

$$Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{2.4}$$

Tem-se ainda que α e β são parâmetros de forma, o que torna a densidade flexível. Dito isso, de acordo com Casella e Berger (2011), os gráficos da função de densidade podem ser classificadas de seis formas diferentes, podendo ser estritamente crescente ($\alpha > 1$, $\beta = 1$), estritamente decrescente ($\alpha = 1$, $\beta > 1$), pode ter um gráfico em formato de U ($\alpha < 1$, $\beta < 1$), unimodal ($\alpha > 1$, $\beta > 1$), simétrica ($\alpha = \beta$) ou como uma distribuição uniforme no intervalo $(0,1)$, isso quando ocorre $\alpha = \beta = 1$. A seguir, na Figura 1 estão dispostos os gráficos referentes a cada situação descrita anteriormente.

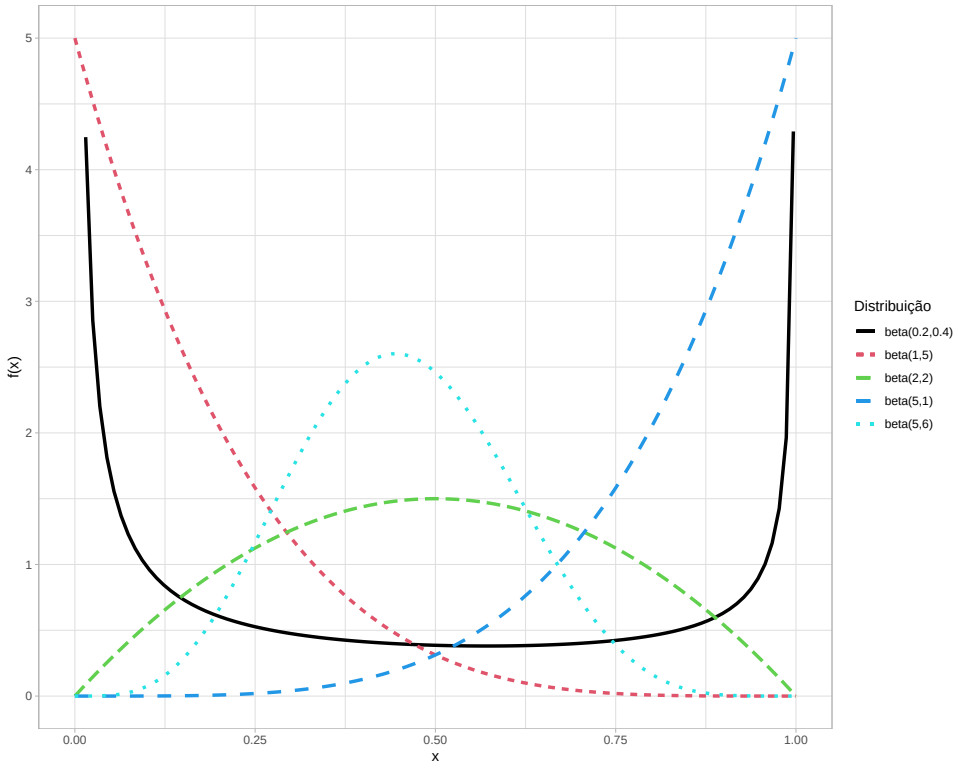


Figura 1 – As diferentes formas da função de densidade da distribuição beta.

2.2 DISTRIBUIÇÃO DA BETA MODULAR

De acordo com Gray e Alava (2018), as variáveis resposta contínuas limitadas em suas extremidades surgem em muitas áreas, já aquelas que medem proporções e taxas são muito

comuns em estudos empíricos. Com base na afirmativa dos autores, as aplicações de variáveis limitadas em intervalos alternativos se transformam linearmente em variáveis dependentes para o intervalo $(0,1)$.

Com base nos trabalhos de Gray e Alava (2018), Pereira, Botter e Sandoval (2012) e Johnson, Kotz e Balakrishnan (1995), é licito dizer que a beta modular é um caso particular da beta truncada, uma distribuição que está no intervalo $(a,1)$, ou ainda, um caso particular da fórmula geral da beta no intervalo (a,b). Sabendo disso, pode-se modelar variáveis que estão no intervalo $(-1,1)$, propondo a transformação $Y = 2X - 1$, no qual $X \sim beta(\alpha,\beta)$. O desenvolvimento para se chegar a função de distribuição de probabilidade da variável Y está localizada no Apêndice A, aqui nomeada de distribuição beta modular (BM), sinalizado por BM(α,β), com parâmetros α e β, denotemos por $Y \sim BM(\alpha,\beta)$, no qual a sua densidade está a seguir

$$f_y(y) = \frac{1}{2^{\alpha+\beta-1}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} (1+y)^{\alpha-1} (1-y)^{\beta-1}, \quad -1 < y < 1. \tag{2.5}$$

Como a transformação de Y é conhecida, podemos encontrar o valor esperado e a variância da distribuição BM. O valor esperado é dado por

$$E(Y) = \frac{\alpha-\beta}{\alpha+\beta}. \tag{2.6}$$

De maneira análoga, pode-se dizer que a variância da BM, obtida com base na equação (2.4), é dada por

$$Var(Y) = \frac{4\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}, \tag{2.7}$$

os cálculos para obtenção da densidade da distribuição BM, média e variância estão dispostos no Apêndice A.

2.2.1 Reparametrização

Para obter uma estrutura de regressão para a média da resposta juntamente com um parâmetro de precisão, é necessário reparametrizar a densidade (2.5) em termos dessas quantidades. Portanto, considere uma variável aleatória $Y \sim BM(\alpha,\beta)$. Agora, tomemos $\mu = (\alpha-\beta)/(\alpha+\beta)$ e $\phi = \alpha+\beta$, logo

$$\mu = \frac{\alpha-\beta}{\alpha+\beta} = \frac{2\alpha-\phi}{\phi} = \frac{2\alpha}{\phi} - 1.$$

E a partir deste ponto, têm-se como novos parâmetros

$$\alpha = \frac{\phi}{2}(\mu+1), \tag{2.8}$$

e

$$\beta = \frac{\phi}{2}(1-\mu). \tag{2.9}$$

Note que $-1 < \mu < 1$ e $\phi > 0$. Baseadas nas equações (2.8) e (2.9), pode-se obter a densidade da distribuição BM em função da média e do parâmetro de precisão, que é dada por

$$f_y(y) = \frac{1}{2^{\phi-1}} \frac{\Gamma(\phi)}{\Gamma\left(\frac{\phi}{2}(\mu+1)\right)\Gamma\left(\frac{\phi}{2}(1-\mu)\right)} (1+y)^{\frac{\phi}{2}(\mu+1)-1} (1-y)^{\frac{\phi}{2}(1-\mu)-1}. \tag{2.10}$$

A função de distribuição da BM pode ser escrita como uma derivação de função de distribuição da beta, dada na equação (2.2), ficando na forma

$$F(y;\mu,\phi) = I_{\left(\frac{y+1}{2}\right)}\left(\frac{\phi}{2}(\mu+1), \frac{\phi}{2}(1-\mu)\right).$$

Dessa forma, temos que a média e a variância são dadas respectivamente por

$$E(Y) = \mu \quad \text{e} \quad Var(Y) = \frac{1-\mu^2}{\phi^2(\phi+1)}.$$

2.3 INFERÊNCIA SOBRE OS PARÂMETROS

Existem diversos métodos de estimação para encontrar estimadores de parâmetros, dentre eles, o mais popular, o método da máxima verossimilhança (MV). Seja $\mathbf{X} = (X_1, \cdots, X_n)^\top$ uma amostra aleatória de uma variável aleatória X, com vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)^\top$, no qual $\mathbf{x} = (x_1, \cdots, x_n)^\top$ representa o vetor observado da variável aleatória. A função de verossimilhança é definida como:

$$L(\boldsymbol{\theta}|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta_1, \cdots, \theta_k),$$

no qual $\theta_1, \ldots, \theta_k$ são parâmetros desconhecidos. Casella e Berger (2011) definem um estimador de máxima verossimilhança (EMV) como: o ponto amostral $\mathbf{x}$ em que $L(\boldsymbol{\theta}|\mathbf{x})$ atinge seu máximo como função de $\boldsymbol{\theta}$, nesse ponto se encontra o EMV do vetor de parâmetros, representado por $\widehat{\boldsymbol{\theta}}(\mathbf{X})$. Se o domínio do estimador coincide com o domínio do parâmetro, ele é uma boa escolha de estimador pontual.

Se a função de verossimilhança for diferenciável em relação a $\boldsymbol{\theta}$, então o EMV pode ser encontrado através da solução do seguinte sistema de equações

$$\frac{\partial}{\partial \theta_j} L(\boldsymbol{\theta}|\mathbf{x}) = 0, j = 1, \ldots, k,$$

o resultado obtidos através da solução das derivadas parciais apresenta o candidato a EMV, sendo necessário avaliar a matriz hessiana para garantir que o ponto é de máximo.

Muitas vezes é mais fácil trabalhar com o logaritmo natural da função de log-verossimilhança, ou seja, $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}|\mathbf{x})$, que é conhecida na literatura como função de log-verossimilhança, este método retorna ao pesquisador uma função que será estritamente crescente, e obteremos o EMV através de $\ell(\boldsymbol{\theta})$. Como a função do logaritmo é monótona, os pontos que a maximizam e a função de log-verossimilhança coincidem.

Sejam $Y_1, \ldots, Y_n$ uma amostra aleatória de tamanho n de $Y \sim BM(\mu, \phi)$ e $\boldsymbol{\theta} = (\mu, \phi)^T$ os parâmetros de interesse para o caso. Tem-se que a função de log-verossimilhança para a fdp da distribuição BM é dada por

$$\ell(\boldsymbol{\theta}) = -n \log B\left(\frac{\phi}{2}(\mu_i+1), \frac{\phi}{2}(1-\mu_i)\right) - n(\phi-1)\log(2) + \left(\frac{\phi}{2}(\mu_i+1)-1\right)\sum_{i=1}^{n}\log(1+y_i) + \left(\frac{\phi}{2}(1-\mu_i)-1\right)\sum_{i=1}^{n}\log(1-y_i). \tag{2.11}$$

Para mais detalhes ver o Apêndice A.

O EMV de $\boldsymbol{\theta}$ é obtidos pela solução do sistema de equações gerado quando a primeira derivação parcial de (2.11), igualada a zero. Este sistema não possui solução analítica fechada. Portanto, é necessário utilizar métodos numéricos para obter uma solução aproximada.

3 MODELOS DE REGRESSÃO

Os modelos de regressão são utilizados quando se quer entender o comportamento de uma variável com relação à variação de outras variáveis. De acordo com Paula (2004), os modelos lineares normais foram usados por muito tempo para descrever alguns fenômenos aleatórios, mesmo quando a suposição de normalidade da variável resposta não era razoável. Nesses casos, eram e ainda são sugeridas aplicações de transformações que venham alcançar a normalidade. Além disso, o uso do modelo de regressão linear para a modelagem de varáveis limitadas nas extremidades apresenta alguns problemas, pois as previsões do modelo podem estar fora do campo de variação da variável resposta. Uma alternativa, é transformar a variável resposta em valores na reta real e em seguida, usar modelos de regressão padrão na variável dependente transformada como solução para este problema (GRAY; ALAVA, 2018).

Rodrigues (2012) define um modelo de regressão linear simples (MRLS) como uma relação de características lineares entre a variável dependente Y e uma variável independente X, enquanto que o modelo de regressão linear múltiplo (MRLM) é a relação linear entre a variável Y dependente e p variáveis independentes, isto é, $(X_1, \cdots, X_p)$. Com base nisso, pode-se dizer que um MRLS, nada mais é do que um caso de MRLM, com $p = 1$. Dessa forma, temos que o MRLM é pode ser representado como:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i,$$

para $i = 1, 2, \ldots, n$ e $j = 1, 2 \ldots, p$, com Y_i sendo a i-ésima observação da variável resposta, β_0 e β_j's são os parâmetros dos modelos, x_{ij} a i-ésima observação da j-ésima variável independente e ε_i corresponde aos erros do modelo e possui distribuição normal com média 0 e variância σ^2.

A escolha de um modelo deve ser baseada de acordo com os objetivos do pesquisador em relação ao seu experimento, nas características de seus dados, em experimentos anteriores relacionados a pesquisa de interesse e na análise descritiva. A partir da escolha, se necessário, deve-se ajustar o modelo até encontrar uma ajuste que descreva bem o comportamento entre as variáveis preditoras e a variável resposta (MONTGOMERY; PECK; VINING, 2021).

Na próxima seção está apresentado o modelo de regressão com a utilização da distribuição BM e suas respectivas funções de ligação.

3.1 MODELO DE REGRESSÃO BETA MODULAR

Sejam $Y_1,\ldots,Y_n$ uma amostra da variável $Y \sim BM(\mu,\phi)$. O modelo é obtido assumindo que a média e a precisão satisfaçam as seguintes relações funcionais

$$g_1\left(\frac{\mu_i+1}{2}\right) = \mathbf{x}_i^\top \boldsymbol{\beta} = \eta_{1i}, \tag{3.1}$$

e

$$g_2(\phi_i) = \mathbf{z}_i^\top \boldsymbol{\gamma} = \eta_{2i}, \tag{3.2}$$

com $i = 1,\ldots,n$, as quantidades $g_1(\cdot)$ e $g_2(\cdot)$ são funções de ligação estritamente monótonas e pelo menos duas vezes diferenciáveis, em que $g_1 : (-1,1) \to \mathbb{R}$ e $g_2 : \mathbb{R}^+ \to \mathbb{R}$, com $\mathbf{x}_i^\top = (x_{i0}, x_{i1}, \ldots, x_{ip})$ e $\mathbf{z}_i = (z_{i0}, z_{i1}, \ldots, z_{iq})^\top$ sendo os vetores de covariáveis relacionadas à i-ésima resposta, $\boldsymbol{\beta} = (\beta_0,\ldots,\beta_p)^\top$ e $\boldsymbol{\gamma} = (\gamma_0,\ldots,\gamma_q)^\top$ são os vetores de parâmetros desconhecidos. Ademais, η_{1i} e η_{2i} são os preditores lineares para ambas as funções e $p+q+2 < n$ (BOURGUIGNON; LEÃO; GALLARDO, 2020).

As funções descritas nas equações (3.1) e (3.2) tornam o modelo mais maleável, no sentido de comportar relações não lineares. Para facilitar a notação, considere $\mu_i{}^* = (\mu_i+1)/2$ e as funções de ligação que podem ser utilizadas no modelo de regressão BM são:

- Função Logito:

$$\log\left(\frac{\mu_i{}^*}{1-\mu_i{}^*}\right) = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

- Função Probito:

$$\Phi^{-1}(\mu_i{}^*) = \mathbf{x}_i^\top \boldsymbol{\beta},$$

  em que $\Phi(.)$ é a função de distribuição acumulada de uma variável aleatória de distribuição normal padrão.

- Complemento log-log:

$$\log\left(-\log\left(1-\mu_i{}^*\right)\right) = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

- Logaritmo:

$$\log(\phi_i) = \mathbf{z}_i^\top \boldsymbol{\gamma}$$

- Raiz:

$$\sqrt{\phi_i} = \mathbf{z}_i^\top \boldsymbol{\gamma}.$$

Pode-se também modelar diretamente a média $\mu_i \in (-1,1)$, através de funções trigonométricas, para caso a resposta apresente um comportamento periódico. As funções de ligação para esse caso seriam,

- Arco seno

$$arcsen(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

- Arco cosseno

$$arccos(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

- Arco tangente

$$arctg(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

A Tabela 1 possui as funções inversas para cada função de ligação mencionada.

Tabela 1 – Funções inversas.

Nome da função de ligação	Função inversa
Logito	$\mu = \frac{2e^{\mathbf{x}^\top\boldsymbol{\beta}}}{1+e^{\mathbf{x}^\top\boldsymbol{\beta}}} - 1$
Probito	$\mu = 2\Phi(\mathbf{x}^\top\boldsymbol{\beta}) - 1$
Complemento log-log	$\mu = 2(1 - e^{-e^{\mathbf{x}^\top\boldsymbol{\beta}}}) - 1$
Arco seno	$sen(\mathbf{x}^\top\boldsymbol{\beta})$
Arco cosseno	$cos(\mathbf{x}^\top\boldsymbol{\beta})$
Arco tangente	$tan(\mathbf{x}^\top\boldsymbol{\beta})$
Logaritmo	$\phi = e^{\mathbf{z}^\top\boldsymbol{\gamma}}$
Raiz	$\phi = (\mathbf{z}^\top\boldsymbol{\gamma})^2$

3.1.1 Inferência para os parâmetros do modelo

Sejam $Y_1,\ldots,Y_n$ uma amostra aleatória de um modelo de regressão BM e $\boldsymbol{\theta} = (\boldsymbol{\beta},\boldsymbol{\phi})^\top$ o vetor de parâmetros do modelo. Dada a densidade da BM (ver equação (2.10)), tem-se que a função de log-verossimilhança é dada por

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\mu_i,\phi_i),$$

na qual

$$\ell_i(\mu_i,\phi_i) = -(\phi_i-1)\log(2)+\log(\Gamma(\phi_i)) - \log\left(\Gamma\left(\frac{\phi}{2}(\mu_i+1)\right)\right) - \log\left(\Gamma\left(\frac{\phi}{2}(1-\mu_i)\right)\right) + \left(\frac{\phi}{2}(\mu_i+1)-1\right)\log(1+y_i) + \left(\frac{\phi}{2}(1-\mu_i)-1\right)\log(1-y_i). \quad (3.3)$$

A função escore é obtida pela diferenciação da equação (3.3) com respeito a cada um dos parâmetros desconhecidos, isto é $U(\boldsymbol{\theta}) = (U_\beta(\boldsymbol{\theta})^\top, U_\gamma(\boldsymbol{\theta})^\top)^\top$, em que

$$U_{\beta_t}(\boldsymbol{\theta}) = \frac{\partial}{\partial\beta_t}\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n}\frac{\partial}{\partial\mu_i}\ell_i(\mu_i,\phi_i)\frac{d\mu_i}{d\eta_{1i}}\frac{\partial\eta_{1i}}{\partial\beta_t}$$
$$U_{\gamma_j}(\boldsymbol{\theta}) = \frac{\partial}{\partial\gamma_t}\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n}\frac{\partial}{\partial\phi_i}\ell_i(\mu_i,\phi_i)\frac{d\phi_i}{d\eta_{2i}}\frac{\partial\eta_{2i}}{\partial\gamma_j}.$$

Para mais detalhes sobre essas expressões ver Souza (2011). Portanto, obtém-se

$$U_{\beta_t}(\boldsymbol{\theta}) = \sum_{t=1}^{n}\frac{\phi_i}{2}\left[-\psi\left(\frac{\phi_i}{2}(\mu_i+1)\right)+\psi\left(\frac{\phi_i}{2}(1-\mu_i)\right)+\log\left(\frac{1+y_i}{1-y_i}\right)\right]\frac{2x_{it}}{g_1'(\mu^*{}_i)}, \quad (3.4)$$

no qual $d\mu_i/d\eta_{1i} = (g_1'(\mu^*{}_i))^{-1}$, $g_1'(.)$ sendo a representação da primeira derivada de $g_1(.)$, $\partial\eta_{1i}/\partial\beta_t = x_{it}$, $\psi(\cdot)$ é a função digama, em que, $\psi(c) = d\log(\Gamma(c))/dc$ com $c>0$, para $t=0,\ldots,p$, sendo que para $t=0$ se tem $x_{i0}=1$.

Calculando as derivadas com relação ao vetor $\boldsymbol{\gamma}$, tem-se que

$$U_{\gamma_j}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\left[-\log(2)+\psi(\phi_i) - \frac{(\mu_i+1)}{2}\psi\left(\frac{\phi_i}{2}(\mu_i+1)\right) - \frac{(1-\mu_i)}{2}\psi\left(\frac{\phi_i}{2}(1-\mu_i)\right) + \frac{\mu_i+1}{2}\log(1+y_i)+\frac{(1-\mu_i)}{2}\log(1-y_i)\right]\frac{z_{ij}}{g_2'(\phi_i)}, \quad (3.5)$$

no qual $d\phi_i/d\eta_{2i} = 2(g_2'(\phi_i))^{-1}$, $g_2'(.)$ é a representação da primeira derivada de $g_2(.)$, $\partial\eta_{2i}/\partial\gamma_j = x_{ij}$, para $j=0,\ldots,q$, e quando $j=0$, se tem $z_{i0}=1$. A fim de simplificar a notação, definimos as seguintes quantidades

$$a_i = \frac{\phi_i}{2}\left[-\psi\left(\frac{\phi_i}{2}(\mu_i+1)\right)+\psi\left(\frac{\phi_i}{2}(1-\mu_i)\right)+\log\left(\frac{1+y_i}{1-y_i}\right)\right],$$
$$b_i = \log(2)+\psi(\phi_i) - \frac{(\mu_i+1)}{2}\psi\left(\frac{\phi_i}{2}(\mu_i+1)\right) - \frac{(1-\mu_i)}{2}\psi\left(\frac{\phi_i}{2}(1-\mu_i)\right) + \frac{\mu_i+1}{2}\log(1+y_i) + \frac{(1-\mu_i)}{2}\log(1-y_i).$$

Feito isso, pode-se escrever o vetor escore na forma matricial que é dado por

$$U_{\beta_t}(\boldsymbol{\theta}) = \mathbf{X}^\top\mathbf{M}\mathbf{a}$$
$$U_{\gamma_t}(\boldsymbol{\theta}) = \mathbf{Z}^\top\mathscr{M}\mathbf{b},$$

na qual $\mathbf{X}$ é uma matriz de dimensão $n \times (p+1)$ com $\mathbf{x}_i$ sendo os elementos da i-ésima linha com $i = 1, \cdots, n$ e $\mathbf{Z}$ uma matriz de dimensão $n \times (q+1)$ com $\mathbf{z}_j$ sendo os elementos da j-ésima linha com $j = 1, \cdots, n$, nos quais

$$\begin{aligned}
&\mathbf{b} = (b_1, \cdots, b_n)^\top, \\
&\mathbf{a} = (a_1, \cdots, a_n)^\top, \\
&\mathbf{M} = diag(2/g_1'(\mu^*{}_1), \cdots, 2/g_1'(\mu^*{}_n)), \\
&\mathscr{M} = diag\left(1/g_2'(\phi_1), \cdots, 1/g_2'(\phi_n)\right).
\end{aligned}$$

A matriz de informação de Fisher, que posteriormente ira auxiliar na construção dos intervalos de confiança e na realização dos testes de hipóteses para os parâmetros do modelo, é calculada por

$$\begin{aligned}
-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \beta_t \partial \beta_k}\right] &= -\sum_{i=1}^{n} E\left[\frac{\partial^2}{\partial \mu_i^2}\ell_i(\mu_i, \phi_i)\right]\left(\frac{d\mu_i}{d\eta_{1i}}\right)^2 x_{it} x_{ik} \\
&= \mathbf{X}^\top \mathbf{Q}\mathbf{M}^2\mathbf{X},
\end{aligned}$$

em que $k = 0, 1, \ldots, p$, $\mathbf{Q} = diag(q_1, \ldots, q_n)^\top$ e $\psi'(.)$ é chamada de função trigama, no qual $\psi'(c) = \frac{d}{dc}\psi(c)$ com $c > 0$. Tem-se também que

$$\begin{aligned}
-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \beta_t \partial \gamma_j}\right] &= -\sum_{i=1}^{n} E\left[\frac{\partial^2 \ell_i(\mu_i, \phi_i)}{\partial \mu_i \partial \phi_i}\right]\frac{d\phi_i}{d\eta_{2i}}\frac{d\mu_i}{d\eta_{1i}} z_{ij} x_{it} \\
&= \mathbf{X}^\top \mathbf{M}\mathscr{M}\mathbf{D},
\end{aligned}$$

no qual $\mathbf{D} = diag(d_1, \ldots, d_n)^\top$. Além disso,

$$\begin{aligned}
-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \gamma_i \partial \gamma_j}\right] &= -\sum_{i=1}^{n} E\left[\frac{\partial^2}{\partial \phi_i^2}\ell_i(\mu_i, \phi_i)\right].\left(\frac{d\phi_i}{d\eta_{2i}}\right)^2 z_{ij} z_{it} \\
&= \mathbf{Z}^\top \mathbf{S}\mathscr{M}^2\mathbf{Z},
\end{aligned}$$

em que $\mathbf{S} = diag(s_1, \ldots, s_n)^\top$.

Com isso pode-se dizer que a matriz de informação de Fisher é dada por

$$K(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{X}^\top \mathbf{Q}\mathbf{M}^2\mathbf{X} & \mathbf{X}^\top \mathbf{M}\mathscr{M}\mathbf{D} \\ \mathbf{X}^\top \mathbf{M}\mathscr{M}\mathbf{D} & \mathbf{Z}^\top \mathbf{S}\mathscr{M}^2\mathbf{Z} \end{pmatrix}$$

Note que os parâmetros betas e gamas não são ortogonais. Logo, sob as condições de regularidade, se tem

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{pmatrix} \sim N_{p+q+2}\left(\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix}, K(\boldsymbol{\theta})^{-1}\right),$$

em que N_{p+q+2} denota a distribuição normal multivariada e $\widehat{\boldsymbol{\beta}}$ e $\widehat{\boldsymbol{\gamma}}$ são os EMVs, respectivamente. E, como observado em Bourguignon, Leão e Gallardo (2020), $\mathbf{K}(\boldsymbol{\theta})^{-1}$ pode ser aproximado de $(-\mathbf{H})^{-1}$, em que $\mathbf{H}$ representa a matriz Hessiana.

Suponha que desejamos testar as hipótese $\mathrm{H}_0 : \theta_m = \theta_m^0$ versus $\mathrm{H}_1 : \theta_m \neq \theta_m^0$, no qual θ_m^0 é um valor especificado para o parâmetro desconhecido θ_m. Uma estatística de teste muito utilizada para testar essas hipóteses é a raiz quadrada da estatística de Wald, que é dada por

$$Z = \frac{\widehat{\theta}_m - \theta_m}{\sqrt{k^{mm}}},$$

em que k^{mm} é a representação do m-ésimo termo da diagonal principal da matriz $\mathbf{K}(\widehat{\boldsymbol{\theta}})^{-1}$. Sob H_0 e para $n \to \infty$, Z converge em distribuição para a normal padrão (PAWITAN, 2001). Além disso, essa estatística pode ser utilizada para construir intervalos de confiança $100(1-\alpha)\%$ assintóticos para os parâmetros do modelo, ou seja,

$$\left[\widehat{\theta}_m - z_{1-\alpha/2}\sqrt{k^{mm}}; \widehat{\theta}_m + z_{1-\alpha/2}\sqrt{k^{mm}}\right],$$

em que $z_{1-\alpha/2}$ é um valor tal que $\Phi(z_{1-\alpha/2}) = 1-\alpha/2$.

4 ANÁLISE DE DIAGNÓSTICO

Para utilizar um modelo de regressão, é desejável que as variáveis explicativas possuam algum tipo de relação com a variável resposta, para que seja possível definir quais as variáveis influenciam na variável de interesse e quais devem ser descartadas por não terem relação com ela. Após o ajuste do modelo, pode acontecer que o mesmo não seja o ideal para determinado conjunto de dados e, assim, gerar problemas ao pesquisador. Por isso a análise de diagnóstico é fundamental no ajuste de modelos de regressão, pois é feita uma avaliação do modelo ajustado para averiguar o nível de adequabilidade do modelo e se o ajuste vai de fato relacionar teoria e realidade.

Uma maneira bastante utilizada para verificar a qualidade do ajuste é criar um gráfico de dispersão entre $g_1(y_i)$ e $g_1(\widehat{\mu}_i)$ e verificar se existe uma relação linear entre essas duas quantidades. Uma medida que pode ser utilizar para quantificar essa relação, é o pseudo R^2 (denotado por R^2_p) que é o quadrado do coeficiente de correlação entre $g_1(y_i)$ e $g_1(\widehat{\mu}_i)$, na qual $0 \leq R^2_p \leq 1$ e quanto mais próximo de 1, melhor o ajuste (FERRARI; CRIBARI-NETO, 2004; PAULA, 2004).

Como explicado por Emiliano *et al.* (2009) a escolha do melhor modelo é importante na modelagem dos dados e o fato de ser comedido é uma das características que mais é levada em consideração. Para a seleção do modelo foram utilizados como ferramente os critérios de comparação o Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiano (BIC), para que um modelo mais comedido seja selecionado. Silva (2020) afirma que devido à forma como foram construídos, será considerado o melhor ajuste de modelo quando os valores de AIC e/ou BIC forem os menores.

Os critérios definidos acima são dados, respectivamente, por

$$AIC = -2\ell(\widehat{\boldsymbol{\theta}}) + 2d,$$
$$BIC = -2\ell(\widehat{\boldsymbol{\theta}}) + d\log(n),$$

em que d a dimensão do vetor $\boldsymbol{\theta}$ (SILVA, 2020), no nosso caso $d = p + q + 2$.

Outro ponto importante na análise de diagnóstico é a análise dos resíduos. Caso o modelo esteja bem ajustado, então espera-se que os resíduos variem de forma aleatória em torno de zero e que possuam uma variação constante. Além disso, quando não se conhece a distribuição dos resíduos, pode-se utilizar o gráfico quantil-normal ou o gráfico quantil-meio-normal, no qual são simulados envelopes e caso os resíduos estejam distribuídos dentro desses envelopes, é um sinal que o modelo está bem ajustado (ATKINSON, 1985).

Existem vários tipos de resíduos, por exemplo: ordinário, deviance, de Pearson, de Cox-Snell, quantílico aleatorizado etc. O resíduo quantílico aleatorizado vem sendo utilizado em diversos trabalhos envolvendo modelos de regressão como em Bourguignon, Leão e Gallardo (2020), Altun (2020), Gómez *et al.* (2020), Silva (2020), Dias (2014), Silva (2020), Júnior e Zeviani (2020) e em Melo, Loose e Carvalho (2021). Esses resíduos são obtidos fazendo uma conversão da variável resposta para uma distribuição normal padrão, isso ocorre, quando o modelo especificado se ajusta bem aos dados. Esses resíduos foram propostos por Dunn e Smyth (1996) e são calculados da seguinte forma

$$r_i^q = \Phi^{-1}(F(y_i;\widehat{\boldsymbol{\theta}})), \quad \text{com } i = 1,2,\ldots,n, \tag{4.1}$$

em que y_i é o i-ésimo valor observado da variável resposta, $F(y_i;\widehat{\boldsymbol{\theta}})$ é a função de distribuição acumulada da variável resposta, $\widehat{\boldsymbol{\theta}}$ é o EMV de $\boldsymbol{\theta}$ e $\Phi^{-1}(.)$ corresponde à função quantil da distribuição normal padrão. Dias (2014) afirma que este tipo de procedimento tem o objetivo principal de evitar massas de pontos na distribuição de resíduos. De acordo com o que está descrito em Ferrari e Cribari-Neto (2004), a ideia principal é aprimorar o gráfico quantil-normal (ou quantil-meio-normal) adicionando envelopes simulados, este processo auxila a decidir se os resíduos observados são consistentes com o modelo ajustado. Ferrari e Cribari-Neto (2004) também expõem o seguinte passo-a-passo para produzir os gráficos de envelopes:

- Ajustar o modelo e gerar uma amostra simulada de n observações independentes usando o modelo ajustado como verdadeiro.
- Ajustar o modelo da amostra gerada e calcular os valores absolutos ordenados de resíduos.
- Repetir as etapas anteriores k vezes.
- Considere os n conjuntos das k estatísticas de ordem para cada conjunto, calcule sua média, valores de mínimo e máximo.
- Plotar esses valores e os resíduos ordenados da amostra original contra $\Phi^{-1}((t+n-1/8)/(2n+1/2))$.

Assim, os valores de mínimo e máximo das estatísticas de ordem k geram o envelope. Além do gráfico de probabilidade normal, também pode-se criar gráficos de dispersão utilizando esses resíduos e verificar alguns comportamentos desejáveis como feito em Bayer (2011). A seguir, são listadas as variáveis envolvidas para averiguar os resíduos,

- Resíduos *versus* os índices: espera-se que os resíduos fiquem distribuídos de forma aleatória em torno do zero e pode-se verificar a possível existência de *outliers* (pontos fora do intervalo $[-3,3]$);

- Resíduos *versus* os valores ajustados: espera-se um comportamento aleatório em torno do zero, isto é, sem padrão; e
- Valores observados da variável resposta *versus* valores ajustados: espera-se que os pontos estejam simetricamente distribuídos em torno da reta identidade.

5 SIMULAÇÃO

Foi realizado um estudo de simulação para averiguar o desempenho do EMV para o modelo de regressão BM. Foram geradas n observações da variável aleatória $Y \sim BM(\mu_i, \phi_i)$, considerando

$$g_1(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$
$$g_2(\phi_i) = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \cdots + \gamma_q z_{qi},$$

em que cada covariável foi gerada de forma independente através de uma distribuição uniforme no intervalo $(0,1)$. Foram considerados três tamanhos amostrais $n = 50, 100, 500$, dois tamanhos para o vetor $\boldsymbol{\beta}$ com $p = 3$ e $p = 4$, dois tamanhos para o vetor $\boldsymbol{\gamma}$ com $q = 1$ (ϕ fixo) e $q = 2$, utilizando as funções de ligação: probito, logito e complemento log-log para a média e a função de ligação logarítmica para a precisão. Foram realizadas $N = 10000$ réplicas de Monte Carlo e, com base nelas, foram calculados

- A média:

$$\widehat{\theta}_{mc} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\theta}_i \tag{5.1}$$

- O Viés relativo:

$$VR(\widehat{\theta}_{mc}) = \frac{\widehat{\theta}_{mc} - \theta}{\theta} \tag{5.2}$$

- O Desvio-padrão:

$$DP(\widehat{\theta}_{mc}) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\widehat{\theta}_i - \widehat{\theta}_{mc})^2} \tag{5.3}$$

em que θ é o valor verdadeiro do parâmetro e $\widehat{\theta}_i$ é a estimativa de máxima verossimilhança da i-ésima réplica de Monte Carlo. Neste trabalho, as estimativas de máxima verossimilhança são obtidas pelos métodos de maximização do tipo Newton através da função `nlm`. Além disso, os valores iniciais utilizados para a obtenção das estimativas dos parâmetros foram calculados através do modelo beta usual, utilizando a transformação $Y^* = (Y + 1)/2$, através da função `betareg` do pacote `betareg` do *software* R Core Team (2022).

Nas Tabelas 2, 3 e 4 apresentam as configurações utilizadas na simulação, as estimativas dos parâmetros, o viés relativo e o desvio-padrão para os modelos logito, probito e complemento log-log, considerando a precisão fixa, respectivamente. Observa-se que o desvio-padrão e o viés relativo diminuem na medida que o tamanho amostral aumenta para todos os

cenários. Pode-se notar também que o viés para o parâmetro ϕ é alto (em média 14% para $n = 50$) e vai para zero mais lentamente, em comparação com os demais parâmetros.

No Apêndice B pode-se verificar os seguintes cenários para a simulação, i: Tabelas 14, 15 e 16 estão os resultados considerando a precisão $\phi = 100$ e as três funções de ligação para a média. Pode-se ver que o comportamento tanto do viés quanto do desvio-padrão são similares em comparação com o cenários no qual $\phi = 50$; ii: Tabelas 17, 18 e 19 possuem os resultados considerando a dimensão $p = 3$ do vetor $\boldsymbol{\beta}$ para $\phi = 50$ e nas Tabelas 20, 21 e 22 com a mesma configuração do vetor $\boldsymbol{\beta}$ considerando $\phi = 100$, e as conclusões são similares às feitas anteriormente.

Tabela 2 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 1: $g_1(\boldsymbol{\mu}^*) = \log(\frac{\boldsymbol{\mu}^*}{1-\boldsymbol{\mu}^*})$**,** $n = 50, 100, 500$ **e** $\phi = 50$**.**

n	Parâmetro	θ	$\widehat{\theta}_{mc}$	$VR(\widehat{\theta}_{mc})$	$DP(\widehat{\theta}_{mc})$
50	β_0	0,5	0,4983	-0,0035	0,1544
	β_1	1,0	1,0041	0,0041	0,1749
	β_2	-1,0	-0,9995	-0,0005	0,1756
	β_3	1,5	1,5035	0,0023	0,1765
	ϕ	50,0	56,9290	0,1386	12,3578
100	β_0	0,5	0,5002	0,0004	0,1057
	β_1	1,0	1,0013	0,0013	0,1199
	β_2	-1,0	-1,0009	0,0009	0,1188
	β_3	1,5	1,5026	0,0017	0,1231
	ϕ	50,0	53,1721	0,0634	7,7485
500	β_0	0,5	0,4997	-0,0006	0,0462
	β_1	1,0	1,0005	0,0005	0,0529
	β_2	-1,0	-0,9999	-0,0001	0,0526
	β_3	1,5	1,5004	0,0003	0,0532
	ϕ	50,0	50,6754	0,0135	3,2383

Tabela 3 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 2 : $g_1(\boldsymbol{\mu}^*) = \Phi^{-1}(\boldsymbol{\mu}^*)$**,** $n = 50, 100, 500$ **e** $\phi = 50$**.**

n	Parâmetro	θ	$\widehat{\theta}_{mc}$	$VR(\widehat{\theta}_{mc})$	$DP(\widehat{\theta}_{mc})$
50	β_0	0,5	0,5006	0,0011	0,1020
	β_1	1,0	1,0048	0,0048	0,1190
	β_2	-1,0	-1,0056	0,0056	0,1186
	β_3	1,5	1,5071	0,0047	0,1247
	ϕ	50,0	57,1603	0,1432	12,8369
100	β_0	0,5	0,4988	-0,0023	0,0708
	β_1	1,0	1,0030	0,0030	0,0829
	β_2	-1,0	-1,0006	0,0006	0,0816
	β_3	1,5	1,5034	0,0023	0,0841
	ϕ	50,0	53,3070	0,0661	7,9614
500	β_0	0,5	0,5002	0,0004	0,0308
	β_1	1,0	1,0004	0,0004	0,0362
	β_2	-1,0	-1,0006	0,0006	0,0359
	β_3	1,5	1,5004	0,0003	0,0366
	ϕ	50,0	50,6193	0,0124	3,3057

Tabela 4 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 3 : $g_1(\boldsymbol{\mu}^*) = \log(-\log(1-\boldsymbol{\mu}^*))$**,** $n = 50, 100, 500$ **e** $\phi = 50$**.**

n	Parâmetro	θ	$\widehat{\theta}_{mc}$	$VR(\widehat{\theta}_{mc})$	$DP(\widehat{\theta}_{mc})$
50	β_0	0,5	0,5016	0,0031	0,0841
	β_1	1,0	1,0020	0,0020	0,1855
	β_2	-1,0	-1,0063	0,0063	0,1866
	β_3	1,5	1,5068	0,0045	0,1891
	ϕ	50,0	56,9170	0,1383	12,5158
100	β_0	0,5	0,4992	-0,0016	0,0579
	β_1	1,0	1,0027	0,0027	0,1281
	β_2	-1,0	-0,9999	-0,0001	0,1273
	β_3	1,5	1,5048	0,0032	0,1298
	ϕ	50,0	53,3379	0,0668	7,8866
500	β_0	0,5	0,4999	-0,0002	0,0255
	β_1	1,0	1,0002	0,0002	0,0567
	β_2	-1,0	-0,9998	-0,0002	0,0561
	β_3	1,5	1,5007	0,0005	0,0572
	ϕ	50,0	50,5910	0,0118	3,3019

Nas Tabelas 5, 6 e 7 possuem as configurações da simulação, bem como as estimativas dos parâmetros, o viés relativo e o desvio-padrão para os modelos logito, probito e complemento log-log, considerando a precisão variável, respectivamente. Nesse caso, temos que a dimensão utilizada para o vetor de $\boldsymbol{\beta}$ é $p = 4$ e a para o vetor $\boldsymbol{\gamma}$ é $q = 2$. Pode-se observar que o desvio-padrão e o viés relativo diminuem à medida que o tamanho amostral aumenta para todos os cenários. Além disso, nota-se que o viés relativo para o parâmetro γ_0 é alto, chegando em torno de 30% para o modelo com função de ligação logito e em torno de 20% para os modelos com função de ligação probito e complemento log-log. O Apêndice B apresenta outros cenários de simulação e os resultados são similares com os apresentados. De forma geral, o EMV se comporta de forma desejável, ou seja, é assintoticamente não viesado e a variância diminui à medida que o tamanho amostral aumenta, indicando também a consistência do estimador.

Tabela 5 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 4: $g_1(\boldsymbol{\mu}^*) = \log(\frac{\boldsymbol{\mu}^*}{1-\boldsymbol{\mu}^*})$**,** $g_2(\phi_i) = \log(\phi_i)$ **e** $n = 50, 100, 500$**.**

n	Parâmetro	θ	$\widehat{\theta}_{mc}$	$VR(\widehat{\theta}_{mc})$	$DP(\widehat{\theta}_{mc})$
50	β_0	0,5	0,5120	0,0241	0,4880
	β_1	1,0	1,0176	0,0176	0,5405
	β_2	-1,0	-1,0254	0,0254	0,5356
	β_3	-0,5	-0,5048	0,0096	0,5313
	γ_0	0,5	0,6521	0,3043	0,5891
	γ_1	1,0	1,1168	0,1168	0,8045
100	β_0	0,5	0,5096	0,0193	0,3316
	β_1	1,0	1,0049	0,0049	0,3634
	β_2	-1,0	-1,0113	0,0113	0,3639
	β_3	-0,5	-0,5089	0,0179	0,3610
	γ_0	0,5	0,5544	0,1088	0,2850
	γ_1	1,0	1,0331	0,0331	0,4727
500	β_0	0,5	0,5007	0,0014	0,1420
	β_1	1,0	1,0010	0,0010	0,1558
	β_2	-1,0	-1,0009	0,0009	0,1583
	β_3	-0,5	-0,4996	-0,0009	0,1555
	γ_0	0,5	0,5106	0,0212	0,1079
	γ_1	1,0	1,0026	0,0026	0,1906

Tabela 6 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 5: $g_1(\boldsymbol{\mu}^*) = \Phi^{-1}(\boldsymbol{\mu}^*)$**,** $g_2(\phi_i) = \log(\phi_i)$ **e** $n = 50, 100, 500$**.**

n	Parâmetro	θ	$\widehat{\theta}_{mc}$	$VR(\widehat{\theta}_{mc})$	$DP(\widehat{\theta}_{mc})$
50	β_0	0,5	0,4930	-0,0140	0,2953
	β_1	1,0	1,0074	0,0074	0,3220
	β_2	-1,0	-0,9951	-0,0049	0,3274
	β_3	-0,5	-0,4914	-0,0172	0,3229
	γ_0	0,5	0,6007	0,2013	0,4124
	γ_1	1,0	1,0593	0,0593	0,7055
100	β_0	0,5	0,5016	0,0033	0,2005
	β_1	1,0	1,0026	0,0026	0,2206
	β_2	-1,0	-0,9993	-0,0007	0,2225
	β_3	-0,5	-0,5019	0,0038	0,2216
	γ_0	0,5	0,5497	0,0993	0,2523
	γ_1	1,0	1,0224	0,0224	0,4500
500	β_0	0,5	0,5013	0,0027	0,0868
	β_1	1,0	1,0002	0,0002	0,0959
	β_2	-1,0	-1,0009	0,0009	0,0963
	β_3	-0,5	-0,5010	0,0019	0,0942
	γ_0	0,5	0,5089	0,0178	0,1080
	γ_1	1,0	1,0058	0,0058	0,1889

Tabela 7 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 6: $g_1(\boldsymbol{\mu}^*) = \log(-\log(1-\boldsymbol{\mu}^*))$**,** $g_2(\phi_i) = \log(\phi_i)$ **e** $n = 50, 100, 500$**.**

n	Parâmetro	θ	$\widehat{\theta}_{mc}$	$VR(\widehat{\theta}_{mc})$	$DP(\widehat{\theta}_{mc})$
50	β_0	0,5	0,4938	-0,0124	0,2619
	β_1	1,0	0,9949	-0,0051	0,2851
	β_2	-1,0	-0,9890	-0,0110	0,2847
	β_3	-0,5	-0,4944	-0,0111	0,2834
	γ_0	0,5	0,6017	0,2035	0,4401
	γ_1	1,0	1,0646	0,0646	0,7091
100	β_0	0,5	0,4978	-0,0043	0,1766
	β_1	1,0	0,9866	-0,0134	0,1943
	β_2	-1,0	-0,9911	-0,0089	0,1933
	β_3	-0,5	-0,4893	-0,0215	0,1885
	γ_0	0,5	0,5494	0,0989	0,2612
	γ_1	1,0	1,0195	0,0195	0,4380
500	β_0	0,5	0,4951	-0,0098	0,0774
	β_1	1,0	0,9871	-0,0129	0,0837
	β_2	-1,0	-0,9876	-0,0124	0,0836
	β_3	-0,5	-0,4911	-0,0178	0,0818
	γ_0	0,5	0,5140	0,0280	0,1076
	γ_1	1,0	0,9917	-0,0083	0,1837

6 APLICAÇÃO EM DADOS REAIS

Neste capítulo são apresentadas duas aplicações, a primeira é referente as eleições estadunidenses no ano de 2016, já a segunda avalia os impactos da criação do programa Bolsa Família nos resultados das eleições presidenciais em 2006 no Brasil.

6.1 ELEIÇÕES ESTADUNIDENSES 2016

O presente estudo tem como objetivo fazer uma análise da diferença da taxa de aceitação dos partidos democratas e republicanos em cada um dos estados, ajustando os dados ao modelo de regressão da BM, levando em consideração que o voto é optativo nos Estados Unidos da América (EUA). Embora existam vários partidos nos EUA, geralmente, a eleição presidencial é disputada pelos partidos mais fortes, que são os partidos Democrata e Republicano. Porém, nem sempre é resultado da votação popular que prevalece, o tipo de eleição que ocorre na região norte do continente americano, permite que os delegados, representantes do estado no colégio eleitoral, decidam o rumo das votações, e quanto mais populoso o estado maior é o peso de seus votos na decisão.

O partido Democrata é aquele cuja ideologia defende igualdade social e econômica, também procura melhorar as regulamentações do mercado e preservação ambiental. Atualmente, a base eleitoral e política do partido Democrata é composta basicamente por progressistas e centristas, com uma pequena parcela de democratas conservadores. Já os Republicanos apresentam uma plataforma baseada no conservadorismo norte-americano, apoio ao capitalismo, forte defesa nacional, desregulamentação e restrições aos sindicatos, buscando ao máximo voltar aos valores tradicionais baseados principalmente na ética judaico-cristã (TOTA, 2008).

A base de dados usada neste trabalho está disponível em Jones (2016), de acordo com os números, houve uma mudança nas eleições de 2016, pois foi uma das poucas vezes em que houve mais estados republicanos, mesmo que os democratas sejam majoritários no país. Os dados dessa pesquisa foram obtidos através de entrevistas feitas por telefone no ano de 2015, com uma amostra aleatória de 177.991 adultos dos 50 estados pertencentes ao país norte-americano, com margem de erro de mais ou menos um ponto percentual com 95% de nível de confiança.

A variável resposta é a diferença entre a taxa de aceitação entre os partidos Democrata e Republicado (DDR), também se tem como variáveis a renda per capita em 1 mil dólares (RPC), a porcentagem de graduados do ensino médio (PGE), a representação percentual de graduados com o ensino superior (PGR) e a porcentagem de pessoas com diplomas avançados (PDA).

Na Figura 2 tem-se um gráfico de barras com as diferenças percentuais referentes às aceitações dos partidos por estado, com base neste, nota-se que houve maior taxa de aceitação para o partido republicano em comparação com a taxa de aceitação para o partido democrata. Além disso, pode-se ver o então candidato Donald Trump acabou vencendo em quase todos os estados que apresentaram uma preferência maior para o partido Republicano. Na Figura 3 tem-se o histograma da variável resposta sobreposto pela densidade estimada da distribuição BM. Note que a distribuição apresenta um bom ajuste aos dados.

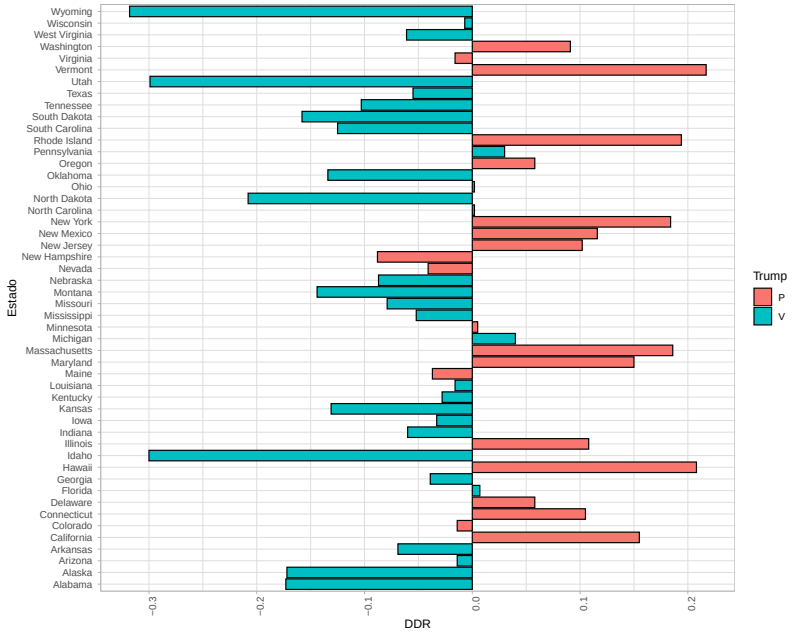


Figura 2 – Gráfico de barras para a variável resposta por estado. A cor azul (V) indica que o Donald Trump venceu e a cor rosa (P) indica que o Donald Trump perdeu.

Na Figura 4 tem-se o gráfico de dispersão, entre todas as variáveis envolvidas no estudo. Nota-se que a variável resposta possui um relação não linear com as variáveis: RPC e PGE, e aparentemente, uma relação linear com variáveis PGR e PDA.

A estrutura de regressão considerada para essa aplicação é dada por

$$g_1(\mu_i^*) = \beta_0 + \beta_1 \log(\text{RPC})_i + \beta_2 \text{PGE}_i + \beta_3 \text{PGR}_i + \beta_4 \text{PDA}_i,$$

e

$$\log(\phi_i) = \gamma_0 + \gamma_1 \log(\text{RPC})_i + \gamma_2 \text{PGE}_i + \gamma_3 \text{PGR}_i + \gamma_4 \text{PDA}_i,$$

para $i = 1, 2, \ldots, 50$. No qual,

- RPC_i: a renda per capita para o i-ésimo estado;
- PGE_i: a porcentagem de graduados do ensino médio para o i-ésimo estado;

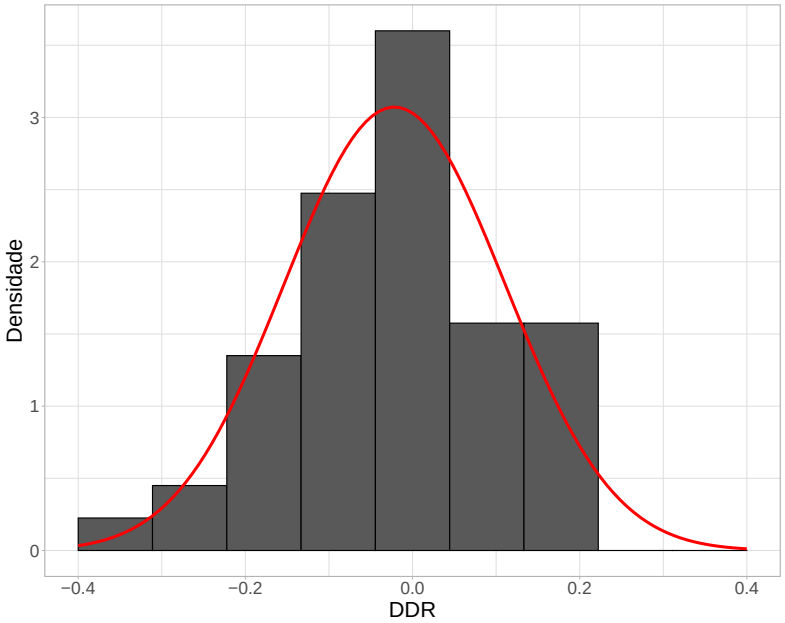


Figura 3 – Histograma para a variável resposta sobreposto pela densidade da distribuição BM (em vermelho).

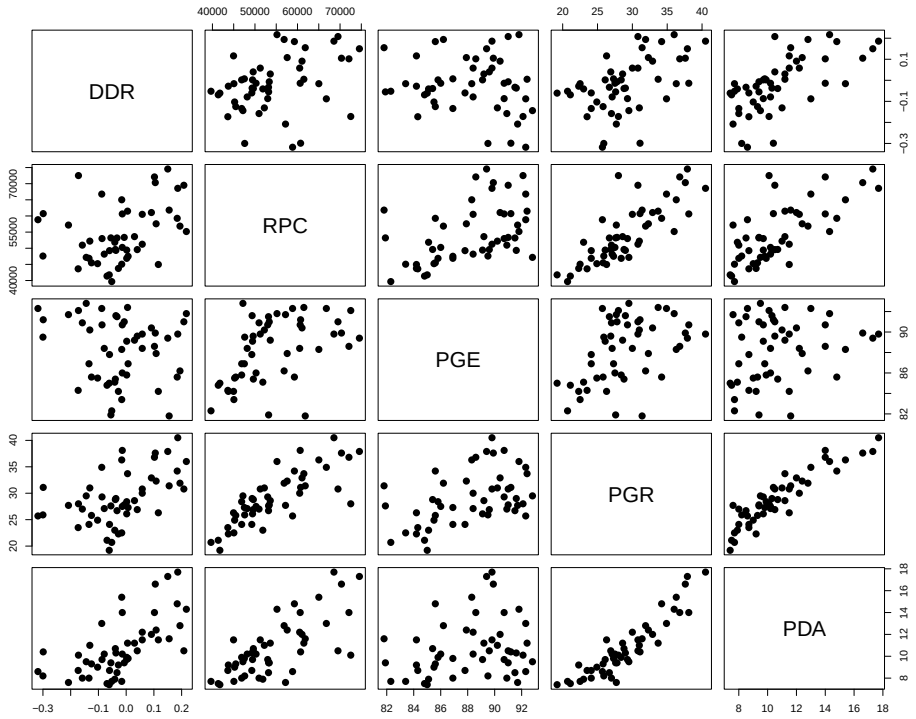


Figura 4 – Gráficos de dispersão para as variáveis no estudo.

- PGR_i: a porcentagem de graduados do ensino superior para o i-ésimo estado; e
- PDA_i: a porcentagem com diplomas avançados para o i-ésimo estado.

Serão utilizadas as funções de ligação probito, logito e complemento log-log para μ_i. Na Tabela 8, tem-se as estimativas dos parâmetros para a estrutura de regressão estabelecida e as medidas de qualidade de ajuste: AIC, BIC e R^2_p. Note que muitas variáveis não foram significativas

e que os modelos forneceram estimativas muito parecidas. Para a estrutura da média, os parâmetros relacionados às variáveis PGE e PDA foram estatisticamente significativos a 10% e 5%, respectivamente; em relação à estrutura da precisão apenas o intercepto foi significativo a 5% para todos modelos. Além disso, o modelo complemento log-log apresentou o menor R_p^2 e os maiores valores de AIC e BIC, comparando com os demais modelos.

Tabela 8 – Estimativas dos parâmetros para o modelo de regressão BM com dispersão variável para os dados sobre a eleição estadunidenses em 2016,

Parâmetro	Probito		Logito		Complemento log-log	
	Estimativa	EP	Estimativa	EP	Estimativa	EP
β_0	0,4550	1,9186	0,7067	3,0678	0,3874	2,2435
β_1	0,0067	0,1923	0,0124	0,3074	-0,0113	0,2245
β_2	-0,0105 **	0,0062	-0,0168 **	0,0098	-0,0125 **	0,0072
β_3	-0,0118	0,0100	-0,0190	0,0159	-0,0124	0,0116
β_4	0,0668 *	0,0155	0,1070 *	0,0248	0,0746 *	0,0178
γ_0	50,5208 *	22,3396	50,6310 *	22,3393	49,8122 *	22,3393
γ_1	-3,7007	2,2577	-3,7106	2,2576	-3,6140	2,2576
γ_2	-0,0573	0,0893	-0,0575	0,0893	-0,0592	0,0893
γ_3	-0,1015	0,1426	-0,1010	0,1426	-0,1042	0,1426
γ_4	0,2632	0,2275	0,2629	0,2275	0,2626	0,2275
R_p^2	0,5012		0,5004		0,4942	
AIC	-160,4264		-160,4614		-159,8034	
BIC	-141,3062		-141,3411		-140,6832	

EP: erro-padrão; *: parâmetro é estatisticamente significativo ao nível de 5%;
** parâmetro é estatisticamente significativo ao nível de 10%.

Foi realizada uma seleção de variáveis tendo com critério de seleção o R_p^2. Na Tabela 9 estão apresentados os ajustes para os modelos após a seleção de variáveis. Note que, todos os parâmetros são significativos ao nível de 5% em todos os modelos. Os modelos probito e logito apresentaram os maiores valores do R_p^2 e os menores valores em relação ao AIC e BIC, em comparação com o modelo com função de ligação complemento log-log. Além disso, percebe-se que os sinais das estimativas são os mesmos para todos os modelos e as estimativas, exceto para o parâmetro β_0, são muito próximas.

Tabela 9 – Estimativas dos parâmetros para o modelo de regressão BM com dispersão variável para os dados sobre a eleição estadunidenses em 2016, após a seleção de variáveis.

Parâmetro	Probito		Logito		Complemento log-log	
	Estimativa	EP	Estimativa	EP	Estimativa	EP
β_0 (intercepto)	0,9779*	0,3855	1,5599*	0,6162	0,7891**	0,4494
β_2 (PGE)	-0,0170*	0,0046	-0,0272*	0,0074	-0,0199*	0,0054
β_4 (PDA)	0,0465*	0,0060	0,0744*	0,0096	0,0527*	0,0067
γ_0 (intercepto)	21,3093*	5,7517	21,3181*	5,7517	21,5655*	5,7514
γ_2 (PGE)	-0,1853*	0,0651	-0,1854*	0,0651	-0,1883*	0,0651
R_p^2	0,5084		0,5077		0,5001	
AIC	-166,1568		-166,1717		-165,8312	
BIC	-156,5966		-156,6116		-156,2711	

EP: erro-padrão; *: parâmetro é estatisticamente significativo ao nível de 5%;
** parâmetro é estatisticamente significativo ao nível de 10%.

Como o modelo probito apresentou o melhor ajuste em termos de R_p^2, embora não tão diferente do modelo logito, tomemos esse modelo para entendemos o relacionamento das covariáveis com a variável resposta. O modelo ajustado com função de ligação probito é dado por

$$\widehat{y} = 2\Phi\left(0,9779 - 0,0170 \cdot \text{PGE} + 0,0465 \cdot \text{PDA}\right) - 1$$
$$\widehat{\phi} = e^{21,3093 - 0,1853 \cdot \text{PGE}}.$$

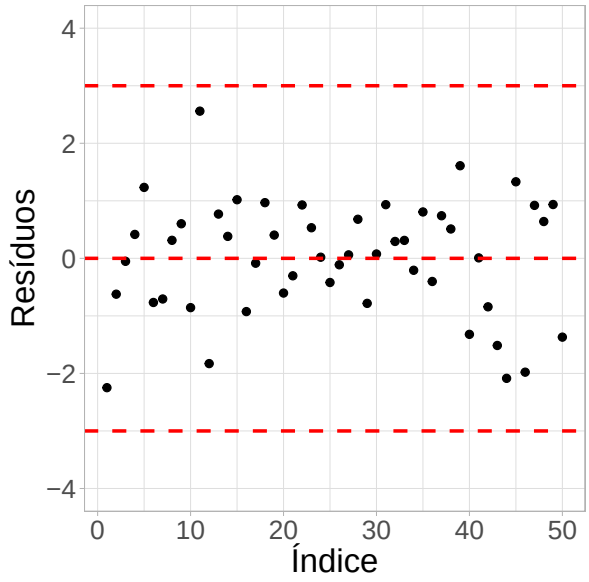
Uma característica da função de ligação probito é a difícil tarefa de interpretar as estimativas dos parâmetros. O que pode ser feito é averiguar o sinal da estimativa e, com isso, verificar se o aumento ou a diminuição do valor contribui para aumentar ou diminuir o valor da variável resposta. Portanto, temos que

- Interpretando a média:
  - Estima-se que, quando a PGE e a PDA são iguais a zero, o valor médio da diferença entre a taxa de aceitação entre os partidos democrata e republicado (DDR) é igual a 0.6719 ($2\Phi(0,9779) - 1$);
  - Fixando a variável PDA, estima-se que, para cada aumento na PGE, o valor médio da diferença entre a taxa de aceitação entre os partidos democrata e republicado aumenta (de forma negativa), ou seja, contribui para que a diferença seja negativa. Portanto, quanto maior a PGE, maior a taxa de aprovação para o partido republicano; e
  - Fixando a variável PGE, estima-se que, para cada aumento na PDA, o valor médio da diferença entre a taxa de aceitação entre os partidos democrata e republicado

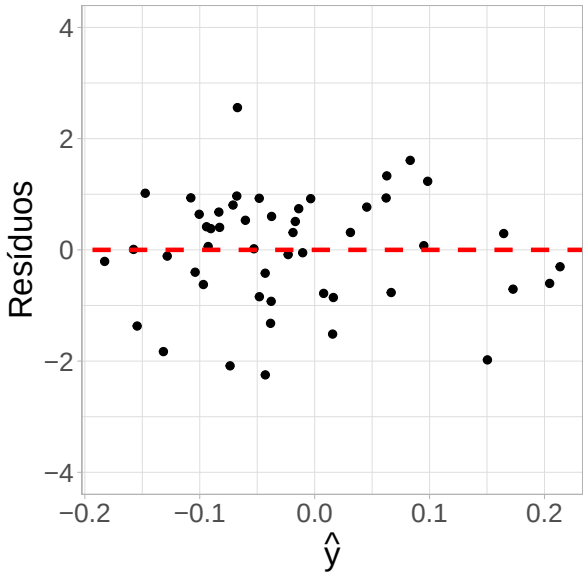
aumenta (de forma positiva), ou seja, contribui para que a diferença seja positiva. Portanto, quanto maior a PDA, maior a taxa de aprovação para o partido democrata.

- Interpretando a precisão:
  - Estima-se que, quando a PGE é igual a zero, a precisão média é igual a 1796915199 ($e^{21,3093}$); e
  - Estima-se que, para cada aumento de uma unidade na PGE, a precisão média diminui em 16.92% ($e^{-0,1853} = 0,8308$). Portanto, quanto maior a PGE, maior a variância.

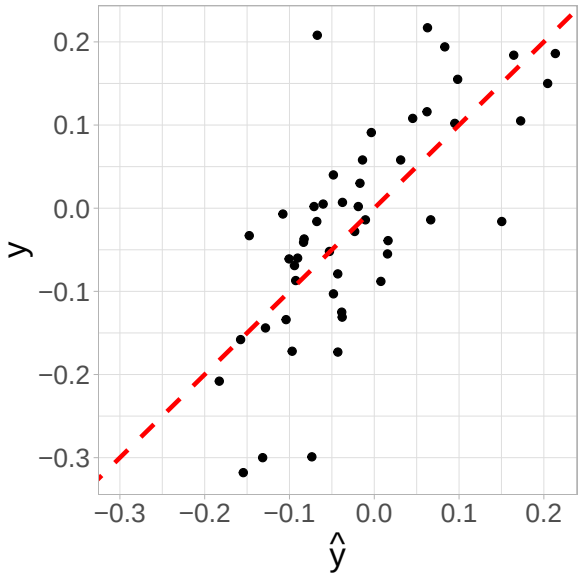
Nas Figuras 5, 6 e 7 temos as análises de resíduos para os modelos probito, logito e complemento log-log, respectivamente. Nas Figuras 5(a), 6(a) e 7(a), percebe-se que não existem *outliers* e que os pontos estão distribuídos de forma aleatória em torno do zero. Nas Figuras 5(b), 6(b) e 7(b), nota-se que os pontos estão distribuídos de forma aleatória em torno do zero sem nenhum padrão. Em relação às Figuras 5(c), 6(c) e 7(c), observa-se um bom ajuste dos modelos em todos os gráficos, pois os pontos estão distribuídos simetricamente em torno da reta identidade. Por fim, nas Figuras 5(d), 6(d) e 7(d), temos o gráfico de probabilidade normal para o valor absoluto dos resíduos quantílicos (ver equação (4.1)) para os modelos probito, logito e complemento log-log, respectivamente. Pode-se perceber que, em todos os gráficos, os resíduos estão variando dentro dos envelopes. Portanto, não temos indícios de afastamento da normalidade dos resíduos para todos os modelos considerados na análise. Portanto, pode-se concluir que todos os modelos apresentaram um bom ajuste.



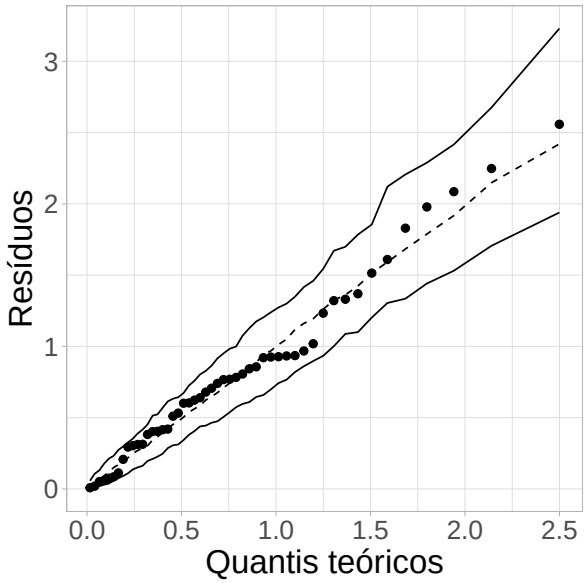
(a) Resíduos versus índices.



(b) Resíduos versus valores ajustados.

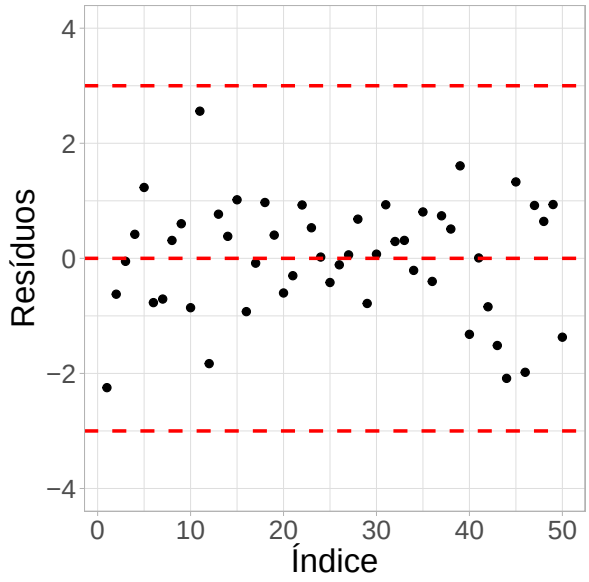


(c) Valores observados versus valores ajustados.

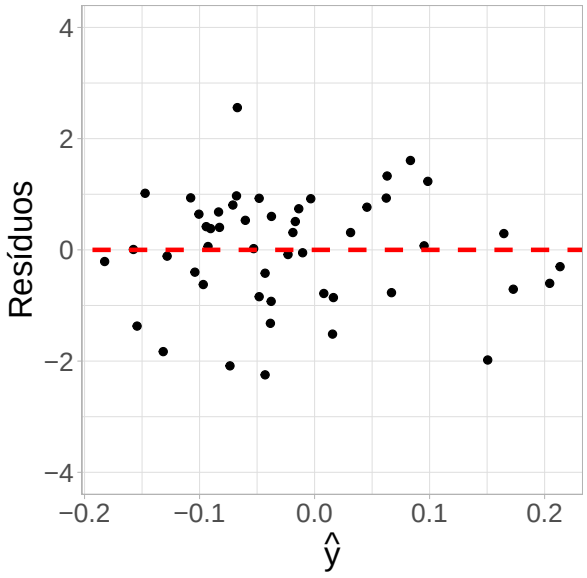


(d) Gráfico de probabilidade normal.

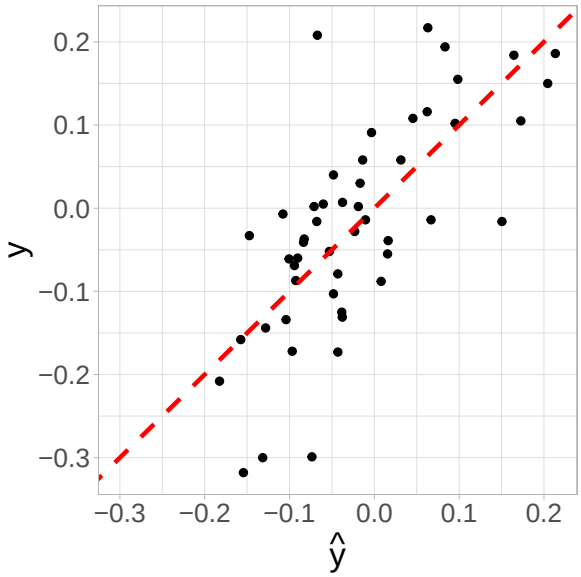
Figura 5 – Análise de diagnóstico para o modelo com função de ligação probito.



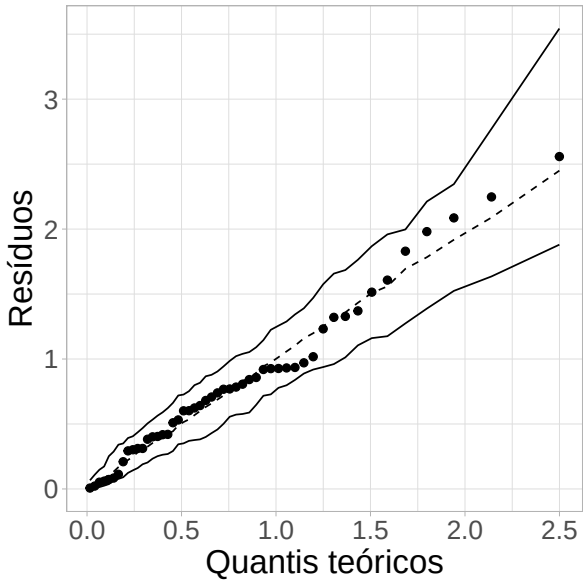
(a) Resíduos versus índices.



(b) Resíduos versus valores ajustados.

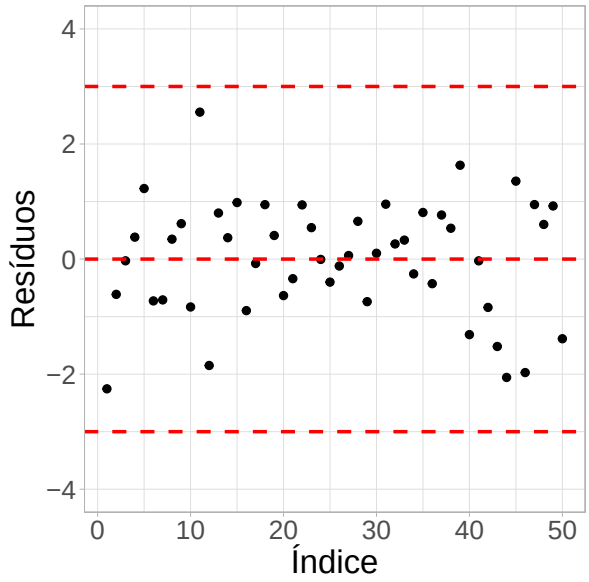


(c) Valores observados versus valores ajustados.

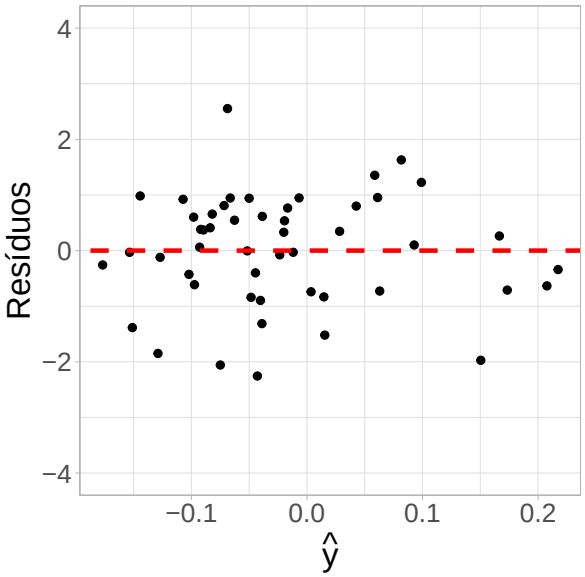


(d) Gráfico de probabilidade normal.

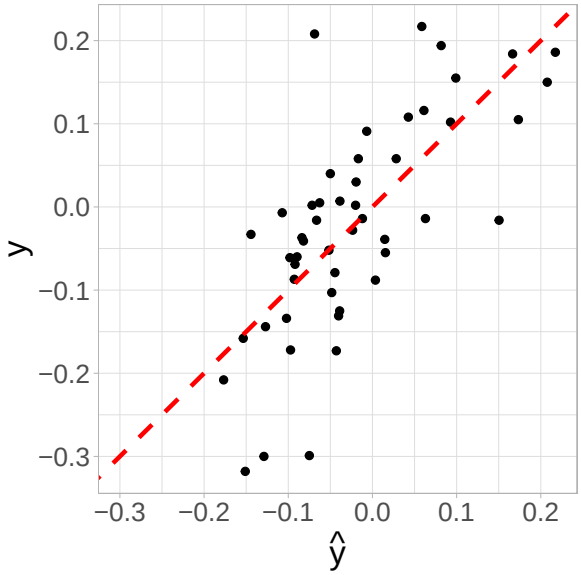
Figura 6 – Análise de diagnóstico para o modelo com função de ligação logito.



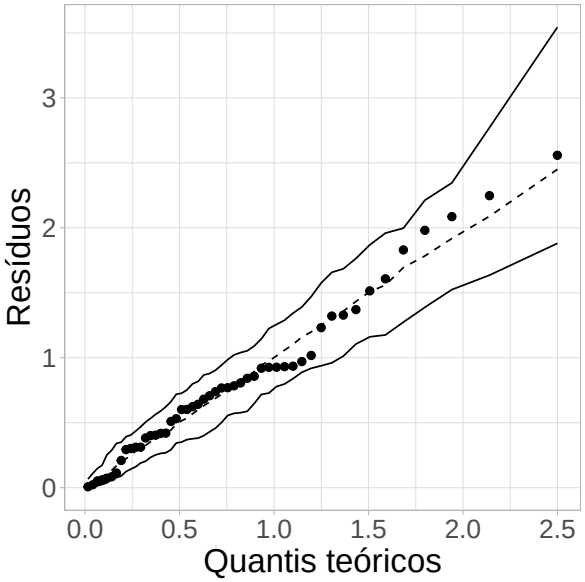
(a) Resíduos versus índices.



(b) Resíduos versus valores ajustados.



(c) Valores observados versus valores ajustados.



(d) Gráfico de probabilidade normal.

Figura 7 – Análise de diagnóstico para o modelo com função de ligação complemento log-log.

6.2 DIFERENÇA DAS TAXAS DE VOTAÇÃO DO CANDIDATO LULA DE 2002 E 2006

Esse estudo tem importância para a comunidade, pois faz uma analise do público eleitor, regiões de maior incidência dos votos no candidato Luiz Inácio Lula da Silva e a criação do programa Bolsa Família.

As eleições para presidência do Brasil ocorrem de 4 em 4 anos, em que os eleitores vão perante as urnas eletrônicas para votar secretamente, porém, diferente do sistema norte-americano, o ato de votar é facultativo para jovens de 16 a 17 anos e obrigatório a partir da idade de 18 até os 70 anos. Além disso, o voto não é obrigatório para quem não sabe se comunicar em português ou quem esteja privado dos direitos políticos, ou seja, analfabetos, como descrito em Brasil (1965).

De acordo com Oliveira e Carneiro (2012), o ex-presidente Luiz Inácio Lula da Silva, durante o seu primeiro mandato, que durou de 2002 a 2006, atendeu a uma série de lutas históricas das classes trabalhadoras, fez uma mudança no ambiente da rede federal de ensino, iniciou a expansão da educação profissional e tecnológica, criou programas e instituições que beneficiavam a comunidade e, principalmente, as pessoas que pertencem às classes sociais mais baixas.

Castro *et al.* (2009) definem políticas sociais como aquelas que constituem um subconjunto das políticas públicas, estas relacionam as ações que determinam qual será o tipo de padrão de proteção social implementado pelo Estado e estão diretamente relacionadas a distribuição de benefícios para os integrantes da sociedade, afim de reduzir as desigualdades estruturais, que são consequências do desenvolvimento socioeconômico, refletidos nas áreas de educação, saúde, previdência, habitação, saneamento etc., e executadas sob a responsabilidade do Estado.

Outra forma de abordar a pobreza mencionada por Castro *et al.* (2009), no qual se tem que este problema não é individual, mas coletivo, isso porque o nível de pobreza de determinado país afeta diretamente o Índice de Desenvolvimento Humano. Um dos principais fundamentos para criação deste conceito foi o alto número de desempregados, da ruptura dos laços sociais e o consequente enfraquecimento da coesão e da solidariedade, desse modo, as compensações financeiras entraram em pauta como forma de reduzir o problema em situações específicas.

Com base nisso e no exposto por Kerstenetzky (2009), durante a administração de Lula houve implementações de programas sociais que tinham como objetivo transferir renda

para os desafortunados, reduzindo assim a pobreza e diminuindo a desigualdade social, assim foi criado o programa Bolsa Família. Com o tempo houve a consolidação, ampliação, redefinição e unificação do programa, que aconteceu inicialmente em algumas cidades, para expandir e se tornar nacional, atendendo famílias pobres com crianças de até 15 anos, provendo assistência preventiva à saúde, especialmente de crianças pequenas e mulheres grávidas, e se baseando nas frequências regulares da prole na escola.

Tendo isso em mente, observou-se por Abensur, Cribari-Neto e Menezes (2007) que nas eleições de 2006, o até então presidente obteve a maior votação da história em um segundo turno. Como seu governo foi voltado para o público mais carente, seu maior índice de eleitorado pertenceu aos locais menos desenvolvidos do país, com isso pode-se dizer que os eleitores se mostraram a favor de tal candidato são aqueles que tem um acesso mais restrito a instrução ou foram beneficiados pelos programas.

Por outro lado, existem aqueles que acreditam que na verdade o que a população carente está recebendo não passa de uma esmola e que esta causa uma acomodação por parte daqueles que a recebem, pois de acordo com as pessoas da oposição, os beneficiados não têm necessidade de trabalharem pois recebem o seu dinheiro sem nenhum esforço. Além disso, existem ainda os casos de fraude e o incentivo à natalidade, já que isso ajuda a perpetuar o benefício, que incentivem o sentimento de rejeição em relação ao programa, sabe-se que esta não é a realidade de todos os que recebem este incentivo, mas sim a visão dos opositores.

Com o intuito de avaliar o impacto do programa Bolsa Família na eleição de 2006, a variável resposta é a diferença entre a proporção de votos do candidato Lula no segundo turno nas eleições no ano de 2006 com a proporção de votos de 2002, na qual denotaremos por pela sigla D62, que significa a diferença entre o percentual de votação entre os anos de 2006 e 2002. Além disso, vamos considerar as seguintes covariáveis: a renda domiciliar per capita média (renda), o índice de gini, densidade demográfica (des.demo.), índice de mortalidade infantil (mor.inf), taxa de analfabetismo (taxa.analfa) e número de famílias beneficiadas com o Programa Bolsa Família (bolsa).

Na Figura 8, pode-se ver o histograma da variável resposta simbolizada por D62, sobreposto pela densidade estimada da distribuição BM, no qual os parâmetros foram estimados via MV. Pode-se perceber que o modelo se ajusta razoavelmente bem aos dados.

A Figura 9 possui os gráficos de dispersão para as variáveis do estudo. Percebe-se que a variável reposta D62 possui uma relação não linear com as variáveis bolsa, renda e des.demo, não apresenta relação com o índice de gini e apresenta uma relação linear com as

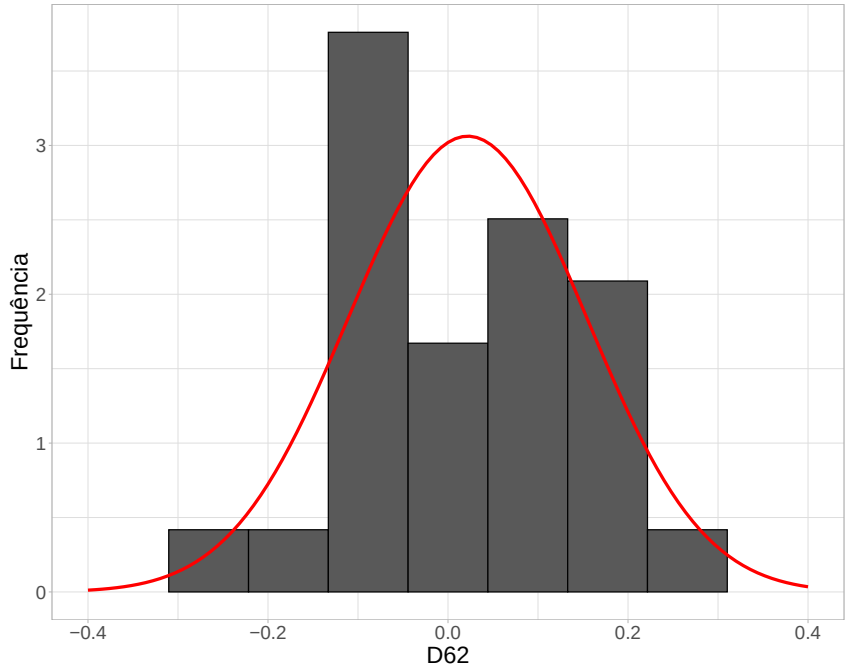


Figura 8 – Histograma para a variável resposta sobreposto pela densidade estimada da distribuição BM.

variáveis mor.inf e taxa.analfa.

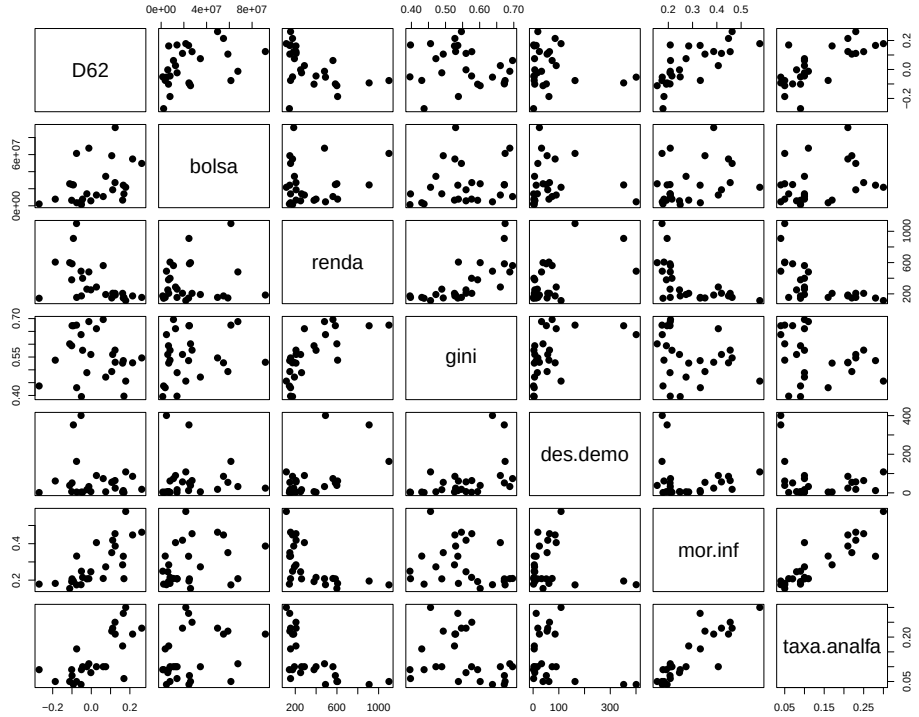


Figura 9 – Gráfico de dispersão para as variáveis em estudo.

Os modelos que serão utilizados nessa aplicação estão disposto a seguir:

$$g_1(\mu_i^*) = \beta_0 + \beta_1 \log(\text{bolsa})_i + \beta_2 \log(\text{renda})_i + \beta_3 \log(\text{des.demo.})_i + \beta_4 \text{mor.inf}_i + \beta_5 \text{taxa.analfa}_i$$

e

$$\log(\phi_i) = \gamma_0 + \gamma_1 \log(\text{bolsa})_i + \gamma_2 \log(\text{renda})_i + \gamma_3 \log(\text{des.demo.})_i + \gamma_4 \text{mor.inf}_i + \gamma_5 \text{taxa.analfa}_i$$

para $i = 1, 2, \ldots, 27$. No qual,

- renda_i: a renda domiciliar per capita para o i-ésimo estado, extraída da PNAD de 2005;
- des.demo_i: a densidade demográfica para o i-ésimo estado em 2005, medida em habitantes por quilômetros quadrados;
- mor.inf_i: o índice de mortalidade infantil para o i-ésimo estado em 2002;
- taxa.analfa_i: a taxa de analfabetismo para o i-ésimo estado em 2003; e
- bolsa_i: o número de famílias beneficiadas com o Programa Bolsa Família, dados do TSE de 2006.

O índice de gini foi removido do modelo devido a complicações numéricas, isto é, a não convergência do método de estimação.

Pode-se observar na Tabela 10 as estimativas dos parâmetros para os modelos considerados e os seus respectivos erros-padrão (EP). Percebe-se que muitos parâmetros não são significativos ao nível de significância de 5%.

- Modelo probito
  - Estrutura para média: o parâmetro relacionado à variável logaritmo do número de beneficiados no programa bolsa família (bolsa) foi significativa a 5% e o parâmetro relacionado à mortalidade infantil (mor.inf) foi significativo a 10%.
  - Estrutura para precisão: o intercepto e o parâmetro relacionado à variável logaritmo do número de beneficiados no programa bolsa família (bolsa) foram significativos a 5%.
- Modelo logito:
  - Estrutura para média: o parâmetro relacionado à variável logaritmo do número de beneficiados no programa bolsa família (bolsa) foi significativa a 5%.
  - Estrutura para precisão: o intercepto e o parâmetro relacionado à variável logaritmo do número de beneficiados no programa bolsa família (bolsa) foram significativos a 5%.
- Modelo complemento log-log:

- Estrutura para média: os parâmetros relacionado às variáveis logaritmo do número de beneficiados no programa bolsa família (bolsa) e o logaritmo da renda foram significativos a 5%.
- Estrutura para precisão: os parâmetros relacionados às variáveis índice de mortalidade infantil (mor.inf) e taxa de analfabetismo (taxa.analf) foram significativo a 5%.

Em relação às medidas de qualidade de ajuste, temos que o modelo com função de ligação probito apresentou o maior valor de R_p^2, enquanto que o modelo com função de ligação complemento log-log apresentou os menores valores de AIC e BIC.

Tabela 10 – Ajuste dos modelos para a avaliação da diferença das taxas de votação em 2002 e 2006.

Parâmetro	Probito		Logito		Complemento loglog	
	Estimativa	EP	Estimativa	EP	Estimativa	EP
β_0	-0,3840	0,3763	-0,6108	0,6013	-0,9005	0,4595
β_1	0,0388 *	0,0165	0,0616 *	0,0264	0,0784 *	0,0185
β_2	-0,0701	0,0559	-0,1115	0,0893	-0,1475*	0,0693
β_3	-0,0083	0,0176	-0,0131	0,0281	-0,0049	0,0201
β_4	0,5926 **	0,3203	0,9471	0,5123	0,3403	0,3025
β_5	0,1907	0,4427	0,3105	0,7083	0,1225	0,3374
γ_0	-14,2504 *	6,8632	-14,2170 *	6,8632	-11,7638	6,8690
γ_1	0,7867 *	0,3233	0,7876 *	0,3233	0,3017	0,3237
γ_2	1,0991	1,0583	1,0918	1,0583	1,7456	1,0573
γ_3	-0,0781	0,3160	-0,0745	0,3160	-0,2850	0,3159
γ_4	4,4538	6,1718	4,3346	6,1718	23,3780 *	6,1152
γ_5	-2,7073	8,1846	-2,5814	8,1846	-24,0287 *	8,0758
R_p^2	0,6418		0,6410		0,6312	
AIC	-89,4618		-89,4596		-89,8383	
BIC	-73,9118		-73,9095		-74,2883	

EP: erro-padrão; *: parâmetro é estatisticamente significativo ao nível de 5%;
** parâmetro é estatisticamente significativo ao nível de 10%.

Uma seleção de variáveis foi realizada para todos os modelos, considerando como critério de seleção o valor do R_p^2. A Tabela 11 possui os melhores modelos após a seleção de variáveis. Note que os modelos possuem estimativas próximas. Em termos de R_p^2, os modelos probito e logito apresentaram os maiores valores, já em termos de AIC e BIC o modelo logito apresentou os menores valores, mas não tão diferentes do modelo probito. Em linhas gerais, o modelo com função de ligação complemento log-log obteve o pior desempenho, em comparação com os demais modelos.

O modelo ajustado com função de ligação probito é dado por

Tabela 11 – Ajuste dos modelos, após a seleção de variáveis, para a avaliação da diferença das taxas de votação em 2002 e 2006.

Parâmetro	Probito		Logito		Complemento loglog	
	Estimativa	EP	Estimativa	EP	Estimativa	EP
β_1 (log(bolsa))	0,0310 *	0,0114	0,0494 *	0,0182	0,0212 **	0,0127
β_2 (log(renda))	-0,1104 *	0,0269	-0,1762 *	0,0431	-0,1429 *	0,0305
β_4 (mor.inf)	0,5241 *	0,1747	0,8396 *	0,2801	0,5108 *	0,1928
γ_0 (intercepto)	-9,3377 *	4,2052	-9,3372 *	4,2052	-10,3170 *	4,2002
γ_1 (log(bolsa))	0,8963 *	0,2534	0,8963 *	0,2534	0,9522 *	0,2531
R_p^2	0,6109		0,6101		0,5363	
AIC	-100,0885		-100,0925		-98,6115	
BIC	-93,6093		-93,6134		-92,1323	

EP: erro-padrão; *: parâmetro é estatisticamente significativo ao nível de 5%;
** parâmetro é estatisticamente significativo ao nível de 10%.

$$\widehat{y} = 2\Phi\left(0,0310 \cdot \log(\text{bolsa}) - 0,1104 \cdot \log(\text{renda}) + 0,5241 \cdot \text{mor.inf}\right) - 1$$

$$\widehat{\phi} = e^{-9,3372+0,8963\log(\text{bolsa})}.$$

Interpretações:

- Interpretando a média:
  - Fixando as variáveis log(renda) e mor.inf, estima-se que, para cada aumento no log(bolsa), o valor médio da diferença entre a proporção de votos nas eleições de 2006 e 2002 (de forma positiva), ou seja, contribui para que a diferença seja positiva. Portanto, quanto maior o valor do log(bolsa), maior a proporção de votos no ano de 2006 em comparação com 2002;
  - Fixando as variáveis log(bolsa) e mor.inf, estima-se que, para cada aumento no log(renda), o valor médio da diferença entre a proporção de votos nas eleições de 2006 e 2002 (de forma negativa), ou seja, contribui para que a diferença seja negativa. Portanto, quanto maior o valor do log(renda), menor a proporção de votos no ano de 2006 em comparação com 2002; e
  - Fixando as variáveis log(bolsa) e log(renda), estima-se que, para cada aumento na taxa de mortalidade infantil, o valor médio da diferença entre a proporção de votos nas eleições de 2006 e 2002 (de forma positiva), ou seja, contribui para que a diferença seja positiva. Portanto, quanto maior o valor da taxa de mortalidade, maior a proporção de votos no ano de 2006 em comparação com 2002;
- Interpretando a precisão:

- Estima-se que, quando o log(bolsa) é igual a zero, a precisão média é igual a $8,804165 \cdot 10^{-5}$ ($e^{-9,3377}$); e
- Estima-se que, para cada aumento de uma unidade no log(bolsa), a precisão aumenta em 2.4506 ($e^{0,8963} = 2,4506$). Portanto, quanto maior o log(bolsa), menor a variância.

Nas Figuras 10, 11 e 12 temos as análises de resíduos para os modelos probito, logito e complemento log-log, respectivamente. Nas Figuras 10(a), 11(a) e 12(a), pode-se perceber que não existem *outliers* e os pontos estão distribuídos em torno de zero de forma aleatória. Nas Figuras 10(b), 11(b) e 12(b), percebe-se que os pontos estão distribuídos de forma aleatória em torno do zero sem nenhum padrão. Além disso, nas Figuras 10(c), 11(c) e 12(c), observa-se um bom ajuste dos modelos em todos os gráficos, pois os pontos estão distribuídos simetricamente em torno da reta identidade. Por fim, nas Figuras 10(d), 11(d) e 12(d), temos o gráfico de probabilidade meio-normal para o valor absoluto dos resíduos quantílicos para os modelos probito, logito e complemento log-log, respectivamente. Percebe-se que em todos os gráficos, os resíduos estão variando dentro dos envelopes. Portanto, não temos indícios de afastamento da normalidade dos resíduos para todos os modelos considerados na análise e, pode-se, concluir que todos os modelos apresentaram um bom ajuste.

Abensur, Cribari-Neto e Menezes (2007) utilizaram um modelo de regressão beta usual com função de ligação logito e com a precisão fixa, para os dados descritos aqui, e utilizaram a proporção de votos do ex-presidente Lula em 2002 como uma covariável para tentar explicar a proporção de votos em 2006. O ajuste feito por Abensur, Cribari-Neto e Menezes (2007) está descrito na Tabela 13 (obs: os valores diferiram um pouco, acredita-se que essa diferença é devida ao método numérico utilizado). Dessa forma, foram ajustados modelos de regressão BM com precisão fixa. O melhor modelo encontrado foi o modelo com função de ligação logito e está descrito na Tabela 12.

Comparando os dois modelos, pode-se perceber que o modelo regressão beta apresentou um valor do R_p^2 um pouco maior (diferença de 0,0036, irrelevante). Outro ponto para se destacar é a estimativa de ϕ, o modelo de regressão BM apresentou uma estimativa bem maior (diferença de 138,762). Em termos de AIC e BIC o modelo de regressão BM apresentou valores bem menores, isto é, o ajuste foi melhor para o modelo de regressão com base nesses dois critérios. Na Figura 13 pode-se ver a análise diagnóstica para o modelo de regressão BM e pode-se ver um comportamento adequado em todos os gráficos.

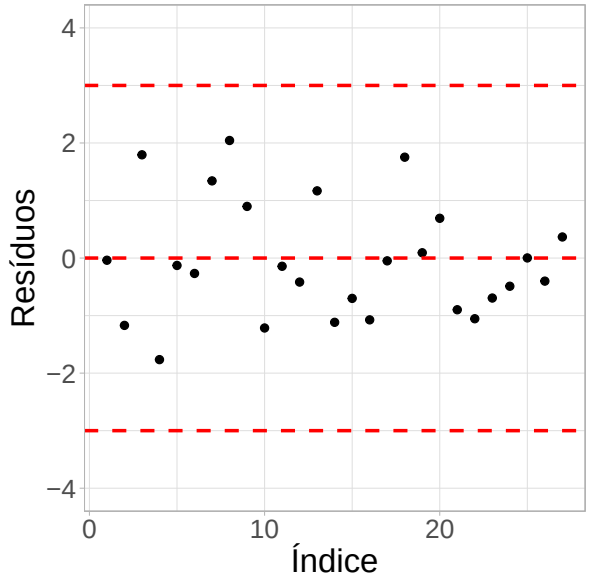
Tabela 12 – Ajuste do modelo de regressão BM com função de ligação logito e precisão fixa para a avaliação da diferença das taxas de votação em 2002 e 2006.

Parâmetro	Estimativa	EP	Valor-p
β_0 (Intercepto)	-1,0122	0,5162	0,0499
β_1 (log(bolsa))	0,0984	0,0327	0,0026
β_2 (log(renda))	-0,1453	0,0691	0,0355
β_4 (mor.inf)	0,8483	0,4070	0,0371
ϕ	174,1428	47,2613	-
R_p^2	0,6589		
AIC	-90,5962		
BIC	-84,1170		

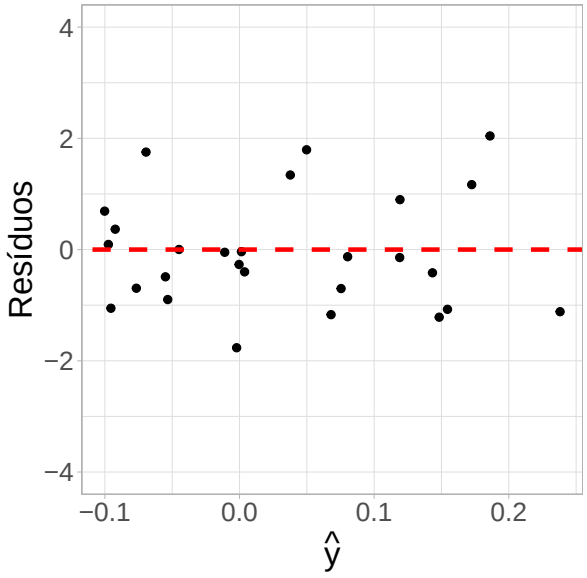
Tabela 13 – Ajuste do modelo de regressão beta com função de ligação logito e precisão fixa para a avaliação da diferença das taxas de votação em 2002 e 2006.

Parâmetro	Estimativa	EP	Valor-p
β_0 (Intercepto)	-3,5267	1,3272	<0,01
β_1 (log(bolsa))	0,2993	0,0661	<0,01
β_2 (log(renda))	-0,5524	0,1103	<0,01
β_6 (Pvotos2002)	3,6767	0,9505	<0,01
ϕ	35,3808	9,5112	-
R_p^2	0,6625		
AIC	-51,6287		
BIC	-45,1496		

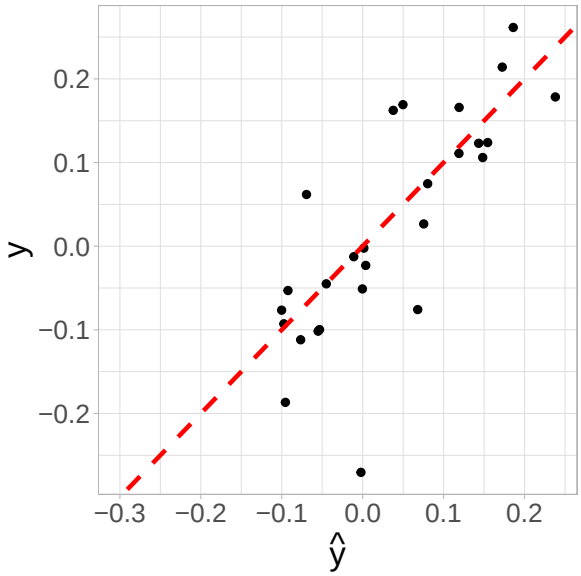
Pvotos2002: proporção de votos em 2002.



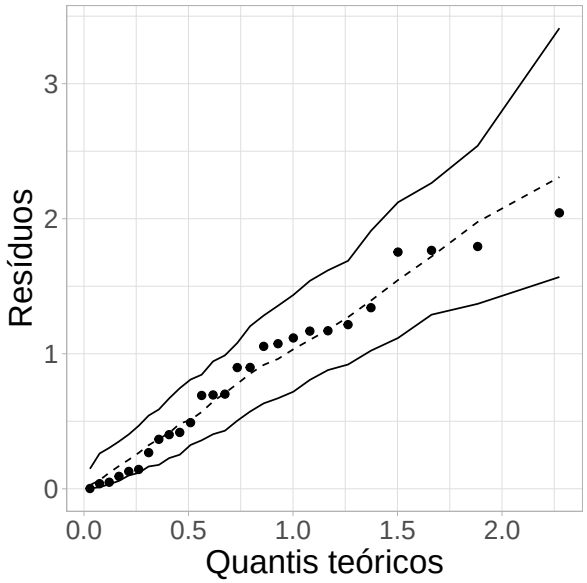
(a) Resíduos versus índices.



(b) Resíduos versus valores ajustados.

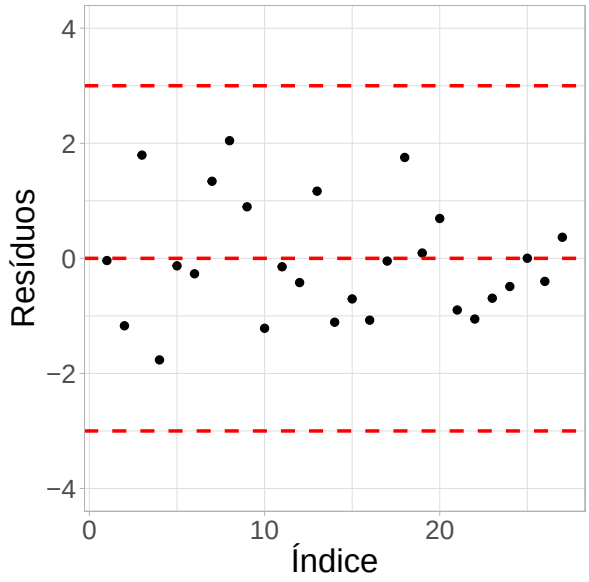


(c) Valores observados versus valores ajustados.

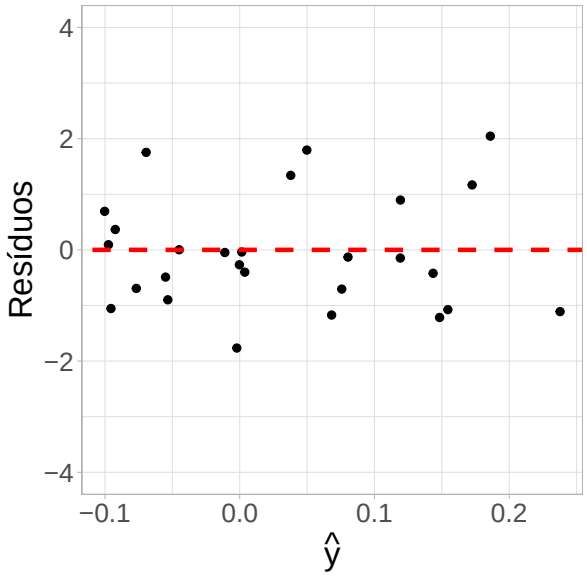


(d) Gráfico de probabilidade normal.

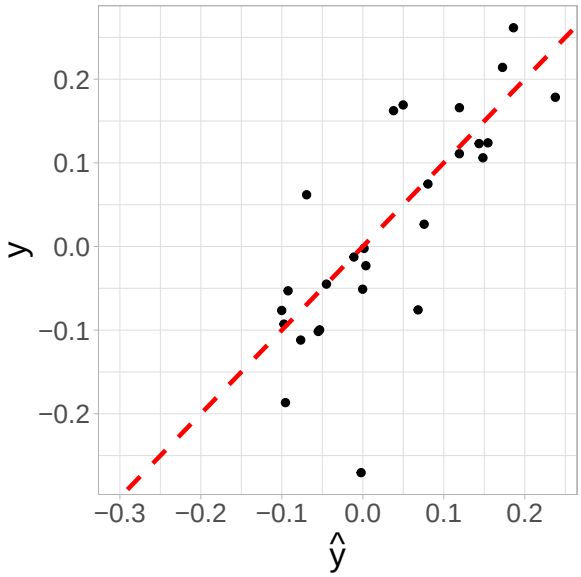
Figura 10 – Análise de diagnóstico para o modelo com função de ligação probito.



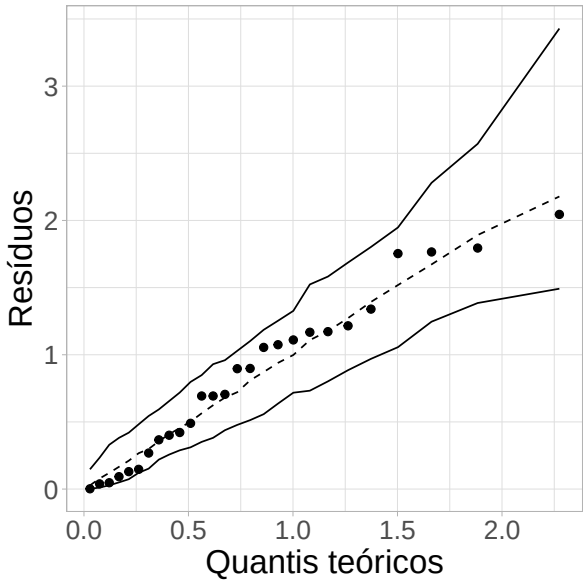
(a) Resíduos versus índices.



(b) Resíduos versus valores ajustados.

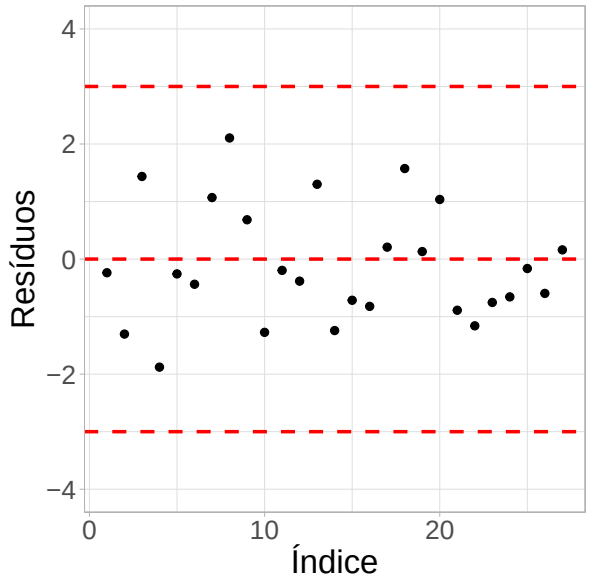


(c) Valores observados versus valores ajustados.

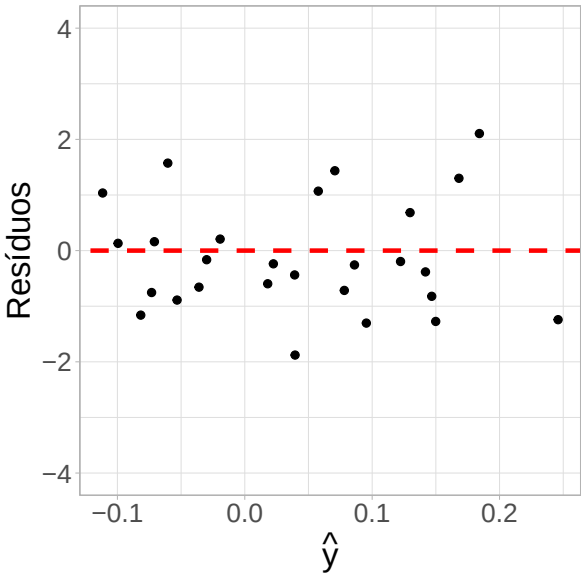


(d) Gráfico de probabilidade normal.

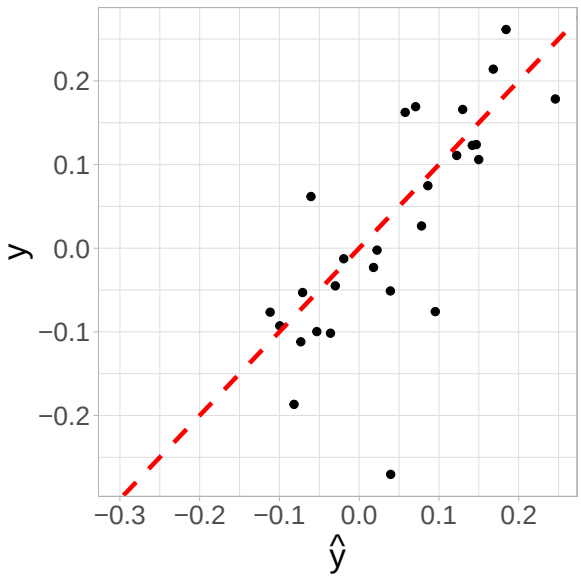
Figura 11 – Análise de diagnóstico para o modelo com função de ligação logito.



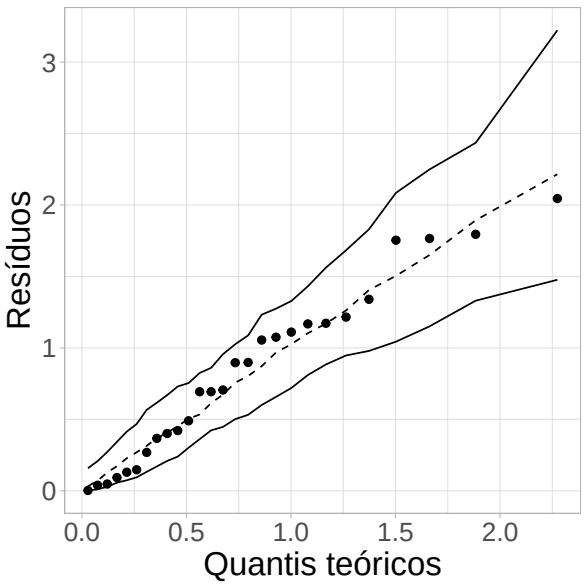
(a) Resíduos versus índices.



(b) Resíduos versus valores ajustados.

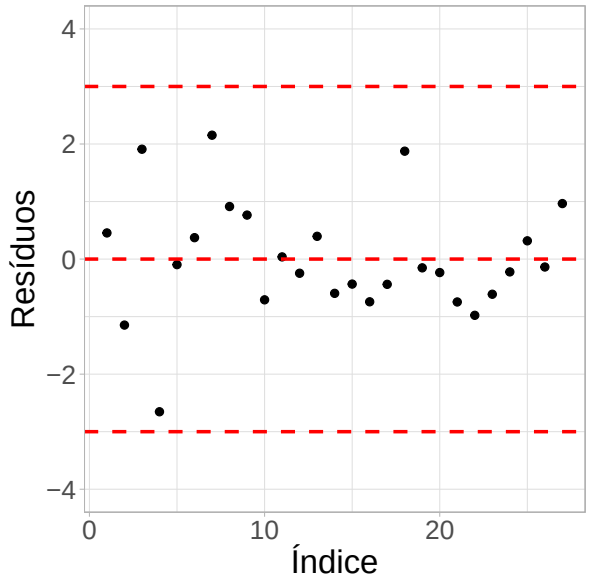


(c) Valores observados versus valores ajustados.

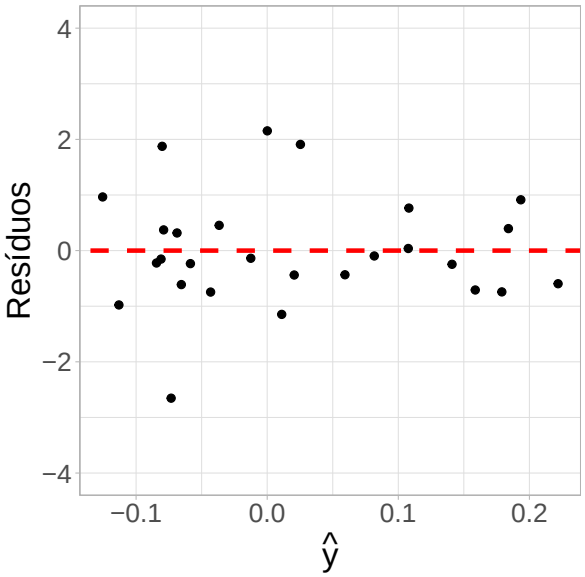


(d) Gráfico de probabilidade normal.

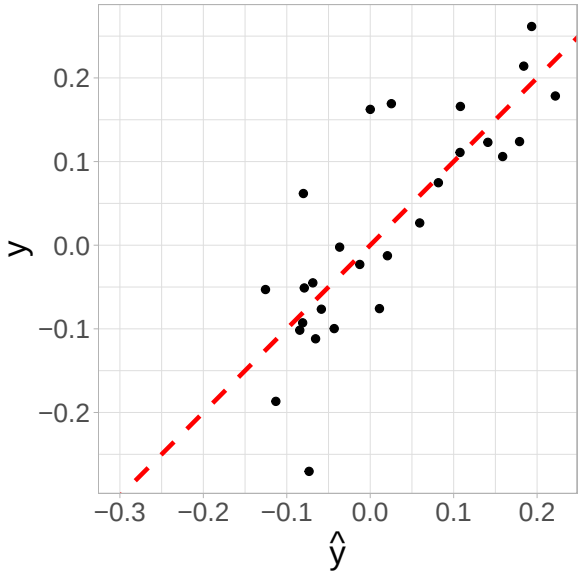
Figura 12 – Análise de diagnóstico para o modelo com função de ligação complemento log-log.



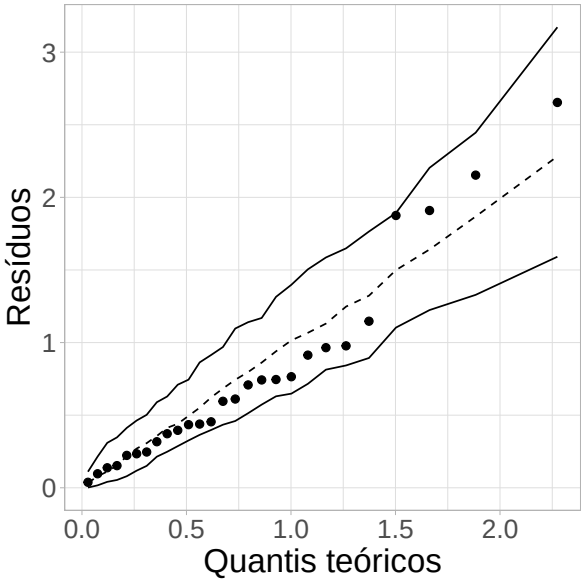
(a) Resíduos versus índices.



(b) Resíduos versus valores ajustados.



(c) Valores observados versus valores ajustados.



(d) Gráfico de probabilidade normal.

Figura 13 – Análise de diagnóstico para o modelo com função de ligação logito para o modelo de regressão BM com precisão fixa.

7 CONSIDERAÇÕES FINAIS

Neste trabalho, foi estudada a distribuição beta reparametrizada chamada de beta modular, na qual considera-se variáveis presentes no intervalo $(-1, 1)$, cuja reparametrização foi feita em relação a média e precisão da variável resposta. Foram expostas algumas propriedades dessa distribuição e a estimação dos parâmetros foi realizada via máxima verossimilhança.

Foram realizados estudos de simulação utilizando o método de Monte Carlo, para avaliar o comportamento do estimador de máxima verossimilhança para os parâmetros do modelo. A partir dessas simulações, conclui-se que os estimadores de máxima verossimilhança do modelo de regressão beta modular apresentam boas propriedades em relação ao viés relativo e desvio-padrão.

Foram realizadas duas aplicações em dados reais, para as quais foram feitos ajustes no modelo beta modular, a primeira aplicação da regressão foi em dados referentes às eleições estadunidenses de 2016, tal conjunto de dados é composto por 177.991 respostas de adultos pertencentes aos 50 estados norte-americanos, em que a variável resposta diz respeito à diferença da taxa de aprovação entre os partidos Democrata e Republicano para cada um dos 50 estados. O modelo de regressão que apresentou o melhor ajuste foi o modelo de regressão BM com função de ligação probito. Pode-se concluir que as pessoas com maior grau de instrução tendem a avaliar melhor o partido Democrata. A segunda aplicação foi realizada utilizando como variável reposta à diferença entre as proporções de votos do ex-presidente Luiz Inácio Lula da Silva nas eleições de 2002 e 2006, período em que foi implantado o programa Bolsa Família, para todas as unidades da federação. Três modelos de regressão BM foram ajustados a esses dados e concluiu-se que os melhores ajustes foram os modelos com as funções de ligação logito e probito com precisão variável. Pode-se constatar que, quanto maior o investimento no programa, maior a proporção de votos em 2006, a taxa de mortalidade também influencia para esse aumento da proporção de votos, com isso, pode-se concluir que a tendência é que o ex-presidente tenha aumentado a proporção de votos nos estados mais pobres. Em contra partida, a renda é um fator que diminui a proporção de votos, isto é, os estados mais ricos tendem a não ter uma votação maior na eleição de 2006 comparado com a eleição de 2002. Além disso, foi comparado o modelo proposto com o modelo de regressão beta usual utilizado por Abensur, Cribari-Neto e Menezes (2007), e através dos critérios AIC e BIC o modelo de regressão BM apresentou os menores valores em comparação ao modelo de regressão beta usual, além de um valor maior para a precisão, o que ocasionou em um ajuste melhor.

Com base no exposto, concluiu-se que o modelo de regressão BM pode ser utilizado para modelar diretamente variáveis limitados ao intervalo $(-1,1)$, em especial diferenças entre taxas ou proporções, o que evita o uso de transformações nessas variáveis, como feito em Souza (2011), para modelar a diferença entre a proporção de votos de 2006 e 2002 do ex-presidente Lula considerando todos os municípios brasileiros.

Trabalhos futuros

Algumas ideias de linhas de pesquisas que podem ser desenvolvidas no futuro:

1. Propor medidas de influência, como a distância de Cook;
2. Propor novas distribuições que possam ser utilizadas no intervalo $(-1,1)$; e
3. Modelar dados mais recentes com sobre as eleições.

REFERÊNCIAS

ABENSUR, T. C.; CRIBARI-NETO, F.; MENEZES, T. A. Impactos do programa bolsa família nos resultados das eleições presidenciais no brasil em 2006. **Anais do XXXV Encontro Nacional de Economia**, Anpec Recife, v. 51, 2007.

ALTUN, E. The lomax regression model with residual analysis: an application to insurance data. **Journal of Applied Statistics**, Taylor & Francis, p. 1–10, 2020.

ATKINSON, A. C. **Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis.** [S.l.]: New York: Oxford University Press, 1985.

BAYER, F. M. Modelagem e inferência em regressão beta. 2011.

BOURGUIGNON, M.; LEÃO, J.; GALLARDO, D. I. Parametric modal regression with varying precision. **Biometrical Journal**, Wiley Online Library, v. 62, n. 1, p. 202–220, 2020.

BRASIL. **LEI Nº 4.737, DE 15 DE JULHO DE 1965.** Brasília, DF, 1965. Acesso em: 23 de maio de 2022. Disponível em: <http://www.planalto.gov.br/ccivil_03/Leis/L4737compilado.htm>.

CASELLA, G.; BERGER, R. L. Inferência estatística-tradução da 2ª edição norte-americana. **Centage Learning**, 2011.

CASTRO, H. C. d. O. de; WALTER, M. I. M. T.; SANTANA, C. M. B. de; STEPHANOU, M. C. Percepções sobre o programa bolsa família na sociedade brasileira. **Opinião pública**, SciELO Brasil, v. 15, p. 333–355, 2009.

DIAS, M. F. Modelos assimétricos inflacionados de zero. 2014.

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.

EMILIANO, P. C.; VIVANCO, M. J. F.; MENEZES, F.; AVELAR, F. G. Fundamentos e comparação de critérios de informação: Akaike and bayesian. **Revista Brasileira Biomatemática**, v. 27, n. 3, p. 394–411, 2009.

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.

GARCIA-MARQUES, T.; QUELHAS, A. C.; GOMES, J. F. Os modelos log-lineares em investigação psicológica. **Análise Psicológica**, ISPA-CRL, v. 15, n. 1, p. 29–48, 1997.

GÓMEZ, Y. M.; GALLARDO, D. I.; LEÃO, J.; GÓMEZ, H. W. Extended exponential regression model: Diagnostics and application to mineral data. **Symmetry**, Multidisciplinary Digital Publishing Institute, v. 12, n. 12, p. 2042, 2020.

GRAY, L. A.; ALAVA, M. H. A command for fitting mixture regression models for bounded dependent variables using the beta distribution. **The Stata Journal**, SAGE Publications Sage CA: Los Angeles, CA, v. 18, n. 1, p. 51–75, 2018.

JAMES, B. R. **Probabilidade: um curso em nível intermediário**. [S.l.: s.n.], 1996.

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

JONES, J. M. **Red States Outnumber Blue for First Time in Gallup Tracking**. [S.l.], 2016. Acesso em: 23 de maio de 2022. Disponível em: <https://news.gallup.com/poll/188969/red-states-outnumber-blue-first-time-gallup-tracking.aspx>.

JÚNIOR, A. C. d. S.; ZEVIANI, W. M. Classificação de churn utilizando um modelo de regressão logística. 2020.

JÚNIOR, S. M.; VALADÃO, L. T.; VIEIRA, A. R. R.; MOURA, M. V. T. de. Análise de dados de vento para a região de botucatu-SP utilizando a distribuição beta. **Revista Brasileira de Agrometeorologia, Santa Maria**, v. 3, p. 129–132, 1995.

KANNEBLEY, S.; PRINCE, D. d. Restrição financeira e financiamento público à inovação no brasil: uma análise com base em microdados da pintec. **Nova Economia**, SciELO Brasil, v. 25, p. 553–574, 2015.

KERSTENETZKY, C. L. Redistribuição e desenvolvimento? a economia política do programa bolsa família. **Dados**, SciELO Brasil, v. 52, p. 53–83, 2009.

MELO, M. d. S.; LOOSE, L. H.; CARVALHO, J. B. d. Lomax regression model with varying precision: Formulation, estimation, diagnostics, and application. **Chilean Journal of Statistics (ChJS)**, v. 12, n. 2, 2021.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. [S.l.]: John Wiley & Sons, 2021.

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.

OLIVEIRA, J. F. A. C. de; CARNEIRO, M. E. F. As políticas neoliberais para a educação profissional: analisando o governo fernando henrique cardoso e luís inácio lula da silva. **Anais. III Seminário Nacional da Educação Profissional e Tecnológica. Minas Gerais: CEFETMG**, 2012.

ORTIZ, C. H.; URIBE, J. I.; GARCÍA, G. A. Informalidad y subempleo: un modelo probit bivariado aplicado al valle del cauca. **Sociedad y Economía**, Universidad del Valle, n. 13, p. 104–131, 2007.

PAULA, G. A. **Modelos de regressão: com apoio computacional**. [S.l.]: IME-USP São Paulo, 2004.

PAWITAN, Y. **In all likelihood: statistical modelling and inference using likelihood**. [S.l.]: Oxford University Press, 2001.

PEREIRA, G. H.; BOTTER, D. A.; SANDOVAL, M. C. The truncated inflated beta distribution. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 41, n. 5, p. 907–919, 2012.

PESCIM, R. R. **A distribuição beta generalizada semi-normal**. Tese (Doutorado) — Universidade de São Paulo, 2009.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2022. Disponível em: <https://www.R-project.org/>.

RODRIGUES, S. C. A. **Modelo de regressão linear e suas aplicações**. Tese (Doutorado) — Universidade da Beira Interior, 2012.

SILVA, E. R. F. da. **Modelo de regressão beta modal**. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Norte, 2020.

SOUZA, T. C. de. **Ensaios sobre modelos de regressão com dispersão variável**. Tese (Doutorado) — Universidade Federal de Pernambuco, 2011.

TOTA, A. P. Origens do bipartidarismo: uma tentativa de entender as eleições norte-americanas. **Novos estudos CEBRAP**, SciELO Brasil, p. 69–76, 2008.

TURKMAN, M. A. A.; SILVA, G. L. Modelos lineares generalizados–da teoria à prática. In: **VIII Congresso Anual da Sociedade Portuguesa de Estatística, Lisboa**. [S.l.: s.n.], 2000.

APÊNDICE A – Cálculos

Sabe-se pelo capítulo 2 que a função de distribuição acumulada para beta é dada por

$$F(x;\alpha,\beta) = \frac{B(x;\alpha,\beta)}{B(\alpha,\beta)} = \frac{\int_0^x t^{\alpha-1}(1-t)^{\beta-1}dt}{B(\alpha,\beta)},$$

na qual $\mathrm{F}(x;\alpha,\beta) = \mathrm{I}_x(\alpha,\beta)$. Sabendo disso, pode-se modelar variáveis que estão no intervalo $(-1,1)$, propondo uma transformação na distribuição beta. Para realizar esta, foi utilizado o método do Jacobiano, para mais esclarecimentos sobre o método, recomenda-se a leitura do livro James (1996). Com isso em mente, considerar-se-a a transformação $Y = 2X - 1$, que por sua vez tem como inversa a equação a seguir

$$X = \frac{Y+1}{2} = g^{-1}(y).$$

Assim, é necessário calcular a derivada da função inversa, que é de fácil obtenção, pois se tratar que uma função relativamente simples. Portanto temos que

$$\frac{d}{dy}g^{-1}(y) = \frac{1}{2} = J,$$

com isso tem-se que a densidade da transformação Y é dada por

$$f_y(y) = f_x(g^{-1}(y))|J| = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{y+1}{2}\right)^{\alpha-1}\left[1-\left(\frac{y+1}{2}\right)\right]^{\beta-1}\frac{1}{2}.$$

O próximo passo é realizar manipulações algébricas, cujo objetivo é o de organizar as quantidades evolvidas na equação, resultando em

$$f_y(y) = \frac{1}{2^{\alpha+\beta-1}}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(1+y)^{\alpha-1}(1-y)^{\beta-1}, \quad -1 < y < 1,$$

no que resulta na densidade da distribuição BM.

Como a transformação de Y é conhecida, podemos encontrar o valor esperado e a variância da nova distribuição. Sabe-se, pelas propriedades da esperança, que

$$E(Y) = 2E(X) - 1.$$

Porém a equação (2.3) representa a $\mathrm{E}(X)$ e por esse motivo, realizando a substituição e efetivando os cálculos, se tem como média da BM:

$$E(Y) = 2\frac{\alpha}{\alpha+\beta} - 1 = \frac{2\alpha-\alpha-\beta}{\alpha+\beta} = \frac{\alpha-\beta}{\alpha+\beta}.$$

De maneira análoga, pode-se dizer que a variância da BM, obtida com base na equação (2.4), é dada por:

$$Var(Y) = 4Var(X) = \frac{4\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Cálculos da reparametrização

A seguir está disposto o desenvolvimento para se obter a densidade da beta modular

$$\begin{aligned} f_y(y) &= \frac{1}{2^{\alpha+\beta-1}}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(1+y)^{\alpha-1}(1-y)^{\beta-1} \\ &= \frac{1}{2^{\frac{\phi}{2}(\mu+1)+\frac{\phi}{2}(1-\mu)-1}}\frac{\Gamma(\frac{\phi}{2}(\mu+1)+\frac{\phi}{2}(1-\mu))}{\Gamma(\frac{\phi}{2}(\mu+1))\Gamma(\frac{\phi}{2}(1-\mu))}(1+y)^{\frac{\phi}{2}(\mu+1)-1}(1-y)^{\frac{\phi}{2}(1-\mu)-1} \\ &= \frac{1}{2^{\phi-1}}\frac{\Gamma(\phi)}{\Gamma\left(\frac{\phi}{2}(\mu+1)\right)\Gamma\left(\frac{\phi}{2}(1-\mu)\right)}(1+y)^{\frac{\phi}{2}(\mu+1)-1}(1-y)^{\frac{\phi}{2}(1-\mu)-1}. \end{aligned}$$

Dados os novos parâmetros, se têm como média

$$E(Y) = \frac{\alpha-\beta}{\alpha+\beta} = \frac{\frac{\phi}{2}(\mu+1)-\frac{\phi}{2}(1-\mu)}{\frac{\phi}{2}(\mu+1)+\frac{\phi}{2}(1-\mu)} = \frac{\phi\mu}{\phi} = \mu,$$

e variância

$$\begin{aligned} Var(Y) &= \frac{4\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= \frac{4\left(\frac{\phi}{2}(\mu+1)\right)\left(\frac{\phi}{2}(1-\mu)\right)}{(\frac{\phi}{2}(\mu+1)+\frac{\phi}{2}(1-\mu))^2(\frac{\phi}{2}(\mu+1)+\frac{\phi}{2}(1-\mu)+1)} \\ &= \frac{1-\mu^2}{\phi^2(\phi+1)}. \end{aligned}$$

Cálculos Inferência sobre os Parâmetros

Como exposto com capítulo 2.3 a função de verossimilhança é obtida através do seguinte procedimento

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n}\left[\frac{1}{2^{\phi-1}}\frac{\Gamma(\phi)}{\Gamma\left(\frac{\phi}{2}(\mu_i+1)\right)\Gamma\left(\frac{\phi}{2}(1-\mu_i)\right)}(1+y_i)^{\frac{\phi}{2}(\mu_i+1)-1}(1-y_i)^{\frac{\phi}{2}(1-\mu_i)-1}\right],$$

logo, tem-se que a função de log-verossimilhança para a FDP da distribuição beta é dada por

$$\begin{aligned} &\ell(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta}|\mathbf{y})) = \\ &= \log\left[\left(\frac{1}{B\left(\frac{\phi}{2}(\mu_i+1),\frac{\phi}{2}(1-\mu)\right)}\right)^n\left(\frac{1}{2^{\phi-1}}\right)^n\prod_{i=1}^{n}\left((1+y_i)^{\frac{\phi}{2}(\mu_i+1)-1}(1-y_i)^{\frac{\phi}{2}(1-\mu_i)-1}\right)\right] \\ &= -n\log B\left(\frac{\phi}{2}(\mu_i+1),\frac{\phi}{2}(1-\mu_i)\right) - n(\phi-1)\log(2) + \left(\frac{\phi}{2}(\mu_i+1)-1\right)\sum_{i=1}^{n}\log(1+y_i) \\ &\quad + \left(\frac{\phi}{2}(1-\mu_i)-1\right)\sum_{i=1}^{n}\log(1-y_i). \end{aligned}$$

Cálculo para obtenção da Matriz de Informação de Fisher

A matriz de informação de Fisher é obtida tomando o valor esperado do negativo da matriz Hessiana, ou seja, $K(\boldsymbol{\theta}) = -E(J(\boldsymbol{\theta}))$. Dessa forma, temos que

$$\begin{aligned}
-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \beta_t \partial \beta_k}\right] &= -\sum_{i=1}^{n} E\left[\frac{\partial^2}{\partial \mu_i^2}\ell_i(\mu_i,\phi_i)\right]\left(\frac{d\mu_i}{d\eta_{1i}}\right)^2 x_{it}x_{ik} \\
&= \sum_{i=1}^{n} \frac{\phi_i^2}{4} E\left[\psi'\left(\frac{\phi_i}{2}(\mu_i+1)\right) + \psi'\left(\frac{\phi_i}{2}(1-\mu_i)\right)\right]\left(\frac{2}{g_1'(\mu_i)}\right)^2 x_{it}x_{ik} \\
&= \sum_{i=1}^{n} \frac{\phi_i^2}{4} \left[\psi'\left(\frac{\phi_i}{2}(\mu_i+1)\right) + \psi'\left(\frac{\phi_i}{2}(1-\mu_i)\right)\right]\left(\frac{2}{g_1'(\mu_i)}\right)^2 x_{it}x_{ik} \\
&= \sum_{i=1}^{n} q_i \left(\frac{2}{g_1'(\mu_i)}\right)^2 x_{it}x_{ik} \\
&= \mathbf{X}^\top \mathbf{Q}\mathbf{M}^2\mathbf{X},
\end{aligned}$$

em que $k = 0,1,\ldots,p$ e $t = 0,1,\ldots,p$,

$$q_i = \frac{\phi_i^2}{4}\left[\psi'\left(\frac{\phi_i}{2}(\mu_i+1)\right) + \psi'\left(\frac{\phi_i}{2}(1-\mu_i)\right)\right],$$

$\mathbf{Q} = diag(q_1,\ldots,q_n)^\top$ e $\psi'(.)$ é chamada de função trigama, no qual tem-se que $\psi'(c) = \frac{d}{dc}\psi(c)$ com $c > 0$.

$$-E\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \beta_t \partial \gamma_j}\right] = -\sum_{i=1}^{n} E\left[\frac{\partial^2 \ell_i(\mu_i,\phi_i)}{\partial \mu_i \partial \phi_i}\right]\frac{d\phi_i}{d\eta_{2i}}\frac{d\mu_i}{d\eta_{1i}} z_{ij}x_{it}$$

$$\begin{aligned}
-\sum_{i=1}^{n} E\left[\frac{\partial^2 \ell_i(\mu_i,\phi_i)}{\partial \mu_i \partial \phi_i}\right] &= \sum_{i=1}^{n} E\left\{\frac{\phi_i}{2}\left[-\frac{(\mu_i+1)}{2}\psi'\left(\frac{\phi_i}{2}(\mu_i+1)\right) + \frac{(1-\mu_i)}{2}\psi'\left(\frac{\phi_i}{2}(1-\mu_i)\right)\right]\right. \\
&\quad \left. + \frac{1}{2}\left[-\psi\left(\frac{\phi_i}{2}(\mu_i+1)\right) + \psi\left(\frac{\phi_i}{2}(1-\mu_i)\right) + \log\left(\frac{1+Y_i}{1-Y_i}\right)\right]\right\} \\
&= \sum_{i=1}^{n} -\left\{\frac{\phi_i}{2}\left[-\frac{(\mu_i+1)}{2}\psi'\left(\frac{\phi_i}{2}(\mu_i+1)\right) + \frac{(1-\mu_i)}{2}\psi'\left(\frac{\phi_i}{2}(1-\mu_i)\right)\right]\right. \\
&\quad \left. + \frac{1}{2}\left[-\psi\left(\frac{\phi_i}{2}(\mu_i+1)\right) + \psi\left(\frac{\phi_i}{2}(1-\mu_i)\right) + E\left[\log\left(\frac{1+Y_i}{1-Y_i}\right)\right]\right]\right\} \\
&= \sum_{i=1}^{n} -\left\{\frac{\phi_i}{2}\left[-\frac{(\mu_i+1)}{2}\psi'\left(\frac{\phi_i}{2}(\mu_i+1)\right) + \frac{(1-\mu_i)}{2}\psi'\left(\frac{\phi_i}{2}(1-\mu_i)\right)\right]\right. \\
&\quad + \frac{1}{2}\left[-\psi\left(\frac{\phi_i}{2}(\mu_i+1)\right) + \psi\left(\frac{\phi_i}{2}(1-\mu_i)\right) + \psi\left(\frac{\phi_i}{2}(\mu_i+1)\right)\right. \\
&\quad \left.\left. -\psi\left(\frac{\phi_i}{2}(1-\mu_i)\right)\right]\right\} \\
&= \sum_{i=1}^{n} -\frac{\phi_i}{2}\left[-\frac{(\mu_i+1)}{2}\psi'\left(\frac{\phi_i}{2}(\mu_i+1)\right) + \frac{(1-\mu_i)}{2}\psi'\left(\frac{\phi_i}{2}(1-\mu_i)\right)\right],
\end{aligned}$$

logo,

$$-E\left[\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\beta_t\partial\gamma_j}\right] = \sum_{i=1}^{n} d_i \frac{2}{g_1'(\mu^*{}_i)}\frac{1}{g_2'(\phi_i)}x_{ij}z_{it}$$
$$= \mathbf{X}^\top\mathbf{M}\mathscr{M}\mathbf{D},$$

em que

$$d_i = -\left\{\frac{\phi_i}{2}\left[-\frac{(\mu_i+1)}{2}\psi\left(\frac{\phi_i}{2}(\mu_i+1)\right)+\frac{(1-\mu_i)}{2}\psi\left(\frac{\phi_i}{2}(1-\mu_i)\right)\right]\right.$$
$$+\frac{1}{2}\left[-\psi'\left(\frac{\phi_i}{2}(\mu_i+1)\right)+\psi'\left(\frac{\phi_i}{2}(1-\mu_i)\right)+\psi\left(\frac{\phi_i}{2}(\mu_i+1)\right)\right.$$
$$\left.\left.-\psi\left(\frac{\phi_i}{2}(1-\mu_i)\right)\right]\right\},$$

e $\mathbf{D} = diag(d_1,\ldots,d_n)^\top$. Além disso, tem-se que

$$-E\left[\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\gamma_i\partial\gamma_j}\right] = -\sum_{i=1}^{n} E\left[\frac{\partial^2}{\partial\phi_i^2}\ell_i(\mu_i,\phi_i)\right]\left(\frac{d\phi_i}{d\eta_{2i}}\right)^2 z_{ij}z_{it}$$
$$= -\sum_{i=1}^{n} E\left[\psi'(\phi_i)-\frac{(\mu_i+1)^2}{4}\psi'\left(\frac{\phi_i}{2}(\mu_i+1)\right)\right.$$
$$\left.-\frac{(1-\mu_i)^2}{4}\psi'\left(\frac{\phi_i}{2}(1-\mu_i)\right)\right]\left(\frac{1}{g_2'(\phi_n)}\right)^2 z_{ij}z_{it}$$
$$= \sum_{i=1}^{n} s_i\left(\frac{1}{g_2'(\phi_n)}\right)^2 z_{ij}z_{it}$$
$$= \mathbf{Z}^\top\mathbf{S}\mathscr{M}^2\mathbf{Z},$$

em que

$$s_i = E\left[\psi'(\phi_i)-\frac{(\mu_i+1)^2}{4}\psi'\left(\frac{\phi_i}{2}(\mu_i+1)\right)\right.$$
$$\left.-\frac{(1-\mu_i)^2}{4}\psi'\left(\frac{\phi_i}{2}(1-\mu_i)\right)\right],$$

e $\mathbf{S} = diag(s_1,\ldots,s_n)^\top$.

APÊNDICE B – Simulação

Tabelas - Cenários de simulação para a precisão fixa e três tipos de funções de ligação para a média.

Tabela 14 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 4: $g_1(\boldsymbol{\mu}^*) = \log(\frac{\boldsymbol{\mu}^*}{1-\boldsymbol{\mu}^*})$ **e** $n = 50, 100, 500$**.**

n	Parâmetro	θ	$\widehat{\theta}_{mc}$	$VR(\widehat{\theta}_{mc})$	$DP(\widehat{\theta}_{mc})$
50	β_0	0,5	0,5005	0,0010	0,1101
	β_1	1,0	1,0014	0,0014	0,1235
	β_2	-1,0	-1,0019	0,0019	0,1246
	β_3	1,5	1,5013	0,0009	0,1268
	ϕ	100,0	113,5305	0,1353	24,7620
100	β_0	0,5	0,4995	-0,0010	0,0759
	β_1	1,0	1,0017	0,0017	0,0849
	β_2	-1,0	-1,0003	0,0003	0,0865
	β_3	1,5	1,5008	0,0005	0,0876
	ϕ	100,0	106,5327	0,0653	15,7221
500	β_0	0,5	0,5004	0,0008	0,0333
	β_1	1,0	1,0000	0,0000	0,0381
	β_2	-1,0	-1,0006	0,0006	0,0376
	β_3	1,5	1,5001	0,0000	0,0382
	ϕ	100,0	101,2534	0,0125	6,4542

Tabela 15 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 5 : $g_1(\boldsymbol{\mu}^*) = \Phi^{-1}(\boldsymbol{\mu}^*)$ **e** $n = 50, 100, 500$**.**

n	Parâmetro	θ	$\widehat{\theta}_{mc}$	$VR(\widehat{\theta}_{mc})$	$DP(\widehat{\theta}_{mc})$
50	β_0	0,5	0,5001	0,0002	0,0746
	β_1	1,0	1,0032	0,0032	0,0862
	β_2	-1,0	-1,0033	0,0033	0,0870
	β_3	1,5	1,5046	0,0030	0,0897
	ϕ	100,0	113,7158	0,1372	24,8941
100	β_0	0,5	0,5000	-0,0001	0,0509
	β_1	1,0	1,0015	0,0015	0,0605
	β_2	-1,0	-1,0011	0,0011	0,0595
	β_3	1,5	1,5019	0,0013	0,0620
	ϕ	100,0	106,5378	0,0654	15,.7558
500	β_0	0,5	0,5001	0,0001	0,0221
	β_1	1,0	1,0001	0,0001	0,0262
	β_2	-1,0	-1,0006	0,0006	0,0258
	β_3	1,5	1,5008	0,0005	0,0269
	ϕ	100,0	101,2108	0,0121	6,5227

Tabela 16 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 6 : $g_1(\boldsymbol{\mu}^*) = \log(-\log(1-\boldsymbol{\mu}^*))$ **e** $n = 50, 100, 500$**.**

n	Parâmetro	θ	$\widehat{\theta}_{mc}$	$VR(\widehat{\theta}_{mc})$	$DP(\widehat{\theta}_{mc})$
50	β_0	0,5	0,4997	-0,0006	0,0604
	β_1	1,0	1,0022	0,0022	0,1380
	β_2	-1,0	-1,0031	0,0031	0,1345
	β_3	1,5	1,5061	0,0040	0,1362
	ϕ	100,0	113,7080	0,1371	25,3384
100	β_0	0,5	0,5004	0,0008	0,0423
	β_1	1,0	1,0007	0,0007	0,0936
	β_2	-1,0	-1,0014	0,0014	0,0950
	β_3	1,5	1,5007	0,0005	0,0956
	ϕ	100,0	106,3043	0,0630	15,7212
500	β_0	0,5	0,5002	0,0004	0,0184
	β_1	1,0	1,0008	0,0008	0,0413
	β_2	-1,0	-1,0007	0,0007	0,0411
	β_3	1,5	1,5001	0,0000	0,0415
	ϕ	100,0	101,2416	0,0124	6,5371

Tabela 17 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 7 : $g_1(\boldsymbol{\mu}^*) = \log(\frac{\boldsymbol{\mu}^*}{1-\boldsymbol{\mu}^*})$ **e** $n = 50, 100, 500$**.**

n	Parâmetro	θ	$\widehat{\theta}_{mc}$	$VR(\widehat{\theta}_{mc})$	$DP(\widehat{\theta}_{mc})$
50	β_0	0,5	0,4992	-0,0017	0,1114
	β_1	1,0	1,0027	0,0027	0,1492
	β_2	-1,0	-1,0017	0,0017	0,1503
	ϕ	50,0	55,5165	0,1103	11,9171
100	β_0	0,5	0,4998	-0,0004	0,0778
	β_1	1,0	1,0019	0,0019	0,1038
	β_2	-1,0	-1,0013	0,0013	0,1031
	ϕ	50,0	52,6984	0,0540	7,7036
500	β_0	0,5	0,4997	-0,0007	0,0345
	β_1	1,0	1,0006	0,0006	0,0464
	β_2	-1,0	-0,9999	-0,0001	0,0456
	ϕ	50,0	50,5263	0,0105	3,2147

Tabela 18 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 8 : $g_1(\boldsymbol{\mu}^*) = \Phi^{-1}(\boldsymbol{\mu}^*)$ **e** $n = 50, 100, 500$**.**

n	Parâmetro	θ	$\widehat{\theta}_{mc}$	$VR(\widehat{\theta}_{mc})$	$DP(\widehat{\theta}_{mc})$
50	β_0	0,5	0,5004	0,0008	0,0719
	β_1	1,0	1,0008	0,0008	0,0962
	β_2	-1,0	-1,0005	0,0005	0,0949
	ϕ	50,0	55,4127	0,1083	11,7535
100	β_0	0,5	0,5006	0,0011	0,0498
	β_1	1,0	0,9999	-0,0001	0,0671
	β_2	-1,0	-1,0008	0,0008	0,0670
	ϕ	50,0	52,5494	0,0510	7,6678
500	β_0	0,5	0,5004	0,0009	0,0218
	β_1	1,0	1,0000	-0,0000	0,0295
	β_2	-1,0	-1,0009	0,0009	0,0292
	ϕ	50,0	50,5169	0,0103	3,1931

Tabela 19 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 9 : $g_1(\boldsymbol{\mu}^*) = \log(-\log(1-\boldsymbol{\mu}^*))$ **e** $n = 50, 100, 500$**.**

n	Parâmetro	θ	$\widehat{\theta}_{mc}$	$VR(\widehat{\theta}_{mc})$	$DP(\widehat{\theta}_{mc})$
50	β_0	0,5	0,5011	0,0021	0,0668
	β_1	1,0	0,9989	-0,0011	0,1765
	β_2	-1,0	-1,0007	0,0007	0,1772
	ϕ	50,0	55,5504	0,1110	11,9375
100	β_0	0,5	0,5004	0,0008	0,0465
	β_1	1,0	1,0011	0,0011	0,1234
	β_2	-1,0	-1,0020	0,0020	0,1221
	ϕ	50,0	52,6926	0,0539	7,6291
500	β_0	0,5	0,5001	0,0002	0,0205
	β_1	1,0	1,0003	0,0003	0,0540
	β_2	-1,0	-1,0003	0,0003	0,0537
	ϕ	50,0	50,4691	0,0094	3,2220

Tabela 20 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 10: $g_1(\boldsymbol{\mu}^*) = \log(\frac{\boldsymbol{\mu}^*}{1-\boldsymbol{\mu}^*})$ **e** $n = 50, 100, 500$**.**

n	Parâmetro	θ	$\widehat{\theta}_{mc}$	$VR(\widehat{\theta}_{mc})$	$DP(\widehat{\theta}_{mc})$
50	β_0	0,5	0,5006	0,0013	0,0799
	β_1	1,0	1,0002	0,0002	0,1057
	β_2	-1,0	-1,0013	0,0013	0,1075
	ϕ	100,0	111,0706	0,1107	23,7142
100	β_0	0,5	0,5007	0,0013	0,0559
	β_1	1,0	0,9991	-0,0009	0,0725
	β_2	-1,0	-1,0006	0,0006	0,0736
	ϕ	100,0	105,3506	0,0535	15,3568
500	β_0	0,5	0,4999	-0,0001	0,0247
	β_1	1,0	1,0003	0,0003	0,0327
	β_2	-1,0	-1,0002	0,0002	0,0330
	ϕ	100,0	101,0045	0,0100	6,4136

Tabela 21 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 11 : $g_1(\boldsymbol{\mu}^*) = \Phi^{-1}(\boldsymbol{\mu}^*)$ **e** $n = 50, 100, 500$**.**

n	Parâmetro	θ	$\widehat{\theta}_{mc}$	$VR(\widehat{\theta}_{mc})$	$DP(\widehat{\theta}_{mc})$
50	β_0	0,5	0,5008	0,0015	0,0506
	β_1	1,0	1,0006	0,0006	0,0682
	β_2	-1,0	-1,0012	0,0012	0,0681
	ϕ	100,0	111,0659	0,1107	23,4802
100	β_0	0,5	0,5003	0,0007	0,0351
	β_1	1,0	0,9996	-0,0004	0,0477
	β_2	-1,0	-1,0001	0,0001	0,0472
	ϕ	100,0	105,1743	0,0517	15,2851
500	β_0	0,5	0,5001	0,0002	0,0156
	β_1	1,0	1,0001	0,0001	0,0209
	β_2	-1,0	-1,0002	0,0002	0,0209
	ϕ	100,0	101,1085	0,0111	6,3664

Tabela 22 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 12 : $g_1(\boldsymbol{\mu}^*) = \log(-\log(1-\boldsymbol{\mu}^*))$ **e** $n = 50, 100, 500$**.**

n	Parâmetro	θ	$\widehat{\theta}_{mc}$	$VR(\widehat{\theta}_{mc})$	$DP(\widehat{\theta}_{mc})$
50	β_0	0,5	0,5004	0,0008	0,0479
	β_1	1,0	1,0014	0,0014	0,1260
	β_2	-1,0	-1,0013	0,0013	0,1254
	ϕ	100,0	111,2590	0,1126	24,0859
100	β_0	0,5	0,4998	-0,0005	0,0327
	β_1	1,0	1,0007	0,0007	0,0877
	β_2	-1,0	-0,9993	-0,0007	0,0865
	ϕ	100,0	105,2355	0,0524	15,4550
500	β_0	0,5	0,5001	0,0003	0,0147
	β_1	1,0	1,0000	0,0000	0,0391
	β_2	-1,0	-1,0002	0,0002	0,0388
	ϕ	100,0	101,0142	0,0101	6,3166

Tabelas - Cenários de simulação para a precisão variável e três tipos de funções de ligação para a média.

Tabela 23 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 4: $g_2(\boldsymbol{\mu}^*) = \log(\frac{\boldsymbol{\mu}^*}{1-\boldsymbol{\mu}^*})$**,** $g(\phi_i) = \log(\phi_i)$ **e** $n = 50, 100, 500$**.**

n	Parâmetro	θ	$\widehat{\theta}_{mc}$	$VR(\widehat{\theta}_{mc})$	$DP(\widehat{\theta}_{mc})$
50	β_0	0,5	0,5044	0,0088	0,4067
	β_1	1,0	1,0293	0,0293	0,5283
	β_2	-1,0	-1,0173	0,0173	0,5303
	γ_0	0,5	0,6568	0,3135	0,6345
	γ_1	1,0	1,1123	0,1123	0,8128
100	β_0	0,5	0,5027	0,0053	0,2788
	β_1	1,0	1,0066	0,0066	0,3639
	β_2	-1,0	-1,0048	0,0048	0,3655
	γ_0	0,5	0,5493	0,0985	0,2928
	γ_1	1,0	1,0222	0,0222	0,4607
500	β_0	0,5	0,4979	-0,0041	0,1192
	β_1	1,0	1,0045	0,0045	0,1553
	β_2	-1,0	-0,9986	-0,0014	0,1572
	γ_0	0,5	0,5078	0,0156	0,1073
	γ_1	1,0	1,0037	0,0037	0,1884

Tabela 24 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 5: $g_2(\boldsymbol{\mu}^*) = \Phi^{-1}(\boldsymbol{\mu}^*)$**,** $g(\phi_i) = \log(\phi_i)$ **e** $n = 50, 100, 500$**.**

n	Parâmetro	θ	$\widehat{\theta}_{mc}$	$VR(\widehat{\theta}_{mc})$	$DP(\widehat{\theta}_{mc})$
50	β_0	0,5	0,4979	-0,0043	0,2443
	β_1	1,0	1,0119	0,0119	0,3184
	β_2	-1,0	-0,9978	-0,0022	0,3172
	γ_0	0,5	0,6059	0,2119	0,4425
	γ_1	1,0	1,0387	0,0387	0,6896
100	β_0	0,5	0,5015	0,0031	0,1671
	β_1	1,0	1,0008	0,0008	0,2173
	β_2	-1,0	-0,9995	-0,0005	0,2185
	γ_0	0,5	0,5444	0,0888	0,2531
	γ_1	1,0	1,0150	0,0150	0,4367
500	β_0	0,5	0,4995	-0,0009	0,0733
	β_1	1,0	0,9997	-0,0003	0,0941
	β_2	-1,0	-0,9977	-0,0023	0,0952
	γ_0	0,5	0,5083	0,0165	0,1067
	γ_1	1,0	1,0039	0,0039	0,1840

Tabela 25 – Resultados da simulação de Monte Carlo para os estimadores do modelo BM - Cenário 6 : $g_2(\boldsymbol{\mu}^*) = \log(-\log(1-\boldsymbol{\mu}^*))$**,** $g(\phi_i) = \log(\phi_i)$ **e** $n = 50, 100, 500$**.**

n	Parâmetro	θ	$\widehat{\theta}_{mc}$	$VR(\widehat{\theta}_{mc})$	$DP(\widehat{\theta}_{mc})$
50	β_0	0,5	0,5025	0,0050	0,2033
	β_1	1,0	0,9678	-0,0322	0,2457
	β_2	-1,0	-0,9679	-0,0321	0,2507
	γ_0	0,5	0,6366	0,2732	0,5377
	γ_1	1,0	1,0343	0,0343	0,7267
100	β_0	0,5	0,4991	-0,0018	0,1377
	β_1	1,0	0,9573	-0,0427	0,1666
	β_2	-1,0	-0,9602	-0,0398	0,1682
	γ_0	0,5	0,5596	0,1192	0,2663
	γ_1	1,0	0,9767	-0,0233	0,4298
500	β_0	0,5	0,4947	-0,0105	0,0602
	β_1	1,0	0,9587	-0,0413	0,0721
	β_2	-1,0	-0,9574	-0,0426	0,0727
	γ_0	0,5	0,5243	0,0485	0,1076
	γ_1	1,0	0,9616	-0,0384	0,1779