UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA

Vitor Hugo Floriano Pereira

Inferência via bootstrap no modelo de regressão Gama Unitária: desenvolvimento e aplicação do pacote `UnitGammaReg` em R

Goiânia
2025

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): Vitor Hugo Floriano Peixoto.

Título do trabalho: Inferência via bootstrap no modelo de regressão Gama Unitária: desenvolvimento e aplicação do pacote UnitGammaReg em R.

2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;

- Submissão de artigo em revista científica;

- Publicação como capítulo de livro.

Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Tatiane Ferreira Do Nascimento Melo Da Silva** , **Professor do Magistério Superior**, em 02/12/2025, às 20:47, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Vitor Hugo Floriano Pereira**, **Discente**, em 04/12/2025, às 10:29, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA

Vitor Hugo Floriano Pereira

Inferência via bootstrap no modelo de regressão Gama Unitária: desenvolvimento e aplicação do pacote `UnitGammaReg` em R

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Estatística da Universidade Federal de Goiás para aprovação no componente curricular TCC, como parte das exigências para a obtenção do título de bacharel em Estatística. **Orientadora**: Profa. Dra. Tatiane F. N. Melo da Silva

Goiânia

2025

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Pereira, Vitor Hugo Floriano
Inferência via bootstrap no modelo de regressão Gama Unitária [manuscrito] : desenvolvimento e aplicação do pacote UnitGammaReg em R / Vitor Hugo Floriano Pereira. - 2025.
34 f.

Orientador: Prof. Tatiani F. N. Melo da Silva.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Instituto de Matemática e Estatística (IME), Estatística, Goiânia, 2025.
Bibliografia. Anexos. Apêndice.
Inclui abreviaturas, símbolos, tabelas, lista de tabelas.

1. Bootstrap. 2. Índice de desenvolvimento humano municipal. 3. Modelo de regressão beta. 4. Modelo de regressão gama unitária. 5. Teste de hipótese. I. Silva, Tatiani F. N. Melo da, orient. II. Título.

CDU 519.22

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

Aos vinte e oito dias do mês de novembro do ano de 2025 iniciou-se a sessão pública de defesa do Trabalho de Conclusão de Curso (TCC) intitulado “Programação Estatística em R: Desenvolvimento de Scripts para Inferência via Bootstrap no Modelo de Regressão Gama Unitária”, de autoria de Vitor Hugo Floriano Pereira, do curso de Estatística, do Instituto de Matemática e Estatística da UFG. Os trabalhos foram instalados pelo Profa. Dra. Tatiane Ferreira do Nascimento Melo da Silva com a participação dos demais membros da Banca Examinadora: Amanda Buosi Gazon Milani (IME/UFG) e Renata Mendonça Rodrigues Vasconcelos (IME/UFG). Após a apresentação, a banca examinadora realizou a arguição do estudante. Registrou-se a solicitação de alteração do título do trabalho. Após deliberação, a orientadora definiu como novo título: "Inferência via bootstrap no modelo de regressão Gama Unitária: desenvolvimento e aplicação do pacote UnitGammaReg em R". Posteriormente, de forma reservada, a Banca Examinadora atribuiu a nota final de 9,0, tendo sido o TCC considerado aprovado.

Proclamados os resultados, os trabalhos foram encerrados e, para constar, lavrou-se a presente ata que segue assinada pelos Membros da Banca Examinadora.

Documento assinado eletronicamente por **Amanda Buosi Gazon Milani**, **Professor do Magistério Superior**, em 02/12/2025, às 17:04, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Renata Mendonca Rodrigues Vasconcelos**, **Professor do Magistério Superior**, em 02/12/2025, às 18:15, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Tatiane Ferreira Do Nascimento Melo Da Silva** , **Professor do Magistério Superior**, em 02/12/2025, às 20:41, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5807999** e o código CRC **4525FF20**.

Referência: Processo nº 23070.060780/2025-85 SEI nº 5807999

Agradecimentos

Obrigado Professora Tatiane por todo o apoio e suporte neste trabalho e em nossa iniciação cientifica, sem a senhora nada disso seria possível, talvez eu nem estivesse escrevendo essa seção se não fosse pela senhora, muito obrigado. Obrigado a todos os meus Amigos, saibam que vocês tornaram cada momento dentro do campus da universidade, tive momentos inesquecíveis que sempre irei guardar com muito carinho. Obrigado pai, o caminho foi difícil, mas mesmo distante o senhor sempre me apoiou e me encorajou a seguir em frente, e vamos fazer dar certo nosso negócio e continuar subindo, obrigado por tudo, sei que não foi barato, mas, eu fiz valer a pena . Obrigado a você mãe, mesmo sendo dura às vezes nunca saiu do meu lado e sempre desejou meu bem, e espero poder dizer a você "Mãe eu consegui", e poder realizar mais um sonho da senhora, e você ter dois filhos graduados. E por fim, obrigado a você Arthur meu irmão, você é como um herói para mim, é a pessoa que eu mais admiro e nunca vou esquecer de que você foi a primeira pessoa a dizer que se orgulhava de mim, e essa mera palavra mudou tudo para mim e deu um sentido para minha vida que me fez chegar até aqui.

Resumo

O modelo de regressão gama unitária é uma alternativa ao modelo de regressão beta, que é usado na modelagem de dados restritos ao intervalo unitário (0,1). Existem várias situações práticas em que estes modelos podem ser usados, a saber: na modelagem de taxas de juros; taxas de descontos; proporção de indivíduos que sobrevivem a um determinado período de tempo, com certa doença; taxa de incidência de doenças respiratórias em uma determinada população; proporções de respostas em uma pesquisa; taxas de aprovação ou reprovação dos alunos em exames; taxa de vitória de times em esportes, entre outras. Neste trabalho, usamos um conjunto de dados reais, referente ao Índice de Desenvolvimento Humano Municipal (IDHM) dos estados brasileiros no ano de 2021, adivindos da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) para ilustrar a aplicabilidade do modelo de regressão gama unitária com inferência via *bootstrap*, propusemos um pacote em linguagem R que permita aos usuários utilizarem essa metodologia, além disso permitir ao usuário a realização de teste de hipóteses para os coeficientes do modelo.

Palavras-chave: Modelo de regressão beta. Modelo de regressão gama unitária. *Bootstrap*. Teste de hipóteses. Índice de Desenvolvimento Humano Municipal.

Abstract

The unit gamma regression model is an alternative to the beta regression model, which is used for modeling data restricted to the unit interval (0,1). There are several practical situations in which these models can be applied, such as: modeling interest rates; discount rates; the proportion of individuals who survive a given period of time with a certain disease; the incidence rate of respiratory diseases in a given population; proportions of responses in a survey; student pass or fail rates in exams; teams' win rates in sports, among others. In this study, we use a real dataset referring to the Municipal Human Development Index (IDHM) of Brazilian states in the year 2021, obtained from the Continuous National Household Sample Survey (PNAD Continua), to illustrate the applicability of the unit gamma regression model with bootstrap-based inference. We also propose an R package that allows users to implement this methodology and perform hypothesis testing for the model coefficients.

Keywords: Beta regression model. Unit gamma regression model. Bootstrap. Hypothesis testing. Municipal Human Development Index.

Lista de tabelas

Lista de abreviaturas e siglas

PNAD	Pesquisa Nacional por Amostra de Domicílio
f.d.p	Função densidade de probabilidade
IDMH	Indice de Desenvolvimento Humano Municipal
EMV	Estimador de Máxima verossimilhança
MV	Máxima verossimilhança.
LR	Estatistica do teste de Razão de Verossimilhança (*Likelihood Ratio*)
W	Estatística do teste de Wald
G	Estatística do teste de Gradiente
S	Estatística do teste de escore (*score*)

Lista de símbolos

μ	Média da distribuição
ϕ	Parâmetro de dispersão
β	Coeficientes do modelo
K	Matriz de Informação de Fisher
U	Vetor Escore
θ	Parâmetros da distribuição
α	Parâmetro de escala da distribuição
ℓ	Função de log-verossimilhança
Φ	Função de densidade acumulada da distribuição normal padrão
$\Re$	Conjunto dos numeros reais

Sumário

Introdução

Variáveis contínuas restritas ao intervalo $(0,1)$ são frequentemente observadas em diversas áreas do conhecimento, como saúde, economia e ciências ambientais. Para esse tipo de dado, o modelo de regressão beta (FERRARI; CRIBARI-NETO, 2004) consolidou-se como uma das principais ferramentas estatísticas, sendo amplamente estudado e aplicado em diferentes contextos. Mais recentemente, o modelo de regressão gama unitária (MOUSA; EL-SHEIKH; ABDEL-FATTAH, 2016) tem se destacado como alternativa, recebendo importantes contribuições na literatura.

Um desafio frequente nesses modelos ocorre quando o tamanho amostral é pequeno ou mesmo de tamanho moderado, pois os estimadores de máxima verossimilhança apresentam vieses e os testes clássicos, como razão de verossimilhança, Wald, Escore e Gradiente, tornam-se pouco precisos. Para contornar esse problema, Barroso (2022) propôs correções de viés e versões modificadas desses testes, utilizando o método de reamostragem *bootstrap*. Contudo, tais procedimentos ainda não estão disponíveis de forma implementada e acessível no software R Core Team (2023), amplamente utilizado pela comunidade estatística e científica.

Assim, o presente trabalho tem como objetivo implementar essas metodologias no ambiente R Core Team (2023), por meio do desenvolvimento de um pacote específico a ser disponibilizado no GitHub. Além disso, será apresentada uma aplicação a dados reais, ilustrando a utilidade das correções e do pacote desenvolvido.

Este trabalho está organizado da seguinte forma[1]: o Capítulo 1 apresenta a revisão bibliográfica dos principais conceitos relacionados aos modelos de regressão e às técnicas de correção de viés e estatísticas de testes. O Capítulo 2 descreve a metodologia adotada para o desenvolvimento e implementação das propostas. No Capítulo 3 são expostos os resultados referentes ao pacote desenvolvido e aplicação a dados reais. Por fim, o Capítulo 4 traz as conclusões e sugestões para trabalhos futuros.

[1] Parte da revisão textual deste trabalho foi auxiliada pelo assistente ChatGPT (OpenAI, 2025), sem interferir no conteúdo técnico ou científico.

1 Revisão Bibliográfica

O modelo de regressão beta foi proposto por Ferrari e Cribari-Neto (2004). Este modelo de regressão é utilizado em situações onde a variável de interesse tem natureza contínua e está restrita a um intervalo (0,1). Existem vários trabalhos na literatura onde o modelo beta é utilizado, como por exemplo, Espinheira, Ferrari e Cribari-Neto (2008) que propuseram medidas de diagnósticos, como a distância de Cook. Zeileis e Cribari-Neto (2009) desenvolveram um pacote no software R Core Team (2023), denominado *betareg*. Neste pacote é possível estimar os parâmetros do modelo, entre outras funções. Ospina e Ferrari (2012) generalizaram o modelo de regressão beta tradicional definido no (0,1) para uma classe mais geral, onde os dados podem estar contidos no intervalo fechado [0,1]. Chien (2013) propôs medidas de diagnóstico alternativas para identificação de outliers nestes modelos. Carrasco, Ferrari e Arellano-Valle (2014) propõem o modelo de regressão beta com erros nas variáveis. Fabrizi, Ferrante e Trivisano (2016) utilizaram um modelo de regressão beta Bayesiano para estimar parâmetros que descrevem a pobreza e desigualdade social. Haitao *et al.* (2018) propõem um modelo de regressão beta modificado para dados composicionais em microbioma. Espinheira *et al.* (2019) desenvolveram critérios de seleção para o modelo beta. Petterle *et al.* (2021) estudaram um modelo de regressão multivariada quase-beta, levando em consideração a correlação entre as componentes da variável de resposta. Geissinger *et al.* (2022) estudaram uma aplicação do modelo de regressão beta nas ciências naturais. Koç e Dünder (2023) propuseram o estimador Kibria-Lukman (KL) e sua versão Jackknife para reduzir os efeitos da multicolinearidade no modelo de regressão beta.

Na literatura, existem vários modelos alternativos ao beta, cujo objetivo é estudar dados restritos ao intervalo unitário (0,1). Um destes modelos é o modelo de regressão Gama Unitária, proposto por Mousa e Abdel-Fattah (2016), baseado na distribuiçao gama unitária (GRASSIA, 1977). Nos últimos anos, vários autores têm estudado este modelo, como por exemplo, Guedes, Cribari-Neto e Espinheira (2020) que propuseram testes da razão de verossimilhanças modificados para o modelo de regressão Gama Unitária. Rocha, Espinheira e Cribari-Neto (2021) obtiveram medidas de influência local e resíduos para este modelo. Barroso (2022) propôs simulações de Monte Carlo, via *bootstrap*, que resolvem problemas relacionados à inferência estatística, quando o tamanho da amostra é pequeno ou mesmo moderado, no modelo de regressão gama unitário. Freitas *et al.* (2023) propuseram um modelo de regressão Gama Unitária, adequado para modelar dados restritos ao intervalo unitária (0,1), considerando medidas repetidas, ou seja, mais de uma observação por unidade experimental. Pachenco (2023) estudou os efeitos da especificação incorreta da função de ligação no modelo de regressão gama unitária.

Os estimadores de máxima verossimilhança (EMVs), em modelos estatísticos, podem apresentar vieses em amostras pequenas ou moderadas, geralmente de ordem n^{-1}, onde n é o tamanho da amostra. Para reduzir esse problema, utilizam-se correções de viés, como o método

de Cox e Snell (1968), ou, de forma mais prática, o *bootstrap* (EFRON, 1979) .

Outro desafio em amostras pequenas é a baixa precisão das aproximações assintóticas de testes clássicos, como razão de verossimilhança, Wald, Escore e Gradiente. Métodos de refinamento, como os de Bartlett (1937), BARNDORFF-NIELSEN (1991) e Skovgaard (2001), podem ser usados, mas, na maioria das vezes, precisam de cálculos complexos, o que também motiva o uso do *bootstrap*.

Barroso (2022) apresentou estudos de simulações no modelo de regressão Gama Unitária, adequado para variáveis definidas no intervalo $(0,1)$, especialmente em situações de pequeno tamanho amostral. Nesses estudos, foram propostas correções de viés para os estimadores de máxima verossimilhança (EMVs) e versões modificadas de testes clássicos, todas baseadas no método *bootstrap*. No entanto, o autor não disponibilizou uma implementação prática dessas metodologias no software R Core Team (2023), amplamente utilizado pela comunidade científica por ser livre e de código aberto.

Diante disso, este trabalho tem como objetivo implementar as metodologias propostas no ambiente R, por meio do desenvolvimento de um pacote específico, disponibilizado na plataforma GitHub e, posteriormente, em versões estáveis no CRAN. Além do desenvolvimento e divulgação do pacote, será apresentada uma aplicação a dados reais, na qual serão utilizadas as estimativas e os testes corrigidos, mais adequados em cenários de pequeno tamanho amostral.

2 Metodologia

2.1 Modelo de regressão beta

Ferrari e Cribari-Neto (2004) propuseram o modelo de regressão beta, baseado na distribuição beta, que é definida no intervalo unitário (0,1). Primeiramente, será definida a função densidade de probabilidade (f.d.p.) da distribuição beta, ou seja, suponha que a variável aleatória Y tem distribuição de probabilidade beta com parâmetros $p, q > 0$, denotada por $Y \sim \mathcal{B}(p, q)$, temos

$$f(y; p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}; \;\; 0 < y < 1, \tag{2.1}$$

onde $\Gamma(.)$ é a função gamma. O valor esperado e a variância Y são dados, respectivamente, por:

$$E(Y) = \frac{p}{p+q} \;\text{ e }\; Var(Y) = \frac{pq}{(p+q)^2(p+q+1)}.$$

A parametrização proposta por Ferrari e Cribari-Neto (2004) é definida pela média (μ) e por um parâmetro de dispersão (ϕ), com

$$\mu = E(Y) = \frac{p}{p+q} \quad \text{e} \quad \phi = p + q. \tag{2.2}$$

Isolando p e q em (2.2), temos

$$p = \mu\phi \quad \text{e} \quad q = (1-\mu)\phi. \tag{2.3}$$

Agora, substituindo (2.3) em (2.1), temos a f.d.p. reparametrizada em função da média:

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

com $0 < \mu < 1$ e $\phi > 0$. Neste caso, podemos reescrever o valor esperado e a variância de Y, respectivamente, como

$$E(Y) = \mu \quad \text{e} \quad \text{Var}(Y) = V(\mu)/(1+\phi),$$

onde $V(\mu) = \mu(1-\mu)$ é chamada por função de variância. Observe que, mantendo μ fixo e aumentando o valor de ϕ, a variância de Y diminui.

O modelo proposto por Ferrari e Cribari-Neto (2004) é definido através de uma função de ligação entre uma amostra aleatória $Y_1, \ldots, Y_n$, com $Y_i \sim \mathcal{B}(\mu_i, \phi)$, $i = 1,...,n$ e variáveis explicativas. Esta função de ligação é definida por

$$g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \eta_i, \quad i = 1,...,n, \tag{2.4}$$

onde $g : (0,1) \to \Re$ é conhecida, estritamente monótona e duas vezes diferenciável. O vetor $\boldsymbol{\beta} = (\beta_1,...,\beta_k)^\top$ é o vetor de parâmetros desconhecidos $(k \times 1)$, o vetor de k covariáveis é $\boldsymbol{x}_i^\top = (x_{i1}, ..., x_{ik})$, e η_i é o preditor linear. Na literatura, existem algumas funções de ligação que podem ser usadas, a saber:

(i) Função logito: $g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$;

(ii) Função probito: $g(\mu_i) = \Phi^{-1}(\mu_i)$;

(iii) Função complemento log-log: $g(\mu_i) = \log\{-\log(1-\mu_i)\}$,

para $i = 1,...,n$, com $\Phi(\cdot)$ sendo a função de distribuição acumulada de uma variável normal padrão. De (2.4), temos

$$\mu_i = g^{-1}(\eta_i), \;\; i = 1,...,n, \tag{2.5}$$

logo, podemos isolar μ_i nas funções de ligação supracitadas:

(i) Função logito: $\mu_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$,;

(i) Função probito: $\mu_i = \Phi(\eta_i)$;

(i) Função complemento log-log: $\mu_i = 1 - \exp\{-\exp(\eta_i)\}$

O logaritmo da função de verossimilhança do modelo de regressão beta (FERRARI; CRIBARI-NETO, 2004) é dado por:

$$\begin{aligned}\ell(\boldsymbol{\beta}, \phi) \;\; = \;\; & \sum_{i=1}^{n} \Big\{ \log\Gamma(\phi) - \log\Gamma(\mu_i\phi) - \log\Gamma\left[(1-\mu_i)\phi\right] + (\mu_i\phi - 1)\log(y_i) \\ & + \left[(1-\mu_i)\phi - 1\right]\log(1-y_i)\Big\}.\end{aligned}$$

A estimação por máxima verossimilhança dos parâmetros $\boldsymbol{\beta}$ e ϕ, deve ser feita utilizando métodos numéricos, já que, a mesma não é possível de forma analítica. Neste caso, é feita a maximização numérica da função de log-verossimilhança, para isso Ferrari e Cribari-Neto (2004) optaram por usar um algoritimo de otimização não-linerar, como o algoritimo de Newton ou um quase-Newton.

2.2 Modelo de regressão gama unitária

O modelo de regressão gama Unitária é baseado na distribuição gama unitária, e semelhantemente ao modelo beta descrito acima, o modelo Gama unitária foi proposto para casos em que a variável de interesse assume valores no intervalo $(0,1)$.

A distribuição gama foi proposta por Grassia (1977). Considere a variável aleatória W, cuja distribuição é a gama com parâmetros $\alpha > 0$ e $\phi > 0$, ou seja, $W \sim Gama(\alpha, \phi)$. A função densidade de probabilidade, neste caso, é dada por:

$$f^*(\omega; \alpha,\phi) = \frac{\alpha^\phi}{\Gamma(\phi)} e^{-\alpha\cdot\omega} \omega^{\phi-1},\ \omega > 0 \tag{2.6}$$

A esperança e variância de W são dadas, respectivamente, por

$$E(W) = \frac{\phi}{\alpha} \text{ e } Var(W) = \frac{\phi}{\alpha^2}.$$

Para obter a f.d.p. da distribuição gama unitária, proposta por Mousa e Abdel-Fattah (2016), devemos usar a transformação $w = log(1/y)$, que implica em $y = e^{-w}$, $0 < y < 1$. Logo, y tem distribuição Gama Unitária ($Y \sim UG(\alpha,\phi)$) com parâmetros α e ϕ e sua f.d.p. é:

$$f(y; \alpha,\phi) = \frac{\alpha^\phi}{\Gamma(\phi)} y^{\alpha-1}(-\log y)^{\phi-1},\ \ 0 < y < 1. \tag{2.7}$$

Neste caso,

$$E(Y) = \left(\frac{\alpha}{\alpha+1}\right)^\phi \text{ e } Var(Y) = \left(\frac{\alpha}{\alpha+2}\right)^\phi - \left(\frac{\alpha}{\alpha+1}\right)^{2\phi}. \tag{2.8}$$

Usando a mesma ideia de Ferrari e Cribari-Neto (2004) , podemos reparametrizar a distribuição em função da média, ou seja, isolamos α em $\mu = E(Y) = [\alpha/(\alpha+1)]^\phi$, resultando em $\alpha = \mu^{1/\phi}/(1-\mu^{1/\phi})$. Substituindo α em (2.7) temos:

$$f(y; \mu, \phi) = \frac{1}{\Gamma(\phi)} \left(\frac{\mu^{1/\phi}}{1-\mu^{1/\phi}}\right)^\phi y^{[\mu^{1/\phi}/(1-\mu^{1/\phi})]-1}(-\log y)^{\phi-1},\ \ 0 < y < 1, \tag{2.9}$$

com $0 < \mu < 1$ e $\phi > 0$. Considerando esta reparametrização, podemos reescrever (2.8) como:

$$E(Y) = \mu \text{ e } Var(Y) = \mu\left[\frac{1}{(2-\mu^{1/\phi})^\phi} - \mu\right].$$

Considere uma amostra aleatória $Y_1, Y_2, \ldots, Y_n$, em que $Y_i \sim UG(\mu_i, \phi)$, com $i = 1, 2, \ldots, n$. Baseados nesta amostra, Mousa e Abdel-Fattah (2016) propuseram o modelo de regressão Gama Unitária

$$g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \eta_i,\ \ i = 1, \ldots, n,$$

em que $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ é o vetor p-dimensional de parâmetros desconhecido, $\boldsymbol{x}_i^\top = (x_{t1}, \ldots, x_{tp})$ são observações de p variáveis explicativas ($p < n$) e $g : (0,1) \to \Re$ é a função de ligação, que é estritamente monótona e duas vezes diferenciável. Aqui, pode-se considerar as mesmas funções de ligação apresentadas na subseção anterior.

Os estimadores dos parâmetros do modelo de regressão gama unitária podem ser obtidos análogos aos do modelo de regressão beta, através de maximização numérica do logaritmo da função de verossimilhança:

$$\ell(\boldsymbol{\beta},\phi) = \sum_{i=1}^{n} \{\phi \log(\alpha_i) - \log[\Gamma(\phi)] + (\alpha_i - 1)\log(y_i) + (\phi - 1)\log[-\log(y_i)]\}.$$

Neste processo de estimação é necessário um valor (chute) inicial para os parâmetros do modelo. Então usamos os mesmos valores propostos por Ferrari e Cribari-Neto (2004) no modelo de regressão beta, dados por:

$$\boldsymbol{\beta}^{(0)} = (X^\top X)^{-1} X^\top z$$

em que $X^\top = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_i)$

$$\phi^{(0)} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\breve{\mu}_i(1-\breve{\mu}_i)}{\breve{\sigma}_i^2} \right] - 1,$$

onde

$$\breve{\mu}_i = g^{-1}(x_i^\top (X^\top X)^{-1} X^\top z) = \frac{e^{x_i^\top (X^\top X)^{-1} X^\top z}}{1 + e^{x_i^\top (X^\top X)^{-1} X^\top z}},$$

$$\breve{\sigma}_i^2 = \frac{\breve{e}^\top \breve{e}}{(n-p)[g'(\breve{\mu}_i)]^2},$$

em que $z = g(y) = log[y/(1-y)]$, $\boldsymbol{x_i^\top}$ é i-ésima linha de X, $g'(\breve{\mu}_i) = 1/[\breve{\mu}_i(1-\breve{\mu}_i)]$ e $\breve{e} = z - X(X^T X)^{-1} X^T z$.

2.3 Correção de viés

O viés é uma medida que indica o quanto um estimador $\hat{\boldsymbol{\theta}}$ se afasta, em média, do valor verdadeiro do parâmetro $\boldsymbol{\theta}$. Ele é definido como a diferença entre o valor esperado do estimador e o parâmetro:

$$vi\acute{e}s(\hat{\boldsymbol{\theta}}) = E[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}. \tag{2.10}$$

Em geral, o verdadeiro valor do parâmetro é desconhecido. Por isso, não é possível calcular diretamente o viés de um estimador. No entanto, com o conhecimento sobre amostragem e inferência estatística, podemos estudar o viés de um estimador analisando suas propriedades teóricas ou por meio de simulações, que permitem aproximar a diferença entre o valor esperado do estimador e o parâmetro.

Com o método de *bootstrap* podemos aproximar a distribuição amostral de um estimador por meio da técnica de reamostragem com reposição a partir da amostra observada. A vantagem é que, ao realizar um grande número de reamostragens, conseguimos estimar com maior precisão medidas como viés, variância e intervalos de confiança.

Para estimar o viés com *bootstrap*, realizamos B réplicas de $\hat{\boldsymbol{\theta}}$ a partir da técnica de reamostragem com reposição, que permite aproximar a distribuição amostral do estimador. Assim, dada uma amostra de tamanho n,

$$x = (x_1, \ldots, x_n)^\top,$$

geramos B reamostragens aleatórias de tamanho n (com reposição) dessa amostra. Para cada amostra

$$x^{*(b)} = (x_1^{*(b)}, \ldots, x_n^{*(b)}),$$

calculamos um estimador $\hat{\boldsymbol{\theta}}^{(b)}$, com $b = 1, \ldots, B$.

Com isso, podemos então calcular a estimativa *bootstrap* para $\hat{\boldsymbol{\theta}}$ e a estimativa do viés via *bootstrap*. Essas estatísticas são dadas por:

$$\hat{\boldsymbol{\theta}}^{*}_{boot} = \frac{1}{B}\sum_{b=1}^{B}\hat{\boldsymbol{\theta}}^{(b)}$$

e

$$\widehat{vi\acute{e}s}_{boot}(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\theta}}^{*}_{boot} - \hat{\boldsymbol{\theta}}.$$

Enfim, a correção do viés pode ser obtida por meio das reamostragens do método *bootstrap*. Mesmo quando a amostra inicial é pequena, ao gerar um grande número de reamostragens conseguimos estimar a distribuição amostral do estimador e, assim, obter resultados mais robustos e confiáveis.

2.4 Testes de hipóteses

Consideremos um caso multiparamétrico em que o vetor de parâmetros $\boldsymbol{\theta}$ é formado por dois subconjuntos: um vetor de parâmetros de interesse e outro de parâmetros de perturbação. Denotamos por $\boldsymbol{\psi}$ o vetor de dimensão q de parâmetros de interesse e por $\boldsymbol{\omega}$ o vetor de dimensão $(p+1-q)$ correspondente aos parâmetros de perturbação:

$$\boldsymbol{\theta} = (\boldsymbol{\beta}^{\top}, \phi)^{\top} = (\boldsymbol{\psi}^{\top}, \boldsymbol{\omega}^{\top}), \qquad \boldsymbol{\psi}^{\top} = (\beta_1, \beta_2, \ldots, \beta_q), \qquad \boldsymbol{\omega}^{\top} = (\beta_{q+1}, \beta_{q+2}, \ldots, \beta_p, \phi).$$

Quando realizamos um teste de hipótese, buscamos aproximar a distribuição da estatística de teste sob a hipótese nula H_0:

$$H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0 \ \text{ versus } \ H_1 : \boldsymbol{\psi} \neq \boldsymbol{\psi}_0,$$

onde $\boldsymbol{\psi}_0$ é um vetor fixo de dimensão q.

Em situações com número reduzido de observações, essa aproximação pode ser insatisfatória, tornando a inferência pouco confiável. Uma alternativa para contornar esse problema é o método *bootstrap*, que compensa o baixo tamanho amostral pelo elevado número de reamostragens.

Na aplicação do *bootstrap*, a partir de uma amostra inicial e sob H_0, geramos B amostras reamostradas com reposição. Para cada amostra $y^{*(b)}$ calculamos a estatística de teste $T^{*(b)}$, e a distribuição empírica obtida serve como base para determinar regiões críticas e valores críticos. Assim, dada uma estatística de teste $T = T(y)$, o valor crítico pode ser definido como o quantil $T^{*}_{1-\alpha}$ da distribuição empírica, e rejeitamos H_0 se $T > T^{*}_{1-\alpha}$.

Além disso, o método *bootstrap* permite a obtenção de um valor-p estimado:

$$\widehat{\alpha}^* = \frac{\#\{T^{*(b)} > T\}}{B},$$

onde $T^{*(b)}$ é a estatística de teste obtida na b-ésima reamostragem. Neste caso, rejeitamos H_0 se $\widehat{\alpha}^* < \alpha$.

Dado que a distribuição de interesse é a Gama Unitária, para a construção das estatísticas de teste utilizamos sua log-verossimilhança, a partir da qual obtemos o vetor escore e a matriz de informação de Fisher (e sua inversa). Aqui, é usada a mesma partição do vetor de parâmetros desconhecidos $\boldsymbol{\theta}$ em função de $\boldsymbol{\psi}$ e $\boldsymbol{\omega}$ para a obtenção do vetor escore $U(\boldsymbol{\theta})$ e matriz de informação de Fisher $K(\boldsymbol{\theta})$ e sua inversa $K(\boldsymbol{\theta})^{-1}$, ou seja,

$$U(\boldsymbol{\theta}) = \begin{bmatrix} U_{\boldsymbol{\psi}}(\boldsymbol{\theta}) \\ U_{\boldsymbol{\omega}}(\boldsymbol{\theta}) \end{bmatrix},$$

$$K(\boldsymbol{\theta}) = \begin{bmatrix} K_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\theta}) & K_{\boldsymbol{\psi}\boldsymbol{\omega}}(\boldsymbol{\theta}) \\ K_{\boldsymbol{\omega}\boldsymbol{\psi}}(\boldsymbol{\theta}) & K_{\boldsymbol{\omega}\boldsymbol{\omega}}(\boldsymbol{\theta}) \end{bmatrix} \quad \text{e} \quad K(\boldsymbol{\theta})^{-1} = \begin{bmatrix} K^{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\theta}) & K^{\boldsymbol{\psi}\boldsymbol{\omega}}(\boldsymbol{\theta}) \\ K^{\boldsymbol{\omega}\boldsymbol{\psi}}(\boldsymbol{\theta}) & K^{\boldsymbol{\omega}\boldsymbol{\omega}}(\boldsymbol{\theta}) \end{bmatrix}.$$

Os detalhes dos cálculos destas quantidades podem ser vistos em Barroso (2022).

Os testes de hipóteses assintóticos considerados neste trabalho são apresentados a seguir:

- Razão de verossimilhança(*likehood ratio*:

$$LR = 2\big[\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widetilde{\boldsymbol{\theta}})\big].$$

- Escore (*score*):

$$S = U_{\boldsymbol{\psi}}(\widetilde{\boldsymbol{\theta}})^\top \, K^{\boldsymbol{\psi}\boldsymbol{\psi}}(\widetilde{\boldsymbol{\theta}}) \, U_{\boldsymbol{\psi}}(\widetilde{\boldsymbol{\theta}}).$$

- Gradiente:

$$G = U_{\boldsymbol{\psi}}(\widetilde{\boldsymbol{\theta}})^\top \, (\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0).$$

- Wald:

$$W = (\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^\top \, \big[K^{\boldsymbol{\psi}\boldsymbol{\psi}}(\widehat{\boldsymbol{\theta}})\big]^{-1} \, (\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0).$$

Sob H_0, essas estatísticas são assintoticamente equivalentes e seguem uma distribuição qui-quadrado com q graus de liberdade (χ^2_q). Contudo, em amostras pequenas, a aproximação assintótica pode não ser adequada. Nesse cenário, a utilização do método *bootstrap* surge como alternativa viável para garantir inferências mais confiáveis. Assim, para a construção dos testes de hipóteses, via *bootstrap*, seguimos os seguintes passos:

1. Calcule a estatística de teste $T = T(y)$ na amostra observada.
2. Estime $\boldsymbol{\theta}$ sob H_0.

3. Gere B reamostragens $y^{*(1)}, \ldots, y^{*(B)}$ a partir dessa suposição.

4. Calcule $T^{*(b)} = T(y^{*(b)})$ para cada $b = 1, \ldots, B$.

5. Obtenha regiões críticas ou valores-p a partir da distribuição empírica de T^*.

Em resumo, neste capítulo foram apresentados os fundamentos teóricos e metodológicos necessários para o desenvolvimento do trabalho, incluindo o modelo de regressão Gama Unitária, os procedimentos de estimação dos parâmetros e as estratégias para correção de viés via método *bootstrap*. Além disso, foram discutidos os principais testes de hipóteses utilizados para avaliar os parâmetros do modelo, tanto sob a abordagem assintótica quanto sob a abordagem empírica baseada em reamostragem.

3 Resultados

3.1 Apresentação geral do pacote *UnitGammaReg*

O pacote desenvolvido neste trabalho tem como objetivo disponibilizar, de forma acessível e reprodutível, as metodologias propostas por Barroso (2022) para o modelo de regressão gama unitária, utilizando o método de reamostragem *bootstrap*, uma vez que não há nenhum script ou função proposta que contemple ambas as metodologias. O pacote *UnitGammaReg* foi implementado em linguagem **R** e encontra-se disponível publicamente no repositório GitHub (<https://github.com/vitorfloriano-prog/Unit_Gamma_Reg>).

O pacote contém funções voltadas à estimação dos parâmetros do modelo, à correção de viés dos estimadores de máxima verossimilhança por meio da metodologia de *bootstrap*, e ao cálculo das versões modificadas dos testes da razão de verossimilhança, Wald, Escore e Gradiente.

Vale ressaltar que que os testes implementados podem ser uniparamétricos, quando avaliam uma única hipótese do tipo $\beta_1 = 0$ (o intercepto não é testado), ou multiparamétricos, quando testam simultaneamente hipóteses do tipo $\beta_1 = \beta_2 = \ldots = \beta_q = 0$, respeitando a ordem dos parâmetros. Além disso, o pacote disponibiliza uma função auxiliar para a visualização dos resultados.

Os arquivos inseridos na Plataforma GitHub foram estruturados da seguinte maneira:

- **R**: contém os scripts com as funções implementadas, neste arquivo existem duas funções, cada uma com um objetivo descrito no arquivo “README.md".
- **DESCRIPTION**: armazena as informações gerais do pacote, como nome, versão, autor e dependências.
- **NAMESPACE**: Controla a visibilidade das funções e objetos do pacote. Ele atua como um mecanismo de isolamento e gerenciamento de dependências.
- **README.md**: arquivo explicativo com instruções básicas de instalação e exemplos de uso, além de uma breve explicação das funcionalidades das funções implementadas;

As principais funções implementadas no pacote são descritas a seguir:

- `ugamma.fit()`: ajusta o modelo de regressão Gama Unitária por meio do método da máxima verossimilhança, realiza os testes de hipóteses conforme descrito no capítulo 2 e corrige o viés dos estimadores de máxima verossimilhança utilizando a metodologia de

bootstrap. Como saída, a função retorna ao usuário as estimativas de máxima verossimilhança, tanto nas versões usuais quanto nas versões corrigidas via *bootstrap*, além dos valores das estatísticas de teste tradicionais (LR, W, G e S) e de suas correspondentes versões corrigidas também pelo procedimento de *bootstrap*.

- `summary.Unit.gamma`: esta função permite a impressão dos resultados do modelo por meio do comando `summary()` nativo do `R`. Sua implementação garante ao usuário uma saída limpa, didática e objetiva, facilitando a interpretação dos resultados obtidos.

3.2 Instalação e utilização do pacote

A instalação e a utilização do pacote são descritas a seguir. No ambiente `R`, digite:

```
# Instalação a partir do GitHub
> install.packages("devtools")
> devtools::install_github("vitorfloriano-prog/Unit_Gamma_Reg")
> library(UnitGammaReg)
```

Após a instalação, as funções do pacote podem ser utilizadas normalmente para o ajuste do modelo. Um ponto importante a ser destacado é que a função `summary.Unit.gamma()` existe apenas para garantir que o comando `summary()` nativo do **R** imprima os resultados nas formas de tabelas, que contém em si todas as informações obtidas com o ajuste, sendo assim, não há a necessidade do uso da função explicitamente, Uma vez que o pacote tenha sido carregado por meio do comando `library(UnitGammaReg)`, a função já estará ativa e pronta para uso, a partir do comando nativo citado anteriormente.

3.3 Aplicação a dados reais

Com o objetivo de ilustrar a aplicabilidade prática do pacote desenvolvido, foi utilizada uma base de dados reais. Inicialmente, apresentamos o conjunto de dados, identificando a variável resposta (Y) e as covariáveis consideradas no estudo. Em seguida, ajustamos o modelo e apresentamos a interpretação dos parâmetros estimados. Por fim, é realizado um teste de hipóteses para verificar a significância de uma das covariáveis sobre a variável resposta.

3.3.1 Apresentação dos dados

Na presente aplicação desejamos investigar a relação entre o desenvolvimento humano juntamente com distribuição de renda e o nível de desenvolvimento da escolaridade nas unidades da federação (UF) do Brasil. Tal escolha nos resulta em um tamanho de amostra relativamente

baixo ($n = 27$), porém isso é de extremo interesse, pois, assim podemos analisar os resultados obtidos por meio da reamostragens de *bootstrap* quando temos uma amostra não muito grande.

Como variável reposta temos o subíndice de escolaridade - (IDHM Educação), este é um dos principais indicadores quando tratamos de qualidade educacional, medindo o nível de educação da população de pessoas com 18 anos ou mais, sendo calculado com base na porcentagem dessas pessoas que completaram o ensino fundamental. Este índice varia entre 0 e 1 sendo ideal para o modelo para o qual desejamos ajustar. Os dados têm como referência o ano de 2021 e foram providos pela PNAD Contínua.

Os dados utilizados são de domínio público e originários do Instituto Brasileiro de Geografia e Estatistica (IBGE) e foram levantados por meio da Pesquisa Nacional por Amostra de Domicílios Contínua(PNAD Contínua), veja Apêndice A.

Como variáveis explicativas, selecionamos o Índice de Desenvolvimento Humano Municipal geral (IDHM), renda per capta, taxa de analfabetismo a população com 18 anos ou mais de idade. O IDHM é um indicador, cujo cálculo se baseia na saúde (longevidade), educação e na renda dos municípios para avaliar a qualidade de vida e progresso nos mesmos. A renda per capta é a renda media de um indivíduo em determinado grupo, sendo então a renda total de cada estado dividida pelo seu número de habitantes. Por fim, a taxa de analfabetismo (18 anos ou mais) é dada pela proporção de pessoas nessa faixa etária que não sabem ler e escrever, servindo para medir o grau de escolaridade de adultos e idosos.

Para a obtenção dos dados recorremos ao site Atlas Brasil (<http://atlasbrasil.org.br/acervo/biblioteca.>), que possui a funcionalidade de consultar tabelas com informações confiáveis de fontes como IBGE e DataSUS por exemplo, além de permitir selecionar os indicadores de interesse e a abrangência de pesquisa.

Nesta aplicação iremos modelar o IDHM Educação em função do IDMH, da renda per capta e da taxa de analfabetismo (18 anos ou mais de idade).

3.3.2 Modelagem e inferência

Com base nos dados descritos na seção 3.3.1 e com base na teoria apresentada em 2.1 e 2.2, o modelo proposto busca explicar o IDHM Educação como um função da taxa de analfabetismo, renda per capta e IDHM, e para a correção de viés aplicaremos a teoria presente em 2.3 e 2.4, com 1000 repetições de *bootstrap*.

Usando a função de ligação logito, consideramos o seguinte modelo:

$$g(\mu_t) = \log\left(\frac{\mu_t}{1-\mu_t}\right) = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \beta_3 X_{t3}. \tag{3.1}$$

Neste modelo, a variável resposta Y_t representa o IDHM Educação, X_{t1} o IDHM , X_{t2} a renda per capta e X_{t3} a taxa de analfabetismo de pessoas com 18 anos ou mais de idade . A estimação de $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^\top$ é realizada por máxima verossimilhança e está presente no código

disponibilizado na plataforma GitHub, os estimadores com o viés corrigido por *boostrap* também está integrada no mesmo código.

No R seguimos a seguinte aplicação :

```
# Instalação a partir do GitHub
install.packages("devtools")
devtools::install_github("vitorfloriano-prog/Unit_Gamma_Reg")

# Executa o código após instalação do pacote.
library(UnitGammaReg)

# Y representa o Subindice de escolaridade - IDHM Educação.
Y <-c(0.688,0.599,0.736,0.739,0.617,0.644,0.857,0.687,0.704,
       0.618,0.684,0.706,0.665,0.573,0.704,0.647,0.640,0.574,
       0.776,0.636,0.708,0.639,0.755,0.740,0.783,0.625,0.667)

# IDHM
x1<-c(0.710,0.684,0.688,0.700,0.691,0.734,0.814,0.771,0.737,
       0.676,0.742,0.736,0.774,0.698,0.769,0.690,0.719,0.690,
       0.762,0.728,0.771,0.700,0.699,0.792,0.806,0.702,0.731)

# Renda per capta
x2 <- c(471.54,404.28,451.27,432.99,451.59,480.55,1326.87,684.63,
         679.62,341.32,767.68,707.48,699.24,465.74,817.79,442.82,
         447.97,452.75,901.42,593.46,944.53,541.74,549.54,901.20,
         971.82,492.78,564.61)

# Taxa de Analfabetismo- 18 anos ou mais
x3<- c(9.04,13.83,4.21,4.64,11.05,12.06,1.9,5.24,4.8,12.4,
        4.05,4.56,4.63,13.68,3.27,6.46,10.36,14.14,1.95,10.29,
        2.14,5.6,5.53, 1.83,1.97,12.18,8.21)

# Data frame com as 3 variaveis explicativas
dat <- data.frame(Y = Y,
                  IDHM = x1,
                  renda.per.capta = x2,
                  T.analfabetismo_18mais = x3)

res_mod.1 <- ugamma.fit(Y ~ ., data = dat, q =1 , B = 1000)
```

```
summary(res_mod.1)

============================================
Model Summary - Unit Gamma (Ugamma.fit)
============================================
sample size: 27

coefficients:
                       Estimate Std.Error
(Intercept)               2.481     0.773
IDHM                     -2.754     1.196
renda.per.capta           0.001     0.000
T.analfabetismo_18mais   -0.043     0.007
phi                     148.164    40.280

coefficients corrected by bootstrap:
                       Estimate Std.Error.boot
(Intercept)               2.517          0.772
IDHM                     -2.851          1.196
renda.per.capta           0.001          0.000
T.analfabetismo_18mais   -0.043          0.007
phi                     148.163         40.280

Hypothesis testing table (LR, G, S, W):
 Statistic p-value
        LR   0.000
         G   1.000
         S   0.000
         W   0.021

Bootstrap hypothesis testing table (LR, G, S, W):
 Statistic p-value.b
        LR     0.412
         G     0.834
         S     0.322
         W     0.019

AIC.Gamma: -73.3919  BIC.Gamma: -68.2085
Number of errors in bootstrap: 0
```

```
==========================================
```

Podemos observar que o IDHM e a taxa de analfabetismo (18 anos ou mais) possuem coeficientes negativos, o que pode sugerir que altos IDHM ou então altas taxas de analfabetismo na fase adulta podem estar relacionadas a baixos subíndices de escolaridade, já a renda per capta ele aparenta influenciar positivamente, porém seu coeficiente é muito baixo.

Quanto ao teste de hipóteses aqui buscamos testar se o IDHM seria estatisticamente significativo para o modelo, ou seja,

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0,$$

e pelos resultado dos testes tradicionais concluímos que a covariável IDHM é significativa, uma vez que em 3 dos 4 testes (LR, S e W), não há evidencias estatísticas para que rejeitemos a hipótese nula. Usando a metodologia de *bootstrap*, temos que as estimativas não mudaram significativamente, IDHM e a taxa de analfabetismo ainda têm coeficientes negativos e bastante próximos aos antigos valores, o mesmo vale para a renda per capta, que se manteve a mesma. As diferenças surgem nos testes de hipóteses, aqui mudamos completamente nossa interpretação com os 4 testes dando a entender que o IDHM não seria estatisticamente significativo para o modelo, sendo a assim optamos por retirá-lo do modelo e realizar uma nova modelagem.

Nesta nova modelagem, consideramos o seguinte modelo:

$$g(\mu_t) = \log\left(\frac{\mu_t}{1-\mu_t}\right) = \beta_0 + \beta_2 X_{t2} + \beta_3 X_{t3}. \tag{3.2}$$

No R, temos:

```
# Data frame com 2 variaveis explicativas
# (Renda per capta Taxa de analfabetismo)

dat_2 <- data.frame(Y = Y,
                    renda.per.capta = x2,
                    T.analfabetismo_18mais = x3)

res_mod.2 <- ugamma.fit(Y ~ ., data = dat_2, q =1 , B = 1000)
summary(res_mod.2)

==========================================
Model Summary - Unit Gamma (Ugamma.fit)
==========================================
sample size: 27
```

```
coefficients:
                       Estimate Std.Error
(Intercept)               0.730     0.123
renda.per.capta           0.001     0.000
T.analfabetismo_18mais   -0.042     0.007
phi                     144.445    39.268

coefficients corrected by bootstrap:
                       Estimate Std.Error.boot
(Intercept)               0.707          0.123
renda.per.capta           0.001          0.000
T.analfabetismo_18mais   -0.042          0.007
phi                     144.444         39.267

Hypothesis testing table (LR, G, S, W):
 Statistic p-value
        LR   1.000
         G   0.036
         S   0.054
         W   0.000

Bootstrap hypothesis testing table (LR, G, S, W):
 Statistic p-value.b
        LR     0.181
         G     0.066
         S     0.059
         W     0.157

AIC.Gamma: -71.7759  BIC.Gamma: -67.8884
Number of errors in bootstrap: 0
==========================================
```

Para estes novos resultados, podemos observar que a renda per capta manteve o mesmo comportamento do modelo anterior, dando a entender que possui uma relação positiva com a variável resposta, porém seu coeficiente é bastante baixo. Quando voltamos nossa atenção para a taxa de analfabetismo na fase adulta, ela também manteve o mesmo comportamento então podemos supor que as evidências de que de fato esta preditora possui uma relação negativa com a resposta estejam se tornando mais forte. O teste de hipóteses aplicado a este novo modelo busca verificar se a renda per capta seria significativa para o mesmo, dessa forma $H_0 : \beta_2 = 0$

contra $H_0 : \beta_2 \neq 0$. Quanto aos resultados dos testes, aqui temos uma decisão dividida, existindo 2 testes que rejeitariam H_0 (testes G e W) e outros 2 que indicam que a renda per capta possa não ser significativa para o modelo (testes LR e S), mas vale ressaltar que o teste de score teve seu valor-p muito próximo à rejeição. Usando *bootstrap* temos que novamente as estimativas não mudaram de forma significativa, porém, os resultados dos testes de hipótese indicam por unanimidade que a renda per capta não teria um efeito significativo no modelo, embora alguns valores-p tenham novamente ficado muito próximos da rejeição da hipótese nula.

Levando isso em consideração, realizamos uma última modelagem, restando apenas a taxa de analfabetismo (18 anos ou mais) como nossa preditora, sendo o modelo dado da seguinte forma

$$g(\mu_t) = \log\left(\frac{\mu_t}{1-\mu_t}\right) = \beta_0 + \beta_3 X_{t3} \tag{3.3}$$

Sua implementação no `R` tem a seguinte forma:

```
# Data frame apenas com a Taxa de analfabetismo
dat_3 <- data.frame(Y = Y,
                    T.analfabetismo_18mais = x3)

res_mod.Final <- ugamma.fit(Y ~ ., data = dat_3, q =1 , B = 1000)
summary(res_mod.Final)

============================================
Model Summary - Unit Gamma (Ugamma.fit)
============================================
sample size: 27

coefficients:
                       Estimate Std.Error
(Intercept)               1.221     0.071
T.analfabetismo_18mais   -0.063     0.009
phi                      40.010    10.844

coefficients corrected by bootstrap:
                       Estimate Std.Error.boot
(Intercept)               1.221          0.078
T.analfabetismo_18mais   -0.063          0.010
phi                      33.188          8.988

Hypothesis testing table (LR, G, S, W):
 Statistic p-value
```

```
        LR       0
         G       0
         S       0
         W       0

Bootstrap hypothesis testing table (LR, G, S, W):
 Statistic p-value.b
        LR         0
         G         0
         S         0
         W         0

AIC.Gamma: -92.7796  BIC.Gamma: -90.1879
Number of errors in bootstrap: 0
============================================
```

Nestes últimos resultados, novamente altas taxas de analfabetismo na fase adulta aparentam estar relacionadas a baixos IDHM Educação, esta relação segundo o resultado dos testes de hipóteses, tem forte evidência de ser verdadeira, uma vez que novamente por unanimidade os teste concordam fortemente que a taxa de analfabetismo na fase adulta tem efeito significativo para o modelo. Quando usamos o metodo de *bootstrap* observamos, por fim, que exceto pela estimativa do parâmetro de dispersão, todos os resultados se mantiveram iguais. A fim de verificar se o modelo final obtido de fato seria o melhor, obtivemos os AIC's e BIC's de cada modelo. Veja,

- para o modelo com as 3 variáveis explicativas temos $AIC = -73{,}3919$ e $BIC = -68{,}2085$;
- para o modelo com renda per capta e taxa de analfabetismo (18 anos ou mais) como preditoras, temos $AIC = -71{,}7759$ e $BIC = -67{,}8884$;
- para o modelo contendo apenas taxa de analfabetismo como preditora temos $AIC = -92{,}7796$ e $BIC = -90{,}1879$.

Observamos que o modelo final (com uma única preditora - taxa de analfabetismo) obtido foi aquele cujo AIC e BIC possuem os menores valores dentre todos os modelos.

Conclusão

O presente trabalho teve como objetivo principal a implementação e a disponibilização, na plataforma `GitHub`, de um *script* em linguagem `R` que permita a aplicação do modelo de regressão gama unitária, apropriado para lidar com variáveis resposta restritas ao intervalo (0,1). Além disso, foi implementado o método de *bootstrap* para a correção do viés dos estimadores de máxima verossimilhança (EMV) e das estatísticas de teste de Wald, Escore, Gradiente e Razão de Verossimilhança, motivado pela inexistência de uma implementação prévia desse tipo.

Este trabalho representa uma implementação inicial, na qual o *script* disponibilizado oferece ao usuário funções úteis e que produzem resultados robustos. Entretanto, ainda existem algumas limitações na implementação. Uma delas é a impossibilidade de escolher qual variável será testada. Sempre que se deseja testar um único parâmetro, o teste é realizado apenas para β_1 e, quando se pretende testar mais de um parâmetro, é executado um teste multiparamétrico, sem a opção de selecionar quais parâmetros incluir, sendo os testes conduzidos sempre em ordem (por exemplo, $\beta_1 = \beta_2 = 0$), fato que foi evidenciado na seção de resultados deste trabalho, uma vez que, foram realizadas 3 modelagens para selecionar as variáveis que seriam significativas. Além disso, o intercepto é mantido no modelo, não sendo submetido a testes de hipótese.

Quanto à aplicação em dados reais, foi investigada a associação entre o subíndice de escolaridade (IDHM Educação) e as variáveis taxa de analfabetismo (18 anos ou mais), renda per capita e IDHM, considerando as 27 unidades da federação brasileiras. O objetivo principal foi verificar qual ou quais variáveis explicativas seriam estatisticamente significativas para o modelo, usando o pacote proposto neste trabalho. Ao final de 3 modelagens constatou-se que o IDHM não seria significativo para o modelo, a renda poderia ser significativa para o modelo a depender do nível de confiança desejado para o teste, já a variável taxa de analfabetismo (18 anos ou mais) seria extremamente significativa para o modelo proposto, com os resultados dos testes de hipóteses, tanto antes quanto após a aplicação de *bootstrap*, indicando uma relação extremamente forte entre o subíndice de escolaridade (IDHM Educação) e a taxa de analfabetismo na fase adulta, além disso, através da comparação das AIC's e do BIC's dos modelos, o modelo final encontrado foi o que apresentou o melhor ajuste dentre os 3 modelos considerados.

Em estudos futuros, sugerimos um aprimoramento nos teste de hipóteses, buscando refiná-los de tal forma a sanar as controvérsias supracitadas, além disso, seriam de interesse implementar uma análise comparativa entre modelos, possibilitado o uso de comando como `Stepwise()` por exemplo.

Referências

BARNDORFF-NIELSEN, O. E. Modified signed log likelihood ratio. **Biometrika**, v. 78, n. 3, p. 557–563, 1991. Disponível em: <https://doi.org/10.1093/biomet/78.3.557>. Citado na página 14.

BARROSO, I. **Inferência estatística via bootstrap no modelo de regressão Gama Unitária**. 2022. Trabalho acadêmico. Citado 5 vezes nas páginas 12, 13, 14, 20 e 22.

BARTLETT, M. S. Properties of sufficiency and statistical tests. **Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences**, v. 160, n. 901, p. 268–282, 1937. ISSN 00804630. Disponível em: <http://www.jstor.org/stable/96803>. Citado na página 14.

CARRASCO, J.; FERRARI, S.; ARELLANO-VALLE, R. Errors-in-variables beta regression models. **Journal of Applied Statistics**, v. 41, n. 7, p. 1530–1547, 2014. Citado na página 13.

CHIEN, L. C. Multiple deletion diagnostics in beta regression models. **Computational Statistics**, v. 28, p. 1639–1661, 2013. Citado na página 13.

COX, D. R.; SNELL, E. J. A general definition of residuals. **Journal of the Royal Statistical Society. Series B (Methodological)**, 1968. Citado na página 14.

EFRON, B. Bootstrap Methods: Another Look at the Jackknife. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 7, n. 1, p. 1 – 26, 1979. Disponível em: <https://doi.org/10.1214/aos/1176344552>. Citado na página 14.

ESPINHEIRA, P. *et al.* Model selection criteria on beta regression for machine learning. **Machine Learning & Knowledge Extraction**, v. 1, p. 427–449, 2019. Citado na página 13.

ESPINHEIRA, P. L.; FERRARI, S. L.; CRIBARI-NETO, F. On beta regression residuals. **Journal of Applied Statistics**, Taylor & Francis, v. 35, n. 4, p. 407–419, 2008. Disponível em: <https://doi.org/10.1080/02664760701834931>. Citado na página 13.

FABRIZI, E.; FERRANTE, M.; TRIVISANO, C. Bayesian beta regression models for the estimation of poverty and inequality parameters in small areas. In: **Analysis of Poverty Data by Small Area Estimation**. [S.l.: s.n.], 2016. p. 299–314. Citado na página 13.

FERRARI, S. L.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of Applied Statistics**, 2004. Citado 6 vezes nas páginas 12, 13, 15, 16, 17 e 18.

FREITAS, J. V. B. *et al.* Unit gamma regression models for correlated bounded data. **Brazilian Journal of Probability and Statistics**, v. 37, p. 693–719, 2023. Citado na página 13.

GEISSINGER, E. *et al.* A case for beta regression in the natural sciences. **Ecosphere**, v. 13, n. 2, 2022. Citado na página 13.

GRASSIA, A. On a family of distributions with argument between 0 and 1 obtained by transformation of the gamma and derived compound distributions. **Australian Journal of Statistics**, 1977. Citado 2 vezes nas páginas 13 e 17.

GUEDES, A. C.; CRIBARI-NETO, F.; ESPINHEIRA, P. L. Modified likelihood ratio tests for unit gamma regressions. **Journal of Applied Statistics**, v. 47, p. 1562–1586, 2020. Citado na página 13.

HAITAO, C. *et al.* A marginalized two-part beta regression model for microbiome compositional data. **PLoS Computational Biology**, v. 14, n. 7, 2018. Citado na página 13.

KOÇ, T.; DÜNDER, E. Jackknife kibria-lukman estimator for the beta regression model. **Communications in Statistics - Theory and Methods**, 2023. Citado na página 13.

MOUSA, A. A. E.-S. A. M.; ABDEL-FATTAH, M. A. A gamma regression for bounded continuous variables. **Advances and Applications in Statistics**, 2016. Citado 2 vezes nas páginas 13 e 17.

OpenAI. **ChatGPT: modelo de linguagem de inteligência artificial**. 2025. Online. Versão GPT-5. Disponível em: <https://chat.openai.com/>. Acesso em: 21 out. 2025. Citado na página 12.

OSPINA, R.; FERRARI, S. L. A general class of zero-or-one inflated beta regression models. **Computational Statistics & Data Analysis**, v. 56, n. 6, p. 1609–1623, 2012. ISSN 0167-9473. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167947311003628>. Citado na página 13.

PACHENCO, W. S. **Efeitos da especificação incorreta da função de ligação do modelo de regressão Gama Unitária**. Dissertação (Dissertação de Mestrado) — Universidade Federal de Goiás, Goiânia, 2023. Citado na página 13.

PETTERLE, R. *et al.* Multivariate quasi-beta regression models for continuous bounded data. **The International Journal of Biostatistics**, v. 17, n. 1, p. 39–53, 2021. Citado na página 13.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2023. Disponível em: <http://www.R-project.org/>. Citado 3 vezes nas páginas 12, 13 e 14.

ROCHA, S.; ESPINHEIRA, P. L.; CRIBARI-NETO, F. Residual and local influence analyses for unit gamma regressions. **Statistica Neerlandica**, v. 75, p. 137–160, 2021. Citado na página 13.

SKOVGAARD, I. M. Likelihood asymptotics. **Scandinavian Journal of Statistics**, v. 28, n. 1, p. 3–32, 2001. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9469.00223>. Citado na página 14.

ZEILEIS, A.; CRIBARI-NETO, F. **Beta Regression in R**. [S.l.], 2009. Citado na página 13.

APÊNDICE A – Tabela de dados

Tabela 1 – Indicadores socioeconômicos selecionados segundo a PNAD Continua (2021)

Territorialidades	**Taxa de analfabetismo - 18 anos ou mais (%)**	**Renda per capita (R$)**	**Subíndice de escolaridade - IDHM Educação**	**IDHM PNAD**
Acre	9,04	471,54	0,698	0,741
Alagoas	14,45	452,41	0,609	0,703
Amapá	4,31	451,27	0,736	0,688
Amazonas	4,46	493,29	0,717	0,711
Bahia	10,21	451,59	0,617	0,711
Ceará	12,05	480,55	0,644	0,734
Distrito Federal	3,65	1320,92	0,792	0,867
Espírito Santo	5,44	634,63	0,807	0,771
Goiás	5,18	758,87	0,780	0,776
Maranhão	14,81	341,32	0,618	0,676
Mato Grosso do Sul	4,27	770,68	0,786	0,785
Mato Grosso	4,18	754,94	0,759	0,777
Minas Gerais	4,36	758,57	0,776	0,787
Paraíba	10,65	465,47	0,646	0,722
Paraná	4,57	817,79	0,793	0,795
Pará	6,94	565,47	0,710	0,741
Pernambuco	10,36	447,97	0,641	0,729
Piauí	11,16	442,58	0,643	0,718
Rio de Janeiro	1,94	901,42	0,776	0,762
Rio Grande do Norte	7,54	553,08	0,740	0,747
Rio Grande do Sul	3,73	842,24	0,776	0,787
Rondônia	5,5	541,74	0,739	0,747
Roraima	4,53	547,97	0,730	0,771
Santa Catarina	1,83	910,24	0,773	0,802
São Paulo	2,57	982,58	0,748	0,805
Sergipe	12,18	492,78	0,625	0,702
Tocantins	8,21	564,61	0,667	0,751

Fonte: Dados do IBGE e de registros administrativos, provenientes da PNAD Contínua, conforme metadados disponíveis em: `http://atlasbrasil.org.br/acervo/biblioteca`.