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

João Marcos Ribeiro Lima

Inflação de Zeros nas Notas da Redação do ENEM: Comparação entre o Modelo Beta Inflacionado em Zero e o Modelo de Barreira

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): João Marcos Ribeiro Lima

Título do trabalho: Inflação de Zeros nas Notas da Redação do ENEM: Comparação entre o Modelo Beta Inflacionado em Zero e o Modelo de Barreira

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 **Ana Carolina Do Couto Andrade**, **Professora do Magistério Superior**, em 27/11/2025, às 09:36, 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 **João Marcos Ribeiro Lima**, **Discente**, em 28/11/2025, às 17:30, 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

João Marcos Ribeiro Lima

Inflação de Zeros nas Notas da Redação do ENEM: Comparação entre o Modelo Beta Inflacionado em Zero e o Modelo de Barreira

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**: Dra. Ana Carolina do Couto Andrade

Goiânia

2025

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

Lima, João Marcos Ribeiro
Inflação de Zeros nas Notas da Redação do ENEM [manuscrito] : Comparação entre o Modelo Beta Inflacionado em Zero e o Modelo de Barreira / João Marcos Ribeiro Lima. - 2025.
75 f.

Orientador: Profa. Dra. Ana Carolina do Couto Andrade.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Instituto de Matemática e Estatística (IME), Estatística, Goiânia, 2025.
Bibliografia. Apêndice.
Inclui siglas, abreviaturas, símbolos, gráfico, tabelas, lista de figuras, lista de tabelas.

1. ENEM. 2. GAMLSS. 3. Inflação em zeros. 4. Modelo de barreira. I. Andrade, Ana Carolina do Couto, 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 cinco dias do mês de novembro do ano de 2025 iniciou-se a sessão pública de defesa do Trabalho de Conclusão de Curso (TCC) intitulado "Inflação de Zeros nas Notas da Redação do ENEM: Comparação entre o Modelo Beta Inflacionado em Zero e o Modelo de Barreira", de autoria de João Marcos Ribeiro Lima, do curso de Estatística, do Instituto de Matemática e Estatística da UFG. Os trabalhos foram instalados pela Profa. Dra. Ana Carolina do Couto Andrade com a participação dos demais membros da Banca Examinadora: Tatiane Ferreira do N.M. da Silva (IME/UFG) e Cynthia Arantes Vieira Tojeiro (IME/UFG). Após a apresentação, a banca examinadora realizou a arguição do estudante. Posteriormente, de forma reservada, a Banca Examinadora atribuiu a nota final de 9,5 , 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 **Cynthia Arantes Vieira Tojeiro**, **Professor do Magistério Superior**, em 26/11/2025, às 14:35, 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 26/11/2025, às 15:54, 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 **Ana Carolina Do Couto Andrade**, **Professora do Magistério Superior**, em 27/11/2025, às 09:37, 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 **5794738** e o código CRC **6EEAD499**.

Referência: Processo nº 23070.059751/2025-71 SEI nº 5794738

Agradecimentos

Agradeço a minha mãe, meu pai e a minha irmã por todo o apoio emocional e financeiro durante toda a jornada da graduação, na produção deste trabalho e ao longo de toda a vida. Amo vocês.

A todos os meus amigos da faculdade, que me acompanharam presencialmente nesta jornada acadêmica e tornaram-na mais leve e divertida e não deixaram a ansiedade tomar de conta de mim. Em especial, Rafael Menezes, Marea e Elis Regina que estão comigo desde o inicio da graduação, e a Akemi e Lucas por terem agregado tanto em minha vida neste último ano. E aos meus amigos que não estavam fisicamente presente, mas que também acompanharam todos os "dramas da academia"e também me ajudaram a manter os pés no chão.

A minha psicologa Regiany por ter sido um pilar essencial para que eu não desistisse do curso e chegasse até aqui.

Aos meus colegas de estágio na SEDUC-GO, em especial meu coordenador Igor Neiva, que sempre foi compreensivo com os desafios deste trabalho e da graduação, e pela ajuda na produção deste trabalho e sugestão do tema.

À minha orientadora, Dra. Ana Carolina do Couto Andrade, pela paciência, dedicação e disposição em me orientar neste trabalho.

A todos os meus colegas do curso de Estatística, que me acompanharam presencialmente nesta jornada acadêmica. Vocês a tornaram mais leve e divertida, ajudaram a dissipar a ansiedade, a resolver questões, estudar para as provas e mantiveram o ambiente amistoso e acolhedor em sala de aula.

A todos os professores do Instituto de Matemática e Estatística (IME), que lapidaram meu conhecimento ao longo destes anos e me fizeram ter certeza de que eu estava no curso certo. Agradeço também à Universidade Federal de Goiás (UFG) pelo ambiente que fomenta a pesquisa e o conhecimento.

A todos que, de forma direta ou indireta, contribuíram para minha formação e para a realização deste sonho, o meu muito obrigado.

'There are no routine statistical questions, only questionable statistical routines'

(Sir David Roxbee Cox,15 Julho 1924 - 18 Janeiro 2022).

Resumo

A nota da Redação do Exame Nacional do Ensino Médio (ENEM), limitada ao intervalo [0, 1000], apresenta uma proporção considerável de notas zero, um fenômeno conhecido como inflação de zeros. Essa característica estrutural dos dados demanda o uso de modelos estatísticos especializados, capazes de lidar com a natureza híbrida da distribuição, que é composta por uma massa pontual em zero e uma componente contínua. O objetivo primário deste trabalho é identificar os fatores que impactam a distribuição das notas da redação para estudantes de escolas públicas e privadas. Como objetivo secundário, de caráter metodológico, busca-se comparar a adequação e a robustez de duas estratégias de modelagem: o modelo Beta Inflacionado em Zero (BEINF0) e o Modelo de Barreira (Hurdle). Para isso, utilizou-se os microdados do ENEM, disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) . A análise foi conduzida em duas populações distintas: uma restrita a Goiânia (2023) e outra ampla, cobrindo o estado de Goiás (2021-2023). Os modelos foram implementados no *framework* GAMLSS no software estatístico `R`, onde o Modelo de Barreira foi especificado com um componente binário (Logístico) e um componente de intensidade contínuo (modelado com a distribuição Box-Cox t). Os resultados da análise descritiva indicam disparidades significativas de desempenho e perfil socioeconômico entre alunos de escolas públicas e privadas. Embora ambas as abordagens de modelagem tenham identificado as notas objetivas e variáveis socioeconômicas como preditores relevantes, a análise de diagnóstico (como *worm plots* e estatísticas de resíduos) demonstrou que o Modelo de Barreira, embora ainda haja inadequações, é metodologicamente mais robusto e conceitualmente mais alinhado à estrutura de avaliação da prova. Ambos os modelos apresentaram dificuldades em capturar a forma da distribuição dos dados, apesar de a abordagem de barreira ter se destacado mais.

Palavras-chave: ENEM. GAMLSS. Inflação em Zeros. Modelo de Barreira

Abstract

The essay score of the National High School Exam (ENEM), bounded to the interval [0, 1000], presents a considerable proportion of zero scores, a phenomenon known as zero inflation. This structural characteristic of the data requires the use of specialized statistical models capable of handling the hybrid nature of the distribution, which consists of a point mass at zero and a continuous component. The primary objective of this work is to identify the factors impacting the distribution of essay scores for students from public and private schools. As a secondary objective, of a methodological nature, this study seeks to compare the adequacy and robustness of two modeling strategies: the Zero-Inflated Beta model (BEINF0) and the Hurdle Model. For this purpose, ENEM microdata made available by the Anísio Teixeira National Institute of Educational Studies and Research (INEP) were utilized. The analysis was conducted on two distinct populations: one restricted to Goiânia (2023) and a broader one covering the state of Goiás (2021-2023). The models were implemented within the GAMLSS framework in the `R` statistical software, where the Hurdle Model was specified with a binary component (Logistic) and a continuous intensity component (modeled with the Box-Cox t distribution). Descriptive analysis results indicate significant disparities in performance and socioeconomic profiles between students from public and private schools. Although both modeling approaches identified objective scores and socioeconomic variables as relevant predictors, diagnostic analysis (such as worm plots and residual statistics) demonstrated that the Hurdle Model, despite remaining inadequacies, is methodologically more robust and conceptually more aligned with the exam's evaluation structure. Both models struggled to capture the shape of the data distribution, although the hurdle approach showed superior performance.

Keywords: ENEM. GAMLSS. Zero Inflation. Hurdle Model.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

ENEM	Exame Nacional do Ensino Médio
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
SISU	Sistema de Seleção Unificada
TRI	Teoria de Resposta ao Item
GAM	Modelos Aditivos Generalizados
GAMLSS	Modelos Aditivos Generalizados para Locação, Escala e Forma
BCT	Distribuição Box-Cox t
BEINF0	Distribuição Beta Inflacionada em Zero
CG	Algoritmo Cole and Green
RS	Algoritmo Rigby and Stasinopoulos
EMV	Estimativa de Máxima Verossimilhança

Lista de símbolos

Y	Variável Aleatória (representando a nota da redação)
α	Parâmetro de forma da distribuição Beta
β	Parâmetro de forma da distribuição Beta
β_k	Vetor de coeficientes (parâmetros) da regressão
$B(.)$	Função Beta
$\Gamma(.)$	Função Gama
μ	Parâmetro de Locação ou média (em todas as distribuições)
ϕ	Parâmetro de precisão (na reparametrização da Beta)
σ	Parâmetro de escala ou dispersão (no GAMLSS, BEINF0 e BCT)
ν	Parâmetro de chance (odds) de zero (na BEINF0); Parâmetro de assimetria (na BCT)
τ	Parâmetro de curtose (na BCT)
p_0	Probabilidade de uma observação ser zero
π_i	Probabilidade de uma observação ser maior que zero (no Modelo de Barreira)
$\mathrm{E}(Y)$	Valor Esperado (média) da variável Y
$\mathrm{Var}(Y)$	Variância da variável Y
$f(y;.)$	Função Densidade de Probabilidade (p.d.f)
$g_k(.)$	Função de Ligação (no GAMLSS)
η_k	Preditores lineares (no GAMLSS)
$l(.)$	Função de log-verossimilhança

Sumário

Introdução

O Exame Nacional do Ensino Médio (ENEM), instituído em 1998 pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP, 2025), visava inicialmente avaliar o desempenho dos estudantes ao término da educação básica. Entretanto, a partir de uma profunda reformulação metodológica em 2009, o ENEM transcendeu sua função avaliativa e consolidou-se como o principal instrumento de acesso ao ensino superior no Brasil. Seus resultados passaram a ser utilizados como critério de seleção unificado para instituições federais, por meio do Sistema de Seleção Unificada (SISU), e como requisito para programas governamentais de bolsas e financiamento, como o Programa Universidade para Todos (PROUNI) e o Fundo de Financiamento Estudantil (FIES) (TERAMATSU; STRAFORINI, 2022).

Atualmente, o ENEM é composto por quatro provas objetivas — Ciências Humanas e suas Tecnologias, Ciências da Natureza e suas Tecnologias, Matemática e suas Tecnologias, e Linguagens, Códigos e suas Tecnologias — e uma Redação. Esta última, única prova discursiva, é avaliada com base em cinco competências, cada uma valendo até 200 pontos, totalizando uma nota máxima de 1000 pontos. Diferentemente das provas objetivas, cujas notas são calculadas pela Teoria de Resposta ao Item (TRI), a Redação permite que o participante atinja a pontuação máxima ou mínima (0) de forma direta. Essa característica confere à redação um caráter estratégico e decisivo, pois uma nota alta pode compensar eventuais dificuldades nas provas objetivas e frequentemente atua como fator determinante na classificação final dos candidatos.

No contexto da prova de Redação, cuja nota varia no intervalo [0, 1000], observa-se uma proporção expressiva de notas iguais a zero, fenômeno conhecido como inflação de zeros. Essa característica demanda o uso de modelos estatísticos capazes de lidar com a natureza mista da distribuição — composta por uma massa pontual em zero e uma componente contínua ao longo do restante do intervalo. Além disso, embora com frequência muito menor, também podem ocorrer notas máximas (1000 pontos). Quando as notas são reescaladas para o intervalo (0, 1), como é comum em modelos baseados na distribuição Beta, esses valores extremos geram um segundo tipo de concentração, denominada inflação de uns. Assim, a análise adequada dessas notas requer modelos que considerem simultaneamente a presença de valores inflacionados em zero (e, eventualmente, em um) e uma parte contínua limitada entre esses extremos.

A literatura recente tem recorrido a modelos de regressão beta inflacionados em zero e/ou um para analisar tais dados. Estudos mais recentes, como os de RÊGO (2021) e LIMA *et al.* (2023), avançaram na aplicação direta desses modelos, utilizando os Modelos Aditivos Generalizados para Localização, Escala e Forma (GAMLSS) para identificar e quantificar fatores que afetam a distribuição das notas em diferentes estados brasileiros.

Uma abordagem alternativa para lidar com a inflação de zeros são os Modelos de Barreira

(MULLAHY ,1986; HEILBRON,1994). Esta metodologia pressupõe que a geração dos dados ocorre em duas etapas: primeiro, um processo binário modela a superação ou não da "barreira"(o evento de nota zero); segundo, um processo distinto modela a magnitude do resultado, condicionado à superação da barreira (a nota, dado que é positiva). No contexto deste trabalho, a barreira é o evento "zerar a redação". Embora essa modelagem seja tradicionalmente aplicada a dados de contagem, ela é extensível a dados de natureza contínua, permitindo a livre escolha da distribuição para o componente pós-barreira.

Nessa linha de investigação, o presente trabalho tem como objetivo primário, identificar os fatores que impactam a distribuição das notas da redação do ENEM para estudantes de escolas públicas e privadas. Contudo, o trabalho avança ao propor um objetivo secundário de caráter metodológico: comparar a adequação e a robustez de diferentes estratégias de modelagem frente a dados de complexidades variadas. Especificamente, busca-se avaliar se o modelo Beta Inflacionado em Zero (BEINF0) (MARTINEZ, 2008), embora plausível, mantém seu bom desempenho ao generalizar de um cenário específico (o município de Goiânia em 2023) para um mais amplo e heterogêneo (o estado de Goiás entre 2021 e 2023), em comparação com a flexibilidade de um Modelo de Barreira.

Para alcançar esses objetivos, a análise será fundamentada em duas populações distintas, extraídas dos microdados do ENEM obtidos junto ao INEP. A primeira população, mais restrita (participantes de Goiânia no ano de 2023), e a segunda população mais ampla (participantes do estado de Goiás durante os anos de 2021, 2022 e 2023). Esta estratégia dual permitirá investigar não apenas quais variáveis são relevantes, mas também qual abordagem de modelagem (Beta Inflacionado ou Barreira) oferece o ajuste mais consistente e generalizável. As análises serão conduzidas no ambiente `R` (R Core Team, 2025), com implementação dos modelos via *framework* GAMLSS, e as variáveis consideradas incluirão notas nas provas objetivas do ENEM e características socioeconômicas e educacionais dos participantes.

Este trabalho está organizado da seguinte forma: após esta Introdução, a Fundamentação Teórica discute o histórico e a estrutura do ENEM, a importância estratégica da redação e uma revisão da literatura correlata. Na sequência, a Metodologia detalha a coleta e o processamento de dados para ambas as populações, as variáveis utilizadas, os critérios de exclusão e a abordagem analítica, com foco nas famílias de distribuições utilizadas, nos modelos GAMLSS e no algoritmo de estimação. Por fim, na seção de Resultados, são apresentados a análise descritiva, os ajustes dos modelos e os diagnósticos obtidos

1 Fundamentação Teórica

1.1 ENEM

O Exame Nacional do Ensino Médio (ENEM) foi criado em 1998 pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), órgão vinculado ao Ministério da Educação (MEC). Inicialmente, seu principal objetivo era de autoavaliação do desempenho dos estudantes ao término do ensino básico e servir também como ferramenta de análise para o sistema educacional, visando influenciar positivamente os currículos do ensino médio no Brasil (INEP, 2021).

Com o passar dos anos, o ENEM passou por transformações significativas, tanto em sua estrutura quanto em sua finalidade. Em 1999, o exame começou a ser utilizado como alternativa ao vestibular em 93 instituições de ensino superior. A partir de 2004, as notas do ENEM passaram a ser aceitas para a concessão de bolsas de estudo parciais e integrais em instituições privadas por meio do Programa Universidade para Todos (PROUNI) (INEP, 2020).

A reformulação mais expressiva ocorreu em 2009, quando o ENEM foi redesenhado para servir como principal instrumento de acesso ao ensino superior público. Essa nova versão, cuja estrutura central permanece em vigor, viabilizou o ingresso em universidades públicas por meio do Sistema de Seleção Unificada (SISU). A consolidação do ENEM como porta de entrada principal para todas as instituições públicas de ensino superior aconteceu em 2013 (INEP, 2020).

Novas mudanças foram implementadas em 2017. Após consultas públicas, o INEP passou a aplicar o exame em dois domingos consecutivos, com a prova de Redação sendo realizada no primeiro dia (INEP, 2020).

Desde a reformulação de 2009, o ENEM é composto por quatro provas objetivas, distribuídas nas seguintes áreas do conhecimento:

- Ciências Humanas e suas Tecnologias
- Ciências da Natureza e suas Tecnologias
- Matemática e suas Tecnologias
- Linguagens, Códigos e suas Tecnologias

Além dessas provas, há também a Redação, a única parte discursiva do exame. Ela é avaliada com base em cinco competências, cada uma valendo até 200 pontos, totalizando uma nota final de até 1000 pontos. As competências são:

- Domínio da norma padrão da língua escrita.
- Compreensão da proposta e aplicação dos conhecimentos para desenvolver o tema dentro dos limites estruturais do texto dissertativo-argumentativo.
- Capacidade de selecionar, relacionar e organizar argumentos coerentes e coesos.
- Demonstração de conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.
- Elaboração de proposta de intervenção para o problema abordado, respeitando os direitos humanos.

A correção da Redação é feita por dois avaliadores independentes, que atribuem notas para cada competência. Caso a diferença entre as notas dos corretores ultrapasse 100 pontos na nota total, ou 80 pontos em qualquer uma das competências, um terceiro corretor é acionado. Persistindo a discrepância, a redação é encaminhada para uma banca composta por especialistas (INEP, 2023b).

Diferentemente das provas objetivas, cujas notas são calculadas por meio da Teoria de Resposta ao Item (ANDRADE; TAVARES; VALLE, 2000), a Redação do ENEM não segue esse modelo. Por isso, é a única parte do exame onde o participante pode atingir a pontuação máxima (1000) ou mínima (0) de forma direta.

Essa característica torna a Redação um componente estratégico na busca por uma vaga no ensino superior. Uma nota alta pode compensar eventuais dificuldades nas provas objetivas e, muitas vezes, representa o diferencial decisivo na classificação dos candidatos. Por esse motivo, tanto estudantes quanto educadores costumam dedicar atenção especial à preparação para essa parte do exame.

1.2 Goiás

A seleção do estado de Goiás como unidade de análise para este estudo é intencional e se justifica pelo o estado representar um dos microcosmos de desigualdades estruturais do Brasil (ROLIM *et al.*, 2022), oferecendo um conjunto de dados ideal para os objetivos desta pesquisa.

Conforme será demonstrado, a heterogeneidade do estado entre as redes pública e privada produz uma distribuição de notas no ENEM com características estatísticas específicas. Notadamente, a ocorrência de um volume de notas zero coexiste com notas de alta performance. Este cenário torna o recorte goiano um laboratório ideal para a aplicação e comparação dos modelos estatísticos, que são o foco deste trabalho.

Para ilustrar o contexto que fundamenta essa escolha, é crucial analisar seus indicadores educacionais. Um dos principais é o Índice de Desenvolvimento da Educação Básica (IDEB).

Criado pelo INEP, o IDEB sintetiza em um único número dois fatores cruciais: a taxa de aprovação dos alunos (o fluxo escolar) e o desempenho médio em exames padronizados (o Saeb). Ao analisar o histórico do estado, o IDEB aponta para uma disparidade de desempenho entre as redes de ensino. Na edição de 2023, por exemplo, a diferença registrada entre as escolas privadas e públicas foi de 0,6 pontos (INEP, 2023a). Embora essa métrica seja apenas um dos indicadores, ela sinaliza a existência de um hiato de qualidade que justifica uma análise mais aprofundada.

As sinopses estatísticas do ano de 2023 (INEP, 2024) revelam a estrutura dessa heterogeneidade. Em Goiás, dos 60.706 estudantes matriculados na 3ª série do ensino médio, a vasta maioria (51.195 alunos, ou 84,3%) pertencia à rede pública. Contudo, ao isolar a capital, Goiânia, essa proporção muda: dos 12.330 alunos da 3ª série, 7.637 (ou 61,9%) estavam na rede pública. Isso demonstra que, enquanto o estado como um todo é dominado pela rede pública, a capital concentra uma participação da rede privada significativamente maior (38,1%) em comparação ao interior (15,7%).

Essa estrutura permite a este trabalho uma análise comparativa robusta. Goiânia constitui um subgrupo que, embora diverso, é menos heterogêneo que o estado completo. A existência dessas duas populações (Goiás total e Goiânia) permitirá avaliar não apenas a adequação dos modelos, mas também a consistência de suas estimativas frente a populações com diferentes graus de homogeneidade. A distribuição das notas de redação do ENEM nesse contexto, por definição heterogênea, justifica plenamente a aplicação de modelos estatísticos avançados, como o BEINF0 e o Modelo de Barreira (*Hurdle Model*), capazes de capturar adequadamente essa complexa estrutura de variabilidade.

1.3 Trabalhos Correlatos

A literatura recente apresenta diversos esforços para modelar dados de desempenho educacional com características de inflação de zeros.

Em abordagens similares, mas focadas em desempenho em Matemática, ALBUQUERQUE (2017) e LOBO; CASSUCE; CIRINO (2017) aplicaram modelos hierárquicos e de efeitos aleatórios para analisar o impacto de características socioeconômicas e do tipo de escola. Em um contexto distinto, OLIVEIRA *et al.* (2017) empregou um modelo de regressão inflacionado de zero para avaliar a eficiência de escolas públicas no estado de Goiás, demonstrando a viabilidade de modelos que lidam com massa de zeros, embora não especificamente para o contexto da redação.

Trabalhos mais recentes avançaram na aplicação direta à prova de redação. RÊGO (2021) e LIMA *et al.* (2023) utilizaram modelos de regressão beta inflacionados em zero para identificar e quantificar fatores que afetam a distribuição das notas da redação nos estados do Rio Grande do Norte e Ceará, respectivamente.

1.4 Modelo BEINF0

Dentre as abordagens estatísticas para dados limitados a um intervalo, o Modelo de Regressão Beta Inflacionado em Zero (BEINF0) tem recebido atenção crescente. Este modelo é especificamente desenhado para lidar com variáveis cuja resposta está limitada ao intervalo semiaberto [0, 1), ou seja, valores que podem ser exatamente zero, mas também assumem valores contínuos entre zero e um.

Conceitualmente, o BEINF0 é um modelo de mistura que assume a existência de dois processos distintos que geram os dados observados. Para ser aplicado a notas como as da redação do ENEM, que originalmente variam de 0 a 1000, exige-se primeiro uma transformação linear para que os escores se situem no intervalo [0, 1], como a divisão por 1000.

O primeiro processo do modelo é discreto, geralmente modelado por uma regressão logística. Ele responde à pergunta: "Este participante recebeu uma nota zero?". Este componente estima a chance (odds) de um aluno obter um zero "verdadeiro", que no contexto do ENEM pode significar fuga ao tema, cópia, ou não atendimento à estrutura dissertativa.

O segundo processo é contínuo e entra em ação apenas para os participantes que não receberam a nota zero. Para este grupo, o modelo utiliza uma Regressão Beta padrão para explicar a variabilidade das notas no intervalo (0, 1). A distribuição Beta é reconhecida por sua flexibilidade, sendo capaz de se adaptar a diferentes formatos de distribuição de dados.

A adequação deste modelo para contextos educacionais, especialmente para as notas do ENEM, é bem documentada. Trabalhos como os de LIMA *et al.* (2023) e RÊGO (2021) como citado anteriormente. Estes estudos validam o BEINF0 como uma ferramenta robusta, pois a nota de redação do ENEM é, por natureza, um dado com inflação de zeros.

No contexto deste estudo, a escolha do BEINF0 se alinha diretamente à justificativa apresentada na seção anterior. O cenário de Goiás, com sua alta heterogeneidade e consequente volume de notas nulas, exige um modelo que possa tratar o zero não como uma nota baixa qualquer, mas como um evento distinto, ao mesmo tempo em que modela com flexibilidade as notas positivas.

1.5 Modelo de Barreira

Uma abordagem alternativa para a modelagem de dados com excesso de zeros é o Modelo de Barreiras, ou *Hurdle Model*. Proposto originalmente em contextos econométricos por CRAGG (1971), este modelo também é estruturado em dois componentes, mas parte de uma premissa conceitual sutilmente diferente daquela do BEINF0.

O Modelo de Barreira assume que os dados são gerados por um processo sequencial de duas etapas:

1. A Barreira (O Processo Binário), que determina se a observação é um "zero"ou um valor "positivo". Ele modela a chance (odds) de um participante "cruzar a barreira", ou seja, conseguir produzir uma nota válida (superior a zero).

2. A Intensidade, apenas para as observações que cruzaram a barreira ($Y > 0$). Este componente utiliza uma distribuição de probabilidade truncada em zero para modelar o nível ou a magnitude da nota positiva.

A principal diferença conceitual em relação ao BEINF0 é como o zero é tratado. No BEINF0, o modelo é uma mistura de zeros verdadeiros (do componente logístico) e zeros potenciais (do limite inferior da distribuição Beta). No Modelo de Barreiras, a separação é estrita: os zeros são gerados apenas pelo primeiro componente (a barreira), e o segundo componente (a distribuição truncada) é, por definição, incapaz de gerar zeros.

A grande vantagem e relevância do Modelo de Barreiras para os dados heterogêneos de Goiás está na sua flexibilidade interpretativa. O modelo assume que o processo que leva um aluno a zerar a redação (ex: fuga ao tema, não letramento) pode ser fundamentalmente distinto do processo que define a nota daqueles que escreveram um texto válido.

Embora seu uso em avaliações educacionais seja menos documentado que o do BEINF0, a sua aplicação clássica em econometria (MULLAHY, 1986) justifica sua adaptação metodológica neste trabalho. A inclusão do Modelo de Barreiras serve como um contraponto teórico-metodológico ao BEINF0, permitindo comparar qual estrutura, a de mistura (BEINF0) ou a de barreira (*Hurdle*), oferece o ajuste mais plausível para explicar a complexa distribuição das notas do ENEM em Goiás.

2 Metodologia

2.1 Base de Dados e população

A pesquisa segue uma abordagem quantitativa, combinando análises descritivas e explicativas e uma avaliação comparativa de diferentes estratégias de modelagem. O objetivo central é aplicar técnicas de modelagem estatística para mensurar e analisar o impacto de variáveis socioeconômicas e educacionais no desempenho dos candidatos na prova de redação do Exame Nacional do Ensino Médio (ENEM). A escolha deste delineamento metodológico justifica-se pela necessidade de quantificar relações e identificar padrões que expliquem as variações nas notas, permitindo a aplicação de métodos inferenciais para generalizar os achados.

A abordagem metodológica adotada neste trabalho alinha-se aos estudos de RÊGO (2021) e LIMA *et al.* (2023), empregando modelos beta inflacionados em zero e extrapola para estratégia de modelos de barreira. O objetivo é identificar fatores relevantes que impactam a distribuição das notas da redação do ENEM para escolas estaduais e privadas de Goiás e comparar a qualidade do ajuste nas duas estratégias abordadas.

A modelagem será conduzida utilizando o *framework* GAMLSS com a família de distribuição beta inflacionada em zero (BEINF0) e familias Binomial (BI) e Box-Cox t (BCT) para a estratégia de Modelo de Barreira. Estas escolhas oferecem maior flexibilidade ao permitir a modelagem simultânea de todos os parâmetros da distribuição. O *framework* permite, ainda, a utilização de funções de ligação (*link functions*) estratégicas, garantindo assim um ajuste mais preciso, robusto e interpretável aos dados.

A população-alvo deste estudo é definida como o conjunto de todos os estudantes concluintes do Ensino Médio da rede de ensino do estado de Goiás que participaram do ENEM nos últimos anos. Para investigar essa população, a pesquisa se baseia nos microdados disponibilizados pelo TEIXEIRA (2024), utilizando as edições mais recentes com dados completos (2021, 2022 e 2023). A edição de 2024 não foi incluída em virtude de limitações na divulgação dos microdados, que impossibilitaram o cruzamento de resultados com informações socioeconômicas dos estudantes e das escolas devido a lei geral de proteção de dados.

Para atender aos objetivos propostos, a análise será conduzida em duas frentes complementares, utilizando duas populações estratégicas:

- Uma de escopo específico, focada nos participantes de Goiânia na edição de 2023, para testar o ajuste inicial dos modelos em um cenário mais homogêneo.
- Uma de escopo amplo, que agrega os dados de todo o estado de Goiás ao longo das edições de 2021, 2022 e 2023, para avaliar a robustez e a generalização dos modelos frente a uma

maior variabilidade de dados.

Esta abordagem dual permite não apenas descrever os fatores que influenciam as notas, mas também comparar a adequação de diferentes modelos estatísticos em contextos distintos, conforme detalhado nas seções subsequentes.

2.1.1 Processamento e Integração dos Dados

A montagem da base de dados consolidada foi realizada no ambiente R (versão 4.3.1) (R Core Team, 2025), com apoio de pacotes especializados em manipulação de dados, como *vroom* (HESTER; WICKHAM, 2023), para importação eficiente de arquivos CSV e funções do ecossistema *tidyverse* (WICKHAM; GROLEMUND, 2019), que otimizam tarefas de filtragem, transformação e junção de dados. Esse *pipeline* assegurou a reprodutibilidade do processo e o tratamento adequado do volume de dados. Ao final do processamento inicial, a base de dados da população considerando Goiânia em 2023 totalizou 6834 registros (3608 de escolas públicas e 3226 de escolas privadas), enquanto a população considerando Goiás em 2021, 2022 e 2023, 68604 registros (53330 e 15274 de escolas privadas).

Na etapa de pré-processamento dos dados, a variável resposta (nota da redação) e todas as covariáveis de natureza contínua foram normalizadas para o intervalo [0, 1]. Este procedimento consistiu na divisão dos valores originais por 1000. Consequentemente, todas as estatísticas descritivas (e.g., média, desvio padrão) e os coeficientes estimados pelos modelos referem-se a esta escala transformada e devem ser interpretados como tal.

2.1.2 Critérios de Exclusão

Os seguintes critérios de exclusão foram aplicados de forma sequencial e idêntica a ambas as populações, com o objetivo de garantir a qualidade, a consistência e a comparabilidade dos dados:

1. **Filtro por Vínculo Escolar:** Selecionaram-se apenas registros de participantes do Ensino Médio, para garantir que a análise se concentre em estudantes no estágio final da educação básica, tornando os grupos comparáveis

2. **Exclusão de Treineiros e Ausentes:** Foram excluídos os registros de candidatos classificados como treineiros e aqueles que não compareceram a todas as provas objetivas do exame.

3. **Tratamento de Dados Ausentes:** Procedeu-se à exclusão de todos os registros que apresentavam dados ausentes em qualquer uma das variáveis selecionadas para a análise, o que resultou na remoção de 1205 registros de Goiânia em 2023 e 12060 registros de Goiás todo nos anos de 2021, 2022 e 2023.

4. **Filtragem de caso raro** Na população de Goiás havia um único participante com nota máxima (1000). O modelo utilizado, BEINF0, considera apenas a inflação em zero e não acomoda automaticamente valores discretos em 1000. De forma semelhante, no modelo de barreira, a presença de um único valor extremo poderia influenciar desproporcionalmente a estimação do componente contínuo. Embora fosse teoricamente possível utilizar um modelo Beta inflacionado em zero e um (BEINF01), a presença de apenas um único indivíduo com nota máxima não seria suficiente para estimar de forma confiável o parâmetro de inflação em 1. Por isso, esse registro foi considerado um outlier e removido para o ajuste dos modelos.

Após a aplicação de todos os critérios, as populações finais da análise ficaram compostas da seguinte forma:

- **População 1 (Goiânia 2023):** 6834 registros, sendo 3608 (52,80%) de escolas publicas e 3226 (47,21%) de escolas privadas.
- **População 2 (Goiás 2021-2023):** 68604 registros, sendo 53330 (77,74%) de escolas públicas e 15274 (22,26%) de escolas privadas.

2.1.3 Variáveis Utilizadas

Tabela 1 – Descrição das variáveis utilizadas no modelo

Variável	**Tipo**	**Descrição**	**Categorias/Range**
NU_NOTA_REDACAO	Contínua	Nota na prova de Redação (Variável Resposta)	[0,1000]
NU_NOTA_MT	Contínua	Nota na prova objetiva de Matemática	[0,983.2]
NU_NOTA_CH	Contínua	Nota na prova objetiva de Ciências Humanas	[0,839.2]
NU_NOTA_CN	Contínua	Nota na prova objetiva de Ciências da Natureza	[0,868.7]
NU_NOTA_LC	Contínua	Nota na prova objetiva de Linguagens e Códigos	[0,793.5]
TP_SEXO	Categórica	Sexo do participante	F: Feminino M: Masculino
TP_COR_RACA	Categórica	Cor/Raca do participante	BRANCA PARDA PRETA OUTROS

Continua na próxima página

Tabela 1 – continuação da página anterior

Variável	Tipo	Descrição	Categorias
ESC_MAE	Categórica	Escolaridade máxima da Mãe do participante	NÃO-FUNDAMENTAL FUNDAMENTAL MÉDIO SUPERIOR
RENDA	Categórica	Renda familiar declarada pelo participante	CLASSE 1 CLASSE 2 CLASSE 3 CLASSE 4
PC	Binária	Indicadora de Computador na casa do participante	0: NÃO 1: SIM

É importante destacar que, no presente trabalho, tanto a variável resposta, quanto as covariáveis de tipo contínua foram escalonadas para o intervalo [0, 1], i.e divididas por mil, os resultados de média, desvio padrão e coeficientes dos modelos estão nessa escala.

2.2 BEINF0

2.2.1 Distribuição Beta

A distribuição Beta é uma distribuição de probabilidade contínua, comumente utilizada para modelar variáveis do tipo taxa, razão ou proporção, cujo suporte é tipicamente o intervalo aberto (0,1). É caracterizada por dois parâmetros positivos, α e β, que determinam sua forma.

Seja Y uma variável aleatória que siga a distribuição Beta, i.e Y $\sim$ Beta(α,β), sua Função Densidade de Probabilidade é comumente representada como

$$f(y;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1} \qquad \text{para } y \in (0,1),$$

onde $B(\alpha,\beta)$ é a função Beta, que pode ser expressa pela função Gama $\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

A média e a variância de uma distribuição Beta são dadas da seguinte forma

$$\begin{aligned} \mathrm{E}(Y) &= \frac{\alpha}{\alpha+\beta} \\ \mathrm{Var}(Y) &= \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}. \end{aligned}$$

É possível fazer uma reparametrização desta distribuição, que foca na média da variável resposta (μ) e em um parâmetro de precisão (ϕ), onde $\mu = \frac{\alpha}{\alpha+\beta}$ e $\phi = \alpha + \beta$ (FERRARI; CRIBARI-NETO, 2004). Se $Y \sim$ Beta(μ,ϕ), então sua f.d.p. é expressa por:

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

dessa forma, a média e a variância passam a ser

$$\begin{aligned} \mathrm{E}(Y) &= \mu \\ \mathrm{Var}(Y) &= \frac{\mu(1-\mu)}{\phi+1}. \end{aligned}$$

É importante notar que esta reparametrização em termos da média (μ) e da precisão (ϕ) é conceitualmente muito utilizada na literatura. Contudo, para a modelagem no *framework* do pacote `gamlss` com a familia BEINF0 adota uma parametrização alternativa para a dispersão, que também deriva dos parâmetros originais α e β, conforme será detalhado.

2.2.2 Distribuição Beta Inflacionada em Zero

Embora a distribuição Beta seja uma alternativa viável para modelagem de vários problemas envolvendo taxas, graças a sua flexibilidade, é notável que ela acaba por não se encaixar bem em muitas das situações, devido a seu suporte restrito ao intervalo (0,1). Pensando nisto, foi proposta por Ospina (OSPINA; FERRARI, 2010) uma mistura de uma distribuição Beta e uma distribuição degenerada em zero, que originou a distribuição Beta Inflacionada em Zero.

Para a implementação no pacote `gamlss`, a família BEINF0 é descrita por Rigby (RIGBY *et al.*, 2019) através de uma reparametrização dos parâmetros clássicos da Beta (α, β) e da probabilidade de zero (p_0), tendo sua densidade expressada por

$$f(y;\alpha,\beta,p_0) = \begin{cases} p_0, & \text{se } y = 0 \\ (1-p_0)\,\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, & \text{se } y \in (0,1) \end{cases}$$

A distribuição Beta Inflacionada em Zero é notavelmente flexível, podendo assumir diversos formatos para se ajustar aos dados, como formas em J, U ou unimodal. Uma propriedade estatística importante, destacada por Ospina (OSPINA; FERRARI, 2010), é a independência assintótica entre o parâmetro de inflação de zeros e os parâmetros da componente Beta, o que simplifica a estimação e a interpretação do modelo.

2.2.3 Modelo de Regressão Beta Inflacionado em Zero

A distribuição Beta Inflacionada em Zero serve de base para os Modelos de Regressão Beta Inflacionados em Zero, que são frequentemente enquadrados na classe dos Modelos Aditivos Generalizados para Localização, Escala e Forma (GAMLSS). Nos GAMLSS, é possível modelar

todos os parâmetros da distribuição da variável resposta (μ, σ,ν) por meio de funções de ligação (como logito para μ e σ, e logarítmica para ν) e variáveis preditoras. Isso permite que a média, a dispersão e a probabilidade de ocorrência de zeros sejam influenciadas por características das observações

Para a implementação no pacote `gamlss`, os parâmetros α e β são reparametrizados em termos de uma média condicional μ e um parâmetro de dispersão σ, onde:

- **A média condicional,** $\mu = \frac{\alpha}{\alpha+\beta}$: Representa o valor esperado da variável resposta, dado que ela é maior que zero. Seu domínio é $0 < \mu < 1$.
- **O parâmetro de dispersão,** $\sigma = (\alpha + \beta + 1)^{-1/2}$: Um parâmetro que controla a variabilidade da componente Beta da distribuição. Valores de σ próximos de zero indicam baixa dispersão (alta precisão), enquanto valores próximos de 1 indicam alta dispersão. Seu domínio é $0 < \sigma < 1$.
- **A chance (odds) de zero,** $\nu = \frac{p_0}{1-p_0}$: Representa a razão entre a probabilidade de uma observação ser zero e a probabilidade de não ser zero. A probabilidade p_0 pode ser recuperada pela fórmula $p_0 = \frac{\nu}{1+\nu}$. Seu domínio é $0 < \nu < \infty$.

Com base nesta parametrização, a média e a variância globais (ou marginais) da variável aleatória $Y \sim \text{BEINF0}(\mu, \sigma, \nu)$ são dadas por:

$$\begin{aligned} \mathrm{E}(Y) &= \frac{\mu}{1+\nu} \\ \mathrm{Var}(Y) &= \frac{\sigma^2\mu(1-\mu) + \mu^2(1+\nu)^{-1}}{1+\nu} \end{aligned}$$

2.2.3.1 Estrutura do Modelo

A principal vantagem da abordagem GAMLSS é que cada um dos três parâmetros da distribuição BEINF0 (μ, σ e ν) é modelado explicitamente em função de um conjunto de variáveis explicativas. Conforme a estrutura detalhada em MARTINEZ (2008) e adaptada para a parametrização do `gamlss`, o modelo de regressão é composto por três sub-modelos simultâneos. Para uma observação i:

- **Parâmetro de Locação (μ_i):** Modelado por $g_1(\mu_i) = \mathbf{x}_{1i}^T\boldsymbol{\beta}$, onde g_1 é uma função de ligação que mapeia o intervalo (0,1) para os números reais, tipicamente a função logit: $g_1(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$. O vetor $\mathbf{x}_{1i}$ corresponde às covariáveis para a média e $\boldsymbol{\beta}$ é o vetor de coeficientes de regressão. Estes coeficientes (β_k) medem o efeito das covariáveis em $\mathbf{x}_1$ sobre a média da variável resposta, dado que ela é positiva ($E[Y|Y > 0]$).

- **Parâmetro de Escala (σ_i):** Modelado por $g_2(\sigma_i) = \mathbf{x}_{2i}^T\boldsymbol{\delta}$, onde g_2 também é comumente uma função de ligação logit, dado que $0 < \sigma < 1$. O vetor $\mathbf{x}_{2i}$ corresponde às covariáveis para a dispersão (que pode ser igual ou diferente de $\mathbf{x}_{1i}$) e $\boldsymbol{\delta}$ é o vetor de coeficientes. Estes coeficientes (δ_k) medem o efeito das covariáveis em $\mathbf{x}_2$ sobre a dispersão da componente Beta.

- **Parâmetro de Forma (ν_i):** Modelado por $g_3(\nu_i) = \mathbf{x}_{3i}^T\boldsymbol{\gamma}$, onde g_3 é uma função de ligação que mapeia o intervalo $(0, \infty)$ para os reais, sendo a função logarítmica a escolha padrão: $g_3(\nu_i) = \log(\nu_i)$. O vetor $\mathbf{x}_{3i}$ corresponde às covariáveis para a inflação de zeros e $\boldsymbol{\gamma}$ é o vetor de coeficientes. Estes coeficientes (γ_k) medem o efeito das covariáveis em $\mathbf{x}_3$ sobre a chance (odds) de a resposta ser zero.

2.2.3.2 Estimação dos Parâmetros

A estimação dos vetores de parâmetros $\boldsymbol{\theta}$ =($\boldsymbol{\beta}$, $\boldsymbol{\delta}$ e $\boldsymbol{\gamma}$) é realizada através do Método da Máxima Verossimilhança. A função de log-verossimilhança $l(\boldsymbol{\theta}; \mathbf{y})$ para o modelo é a soma das contribuições de cada observação y_i, com base na densidade $f(y_i; \mu_i, \sigma_i, \nu_i)$ definida anteriormente, onde os parâmetros são substituídos por suas respectivas estruturas de regressão. Como destacado em MARTINEZ (2008), as equações de estimação resultantes da derivação da função de log-verossimilhança são não-lineares e não possuem uma solução analítica de forma fechada. Portanto, os estimadores de máxima verossimilhança devem ser obtidos por meio de algoritmos numéricos iterativos. MARTINEZ (2008) fornece as expressões analíticas para o vetor escore e a matriz de informação de Fisher esperada, que são a base para algoritmos de otimização como o Escore de Fisher (NELDER; WEDDERBURN, 1972). No contexto do pacote `gamlss`, a estimação é realizada utilizando o algoritmo `RS` (Rigby e Stasinopoulos) RIGBY; STASINOPOULOS (2005), que é uma generalização do algoritmo de Escore de Fisher, permitindo ajustar todos os sub-modelos (para μ, σ, ν) simultaneamente.

2.3 Modelo de Barreira (*Hurdle Model*)

O Modelo de Barreira, também conhecido como *Hurdle Model*, representa uma abordagem alternativa para lidar com dados que contêm uma proporção excessiva de zeros. Proposto inicialmente por CRAGG (1971), este modelo conceitualiza a geração dos dados como um processo de duas etapas, separando a decisão de ocorrência do evento da magnitude desse evento.

Em contraste com os modelos inflacionados em zero, que assumem que os zeros podem surgir de dois processos distintos (um estrutural e outro amostral da distribuição contínua/discreta), o Modelo de Barreira postula que todos os zeros são gerados por um único processo. Uma barreira inicial deve ser superada para que um valor positivo seja observado. Se a barreira não for superada, o resultado é um zero; se for, o valor da observação é determinado por um segundo processo, modelado por uma distribuição de probabilidade truncada em zero.

A escolha entre a abordagem de Barreira e a Inflacionada em Zeros é uma etapa metodológica crucial e depende da fundamentação teórica sobre o processo gerador dos zeros, sendo um tema central na literatura estatística, como visto em FENG (2021) ou em PITSHA; CHIRUKA; MARANGE (2025). A principal vantagem da abordagem de barreiras é a sua interpretação conceitual, que permite uma separação completa entre os fatores que influenciam a presença ou ausência de uma resposta (a barreira) e os fatores que influenciam a magnitude da resposta, dado que ela existe. Esta característica torna o modelo particularmente atraente para fenômenos onde o zero é um estado qualitativamente distinto, e não apenas um valor baixo.

No contexto da redação do ENEM, podemos pensar que a atribuição de nota zero não representa o extremo inferior de uma escala de proficiência, mas sim o resultado de um conjunto de eventos eliminatórios e binários. De acordo com o edital do exame, um candidato recebe nota zero por razões como: fuga total ao tema, não atendimento ao tipo textual dissertativo-argumentativo, folha de redação em branco, texto insuficiente (abaixo de um número mínimo de linhas), cópia de textos motivadores, entre outros (INEP, 2023c). Tecnicamente é possível que um participante tire nota zero apenas pontuando 0 em todas as competências, entretanto, é muito improvável um cenário onde todas as competências da redação sejam pontuadas como zero e o participante não tenha atendido a nenhum dos requisitos de anulação definidos pelo edital.

2.3.1 O Processo de Decisão Binária (A Barreira)

A primeira etapa consiste em um modelo binário que governa a probabilidade de uma observação ser positiva em oposição a ser nula. Essencialmente, este processo modela $P(Y_i > 0|\mathbf{x}_{1i})$, onde Y_i é a variável resposta e $\mathbf{x}_{1i}$ é um vetor de covariáveis que podem influenciar a travessia da barreira.

Seja $\pi_i = P(Y_i > 0|\mathbf{x}_{1i})$ a probabilidade de a i-ésima observação ser maior que zero. A escolha mais comum para modelar essa probabilidade é o modelo de regressão logística, devido à sua simplicidade e interpretação direta em termos de chances (*odds*). A relação é definida por:

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_{1i}'\boldsymbol{\gamma} \quad \pi_i \in (0,1),$$

onde:

- π_i é a probabilidade de sucesso, isto é, $Y_i > 0$.
- $\mathbf{x}_{1i}$ é o vetor de variáveis preditoras para a etapa binária.
- $\boldsymbol{\gamma}$ é o vetor de coeficientes a serem estimados para a parte binária.

Este componente do modelo determina, portanto, se a variável resposta assume um valor positivo, mas não informa nada sobre qual será esse valor.

2.3.2 O Processo de Intensidade (Distribuição BCT)

Uma vez que a barreira é superada ($Y_i > 0$), a segunda etapa do modelo entra em ação para descrever a magnitude da resposta. Este processo é modelado por uma distribuição de probabilidade contínua, que é explicitamente truncada à esquerda em zero para garantir que apenas valores positivos possam ser gerados. A probabilidade de este evento ocorrer é π_i, e a probabilidade de anulação ($Y_i = 0$) é $1 - \pi_i$.

Para este trabalho, a distribuição escolhida para o processo de intensidade é a distribuição Box-Cox t (RIGBY; STASINOPOULOS, 2006), que é uma extensão da distribuição Box-Cox Normal. Esta escolha é justificada pela sua notável flexibilidade, sendo capaz de acomodar assimetria e curtose (caudas pesadas) nos dados. A BCT é uma distribuição de quatro parâmetros, o que a torna ideal para o *framework* GAMLSS.

Para as observações que superam a barreira, é necessário escolher uma distribuição de probabilidade que descreva bem sua dispersão, assimetria e curtose. A escolha da distribuição Box-Cox t (BCT), implementada via GAMLSS, se baseia nos argumentos apresentados nas seções seguintes.

2.3.2.0.1 1. Flexibilidade para Capturar a Forma da Distribuição:

A principal vantagem da BCT reside em seus quatro parâmetros (μ, σ, ν, τ), que permitem modelar a locação, escala, assimetria e peso das caudas da distribuição de forma independente. A distribuição das notas da redação raramente é simétrica e frequentemente apresenta um acúmulo de notas em certas faixas, além de uma proporção não desprezível de notas muito altas ou muito baixas (caudas pesadas), fenômenos que a BCT é especialmente projetada para acomodar.

2.3.2.0.2 2. Endereçando a Incompatibilidade de Suporte:

É crucial reconhecer que, teoricamente, a distribuição BCT possui suporte no intervalo $(0, \infty)$, enquanto os dados das notas são limitados superiormente em 1000. Embora distribuições com suporte limitado, como a Beta, pareçam uma escolha mais natural, sua flexibilidade pode não ser suficiente para alguns contextos, levando a um ajuste insuficiente. A justificativa para o uso da BCT, mesmo com o suporte ilimitado, é pragmática e empiricamente fundamentada:

- **Raridade do Evento de Nota Máxima:** A ocorrência da nota 1000 é um evento extremamente raro na população de participantes do ENEM. A densidade de observações próximas a este limite superior é, na prática, muito baixa. Portanto, a ausência de um limite superior formal na distribuição BCT tem um impacto prático negligenciável na qualidade do ajuste para a vasta maioria dos dados, que se concentram em regiões distantes deste extremo.
- **Priorização do Ajuste Empírico:** O objetivo principal é encontrar um modelo que descreva com a máxima fidelidade a distribuição empírica dos dados. A capacidade

superior da BCT em se moldar à assimetria e à curtose observadas nas notas pode superar a desvantagem teórica de seu suporte. O ganho em adequação do ajuste ao modelar o "corpo"e as caudas da distribuição compensa a pequena imprecisão no limite extremo e pouco povoado.

A densidade da distribuição BCT é obtida através do método da transformação de variáveis (ou mudança de variável). A lógica é que, se Z é uma variável aleatória com uma distribuição t-Student padrão, então a variável Y, que é uma transformação Box-Cox de Z, terá a densidade BCT. A fórmula geral para a mudança de variável é:

$$f_Y(y) = f_Z(z_y) \cdot \left| \frac{dz_y}{dy} \right|,$$

onde $f_Z(z_y)$ é a f.d.p. da distribuição t-Student padrão, z_y é a transformação Box-Cox inversa (expressa z em função de y) e $\left| \frac{dz_y}{dy} \right|$ é o Jacobiano da transformação (o valor absoluto da derivada de z_y em relação a y). Vamos definir cada um desses componentes.

Seja Z uma variável aleatória que segue a distribuição t-Student padrão com $\tau > 0$ graus de liberdade (o parâmetro de curtose). Sua função de densidade de probabilidade, f_t, é dada por:

$$f_t(z|\tau) = \frac{\Gamma\left(\frac{\tau+1}{2}\right)}{\sqrt{\pi\tau}\,\Gamma\left(\frac{\tau}{2}\right)} \left(1 + \frac{z^2}{\tau}\right)^{-\frac{\tau+1}{2}},$$

onde $\Gamma(\cdot)$ é a função Gama. Esta densidade $f_t(z|\tau)$ é simétrica em torno de zero e tem caudas mais pesadas que a distribuição Normal, controladas por τ.

A transformação Box-Cox possui duas formas, uma para quando o parâmetro de assimetria $\nu \neq 0$ e um caso limite para $\nu = 0$.

- Caso $\nu \neq 0$ A transformação inversa é $z_y = \frac{(y/\mu)^\nu - 1}{\sigma\nu}$. O Jacobiano (a derivada de z_y em relação a y) é expresso como $\left| \frac{dz_y}{dy} \right| = \frac{y^{\nu-1}}{\sigma\mu^\nu}$.
- Caso $\nu = 0$ A transformação inversa torna-se logarítmica $z_y = \frac{\log(y/\mu)}{\sigma}$. E o Jacobiano é expresso por $\left| \frac{dz_y}{dy} \right| = \left| \frac{1}{\sigma y} \right| = \frac{1}{\sigma y}$.

A função de densidade de probabilidade $f_{BCT}(y)$ para $y > 0$, com os quatro parâmetros (μ, σ, ν, τ), é obtida substituindo os componentes $f_t(z_y|\tau)$ e $\left| \frac{dz_y}{dy} \right|$ de volta na fórmula original. A densidade BCT é, portanto, definida por partes:

$$f_{BCT}(y|\mu,\sigma,\nu,\tau) = \begin{cases} \frac{1}{\sigma y} \cdot f_t\left(\frac{\log(y/\mu)}{\sigma} \middle| \tau\right) & \text{se } \nu = 0 \\ \\ \frac{y^{\nu-1}}{\sigma\mu^\nu} \cdot f_t\left(\frac{(y/\mu)^\nu - 1}{\sigma\nu} \middle| \tau\right) & \text{se } \nu \neq 0 \end{cases}$$

Onde $f_t(\cdot|\tau)$ é a densidade t-Student padrão definida anteriormente. Esta é a densidade utilizada pelo `gamlss` para estimar os parâmetros (μ, σ, ν, τ) por máxima verossimilhança, e é a densidade que, em nosso modelo de barreira, descreve o comportamento das notas Y_i dado que $Y_i > 0$. Mais detalhes da familia de distribuição BCT podem ser consultados em RIGBY; STASINOPOULOS (2006).

No contexto GAMLSS, cada um dos quatro parâmetros da distribuição BCT pode ser modelado como uma função de variáveis preditoras:

- **Parâmetro de Locação (μ):** Geralmente associado à mediana da distribuição. Pode ser modelado por $g_1(\mu_i) = \mathbf{x}'_{2i}\boldsymbol{\beta}$.
- **Parâmetro de Escala (σ):** Controla a dispersão dos dados, análogo a um coeficiente de variação. Pode ser modelado por $g_2(\sigma_i) = \mathbf{z}'_i\boldsymbol{\delta}$.
- **Parâmetro de Assimetria (ν):** É o parâmetro de transformação de Box-Cox, controlando a assimetria da distribuição. Pode ser modelado por $g_3(\nu_i) = \mathbf{w}'_i\boldsymbol{\zeta}$.
- **Parâmetro de Curtose (τ):** Representa os graus de liberdade da distribuição t subjacente, controlando o peso das caudas. Pode ser modelado por $g_4(\tau_i) = \mathbf{q}'_i\boldsymbol{\eta}$.

É importante notar que os conjuntos de covariáveis para cada um dos parâmetros ($\mathbf{x}_{1i}$, $\mathbf{x}_{2i}$, $\mathbf{z}_i$, $\mathbf{w}_i$, $\mathbf{q}_i$) podem ser diferentes, permitindo uma modelagem extremamente detalhada e flexível da estrutura dos dados. A combinação do Modelo de Barreira com a distribuição BCT oferece um *framework* que é, ao mesmo tempo, conceitualmente mais alinhado à avaliação do ENEM e metodologicamente mais flexível para se ajustar à complexa realidade dos dados.

2.4 GAMLSS

Os GAMLSS (*Generalized Additive Models for Location, Scale and Shape*) foram propostos por Rigby e Stasinopoulos (RIGBY; STASINOPOULOS, 2005) como uma alternativa mais flexível aos Modelos Lineares Generalizados (MLG) (NELDER; WEDDERBURN, 1972) e Modelos Aditivos Generalizados (GAM) (HASTIE; TIBSHIRANI, 1986)

Eles relaxam algumas das premissas para o ajuste dos modelos, notadamente a suposição de que a variável resposta deve pertencer à família exponencial de distribuições. Nos GAMLSS, a variável resposta pode seguir uma família de distribuição geral

A principal inovação é a capacidade de modelar explicitamente não apenas o parâmetro de localização (média), mas também os parâmetros de escala (variância/dispersão) e forma (assimetria e curtose) da distribuição da variável resposta. Cada um desses parâmetros (μ para localização, σ para escala, ν e τ para forma) pode ser modelado por meio de funções paramétricas, não-paramétricas, lineares, não-lineares ou aditivas de variáveis preditoras. Isso confere grande

flexibilidade para se ajustar a uma variedade de comportamentos de dados, incluindo aqueles com caudas pesadas ou leves, e assimetria positiva ou negativa

Para relacionar os preditores lineares com os parâmetros da distribuição, utilizam-se funções de ligação (g_k), que são funções monótonas conhecidas. Para familias de distribuições de até 4 parâmetros, obtemos os seguintes modelos

$$
\begin{aligned}
g_1(\mu_t) &= \eta_{t1} = \mathbf{x}_{t1}^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_p x_{tp} \\
g_2(\sigma_t) &= \eta_{t2} = \mathbf{s}_{t2}^\top \boldsymbol{\gamma} = \gamma_0 + \gamma_1 s_{t1} + \cdots + \gamma_l s_{tl} \\
g_3(\nu_t) &= \eta_{t3} = \mathbf{z}_{t3}^\top \boldsymbol{\lambda} = \lambda_0 + \lambda_1 z_{t1} + \cdots + \lambda_m z_{tm} \\
g_4(\tau_t) &= \eta_{t4} = \mathbf{w}_{t4}^\top \boldsymbol{\delta} = \delta_0 + \delta_1 w_{t1} + \cdots + \delta_m w_{tm},
\end{aligned}
$$

onde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$, $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_l)^\top$, $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \ldots, \lambda_m)^\top$ e $\boldsymbol{\delta} = (\delta_0, \delta_1, \ldots, \delta_m)^\top$ são os vetores de parâmetros desconhecidos a serem estimados. Os vetores $\mathbf{x}_{t1} = (1, x_{t1}, \ldots, x_{tp})^\top$, $\mathbf{s}_{t2} = (1, s_{t1}, \ldots, s_{tl})^\top$, $\mathbf{z}_{t3} = (1, z_{t1}, \ldots, z_{tm})^\top$ e $\mathbf{w}_{t4} = (1, w_{t1}, \ldots, w_{tm})^\top$ são os vetores de covariáveis (que incluem o intercepto) para a observação t, podendo conter as mesmas variáveis ou outras completamente diferentes.

2.4.1 Método de Estimação RS

Os algoritmos RS e CG (RIGBY; STASINOPOULOS, 2005) são dois dos métodos utilizados para maximizar a função de verossimilhança nos GAMLSS, e assim, obter as estimativas dos parâmetros (β_k). O algoritmo RS é o algoritmo padrão para estimação dos coeficientes no pacote `gamlss`, devido a isso, foi escolhido este método para estimação.

O algoritmo RS organiza o algoritmo em dois loops: iteração externa (*backfitting* por parâmetros) e iteração interna (atualização baseada em derivadas até convergência), com critérios de parada e limites ajustáveis

1. Iteração Externa: *Backfitting* por parâmetros

Inicia-se com um vetor de estimativas iniciais para todos os parâmetros ($\hat{\mu}$,$\hat{\sigma}$,$\hat{\nu}$,$\hat{\tau}$). Em cada ciclo, ajusta o modelo para um parâmetro de cada vez, mantendo os outros fixos em suas últimas estimativas. A ordem é μ, depois σ, ν e τ. Após ajustar todos os parâmetros em um ciclo, o Desvio Global (uma medida de ajuste do modelo, $-2l(\hat{\mu}$,$\hat{\sigma}$,$\hat{\nu}$,$\hat{\tau} \mid Y)$) é calculado. O processo se repete até que o Desvio Global convirja, indicando que as estimativas dos parâmetros alcançaram um máximo ou mínimo local

2. Iteração Interna: Newton–Raphson / Fisher Scoring

Esta etapa é acionada pela iteração externa para ajustar cada parâmetro (θ_k) individualmente. Ajusta repetidamente modelos ponderados por pesos (w_k) para uma variável resposta modificada (ou variável de trabalho) z_k, até a convergência. Este processo é análogo à estimação

iterativa pelos mínimos quadrados reponderados (IRLS) em MLG. A variável de trabalho (z_k) e os pesos (w_k) são derivados das primeiras e segundas derivadas da função de log-verossimilhança, respectivamente. Pesos negativos não são permitidos. A iteração interna incorpora métodos de afinamento (*finetuning*) para evitar que as estimativas saltem demais (*overjumping*), garantindo uma convergência mais suave e estável.

GAMLSS oferece um arcabouço flexível para modelagem estatística, e o algoritmo RS é um método robusto para estimar seus parâmetros, lidando eficientemente com a complexidade de modelar múltiplos aspectos da distribuição da variável resposta.

2.4.2 Seleção de Covariáveis

A complexidade dos modelos GAMLSS, como o BEINF0 e o BCT, reside no fato de que cada um dos seus parâmetros pode ser modelado em função de covariáveis. A seleção de quais variáveis incluir em cada submodelo simultaneamente é uma tarefa desafiadora.

No presente trabalho, adotou-se um procedimento de seleção stepwise automatizado, utilizando a função `stepGAICALL.A()` do pacote `gamlss` no *software* `R` (R Core Team, 2025). Esta função implementa um algoritmo de busca *forward* que seleciona o melhor conjunto de preditores com base no Critério de Informação de Akaike Generalizado (GAIC) (RIGBY; STASINOPOULOS, 2005).

O procedimento metodológico foi dividido em duas etapas principais, utilizando as populações de forma independente.

2.4.2.1 Descoberta e Seleção do Modelo (População 1)

O processo de seleção de variáveis foi conduzido exclusivamente com os dados de Goiânia em 2023. Para cada um dos modelos principais (BEINF0, *Hurdle*-Logístico e *Hurdle*-BCT), e separadamente para cada tipo de escola (Pública e Privada), o processo seguiu os seguintes passos:

- Iniciou-se com um modelo nulo, contendo apenas o intercepto para cada parâmetro a ser modelado.
- Um escopo de covariáveis candidatas (incluindo termos lineares e interações pré-definidas) foi especificado para cada parâmetro (μ, σ, ν, etc.).
- A função ‘stepGAICALL.A()‘ foi executada, testando iterativamente a adição de termos em cada submodelo e retendo apenas as variáveis que resultavam em uma melhoria no ajuste do modelo global, definida pela redução do valor do GAIC.

2.4.2.2 Validação e Reajuste (População 2)

Para validar a robustez dos achados e evitar o sobreajuste aos dados da população 1, bem como evitar deixar o método selecionar várias covariáveis com efeito irrelevante devido a alta dimensionalidade da população 2, o processo de seleção das variáveis não foi repetido.

Em vez disso, os modelos finais definidos na etapa anterior foram mantidos com o mesmo conjunto de covariáveis para cada parâmetro. Estes modelos foram, então, reajustados utilizando os dados de Goiás nos anos de 2021 à 2023. Os resultados e coeficientes apresentados nas seções seguintes são, portanto, referentes a estes modelos reajustados na população maior, permitindo avaliar a estabilidade dos efeitos encontrados.

2.4.3 Interpretação dos Parâmetros

A interpretação dos coeficientes nos GAMLSS é mais complexa do que em modelos lineares clássicos. O efeito de uma covariável não é, em geral, aditivo na escala original da resposta, mas sim na escala da função de ligação do parâmetro que está sendo modelado (seja μ, σ, ν ou τ). Neste trabalho, foram empregadas três funções de ligação distintas, cuja interpretação detalhamos a seguir.

2.4.3.0.1 1. Ligação Identidade

É análoga à regressão linear, a função de ligação é $g(\theta) = \theta$, de modo que $\theta_i = \mathbf{x}_i'\boldsymbol{\beta}$. Uma variação unitária na covariável X_j acarreta uma variação aditiva de β_j unidades no parâmetro θ, mantendo as demais covariáveis constantes.

2.4.3.0.2 2. Ligação Logarítmica (Log)

É usada para garantir que um parâmetro permaneça positivo (ex: $\sigma > 0, \tau > 0$). A função de ligação é $g(\theta) = \log(\theta)$, de modo que $\log(\theta_i) = \mathbf{x}_i'\boldsymbol{\beta}$, o que é equivalente a $\theta_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$. O efeito é multiplicativo sobre o parâmetro. Uma variação unitária na covariável X_j multiplica o valor do parâmetro θ pelo fator e^{β_j}. Se $\beta_j > 0$, $e^{\beta_j} > 1$, indicando um aumento percentual, se $\beta_j < 0$, $e^{\beta_j} < 1$, indicando um decréscimo percentual.

2.4.3.0.3 3. Ligação Logística (Logit)

É usada para parâmetros confinados ao intervalo $(0, 1)$, como μ e σ da distribuição Beta. A função de ligação é $g(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$, que modela o logaritmo da chance (log-odds) do parâmetro. O modelo é $\log\left(\frac{\theta_i}{1-\theta_i}\right) = \mathbf{x}_i'\boldsymbol{\beta}$. O coeficiente e^{β_j} é uma razão de chances (Odds Ratio). Uma variação unitária na covariável X_j multiplica a chance (odds) do parâmetro, (i.e., $\frac{\theta}{1-\theta}$), pelo fator e^{β_j}.

A Interpretação do parâmetro ν do BEINF0 merece uma atenção especial. Conforme a seção 2.2.3, o `gamlss` não modela diretamente a probabilidade de zero (p_{0i}), mas sim a sua chance (odds), $\nu_i = \frac{p_{0i}}{1-p_{0i}}$, e aplica uma ligação log a este parâmetro ν: $g(\nu_i) = \log(\nu_i) = \mathbf{x}_i'\boldsymbol{\beta}$. Substituindo a definição de ν_i na equação da ligação, temos:

$$\log\left(\frac{p_{0i}}{1-p_{0i}}\right) = \mathbf{x}_i'\boldsymbol{\beta},$$

portanto, embora o pacote tecnicamente liste uma ligação log (aplicada ao parâmetro odds ν), o modelo resultante é, de fato, um modelo logístico padrão (*logit link*) para a probabilidade de zero. A interpretação de seus coeficientes é idêntica à de uma regressão logística: e^{β_j} é a razão de chances de um aluno obter nota zero.

2.5 Diagnóstico

2.5.1 Resíduo Quantílico

A avaliação da adequação do modelo foi fundamentada na análise dos resíduos quantílicos (DUNN; SMYTH, 1996). Estes resíduos representam uma generalização dos resíduos padronizados, com a propriedade fundamental de seguirem uma distribuição normal padrão (N(0,1)) se o modelo estiver corretamente especificado, independentemente da distribuição original da variável resposta. Eles são calculados transformando a diferença entre os valores observados e os previstos para a escala da função de distribuição acumulada do modelo.

A formulação do resíduo quantilico é dada por

$$r_i = \Phi^{-1}(p_i),$$

com $p_i = F(y_i; \hat{\theta})$, onde Φ^{-1} é a função função quantil inversa da distribuição normal padrão e $p_i = F(y_i; \hat{\theta})$ é a probabilidade acumulada do valor observado y_i segundo o modelo ajustado.

2.5.2 Métodos de Análise

A verificação da normalidade dos resíduos foi realizada por meio de uma abordagem multifacetada. Uma análise numérica inicial baseou-se no sumário estatístico dos resíduos, avaliando se a média, variância, assimetria e curtose se aproximavam dos valores esperados para uma distribuição normal padrão.

Para uma inspeção visual detalhada, foram utilizadas ferramentas de diagnóstico avançadas, como as implementadas no pacote `gamlss`, que superam as limitações de gráficos tradicionais. Destacam-se o *Worm Plot* (BUUREN; FREDRIKS, 2001) e o *Bucket Plot* (BASTIANI *et al.*, 2022).

O *worm plot* plota os desvios dos resíduos quantílicos em relação aos quantis esperados de uma distribuição normal, é similar a um gráfico quantil-quantil, mas em vez de uma linha

diagonal, o *worm plot* exibe uma linha de referência horizontal com intervalos de confiança. Se o modelo estiver bem ajustado, os pontos, e a linha de referência, devem se distribuir aleatoriamente em torno da linha horizontal zero e dentro das bandas de confiança (geralmente de 95%). Padrões sistemáticos indicam problemas, especialmente em relação a assimetria e curtose dos resíduos.

O *bucket plot* oferece uma avaliação robusta da assimetria e curtose. Neste gráfico, a distribuição dos resíduos é reamostrada (via bootstrap) múltiplas vezes, e para cada amostra, são calculadas as medidas de assimetria e curtose transformadas para uma escala que vai de -1 a 1. O gráfico exibe a nuvem de pontos gerados pelas amostras bootstrap, possibilitando analisar visualmente se a nuvem se aproxima bem do ponto central (que seria o par ordenado ideal de assimetria e curtose de uma normal padrão). É possivel ainda escolher amostrar apenas dos residuos centrais, das caudas, ou de todo o conjunto.

3 Resultados

3.1 Análise Descritiva

Tabela 2 – Estatísticas descritivas por tipo de escola e população

População	Tipo de Escola	Tamanho Populacional	Número de zeros	% de Zeros	Média (≠0)	Desvio Padrão (≠0)
(Goiânia, 2023)	Pública	3608	123	3.41	0.67	0.17
	Privada	3226	17	0.53	0.83	0.13
	Total	6834	140	2.05	0.75	0.17
(Goiás, 2021-2023)	Pública	53330	2375	4.45	0.62	0.16
	Privada	15274	78	0.51	0.79	0.14
	Total	68604	2453	3.58	0.66	0.17

Fonte: Dados do INEP. Elaboração própria.

A Tabela 2 apresenta as estatísticas descritivas das duas populações utilizadas neste estudo, segmentadas por tipo de escola (pública e privada). Observa-se que, em ambos os grupos, a maior parte dos participantes é oriunda de escolas públicas — representando aproximadamente 52,7% na população 1 (Goiânia, 2023) e 77,7% na população 2 (Goiás, 2021–2023). Essa predominância reflete, em parte, a distribuição real dos estudantes na rede educacional do estado.

Em relação ao desempenho, constata-se que as médias das notas positivas são maiores entre estudantes de escolas privadas: cerca de 0,83 contra 0,67 na população 1 e 0,79 contra 0,62 na população 2. Essa diferença corresponde, em termos percentuais, a um ganho relativo de aproximadamente 24% e 27%, respectivamente, em relação às médias da rede pública, indicando uma vantagem para alunos oriundos da rede privada.

Além disso, verifica-se que a variabilidade das notas entre estudantes de escolas privadas é menor — com desvios-padrão de 0,13 e 0,14, em comparação com 0,17 e 0,16 nas escolas públicas. Essa menor dispersão sugere maior homogeneidade no desempenho dos alunos da rede privada, possivelmente associada a condições educacionais mais uniformes.

Outro aspecto relevante é a proporção de notas zeradas: nas escolas públicas, os percentuais são de 3,41% (população 1) e 4,45% (população 2), enquanto nas privadas os valores caem para 0,53% e 0,51%, respectivamente. Isso reforça a ideia de que fatores estruturais, como condições socioeconômicas, influenciam de forma significativa o desempenho dos estudantes.

Em síntese, os resultados descritivos indicam que a origem escolar constitui um marcador importante de desigualdade no desempenho, com vantagens sistemáticas para alunos provenientes da rede privada, tanto em termos de médias mais elevadas quanto de menor dispersão e incidência de notas zeradas

Conforme antecipado na seção de Metodologia, a escolha do Modelo de Barreiras (*Hurdle*

Model) fundamenta-se na premissa de que a nota zero na redação do ENEM não é um indicador de proficiência nula, mas sim o resultado de uma penalidade ou anulação. Para validar esta premissa, na População 2, referente a Goiás nos anos 2021-2023, foi realizada uma análise de frequência dos motivos de anulação.

A Tabela 3 apresenta a distribuição das redações com nota zero, segundo o tipo de escola (pública ou privada), com base na variável TP_STATUS_REDACAO dos microdados do INEP.

A análise confirma de maneira inequívoca a hipótese central: não há nenhuma ocorrência de nota zero com o status "Tipo 1: Sem problemas". Em 100% dos casos, a nota zero está associada a um dos motivos de anulação previstos no edital.

Em ambos os grupos, os motivos predominantes são "Redação em Branco"(39,3% nas públicas e 45,5% nas privadas), "Fuga ao Tema"(28,4% e 17,9%, respectivamente) e "Cópia do Texto Motivador"(20,8% e 25,2%, respectivamente). Esta constatação empírica fortalece a escolha do Modelo de Barreiras, pois valida que o processo gerador do "zero"(a barreira) é, de fato, qualitativamente distinto do processo que atribui as notas positivas (o desempenho).

Tabela 3 – Distribuição dos motivos de anulação para redações com nota zero, por tipo de escola (População 2: Goiás 2021-23)

		Públicas ($N_{zeros} = 2287$)		Privadas ($N_{zeros} = 123$)	
Código	Motivo da Anulação	n	%	n	%
4	Redação em Branco	899	39,3%	56	45,5%
6	Fuga ao Tema	649	28,4%	22	17,9%
3	Cópia Texto Motivador	475	20,8%	31	25,2%
8	Texto Insuficiente	145	6,3%	10	8,1%
2	Anulada	48	2,1%	1	0,8%
9	Parte Desconectada	37	1,6%	1	0,8%
7	Não atendimento ao tipo textual	34	1,5%	2	1,6%
Total		**2287**	**100,0%**	**123**	**100,0%**

Fonte: Dados do INEP. Elaboração própria.

Tabela 4 – Perfil Sociodemográfico das populações por Tipo de Escola

		Pop. 1 (Goiânia 2023)				Pop. 2 (Goiás 2021-2023)			
		Pública ($n = 3608$)		Privada ($n = 3226$)		Pública ($n = 53330$)		Privada ($n = 15274$)	
Variável	Nível	n	(%)	n	(%)	n	(%)	n	(%)
TP_SEXO	Feminino	1948	(54.0)	1781	(55.2)	29449	(55.2)	8236	(53.9)
	Masculino	1660	(46.0)	1445	(44.8)	23881	(44.8)	7038	(46.1)
TP_COR_RACA	Branca	1464	(40.6)	2194	(68.0)	19706	(36.9)	9944	(65.1)
	Preta	421	(11.7)	130	(4.0)	5792	(10.9)	640	(4.2)
	Parda	1638	(45.4)	849	(26.3)	26006	(48.8)	4369	(28.6)
	Outros	85	(2.4)	53	(1.6)	1826	(3.4)	321	(2.1)
RENDA	Classe 1	721	(20.0)	63	(2.0)	12526	(23.5)	360	(2.4)
	Classe 2	1332	(36.9)	281	(8.7)	20763	(38.9)	1732	(11.3)
	Classe 3	1200	(33.3)	1041	(32.3)	16036	(30.1)	5654	(37.0)
	Classe 4	355	(9.8)	1841	(57.1)	4005	(7.5)	7528	(49.3)
ESC_MAE	Não Fund.	483	(13.4)	59	(1.8)	10441	(19.6)	445	(2.9)
	Fundamental	523	(14.5)	135	(4.2)	8356	(15.7)	806	(5.3)
	Médio	1709	(47.4)	914	(28.3)	22013	(41.3)	4781	(31.3)
	Superior	893	(24.8)	2118	(65.7)	12520	(23.5)	9242	(60.5)
PC (Computador)	Não	1521	(42.2)	395	(12.3)	23689	(44.4)	1838	(12.0)
	Sim	2087	(57.8)	2831	(87.7)	29641	(55.6)	13436	(88.0)
TP_DEP_ADM_ESC	Federal	234	(6.5)	0	(0.0)	3229	(6.1)	3	(0.0)
	Estadual	3031	(84.0)	0	(0.0)	46063	(86.4)	20	(0.1)
	Municipal	5	(0.1)	0	(0.0)	194	(0.4)	0	(0.0)
	Privada	338	(9.4)	3226	(100.0)	3844	(7.2)	15251	(99.8)
ANO	2021	0	(0.0)	0	(0.0)	16719	(31.4)	4305	(28.2)
	2022	0	(0.0)	0	(0.0)	18428	(34.6)	5180	(33.9)
	2023	3608	(100.0)	3226	(100.0)	18183	(34.1)	5789	(37.9)

Fonte: Dados do INEP. Elaboração própria.

Nota: As porcentagens (%) são calculadas dentro de cada coluna.

Tabela 5 – Comportamento da Nota da Redação por Subgrupo

		Pop. 1 (Goiânia 2023)				Pop. 2 (Goiás 2021-23)			
		Pública ($n = 3608$)		Privada ($n = 3226$)		Pública ($n = 53330$)		Privada ($n = 15274$)	
Variável	Nível	% Zeros	Média(Y>0)	% Zeros	Média(Y>0)	% Zeros	Média(Y>0)	% Zeros	Média(Y>0)
TP_SEXO	Feminino	2.5%	0.69	0.3%	0.85	3.7%	0.64	0.5%	0.82
	Masculino	4.5%	0.64	0.8%	0.80	5.4%	0.60	0.5%	0.77
TP_COR_RACA	Branca	2.5%	0.69	0.5%	0.84	3.4%	0.65	0.4%	0.81
	Preta	4.5%	0.63	0.8%	0.79	6.0%	0.60	0.9%	0.75
	Parda	4.0%	0.66	0.6%	0.81	4.9%	0.61	0.7%	0.78
	Outros	2.4%	0.65	0.0%	0.81	4.5%	0.59	0.0%	0.77
RENDA	Classe 1	5.5%	0.62	1.6%	0.74	6.9%	0.58	1.7%	0.71
	Classe 2	3.7%	0.66	1.1%	0.78	4.6%	0.61	0.6%	0.75
	Classe 3	2.3%	0.70	0.9%	0.81	3.1%	0.66	0.6%	0.78
	Classe 4	1.7%	0.73	0.2%	0.85	1.4%	0.71	0.4%	0.82
ESC_MAE	Não Fund.	6.8%	0.61	0.0%	0.77	7.6%	0.57	0.4%	0.75
	Fundamental	4.6%	0.64	0.7%	0.79	5.6%	0.59	1.0%	0.75
	Médio	3.1%	0.67	0.7%	0.80	3.8%	0.63	0.5%	0.77
	Superior	1.5%	0.71	0.5%	0.84	2.3%	0.68	0.5%	0.81
PC (Computador)	Não	3.9%	0.64	0.5%	0.78	5.8%	0.59	0.7%	0.75
	Sim	3.1%	0.69	0.5%	0.83	3.4%	0.65	0.5%	0.80
TP_DEP_ADM_ESC	Federal	1.3%	0.76	—	—	1.2%	0.71	0.0%	0.57
	Estadual	3.6%	0.65	—	—	5.0%	0.61	10.0%	0.61
	Municipal	0.0%	0.49	—	—	2.1%	0.62	—	—
	Privada	3.0%	0.76	0.5%	0.83	1.2%	0.72	0.5%	0.79
ANO	2021	—	—	—	—	3.3%	0.60	0.4%	0.77
	2022	—	—	—	—	5.6%	0.63	0.6%	0.80
	2023	3.4%	0.67	0.5%	0.83	4.3%	0.64	0.5%	0.81

Fonte: Dados do INEP. Elaboração própria.

Nota: Média $(Y > 0)$ indica a média condicional das notas positivas. % Zeros é (N_zeros / N) para cada subgrupo.

As Tabelas 4 e 5 detalham os perfis sociodemográficos das populações e o comportamento da nota da redação por níveis das covariáveis categóricas. Essas informações permitem compreender melhor as diferenças estruturais entre estudantes da rede pública e privada, bem como suas possíveis associações com o desempenho observado.

De modo geral, observa-se que a distribuição de características sociodemográficas difere substancialmente entre os dois tipos de escola. Em termos de sexo, há um leve predomínio do público feminino em todas as populações, com proporções em torno de 54–55% tanto na rede pública quanto na privada. Em relação à cor/raça, a rede privada mostra predomínio de estudantes brancos (68,0% na população 1 e 65,1% na população 2), enquanto na rede pública predominam estudantes pardos (45,4% e 48,8%) e há uma maior diversidade racial.

Em relação a renda, a disparidade é marcante: quase 90% dos estudantes de escolas privadas estão concentrados nas classes de renda mais altas (Classes 3 e 4) na população 1 e na população 2, cerca de 86%, e entre os participantes provenientes da rede pública, predomina a concentração nas classes mais baixas (Classe 1 e 2), cerca de 57% na população 1 e 62,4% na população 2. Essa diferença de composição socioeconômica também se reflete na escolaridade máxima da mãe do participante, aproximadamente 66 - 60% dos alunos de escolas privadas têm mães com ensino superior, contra apenas 25 - 23% na rede pública, nas populações 1 e 2

respectivamente.

Essas desigualdades estruturais também aparecem no acesso a recursos educacionais, enquanto mais de 80% dos participantes de escolas privadas declaram possuir computador em ambas as populações, esse percentual cai para aproximadamente 58-56% na rede pública, nas populações 1 e 2 respectivamente. Além disso, a rede pública é majoritariamente estadual, com baixa presença de escolas federais ou municipais, ao passo que a rede privada concentra quase integralmente os alunos nas instituições particulares.

Essas diferenças de perfil estão associadas a diferenças significativas no desempenho em redação. Em todas as covariáveis consideradas, a proporção de notas zero é sempre maior na rede pública. Na variável "Renda", é notável o decréscimo de porcentagem de notas zeradas em cada nível para a rede pública em ambas as populações. O mesmo efeito é observado para a rede privada, entretanto a queda é drasticamente menor, mesmo efeito é observado para a variável de escolaridade da mãe, notando aí que as questões socioeconômicas distinguem muito bem participantes provenientes de escolas públicas, mas provavelmente não há um efeito tão acentuado entre participantes de escolas privadas.

As médias das notas positivas também apresentam gradientes, estudantes de renda mais alta, cor branca, mães com maior escolaridade e acesso a computador tendem a apresentar desempenhos médios mais elevados, tanto na rede pública quanto na privada, entretanto observa-se que mesmo dentro destes subgrupos mais privilegiados, os estudantes da rede privada apresentam médias superiores às da rede pública.

É importante salientar que há escolas declaradas como pertencentes à rede pública, mas que possuem dependência administrativa privada. Nesses casos, tratam-se de instituições conveniadas, que firmam acordos de cooperação com o poder público para oferta de ensino. Por outro lado, registros de escolas classificadas como de rede privada, mas com dependência administrativa estadual, municipal ou federal, representam inconsistências ou possíveis erros de preenchimento na base de dados. Como o número de participantes provenientes dessas escolas privadas com dependência administrativa não privada é reduzido, optou-se por ignorar tais casos e prosseguir a análise.

Tabela 6 – Estatísticas Descritivas das Covariáveis Contínuas (Notas)

		Pop. 1 (Goiânia 2023)		Pop. 2 (Goiás 2021-23)	
Variável	Estatística	Pública	Privada	Pública	Privada
NU_NOTA_MT	Média	0.53	0.66	0.51	0.63
	Desvio Padrão	0.12	0.13	0.11	0.12
	Correlação (c/ Y)	0.51	0.49	0.48	0.50
NU_NOTA_CH	Média	0.52	0.59	0.50	0.58
	Desvio Padrão	0.08	0.08	0.08	0.08
	Correlação (c/ Y)	0.53	0.49	0.50	0.52
NU_NOTA_CN	Média	0.49	0.58	0.48	0.55
	Desvio Padrão	0.08	0.09	0.07	0.08
	Correlação (c/ Y)	0.45	0.45	0.42	0.46
NU_NOTA_LC	Média	0.52	0.57	0.50	0.56
	Desvio Padrão	0.07	0.07	0.07	0.07
	Correlação (c/ Y)	0.54	0.49	0.53	0.52

Fonte: Dados do INEP. Elaboração própria.

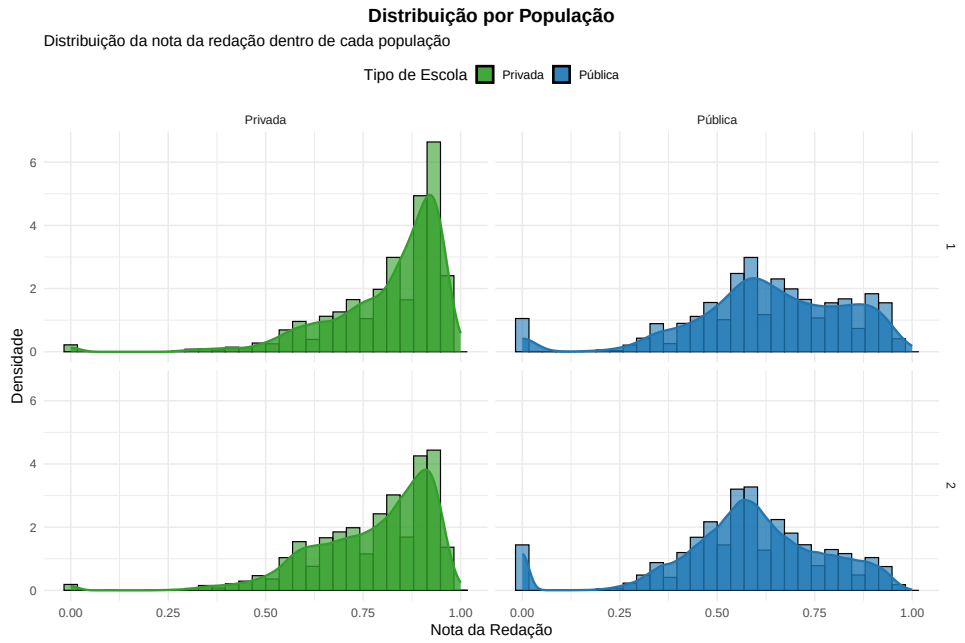
Nota: Correlação (c/ Y) refere-se à Correlação de Pearson com a nota da redação.

A Tabela 6 apresenta as estatísticas descritivas das notas das provas objetivas do ENEM, separadas por tipo de rede de ensino e por população, bem como o coeficiente de correlação de Pearson entre cada nota e o desempenho na redação.

De modo geral, o padrão de desempenho superior entre alunos da rede privada se mantém de forma consistente em todas as áreas do conhecimento e em ambas as populações. As diferenças médias entre as redes variam entre 0,05 e 0,13 pontos, com destaque para a prova de Matemática, onde o diferencial é o mais acentuado. Isso indica que a desigualdade de desempenho entre as redes não se restringe à produção textual, mas se manifesta também nas habilidades cognitivas avaliadas pelas provas objetivas, especialmente nas de raciocínio lógico e quantitativo.

Em termos de dispersão, nota-se que os desvios padrão das notas são muito semelhantes entre as redes, o que sugere que, embora as médias sejam consistentemente maiores nas escolas privadas, a variabilidade interna dos desempenhos é comparável, ou seja, dentro de cada rede, os níveis de heterogeneidade no desempenho individual permanecem próximos, indicando, mais uma vez, que a desigualdade principal ocorre entre redes, e não dentro delas.

Ainda que as magnitudes das correlações sejam moderadas, sua consistência entre as populações e redes indica que o desempenho na redação está associado de forma significativa ao domínio das competências gerais medidas nas demais áreas do ENEM. Tal padrão é esperado, uma vez que a escrita dissertativo-argumentativa demanda não apenas domínio da norma padrão da língua, mas também competências cognitivas transversais, como capacidade analítica, leitura crítica e organização lógica de ideias.



Fonte: Dados do INEP. Elaboração própria.

Figura 1 – Distribuição da nota da redação por tipo de escola, na população 1 (superior) e 2 (inferior)

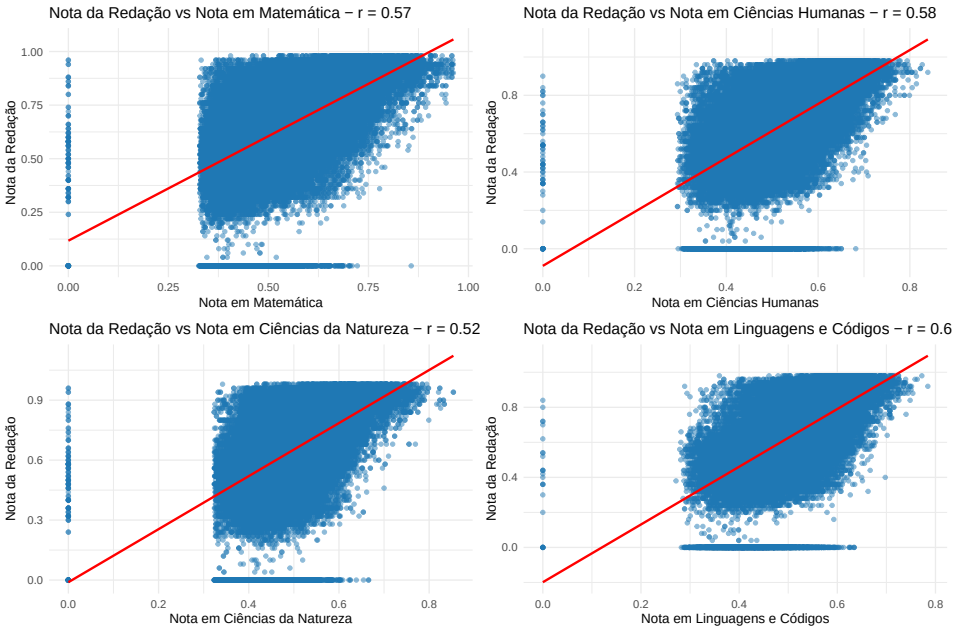
Na Figura 1 observa-se a distribuição da nota da redação segmentada por tipo de escola (Privada e Pública) e pelas duas populações (1 e 2).

Para os participantes de escolas privadas, a distribuição é fortemente assimétrica à esquerda, com um pico de alta densidade concentrado em notas elevadas, próximas a 1. Na população 1, esse pico é mais pronunciado, com a densidade ultrapassando o valor de 5. Já na população 2, o pico correspondente é de menor densidade e a distribuição exibe uma bimodalidade, antecedendo o pico principal (aproximadamente em 0,80), sugerindo maior variabilidade nesse grupo.

Em contrapartida, a distribuição dos participantes de escolas públicas é marcadamente diferente. Nota-se a presença de um segundo pico pronunciado em nota zero (0) em ambas as populações. A distribuição principal das notas (excluindo o zero) é mais centrada, com o modo principal localizado em notas intermediárias (próximo a 0,60 - 0,65). Similarmente ao observado nas escolas privadas, uma bimodalidade é visível após o pico principal (por volta de 0,75), sendo esta característica mais acentuada na população 2.

É notável a complexidade das curvas de densidade que a nota da redação assume. As várias regiões de picos sugerem que modelos que assumem distribuições mais flexíveis sejam necessários para captar estes nuances.

Fica evidente, tanto visualmente quanto pelas tabelas, que o tipo da escola é um fator que diferencia fortemente os alunos. Entretanto, mesmo segregando os participantes, é notável que dentro dos grupos de tipo de escola (especialmente na população 2) há algum fator latente ou não medido que provoca essa aparente multimodalidade nas densidades.



Fonte: Dados do INEP. Elaboração própria.

Figura 2 – Associação entre notas da redação e notas das provas objetivas

A Figura 2 apresenta os gráficos de dispersão entre as notas dos participantes nas provas objetivas e suas respectivas notas na redação. Observa-se elevada variabilidade entre as provas, embora seja visível uma tendência linear, isso indica que, o desempenho em uma prova objetiva pode não ser um preditor determinístico da nota da redação, restando uma alta variabilidade não explicada.

Tal comportamento sugere que as notas das provas objetivas atuam como bons preditores da variabilidade na nota da redação, ainda que não expliquem integralmente seu comportamento.

O coeficiente de correlação de Pearson, que quantifica o grau e a direção da associação linear entre duas variáveis contínuas, evidencia uma correlação de aproximadamente 0,6 entre a nota da redação e a nota em Linguagens e Códigos. Valores semelhantes são observados para as provas de Matemática e suas Tecnologias e Ciências Humanas e suas Tecnologias. Esses resultados reforçam a existência de uma associação linear moderada entre as avaliações, especialmente entre aquelas que mensuram competências também demandadas na elaboração do texto dissertativo-argumentativo, tais como leitura crítica, pensamento analítico, organização lógica das ideias e domínio de estratégias de escrita.

Ressalta-se a presença de uma densa concentração de pontos ao longo do eixo Y = 0, correspondendo a candidatos que obtiveram nota zero na redação, mas apresentaram notas diversas nas provas objetivas. Esse padrão sugere que a ocorrência de nota zero na redação pode não refletir simplesmente o nível mais baixo de habilidade, mas sim um evento qualitativamente distinto — possivelmente relacionado a penalizações específicas (como fuga ao tema ou redação em branco). De modo análogo, observa-se uma faixa de pontos sobre o eixo X = 0, correspondente

a participantes que zeraram alguma prova objetiva, mas obtiveram notas positivas na redação. Esse fenômeno se explica, em geral, por erros de preenchimento do gabarito, como a marcação múltipla de alternativas em uma mesma questão, o que acarreta a anulação da prova. Como as provas objetivas são avaliadas segundo a Teoria de Resposta ao Item (TRI), é improvável que um participante obtenha nota zero por desempenho real, reforçando a interpretação de que tais casos representam eventos atípicos e não extremos legítimos de habilidade.

As evidências descritivas sugerem, portanto, que as diferenças de desempenho estão fortemente associadas a fatores estruturais e contextuais, um padrão que será formalmente investigado na próxima seção por meio de modelos de regressão com inflação de zeros.

3.2 Resultados dos Modelos

Nesta seção são apresentados os resultados dos modelos ajustados para investigar o desempenho dos estudantes. Foram considerados dois tipos de modelos: o modelo beta inflacionado em zero (BEINF0) e o modelo de barreira (*hurdle*). Cada modelo foi ajustado separadamente para escolas públicas e privadas e, dentro de cada grupo, para os conjuntos de dados referentes a Goiânia e ao estado de Goiás. Para o modelo BEINF0, ajustam-se diretamente os parâmetros μ, σ e ν, considerando a possibilidade de inflacionamento no zero. No caso do modelo de barreira, a modelagem ocorre em duas etapas: inicialmente, ajusta-se um modelo logístico para a probabilidade de um aluno obter nota maior que zero; em seguida, para os estudantes com nota positiva, ajusta-se um modelo para os parâmetros de localização, escala, assimetria e curtose (μ, σ, ν e τ). Essa estrutura permite avaliar não apenas o comportamento central da distribuição das notas, mas também sua dispersão e assimetria, além da probabilidade de ocorrências em zero.

3.2.1 População 1

3.2.1.1 BEINF0

Os resultados podem ser conferidos na Tabela 7. O submodelo de locação (μ) controla o aumento (em escala logistica) na média da nota da redação. A partir deste submodelo, os submodelos para outros parâmetros são ajustados.

A análise dos coeficientes deste submodelo para média indica que as notas obtidas nas provas objetivas constituem os principais fatores explicativos da nota da redação, tanto entre participantes de escolas públicas quanto privadas. Observa-se, contudo, uma diferença quanto à prova de maior influência: enquanto, entre os alunos de escolas públicas, a nota em Matemática apresenta o maior impacto positivo sobre a média da redação, entre os alunos de escolas privadas, o destaque recai sobre Ciências da Natureza.

No modelo ajustado para participantes de escolas públicas, o efeito isolado do sexo não se mostrou estatisticamente significativo, sugerindo que, nesse grupo, homens e mulheres obtêm,

em média, desempenhos equivalentes na redação. Esse padrão não se repete entre os participantes de escolas privadas, para os quais o coeficiente associado ao sexo masculino é significativamente negativo, indicando desempenho médio inferior dos homens em comparação às mulheres.

Ainda entre os participantes de escolas públicas, o tipo de dependência administrativa mostrou efeito diferenciado apenas para escolas conveniadas (categoria privada), cujo coeficiente foi positivo e estatisticamente significativo, indicando notas médias mais elevadas em comparação à categoria de referência (escolas federais). As demais categorias, estaduais e municipais, não apresentaram diferenças significativas em relação à referência, o que sugere homogeneidade de desempenho entre essas redes.

No grupo de escolas privadas, a renda familiar mostrou-se um importante determinante da nota da redação. Comparativamente à classe 1 (categoria de referência), todas as demais classes apresentaram coeficientes positivos e estatisticamente significativos, indicando aumento sistemático do desempenho médio conforme o nível de renda.

Entretanto, a interação entre renda e nota em Ciências da Natureza revelou um padrão interessante, o efeito positivo da nota em Ciências da Natureza sobre a redação é atenuado à medida que a renda aumenta. Em outras palavras, o ganho em desempenho na redação associado a um bom desempenho em Ciências da Natureza é mais pronunciado entre alunos de baixa renda do que entre os de renda mais alta.

Por fim, no modelo referente às escolas públicas, embora o efeito isolado do sexo não tenha sido significativo, a interação entre sexo e nota em Matemática apresentou significância estatística. Esse resultado indica que o efeito positivo de Matemática sobre a redação é menor para os participantes do sexo masculino, isto é, homens tendem a se beneficiar menos de um bom desempenho em Matemática na determinação de sua nota de redação.

O intercepto nos dá condições de calcular a nota média de um participante com todas as covariáveis contínuas como zero e as categóricas no nível de referência (isto é, auto declarado branco, renda classe 1, sexo feminino e pertencente a escola com dependência administrativa federal no caso do grupo de participantes de escola pública), no modelo para participantes de escolas públicas, o intercepto foi de -2,69, então, a média da nota da redação de um participante nestas condições é de aproximadamente 63 pontos. Já para participantes provenientes de escolas privadas, o intercepto foi de -3,17, então a nota média nas mesmas condições é de aproximadamente 40 pontos. Este resultado é contraintuitivo à primeira vista, pois a nota basal prevista para o aluno de escola privada é inferior. Isso indica que a grande discrepância no desempenho (observada na análise descritiva) não é explicada pelo ponto de partida dos modelos, mas na verdade, ela é impulsionada pelos coeficientes das covariáveis, que são substancialmente maiores no modelo privado. O modelo das escolas privadas começa mais baixo, mas cresce muito mais rapidamente com o aumento das notas objetivas e da renda, justificando as médias superiores desse grupo.

O submodelo de dispersão (σ) representa a variabilidade residual das notas de redação condicionalmente às covariáveis incluídas no submodelo de média (μ). Em termos interpretativos, coeficientes positivos em σ indicam maior heterogeneidade das notas dentro de determinado grupo ou condição, enquanto coeficientes negativos indicam menor variabilidade, isto é, respostas mais concentradas em torno da média esperada.

No caso das escolas públicas, o submodelo de dispersão selecionou as notas em Ciências Humanas, Ciências da Natureza e a variável indicadora de posse de computador em casa (esta última sem significância estatística). Observa-se que o coeficiente para Ciências Humanas é positivo e significativo, sugerindo maior heterogeneidade entre alunos com bom desempenho nessa área. Em contrapartida, Ciências da Natureza apresenta coeficiente negativo, indicando menor variabilidade entre os estudantes com melhor desempenho nesse domínio.

Para as escolas privadas, os coeficientes associados às notas em Matemática, Ciências Humanas e Ciências da Natureza são fortemente negativos, o que sugere maior consistência no desempenho linguístico entre os alunos com boas habilidades analíticas e interpretativas. Em relação à escolaridade materna, observa-se aumento dos coeficientes conforme o nível de instrução da mãe cresce, em comparação aos filhos de mães que não completaram o ensino fundamental. Esse resultado, contudo, pode refletir desbalanceamento amostral, uma vez que apenas cerca de 1,8% dos indivíduos pertencem à categoria de referência (conforme tabela 4), o que pode inflar artificialmente a dispersão e a significância estatística desses efeitos. Por fim, o termo de interação entre Matemática e sexo masculino apresenta efeito positivo sobre σ, indicando que, entre os homens de escolas privadas, um melhor desempenho em Matemática está associado a maior variabilidade nas notas de redação. Esse resultado reforça a ideia de que o ganho em desempenho linguístico não é linearmente compartilhado dentro desse grupo.

O submodelo de inflação em zero (ν), por sua vez, modela a probabilidade de ocorrência de notas nulas na redação, ou seja, a odds de o participante obter nota igual a zero. Entre os participantes de escolas públicas, observam-se coeficientes negativos expressivos nas provas de Linguagens e Códigos (principal fator associado à redução da probabilidade de zerar) e Matemática, indicando que um bom desempenho global nas provas objetivas está associado à menor propensão à anulação da redação. Já entre os alunos de escolas privadas, o efeito mais relevante ocorre na nota de Ciências Humanas, que nem foi significativa no modelo para escolas públicas. Esse resultado sugere que, nesse grupo, as habilidades interpretativas e argumentativas capturadas por essa prova são o principal fator protetor contra a ocorrência de notas nulas na redação.

3.2.1.2 *Hurdle* - Logistico

Nesta primeira parte do modelo de barreira, apresentada na Tabela 8, a variável resposta não é a nota da redação em si, mas um indicador binário que assume valor 1 caso a redação tenha obtido nota positiva e 0 caso tenha sido anulada. Assim, o modelo logístico estima a odds de um

Tabela 7 – Parâmetros Comparativos dos Modelos BEINF0 por Tipo de Escola (População 1)

	Públicas ($n = 3608$)			Privadas ($n = 3226$)		
Parâmetro	EMV	Viés	p-valor	EMV	Viés	p-valor
Modelo para μ ligação logit						
Intercepto	-2.69	0.12	$< 2e$-16	-3.17	0.57	2.45×10^{-8}
NU_NOTA_LC	1.92	0.24	1.77×10^{-15}	1.44	0.28	1.94×10^{-7}
NU_NOTA_MT	2.41	0.16	$< 2e$-16	1.98	0.18	$< 2e$-16
TP_SEXOM	-0.08	0.10	0.41837	-0.56	0.13	2.34×10^{-5}
NU_NOTA_CH	1.68	0.21	5.43×10^{-16}	1.39	0.27	2.16×10^{-7}
NU_NOTA_CN	0.85	0.21	3.73×10^{-5}	3.43	1.08	0.00155
TP_DEP_ADM_ESTADUAL	-0.01	0.05	0.75532	—	—	—
TP_DEP_ADM_MUNICIPAL	-0.28	0.28	0.32032	—	—	—
TP_DEP_ADM_PRIVADA	0.20	0.06	0.00030	—	—	—
TP_COR_RACAPRETA	-0.11	0.03	0.00213	—	—	—
TP_COR_RACAPARDA	-0.02	0.02	0.28522	—	—	—
TP_COR_RACAOUTROS	0.05	0.07	0.49115	—	—	—
PCSIM	0.04	0.02	0.05103	—	—	—
RENDACLASSE 2	—	—	—	1.39	0.62	0.02564
RENDACLASSE 3	—	—	—	1.34	0.58	0.02109
RENDACLASSE 4	—	—	—	1.47	0.57	0.01043
NU_NOTA_MT:TP_SEXOM	-0.48	0.19	0.01072	0.19	0.19	0.31792
NU_NOTA_CN:RENDA2	—	—	—	-2.41	1.17	0.03958
NU_NOTA_CN:RENDA3	—	—	—	-2.24	1.10	0.04127
NU_NOTA_CN:RENDA4	—	—	—	-2.37	1.08	0.02876
Modelo para σ ligação logit						
Intercepto	-0.78	0.11	1.01×10^{-11}	0.40	0.20	0.04538
NU_NOTA_MT	—	—	—	-1.57	0.20	2.65×10^{-14}
NU_NOTA_CH	0.51	0.24	0.03549	-1.04	0.29	0.000275
NU_NOTA_CN	-0.74	0.24	0.00255	-0.50	0.30	0.095396
PCSIM	-0.04	0.03	0.15323	—	—	—
ESC_MAEFUNDAMENTAL	—	—	—	0.30	0.13	0.022802
ESC_MAEMEDIO	—	—	—	0.17	0.11	0.125930
ESC_MAESUPERIOR	—	—	—	0.28	0.11	0.014263
RENDACLASSE 2	—	—	—	0.16	0.12	0.203130
RENDACLASSE 3	—	—	—	0.05	0.12	0.689109
RENDACLASSE 4	—	—	—	0.00	0.12	0.978436
NU_NOTA_MT:TP_SEXOM	—	—	—	0.38	0.05	9.73×10^{-16}
Modelo para ν ligação log						
Intercepto	5.12	0.64	2.53×10^{-15}	2.19	1.25	0.079985
NU_NOTA_LC	-11.19	1.75	1.78×10^{-10}	—	—	—
NU_NOTA_MT	-4.08	0.94	1.33×10^{-5}	-3.80	2.01	0.059196
NU_NOTA_CH	-2.33	1.53	0.12900	-9.54	2.69	0.000404

Fonte: Dados do INEP. Elaboração própria.

Nota: O símbolo '—' indica que a variável não foi incluída no modelo final para aquele grupo.

participante não zerar a redação, condicionalmente às covariáveis incluídas.

O comportamento desse submodelo é muito similar ao parâmetro ν da modelagem Beta Inflacionada em Zero (BEINF0), com a diferença de que aqui os coeficientes indicam diretamente o efeito sobre a probabilidade de obter nota não nula, e que o modelo para escolas privadas contem a covariável de sexo do participante. Dessa forma, valores positivos dos coeficientes estão associados a maiores chances de não zerar, enquanto coeficientes negativos indicam maior propensão à nota zero.

Tabela 8 – Parâmetros Comparativos do Modelo de Barreira (Logístico) por Tipo de Escola (População 1)

	Públicas ($n = 3608$)			Privadas ($n = 3226$)		
Parâmetro	EMV	Viés	p-valor	EMV	Viés	p-valor
Modelo Logístico (Parte 1 do *Hurdle*)						
Intercepto	-5.12	0.64	2.53e-15	-1.79	1.24	0.150604
NU_NOTA_LC	11.19	1.75	1.78e-10	—	—	—
NU_NOTA_MT	4.08	0.94	1.33e-05	4.07	1.94	0.036315
TP_SEXOM	—	—	—	-1.03	0.52	0.048863
NU_NOTA_CH	2.33	1.53	0.129	9.55	2.61	0.000261

Fonte: Dados do INEP. Elaboração própria.

Nota: O símbolo '—' indica que a variável não foi incluída no modelo final para aquele grupo.

Observa-se que, para os participantes de escolas públicas, o desempenho em linguagens e códigos se mantém como o principal fator redutor da chance de anulação da redação. A nota em Matemática também apresenta coeficiente positivo.

Entre os participantes de escolas privadas, o efeito mais expressivo é observado na nota em Ciências Humanas, que atua como o principal fator protetor contra notas nulas. Além disso, o sexo masculino mostrou coeficiente negativo e significativo, indicando que, dentro desse grupo, homens têm menor chance de obter nota não nula em comparação às mulheres.

Considerando todas as covariáveis zeradas, a chance estimada de não zerar a redação é de aproximadamente 0,59% para estudantes de escolas públicas. Para o grupo de escolas privadas, o intercepto do modelo não foi estatisticamente significativo, não permitindo inferir uma probabilidade base para este grupo.

3.2.1.3 *Hurdle* - BCT

Na segunda parte do modelo de barreira (*hurdle*), na Tabela 9, é apresentada as estimativas dos submodelos para os parâmetros de locação (mediana), dispersão (aproximadamente o coeficiente de variação), assimetria e curtose supondo uma distribuição Box-Cox t para a parte

Tabela 9 – Parâmetros Comparativos do Modelo de Barreira (BCT) por Tipo de Escola (População 1)

	Públicas ($n = 3485$)			Privadas ($n = 3209$)		
Parâmetro	EMV	Viés	p-valor	EMV	Viés	p-valor
Modelo para μ ligação logit						
Intercepto	-2.45	0.10	< 2e-16	-2.60	0.08	< 2e-16
NU_NOTA_LC	2.17	0.22	< 2e-16	2.29	0.21	< 2e-16
NU_NOTA_MT	2.35	0.11	< 2e-16	2.45	0.09	< 2e-16
TP_SEXOM	-0.34	0.02	< 2e-16	-0.35	0.02	< 2e-16
NU_NOTA_CH	1.84	0.18	< 2e-16	1.95	0.18	< 2e-16
TP_DEP_ADM_ESTADUAL	-0.05	0.04	0.206674	—	—	—
TP_DEP_ADM_MUNICIPAL	-0.33	0.19	0.086189	—	—	—
TP_DEP_ADM_PRIVADA	0.19	0.05	0.000597	—	—	—
TP_COR_RACAPRETA	-0.08	0.03	0.015770	—	—	—
TP_COR_RACAPARDA	-0.01	0.02	0.504551	—	—	—
TP_COR_RACAOUTROS	0.06	0.07	0.349660	—	—	—
ESC_MAEFUNDAMENTAL	0.05	0.04	0.193584	—	—	—
ESC_MAEMEDIO	0.06	0.03	0.050008	—	—	—
ESC_MAESUPERIOR	0.10	0.03	0.004515	—	—	—
Modelo para σ ligação log						
Intercepto	0.14	0.12	0.25482	0.15	0.10	0.115493
NU_NOTA_LC	-1.50	0.28	1.01e-07	-1.81	0.28	7.27e-11
NU_NOTA_MT	-1.06	0.15	1.10e-12	-0.74	0.13	1.37e-08
NU_NOTA_CH	-0.43	0.24	0.06908	-0.81	0.23	0.000588
TP_SEXOM	0.14	0.02	1.72e-08	—	—	—
NU_NOTA_CN	-0.59	0.23	0.00698	—	—	—
TP_DEP_ADM_ESTADUAL	0.06	0.05	0.24336	—	—	—
TP_DEP_ADM_MUNICIPAL	-0.07	0.32	0.82222	—	—	—
TP_DEP_ADM_PRIVADA	-0.06	0.06	0.36327	—	—	—
PCSIM	-0.04	0.02	0.10874	—	—	—
Modelo para ν ligação identidade						
Intercepto	1.35	0.06	< 2e-16	-4.54	0.42	< 2e-16
NU_NOTA_MT	—	—	—	12.29	0.79	< 2e-16
TP_SEXOM	—	—	—	2.68	0.54	5.77e-07
NU_NOTA_CH	—	—	—	1.79	0.78	0.0227
NU_NOTA_MT:TP_SEXOM	—	—	—	-8.21	1.05	6.92e-15
Modelo para τ ligação log						
Intercepto	37.80	0.00	< 2e-16	54.50	0.00	< 2e-16

Fonte: Dados do INEP. Elaboração própria.

Nota: O símbolo '—' indica que a variável não foi incluída no modelo final para aquele grupo.

contínua da distribuição das notas da redação, ou seja, somente para os participantes que não zeraram a redação.

Os resultados gerais para mediana e dispersão são consistentes com os achados do modelo BEINF0, embora o modelo BCT tenha convergido para um conjunto mais parcimonioso de covariáveis.

No submodelo para a mediana (μ), que manteve a ligação logit, as notas objetivas continuam sendo os principais preditores do desempenho. Para participantes oriundos de escolas públicas, nota-se que os maiores efeitos e mais significativos continuam sendo as notas nas provas objetivas, com a nota na prova de matemática sendo o fator que mais influencia positivamente a nota da redação. O aumento na escolaridade da mãe e o fato de estudar em escola conveniada, da mesma forma como atestado no modelo BEINF0, associam-se a um aumento significativo na odds da nota. Em contrapartida, ser autodeclarado preto está associado a uma diminuição na odds da mediana da nota da redação, o que reforça que, mesmo dentro do grupo da rede pública, há desigualdades estruturais, econômicas e raciais significativas.

Para as escolas privadas, o modelo de mediana é mais enxuto, selecionando apenas as notas objetivas (com excessão de nota em ciências da natureza) e o sexo, que sugere que dentro deste grupo, estas covariáveis superem as desigualdades estruturais que são atestadas entre os participantes de escolas públicas. O "efeito ambíguo"do intercepto também se repete, com a nota base das escolas privadas (aproximadamente 69) sendo inferior à das públicas (aproximadamente 79). Isso corrobora a tese de que a vantagem da rede privada não advém de um ponto de partida superior, mas sim de coeficientes mais elevados associados às covariáveis de desempenho.

O submodelo de dispersão (σ) oferece insights particularmente ricos sobre a homogeneidade do desempenho, onde coeficientes negativos indicam menor variabilidade.

Em ambos os grupos, os coeficientes para as notas em linguagens e codigos, matemática e ciências humanas são negativos e grandes em valor absoluto, o que é um achado crucial: alunos com melhor desempenho nestas áreas não apenas alcançam notas maiores, mas o fazem de forma mais consistente e previsível.

Aumentos na nota de linguagens e códigos estão associados a uma maior homogeneidade nas notas da redação, sendo este efeito maior entre os participantes de escolas privadas, o mesmo efeito se observa na nota de ciências humanas.

A nota em matemática também, porém o efeito de homogeneidade é maior entre participantes de escolas públicas, o que sugere que, as habilidades mais associadas a notas altas em redação e homogeneidade dos resultados entre participantes de escolas públicas, são as habilidades medidas na prova de matemática (estratégia, pensamento analítico), e para os participantes de escolas privadas, as habilidades medidas em linguagens e códigos e ciências humanas (repertório, melhor escrita etc).

Adicionalmente, para os participantes de escolas públicas, ser do sexo masculino está

associado a uma maior dispersão nas notas.

O submodelo de assimetria ν captura a forma da distribuição. Para as escolas privadas, o intercepto fortemente negativo confirma perfeitamente a observação visual na Figura 1, uma distribuição com forte assimetria à esquerda, refletindo a alta concentração de notas próximas ao limite superior.

Neste grupo, observou-se um efeito fortemente dependente das notas de matemática e do sexo masculino, com interação significativa entre ambos. O coeficiente positivo e grande de Matemática e o termo negativo (e grande em valor absoluto) da interação desta com sexo indicam que, embora o aumento em Matemática tenda a alongar a cauda direita, esse efeito é atenuado entre os homens. O efeito positivo de Ciências Humanas sugere leve deslocamento da assimetria em função da capacidade interpretativa, reduzindo a ocorrência de notas extremamente baixas e corroborando com as interpretações anteriores para participantes de escolas privadas.

Para as escolas públicas, o modelo estimou um intercepto constante, sugerindo uma distribuição ligeiramente simétrica ou com cauda à direita. Este achado é, à primeira vista, contraditório com o histograma, que exibe um pico principal em notas intermediárias-altas e um segundo pico à direita. Esta discrepância indica que a forma da distribuição das escolas públicas não é uma simples assimetria, mas sim uma multimodalidade, como se suspeitava, que um modelo de assimetria e curtose como o BCT não é projetado para capturar, evidenciando a profunda heterogeneidade deste grupo.

O parâmetro de curtose τ regula o grau de concentração de probabilidades nas caudas. Valores mais altos de τ indicam distribuições com caudas mais pesadas e concentração central mais aguda. Observa-se que o modelo estimou valores de τ bastante elevados para ambos os grupos (37,8 e 54,5), com significância estatística máxima, sugerindo distribuições com caudas longas, coerentes com o que vemos na Figura 1 e com a natureza da variável nota, que apresenta tanto casos de desempenho muito baixo quanto muito alto.

O fato do grupo de escolas privadas apresentar τ maior indica que, embora haja maior média e menor dispersão, a distribuição é mais concentrada em torno de valores altos, mas ainda com presença de outliers extremos, o que pode ser também observado na longa cauda da distribuição presente na Figura 1

Em síntese, o modelo BCT complementa a análise realizada via BEINF0 ao dissecar a forma da distribuição das notas. Os resultados indicam consistência estrutural. A inclusão de ν e τ apenas reforça que o desempenho de alunos de escolas privadas tende a ser mais homogêneo e concentrado, enquanto o de escolas públicas é mais heterogêneo e estruturalmente complexo, refletindo as desigualdades do sistema educacional.

3.2.2 População 2

3.2.2.1 BEINF0

Os resultados apresentados na Tabela 10 estão alinhados com aqueles observados na Tabela 7, indicando coerência entre as estimativas obtidas nos diferentes ajustes. Essa convergência reforça a consistência dos achados e aumenta a confiança nas interpretações realizadas. De forma resumida, os principais pontos comparativos podem ser destacados a seguir.

Observa-se variação na magnitude dos coeficientes. Agora em ambas as redes, as notas elevadas em matemática permanecem como o fator que mais influencia positivamente o desempenho na redação.

Nota-se também alterações nos padrões de significância estatística de alguns termos, a variável cor/raça, por exemplo, passou a apresentar todos os níveis significativamente distintos da categoria de referência (Branca). Esse resultado decorre, em grande parte, do aumento expressivo do tamanho populacional, o que tende a reduzir viéses e aumentar o poder dos testes, resultando em maior número de coeficientes estatisticamente significativos, ainda que de pequena magnitude. No caso da interação entre nota em ciências da natureza e renda, no submodelo da média para escolas privadas, observa-se o contrário, nenhum termo manteve significância estatística.

Por fim, houve mudanças no sinal dos coeficientes do submodelo referente ao parâmetro de dispersão. Entre participantes de escolas públicas, aumentos na nota em ciências da natureza passaram a estar associados a maior dispersão nas notas de redação, o mesmo para estudantes que possuem computador em casa.

Tabela 10 – Parâmetros Comparativos dos Modelos BEINF0 por Tipo de Escola (População 2)

	Públicas ($n = 53330$)			Privadas ($n = 15274$)		
Parâmetro	EMV	Viés	p-valor	EMV	Viés	p-valor
Modelo para μ ligação logit						
Intercepto	-2.61	0.03	< 2e-16	-2.75	0.23	< 2e-16
NU_NOTA_LC	2.06	0.05	< 2e-16	1.51	0.12	< 2e-16
NU_NOTA_MT	2.17	0.04	< 2e-16	1.93	0.08	< 2e-16
TP_SEXOM	0.12	0.03	2.07e-06	-0.36	0.05	6.06e-11
NU_NOTA_CH	1.34	0.05	< 2e-16	1.81	0.11	< 2e-16
NU_NOTA_CN	0.97	0.05	< 2e-16	1.84	0.44	2.98e-05
TP_DEP_ADM_ESTADUAL	-0.09	0.01	4.88e-16	—	—	—
TP_DEP_ADM_MUNICIPAL	-0.02	0.04	0.67755	—	—	—
TP_DEP_ADM_PRIVADA	0.14	0.01	< 2e-16	—	—	—
TP_COR_RACAPRETA	-0.06	0.01	3.12e-13	—	—	—
TP_COR_RACAPARDA	-0.05	0.01	< 2e-16	—	—	—
TP_COR_RACAOUTROS	-0.07	0.01	6.57e-08	—	—	—
PCSIM	0.08	0.01	< 2e-16	—	—	—
RENDACLASSE 2	—	—	—	0.33	0.25	0.17845
RENDACLASSE 3	—	—	—	0.48	0.23	0.03791
RENDACLASSE 4	—	—	—	0.67	0.23	0.00326
NU_NOTA_MT:TP_SEXOM	-0.74	0.05	< 2e-16	-0.07	0.09	0.43992
NU_NOTA_CN:RENDA2	—	—	—	-0.42	0.48	0.37806
NU_NOTA_CN:RENDA3	—	—	—	-0.57	0.45	0.20149
NU_NOTA_CN:RENDA4	—	—	—	-0.83	0.44	0.06040
Modelo para σ ligação logit						
Intercepto	-1.47	0.03	< 2e-16	0.14	0.08	0.0872
NU_NOTA_MT	—	—	—	-0.90	0.09	< 2e-16
NU_NOTA_CH	0.38	0.05	2.93e-12	-0.81	0.13	2.21e-10
NU_NOTA_CN	0.25	0.06	5.69e-05	-0.70	0.13	1.95e-07
PCSIM	0.04	0.01	1.62e-06	—	—	—
ESC_MAEFUNDAMENTAL	—	—	—	0.08	0.05	0.0932
ESC_MAEMEDIO	—	—	—	0.05	0.04	0.1927
ESC_MAESUPERIOR	—	—	—	0.06	0.04	0.1535
RENDACLASSE 2	—	—	—	0.06	0.05	0.2576
RENDACLASSE 3	—	—	—	0.03	0.05	0.5452
RENDACLASSE 4	—	—	—	-0.01	0.05	0.8215
NU_NOTA_MT:TP_SEXOM	—	—	—	0.24	0.02	< 2e-16
Modelo para ν ligação log						
Intercepto	5.01	0.18	< 2e-16	1.73	0.64	0.00695
NU_NOTA_LC	-9.00	0.40	< 2e-16	—	—	—
NU_NOTA_MT	-4.93	0.28	< 2e-16	-4.64	0.99	2.99e-06
NU_NOTA_CH	-4.26	0.36	< 2e-16	-9.02	1.15	4.47e-15

Fonte: Dados do INEP. Elaboração própria.

Nota: O símbolo '—' indica que a variável não foi incluída no modelo final para aquele grupo.

3.2.2.2 *Hurdle* - Logístico

Tabela 11 – Parâmetros Comparativos do Modelo de Barreira (Logístico) por Tipo de Escola (População 2)

	Públicas ($n = 53330$)			Privadas ($n = 15274$)		
Parâmetro	EMV	Viés	p-valor	EMV	Viés	p-valor
Modelo Logístico (Parte 1 do *Hurdle*)						
Intercepto	-5.29	0.17	$< 2e$-16	-1.91	0.62	0.00192
NU_NOTA_LC	9.03	0.40	$< 2e$-16	—	—	—
NU_NOTA_MT	4.99	0.27	$< 2e$-16	4.77	0.99	$1.31e$-06
TP_SEXOM	—	—	—	-0.20	0.23	0.38281
NU_NOTA_CH	3.89	0.36	$< 2e$-16	8.64	1.13	$2.77e$-14

Fonte: Dados do INEP. Elaboração própria.

Nota: O símbolo '—' indica que a variável não foi incluída no modelo final para aquele grupo.

Para a primeira parte do modelo de barreira, os coeficientes apresentados na tabela 11 mantêm a consistência direcional (sem inversão de sinal) em relação aos apresentados na tabela 8, embora com alterações na magnitude, e novamente mudanças na significância de alguns coeficientes.

No modelo para privadas, o intercepto foi significativo e o coeficiente associado ao sexo deixou de ser.

Agora, podemos dizer que, participantes oriundos de escolas públicas tem a probabilidade base de não zerar a redação de aproximadamente 0,50%, enquanto que para participantes oriundos de escolas privadas nas mesmas condições, essa probabilidade sobe para 12,89%.

Esse achado quantifica a enorme diferença na vantagem inicial entre os dois grupos, mesmo antes de se considerar o impacto positivo das notas objetivas.

3.2.2.3 *Hurdle* - BCT

No submodelo da mediana, observa-se que as notas das provas objetivas continuam representando o principal efeito positivo sobre a odds, tanto em escolas públicas quanto privadas. O sexo masculino permanece associado a menores valores medianos na redação. Entre participantes de escolas públicas, a escolaridade materna mantém efeito relevante, de modo que níveis mais elevados de instrução da mãe estão associados a aumentos na odds mediana da nota de redação.

No que se refere à variável cor/raça, na população 1 apenas a categoria "Preta" apresentava diferença significativa em relação à categoria de referência (Branca). Já na população 2, embora os efeitos permaneçam de pequena magnitude, todas as categorias mostram-se estatisticamente

Tabela 12 – Parâmetros Comparativos do Modelo de Barreira (BCT) por Tipo de Escola (População 2)

	Públicas ($n = 50955$)			Privadas ($n = 15197$)		
Parâmetro	EMV	Viés	p-valor	EMV	Viés	p-valor
Modelo para μ ligação logit (não-canonica)						
Intercepto	-2.24	0.02	$< 2e$-16	-2.18	0.05	$< 2e$-16
NU_NOTA_LC	2.25	0.05	$< 2e$-16	1.92	0.14	$< 2e$-16
NU_NOTA_MT	1.86	0.03	$< 2e$-16	2.16	0.06	$< 2e$-16
TP_SEXOM	-0.23	0.00	$< 2e$-16	-0.29	0.01	$< 2e$-16
NU_NOTA_CH	1.56	0.04	$< 2e$-16	2.25	0.12	$< 2e$-16
TP_DEP_ADM_ESTADUAL	-0.11	0.01	$< 2e$-16	—	—	—
TP_DEP_ADM_MUNICIPAL	-0.03	0.04	0.407	—	—	—
TP_DEP_ADM_PRIVADA	0.13	0.01	$< 2e$-16	—	—	—
TP_COR_RACAPRETA	-0.06	0.01	1.82e-12	—	—	—
TP_COR_RACAPARDA	-0.05	0.01	$< 2e$-16	—	—	—
TP_COR_RACAOUTROS	-0.07	0.01	1.20e-08	—	—	—
ESC_MAEFUNDAMENTAL	0.04	0.01	8.09e-09	—	—	—
ESC_MAEMEDIO	0.10	0.01	$< 2e$-16	—	—	—
ESC_MAESUPERIOR	0.16	0.01	$< 2e$-16	—	—	—
ANO2022	0.03	0.01	6.60e-09	0.07	0.01	2.58e-07
ANO2023	0.08	0.01	$< 2e$-16	0.12	0.02	2.96e-14
Modelo para σ ligação log						
Intercepto	-0.56	0.03	$< 2e$-16	0.97	0.09	$< 2e$-16
NU_NOTA_LC	-1.18	0.07	$< 2e$-16	-2.04	0.25	$< 2e$-16
NU_NOTA_MT	-0.53	0.04	$< 2e$-16	-1.45	0.12	$< 2e$-16
NU_NOTA_CH	-0.79	0.07	$< 2e$-16	-1.63	0.22	8.5e-14
TP_SEXOM	0.06	0.01	3.72e-16	—	—	—
NU_NOTA_CN	0.11	0.06	0.069371	—	—	—
TP_DEP_ADM_ESTADUAL	0.05	0.01	0.000313	—	—	—
TP_DEP_ADM_MUNICIPAL	0.04	0.06	0.500139	—	—	—
TP_DEP_ADM_PRIVADA	-0.04	0.02	0.012509	—	—	—
PCSIM	0.00	0.01	0.558856	—	—	—
ANO2022	0.05	0.01	6.13e-10	-0.10	0.03	7.7e-05
ANO2023	0.16	0.01	$< 2e$-16	-0.01	0.03	0.843
Modelo para ν identidade						
Intercepto	0.95	0.02	$< 2e$-16	-14.32	0.71	$< 2e$-16
NU_NOTA_MT	—	—	—	21.29	1.12	$< 2e$-16
TP_SEXOM	—	—	—	4.35	0.73	3.28e-09
NU_NOTA_CH	—	—	—	15.60	1.15	$< 2e$-16
NU_NOTA_MT:TP_SEXOM	—	—	—	-13.43	1.29	$< 2e$-16
Modelo para τ ligação log						
Intercepto	4.05	0.21	$< 2e$-16	13.11	0.53	$< 2e$-16

Fonte: Dados do INEP. Elaboração própria.

Nota: O símbolo '—' indica que a variável não foi incluída no modelo final para aquele grupo.

distintas, o que se explica pelo maior tamanho populacional e, consequentemente, pela redução do erro-padrão das estimativas.

Apenas as escolas municipais não diferem significativamente das federais, as estaduais estão associadas a uma diminuição da odds, enquanto as conveniadas apresentam efeito positivo. Embora o efeito da variável temporal seja pequeno, é estatisticamente significativo e positivo, indicando tendência de crescimento na odds mediana da nota da redação ao longo do tempo.

No submodelo de dispersão, observa-se que, para escolas públicas, o sexo, anteriormente significativo, deixou de apresentar efeito estatisticamente distinto, assim como a nota em Ciências da Natureza. As categorias de dependência administrativa tornaram-se significativas, porém com coeficientes de magnitude bastante reduzida, padrão que se repete para a variável posse de computador em casa, cujo coeficiente aproximou-se de zero.

Para o modelo referente às escolas privadas, com exceção do intercepto, todos os coeficientes apresentaram aumento em valor absoluto, refletindo padrões mais homogêneos na população 2 entre participantes com bom desempenho nas provas objetivas. Além disso, verifica-se que, com o avanço temporal, há maior variabilidade nos resultados de redação.

No submodelo da assimetria, os coeficientes exibiram redução expressiva de magnitude, exceto o coeficiente associado ao sexo, no modelo das escolas privadas, que apresentou pequeno aumento.

Por fim, o parâmetro de curtose (τ) foi ligeiramente menor em ambas as redes na população 2, com redução mais acentuada entre participantes de escolas públicas, indicando que essa população tem menor presença de valores extremos.

3.3 Diagnóstico dos Modelos

Tabela 13 – Resumo das Estatísticas dos Resíduos Quantílicos (População 1)

Modelo	Média	Variância	Assimetria	Curtose
BEINF0 (Pública)	-0.005	1.019	0.294	3.656
Hurdle-Logístico (Pública)	0.011	1.004	-0.006	2.921
Hurdle-BCT (Pública)	0.000	1.002	0.200	3.130
BEINF0 (Privada)	0.004	0.992	-0.282	2.861
Hurdle-Logístico (Privada)	-0.009	0.987	-0.021	2.905
Hurdle-BCT (Privada)	0.004	0.981	0.053	2.579

Fonte: Elaborado pelo autor

Nota: O valor esperado para um ajuste perfeito (resíduos $\sim$ N(0,1)) é Média=0, Variância=1, Assimetria=0 e Curtose=3.

O Modelo BEINF0 mostrou-se adequado para a População 1, apesar de apresentar uma assimetria e curtose mais alta que a esperada para o modelo no grupo de participantes prove-

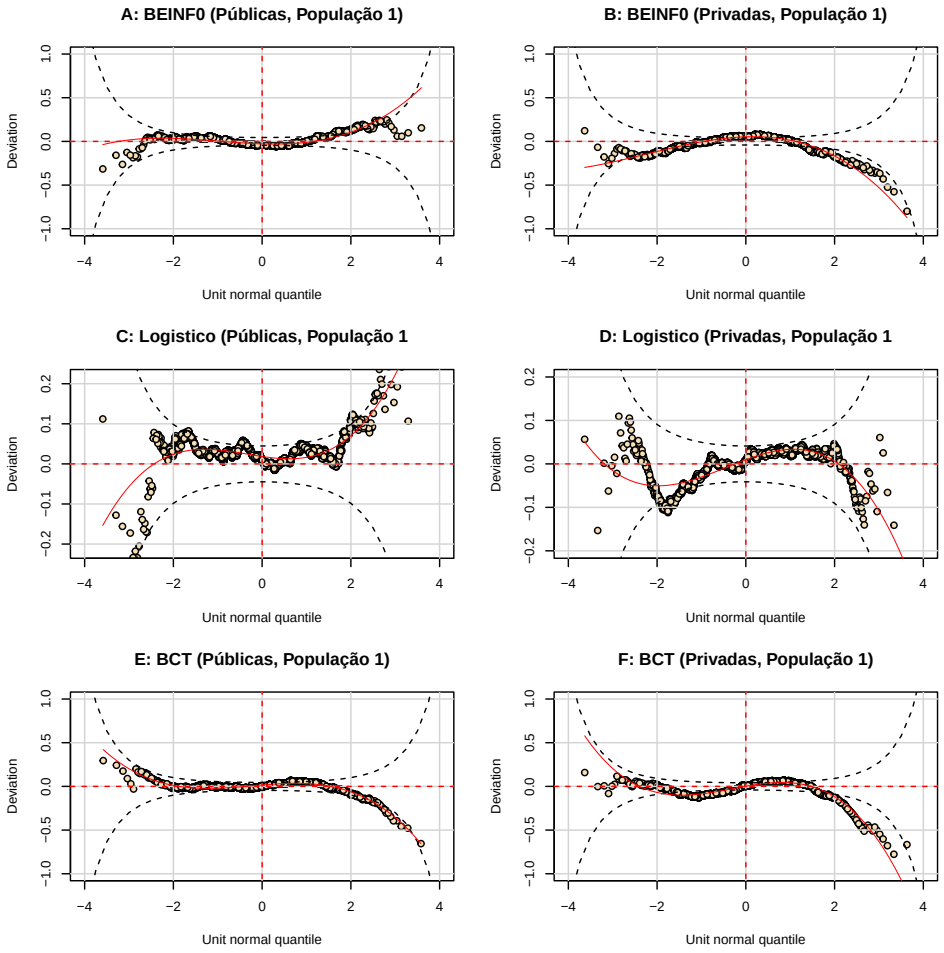


Figura 3 – *Worm Plots* (População 1)

Tabela 14 – Resumo das Estatísticas dos Resíduos Quantílicos (População 2)

Modelo	Média	Variância	Assimetria	Curtose
BEINF0 (Pública)	-0.005	1.019	0.295	3.655
Hurdle-Logístico (Pública)	-0.004	1.001	0.013	2.942
Hurdle-BCT (Pública)	0.004	0.999	-0.083	2.911
BEINF0 (Privada)	0.002	0.994	-0.189	3.033
Hurdle-Logístico (Privada)	0.007	0.987	0.003	3.042
Hurdle-BCT (Privada)	-0.014	1.000	-0.009	2.603

Fonte: Elaborado pelo autor

Nota: O valor esperado para um ajuste perfeito (resíduos $\sim$ N(0,1)) é Média=0, Variância=1, Assimetria=0 e Curtose=3.

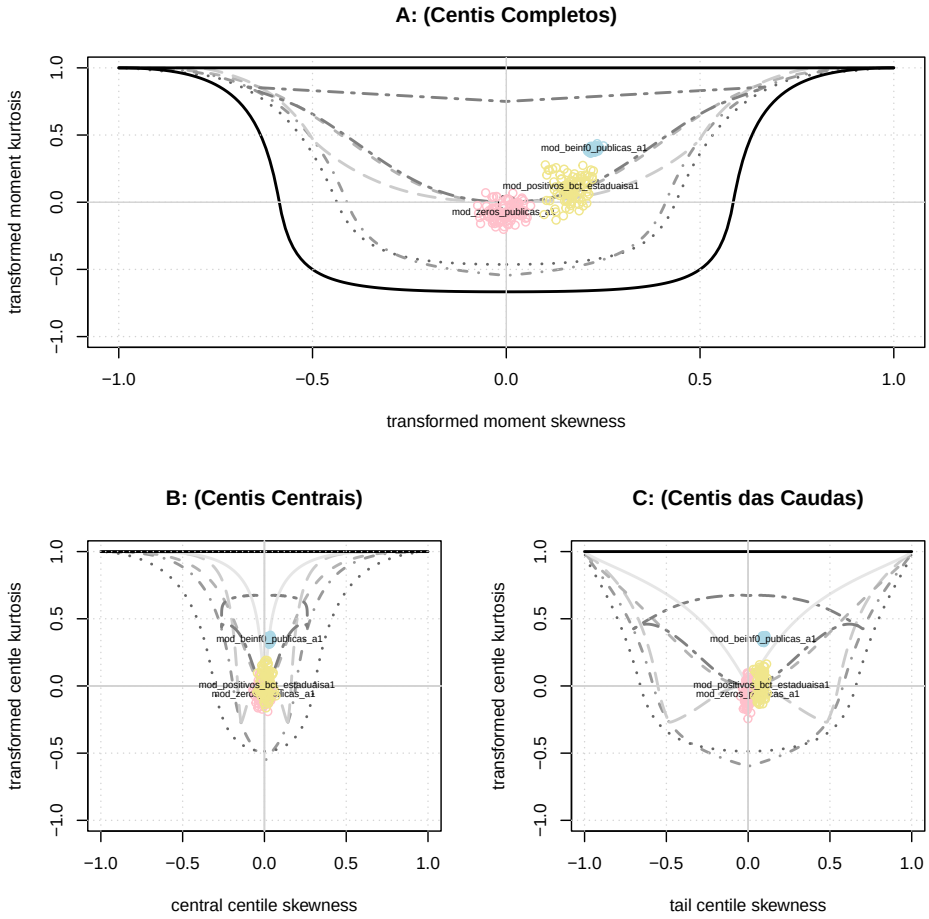


Figura 4 – *Bucket Plot* da População 1 para Escolas Públicas

nientes de escolas públicas. Para o grupo de participantes de escolas privadas, os coeficientes de assimetria e curtose foram menores que os esperados. É possivel ver no painel A da Figura 3, algumas inadequações referentes a assimetria dos resíduos, olhando o qqplot na Figura 9, é possivel ver o desvio na cauda superior. No painel B da Figura 3, é possivel notar um padrão de decrescimento no modelo para privadas. Na Figura 4 é possível notar que em todas as partições, o modelo BEINF0 está bem distante do centroide das figuras. Na Figura 7 é possivel notar que considerando os centis das caudas, a nuvem de pontos se descola para a esquerda do centro esperado, o que corrobora com o que foi atestado com as demais análises. Há um padrão estranho de decrescimento do gráfico de residuos vs valores ajustados, observado na Figura 12, bem como desvio na cauda superior do qqplot.

Entretando na População 2, este se mostrou-se inadequado nas frentes de diagnóstico. Os *worm plots* (Painéis A e B na Figura 5) revelam desvios sistemáticos claros, indicando um mau ajuste nas caudas da distribuição. As estatísticas de resíduos (Tabelas 14) apresentam valores de curtose e assimetria distantes dos ideais, o que fortalece o diagnóstico do *worm plot*. A inadequação é fortemente corroborada pelos *bucket plots* (Figuras 4 7 6 e 8). Em quase todos os cenários (Centis Completos, Centrais e das Caudas), a nuvem de pontos referente ao modelo

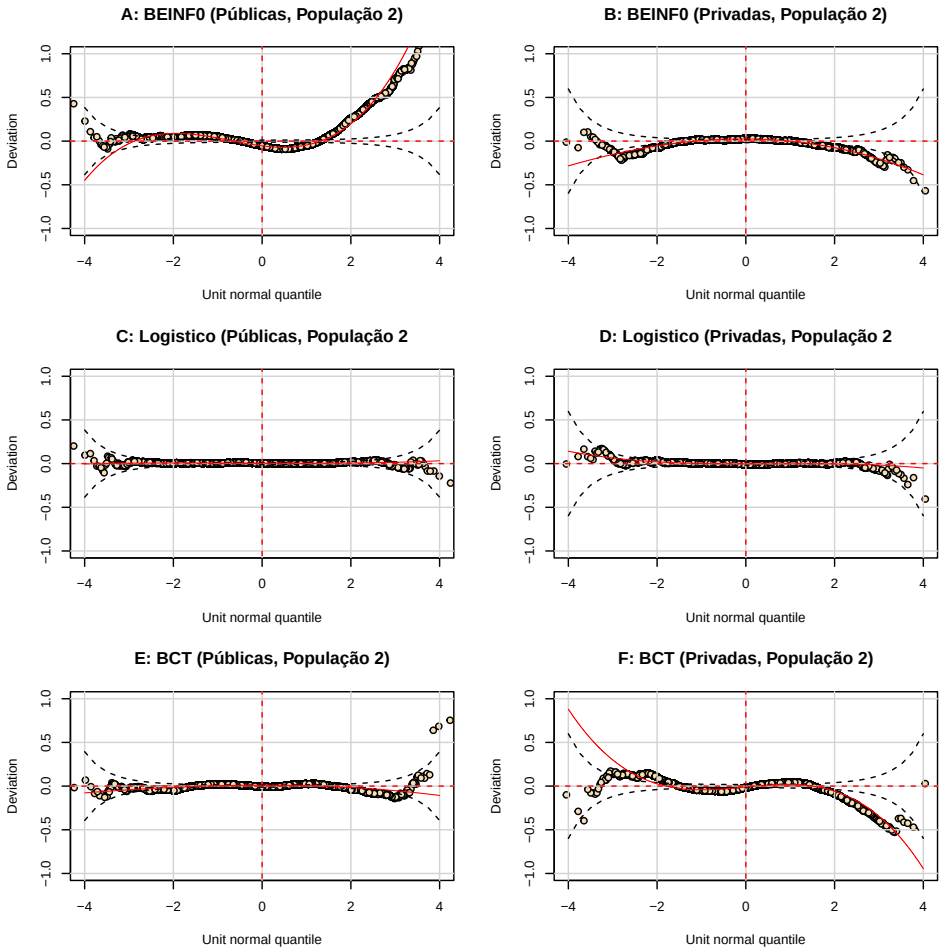


Figura 5 – *Worm Plots* (População 2)

BEINF0 (em azul claro) encontra-se visivelmente distante da origem (0,0), confirmando um mal ajuste em relação à assimetria e curtose. O qqplot visto na Figura 15 continua apresentando o desvio na cauda superior, e na Figura 18 continuamos vendo um padrão de decrescimento no gráfico de residuos vs valores ajustados, apesar do qqplot parecer melhor.

Já para a abordagem de modelo de barreira (Logístico e BCT), demonstram um ajuste mais adequado. Visualmente, os *worm plots* (Gráficos C, D, E e F nas Figuras 3 e 5) mostram que os resíduos estão, em sua maioria, contidos dentro das bandas de confiança, com a linha de tendência próxima de zero, apenas nos modelos para o grupo de participantes de escolas privadas o diagnóstico visual parece não ser tão bom. Numericamente, as Tabelas 13 e 14 mostram que ambos os modelos produzem estatísticas muito próximas dos valores de referência (Média ≈ 0, Variância ≈ 1, Assimetria ≈ 0, Curtose ≈ 3). Os *bucket plots* reforçam essa conclusão, nas Figuras 4 e 6, a nuvem de pontos referente aos componentes dos modelos de barreira (parte logística em rosa e parte BCT em amarelo) estão agrupadas próximas da origem (0,0) em relação ao modelo BEINF0. Entretanto na Figura 7 ambas as estratégias parecem iadequadas em pontos distintos (BEINF0 em assimetria, *Hurdle*-BCT em curtose) e na Figura 8 observa-se que o modelo *Hurdle*-BCT performou pior que o modelo BEINF0. Os demais gráficos de resíduo

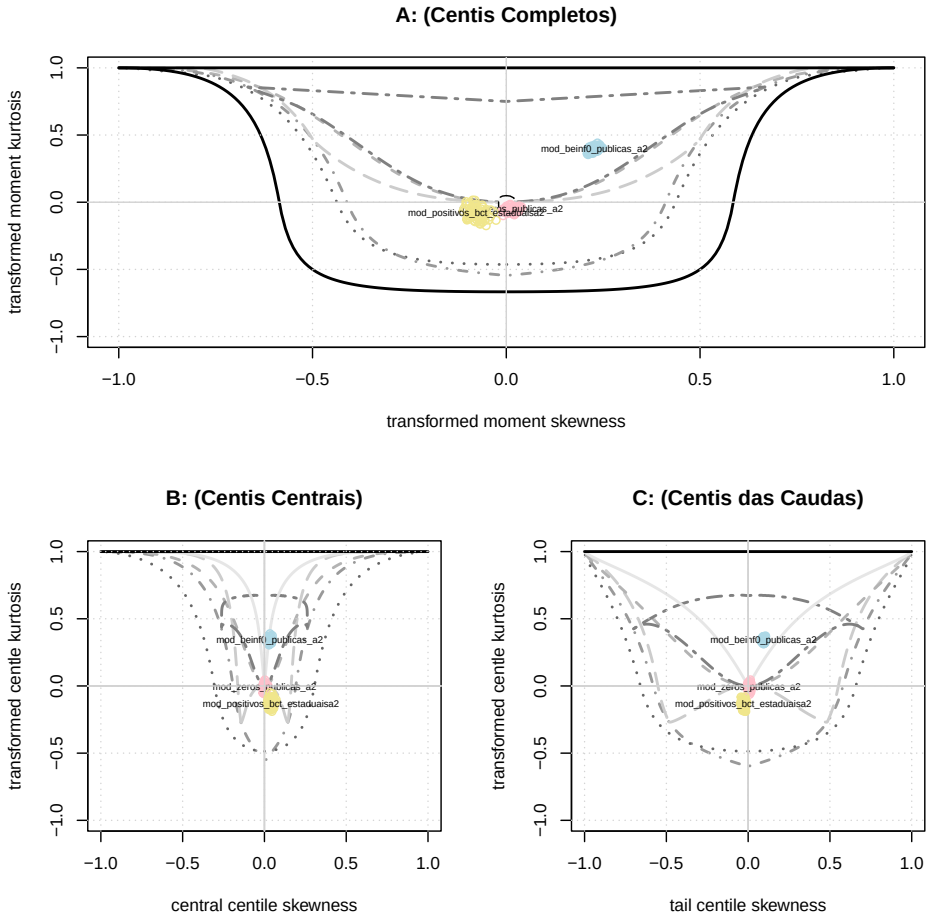


Figura 6 – *Bucket Plot* da População 2 para Escolas Públicas

apontam para um melhor ajuste da estratégia com modelo de barreira, nas Figuras 10 e 11 nota-se um bom comportamento dos residuos em todas as frentes, o mesmo se observa nas Figuras 16 e 17. Para o grupo de participantes de escolas privadas, na população 1, o modelo de barreira apresentou um ajuste ruim, visto pelo qqplot, com ambas as caudas saindo da linha de referencia e a densidade dos resíduos apresentando uma bimodalidade na Figura 14. Nas demais Figuras 13, 19 e 20 é possivel notar um bom comportamento dos resíduos.

A análise conjunta dos *worm plots*, das estatísticas de resíduos e dos *bucket plots* apresenta um veredito claro. O modelo BEINF0 falha em capturar adequadamente as características da distribuição dos dados, notadamente a assimetria e a curtose. Ambos os modelos de barreira (Logístico e BCT) mostram-se mais robustos e adequados. Em ambas as populações, seus valores de assimetria e curtose são, de forma geral, os mais próximos dos valores ideais de 0 e 3, respectivamente. Portanto, a análise de diagnóstico corrobora a escolha da estratégia de modelos de barreira. Os demais gráficos de análise de residuos podem ser conferidos no Apêndice A

Conclusão

A análise comparativa das abordagens utilizando BEINF0 e Modelo de Barreira revelou que, no que diz respeito à interpretação dos resultados, ambas as abordagens são coerentes. Os coeficientes estimados mostraram-se consistentes e levaram a conclusões muito similares, mesmo em diferentes populações, conferindo robustez à compreensão dos fatores que influenciam as notas.

Ambas as abordagens concordam que os grupos de participantes provenientes de escolas públicas e privadas são distintos entre si, no que tange à intensidade pela qual os fatores (às vezes até os mesmos) influenciam a nota da redação do ENEM. Observou-se que dentro do grupo oriundo de escolas privadas, os estudantes apresentam maior homogeneidade e resultados superiores em relação aos participantes oriundos de escolas públicas. Observa-se que os fatores que mais influenciam em ambas as abordagens e grupos sempre são as notas nas provas objetivas, que mensuram habilidades valorizadas na produção de um bom texto, como um bom repertório e pensamento analítico. Os fatores mais associados à prevenção de uma nota zero são justamente os mesmos associados a aumentos na nota da redação. Homens em ambos os grupos são associados a menores notas.

No entanto, ao avaliar a adequação metodológica, bem como a generalização do modelo em uma população mais heterogênea, a abordagem utilizando o modelo de barreira destacou-se como mais realista e com melhor ajuste ao problema. A melhor performance desta abordagem reside na sua flexibilidade, pois permite o uso de uma distribuição mais complexa e aderente às nuances das notas não zeradas da redação. Neste contexto específico, ficou evidente que a distribuição beta, utilizada no modelo BEINF0, apresentou limitações para capturar toda a complexidade da parte de notas não zeradas. Embora a distribuição BCT também tenha demonstrado certo desajuste de curtose nos grupos de escolas privadas, o modelo de barreira, como um todo, mostrou-se mais aderente às nuances dos dados e melhor em termos de ajuste de assimetria.

Os resultados obtidos demonstram a utilidade de modelos como o BEINF0 e o de barreira para lidar com dados inflacionados em zero, comuns em avaliações educacionais como a redação do ENEM. A distinção entre a modelagem da ocorrência de notas zero e do desempenho condicional dos alunos com pontuação não zerada é essencial para estimativas mais precisas de cada componente da distribuição. A análise separada entre escolas públicas e privadas permite identificar diferenças estruturais entre os grupos, oferecendo subsídios importantes para a formulação de políticas educacionais, como programas de reforço direcionados a estudantes com maior risco de baixo desempenho, iniciativas de capacitação docente e estratégias de incentivo ao engajamento dos alunos. A comparação entre uma população mais restrita (Goiânia) e uma mais ampla (Goiás) mostra que, em termos de diagnóstico, o ganho é maior no modelo de barreira,

visto que em uma população maior e mais heterogênea os resíduos apresentam comportamento próximo do que se espera, caso o modelo esteja bem ajustado. Como encaminhamento e sugestão para trabalhos futuros, uma investigação utilizando uma mistura de distribuições beta (beta mixture) para a parte contínua poderia acomodar de forma ainda mais eficaz a heterogeneidade encontrada nos dados, refinando o ajuste do modelo.

Em suma, este trabalho contribui ao validar metodologias estatísticas robustas para a análise de dados educacionais complexos, marcados pela inflação de zeros. Ao demonstrar a importância de modelar separadamente a probabilidade de nota zero e o desempenho, e ao confirmar a heterogeneidade significativa entre grupos de escolas públicas e privadas, a pesquisa oferece um diagnóstico mais preciso dos fatores de desempenho. A principal contribuição deste estudo é, portanto, fornecer subsídios técnicos e analíticos que podem auxiliar gestores na formulação de políticas educacionais mais eficazes e direcionadas.

Referências

ALBUQUERQUE, M. M. d. Desempenho escolar dos estudantes da região sudeste que realizaram o enem: uma análise com modelos hierárquicos. 2017. Citado na página 20.

ANDRADE, D. F. de; TAVARES, H. R.; VALLE, R. da C. Teoria da resposta ao item: conceitos e aplicações. **ABE, Sao Paulo**, 2000. Citado na página 19.

BASTIANI, F. D. *et al.* Bucket plot: A visual tool for skewness and kurtosis comparisons. **Brazilian Journal of Probability and Statistics**, Brazilian Statistical Association, v. 36, n. 3, p. 421–440, 2022. Citado na página 37.

BUUREN, S. v.; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. **Statistics in medicine**, Wiley Online Library, v. 20, n. 8, p. 1259–1277, 2001. Citado na página 37.

CRAGG, J. G. Some statistical models for limited dependent variables with application to the demand for durable goods. **Econometrica: journal of the Econometric Society**, JSTOR, p. 829–844, 1971. Citado 2 vezes nas páginas 21 e 29.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and graphical statistics**, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996. Citado na página 37.

FENG, C. X. A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. **Journal of statistical distributions and applications**, Springer, v. 8, n. 1, p. 8, 2021. Citado na página 30.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of applied statistics**, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004. Citado na página 26.

HASTIE, T.; TIBSHIRANI, R. Generalized additive models. **Statistical science**, Institute of Mathematical Statistics, v. 1, n. 3, p. 297–310, 1986. Citado na página 33.

HEILBRON, D. C. Zero-altered and other regression models for count data with added zeros. **Biometrical Journal**, Wiley Online Library, v. 36, n. 5, p. 531–547, 1994. Citado na página 17.

HESTER, J.; WICKHAM, H. **vroom: Read and Write Rectangular Text Data Quickly**. [S.l.], 2023. R package version 1.6.5. Disponível em: <https://CRAN.R-project.org/package=vroom>. Citado na página 24.

INEP. **Histórico do Exame Nacional do Ensino Médio (ENEM)**. 2020. Portal INEP / Avaliação e Exames Educacionais. Acesso em: 04 nov. 2025. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem/historico>. Citado na página 18.

INEP. **Exame Nacional do Ensino Médio – Enem: procedimentos de análise**. Brasília, DF: [s.n.], 2021. Publicação institucional. Citado na página 18.

INEP. **Base de Dados do IDEB – Resultados**. 2023. Acesso em: 30 abr. 2024. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb/resultados>. Citado na página 20.

INEP. **A Redação do Enem 2023: cartilha do participante**. Brasília: [s.n.], 2023. Citado na página 19.

INEP. **A Redação do Enem 2023: cartilha do participante**. Brasília: [s.n.], 2023. Cartilha institucional. Citado na página 30.

INEP, I. N. de Estudos e P. E. A. T. **Sinopse Estatística da Educação Básica 2023**. Brasília, DF, 2024. Acesso em: dd mm. aaaa. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/sinopses-estatisticas/educacao-basica>. Citado na página 20.

INEP, I. N. de Estudos e P. E. A. T. **Censo da Educação Básica 2023-2024**. 2025. <https://www.gov.br/inep>. Microdados acessados em 07/2025. Citado na página 16.

LIMA, J. W. P. *et al.* Avaliação da nota de redação do enem no estado do ceará via modelo de regressão beta inflacionado. Universidade Federal da Paraíba, 2023. Citado 4 vezes nas páginas 16, 20, 21 e 23.

LOBO, G. D.; CASSUCE, F. C. d. C.; CIRINO, J. F. Avaliação do desempenho escolar dos estudantes da região nordeste que realizaram o enem: uma análise com modelos hierárquicos. **Revista Espacios**, v. 38, n. 5, p. 12, 2017. Citado na página 20.

MARTINEZ, R. O. **Modelos de regressão beta inflacionados**. Tese (Doutorado) — Universidade de São Paulo, 2008. Citado 3 vezes nas páginas 17, 28 e 29.

MULLAHY, J. Specification and testing of some modified count data models. **Journal of econometrics**, Elsevier, v. 33, n. 3, p. 341–365, 1986. Citado 2 vezes nas páginas 17 e 22.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. **Journal of the Royal Statistical Society Series A: Statistics in Society**, Oxford University Press, v. 135, n. 3, p. 370–384, 1972. Citado 2 vezes nas páginas 29 e 33.

OLIVEIRA, G. R. *et al.* Avaliação de eficiência das escolas públicas de ensino médio em goiás: uma análise de dois estágios. **Economia Aplicada**, Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto-USP . . . , v. 21, n. 2, p. 163, 2017. Citado na página 20.

OSPINA, R.; FERRARI, S. L. Inflated beta distributions. **Statistical papers**, Springer, v. 51, n. 1, p. 111–126, 2010. Citado na página 27.

PITSHA, P.; CHIRUKA, R. T.; MARANGE, C. S. A comparison of the robust zero-inflated and hurdle models with an application to maternal mortality. **Mathematical and Computational Applications**, MDPI, v. 30, n. 5, p. 95, 2025. Citado na página 30.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2025. Disponível em: <http://www.R-project.org/>. Citado 3 vezes nas páginas 17, 24 e 35.

RÊGO, F. E. D. d. **Fatores que influenciam na nota da redação do ENEM no Rio Grande do Norte**. Dissertação (B.S. thesis) — Universidade Federal do Rio Grande do Norte, 2021. Citado 4 vezes nas páginas 16, 20, 21 e 23.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. **Journal of the Royal Statistical Society Series C: Applied Statistics**, Oxford University Press, v. 54, n. 3, p. 507–554, 2005. Citado 4 vezes nas páginas 29, 33, 34 e 35.

RIGBY, R. A.; STASINOPOULOS, D. M. Using the box-cox t distribution in gamlss to model skewness and kurtosis. **Statistical Modelling**, Sage Publications Sage CA: Thousand Oaks, CA, v. 6, n. 3, p. 209–229, 2006. Citado 2 vezes nas páginas 31 e 33.

RIGBY, R. A. *et al.* **Distributions for modeling location, scale, and shape: Using GAMLSS in R**. [S.l.]: Chapman and Hall/CRC, 2019. Citado na página 27.

ROLIM, T. M. *et al.* Riqueza, desigualdade e pobreza no brasil: o caso da região centro-oeste brasileira. Brasil, 2022. Citado na página 19.

TANIGUCHI, M.; HIRUKAWA, J. Generalized information criterion. **Journal of Time Series Analysis**, Wiley Online Library, v. 33, n. 2, p. 287–297, 2012. Nenhuma citação no texto.

TEIXEIRA, I. N. de Estudos e P. E. A. **Microdados do Enem 2023**. Brasília: [s.n.], 2024. Base de microdados. Acesso em: 30 abr. 2024. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>. Citado na página 23.

TERAMATSU, G.; STRAFORINI, R. Do enem ao sisu: cartografia da interiorização do acesso à educação superior no brasil. Instituto de Pesquisa Econômica Aplicada (Ipea), 2022. Citado na página 16.

WICKHAM, H.; GROLEMUND, G. Welcome to the tidyverse. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 2019. Citado na página 24.

APÊNDICE A – Gráficos de Diagnóstico dos Modelos

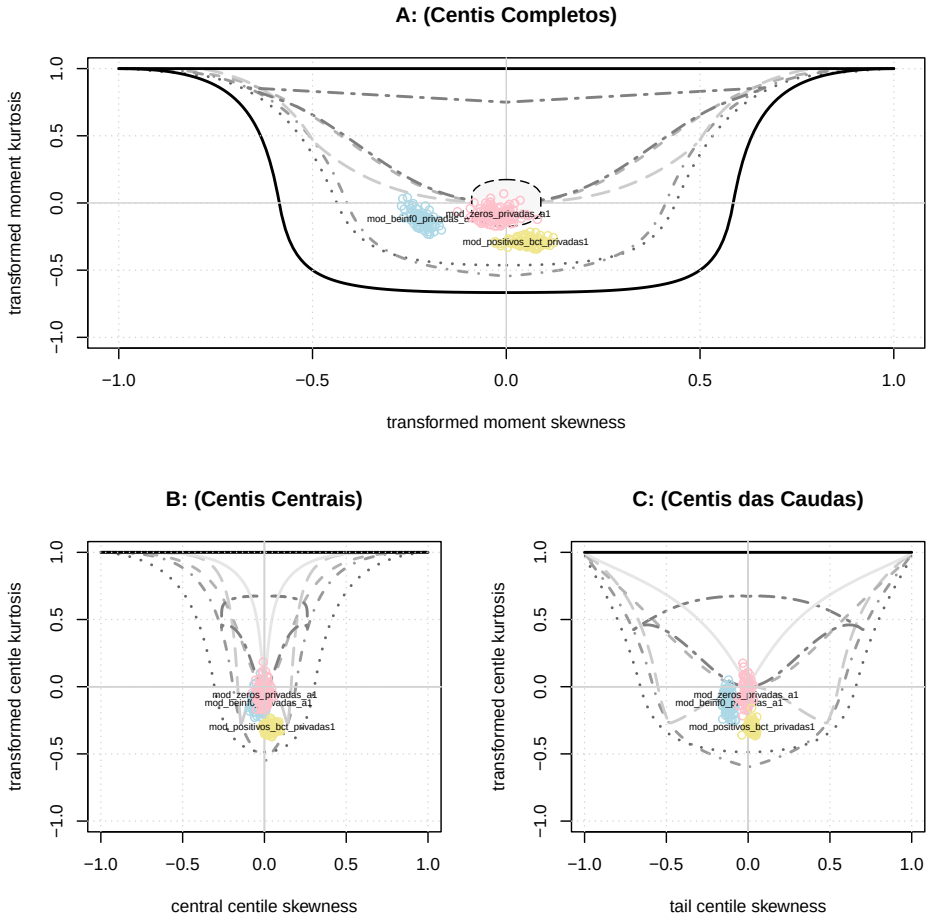


Figura 7 – *Bucket Plot* da População 1 para Escolas Privadas

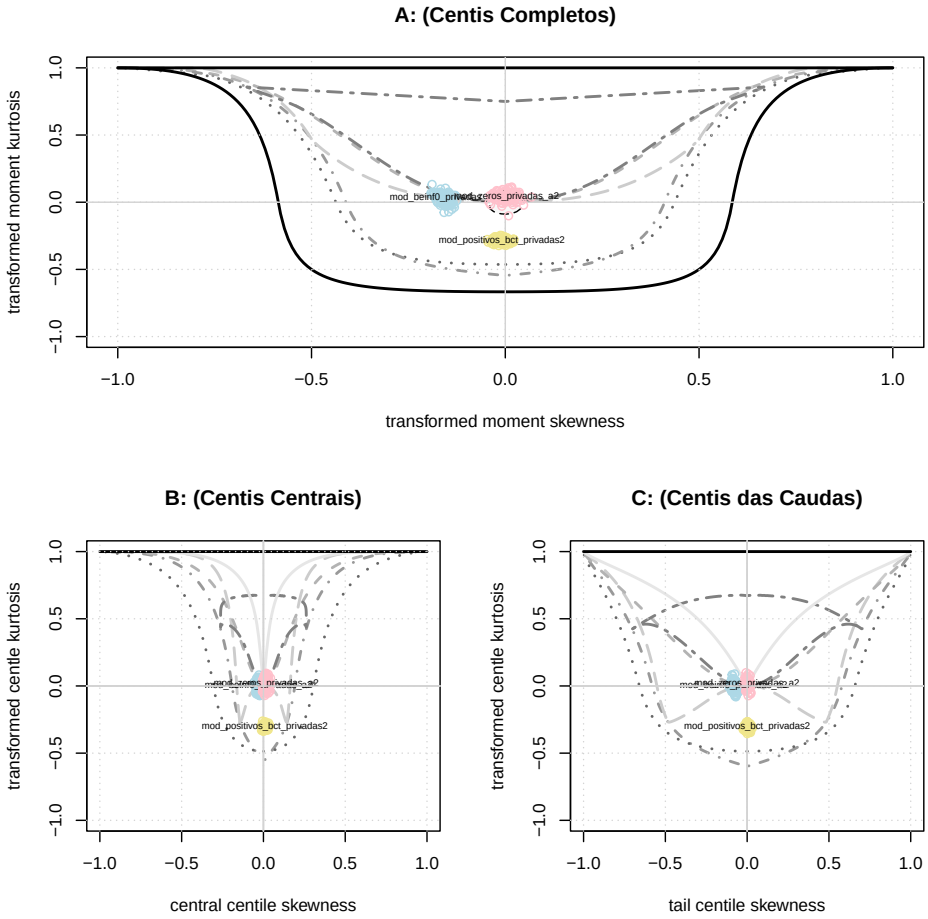


Figura 8 – *Bucket Plot* da População 2 para Escolas Privadas

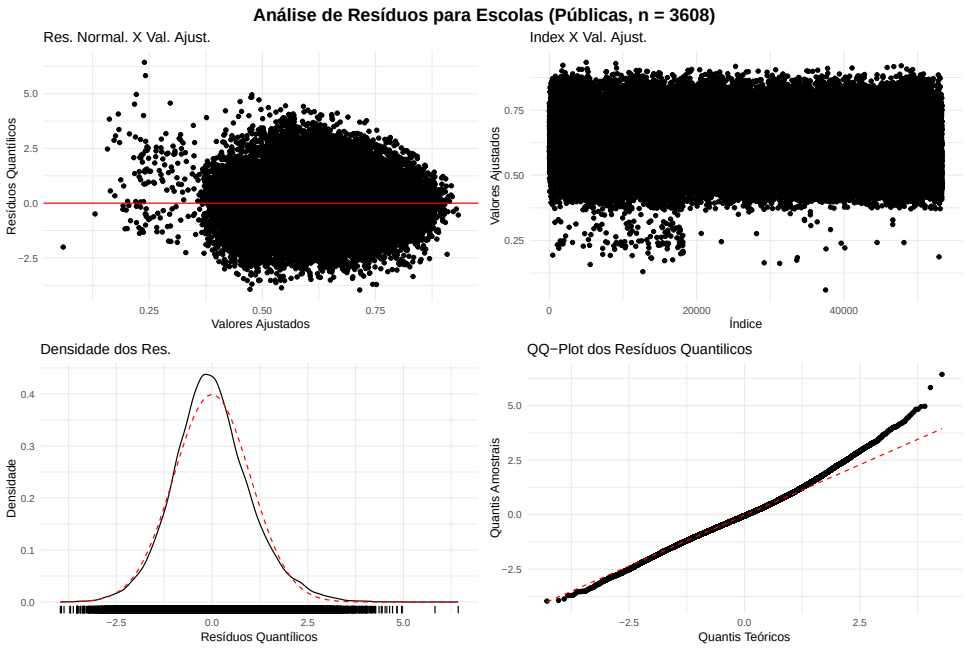


Figura 9 – Diagnóstico Residuos Quantilicos (BEINF0, Públicas, População 1)

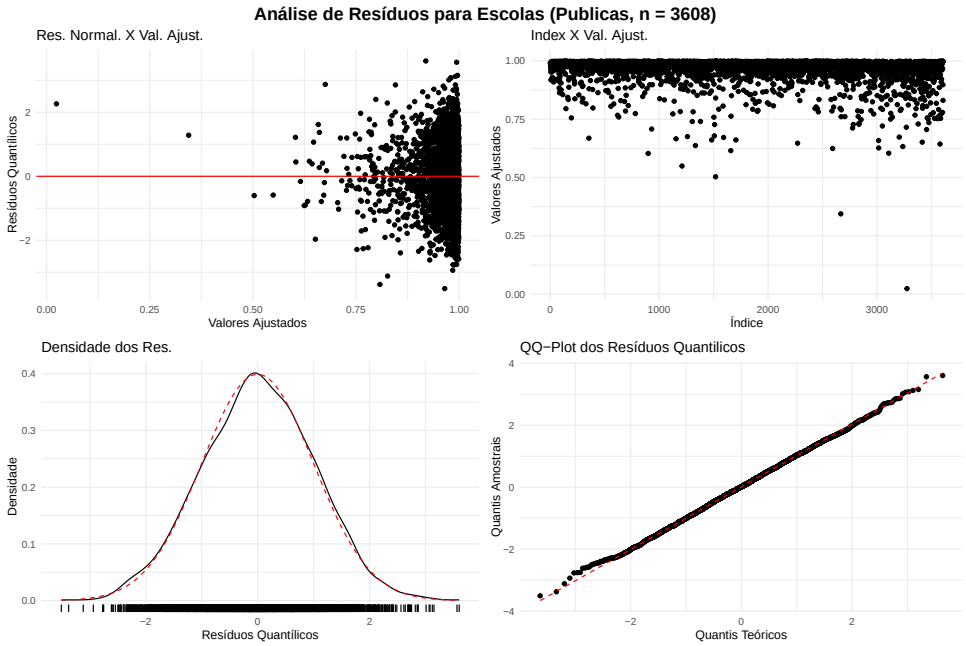


Figura 10 – Diagnóstico Residuos Quantilicos (*Hurdle*-Logístico, Públicas, População 1)

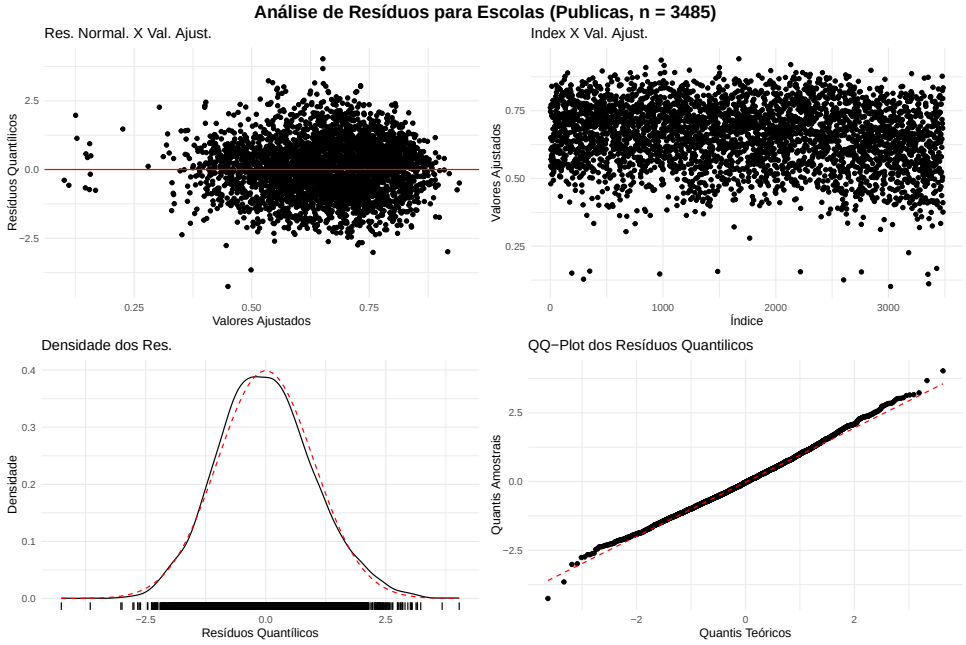


Figura 11 – Diagnóstico Residuos Quantilicos (*Hurdle*-BCT, Públicas, População 1)

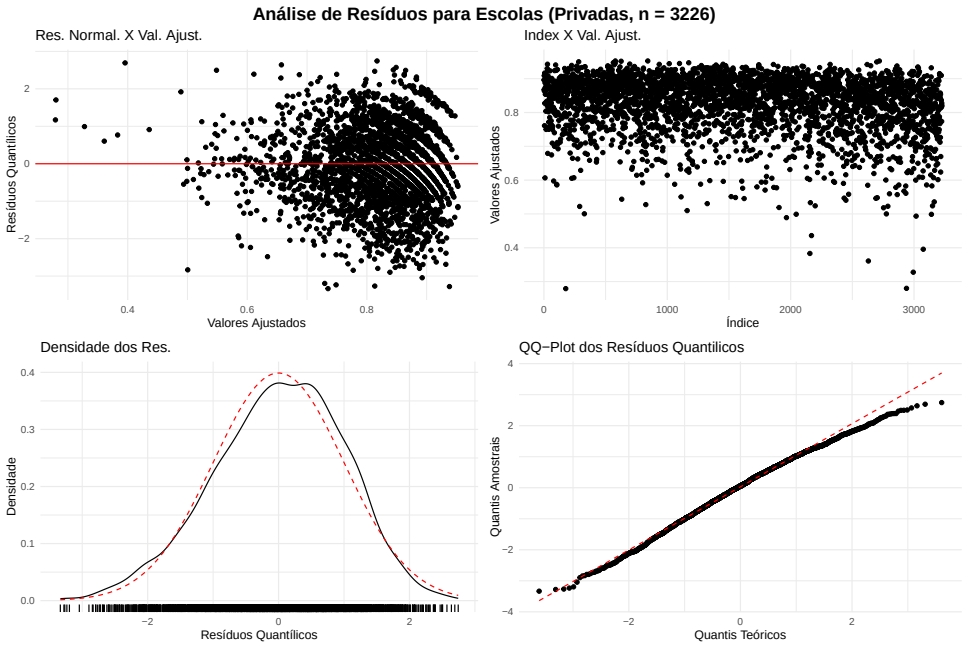


Figura 12 – Diagnóstico Residuos Quantilicos (BEINF0, Privadas, População 1)

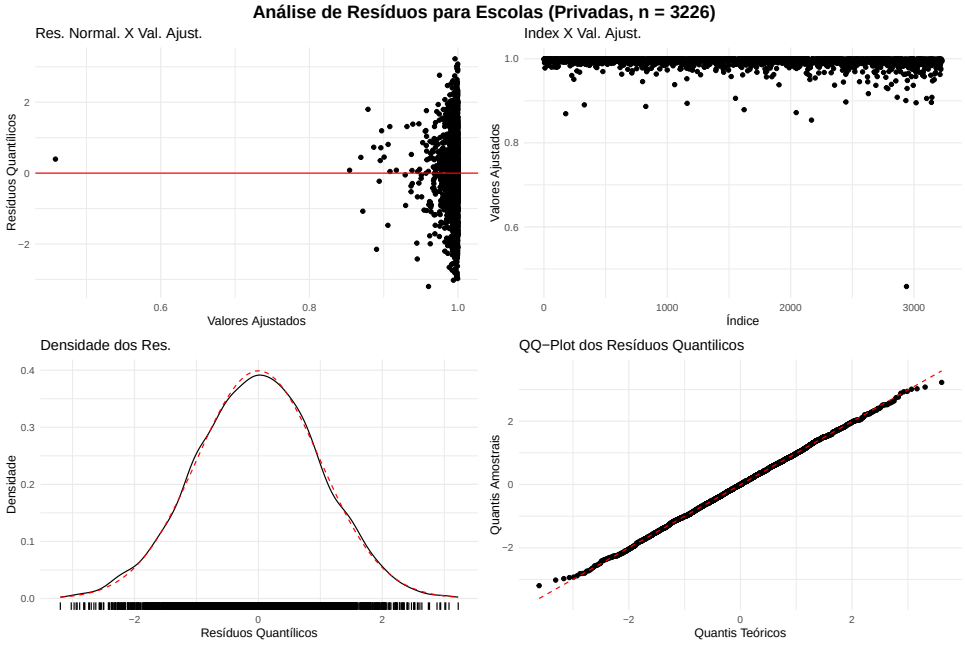


Figura 13 – Diagnóstico Residuos Quantilicos (*Hurdle*-Logístico, Privadas, População 1)

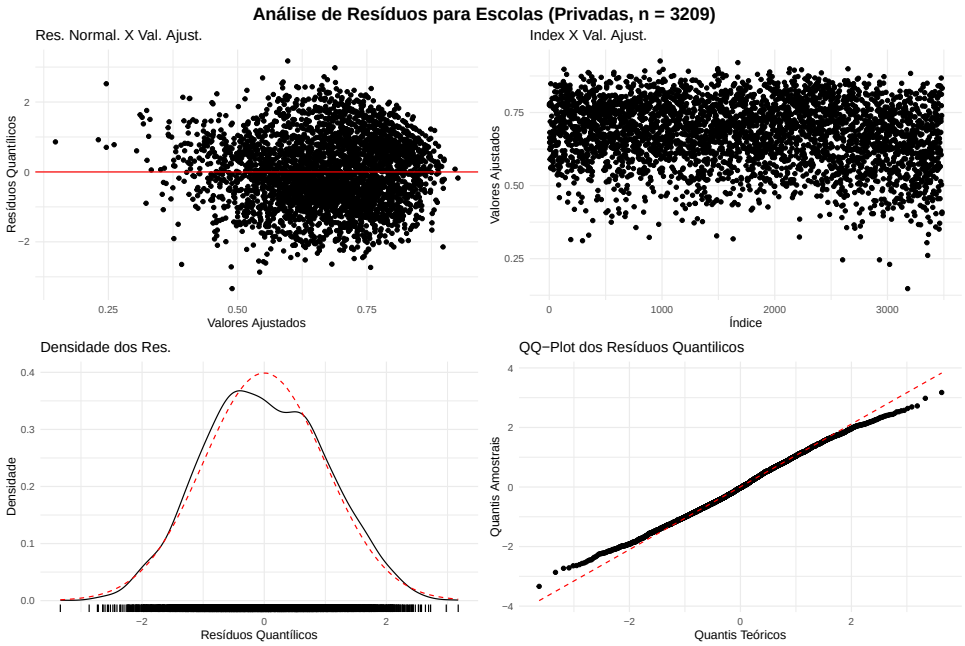


Figura 14 – Diagnóstico Residuos Quantilicos (*Hurdle*-BCT, Privadas, População 1)

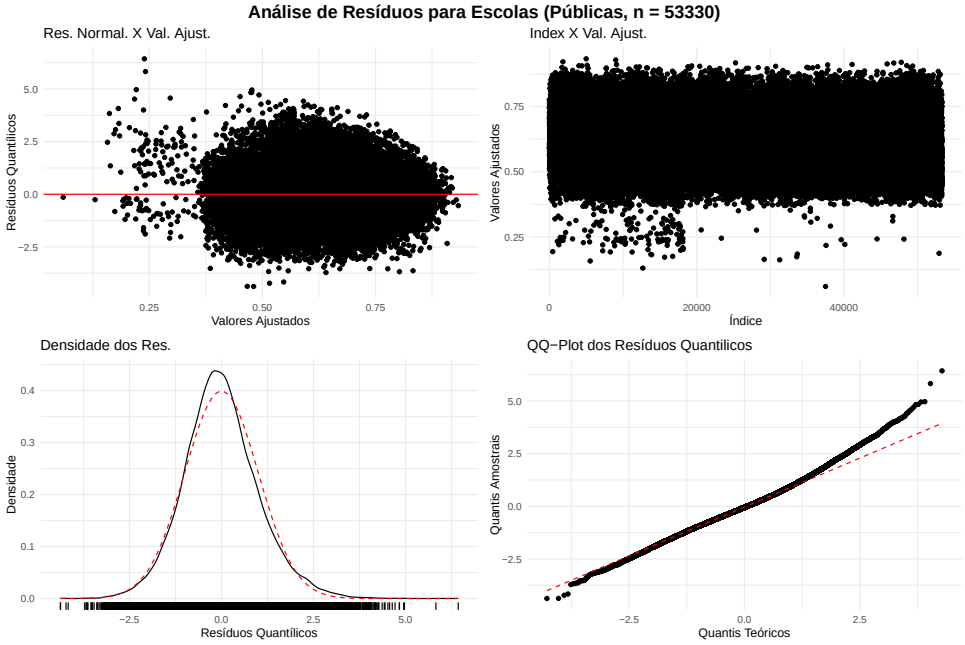


Figura 15 – Diagnóstico Residuos Quantilicos (BEINF0, Públicas, População 2)

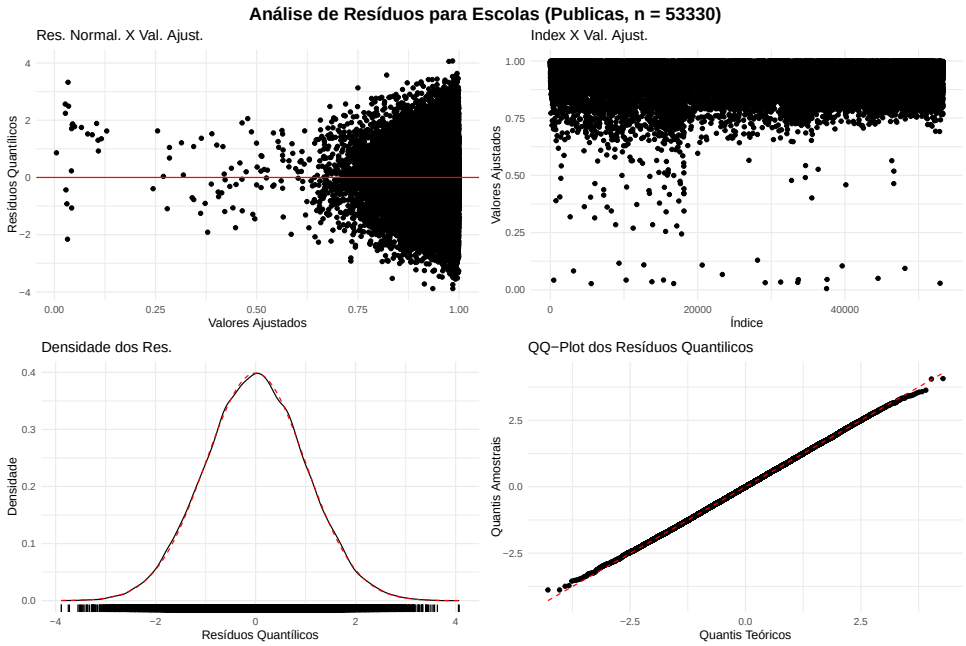


Figura 16 – Diagnóstico Residuos Quantilicos (*Hurdle*-Logístico, Públicas, População 2)

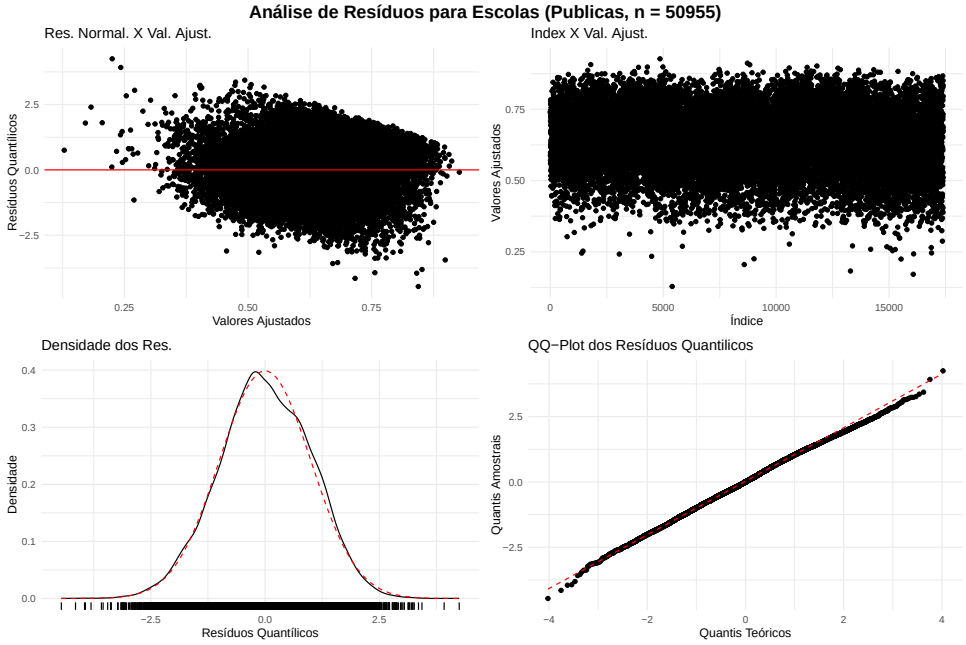


Figura 17 – Diagnóstico Residuos Quantilicos (*Hurdle*-BCT, Públicas, População 2)

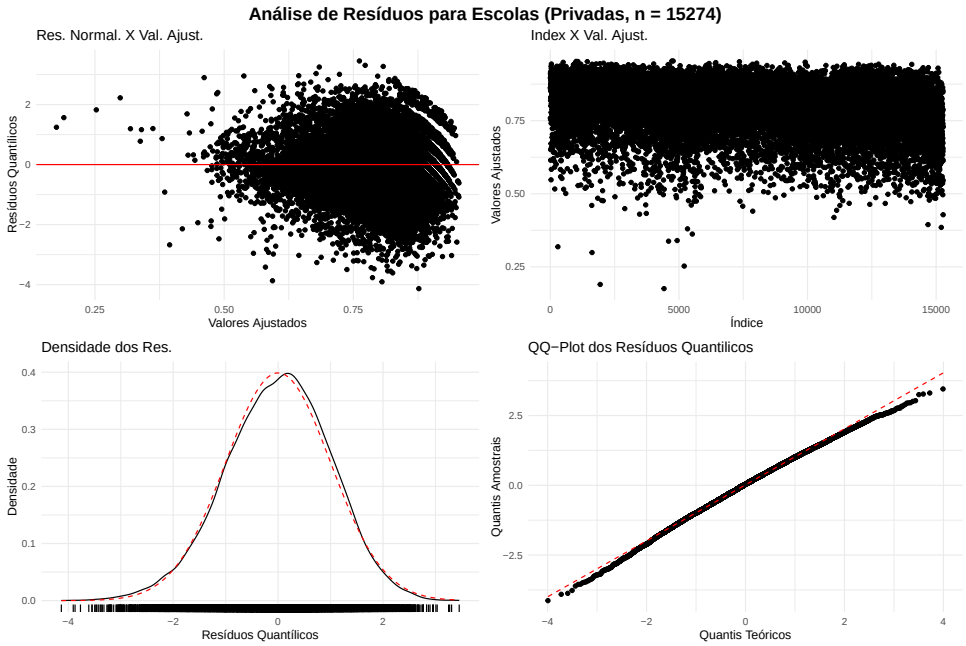


Figura 18 – Diagnóstico Residuos Quantilicos (BEINF0, Privadas, População 2)

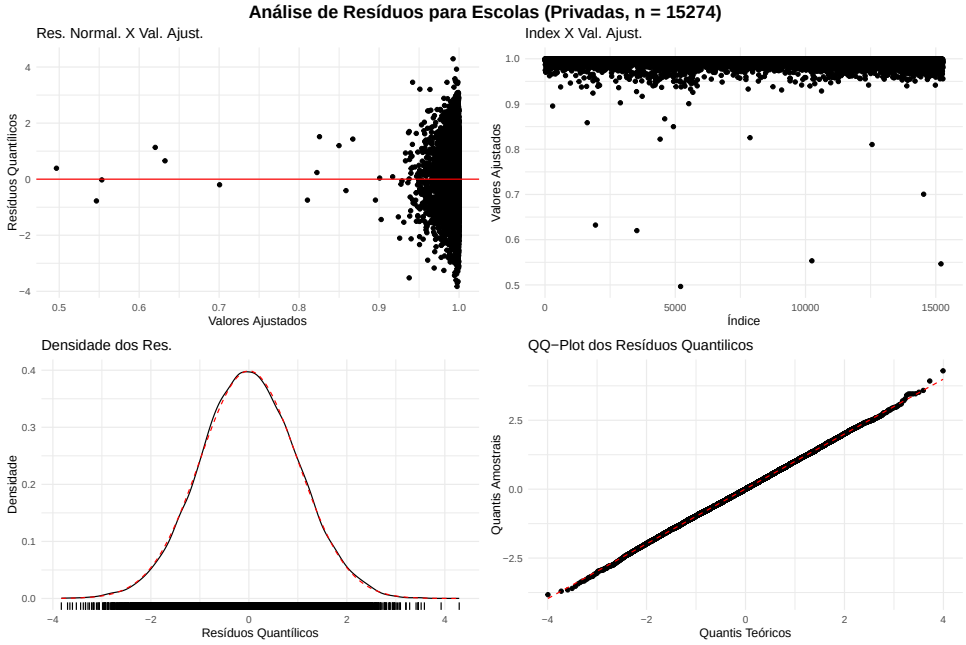


Figura 19 – Diagnóstico Residuos Quantilicos (*Hurdle*-Logístico, Privadas, População 2)

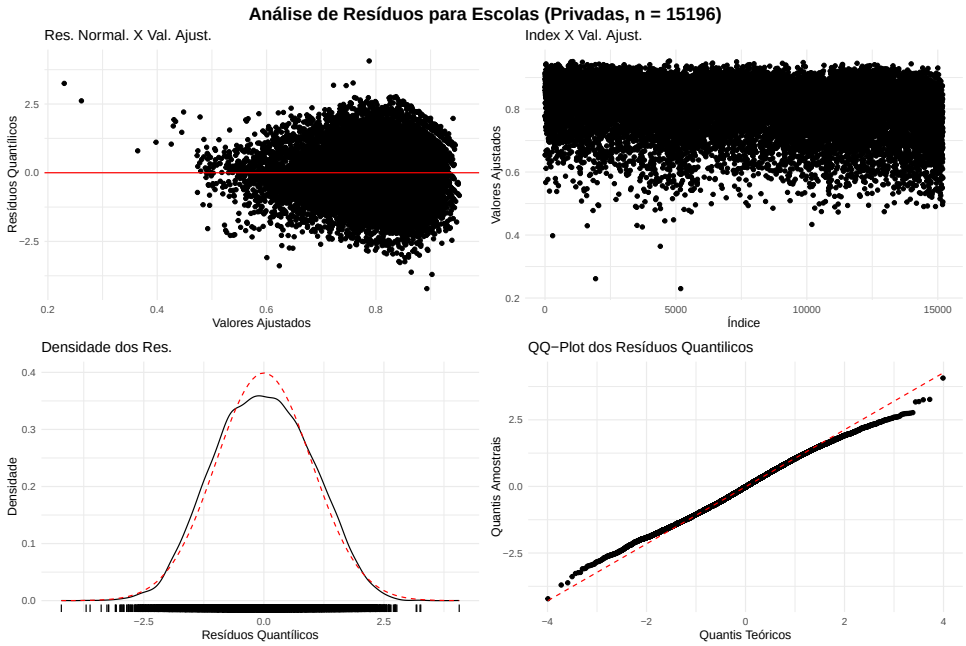


Figura 20 – Diagnóstico Residuos Quantilicos (*Hurdle*-BCT, Privadas, População 2).