Universidade Federal da Paraíba
Centro de Informática
Programa de Pós-Graduação em Modelagem Matemática e Computacional

José Wenes Pereira Lima

AVALIAÇÃO DA NOTA DE REDAÇÃO DO ENEM NO ESTADO DO CEARÁ VIA MODELO DE REGRESSÃO BETA INFLACIONADO

João Pessoa
Maio de 2023

Universidade Federal da Paraíba
Centro de Informática
Programa de Pós-Graduação em Modelagem Matemática e Computacional

AVALIAÇÃO DA NOTA DE REDAÇÃO DO ENEM NO ESTADO DO CEARÁ VIA MODELO DE REGRESSÃO BETA INFLACIONADO

José Wenes Pereira Lima

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional, UFPB, da Universidade Federal da Paraíba, como parte dos requisitos necessários à obtenção do título de Mestre em Modelagem Matemática e Computacional.

Orientadores: Tatiene Correia de Souza
Tarciana Liberal Pereira

João Pessoa
Maio de 2023

AVALIAÇÃO DA NOTA DE REDAÇÃO DO ENEM NO ESTADO DO CEARÁ VIA MODELO DE REGRESSÃO BETA INFLACIONADO

José Wenes Pereira Lima

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL (PPGMMC) DO CENTRO DE INFORMÁTICA DA UNIVERSIDADE FEDERAL DA PARAÍBA COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM MODELAGEM MATEMÁTICA E COMPUTACIONAL.

Examinada por:

gov.br Documento assinado digitalmente
TATIENE CORREIA DE SOUZA
Data: 03/07/2023 20:00:34-0300
Verifique em https://validar.iti.gov.br

Prof. Tatiene Correia de Souza, D.Sc.

Prof. Tarciana Liberal Pereira, D.Sc.

gov.br Documento assinado digitalmente
MARCELO RODRIGO PORTELA FERREIRA
Data: 06/07/2023 17:45:47-0300
Verifique em https://validar.iti.gov.br

Prof. Marcelo Rodrigo Portela Ferreira, D.Sc.

gov.br Documento assinado digitalmente
ANA HERMINIA ANDRADE E SILVA
Data: 05/07/2023 09:01:37-0300
Verifique em https://validar.iti.gov.br

Prof. Ana Hermínia Andrade e Silva, D.Sc.

JOÃO PESSOA, PB – BRASIL
MAIO DE 2023

Catalogação na publicação
Seção de Catalogação e Classificação

L732a Lima, José Wenes Pereira.
Avaliação da nota de redação do Enem no estado do Ceará via modelo de regressão beta inflacionado / José Wenes Pereira Lima. - João Pessoa, 2023.
43 f. : il.

Orientação: Tatiene Correia de Souza.
Coorientação: Tarciana Liberal Pereira.
Dissertação (Mestrado) - UFPB/CI.

1. Matemática computacional - Modelagem. 2. Enem - Exame Nacional do Ensino Médio. 3. Modelo de regressão beta inflacionado. I. Souza, Tatiene Correia de. II. Pereira, Tarciana Liberal. III. Título.

UFPB/BC CDU 519.6(043)

Elaborado por MAGNOLIA FELIX DE ARAUJO - CRB-15/883

Agradecimentos

Com estas palavras, quero, em primeiro lugar, expressar minha gratidão a Deus por me conceder força e coragem para alcançar este momento significativo em minha carreira profissional e em minha vida pessoal. Agradeço ao Pai misericordioso por estar sempre presente em todos os momentos da minha vida, por me guiar com sabedoria, paciência, saúde e dedicação, e por me ajudar a superar todos os obstáculos ao longo desta longa jornada.

Gostaria de agradecer também à minha esposa, Amanda e minha filha Maria Alice por estarem sempre ao meu lado, apoiando-me em todos os desafios desta jornada, com seu amor e carinho sem limites.

Aos meus pais, Selma e Francisco, agradeço pela confiança depositada em mim e pelo apoio incondicional que sempre me deram.

A minha vó Clara (Didi), meu sincero agradecimento por todo o amor e carinho que sempre me proporcionou.

Às professoras Tatiene e Tarciana, minhas orientadoras, mulheres de fibra e profissionais inspiradoras que acolheram e guiaram-me nessa caminhada, expresso todo o meu carinho e apreço.

Aos professores do PPGMMC-UFPB, em nome do professor Marcelo e da professora Ana Hermínia, agradeço a esses mestres espetaculares que me conduziram na busca pelo conhecimento.

Aos meus colegas, Joaquim, Rafael, Erissandro e José Arnaldo, expresso minha gratidão pela amizade e companheirismo.

Tenho a certeza de que este trabalho seria insuficiente em suas páginas para agradecer a todos os que, direta ou indiretamente, contribuíram para tornar meu sonho realidade. Por isso, expresso aqui meus sinceros agradecimentos a todos."

Resumo da Dissertação apresentada ao PPGMMC/CI/UFPB como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

AVALIAÇÃO DA NOTA DE REDAÇÃO DO ENEM NO ESTADO DO CEARÁ VIA MODELO DE REGRESSÃO BETA INFLACIONADO

José Wenes Pereira Lima

Maio/2023

Orientadores: Tatiene Correia de Souza
Tarciana Liberal Pereira

Programa: Modelagem Matemática e Computacional

O Exame Nacional do Ensino Médio (ENEM) é atualmente o mais importante instrumento de avaliação da educação básica de nível médio no Brasil, além disso, tal instrumento de avaliação tem nas notas obtidas pelos participantes um critério preponderante para o acesso às universidades públicas e privadas do país. A prova avalia objetivamente os participantes em quatro áreas, são elas: ciências humanas e suas tecnologias, ciências da natureza e suas tecnologias, matemática e suas tecnologias, linguagens e códigos e suas tecnologias. Além destas, o exame conta ainda com a prova de redação, sendo esta a única prova não objetiva do exame. A nota na prova de redação é importantíssima na aprovação do candidato em diversos cursos universitários. Haja vista que não se limita apenas à escrita, mas também avalia a capacidade do estudante de analisar um tema, construir argumentos consistentes, fundamentar suas ideias e propor soluções. Essas habilidades são relevantes para o ambiente acadêmico, onde é necessário realizar pesquisas, debater ideias e participar de discussões. A presente dissertação tem por objetivo identificar os fatores que influenciam na nota de redação dos participantes que realizaram o ENEM em 2019 no estado do Ceará. Para isso foi utilizado o modelo de regressão beta inflacionado, uma vez que a variável resposta apresenta assimetria e assume valores no intervalo [0,1]. Os dados foram obtidos por meio do portal do INEP, na página relacionada ao ENEM. O número de participantes observados foram, respectivamente, 74.943 que estudaram em escolas públicas e 5.279 que estudaram em escolas privadas. Por meio da análise dos modelos propostos, constatamos que, tanto em escolas públicas quanto privadas, as notas nas provas objetivas influenciam as notas de redação dos alunos. Adicionalmente, as notas de ciências Humanas, linguagens e códigos

e matemática e suas tecnologias apresentam forte influência na probabilidade dos participantes tirar nota 1000 na redação.

Palavras-Chave: ENEM, Modelo de Regressão Beta Inflacionado.

Abstract of Dissertation presented to PPGMMC/CI/UFPB as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

AVALIAÇÃO DA NOTA DE REDAÇÃO DO ENEM NO ESTADO DO CEARÁ VIA MODELO DE REGRESSÃO BETA INFLACIONADO

José Wenes Pereira Lima

May/2023

Advisors: Tatiene Correia de Souza
Tarciana Liberal Pereira

Program: Computational Mathematical Modelling

The National High School Exam (ENEM) is currently the most important instrument for assessing basic secondary education in Brazil. Furthermore, this assessment tool heavily relies on the scores obtained by participants as a predominant criterion for accessing public and private universities in the country. The exam objectively evaluates participants in four areas: humanities and their technologies, natural sciences and their technologies, mathematics and their technologies, and languages and codes and their technologies. In addition to these, the exam also includes a writing test, which is the only non-objective part of the exam. The score on the writing test is crucial for a candidate's approval in various university courses, as it goes beyond mere writing skills and assesses the student's ability to analyze a topic, construct sound arguments, provide evidence for their ideas, and propose solutions. These skills are relevant in an academic environment where conducting research, debating ideas, and participating in discussions are necessary. This dissertation aims to identify the factors that influence the scores on the writing test for participants who took the ENEM in 2019 in the state of Ceará, Brazil. To achieve this, the inflated beta regression model was used, given that the response variable exhibits asymmetry and takes values in the interval [0, 1]. The data was obtained through the INEP portal, specifically the page related to the ENEM. The number of observed participants was 74,943 from public schools and 5,279 from private schools. Through the analysis of the proposed models, it was found that both in public and private schools, the scores on the objective tests influence the scores on the writing test for students. Additionally, the scores on humanities, languages and codes, and mathematics and their technologies strongly influence the probability of participants achieving a score

of 1000 score on the writing test.

Keywords: ENEM, Inflated Beta Regression Model.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A educação desempenha um papel preponderante na transformação da sociedade não só reduzindo as desigualdades sociais, mas também seus impactos econômicos; desenvolvendo nas pessoas aspectos para o exercício da cidadania e qualificação para o trabalho. A constituição federal de 1988, em seu artigo 206, incumbe a União o dever de garantir a qualidade do ensino em todo o território nacional. Sob esse viés, a busca pela ampla qualidade e o enfrentamento dos desafios para atingir patamares educacionais de países desenvolvidos requer do país uma série de investimentos em educação e a elaboração de políticas públicas consistentes.

Uma ferramenta bastante utilizada na área de educação pelos governos mundiais são as avaliações em larga escala, que visa avaliar a qualidade dos sistemas de ensino e os contrastes entre seus grupos sociais. Segundo NÓVOA (2002) as avaliações em larga escala devem ser utilizadas para entender os processos educativos e apoiar a melhoria da qualidade da educação. O autor ressalta a importância de utilizar essas avaliações como instrumentos de compreensão e aprimoramento, em vez de focar exclusivamente nos resultados de classificação e ranking de escolas e estudantes.

Neste sentido, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) criou, em 1998, o Exame Nacional do Ensino Médio (ENEM), com o objetivo de avaliar o desempenho dos estudantes de escolas públicas e particulares na última etapa da educação básica. Ao longo dos anos, o ENEM vem se modificando e tornou-se um instrumento central na política educacional brasileira, configurando-se como porta de entrada no ensino superior para muitos estudantes brasileiros. Dentre as muitas funções do ENEM, destacamos a coleta de informações sobre os participantes e a categorização desses dados, favorecendo sua análise.

O INEP organiza os dados do ENEM em formato de microdados para cada ano da prova. Os dados são disponibilizados para acesso no site do instituto, a partir dos microdados, conseguimos realizar análises e obter padrões de desempenho dos participantes.

A exploração dos dados do ENEM, com ênfase na modelagem preditiva, é necessária como ferramenta de apoio à elaboração de políticas educacionais, visando direcionar os investimentos e potencializar a qualidade dos resultados. A prova de redação, considerada uma das partes mais desafiadoras do ENEM para muitos estudantes, desempenha um papel crucial na consolidação da nota do candidato, pois requer habilidades específicas de escrita e análise. No contexto educacional, destaca-se a ocorrência frequente de notas zero na redação em todas as edições, sendo então de extrema importância na elaboração de políticas na educação básica. Diante desse cenário, é comum recorrer a modelos de regressão para analisar as características que influenciam as notas de redação e, até mesmo, utilizar modelos preditivos como ferramentas inovadoras no campo educacional.

Alguns trabalhos utilizando modelos regressão para dados do ENEM vêm sendo desenvolvidos. VIGGIANO e MATTOS (2013) estudaram o desempenho dos estudantes no ENEM em 2010 em diferentes regiões do Brasil e concluíram que os participantes das regiões sudeste e nordeste têm melhor desempenho na redação do que os das demais regiões. SANTOS (2018) buscou desenvolver modelos para investigar a influência de características socioeconômicas nas notas de redação e matemática do ENEM 2018, o estudo utilizou técnicas de regressão para construir os modelos. Os resultados destacam a importância de características como renda familiar, escolaridade dos pais, tipo de escola e acesso à internet para explicar as notas obtidas pelos candidatos. PORTO (2019) realizou uma mineração de dados educacionais para avaliar possíveis influências no desempenho dos candidatos do ENEM. Os autores mostram que a partir da técnica correta de mineração de dados é possível analisar e caracterizar as escolas através do desempenho dos alunos. No trabalho de RÊGO (2021) o autor avaliou a influência de alguns fatores na nota de redação do ENEM 2019, aplicando o modelo de regressão beta inflacionado em zero, ao qual o autor se refere como beta ajustado em zero. A análise foi realizada para as notas de candidatos que realizaram a prova do ENEM no ano de 2019 no estado do Rio Grande do Norte. SANTOS (2018) buscou desenvolver modelos para investigar a influência de características socioeconômicas nas notas de redação e matemática do ENEM 2018, o estudo utilizou técnicas de regressão para construir os modelos. Os resultados destacam a importância de características como renda familiar, escolaridade dos pais, tipo de escola e acesso à internet para explicar as notas obtidas pelos candidatos.

Este trabalho tem como objetivo geral investigar e analisar a influência de fatores, na nota de redação do ENEM no estado do Ceará, uma vez que as notas de redação são variáveis contínuas, distribuídas no intervalo (0,1) mas que podem conter excesso de zeros e/ou uns, o modelo de regressão linear normal não é adequado. Adicionalmente, identificamos a necessidade de encontrar fatores que influenciam na

ocorrência de notas iguais a zero. Um modelo de regressão beta inflacionado, para variáveis limitadas ao intervalo unitário padrão, mas que podem conter zeros e/ou uns foi proposto por OSPINA e FERRARI (2012). Na literatura, já existem vários trabalhos sobre análise de problemas cotidianos utilizando modelos de regressão beta inflacionados, vale destacar: PEREIRA *et al.* (2014) utilizaram o modelo de regressão beta inflacionado para explicar a eficiência administrativa dos municípios brasileiros por Região. O objetivo do trabalho era avaliar e comparar os desempenhos das regiões brasileiras no que se refere ao gerenciamento de recurso público. DA SILVA *et al.* (2018) avaliaram a eficiência da atenção básica à saúde em municípios da Paraíba, utilizando um modelo de regressão beta inflacionado. Foi observado que a eficiência média foi maior em municípios menores e influenciada positivamente pelo tamanho da população, condições de saneamento básico e índice FIRJAN de desenvolvimento municipal na saúde. Por outro lado, o número de consultas médicas por estabelecimento de saúde exerceu efeito negativo sobre a eficiência. DINIZ e MELO (2019) utilizaram o modelo de regressão beta inflacionado em zero e um, para verificar quais fatores influenciam a proporção de mulheres nas empresas. Os autores concluíram que a medida que o tempo médio de estudo da empresa aumenta, a proporção de mulheres nas empresas também aumenta. Por outro lado, à medida que a renda média da empresa aumenta, a proporção de mulheres nas empresas diminui.

O presente trabalho está organizado da seguinte maneira: No Capítulo 2, apresentamos um breve histórico do ENEM e as características gerais da prova. O Capítulo 3 apresenta alguns conceitos básicos das distribuições, beta e beta inflacionada, como também situa o leitor sobre o modelo de regressão utilizado nesse trabalho. O Capítulo 4 desta dissertação dedica-se à apresentação dos dados e estatísticas descritivas da base de dados utilizada. Além disso, abordamos os resultados do ajuste do modelo de regressão beta inflacionado para analisar as notas de redação dos candidatos ao ENEM no estado do Ceará em 2019, considerando tanto alunos de escolas públicas quanto privadas. Por fim, no Capítulo 6, apresentamos as conclusões e considerações finais.

Capítulo 2

Revisão de Literatura

Neste Capítulo apresentamos um breve histórico do ENEM e sua características, enfatizando a prova de redação que é foco desta pesquisa.

2.1 O ENEM e suas Características

Considerado por muitos pesquisadores como a maior avaliação em larga escala do Brasil, o ENEM foi criado em 1998 (Portaria MEC n.º 438/1998) pelo INEP, um órgão vinculado ao Ministério da Educação (MEC) responsável pela produção de estudos e pesquisas sobre a educação básica e o ensino superior do país. Dentre os objetivos do ENEM, destaca-se a análise do desempenho escolar ao fim do ciclo básico de educação e a colaboração no ingresso no ensino superior; substituindo os antigos vestibulares em muitos cursos nas instituições públicas e até mesmo em instituições privadas, outrossim tem como base viabilizar o acesso a programas de bolsas (ProUni) e de financiamento estudantil (Fies) (INEP, 2022).

Nessa perspectiva, os resultados do ENEM fornecem parâmetros de comparação e análise de melhorias na qualidade do ensino no país. Em 2009, o MEC realizou uma grande reformulação metodológica no ENEM, o exame passou a ser usado como forma de seleção unificada nos processos seletivos das Universidades Federais, além de continuar a ser utilizado para a seleção dos bolsistas do ProUni. Soma-se a isso o processo seletivo de instituições particulares de ensino superior.

A partir de 2009, o ENEM passou a ter quatro áreas, com provas de múltipla escolha e uma prova discursiva de redação. Às quatro áreas do conhecimento abordadas na prova são: (1) linguagens, códigos e suas tecnologias; (2) ciências humanas e suas tecnologias; (3) ciências da natureza e suas tecnologias e (4) matemática e suas tecnologias.

A redação do ENEM é uma prova totalmente subjetiva e exige que o candidato elabore um texto dissertativo-argumentativo sobre o tema proposto, seguindo a norma padrão da língua portuguesa. Ademais, o tema gera muita expectativa nos

candidatos durante toda sua preparação para o exame, mas ele só é revelado na hora da prova e está normalmente contextualizado com assuntos importantes, os quais, em sua maioria, abordam problemas sociais.

A nota atribuída ao candidato é uma variável contínua, com valor mínimo 0 e valor máximo igual a 1000. A correção da prova de redação dos candidatos é feita por dois professores, que avaliam, de forma independente, as cinco competências listadas na Tabela 2.1,

Tabela 2.1: Competências para avaliação da prova de Redação do ENEM 2019.

Número da competência	Nome da competência	Nota Máxima
1	Demonstrar domínio da modalidade escrita formal da língua portuguesa.	200
2	Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa.	200
3	Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.	200
4	Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.	200
5	Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.	200
	Nota Máxima da redação do ENEM	1000

Fonte: Cartilha da Redação do Enem 2019.

Aqui cabe ressaltar que a nota final atribuída ao aluno é dada pela média das notas dos dois avaliadores. Caso haja diferença superior a 100 entre as notas dos dois avaliadores, a redação do candidato é analisada por um terceiro avaliador e se a divergência continuar, será analisada por uma banca (INEP, 2022). Além disso, cada competência avaliada tem a nota máxima de 200 pontos, e a soma das notas em cada competência forma a nota total da redação.

Capítulo 3

Conceitos Básicos

3.1 Distribuições de Probabilidade

Nesta seção, apresentamos as distribuições beta e beta inflacionada, em seguida, apresentamos o modelo de regressão beta inflacionado, que utiliza essa distribuição como base para análise de dados.

3.1.1 Distribuição Beta

A distribuição beta é uma distribuição de probabilidade contínua com dois parâmetros positivos que determinam sua forma. Ela tem como suporte o intervalo aberto $(0, 1)$. MOOD *et al.* (1974) definiram a distribuição beta como: dado uma variável aleatória Y que segue distribuição beta, com parâmetros $p, q > 0$, sua função densidade pode ser escrita na forma:

$$f_y(y; p, q) = \frac{1}{B(p, q)} y^{p-1}(1-y)^{q-1} I_{(0,1)}(y),$$

em que $0 < y < 1$, e $B(p, q) = \int_0^1 y^{p-1}(1-y)^{q-1} dy$.

A notação usada para indicarmos que Y é uma variável aleatória com distribuição beta é $\mathrm{Y} \sim Beta(p, q)$.

A média e a variância de Y são dadas, respectivamente, por:

$$\mathrm{E(Y)} = \mu_y = \frac{p}{p+q} \qquad \text{e}$$

$$\mathrm{Var(Y)} = \sigma_y^2 = \frac{pq}{(p+q+1)(p+q)^2}.$$

Vale destacar a relação entre a função beta e função gama $\Gamma(\cdot)$, em que podemos expressar:

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}.$$

Assim a função de distribuição acumulada (f.d.a) de uma variável Y com distribuição beta é dada por:

$$F(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \int_0^y y^{p-1}(1-y)^{q-1}dy, \qquad 0 < y < 1$$

$$= \frac{B_y(y;p,q)}{B(p,q)}.$$

A distribuição beta destaca-se como uma distribuição bastante flexível, sendo uma das mais usadas em problemas de modelagem que envolve taxas, razões e proporções. PEREIRA *et al.* (2014) realizaram a análise das eficiências administrativas dos municípios do estado de São Paulo com base em modelos de regressão beta e beta inflacionado com efeitos espaciais. DE LIMA LEITE e DAS VIRGENS FILHO (2007) utilizaram a distribuição beta como modelo probabilístico para análise de dados de velocidade do vento em Ponta Grossa, Paraná. O objetivo do estudo foi avaliar a adequação da distribuição beta como modelo para descrever a distribuição de velocidades do vento e compará-la com outras distribuições comumente utilizadas na literatura. DE ARAÚJO SOARES *et al.* (2019) realizaram uma avaliação epidemiológica da esquistossomose no estado de Pernambuco, Brasil, utilizando um modelo de regressão beta. No estudo o modelo de regressão beta é utilizado para analisar a relação entre a prevalência da esquistossomose e variáveis socioeconômicas e ambientais, como idade, sexo, escolaridade, acesso à água potável e saneamento básico.

É muito comum nos estudos de análise de regressão que o objetivo do pesquisador seja modelar a resposta média. Desta forma, FERRARI e CRIBARI-NETO (2004) propuseram uma nova reparametrização da distribuição beta, que objetiva obter uma estrutura de regressão para a média da resposta, incluindo um parâmetro de precisão. Logo, seja $\mu = \dfrac{p}{(p+q)}$ e $\phi = p + q$, temos que $p = \mu\phi$ e $q = (1-\mu)\phi$. Sob a nova reparametrização a função densidade de Y é:

$$f_y(y,\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad \forall y \in (0,1), \tag{3.1}$$

em que $0 < \mu < 1$ e $\phi > 0$. Agora dizemos que Y tem distribuição beta com média μ e precisão ϕ e usamos como notação $Y \sim B(\mu,\phi)$. A média e variância agora podem escritas, respectivamente, como:

$$\mathrm{E(Y)} = \mu \quad \text{e}$$
$$\mathrm{V(Y)} = \frac{\mu(1-\mu)}{\phi+1}.$$

3.1.2 Distribuição Beta Inflacionada

Com vários avanços em problemas na ciência resolvidos pelo uso de ferramentas estatísticas e das distribuições de probabilidade, é comum encontrar situações em que as taxas, índices ou proporções não estejam distribuídas no intervalo $(0,1)$ mas sim nos intervalos [0,1), (0,1] ou [0,1]. Como solução surge uma família de distribuições, denominada distribuição beta inflacionada proposta por OSPINA e FERRARI (2010). Para o autor, o termo *inflacionado* aplica-se nesse caso ao fato de que a massa de probabilidades de alguns pontos excede o que é permitido pelo modelo.

Conforme PEREIRA *et al.* (2014) a distribuição inflacionada em zero ou um é uma mistura entre uma distribuição degenerada em c, em que c é igual a 0 ou 1, e uma variável distribuída de forma contínua no intervalo (0,1). Segundo OSPINA e FERRARI (2010) a função densidade de probabilidade da distribuição beta inflacionada em zero ou um é dada por:

$$\mathrm{bi}_c(y;\alpha,\mu,\phi) = \begin{cases} \alpha, & \text{se } y = c, \\ (1-\alpha)f(y;\mu,\phi), & \text{se } y \in (0,1), \end{cases} \tag{3.2}$$

em que $0 < \alpha, \mu < 1$ e $\phi > 0$, sendo $f(y;\mu,\phi)$ a função densidade (3.1). Definida a função densidade de uma variável aleatória beta inflacionada no ponto c, em que c igual a 0 ou 1, note que: se $c = 0$, a distribuição é um beta inflacionada em 0, com notação usual $Y \sim BIZ(\alpha,\mu,\phi)$. Se $c = 1$, a distribuição é uma beta inflacionada em 1, a qual escrevemos $Y \sim BIU(\alpha,\mu,\phi)$.

A média e a variância da distribuição $BIZ(\alpha,\mu,\phi)$ são dadas, respectivamente, por:

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

A média e a variância da distribuição $BIU(\alpha,\mu,\phi)$ são dadas, respectivamente, por:

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

É possível verificar então que μ não é o valor esperado da variável resposta e sim a média condicional, ou seja $\mu = E(Y \mid Y \in (0,1))$.

Seja Y uma variável aleatória que assume valores no intervalo fechado [0,1]. Dizemos que Y tem distribuição beta inflacionada em zero e um, com parâmetros (α, p, μ, ϕ) se tem função densidade

$$bizu(y;\alpha,p,\mu,\phi) = \begin{cases} \alpha p, & se\ y = 1 \\ \alpha(1-p), & se\ y = 0, \\ (1-\alpha)f(y;\mu,\phi), & se\ y \in (0,1), \end{cases} \tag{3.3}$$

em que $0 < \alpha, p, \mu < 1$ e $\phi > 0$, sendo $f(y;\mu,\phi)$ a função densidade da distribuição beta. Usamos a notação $Y \sim BIZU(\alpha, p, \mu, \phi)$ para indicar que Y segue distribuição beta inflacionada em zero e um. É possível verificar que $P(Y = 0) = \alpha(1-p)$ e $P(Y = 1) = \alpha p$. A seguir apresentamos a esperança e variância de uma variável aleatória com distribuição beta inflacionada em zero e um, dadas, respectivamente, por:

$$\begin{aligned} &\mathrm{E(Y)} = \alpha p + (1-\alpha)\mu \qquad \text{e} \\ &\mathrm{Var(Y)} = \alpha p(1-p)\frac{(1-\alpha)\mu(1-\mu)}{(\phi+1)} + \alpha(1-\alpha)(p-\mu)^2. \end{aligned}$$

3.2 Modelo de Regressão Beta Inflacionado

A solução de problemas em diversos ramos da ciência tem como ferramenta principal os modelos estatísticos. Destacamos a utilização dos modelos de regressão, uma técnica estatística muito utilizada para avaliar a relação entre uma variável aleatória de interesse, denominada de variável dependente (Y) e um conjunto de variáveis independentes $(\mathrm{X_1}, \cdots, \mathrm{X_k})$. Segundo ALTMAN e KRZYWINSKI (2015) o modelo de regressão linear é o mais básico e o mais utilizado em análises experimentais. Contudo, o modelo clássico torna-se inadequado diante de algumas situações, por exemplo, quando a variável de interesse apresenta-se na forma de taxas, frações ou proporções e é distribuída no intervalo unitário padrão.

O modelo de regressão beta foi proposto por FERRARI e CRIBARI-NETO (2004) com o intuito de superar algumas limitações dos modelos clássicos. Nesse modelo a variável resposta segue distribuição beta e relaciona a resposta média com um preditor linear por meio de uma função de ligação. Os modelos de regressão beta se assemelham aos Modelos Lineares Generalizados (MLGs) propostos por NELDER e WEDDERBURN (1972). Porém, vale ressaltar que podemos ter situações em que a variável resposta está distribuída no intervalo [0,1), (0,1] ou [0,1]. Diante de tais situações, surgem os modelos beta inflacionados, dentre as quais destaca-se OSPINA e FERRARI (2012), que propuseram o modelo beta inflacionado em zero ou um e zero e um.

Conforme OSPINA e FERRARI (2012) sejam $y_1, ..., y_n$ variáveis aleatórias independentes cada uma com distribuição beta inflacionada no ponto c ($c = 0$ *ou* $c = 1$) dada por 3.2. Supondo as seguintes relações funcionais para a média condicional de y_t, a massa de probabilidade em c e o parâmetro de precisão, o modelo de regressão beta inflacionado em c com dispersão variável é definido.

$$h(\alpha_t) = \sum_{i=1}^{M} z_{ti}\gamma_i = \zeta_t,$$
$$g(\mu_t) = \sum_{i=1}^{m} x_{ti}\beta_i = \eta_t,$$
$$b(\phi_t) = \sum_{i=1}^{q} s_{ti}\lambda_i = \kappa_t,$$

em que $\gamma = (\gamma_1, ..., \gamma_M)^\top$,$\beta = (\beta_1, ..., \beta_m)^\top$ e $\lambda = (\lambda_1, ..., \lambda_q)^\top$ são vetores de parâmetros de regressão desconhecidos, tais que $\gamma \in \mathbb{R}^M$, $\beta \in \mathbb{R}^m$ e $\lambda \in \mathbb{R}^q$, $x_{t1}, ..., x_{tm}$, $z_{t1}, ..., z_{tM}$ e $s_{t1}, ..., s_{tq}$ são observações de covariáveis $(m + M + q < n)$ que podem coincidir total ou parcialmente. As funções de ligação $h : (0, 1) \longrightarrow \mathbb{R}$, $g : (0, 1) \longrightarrow \mathbb{R}$ e $b : (0, \infty) \longrightarrow \mathbb{R}$ são estritamente monótonas e duas vezes diferenciáveis. Entre as funções de ligação mais utilizadas para α e μ estão a função logit com $g(\mu) = log(\mu/(1 - \mu))$, a função probit com $g(\mu) = \Phi^{-1}(\mu)$, em que $\Phi(.)$ é a função de distribuição da normal padrão, a especificação log-log complementar com $g(\mu) = log(-log(1-\mu))$, a ligação log-log com $g(\mu) = -log(-log(\mu))$ e a função de ligação de Cauchy com $g(\mu) = tan(\pi(\mu - 0,5))$. Já para ϕ é possível utilizar a função logarítmica com $b(\phi) = log(\phi)$ ou raiz quadrada com $b(\phi) = \sqrt{\phi}$.

A função de log-verossimilhança para o modelo de regressão beta inflacionado em c é da forma:

$$\ell(\theta) = \ell_1(\gamma) + \ell_2(\beta, \lambda)$$

em que

$$\ell_1(\gamma) = \sum_{t=1}^{n} \ell_1(\alpha_t) \quad \text{e} \quad \ell_2(\beta, \lambda) = \sum_{t:y_t \in (0,1)} \ell_t(\mu_t, \phi_t).$$

Considere agora que $y_1, y_2, ..., y_n$ seguem distribuição beta inflacionada em zero e um. O modelo de regressão beta inflacionado em zero e um com precisão constante é definido por 3.3 e pelos componentes sistemáticos:

$$g(\mu_t) = \sum_{i=1}^{n} x_{ti}\beta_i = \eta_t$$
$$H(\delta_{0t}, \delta_{1t}) = (h_0(\delta_{0t}, \delta_{1t}), h_1(\delta_{0t}, \delta_{1t})) = (\zeta_{0t}, \zeta_{1t}),$$

em que $\mu_t = E(y_t|y_t \in (0,1)), \delta_{0t} = P(y_t = 0), \delta_{1t} = P(y_t = 1)$ e $1-\delta_{0t}-\delta_{1t} = P(y_t \in (0,1))$. $\eta_t = x_t^\top \beta$, $\zeta_{0t} = v_t^\top \rho$ e $\zeta_{1t} = z_t^\top \gamma$ são preditores lineares; $\beta = (\beta_1, \beta_2, ..., \beta_k)^\top$, $\rho = (\rho_1, \rho_2, ..., \rho_{k0})^\top$e $\gamma = (\gamma_1, \gamma_2, ..., \gamma_{k1})^\top$ são vetores de parâmetros da regressão a serem estimados,tais que $\beta \in \mathbb{R}^k$, $\rho \in \mathbb{R}^{k_0}$e $\gamma \in \mathbb{R}^{k_1}$. $x_t = (x_{t1}, x_{t2}, ..., x_{tk})^\top$, $v_t = (v_{t1}, v_{t2}, ..., v_{tk_0})^\top$ e $z_t = (z_{t1}, z_{t2}, ..., z_{tk_1})^\top$ os valores observados de k,k_0 e k_1 variáveis exógenas conhecidas respectivamente.

Assume-se que a função de ligação $g : (0,1) \longrightarrow \mathbb{R}$ é uma função estritamente monótona e duplamente diferenciável. A função H é uma transformação bijetora do conjunto $C = \{(\delta_{0t}, \delta_{1t}) : 0 < \delta_{0t} < 1, 0 < \delta_{1t} < 1 - \delta_{0t}\}$ a $\mathbb{R}^2$duplamente diferenciável. Sob condições impostas para H, garante-se que as derivadas parciais de $\delta_{0t} = h_0^*(\zeta_{0t}, \zeta_{1t})$ e de $\delta_{1t} = h_1^*(\zeta_{0t}, \zeta_{1t})$ são contínuas em $\mathbb{R}^2$ e δ_{0t},δ_{1t} podem ser escritos em termos de ζ_{0t}e ζ_{1t} de forma única. Logo a função H pode ser escolhida de forma geral para satisfazer as condições exigidas acima. Um exemplo é que, considerando H tal que

$$H(\delta_{0t}, \delta_{1t}) = (h_0(\delta_{0t}, \delta_{1t}), h_1(\delta_{0t}, \delta_{1t})) = \left(h\left(\frac{\delta_{0t}}{1-\delta_{0t}-\delta_{1t}},\right), h\left(\frac{\delta_{1t}}{1-\delta_{0t}-\delta_{1t}},\right)\right),$$

em que a função $h : \mathbb{R}^+ \to \mathbb{R}$ estritamente monótona e duas vezes diferenciável. Veja que h_0 e h_1 são funções de $\mathbb{R}^+$ em $\mathbb{R}$.

A função de log-verossimilhança para o modelo de regressão beta inflacionado em zero e um é dada por:

$$l(\theta) = l_1(\rho, \gamma) + l_2(\beta, \phi)$$

no qual,

$$l_1(\rho, \gamma) = \sum_{t=1}^{n} l_t(\delta_{0t}, \delta_{1t}) \text{ e } l_2(\beta, \phi) = \sum_{t:y_t \in (0,1)} l_t(\mu_t, \phi)$$

em que,

$$\begin{aligned} l_t(\delta_{0t}, \delta_{1t}) &= I_{\{0\}}(y_t) + log\delta_{0t} + I_{\{1\}}(y_t) log\delta_{1t} \\ &\quad +(1 - I_{\{0\}}(y_t) - I_{\{1\}}(y_t)) log(1 - \delta_{0t} - \delta_{1t}), \\ l_t(\mu_t, \phi) &= log\Gamma(\phi) - log\Gamma(\mu_t\phi) - log((1-\mu_t)\phi) + (\mu_t\phi - 1) logy_t \\ &\quad +\{(1-\mu_t)\phi - 1\} log(1 - y_t). \end{aligned}$$

Os modelos beta inflacionados podem ser classificados numa classe bem mais completa para a modelagem, a classe dos Modelos Aditivos Generalizados para Locação, Escala e Forma (GAMLSS) (RIGBY *et al.*, 2019). Esses modelos permitem modelar não só a média, mas também a precisão e além da função de ligação podemos usar funções de suavização melhorando o processo de modelagem e a qualidade dos ajustes. Nessa dissertação iremos abordar o modelo de regressão beta inflacionado em zero e um com precisão variável.

Capítulo 4

Resultados e Discursões

Neste Capítulo apresentamos o processo de obtenção dos dados e posteriormente uma análise descritiva das variáveis usadas no estudo. Adicionalmente, apresentamos os resultados dos modelos de regressão beta inflacionados utilizados para alcançar o objetivo desse estudo que é encontrar os fatores que influenciam a nota de redação.

4.1 Descrição dos Dados

O banco de dados é composto por informações dos candidatos que realizaram o ENEM em 2019 no estado do Ceará e se o candidato não está fazendo o exame por experiência. Foram considerados critérios de inclusão, o local de nascimento e residência que nesse estudo foi o estado do Ceará.

O estado em foco neste estudo é o Ceará, localizado na região nordeste do Brasil, caracterizado por sua extensão territorial no semiárido brasileiro. Atualmente, o Ceará se destaca como o estado que mais cresceu no Índice de Desenvolvimento da Educação Básica (Ideb), sendo um exemplo para o Brasil em termos de resultados educacionais alcançados, apesar das condições adversas de escassez de recursos públicos. Esse crescimento na área da educação pode ser atribuído à eficiência nos investimentos realizados pelo estado. Conforme SOUSA e OLIVEIRA (2010) o Ceará implantou, em 1992 o Sistema Permanente de Avaliação da Educação Básica do Ceará (SPAECE), mas só a partir de 2007 o SPAECE foi incorporado ao ensino médio, aplicando a avaliação aos alunos do 3º ano. Dentre seus objetivos, o SPAECE promove aos governantes e gestão escolar uma visão clara e concreta do processo de ensino e aprendizagem e os desafios para melhoria da qualidade do ensino (DO AMARAL SOARES e WERLE, 2016). Conforme LOUREIRO *et al.* (2020) o sucesso da proposta educacional do Ceará está fundamentada na ciência da aprendizagem, destacando entre seus princípios básicos, o uso de um sistema robusto e confiável de avaliação da aprendizagem.

Os dados utilizados nesse trabalho foram extraídos no portal do INEP, na página relacionada ao ENEM, pelo link: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem. Os microdados do ENEM são arquivos que contêm os questionários respondidos pelos participantes durante o processo de inscrição do exame, armazenando todas as informações disponibilizadas pelos inscritos. A Tabela 4.1 apresenta as variáveis utilizadas para análise de dados nessa pesquisa. Para as edições do ENEM nos anos 2020 e de 2021, o INEP não divulgou o local de nascimento e residência do candidato. A ausência dessas informações na base de dados, inviabilizou o uso dos microdados dos anos de 2020 e 2021, haja vista os critérios de seleção preconizados para este trabalho. A variáveis SEXO, RENDA, INTERNET e COMPUTADOR são qualitativas e as variáveis RD, LC, MT, CN, CH e IDADE são quantitativas. A amostra utilizada nesse trabalho foi de 80.222 candidatos, sendo 74.943 de escolas públicas e 5.279 de escolas privadas. Todo o processo de organização e análise de dados dessa pesquisa foi realizado no *software* R CORE TEAM (2022).

Tabela 4.1: Descrição das variáveis utilizadas.

Variável	**Descrição**
RD	Nota do participante em Redação
LC	Nota do participante em linguagens e códigos.
MT	Nota do participante em Matemática e sua tecnologias.
CN	Nota do participante em Ciências da Natureza e suas Tecnologias.
CH	Nota do participante em Ciências Humanas e suas Tecnologias.
SEXO	Sexo do participante:(M) Masculino e (F) Feminino.
IDADE	Idade do participante.
RENDA	Varável Categórica: A renda do participante dividida em quatro classes: Classe 1- $\leqslant 998,00; Classe2 - 998,00 \mid - 1996,00; classe3 - 1996,00 \mid - 4990,00; classe4 - \geqslant 4990,00$
INTERNET	Situação de acesso a internet na residência do participante:(A) Sim e (B) Não.
COMPUTADOR	Variável *dummy*: 1 se o participante tem computador, 0 caso contrário.

4.2 Estatística Descritiva das Variáveis

4.2.1 Análise descritiva das notas de redação dos candidatos de escolas públicas no Ceará em 2019.

A Tabela 4.2 apresenta algumas medidas descritivas das notas das provas objetivas e de redação dos alunos de escolas públicas. As notas foram convertidas para o intervalo [0,1) pela divisão pelo valor máximo (1000). Para cada uma das provas objetivas - Ciências da Natureza (CN), Linguagens e Códigos (LC), Matemática (MT) e Ciências Humanas (CH) e para a nota de Redação foram calculados os valores: da média (μ), desvio-padrão (σ), mediana, mínimo, máximo, percentual de zeros ($\%Zeros$) e coeficiente de variação (CV).

Tabela 4.2: Média (μ), desvio-padrão (σ), mediana, mínimo, máximo, percentual de zeros ($\%Zeros$), coeficiente de variação (CV) das notas dos candidatos de escolas públicas no Ceará em 2019.

	RD	**CN**	**LC**	**CH**	**MT**
μ	0,52	0,44	0,49	0,47	0,49
σ	0,22	0,06	0,06	0,07	0,09
Mediana	0,56	0,43	0,49	0,46	0,46
min	0,00	0,00	0,00	0,00	0,00
max	0,98	0,78	0,70	0,75	0,95
$\%Zeros$	8,12	0,01	0,04	0,08	0,01
CV	0,42	0,14	0,13	0,15	0,18

É possível constatar, na Tabela 4.2, que a nota de redação apresenta um maior percentual de zeros, bem como um maior coeficiente de variação. Vale destacar também que as notas de ciências e matemática apresentam o menor percentual de zeros (1%).

A Figura 4.1 apresenta o histograma das notas de redação para os alunos de escola pública. No gráfico as barras com pontos acima representam as quantidades de zeros na amostra. Observamos claramente o inflacionamento em zero, neste caso a distribuição beta inflacionada em zero é ideal para o conjunto de dados.

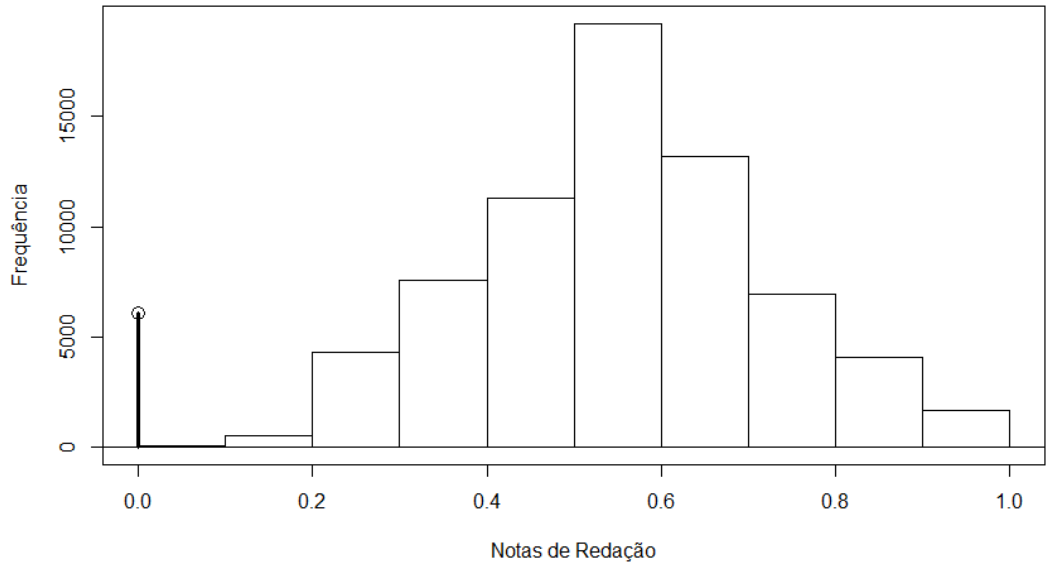
A Figura 4.2 apresenta os *boxplots* das notas das provas objetivas e da nota de redação. Observamos que a nota de redação tem um comportamento assimétrico. A Figura 4.3 apresenta os *boxplots* por sexo das notas das provas objetivas e de redação. Observamos que as notas de redação apresentam média maior para os candidatos do sexo feminino e que para esse grupo os dados apresentam-se mais simétricos. Para verificar se existe efeito do sexo do candidato nas notas de redação, aplicamos o teste de Mann-Whitney, para o qual adotamos as seguintes hipóteses:

H_0: as notas medianas para candidatos do sexo masculino e feminino são iguais.
H_1: as notas medianas para candidatos do sexo masculino e feminino são diferentes.

Ao nível de significância de 5%, rejeitamos a hipótese nula (p-valor<0,0001), ao rejeitarmos a hipótese nula significa dizer que existe uma relação entre a nota de redação e o sexo do candidato.

A Figura 4.4 apresenta o gráfico da matriz de correlação de Pearson entre as variáveis contínuas. Observamos que a variável que apresenta maior correlação com a nota de redação é a nota de linguagens e códigos. Vale destacar uma forte correlação entre a variáveis: nota de linguagens e códigos (LC_PU) e as nota de ciências humanas (CH_PU) com correlação positiva.

Figura 4.1: Histograma das notas da redação dos candidatos de escolas públicas no Ceará em 2019.



A Figura 4.5 ilustra o mapa das médias das notas de redação por município do Ceará, considerando as escolas públicas. Observamos que nenhum município apresentou média superior a 0,8, bem como verificamos uma predominância de médias no intervalo 0,4|-0,6. Apenas em sete municípios cearenses a média das notas de redação está contida no intervalo de 0,6|-0,8, com destaque para o município de Abaiara que obteve a melhor média de 0,648.

Figura 4.2: *Boxplot* das notas dos candidatos de escolas públicas no Ceará em 2019.

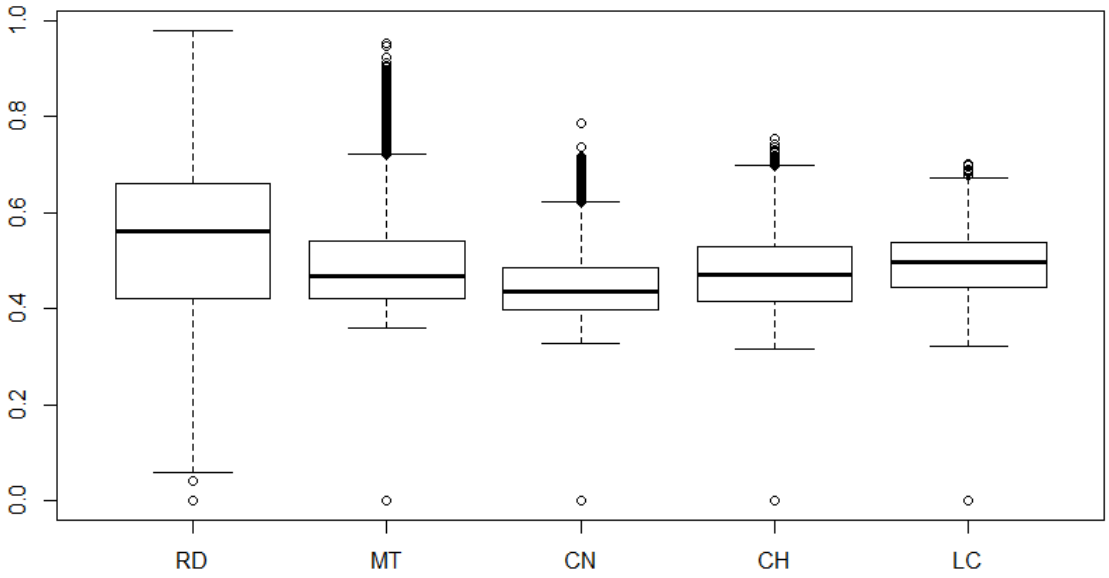


Figura 4.3: Boxplots das notas dos candidatos de escolas públicas por sexo no Ceará em 2019.

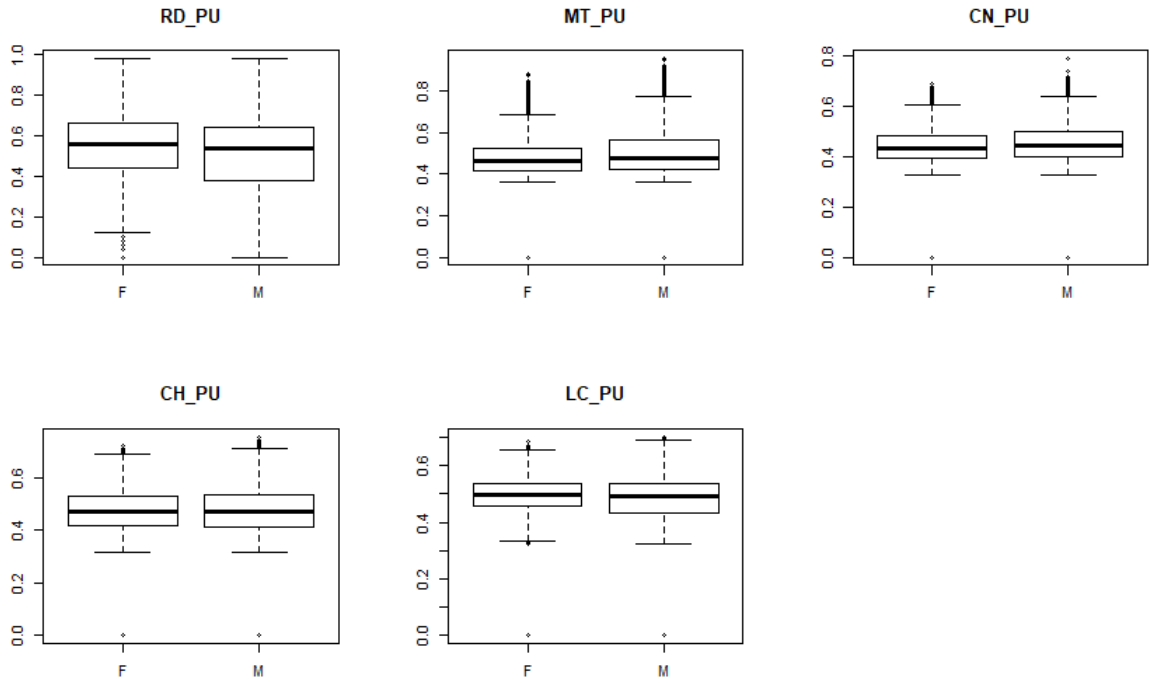


Figura 4.4: Matriz de correlação da Nota de Redação dos candidatos de escola pública no Ceará em 2019.

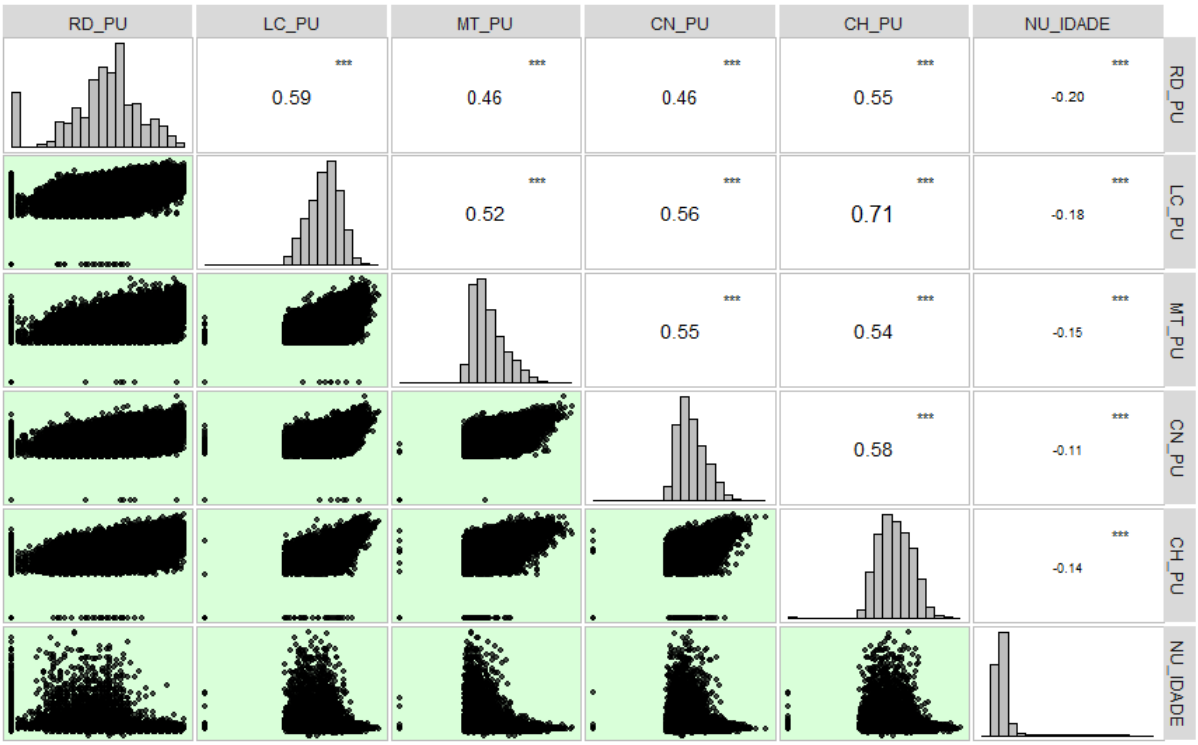
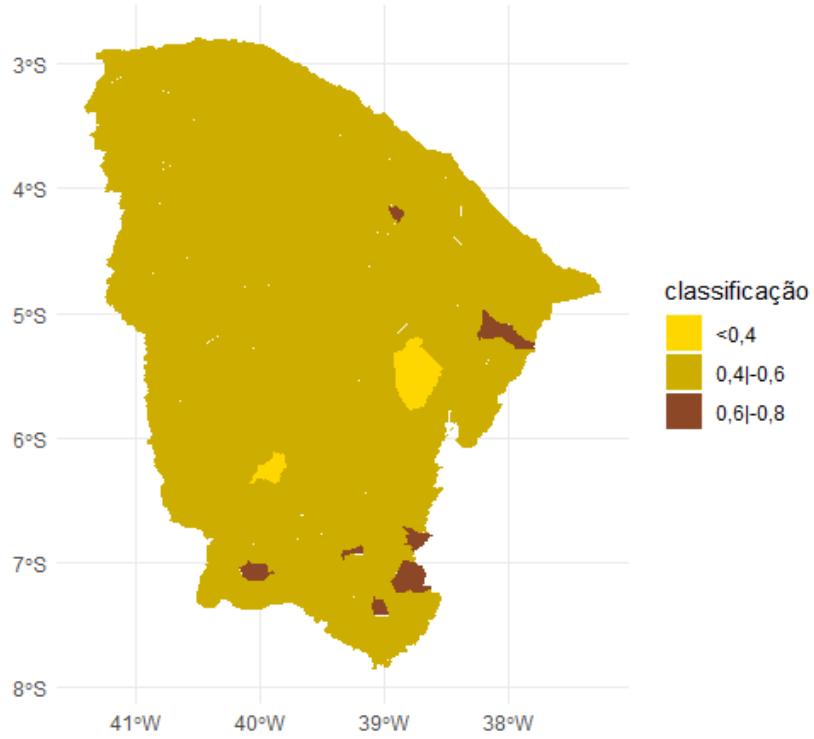


Figura 4.5: Mapa das médias de redação de alunos de escolas públicas por municípios do Ceará.



4.2.2 Análise descritiva das notas de redação dos candidatos de escola privada no Ceará em 2019.

A Tabela 4.3 apresenta: a média (μ), o desvio-padrão (σ),mediana, mínimo, máximo, proporção de zeros ($\%Zeros$) e o coeficiente de variação (CV) das notas das provas objetivas e de redação dos alunos de escolas privadas. Adotamos as seguintes abreviações para as variáveis avaliadas no conjunto de dados: nota de Redação (RD_PR), nota de Ciências da Natureza (CN_PR), nota de Linguagens e Códigos (LC_PR), nota de Matemática (MT_PR) e nota de Ciências Humanas (CH_PR).

Os valores expressos na Tabela 4.3 mostram a grande diferença existente entre os sistemas público e privado de ensino no Ceará, basta observar a média da nota de redação.Vale destacar também a baixa variabilidade das notas de redação. Para melhor ilustrar esse contexto, a Figura 4.6 apresenta um histograma das notas de redação dos candidatos de escolas privadas, ao qual constatamos uma grande assimetria, observa-se que apenas um candidato tirou 0 e outro tirou 1.

Tabela 4.3: Média (μ), desvio-padrão (σ),mediana, mínimo, máximo,proporção de zeros ($\%Zeros$), coeficiente de variação (CV) das notas dos candidatos de escolas privadas no Ceará em 2019.

	RD_PR	**CN_PR**	**LC_PR**	**CH_PR**	**MT_PR**
μ	0,78	0,55	0,57	0,58	0,65
σ	0,14	0,08	0,05	0,07	0,12
mediana	0,82	0,56	0,57	0,58	0,65
min	0,00	0,33	0,32	0,00	0,36
max	1,00	0,80	0,73	0,79	0,98
$\%Zeros$	0,24	0,00	0,00	0,03	0,00
CV	0,18	0,26	0,09	0,13	0,19

Na Figura 4.7 observamos que quando comparados os dois sistemas de ensino, no sistema privado as notas de redação estão concentradas num intervalo mais amplo, com valor mínimo de 0 e máximo de 1.

Na Figura 4.8 apresentamos os boxplots das notas das cinco provas distribuídas por sexo. Observamos um melhor desempenho dos candidatos do sexo feminino na redação, assim como na escola pública. Para verificar a relação entre as notas de redação e o sexo do candidato aplicamos o teste de Mann-Whitney. Utilizamos as hipóteses da seção anterior, ao nível de 5% de significância, rejeitamos a hipótese nula, ou seja, podemos afirmar que há relação entre a nota de redação e o sexo do candidato.

A Figura 4.9 apresenta o gráfico da matriz de correlações entre as variáveis contínuas. A variável que apresenta maior correlação com a nota de redação (RD_PR) é a nota de ciências humanas (CH_PR)

Figura 4.6: Histograma das notas da redação dos candidatos de escolas privadas no Ceará em 2019.

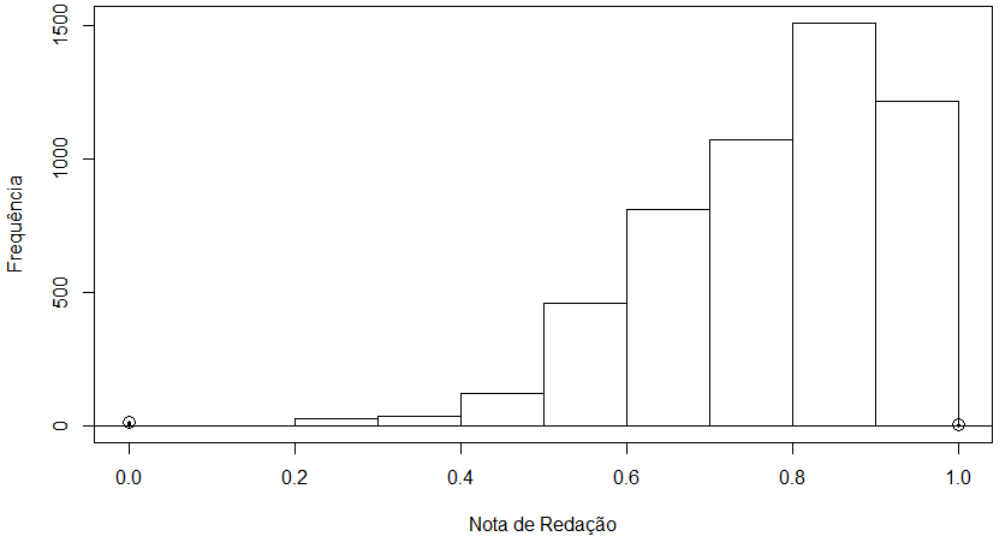


Figura 4.7: Boxplot das notas dos candidatos de escolas privadas no Ceará em 2019.

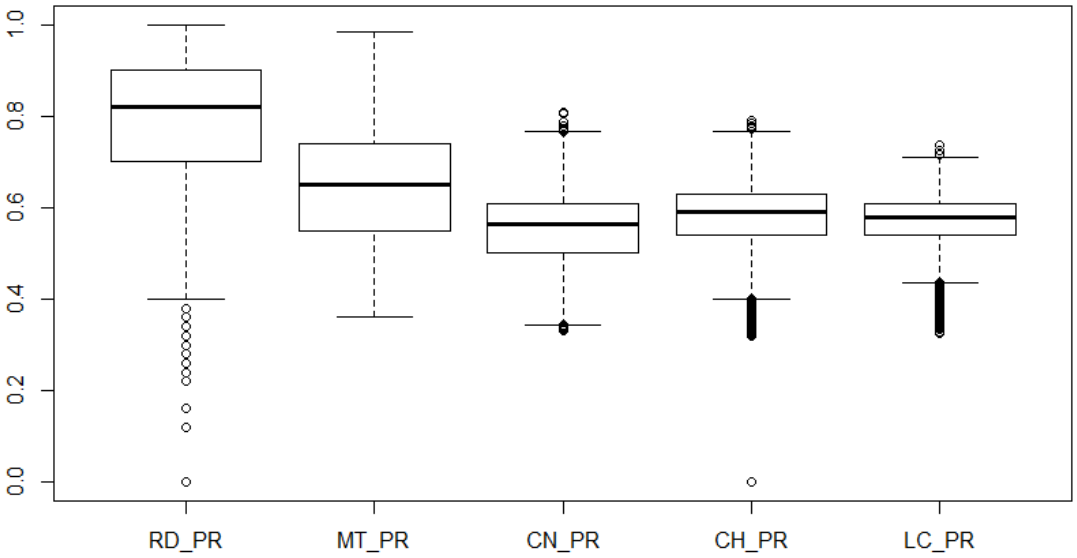


Figura 4.8: Boxplot das notas dos candidatos de escolas privadas por sexo no Ceará em 2019.

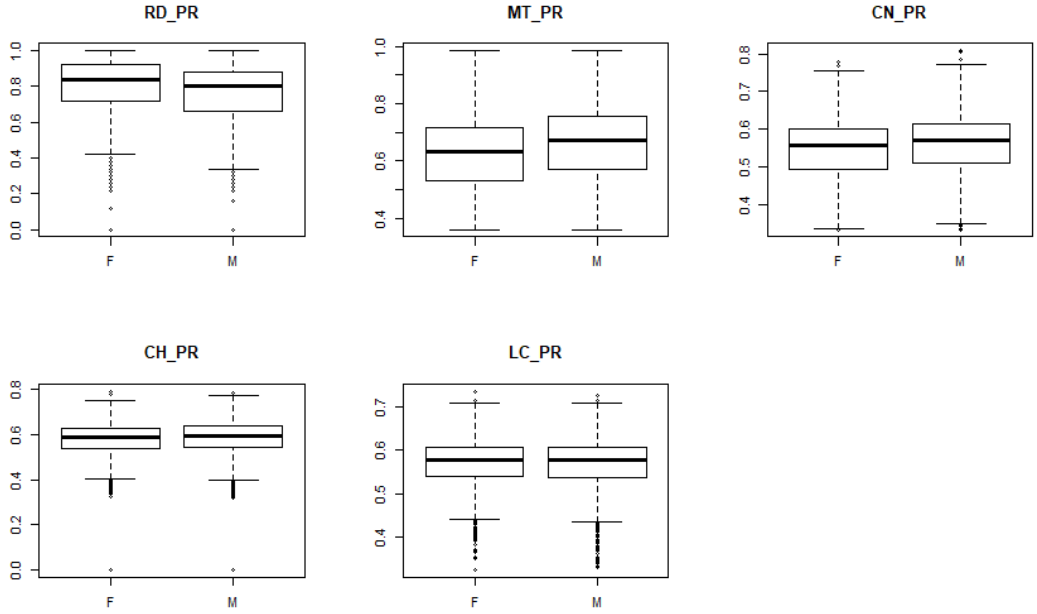
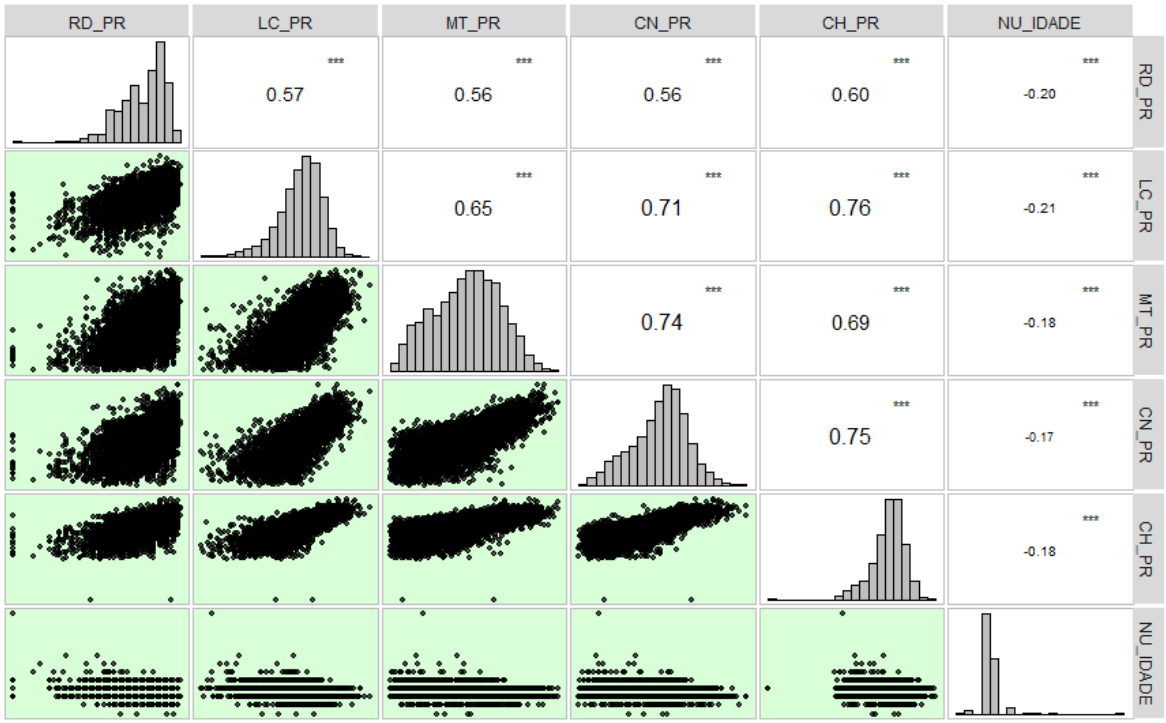


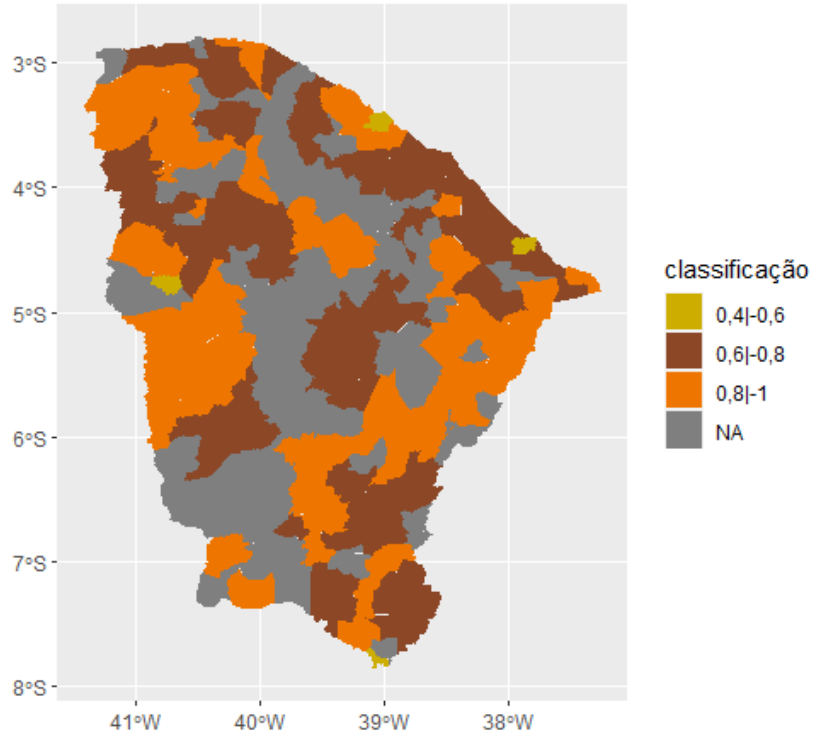
Figura 4.9: Matriz de correlação da Nota de Redação dos candidatos de escola privada com outras variáveis no Ceará em 2019.



A Figura 4.10 apresenta o mapa das médias das notas de redação dos alunos de escolas privadas por município no Ceará. Dos 184 municípios do estado, apenas 110 possuem instituições de ensino privadas de nível médio. Ao analisar o mapa com

as médias das notas de redação, é possível destacar o excelente desempenho dos alunos nas regiões metropolitanas de Fortaleza e noroeste cearense. Essas regiões se destacam por abrigar grandes centros de educação privada, muitos dos quais são referências em todo o país. Podemos observar que no mapa uma predominância de médias nos intervalos 0,6|-0,8 e 0,8|-1, refletindo uma grande diferença entre os sistemas de ensino público e privado. Essa disparidade sugere a existência de desigualdades educacionais que podem estar associadas a fatores socioeconômicos e infraestruturais.

Figura 4.10: Mapa das médias de redação de alunos de escolas privadas por município no Ceará.



4.3 Ajuste do modelo de regressão beta inflacionado para as notas de redação dos candidatos de escolas públicas no estado do Ceará ENEM 2019.

Nesta seção apresentamos uma modelagem de regressão para a nota de redação do ENEM 2019 para alunos de escolas públicas. Utilizamos o modelo de regressão beta inflacionado em zero (BIZI) proposto por OSPINA e FERRARI (2012), visto sua adequabilidade a dados dessa natureza. O procedimento computacional foi desenvolvido utilizando os pacotes *gamlss*(RIGBY e STASINOPOULOS, 2005) e

betareg (CRIBARI-NETO e ZEILEIS, 2010) do *software* estatístico R 2022 (TEAM).

Para selecionarmos as covariáveis utilizadas para explicar a nota de redação dos alunos de escolas públicas do Ceará no ENEM 2019 utilizamos os critérios de seleção de modelo AIC (*Akaike's information criterion)* proposto por Akaike (1974) e o BIC (*Critério de informação Bayesiano*), proposto por Schwarz (1978). Inicialmente, para ajustar um modelo de regressão beta inflacionado é essencial avaliar se a precisão é fixa ou variável. Para realizar essa avaliação, utilizamos o teste da razão de verossimilhança. Ao nível de significância de 5%, (p-valor<0,001) rejeitamos a hipótese nula de que a precisão é constante. Portanto, podemos concluir que a precisão varia no modelo em questão. Isso significa que, além de modelar a média condicional e a probabilidade das notas serem zero, é necessário incluir um componente que modele a precisão. Dessa forma, o modelo pode capturar adequadamente a relação entre as notas de redação e as covariáveis significativas, considerando a variação na precisão em diferentes pontos do conjunto de dados. É importante destacar que a inclusão de uma estrutura de regressão para o parâmetro da precisão pode levar a resultados mais precisos e confiáveis na análise dos dados.

A Tabela 4.4 apresenta os coeficientes estimados, os valores da estatística de teste t, os erros-padrão e os p-valores dos testes de hipótese. É relevante salientar que dentre as opções de função de ligação adotada nos parâmetros da média condicional e da probabilidade em zero a que melhor se adequou ao modelo selecionado foi a função logit, enquanto que no parâmetro da precisão, foi a função log.

Com o modelo selecionado, procedeu-se à realização de um teste de especificação RESET (RAMSEY (1969);CRIBARI-NETO e PEREIRA (2013)), proposto no contexto do modelo de regressão beta inflacionado, no qual foi considerada a segunda potência do preditor linear como variável teste. Esse teste tem como hipótese nula que o modelo selecionado está bem especificado, em contraposição à hipótese alternativa é de má especificação do modelo selecionado. Constatou-se, com um nível de significância de 5%, que não rejeitamos a hipótese nula (p-valor = 0,190), o que indica que o modelo selecionado está bem especificado. Para avaliar a adequação do modelo, é comum utilizar um pseudo-R^2. MCFADDEN (1974) propôs uma medida baseada no logaritmo da função de verossimilhanças. O pseudo-R^2, PR^2, é calculado da seguinte forma:

$$PR^2 = 1 - \frac{\hat{l}_N}{\hat{l}_F},$$

em que $\hat{l}_F$ é a log-verossimilhança maximizada do modelo ajustado e $\hat{l}_N$ é a log-verossimilhança maximizada do modelo nulo, que é o modelo sem a estrutura de regressão. O pseudo-R^2 obtido para o modelo selecionado é igual a 0,8551.

No que tange ao modelo utilizado para explicar a nota de redação dos estudantes de escolas públicas do Ceará no ENEM 2019, verificamos, por meio da análise

dos coeficientes estimados, que as variáveis *LC*, *MT*, *CN*, *CH* e *Renda* exerceram influência positiva na variável resposta. Em outras palavras, estudantes com maiores notas nas provas objetivas do ENEM e estudantes cuja renda familiar é de no mínimo R$ 998,00 têm notas de redação maiores do que aqueles cuja renda familiar é menor que R$ 998,00. É importante destacar que a influência positiva da variável renda sugere que alunos de escolas públicas que possuem condições financeiras melhores têm, em geral, notas maiores de redação.

Ao considerar a estrutura de regressão para a precisão, observamos que a covariável referente à nota de Ciências Humanas (CH) apresentou efeito positivo. Isso indica que quanto maior a pontuação do aluno nessa área do conhecimento, maior a precisão, ou seja, menor a dispersão das notas de redação. Por outro lado, a covariável referente à nota de Matemática (MT) apresentou efeito negativo na precisão. Ou seja, quanto maior a pontuação do aluno em matemática, mais dispersas são as notas de redação.

Ao analisar a estrutura de regressão para o α que é a probabilidade do aluno tirar nota zero na redação, constatou-se que todas as variáveis incluídas no modelo apresentaram efeito inversamente proporcional a essa probabilidade. Em outras palavras, a medida que as notas nas provas objetivas do ENEM aumentam, a chance de o aluno tirar nota zero na redação diminui.

Por intermédio da análise dos gráficos apresentados na Figura 4.11, concluímos que o modelo de regressão beta inflacionado selecionado para explicar a nota de redação para alunos de escolas públicas do Ceará no ENEM 2019 parece estar bem ajustado, visto que os resíduos permanecem dentro do intervalo (-4,4), e encontra-se, em geral, dentro das bandas de confiança dos envelopes simulados.

Tabela 4.4: Estimativa dos parâmetros do modelo de regressão beta inflacionado para escolas públicas ENEM 2019.

Modelo para μ				
Variáveis	**Estimativa**	**Valor t**	**Erro padrão**	**p-valor**
(intercepto)	-3,649	-187,490	0,019	<0,001
LC	3,322	70,304	0,047	<0,001
MT	1,427	48,196	0,029	<0,001
CN	1,350	31,558	0,042	<0,001
CH	1,997	48,005	0,041	<0,001
RENDAClasse 2	0,023	3,099	0,007	<0,001
RENDAClasse 3	0,080	7,175	0,011	<0,001
RENDAClasse 4	0,087	3,398	0,025	0,001
Modelo para ϕ				
Variáveis	**Estimativa**	**Valor t**	**Erro padrão**	**p-valor**
(intercepto)	3,026	87,610	0,345	<0,001
CH	0,150	1,981	0,076	0,047
MT	-0,963	-15,023	0,064	<0,001
Modelo para α				
Variáveis	**Estimativa**	**Valor t**	**Erro padrão**	**p-valor**
(intercepto)	9,934	48,210	0,143	<0,001
LC	-10,974	-39,640	0,276	<0,001
MT	-5,731	-21,080	0,271	<0,001
CH	-3,758	-15,460	0,243	<0,001

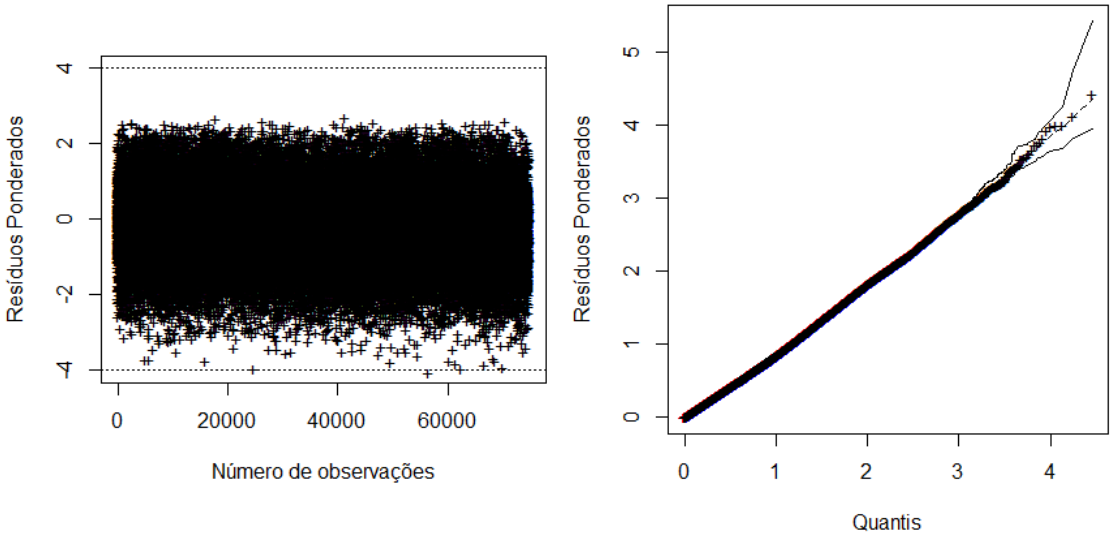


Figura 4.11: Gráfico dos resíduos quantis aleatorizados do modelo de regressão beta inflacionado para as notas de redação. Escolas Públicas - ENEM 2019.

4.4 Ajuste do modelo de regressão beta inflacionado para as notas de redação dos candidatos de escolas privadas no estado do Ceará- ENEM 2019

Diferentemente dos dados para escolas públicas, nos dados para escolas privadas a nota de redação apresenta inflacionamento em zero e um, desta forma, utilizamos o modelo de regressão beta inflacionado em zero e um (BIZU) proposto por OSPINA e FERRARI (2012). Para a seleção das covariáveis, utilizamos os critérios de seleção AIC e BIC. A Tabela 4.5 apresenta os coeficientes estimados, os valores da estatística de teste t, erro padrão e p-valores dos testes de hipóteses.

Para verificarmos se a precisão do modelo selecionado é fixa ou variável aplicamos o teste da razão de verossimilhança e constatamos, ao nível de significância de 5% (p-valor<0,001) que devemos rejeitar a hipótese nula de que a precisão é constante, logo, podemos concluir que a precisão é variável.

A Tabela 4.5 apresenta os coeficientes estimados, os valores da estatística t, os erros-padrão e os p-valores dos testes de hipótese. É relevante salientar que dentre as opções de função de ligação a que melhor se adequou nos parâmetros da média condicional e da probabilidade em zero foi a função logit, enquanto que para os parâmetros da precisão e probabilidade em um, empregou-se a função log.

Para verificarmos a qualidade do modelo proposto explicar as notas de redação de alunos de escolas privadas no ENEM 2019, utilizamos o coeficiente de determinação ajustado (pseudo-R^2) e o teste de RESET (RAMSEY (1969);CRIBARI-NETO e PEREIRA (2013)). O resultado do p-valor para o teste RESET foi 0,300, ao nível de significância de 5% não rejeitamos a hipótese nula de que o modelo encontra-se bem especificado. O pseudo-R^2 (MCFADDEN, 1974) do modelo final foi igual a 0,315. O baixo valor para o pseudo-R^2 possa ser devido a uma maior heterogeneidade existente nos dados.

Por meio da análise dos coeficientes do modelo proposto, é possível verificar que as variáveis *LC, MT, CN* e *CH* apresentaram efeito positivo na variável resposta, ou seja, os participantes que apresentaram maiores notas para estas competências têm uma maior nota na redação. Por outro lado,as variáveis SEXOM e IDADE apresentam efeito negativo. Candidatos do sexo masculino e mais velhos tendem a apresentar notas menores na redação. Considerando a estrutura de regressão para o parâmetro de precisão, temos que as covariáveis *CH* e *INTERNET* influenciam de forma negativa na precisão. Indivíduos com notas maiores em ciências humanas e que não têm internet em casa apresentam maior dispersão nas notas de redação. Por outro lado, candidatos do sexo masculino apresentam notas mais precisas, ou seja, menos dispersas.

Tabela 4.5: Estimativa dos parâmetros do modelo de regressão beta inflacionado para escolas privadas ENEM 2019.

Modelo para μ				
Variáveis	**Estimativa**	**Valor t**	**Erro padrão**	**p-valor**
(intercepto)	-1,973	-7,230	0,273	<0,001
LC	1,355	5,500	0,246	<0,001
MT	1,639	15,760	0,104	<0,001
CN	1,480	8,170	0,181	<0,001
CH	2,810	13,800	0,203	<0,001
IDADE	-0,042	-3,030	0,013	<0,001
SEXOM	-0,327	-17,940	0,018	<0,001
Modelo para ϕ				
Variáveis	**Estimativa**	**Valor t**	**Erro padrão**	**p-valor**
(intercepto)	-0,128	-1,170	0,109	0,239
CH	-1,417	-8,310	0,170	<0,001
SEXOM	0,089	3,680	0,024	<0,001
INTERNETB	-0,192	-3,440	0,055	<0,001
Modelo para δ_0				
Variáveis	**Estimativa**	**Valor t**	**Erro padrão**	**p-valor**
(intercepto)	1,699	1,050	1,611	0,292
MT	-14,237	-4,170	3,411	<0,001
Modelo para δ_1				
Variáveis	**Estimativa**	**Valor t**	**Erro padrão**	**p-valor**
(intercepto)	-31,512	8,420	-3,740	<0,001
CH	37,075	12,010	3,087	0,002

No tocante aos parâmetro δ_0 e δ_1,que são, respectivamente, a proporção de zeros e de uns, o modelo selecionou apenas a variável nota de matemática, que apresenta efeito inversamente proporcional, para estrutura do δ_0 e a nota de ciências humanas que apresenta efeito diretamente proporcional, para estrutura do δ_1, ou seja, alunos com maiores notas de matemática têm uma probabilidade menor de tirar zero em redação, ao passo que alunos com maiores notas de ciências tendem a ter uma probabilidade maior de tirar 1 em redação.

Na Figura 4.12 apresentamos os gráficos dos resíduos em comparação aos índices das observações e o gráfico de probabilidade normal com envelopes simulados. No primeiro gráfico é possível visualizar que os resíduos estão distribuídos de forma aleatória entre os limites (-4,4); não tivemos observações que ultrapassaram esse intervalo. Porém, há resíduos que se encontram fora das bandas de confiança dos envelopes simulados, mas não há fortes indícios de afastamento da suposição de que o modelo de regressão beta inflacionado selecionado é adequado para os dados.

As análises revelaram que, tanto em escolas públicas quanto privadas, as provas objetivas influenciam as notas de redação dos alunos. Em escolas públicas, o desempenho nas provas objetivas e a renda familiar foram determinantes. Já nas escolas

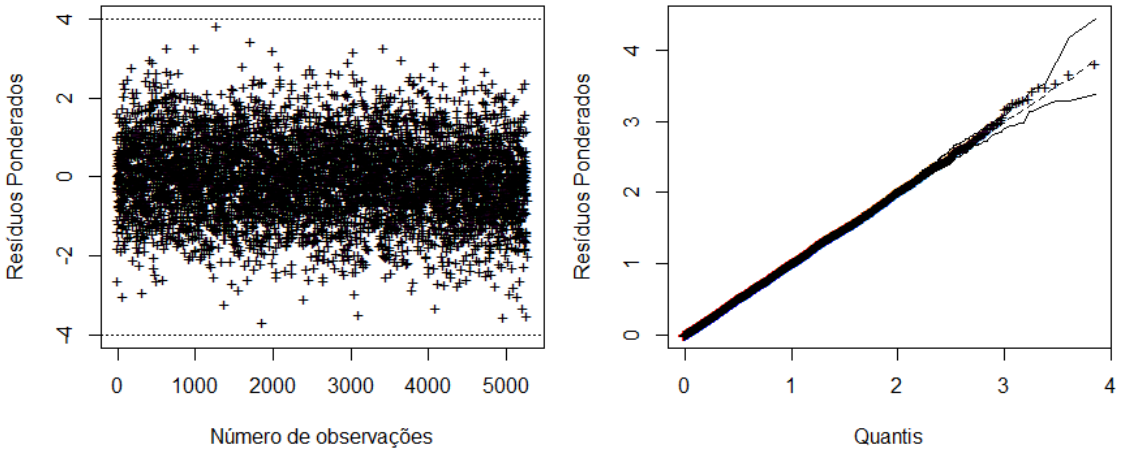


Figura 4.12: Gráfico dos resíduos quantis aleatorizados do modelo de regressão beta inflacionado para as notas de redação. Escolas privadas - ENEM 2019.

privadas, além das notas, o sexo e a idade também desempenharam papéis importantes. Notavelmente, alunos de escolas públicas com melhores notas em matemática, linguagem e ciências humanas têm menores probabilidades de tirar zero em redação, enquanto que para escolas privadas, apenas a nota de matemática influencia negativamente na probabilidade do aluno tirar zero em redação. Adicionalmente, alunos com maiores notas em ciências humanas, têm maiores chances de tirar notas maiores em redação.

Capítulo 5

Conclusões

No presente trabalho, avaliamos os fatores que influenciam na nota de redação do ENEM dos alunos que realizaram a prova no estado do Ceará no ano de 2019, por meio de modelos de regressão beta inflacionados.

Com base na análise descritiva é possível destacar a heterogeneidade dos dados. Observamos uma grande diferença entre as observações para candidatos de escolas públicas e os de escolas privadas. Entretanto, destacamos nos dois grupos observados, variáveis com correlações bem semelhantes, fomentando a aplicação do modelo regressão para que tivéssemos uma avaliação mais precisa dos impactos dessas variáveis na variável resposta (nota de redação).

A análise do modelo revelou que tanto nas escolas públicas quanto nas escolas privadas, as notas das provas de Ciências Humanas e suas Tecnologias, Ciências da Natureza e suas Tecnologias, Linguagens e Códigos e suas Tecnologias e Matemática e suas Tecnologias exercem influência positiva na nota de redação. Vale destacar na análise de ambos os modelos as notas nas provas objetivas tem fortes influências na nota do candidato, destacamos a nota de Ciências humanas e suas tecnologias como fator preponderante na produção de textos dissertativos-argumentativos, modelo este base na redação do ENEM atual. As ciências humanas é parte crucial no currículo escolar, desenvolvendo no aluno o senso crítico e o pensamento reflexivo, contribuindo para a formação de cidadãos conscientes e engajados na transformação social. Além disso, é relevante ressaltar a influência da renda dos candidatos das escolas públicas, como uma variável significativa no modelo. No caso das escolas privadas, as variáveis sexo e idade também se mostraram como fatores preponderantes na nota do candidato. Esses resultados destacam a complexidade das disparidades socioeconômicas e demográficas na pontuação da redação, e a necessidade de ações que visem incentivar e melhorar o desempenho dos alunos nos conteúdos de matemática, ciências humanas, da natureza e liguagens e códigos.

Adicionalmente a modelagem da média, o modelo de regressão beta inflacionado modela as probabildade da variável assumir o valor zero e o valor 1. Nesse contexto,

verificamos que quanto maior a notas de linguagens e códigos, matemática e ciências humanas, maior a nota de redação para alunos de escola pública. Contudo, para alunos de escola privada, apenas a nota de matemática influencia negativamente, ou seja, quanto maior a nota de matemática, menor a probabilidade do aluno tirar zero em redação. É importante ressaltar que a nota máxima de redação foi obtida apenas nas escolas privadas onde concluímos que quanto maior a nota de ciências humanas, maior a probabilidade do aluno tirar um em redação.

Os resultados encontrados neste trabalho fornecem uma visão geral do desempenho dos alunos do estado do CE na prova de redação, bem como instrumentos que podem ser utilizados na promoção de políticas e tomadas de decisão do ensino no estado do CE, reforçando a importância de um olhar multidisciplinar na preparação para o ENEM.

Referências Bibliográficas

ALTMAN, N., KRZYWINSKI, M., 2015, "Points of Significance: Association, correlation and causation." *Nature methods*, v. 12, n. 10.

CORE TEAM, R., 2022. "R Core Team R. R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria; 2022". Disponível em: <https://www.r-project.org/>.

CRIBARI-NETO, F., PEREIRA, T. L., 2013, "Avaliação da eficiência de administrações municipais no estado de Sao Paulo: uma nova abordagem via modelos de regressão beta", *Revista Brasileira de Biometria*, v. 31, n. 2, pp. 270–294.

CRIBARI-NETO, F., ZEILEIS, A., 2010, "Beta regression in R", *Journal of statistical software*, v. 34, pp. 1–24.

DA SILVA, C. R., DE SOUZA, T. C., LIMA, C. M. B. L., 2018, "Avaliação da eficiência na atenção básica à saúde no Estado da Paraíba: Uma análise via modelo de regressão beta inflacionado", *Ciência e Natura*, v. 40, n. 21, pp. 1–11.

DE ARAÚJO SOARES, D., DE AZEVEDO SOUZA, S., DA SILVA, D. J., et al., 2019, "Avaliação epidemiológica da esquistossomose no estado de Pernambuco através de um modelo de regressão beta", *Archives of Health Sciences*, v. 26, n. 2, pp. 116–120.

DE LIMA LEITE, M., DAS VIRGENS FILHO, J. S., 2007, "AVALIAÇÃO DA DISTRIBUIÇÃO BETA COMO MODELO PROBABILÍSTICO PARA ANÁLISE DE DADOS DE VELOCIDADE DO VENTO PARA PONTA GROSSA-PR", *Publicatio UEPG: Ciências Exatas e da Terra, Agrárias e Engenharias*, v. 13, n. 01.

DINIZ, J. S. M., MELO, D. L. M., 2019, "Modelo de regressão beta inflacionado em zero e um: uma aplicação à proporção de mulheres nas empresas", .

DO AMARAL SOARES, E., WERLE, F. O. C., 2016, "Sistema de avaliação da Educação Básica do Ceará: a importância do foco na aprendizagem", *Revista Exitus*, v. 6, n. 2, pp. 159–179.

FERRARI, S., CRIBARI-NETO, F., 2004, "Beta regression for modelling rates and proportions", *Journal of applied statistics*, v. 31, n. 7, pp. 799–815.

INEP, 2022. "Apresentação do Exame Nacional do Ensino Médio (Enem)." Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem>.

LOUREIRO, GROPELLO, D., ARAIS, 2020. "Não há mágica: a fórmula para o sucesso do Ceará e de Sobral para reduzir a pobreza de aprendizagem". Disponível em: <https://blogs.worldbank.org/pt/education/nao-ha-magica-formula-para-o-sucesso-do-ceara-e-de-sobral-para-reduzir-pob

MCFADDEN, D., 1974, "The measurement of urban travel demand", *Journal of public economics*, v. 3, n. 4, pp. 303–328.

MOOD, A., GRAYBILL, F., BOES, D., 1974, "Tests of hypotheses", *Introduction to the theory of statistics. Tokio: McGraw-Hill*, pp. 401–470.

NELDER, J. A., WEDDERBURN, R. W., 1972, "Generalized linear models", *Journal of the Royal Statistical Society: Series A (General)*, v. 135, n. 3, pp. 370–384.

NÓVOA, A., 2002, *Para uma Educação de Qualidade*. Porto Editora.

OSPINA, R., FERRARI, S. L., 2010, "Inflated beta distributions", *Statistical papers*, v. 51, n. 1, pp. 111–126.

OSPINA, R., FERRARI, S. L., 2012, "A general class of zero-or-one inflated beta regression models", *Computational Statistics & Data Analysis*, v. 56, n. 6, pp. 1609–1623.

PEREIRA, T. L., SOUZA, T. C., CRIBARI-NETO, F., 2014, "Uma avaliação da eficiência do gasto público nas regiões do Brasil", *Ciência e Natura*, v. 36, pp. 23–36.

PORTO, K. G. S., 2019, "Descoberta de conhecimento através da análise e mineração em dados do Enem", .

RAMSEY, J. B., 1969, "Tests for specification errors in classical linear least-squares regression analysis", *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 31, n. 2, pp. 350–371.

RÊGO, F. E. D. D., 2021, *Fatores que influenciam na nota da redação do ENEM no Rio Grande do Norte*. B.S. thesis, Universidade Federal do Rio Grande do Norte.

RIGBY, R. A., STASINOPOULOS, D. M., 2005, "Generalized additive models for location, scale and shape", *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, v. 54, n. 3, pp. 507–554.

RIGBY, R. A., STASINOPOULOS, M. D., HELLER, G. Z., et al., 2019, *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. CRC press.

SANTOS, M. C. D., 2018, "Desenvolvimento de modelos para estimar principais características socioeconômicas para as notas de redação e matemática para ENEM 2018", *Revista de Estudos e Pesquisas sobre Ensino Superior*, v. 4, n. 3, pp. 88–104.

SOUSA, S. Z., OLIVEIRA, R. P. D., 2010, "Sistemas estaduais de avaliação: uso dos resultados, implicações e tendências", *Cadernos de Pesquisa*, v. 40, n. 141, pp. 793–822.

TEAM, R. C. "R: A language and environment for statistical computing. Vienna, Austria, 2022, year = 2022, url = https://www.r-project.org/". .

VIGGIANO, E., MATTOS, C., 2013, "O desempenho de estudantes no Enem 2010 em diferentes regiões brasileiras", *Revista Brasileira de Estudos Pedagógicos*, v. 94, pp. 417–438.