Casi 17ª Edição
Congresso de Administração, Sociedade e Inovação

Artigo completo
De 03 a 05 de dezembro de 2025

Determinantes da nota de redação no Enem: uma análise multinomial dos fatores individuais e contextuais

Matheus Hideki Shishido[1]; Felipe Nathan Ferreira dos Santos[2]

[1] Discente do MBA em Data Science e Analytics – USP/ESALQ. E-mail: matheus.shishido@outlook.com
[2] Professor Orientador do MBA em Data Science e Analytics – USP/ESALQ. Doutorando e Mestre em Economia Aplicada pela Universidade Federal de Viçosa (UFV). Analista Pesquisador em Economia do DEE/SPGG – RS. E-mail: felipe.nathan@ufv.br

Resumo: O Exame Nacional do Ensino Médio é um dos principais mecanismos para o ingresso ao ensino superior no Brasil. Além das provas objetivas, a prova de redação é um componente de peso na nota final dos candidatos. Dada a escassez de estudos dedicados a esta parte da prova, este trabalho busca analisar os determinantes individuais, escolares e regionais associados à situação da nota da redação. A partir de uma abordagem descritiva e documental com base nos microdados do Enem 2023, foram aplicados uma regressão logística multinomial, para investigar os fatores que levam a anulação da redação; e uma regressão linear múltipla para analisar se os mesmos preditores impactam a pontuação dos candidatos entre as redações válidas. Os resultados da regressão multinomial indicam que um menor desempenho das provas objetivas, ser aluno de escolas públicas, residir nas regiões norte ou nordeste e possuir um baixo nível socioeconômico aumentam a probabilidade de o participante ter a redação anulada. A regressão linear apresenta resultados similares, com alunos de escolas estaduais e municipais tendo pontuações, em média, inferiores àqueles da rede privada ou federal. Fatores como nota nas provas objetivas e escolaridade dos pais também se mostraram estatisticamente significativos.

Palavras-Chave: Enem, Redação, Desigualdade Educacional, Regressão Logística Multinomial, Determinantes da Nota.

1. Introdução

O Exame Nacional do Ensino Médio (Enem) foi criado pelo Governo Federal em 1998 com o objetivo de avaliar o desempenho escolar dos estudantes ao término da educação básica (Brasil, 2025). Desde então, o desempenho dos participantes tem sido amplamente investigado, com ênfase em características individuais, fatores socioeconômicos e desigualdades regionais, a fim de compreender os diferentes contextos educacionais (Gomes e Borges ,2009; Lucena e Santos, 2020). No entanto, a maioria das análises se concentra na nota média dos participantes no Exame, deixando em segundo plano as análises para cada uma das provas objetivas[1] assim como da prova dissertativa (Redação), que possui componentes específicos.

Gomes e Borges (2009) investigaram os resultados das provas objetivas do Enem de 2001 em conjunto com testes de inteligência aplicados aos mesmos estudantes, utilizando regressão múltipla para selecionar variáveis. Segundo eles, as habilidades cognitivas como resolução de problemas, compreensão verbal e rapidez cognitiva estão fortemente relacionadas ao desempenho global no exame. Por outro lado, Lucena e Santos (2020) analisaram os dados de 2016 para verificar o impacto do perfil socioeconômico dos estudantes nas notas do Enem, com ênfase em variáveis como escolaridade dos pais e tipo de escola. Foi constatado que estudantes com maior renda familiar, oriundos de escolas privadas e que não trabalham tendem a apresentar desempenho acima da média.

Além dessas abordagens, há estudos que consideraram agrupamentos como escolas e municípios para investigar as diferenças de desempenho. Jaloto e Primi (2021) destacaram que fatores como atraso escolar, raça/cor e nível socioeconômico influenciam as notas, e que, em média, escolas privadas apresentam melhor desempenho do que escolas estaduais. De modo semelhante, Carvalho Junior, Mendes e Ferreira (2023) observaram uma tendência de melhores resultados entre alunos cujos docentes tinham maior titulação. Travitzki, Ferrão e Couto (2016) evidenciaram que a maior parte da variação no desempenho dos alunos ocorre dentro dos próprios municípios, sugerindo que existem escolas com desempenhos muito distintos mesmo em contextos geográficos semelhantes.

Apesar do interesse crescente, ainda são poucos os estudos que analisam especificamente a nota da redação. Essa lacuna é relevante, considerando o impacto direto desse componente nos processos seletivos de acesso ao ensino superior. Primeiramente, candidatos que obtém nota zero na redação são automaticamente excluídos de programas como como o Sistema de Seleção Unificada (SISU) (BRASIL, 2025c), Programa Universidade Para Todos (Prouni) (BRASIL, 2025d) e Fundo de Financiamento Estudantil (FIES) (Brasil, 2025d). Além disso, por não ser corrigida pela Teoria de Resposta ao Item (TRI) e por ser é a única com pontuação estabelecida entre 0 e 1000 pontos, a prova de redação tem um peso substancial na composição da média final dos candidatos, podendo elevar significativamente a pontuação geral e, consequentemente, as chances de aprovação. Essas particularidades reforçam a necessidade de uma investigação mais aprofundada sobre os determinantes da nota de redação no Enem.

[1] Ciências da Natureza e suas Tecnologias (CN), Ciências Humanas e suas Tecnologias (CH), Linguagens, Códigos e suas Tecnologias (LC), Matemática e suas Tecnologias (MT).

Viggiano e Mattos (2013) destacam que, em 2010, cerca de 4% dos participantes obtiveram nota zero na redação, por motivos como fuga ao tema, não entrega do texto ou descumprimento dos critérios mínimos de avaliação. Essa distribuição assimétrica da nota – com alta concentração em pontuações muito baixas – indica que a redação pode responder a determinantes distintos em comparação com as demais provas. Além disso, os dados mostram uma concentração da nota máxima em contextos específicos: em 2023, apenas 6% das redações com nota máxima foram produzidas por estudantes de escolas públicas (Lima, 2024). Em 2024, dos 12 participantes que alcançaram a pontuação máxima, apenas 1 era oriundo da rede pública, representando 0,00008% do total (Gomes, 2025). Esses dados sugerem a existência de um padrão de desigualdade educacional que merece investigação mais aprofundada.

Diante disso, o presente estudo busca compreender os fatores associados às diferentes categorias de nota na redação do Enem, considerando variáveis individuais, escolares e regionais. Ao analisar esses determinantes, pretende-se oferecer subsídios para a formulação de políticas públicas voltadas à redução das desigualdades no desempenho educacional, especialmente em um componente tão decisivo como a redação.

2. Fundamentação teórica

A análise do desempenho na prova de redação do Enem pode ser compreendida à luz de diferentes perspectivas teóricas. A teoria do capital humano, proposta por Becker (1964), entende a educação como investimento em habilidades cognitivas e produtivas, que se traduzem em melhores oportunidades no mercado de trabalho e em maior rendimento econômico. Nesse sentido, a redação constitui uma dimensão relevante da mensuração das competências adquiridas ao longo da trajetória escolar, refletindo o acúmulo de capital humano associado à qualidade do ensino, ao tempo de estudo e ao apoio familiar.

No entanto, compreender os determinantes da nota da redação apenas sob a ótica econômica seria insuficiente. Bourdieu e Passeron (1970) destacam que a escola opera como espaço de reprodução social, no qual o desempenho dos estudantes é fortemente influenciado pelo capital cultural herdado da família. Isso significa que habilidades linguísticas e discursivas, avaliadas na redação, tendem a reproduzir desigualdades pré-existentes, beneficiando alunos oriundos de famílias com maior escolaridade e repertório cultural.

Além disso, o Relatório Coleman (1966) introduziu a noção de efeito-escola, segundo a qual o desempenho educacional não depende apenas de características individuais, mas também das condições institucionais, tais como infraestrutura, qualificação docente e tipo de gestão. Essa perspectiva contribui para compreender as diferenças observadas entre estudantes de escolas públicas e privadas, bem como entre redes estadual, municipal e federal.

Por fim, a abordagem das capacitações de Sen (2018) amplia o debate ao conceber a educação não apenas como meio de acumulação de habilidades, mas como instrumento de expansão das liberdades individuais. Assim, a redação do Enem pode ser vista não somente como um critério seletivo, mas também como espaço de exercício da capacidade de argumentação e expressão crítica, fundamentais para a participação social e cidadã.

Dessa forma, as teorias apresentadas permitem articular a dimensão econômica, social e institucional da educação, fornecendo uma base sólida para interpretar os determinantes individuais e contextuais da nota de redação no Enem.

3. Método de pesquisa

A presente pesquisa possui uma natureza descritiva, pois seu objetivo está na análise dos padrões das diferentes classificações da nota da redação do ENEM e de suas associações com variáveis explicativas. Essas variáveis abrangem tanto características individuais dos participantes quanto fatores socioeconômicos e regionais. A abordagem descritiva é apropriada para este estudo, pois permite identificar e caracterizar tendências nos dados sem estabelecer relações causais diretas.

3.1. Descrição da base e fonte de dados

Para a obtenção das informações necessárias à caracterização da população-alvo, a pesquisa se baseará em uma abordagem documental, tendo como principal fonte os microdados do Enem, disponibilizados publicamente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Os microdados contêm informações detalhadas sobre os participantes da prova, incluindo seu desempenho nas provas objetivas, características socioeconômicas e regionais, além da classificação atribuída à redação. O uso dessa base de dados possibilita uma análise em larga escala, permitindo examinar padrões e tendências no desempenho dos candidatos à luz de diferentes fatores contextuais. O ano de 2023 foi escolhido pelo fato de ser o mais recente disponibilizado pelo MEC.

A seleção das variáveis utilizadas na análise foi orientada pelos objetivos do estudo, que busca compreender os fatores associados à classificação das notas de redação dos participantes do Enem. Foram consideradas variáveis relacionadas ao desempenho acadêmico nas demais áreas do Exame, características socioeconômicas e educacionais dos participantes, bem como informações sobre a escola frequentada. Essas variáveis permitem captar tanto aspectos individuais quanto contextuais que podem influenciar o desempenho na redação. A Tabela 1 e a Tabela 2 mostram o descritivo das variáveis.

Tabela 1. Descritivo dos valores possíveis da variável preditora.

Categoria	Variável	Descrição	Nome	Valor
Dados da redação	tp_status_redacao	Situação da redação do participante. Variável dependente do estudo.	1	Sem problemas.
			2	Anulada.
			3	Cópia Texto Motivador.
			4	Em branco.
			6	Fuga ao tema.
			7	Não atendimento ao tipo textual.
			8	Texto insuficiente.
			9	Parte desconectada.

Fonte: Adaptado de Brasil (2025e).

Tabela 2. Descritivo das variáveis explicativas.

Categoria	Variável	Descrição	Nome	Valor
Dados da escola	tp_ dependencia_ adm_esc	Dependência administrativa da escola.	1 2 3 4	Federal. Estadual. Municipal. Privada.
	sg_uf_esc	Sigla da Unidade da Federação (UF) da escola.	NA	Código das UFs.
Dados das provas objetivas	nu_nota_cn	Nota da prova de ciências da natureza.	NA	Não se aplica.
	nu_nota_ch	Nota da prova de ciências humanas.	NA	Não se aplica.
	nu_nota_mt	Nota da prova de Matemática.	NA	Não se aplica.
Dados Socioeconômicos	q001/q002	Escolaridade do pai ou homem responsável (*q001*) e mulher ou homem responsável (*q002*) pelo participante.	A	Nunca estudou.
			B	Não completou a 4ª série/5º ano do Ensino Fundamental.
			C	Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.
			D	Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.
			E	Completou o Ensino Médio, mas não completou a Faculdade.
			F	Completou a Faculdade, mas não completou a Pós-graduação.
			G	Completou a Pós-graduação.
			H	Não sabe.
	q006	Renda familiar mensal.	A	Nenhuma Renda.
			B	Até R$ 1.320,00.
			C	De R$ 1.320,01 até R$ 1.980,00.
			D	De R$ 1.980,01 até R$ 2.640,00.
			E	De R$ 2.640,01 até R$ 3.300,00.
			F	De R$ 3.300,01 até R$ 3.960,00.
			G	De R$ 3.960,01 até R$ 5.280,00.
			H	De R$ 5.280,01 até R$ 6.600,00.
			I	De R$ 6.600,01 até R$ 7.920,00.
			J	De R$ 7.920,01 até R$ 9240,00.

K	De R$ 9.240,01 até R$ 10.560,00.
L	De R$ 10.560,01 até R$ 11.880,00.
M	De R$ 11.880,01 até R$ 13.200,00.
N	De R$ 13.200,01 até R$ 15.840,00.
O	De R$ 15.840,01 até R$19.800,00.
P	De R$ 19.800,01 até R$ 26.400,00.
Q	Acima de R$ 26.400,00.

Fonte: Adaptado de Brasil (2025e).

A variável tp_dependencia_adm_esc foi incluída no modelo com o objetivo de captar os efeitos do tipo de gestão escolar sobre a situação da redação. Conforme demonstrado por Lucena e Santos (2020), o tipo de escola é um dos principais determinantes do desempenho no ENEM, mesmo após o controle por fatores socioeconômicos. Isso se justifica pela desigualdade na qualidade de infraestrutura, formação docente e práticas pedagógicas entre escolas públicas e privadas.

A inclusão da variável sg_uf_esc busca captar efeitos contextuais regionais que possam influenciar a situação da redação, como diferenças estruturais nos sistemas educacionais estaduais. Jaloto e Primi (2021) mostraram que parte significativa da variabilidade no desempenho no ENEM pode ser atribuída a fatores regionais, o que justifica a consideração da UF como um fator explicativo potencialmente relevante.

De forma análoga ao tratamento dado às variáveis socioeconômicas, optou-se por realizar o agrupamento das Unidades da Federação (UFs) em suas respectivas macrorregiões geográficas. A variável original de UF possui uma alta cardinalidade de 27 níveis, o que dificulta a análise estatística inferencial, principalmente pela grande disparidade na quantidade de observações entre os estados. UFs com grande população, como São Paulo, possuem um volume de dados massivo, enquanto estados como Roraima ou Amapá têm uma representação amostral consideravelmente inferior. Essa heterogeneidade resulta em estimativas de efeito com variâncias muito distintas, fazendo com que os coeficientes associados aos estados menores tendam a apresentar erros-padrão elevados e, consequentemente, dificuldade em se obter significância estatística (Fávero; Belfiore, 2024).

A agregação das 27 UFs nas 5 macrorregiões consolida os dados em categorias com um volume amostral mais robusto, conferindo um duplo benefício à análise. Isso porque a medida não só eleva a robustez estatística do estudo, pois ao aumentar o tamanho da amostra de cada categoria, permite que as estimativas para cada região se tornem mais precisas e estáveis, como também permite um melhor embasamento teórico e interpretativo. No mais, o agrupamento não é arbitrário, mas baseado em uma classificação oficial do IBGE que reflete similaridades históricas, culturais e econômicas.

Dessa forma, a decisão de analisar os dados por região é uma escolha que troca uma granularidade estatisticamente instável por um nível de agregação mais robusto, o que contribui para a validade e a clareza dos resultados inferenciais deste estudo.

As variáveis relativas à escolaridade dos pais ou responsáveis (*q001* e *q002*) e à renda familiar (*q006*) foram incluídas no modelo como indicadores do nível socioeconômico do estudante. Essa escolha é respaldada por Lucena e Santos (2020), que mostraram que maior escolaridade dos pais e renda familiar mais elevada estão positivamente associadas às notas do ENEM, mesmo após o controle por tipo de escola. Embora este estudo não modele diretamente a nota da redação, entende-se que essas variáveis socioeconômicas afetam também a situação da nota obtida, uma vez que refletem desigualdades estruturais de acesso à formação adequada em leitura, escrita e interpretação de textos.

Pelas mesmas razões da necessidade de redução de variáveis na implementação do modelo, foi criada a variável faixa_sm para a inclusão da renda familiar do candidato, que agrupou os itens de q006 em 6 categorias (sem renda, até 1, mais de 1 a 2, mais de 2 a 5, mais de 5 a 10 e acima de 10 salários-mínimos).

A inclusão das notas das provas objetivas como variáveis preditoras da situação da redação fundamenta-se na compreensão de que o ENEM, como um todo, visa avaliar um conjunto de habilidades e competências cognitivas essenciais. O estudo de Gomes e Borges (2009) sobre a validade de construto do ENEM mostrou que o desempenho na prova objetiva é fortemente explicado por habilidades cognitivas, destacando-se a resolução de problemas, a rapidez de raciocínio e a compreensão verbal.

3.2. Estratégia empírica

A modelagem estatística será conduzida por meio de um modelo de regressão logística multinomial. Esta modelagem é realizada para casos em que há mais de duas categorias na variável dependente. Diferentemente do modelo de regressão logística binária, é necessário realizar o cálculo de mais de um logito Z_{ij}.(leia-se o logito da observação i para a categoria j). Neste modelo, adota-se uma categoria como o de referência, tomando o logito desta como zero e os demais calculados pela equação (1) (Fávero; Belfiore, 2024):

$$Z_{ij} = \alpha_j + X_i^T \beta_j \quad (1)$$

Em que Z_{ij} é o logito da observação i para a categoria j, α_j é o intercepto da categoria j, X_i^T é o vetor de preditores para o indivíduo i e β_j são os coeficientes associados às variáveis independentes do modelo para a categoria j.
Os valores de Z_{ij} são utilizados para o cálculo das probabilidades de cada uma das J categorias possíveis pela equação (2):

$$p_{ij} = \frac{e^{Z_{ij}}}{1 + \sum_{k=1}^{J-1} e^{Z_{ik}}} \quad (2)$$

Em que p_{ij} representa a probabilidade da observação i pertencer à categoria j, e o termo de somatória corresponde à soma da constante de Neper elevada aos logitos das demais categorias, exceto a de referência (que é a unidade da equação, derivado de e^0, uma vez que $Z_{i0} = \beta_{i0} = 0 \rightarrow e^{Z_{i0}} = 1$). Para a obtenção dos parâmetros, aplica-se a razão de chance entre p_{ij} e p_{i0} para a linearização da equação. Com isso, chega-se na equação (3), que pode ser obtida a partir de (1) e (2):

$$\ln\left(\frac{p_{ij}}{p_{i0}}\right) = \alpha_j + X_i^T \beta_j \quad (3)$$

A estimação dos vetores β_j é realizada por meio do método de máxima verossimilhança L, cuja função de verossimilhança conjunta para n observações e J categorias é dada pela equação (4):

$$L = \Pi_{i=1}^{n} \Pi_{j=0}^{J-1} (p_{ij})^{Y_{ij}} \quad (4)$$

Em que Y_{ij} é uma variável indicadora que admite o valor 1 se a observação i pertence à categoria j, e 0 caso contrário. Ao substituir p_{ij} pela expressão equivalente obtida pela equação (2), obtém-se a equação (5):

$$L = \Pi_{i=1}^{n} \Pi_{j=0}^{J-1} \left(\frac{e^{Z_{ij}}}{1 + \sum_{k=1}^{J-1} e^{Z_{ik}}}\right)^{Y_{ij}} \quad (5)$$

A equação (5) é linearizada, criando o *log likelihood*, LL, equação (6):

$$LL = \sum_{i=1}^{n} \sum_{j=0}^{J-1} \left[Y_{ij} \ln\left(\frac{e^{Z_{ij}}}{1 + \sum_{k=1}^{J-1} e^{Z_{ik}}}\right)\right] \quad (6)$$

Esta equação é usada como a função objetivo para a estimação da matriz de parâmetros ótimos $\hat{\beta}$ por meio da obtenção do valor máximo de LL com o auxílio de métodos numéricos para a descoberta dos parâmetros ótimos. Ou seja, matematicamente, tem-se o a seguinte expressão de otimização, equação (7):

$$\hat{\beta} = \arg\max_{\beta} LL(\beta) \quad (7)$$

A análise direta dos coeficientes em modelos logit multinomiais não é intuitiva, pois eles representam os parâmetros de uma função logit que envolve exponenciação com base neperiana. Ou seja, os coeficientes estimados indicam variações na razão de chances relativas

entre categorias, e não nas probabilidades diretamente observáveis. Essa interpretação, embora válida, pode ser pouco acessível quando se deseja avaliar o impacto absoluto de uma variável sobre as diferentes categorias da variável dependente (Greene, 2008).

Por isso, uma prática comum na literatura, conforme apontado por Greene (2008), é realizar a análise dos efeitos marginais, os quais representam a variação da probabilidade p_{ij} em relação a uma variável preditora. Para variáveis contínuas, o efeito marginal é a derivada parcial de p_{ij} em relação a uma variável preditora X_r, equação (8):

$$\frac{\delta p_{ij}}{\delta X_r} = p_{ij}\left(\beta_{jr} - \sum_{k=0}^{J-1} p_{ik}\beta_{kr}\right) \quad (8)$$

Em que X_r é a variável preditora em que se deseja avaliar o efeito marginal, β_{jr} o coeficiente para a categoria j do preditor r e β_{kr} todos os coeficientes dos preditores de r para todas as categorias. Este resultado é interpretado como a variação em p_{ij} ao realizar uma variação infinitesimal na variável preditora.

Para as variáveis *dummy*, o efeito marginal não é a derivada, mas a diferença discreta nas probabilidades estimadas, dada a natureza dessas variáveis. Em outras palavras, o efeito marginal indica quanto a probabilidade de pertencimento à categoria *j* se altera ao passar da condição "pertencente" (1) para "não pertencente" (0) em uma variável binária. (Greene, 2008). A equação (9) apresenta o cálculo:

$$\Delta p_{ij} = p_{ij}(X_k = 1) - p_{ij}(X_k = 0) \quad (9)$$

Em que p_{ij} é a probabilidade prevista para uma dada categoria j quando X_k é igual a zero ou um, mantidas todas as demais variáveis de X_i constantes. Além disso, será conduzida uma análise complementar utilizando um modelo de regressão linear múltipla para investigar os determinantes das notas de redação entre os participantes cuja prova não foi anulada. Esse modelo permitirá avaliar como características individuais, como o desempenho nas provas objetivas, e fatores contextuais, como indicadores socioeconômicos e regionais, influenciam a variação da pontuação final. Dessa forma, será possível examinar se os mesmos fatores que impactam a situação da nota da redação no objetivo primário também afetam a distribuição das notas válidas, possibilitando uma compreensão mais detalhada dos mecanismos subjacentes ao desempenho dos candidatos.
A forma geral está representada pela equação (10) (Fávero; Belfiore, 2024):

$$Y = X\beta + \varepsilon \quad (10)$$

Em que Y é o vetor contendo os valores reais da variável dependente, X é a matriz com os valores das variáveis independentes, β o vetor com os coeficientes associados a cada variável explicativa e ε o vetor dos termos de erro idiossincrático.

A estimação dos parâmetros ótimos $\hat{\beta}$ é obtido por meio do método dos mínimos quadrados ordinários (MQO), que busca minimizar a soma dos quadrados das diferenças entre os valores observados da variável dependente Y e os ajustados pelo modelo. A partir disso, chega-se na equação (11) para a obtenção dos estimadores $\hat{\beta}$ (Fávero; Belfiore, 2024):

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{11}$$

Para a realização das análises, serão utilizadas técnicas de estatística descritiva e inferencial, de modo a sintetizar as informações e identificar padrões relevantes nos dados. O tratamento e a modelagem estatística serão conduzidos no software *R*®, que possibilita a aplicação de métodos estatísticos avançados com o auxílio de bibliotecas *open source* para este fim, além da manipulação eficiente de grandes bases de dados.

Dessa forma, a metodologia adotada viabiliza uma análise robusta dos fatores que influenciam a situação da nota da redação do Enem, garantindo que os resultados sejam apresentados de maneira sistemática e alinhada ao caráter descritivo da pesquisa. A validade e a confiabilidade dos resultados obtidos por meio do modelo de regressão logística multinomial dependem da observância de determinados pressupostos fundamentais. Em primeiro lugar, o modelo exige que a variável dependente seja de natureza categórica nominal, com suas categorias sendo mutuamente exclusivas e exaustivas, o que significa que cada observação deve ser classificada em uma, e apenas uma, das categorias possíveis. Este pressuposto é intrinsecamente satisfeito pela variável dependente principal, *tp_status_redacao*, cujas categorias que descrevem a situação da redação do participante no ENEM, que são nominais e mutuamente exclusivas.

Outro pressuposto é a independência das observações, o que implica que o resultado da variável dependente para um participante não exerce influência sobre o resultado de outro. Esta condição é importante para evitar vieses nos erros padrão dos coeficientes estimados. Considera-se que tal pressuposto é atendido, uma vez que os dados são de corte transversal, ou seja, cada registro na base de dados corresponde a um participante individual e distinto do ENEM, não havendo, portanto, medidas repetidas que pudessem comprometer a independência amostral.

Ademais, a ausência de multicolinearidade perfeita ou forte entre as variáveis independentes é uma exigência para a estabilidade do modelo. No desenvolvimento deste estudo, a verificação da multicolinearidade será realizada por meio do cálculo do Fator de Inflação da Variância (VIF) para todas as variáveis preditoras, incluindo as contínuas e as variáveis *dummy*. Valores de VIF superiores a 5 ou 10 são geralmente considerados indicativos de multicolinearidade problemática (Fávero; Belfiore, 2024). Atenção particular será dedicada à criação de variáveis *dummy* a partir de preditores categóricos, com a omissão de uma categoria de referência para cada um, de modo a evitar a armadilha da variável *dummy* (multicolinearidade perfeita), e a matriz de correlação entre os preditores poderá ser utilizada como ferramenta de diagnóstico preliminar.

Por fim, discute-se o pressuposto da Independência de Alternativas Irrelevantes (IIA), uma característica inerente ao modelo logístico multinomial. Este pressuposto estabelece que a razão

das probabilidades de uma observação pertencer a duas categorias distintas da variável dependente é independente da existência ou ausência de outras categorias "irrelevantes" no conjunto de desfechos. Embora classicamente relevante em modelos de escolha discreta, em que indivíduos selecionam uma opção dentre alternativas substitutas, sua aplicação direta e crítica pode ser ponderada no contexto deste estudo.

A variável dependente *tp_status_redacao* representa diferentes status ou classificações da redação (como "Sem problemas", "Anulada", "Em branco"), que são a conclusão observada da avaliação do texto, e não escolhas ativas dos participantes entre opções intercambiáveis. Dado que estas categorias são classificações qualitativamente distintas e não alternativas substitutas, a interpretação tradicional do IIA, e a necessidade de testes formais como o de Hausman-McFadden, podem ser consideradas menos centrais. O foco recai, portanto, na adequação global do modelo e na interpretabilidade dos fatores associados a cada status da redação em relação à categoria de referência, considerando a natureza específica dos desfechos da variável dependente.

4. Resultados e Discussão

Antes da estimação do modelo estatístico, foi realizada uma análise exploratória da amostra, com o objetivo de compreender a distribuição das variáveis envolvidas. A Tabela 3 apresenta a frequência absoluta e relativa das de todas as categorias utilizadas na modelagem. Variáveis marcadas com * indicam que ela foi usada como a referência no modelo.

Tabela 3. Frequência das variáveis categóricas

Variável/Categoria	Frequência Absoluta	Frequência Relativa
Status da redação (*tp_status_redacao*)	721.429	100%
Sem problemas (1)	692.666	96,01%
Anulada (2)	382	0,05%
Cópia Texto Motivador (3)	6.948	0,96%
Em Branco (4)*	12.704	1,76%
Fuga ao tema (6)	5.666	0,79%
Não atendimento ao tipo textual (7)	257	0,04%
Texto insuficiente (8)	2.389	0,33%
Parte desconectada (9)	417	0,06%
Tipo de escola (*tp_dependencia_adm_esc*)	721.429	100%
Federal (1)*	43.946	6,09%
Estadual (2)	456.911	63,33%
Municipal (3)	5.832	0,81%
Privada (4)	214.740	29,77%
Região *(regiao)*	721.429	100%
Norte (NN)	62.602	8,68%
Nordeste (NE)	236.850	32,83%
Sudeste (SE)*	259.666	35,99%
Sul (SS)	95.683	13,26%
Centro-Oeste (CO)	66.628	9,24%
Escolaridade do pai (*q001*)	721.429	100%
Nunca estudou (A)*	16.831	2,33%

Variável/Categoria	Frequência Absoluta	Frequência Relativa
Até o 5º ano do Ensino Fundamental (B)	77.719	10,77%
Até o 9º ano do Ensino Fundamental (C)	80.945	11,22%
Ensino Médio Incompleto (D)	83.630	11,59%
Ensino Médio Completo (E)	230.923	32,01%
Graduação Incompleta (F)	85.681	11,88%
Pós-Graduação completa (G)	70.334	9,75%
Não sabe (H)	16.831	2,33%
Escolaridade da mãe (*q002*)	721.429	100%
Nunca estudou (A)*	8.764	1,21%
Até o 5º ano do Ensino Fundamental (B)	48.996	6,79%
Até o 9º ano do Ensino Fundamental (C)	62.130	8,61%
Ensino Médio Incompleto (D)	82.847	11,48%
Ensino Médio Completo (E)	266.448	36,93%
Graduação Incompleta (F)	110.222	15,28%
Pós-Graduação completa (G)	113.356	15,71%
Não sabe (H)	28.666	3,97%
Renda (*faixa_sm*)	721.429	100%
Acima de 10 salários-mínimos (A)	50.953	7,06%
Mais de 5 a 10 salários-mínimos (B)	79.600	11,03%
Mais de 2 a 5 salários-mínimos (C)	196.390	27,22%
Mais de 1 a 2 salários-mínimos (D)	178.454	24,74%
Até 1 salário-mínimo (E)	186.733	25,88%
Sem renda (F)	29.299	4,06%

Fonte: Os autores (2025).

4.1. Análise do balanceamento das variáveis categóricas

A análise descritiva da amostra revela um forte desbalanceamento nas variáveis categóricas utilizadas no modelo. A variável dependente tp_status_redacao, por exemplo, mostra que 96,01% dos participantes obtiveram nota válida na redação (código 1), ao passo que todas as demais categorias representam menos de 2% da amostra cada. Esse desequilíbrio é particularmente relevante, dado que pode impactar a capacidade preditiva do modelo multinomial para classes raras, como as provas anuladas por estarem em branco (código 4, com 1,76%). Essa observação pode comprometer a significância dos parâmetros, especialmente para as categorias mais raras, como “Anulada” (código 2), “Não atendimento ao tipo textual” (código 7) e “Parte desconectada” (código 9).

Além disso, percebe-se que, nesta edição em específico, cerca de 4% dos estudantes que participaram dos dois dias de prova tiveram a redação anulada, uma proporção semelhante àquela encontrada por Viggiano e Mattos (2013) na edição de 2010. A estabilidade desse índice sugere que os fatores associados a esse desfecho são de natureza estrutural e persistente, e não meramente conjunturais ou específicos de uma única edição do exame, reforçando a importância de investigar as características socioeconômicas e educacionais que podem explicar este fenômeno. A estabilidade desse índice nos últimos anos, conforme ilustrado na Figura 1, sugere que os fatores associados a esse desfecho são de natureza estrutural e persistente. A Figura 2 mostra a abertura das categorias relacionadas à nulidade.

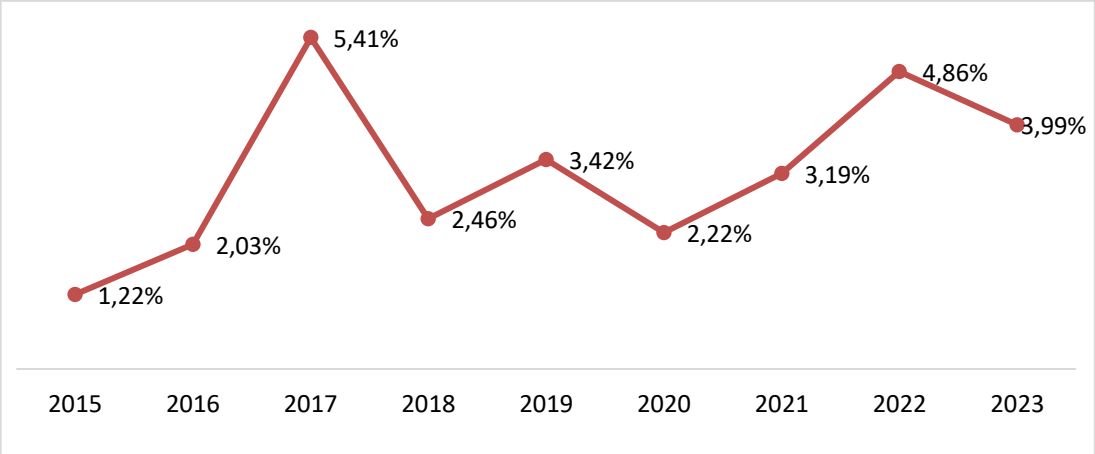

Figura 1. Percentual de redações anuladas das edições de 2015 a 2023.
Fonte: Adaptado de Brasil (2025a).

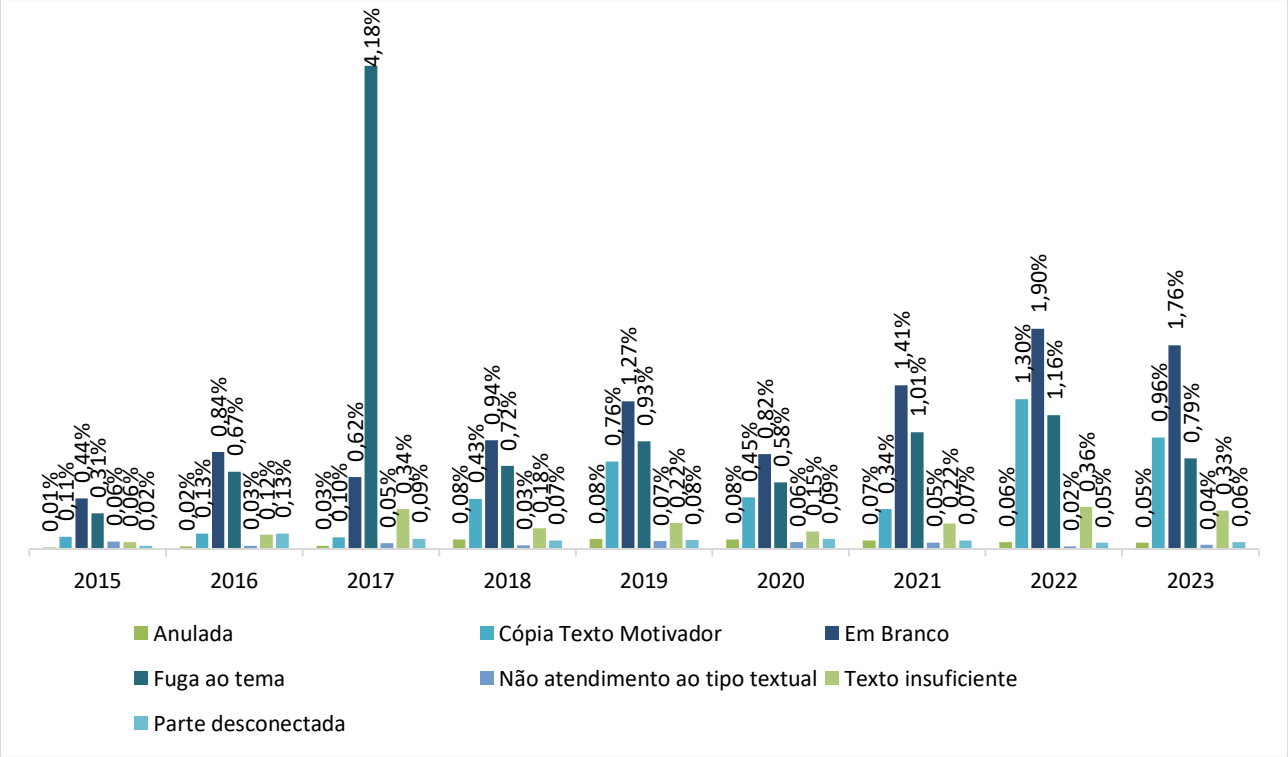

Figura 2. Abertura do percentual das redações anuladas por categoria das edições de 2015 a 2023.
Fonte: Adaptado de Brasil (2025a).

A distribuição dos status da Figura 2 mostra uma distribuição não homogênea entre as categorias que levam a uma redação ser anulada. É possível argumentar que esses três principais motivos discutidos a seguir refletem, em diferentes níveis, uma mesma defasagem cognitiva. A entrega da redação "Em Branco" sinaliza ou a desistência do aluno diante da tarefa ou a falta de planejamento para a preparação do texto frente a necessidade da realização das provas objetivas

(o que não deixa de ser um problema, uma vez que é necessário respeitar o tempo de aplicação do exame). Por sua vez, a "Fuga ao Tema" demonstra uma falha associada tanto a leitura quanto a compreensão do recorte temático. Por fim, a cópia sugere a carência de criatividade e de repertório para sustentar uma dissertação argumentativa. Portanto, tais categorias se destacam para uma parcela de estudantes que não desenvolveu as competências mínimas para interpretar uma proposta e articular um pensamento crítico por escrito, elementos centrais do exame.

Da mesma forma, a variável tp_dependencia_adm_esc apresenta forte concentração em escolas públicas estaduais (código 2, 63,33%), com número expressivamente menor de estudantes oriundos de escolas privadas (código 4, 29,77%) e escolas federais ou municipais somando menos de 7% juntos.

Nas variáveis socioeconômicas, a análise da distribuição de renda familiar a partir de faixa_sm revela um padrão de forte concentração. Conforme os dados, a esmagadora maioria dos participantes se agrupa nas faixas E, D e C, que, somadas, representam 77,84% de toda a amostra. Isso indica uma representatividade massiva de grupos com renda familiar de até 5 salários-mínimos, o que é condizente com o perfil socioeconômico do país. Em contrapartida, os grupos nos extremos da distribuição, como a faixa F (Sem Rendimento), com 4,06%, e a faixa A (Acima de 10 SM), com 7,06%, são significativamente minoritários.

Tal desbalanceamento dentro das categorias deve ser considerado tanto na especificação do modelo quanto na interpretação dos resultados, uma vez que pode reduzir a precisão das estimativas para grupos minoritários e amplificar a influência de categorias majoritárias.
A Tabela 4 apresenta as estatísticas descritivas das notas dos participantes nas notas objetivas da prova.`

Tabela 4 - Estatísticas descritivas dos preditores quantitativos.

Variável	N	Média	Desvio padrão	1Q	Mediana	3Q
nu_nota_cn	721.429	498,18	87,30	441,5	497,2	555,1
nu_nota_ch	721.429	526,96	86,62	472,0	534,7	588,5
nu_nota_mt	721.429	547,30	132,82	441,7	542,2	647,7
nu_nota_lc	721.429	522,05	74,47	475,8	528,2	574,6

Fonte: Os autores (2025).

Os resultados descritivos indicam que as notas das provas objetivas apresentam padrões distintos de dispersão, com destaque para a prova de matemática, cuja maior variabilidade pode refletir uma heterogeneidade mais ampla entre os participantes. Ao utilizar essas notas como variáveis explicativas para a situação da nota da redação, parte-se do pressuposto de que há transferência de competências cognitivas entre as áreas avaliadas conforme Gomes e Borges (2009).

Em especial, as provas de Linguagens e Códigos e Ciências Humanas são as que possuem maior relação com as habilidades exigidas na redação, como interpretação de textos, construção argumentativa e repertório sociocultural. Assim, espera-se que maiores notas nessas áreas estejam associadas a uma menor probabilidade de ocorrência de problemas na redação, como anulação, cópia do texto motivador ou fuga ao tema.

Contudo, o efeito dessas notas sobre o desempenho na redação pode variar conforme o tipo de escola frequentada pelo estudante, dado que escolas privadas e federais tendem a apresentar estruturas pedagógicas mais robustas e maior foco no desenvolvimento da escrita. Dessa forma, a inclusão de interações entre as notas das provas e a variável tp_dependencia_adm_esc pode revelar efeitos diferenciais, indicando, por exemplo, que o mesmo nível de desempenho em Linguagens e Códigos resulta em diferentes probabilidades de nota válida na redação a depender do tipo de escola.

Além disso, o perfil socioeconômico dos estudantes, capturado pelas variáveis q001, q002 e regiao, pode atuar como preditores diretos para influenciar tanto o desempenho nas áreas objetivas quanto na própria redação. Assim, a modelagem multinomial permite explorar essas interações complexas, possibilitando uma análise mais rica das desigualdades educacionais presentes na avaliação.

4.2. Verificação da presença de multicolinearidade

A análise de multicolinearidade revelou que todas as variáveis do modelo apresentaram valores de VIF consideravelmente baixos, indicando a ausência de colinearidade crítica entre os preditores. Este resultado sugere a estabilidade dos coeficientes estimados no modelo, fortalecendo a validade e a confiabilidade na interpretação individual dos efeitos de cada variável sobre a nota da redação.

4.3. Significância do modelo e das variáveis preditoras

Para avaliar a significância conjunta das variáveis explicativas no modelo de regressão logística multinomial, foi realizado um teste de razão de verossimilhança entre dois modelos aninhados: um modelo nulo (modelo 1), contendo apenas o intercepto, e um modelo completo (modelo 2), que inclui as variáveis *tp_dependencia_adm_esc*, *regiao*, *nu_nota_cn*, *nu_nota_ch*, *nu_nota_mt*, *nu_nota_lc*, *q001*, *q002* e *faixa_sm*.

O teste apresentou uma estatística de razão de verossimilhança com p-valor < 0,05, o que indica que o modelo completo apresenta ajuste significativamente melhor do que o modelo nulo. Deste modo, pode-se afirmar que, a 5% de significância, o conjunto das variáveis independentes contribui de forma significativa para explicar a variável resposta *tp_status_redacao*.

O pseudo-R^2 de Nagelkerke obtido foi de aproximadamente 0,177, indicando que o modelo explica cerca de 17,7% da variabilidade observada na situação da nota da redação. Embora esse valor não seja alto, ele é coerente com modelos de resposta categórica com múltiplas categorias, nos quais geralmente é mais difícil alcançar altos valores de R^2. Isso sugere que, embora as variáveis utilizadas contribuam significativamente para a explicação da variável dependente, outros fatores não observados também podem influenciar a situação da nota da redação.

Todavia, esse valor não deve ser interpretado de forma análoga ao R^2 da regressão linear tradicional. Isso ocorre porque o pseudo-R^2 é calculado com base na função de log-verossimilhança, e não na variância explicada dos resíduos. Como apontam Louviere; Hensher e Swait (2000), mesmo valores aparentemente baixos de pseudo-R^2 em modelos logit podem representar melhorias substanciais em relação ao modelo nulo (o que confere com o teste de

razão) e indicar um ajuste significativo. Portanto, o resultado ainda pode ser considerado satisfatório no contexto da modelagem de escolhas discretas, especialmente quando acompanhado por testes de verossimilhança estatisticamente significativos.

Para avaliar a contribuição individual de cada preditor, foi conduzida uma análise de desvio por meio do teste de Wald. Os resultados demonstram que todas as variáveis incluídas no modelo - tipo de escola (*tp_dependencia_adm_esc*), região (*regiao*), as notas das diferentes áreas de conhecimento (*nu_nota_cn*, *nu_nota_ch*, *nu_nota_mt*, *nu_nota_lc*) e os indicadores socioeconômicos (*q001, q002, faixa_sm*) - possuem uma associação altamente significativa com a variável dependente. Para todas as variáveis, o p-valor foi inferior ao nível de significância de 5%, reforçando a pertinência de sua inclusão no modelo final e indicando que cada uma delas agrega poder explicativo relevante.

4.3. Análise da significância dos parâmetros e dos efeitos marginais

O desempenho prévio dos estudantes, medido pelas notas nas provas objetivas, revelou-se um preditor significativo para a situação da redação. A análise dos efeitos marginais demonstra que notas mais altas nas provas objetivas estão associadas a uma maior probabilidade de o estudante obter uma nota válida e a uma menor de ser classificado em qualquer um dos motivos de anulação. A nota de Linguagens e Códigos apresentou o impacto mais expressivo, de modo que um incremento infinitesimal na prova aumenta a probabilidade de a redação ser classificada como "Sem problemas" (pela derivada ser positiva) e reduz significativamente a chance de anulação por "Cópia do Texto Motivador", "Fuga ao Tema" e por ser deixada "Em Branco" pelo motivo oposto.

Desta forma, o desempenho nas provas objetivas pode ser entendido como um fator de proteção à nulidade da prova escrita: quão maior for, menor a probabilidade nas categorias que configura nota zero à redação. Notas mais baixas, por outro lado, elevam tal risco. A deficiência em linguagens e códigos (*nu_nota_lc*) é crítica: os efeitos marginais indicam que notas baixas aumentam a probabilidade de a redação ser anulada por "Cópia do Texto Motivador" (categoria 3), "Fuga ao Tema" (categoria 6) e por ser entregue "Em Branco" (categoria 4). Um baixo desempenho em ciências humanas (*nu_nota_ch*) também se mostra um fator de risco, elevando a chance de anulação por esses mesmos motivos, o que sugere que a falta de repertório e de capacidade interpretativa se manifesta diretamente nos erros que levam à nota zero.

Vale acrescentar que, para alguns casos, os efeitos marginais não foram estatisticamente significativos, como os efeitos marginais da nota em ciências da natureza para as categorias “Anulada”, “Não atendimento ao texto” e “Parte desconectada”, não sendo possível inferir com respaldo estatístico os efeitos para essas variáveis. Todavia, era de se esperar a não significância em situações em que houvesse baixas representações de determinadas categorias dado o forte desbalanceamento dos dados, vide o que foi discutido na metodologia. Esses resultados, portanto, possuem uma associação com esta característica do banco de dados utilizado no estudo.

As observações acima corroboram a observação de Gomes e Borges (2009) sobre a transferência de habilidades cognitivas gerais avaliadas no exame, uma vez que bons

desempenhos nas provas objetivas aumentam a chance de ausência de problemas na redação e vice-versa.

Da perspectiva dos efeitos do tipo de escola frequentado pelo participante tomando as escolas Federais como referência, observa-se uma desvantagem para os alunos da rede pública estadual e municipal. Estudantes de escolas estaduais, em termos dos efeitos marginais, têm uma probabilidade 2,04 pontos percentuais menor de obter uma redação válida e chances percentuais maiores de incorrer em problemas como "Cópia do Texto Motivador" e "Fuga ao Tema". Em contraste, alunos de escolas privadas apresentam uma probabilidade maior de ter uma redação "Sem problemas" em comparação aos estudantes de escolas federais, controlando pelos demais fatores.

Sob o ponto de vista das provas anuladas, o impacto também é visível: estudantes de escolas estaduais, quando comparados aos de escolas Federais, apresentam, em média, mantidas todas as demais condições constantes, uma probabilidade maior de terem a redação anulada por "Cópia do Texto Motivador" (0,67%) e "Fuga ao Tema" (0,56%). Desta forma, ambos os resultados refletem a desigualdade no desempenho sob a perspectiva das escolas frequentadas pelos participantes, uma vez que as privadas tendem a incrementos positivos na probabilidade de ter a redação validada, ao passo que as públicas e estaduais tendem a ir ao caminho oposto.

A dinâmica acima se conecta às disparidades regionais, onde, com o sudeste como referência: participantes das regiões nordeste e norte apresentam uma probabilidade menor de obter uma redação sem problemas. Adicionalmente, estudantes do nordeste têm uma chance maior de deixar a redação "Em Branco" e de incorrer em "Cópia do Texto Motivador" do que os do Sudeste, sugerindo que as desigualdades educacionais regionais persistem e impactam diretamente um componente decisivo do Enem.

Sob a perspectiva regional, participantes do nordeste, em comparação com os do Sudeste, possuem uma chance de 0,56 ponto percentual maior de deixar a redação "Em Branco" e 0,36 ponto percentual maior de ser categorizado como "Em Cópia". Esses dados indicam que as condições estruturais das redes de ensino e as desigualdades regionais expõem certos grupos de estudantes a um risco maior de fracasso na redação.

O capital social e econômico das famílias, representado pela escolaridade dos pais e pela renda familiar, também se mostrou um fator determinante. Os resultados indicam que, quanto maior a escolaridade dos pais, especialmente a materna (*q002*), maior a probabilidade de o filho obter uma nota válida na redação e menor a chance de anulação por motivos como "Fuga ao Tema". A renda familiar (*faixa_sm*) funciona como um preditor igualmente robusto. Comparados aos participantes "Sem renda", estudantes de todas as outras faixas de renda têm um aumento na probabilidade de obter uma redação sem problemas. Inversamente, rendas mais altas estão associadas a maiores reduções na probabilidade de anulação; por exemplo, pertencer à faixa de renda mais elevada reduz significativamente a probabilidade de anulação por "Cópia do Texto Motivador" em relação a não possuir renda.

A menor escolaridade dos pais ou responsáveis (*q001* e *q002*) aumenta consistentemente a probabilidade de a redação ser zerada por diferentes motivos. Em comparação com filhos de mães com Ensino Médio completo, aqueles cujas mães nunca estudaram têm um risco maior de ter o texto anulado por "Fuga ao Tema". Da mesma forma, a baixa renda familiar (*faixa_sm*) eleva a vulnerabilidade. Participantes na condição "Sem renda" (a categoria de referência) apresentam uma probabilidade maior de ter a redação anulada por "Cópia do Texto Motivador" ou "Fuga ao Tema" quando comparados àqueles de faixas de renda mais altas. Isso demonstra que a carência de capital econômico e social se traduz em um maior risco de cometer os erros que levam à anulação da prova.

4.4. Regressão linear complementar

Conforme descrito na metodologia, foi realizada uma análise de regressão linear múltipla para aprofundar a compreensão dos fatores determinantes da pontuação na redação, considerando apenas os participantes com notas válidas. O modelo ajustado mostrou-se robusto, alcançando um R^2 de 0,39. Isso indica que o conjunto de variáveis selecionadas é capaz de explicar aproximadamente 39% da variação nas notas de redação. A significância global do modelo foi confirmada pelo teste F (p-valor < 0,001), validando a relação entre os preditores e o desempenho na prova.

A análise dos coeficientes revela o impacto de diferentes fatores na nota final da redação. Os fatores escolares e regionais mostraram-se preditores de grande impacto, pois, tomando as escolas federais como referência, os estudantes de escolas estaduais e municipais obtiveram, em média, notas 48,6 e 49,3 pontos inferiores, respectivamente. Em contrapartida, alunos de escolas privadas apresentaram um desempenho médio 20,4 pontos superior, descobertas estas que corroboram a literatura sobre desigualdade na qualidade entre redes de ensino.

Regionalmente, e controlando pelos demais fatores, participantes do nordeste e do norte tiveram as maiores notas em comparação com a região sudeste. Tal resultado sugere que, apesar dessas regiões terem maiores problemas com alunos que não conseguem atender aos requisitos básicos da redação de modo que ela não seja anulada, aqueles que a realizam possuem desempenhos acima da média.

Adicionalmente, o desempenho acadêmico prévio é um forte preditor, com a nota de linguagens e códigos (*nu_nota_lc*) apresentando o maior efeito positivo, onde cada ponto adicional nesta prova está associado a um aumento de 0,60 ponto na redação. As notas de ciências humanas (*nu_nota_ch*) e matemática (*nu_nota_mt*) também se mostraram preditores altamente significativos. Por fim, os fatores socioeconômicos, como a escolaridade dos pais (*q001* e *q002*) e a renda familiar *(faixa_sm*), também foram determinantes significativos, com níveis mais elevados associando-se positivamente a maiores notas e reforçando o impacto do capital social e econômico no desempenho educacional.

6. Conclusões

Este estudo mostrou que a capacidade de desenvolver os requisitos mínimos redação do Enem para a garantia de uma nota não nula é um reflexo de um complexo conjunto de desigualdades estruturais. Os resultados apontam que estudantes com menor desempenho nas provas objetivas, oriundos de escolas públicas, residentes nas regiões norte e nordeste e com menor capital socioeconômico enfrentam um duplo prejuízo: um risco significativamente maior de ter a redação anulada e, quando a prova é validada, a tendência a notas consideravelmente inferiores. Fica evidente, portanto, que o capital social, econômico e cultural acumulado ao longo da trajetória educacional se mostra como um fator decisivo para o sucesso neste componente do exame.

As implicações não só destes resultados, como também a literatura consultada, são diretas para o debate sobre políticas públicas, indicando que intervenções focadas apenas em aspectos pedagógicos são insuficientes se não abordarem as disparidades estruturais entre as redes de ensino e as desigualdades regionais. Embora este estudo identifique fortes associações, sua natureza transversal não permite estabelecer causalidade direta. Para aprofundar a análise, sugere-se a condução de um novo estudo com modelos de regressão multinível, abordagem ideal para investigar como o contexto escolar e regional influencia os resultados individuais e, assim, qualificar o desenho de políticas mais eficazes.

Conclui-se, assim, que a redação do Enem transcende um simples teste de escrita para atuar como um espelho que reflete e amplifica as profundas diferenças sociais e educacionais do Brasil. Este cenário reforça a necessidade de um compromisso contínuo e efetivo com a promoção da equidade em todo o sistema de ensino nacional.

Referências

ABREU DE CARVALHO JUNIOR, José Roberto; DE ALMEIDA MENDES, Wanderson; MARQUES FERREIRA, Marco Aurélio. Influência da Qualificação Docente Sobre o Desempenho Discente no Enem. Administração: Ensino e Pesquisa, v. 24, n. 1, p. 72–97, 31 maio 2023.

BECKER, Gary S. **Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education.** Chicago: University of Chicago Press, 1964.

BOURDIEU, Pierre; PASSERON, Jean-Claude. **A reprodução: elementos para uma teoria do sistema de ensino**. Rio de Janeiro: Francisco Alves, 1970.

COLEMAN, James S. et al. **Equality of Educational Opportunity**. Washington, DC: U.S. Government Printing Office, 1966.

BRASIL. Microdados. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>. Acesso em: 5 jun. 2025a.

BRASIL. 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>. Acesso em: 28 fev. 2025b.

BRASIL. Sistema Único de Seleção Unificada (SISU). Disponível em: <https://acessounico.mec.gov.br/sisu>. Acesso em: 30 maio. 2025c.

BRASIL. Programa Universidade para Todos. Disponível em: <https://acessounico.mec.gov.br/prouni>. Acesso em: 30 maio. 2025d.

BRASIL. Fundo de Financiamento Estudantil. Disponível em: <https://acessounico.mec.gov.br/fies>. Acesso em: 4 jun. 2025e.

FÁVERO, Luiz Paulo; BELFIORE, Patrícia. Manual de Análise de Dados: Estatística e Machine Learning com Excel®, SPSS®, Stata®, R® e Python®. 2. ed. Rio de Janeiro: GEN LTC, 2024.

GOMES, Cristiano Mauro Assis; BORGES, Oto. O Enem é uma avaliação educacional construtivista? Um estudo de validade de construto. Estudos em Avaliação Educacional, v. 20, n. 42, p. 73–88, abr. 2009.

GOMES, Tamiris. De onde são os estudantes que tiraram nota mil no Enem 2024? | CNN Brasil. Disponível em: <https://www.cnnbrasil.com.br/educacao/de-onde-sao-os-estudantes-que-tiraram-nota-mil-no-enem-2024/>. Acesso em: 4 mar. 2025.

GREENE, William H. .. Econometric analysis. 6. ed. *[S.l.]*: Prentice Hall, 2008.

JALOTO, Alexandre; PRIMI, Ricardo. Fatores socioeconômicos associados ao desempenho no Enem. Em Aberto, v. 34, n. 112, p. 125–141, 30 dez. 2021.

LIMA, Bruna. Indicadores do Enem evidenciam diferenças entre escolas públicas e privadas.

LOUVIERE, J. J.; HENSHER, D. A.; SWAIT, J. D. Stated Choice Methods. Cambridge: Cambridge University, 2000.

LUCENA, João Paulo Oliveira; DOS SANTOS, Heric Nero Lisboa. A relação entre desempenho no Exame Nacional do Ensino Médio e o perfil socioeconômico: um estudo com os microdados de 2016. Revista de Gestão e Secretariado, v. 11, n. 2, p. 1–23, 5 ago. 2020.

TRAVITZKI, Rodrigo; FERRÃO, Maria Eugénia; COUTO, Alcino Pinto. Desigualdades educacionais e socioeconómicas na população Brasileira pré-universitária: Uma visão a partir da análise de dados do ENEM. Education Policy Analysis Archives, v. 24, n. 74, p. 1–36, 11 jul. 2016.

VIGGIANO, Esdras; MATTOS, Cristiano. O desempenho de estudantes no Enem 2010 em diferentes regiões brasileiras. Revista Brasileira de Estudos Pedagógicos, v. 94, n. 237, p. 417–438, 20 ago. 2013.