FUNDAÇÃO GETULIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA

Análise da influência do tema da redação do Enem 2022 sobre indivíduos de cor/raça indígena

---

Projeto de Monografia apresentado à Escola de Matemática Aplicada – FGV/EMAp como requisito parcial para continuidade ao trabalho de monografia

Orientador: Rodrigo dos Santos Targino

Aluno: Rafael Felipe dos Santos

Rio de Janeiro, 2025

FUNDAÇÃO GETULIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA

"Declaro ser o único autor do presente projeto de monografia que refere-se ao plano de trabalho a ser executado para continuidade da monografia e ressalto que não recorri a qualquer forma de colaboração ou auxílio de terceiros para realizá-lo a não ser nos casos e para os fins autorizados pelo professor orientador."

_____

Rafael Felipe dos Santos

_____

Professor Orientador: Rodrigo dos Santos Targino

Aprovado em _____ de _____ de _____
Grau atribuído ao Projeto de Monografia: _____

Índice de conteúdo

Índice de imagens

1. Introdução

1.1. O problema e sua relevância

O Exame Nacional do Ensino Médio (Enem) é um dos principais instrumentos de avaliação educacional do Brasil, desempenhando um papel crucial na seleção de candidatos para o ingresso em instituições de ensino superior. A prova de redação, em particular, tem um peso significativo na composição da nota final e frequentemente aborda temas sociais relevantes. Em 2022, especificamente, o tema da redação foi "desafios para a valorização de comunidades e povos tradicionais no Brasil". Tal fato levanta uma questão importante: o quanto o tema da prova pode influenciar o desempenho de certos grupos sociais, especialmente daqueles que se identificam com a temática abordada?

Este trabalho busca investigar se candidatos que se autodeclararam indígenas foram favorecidos pela escolha do tema da redação de 2022. A relevância da investigação está em compreender como a formulação do tema pode afetar a equidade no desempenho entre diferentes grupos sociais. Caso se comprove uma influência significativa, os resultados poderão subsidiar reflexões sobre a elaboração de temas mais inclusivos, especialmente para populações historicamente marginalizadas.

Para explorar essa questão, utilizaremos um conjunto abrangente de variáveis, incluindo gênero, idade, nível de educação dos pais, local de residência, habilidade de escrita, entre outros. Essas variáveis permitirão controlar possíveis fatores de confusão e obter uma ideia mais precisa do efeito da cor/raça indígena no desempenho da redação de 2022.

Inicialmente, o foco será na análise da prova de 2022. No entanto, posteriormente, podemos explorar outros temas de redação que abordaram diferentes grupos sociais.

1.2. Dados

Os dados utilizados neste estudo foram obtidos a partir do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Cada ano da prova do Enem possui uma planilha com aproximadamente 4 milhões de linhas e 76 colunas, contendo informações sobre características socioeconômicas dos alunos além dos resultados em todas as provas do exame. A redação do Enem é avaliada com base em cinco competências. Para este estudo, focaremos nas competências que avaliam a compreensão e desenvolvimento do tema, pois são diretamente influenciadas pelo tema da redação. Nesse sentido, as competências 1 e 4 poderão utilizadas como parâmetros preditivos, pois avaliam aspectos técnicos da escrita e não possuem relação com o tema. Para este estudo, serão analisados os dados dos anos de 2021, 2022 e 2023.

1.2.1. Tratamento de dados

É importante mencionar que, durante o tratamento dos dados, foi observado que apenas os alunos que estavam cursando o último ano do ensino médio informaram o tipo de escola (pública ou privada) que frequentavam/frequentaram. Portanto, a análise será focada nesses alunos, excluindo treineiros e egressos e garantindo uma maior homogeneidade nos dados. Essa decisão também foi tomada tendo em vista que o tipo de escola é uma variável mais relevante para determinar a situação socioeconômica de um aluno do que o tempo desde que concluiu o Ensino Médio.

Além disso, para garantir uma comparação justa entre indígenas e não indígenas, realizamos um matching exato, o que significa que, para cada aluno indígena, identificamos um aluno não indígena com os mesmos valores para várias características socioeconômicas e demográficas. Essas características incluem faixa etária, tipo de escola, número de computadores em casa, acesso à internet, renda familiar per capita, nível de escolaridade do responsável mais educado, nível de emprego do responsável melhor empregado, se o aluno realizou a prova em uma capital e qual região do país ele reside.

Esse procedimento foi adotado para reduzir vieses, igualando as situações socioe-

Variáveis	Tipo	Descrição	Domínio
NU_ANO	Categórica Ord.	Ano do Enem	0,1 ou 2 (2021, 2022 ou 2023)
TP_FAIXA_ETARIA	Categórica Ord.	Faixa etária	0 a 6 (17 a 20 anos)
TP_SEXO	Categórica Bin.	Sexo	0 = feminino, 1 = masculino
TP_COR_RACA_1	Categórica Bin.	Indica cor/raça branca	0 = False, 1 = True
TP_COR_RACA_2	Categórica Bin.	Indica cor/raça preta	0 = False, 1 = True
TP_COR_RACA_3	Categórica Bin.	Indica cor/raça parda	0 = False, 1 = True
TP_COR_RACA_4	Categórica Bin.	Indica cor/raça amarela	0 = False, 1 = True
TP_COR_RACA_5	Categórica Bin.	Indica cor/raça indígena	0 = False, 1 = True
TP_ESCOLA	Categórica Ord	Tipo de escola no E.M.	0 = pública, 1 = privada
Regiao_Centro-Oeste	Categórica Bin.	indica região onde fez a prova	0 = False, 1 = True
Regiao_Nordeste	Categórica Bin.	indica região onde fez a prova	0 = False, 1 = True
Regiao_Norte	Categórica Bin.	indica região onde fez a prova	0 = False, 1 = True
Regiao_Sudeste	Categórica Bin.	indica região onde fez a prova	0 = False, 1 = True
Regiao_Sul	Categórica Bin.	indica região onde fez a prova	0 = False, 1= True
CAPITAL	Categórica Bin.	fez a prova em uma capital	0 = False, 1 = True
NU_NOTA_CN	Num. Contínuo	nota em Ciências da Natureza	0 a 1000
NU_NOTA_CH	Num. Contínuo	nota em Ciências Humanas	0 a 1000
NU_NOTA_LC	Num. Contínuo	nota em Linguagens	0 a 1000
NU_NOTA_MT	Num. Contínuo	nota em Matemática	0 a 1000
NU_NOTA_COMP1	Num. Contínuo	nota na Competência 1	0 a 200
NU_NOTA_COMP4	Num. Contínuo	nota na Competência 4	0 a 200
renda_fam	Categórica Ord.	Renda Familiar Per Capita	0 a 12
maior_escolaridade	Categórica Ord.	nível de escolaridade do responsável mais educado	0 a 6
maior_emprego	Categórica Ord.	nível de emprego do responsável melhor empregado	0 a 4
Q022	Categórica Ord.	número de celulares em casa	0 a 4
Q025	Categórica Bin.	acesso à internet	0 = False, 1 = True
tema_relevante	Categórica Bin.	Se o tema da redação é relevante para índigenas	1, caso ano seja 2022, 0 c.c
NU_NOTA_REDACAO	Num. Contínuo	Nota na redação	0 a 1000

Tabela 1: Variáveis selecionadas como relevantes (já tratadas)

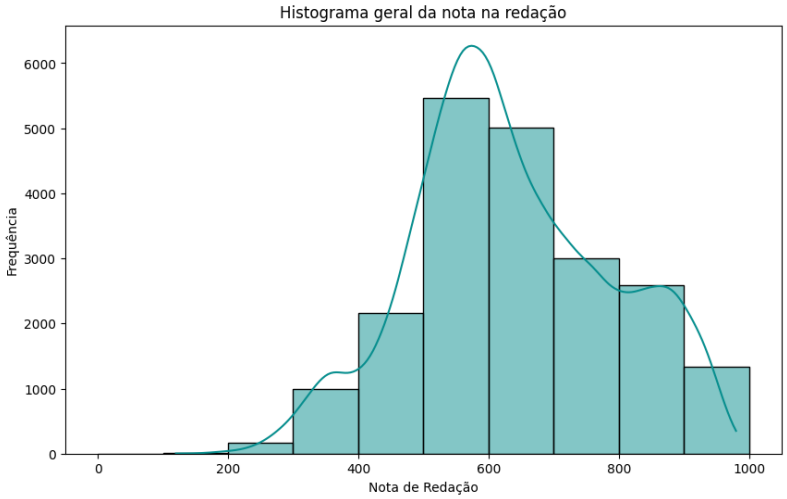
conômicas dos grupos comparados. Com o matching, as correlações entre variáveis socioeconômicas e a nota da redação deve ser menor, refletindo a equiparação das condições de base.

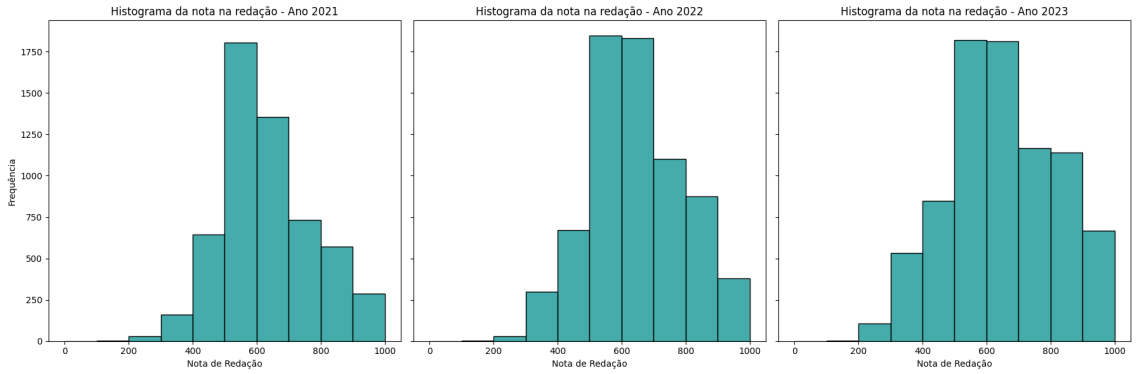
Os alunos indígenas apresentam diferenças significativas em relação à população geral, como menor acesso à internet, menor presença em capitais e uma distribuição geográfica distinta. Além disso, há uma grande disparidade na quantidade de dados entre indígenas e não indígenas (aproximadamente 200 não-índios para cada índio). Portanto, o matching deve proporcionanr uma melhor compreensão do impacto do tema da redação sobre esse grupo.

1.3. Análise exploratória

1.3.1. Distribuição geral nas notas de redação

A distribuição das notas da redação segue um padrão aproximadamente normal. Isso é esperado em uma amostra grande e diversa como a do Enem e é favorável para a aplicação de modelos de regressão linear.





1.3.2. Comparação das notas entre indígenas e não-indígenas

- **Pré-matching:** Ao observar o KDE das notas de redação entre indígenas e não indígenas antes do matching, verifica-se que os indígenas consistentemente apresentaram notas inferiores aos não indígenas. No entanto, essa diferença nas notas diminuiu a cada ano. Em 2023, a diferença foi significativamente menor do que em 2021, apesar de os temas das redações não serem diretamente relacionados à questão racial em ambos os anos. Essa redução na disparidade pode ser atribuída a vários fatores. Uma possível explicação é que a melhoria da qualidade da internet em regiões remotas podem ter contribuído para uma melhor preparação para o exame, especialmente para os estudantes indígenas que, quando comparados com os não-indígenas, costumam morar em locais mais afastados e com menor acesso à internet. Independente da explicação para esse fenômeno, que não vamos nos aprofundar, é fato que o ano de aplicação da prova será uma variável relevante para o nosso modelo.

- **O efeito do matching:** Em 2022, a diferença entre as curvas de indígenas e não-indígenas teve uma redução após o matching, algo que não foi observado nos outros anos. Tal fato pode indicar que há fatores específicos nesse ano que influenciaram a distribuição das notas de redação entre os grupos.

  O tema da redação de 2022, pode ter sido particularmente relevante para os estudantes indígenas, mais do que para os não indígenas. Isso pode ter reduzido as diferenças nas habilidades de redação quando os fatores sociais foram controlados pelo matching. Em outras palavras, o tema pode ter nivelado o campo de jogo de uma maneira que os outros temas de 2021 e 2023 não fizeram.

  Importante mencionar que não estamos afirmando que isso de fato ocorreu, mas que esse resultado corrobora com a nossa tese inicial.

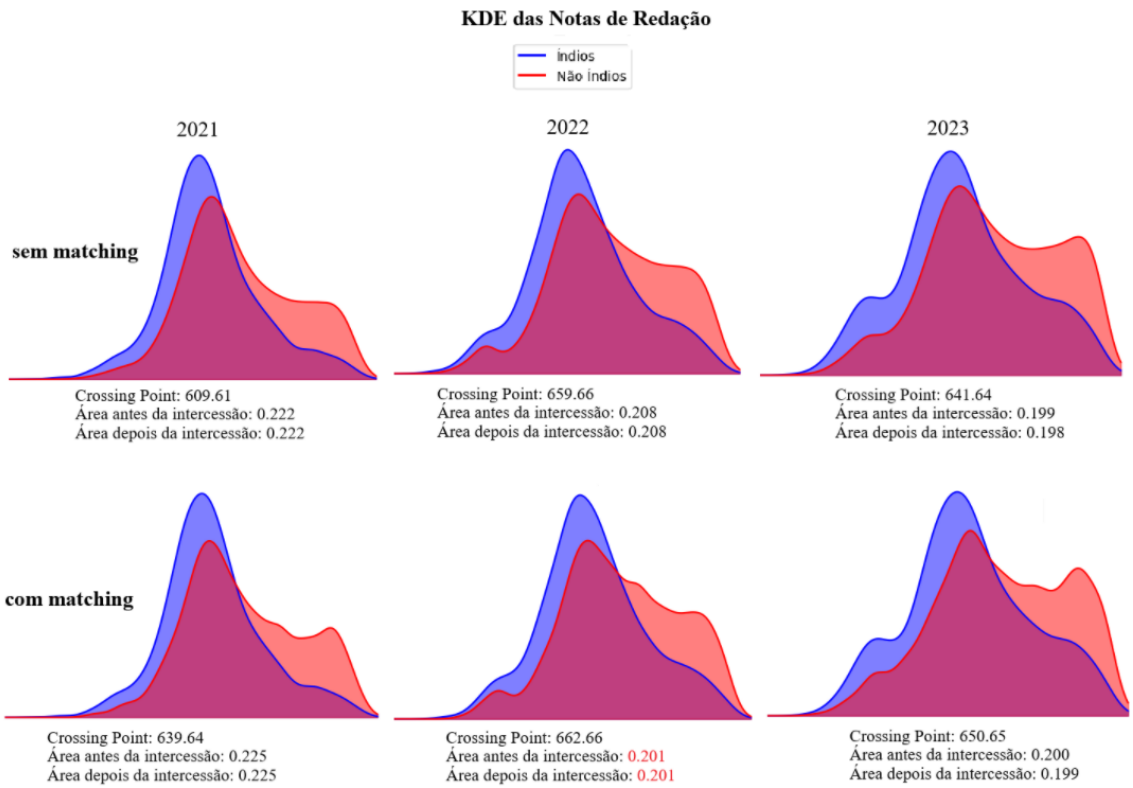
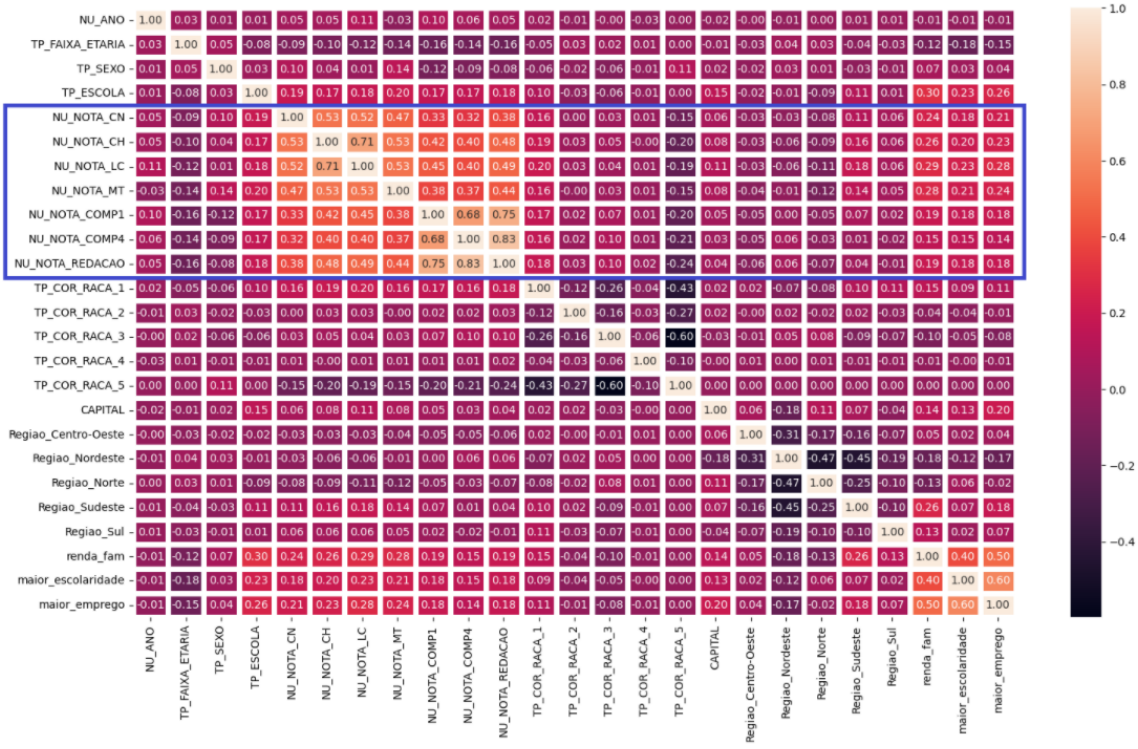


Imagem 1: Área entre as curvas antes e depois do ponto de interseção para os diferentes anos. Em 2022, o uso de matching resultou na redução da área da diferença entre as curvas, ao contrário dos outros anos, onde essa redução não foi observada.

1.3.3. Análise de Correlação entre Variáveis

Uma etapa crucial da análise exploratória é entender como as variáveis se relacionam entre si, especialmente em relação à nossa variável de interesse: a nota da redação. Para isso, construímos uma matriz de correlação, que nos permite visualizar essas relações de maneira clara e intuitiva (a partir de agora, irei me referir aos dados já tratados, ou seja, com o matching já feito).

A matriz de correlação abaixo mostra a correlação entre todas as variáveis do conjunto de dados, incluindo notas das diversas provas (Matemática, Linguagens, Ciências Humanas, etc.) e variáveis socioeconômicas.



Ao analisar a matriz, podemos observar que as notas das outras provas apresentam valores de correlação similares com a nota da redação. Isso sugere que fatores sociais e econômicos tem um impacto similar independente de qual prova estejamos comparando.

Notas como Proxies: Esse fenômeno pode ser entendido pelo conceito de "proxies". Proxies são variáveis que, apesar de não medirem diretamente o que queremos, possuem uma relação forte com a variável de interesse, permitindo inferências sobre ela. No nosso caso, as notas em Matemática, Linguagens, Ciências Humanas e outras disciplinas atuam como proxies para a nota de redação. Elas encapsulam informações sobre as habilidades gerais dos candidatos, além das características socioeconômicas de cada um.

Essa alta correlação entre as notas em cada prova pode introduzir o problema de multicolinearidade nos modelos de regressão. A multicolinearidade afeta principalmente as estimativas dos coeficientes das variáveis correlacionadas, tornando-os menos precisos (ou seja, com maiores erros padrão). Isso pode dificultar a interpretação dos coeficientes individuais dessas variáveis, pois é difícil determinar o efeito

isolado de cada uma delas. No entanto, como nosso objetivo principal estimar o impacto da variável 'TP_COR_RACA_5' no ano de 2022, se incluirmos no modelo as notas de outras provas, a multicolinearidade entre as notas das provas não deve afetar significativamente o coeficiente da variável que identifica participantes indígenas. Isso porque a variável 'TP_COR_RACA_5' não é altamente correlacionada com as notas das provas. Nesse sentido, a construção de um modelo de regressão linear deve permitir observar os seguintes aspectos: i) Os coeficientes das notas das provas podem ser menos precisos devido à multicolinearidade, mas isso não impede a inclusão dessas variáveis no modelo; ii) O coeficiente da variável 'TP_COR_RACA_5' deve permanecer relativamente estável e interpretável. Este coeficiente nos dará uma estimativa do impacto de ser indígena na nota da redação, controlando pelas outras variáveis.

2. Próximos Passos

Para investigar os fatores que influenciam o desempenho na redação e realizar previsões, pretendo utilizar diferentes abordagens estatísticas e comparar seus resultados.

2.1. Modelagem

Para analisar os fatores que influenciam o desempenho dos alunos na redação, planejo utilizar inicialmente modelos de regressão linear. Essa escolha se baseia na capacidade desse método de estimar de forma direta o impacto de variáveis explicativas, como as notas em outras áreas e a identificação como indígena, sobre a nota da redação, além de permitir o controle por variáveis correlacionadas.

A definição do modelo final será feita após uma análise exploratória detalhada e testes comparativos. Serão consideradas variações dentro da abordagem de regressão linear, como a inclusão de termos de interação, ponderação ou outras extensões que garantam a adequação estatística do modelo aos dados analisados.

2.2. Diagnósticos

Posteriormente, discutiremos os diagnósticos empregados para avaliar a adequação dos modelos. Assim, será possível verificar se os pressupostos estatísticos foram atendidos e identificar possíveis problemas que comprometam a validade das inferências.

2.3. Conclusão

Assim, estamos dando um passo importante em direção à resposta da pergunta inicial: "Como o tema da prova pode afetar o desempenho de grupos sociais que se identificam com o tema?"