SciELO Preprints

Estado da publicação: O preprint não foi submetido para publicação

ESTUDO DA RELAÇÃO ENTRE O DESEMPENHO NO EXAME NACIONAL DO ENSINO MÉDIO E ALGUNS INDICADORES SOCIAIS E ECONÔMICOS

Elisa Thomé Sena, Ewout der Harr, Otaviano Helene

https://doi.org/10.1590/SciELOPreprints.12047

Submetido em: 2025-05-21
Postado em: 2025-05-26 (versão 1)
(AAAA-MM-DD)

A moderação deste preprint recebeu o endosso de:
Sonia Maria Portella Kruppa (ORCID: https://orcid.org/0000-0001-6195-3436)

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

ESTUDO DA RELAÇÃO ENTRE O DESEMPENHO NO EXAME NACIONAL DO ENSINO MÉDIO E ALGUNS INDICADORES SOCIAIS E ECONÔMICOS

ELISA THOMÉ SENA[1]
elisa.sena@unifesp.br
https://orcid.org/0000-0003-0939-0036

EWOUT TER HAAR[2]
ewout@usp.br
https://orcid.org/0000-0002-0951-525X

OTAVIANO HELENE[2]
otaviano@if.usp.br
https://orcid.org/0000-0002-7689-1706

[1]Universidade Federal de São Paulo, Osasco (SP), Brasil
[2]Universidade de São Paulo, São Paulo (SP), Brasil

RESUMO: Este trabalho apresenta uma análise da relação entre o desempenho dos estudantes nas provas de Exame Nacional de Ensino Médio, ENEM, com diversos indicadores escolares, socioeconômicos, individuais e geográficos. Foram considerados os dados do ENEM de 2019, último ano disponível antes do início da pandemia de COVID-19. Os indicadores usados como variáveis explicativas foram: a renda domiciliar per capita, a escolaridade dos pais, a etnia/cor da pele, a idade, a unidade da federação onde o exame foi realizado e o vínculo administrativo da escola onde o candidato cursou o ensino médio. Os parâmetros que relacionam as variáveis explicativas com as notas do ENEM foram ajustados por meio de uma regressão linear usando-se o método dos mínimos quadrados ordinário em um procedimento equivalente ao hierárquico multinível. Os ajustes dos parâmetros foram feitos considerando-se tanto as notas médias como as notas em cada uma das provas. Os resultados obtidos indicam uma forte relação entre o desempenho nas provas e aqueles indicadores. Todas as variáveis independentes consideradas mostraram-se estatisticamente significativas. Entretanto, elas são insuficientes para explicar a totalidade dos resultados obtidos pelos estudantes. Os resultados obtidos podem ser considerados na elaboração de políticas públicas na área educacional e para indicar pesquisas posteriores.

Palavras-chave: Indicadores socioeconômicos - Desempenho estudantil –Exame Nacional do Ensino Médio (ENEM)

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

STUDY OF THE RELATIONSHIP BETWEEN PERFORMANCE IN THE NATIONAL HIGH SCHOOL EXAM AND SOME SOCIAL AND ECONOMIC INDICATORS

ABSTRACT: This paper presents an analysis of the relationship between students' performance on the National High School Exam (ENEM) and various school, socioeconomic, individual, and geographic indicators. The data analyzed were from the 2019 edition of ENEM, the last available year before the onset of the COVID-19 pandemic. The explanatory variables included per capita household income, parents' educational attainment, race/skin color, age, the state where the exam was taken, and the administrative affiliation of the high school the student attended. The parameters linking the explanatory variables to ENEM scores were estimated using linear regression via the ordinary least squares method, in a procedure equivalent to multilevel hierarchical modeling. The parameters were estimate considering both the average scores and the scores for each individual subject test. The results indicate a strong relationship between exam performance and the explanatory variables we selected. All the independent variables analyzed were statistically significant. However, they are not sufficient to fully explain students' performance. The results obtained can be considered in the development of public policies in the education field and to guide further researches.

Keywords: Socioeconomic indicators – Student performance – National High School Exam (ENEM)

RESUMEN Este trabajo presenta un análisis de la relación entre el desempeño de los estudiantes en las pruebas del Examen Nacional de Enseñanza Media (ENEM) y diversos indicadores escolares, socioeconómicos, individuales y geográficos. Para ello, se consideraron los datos del ENEM de 2019, el último año disponible antes del inicio de la pandemia de COVID-19.

Los indicadores utilizados como variables explicativas incluyen el ingreso familiar per cápita, el nivel educativo de los padres, la etnia/color de piel, la edad, la unidad federativa donde se realizó el examen y la naturaleza administrativa de la escuela en la que el candidato cursó la educación media.

Para analizar la relación entre estas variables explicativas y las calificaciones del ENEM, se empleó un modelo de regresión lineal basado en el método de los mínimos cuadrados ordinarios, siguiendo un procedimiento análogo al modelo jerárquico multinivel. Los parámetros fueron ajustados considerando tanto las calificaciones promedio como las correspondientes a cada una de las pruebas individuales.

Los resultados obtenidos revelan una estrecha relación entre el desempeño en los exámenes y los indicadores analizados. Todas las variables independientes consideradas resultaron ser

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

estadísticamente significativas. Sin embargo, estos factores por sí solos no explican completamente los resultados obtenidos por los estudiantes. Los hallazgos de este estudio pueden contribuir a la formulación de políticas públicas en el ámbito educativo y servir como referencia para futuras investigaciones.

.

Palabras clave: Indicadores socioeconómicos – Desempeño estudiantil – Examen Nacional de Enseñanza Media

INTRODUÇÃO

Conhecer as relações entre o desempenho estudantil e indicadores socioeconômicos, geográficos, individuais e escolares pode contribuir para o estabelecimento de políticas públicas ligadas ao setor educacional. Com esse objetivo, apresentamos um estudo da relação entre o desempenho dos estudantes brasileiros em provas do Exame Nacional do Ensino Médio (ENEM) com as seguintes variáveis independentes consideradas como explicativas: renda domiciliar per capita, nível escolar de mães e pais, unidade federativa, idade, vínculo administrativo da escola frequentada no ensino médio, cor/etnia e sexo. Os resultados são equivalentes a uma análise multinível, como ilustrado no Apêndice C. A técnica adotada, ajustar parâmetros que relacionam as variáveis independentes com as variáveis dependentes usando o método dos mínimos quadrados, permite estimar a relação entre as notas e cada uma das variáveis independentes quando todas as outras são mantidas constantes.

As variáveis dependentes consideradas, ou seja, os indicadores do desempenho estudantil, foram as notas nas diversas provas do Exame Nacional do Ensino Médio (ENEM) de 2019 bem como a nota média naquele exame. A amostra analisada continha, inicialmente, pouco mais do que 5 milhões de estudantes. Excluindo treineiros e pessoas que não compareceram a uma das provas, a amostra foi reduzida para pouco mais do que três milhões. Finalmente, foram excluídos da análise registros em que havia valores faltantes para quaisquer das variáveis explicativas, bem como casos em que o candidato respondeu que não sabia a renda domiciliar ou escolaridade dos pais. Feitas essas exclusões, a amostra analisada continha um total de 757712 estudantes. Vale ressaltar que, a base de dados utilizada fornecia a informação sobre vínculo administrativo da escola apenas para os estudantes que concluiriam o ensino médio em 2019. Desta forma, o modelo foi ajustado para este grupo.

Como consequência das exclusões, como ocorrem em qualquer análise com dados submetidos a filtros, os resultados obtidos são válidos apenas para o conjunto analisado, sendo enviesados em relação a outro subconjunto ou ao universo dos concluintes do ensino médio. Entretanto, podemos supor que as conclusões qualitativas são válidas para o universo, indicando possíveis análises mais específicas quando elas se

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

mostrarem necessárias quanto aos aspectos quantitativos. Igualmente, os resultados podem indicar a necessidade de inclusão de novas variáveis explicativas.

É necessário observar que o fato de terem sido considerados os exames de 2019 não limita o alcance das conclusões, uma vez que as variações do desempenho estudantil ao longo do tempo bem como suas relações com as variáveis independentes são bastante lentas. Além disso, o exame de 2019 é o último a não ser afetado pela pandemia da COVID-19.

O vínculo administrativo da escola, a unidade da federação, o sexo e a cor/etnia foram considerados como variáveis fictícias (também chamadas de variáveis binárias ou dummy). A técnica adotada é de uso comum em ciências sociais aplicadas (CHEIN, 2019).

Considerando uma perspectiva de análise multinível, teríamos um nível associado à unidade da federação e outro, à pessoa, incluindo os aspectos socioeconômicos (renda per capita, cor/etnia, escolaridade dos pais, gênero e idade), e as características da escola.

VARIÁVEIS EXPLICATIVAS

Nesta seção examinamos possíveis relações entre as variáveis explicativas e as notas médias nas provas do ENEM. Nesta análise exploratória foram excluídos apenas treineiros e candidatos que não compareceram a pelo menos uma das provas. Porém, como nesta etapa cada variável explicativa foi analisada individualmente, o número de dados faltantes em cada uma das variáveis analisadas não é sempre o mesmo.

Esse exame preliminar serve apenas como guia ou justificativa para a escolha das variáveis explicativas, uma vez que não considera as possíveis correlações entre elas. A regressão múltipla, por sua vez, apresentada na Seção III, permite estimar o efeito de cada variável quando todas as demais permanecem inalteradas.

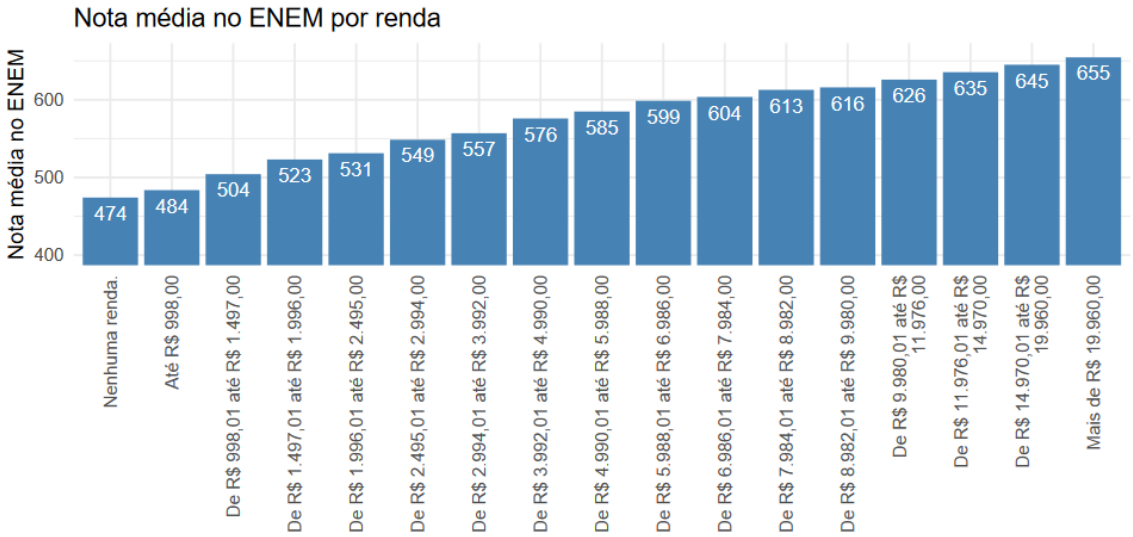
Dependência com a renda

As condições socioeconômicas estão relacionadas ao desempenho de estudantes no ENEM [ver, p. ex., JALOTO, 2021], entre elas a renda familiar. A figura 1 mostra as notas médias no ENEM de 2019 da amostra analisada para diferentes faixas de renda domiciliares, evidenciando a relevância desse fator no desempenho estudantil. Como pode-se observar, a variação do desempenho com um mesmo incremento da renda é mais significativa nos casos de baixa renda do que nos casos de rendas mais altas. Por exemplo, a nota média de estudantes nas faixas de menores rendas cresce cerca de 15 a 20 pontos para um aumento da renda domiciliar mensal da ordem de R$ 500 (a valores de 2019); entretanto, nas faixas de renda mais altas, esse mesmo aumento da renda implica em uma pequeníssima variação da nota, inferior a um ponto, fato que ilustra uma das consequências negativas no processo educacional de uma má distribuição de renda.

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

Como a renda da uma pessoa tende a ser maior quanto maior for sua escolaridade [BALASSIANO, 2005], o desempenho escolar de um estudante hoje terá grande influência em sua renda futura, processo que contribui a reproduzir a atual concentração de renda no país.

Figura 1: Nota média do ENEM 2019 em função da renda familiar mensal.



Por causa da característica não linear da dependência do desempenho com a renda, este trabalho usa uma função logarítmica da renda, como é usual na literatura. Além disso, é importante considerar a renda em termo per capita [Magalhães, 2009; DATTA, 1980], por refletir melhor o padrão de vida: um domicílio com determinada renda, mas com apenas duas pessoas oferece condições de vida bem diferentes de outro com a mesma renda, mas muito mais pessoas.

Dependência com a idade

A dependência das notas dos estudantes com a idade é mostrada na figura 2. Como pode ser observado, essa dependência é mais marcante em Matemática e Redação. Entretanto, em Matemática, parece não ser um efeito direto da idade, pois isso desaparece quando da análise multinível; possivelmente, essa aparente relação aparece por causa da correlação entre as várias variáveis explicativas. Nos grupos mais jovens, a nota média cresce na medida em que a idade aumenta até os 17 anos. Isso pode ser devido ao fato que os grupos de idades menores incluem pessoas que não completariam o ensino médio em 2019, mas não forneceram essa informação (caso tivessem fornecido, seriam excluídos da amostra por serem "treineiros").

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

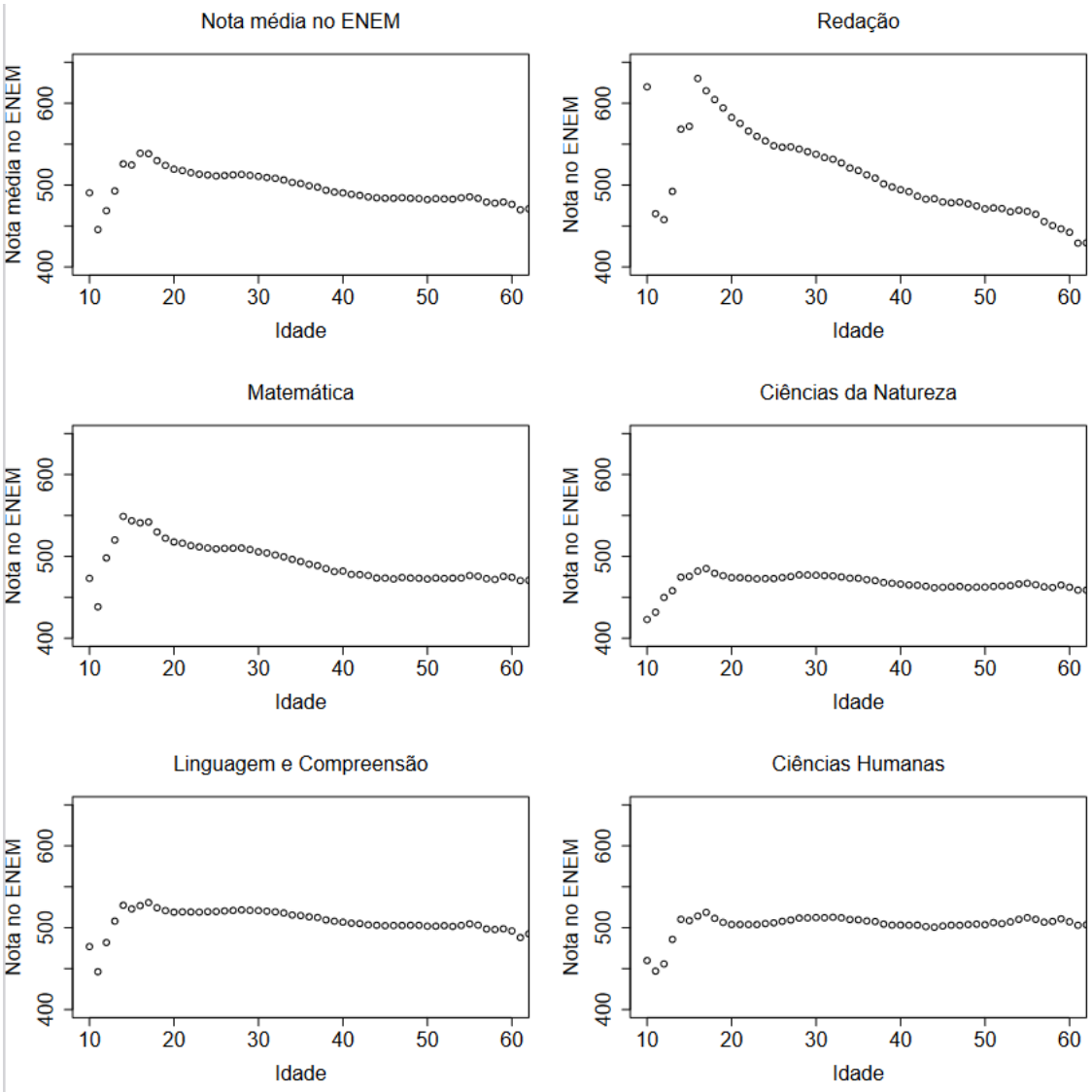
A partir dos 17 anos, a nota tende a se reduzir com o aumento da idade. Esse comportamento também é mais marcante nas provas de Redação e Matemática, indicando a necessidade de se considerar na análise, além das notas médias de todas as provas, as notas nas diferentes disciplinas.

Embora a análise feita não permita testar hipóteses explicativas para a redução das notas com o aumento da idade, isso pode refletir o fato que pessoas que completam o ensino médio com idades mais elevadas possivelmente acumulam uma defasagem idade-série maior, fator que poderia estar relacionado ao desempenho nas provas [SOARES, 2024].

Optou-se por incluir apenas pessoas com 17 a 20 anos, tendo em vista o fato que a partir dos 21 anos há uma mudança na tendência de variação das notas com a idade, com um aparente ressalto próximo aos 30 anos, característica perceptível nos gráficos da figura 2.

Figura 2: Nota média e nas provas específicas do ENEM 2019 em função da idade do candidato.

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047



Relação com a escolaridade dos pais

O número de anos de estudo completado pelos pais foi estimado usando os critérios da tabela 1. Casos em que havia informação de apenas um dos pais não foram incluídos na amostra. Embora a estimativa indicada na Tabela 1 seja aproximada, havendo várias outras formas para obter o mesmo indicador [RIGOTTI; 2013, UNESCO; 2013], ela se mostra adequada e suficientemente precisa para os objetivos deste trabalho.

Tabela 1 – Número de anos de estudo dos pais estimado com base no nível mais elevado de ensino frequentado adotado neste trabalho.

Nível de ensino mais elevado que foi completado (informado pelo estudante)	Anos de estudo considerados
Nunca estudou	0

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

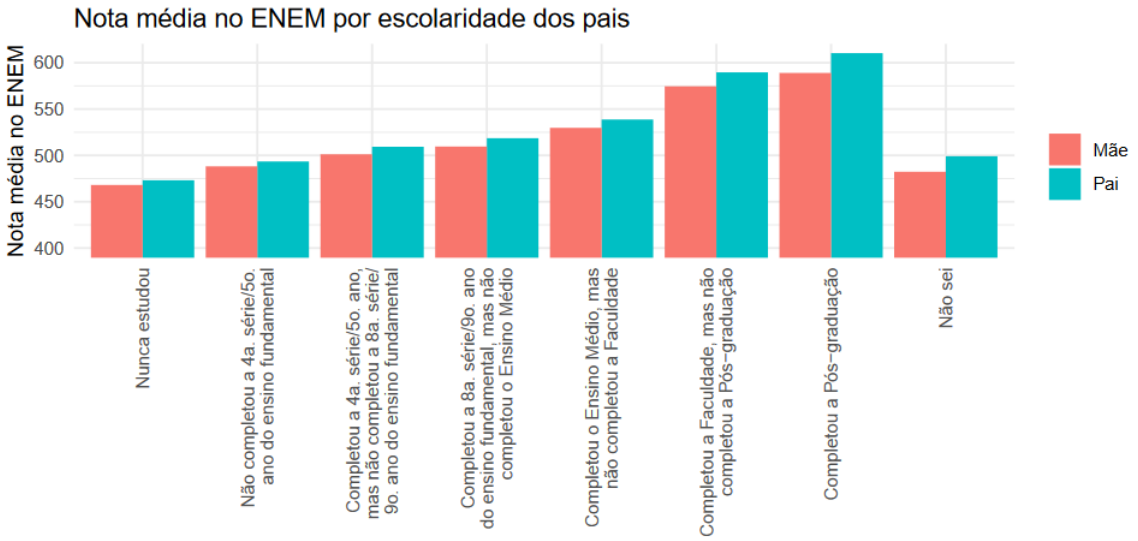
Não completou a 4a. série/5o. ano do ensino fundamental	2
Completou a 4a. série/5o. ano, mas não completou a 8a. série/9o. ano do ensino fundamental	6
Completou a 8a. série/9o. ano do ensino fundamental, mas não completou o ensino médio	9,5
Completou o ensino médio, mas não completou o ensino superior	11
Completou o ensino superior, mas não completou a pós-graduação	15
Completou a pós-graduação	17

A dependência da nota média com o nível escolar dos pais é ilustrada na figura 3. Embora o número de anos de estudo dos pais e mães sejam correlacionados, optou-se em adotar sua soma das duas como variável explicativa. Vale observar que a dependência do desempenho estudantil com a escolaridade dos pais cria uma dificuldade para a mobilidade social, sendo o Brasil o 4º país entre 41 com maior correlação entre a escolarização em diferentes gerações [HERTZ; 2008], ou seja, com menor mobilidade social: filhos de pais pouco escolarizados tendem a ser também pouco escolarizados, processo que contribui para reproduzir no futuro as desigualdades sociais atuais.

A relação do desempenho de estudantes e a escolaridade dos pais [MENDES, 2015], um dos principais capitais culturais, é um aspecto importante que, por afetar negativamente o desempenho estudantil, pode e deve ter seu efeito amenizado por meio de políticas públicas.

Figura 3: Nota média no ENEM 2019 em função da escolaridade dos pais.

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047



Os efeitos dos aspectos familiares no desenvolvimento educacional de crianças e jovens tem sido bastante estudado e as conclusões aqui obtidas, tanto no que diz respeito à renda e à escolarização dos pais, estão de acordo com outros estudos (HARTAS, 2011; ROCHA et al. 2018).

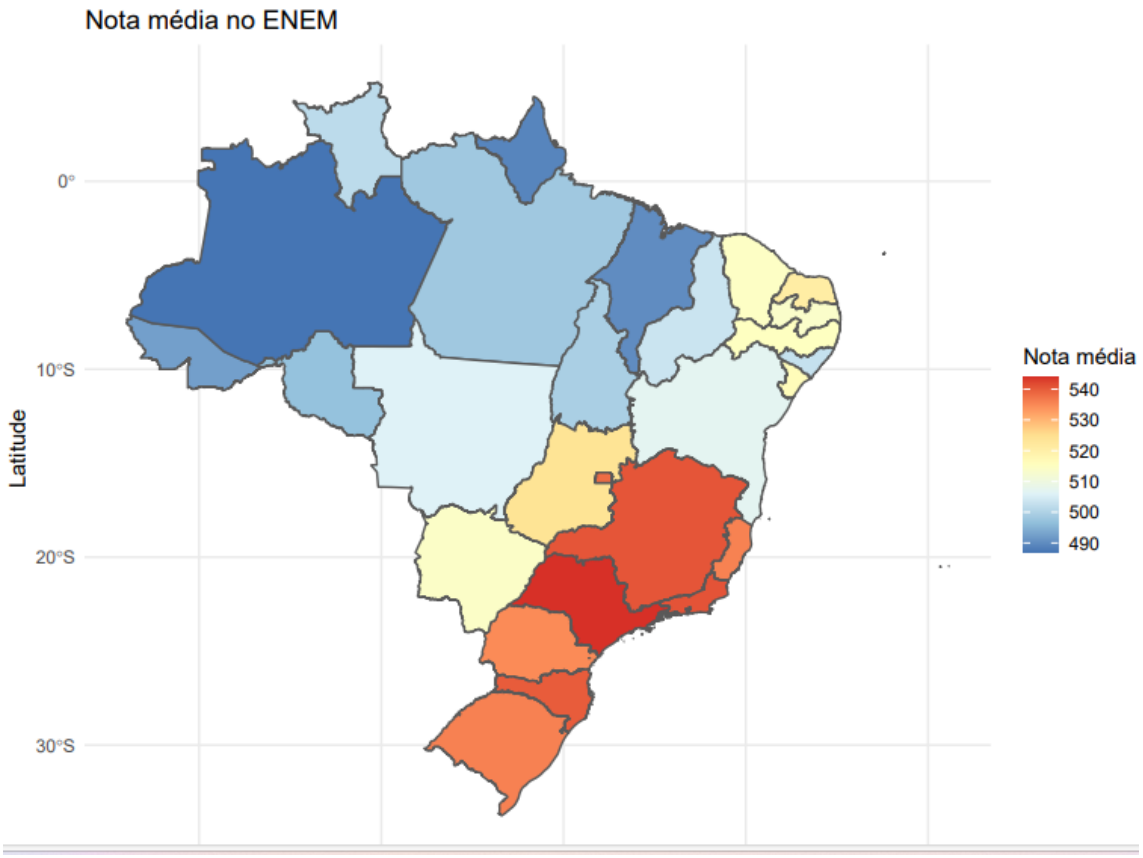
Variação nas diferentes unidades federativas

O mapa da figura 4 mostra que as notas médias do ENEM 2019 variaram significativamente entre os diversos estados, atingindo cerca de 50 pontos entre os casos extremos. Dadas as diferenças educacionais entre os estados [IBGE, 2023], esse fato é esperado e indica ser necessário usar a unidade da federação como uma variável fictícia no modelo.

É importante notar, entretanto, que para entender melhor a característica intrínseca de cada unidade da federação seria necessário realizar uma regressão que considerasse a possibilidade dos efeitos das demais variáveis explicativas serem diferentes nas várias unidades da federação. Esse desacoplamento é necessário uma vez que a distribuição da população pelas diferentes etnias, tipos de escola ou faixa de renda não é a mesma em todo o país.

Figura 4: Distribuição das notas médias do ENEM 2019 por unidade federativa.

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047



Variação das notas com o vínculo administrativo das escolas

A figura 5 mostra a média das notas segundo o vínculo administrativo da escola de ensino médio. (Não foram incluídas escolas rurais.) Essa diferença é significativa em todas as provas, embora seja mais importante nas notas de Matemática e Redação. Como esta variável consta na base de dados apenas para estudantes que concluíram o ensino médio em 2019, a população mostrada na figura 5 se restringe a este grupo específico.

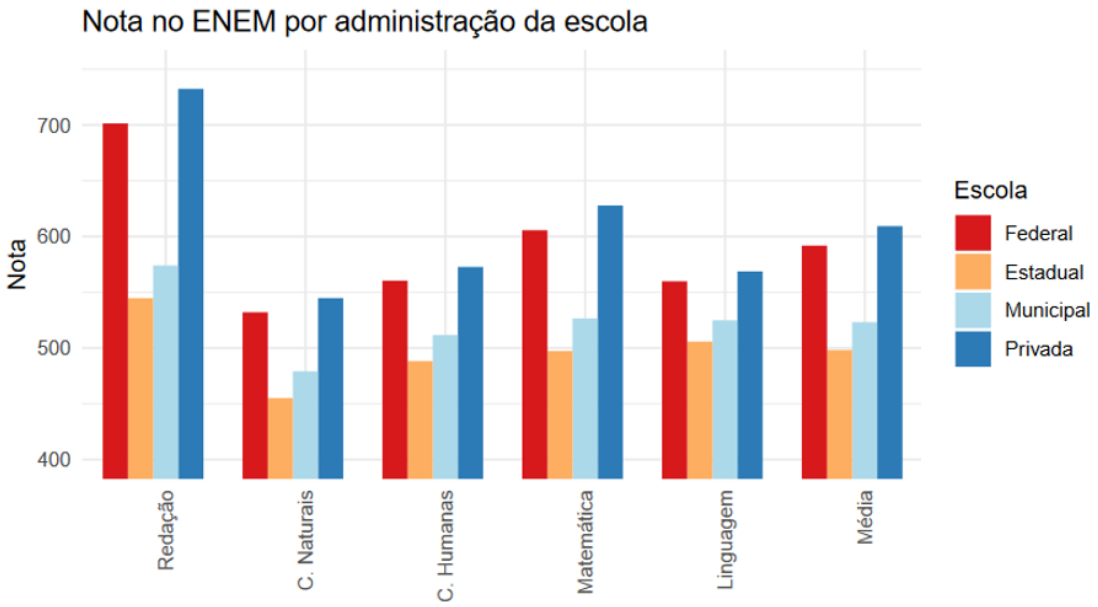
Um aspecto importante da análise incluindo todas as variáveis é que o fato de alunos de escolas privadas terem, em média, desempenho melhor do que seus colegas das instituições federais não indica que elas sejam intrinsecamente melhores. A diferença observada é influenciada pelo perfil dos estudantes. Nas análises completas, veremos que o desempenho das pessoas provenientes das instituições federais é mais elevado do que os das escolas privadas quando as demais condições são mantidas iguais, o inverso do que a figura 5 poderia sugerir. Esse resultado está de acordo com o obtido em análise do desempenho de estudantes oriundos dos Institutos Federais de Educação, Ciência e Tecnologia no ENEM [Souza, 2019] e mostra a importância de se usar uma regressão com as várias variáveis simultaneamente; não fazer isso poderia levar a conclusões erradas.

De forma análoga, a diferença da média dos estudantes das escolas privadas e estaduais, superior a 100 pontos como mostra a figura 5, é reduzida para cerca de 60 pontos (veja seção III). Isso mostra que grande parte da diferença entre esses dois grupos

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

de estudantes não é devida ao vínculo da escola, mas, sim, a questões regionais e a fatores socioeconômicos.

Figura 5: Nota média e das provas específicas do ENEM 2019 segundo o vínculo administrativo da escola onde cursou o ensino médio.



Homens e mulheres

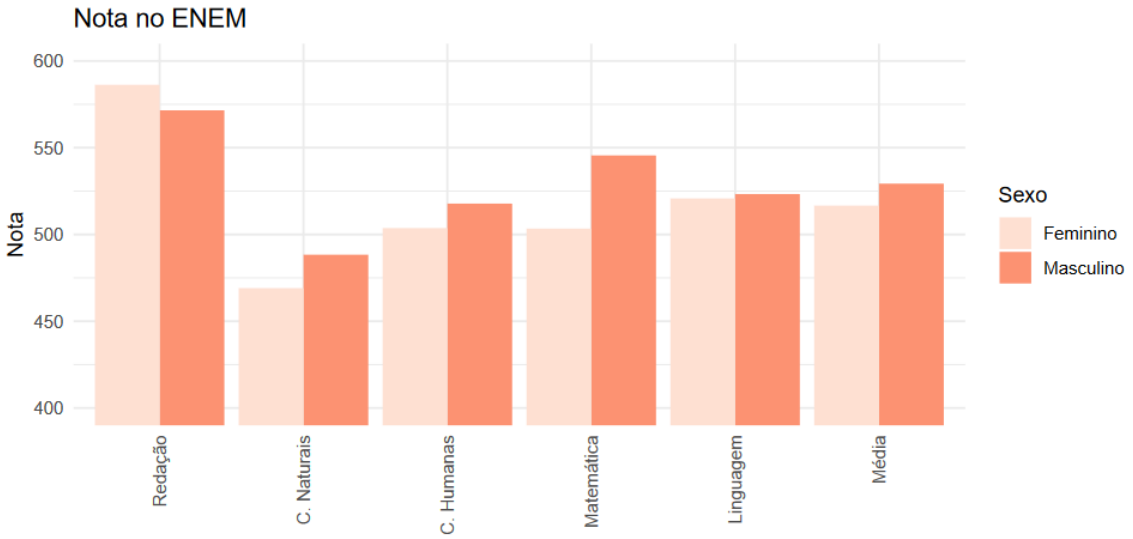
Como mostrado na figura 6, há uma pequena diferença entre as notas médias de homens e mulheres. Entretanto, essa diferença não é a mesma nas diferentes disciplinas. Por exemplo, em Redação as mulheres têm um desempenho significativamente melhor do que os homens; nas demais disciplinas, essa tendência é invertida, com homens tendo notas maiores do que mulheres.

A análise das notas médias mostrou uma vantagem de 11 pontos na média para candidatos do sexo masculino. Entretanto, verificando as notas de provas específicas, observamos que os homens obtiveram uma nota 42 pontos maior em Matemática, sendo a nota de Redação 19 pontos menor do que a das mulheres. Esse fato indica que para um entendimento mais amplo dos efeitos das demais variáveis explicativas é fundamental analisar não apenas as médias, mas, também, as notas nas diferentes provas.

Vale observar que a diferença de notas entre homens e mulheres nas várias disciplinas também existe em outros países. Uma análise dos resultados do PISA mostra que, com poucas exceções, homens apresentam melhores desempenhos em Matemática e mulheres em leitura. [GUISO, 2008], concordando com os resultados aqui obtidos.

Figura 6: Nota média e das provas específicas do ENEM 2019 em função do sexo.

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047



Assim como nos demais casos, essas diferenças de notas, quando analisadas em conjunto com as demais variáveis, são diferentes, pois parte delas pode estar relacionada à idade, renda etc. Quando a análise mais geral é feita, na seção III, vemos, por exemplo, que a diferença da nota em Matemática é de da ordem de 33 pontos, significativamente diferente dos 41,7 quando a análise é feita não considerando o efeito das demais variáveis. Por outro lado, a diferença de 18,7 pontos a favor das mulheres no caso da nota de Redação é ampliada para quase 31 pontos na análise mais geral. Essa observação, como já feita em outros casos, mostra a importância de uma análise que inclua um nível de desagregação que permita avaliar o efeito individual de cada variável. Isso pode ser feito usando-se o método dos mínimos quadrados, adotado neste trabalho, bem como uma análise multinível. A equivalência dos dois procedimentos quando há uma grande quantidade de dados, como é o caso, é discutida no Apêndice C.

Diferença segundo a cor da pele/etnia

Outro fator relacionado ao desempenho no ENEM é a etnia ou cor da pele, como mostrado na figura 7. Essas diferenças têm sido observadas por diversos autores [veja, p. ex., SOARES, 2006 e referências lá citadas] e as conclusões não discordam das aqui obtidas.

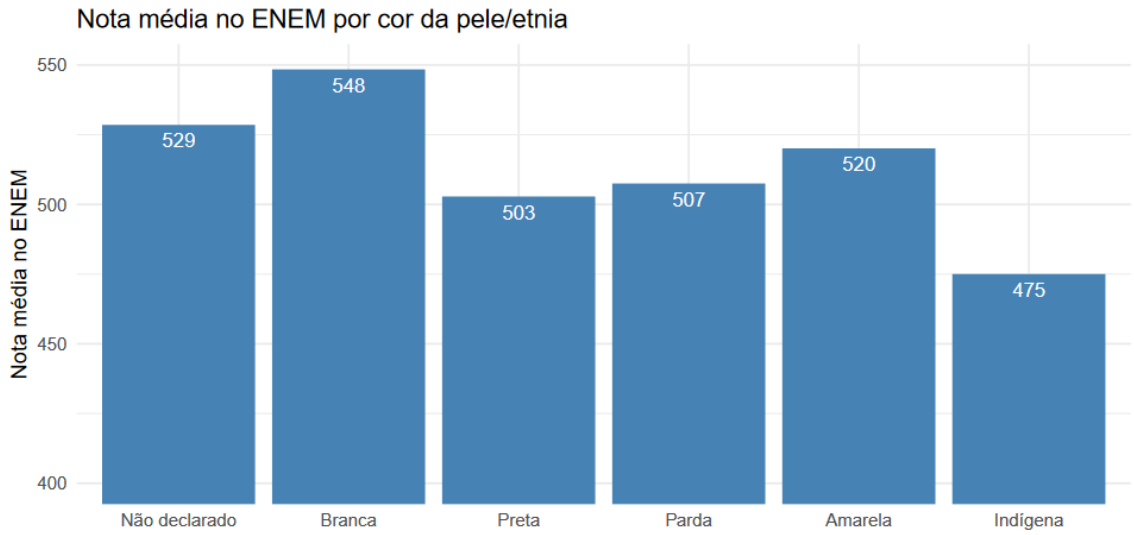
Embora a diferença entre a média das pessoas de cor branca e de cor preta seja de 45 pontos, como pode ser observado na figura 7, quando as demais variáveis são incluídas, ou seja, a comparação é feita considerando-se pessoas com todas as demais características iguais, ela é reduzida para pouco mais do que 12 pontos (ver tabela 3). Esse é mais um exemplo para justificar a importância de uma análise simultânea com todas as variáveis explicativas.

Como ainda resta uma diferença de 12 pontos entre as médias de pessoas brancas e pretas, podemos supor que há outras variáveis importantes – como, por exemplo, a distribuição das diferentes etnias pelas unidades da federação, o tipo e a qualidade de

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

escolarização dos pais, além dos anos completos de estudo, entre outras variáveis não incluídas nesta análise.

Figura 7: Nota média e das provas específicas do ENEM 2019 segundo a cor/etnia.



MODELO ADOTADO E AJUSTE DA FUNÇÃO

É importante observar, como já afirmado, que a relação entre a variável dependente (no caso, as notas nas provas do ENEM) com uma das variáveis independentes pode ser devida ao fato que cada variável independente pode estar correlacionada com outras variáveis também independentes. Sendo mais explícito, a diferença de nota entre estudantes de escolas privadas e escolas federais pode ser afetada pelo diferente perfil socioeconômico desses dois conjuntos de estudantes e não apenas pelo vínculo administrativo da escola, hipótese que os resultados confirmam. Consequentemente, a dependência do desempenho nas provas do ENEM com as variáveis explicativas não deve ser examinada caso a caso, mas, sim, em conjunto. Além disso, para levar em conta corretamente as diferenças de perfil socioeconômico ou de cor / etnia entre unidades da federação, é necessário fazer uma análise multinível incluindo variáveis representando as unidades da federação no modelo.

Com esse propósito, a nota yi do estudante identificado pelo índice i foi relacionada com as variáveis independentes segundo a equação

$$y_i = \beta_1 \cdot log(rdpc_i) + \beta_2 \cdot I_i + \beta_3 \cdot EscPais_i + \beta_4 \cdot Estado_i + \beta_5 \cdot Vinc_i + \beta_6 \cdot S_i + \beta_7 \cdot Cor_i + e_i \quad (1) .$$

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

Nessa equação, o índice i identifica a pessoa, yi sua nota, Ii sua idade, $EscPais_i$ a escolaridade dos pais, $Estado_i$ a unidade da federação onde a prova foi feita, Vinci o vínculo administrativo da escola na qual foi cursado o ensino médio, Si o sexo e Cori a cor/etnia. Os parâmetros ajustados usando-se o método dos mínimos quadrados, são identificados por β_j.

As variáveis Estado, vínculo administrativo, sexo e cor/etnia foram trabalhadas na forma de variáveis binárias. O fator e_i é o erro, por suposição com a propriedade que $\langle e_i \rangle = 0$, onde $\langle \quad \rangle$ indica valor esperado da grandeza que está entre as chaves. Para estimar as incertezas nos parâmetros ajustados, foi suposto, como já dito, que $\langle e_i^2 \rangle$ assume o mesmo valor para qualquer valor de i. Note-se que o uso do método dos mínimos quadrados ordinário é idêntico à regressão linear comumente adotada por pesquisadores nas áreas de ciências humanas e disponíveis em pacotes fechados de softwares. Tais pacotes estimam os parâmetros dos modelos de forma relativamente opaca que, no caso de pacotes de modelagem multinível, resultam em parâmetros que requerem experiência para interpretar corretamente.

Tabela 2 – Variáveis explicativas incluídas na análise.

Nome da variável	Descrição da variável
rdpc	Renda domiciliar per capita
I	Idade
EscPais	Soma das escolaridades do pai e da mãe
Estado	Unidade da federação onde foram realizadas as provas
Vinc	Vínculo administrativo da escola de ensino médio
S	Sexo
Cor/etnia	Cor

Para evitar o aparecimento de matrizes singulares, nos casos das variáveis fictícia (vínculo administrativo da escola, sexo e cor/etnia) adotou-se, respectivamente, as seguintes características como referência: municipal, mulher e indígena. Note que a análise dos resultados obtidos não depende dos valores absolutos dos parâmetros, mas, sim, das diferenças entre eles. Portanto, é irrelevante quais variáveis são usadas como referência.

Os valores indicados na tabela 3 são, portanto, as diferenças de notas em relação a estas características referenciais. Por exemplo, estudantes de escolas estaduais tiveram uma média 17,6 pontos abaixo da média de seus colegas de escolas municipais; homens tiveram uma nota média em Matemática 32,75 pontos acima da nota média das mulheres e 30,61 pontos abaixo em Redação.

RESULTADOS

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

A tabela 3 mostra os resultados dos parâmetros ajustados nos casos em que as variáveis dependentes foram as médias e as notas nas provas de Matemática e de Redação (bem como os respectivos desvios padrões, entre parênteses). Os resultados para Ciências Naturais, Ciências Humanas e Linguagem aparecem no Apêndice A. Os desvios padrões dos parâmetros foram estimados na forma descrita no Apêndice B.

Embora algumas entidades optem por escolher pesos diferentes para as diferentes provas no cálculo da nota a ser usada em um processo seletivo, optamos pela média simples, tanto por ser comumente adotada como pelo fato que selecionar estudantes com pesos diferentes pouco altera a lista dos incluídos, como discutido na seção Conclusões.

Tabela 3 – Resultado dos parâmetros ajustados e respectivos desvios padrões.

	Nota média		Nota de Matemática		Nota de Redação	
Variável	Valor ajustado	Desvio padrão	Valor ajustado	Desvio padrão	Valor ajustado	Desvio padrão
Rdpc	22,91	0,13	29,45	0,17	34,19	0,32
Idade	-15,12	0,11	-15,03	0,15	-32,84	0,27
EscPais	1,369	0,013	1,335	0,017	2,327	0,03
RO	-6,2	2,7	-7,0	3,5	-5,3	6,4
AC	-5,8	2,8	-14,2	3,7	4,6	6,7
AM	-16,0	2,6	-16,7	3,4	-35,1	6,2
RR	-19,5	3,1	-17,8	4,1	-55,8	7,5
PA	-2,8	2,5	-12,9	3,4	17,3	6,2
AP	-15,0	2,7	-26,5	3,6	-11,4	6,7
TO	-9,4	2,7	-11,2	3,5	-4,2	6,4
MA	-12,8	2,5	-13,2	3,4	-14,0	6,2
PI	6,9	2,6	6,0	3,4	33,5	6,2
CE	10,2	2,5	18,4	3,4	18,1	6,1
RN	11,5	2,6	9,4	3,4	30,8	6,3
PB	8,6	2,6	5,4	3,4	31,6	6,2
PE	4,6	2,5	11,8	3,4	4,6	6,1
AL	-6,9	2,6	-6,3	3,4	-1,9	6,3
SE	14,7	2,6	7,9	3,5	52,6	6,4
BA	4,2	2,5	3,2	3,4	9,4	6,2
MG	18,5	2,5	27,3	3,4	23,3	6,1
ES	15,6	2,6	22,5	3,4	24,8	6,3
RJ	9,4	2,6	8,3	3,4	8,3	6,2
SP	-1,3	2,5	5,6	3,3	-31,8	6,1
PR	-5,8	2,5	0,1	3,4	-41,2	6,1

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

SC	-2,1	2,6	3,5	3,4	-25,0	6,2
RS	10,0	2,6	14,6	3,4	3,8	6,2
MS	-4,5	2,6	-5,9	3,4	-12,8	6,3
MT	-17,7	2,6	-20,7	3,4	-29,7	6,3
GO	6,0	2,6	6,3	3,4	11,2	6,2
DF	4,4	2,6	0,8	3,4	-5,3	6,3
Privada	43,8	0,9	52,2	1,2	85,7	2,1
Estadual	-17,6	0,8	-19,9	1,1	-25,1	2,1
Federal	61,2	0,9	72,3	1,2	105,8	2,2
Homem	3,64	0,16	32,75	0,21	-30,61	0,38
Branca	24,6	1,1	23,6	1,4	41,2	2,6
Preta	12,1	1,1	2,9	1,4	25,2	2,6
Parda	13,7	1,0	10,5	1,4	24,9	2,5
Amarela	19,9	1,2	23,3	1,5	29,6	2,8

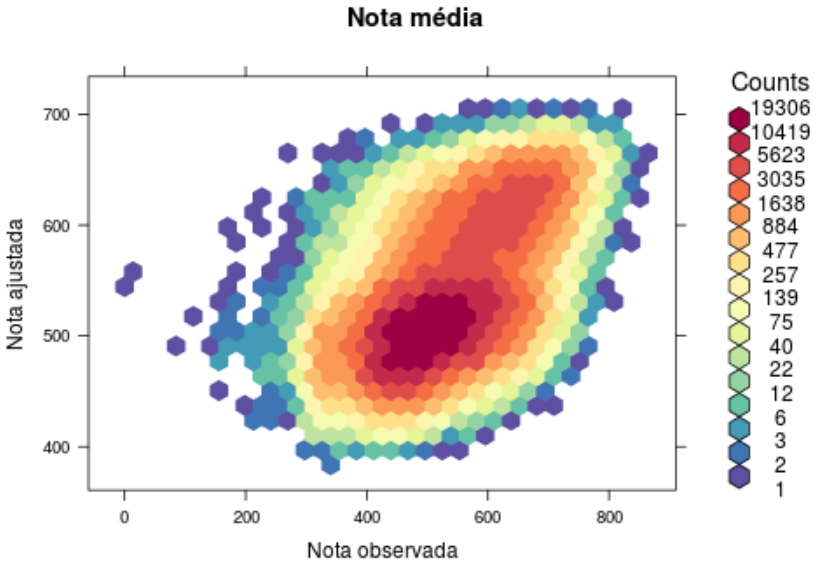
Os resultados obtidos são praticamente idênticos àqueles obtidos usando-se um procedimento de análise multinível utilizando-se os pacotes usuais de software disponíveis, como mostrado no Apêndice C. Isso confirma que os valores ajustados dos parâmetros são equivalentes aos obtidos com uma análise multinível, fato esperado considerando a grande quantidade de dados usados na análise.

ANÁLISE DOS RESULTADOS

Todas as variáveis independentes consideradas neste trabalho mostraram-se significativas, fato evidenciado pelos pequenos valores estimados dos desvios padrões em relação aos valores ajustados dos parâmetros. Não há dúvidas, portanto, quanto à relevância de cada uma delas. Entretanto, como será discutido mais adiante, elas não são suficientes para explicar a totalidade das notas.

Figura 8 – Comparação entre as notas médias observadas e as notas calculadas em função das variáveis consideradas.

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047



A nota média de cada estudante foi calculada com base nos seus indicadores correspondentes às variáveis independentes e usando-se os parâmetros ajustados na equação (1). As notas observadas se aproximam daqueles valores calculados usando-se o modelo, como é mostrado na figura 8. O coeficiente de correlação entre as notas observadas e as calculadas pelo modelo adotado é 0,62. No que segue, vamos analisar alguns resultados obtidos.

Unidades da federação

Os parâmetros correspondentes aos interceptos estaduais, tabela 4, indicam fatores que podem afetar o desempenho estudantil e que não são explicados pelas demais variáveis consideradas. Ou seja, dadas as mesmas condições e características pessoais dos estudantes e vínculos administrativos da escola, há fatores intrínsecos aos estados cujo efeito não é detectado pelo modelo.

Para facilitar a interpretação dos coeficientes dos estados, apresentamos as diferenças entre os interceptos ajustados e as médias nacionais, 582,9, 533,0 e 867,3 para a médias, matemática e na redação, respectivamente"

Vale observar que usar um intercepto único para todas as unidades da federação – ou seja, um intercepto para o Brasil – poderia esconder características específicas daquelas unidades.

"

Tabela 4 – Estados com os valores mais baixos e mais altos do parâmetro β4 (relacionado às unidades da federação) nas notas médias e de Matemática e Redação.

Média		Matemática		Redação		
RR	-20	AP	27	RR	-56	

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

MT	-18	MT	-21	PR	-41	
AM	-16	RR	-18	AM	-35	
SE	15	CE	18	PB	32	
ES	16	ES	23	PI	34	
MG	19	MG	27	SE	53	

Os estados de Roraima, especialmente, mas também, Mato Grosso e Amazônia apresentam características intrínsecas, indicadas pelos valores dos interceptos, bastante inferiores às dos demais estados. Por outro lado, Minas Gerais, Sergipe e Espírito Santo apresentam características intrínsecas mais elevadas. Essas observações sugerem a necessidade de um estudo mais profundo para entender as razões dessas diferenças. É possível que elas tenham origem em aspectos históricos relacionados aos sistemas educacionais, a orçamentos por aluno, especialmente nas escolas públicas, à renda per capita estadual, e aos investimentos por estudante nas redes estaduais, onde está a maioria deles e ao índice de Gini, cujo efeito nos resultados educacionais já foi observado [Helene, 2020].

Vínculo administrativo da escola

O efeito do vínculo administrativo das escolas, inclusive no ENEM, tem sido investigado [MORAES, 2022]. Como no caso das demais variáveis explicativas, é necessário fazer uma análise do tipo multinível, para desacoplar o efeito intrínseco do vínculo das demais variáveis explicativas. As diferenças negativas das escolas estaduais na média das provas, 17,6 pontos abaixo da pontuação média das escolas municipais (que foram tomadas como referência), é ainda mais grave quando analisados os desempenhos em Matemática e Redação, 19,9 e 25,1 pontos, respectivamente. Isso indica que as demais provas, Ciências Naturais e Humanas e Linguagem, tendem a reduzir as deficiências de desempenho dos estudantes provenientes das escolas estaduais. Essa diferença pode ser devida aos investimentos por aluno nas escolas estaduais, investimentos esses correspondentes majoritariamente à remuneração de docentes e demais trabalhadores e despesas correntes, de capital e de manutenção.

As maiores diferenças referentes ao vínculo administrativo ocorrem entre estudantes das escolas federais e das escolas estaduais, da ordem de 80 e 90 pontos aproximadamente na média e em Matemática, respectivamente, e cerca de 130 pontos em Redação. Seria interessante um estudo detalhado para descobrir quais fatores provocam esse efeito, sendo os processos seletivos para ingresso nas instituições federais, orçamentos por aluno, condições salariais e existência de carreira docente bem definida [Souza, 2019; Madeira, 2024] possíveis fatores explicativos.

Etnia/cor da pele

A diferença de desempenho associado à cor da pele/etnia na média é uma combinação das grandes diferenças das notas em Matemática e Redação, atenuada pelas

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

menores diferenças nas notas nas demais disciplinas. (As diferenças entre brancos e pretos, de cerca de 12 pontos na média geral, sobe para 16 pontos em Redação e cerca de 21 pontos em Matemática.) Esse fato mostra que as variáveis consideradas na análise não contemplam todos os aspectos importantes. É possível, por exemplo, que as notas de estudantes de escolas com um mesmo tipo de vínculo administrativo variem segundo suas localizações nas cidades – escolas periféricas e centrais, por exemplo: se a distribuição dos estudantes de diferentes etnias depender também da localização das escolas frequentadas, o efeito detectado pode não ter nenhuma relação com a etnia. Além disso, vieses criados por alguns dos filtros – como, por exemplo, não saber a renda familiar ou a escolaridade de um dos pais – podem ser responsáveis por parte das notas não serem explicadas pelas variáveis consideradas. O tipo e a qualidade da educação dos pais, além da soma simples do número de anos de frequência escolar, pode ser outro fator importante na definição do desempenho escolar dos filhos.

Idade

As notas tendem a se reduzir com a idade após os 17 anos cerca de 15 pontos a cada ano adicional, tanto na média como em Matemática. Essa piora do desempenho é mais marcante em Redação, com redução de 32 pontos para cada ano adicional de idade. Vale observar que essa tendência é muito menos significativa nas provas de Ciências Naturais e Humanas e Linguagem do que nas provas de Matemática e Redação, como mostrado no Apêndice A (reduções de 8, 10 e 10 pontos a menos para cada ano, respectivamente). Esse resultado ilustra o efeito negativo da defasagem idade-série [SOARES, 2024].

Escolarização dos pais

O efeito da escolarização dos pais, parâmetro β3 na equação (1) é bastante significativo. Oito anos a mais de escolaridade somada (como, p. ex., aquela correspondente a ambos os pais terem concluído o ensino médio versus terem curso superior em nível de graduação) pode levar a uma diferença de 11, 11 e 19 pontos na nota média e nas notas em Matemática e em Redação, respectivamente. As diferenças em Matemática e Redação são ligeiramente maiores do que as observadas nas demais provas.

Renda

Um pequeno aumento de renda nos grupos mais desfavorecidos leva a um ganho educacional bastante significativo, o que não ocorre nos grupos mais ricos, casos em que aumento de renda não em nenhum efeito significativo. Isso sugere que uma melhor distribuição de renda bem como instrumentos de transferência de renda como "Bolsa Família" e o "Pé-de-Meia", federais e programas estaduais equivalentes podem ter impacto importante no desempenho estudantil

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

DISCUSSÃO E CONCLUSÃO

Este trabalho mostra que quando há muitos dados o método de mínimos quadrados ordinário (MQO) é suficiente, não havendo necessidade do uso de pacotes fechados para a aplicação de métodos hierárquicos. Por ser mais fácil interpretar e controlar os parâmetros, o MQO substitui, com vantagens, os pacotes prontos que usam métodos multiníveis.

O ENEM é um exame nacional cujo resultado tem sido usado nos processos seletivos para o ingresso em cursos superiores, inclusive em instituições públicas. Uma consequência prática de um resultado neste exame é extremamente importante para definir o futuro de uma pessoa jovem. Assim, é interessante avaliar as consequências de uma diferença de nota no ENEM.

Uma nota de corte típica para cursos relativamente disputados é 700 pontos e apenas cerca de 3% dos estudantes atingem esse valor. Entretanto, perto de 10% atingem a nota 650, o que mostra que uma diferença de 50 pontos é extremamente relevante. Ora, 50 pontos é menos do que a diferença da nota média entre estudantes das escolas estaduais e seus colegas das escolas federais e privadas de uma mesma unidade da federação e com todos os demais indicadores iguais. (Apenas 1,5% dos estudantes das escolas estaduais obtêm nota igual ou superior a 700 pontos e quase a totalidade deles matriculados em escolas com algum diferencial, como escala de aplicação ou escola técnica.) Esse resultado mostra o efeito discriminatório de um sistema escolar privatizado de forma mercantil como o brasileiro, que concentra seu potencial educacional em escolas que atendem os grupos mais favorecidos.

Muitas instituições, na seleção de seus estudantes, costumam calcular a média das notas do ENEM dando peso maior para uma ou algumas das provas. Entretanto, a mudança dos aprovados usando pesos diferentes é, talvez, irrelevante. Por exemplo, vamos comparar os 5% com melhores notas quando a prova de matemática recebe peso 2 e as demais, peso um, com os 5% com melhores notas quando todas as provas têm o mesmo peso. Nesse caso, apenas um em cada cerca de 15 dos excluídos pelo primeiro critério seriam incluídos pelo segundo, sendo as diferenças entre as notas dos excluídos e dos últimos incluídos praticamente irrelevante.

Todas as variáveis explicativas consideradas neste trabalho se mostraram estatisticamente significativas, explicando grande parte dos resultados observados nas diversas provas do Exame Nacional de Ensino Médio. Entretanto, vários aspectos não são explicados por elas.

As causas das diferenças no desempenho de estudantes dos diferentes tipos de vínculo administrativo das escolas é uma das características que merecem ser estudadas. Como sugestão de possíveis variáveis não consideradas e que podem explicar essa diferença estão os orçamentos por estudante, orçamento este correspondente às despesas com professores e demais trabalhadores, manutenção predial, despesas de consumo etc., ou seja, despesas diretamente relacionadas às atividades estudantis. Tais despesas são

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

reconhecidamente ligadas às condições de funcionamento das escolas, de trabalho dos docentes e de condições de estudo dos alunos [Souza, 2019; Madeira, 2024]. Inclusive, as condições de docência são reconhecidamente importantes no desempenho dos estudantes no ENEM [Feijó, 2021]. A variável "orçamento por estudante" pode, também, estar relacionada à variação das notas com a unidade da federação onde a prova foi realizada.

A variável idade está relacionada à defasagem idade-série. A inclusão de uma variável que permitisse separar o efeito desse fator pode ser esclarecedora e entender as consequências práticas da defasagem.

O efeito discriminatório das provas de Matemática e Redação é bem maior do que o mesmo efeito nas demais provas, fato perceptível quando comparamos os parâmetros relacionados à renda per capita domiciliar e ao tipo de escola frequentada. Políticas públicas que pretendam enfrentar essa questão da discriminação socioeconômica sem atuar sobre suas causas podem considerar a utilização de uma ponderação das notas para o cálculo da média final que preserve o objetivo seletivo, mas reduza os efeitos das desigualdades econômicas.

Outra variável não considerada e que poderia explicar parte das diferenças de notas é a renda per capita das diferentes regiões do país. Por exemplo, uma mesma renda domiciliar per capita pode ter efeitos diferentes nas diferentes unidades da federação. Tal análise pode ser feita incluindo-se entre as variáveis explicativas o PIB per capita das unidades da federação.

A desagregação dos dados pelos municípios também poderia contribuir para melhorar o modelo. A localização dos estudantes dentro do município, em especial no caso das grandes cidades, também poderia melhorar os resultados obtidos e revelar informações importantes.

Como já observado, uma diferença aparentemente pequena na nota de uma prova usada como critério para seleção pode ter uma consequência muito grande no efeito de exclusão. Assim, as aparentemente pequenas diferenças de notas segundo as variáveis explicativas podem ter, quando adicionadas, grandes consequências no poder de exclusão.

Como conclusão final, podemos afirmar que os resultados aqui apresentados, uma vez definido o objetivo do sistema educacional na construção do país, permitem estabelecer e otimizar as políticas públicas que melhor respondam àquele objetivo.

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

APÊNDICE A

Valores dos parâmetros ajustados para as notas nas provas de Ciência Naturais, Ciências Humanas e Linguagem.

	Ciências Naturais		Ciências Humanas		Linguagem	
Variável	Valor ajustado	Desv. pad.	Valor ajustado	Desv. pad.	Valor ajustado	Desv. pad.
Rdpc	17,79	0,12	18,78	0,13	14,35	0,10
Idade	-7,73	0,10	-10,49	0,11	-9,53	0,09
EscPais	0,99	0,01	1,11	0,01	1,09	0,01
RO	-5,8	2,5	-7,1	2,7	-5,6	2,1
AC	-6,7	2,6	-7,1	2,8	-5,4	2,2
AM	-10,9	2,4	-10,0	2,6	-6,9	2,0
RR	-8,7	2,9	-7,1	3,1	-7,7	2,4
PA	-4,0	2,4	-5,0	2,6	-9,3	2,0
AP	-13,8	2,6	-11,2	2,8	-12,1	2,2
TO	-7,9	2,5	-13,2	2,7	-10,3	2,1
MA	-10,1	2,4	-13,9	2,6	-12,5	2,0
PI	1,4	2,4	-2,5	2,6	-3,5	2,0
CE	6,8	2,4	5,7	2,6	2,4	2,0
RN	6,5	2,4	4,8	2,6	6,2	2,1
PB	2,9	2,4	2,6	2,6	0,6	2,0
PE	3,6	2,4	1,6	2,6	1,5	2,0
AL	-8,9	2,4	-9,2	2,6	-8,1	2,0
SE	5,5	2,5	4,6	2,7	3,2	2,1
BA	3,2	2,4	3,2	2,6	2,3	2,0
MG	16,2	2,4	14,4	2,6	11,6	2,0
ES	12,3	2,4	11,2	2,6	7,4	2,0
RJ	7,6	2,4	10,5	2,6	12,3	2,0
SP	3,9	2,3	6,7	2,5	9,4	2,0
PR	2,7	2,4	5,3	2,6	4,4	2,0
SC	1,2	2,4	6,0	2,6	4,2	2,0
RS	4,9	2,4	15,0	2,6	12,2	2,0
MS	-0,9	2,4	-2,6	2,6	-0,1	2,1
MT	-11,7	2,4	-13,3	2,6	-13,2	2,0
GO	4,7	2,4	3,5	2,6	4,3	2,0
DF	6,6	2,4	7,7	2,6	12,3	2,1
Privada	35,5	0,8	28,7	0,9	17,0	0,7
Estadual	-17,2	0,8	-14,9	0,9	-11,1	0,7
Federal	48,6	0,9	45,8	0,9	33,7	0,7

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

Homem	13,9	0,15	5,67	0,16	-3,5	0,12
Branca	18,2	1,0	22,4	1,1	17,8	0,8
Preta	7,6	1,0	13,6	1,1	11,2	0,9
Parda	9,6	1,0	13,1	1,1	10,7	0,8
Amarela	15,8	1,1	16,7	1,2	14,2	0,9

APÊNDICE B

O desvio padrão de cada dado, σ, foi estimado pela equação

$$\sigma^2 = \frac{1}{N-n}\sum\left(y_i - y_{ajustado,i}\right)^2 \quad , \quad \text{(B1)}$$

onde *N* é a quantidade de dados usados na análise, 757.712, e *n* é a quantidade de parâmetros ajustados, no caso, 38.

A partir da estimativa do desvio padrão de cada dado, o método dos mínimos quadrados permite o cálculo dos desvios padrões dos dados pela equação [HELENE; 2012]

$$V_{parâmetros} = \sigma^2 \cdot [X^t \cdot X]^{-1} \quad . \quad \text{(B2)}$$

Os elementos da matriz **X** são dados por

$$X_{i,j} = \frac{\partial y_i}{\partial \beta_j} \quad , \text{(B3)}$$

sendo y_i dado pela equação (1).

APÊNDICE C – MÉTODO MULTINÍVEL

No modelo adotado neste trabalho, deixamos o intercepto variar por estado, o que permite levar em conta, e analisar, a variação do desempenho dos estudantes no ENEM entre unidades da federação quando todas as demais variáveis explicativas são iguais. Este é um exemplo particular de uma abordagem geral chamada "modelagem multinível" (GELMAN; HILL, 2007; GOLDSTEIN, 2011; SNIJDERS; BOSKER, 2012), que permite levar em conta o fato que indivíduos pertencem a grupos diferentes (estados, no nosso caso). A intuição que justifica esta abordagem é que há uma possibilidade que indivíduos que pertencem ao mesmo grupo são similares e que é proveitoso levar esta possibilidade em conta no modelo, tanto para fins de melhorar as predições para indivíduos, como para descobrir características e variação entre os grupos.

Na abordagem adotada neste trabalho, usando regressão tradicional, os coeficientes que variam por grupo (os parâmetros indicadores ou "dummies" dos estados, ou seja, os interceptos) são considerados fixos e desconhecidos, a serem estimados a partir dos dados. É esperado que essa abordagem funcione bem quando o número de indivíduos em cada grupo é suficientemente grande. Pela quantidade de dados do ENEM, mesmo para os estados com menos candidatos, não é esperada nenhuma diferença ao usar modelagem multinível para o caso do nosso modelo. Fizemos a regressão com uma

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

biblioteca popular (JOLLY, 2018) para modelagem multinível e, como esperado, os valores dos coeficientes foram iguais, dentro do erro padrão, aos valores obtidos com o método dos mínimos quadrados, como mostrado na Tabela C1. (Aqui, diferentemente do que foi mostrado na Tabela 3, aparecem os parâmetros ajustados sem a subtração da média nacional.)

Tabela C1 – Parâmetros correspondentes às unidades da federação: comparação entre os resultados obtidos pelo método dos mínimos quadrados ordinário com os obtidos adotando-se um procedimento multinível padrão (JOLLY, 2018).

	Multinível	MMQ	Diferença		Multinível	MMQ	Diferença
AC	599,6	599,57	0,03	PB	613,94	614,00	-0,05
AL	598,48	598,48	-0,01	PE	609,95	609,98	-0,04
AM	589,44	589,43	0,01	PI	612,26	612,31	-0,05
AP	590,48	590,36	0,12	PR	599,62	599,65	-0,03
BA	609,57	609,6	-0,04	RJ	614,73	614,77	-0,04
CE	615,6	615,64	-0,04	RN	616,82	616,89	-0,07
DF	609,76	609,81	-0,05	RO	599,2	599,21	0
ES	620,95	621,01	-0,07	RR	586,42	585,93	0,49
GO	611,32	611,36	-0,04	RS	615,4	615,45	-0,05
MA	592,61	592,62	-0,01	SC	603,3	603,34	-0,03
MG	623,87	623,91	-0,04	SE	620,02	620,12	-0,1
MS	600,89	600,9	-0,02	SP	604,09	604,12	-0,03
MT	587,67	587,65	0,02	TO	596,03	596	0,03
PA	602,54	602,57	-0,03				

Na abordagem multinível mais geral, os próprios coeficientes são modelados como vindo de alguma distribuição, por exemplo uma distribuição normal, com média e desvio padrão que por sua vez são estimados pelos dados (Gelman; Hill, 2007, p1). A modelagem multinível é apropriada quando, por exemplo, desejamos levar em conta a variação entre escolas. Neste caso teríamos que incluir dezenas de milhares de coeficientes, algo difícil de fazer por meio de indicadores. Na abordagem tradicional também é difícil incluir variáveis explicativas em nível de estado (por exemplo uma variável "conservador / progressivo" para cada estado). Não seria possível fazer isso por meio de indicadores, porque já há indicadores dos próprios estados. Seria possível fazer a regressão das médias

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

dos 27 estados contra variáveis explicativas características dos estados. A modelagem multinível levaria em conta os dados nos dois níveis simultaneamente, além de tratar melhor casos em que há poucos indivíduos em alguns dos grupos.

Outro caso onde a modelagem multinível mais geral é apropriada é quando há muitos coeficientes a serem estimados no modelo. Por exemplo, podemos estar interessados em como os coeficientes da renda domiciliar por capita, ou a diferença entre homens e mulheres variam entre estados e por tipo de escola (particular, municipal, federal ou estadual). Isso levaria a 27×4 coeficientes (inclinações no caso da renda, ou diferenças entre homens e mulheres), um para cada combinação de estado / tipo de escola. Como já mencionado, a situação fica mais complexa ainda se usamos as escolas como grupo ao invés de estados. Nestas situações onde em alguns grupos há poucos dados e o modelo tem muitos coeficientes para estimar, a modelagem multinível pode ajudar para controlar o risco de "overfitting" por meio de "regularização" ou o uso de funções densidade de probabilidades a priori (priors) informativas.

Finalmente, podemos quantificar um pouco a afirmação acima, de que quando há indivíduos suficientes em cada grupo, não é necessário usar a modelagem multinível ou outros métodos de estimação mais complexos do que o método dos mínimos quadrados ordinário. Vamos considerar o caso mais simples possível, o das médias do ENEM por estado, sem nenhuma variável explicativa. Vamos chamar a variância do ENEM entre candidatos dentro de um estado de σ^2 e a variância entre estados de τ^2. O coeficiente de correlação intraclasse (ICC) é

$$\frac{\tau^2}{\tau^2+\sigma^2} \quad .$$

Se aproximamos as variâncias na população pelas suas estimativas amostrais, temos que no nosso caso τ^2 é da ordem de 17^2 e σ^2 é da ordem de 71^2, dando um ICC de 5%. Uma interpretação deste número é que 5% da variância na nota ENEM é associada ao estado e 95% da variância restante, aos indivíduos. Em modelos multinível em geral é possível que as estimativas dos coeficientes de grupos com poucos dados sejam "puxados" na direção da média geral (um efeito chamado também de "shrinkage").

Para o modelo mais simples de somente médias, um estimador da média de um estado seria (GELMAN; HILL, 2007, p253) um valor intermediário entre a média geral (do Brasil) e a média simples de um estado. O peso entre um e outro extremo seria dado por

$$\frac{\tau^2}{\tau^2+\frac{\sigma^2}{n}} \ ,$$

onde *n* é o número de dados de um estado. Ou seja, no caso extremo e hipotético de um estado com somente 1 dado, usaríamos uma média ponderada com peso de 0,05 para a nota do estado em 0,95 da nota média do Brasil. Mas no nosso caso, o estado com o menor número de dados, Roraima, tem 1800 notas na base, após todos os filtros. Neste caso, o peso na média ponderada das 1800 notas de Roraima é 0.99 e somente 0.01 da

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

média nacional, explicando por que não precisamos dos métodos mais complexos para estimar nosso modelo multinível.

REFERÊNCIAS

BALASSIANO, Moisés; SEABRA, Alexandre Alves de; LEMOS, Ana Heloisa. Escolaridade, salários e empregabilidade: tem razão a teoria do capital humano?. Revista de Administração Contemporânea, v. 9, p. 31-52, 2005.

Brasil, 2024. índice de Gini das UFs brasileiras, sidra.ibge.gov.br/tabela/7435#/n1/all/n3/all/v/10681/p/last%201/d/v10681%203/l/v,p,t/cfg/uf,2comp,1comp,cod,signiv,idrot,abvrot,nt,agpcab,/resultado, consultado em dezembro de 2024.

CHEIN, Flávia. Introdução aos modelos de regressão linear: um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas. 2019.

DATTA, Gautam; MEERMAN, Jacob. Household income or household income per capita in welfare comparisons. Review of Income and Wealth, v. 26, n. 4, p. 401-418, 1980.

FEIJÓ, Janaína Rodrigues; FRANÇA, João Mário Santos de. Diferencial de desempenho entre jovens das escolas públicas e privadas. Estudos Econômicos (São Paulo), v. 51, p. 373-408, 2021.made

GELMAN, A.; HILL, J. Data analysis using regression and multilevel/hierarchical models. 23rd printing ed. Cambridge: Cambridge Univ. Press, 2007.

GOLDSTEIN, H. Multilevel statistical models. 4th ed ed. Hoboken, N.J: Wiley, 2011.

GUISO, Luigi et al. Culture, gender, and math. Science, v. 320, n. 5880, p. 1164-1165, 2008.

HARTAS, Dimitra. Families' social backgrounds matter: Socio-economic factors, home learning and young children's language, literacy and social outcomes. British Educational Research Journal, v. 37, n. 6, p. 893-914, 2011.

HELENE, O., Método dos Mínimos Quadrados com Formalismo Matricial,2ª. Edição, São Paulo, Editora Livraria da Física, 2012

HELENE, Otaviano; MARIANO, Leandro. Educação e desigualdade na distribuição de rendas. Educação & Sociedade, v. 41, p. e223485, 2020.

HERTZ, Tom et al. The inheritance of educational inequality: International comparisons and fifty-year trends. The BE Journal of Economic Analysis & Policy, v. 7, n. 2, 2008.

IBGE/PNAD-C sidra.ibge.gov.br. Tabela 7127: Número médio de anos de estudo das pessoas de 15 anos ou mais, por cor ou raça e grupo de idade». Consultado em 20 de janeiro de 2023

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

JALOTO, A.; PRIMI, R. Fatores socioeconômicos associados ao desempenho no Enem. Em Aberto, v. 34, n. 112, 30 dez. 2021.

JOLLY, E. Pymer: Connecting R and Python for Linear Mixed Modeling. Journal of Open Source Software, v. 3, n. 31, p. 862, 26 nov. 2018.

BOSKER, 2012 EWOUT: PRECISA COMPLETAR A REGFERÊNCIA

MADEIRA, Fellipe; CAPRARA, Bernardo Mattes. Condições de docência, origens sociais e resultados educacionais: análise quantitativa sobre o desempenho de estudantes no Saeb 2021. Inter-ação: revista da Faculdade de Educação da UFG. Goiânia. Vol. 49, n. 3 (set./dez. 2024), p.[1533]-1550, 2024.

MAGALHÃES, João Carlos Ramos; MIRANDA, Rogério Boueri. Dinâmica da renda per capita, longevidade e educação nos municípios brasileiros. Estudos Econômicos (São Paulo), v. 39, p. 539-569, 2009.

MENDES, Igor A. Assaf; COSTA, Bruno Lazzarotti D. Considerações sobre o papel do Capital Cultural e acesso ao ensino superior: uma investigação com dados de Minas Gerais. Educação em Revista, v. 31, p. 71-95, 2015.

MORAES, C. P. D. et al. Efeito escola a partir de indicadores educacionais: análise entre escolas públicas e privadas no ENEM. Revista Meta: Avaliação, v. 14, n. 42, p. 67, 31 mar. 2022.

RIGOTTI, José Irineu Rangel et al. A re-examination of the expected years of schooling: what can it tell us. Brasília: International Policy Centre for Inclusive Growth (IPC-IG), 2013.

ROCHA, Aline Lemes da Paixão; LELES, Claudio Rodrigues; QUEIROZ, Maria Goretti. Fatores associados ao desempenho acadêmico de estudantes de Nutrição no Enade. Revista brasileira de Estudos pedagógicos, v. 99, n. 251, p. 74-94, 2018.

SNIJDERS, T. A. B.; BOSKER, R. J. Multilevel analysis: an introduction to basic and advanced multilevel modeling. 2nd ed ed. Los Angeles: Sage, 2012.

SOARES, Denilson Junio Marques; SANTOS, Wagner dos. Indicadores de avaliação de contexto e resultados educacionais no Ideb: uma análise das escolas estaduais de ensino médio no Espírito Santo. Revista Brasileira de Estudos Pedagógicos, v. 105, p. e5872, 2024.

SOARES, José Francisco. Measuring cognitive achievement gaps and inequalities: The case of Brazil. International Journal of Educational Research, v. 45, n. 3, p. 176-187, 2006.

DE SOUSA, L. A. et al.. Desempenho das Instituições Federais de Educação Profissional, Científica e Tecnológica Brasileiras no ENEM, Tendencias Pedagógicas, 34, p. 128-138. doi: 10.1, 2019

UNESCO. UIS METHODOLOGY FOR ESTIMATION OF MEAN YEARS OF SCHOOLING. https://uis.unesco.org/sites/default/files/documents/uis-methodology-for-estimation-of-mean-years-of-schooling-2013-en_0.pdf, 2013

SciELO Preprints - Este documento é um preprint e sua situação atual está disponível em: https://doi.org/10.1590/SciELOPreprints.12047

CONTRIBUIÇÃO DOS AUTORES

Autor 1 – Idealização do estudo, metodologia, escrita e revisão do manuscrito.

Autor 2 – Revisão, análise dos dados, metodologia e análise do estudo.

Autor 3 – Levantamento e análise dos dados, metodologia e revisão do manuscrito.

DECLARAÇÃO DE CONFLITO DE INTERESSE

Os autores declaram que não há conflito de interesse com o presente artigo.

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.