Uma análise em perspectiva multinível sobre os microdados do ENEM de 2020

Eliabe Rocha Da Cruz[1]*; Gabrielle Maria Romeiro Lombardi[2]

[1] Universidade do Estado do rio de Janeiro. Bacharel em Ciências Sociais – Instituto de Ciências Sociais. Rua Juqueri, nº 194 – Irajá; 21371-370 Rio de Janeiro, Rio de Janeiro, Brasil
[2] PECEGE, Doutora em Genética e Melhoramento de Plantas. Rua Alexandre Herculano, 120 - Vila Monteiro; CEP 13418-445; Piracicaba, São Paulo, Brasil
*autor correspondente: Eliabe-rocha@hotmail.com

Uma análise em perspectiva multinível sobre os microdados do ENEM de 2020

Resumo

Este trabalho tem como objetivo a investigação analítica da base de dados do Exame Nacional do Ensino Médio [ENEM] de 2020, considerando fatores como renda, escolaridade dos responsáveis, raça, tipo de escola, região e desempenho no exame por parte dos candidatos. Para tal feito, foi utilizado a técnica de aprendizado de máquina supervisionada chamada de modelagem multinível, que considera aspectos hierárquicos para melhor estimação e parâmetros conforme o conjunto de dados trabalhado. Para além modelagem hierárquica, também foi realizada uma análise com foco descritivo dos dados, procurando compreender como se distribuem os participantes do ENEM conforme suas características socioeconômicas. Ao final do estudo foi possível responder perguntas como qual o impacto da renda sobre o desempenho? Existem diferenças nas médias do exame entre alunos de escolas públicas e privadas? O grau de escolaridade da mãe pode influenciar no desempenho dos candidatos? Além disso, o estudo gerou insumos para outras pesquisas na área de educação e políticas públicas.

Palavras-chave: Modelo hierárquico; Educação; Exame Nacional do Ensino Médio; Desempenho.

Introdução

Em 1998 o Exame Nacional do Ensino médio [ENEM] surge como ferramenta para avaliação do desempenho de estudantes e de qualidade escolar, fundamental para compreensão dos rumos que a educação do país tomaria após a aprovação da primeira Lei de Diretrizes e Bases da Educação Nacional [LDB] (lei 9.394/96) pós constituição de 1988.

A partir da lei 9.394, que estabelece as diretrizes e bases da educação nacional, o ensino médio passou a ser considerado como educação básica e a universalizá-lo, permitiu a criação de um novo currículo baseado nas competências básicas necessárias para o avanço das séries escolares, além de remodelar e reinserir no currículo a formação profissional. Todas essas mudanças, representaram um grande marco social e histórico que permitiu o aprimoramento da qualidade da educação básica e o aumento da oferta de vagas e de acesso ao ensino médio e superior (Castro e Tiezzi, 2005).

A princípio o objetivo do ENEM era permitir uma visão da adesão e impacto das novas diretrizes curriculares oriundas da LDB, que visavam integrar o ensino fundamental e médio de forma continuada e formar um indivíduo para além das capacidades técnicas exigidas no mercado de trabalho. Desde sua criação, o exame sofreu inúmeras mudanças que perpassam pelo formulário sócio econômico, o número de questões aplicadas na prova e até sua finalidade, sendo atualmente o principal instrumento de avaliação para acesso ao ensino superior. O número de inscritos saltou de 157.221 em 1998[1] para mais de 4,3 milhões em

1 Inscrições no Enem crescem 20 vezes desde 1998. Portal do Ministério da Educação [MEC]. 23 Ago. 2006. Fonte: INEP: Disponível em: <http://portal.mec.gov.br/ultimas-noticias/201-266094987/6881-sp-1649249425>. Acesso em: 12 de nov. 2021

2020[2], de acordo com o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Freitas, 2004; Castro E Tiezzi, 2005).

É importante destacar que tal crescimento decorre da implementação de outros mecanismos que universalizam o acesso ao ensino superior, como o Sistema de Seleção Unificada [SISU], criado em 2010. O SISU permite que os participantes do ENEM usassem suas notas para concorrerem á vagas em unidades públicas em todo o Brasil e em algumas universidades de Portugal (Felicetti e Alberto, 2017), e em algumas faculdades e universidade particulares na modalidade de bolsistas pelo Programa Universidade para Todos [Prouni], criado em 2004 (Nogueira, 2017).

Além de ser um instrumento de avaliação e acesso de alunos ao ensino superior, o INEP, fornece anualmente uma rica base de dados com informações referente ao desempenho e a situação socioeconômica dos participantes (Ferrari et al, 2019). Essa base de dados reúne informações de localidade, tipo de ensino e escola, notas dos participantes, faixa de renda, escolaridade dos familiares, raça etc., o que permite conhecer um pouco sobre a realidade dos participantes e até realizar algumas inferências quanto aos grupos que participam do exame.

Esses dados vêm se acumulando nos últimos 23 anos e embora sofram mudanças, como a inserção de novas perguntas ou retirada de outras, o levantamento de informações que permitam a avaliação do modelo educacional, bem como fermenta de acesso ao ensino superior se mantêm quase que inalterada. Dificilmente, todos os dados são explorados de forma conjunta dado o caráter complexo e abrangente que as bases de dados do ENEM apresentam, com variáveis que permitem perpassar entre o participante, escola, tipo da escola, bairro, município e estado, além de informações que permitem a categorização e agrupamentos desses mesmos participantes.

Assim, o uso da modelagem multinível como ferramenta principal de análise e obtenção de respostas, pode ser empregada para realização de uma avaliação mais ampla considerando a influência que os contextos, ou níveis, geram sobre os participantes e vice-versa (Courgeau, 2003)

Dessa forma, este trabalho visa entender o impacto das características sociais e econômicas no desempenho médio dos participantes do Enem a partir da análise descritiva dos dados e da modelagem multinível. Além de gerar perguntas para pesquisas posteriores e proporcionar material de consulta ou apoio para pesquisas de políticas públicas e modelagem multinível com foco em educação.

2 Enem 2020 tem mais de 4,3 milhões de inscritos. Portal do Ministério da Educação [MEC]. 20 de Mai. 2020. Disponível em: <http://portal.mec.gov.br/pronatec/oferta-voluntaria/418-noticias/enem-946573306/90211-enem-2020-tem-mais-de-4-3-milhoes-de-inscritos.> Acesso em: 12 nov. 2021.

Material e Métodos

Descrição da Base de dados

Como matéria-prima desse estudo, foi utilizada a base de microdados de 2020, fornecidas pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [Inep]. Embora seja possível obter dados referentes aos anos anteriores a 2020, é importante ressaltar que não há regularidade na estrutura na forma como os dados foram coletados, variando o número de perguntas e respostas para cada pergunta. Além disso, o fato do identificador do participante não ser o mesmo entre os anos, um painel longitudinal temporal se torna inviável para o escopo deste estudo. A escolha da base com dados de 2020 se baseia, principalmente, por ter sido a última publicada até a data de finalização deste estudo.

Conforme o Inep informa em seu portal, os microdados são a menor unidade de informação colhida por meio da prova realizada anualmente e supre a carência de dados relacionados ao desempenho dos participantes por meio das notas e de informações socioeconômicas.

As bases são disponibilizadas pelo órgão de forma aberta em seu site, sendo possível baixá-las comprimidas no formato ZIP (arquivo comprimido) junto do dicionário de dados, arquivos de carregamento nos formatos R, SAS e SPS, além de documentos como editais, provas e gabaritos do referido ano. A disponibilidade desses arquivos pode variar de ano para ano.

Sobre a base, no ano de 2020 foram contabilizadas 76 variáveis, distribuídas entre seis grupos de perguntas.

- Dados do participante – que concentra informações como número de inscrição do participante, município de residência, idade, estado raça/cor, categoria de escola (pública, privada, N/R*, exterior), tipo de Ensino (regular, educação especial ou educação de jovens e adultos), etc.
- Dados da escola – tratando de informações referentes a dependência administrativa (estadual, municipal, feral, etc.), localização (rural ou urbana) e a situação de funcionamento da escola.
- Dados do local de aplicação da prova – referente ao município e estado onde a prova foi realizada.
- Dados da prova objetiva – traz informações como nota dos alunos nos eixos curriculares (ciências da natureza, ciências humanas, matemática e linguagens), a

presença na prova de cada eixo, o tipo da prova dividido por cores, além dos vetores de gabarito e de respostas doo participante.

- Dados da redação – contendo a situação da redação do participante (anulado, sem problemas, em branco, etc.), a nota de cada componente (sendo a redação avaliada sob a ótica de cinco componentes) e a nota final.
- Dados do questionário socioeconômico – reúne informações sociais e econômicas dos alunos, como faixa de renda, escolaridade e ocupação dos responsáveis, além de acesso à internet e equipamentos do cotidiano.

É importante destacar que o número de variáveis oscila entre os anos, bem como a divisão dessas categorias. Por exemplo, em 2015 a base de dados apresentou 166 variáveis distribuídas entre dez grupos e em 2018 foram 137 variáveis, divididas em 9 grupos.

A diversidade, bem como a quantidade de informações coletadas se destacam como os principais motivos para a escolha do material e tema. Tais dados permitem que a partir dos diferentes níveis e características, formem agrupamentos que expliquem o comportamento de determinado fenômeno como, por exemplo, a variabilidade entre notas por região, o impacto do capital cultural e da renda familiar sobre as notas ou a concentração de participantes com características sócios econômicas agrupadas geograficamente. As possibilidades de perguntas e constructos se tornam quase infinitas, limitando-se apenas a imaginação do investigador.

Outra característica interessante é o pré-tratamento que as bases oferecem através de arquivos de suporte para carregamento e transformação das variáveis. Os arquivos encontram-se nos formatos dos principais softwares de análise de dados do mercado, como o SPSS e o R, requisitando pouca preocupação com pré-processamento, a não ser para uso de técnicas específicas de modelagem ou perguntas de hipóteses mais específicas.

Técnicas para análise dos dados

Análise descritiva

Visando extrair os melhores resultados possíveis, optou-se primeiramente pelo uso da análise descritiva da base de dados do ENEM. Nesse ponto, analisar os dados em busca das regiões e estados com melhor ou pior desempenho, a média por categoria de escola ou tipo de Ensino, a frequência de ocorrência do grupo de profissões dos responsáveis e as correlações positivas ou negativas entre as variáveis permitirá a obtenção de material que responda perguntas como a região com maior ou menor média e sua possível relação entre com as taxas de indicadores sociais como IDH, número de participantes que deixaram pelo

menos uma prova em branco, a faixa salarial que mais se repete entre os alunos e se há correlação com as notas maiores ou menores, o impacto do grau de escolaridade dos responsáveis sobre o desempenho médio dos participantes, etc. Essas perguntas, embora pareçam genéricas, podem ajudar na criação de um cenário para análise mais aprofundada e com técnicas mais específicas.

Modelagem Multinível

Após a análise por estatística descritiva, foi utilizado o pacote nlme v. 3.1-157 para linguagem de programação R. O pacote permite o treino de modelos mistos lineares e não lineares, para o estudo a técnica usada foi a modelagem multinível, visando compreender a influência de variáveis como escolaridade dos responsáveis, renda, idade, autodeclaração de raça ou tipo de escola na variabilidade do desempenho dos alunos. Além disso, a técnica permitirá calcular qual a proporção dessa variabilidade é explicada pelo nível.

A característica da base de dados, como o fato de apresentar naturalmente uma disposição hierárquica, faz com que a melhor técnica seja a modelagem multinível, dado os ganhos em capacidade descritiva e até preditiva, embora este último não se aplique ao estudo. Esses ganhos são decorrentes da especificação dos componentes aleatórios de cada um dos níveis, trazendo melhor ajuste dos dados se comparado a um modelo “Ordinary least squares [OLS]”, pois este considera apenas um nível no processo de modelagem (Favero et al, 2017).

Limpeza e análise descritiva dos dados de 2020

Para análise descritiva dos dados e apresentação da população estudada, foram selecionadas variáveis de localização, desempenho, renda, idade, sexo, tipo de ensino e de escola, Raça/Cor, Escolaridade dos pais entre outras.

Para análise, bem como para aplicação do modelo, foi desconsiderado todos os dados daqueles que faltaram a pelo menos um dia de prova, isso porque na construção da base essas informações foram registradas como valores faltantes ou nulos. O mesmo se aplica para aqueles cuja nota na redação foi igual a zero ou foram desclassificados, os que não informaram o tipo de escola onde cursaram a maior parte do ensino médio e aqueles que apresentaram dados faltantes na base e aqueles com idades inferiores a dez e superiores a 100 anos e de perguntas em que era permitido ao candidato optar pela resposta “Não Respondeu” ou “Não informado”.

Somado a isto, foram removidos dados com informações divergentes entre a variável do tipo escola e da dependência administrativa, 33.990 candidatos marcaram na questão

sobre o tipo de escola a opção pública, entretanto, quando questionados sobre a dependência administrativa (municipal, estadual, federal e privada) os mesmos marcaram a opção que referenciavam a escola privada. Outros 76 declararam pertencer a escola particular, porém declararam a dependência administrativa como municipal, estadual ou federal, por esse cruzamento equivocado de informações, 34.066 observações foram retiradas da base.

Além disso, duas variáveis foram criadas para fins do estudo, uma variável contendo a média das notas dos quatro eixos temáticos e da redação, uma contendo as regiões a qual o candidato pertence. Uma relação dos filtros, bem como a redução da base de análise pode ser vista na Tabela 1 abaixo:

Tabela 1. Número de observações deletadas por filtro

Filtro	Número de observações deletadas
Faltaram nos dias de exame	3.192.751
Desclassificados na redação ou nota igual a 0	661.625
Tipo de escola não informado	1.736.348
Estudantes com dados ausentes	284.907
Candidatos com informações divergentes	34.066
Base válida para análise	475.805
Redução total	5.307.304

Fonte: Dados originais da pesquisa

Ao todo foram contabilizados 5.783.109 participantes, após a limpeza da base somente 475.805 (8%) observações se mostraram válidas para o prosseguimento do estudo. A maior redução na base foi ocasionada devido ao não comparecimento a pelo menos um dia de prova, indicando uma redução em aproximadamente 55% dos dados.

O estudo decorrerou com amostra de 475.805 observações, tanto para a análise exploratória, quanto para modelagem multinível. A partir disso, deve ser considerado que a redução da base observável pode impactar nos resultados da modelagem, bem como na distribuição e a média do desempenho dos candidatos pelos grupos na análise exploratória, porém devido a construção do modelo multinível não necessitar do balanceamento dos dados entre escolas de diferentes dependências administrativas (Tabachnick e Fidell, 2012), o efeito da redução foi desconsiderada para o trabalho.

Modelagem dos dados

Como mencionado inicialmente, para este estudo optou-se pela modelagem multinível para verificar-se as relações existentes nos dados do Enem de 2020, tal escolha se pauta tanto na capacidade desse tipo de técnica, quanto pela própria natureza hierárquica que os dados apresentam.

Breve resumo histórico: vantagens e desvantagens de um HLM

Os modelos multiníveis aparecem em distintas literaturas com diferentes nomes, “multilevel linear model”, “mixed-effects models”, “random-effects models”, “random-coefficient regression models”, “covariance components models” e, mais comumente, “Hierarchical Linear Models”, ou modelos lineares hierárquicos em português, dado suas características de estruturação em níveis. (Raudenbush e Bryk, 2022)

A característica de destaque de um modelo multinível, se comparado a outros modelos como o OLS, é sua capacidade de considerar contextos no processo de análise. Em outros termos, significa que é possível acrescentar variáveis explicativas e erros-padrão referentes a cada novo nível implementado no modelo. Esse tipo de estrutura de erros complexa, permite que os “Hierarchical Linear Models” [HLM] apresentem um melhor ajuste se comparado aos modelos OLS tradicionais (Santos et al, 2000)

Além disso, por considerar diferentes contextos, o modelo multinível ao realizar o produto do termo de erro pela variável explicativa, pode minimizar ou eliminar o problema da heterocedasticidade, sendo este um problema recorrente dos modelos não hierárquicos devido à alta variância dos termos de erro. (Fávero e Belfiore, 2017)

Em contraponto as vantagens, a modelagem multinível tem como necessidade principal um poder computacional mais robusto (Lazega e Snijders, 2016), dado o uso de variáveis explicativas e acréscimo de níveis na estimação do modelo, como consequência direta dessa limitação o uso desse tipo de técnica ficou restrito até finais de 1970 (Raudenbush e Bryk, 2002; Santos et al, 2000).

Embora a limitação da capacidade de processamento em dados de larga escala ainda se faça presente, principalmente com o constante aumento no volume de dados, o desenvolvimento de processadores mais potentes e o advento da computação em nuvem minimizam esse problema para que modelos HLM possam ser estimados com mais precisão e eficiência.

Estimação do modelo

Os modelos lineares apresentam uma estrutura básica geral, descrita por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + e_i \quad (1)$$

em que Y_i representa o fenômeno a ser estudado, ou variável dependente, para cada observação i ($i = 1,2,\dots,n$) na base de dados. β_0 é o intercepto, $\beta_1, \beta_2, \dots, \beta_k$representam os

coeficientes de inclinação de cada variável, X_{i1}, X_{i2}, ..., X_{ik} são as variáveis explicativas adicionadas ao modelo e, por último, e_i representa o termo de erro do modelo. É necessário considerar ainda que, para que o modelo acima proposto tenha validade, os termos de erro devem apresentar aderência a uma distribuição normal, média igual a zero e que as variáveis explicativas e os erros-padrão sejam independentes entre si, excluindo a característica de multicolinearidade do modelo.

Tal como a estrutura de um modelo linear tradicional, um modelo multinível segue a mesma regra de formação, porém considerando variáveis explicativas e erros-padrão para cada nível acrescido ao modelo, além da exclusão da regra de independência entre os erros e as variâncias não necessariamente serão constantes (Raudenbush e Bryk, 2002).

De maneira geral, a estrutura básica de um modelo de j níveis é:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} \quad (2)$$

sendo:

Y_{ij} a variável dependente, ou variável resposta, obtida da modelagem sobre cada indivíduo do nível j.

β_{0j} e β_{1j} são, respectivamente, os valores de intercepto e de inclinação do modelo.

r_{ij} representa o erro idiossincrático para cada indivíduo em cada nível.

Se tratando de um modelo de nível II, substitui-se β_{0j} e β_{1j} pelas equações abaixo. Dessa forma, há:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \quad (3)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j} \quad (4)$$

em que:

γ_{00} é o intercepto geral;

γ_{01} é o valor esperado quando se altera uma unidade na característica W.

γ_{10} representa a mudança de inclinação em detrimento da variável X.

γ_{11} é a variação de inclinação em função de W.X.

W_j representa a variável característica do grupo j, sendo invariável entre os indivíduos do mesmo grupo.

u_{0j} e u_{1j} são os termos de erro que indicam a existência de aleatoriedade nos interceptos e nas inclinações do segundo nível, respectivamente. Aqui assume-se que os termos de erro, tanto de primeiro, quanto de segundo nível apresentam distribuição normal.

$$u_{ij}\ N\left(0, \sigma_{ui}^2\right) \quad (5)$$

$$r_{ij}\ N(0, \sigma_r^2) \quad (6)$$

Caso a variância entre os efeitos aleatórios do intercepto (III) e de inclinação (IV) sejam iguais a zero, então o modelo adequado é o OLS e não um “Hierarchical Linear Model [HLM]”. (Tabachnick e Fidell, 2012).

Reordenando as equações, o modelo final de nível dois ficaria como descrito na fórmula VII.

$$Y_{ij} = \gamma_{00} + \gamma_{01}W_j + \gamma_{10}.X_{ij} + \gamma_{11}W_j.X_{ij} + u_{0j} + u_{1j}.X_{ij} + r_{ij} \quad (7)$$

sendo a primeira parte, até $W_j.X_{ij}$, correspondente aos efeitos fixos do modelo e a segunda parte aos efeitos aleatórios.

Estimação dos parâmetros e modelagem

Para fins deste estudo, foi feito uso da Máxima verossimilhança restrita, “restricted estimation of maximum likelihood – [REML]”, para estimação dos parâmetros, embora a estimação dos termos de erro possa ser realizada por meio somente da máxima verossimilhança [MLE], a opção restrita permitiu ganho de tempo de computacional durante o processo de estimação.

De acordo com Fávero (2016), ambos procedimentos usam iterações computacionalmente demandantes para se estimar os parâmetros através da maximização da função de verossimilhança, porém em grandes amostras e dados, quase não há diferença nas estimações dos erros entre o REML e o MLE. Em outras palavras, implica-se entender quais os valores referentes a cada parâmetro do modelo que maximiza o valor da função de verossimilhança e, consequentemente, gera um modelo mais eficiente.

Ainda sobre a estimação dos parâmetros e escolha das variáveis, diferente da técnica de mínimos quadrados ordinários [MQO] ou OLS, em inglês, que faz uso da técnica de “Stpwise” para definir quais variáveis farão parte do modelo final, não existe algoritmo que permita esse procedimento na modelagem multinível. Sendo assim, a construção do modelo seguirá a estratégia “Step-up”, em que serão acrescidas variáveis a cada etapa para que seja possível analisar qual o impacto sobre o modelo (Raudenbush e Bryk, 2002).

Construção do modelo

Construção do modelo nulo

Como primeira etapa da modelagem, foi estimado o modelo nulo de dois níveis, sendo a média do aluno em função da dependência administrativa da escola do candidato (Municipal, Estadual, Federal ou Particular) com o objetivo de avaliar se há variação no "Log-likelihood" [LogLik] entre um modelo OLS e um modelo multinível, onde foi constatada a diferença nos valores de Máxima Verossimilhança ["LogLikelihood"], em favor do modelo multinível, conforme Tabela 2.

Tabela 2. Comparação de LogLik entre modelos

Modelo	LogLik
Modelo Nulo HLM 2	- 2695116
Modelo Nulo OLS	- 2772310

Fonte: Dados originais da pesquisa

Quando comparado os valores através teste da razão de verossimilhança, o modelo multinível apresenta significância estatística, a 95% do intervalo de confiança, apontando para seu desempenho superior. Cabe ressaltar, porém, que embora o teste de razão tenha sido favorável ao modelo multinível, a significância estatística do termo de erro ν_{0j} foi de 15.6%, não sendo possível rejeitar a hipótese nula e orientando uma análise segundo um modelo não hierárquico, visto não haver diferença estatisticamente significante entre os interceptos aleatórios do modelo. A discussão sobre a significância dos termos de erro do intercepto e das inclinações é levantada pelas professoras e pesquisadoras da California State University, Barbara G. Tabachnick e Linda S. Fidell, em seu livro "Using Multivariate Statistics" (2012).

Dada a contradição entre os testes de razão de verossimilhança do modelo e de significância do termo de erro do intercepto aleatório, optou-se por continuar a modelagem multinível e adicionar os termos de inclinação aleatório. Caso o termo de erro de inclinação apresente significância estatística de até 10% ($p < 0,10$) e ao mesmo tempo o teste de razão seja favorável ao procedimento multinível, a abordagem hierárquica será a escolhida para este estudo. A liberdade para que o teste de significância passe de 5% para 10% decorre da própria divergência entre os testes para os modelos nulos.

Ainda no modelo multinível nulo, outro parâmetro que pode ser estimado foi a correlação intraclasse, ou seja, o impacto da dependência administrativa no desempenho dos alunos. Conforme aponta Tabachnick e Fidell (2012). A fórmula de obtenção do valor se dá por

$$\rho = \frac{s_{bg2}^2}{s_{bg2}^2 + s_{wg}^2} \quad (8)$$

sendo o numerador a estimação da variância do termo de erro do intercepto e o denominador a soma das variâncias dos termos erro de intercepto e resíduos.

Após a aplicação da fórmula constatou-se que 23% da variância do desempenho dos alunos é explicada pelo fato dos mesmos estudarem em escola pública (municipal, estadual ou federal) ou privada, mostrando o impacto sobre o modelo e indicando a possibilidade de prosseguimento com o modelo de dois níveis.

Modelo final

Após a escolha do número de níveis do modelo, bem como a verificação do impacto sobre o desempenho dos estudantes no exame, foram inseridas as variáveis de efeito aleatórios e as demais preditoras do modelo. Para fins deste estudo não foram utilizadas interações entre variáveis de nível um e nível dois, devido ao tamanho da base de dados e o custo computacional necessário para processamento.

Após a modelagem dos dados, observou-se que as variáveis renda, grau de escolaridade da mãe, cor ou raça apresentaram significância estatística para o modelo proposto com exceção da reposta 4, referente a variável TP_COR_RACA. Por padrão, o pacote nlme, transforma variáveis categóricas do tipo “fator” com mais de um nível em dummies, possibilitando a estimação dos parâmetros do modelo. Nesse caso, a “dummies” equivalente a resposta 4 de TP_COR_RACA representa 10.078 (0.4%) participantes que se auto identificaram com a cor amarela no questionário socioeconômico e para o modelo final não foram consideradas, devido a não significância estatística a 5%.

Para fins de verificação de validade do modelo, foi realizado novamente o teste de significância com os termos de erro aleatórios, onde foi possível constatar que ν_{0j} teve um salto no valor p, indo de 15.6% para 19%, porém o termo de erro ν_{1j} referente a inclinação aleatória apresentou significância estatística de 1,7% (0.017), indicando que há influência da idade sobre inclinação das retas do modelo.

Também para fins de comparação de eficiência do modelo e sua validação, foi realizado um teste de razão de verossimilhança, permitindo assim testar a significância estatística do log Likelihood para um modelo OLS e um modelo multinível. Para o teste, foram formuladas as seguintes hipóteses:

H_0: Não há diferença estatística entre ambos modelos.

H_1: Há diferença estatística entre os modelos.

Não sendo possível rejeitar a hipótese nula, o modelo OLS seria mais adequado para o estudo. Em contraponto, foi assumido que a hipótese alternativa (H_1) é a prevalência da significância estatística do modelo multinível em detrimento do modelo OLS.

Após realização do teste, o modelo hierárquico de dois níveis apresentou significância estatística (p < 0,05), conforme output da Tabela 3, dessa forma rejeitando a hipótese nula.

Tabela 3. Teste de razão de verossimilhança dos modelos finais

	#DF	LogLik	DF	Chisq	Pr(>Chisq)
Modelo OLS	31	-2683605	NA	NA	NA
Modelo HLM	34	-2658151	3	50908.3	0

Fonte: Dados originais da pesquisa

Conforme mencionado anteriormente, houve uma flexibilização do limite de significância do teste com termos aleatórios, e caso um dos termos estivesse abaixo de 10% (p < 0,10), além do favorecimento do modelo multinível em detrimento do modelo OLS, seria dada continuidade na análise e interpretação à luz do modelo hierárquico.

As tabelas 4 e 5 apresentam os principais resultados do modelo hierárquico final.

Tabela 4 Likelihood e resultados dos efeitos aleatórios

Treinamento do modelo Linear mixed-effects por REML	
log Likelihood	-2658151

Efeitos Aleatórios: ~TP_FAIXA_ETARIA | TP_DEPENDENCIA_ADM_ESC	
Elemento	Desvio Padrão
(Intercept)	25.410903
TP_FAIXA_ETARIA	6.843957

Fonte: Dados originais da pesquisa

Tabela 5 Resultados dos efeitos fixos

Efeitos Fixos: NU_MEDIA ~ TP_FAIXA_ETARIA + Q006 + Q002 + TP_COR_RACA		
Elemento	Valor	p-value
(Intercept) - Referência	532.9544	0.0000
TP_FAIXA_ETARIA	-11.2131	0.0011
Q006D (Até dois salários - R$2.090,00)	30.9430	0.0000
Q006Q (Acima de 20 salários - R$20.900,00)	103.0526	0.0000
Q002B (Não completou Ens. Fundamental 1)	6.0878	0.0000
Q002F (Ensino Superior completo)	30.1888	0.0000
TP_COR_RACA1 (Brancos)	6.6402	0.0000
TP_COR_RACA2 (Pretos)	- 11.8109	0.0000
TP_COR_RACA5 (Indígenas)	- 28.0443	0.0000

Fonte: Dados originais da pesquisa

Resultados e Discussão

Análise descritiva

De acordo com o gráfico 1 que mostra a densidade do desempenho por tipo de escola e a Tabela 6 que agrupa a média dos participantes do tipo de escola e região, nota-se que as escolas particulares apresentam média superior às escolas públicas, a nível regional e nacional. Estendendo ainda essa análise, por meio da Tabela 6, é possível comparar as notas nas diferentes regiões do país, tal como a distribuição de alunos por tipo de escola.

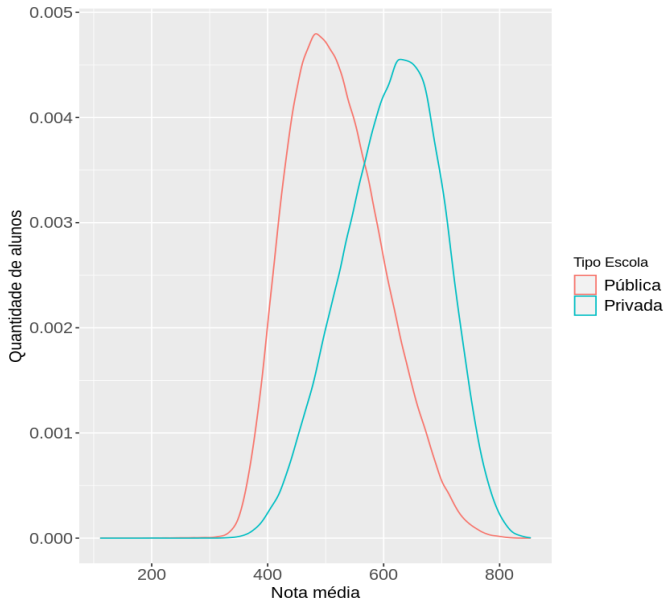

Figura 1. Gráfico de Densidade
Fonte: Resultados originais da pesquisa

Tabela 6. Médias por região e tipo de escola

Variável	Níveis	Quantidade de alunos	Média geral
Pública	Centro Oeste	31.165	518.92
	Nordeste	125.280	503.26
	Norte	33.666	495.42
	Sudeste	106.880	538.65
	Sul	39.958	535.03
	Nacional	336.949	518,26
Privada	Centro Oeste	12.248	617.85
	Nordeste	37.454	604.58
	Norte	6.953	599.31
	Sudeste	66.192	618.02
	Sul	16.009	618.44
	Nacional	138.856	611.64

Fonte: Resultados originais da pesquisa

Embora mais abrangente, é possível notar já de antemão e sem a necessidade de modelagem, as variações de média entre o tipo de escola, corroborando com estudos de educação que apontam que tal diferença é crucial no processo formativo e de acesso a oportunidades por parte do jovem (Belluzzo e Moraes, 2014).

Observa-se, também, que a região norte é a que apresenta o menor desempenho geral no Enem 2020, indo de encontro com dado da mesma região apresentar o menor IDH, segundo o Instituto de Pesquisa Econômica Aplicada (IPEA e FJP, 2016).

Aprofundando no nível escolar, a base ainda fornece a variável de dependência administrativa, que subdivide escolas públicas em federais, estaduais e municipais, além da escola privada. Conforme mostra a Tabela 7, o melhor desempenho médio no exame continua sendo o de escolas privadas, porém um olhar mais atento sobre as dependências relacionadas as escolas púbicas, permite destacar a discrepância entre as escolas federais e estaduais, principalmente.

Enquanto 90% dos candidatos oriundos de escolas federais tiveram nota inferior ou igual a 697.32, a mesma proporção de alunos em escolas estaduais obtiveram nota inferior ou igual a 611.56, diferença de 85.76. O uso de percentis na análise, ajuda a observar o comportamento das notas nas diferentes faixas de da amostra, além de reduzir o impacto da generalização da média sobre o desempenho do grupo.

Tabela 7 Quartis do desempenho médio por dependência administrativa

Dependência Administrativa	Total	Base	Média Nacional	Percentil 10%	Percentil 50%	Percentil 90%
Federal	37130	8%	595.65	487.560	599.00	697.32
Estadual	296035	62%	509.15	416.420	502.78	611.56
Municipal	3784	1%	530.67	423.464	528.77	637.84
Privada	138856	29%	613.49	499.100	619.20	717.94

Fonte: Resultados originais da pesquisa

Outra análise que mostra sua relevância para o estudo é a distribuição de candidatos por sexo e o desempenho conforme a faixa de renda. Sobre o fator sexo, observa-se que o número de participantes do sexo feminino ultrapassa em números os participantes do sexo masculino, conforme gráfico 2, porém quando analisado o fator desempenho nota-se que candidatos masculinos apresentam média superior em, aproximadamente, 13 pontos.

Ainda sobre o sexo dos participantes, constatou-se que dentre as dez maiores notas do exame, somente três pertenciam as candidatas, todas provenientes de escolas particulares, sendo duas com renda superior a 20 salários e uma com renda acima de cinco salários.

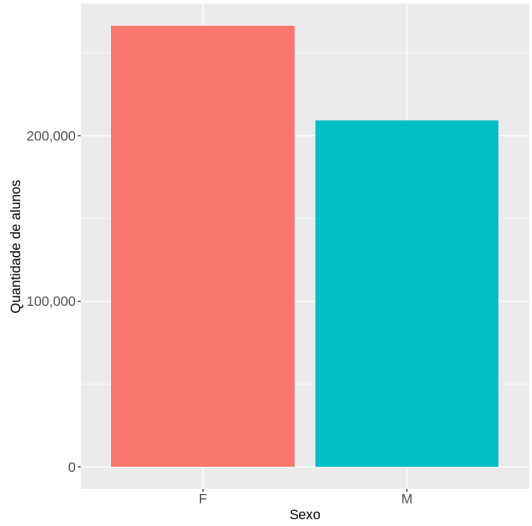

Figura 2 Distribuição de participantes por sexo
Fonte: Resultados originais da pesquisa

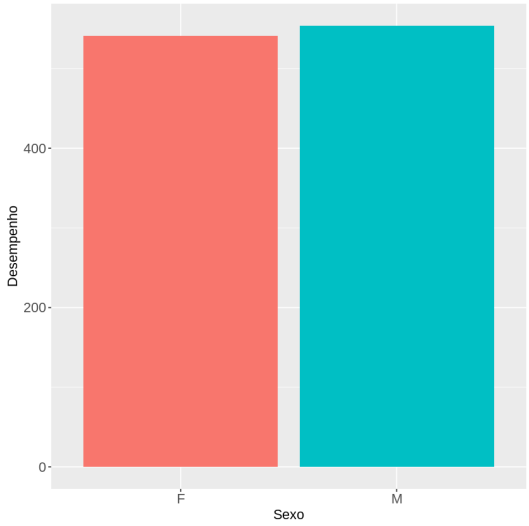

Figura 3 Distribuição da média por sexo
Fonte: Resultados originais da pesquisa

No que tange ao fator renda, 52% da amostra apresentou renda de até dois salários mínimo, à época R$ 2.090,00, conforme tabela 7. Estendendo a análise, dentre as dez maiores notas do exame em 2020, seis candidatos tinham renda familiar acima de R$ 20.900,00 (20 salários) e dois candidatos renda acima de R$ 12.540,01 (12 salários). Todos os dez participantes obtiveram nota superior a 840 pontos e frequentavam ou frequentaram o ensino médio em escola privada. Em contraponto, dentre as dez menores notas do exame, sete eram de participantes com renda de até R$ 1.045,00, somente dois eram de escolas privadas, a maior e menor nota do grupo foram 220.28 e 112 pontos, respectivamente.

Tabela 8 Distribuição de alunos e média por renda

Faixa de renda?	Total	% Base	Média
Acima de R$ 20.900,00	11.626	2%	661,97
De R$ 15.675,01 até R$ 20.900,00	8.989	2%	647,67
De R$ 12.540,01 até R$ 15.675,00	8.689	2%	638,07
De R$ 10.450,01 até R$ 12.540,00	9.200	2%	631,12
De R$ 9.405,01 até R$ 10.450,00	9.506	2%	624,72
De R$ 8.360,01 até R$ 9.405,00	6.541	1%	618,16
De R$ 7.315,01 até R$ 8.360,00	8.430	2%	612,79
De R$ 6.270,01 até R$ 7.315,00	10.624	2%	606,48
De R$ 5.225,01 até R$ 6.270,00	16.687	4%	598,75
De R$ 4.180,01 até R$ 5.225,00	26.004	5%	588,37
De R$ 3.135,01 até R$ 4.180,00	32.747	7%	574,01
De R$ 2.612,51 até R$ 3.135,00	37.736	8%	559,68
De R$ 2.090,01 até R$ 2.612,50	32.226	7%	549,73
De R$ 1.567,51 até R$ 2.090,00	53.760	11%	534,71
De R$ 1.045,01 até R$ 1.567,50	64.107	13%	521,50
Até R$ 1.045,00	116.519	24%	495,46
Nenhuma Renda	22.414	5%	479,38

Fonte: Resultados originais da pesquisa

Semelhante a relação feita entre IDH e a Média do ENEM, é possível notar que com o aumento da faixa de renda, há também o aumento do desempenho no exame. O mesmo movimento de crescimento foi observado com a escolaridade e o grupo de ocupação materno, ou seja, quanto maior a especialização educacional e laboral, maior a média dos filhos(a) na prova, conforme Figura 4 e Figura 5. Os grupos F e H, grau laboral e escolaridade materna respectivamente, correspondem a candidatos que não souberam opinar no questionário. Os demais graus de ambas variáveis estão ordenadas de forma crescente, do menor para o maior grau de especialização laboral e escolar materna.

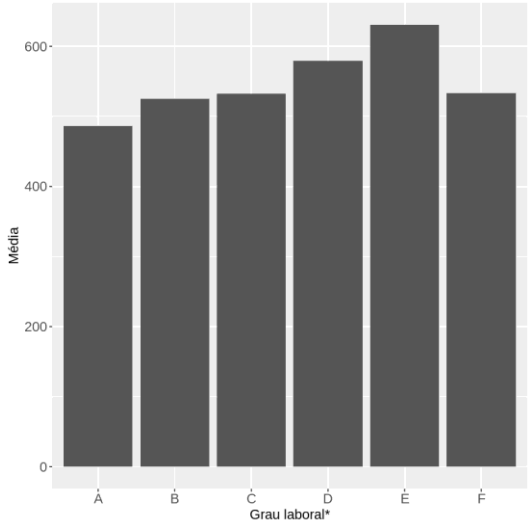

Figura 4 Desempenho em função da especialização materna
Fonte: Resultados originais da pesquisa

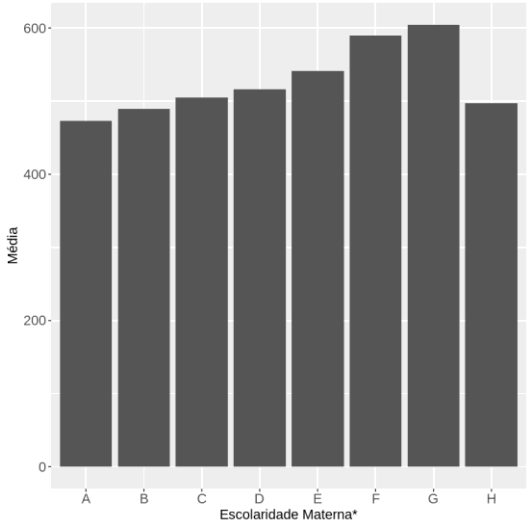

Figura 5 Desempenho em função do grau educacional materno
Fonte: Resultados originais da pesquisa

A partir desses breves resultados, foram levantados os seguintes questionamentos: Pode variáveis como a escolaridade dos pais, a idade e autodeclaração de cor ou raça e renda influenciarem no desempenho dos alunos? Qual seria o impacto dessas variáveis sobre o modelo e o quanto elas contribuem para explicar o fenômeno de aumento ou baixa das médias entre os alunos? Essas serão algumas das perguntas que nortearão esse estudo.

Resultados do Modelo

A partir das estimativas geradas pelo modelo, conforme Tabelas 4 e 5, é possível explorar algumas conclusões sobre o desempenho médio dos alunos e sobre alguns dos fatores que o impactaram.

O primeiro ponto, e que merece maior destaque, é o impacto que a renda apresenta sobre o desempenho no ENEM. Estudantes cuja família recebia entre 1 ½ salários mínimos, apresentavam um desempenho superior em 31 pontos na média. Essa diferença se aproxima

dos 50 pontos para aqueles que possuem renda familiar de até quatro salários mínimos (R$ 4.180).

O segundo ponto importante é o impacto que o grau de escolaridade da mãe sobre o desempenho do aluno, o participante do exame cuja mãe concluiu graduação, porém não a pós-graduação, apresentou aumento de aproximadamente 30 pontos na média se comparado àqueles cuja mãe não concluiu a 4º série do ensino fundamental. Tal interação entre desempenho e escolaridade materna pode decorrer da influência cultural, entendendo aqui toda estrutura comportamental e social necessária para progredir academicamente. Esse processo é discutido por Bourdieu “A Escola conservadora: as desigualdades frente à escola e à cultura” (1966), além de confirmar levantamentos sobre realizados pelo Instituto de Pesquisa Econômica Aplicada [Ipea] (Barros et al., 2001).

Outro aspecto é a correlação entre auto declaração e desempenho, sendo os grupos que se autodeclararam pretos ou indígenas, apresentaram impacto negativo, 16 pontos e 36 pontos respectivamente, com relação ao desempenho daqueles que não declararam nenhuma resposta ou não souberam informar. Tal discrepância é elevada se comparada com candidatos que se auto declararam brancos, nesse caso pessoas negras apresentam uma diferença de aproximadamente 19 pontos, enquanto aqueles que se identificaram como indígenas tiveram uma diferença de 35 pontos.

Aqui, fatores como distribuição de renda e racismo estrutural podem apresentar um papel fundamental nessa comparação (Cacciamali e Hirata, 2005; Carvalho, 2005), porém apenas observando o arranjo dos dados desse estudo mais detidamente, foi possível concluir que as mães de 40% dos alunos que se autodeclararam pretos terminaram o ensino médio, porém não chegaram a terminar a faculdade. Seguindo a mesma lógica, 37% dos candidatos que se autodeclararam indígenas tinham mães que também terminaram o ensino médio, porém não a faculdade.

Tratando sobre os efeitos aleatórios do modelo, foi possível constatar que a cada um ano do candidato, a partir dos 17 anos, incrementa em média 10.11 pontos na média final. O efeito aleatório apresentou correlação positiva de 0.43, explicando o incremento na nota.

Por fim, é possível analisar o comportamento dos efeitos aleatórios de intercepto e inclinação a partir da variável que hierarquiza a amostra. Assim como apontado na Tabela 7, a diferença entre escolas federais e estaduais se mostrou marcante no modelo. Conforme aponta a Figura a 6, escolas federais[3] apresentam uma correlação de inclinação aleatória positiva superior a 11 pontos, quando comparadas as escolas públicas, sendo o fator que se destaca para explicar a diferença estatística significante do modelo.

3 Os rótulos do eixo y do gráfico de inclinações aleatórias significam: 1 – Escolas Federais; 2 – Escolas Estaduais; 3 – Escolas Municipais e 4 – Escolas privadas.

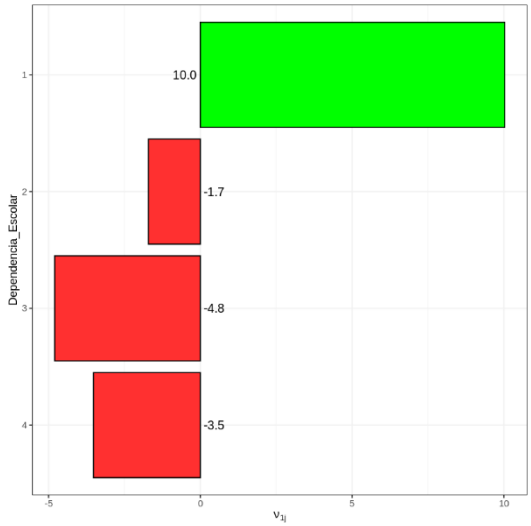

Figura 6 Comportamento das inclinações aleatórias
Fonte: Resultados originais da pesquisa

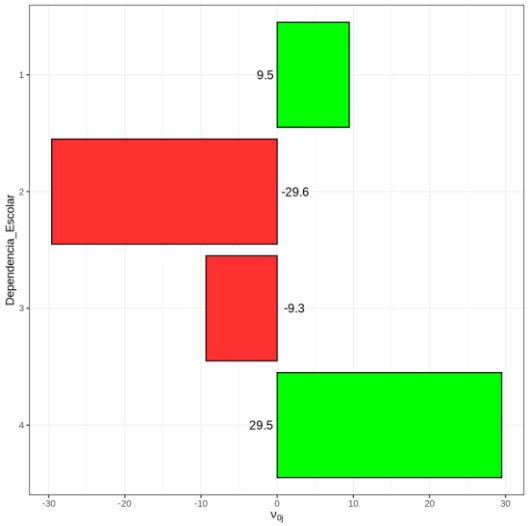

Figura 7 Comportamento dos interceptos aleatórios[4]
Fonte: Resultados originais da pesquisa

Tratando sobre o intercepto aleatório, é importante lembrar que o teste de significância apontou p-valor superior a 0.10, indicando que não há diferença estatística significante para o modelo. Porém, para fins de análise e em consonância com o estudo sobre as médias das dependências administrativas realizado anteriormente, verifica-se na Figura 7 a diferença entre escolas privadas e públicas, destacando a discrepância entre escolas estaduais e federais, e entre escolas estaduais e privadas, ambos com diferença de 39.1 e 59.1 pontos, respectivamente. Cabe pontuar que a depender do objetivo do estudo, as diferenças apontadas nos interceptos poderiam apresentar relevância para análise, sendo necessário apenas o ajuste do p-valor no teste de significância.

Comparação entre o modelo OLS e o modelo multinível

Concomitante ao desenvolvimento do modelo multinível, foi executado um modelo OLS para fins de comparação. Na figura 3 é possível observar o ganho de LogLik entre os modelos nulos e finais. Embora o ganho não tenha sido alto em escala, os testes de razão de verossimilhança apresentaram significância estatística em favor dos modelos multiníveis nulo e finais, quando comparados aos seus pares em modelagem não hierarquizada.

4 Os rótulos do eixo y do gráfico de interceptos aleatórios significam: 1 – Escolas Federais; 2 – Escolas Estaduais; 3 – Escolas Municipais e 4 – Escolas privadas.

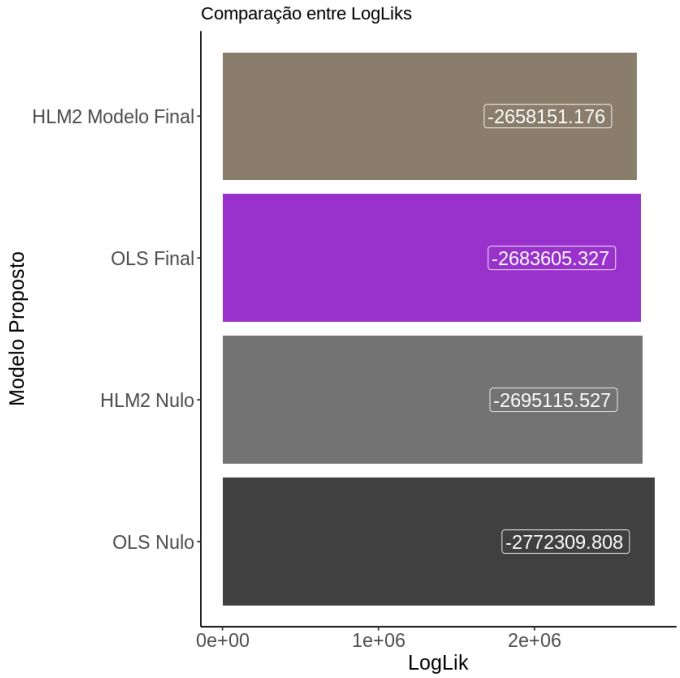

Figura 8. Comparação entre os LogLiks
Fonte: Resultados originais da pesquisa

Uma outra forma comparativa entre os modelos é por meio dos “fitted values”, que dispostos sobre gráfico tendem a apresentar o modelo com melhor ajuste dos dados estimados, conforme Figura 9.

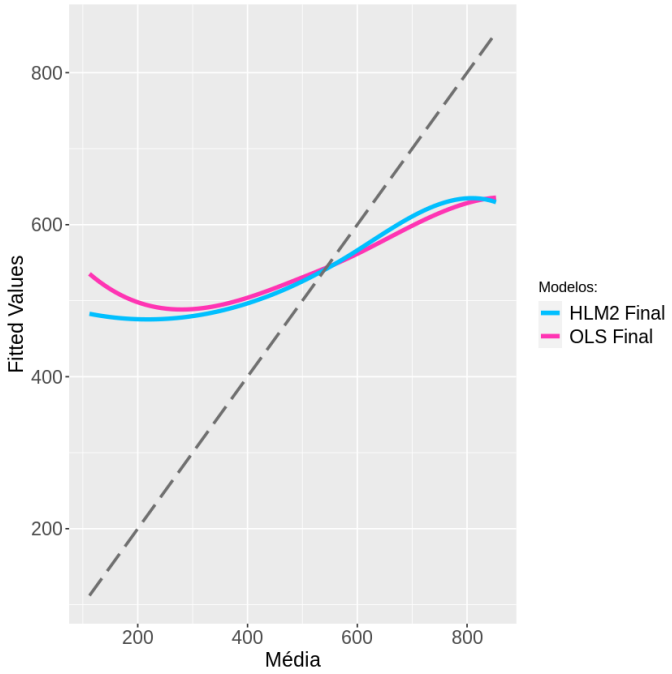

Figura 9. Curva de ajuste dos modelos
Fonte: Resultados originais da pesquisa

É válido pontuar que o ajuste do modelo é dependente da escolha das variáveis fixas, aleatórias e da quantidade de níveis implementados na modelagem, fato esse comprovado por meio da estratégia “step-up” e incrementação do LogLik. Além disso, o acréscimo de variáveis externa a base, bem como outras perguntas como tempo de experiência médio dos professores de cada escola ou ranking das escolas no município, estado ou região poderiam acrescentar mais poder de análise ainda ao modelo estudado, porém essas questões se estendem ao propósito do presente estudo.

Considerações Finais

A aplicação da técnica de modelagem multinível, ou hierárquica, lançou luz sobre alguns pontos muito debatidos nos estudos sobre educação, principalmente no que tange a relação entre desempenho escolar, renda, raça e nível de escolaridade dos responsáveis, onde foi constatado o impacto de tais características sociais e econômicas sobre o desempenho médio dos participantes do Enem em 2020. Outros fatores como impacto de políticas públicas, localização das escolas, gênero, tempo médio de estudo semanal ou composição familiar em números podem apresentar correlações interessantes com o desempenho escolar individual, entretanto o escopo deste estudo não daria conta de abordar tais questões, seja por falta de dados ou limites na capacidade computacional para estimar os modelos. De qualquer forma, a abordagem multinível se mostrou um grande aliado concernente a estudos educacionais, apontando para múltiplas possibilidades.

Agradecimento

Gostaria de agradecer a meus pais que me ensinaram a sempre lutar pelos meus sonhos, a Andreza Melo, minha companheira, e meu grande amigo Richard Ovanovik por me apoiarem mesmo em momentos de pessimismo.

Referências

Barros, R. P., Mendonça, R., Santos, D. D., & Quintaes, G. 2001. Determinantes do desempenho educacional no brasil. Rio de Janeiro, Brasil. Disponível em:<http://repositorio.ipea.gov.br/bitstream/11058/2160/1/TD_834.pdf>. Acesso em: 10 mar. 2022.

Belluzzo, W., & Moraes, A. 2014. O diferencial de desempenho escolar entre escolas públicas e privadas no Brasil. Nova economia 24: 409 - 430.

Bourdieu, P.; et al. 1998. Escritos de Educação. 1ed. Editora Vozes. Petrópolis, RJ, Brasil.

Cacciamali, M. C., & Hirata, G. I. 2005. A Influência da Raça e do Gênero nas Oportunidades de Obtenção de Renda: Uma Análise da Discriminação em Mercados de Trabalho Distintos: Bahia e São Paulo. Estudos Econômicos 35: 767-795.

Carvalho, M. 2005. Quem é negro, quem é branco: desempenho escolar e classificação racial de alunos. Revista Brasileira de Educação 28: 77 - 95.

Castro, Maria Helena Guimarães de; Tiezzi, Sergio. 2004. A reforma do ensino médio e a implantação do Enem no Brasil. Desafios 65: 46-115.

Courgeau, D. 2012. Methodology and Epistemology of Multilevel Analysis: Approaches from Different Social Sciences. 2ed. Springer. Netherlands, Alemanha.

Fávero, L. P., & Belfiore, P. 2017. Manual de análise de dados. 1ed. Elsevier, Rio de Janeiro, RJ, Brasil.

Felicetti, Vera Lucia e Cabrera, Alberto F. 2017. Resultados da Educação Superior: o ProUni em Foco. Avaliação: Revista da Avaliação da Educação Superior (Campinas) 22: 871-893.

Ferrari Bravin, g.; Lee, l.; Rissino, s. Das d. 2019. Mineração de dados educacionais na base de dados do enem 2015. Brazilian Journal of Production Engineering – BJPE 5: 186–201.

Freitas, Dirce Nei Teixeira de. 2004. Avaliação da educação básica e ação normativa federal. Cadernos de Pesquisa 34, 663-689.

Instituto de Pesquisa Econômica Aplicada [IPEA]; Fundação João Pinheiro [FJP]. Desenvolvimento humano nas macrorregiões brasileiras. 2016. Disponível em: <https://www.ipea.gov.br/portal/images/stories/PDFs/livros/livros/20160331_livro-idhm.pdf>. Acesso em: 25 mar. 2022.

Lazega, E., & A.B. Snijders, T. 2016. Multilevel Network Analysis: Sciences: Theory, Methods and Applications. 12ed. Springer. Netherlands, Alemanha.

Nogueira, Cláudio Marques Martins et al. 2017. Promessas e limites: o sisu e sua implementação na universidade federal de minas gerais. Educação em Revista 33.

Morettin, P., & Bussab, W. 2010. Estatística Básica. 6ed. Saraiva. São Paulo, SP, Brasil.

Raudenbush, S. W., & Bryk, A. S. 2002. Hierarchical linear models: Applications and data analysis methods. 2ed. sage. Thousand Oaks, CA, EUA.

Santos, C. A. S. T., Ferreira, L., Oliveira, N. F., Dourado, M. I. C., & Barreto, M. L. 2000. Modelagem multinível. Sitientibus 22: 89-98.

Tabachnick, Barbara G. e Fidell, Linda S. 2012. Using Multivariate Statistics. 6ed. Pearson Education. Upper Saddle River, New Jersey, EUA.