

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
Programa de Pós-Graduação em Estatística Aplicada

Tiago Henrique de Pinho Marques França

“NÃO SABE”, “NÃO RESPONDEU” NA INSCRIÇÃO DO ENEM: o potencial interpretativo das respostas imprecisas e ausentes

Belo Horizonte
2024

Tiago Henrique de Pinho Marques França

“NÃO SABE”, “NÃO RESPONDEU” NA INSCRIÇÃO DO ENEM: o potencial interpretativo das respostas imprecisas e ausentes

Monografia apresentada ao Programa de Pós-Graduação em Estatística Aplicada do Departamento em Estatística da Universidade Federal de Minas Gerais como requisito para obtenção do título de Especialista em Estatística.

Orientadora: Jussiane Nader Gonçalves

Belo Horizonte
2024

França, Tiago Henrique de Pinho Marques.

F814n Não sabe, Não respondeu na inscrição do ENEM: [recurso eletrônico] o potencial interpretativo das respostas imprecisas e e ausentes / Tiago Henrique de Pinho Marques França – 2024.
1 recurso online (63 f. il., color.) : pdf.

Orientadora: Jussiane Nader Gonçalves.
Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.

Referências: f. 53-55

1. Estatística. 2. Análise de regressão. 3. Ausência de dados (Estatística). 4. Exame Nacional do Ensino Médio (Brasil)
I. Gonçalves, Jussiane Nader. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2(043)



Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 333º. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE TIAGO HENRIQUE DE PINHO MARQUES FRANÇA.

Aos dezesseis dias do mês de agosto de 2024, às 08:30 horas por meio de videoconferência, os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso do aluno **Tiago Henrique de Pinho Marques França**, intitulado: “*“Não sabe”, “Não respondeu” na inscrição do ENEM: o potencial interpretativo das respostas imprecisas e ausentes.*”, como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, a Presidente da Comissão, Professora Jussiane Nader Gonçalves – Orientadora, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do candidato. Após a defesa, os membros da banca examinadora reuniram-se sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o candidato foi considerado Aprovado condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente ao candidato pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 16 de agosto de 2024.

Documento assinado digitalmente
 **JUSSIANE NADER GONCALVES**
Data: 16/08/2024 15:16:14-0300
Verifique em <https://validar.iti.gov.br>

Profa. Jussiane Nader Gonçalves (Orientadora)
DEST/UFMG

Documento assinado digitalmente
 **JULIANA FREITAS DE MELLO E SILVA**
Data: 17/08/2024 16:36:34-0300
Verifique em <https://validar.iti.gov.br>

Profa. Juliana Freitas de Mello e Silva
CIDACS / Fiocruz - BA

AGRADECIMENTOS

Ao corpo de professores da Especialização em Estatística Aplicada, em especial a Profa. Jussiane Nader Gonçalves pela orientação.

Aos colegas que compartilharam comigo o tempo e as ideias.

RESUMO

Este trabalho tem por objetivo identificar o potencial analítico dos dados imprecisos (“Não sabe”, “Não respondeu”, etc.) e dados ausentes (*missing data*) do questionário dos inscritos no Exame Nacional do Ensino Médio, em sua edição de 2022, a fim de levantar hipóteses acerca da opção por esse tipo de declaração, bem como contribuir com processos decisórios frente a dados com tais limitações. Metodologicamente, optou-se por análises descritivas e inferenciais, por meio de modelos de regressão linear, para buscar similaridades e distinções estatisticamente significativas entre os candidatos que declararam dados imprecisos e as demais categorias presentes nos quesitos. Os resultados indicam que duas circunstâncias sintetizam as tendências observadas nos dados: a dificuldade para compreender alguma pergunta do questionário, o que se associa a notas consistentemente mais baixas no exame; a presença de fatores intervenientes ou categorias ocultas nos dados. Ademais, este trabalho apresenta interpretações individuais para cada quesito com dados imprecisos ou ausentes, com destaque à cor ou raça, estado civil, dependência administrativa e localização da escola, e escolaridade e ocupação dos pais.

Palavras-chave: *missing data*; dados imprecisos; perfil socioeconômico; eficácia escolar.

ABSTRACT

This work aims to identify the analytical potential of imprecise data (“Do not know”, “Do not answer” types) and missing data from the ENEM 2022 Examination registration questionnaire, in order to raise hypothesis related to the choice of this type of declaration. Thus, it also aims to contribute to decision-making processes with data with such limitations. Methodologically, we opted for descriptive and inferential analyses, using linear regression models, to look for similarities and statistically significant distinctions between the candidates who declared inaccurate answers and the other categories present in the questions. The results indicate that two situations summarize the trends observed in the data: the difficulty in understanding a question in the questionnaire, which is associated with consistently lower grades in the exam; the presence of intervening factors or hidden categories in the data. Furthermore, this work presents interpretations for each question with imprecise or missing data, with emphasis on color/race, marital status, school administrative dependency and location, and parents' education and occupation.

Keywords: missing data; inaccurate data; socioeconomic profile; school effectiveness.

SUMÁRIO

1. DADOS IMPRECISOS E AUSENTES NO CONTEXTO DOS ESTUDOS SOBRE EDUCAÇÃO: UMA INTRODUÇÃO.....	8
2. DADOS E MÉTODOS: A AUSÊNCIA DE INFORMAÇÃO E COMO INVESTIGÁ-LA.....	9
2.1. Modelos de Regressão Linear	17
3. ANÁLISE DESCRITIVA: AS DESVANTAGENS APARENTES	18
4. ELABORAÇÃO DE MODELOS: BUSCA POR GANHO INTERPRETÁVEL	23
4.1. Exame dos modelos simples	23
4.2. Elaboração e validação dos modelos múltiplos	25
4.3. Validação e caracterização dos modelos finais	28
5. ANÁLISES INFERENCIAIS	32
5.1. Análise das covariáveis	33
5.2. Cor ou raça	35
5.3. Nacionalidade	37
5.4. Estado civil	38
5.5. Dependência Administrativa, região e localização da escola	40
5.6. Tipo de Ensino	44
5.7. Pai: escolaridade e ocupação	46
5.8. Mãe: escolaridade e ocupação	47
6. SÍNTESE DOS RESULTADOS E CONSIDERAÇÕES	51
REFERÊNCIAS	53
APÊNDICE I – Modelos lineares múltiplos desenvolvidos	56

1. DADOS IMPRECISOS E AUSENTES NO CONTEXTO DOS ESTUDOS SOBRE EDUCAÇÃO: UMA INTRODUÇÃO

A investigação científica com base no desempenho de estudantes em exames padronizados é parte fundamental de uma agenda de pesquisa que busca promover a compreensão dos determinantes e fatores associados a um bom processo de aprendizagem. Reconhecidamente, parte de Coleman (1966) investigar como as condições socioeconômicas e escolares contribuem para aquilo que se convencionou denominar por eficácia escolar.

Essa perspectiva tem grande conformidade com a elaboração teórica que a Economia da Educação estabeleceu como Função de Produção Educacional, formalizada por Todd & Wolpin (2003). Segundo essa proposição, a educação corresponde ao acúmulo de conhecimento que é produzido pelos sujeitos com base em insumos de origem familiar, escolar e comunitária. Nesse intento, a academia passou a buscar modelos que pudessem distinguir a dimensão do efeito de cada um desses fatores e suas variáveis correspondentes.

Por outro lado, estudos direcionados às desigualdades na educação baseiam-se em grande parte em dados sistematizados de pesquisas de base populacional, para promover o entendimento sobre questões como atendimento e evasão escolar, especialmente associados a clivagens sociais, com a racial (Barbosa, 2005). Por vezes, tais estudos incorporam em suas análises o desempenho dos estudantes por meio de dados de exames padronizados.

Seja pelo interesse de identificar e mensurar as desigualdades e mecanismos para superá-las, ou pela busca de fatores que desvendem as origens do acúmulo de conhecimento e da consolidação das habilidades escolares, esses intentos analíticos dependem de conjuntos de informações que contenham marcadores de distintas categorias de análise sejam elas individuais, familiares ou escolares. Logo, respostas imprecisas (como “não respondeu”, “não sei”, “não declarou” ou “não informado”) ou respostas ausentes (*missing data*) podem representar um empecilho para a obtenção de resultados substancialmente mais assertivos.

Em levantamento recente sobre estudos que empregam os dados da Prova Brasil (Ferrão; Prata; Alves, 2020), publicados entre 2000 e 2018, dos 60 artigos científicos identificados que contêm qualquer referência aos dados ausentes, incompletos ou “sem resposta”, apenas 23 mencionam explicitamente a existência de informações nessas condições em sua metodologia ou resultados, e somente em 18 houve aplicação de algum método para lidar com isso. Portanto, se bem representadas por esse cenário, as pesquisas de avaliação educacional apresentariam evidente limitação a respeito do reconhecimento e tratamento de dados nessas condições.

Embora seja uma elaboração restrita a dados ausente, atitude semelhante é percebida em produções que simplesmente ocultam ou não oferecem tratamento metodológico ou analítico para as variáveis categóricas que apresentam respostas imprecisas, conforme o quesito que pretendem inquirir.

Nesse contexto, com a finalidade de contribuir com o entendimento acerca dessa questão e promover análises que contemplem dados com essa característica, este estudo tem por objetivo identificar a possibilidade interpretativa das respostas imprecisas e ausente nos dados produzidos pela inscrição dos estudantes que prestaram o Exame Nacional do Ensino Médio (ENEM) em sua edição de 2022, por meio da relação de tais respostas com o desempenho dos candidatos.

2. DADOS E MÉTODOS: A AUSÊNCIA DE INFORMAÇÃO E COMO INVESTIGÁ-LA

O conjunto de dados selecionados para a investigação do potencial interpretativo e analítico das respostas imprecisas e ausentes nas declarações dos participantes do ENEM, em sua edição de 2022, tem origem nos microdados do exame disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP, 2023).

Em sua versão original, o conjunto de informações registradas do exame para aquele ano totaliza 3.476.105 casos, que corresponde ao número de inscritos, e 76 variáveis organizadas em seis blocos¹, segundo o dicionário que acompanha, a saber: Dados do Participante, Dados da Escola, Dados do Local de Aplicação da Prova, Dados da Prova Objetiva, Dados da Redação e Dados do Questionário Socioeconômico.

São consideradas para este estudo respostas imprecisas aquelas que não apontam de fato uma categoria a qual o indivíduo pertence dado o conteúdo original da questão, a exemplo do registro como “Não informado” para a questão do Estado civil, que não denota de fato um estado civil, que se torna assim desconhecido. Associado a registros de fato ausentes no banco de dados, denominados aqui também por *missing*. Essas duas condições se encontram, ao todo, em 17 variáveis em três diferentes blocos de informação.

A tipificação de tais respostas e sua prevalência nos dados será apresentada a seguir, tendo em consideração diferentes recortes no conjunto de dados, nos quais as análises realizadas estão circunscritas. Assim, são apresentadas tais respostas para quatro desses recortes: i. o

¹ Ademais, para o ano de 2022 também está disponível um questionário denominado Hábitos de Estudo na Pandemia, que não virá a compor as análises deste estudo pelo seu caráter excepcional, de pouca relação com os dados dos anos anteriores.

conjunto completo dos dados, que diz respeito aos sujeitos inscritos no exame; ii. os presentes de fato em ambos os dias, uma vez que só para esses se tem as notas obtidas nas provas; iii. os alunos concluintes do Ensino Médio que comparecerem de fato ao exame, uma vez que somente a eles foi requisitada declaração que gerou o bloco Dados da Escola, iv. o conjunto desses alunos que não obtiveram nenhuma nota zero nas cinco provas que realizaram.

Quadro 1 – Recortes aplicados ao banco de dados.

Recortes	Denominação	Quantidade	Percentual dos inscritos	Percentual do recorte anterior
I	Inscritos no exame	3.476.105	100%	-
II	Presentes nas provas	2.343.148	67,4%	67,4%
III	Concluintes do EM presentes nas provas	957.130	27,5%	40,8%
IV	Concluintes do EM presentes nas provas sem nota zero	908.669	26,1%	94,9%

Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

O último recorte foi adotado somente na composição dos modelos de regressão múltiplos, pois tais valores impuseram desafios para a validação dos pressupostos dos modelos. A análise exploratória foi realizada tendo como referência os concluintes do Ensino Médio presentes no exame. Assim, as análises descritivas abarcam 957.130 casos (27,5% do conjunto original dos dados), enquanto os modelos de regressão se baseiam em 908.669 casos (94,9% dos casos das análises descritivas).

Definido o escopo demarcado nos dados, faz-se necessário apresentar as variáveis presentes nos microdados que contam com alguma resposta imprecisa ou ausente, e sobre elas oferecer alguma reflexão ou mesmo hipótese relacionada ao seu registro de interesse. Para tal, será apresentado esse conjunto de informações de modo sistematizado, seguindo os blocos por meio dos quais os dados são originalmente organizados.

Quadro 2 – Variáveis com respostas imprecisas ou ausentes nos Dados do Participante.

Codificação	Definição	Tipificação	Percentual inscritos	Percentual presentes no exame	Percentual concluintes presentes
TP_ESTADO_CIVIL	Estado civil	Não informado	3,7%	3,1%	2,8%
TP_COR_RACA	Cor ou raça	Não declarado	1,9%	1,7%	1,8%
TP_NACIONALIDADE	Nacionalidade	Não informado	0,05%	0,04%	0,05%
TP_ANO_CONCLUIU	Ano da conclusão do EM	Não informado	57,7% (9,9% entre EM completo)	62,3% (8,2% entre EM completo)	-
TP_ESCOLA	Tipo de escola (pública ou privada)	Não respondeu	62,1%	59,2%	-
TP_ENSINO	Tipo de ensino (regular ou especial)	Missing	63,7%	59,9%	1,9%

Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

Nos denominados Dados do Participante, há uma distinção entre variáveis de cunho pessoal (estado civil, cor/raça e nacionalidade) e outras sobre a carreira escolar (ano da conclusão e tipo de escola e ensino que cursou o Ensino Médio) que apresentam dados imprecisos ou ausentes.

Os do primeiro tipo têm relativa baixa prevalência entre os inscritos, e ainda mais reduzida entre os sujeitos que de fato chegaram a prestar o exame, não ultrapassando 4% na variável em que esse tipo de resposta foi mais recorrente, o estado civil. A nacionalidade, no outro extremo, não foi declarada somente por 5 pessoas a cada 10.000 inscritos aproximadamente.

Por outro lado, as variáveis relativas à carreira escolar estão majoritariamente sem informação precisa entre os inscritos. No entanto, isso é em grande parte efeito da variedade de situações em que os inscritos se encontram em relação ao Ensino Médio (EM), havendo estudantes não concluintes, estudantes do último ano do EM, os que já concluíram esse grau de estudo e, em muito menor caso, aqueles que não concluíram e não são estudantes. Para os estudantes e os que ainda não concluíram essa etapa dos estudos, não há de fato o que se declarar.

Diversos podem ser os motivos que levam as pessoas a não declarar essas informações como o desconhecimento de conceitos (no caso de estado civil, ensino regular ou especial), dificuldade de se enquadrar nas categorias (no caso da cor ou raça) ou mesmo dificuldade de se

lembrar (como no ano de conclusão do EM), ou mesmo deliberada recusa de fornecer a informação por determinado motivo, seja ele pessoal ou social.

Quadro 3 – Variáveis com respostas ausentes nos Dados da Escola.

Codificação	Definição	Tipificação	Percentual inscritos	Percentual presentes no exame	Percentual concluintes presentes
CO_MUNICIPIO_ESC	Código do município	<i>Missing</i>	72,6%	70,4%	27,6%
NO_MUNICIPIO_ESC	Nome do município	<i>Missing</i>	72,6%	70,4%	27,6%
CO_UF_ESC	Código da UF	<i>Missing</i>	72,6%	70,4%	27,6%
SG_UF_ESC	Sigla da UF	<i>Missing</i>	72,6%	70,4%	27,6%
TP_DEPENDENCIA_ADM_ESC	Dependência administrativa	<i>Missing</i>	72,6%	70,4%	27,6%
TP_LOCALIZACAO_ESC	Localização (rural ou urbano)	<i>Missing</i>	72,6%	70,4%	27,6%
TP_SIT_FUNC_ESC	Situação funcional	<i>Missing</i>	72,6%	70,4%	27,6%

Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

Os Dados da Escola apresentam uma precisa consistência da prevalência de dados ausentes, sendo possível atribuir isso à identificação (ou não) da escola em que os sujeitos cursam o EM. Observa-se que informações sobre esses quesitos foram obtidos somente dos candidatos que ainda eram estudantes. Assim, há mais alta prevalência de informações ausentes entre os inscritos e os candidatos que compareceram às provas, quando comparado aos que compareceram, mas que estão no último ano do EM, com expectativas de conclusão no ano de aplicação do exame, no caso, 2022. No entanto, mesmo sobre esse grupo, 27,6% de seus componentes não têm informações sobre sua escola declaradas.

Quadro 4 – Variáveis com respostas imprecisas nos Dados do Questionário Socioeconômico.

Codificação	Definição	Tipificação	Percentual inscritos	Percentual presentes no exame	Percentual concluintes presentes
Q001	Escolaridade do pai	Não sei	9,6%	8,6%	10,0%

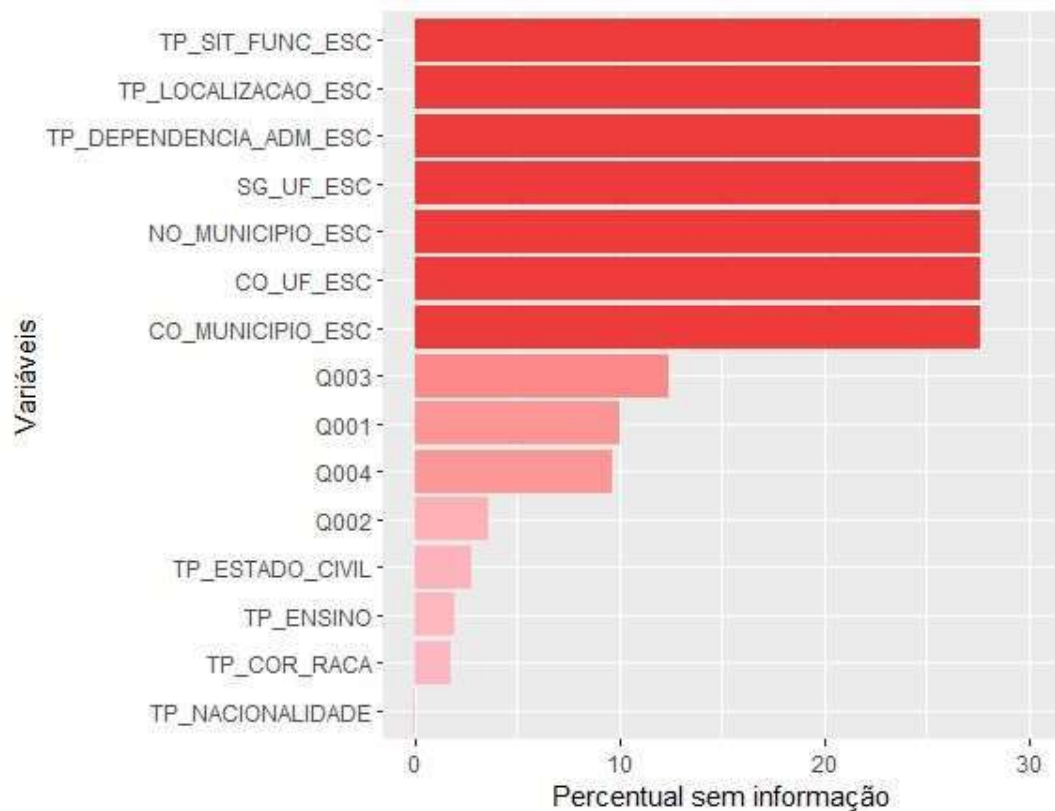
Q002	Escolaridade da mãe	Não sei	3,2%	2,7%	3,6%
Q003	Ocupação do pai	Não sei	12,1%	11,1%	12,4%
Q004	Ocupação da mãe	Não sei	8,7%	8,4%	9,6%

Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

A respeito do questionário socioeconômico, apenas quatro questões apresentam respostas imprecisas, porém potencialmente reveladoras. Essas variáveis se referem à escolaridade e ocupação dos pais ou pessoas responsáveis do candidato, e a falta de conhecimento são mais prevalentes entre os estudantes concluintes participantes do exame que dos inscritos em geral, sobretudo em relação aos seus pais (homens). Há uma variedade de possibilidades em torno desses casos que devem ser consideradas, muitas que sugerem alguma fragilidade social, como não conhecer os pais, terem os pais falecidos, não ter proximidade com eles, além de outros fatores como desconhecer de fato a informação mesmo convivendo com eles, ou simplesmente ter dificuldade de enquadrar os pais nas categorias apresentadas, sobretudo no que diz respeito à ocupação. Essas questões serão discutidas junto à análise desenvolvida.

Em perspectiva comparativa, a não declaração da escola é a resposta imprecisa ou ausente mais frequente, consequentemente todos os dados escolares advindos dessa informação também são desconhecidos. Em seguida então as informações de escolaridade e ocupação do pai e da mãe, nessa ordem. E, por fim, as informações dos dados dos participantes, sendo o estado civil não informado o mais comum, e a nacionalidade não informada menos comum.

Gráfico 1 – Percentual de respostas imprecisas ou ausentes por variável

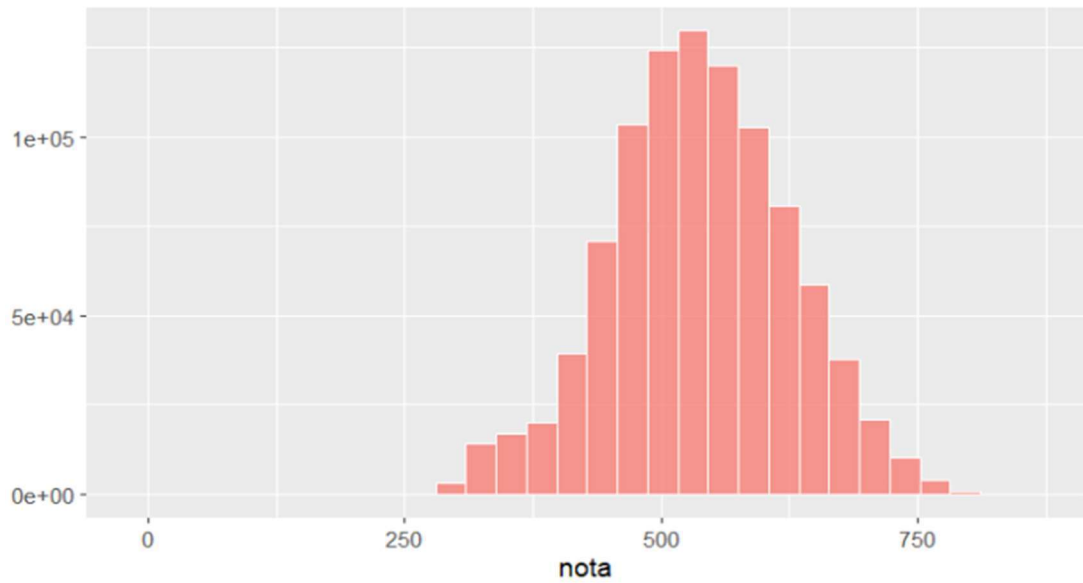


Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

É importante ressaltar que outras variáveis também estão presentes no conjunto de dados socioeconômicos declarados pelos indivíduos, como renda mensal familiar e acesso a determinados bens duráveis. Cada uma delas foi analisada para composição dos modelos empregados na análise central desta investigação. Além dessas, foi necessário elaborar uma variável que exprimisse a ideia de desempenho geral no exame.

Como as instituições de Ensino Superior têm autonomia para decidir como empregarão as notas das distintas provas que compõem o ENEM para a classificação dos candidatos pelas suas vagas ofertadas, não há um parâmetro universal para de auferir o resultado geral do exame. Entretanto, o INEP (2021) indica que a “nota média” dos candidatos pode ser obtida pela média aritmética das notas das cinco provas que compõem o exame. Esta será a variável dependente do estudo, denominada aqui simplesmente por “nota”. Potencialmente, essa pode variar entre 0 e 100 pontos, e no conjunto de dados observados para a população de interesse está circunscrita entre 0 e 855,8 pontos. Sua média é 537,4 e mediana 536. Sua distribuição se encontra no gráfico a seguir.

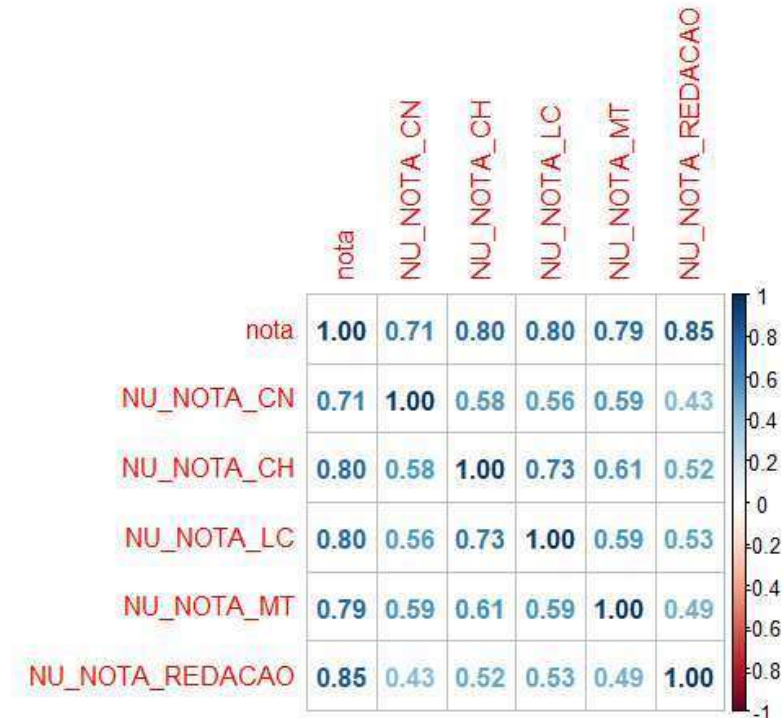
Gráfico 2 – Distribuição da variável nota



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

A correlação da média das notas do candidato (nota) com as demais notas de cada prova é necessariamente positiva, sendo mais correlacionada com a prova de Redação, Linguagens, Códigos e suas Tecnologias e Ciências Humanas e suas Tecnologias, que as provas de Matemática e suas Tecnologias e Ciências da Natureza e suas Tecnologias, sendo essa a menos correlacionada.

Gráfico 3 – Matriz de correlação da média das notas com as demais notas



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

Por sua vez, os procedimentos analíticos aplicados para compreender a possibilidade interpretativa das categorias imprecisas e respostas ausentes dos estudantes finalistas que prestaram o ENEM, bem como relacionar essas condições com o desempenho desses no exame, podem ser sistematizados em três etapas progressivas.

Inicialmente, um processo descritivo foi realizado para ilustrar a relação das respostas de interesse deste estudo com a renda familiar declarada pelo estudante e a relação com o desempenho no exame. O primeiro procedimento pode auxiliar a entender as origens socioeconômicas da não resposta. O segundo, por sua vez, auxilia a entender a posição das categorias imprecisas ou ausente dentre as demais categorias das variáveis, no que diz respeito à nota obtida.

A segunda etapa, breve e intermediária, resume-se ao emprego de cada variável potencialmente explicativa em modelos de regressão linear simples com a nota, para subsidiar estratégias e avaliar a pertinência de caminhos adotados na composição dos modelos múltiplos.

Por fim, são desenvolvidos modelos de regressão lineares múltiplos para avaliar o desempenho das variáveis com respostas imprecisas ou ausentes num contexto explicativo

multidimensional, e assim inferir o potencial analítico das realidades ocultas em tais dados, especificamente no que diz respeito ao desempenho no exame.

2.1. Modelos de Regressão Linear

A regressão linear consiste num método estatístico pelo qual se estabelece uma relação linear entre uma ou mais variáveis explicativas (independentes) e com a variável resposta (dependente), para dessa forma identificar o que se espera de alterações nessa variável, em função de mudanças naquelas (Costa, 2019).

No âmbito deste estudo, há o emprego de regressões lineares simples, isto é, quando há uma única variável explicativa, para desenvolver análises preliminares que auxiliam tomadas de decisão sobre a elaboração de modelos múltiplos que possibilitam, por sua vez, a análise que de fato corresponde aos objetivos estabelecidos. Para a composição dos modelos são ajustados, por sua vez, regressões lineares múltiplas, isto é, elaborações nas quais mais de uma variável explicativa se faz presente. Dessa forma, considerando que a amostra é composta por n indivíduos, a equação que sintetiza o modelo de regressão, considerando k variáveis explicativas, é dada por:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i, \quad i = 1, \dots, n,$$

onde:

y_i : medida da variável dependente para o i -ésimo indivíduo,

x_{ij} : medida da j -ésima variável dependente, com $j = 1, \dots, k$, para o i -ésimo indivíduo,

β_0 : coeficiente linear da reta de regressão (intercepto),

β_j : coeficiente associado à j -ésima variável explicativa, com $j = 1, \dots, k$,

e_i : termo de erro aleatório, assumindo $e_i \sim Normal(0, \sigma^2)$.

Assim, as análises centrais recaem principalmente sobre uma interpretação dos coeficientes β 's obtidos por meio dos estimadores de mínimos quadrados que sintetizam, no caso de variáveis categóricas, a diferença esperada na nota dos sujeitos entre declarar pertencimento a determinada categoria, seja em ganho ou perda, em relação à categoria de referência como inseridas no modelo (Triola, 1999). Estrategicamente, as categorias de

respostas imprecisas ou ausentes foram arranjadas como categorias de referência para que a diferença entre essa e cada uma das outras se fizesse evidente.

O processo de elaboração dos modelos, com explicação a respeito da estratégia, bem como suas principais características e processo de validação se encontra em seção posterior a esta, após a análise descritiva das variáveis de interesse.

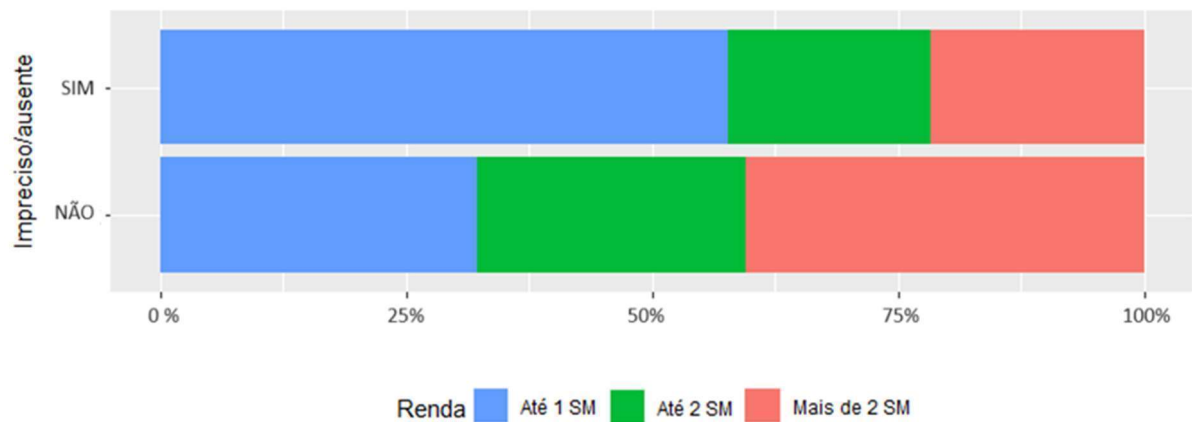
3. ANÁLISE DESCRITIVA: AS DESVANTAGENS APARENTES

A primeira etapa da análise descritiva tem por objetivo obter um entendimento simples e inicial da origem social dos estudantes que recorreram às categorias imprecisas ou que não declararam onde estudam. Para tal, em cada uma das variáveis que contém esses casos, observou-se como se encontrava a distribuição dos indivíduos pelas duas escalas categorias de renda familiar mensal², que destacam os extremos dessa distribuição, separados em dois grupos, aqueles que tinha recorrido às categorias imprecisas e aqueles que declararam qualquer uma das outras categorias.

Os resultados apontam que, de modo geral, há maior participação proporcional de pessoas de menor renda entre os sujeitos que deliberadamente marcaram não saber, não informar ou não declarar os quesitos, com alguma distinção entre as diferentes variáveis. No outro extremo da distribuição de renda, a tendência também se verifica, isto é, os candidatos que recorreram às categorias pouco elucidativas ocupam menos as faixas superiores de renda. Apenas o estado civil parece não ter distinção na proporção de faixas superiores de renda, embora ainda tenha nas superiores. Os gráficos a seguir ilustram a maior e a menor distinção observadas nessa ordem.

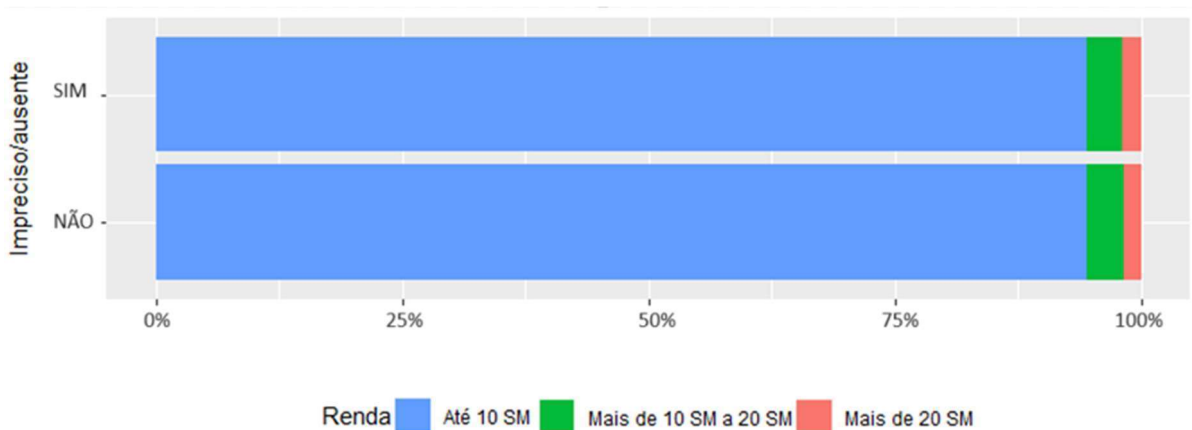
² Originalmente a variável de renda familiar mensal possui 17 faixas, e para evidenciar possíveis distinções de maior interesse, essa foi recodificada para a análise descritiva em duas variáveis. A primeira para as faixas de renda mais baixas: até um salário mínimo, mais de um salário mínimo a dois salários mínimos, e mais de dois salários mínimos. A segunda tratou de distinguir as rendas mais elevadas: até 10 salários mínimos, mais de 10 até 20 salários mínimos, e mais de 20 salários mínimos. Ambas as recodificações foram empregadas com o intuito de vislumbrar se haveria maior concentração de candidatos com mais ou com menos renda entre os que recorreram às categorias imprecisas sobre si.

Gráfico 04 – Proporção de estudantes finalistas pelas faixas de renda mensal familiar, por tipo de resposta à nacionalidade



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

Gráfico 05 – Proporção de estudantes finalistas pelas faixas de renda mensal familiar, por tipo de resposta ao estado civil



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

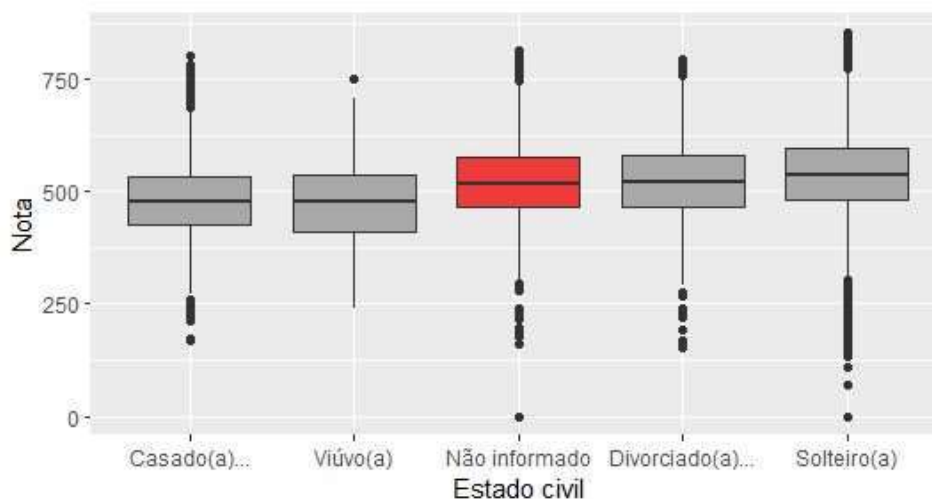
Esse resultado indica uma tendência de estudantes de famílias mais pobres recorrerem com maior frequência ao recurso de não se categorizar, contudo, não é possível afirmar se isso adviria de modo preponderante de uma decisão por ocultar informações de forma deliberada. Por outro lado, não há indício que haja especial tendência entre os candidatos advindos das famílias com maior renda de não se categorizar por meio das respostas imprecisas ou ausentes.

Frente a esse cenário, é razoável estabelecer a hipótese de que a dificuldade para o entendimento das perguntas ou opções de resposta das questões apresentadas aos candidatos no processo de inscrição do exame esteja relacionado a um desempenho mais fraco nas notas das provas, sendo ambos, o desempenho e o entendimento relacionados com indicadores de status

socioeconômicos desfavorecido, como a baixa renda familiar. O que também pode ser investigado por meio de análise descritiva no mesmo conjunto de dados.

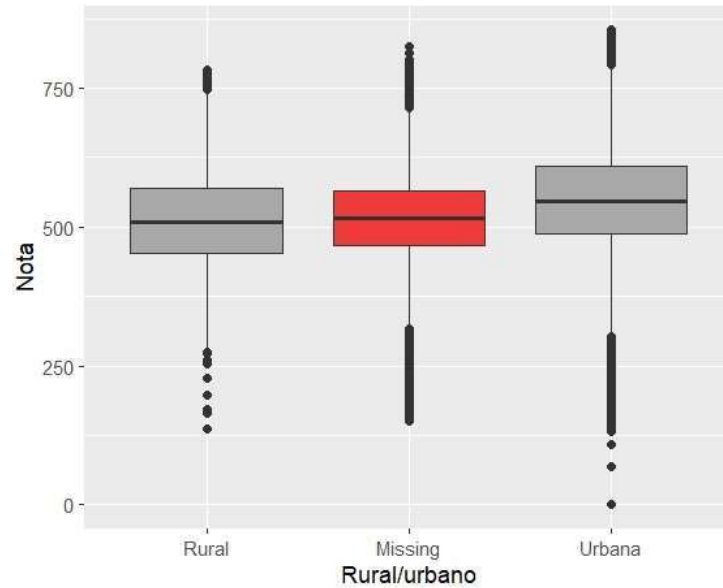
O cruzamento do desempenho no exame com as diferentes variáveis revela, de modo geral, que o registro desses casos está associado a alguma desvantagem ou das variáveis de maior prevalência. Contudo, nem sempre se aproxima das categorias de menor desempenho. Por exemplo, tem-se a média da nota das respostas imprecisas do estado civil, sendo o desempenho desse grupo localizado na posição 3 de 5 categorias, com os solteiros e divorciados com nota média superior e os viúvos e casados com nota inferior. De modo similar, os que não declararam a escola, e consequentemente não tiveram sua localização determinada, apresentam média das notas entre os estudantes de zona rural e urbana.

Gráfico 06 – Desempenho dos estudantes finalistas no exame, por estado civil



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

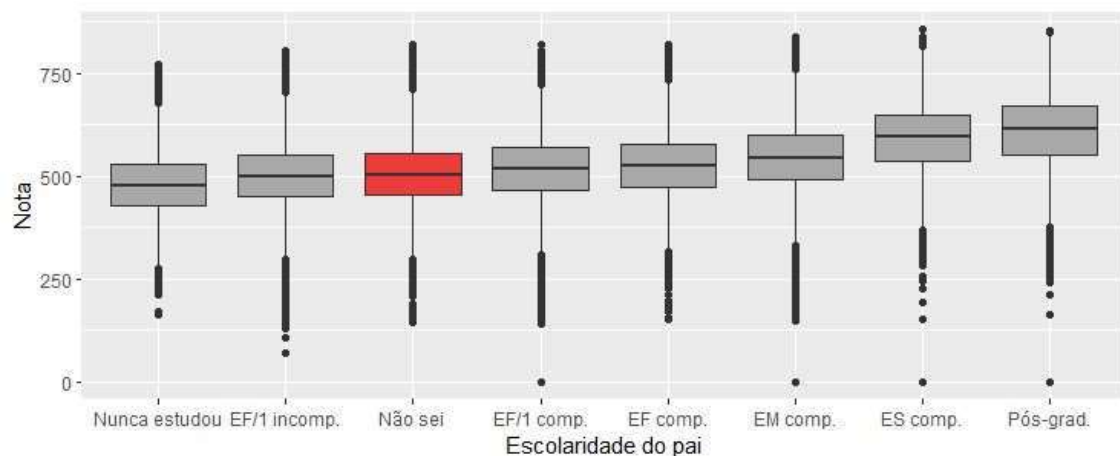
Gráfico 07 – Desempenho dos estudantes finalistas no exame, por localização da escola



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

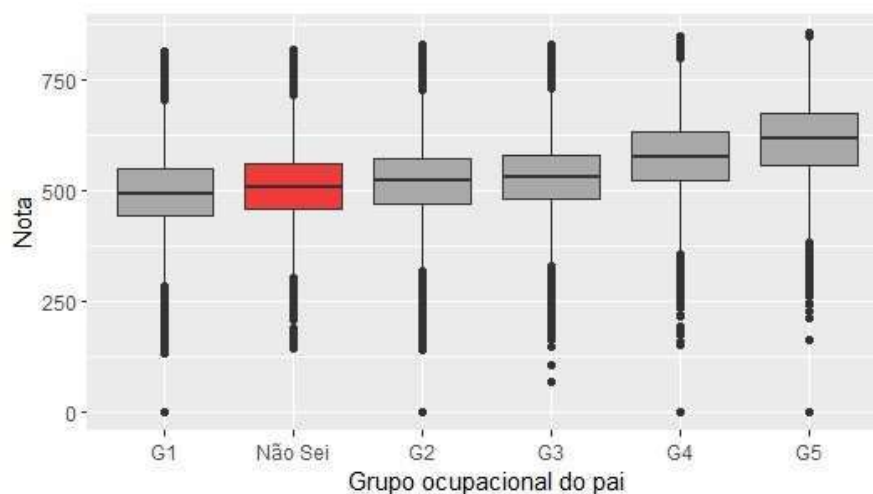
Em outros casos, aparentemente, se encontram em uma posição mais próxima da desvantagem, como nas informações sobre os pais dos candidatos, cuja média das notas das respostas imprecisas se aproximam das categorias de posição inferior na distribuição ordinal das categorias de tais variáveis.

Gráfico 08 – Desempenho dos estudantes finalistas no exame, por escolaridade do pai



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

Gráfico 09 – Desempenho dos estudantes finalistas no exame por grupo ocupacional do pai



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

O desempenho no exame, obtido em conjunto a partir da nota dos estudantes, para as questões relacionados a seus pais, revela que as circunstâncias associadas à declaração de não saber a escolaridade do pai, se aproxima do desempenho médio de quem tem o pai com a primeira etapa do Ensino Fundamental³ incompleta. No que diz respeito ao grupo ocupacional, se encontraria entre o Grupo 1, que conta com trabalhadores do campo (como lavrador, bóia fria, pescador e extrativista) e o Grupo 2, de ocupações de serviços que não exigem qualificação (como faxineiro, repositor de mercadoria, atendente de loja, etc.)⁴. Situação similar ocorre com os resultados de desempenho em relação às declarações sobre as mães. Mas, no que tange à escolaridade, o “não saber” apresenta um desempenho médio ainda inferior, entre as categorias “não estudou” e a primeira etapa do Ensino Fundamental incompleta.

Ao se levar em consideração o cruzamento das respostas ausentes imprecisas com as faixas de renda, bem como com o desempenho médios dos estudantes, torna-se evidente que alguma desvantagem seja econômica ou educacional se associa às opções por essas respostas, e que em nenhuma das treze variáveis investigadas descritivamente essas condições aparecem como associada a categorias vantajosas ao desempenho do aluno.

Embora sejam importantes para indicar as tendências associadas às respostas imprecisas ou ausentes nos dados declarados pelos candidatos do ENEM, esse conjunto de resultados

³ Denominada 4ª série ou 5º ano do Ensino Fundamental.

⁴ O Grupo 3 trata de profissionais de serviços e comércio (como padeiro, operário de fábrica, caminhoneiro, pedreiro, etc.), o Grupo 4 de prestadores de serviço com mais qualificados (professor da educação básica, policial militar de baixa patente, pequeno comerciante, etc.), e o Grupo 5 de profissionais ainda mais distintos (médico, engenheiro, advogado, professor universitário, diretor de empresa, etc.).

descritivos não são suficientes para sustentar interpretações dedutivas acerca da interferência dessas declarações sobre o desempenho no exame, mas, certamente apontam para a pertinência de se dar sequência a essa investigação por meio de análise inferencial, neste caso, com o desenvolvimento de modelos de regressão linear.

4. ELABORAÇÃO DE MODELOS: BUSCA POR GANHO INTERPRETÁVEL

Esta seção se dedica ao processo de elaboração de modelos de regressão linear, seguido de análise de consistência para validação desses e caracterização ampla dos modelos obtidos. Inicialmente, serão tratados os resultados dos modelos de regressão linear simples que empregam cada uma das variáveis que podem contribuir para elucidar o desempenho no exame, com ênfase interpretativa naquelas que tem categorias de repostas imprecisas ou ausentes. Esse processo tem por objetivo subsidiar decisões para a composição de um modelo que serve de ponto de partida (Modelo Base), e para a confecção dos modelos finais (Modelo Região e Modelo Localização), que de fato subsidiam as análises centrais deste trabalho. O processo de desenvolvimento para obtenção de tais modelos é, em seguida, apresentado conjuntamente com as especificações que permitem a sua validação. Todas as etapas foram realizadas tendo como recorte o conjunto de dados estudantes concluintes do Ensino Médio presentes em todas as provas do exame do ano de 2022, com exceção da validação, e consequentemente análise, dos modelos finais de excluem os concluintes que obtiveram nota zero em qualquer uma das notas⁵. Por fim, é apresentada a análise de resultados obtidos a partir dos modelos finais.

4.1. Exame dos modelos simples

Os modelos simples das variáveis de potencial explicativo foram utilizados para que se tivesse possibilidades de antever problemas e traçar estratégias quanto à inserção, exclusão e transformações necessárias de variáveis.

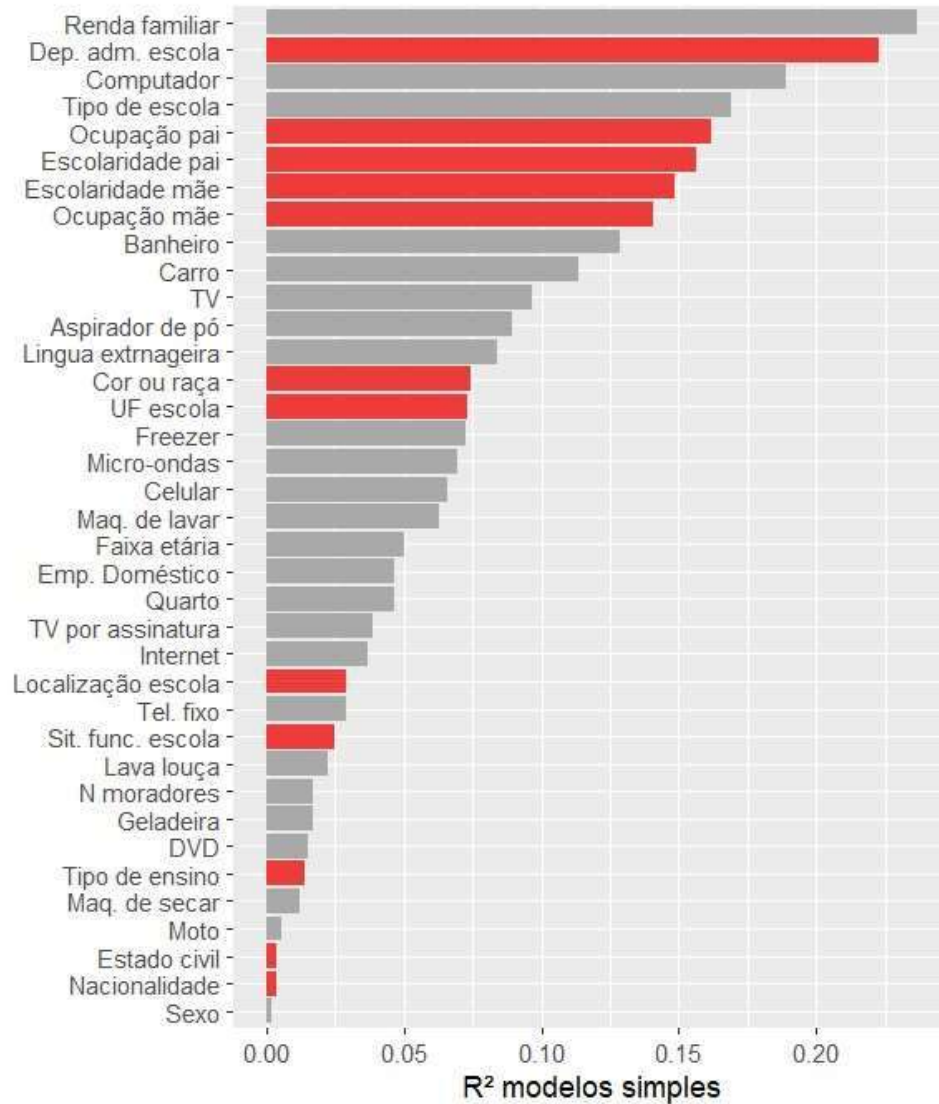
O conjunto de resultados das variáveis de interesse indica que, de modo geral, há indícios significativos de distinção da não resposta ou das repostas imprecisas em relação às

⁵ Esse novo delineamento foi necessário uma vez que os modelos que partem da nota média não tiveram a capacidade de oferecer valores ajustador que pudessem aproximar de extremos como zero. Essa disparidade entre valores zero observados no banco e os valores ajustado levou na prática à grandes limitações no processo de validação, que resultou na exclusão desses casos (5,1% dos casos) na elaboração dos modelos finais, bem como sua validação e análise de resultados.

demais categorias para a definição da nota, o que corroboram com algum potencial explicativo dessas. No entanto, para alguns casos, esse tipo de resposta não se distingue significativamente de uma das demais categorias presentes, são elas: afirmar-se divorciado (quanto ao estado civil), ser declaradamente pardo (nas definições de cor ou raça) e estudar numa escola do Amapá (no conjunto das unidades federativas). Essas três categorias reiteram a aproximação da distribuição das notas dos sujeitos que recorreram a respostas imprecisas ou ausentes daqueles associados a condições mais comuns (ser pardo) ou desvantajosas (ser divorciado ou estudar no Amapá).

Ademais, a análise também levou a algumas decisões estratégicas para dar sequência à elaboração do modelo para a análise central, são elas: i. o abandono de empregar o município da escola do estudante na análise pela enorme variação categórica nos dados; ii. o abandono da variável sobre a situação funcional da escola por apresentar registros incongruentes; iii. considerar a recategorização das unidades federativas da escola em regiões geográficas do país (especialmente em combinação com outras variáveis); iii. o tratamento da variável que assinala a quantidade de moradores na residência passa a ser numérico.

Os modelos simples, através do conjunto de coeficientes de determinação também foram utilizados para definir a ordem e a estratégia para a elaboração dos modelos de regressão linear múltipla. Os valores dessa medida apresentam variação, conforme ilustra o gráfico a seguir.

Gráfico 10 – Coeficiente de determinação (R^2) dos modelos lineares simples⁶

Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

4.2. Elaboração e validação dos modelos múltiplos

Para a definição dos modelos múltiplos adotou-se a estratégia de inicialmente elaborar um modelo base a partir das variáveis de interesse do estudo para, em seguida, carregá-lo com mais variáveis explicativas. O principal desafio para a elaboração de tal modelo consistiu em lidar com a multicolinearidade presente nos dados da escola, uma vez que todos os casos assinalados como *missing* eram coincidentes entre as variáveis desse grupo. Ou seja, o estudante que não declarou sua escola não tem os dados de dependência administrativa, unidade da

⁶ Com destaque, em vermelho, às variáveis que contêm respostas imprecisas ou ausentes no banco de dados.

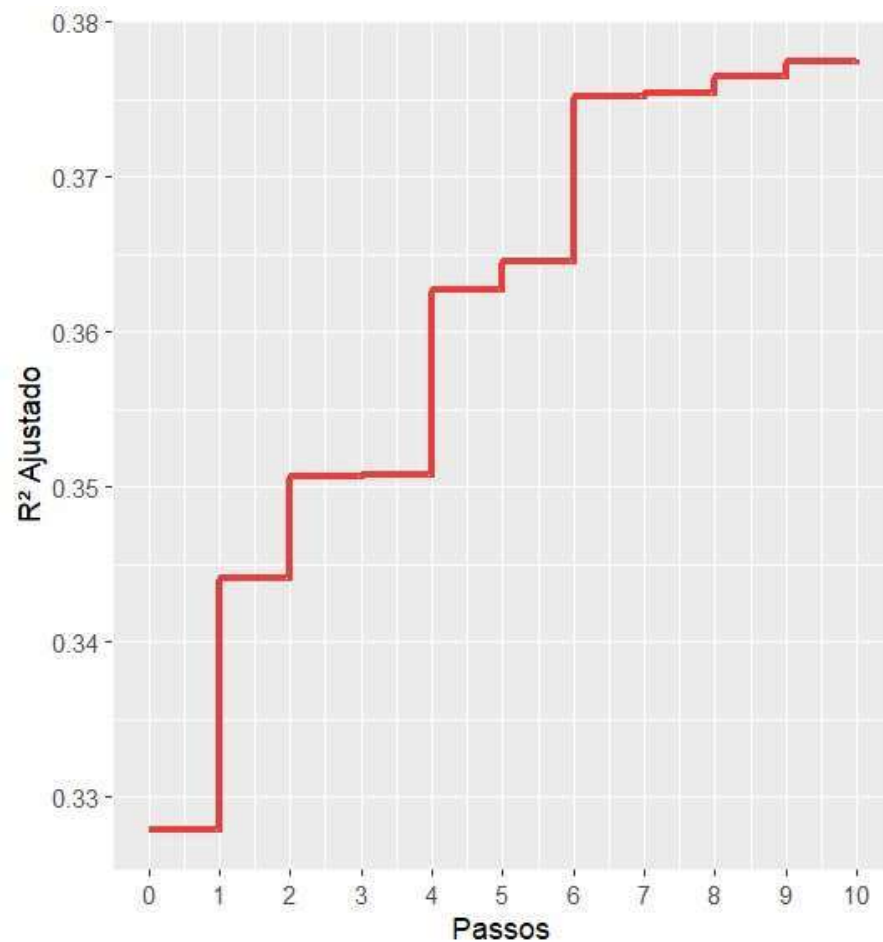
federação, localização (entre rural e urbana) ou da situação funcional, e município da instituição, embora as duas últimas já tenham sido descartadas.

Para sanar a questão, e obter resultados das três variáveis relacionada à escola, optou-se por agregar duas dessas variáveis em uma só, associando a dependência administrativa à região (com base na UF da instituição), assim obtendo uma variável com 21 categorias que combinam cada dependência administrativa (federal, estadual, municipal e privada) com cada região do país (norte, nordeste, centro-oeste, sudeste e sul), além da manutenção do *missing*. Essa variável, denominada a partir de então *admreg*, foi utilizada com as demais variáveis com categorias imprecisas para compor o modelo base. E depois de alcançado o primeiro modelo múltiplo (Modelo Região) foi substituída pela combinação da dependência administrativa com a localização, em nova variável (*admloc*), acrescida de unidade da federação onde a prova foi realizada para a consolidação de outro modelo múltiplo (Modelo Localização), para que a distinção entre rural e urbana também pudesse ser contemplada na análise.

O processo para se alcançar o Modelo Região se deu pela tentativa de inserção consecutivas de cada uma das demais variáveis explicativas ao Modelo Base, segundo a ordem em que possivelmente apresentariam maior potencial de contribuição de explicação da variabilidade da nota, como explicita o Gráfico 10. A pertinência de cada variável elegível a inclusão foi avaliada tendo em consideração dois critérios que deveriam ser necessariamente contemplados. O primeiro consiste em ter ganho observado no alcance explicativo do modelo, verificado por meio do coeficiente de determinação ajustado a cada inclusão pleiteada. O segundo é a ausência de qualquer alteração dos coeficientes das variáveis anteriormente inseridas que levasse ao prejuízo ou contrassenso interpretativo.

Ao fim do processo, além das variáveis de interesse para este estudo, foram incluídas outras dez variáveis cujos ganhos na capacidade explicativa do modelo encontram-se ilustrados no Gráfico 11.

Gráfico 11 – Ganho com a inclusão de cada uma das covariáveis adicionadas ao Modelo Região



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

Por ordem de inclusão, foram adicionadas ao modelo: renda domiciliar mensal, quantidade de computadores, quantidade de banheiros, língua estrangeira de opção do candidato (entre inglês e espanhol), quantidade de telefones celulares, faixa etária do candidato⁷, se há acesso à internet na residência, quantidade de residentes no momento da inscrição, quantidade de motocicletas e sexo. Assim, foram elaborados o Modelo Região e, com a alteração previamente descrita, o Modelo Localização, que se encontram na íntegra em apêndice (Apêndice I), assim como o Modelo Base.

⁷ A variável conta com idades unitárias entre 17 e 25 anos, e se refere à idade do inscrito em 31/12/2022.

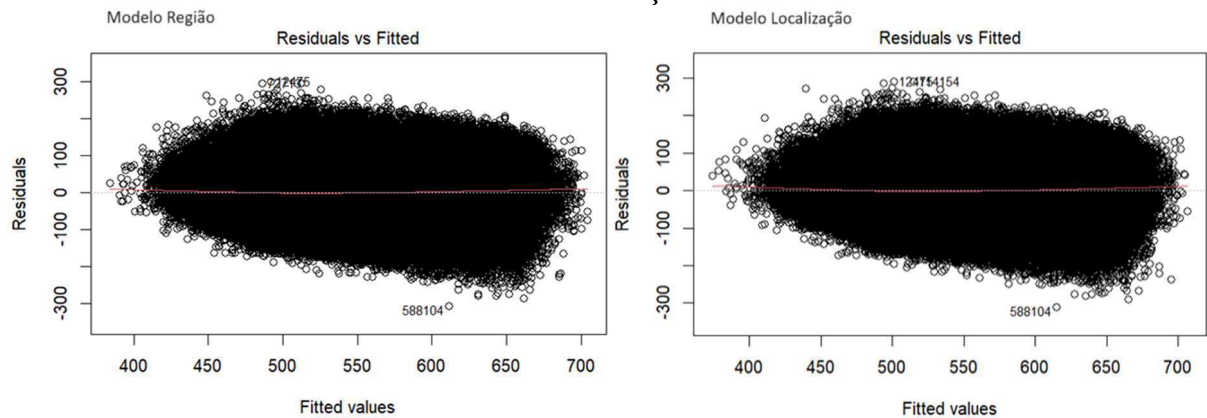
4.3. Validação e caracterização dos modelos finais

Os modelos finais obtidos foram verificados quanto aos requisitos para um ajuste adequado por modelagem linear. Isso inclui a verificação do *Variance Inflation Factor* (VIF) para atestar a possibilidade de redundância explicativa entre as variáveis selecionadas, isto é, a ocorrência de multicolinearidade.

O exame dessa medida aponta duas variáveis com VIF superiores à marca 5, que indica um grau relevante de multicolinearidade. Essas correspondem à renda familiar total mensal (5,68 no Modelo Região, e 5,81 no Modelo Localização) e ao grupo ocupacional do pai (5,59 no primeiro modelo, e 5,65 no segundo). Era esperado que houvesse alguma sobreposição ou redundância no modelo, sobretudo na variável relativa à renda, uma vez que o conjunto das demais variáveis, em grande parte, está relacionada à sua obtenção, como o grupo ocupacional, ou acesso a bens e serviços relacionadas com ela, como computadores e internet. No entanto, optou-se por manter ambas as variáveis pela sua importância na composição do modelo, sendo a renda a de maior potencial explicativo, e o grupo ocupacional do pai uma das variáveis de interesse deste estudo.

Por sua vez, a homoscedasticidade pode ser ilustrada a partir da análise gráfica dos valores ajustados conjuntamente com seus resíduos. Nos casos dos modelos obtidos, é notada uma flutuação dos resíduos em torno do valor zero, mas que apresenta algum nível de desbalanceamento. Por outro lado, o grande volume de dados dispostos no gráfico não permite uma impressão justa a respeito de concentrações. Especialmente, há indício de heteroscedasticidade mais evidenciados em valores ajustados abaixo de 450 e por volta de 650, o que se relaciona com as tendências do modelo apresentadas mais à frente ainda nesta seção.

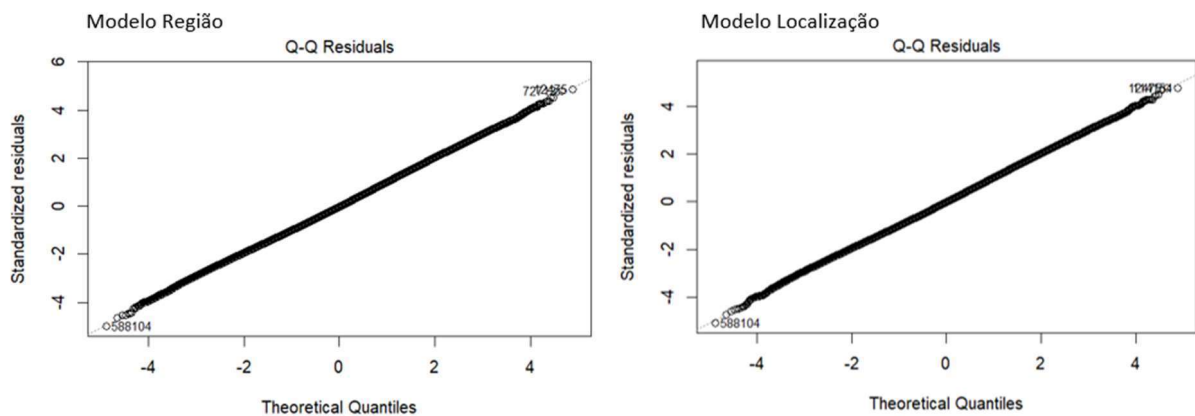
Gráficos 12 e 13 – Valores ajustados *versus* resíduos dos modelos Região e Localização



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

Determinante para a validação do ajuste do modelo, a normalidade dos resíduos também foi investigada graficamente e por meio do teste estatístico. A expressão gráfica do *Quantile-Quantile Plot* (Q-Q Plot) para os resíduos indica uma grande concentração de pontos em padrão diagonal que reforçam uma tendência à normalidade, para ambos os modelos, embora sejam desafiador atestar visualmente algum serpentar entorno do eixo.

Gráficos 14 e 15 – Q-Q plot dos modelos Região e Localização

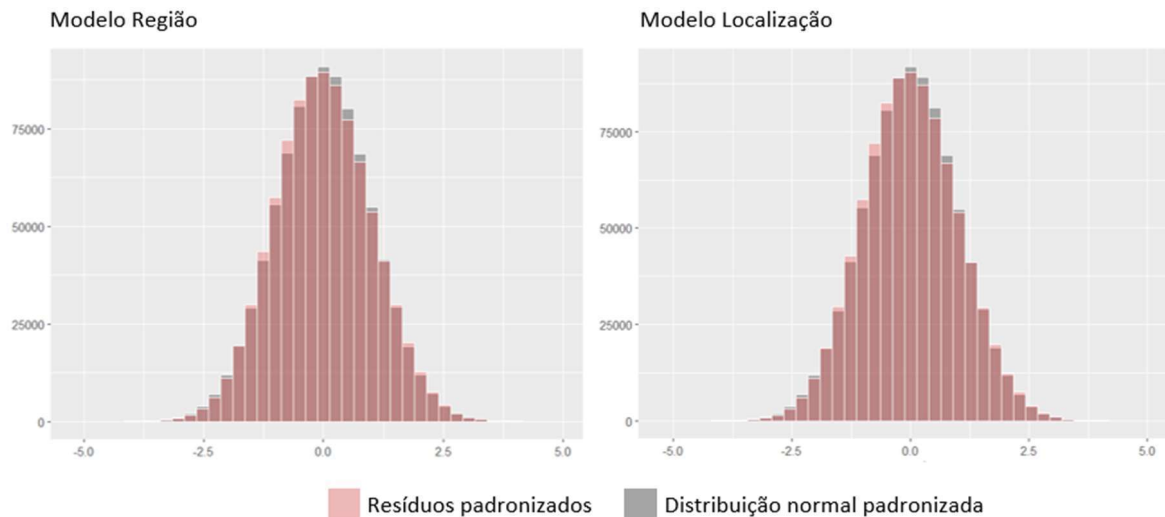


Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

Alternativamente, pode-se buscar identificar a normalidade por meio de disposição dos valores padronizados para os resíduos em comparação com uma curva normal padronizada, elaborada com o mesmo volume de casos. Esse esforço se encontra representado nos

histogramas a seguir, nos quais uma aproximação à normalidade é evidente. Essa configuração é semelhante nos dois modelos.

Gráficos 16 e 17 – Histogramas dos resíduos padronizados e da distribuição normal padronizada



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

O teste de Shapiro-Wilk amplamente empregado com a finalidade de verificar normalidade em determinada distribuição também foi aplicado. Contudo, dada a limitação da execução do teste por meio de rotina empregada no software R, que se restringe a um conjunto máximo de 5.000 observações, foram executados mil testes para cada modelo com subamostras aleatórias cujo tamanho foi fixado nesse limite máximo. Os resultados indicam que a maior parte das amostras de resíduos poderiam ser enquadradas como uma aproximação da distribuição normal, mas isso se verifica sobretudo no Modelo Região.

Quadro 05 – Síntese dos resultado dos testes de Shapiro-Wilk.

Condição	Modelo Região	Modelo Localização
Amostra P-valor > 0,05	635 (63,5%)	515 (51,5%)
Amostra P-valor > 0,01	398 (39,8%)	275 (27,5%)

Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

De modo geral, ao se considerar o resultado dos testes e a associação com os resultados gráficos, pode-se afirmar que há uma aproximação à distribuição normal que possibilita a validação do modelo para fins interpretativos de seus parâmetros.

Os modelos desenvolvidos apresentam uma capacidade explicativa da variabilidade da nota no exame que se encontra entre 38% e 40%, como denotam os coeficientes de determinação obtidos.

Quadro 06 – Coeficientes de determinação dos modelos Região e Localização.

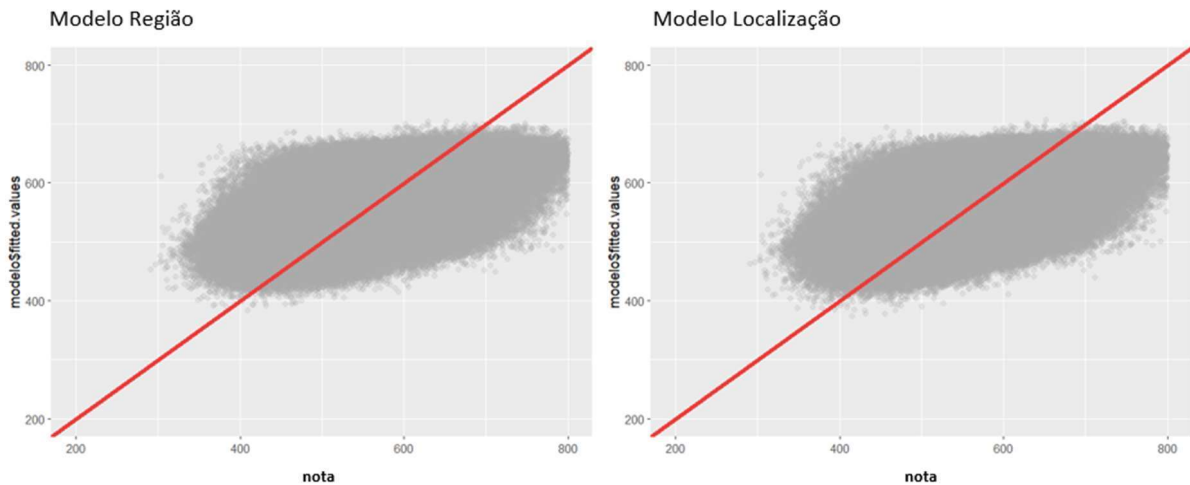
Coeficiente	Modelo Região	Modelo Localização
R^2	0,3847	0,3916
R^2 Ajustado	0,3846	0,3915

Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

Dado o esforço de verificação dos ganhos na possível inclusão de cada uma das variáveis socioeconômicas presente no banco de dados do ENEM para o ano de 2022 na elaboração do modelo, é possível que esse patamar seja um indicativo de proximidade a um platô para esse ano do potencial explicativo da nota média no exame para aquele ano. Dessa maneira, outros fatores socioeconômicos que escapam aos dados, e sobretudo outros fatores que não podem ser enquadrados nessa classificação explicariam a maior parte da variação do desempenho dos estudantes finalistas no exame. Contudo, essa afirmação carece de mais investigação para ser atestada ou generalizada.

Por fim, outro fator que caracteriza os modelos obtidos é sua tendência a regredir notas de valores extremos a uma aproximação da nota média, ou seja, o modelo tende a reduzir as notas daqueles que obtiveram notas relativamente altas e, sobretudo, a ajustar notas mais elevadas para quem de fato obteve notas especialmente baixas.

Gráficos 18 e 19 – Nota *versus* valores ajustados pelos modelos Região e Localização



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

Essas são as tendências a respeito dos modelos obtidos, e permitem se levantar hipóteses sobre a excepcionalidade (seja de notas altas, mas sobretudo de notas baixas) estarem especialmente fora do alcance explicativo dos fatores socioeconômicos, ou pelo menos daqueles declarados pelos candidatos. Outro fator que pode colaborar para que esse resultado ocorra consiste numa concentração de categorias mais frequentes e esvaziamento de categorias menos frequentes, ao se limitar os dados ao recorte dos estudantes finalistas do Ensino Médio, reduzindo de modo evidente a variabilidade etária, conjugal, entre outras.

Embora a tendência para a concentração das notas ajustadas mais próximas à média se faça expressamente evidente nos dados, isso não desqualifica o modelo obtido em sua interpretação sobre as variáveis que os compõem. Isso é especialmente recomendado ao se contemplar quão significativas foram das diferenças observadas nos coeficientes das variáveis, e tendo em conta que o objetivo central deste trabalho não consiste na busca pela explicação global do desempenho no exame, mas na análise dos dados imprecisos e ausentes.

5. ANÁLISES INFERENCIAIS

Nessa seção são apresentados os resultados dos modelos e oferecidas interpretações sobre o quão significativo são as categorias sociais expressas nas variáveis que os compõem. Um breve comentário sobre os resultados das covariáveis é tecido, para se conhecer as

tendências que elas representam, tendo como parâmetro os coeficientes obtidos no Modelo Região. Em seguida, o resultado de cada variável que contém respostas imprecisas ou ausente é tratada também segundo sua contribuição para a modulação do desempenho do estudante no exame, considerando os coeficientes de ambos os modelos, apresentados aqui em seus valores inteiros mais próximos.

5.1. Análise das covariáveis

O sexo do candidato, conforme declarado no ato da inscrição no exame, tem baixa importância segundo a sua expressão, tal como inserido no modelo, ou seja, sem qualquer arranjo interativo com outro quesito. Seu resultado aponta para uma diferença pouco significativa, estimada como inferior a um ponto no resultado do exame, a considerar a nota média das provas. Essa pouco evidente vantagem de menos de 1 ponto seria para a mulheres.

A idade dos estudantes, por sua vez, aparece como um elemento dotado forte significância, como as demais variáveis a seguir. Desse modo, qualquer ganho de idade a partir dos 17 representa uma perda no desempenho, o que se agrava, sobretudo, a partir dos 19 anos. Enquanto a diferença de quem tem 17 e 18 anos se aproxima de 4 e 8 pontos a menos, em comparação a quem tem a idade abaixo dessas, essa perda salta para 29 pontos para quem tem 19 anos e 35 pontos para quem tem 20 anos. Contudo, é necessário retomar que essa seria condição entre estudantes finalistas do Ensino Médio, de modo que essa distinção na relação idade-série entre estudantes frequentemente pode estar associada à reprovação na carreira escolar, bem como abandono e reingresso, quando jovem ou mesmo adulto. Assim, tais resultados podem apresentar divergências com aqueles esperados para o conjunto de todos os inscritos no exame.

Por seu turno, a opção pela língua espanhola no ENEM representaria, segundo o modelo, uma associação a um resultado inferior na grandeza aproximada de 19 pontos, em comparação com a opção pelo inglês. Em estudo incipiente em uma escola municipal de Linhares (ES) (Mendes; Nunes, 2019), com 555 estudantes finalistas no Ensino Médio, indica que, mesmo tendo a língua inglesa ocupando mais espaço no currículo escolar, a predileção pela língua espanhola no contexto do exame se dá principalmente pela acreditada facilidade na leitura e interpretação de texto, enquanto os optantes pelo inglês indicavam o contato e o gosto pela língua, derivados de maior proximidade e estudo com o idioma. Ou seja, aqueles que optaram pela língua inglesa tinham, em geral, mais oportunidades ou dedicação no estudo da língua estrangeira que optaram.

Também se faz significativa a quantidade de moradores no domicílio do candidato. A única variável inserida no modelo como numérica indica que a partir da média observada (3,88) espera-se uma redução próxima de 2 pontos para cada pessoa a mais que vir a compor o domicílio.

Identificada como a variável com maior relação com a nota no ENEM, a faixa da renda familiar total mensal está configurada em 17 categorias que variam progressivamente de “Nenhuma renda” a “Acima de R\$ 24.240,00”. Tendo a primeira como referência no modelo, todas se apresentaram significativas, com uma diferença mínima aproximada de 8 pontos (entre quem não declarou renda alguma e quem declarou até R\$ 1.212,00), e a diferença máxima de 55 pontos em comparação com a faixa mais elevada de rendimentos.

A quantidade de banheiros instalados na residência também segue uma leitura intuitiva de suas categorias ordinais que variam segundo o ganho de uma unidade, desde não haver banheiro a quatro ou mais⁸. A correspondência na nota, segundo o modelo ajustado, seria de um ganho de 3, 4, 5 e 8 pontos na nota média do exame respectivamente, em relação a quem não os têm.

Por outro lado, à primeira vista expressando coeficientes contraintuitivos, a quantidade de motocicletas, seguindo as mesmas faixas de valores, entre não ter e ter quatro ou mais, indica que quanto maior a quantidade de motos no domicílio, menor seria a nota do estudante, com redução aproximada de 4, 9, 21 e 30 pontos, consecutivamente. É possível que esses valores, bastante expressivos, possam estar relacionados a questões socioeconômicas intervenientes relacionados à motocicleta, mais do que o impacto do veículo em si sobre a vida dos estudantes. É sabido que, além do transporte particular de elevado risco (Corgozinho; Montagner; Rodrigues, 2018)⁹, esses veículos são frequentemente relacionados a inserções laborais precárias (Souza, 2022), como os entregadores de mercadorias de pequeno volume e transporte de passageiros, na função de mototaxista. Esse serviço é mais comum em regiões periféricas e de grande adensamento populacional, como morros e favelas. Ademais, nas áreas rurais, regiões também relacionadas de modo geral a status socioeconômicos desfavorecidos, as motocicletas se fazem fortemente presente como meio de transporte entre os trabalhadores do campo (Fiúza, et al., 2023). Nesse sentido, ter motos em casa, e principalmente ter mais de uma, denotaria, sobretudo, a proximidade com essas realidades sociais e laborais.

⁸ Essas faixas se mantêm nas demais variáveis de acesso a bens duráveis, com as opções de resposta: “Não tem”, “um”, “dois”, “três” ou “quatro ou mais”.

⁹ Este estudo epidemiológico indica que as vítimas fatais de acidentes com motocicletas possuem um perfil predominantemente masculino, jovem, pardo e com baixa escolaridade.

Por fim, compuseram os modelos três variáveis de acesso a aparelhos tecnológicos e informação, com a quantidade de aparelhos celulares e computadores, em variáveis dedicadas, e o acesso à internet no domicílio, em variável dicotômica. Como esperado, ter esses aparelhos em casa tende a elevar a nota do participante, especialmente acesso a computadores com expressivos ganhos aproximados de 11, 18, 23 e 27 pontos a dependente da quantidade de aparelhos. Curiosamente a quantidade de aparelhos celulares também estaria relacionado a uma elevação do desempenho, no entanto, nos domicílios em que se tem apenas dois desses aparelhos (12 pontos aproximados de vantagem em relação a não ter nenhum), o ganho se faz maior em relação aos que tem três (em torno de 10 pontos) ou quatro ou mais celulares (9 pontos). O acesso à internet em casa elevaria o desempenho no teste em um pouco mais de 4 pontos, em relação a quem não a tem. Todavia, pelo conjunto de covariáveis que compõem o controle do modelo, tais como acesso aos aparelhos celulares e a renda familiar, o resultado específico sobre o acesso à internet precisa ser interpretado com cautela, não atribuindo sobre si todas as implicações de se ter acesso à internet em ambiente doméstico relacionadas com o desempenho no exame.

5.2. Cor ou raça

O sistema de classificação social utilizado no ENEM está em uso no Brasil há mais de quatro décadas, e é amplamente adotado em registros administrativos públicos de toda natureza, mas sua origem se não deu a partir da formalização ou reconhecimento das categorias sociais cotidianamente utilizadas pela população à época (Osorio, 2003). Isso pode levar as pessoas a demonstrarem uma dificuldade de enquadrar-se em uma das categorias (branca, preta, parda, amarela ou indígena), que não contemplam noções populares como morena, morena clara, clara ou mulata (Osorio, 2003).

O termo pardo é aquele que goza de menor legitimidade social, sendo por vezes mal compreendido, questionado ou mesmo rechaçado (Souza, 2019). Por outro lado, esse termo que condensa a premissa da miscigenação racial é também o de maior prevalência nos inquéritos nacionais de base populacional na atualidade. Entre os inscritos no exame, os autodeclarados pardos passaram a ser o maior grupo, ultrapassando os brancos na segunda década deste século.

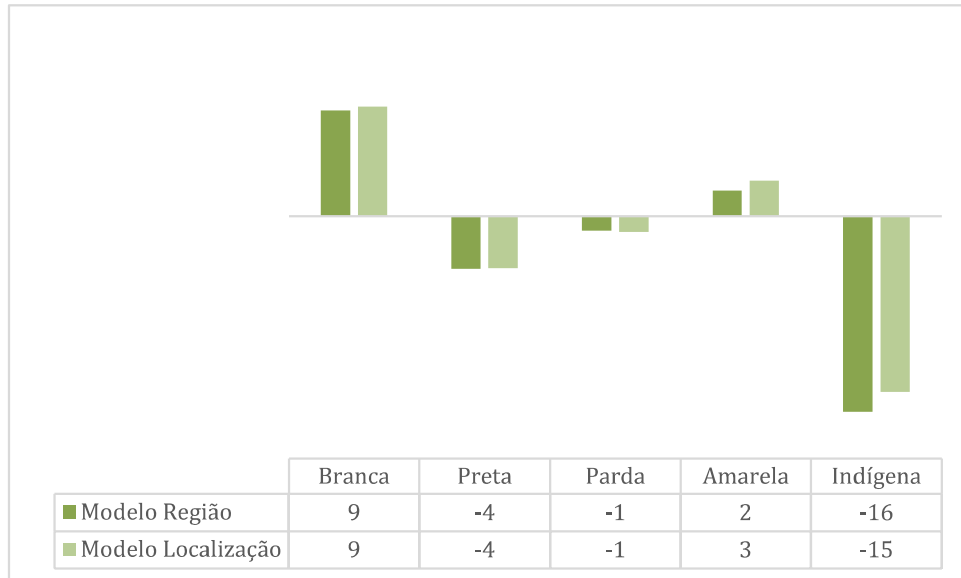
Por outro lado, a classificação adotada permite uma distinção entre pessoas socialmente reconhecidas como negras (pretas e pardas) e brancas que recorriam em outra ocasião a categorias racialmente menos demarcadas, como morena. Embora por convicção ou mesmo má-fé, desde o advento de políticas afirmativas com recorte racial, é possível identificar pessoas

brancas que se valem dessa classificação (Costa; Schucman, 2022). Outro efeito esperado é que o próprio desenvolvimento dessas políticas e os debates sociais em torno delas levaram a um maior conhecimento e apropriação das categorias do sistema de classificação pelos sujeitos.

Fortemente associado ao processo de desenvolvimento de tais políticas, ganhou espaço no país uma agenda de pesquisa interessada nas desigualdades raciais no âmbito educacional (Barbosa, 2005), que trata de fatores como desigualdade na origem social dos estudantes, qualidade das escolas e professores, representações sociais racializadas, preconceitos e processos discriminatórios no ambiente escolar. Sendo o ENEM atualmente o maior mecanismo de ingresso no Ensino Superior no país, a classificação racial dos candidatos ganha contornos ainda mais relevante para as análises, propostas e avaliação das políticas de acesso a esse nível de ensino.

A princípio, pelos resultados do modelo simples, somente com a variável cor/raça como regressora da nota, era esperado que não houvesse distinção estatisticamente significativa da distribuição da nota das pessoas autodeclaradas pardas com aquelas que não declararam sua cor ou raça. Contudo, com o controle pelas demais variáveis socioeconômicas presentes nos modelos múltiplos, todas as categorias de cor ou raça demonstraram contraste significativo com a não declaração, ainda que para os pardos, seja aproximadamente apenas de um ponto na nota média das provas.

Gráfico 20 – Coeficientes dos modelos Região e Localização para a variável cor/raça



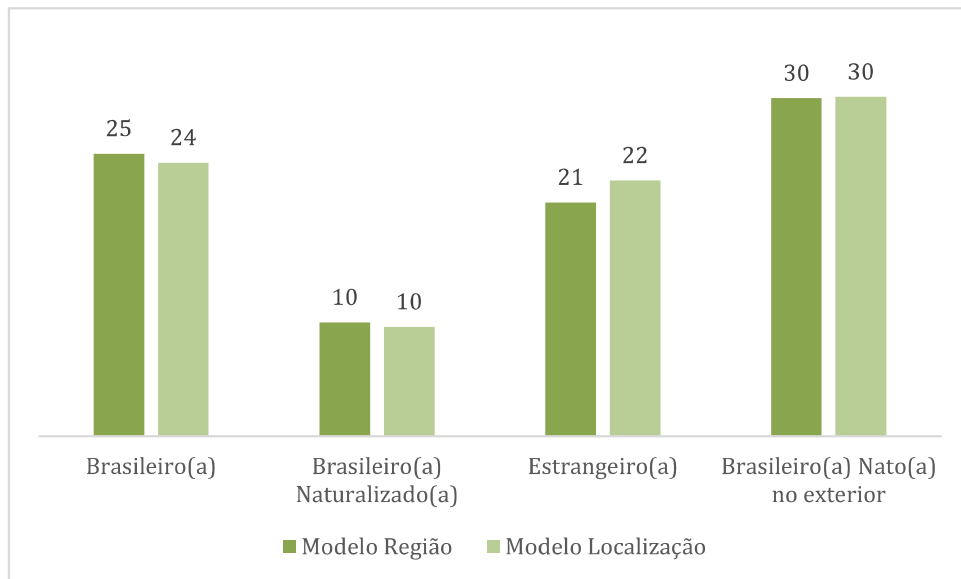
Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

Em suma, para o ano de 2022, o resultado indica relativa desvantagem na nota para as pessoas que não declaram sua cor ou raça, em comparação com as pessoas brancas e, em menor magnitude, com as pessoas amarelas. Por outro lado, relativa vantagem em relação às pessoas pretas e indígenas. Esse resultado reforça a elaboração teórica de alguma não conformidade de parte dos sujeitos com o sistema de classificação racial, daqueles que não se entendem como brancos ou preto, cuja associação racial com a nota muito se aproxima à dos pardos.

5.3. Nacionalidade

A respeito da nacionalidade e o desempenho no exame, não se encontram estudos sistematizados com esse enfoque. Os resultados dos modelos, por sua vez, apresentam significativa distinção entre os estudantes que não informaram em qual das categorias se enquadravam e todas as outras condições declaradas, tendo um desempenho significativamente inferior.

Gráfico 21 – Coeficientes dos modelos Região e Localização para a variável nacionalidade



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

Como se espera que estrangeiros de fato ou brasileiros naturalizados, devido aos seu estatuto frente à burocracia brasileira, tenha conhecimento da sua condição em relação à nacionalidade no Brasil, a hipótese que se sobressai a respeito dessa não definição consiste em ter dificuldades conseguir distinguir a diferença entre as categorias apresentadas, de modo que tanto brasileiros quanto estrangeiros que conhecem seus status teriam vantagem sobre esse candidato.

Dessa maneira, o resultado alcançado ganha sentido, já que não conseguir identificar-se dentre as categorias de nacionalidade seria um indicativo em si de dificuldades para leitura e compreensão de texto, habilidade transversal necessária para um bom desempenho em todas as provas que compõem o exame.

5.4. Estado civil

O casamento na infância e adolescência no Brasil tem forte relação com outros problemas sociais, como gravidez precoce e abandono escolar (Wodon *et al.*, 2019), e afeta majoritariamente meninas. Mesmo podendo ser relacionada a diversos motivos, tais como reduzir o opróbrio sobre o comportamento sexual e gravidez e o desejo pela liberdade ou reconhecimento comunitário como pessoa com agência própria (Costa; Freitas, 2019; Veiga; Loyola, 2020), essas uniões estão fortemente relacionadas com a pobreza, de tal modo que se estima que o incremento na renda familiar advindo com o Programa Bolsa Família tenha efeito significativo para protelá-las (Vasconcelos; Griebeler, 2023).

Por outro lado, não se espera que seja essa a realidade comum dos finalistas do Ensino Médio que compareceram às provas do ENEM e que não se declaram como solteiros, que em sua maioria do sexo masculino. Até os 21 anos completos a amostra contém 25.952 casos (aproximadamente 2,8%) de pessoas que não informaram seu estado civil, o que corresponde a 2,76 vezes o número das pessoas casadas (ou que moram com o companheiro), separadas e viúvas, conforme declarado. Espera-se também que jovens que passaram por esses processos de união e encerramento dessas relações tenha mais conhecimento sobre seu estado civil e menos dificuldade para declará-lo, que aqueles que não passaram, uma vez que, sobretudo quando formalizadas, essas relações implicam na mudança do estado civil e necessidade de lidar burocraticamente com isso.

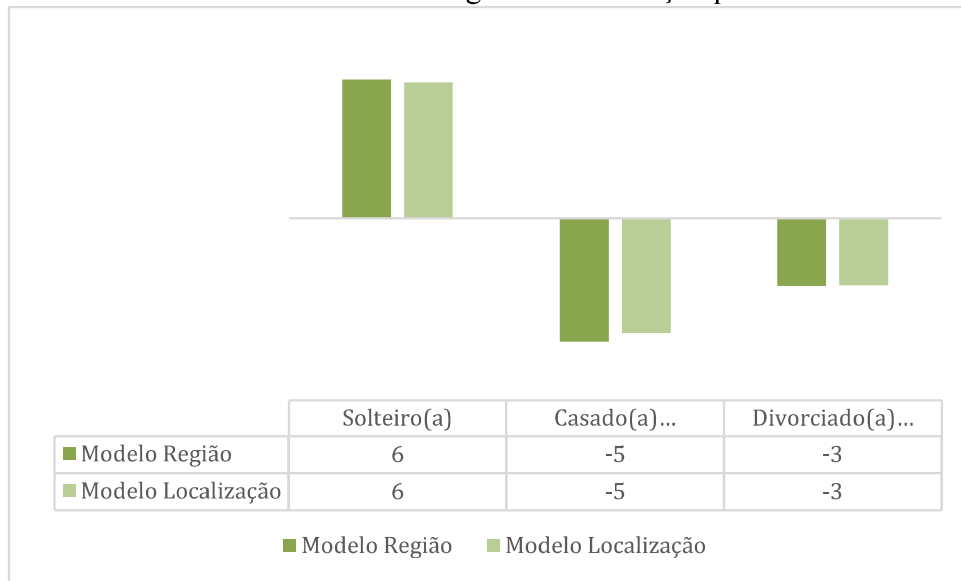
Assim, o estado civil não informado desse público, mesmo que possa cobrir um universo complexo de possibilidades, pode representar uma dificuldade de entender o que seria estado civil (especialmente para quem não teve que lidar com alterações em seu estado) ou dificuldade de conseguir identificar em qual seria seu estado civil frente às opções dadas se estiver em algum relacionamento, como namorados ou noivos, por exemplo.

O modelo de regressão simples indicou que não haveria distinção significativa entre o desempenho no exame das pessoas que não informaram seu estado civil e o dos divorciados ou separados¹⁰. Já, os modelos de regressão múltipla não conseguiram distinguir de formas significativa esse grupo dos viúvos¹¹. Contudo, pelas razões apresentadas anteriormente, seria pouco sensato creditar a não identificação do estado civil a casos de estudantes separados ou viúvos.

¹⁰ P-valor 0,256.

¹¹ P-valores de 0,464 e 0,385, para o Modelo Região e Modelo Localização, respectivamente.

Gráfico 22 – Coeficientes dos modelos Região e Localização para a variável estado civil



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

Os resultados dos coeficientes indicam que aqueles que já passaram por mudança em seu estado civil, deixando de ser solteiro, têm um desempenho inferior aos solteiros e àqueles que não declaram seu estado (casados, em união, divorciados e separados). Porém, os solteiros também teriam vantagem significativa (de 6 pontos) em relação aos que não declaram seu estado civil. O resultado sugere, portanto, um desempenho relativo intermediário ou uma possível perspectiva intermediária sobre o entendimento da conjugalidade.

5.5. Dependência Administrativa, região e localização da escola

Esta seção destina-se à análise dos dados escolares, e tem como ponto central a não declaração da escola onde se estuda. A partir de tal declaração foi possível investigar a região (a partir da unidade federativa registrada no banco de dados), a localização, que contrapõe escolas em endereços rurais e urbanos e a dependência administrativa da escola, que distingue escolas municipais, estaduais, federais e privadas. O Modelo Região conjuga a dependência administrativa com a região do país, e o Modelo Localização associa a mesma variável à localização.

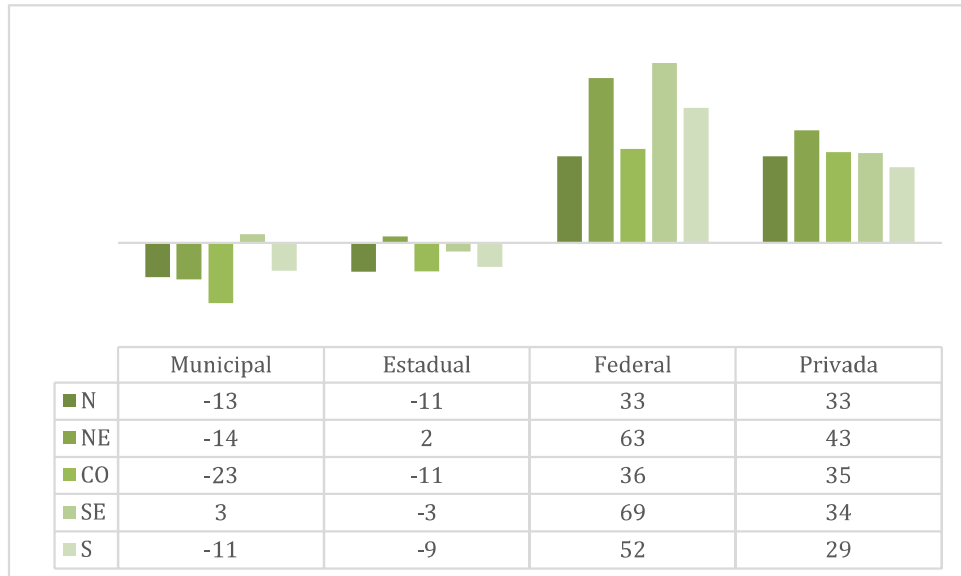
A dependência administrativa é, segundo as análises dos modelos simples, a segunda variável com maior potencial explicativo desse conjunto de dados, atrás somente da renda familiar, enquanto outros estudos indicam que esse seria o fator de maior importância na determinação das notas (Lima; Brighenti, 2023), a depender do recorte geográfico e temporal

empregado. Em análise sintetizada entre escolas públicas e privadas, também se constatou que o hiato (*gap*) atribuído ao sistema de ensino está presente não somente tendo a média como ponto de referência da análise, mas também os percentis 10, 20, 25 e 75, e que essa distância entre públicas e privadas aumentava à medida que se passava dos percentis mais baixos para o intervalo mais elevado, nas notas do exame ocorrido em 2017 (Feijó; França, 2021).

Por sua vez, diversos são os fatores que estão na gênese das desigualdades regionais da educação, mas, sabidamente o desenvolvimento de políticas para tal (e a falta delas) atravessa a todos eles (Komatsu, 2017). As disparidades históricas no investimento no ensino secundário e superior nas diferentes regiões brasileiras ganharam atenção pública de impacto somente nas últimas décadas com a expansão e interiorização das instituições de Ensino Superior (Diniz; Vieira, 2015), com destaque aos institutos e universidades federais.

Os resultados obtidos indicam que todas as categorias provenientes do cruzamento da região da escola com sua dependência administrativa apresentaram distinção significativa do conjunto daqueles que não declararam a sua escola atual. Contudo, os coeficientes do modelo variam entre ganhos e perdas relativas em relação às outras denominações.

Gráfico 23 – Coeficientes do Modelo Região para a combinação da dependência administrativa da escola e sua região



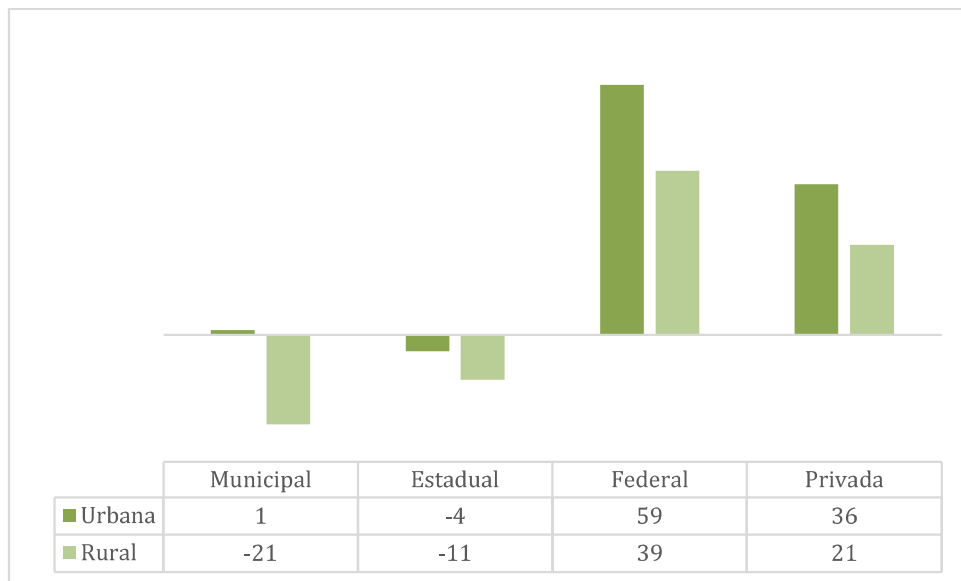
Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

As menores distâncias observadas entre a não declaração e as demais categorias por meio do modelo são do efeito esperado das escolas municipais do sudeste (que têm 3 pontos de vantagem), as estaduais do nordeste (2 pontos a mais) e as estaduais do sudeste (três pontos de desvantagem). No entanto, as maiores distâncias relativas são observadas nos escores que denotam a educação federal (69 pontos a mais no sudeste) e privada (43 pontos a mais no nordeste). Esse cenário aproxima a não declaração da escola às circunstâncias mais comuns em relação aos quesitos dependência administrativa e região, isto é: ser de escola pública localizada no sudeste ou nordeste do país.

Já, sobre a dicotomia rural e urbano, faz-se importante destacar que as escolas rurais apresentam, de modo geral, circunstâncias associadas a piores condições de aprendizado ou rendimento escolar, dentre elas: dificuldades de acesso à instituição, menos professores habilitados, forte adoção de classes multisseriadas, entre outros (INEP, 2007). Além disso, a infraestrutura encontrada nas escolas rurais possui grande defasagem em relação às de área urbana, com menos bibliotecas, laboratórios de informática, acesso à internet e quadras esportivas (Pereira; Castro, 2021).

Além disso, com as estratégias do ensino à distância na educação básica, adotadas especialmente para atender a continuidade do ensino durante a pandemia de Covid-19, detectou-se uma grande limitação dos estudantes das escolas rurais em fazer essa adaptação (Souza; Ramos, 2020). Fator que pode agravar ainda mais as disparidades esperadas entre as escolas rurais e urbanas em seu desempenho no ENEM ocorrido em 2022.

Gráfico 24 – Coeficientes do Modelo Localização para a combinação da dependência administrativa da escola e sua localização



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

O Modelo Localização produziu resultados condizentes com as desigualdades previamente documentadas entre as escolas rurais e urbanas, com disparidades evidentes no desempenho dos alunos oriundo dos sistemas públicos e privados. O modelo não identificou distinção significativa entre os estudantes que não declararam sua escola e aqueles em escolas municipais de áreas urbanas¹². Por outro lado, encontraria sua maior desvantagem com as escolas federais urbanas (que teriam 59 pontos de vantagem), e maior vantagem em contraste com as rurais municipais (com teriam 21 pontos a menos).

Esse resultado reforça uma tendência de posição mais ordinária para aqueles que não declararam a escola onde estuda, tanto por não haver detecção de diferença significativa entre esses e o grupo das escolas municipais de área urbana, tanto pela diferença relativamente pequena, de 4 pontos de vantagem, em comparação aos estudantes das escolas estaduais também de áreas urbanas.

Nesse sentido, esse conjunto de variáveis oferece pouco subsídio para levantar hipóteses sobre possíveis elementos para a promoção de um significado ou razão para não se declarar a escola.

5.6. Tipo de Ensino

¹² P-valor associado à categoria de 0,217.

Além da falta de registro da escola, há também respostas ausentes (*missing*) no banco de dados para o tipo de ensino no qual o candidato estudou durante o EM, cujas opções eram “Ensino Regular” e “Educação Especial - Modalidade Substitutiva”, para os inscritos em 2022. Nesse contexto, Educação Especial diz respeito especificamente a processos e suportes educacionais específicos para pessoas com deficiência, transtornos de desenvolvimento e altas habilidades, nos termos da legislação vigente.

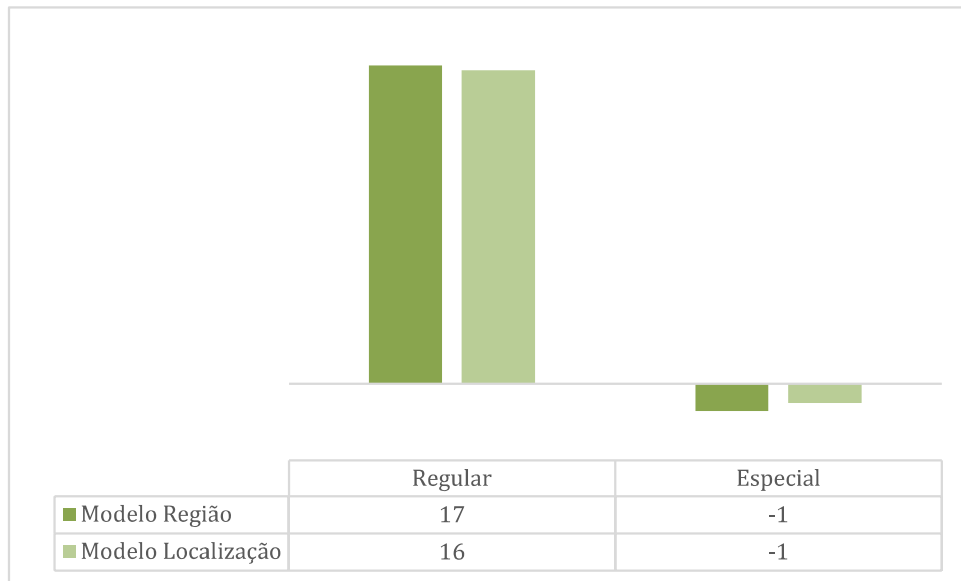
A Lei nº 9.394/1996, que institui as Diretrizes e Bases da Educação Nacional, conceitua a educação especial como modalidade de educação escolar oferecida preferencialmente na rede regular de ensino, para alunos com deficiência, transtornos globais do desenvolvimento e altas habilidades/superdotação. De acordo com a Política Nacional de Educação Especial na Perspectiva da Educação Inclusiva – MEC/2008, a Resolução CNE/CEB nº 04/2009 e o Decreto nº 7611/2011, a educação especial é uma modalidade de ensino que perpassa todos os níveis, etapas e modalidades, realiza o atendimento educacional especializado, disponibiliza os serviços e recursos e orienta quanto a sua utilização no processo de ensino e aprendizagem nas turmas comuns do ensino regular (INEP, 2014).

Aponta estudo anterior (França; Rios-Neto, 2012) que o impacto das deficiências e seus fatores associados sobre a progressão escolar pode ser constatado ao longo da carreira educacional, contudo, seus efeitos, embora tipicamente negativos, de modo geral, são tão diversos quão diversas são as categorias de deficiências. Porém, tais efeitos seriam regressivos à medida do avanço na carreira escolar, pela adaptação dos estudantes com deficiência, assim como pelo elevado abandono escolar. No contexto do ENEM, constatou-se que os candidatos com deficiência advêm de contextos socioeconômicos favorecidos¹³, e se beneficiam (ou dependem) mais desse *background* que seus pares sem deficiência (Duarte, 2020).

Por sua vez, os resultados obtidos com os modelos elaborados para este estudo apresentam evidente distinção entre os que declararam ser oriundos da educação especial em comparação ao ensino regular, entre 17 e 18 pontos, a depender do modelo. No entanto, deve-se elucidar que nem todo estudante com deficiência faz uso ou tem acesso à educação especial, de modo que o desempenho das pessoas com deficiência não pode ser inferido somente pela performance das pessoas egressas dessa modalidade de ensino.

¹³ Predominância de estudantes de cor branca com pais mais escolarizados, renda familiar relativamente mais elevada, com maior acesso a computador em casa, e com maior prevalência de matrículas no ensino privado.

Gráfico 25 – Coeficientes dos modelos Região e Localização para a variável tipo de ensino



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

Por outro lado, os modelos não conseguiram evidenciar significativa distinção¹⁴ entre as pessoas egressas da educação especial e aquelas que não tem a informação declarada. Nesse sentido, uma série de circunstâncias podem ter sido abrigadas tanto nas respostas ausentes como na própria categoria “Educação Especial - Modalidade Substitutiva”, que pode não ser facilmente compreendida por uma parte dos inscritos. Assim, além das pessoas com deficiência que estão inseridas na educação regular, mas têm acesso a dispositivos da educação especial, a dificuldade em se enquadrar nas categorias pode também advir de estudantes de outras modalidades de ensino menos frequentes, como supletivos, Educação de Jovens e Adultos (EJA), Ensino Médio de tempo integral ou mesmo integrado ao ensino técnico, que não se reconheceriam de imediato como parte do “ensino regular”.

Um indício de que essa declaração pode ser alvo de confusão para os inscritos, sobretudo para aqueles que não advêm do modelo de ensino mais comum, consiste na alteração desta questão no processo de inscrição para o exame na sua edição de 2023, em que houve a inclusão da categoria “Educação de Jovens e Adultos”. Desse modo, reduziu-se esse fator de ambiguidade.

5.7. Pai: escolaridade e ocupação

¹⁴ P- valores de 0,198 e 0,356, respectivamente, para os modelos Região e Localização.

Desde a publicação do Relatório Coleman (1966), trabalho seminal para os estudos de eficácia escolar, estabeleceu-se a importância do contexto socioeconômico familiar dos alunos como o principal determinante do aprendizado, quando mensurado a partir de testes padronizados. As características escolares figurariam em segundo plano¹⁵. Desde então, o nível de instrução e ocupação dos pais, bem como a renda familiar, passaram a figurar como elementos incontornáveis nos estudos sobre desempenho educacional.

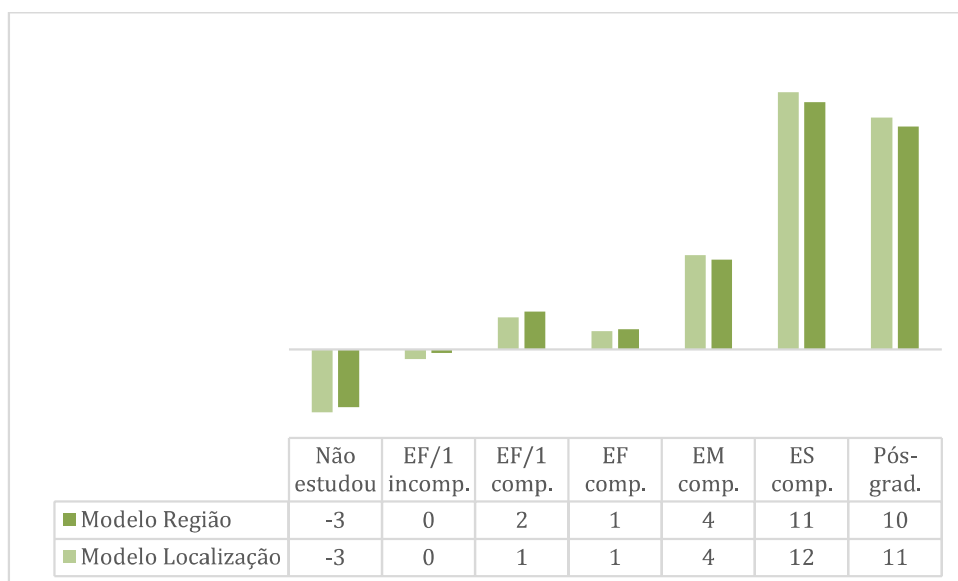
Isso pode se dar, a partir da perspectiva analítica associada ao manejo científico liberal, associada à Teoria do Capital Humano (Becker, 1964), que, em síntese, entende a educação como uma espécie de fundo familiar, no qual, quanto maior o investimento, maior seria o retorno esperado, quando os jovens ingressarem no mercado de trabalho. Ou ainda interpretada numa perspectiva galgada em noções menos quantificáveis, como aquelas baseadas na edificação de um capital cultural e simbólicos, que os pais transmitiriam aos filhos por meio da atitude e habilidades, que moldam seu modo de agir e interpretar as questões educacionais. Essa perspectiva contempla ainda a relação afetiva dos pais com o ambiente escolar, a exemplo do estudo de Souza (2009) sobre a baixa eficácia na transmissão de valores relacionados à educação da mais baixa classe socioeconômica do país.

Desse modo, pelo investimento financeiro, capacitação para orientação educacional ou transmissão de expectativas e familiaridade com as questões escolares, os pais com inserção laboral mais qualificada e maior nível educacional produziram maior sucesso educacional de seus filhos. E os pais com menos recursos laborais e educativos teriam menos soluções econômicas e culturais contra o abandono da carreira escolar de seus filhos. Evidentemente, numa perspectiva simplificada que desconsidera o papel do Estado, das políticas públicas e outros agentes no desenvolvimento da educação.

Os resultados obtidos pelos modelos coadunam com a premissa que é persistente a reprodutibilidade das desigualdades educacionais em âmbito intergeracional, sobretudo quando o pai não tem Ensino Médio e, principalmente, Ensino Superior completo. Isso ocorreria à medida que a escolaridade do pai (ou o homem que lhe é o responsável), assim como sua inserção laboral, impactasse a possibilidade de ingresso dos filhos no Ensino Superior, por meio do ENEM.

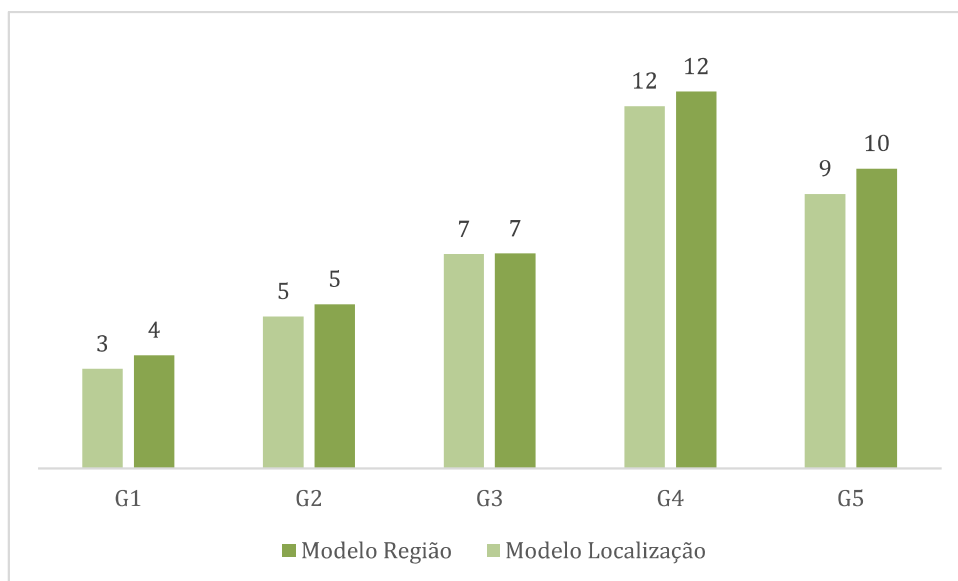
Gráfico 26 – Coeficientes dos modelos Região e Localização para a variável escolaridade do pai ou homem responsável

¹⁵ O fator escola seria mais relevante para o desempenho dos alunos, quanto menos recursos tivessem suas famílias.



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

Gráfico 27 – Coeficientes dos modelos Região e Localização para a variável grupo ocupacional do pai ou homem responsável



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

Sobre contraste entre as pessoas que declararam a escolaridade e ocupação do pai e as que registraram não ter conhecimento a respeito, todas as categorias indicam algum nível significativo de distinção, com exceção da primeira etapa do Ensino Fundamental incompleta¹⁶. Os candidatos que não declararam a escolaridade pai (ou outro homem responsável) teriam vantagem somente em comparação àqueles que declaram que essa figura não estudou (em 3 pontos aproximadamente).

Substancialmente, diversos podem ser os motivos que levam o inscrito a não ter conhecimento da escolaridade seus pais, entre eles: não o conhecer, tê-lo como uma figura distante, ausente ou mesmo falecida, além de simplesmente desconhecer a informação no ato da inscrição. Além dessas circunstâncias, para a ocupação ainda se somam a dificuldade de enquadramento nas categorias e o desemprego de longa duração, além de nunca ter trabalhado ou mesmo ser dedicado ao trabalho doméstico. Tais razões levariam a uma prevalência de “não sei” da ocupação que é superior ao da escolaridade.

Segundo levantamento recente¹⁷, 6,6% das crianças nascidas entre agosto de 2022 e julho 2023 não têm o pai no registro. Situação que leva a uma sobrecarga de cuidado para o restante da família, em especial, das mulheres. Espera-se que nessas circunstâncias as possibilidades de suporte à educação dos filhos sejam possivelmente comprometidas. Ainda que para a declaração do inscrito no exame a resposta possa ser do pai ou do homem responsável

¹⁶ Isto é, pais que não concluíram a 4ª série (hoje 5º ano) do Ensino Fundamental.

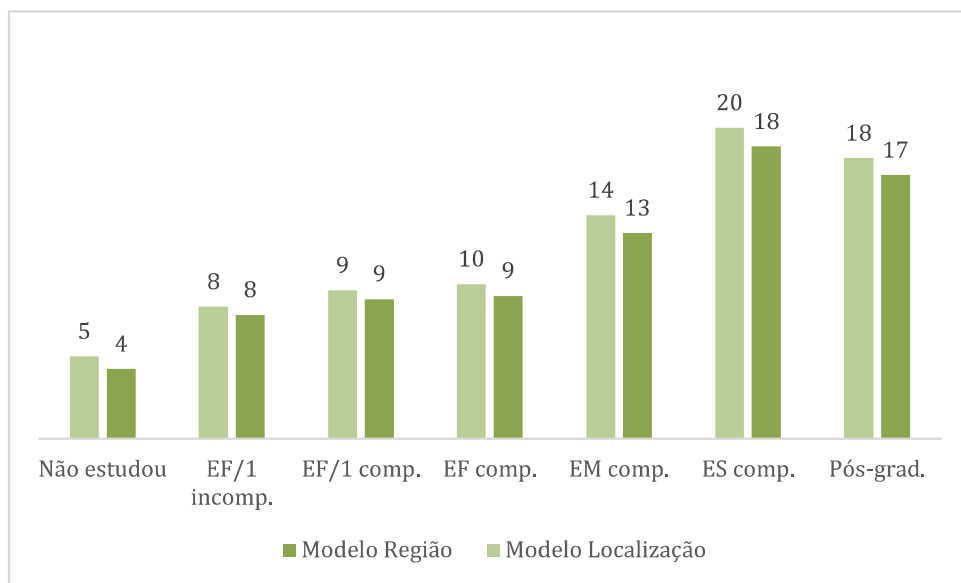
¹⁷ Segundo notícia da Associação Nacional dos Registradores de Pessoas Naturais (ARPEN, 2023).

pelo candidato, há uma diferença ainda evidente da falta de informação deste homem e da mãe ou mulher responsável.

5.8. Mãe: escolaridade e ocupação

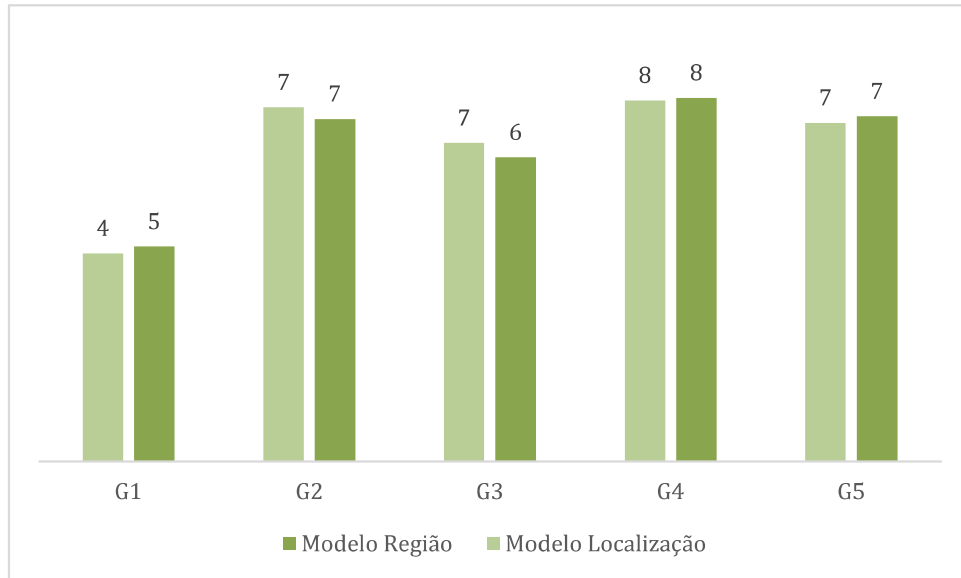
A escolaridade e ocupação das mães (ou da mulher responsável) tiveram em todas as suas categorias algum ganho estatisticamente significativo no desempenho do exame em relação aos candidatos que declararam não saber tais informações. Dessa forma, a condição de não saber tais informações da mãe representaria necessariamente uma perda significativa, frente a qualquer outra escolaridade ou ocupação que essa mulher possa ter.

Gráfico 28 – Coeficientes dos modelos Região e Localização para a variável escolaridade da mãe ou mulher responsável



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

Gráfico 29 – Coeficientes dos modelos Região e Localização para a variável grupo ocupacional da mãe ou mulher responsável



Fonte: Elaboração própria a partir dos dados do ENEM 2022 (INEP, 2023).

Novamente, o ganho da conversão da educação se faz mais expressiva a partir do Ensino Médio e, mais enfaticamente, Ensino Superior completos. Por outro lado, não há progressão evidente entre os grupos ocupacionais G2 e G5, a indicar que existe um patamar de ganho relativo entre não saber a ocupação da mãe (ou mulher de referência) em relação a quem a tem no trabalho no campo (G1), e um patamar um pouco mais elevado para os candidatos filhos de mulheres em qualquer outra ocupação que não essa.

As possíveis razões para se declarar o desconhecimento a respeito do nível de instrução da mãe (ou mulher responsável) podem se assemelhar às listadas para o pai, contudo, menos prevalentes. Assim, diversas categorias ocultas podem compor o não saber, tais como: não conhecer a mãe (ou não ter mulher de referência), desemprego de longa data ou nunca ter trabalhado fora do âmbito doméstico, entre outras. Em comum, todas elas guardariam uma tendência de desvantagem para o estudante.

Por fim, ao se comparar a contribuição dos estudos e ocupação dos homens e mulheres responsáveis pelo estudante, percebe-se a escolaridade da mãe se converte em maior ganho para os filhos em todos os níveis de instrução. Isso reafirma a tendência social de ênfase da responsabilidade da mulher pelo cuidado com os filhos, ao se entender que parte desse trabalho consiste no acompanhamento escolar. Por outro lado, o ganho relativo da ocupação do pai sobressairia ao da mãe nos dois grupos laborais de maior qualificação.

6. SÍNTESE DOS RESULTADOS E CONSIDERAÇÕES

O conjunto dos resultados obtidos com a modelagem elaborada favorece o entendimento sobre a opção por respostas que não declaram de fato o quesito ao qual a pergunta se dedica. Embora para cada questão tratada tenha sido acompanhada por reflexões para se levantar hipótese a respeito das possíveis condições que levaram ao registro das respostas imprecisas, há ao menos duas sínteses que podem ser fundamentadas a partir dos resultados.

A primeira diz respeito ao entendimento precário da questão ou das possibilidades de respostas. A nacionalidade e possivelmente o tipo de ensino seriam exemplos dessas ocorrências. Associados a essa dificuldade estariam as notas marcadamente baixas, uma vez que o não entendimento das perguntas e respostas estaria associado à limitação nas habilidades de leitura e compreensão das questões, bem como de algum nível de conhecimento sobre os temas. Habilidades estas que são também requeridas dos candidatos ao responderem as provas. Assim, a dificuldade em responder ao questionário seria uma antecipação das dificuldades em responder a prova.

A segunda síntese diz respeito às condições ocultas ou fatores intervenientes que levariam a uma dificuldade de enquadramento nas opções presentes. A existência de categorias ocultas seriam o caso das variáveis cor/raça e estado civil, nos quais o sujeito não reconheceria sua condição nas opções ofertadas. E os fatores intervenientes estaria em categorias como a escolaridade e ocupação dos pais, quando esses não são conhecidos ou não têm atividade laboral. Nessas situações, o desempenho desse grupo não necessariamente remeterá a notas das mais baixas nas provas, a depender de qual condição ou fator seria esse.

Por outro lado, tecnicamente há mecanismos que podem vir a reduzir os enganos e promover o melhor enquadramento dos sujeitos, mesmo sem que as categorias definidas sejam alteradas, como a inclusão de notas explicativas das categorias ou mecanismos de enquadramento por aproximação textual para a declaração das ocupações e escolas. Metodologicamente, assegurar a autoaplicação correta dos questionários é dever de quem os elabora, recaindo, portanto, sobre o INEP o papel de aprimorar esse instrumento.

Ademais, este estudo apresenta limitações e outros apontamentos para o desenvolvimento da agenda de pesquisa na qual se insere. Esse é o caso dos sujeitos que tiveram alguma nota zero nas provas e foram excluídos da análise central desenvolvida. Sobre esses casos, bem como os sujeitos que não comparecerem às provas, faz-se necessário desenvolver modelos próprios de análise, tanto para a identificação de fatores associados ou determinantes, quanto para entender como se relacionam com as respostas imprecisas.

Também se recomenda que outros pontos na distribuição da nota sejam investigados, uma vez que o exame se converte em uma competição por melhor desempenho para o ingresso

no Ensino Superior, e valores que variam em torno da nota média não são, de modo geral, suficientes para garantir o sucesso nessa disputa que é relativa à cada curso e modalidade das vagas (ampla concorrência e diversos tipos de cotas).

Outras abordagens sobre o fenômeno das escolhas por respostas imprecisas se fazem também necessárias, inclui-se nisso o desenvolvimento de modelos em que tais respostas figurem como variáveis dependentes, seja uma a uma ou mesmo como uma variável de contagem obtida pela soma dessas.

Espera-se, por fim, que este trabalho contribua para as discussões e tomadas de decisão a respeito de processos de tratamento de dados ausentes ou imprecisos, por meio da visão crítica que promove em torno da possibilidade de se extrair sentido ou distinção interpretável de registros nessa condição.

REFERÊNCIAS

ASSOCIAÇÃO NACIONAL DOS REGISTRADORES DE PESSOAS NATURAIS (ARPEN-BRASIL). Mais 170 mil crianças não receberam o nome do pai no último ano no Brasil. ARPEN BRASIL, Brasília, 23 ago. 2023. Disponível em: <https://arpenbrasil.org.br/press_releases/mais-170-mil-criancas-nao-receberam-o-nome-do-pai-no-ultimo-ano-no-brasil/>. Acesso em: 19 abr. 2023.

BARBOSA, M. As relações entre educação e raça no Brasil: um objeto em construção. In: SOARES, S. *et al.* (orgs.). Os mecanismos de discriminação racial nas escolas brasileiras. Rio de Janeiro: Ipea, 2005.

BECKER, G. Human capital: a theoretical and empirical analysis, with special reference to education. Chicago: University of Chicago Press. 1964.

COLEMAN, J. Desempenho nas escolas públicas, (extraído de 1966). In: BROOKE, N., SOARES, J. (Org.). Pesquisa em eficácia escolar: origem e trajetórias. Belo Horizonte: Editora UFMG, 2008.

CORGOZINHO, M. M.; MONTAGNER, M. Â.; RODRIGUES, M. A. C.. Vulnerabilidade sobre duas rodas: tendência e perfil demográfico da mortalidade decorrente da violência no trânsito motociclístico no Brasil, 2004-2014. Cadernos Saúde Coletiva, v. 26, n. 1, p. 92–99, jan. 2018.

COSTA, Eliane Silvia; SCHUCMAN, Lia Vainer. Identidades, Identificações e Classificações Raciais no Brasil: O Pardo e as Ações Afirmativas. Estudos e Pesquisas em Psicologia, [S. l.], v. 22, n. 2, p. 466–484, 2022.

COSTA, M. Tópicos em Ciência de Dados. Rio de Janeiro: Bonecker, 2019.

COSTA, M.; FREITAS, M. O casamento infantil no Brasil e as questões de gênero. Revista Jurídica em Pauta. v. 1, n. 2, p. 33-44, 2019.

DINIZ, C.; VIEIRA, D. Ensino Superior e Desigualdades Regionais: notas sobre a experiência recente do Brasil. Revista Paranaense de Desenvolvimento - RPD, [S. l.], v. 36, n. 129, p. 99–115, 2015.

DUARTE, H. Estudo sobre o desempenho dos estudantes com deficiência no ENEM 2019. Orientadora: Cristiana Tristão Rodrigues. 2020. Dissertação (Mestrado) - Programa de Pós-Graduação em Economia, UFV, Viçosa, 2020.

FEIJÓ, J. R.; FRANÇA, J. M. S. DE .. Diferencial de desempenho entre jovens das escolas públicas e privadas. Estudos Econômicos (São Paulo), v. 51, n. 2, p. 373–408, abr. 2021.

FERRÃO, M. E.; PRATA, P.; ALVES, M. T. G. Multiple imputation in big identifiable data for educational research: An example from the Brazilian education assessment system. Ensaio: Avaliação e Políticas Públicas em Educação, v. 28, n. 108, p. 599–621, jul. 2020.

FIÚZA, A. *et al.* A motocicleta na vida das famílias de produtores rurais mineiros. Transporte y Territorio, v.29, p. 177-195, jul. 2023.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. Microdados do Enem 2022. Brasília: Inep, 2023. Disponível em: < <https://www.gov.br/inep/ptbr/acesso-a-informacao/dados-abertos/microdados/enem>>. Acesso em: 19 abr. 2023.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. Entenda sua nota do ENEM: Guia do participante. Brasília: Inep, 2021. Disponível em: < <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem> >. Acesso em: 19 abr. 2023.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. Panorama da educação no campo. Brasília: Inep, 2007.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. Orientação para o preenchimento de alunos com deficiência, transtorno global do desenvolvimento e altas habilidades/ superdotação no Censo Escolar 2014. Brasília: Inep, 2014.

KOMATSU, B. A História da Educação e as Origens da Desigualdade Regional no Brasil. São Paulo: Insper, set. 2017. (Policy Papers N° 27).

LIMA, C. C. V. DE .; BRIGHENTI, C. R. G.. Desempenho de estudantes de Minas Gerais no Exame Nacional do Ensino Médio considerando variáveis socioeconômicas. Educação e Pesquisa, v. 49, p. e253303, 2023.

MENDES, M.; NUNES, M. Inglês ou Espanhol: Quais fatores os alunos privilegiam na escolha de uma língua para o ENEM?. Vivências: Revista Eletrônica de Extensão da URI, v. 15, n. 28, p. 124-134, mai. 2019.

OSORIO, R. O sistema classificatório de “cor ou raça” do IBGE. Brasília: Ipea, nov. 2003. (Texto para Discussão N° 996)

PEREIRA, C.; CASTO, C. Educação no meio rural: Diferenciais entre urbano e rural. Rio de Janeiro: Ipea, 2021. (Texto para Discussão N° 2632)

SOUZA, E. C. de; RAMOS, M. D. P. Trabalho docente em escolas rurais: pesquisa e diálogos em tempos de pandemia. Retratos da Escola, [S. l.], v. 14, n. 30, p. 806–822, 2021.

SOUZA, G. Análise do sistema de classificação por cor/ raça no Brasil. RELACult: Revista Latino-Americana de Estudos em Cultura e Sociedade, v. 5, mai. 2019.

SOUZA, H. S. Plataforma da precarização: os conflitos na regulação jurídica do trabalho dos entregadores de aplicativo. Laborare, São Paulo, Brasil, v. 5, n. 8, p. 69–89, 2022.

SOUZA, J. Ralé Brasileira: quem é e como vive. Belo Horizonte: Editora UFMG, 2009.

TODD, P., WOLPIN, K. On the specification and estimation of the production function for cognitive achievement. The Economic Journal, Cambridge, v. 113, n. 481, p. F3-F33, Feb. 2003.

TRIOLA, F. Introdução à estatística. Rio de Janeiro: LTC Editora, 1999 (12ª Edição).

VASCONCELOS, A. M.; GRIEBELER, M. DE C.. Muito jovem ou pobre para casar? Efeito do Programa Bolsa Família sobre o casamento infantil feminino. Revista Brasileira de Economia, v. 77, p. e062023, 2023.

VEIGA, M.; LOYOLA, V. Escolher é Ser Escolhida: Meninice, Pobreza e Casamento Infantil no Brasil. *Psicologia: Teoria e Pesquisa*, v. 36, n. spe, p. e36nspe18, 2020.

WODON, *et al.*. Casamento na Infância e Adolescência: A educação das meninas e a legislação brasileira. Ending Child Marriage Notes Series. Washington: Banco Mundial, abr. 2019.

APÊNCIDE I – Modelos lineares múltiplos desenvolvidos

Modelo Base

Categorias	Coefficiente estimado	Erro Padrão	Valor Teste T	P-valor	Síntese Signif.
Intercepto	395,9	3,35	118,14	< 2e-16	***
Estado civil (ref. Não informado)					
Solteiro(a)	6,9	0,45	15,53	< 2e-16	***
Casado(a)...	-10,6	1,07	-9,93	< 2e-16	***
Divorciado(a)	-3,9	0,99	-3,95	7,87E-05	***
Viúvo(a)	-15,6	5,02	-3,12	0,002	**
Cor ou Raça (ref. Não declarado)					
Branca	16,6	0,56	29,43	< 2e-16	***
Preta	-4,2	0,60	-7,00	2,60E-12	***
Parda	-0,3	0,56	-0,57	0,57	
Amarela	5,0	0,78	6,47	9,88E-11	***
Indígena	-21,8	1,17	-18,60	< 2e-16	***
Nacionalidade (Não declarado)					
Brasileiro(a)	30,2	3,25	9,29	< 2e-16	***
Brasileiro(a) naturalizado(a)	13,1	3,29	3,97	7,16E-05	***
Estrangeiro(a)	11,8	3,71	3,19	0,001	**
Brasileiro(a) nato(a) no exterior	36,0	3,59	10,02	< 2e-16	***
Tipo de ensino (ref. Não respondeu)					
Especial	11,4	1,20	9,56	< 2e-16	***
Regular	38,2	0,55	69,99	< 2e-16	***
Dep. Adm. da Escola e Região (ref. Missing)					
Estadual Centro-oeste	-10,1	0,38	-26,38	< 2e-16	***
Estadual Norte	-21,0	0,38	-55,27	< 2e-16	***
Estadual Nordeste	-8,9	0,24	-37,67	< 2e-16	***
Estadual Sul	-0,1	0,34	-0,26	0,795	
Estadual Sudeste	4,9	0,24	20,34	< 2e-16	***
Federal Centro-oeste	45,8	1,09	42,15	< 2e-16	***
Federal Norte	35,1	1,10	31,99	< 2e-16	***
Federal Nordeste	63,1	0,59	106,63	< 2e-16	***
Federal Sul	62,4	0,93	67,42	< 2e-16	***
Federal Sudeste	82,2	0,64	128,20	< 2e-16	***
Municipal Centro-oeste	-17,0	6,92	-2,45	0,014	*
Municipal Norte	-22,3	2,14	-10,44	< 2e-16	***
Municipal Nordeste	-23,3	3,40	-6,86	6,94E-12	***
Municipal Sul	-3,6	3,76	-0,97	0,334	
Municipal Sudeste	14,9	1,19	12,56	< 2e-16	***
Privada Centro-oeste	55,3	0,55	100,27	< 2e-16	***
Privada Norte	45,0	0,70	64,16	< 2e-16	***
Privada Nordeste	57,3	0,37	153,81	< 2e-16	***
Privada Sul	50,5	0,47	107,20	< 2e-16	***
Privada Sudeste	56,0	0,30	189,98	< 2e-16	***
Escolaridade pai (ref. Não sei)					

Pós-grad.	24,1	0,43	55,70	< 2e-16	***
ES comp.	22,3	0,38	58,05	< 2e-16	***
EM comp.	10,3	0,31	33,59	< 2e-16	***
EF comp.	4,0	0,35	11,63	< 2e-16	***
EF/1 comp.	3,7	0,35	10,70	< 2e-16	***
EF/1 incomp.	-0,8	0,35	-2,36	0,018	*
Não estudou	-8,2	0,59	-13,89	< 2e-16	***
Escolaridade mãe (ref. Não sei)					
Pós-grad.	33,3	0,51	65,32	< 2e-16	***
ES comp.	32,3	0,49	65,54	< 2e-16	***
EM comp.	21,6	0,45	48,15	< 2e-16	***
EF comp.	13,9	0,48	29,06	< 2e-16	***
EF/1 comp.	12,4	0,49	25,12	< 2e-16	***
EF/1 incomp.	8,1	0,52	15,74	< 2e-16	***
Não estudou	-1,3	0,83	-1,61	0,108	
Grupo ocupacional pai (ref. Não sei)					
G5	26,5	0,41	64,16	< 2e-16	***
G4	21,7	0,30	72,49	< 2e-16	***
G3	11,2	0,28	40,06	< 2e-16	***
G2	6,4	0,29	21,95	< 2e-16	***
G1	1,2	0,35	3,42	0,0006	***
Grupo ocupacional mãe (ref. Não sei)					
G5	18,0	0,44	40,53	< 2e-16	***
G4	13,1	0,31	42,25	< 2e-16	***
G3	8,8	0,38	22,92	< 2e-16	***
G2	7,5	0,28	26,75	< 2e-16	***
G1	-1,9	0,39	-4,86	1,20E-06	***

Modelo Região

Categorias	Coefficiente estimado	Erro Padrão	Valor Teste T	P-valor	Síntese Signif.
Intercepto	440,14007	3,29308	133,656	< 2e-16	***
Sexo (ref. Masculino)					
Feminino	0,2263	0,13491	1,677	0,093465	.
Cor ou Raça (ref. Não declarado)					
Branca	8,72922	0,50282	17,361	< 2e-16	***
Preta	-4,34585	0,53215	-8,167	3,18E-16	***
Parda	-1,21737	0,5017	-2,426	0,015247	*
Amarela	2,11586	0,69105	3,062	0,0022	**
Indígena	-16,14978	1,0564	-15,288	< 2e-16	***
Faixa etária (ref. Até 16 anos)					
17 anos	-3,73803	0,68508	-5,456	4,86E-08	***
18 anos	-7,99165	0,68563	-11,656	< 2e-16	***
19 anos	-29,14107	0,72607	-40,135	< 2e-16	***

20 anos	-35,37331	0,85521	-41,362	< 2e-16	***
21 anos	-35,92181	1,17744	-30,508	< 2e-16	***
22 anos	-37,06856	1,61448	-22,96	< 2e-16	***
23 anos	-34,30834	2,05733	-16,676	< 2e-16	***
24 anos	-31,89069	2,45961	-12,966	< 2e-16	***
25 anos	-31,09166	2,78396	-11,168	< 2e-16	***
26 e 30 anos	-25,49054	1,63703	-15,571	< 2e-16	***
31 e 35 anos	-25,65337	1,96069	-13,084	< 2e-16	***
36 e 40 anos	-33,12431	2,18958	-15,128	< 2e-16	***
41 a 45 anos	-33,66707	2,60514	-12,923	< 2e-16	***
46 e 50 anos	-35,78409	3,1336	-11,419	< 2e-16	***
51 e 55 anos	-42,62224	3,89142	-10,953	< 2e-16	***
56 e 60 anos	-44,57895	4,86515	-9,163	< 2e-16	***
61 e 65 anos	-49,47154	9,82075	-5,037	4,72E-07	***
66 e 70 anos	-37,42954	16,53582	-2,264	0,023602	*
Maior de 70 anos	-31,89517	21,98549	-1,451	0,146853	
Nacionalidade (Não declarado)					
Brasileiro(a)	24,80258	2,96831	8,356	< 2e-16	***
Brasileiro(a) naturalizado(a)	9,97984	3,00258	3,324	0,000888	***
Estrangeiro(a)	20,52528	3,37311	6,085	1,17E-09	***
Brasileiro(a) nato(a) no exterior	29,68077	3,25232	9,126	< 2e-16	***
Estado civil (ref. Não informado)					
Solteiro(a)	6,10903	0,39762	15,364	< 2e-16	***
Casado(a)...	-5,43311	1,00663	-5,397	6,77E-08	***
Divorciado(a)	-2,98382	0,88357	-3,377	0,000733	***
Viúvo(a)	-3,52183	4,8108	-0,732	0,464128	
Dep. Adm. da Escola e Região (ref. Missing)					
Estadual Centro-oeste	-10,99155	0,34114	-32,22	< 2e-16	***
Estadual Norte	-11,18648	0,34281	-32,632	< 2e-16	***
Estadual Nordeste	2,34637	0,21394	10,968	< 2e-16	***
Estadual Sul	-9,3349	0,30299	-30,809	< 2e-16	***
Estadual Sudeste	-3,40549	0,21631	-15,744	< 2e-16	***
Federal Centro-oeste	36,15308	0,9409	38,424	< 2e-16	***
Federal Norte	33,25284	0,95303	34,892	< 2e-16	***
Federal Nordeste	63,31043	0,51791	122,242	< 2e-16	***
Federal Sul	51,83773	0,80409	64,468	< 2e-16	***
Federal Sudeste	69,13103	0,5584	123,802	< 2e-16	***
Municipal Centro-oeste	-23,26996	6,08578	-3,824	0,000131	***
Municipal Norte	-13,27334	1,91648	-6,926	4,33E-12	***
Municipal Nordeste	-14,17197	3,06427	-4,625	3,75E-06	***
Municipal Sul	-10,73759	3,29457	-3,259	0,001117	**
Municipal Sudeste	3,2671	1,04137	3,137	0,001705	**
Privada Centro-oeste	34,91326	0,48445	72,068	< 2e-16	***
Privada Norte	33,31451	0,61103	54,522	< 2e-16	***
Privada Nordeste	43,26088	0,3277	132,012	< 2e-16	***
Privada Sul	29,13077	0,41583	70,054	< 2e-16	***

Privada Sudeste	34,4926	0,26907	128,194	< 2e-16	***
Tipo de ensino (ref. Não respondeu)					
Especial	-1,41857	1,10255	-1,287	0,198225	
Regular	16,52158	0,55892	29,56	< 2e-16	***
Língua estrangeira (ref. Inglês)					
Espanhol	-18,86655	0,1424	-132,489	< 2e-16	***
Escolaridade pai (ref. Não sei)					
Pós-grad.	10,14719	0,38467	26,379	< 2e-16	***
ES comp.	11,24488	0,34047	33,027	< 2e-16	***
EM comp.	4,07803	0,27367	14,901	< 2e-16	***
EF comp.	0,92333	0,30889	2,989	0,002798	**
EF/1 comp.	1,71546	0,30975	5,538	3,06E-08	***
EF/1 incomp.	-0,16478	0,31689	-0,52	0,603073	
Não estudou	-2,60246	0,53369	-4,876	1,08E-06	***
Escolaridade mãe (ref. Não sei)					
Pós-grad.	16,54912	0,45997	35,979	< 2e-16	***
ES comp.	18,34538	0,44348	41,366	< 2e-16	***
EM comp.	12,92083	0,4059	31,833	< 2e-16	***
EF comp.	8,96522	0,43358	20,677	< 2e-16	***
EF/1 comp.	8,76809	0,44704	19,613	< 2e-16	***
EF/1 incomp.	7,7781	0,46665	16,668	< 2e-16	***
Não estudou	4,3749	0,76558	5,714	1,10E-08	***
Grupo ocupacional pai (ref. Não sei)					
G5	9,83344	0,37383	26,304	< 2e-16	***
G4	12,36123	0,26772	46,173	< 2e-16	***
G3	7,04464	0,24978	28,204	< 2e-16	***
G2	5,37539	0,25981	20,69	< 2e-16	***
G1	3,70452	0,31077	11,92	< 2e-16	***
Grupo ocupacional mãe (ref. Não sei)					
G5	7,27758	0,39429	18,457	< 2e-16	***
G4	7,666	0,27344	28,035	< 2e-16	***
G3	6,41391	0,34023	18,851	< 2e-16	***
G2	7,22169	0,24799	29,121	< 2e-16	***
G1	4,52922	0,34671	13,064	< 2e-16	***
Quantidade de moradores	-2,2017	0,05795	-37,992	< 2e-16	***
Renda domiciliar mensal (ref. Não tem)					
Até R\$ 1.212,00	7,9737	0,33166	24,042	< 2e-16	***
De R\$ 1.212,01 até R\$ 1.818,00	15,60435	0,36479	42,776	< 2e-16	***
De R\$ 1.818,01 até R\$ 2.424,00	18,14682	0,38539	47,087	< 2e-16	***
De R\$ 2.424,01 até R\$ 3.030,00	20,59768	0,41121	50,09	< 2e-16	***
De R\$ 3.030,01 até R\$ 3.636,00	25,05408	0,4559	54,955	< 2e-16	***
De R\$ 3.636,01 até R\$ 4.848,00	25,41323	0,44383	57,259	< 2e-16	***
De R\$ 4.848,01 até R\$ 6.060,00	27,7934	0,45489	61,099	< 2e-16	***
De R\$ 6.060,01 até R\$ 7.272,00	30,68547	0,5547	55,319	< 2e-16	***
De R\$ 7.272,01 até R\$ 8.484,00	32,37757	0,61082	53,007	< 2e-16	***
De R\$ 8.484,01 até R\$ 9.696,00	35,60448	0,69195	51,455	< 2e-16	***

De R\$ 9.696,01 até R\$ 10.908,00	36,06072	0,63605	56,695	< 2e-16	***
De R\$ 10.908,01 até R\$ 12.120,00	38,09097	0,65163	58,455	< 2e-16	***
De R\$ 12.120,01 até R\$ 14.544,00	41,1772	0,72994	56,412	< 2e-16	***
De R\$ 14.544,01 até R\$ 18.180,00	45,28289	0,71072	63,714	< 2e-16	***
De R\$ 18.180,01 até R\$ 24.240,00	49,39739	0,74158	66,611	< 2e-16	***
Acima de R\$ 24.240,00	54,97557	0,70808	77,64	< 2e-16	***
Tem banheiro (ref. Não)					
Sim, um	2,9127	0,89691	3,247	0,001164	**
Sim, dois	4,1006	0,91033	4,505	6,65E-06	***
Sim, três	5,1136	0,93588	5,464	4,66E-08	***
Sim, quatro ou mais	8,16306	0,97495	8,373	< 2e-16	***
Tem motocicleta (ref. Não)					
Sim, um	-3,59514	0,16023	-22,438	< 2e-16	***
Sim, dois	-8,95015	0,39352	-22,744	< 2e-16	***
Sim, três	-21,35054	1,21744	-17,537	< 2e-16	***
Sim, quatro ou mais	-29,88233	2,74015	-10,905	< 2e-16	***
Tem celular (ref. Não)					
Sim, um	3,84958	0,47613	8,085	6,22E-16	***
Sim, dois	11,65526	0,46084	25,291	< 2e-16	***
Sim, três	10,43481	0,45738	22,814	< 2e-16	***
Sim, quatro ou mais	9,00814	0,46797	19,249	< 2e-16	***
Tem computador (ref. Não)					
Sim, um	10,99677	0,16847	65,273	< 2e-16	***
Sim, dois	18,39397	0,26647	69,028	< 2e-16	***
Sim, três	22,99243	0,38448	59,801	< 2e-16	***
Sim, quatro ou mais	26,92788	0,522	51,586	< 2e-16	***
Tem internet em casa (ref. Não)					
Sim	4,44158	0,28066	15,825	< 2e-16	***

Modelo Localização

Categorias	Coefficiente estimado	Erro Padrão	Valor Teste T	P-valor	Síntese Signif.
Intercepto	439,32247	3,39435	129,428	< 2e-16	***
Sexo (ref. Masculino)					
Feminino	0,38767	0,13434	2,886	0,003906	**
Cor ou Raça (ref. Não declarado)					
Branca	9,05134	0,50052	18,084	< 2e-16	***
Preta	-4,31159	0,52987	-8,137	4,05E-16	***
Parda	-1,30883	0,49911	-2,622	0,008733	**
Amarela	2,9341	0,68758	4,267	1,98E-05	***
Indígena	-14,51102	1,05223	-13,791	< 2e-16	***
Faixa etária (ref. Até 16 anos)					
17 anos	-3,32077	0,6821	-4,868	1,13E-06	***
18 anos	-8,40697	0,68289	-12,311	< 2e-16	***

19 anos	-30,07331	0,72278	-41,608	< 2e-16	***
20 anos	-36,64592	0,85125	-43,049	< 2e-16	***
21 anos	-37,36613	1,17155	-31,895	< 2e-16	***
22 anos	-38,66434	1,60607	-24,074	< 2e-16	***
23 anos	-35,35642	2,04635	-17,278	< 2e-16	***
24 anos	-33,08318	2,44616	-13,525	< 2e-16	***
25 anos	-31,8301	2,76858	-11,497	< 2e-16	***
26 e 30 anos	-26,01266	1,62843	-15,974	< 2e-16	***
31 e 35 anos	-26,27054	1,95029	-13,47	< 2e-16	***
36 e 40 anos	-33,80555	2,1778	-15,523	< 2e-16	***
41 a 45 anos	-34,38002	2,59096	-13,269	< 2e-16	***
46 e 50 anos	-36,93525	3,11654	-11,851	< 2e-16	***
51 e 55 anos	-44,3604	3,86997	-11,463	< 2e-16	***
56 e 60 anos	-44,98302	4,83824	-9,297	< 2e-16	***
61 e 65 anos	-51,38648	9,76554	-5,262	1,43E-07	***
66 e 70 anos	-35,78678	16,44262	-2,176	0,029521	*
Maior de 70 anos	-29,52727	21,86166	-1,351	1,77E-01	
Nacionalidade (Não declarado)					
Brasileiro(a)	24,01397	2,95161	8,136	4,09E-16	***
Brasileiro(a) naturalizado(a)	9,58701	2,98573	3,211	0,001323	**
Estrangeiro(a)	22,46382	3,35441	6,697	2,13E-11	***
Brasileiro(a) nato(a) no exterior	29,80742	3,23404	9,217	< 2e-16	***
Estado civil (ref. Não informado)					
Solteiro(a)	5,97445	0,39545	15,108	< 2e-16	***
Casado(a)...	-5,04561	1,00108	-5,04	4,65E-07	***
Divorciado(a)	-2,95227	0,87863	-3,36	0,000779	***
Viúvo(a)	-4,15484	4,78365	-0,869	3,85E-01	
UF onde realizou a prova (ref. AC)					
AL	1,00859	1,05177	0,959	0,337589	
AM	-14,90952	1,02526	-14,542	< 2e-16	***
AP	-9,41923	1,3079	-7,202	5,95E-13	***
BA	3,54271	0,95441	3,712	0,000206	***
CE	15,08337	0,94782	15,914	< 2e-16	***
DF	-1,19105	1,02545	-1,161	0,245443	
ES	13,24406	1,02211	12,958	< 2e-16	***
GO	-0,9226	0,96748	-0,954	0,340281	
MA	-3,32763	0,98575	-3,376	0,000736	***
MG	12,62093	0,94791	13,314	< 2e-16	***
MS	-12,15675	1,07207	-11,34	< 2e-16	***
MT	-12,21352	1,04349	-11,704	< 2e-16	***
PA	1,62345	0,96858	1,676	0,093718	.
PB	11,11233	1,00613	11,045	< 2e-16	***
PE	9,51317	0,96282	9,881	< 2e-16	***
PI	13,05497	1,02264	12,766	< 2e-16	***
PR	-10,82011	0,96501	-11,212	< 2e-16	***
RJ	4,00743	0,95169	4,211	2,54E-05	***

RN	11,51449	1,0217	11,27	< 2e-16	***
RO	-1,81034	1,14659	-1,579	0,11436	
RR	-4,39334	1,76105	-2,495	0,012606	*
RS	4,79638	0,97211	4,934	8,06E-07	***
SC	-4,89786	0,99369	-4,929	8,27E-07	***
SE	15,60184	1,0628	14,68	< 2e-16	***
SP	-3,29394	0,93479	-3,524	0,000426	***
TO	-10,15529	1,15898	-8,762	< 2e-16	***
Dep. Adm. da Escola e Localização (ref. Missing)					
Rural Estadual	-10,70351	0,54414	-19,671	< 2e-16	***
Rural Federal	38,64648	0,87455	44,19	< 2e-16	***
Rural Municipal	-21,20904	2,74323	-7,731	1,06E-14	***
Rural Privada	21,20197	1,72936	12,26	< 2e-16	***
Urbana Estadual	-3,89049	0,16176	-24,052	< 2e-16	***
Urbana Federal	59,0172	0,34778	169,698	< 2e-16	***
Urbana Municipal	1,0896	0,88251	1,235	0,216959	
Urbana Privada	35,5109	0,22313	159,15	< 2e-16	***
Tipo de ensino (ref. Não respondeu)					
Especial	-1,01189	1,09657	-0,923	0,356122	
Regular	16,26243	0,55602	29,248	< 2e-16	***
Língua estrangeira (ref. Inglês)					
Espanhol	-19,64329	0,14333	-137,051	< 2e-16	***
Escolaridade pai (ref. Não sei)					
Pós-grad.	10,55214	0,38261	27,58	< 2e-16	***
ES comp.	11,70412	0,33865	34,561	< 2e-16	***
EM comp.	4,28919	0,27221	15,757	< 2e-16	***
EF comp.	0,82558	0,30717	2,688	0,007195	**
EF/1 comp.	1,45333	0,30811	4,717	2,39E-06	***
EF/1 incom.	-0,43196	0,31524	-1,37	0,170597	
Não estudou	-2,84546	0,53088	-5,36	8,33E-08	***
Escolaridade mãe (ref. Não sei)					
Pós-grad.	17,62488	0,45864	38,429	< 2e-16	***
ES comp.	19,51676	0,4422	44,135	< 2e-16	***
EM comp.	14,03043	0,40487	34,654	< 2e-16	***
EF comp.	9,71047	0,43178	22,489	< 2e-16	***
EF/1 comp.	9,31798	0,44522	20,929	< 2e-16	***
EF/1 incom.	8,29461	0,46456	17,855	< 2e-16	***
Não estudou	5,16657	0,76168	6,783	1,18E-11	***
Grupo ocupacional pai (ref. Não sei)					
G5	9,00526	0,3719	24,214	< 2e-16	***
G4	11,88521	0,26636	44,62	< 2e-16	***
G3	7,02114	0,24846	28,259	< 2e-16	***
G2	4,97544	0,25844	19,252	< 2e-16	***
G1	3,25827	0,3097	10,521	< 2e-16	***
Grupo ocupacional mãe (ref. Não sei)					
G5	7,13988	0,39216	18,207	< 2e-16	***

G4	7,61305	0,27198	27,992	< 2e-16	***
G3	6,71544	0,33852	19,837	< 2e-16	***
G2	7,46875	0,24668	30,278	< 2e-16	***
G1	4,38321	0,34544	12,689	< 2e-16	***
Quantidade de moradores	-2,0104	0,05797	-34,682	< 2e-16	***
Renda domiciliar mensal (ref. Não tem)					
Até R\$ 1.212,00	7,78074	0,3304	23,55	< 2e-16	***
De R\$ 1.212,01 até R\$ 1.818,00	16,2676	0,36421	44,666	< 2e-16	***
De R\$ 1.818,01 até R\$ 2.424,00	19,17377	0,38508	49,791	< 2e-16	***
De R\$ 2.424,01 até R\$ 3.030,00	21,97542	0,41116	53,448	< 2e-16	***
De R\$ 3.030,01 até R\$ 3.636,00	26,48315	0,45569	58,116	< 2e-16	***
De R\$ 3.636,01 até R\$ 4.848,00	26,96988	0,44407	60,734	< 2e-16	***
De R\$ 4.848,01 até R\$ 6.060,00	29,33972	0,45526	64,447	< 2e-16	***
De R\$ 6.060,01 até R\$ 7.272,00	32,05154	0,5541	57,844	< 2e-16	***
De R\$ 7.272,01 até R\$ 8.484,00	33,71189	0,60973	55,29	< 2e-16	***
De R\$ 8.484,01 até R\$ 9.696,00	36,74825	0,69018	53,244	< 2e-16	***
De R\$ 9.696,01 até R\$ 10.908,00	37,3157	0,63488	58,776	< 2e-16	***
De R\$ 10.908,01 até R\$ 12.120,00	39,31397	0,65031	60,454	< 2e-16	***
De R\$ 12.120,01 até R\$ 14.544,00	42,31999	0,72787	58,143	< 2e-16	***
De R\$ 14.544,01 até R\$ 18.180,00	46,44709	0,70891	65,519	< 2e-16	***
De R\$ 18.180,01 até R\$ 24.240,00	50,46099	0,73944	68,243	< 2e-16	***
Acima de R\$ 24.240,00	56,15775	0,70623	79,518	< 2e-16	***
Tem banheiro (ref. Não)					
Sim, um	1,77101	0,89276	1,984	0,047284	*
Sim, dois	2,73535	0,90616	3,019	0,002539	**
Sim, três	3,9908	0,93153	4,284	1,83E-05	***
Sim, quatro ou mais	7,40632	0,97038	7,632	2,31E-14	***
Tem motocicleta (ref. Não)					
Sim, um	-3,88139	0,16113	-24,089	< 2e-16	***
Sim, dois	-9,23178	0,39279	-23,503	< 2e-16	***
Sim, três	-21,51649	1,21095	-17,768	< 2e-16	***
Sim, quatro ou mais	-30,0806	2,72482	-11,039	< 2e-16	***
Tem celular (ref. Não)					
Sim, um	3,41103	0,47352	7,204	5,87E-13	***
Sim, dois	11,10191	0,45835	24,221	< 2e-16	***
Sim, três	9,88702	0,45493	21,733	< 2e-16	***
Sim, quatro ou mais	8,39588	0,46553	18,035	< 2e-16	***
Tem computador (ref. Não)					
Sim, um	11,08762	0,16818	65,927	< 2e-16	***
Sim, dois	18,25527	0,26565	68,72	< 2e-16	***
Sim, três	22,74915	0,38289	59,414	< 2e-16	***
Sim, quatro ou mais	26,59957	0,51953	51,199	< 2e-16	***
Tem internet em casa (ref. Não)					
Sim	3,95977	0,28054	14,115	< 2e-16	***