Ingrid Trindade Marrocos

Modelando a relação entre fatores sociais e proficiência no ENEM

Niterói - RJ, Brasil

21 de Dezembro de 2022

Universidade Federal Fluminense

Ingrid Trindade Marrocos

Modelando a relação entre fatores sociais e proficiência no ENEM

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador(a): Dr. Rafael Santos Erbisti

Niterói - RJ, Brasil

21 de Dezembro de 2022

Universidade Federal Fluminense

Ingrid Trindade Marrocos

Modelando a relação entre fatores sociais e proficiência no ENEM

Monografia de Projeto Final de Graduação sob o título *"Modelando a relação entre fatores sociais e proficiência no ENEM"*, defendida por Ingrid Trindade Marrocos e aprovada em 21 de Dezembro de 2022, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Prof. Dr. Rafael Santos Erbisti
Departamento de Estatística – UFF

Profa. Dra. Ana Beatriz Monteiro Fonseca
Departamento de Estatística – UFF

Prof. Dr. Victor Eduardo Leite de Almeida Duca
Departamento de Estatística – UFF

Niterói, 21 de Dezembro de 2022

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

M361m Marrocos, Ingrid Trindade
Modelando a relação entre fatores sociais e proficiência no ENEM / Ingrid Trindade Marrocos. - 2022.
73 f.: il.

Orientador: Rafael Santos Erbisti.
Trabalho de Conclusão de Curso (graduação)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2022.

1. Modelo de regressão linear. 2. LASSO. 3. ENEM. 4. Efeito espacial. 5. Produção intelectual. I. Erbisti, Rafael Santos, orientador. II. Universidade Federal Fluminense. Instituto de Matemática e Estatística. III. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

O ENEM tem como principal objetivo analisar a proficiência dos alunos do ensino médio e possibilitar o seu ingresso no ensino superior. A partir de indicadores educacionais e socioeconômicos torna-se possível ter uma melhor compreensão sobre a real situação educacional de uma região e, com isso, auxiliar a discussão e criação de políticas públicas voltadas para ampliação e melhoria do sistema de educação. O objetivo deste trabalho é modelar a relação entre fatores que podem influenciar a proficiência dos alunos no ENEM, considerando a real desigualdade persistente no Brasil e suas diferenças regionais. Foi avaliada a proficiência dos alunos a partir das notas médias de matemática e português no ENEM de 2019 das escolas dos candidatos, em todo o Brasil. Após o agrupamento dos dados que estavam no nível do aluno para o nível da escola, obteve-se informação de 29.181 escolas para o modelo de matemática e 29.298 para o modelo de português. Neste trabalho foi utilizado o modelo de regressão linear normal incorporando a informação da localização espacial das escolas a partir de variáveis de efeitos fixos, identificando, assim, a microrregião na qual a escola pertence. Além disso, foi utilizado o método de regularização LASSO para selecionar os indicadores socioeconômicos e de infraestrutura escolar de maior relevância, bem como os efeitos fixos significativos relacionados às localizações das escolas, sendo usado o $\lambda = 0,446$, pois possui o menor erro de predição. Os resultados encontrados não se diferenciam dos já discutidos na literatura: nota-se que alunos não brancos, de renda baixa, oriundos de escolas públicas das regiões Norte e Nordeste apresentam menor proficiência em ambas as provas. Ademais, a presença de quadra e de candidatas gestantes na escola não demostrou ser relevante para a nota de matemática, e escolas com alunos autistas possuem efeito negativo na nota média das escolas nas provas de matemática (-110 pontos) e português (-66,3 pontos). Observou-se também que a maioria das escolas são públicas (72%) e que as escolas localizadas no Sudeste são as que mais possuem candidatos de cor branca (72,5%). Na análise dos resíduos, o Teste de Breusch–Pagan e Teste de Lilliefors rejeitam os pressupostos de variância σ^2 constante e normalidade, respectivamente.

Palavras-chave: Modelo de regressão linear. LASSO. ENEM. Efeito espacial.

Dedicatória

Dedico este trabalho a minha mãe e ao meu pai que sempre me apoiaram e deram suporte, à toda minha família que torce pelo meu sucesso e a mim, por não ter desistido e sempre ter me dedicado.

Agradecimentos

Agradeço imensamente aos meus pais que sempre estiveram comigo, me incentivando e apoiando desde muito pequena. Minha mãe Lory, que aturou meu mau humor e estresse por muita das vezes e nunca deixou de me auxiliar e dar apoio em todas as minhas decisões. Ao meu pai, Rodrigo, que sempre batalhou muito para me dar as melhores oportunidades e experiencias possíveis, sempre acreditando no meu potencial.

À minha família que sempre esteve muito presente e torcendo pelo meu sucesso. Agradeço à minha irmã Claudia por sempre estar disponível para me ajudar e por enviar mimos nos dias mais tensos. À minha irmã Sofia, que consegue fazer com que eu esqueça de tudo e só aproveite com leveza nossos momentos juntas que são sempre únicos. Ao meu namorado Patrick que sempre foi meu parceiro e respeitou meus momentos de estudo, me aturando e sendo meu refúgio nos dias que eu não aguentava mais. A todos da minha família que de alguma maneira me ajudaram a chegar aqui, meus avós, minha dinda, meus tios e primos.

Aos meus amigos que sempre me trouxeram momentos de distração, conversas e conselhos, em especial a Isabela, Brenda, Carolzinha, Amy, Julia, Bea, Daniel, Roger, Rick, Felizardo e Vik.

A Universidade Federal Fluminense por realizar meu sonho de estudar em uma Federal e por me fazer viver momentos inesquecíveis. Aos meus amigos que a UFF me presenteou, principalmente a Carol, Vitor, Patricia Mello, que passaram madrugadas comigo fazendo trabalho, a Carla, Matheus, Luana, Giovanna, Isaque e Patricia que entraram comigo e me fizeram viver muitos momentos incríveis em união. À Hillary que já estava na minha a vida antes e sempre me apoiou, me aconselhando, emprestado material para estudar e cedendo sua casa para eu dormir.

Por fim, a todos os professores da UFF que foram essenciais para minha formação e a Ana Maria que sempre orientou com os problemas da coordenação. Em especial, um muito obrigada ao meu orientador Rafael Erbisti por ter aceitado entrar nessa loucura comigo. Sempre esteve muito presente, disponível e paciente com minhas diversas mensagens. Serei eternamente grata a você! Também agradeço a banca por terem aceitado meu convite.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Neste primeiro capítulo serão apresentados a motivação deste trabalho, uma breve revisão bibliográfica e os objetivos.

1.1 Motivação

Segundo Nelson Mandela, pai da moderna nação sul-africana e líder na luta contra o regime do *Apartheid*, "A educação é a arma mais poderosa que você pode usar para mudar o mundo". Realmente, a educação é um fator essencial para a construção de uma sociedade mais justa, pois dá oportunidade e permite que as pessoas desprivilegiadas alcancem posições que, em algumas condições habituais, seriam inviáveis. (BRANDÃO; FAGUNDES, 2016)

Nos últimos anos houve um aumento expressivo no interesse de estudo da educação, que está associada ao desenvolvimento de um país, cultural e economicamente. Diversas pesquisas foram realizadas na área e, em todos os casos, pode-se notar a existência de uma forte relação entre os níveis educacionais e o grau de desenvolvimento de um país. (FREITAS, 2022) Nesse contexto, é sabido que investimentos públicos em educação trazem diversos benefícios para a sociedade, como a redução da criminalidade, redução de desigualdade social e de renda, e o aperfeiçoamento do capital humano (ALBERNAZ; FERREIRA; FRANCO, 2002).

As desigualdades socioeconômicas regionais são fatores que persistem no Brasil. O Relatório de Desenvolvimento Humano de 2019 Conceição (2020), feito pelo Programa das Nações Unidas para o Desenvolvimento, indicou que o Brasil está na sétima posição entre os países mais desiguais do mundo. Este fator está diretamente ligado com o nível de oportunidade de acesso à educação, visto que diversos estudos apontam que o espaço no qual o indivíduo vive possui alta relação com sua proficiência escolar. Coerentemente, no Brasil, piores condições escolares são observadas em regiões periféricas, com menor

acesso aos serviços públicos básicos (GOMES; MELO, 2021)

Em especial, ao se discutir, especificamente, sobre proficiência em exames para acesso ao ensino superior, as condições escolares e socioeconômicas do aluno se tornam ponto fundamental de estudo. Nesse sentido, a partir da incorporação de indicadores educacionais e socioeconômicos torna-se possível ter uma melhor compreensão sobre a real situação educacional de uma região. O Exame Nacional do Ensino Médio (ENEM), criado em 1998 e realizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), possui como principal objetivo analisar a proficiência escolar dos alunos do ensino médio e possibilitar o ingresso no ensino superior. Entretanto, diversos estudos retratam os padrões de desigualdade encontrados nos resultados do ENEM, oriundos de um país plural, com vasto território sócio e economicamente desigual (GOMES; MELO, 2021).

A partir da discussão feita, torna-se necessário refletir criticamente sobre fatores que podem influenciar na proficiência dos alunos no ENEM, considerando a real desigualdade persistente no Brasil e suas diferenças regionais. Essas informações são valiosas uma vez que podem apontar caminhos para a geração de políticas públicas.

1.2 Revisão Bibliográfica

Há uma extensa literatura sobre educação no Brasil e um número considerável de publicações referentes à análise sobre a proficiência no ENEM. Muitos desses trabalhos versam sobre discussões de caráter socioeconômico, avaliando fatores que influenciam na proficiência de alunos em testes educacionais realizados nacionalmente. A discussão sobre a proficiência em outras avaliações em larga escala, como a Prova Brasil e o Sistema Nacional de Avaliação da Educação Básica (Saeb), é fundamental, pois são avaliações utilizadas para diagnóstico e que têm o objetivo de avaliar a qualidade do ensino oferecido pelo sistema educacional brasileiro. Diferentemente do ENEM, a Prova Brasil e o Saeb são testes aplicados na quarta e oitava séries (quinto e nono anos) do ensino fundamental, os estudantes respondem a itens (questões) de língua portuguesa, com foco em leitura, e matemática, com foco na resolução de problemas.

Soares, Soares e Santos (2020) avaliaram uma amostra composta por 639 escolas participantes da Prova Brasil de 2017 e consideraram 30 informações sobre características das escolas extraídas do Censo Escolar. Um dos objetivos dos autores foi relacionar essas características com a proficiência das escolas dos anos finais do ensino fundamental na Prova Brasil. A partir da aplicação de um modelo logístico e aplicação do método

Stepwise para a seleção das variáveis significativas do modelo, os autores verificaram que as variáveis coleta periódica de lixo, presença de laboratório de informática, quadra de esportes, dependências e vias adequadas a alunos com deficiência ou mobilidade reduzida (PNE), auditório e internet banda larga, se encontram em menores proporções nas escolas que atendem a estudantes com menor nível socioeconômico, refletindo em sua proficiência escolar. De forma geral, foi visto que boas características, esperadas em qualquer escola ou região do país, estavam relacionadas a alunos com maior nível socioeconômico e em regiões onde há a atuação do poder público.

Já na análise do Saeb, Oliveira (2022) avaliou a relação entre proficiência escolar, infraestrutura e outros indicadores em Educação, sendo empregado o Modelo Linear Normal com as notas do Saeb e as regiões geográficas do país sendo covariáveis do modelo. O autor também utilizou o Método Stepwise para adicionar ou excluir variáveis preditoras no modelo e constatou que, dependendo do segmento da educação básica, a infraestrutura escolar pode influenciar na aprendizagem do aluno, em diferentes graus.

Semelhante a proposta do presente trabalho, recentemente, Melo et al. (2021) avaliaram o impacto das variáveis socioeconômicas na proficiência no ENEM de 2018 ao nível municipal, utilizando uma amostra de 5.548 municípios do Brasil (IBGE, 2019) e informações de 79 variáveis sociodemográficas. Os autores consideraram a média e a variância das notas da prova objetiva e da redação como variáveis dependentes, além das variáveis independentes classificadas em cinco grupos: econômicas, raciais, de perfil instrucional da mãe, de incentivo escolar e de infraestrutura e ensino escolar. A partir de modelos de regressão, incorporando ou não uma estrutura espacial (método STLS) e aplicação do método LASSO (usou-se o método K-FOLD para evitar o overfitting), observou-se que, para a prova objetiva, variáveis como o percentual de estudantes com bolsa de estudos, renda, raça, escolaridade e nível instrucional da mãe são fatores relevantes para o desempenho na prova e para a variabilidade das notas dos estudantes de cada município. Para a nota na redação, as variáveis foram similares às da prova objetiva, mas observou-se um menor impacto na média e na dispersão das notas. Esse fator explicativo aumenta quando é introduzido um componente espacial no modelo para as notas de redação, indicando que outros fatores regionais, diferentes dos socioeconômicos, devem impactar no desempenho e a dispersão das notas. Ademais, eles verificaram que características socioeconômicas, como rendimento familiar e formação da mãe, têm forte relação com o desempenho dos estudantes. A variável geográfica/espacial foi também identificada como fator importante, sendo a região Sul/Sudeste aquela que tem melhor desempenho; e quanto maior condição socioeconômica, melhor foi o desempenho nas avaliações escolares.

Também com o objetivo de avaliar variáveis sociais que explicam a proficiência nos quatro testes da edição do ENEM de 2018, Jaloto e Primi (2021) analisaram a influência do atraso escolar, de sexo, cor/raça, nível socioeconômico e dependência administrativa da escola do aluno. Além de avaliar a associação entre as variáveis e a proficiência, os autores analisaram o comportamento do efeito do nível socioeconômico entre escolas. A partir de modelos de regressão linear multinível foi verificado que o desempenho dos alunos brancos superou o dos demais em menos de 10,0 pontos e o fato do aluno estudar em escola privada aumentou a nota em 83,9 pontos, em média, em relação à escola estadual.

Torres et al. (2020) realizaram um estudo do ENEM do ano de 2017, no qual analisaram as possíveis diferenças entre os ensinos público e privado, no Estado do Rio Grande do Sul, a partir da proficiência dos alunos no Exame Nacional do Ensino Médio (Enem). Aplicou-se a modelagem de regressão quantílica decomposta em quartis à distribuição condicional da proficiência. Os resultados indicaram ser favoráveis a proficiência dos alunos de escolas de natureza privada, de forma crescente, para os quartis, sendo coerente com evidências da literatura. Por outro lado, verificou-se desempenho inferior da escola de natureza pública. Também obteve melhor desempenho aqueles que possuem melhores condições socioeconômicas familiares, como renda e mãe com formação no Ensino Superior.

É evidente o contraste entre qualidade de ensino, infraestrutura, métodos de ensino, organização e cobrança de rendimento dos alunos em instituições e ensino básico públicas e privadas. Nesse sentido, Moraes e Peres (2022) avaliaram a proficiência de alunos de escolas públicas e privadas da região Sudeste do Brasil no ENEM de 2017. Os autores compararam a importância de variáveis individuais, de infraestrutura escolar e indicadores educacionais para alunos de desempenho alto, médio e baixo de escolas públicas e privadas, a partir de modelos de regressão quantílica. Moraes e Peres (2022) verificaram que as diferenças já conhecidas para o ensino fundamental são potencializadas ao final do ensino médio, com grande impacto socioeconômico, especialmente para alunos de escolas públicas com baixo desempenho.

Conforme citado acima, em alguns estudos foi realizada a seleção de variáveis significativas para o modelo, tendo em vista que o trabalho com um grande número de variáveis se torna um desafio, por ser muito custoso e por interferir no processo de encontrar a relação entre um conjunto de covariáveis e uma variável de interesse para modelos lineares encaixados. (TIBSHIRANI, 1996) Com isto, existem diversos métodos e critérios que realizam a seleção de variáveis em duas etapas, sendo elas, a estimação dos coeficientes

associados às covariáveis e depois a seleção das variáveis. Entretanto, métodos mais recentes baseados na verossimilhança penalizada realizam simultaneamente as duas etapas da seleção das variáveis relevantes pro modelo, como o método de regularização LASSO, proposto por Tibshirani (1996), que será utilizado neste trabalho.

1.3 Objetivos

O objetivo geral deste trabalho é modelar a relação entre a proficiência das escolas dos candidatos a partir das notas médias de matemática e português no ENEM de 2019 do Brasil, com fatores sociais e infraestrutura da escola. Os objetivos específicos do trabalho são:

- Utilizar o Modelo de Regressão Linear Normal para avaliar as notas médias de matemática e português no ENEM de 2019 das escolas, em nível nacional.
- Incorporar a informação da localização espacial das escolas a partir de variáveis de efeitos fixos, identificando a microrregião na qual a escola pertence.
- Avaliar se existe associação entre a infraestrutura escolar, fatores sociais e características dos candidatos, com a proficiência das escolas dos alunos, medida através da nota de português e matemática do ENEM;
- Utilizar o método de regularização LASSO para selecionar as variáveis que agregaram informação ao modelo, bem como os efeitos fixos significativos relacionados às microrregiões.
- Realizar o procedimento de estimação sob perspectiva da Inferência Clássica para os modelos lineares.

2 Materiais e Métodos

Neste capítulo será apresentada a descrição dos materiais utilizados nas análises, além da apresentação detalhadas das metodologias que foram adotadas para a execução deste trabalho.

2.1 Bases de dados

Para a realização deste trabalho foi necessário coletar informações de bases distintas. Uma delas é a base de 2019 do Censo Escolar, que contém informações de 228.521 escolas dos candidatos. Esta pesquisa é realizada pelo INEP.(BRASIL, 2020)

As informações do Censo Escolar são construídas a partir de duas fases de coleta de dados de caráter obrigatório. A primeira fase está relacionada à coleta de informações sobre os estabelecimentos de ensino, turmas, alunos e profissionais escolares em sala de aula. A segunda fase leva em consideração dados sobre o rendimento escolar dos alunos ao final do ano letivo, obtendo então informações sobre a Situação do Aluno. Para este trabalho será utilizado apenas os dados de infraestrutura escolar retirado da primeira etapa do Censo Escolar.

A outra base utilizada neste trabalho é a do ENEM de 2019, obtida também pelo INEP, por intermédio da Diretoria de Avaliação da Educação Básica, que contém informações de 5.095.270 candidatos inscritos, a caracterização do participante e da escola que ele declarou ter frequentado, e as notas das provas objetivas e da redação. Desde 2020, a prova do ENEM pode ser realizada de forma impressa ou pelo Enem Digital, que é feito em computadores nos locais definidos pelo INEP, sendo uma escolha que fica a critério do candidato. A aplicação ocorre em dois dias e a prova contém 180 questões objetivas, que são divididas em quatro áreas de conhecimento, são elas: linguagens, códigos e suas tecnologias; ciências humanas e suas tecnologias; ciências da natureza e suas tecnologias; e matemática e suas tecnologias. Além disso, também é necessário realizar uma prova de redação, que é um texto dissertativo-argumentativo sobre um tema escolhido a cada ano.

Apesar da base do ENEM vir no nível do candidato, para este trabalho esses dados foram agrupados para o nível da escola destes alunos, devido o custo computacional de trabalhar com uma base de dados tão grande como é a do ENEM de 2019.

2.1.1 Variáveis do ENEM 2019

A base do ENEM 2019 possui informações sobre dados do participante, dados da escola, dados dos pedidos de atendimento especializado, dados dos pedidos de atendimento específico, dados dos pedidos de recurso especializados e específicos para realização das provas, dados do local de aplicação da prova, dados da prova objetiva, dados da redação e dados do questionário socioeconômico. Na Tabela 1 são apresentadas 14 variáveis desta base que serão analisadas neste trabalho.

Tabela 1: Variáveis relacionadas as escolas dos candidatos do ENEM.

Variável	**Descrição**
nota_mat	Nota média da escola na prova de matemática
nota_port	Nota média da escola na prova de português
prop_mulher	Proporção de candidatas do gênero feminino em cada uma das escolas
prop_solteiro	Proporção de candidatas solteiros em cada uma das escolas
prop_branco	Proporção de candidatos de cor/raça branca em cada uma das escolas
prop_autismo	Proporção de candidatos que possuem autismo em cada uma das escolas
prop_gestante	Proporção de candidatas que são gestantes em cada uma das escolas
prop_maior_18	Proporção de candidatos que são maiores de idade (maior de 18 anos) em cada uma das escolas
prop_possui_empregado	Proporção de candidatos que possuem empregado(a) em casa
prop_pos_mae	Proporção de mães dos candidatos que possuem Pós-graduação por escola
prop_tem_carro	Proporção de candidatos que possuem carro por escola
prop_tem_pc_casa	Proporção de candidatos que possuem computador em casa
prop_qnt_pessoa_mais_10	Proporção de candidatos que moram com mais de 10 pessoas em casa (incluindo ele)
prop_tem_internet_casa	Proporção de candidatos que possuem internet em casa

2.1.2 Variáveis do Censo Escolar 2019

Na Tabela 2 são apresentadas as 8 variáveis que caracterizam as escolas, trazendo informação do nível socioeconômico, e que serão utilizadas ao longo deste trabalho, cujas informações foram coletadas da base de dados do Censo Escolar 2019.

Tabela 2: Variáveis relacionadas a infraestrutura da escola dos candidatos do ENEM 2019.

Variável	**Descrição**
CO_MICRORREGIAO	Código que identifica a qual microrregião a escola pertence
auditorio	Variável binária que indica se a escola possui auditório
agua_potavel	Variável binária que indica se a escola possui água potável
localizacao_Urbana	Variável binária que indica se a escola é localizada na área urbana
lab_ciencia	Variável binária que indica se a escola possui laboratório de ciências
piscina	Variável binária que indica se a escola possui piscina
quadra	Variável binária que indica se a escola possui quadra
internet	Variável binária que indica se a escola possui internet
escola_privada	Variável binária que indica se a Dependência Administrativa da escola é privada ou não

Após descrever o material utilizado e as variáveis presentes neste trabalho, será apresentado o Modelo de Regressão Linear Normal que será utilizado.

2.1.3 Área de estudo

Conforme descrito na seção 1.3, um dos objetivos deste trabalho é tentar identificar como explica as notas de matemática e português do ENEM de 2019 a partir de características das escolas e dos indivíduos.

Além disso, a análise de estudo será no nível da escola dos candidatos, ou seja, foram agrupadas por escola as notas que antes estavam no nível do aluno. Sabe-se também que este trabalho visa levar em consideração a localização (microrregião) destas escolas. Para isto foi utilizada a divisão territorial de microrregião do Brasil, disponibilizada no site do IBGE. (BRASIL, 2021)

A Figura 1 apresenta o *shape file* usado neste trabalho que contém as 560 microrregiões do Brasil, e desta forma foi possível identificar a qual microrregião cada escola pertence.

Figura 1: Divisão territorial do Brasil: Microrregiões.

2.2 Modelo de Regressão Linear Normal

O modelo linear normal é um modelo de regressão que tem como objetivo estudar a relação entre variáveis. Neste sentido, seja $\mathbf{Y}$ o vetor da variável resposta de dimensão $n \times 1$, enquanto tem-se $p-1$ são variáveis explicativas, $X_{i,j}$, sendo $i = 1, \ldots, n$ e $j = 1, 2, \ldots, p-1$. (KUTNER et al., 2005)

Desta forma, a Equação (2.1) descreve como é definida essa relação:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + +\beta_2 X_{i,2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i \tag{2.1}$$

Onde:

- Y_i é o valor da variável resposta para a i-ésima observação;
- $X_{i,j}$ é o valor da j-ésima variável preditora (independente) para a i-ésima observação;
- β_0 é um parâmetro desconhecido, que indica o nível médio global;

- β_j são parâmetros desconhecidos ($j = 1, \cdots, p-1$), que medem os efeitos de cada covariável na média da variável resposta.

- ϵ_i é um erro aleatório

  A Equação (2.1) também pode ser reescrita na forma matricial:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.2}$$

  De forma ilustrativa, considerando n pares do tipo $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, tem-se:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ 1 & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

2.2.1 Pressupostos

O modelo de regressão linear descrito na Equação (2.1) possui os seguintes pressupostos:

- A variável resposta $\mathbf{Y}$ é a soma de dois componentes: (1) $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}$, que define um componente sistemático e é um termo não aleatório e (2) $\boldsymbol{\epsilon}$, que é um termo aleatório. Logo, $\mathbf{Y}$ é uma variável aleatória;
- Os erros seguem uma distribuição normal: $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$;
- Os erros ϵ_i são variáveis aleatórias de média zero, então: $\mathrm{E}[\epsilon_i] = 0$, onde $i = 1, \cdots, n$;
- Os erros aleatórios possuem variância constante desconhecida: $\mathrm{V}[\epsilon_i] = \sigma^2$, onde $i = 1, \cdots, n$;
- O erro ϵ_i indica que a variável resposta Y_i na i-ésima observação, se desvia do valor esperado da função de regressão.
- Sabendo que os erros aleatórios $\boldsymbol{\epsilon}$ são não correlacionados, consequentemente as variáveis $\boldsymbol{Y}$ também são não correlacionadas (independentes).

2.2.2 Efeitos Fixos Espaciais

Neste trabalho, serão avaliados indicadores educacionais, de infraestrutura, sociais, econômicos e demográficos que caracterizam escolas e as regiões onde estão localizadas. Entretanto, a base é construída a partir de informações de todo o Brasil. Devido à grande heterogeneidade das informações em todo o território nacional, é adequado que o modelo de regressão considere esses efeitos no processo de estimação. A forma mais adequada para representar efeitos espaciais é feita incluindo termos de erros aleatórios espacialmente estruturados à equação do modelo descrito na Equação (2.2). Entretanto, o custo computacional no procedimento de estimação é proibitivo nesse contexto de escala espacial (microrregião), uma vez que há 558 áreas no Brasil. Como solução, será proposto um modelo de regressão não espacial, similar ao da Equação (2.3), onde os efeitos que capturam o impacto da escola pertencer à determinada região são representados através de efeitos fixos.

Considerando que as escolas estejam ordenadas por microrregião, pode-se definir o seguinte modelo

$$\boldsymbol{Y} = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\gamma} + \boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{2.3}$$

De maneira ilustrativa, segue sua forma matricial:

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \boldsymbol{\gamma} = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{p-1} \end{bmatrix}$$

$$\boldsymbol{Z} = \begin{bmatrix} \mathbf{1}_{n1} & 0 & \cdots & 0 \\ 0 & \mathbf{1}_{n2} & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \mathbf{1}_{w} \end{bmatrix} \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_w \end{bmatrix} \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Aplicando para atual o trabalho tem-se que:

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_n \end{bmatrix} \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \boldsymbol{\gamma} = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{p-1} \end{bmatrix}$$

$$\boldsymbol{Z}= \begin{bmatrix} \mathbf{1}_{n1} & 0 & \cdots & 0 \\ 0 & \mathbf{1}_{n2} & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \mathbf{1}_{n558} \end{bmatrix} \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_{558} \end{bmatrix} \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

onde:

- $\boldsymbol{Y}$ é o vetor da variável resposta, que será as notas médias de matemática ou português das escolas;
- $\boldsymbol{X}$ é a matriz das variáveis preditoras (independentes);
- $\boldsymbol{\gamma}$ são os parâmetros desconhecidos que medem os efeitos de cada covariável na média da variavél resposta;
- $\boldsymbol{Z}$ é uma matriz de variáveis *dummies* que identifica cada uma das microrregiões em que as escolas estão localizadas. Cada coluna da matriz indica uma microrregião k, ($k = 1, \ldots, 558$) e cada linha representa uma escola i, ($i = 1, \ldots, n$);
- $\boldsymbol{\theta}$ representa os efeitos associados a cada uma dessas microrregiões.
- $\boldsymbol{\epsilon}$ é o termo de erro aleatório.

2.2.3 Estimação dos parâmetros

O objetivo desta seção é apresentar os estimadores para os parâmetros desconhecidos do modelo descrito na Equação (2.3).

Há diversos métodos na literatura para se obter o estimador para os parâmetros de um modelo de regressão linear, mas neste trabalho será utilizado o método de mínimos quadrados, proposto por Gauss (1870), que consiste em minimizar a soma dos quadrados dos erros ϵ_i (SQE).

Considere $\boldsymbol{\beta} = (\boldsymbol{\gamma}, \boldsymbol{\theta})^T$ na realização de todas as estimações. Então, a soma dos quadrados dos erros pode ser definida por:

$$SQE = \sum_{i=1}^{n} {\epsilon_i}^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i,1} - \beta_2 X_{i,2} - \cdots - \beta_{p-1} X_{i,p-1})^2 \tag{2.4}$$

Reescrevendo de forma matricial tem-se :

$$SQE = \sum_{i=1}^{n} {\epsilon_i}^2 = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = (\mathbf{Y} - X\boldsymbol{\beta})^T (\mathbf{Y} - X\boldsymbol{\beta}) = (\mathbf{Y}^T - \boldsymbol{\beta}^T \mathbf{X}^T)(\mathbf{Y} - \boldsymbol{X\beta})$$

$$= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\ \boldsymbol{\beta}$$

$$= \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{X}\ \boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\boldsymbol{X\beta}$$

Com isto, resulta na seguinte equação:

$$SQE = \mathbf{Y}^T\mathbf{Y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\mathbf{Y} + \boldsymbol{\beta}^T\mathbf{X}^T\boldsymbol{X\beta} \tag{2.5}$$

A maneira de encontrar o estimador para $\boldsymbol{\beta}$ por mínimos quadrados é através dos pontos mínimos da função, que para achar precisamos derivar SQE em relação a $\boldsymbol{\beta}$ e igualar a zero.

Logo, tem-se a seguinte equação:

$$\begin{aligned} \frac{SQE}{\partial\boldsymbol{\beta}} &= 0 - 2\mathbf{Y}^T\boldsymbol{X} + 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X} = 0 \\ &\iff -2\boldsymbol{Y}\boldsymbol{X}^T + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}^T = 0 \\ &\iff \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}^T = \boldsymbol{X}^T\mathbf{Y} \\ &\iff \widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{X}^T\mathbf{Y} \end{aligned} \tag{2.6}$$

Com isto, encontra-se o estimador de $\boldsymbol{\beta}$ a partir do método de mínimos quadrados. Agora, serão realizados os cálculos da esperança e variância de $\widehat{\boldsymbol{\beta}}$.

- Esperança de $\widehat{\boldsymbol{\beta}}$

$$\begin{aligned} E[\widehat{\boldsymbol{\beta}}] &= E\left[(\boldsymbol{X}^T\boldsymbol{X})\mathbf{X}^T\mathbf{Y}\right] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E\left[\boldsymbol{Y}\right] \\ &= \underbrace{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}}_{\mathbf{I}}\boldsymbol{\beta} = \boldsymbol{\beta} \end{aligned} \tag{2.7}$$

Nota-se que a matriz $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ é simétrica, o que significa que $\widehat{\boldsymbol{\beta}}$ é um estimador não tendencioso para o vetor de parâmetros $\boldsymbol{\beta}$.

- Variância de $\widehat{\boldsymbol{\beta}}$

$$\begin{aligned} Var[\widehat{\boldsymbol{\beta}}] &= Var\left[(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{X}^T\boldsymbol{Y}\right] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \underbrace{Var\left[\boldsymbol{Y}\right]}_{\sigma^2 I}\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right]^T \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right]^T \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\underbrace{\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}}_{\mathbf{I}} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{aligned} \tag{2.8}$$

Logo, tem-se que a distribuição de probabilidade do vetor $\boldsymbol{Y}$ é Normal Multivariada, ou seja, NMV($\boldsymbol{X\beta}$,$\sigma^2\mathbf{I}$). Aplicando a transformação linear de normal multivariada, conclui-se que $\widehat{\boldsymbol{\beta}}$ é normal multivariada que terá dimensão p com o vetor de esperanças e matriz de variâncias e covariâncias calculadas nas Equações (2.7) e (2.8). Então,

$$\widehat{\boldsymbol{\beta}} \sim \text{NMV}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

2.2.4 Estimação da variância

Um indicador da variabilidade da distribuição de probabilidade de $\boldsymbol{Y}$ é a variância σ^2 do termo de erro ϵ_i. O estimador de σ^2 é obtido a partir da Soma dos Quadrados do Erro (SQE), que é a soma da diferença entre os valores Y_i da amostra e os valores ajustados $\widehat{Y}_i$, elevada ao quadrado. Então sua forma matricial é:

$$SQE = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{Y}^T\mathbf{Y} - \widehat{\boldsymbol{\beta}}\mathbf{X}^T\mathbf{Y} \tag{2.9}$$

Considerando que p parâmetros são estimados no modelo de regressão, a SQE tem $n - p$ graus de liberdade. Além disto, sabe-se que o MQE definido como:

$$MQE = \frac{\boldsymbol{Y}^T\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}\boldsymbol{X}^T\boldsymbol{Y}}{n-p} \tag{2.10}$$

é um estimador não viesado para variância σ^2, então tem-se que:

$$\hat{\sigma}^2 = MQE = \frac{SQE}{n-p} \tag{2.11}$$

2.2.5 Resíduos do modelo

É relevante a análise dos resíduos pois através deles pode-se detectar se há violação das suposições do modelo, avaliando também o seu ajuste. Ele é definido como:

$$e_i = Y_i - \widehat{Y}_i, \qquad i = 1, \cdots, n \tag{2.12}$$

Onde:

- Y_i é o valor observado da variável resposta
- $\widehat{Y}_i$ é o valor ajustado da variável resposta, com $\widehat{\boldsymbol{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$

A partir dele pode-se conferir se as propriedades necessárias realmente estão presentes no modelo de regressão, através de análises gráficas e utilizado os testes de Breusch–Pagan e Lilliefors, para a homocedasticidade e normalidade, respectivamente. As propriedades são:

- Linearidade
- Homocedasticidade: erros com variância σ^2 constante.
- Normalidade: Os resíduos possuem Distribuição Normal.

Define-se o vetor de resíduos como:

$$\mathbf{e}^* = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \tag{2.13}$$

Isto é:

$$\mathbf{e} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} x_1^T\widehat{\beta} \\ x_2^T\widehat{\beta} \\ \vdots \\ x_n^T\widehat{\beta} \end{bmatrix}$$

Ademais, é mais usual a utilização do resíduo padronizado, que é definido por:

$$\mathbf{e}^* = \frac{\mathbf{Y} - \widehat{\mathbf{Y}}}{\sqrt{MQE}} \tag{2.14}$$

Onde $MQE = \frac{\mathbf{e}^T\mathbf{e}}{n-p}$ é o erro quadrático médio, apresentado na seção 2.2.4. Através da análise dos resíduos $\mathbf{e}^*$, é possível diagnosticar os desvios em relação aos pressupostos definidos na seção 2.2.1.

Através da análise gráfica, a Figura 2 representa o que espera-se com base nos pressupostos. Em contrapartida, as Figuras 3 e 4 apresentam exemplos em que há violação dos pressupostos em relação a homocedasticidade e normalidade, respectivamente, conforme indicado por Kutner et al. (2005).

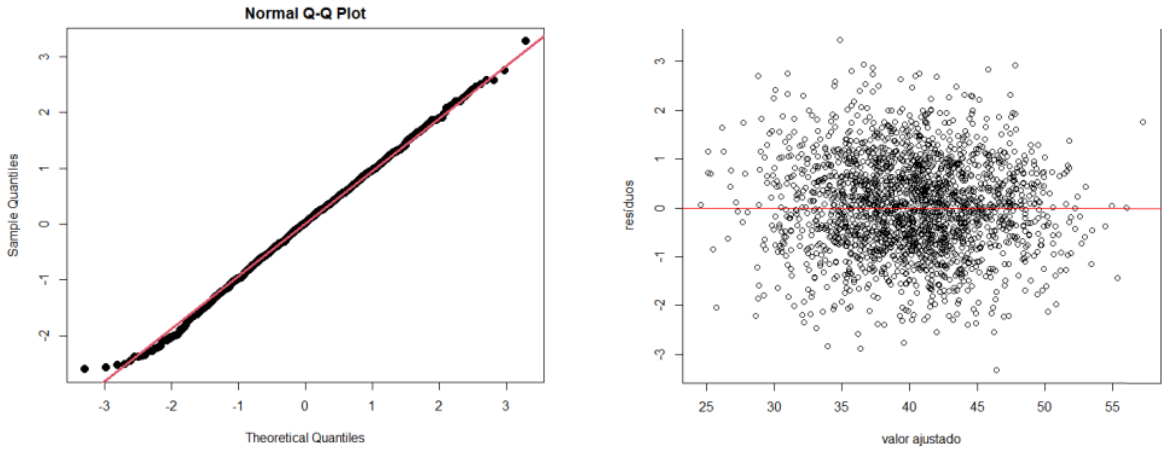


Figura 2: Comportamento de Resíduos que seguem os pressupostos.

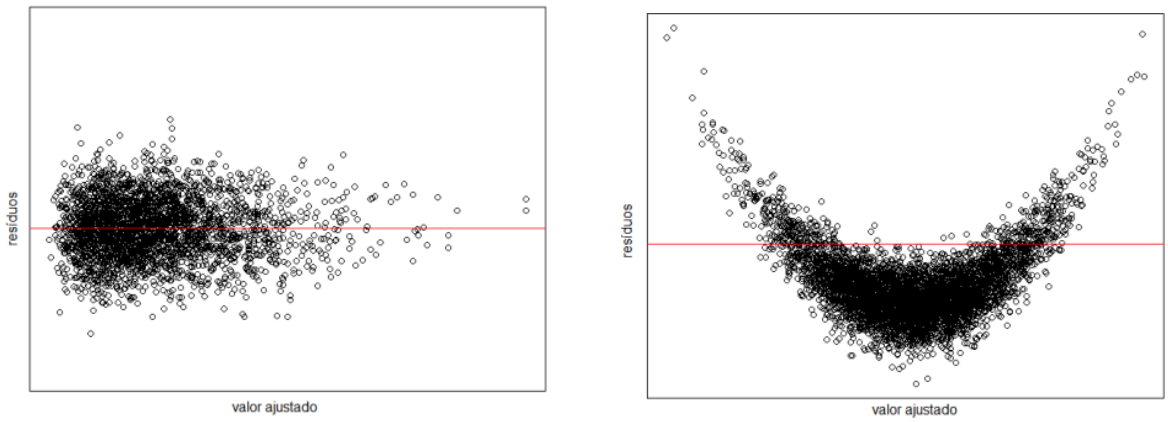


Figura 3: Exemplos de Análise de resíduos nos quais a variância não é constante.

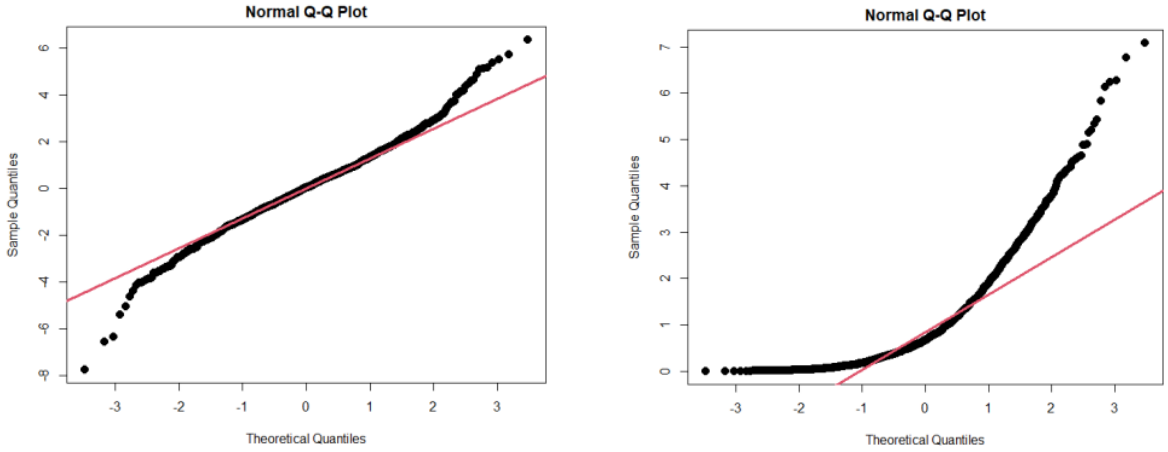


Figura 4: Exemplos de Análise de resíduos nos quais a variância não é constante.

2.2.5.1 Teste de Breusch–Pagan

A realização do teste de Breusch–Pagan é uma das formas de verificar se a variância dos erros é constante.(KUTNER et al., 2005)

- Hipóteses:

$$\begin{cases} H_0 : VAR(\epsilon_i|x_{1i}, x_{2i}, \cdots, x_{ki}) = VAR(\epsilon_i) = \sigma^2, \quad i = 1, 2, \cdots, n \\ H_1 : VAR(\epsilon_i|x_{1i}, x_{2i}, \cdots, x_{ki}) = VAR(\epsilon_i) = \sigma_i^2, \quad i = 1, 2, \cdots, n \end{cases}$$

- Estatística de teste:

$$X^2_{BP} = \frac{\frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y}_i)^2}{2}}{\left(\frac{\sum_{i=1}^{n} y_i - \widehat{y}_i)^2}{n}\right)^2} \sim X^2_1 \tag{2.15}$$

Através da estatística de teste, obtem-se o p-valor que é dado por:

$$p - valor = P[x_1^2 > X_{BP}^2 | H_0] \tag{2.16}$$

A tomada de decisão é feita pela avaliação do p-valor. Quando ele for menor ou igual a α rejeita-se a hipótese nula (H_0), ou seja, indica-se que a variância dos erros não é constante.

2.2.5.2 Teste de Lilliefors

O teste de Lilliefors é utilizado para verificar se a os erros seguem uma distribuição normal. Esse teste é uma adaptação do teste de Kolmogorof-Smirnoff, que é feita com média e variância conhecidas, já o teste de Lilliefors é realizado utilizando as estimativas de média e variância. (LILLIEFORS, 1967)

- Hipóteses:

$$\begin{cases} H_0 : \text{Os dados seguem a distribuição normal} \\ H_1 : \text{Os dados não seguem a distribuição normal} \end{cases}$$

A estatística de teste é a diferença absoluta entre a função de distribuição acumulada padronizadas, denominada por $F(x_i)$ e a teórica, chamada de $F_n(x_i)$. Dessa forma, tem-se:

$$D_n = Max|F(x_i) - F_n(x_i)| \tag{2.17}$$

Desta forma, precisa-se padronizar os dados primeiros, da seguinte forma:

$$z_i = \frac{y_i - \overline{y}}{s}, i = 1, 2, \cdots, n. \tag{2.18}$$

onde,

$$s = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right]} \tag{2.19}$$

Logo, a região crítica do teste é definida por $RC = \left\{Z \in R | D_n \geq z_{\alpha/2}\right\}$. A hipótese nula é rejeitada ao nível de significância de α quando o valor observado para Z pertencer à região crítica.

2.3 LASSO

É notório que os modelos em análise neste estudo possuem muitas covariáveis para serem estimadas, o que gera um elevado custo computacional e um baixo desempenho devido ao super ajuste da estimação de mínimos quadrados. Contudo, sabe-se que em alguns casos existem variáveis no modelo que não são relevantes para a explicação da variável resposta, então aplicou-se o método de regularização LASSO para realizar a seleção de variáveis. Ademais, a exclusão das variáveis que não têm relevância para o modelo ajustado é um processo importante para obter-se o melhor modelo, pois caso ao contrário o modelo terá baixo poder preditivo, já que a variância do estimador é alta devido a quantidade de parâmetros.

Existem diversos métodos na literatura que realizam a seleção de variáveis, mas neste trabalho optou-se por usar o método de regularização LASSO (*Least Absolute Shrinkage and Selection Operator*), proposto por Tibshirani (1996), pois é um método mais atual, bastante utilizado na área de machine learning e possui boas propriedades. Ele também é denominado L1, e é um penalizador da estimação dos parâmetros. Este método minimiza a soma dos quadrados dos resíduos com uma restrição nos parâmetros de regressão $\boldsymbol{\beta}$ a serem estimados. Com isto, a estimativa de $\boldsymbol{\beta}$ usando o LASSO é definida como:

$$arg\min_{\boldsymbol{\beta}} = (\sum_{i=1}^{n}(Y_i - \boldsymbol{X_i}\boldsymbol{\beta})) + \lambda \sum_{j=1}^{p-1} \mid \beta_j \mid \tag{2.20}$$

onde

$$\boldsymbol{X_i} = [X_{1i}, \cdots, X_{p-1}]^T \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

e $\lambda \geq 0$, em que λ representa o parâmetro de penalização. Nesse caso, quanto maior for o valor de λ, maior é a penalização e mais coeficientes são zerados. Nota-se também, pela Equação (2.20), que foi acrescentado um viés às estimativas de mínimos quadrados. A Figura 5 exemplifica esse viés adicionado às estimativas de mínimos quadrados e a penalização em direção a zero.

O LASSO é uma das técnicas encontradas na literatura para redução da dimensão do problema citado. Ele causa um aumento de viés para que em contrapartida resulte

na diminuição da variância da função de predição, buscando um bom balanceamento. A Figura 5 ilustra este procedimento. Considera-se que o modelo possui apenas 2 parâmetros (β_1 e β_2), a curva é a de mínimos quadrados (SQE) e o ponto no meio representa os betas que minimizam a função de erro. Então ao mesmo tempo que o LASSO acrescenta um viés, ele também é responsável por zerar um dos parâmetros do modelo. Através da Figura 5 este viés parece ser alto, mas na prática ao verificar essas estimativas, observa-se que ele não é tão grande quanto parece.

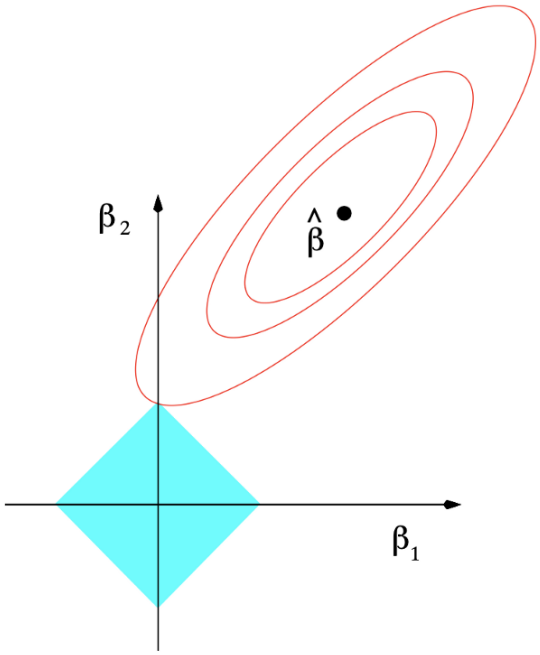


Figura 5: Exemplo do método de regularização LASSO. $\widehat{\boldsymbol{\beta}}$: estimativa por mínimos quadrados.

Fonte: Tibshirani (1996)

Enquanto realiza a penalização dos parâmetros este método também é responsável pela seleção de variáveis significativas do modelo, de tal forma que ele zera determinadas estimativas dos coeficientes, e com isso ocorre a eliminação destas variáveis do modelo.

2.4 Validação Cruzada

O valor escolhido de λ é essencial para que o método de LASSO funcione adequadamente. Vale ressaltar que quando $\lambda = 0$ o método de estimação se torna o de mínimos quadrados, o que não favorece já que ele não seleciona as covariáveis do modelo. Além

disso, escolher um valor muito grande para λ também não é uma boa opção, pois é possível acabar tendo em mãos um modelo sem covariáveis, apenas com o intercepto.

O método que será implementado neste trabalho para escolher o valor de λ é o de validação cruzada. Primeiro, escolhe-se diversos valores para λ e em seguida realiza-se os cálculos do erro para cada valor de λ. Por fim, o valor selecionado para λ será o que tiver obtido o menor valor de erro, que neste caso é o Erro Quadrático Médio, conforme definido na Equação (2.10).

Vale ressaltar que não é necessário o conhecimento dos graus de liberdade do modelo, ou seja, o número de parâmetros para a escolha de λ, pois no caso do método LASSO ele é considerado também como uma quantidade de impacto no ajuste do modelo.

Considerando n_i o número de observações da i-ésima parte dos dados, tem-se que o erro de predição relacionado a λ é dado por:

$$MQE_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (Y_j - \widehat{Y}_j)^2 \tag{2.21}$$

Através do cálculo apresentado na Equação (2.21), escolhe-se o melhor valor para λ, que é aquele que possuir menor erro de predição. (TIBSHIRANI, 1996)

3 Resultados

Este capítulo será dividido em 3 partes. A primeira se refere às manipulações realizadas na base para que pudessem ser realizadas as demais análises. A segunda tratará das análise descritivas das notas de matemática e português do ENEM de 2019, com a informação das escolas com características dos candidatos e infraestrutura escolar, acrescida de efeitos fixos que caracterizam áreas. A última abordará as análises dos dois modelos lineares normais, para a nota de matemática e português, apresentados na seção 2.2.

3.1 Limpeza e Manipulação da base

A base do ENEM de 2019 continha 5.095.270 observações de alunos que foram agrupados pelas escolas dos candidatos, sendo calculado as notas médias das provas de matemática e português. Mesmo após esta manipulação algumas escolas não obtinham informação de nenhum aluno e a medida tomada foi a remoção destas escolas da base, ficando no total com 28.886 escolas com notas para matemática e 28.99 escolas com notas para português.

As informações das características dos candidatos após o agrupamento por escola ficaram de forma proporcional, ou seja, proporção de candidatos na escola com determinada característica, para que não perdêssemos esses registros. Então, por exemplo, na base existe a proporção de quantas mulheres, pessoas brancas e maiores de idade possuem em cada uma das escolas.

Ademais, foram criadas *dummies* da variável *CO_MICRORREGIAO*, ou seja, 558 novas variáveis no modelo, que são os códigos das microrregiões do Brasil, indicando se cada escola pertence ou não àquela região.

3.2 Análise Exploratória

Nessa seção será feita uma breve análise descritiva dos dados utilizados neste trabalho, a fim de entender melhor como é o comportamento dos dados.

3.2.1 Análises das escolas no Brasil

As análises foram iniciadas no nível Brasil, sem distinção da localidade no candidato.

3.2.1.1 Características dos candidatos

A Tabela 3 apresenta as estatísticas descritivas referentes à nota média de matemática das escolas no ENEM de 2019. Conclui-se que a média das notas é de 523,1, sendo a nota média máxima de uma escola 985,5 pontos. Além disso, há indicativos de que as notas médias de matemática das escolas possuem uma distribuição assimétrica positiva, visto que o valor da média é maior do que o da mediana, que é maior do que a moda (nota mais frequente).

Medida	Nota de matemática
Mínimo	0
Q1	435,1
Mediana	501,1
Média	523,1
Q3	597,8
Máximo	985,5
Moda	482,4
Desvio Padrão	73,7
Coeficiente de variação	14,2

Tabela 3: Tabela de medidas descritivas da nota média de matemática das escolas no ENEM de 2019.

A Tabela 4 apresenta as estatísticas descritivas referentes à nota média de português das escolas no ENEM de 2019. Conclui-se que a média das notas é de 520,5, sendo a nota

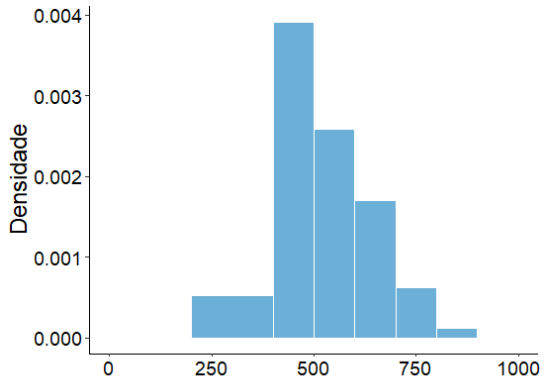
média máxima de uma escola 801,7 pontos. Além disso, há indicativos de que as notas médias de português das escolas possuem uma distribuição assimétrica negativa, visto que o valor da média é menor do que o da mediana, que é menor do que a moda (nota mais frequente).

Medida	Nota de português
Mínimo	0
Q1	483,5
Mediana	526,2
Média	520,5
Q3	565,3
Máximo	801,7
Moda	489,9
Desvio Padrão	42,7
Coeficiente de variação	8,3

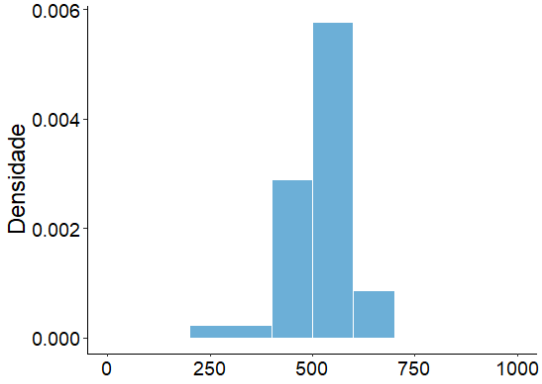
Tabela 4: Tabela de medidas descritivas da nota média de português das escolas no ENEM de 2019.

A Figura 6(a) indica o comportamento da nota média de matemática das escolas no ENEM de 2019. As análises demonstram que obteve-se uma maior frequência de notas entre 400 e 500, e a presença de notas entre 700 e 900. Além disso, a análise gráfica induz mais ainda a suposição de assimetria positiva, visto que há uma maior frequência de notas maiores que a média (523,1). Enquanto a Figura 6(b) indica o comportamento da nota média de português das escolas no ENEM de 2019. Através das análises observa-se que há uma maior frequência de notas entre 500 e 600, com baixa frequência de notas maiores que 700. Além disso, a análise gráfica também induz mais ainda a suposição de assimetria negativa, visto que há uma maior frequência de notas menores que a média (520,5).

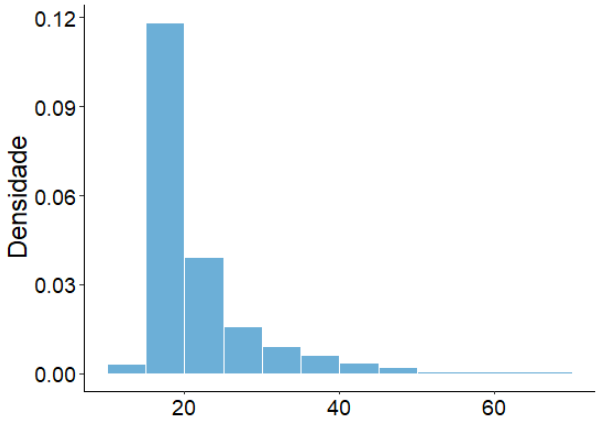
Além disso, a Figura 6(c) mostra que a maioria dos candidatos do ENEM de 2019 possui idade entre 15 e 20 anos, o que é coerente já que neste intervalo tem a idade que geralmente estão no terceiro ano do ensino médio e realizam a prova a fim de ingressar em uma universidade.



(a) Nota na prova de matemática



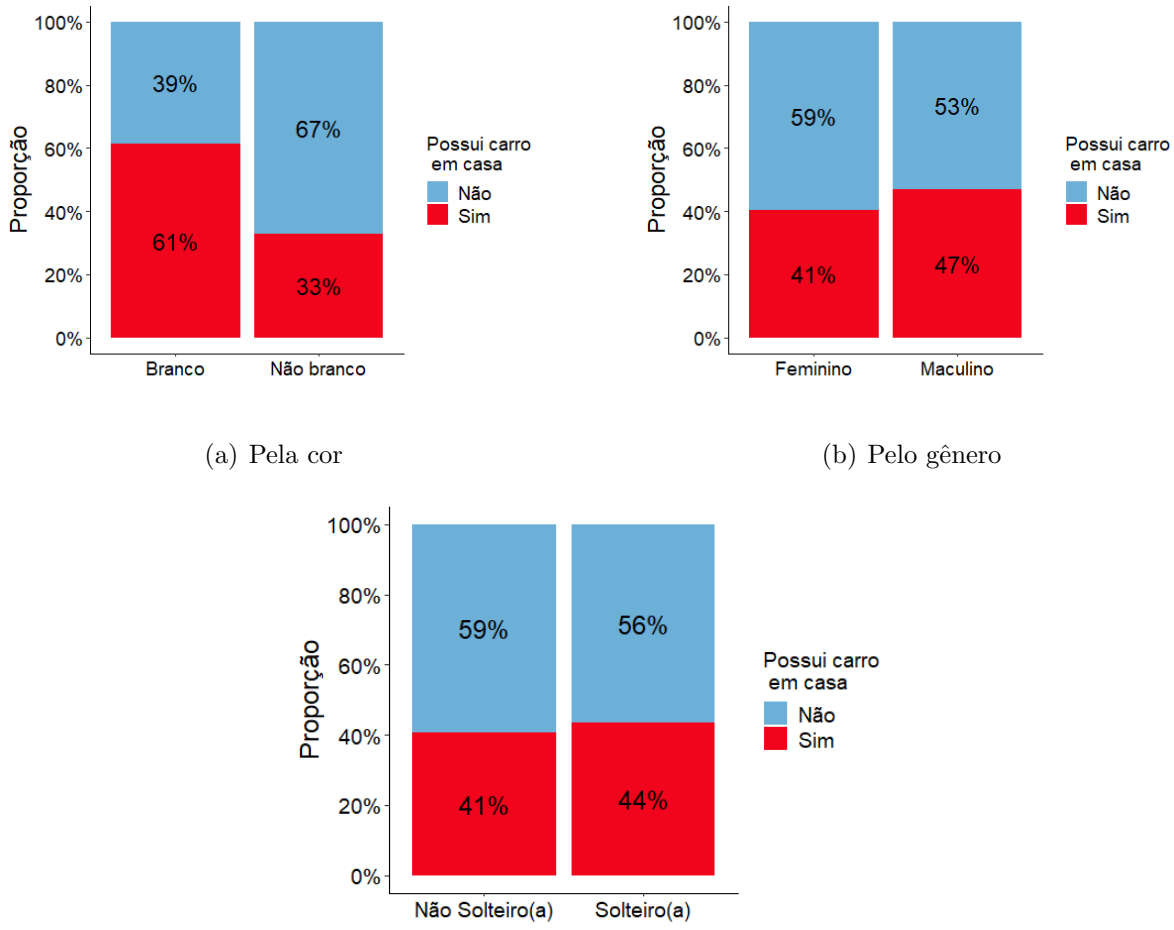
(b) Nota na prova de português



(c) Idade

Figura 6: Características gerais dos candidatos: Notas no ENEM 2019 e idade dos candidatos

A Figura 7 indica as proporções de candidatos que possuem carro por características físicas e pessoais, tais como, cor/raça, gênero e estado civil. Pode-se perceber que há uma diferença relevante dos candidatos que possuem carro quando se comparado pela cor, visto que mais da metade deles são brancos (61%). Quando comparado em relação a gênero e estado civil não se obteve muita diferença entre os jovens que possuem ou não carro, sendo uma diferença de 6 pontos percentuais e 3 pontos percentuais, respectivamente.



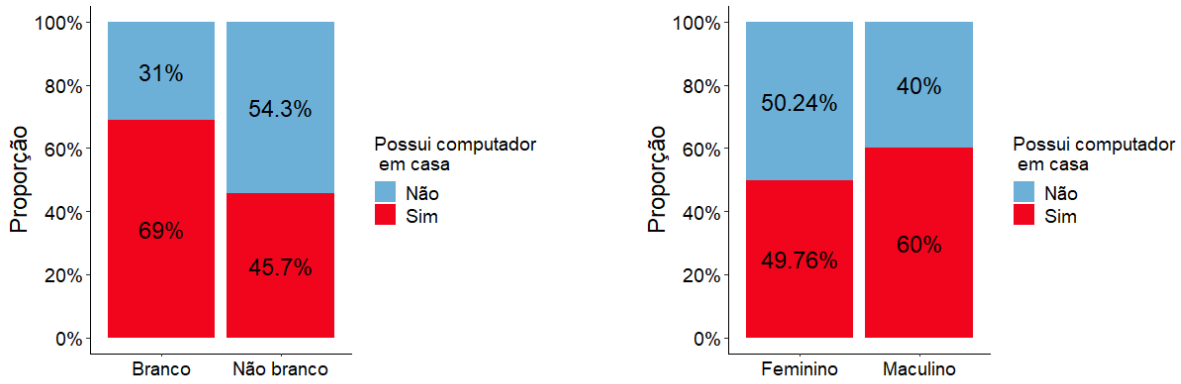
(a) Pela cor

(b) Pelo gênero

(c) Por estado civil

Figura 7: Proporção de candidatos que possuem carro por suas características.

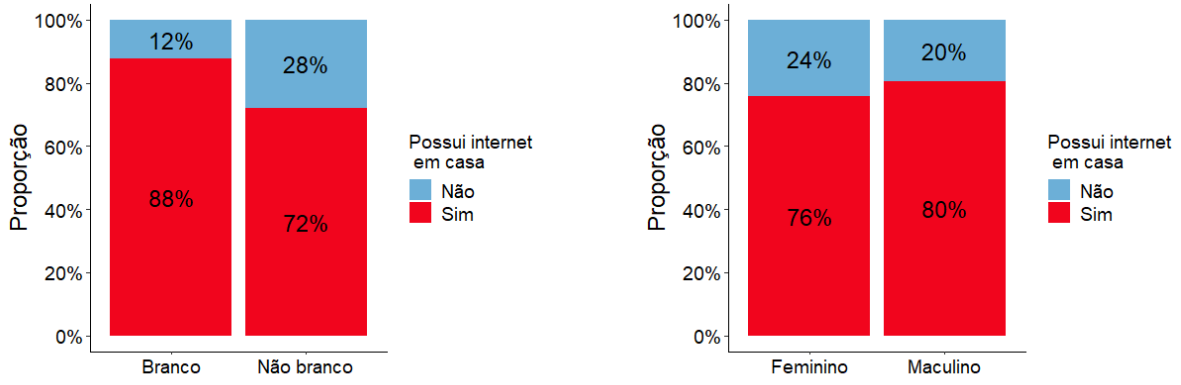
A Figura 8 demonstra a proporção de alunos que possuem computador em casa por suas características físicas. O computador é um elemento que influencia no estudos dos alunos, facilitando a forma de aprendizado. Através da análise pode-se perceber que tanto a cor quanto o gênero são variáveis que são relevantes na presença de computador em casa, já que 69% dos candidatos que têm este bem são de cor branca e 60% do gênero masculino.



(a) Pela cor (b) Pelo gênero

Figura 8: Proporção de candidatos que possuem computador em casa por suas características físicas.

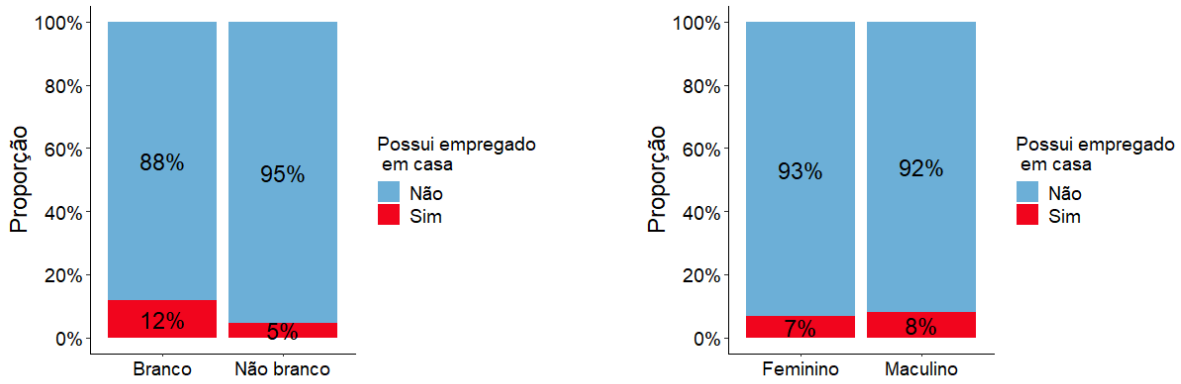
A internet é outro elemento essencial para uma melhor produtividade no estudo. Ela facilita na procura de conteúdos e/ou no esclarecimento de dúvidas. A Figura 9 possui indícios de que a variável da raça do candidato é importante na presença de internet na casa do candidato, sendo 88% brancos. Em contrapartida, o gênero do indivíduo não demonstrou ser relevante para esta questão.



(a) Pela cor (b) Pelo gênero

Figura 9: Proporção de candidatos que possuem internet em casa por suas características físicas.

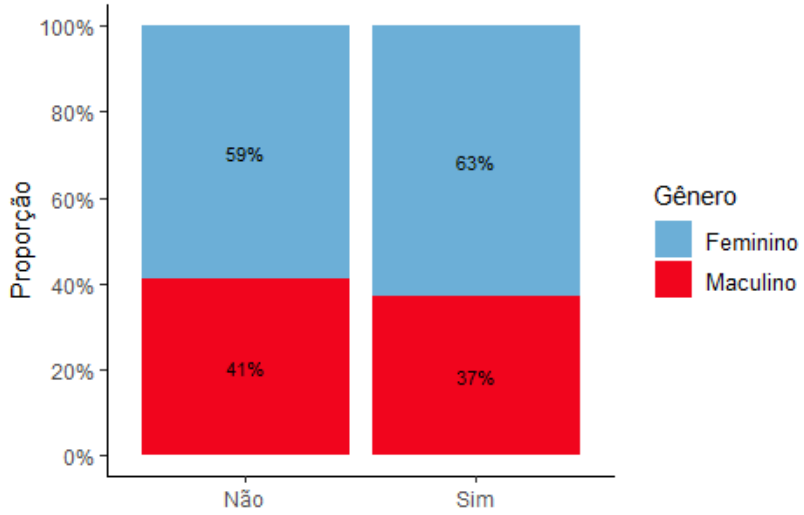
A Figura 10 indica que poucos candidatos possuem empregado(a) em casa, sendo 12% de cor branca e apenas 5% não bancos. Além disso, a diferença entre os gêneros é praticamente desprezível, sendo 4% divergente.



(a) Pela cor (b) Pelo gênero

Figura 10: Proporção de candidatos que possuem empregado(a) em casa por suas características físicas.

A Figura 11 mostra que a maioria dos candidatos que fizeram a prova do ENEM apenas por experiência, pois não estão ainda no terceiro ano do ensino médio, são do gênero feminino, representado por 63% do total de treineiros da prova do ENEM de 2019.



(a) Candidatos treineiros

Figura 11: Proporção de candidatos que para treinar realizaram a prova do ENEM de 2019 por gênero.

3.2.1.2 Infraestrutura Escolar

Ainda no nível Brasil, foram realizadas análises sobre a infraestrutura da escola dos candidatos do ENEM de 2019.

A Figura 12 indica a proporção da dependência administrativa das escolas dos candidatos. Pode-se verificar que a maioria delas são de caráter público, abrangendo escolas municipais, estaduais e federais, representando 72% do total, e somente 28% são escolas privadas.

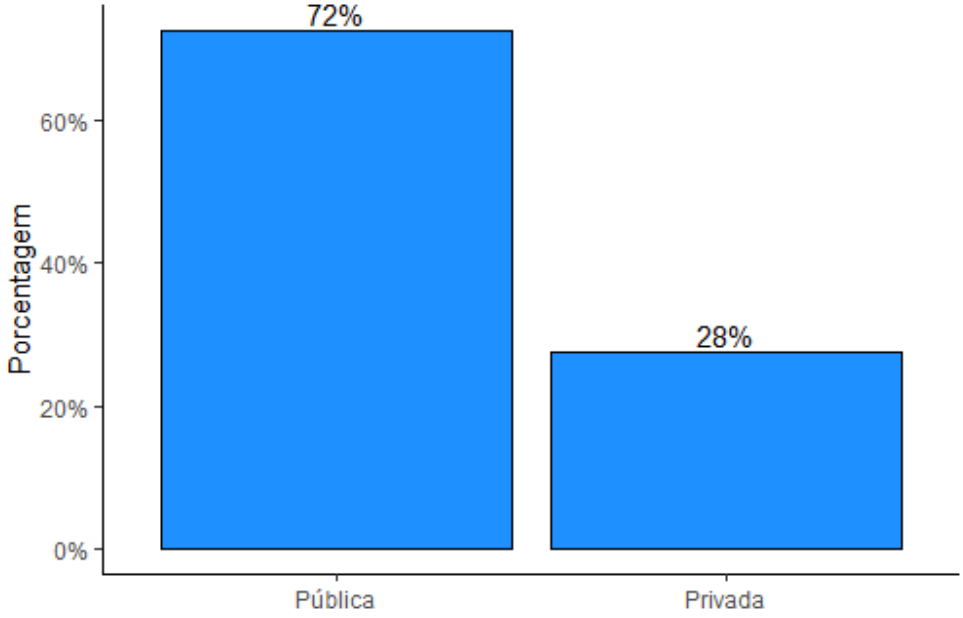
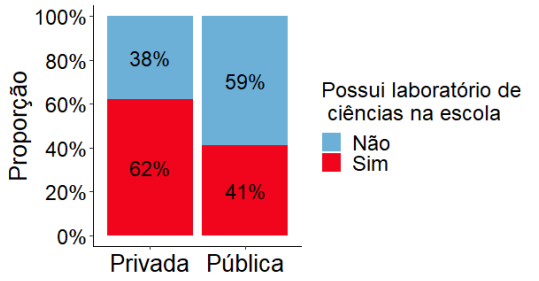


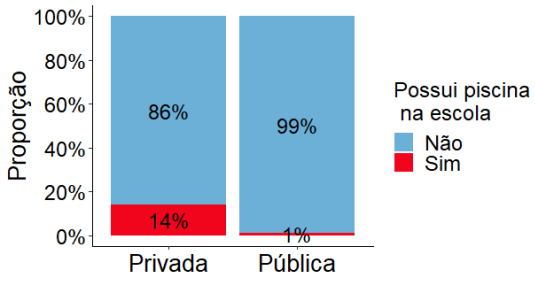
Figura 12: Proporção da dependência administrativa da escola.

A Figura 13 demonstra a distribuição da dependência administrativa da escola por características de infraestrutura. Percebe-se que a maioria das escolas que possuem laboratório de ciência são privadas (62%). Por outro lado, a variável piscina não é presente na maioria das escolas, tanto pública como privada, mas entre os 2 tipos, encontra-se em maior quantidade (14%) das escolas privadas.

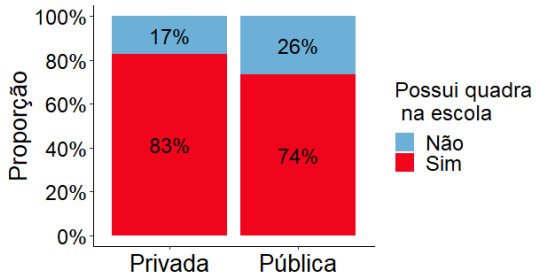
Além disso, a Figura 13(c) indica que mais da metade das escolas dos 2 os tipos de dependência administrativa possuem quadra de esportes na escola, um ponto positivo que apareceu nas análises.



(a) Pela presença de laboratório de ciência



(b) Pela presença de piscina



(c) Pela presença de quadra de esportes

Figura 13: Proporção da Dependência Administrativa da escola por características de infraestrutura.

A Figura 14 mostra que no Brasil a proporção de escolas na localização urbana é bem maior do que na rural, caracterizado por 91% do total das escolas.

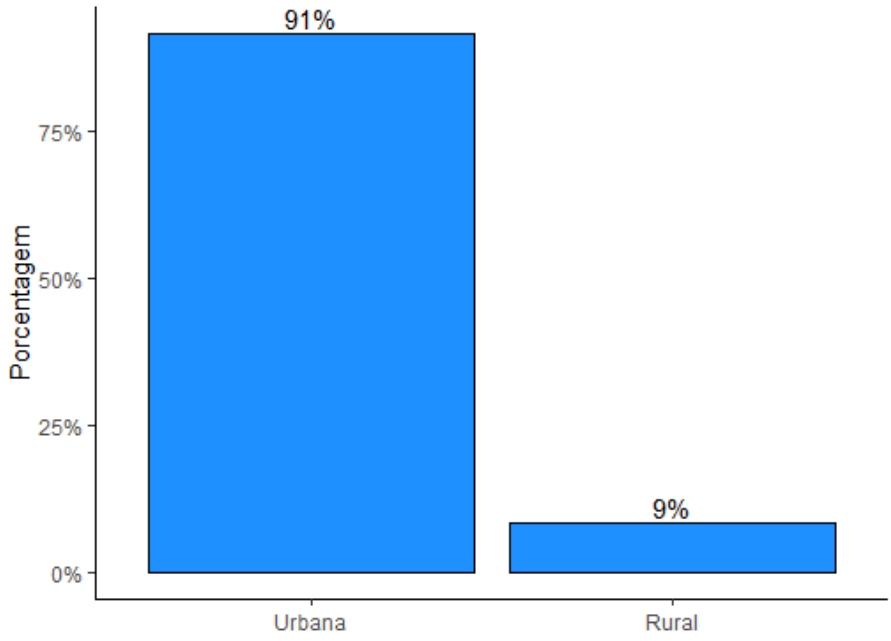
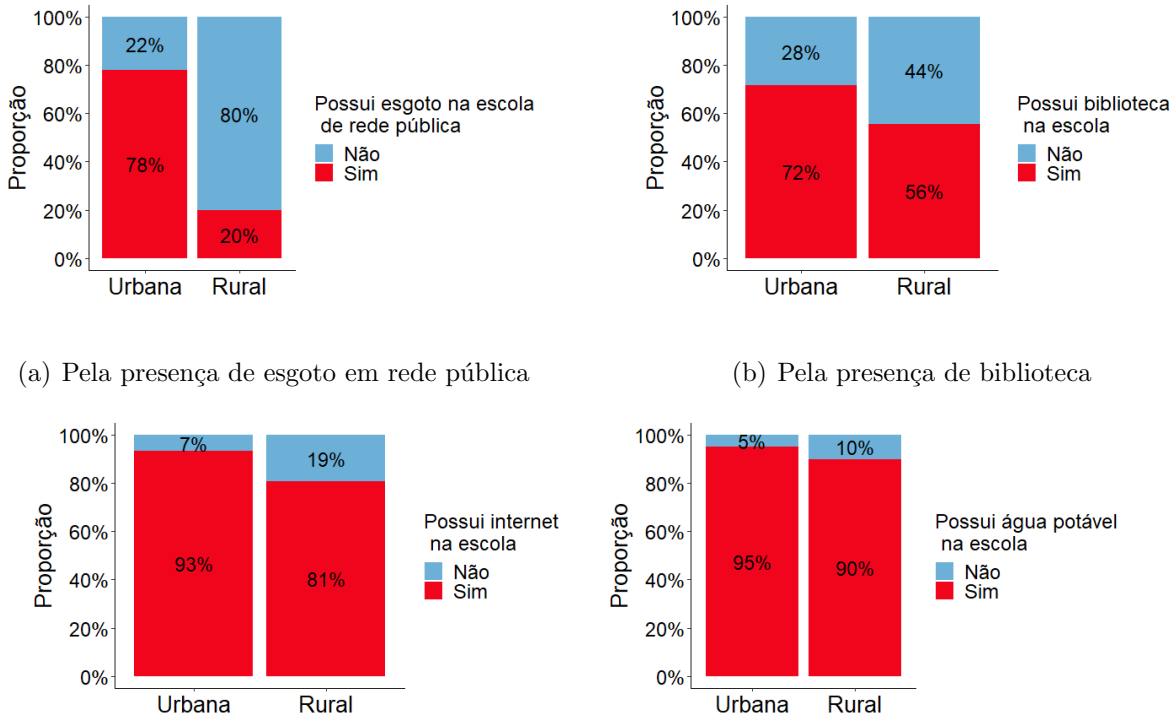


Figura 14: Localização da escola.

A Figura 15 retrata a proporção das características de infraestrutura da escola por localização (urbana ou rural). Através das análises pode-se verificar que a característica que mais é influenciada pela localização é a presença de esgoto em escola de rede pública, visto que em escolas públicas urbanas 78% possuem esgoto, enquanto nas rurais apenas 20% têm.

A presença biblioteca e internet na escola foram variáveis que também demonstraram comportamentos diferentes dependendo da localização, pois em todos os casos as escolas urbanas possuíram uma proporção de pelo menos 12 pontos percentuais a mais do que nas escolas rurais. Por fim, a presença de água potável na escola foi a variável que menos apresentou influência pela localização, obtendo uma diferença de apenas 5 pontos percentuais entre elas.



(a) Pela presença de esgoto em rede pública

(b) Pela presença de biblioteca

(c) Pela presença de internet

(d) Pela presença de água potável

Figura 15: Proporção de características de infraestrutura da escola por localização.

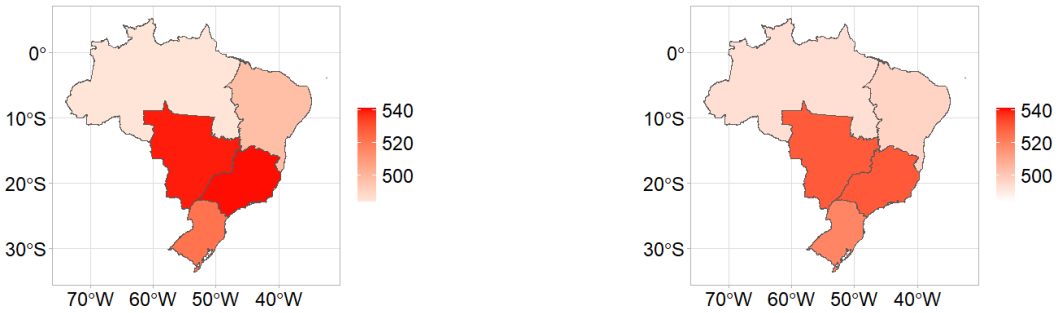
3.2.2 Análises das escolas segundo grandes regiões

As análises ao nível Brasil são válidas para entendermos de forma geral como são os comportamentos das variáveis do estudo. No entanto, sabe-se que existem diferenças regionais entre elas, e por isso serão feitas análises considerando a região dos candidatos

e das escolas.

3.2.2.1 Características dos candidatos

A Figura 16 demonstra a diversidade das notas de português e matemática das escolas nas diferentes regiões do Brasil. O Centro-Oeste e o Sudeste são as regiões que apresentam melhor rendimento dos candidatos nas provas, sendo a nota de matemática um pouco superior.

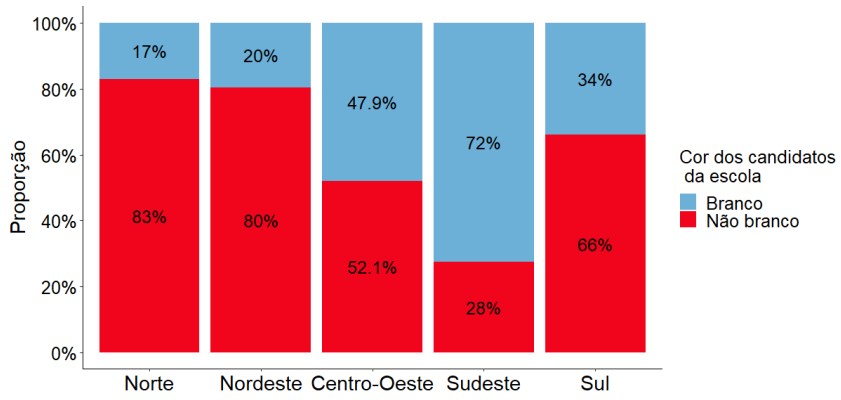


(a) Notas na prova de matemática

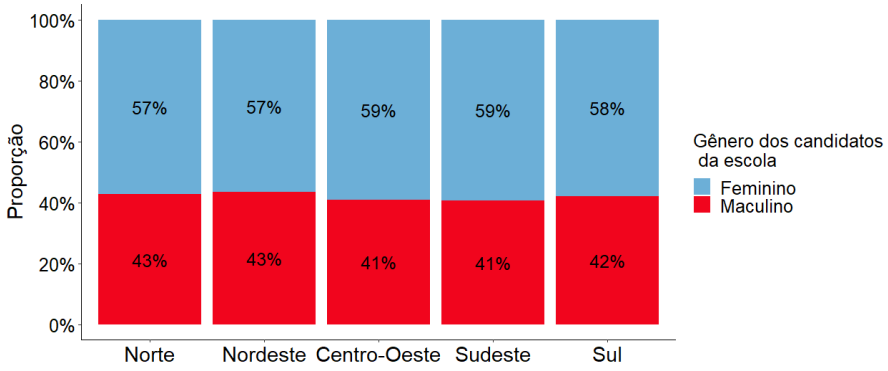
(b) Notas na prova de português

Figura 16: Comportamento da nota de matemática e português pelas grandes regiões do Brasil.

A Figura 17 demonstra a proporção de características demográficas dos candidatos do ENEM de 2019 pela localização das escolas nas grandes regiões. Percebe-se que o Sudeste é a região com escolas que mais possui candidatos de cor branca, sendo mais da metade (72,5%). A proporção de candidatos do gênero feminino não varia muito nas diferentes escolas das regiões, sendo geralmente um pouco mais da metade dos alunos das escolas do gênero feminino.



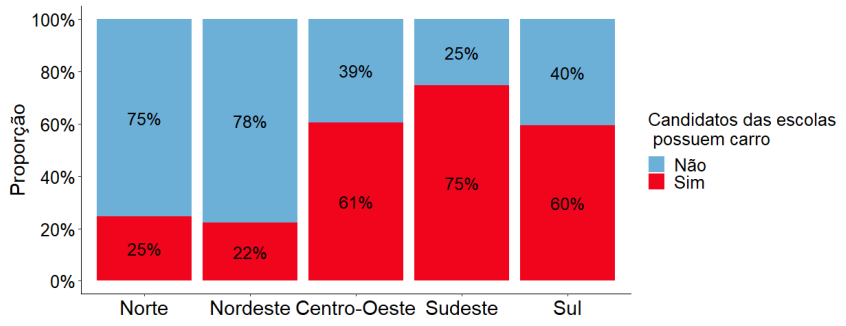
(a) Pela cor do candidato



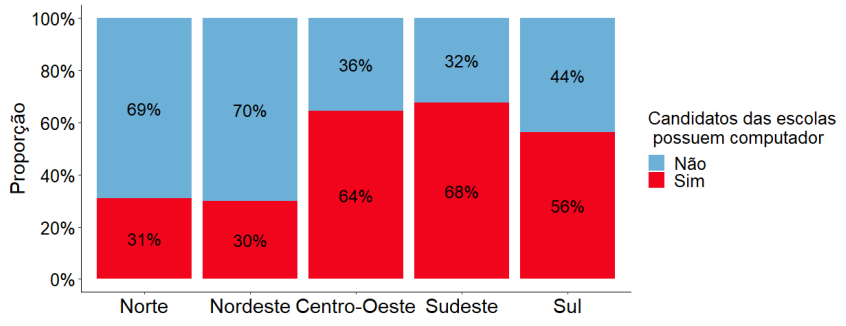
(b) Pelo gênero do candidato

Figura 17: Proporção de características físicas dos candidatos do ENEM de 2019 por região.

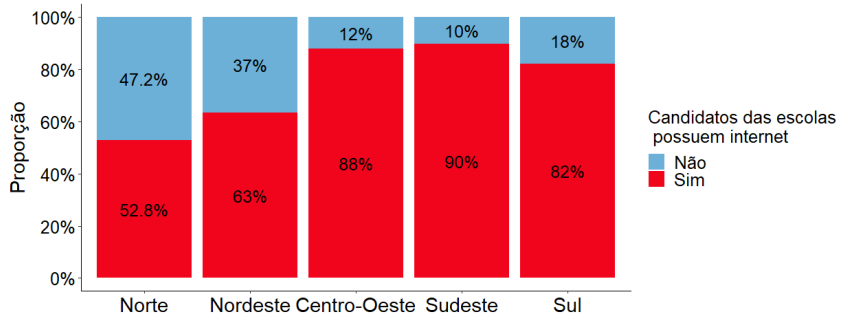
A Figura 18 indica que o Sudeste é a região que mais tem escolas com candidatos que possuem carro, computador e internet, refletido pelo poder socioeconômico da área. Em contrapartida, a maior proporção de escolas que possuem candidatos que possuem empregado(a) em casa é no Sul (10%). Além disso, percebe-se que o Norte e os Nordeste são as regiões nas quais os candidatos das escolas menos possuem bens materiais, como a presença de carro, computador e internet em casa, como também a presença de empregado(a).



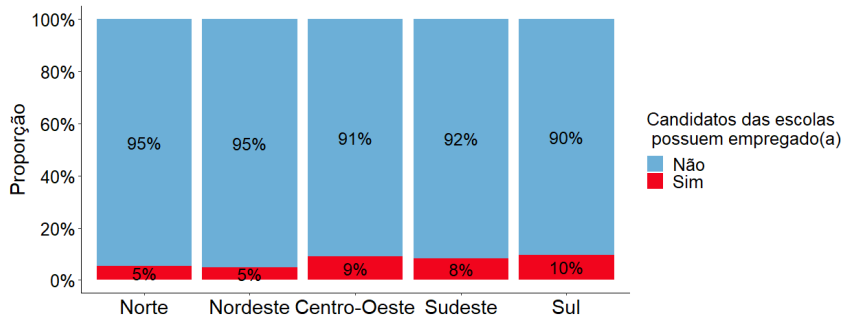
(a) Possui carro



(b) Possui computador



(c) Possui internet



(d) Possui empregado(a)

Figura 18: Proporção de características socieconômicas nas casas dos candidatos do ENEM de 2019 por região.

3.2.2.2 Infraestrutura Escolar

Assim como as características dos indivíduos variam conforme a região das escolas, as características de infraestrutura das escolas dos candidatos também sofrem essas modificações. Por isto, foram feitas análises no nível da escola por região do Brasil.

A Figura 19 indica a proporção da Dependência Administrativa das escolas dos candidatos por região do Brasil. Observa-se que em todas as regiões a maioria das escolas dos candidatos do ENEM de 2019 são escolas públicas (estaduais).

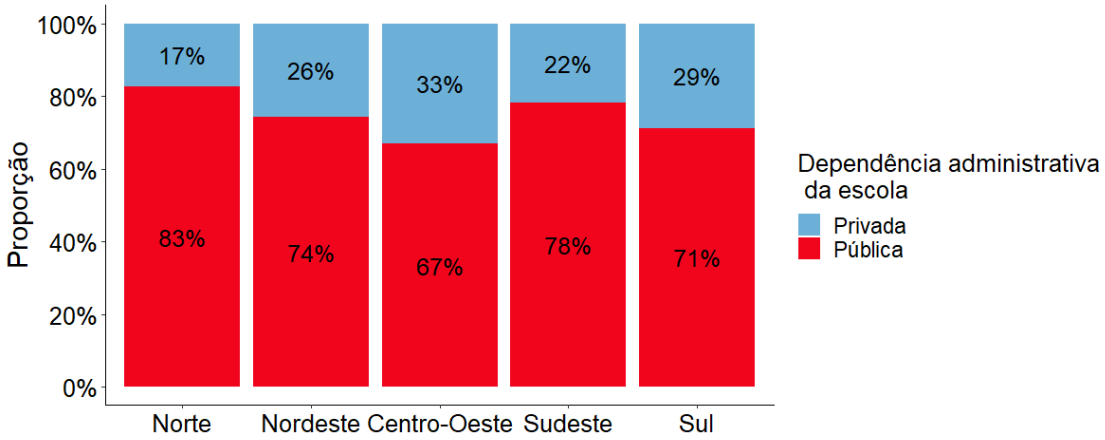
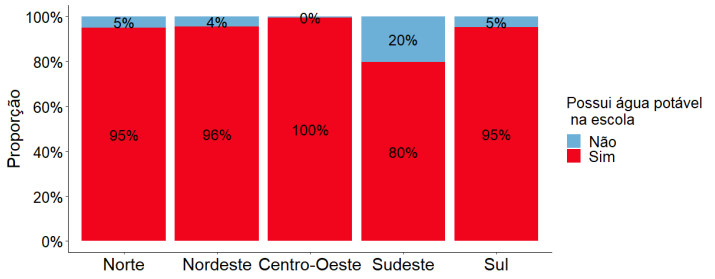


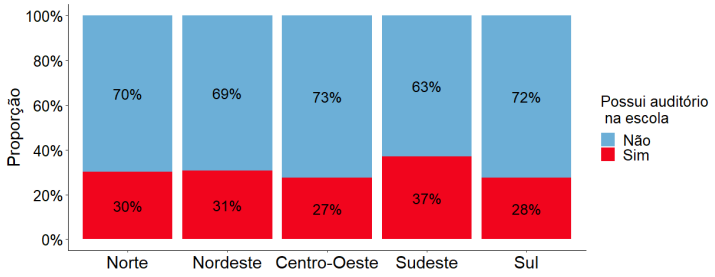
Figura 19: Proporção da dependência administrativa das escolas dos candidatos do ENEM de 2019 por região do Brasil.

A Figura 24 aponta a diferença regional das características de infraestrutura da escola. Através da análise de cinco características das escolas pode-se perceber que em praticamente todas elas apresentam diferenças regionais. A região Sudeste apresentou melhor desenvolvimento escolar das seguintes características: presença de auditório (37%), presença de laboratório de ciência (71%), presença de internet na escola (100%). Enquanto as escolas da região Centro-Oeste possuem uma melhor infraestrutura relacionada à presença de água potável nas escolas (100%) e possui um bom desenvolvimento ligado a presença de internet na escola também (90%).

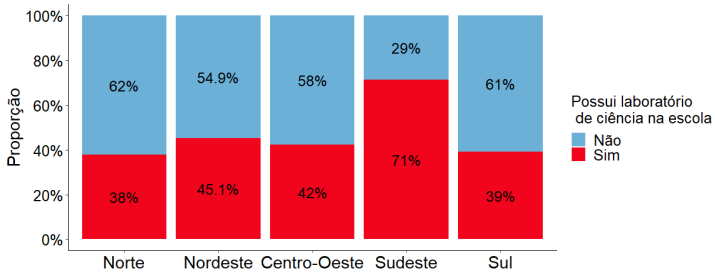
Por fim, percebe-se que a o desenvolvimento das escolas nas regiões varia conforme a característica avaliada. No quesito de possuir piscina na escola, por exemplo, a região que possui maior porcentagem é o Sul, com 7%.



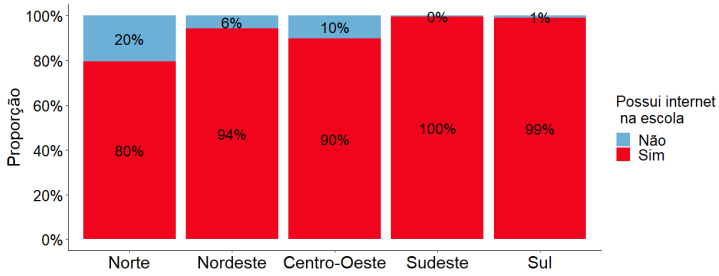
(a) Pela presença de água potável



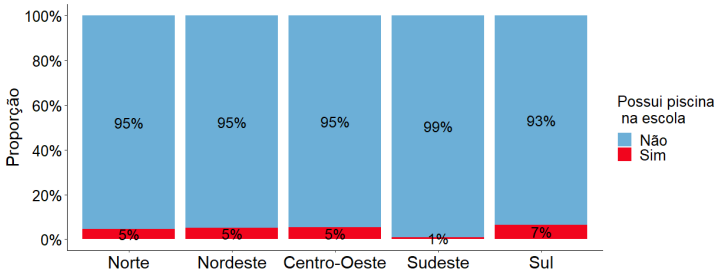
(b) Pela presença de auditório



(c) Pela presença de laboratório de ciência



(d) Pela presença de internet



(e) Pela presença de piscina

Figura 20: Proporção de características de infraestrutura da escola por região do Brasil.

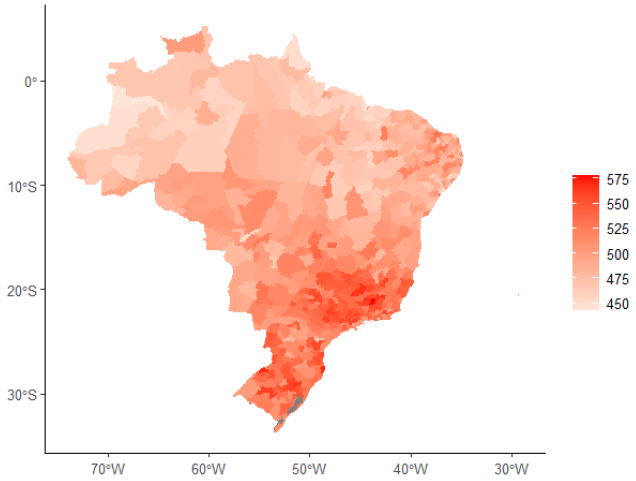
3.2.3 Análises das escolas segundo microrregiões

As análises no nível de microrregião é mais desagregada, e por isto traz informações mais detalhistas.

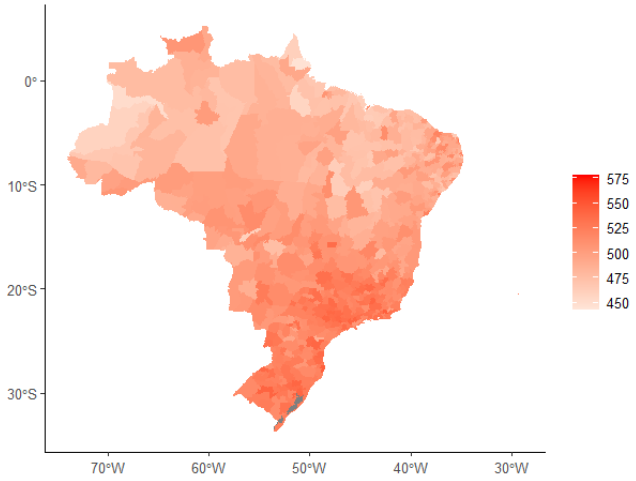
Vale ressaltar que durante as análises verificou-se que 2 microrregiões não possuem nenhuma escola de candidatos do ENEM de 2019, são elas: Lagoa Mirim (RS) e Lagoa dos Patos (RS).

3.2.3.1 Características dos candidatos

A Figura 21 apresenta o comportamento da nota de matemática e português das escolas por microrregião do Brasil. Pode-se perceber que o desenvolvimento dos alunos é um pouco melhor nas escolas das microrregiões pertencentes as regiões Sul e Sudeste. No Centro-Oeste as notas das escolas por microrregiões são distribuídas entre 450 a 550. Enquanto isso, percebemos que as escolas das microrregiões do Norte e Nordeste foram as que obtiveram menores notas na prova de matemática e português do ENEM de 2019.



(a) Notas na prova de matemática



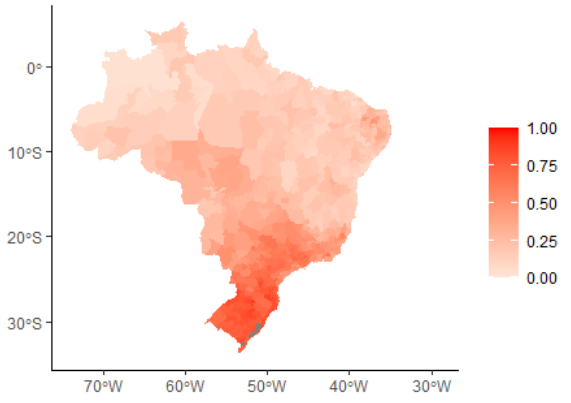
(b) Notas na prova de português

Figura 21: Comportamento da nota média do ENEM em matemática e português por microrregião do Brasil.

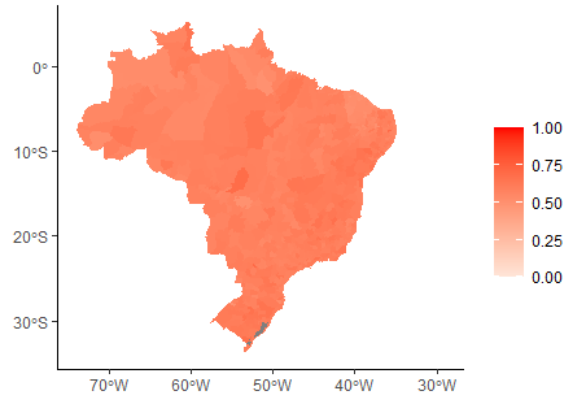
A Figura 22 retrata a proporção de determinadas características físicas dos candidatos considerando a microrregião das escolas. É visível que as escolas das microrregiões do Norte e Nordeste são as que possuem menor quantidade de candidatos brancos, enquanto as microrregiões do Sul possuem uma proporção maior. As escolas das microrregiões do Centro-Oeste ficam no meio termo, variando em torno de 20 a 60 porcento.

Em relação a proporção de candidatos que são do gênero feminino, podemos ver que a microrregião da escola não é um fator que influencia muito, já que o comportamento desta variável é bem parecido dentro todas as elas. Outra variável que segue este comportamento é a a proporção de candidatos maiores de 18 anos nas escolas. Apesar de estar também está bem distribuída pelas escolas das microrregiões do Brasil, percebe-se uma proporção

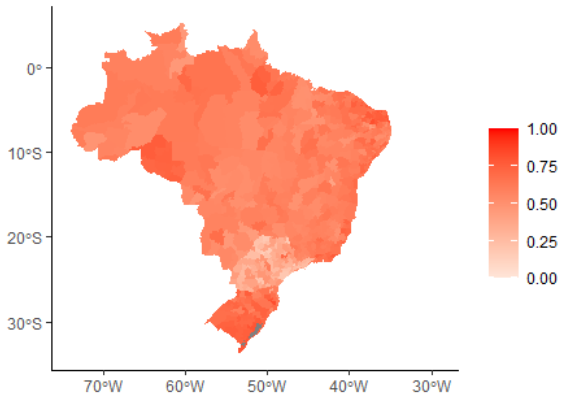
um pouco menor em algumas escolas pertencentes às microrregiões entre o Sul e o Sudeste.



(a) Cor branca



(b) Gênero Feminino



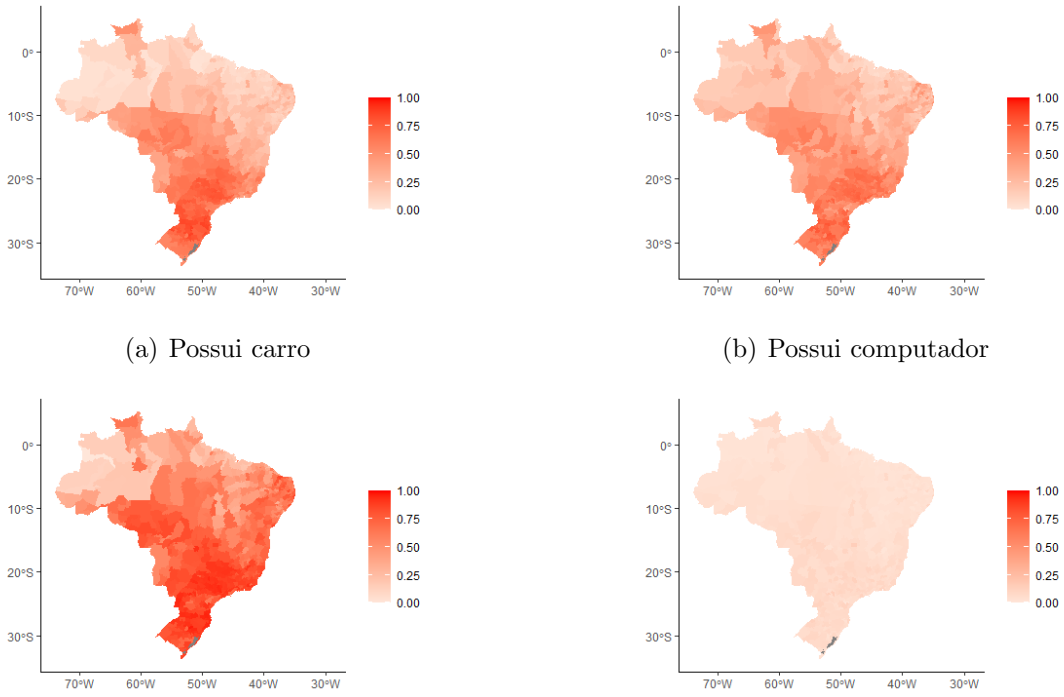
(c) Maior de idade

Figura 22: Proporção de características físicas dos candidatos do ENEM de 2019 por microrregião.

A Figura 23 indica a proporção de alguns bens que os candidatos do ENEM de 2019 possuem por microrregião das escolas do Brasil. Em todas as seguintes análises as escolas das microrregiões da região Sul e Sudeste demostraram uma maior proporção destes bens,

com exceção da presença de empregado(a), que indicou ter proporção bem baixa dentre todas as microrregiões que possuem escolas de candidatos do ENEM 2019.

Além disso, a presença de internet na casa do candidato mostrou variar dentro das escolas das microrregiões pertencentes as regiões Norte e Nordeste. Dentre todas essas análises, de forma geral, essa é a variável que mostrou ser mais presente nos candidatos.



(a) Possui carro

(b) Possui computador

(c) Possui internet

(d) Possui empregado(a)

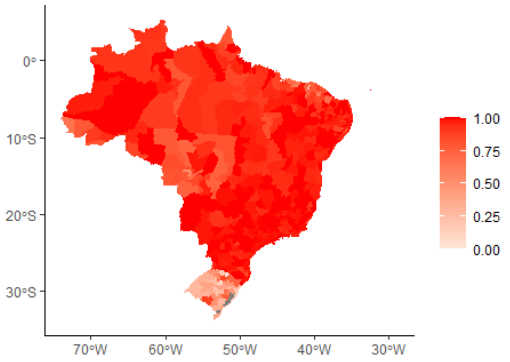
Figura 23: Proporção de características socieconômicas nas casas dos candidatos do ENEM de 2019 por microrregião.

3.2.3.2 Infraestrutura Escolar

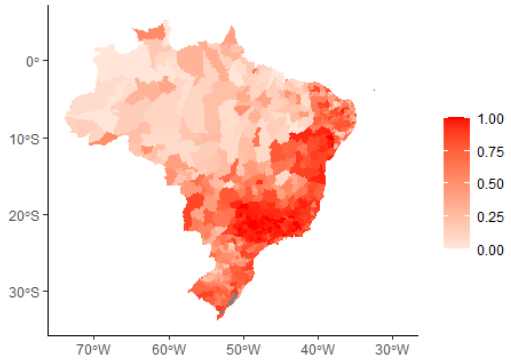
Por fim, foram realizadas as análises da infraestrutura escolar por microrregião das escolas do Brasil. Pode-se perceber que a grande maioria das escolas dos candidatos possuem aguá potável, internet, computador e cozinha, o que é um fator extremamente importante e positivo. Estas variáveis não apresentam muita distinção na variação da microrregião, com exceção da presença de internet na escola que foi visível uma menor proporção nas escolas das microrregiões do Norte.

Por outro lado, observa-se que a presença de auditório nas escolas é a mais baixa entre todas as análises em todas as microrregiões. A proporção da presença de laboratório de ciência é um pouco maior em algumas escolas das microrregiões do Sul e do Nordeste. Já

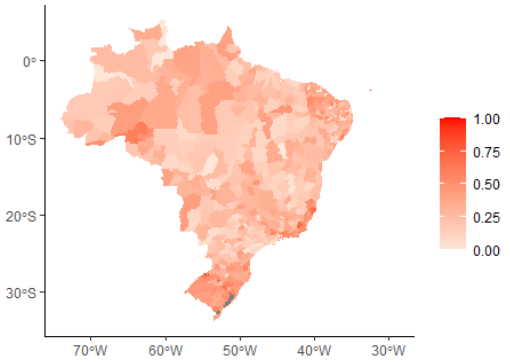
a presença de esgoto em rede pública apresentou uma maior proporção variada entre as escolas das microrregiões do Sul, Sudeste, Norte e Nordeste, com algumas contendo um valor maior do que todas as demais. As escolas das microrregiões pertences do Norte são a que menos possuem essa variável.



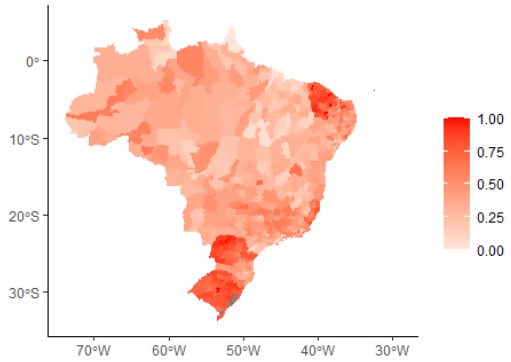
(a) Pela presença de água potável



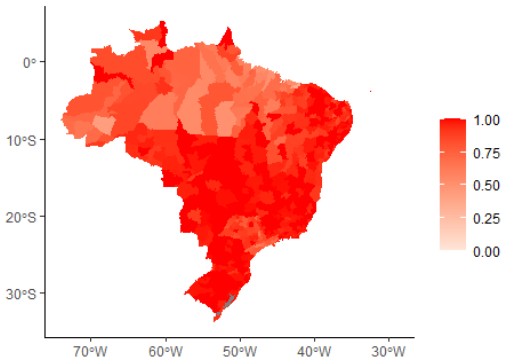
(b) Pela presença de esgoto em rede pública



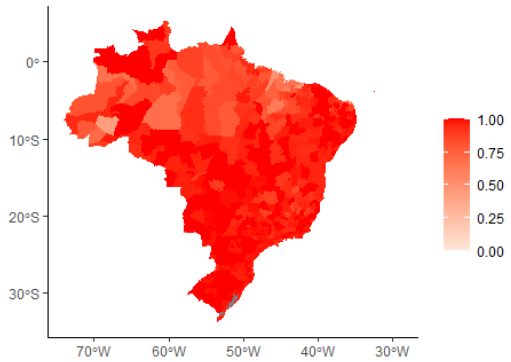
(c) Pela presença de auditório



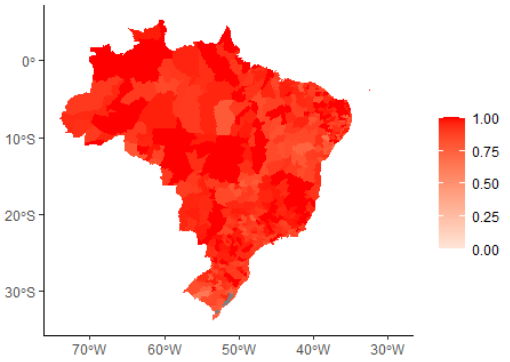
(d) Pela presença de laboratório de ciência



(e) Pela presença de internet



(f) Pela presença de computador



(g) Pela presença de cozinha

Figura 24: Proporção de características de infraestrutura da escola por região do Brasil.

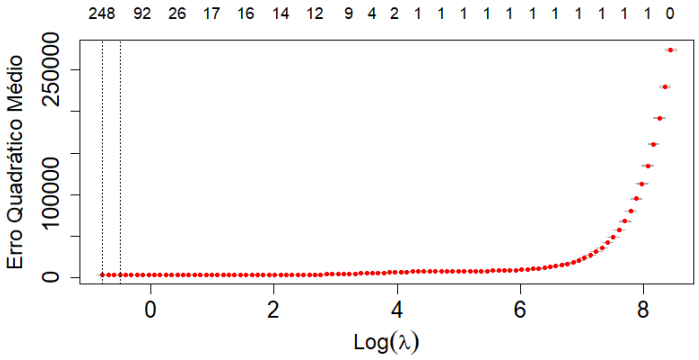
Após todas estas análises, conseguimos entender melhor como é o comportamento dos dados e o que acreditamos ser relevante para os próximos estudos do modelo.

3.3 Análise do Modelo

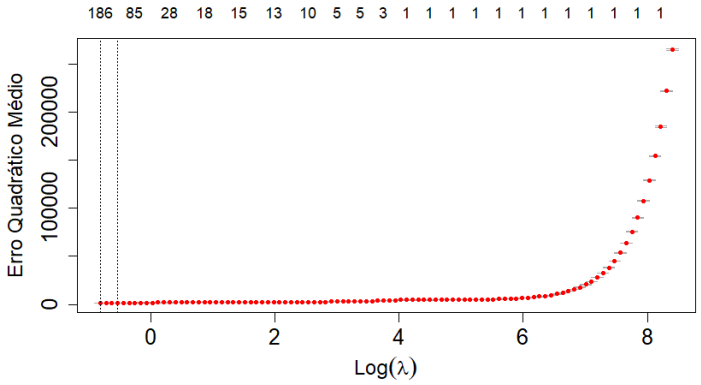
Nesta seção serão apresentados os resultados para o modelo de regressão linear normal, relacionando as notas de matemática e português do ENEM de 2019 das escolas com as demais variáveis da base de dados. Foram utilizadas 29.181 escolas nas análises da nota de matemática, enquanto para a nota de português foram utilizadas 29.298 escolas.

Na aplicação do LASSO foi usado o menor λ, pois é o valor que possui menor erro de predição e desta forma a penalização de $\boldsymbol{\beta}$ será a menor possível. Foram usadas vinte e duas covariáveis, apresentadas nas Tabelas 1 e 2, englobando características dos candidatos e de infraestrutura da escola, e microrregião. O LASSO selecionou as variáveis que são relevantes para o modelo e quais não influenciam.

A Figura 25 apresenta o comportamento da escolha do λ durante a validação cruzada para os dois modelos. Na parte superior da imagem indica a quantidade de variáveis no modelo, mostrando quantas são eliminadas conforme aumentamos o valor de λ. O λ escolhido é aquele que minimiza o erro quadrático médio, ou seja, que minimiza o erro de validação cruzada, que neste caso foi 0,446 para ambos os modelos. Na análise gráfica foi utilizada a escala Log para obter uma melhor ilustração, com isso foi realizada a converção de $\lambda = 0,446$ para $Log(\lambda) = -0,8$, que é o valor retratado nos gráficos.



(a) Modelo de matemática



(b) Modelo de português

Figura 25: Penalização adequada para os modelos.

3.3.1 Modelo de matemática

A Figura 26 indica os coeficientes estimados do modelo em estudo. Podemos perceber que após a aplicação do método de regularização LASSO, dois coeficientes de variáveis presentes no modelo foram zerados, as variáveis são: *quadra* (indica a presença de quadra na escola) e *prop_gestante* (proporção de candidatas gestantes na escola). Ou seja, foi identificado que essas variáveis não são relevantes para o entendimento do comportamento da nota de matemática do ENEM de 2019.

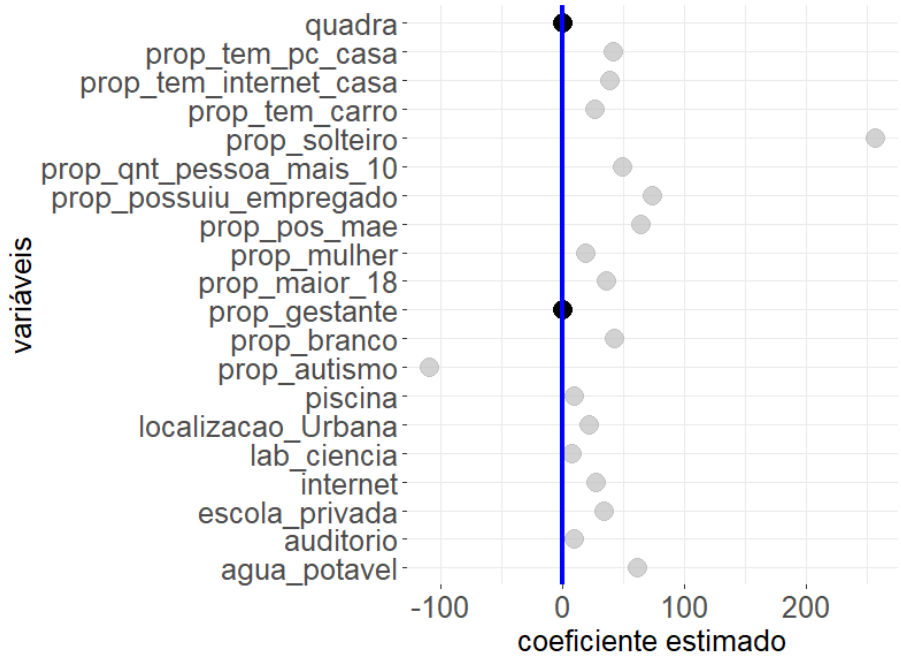


Figura 26: Coeficiente estimado para o Modelo da nota de Matemática.

Além disso, a variável *prop_autismo* demostrou um efeito negativo, ou seja, escolas com mais pessoas que possuem autismo apresentam nota menor. Então, por exemplo: A cada 1 ponto percentual de pessoas autistas na escola, espera-se que a nota media em matemática dessa escola reduza 110 pontos.

As demais variáveis obtiveram a estimação de seus coeficientes positivo, com a seguinte interpretação:

- A proporção de alunos brancos na escola apresentou efeito positivo. A cada 1 ponto percentual de pessoas brancas na escola, espera-se que a nota media de matemática aumente 42,2 pontos.
- A variável que indica a proporção de solteiro na escola foi a que mais se destacou, demonstrando um efeito positivo. A cada 1 ponto percentual de pessoas solteiras na escola, espera-se que a nota media de matemática aumente 256 pontos.
- Escolas com alunos que possuem computador e carro em casa têm, em média, nota de matemática média maior do que candidatos que não possuem esses bens, sendo 41,3 e 26,4 pontos a mais na média, respectivamente.
- Escolas com candidatos que possuem empregado(a) e moram com mais de 10 pessoas em casa demostraram ser relevantes de forma positiva na nota do ENEM, sendo os seguintes coeficientes estimados : 72,9 e 49, respectivamente.

- A proporção de escolas com alunos nos quais a mãe possui Pós-Graduação apresentou efeito positivo. Espera-se que a nota media de matemática aumente 64,1 pontos para cada 1 ponto percentual de escolas com alunos que se encaixam nesse requisito.

- Escolas com laboratório de ciências, internet, piscina, auditório e aguá potável têm, nota média de matemática maior do que escolas que não possuem essa infraestrutura.

- Escolas privadas possuem um efeito positivo, esperando-se que a nota média de matemática aumente 33,7 pontos a cada 1 ponto percentual de alunos que estudam escola privada. Além disso, escolas localizadas na área urbana também apresentam um melhor proficiência no ENEM.

A Figura 27 foi padronizada entre 0 e 1 para identificar o incremento na nota de matemática, sem considerar o sinal desse efeito. Após a aplicação do LASSO, 328 microrregiões demostraram não ser significantes para o comportamento da nota de matemática do ENEM de 2019 e obtiveram seus coeficientes de estimação zerados, consequentemente sendo removidas do modelo.

Pode-se perceber que as microrregiões do Centro-Oeste e Norte foram as que demonstrarem ser menos significativas na nota de matemática. Em contrapartida, uma boa parte das microrregiões do Sul, Sudeste e Nordeste apresentaram relevância. Ademais, a microrregião que apresentou ser mais significante é Baixo Curu, pertencente ao Ceará (Nordeste).

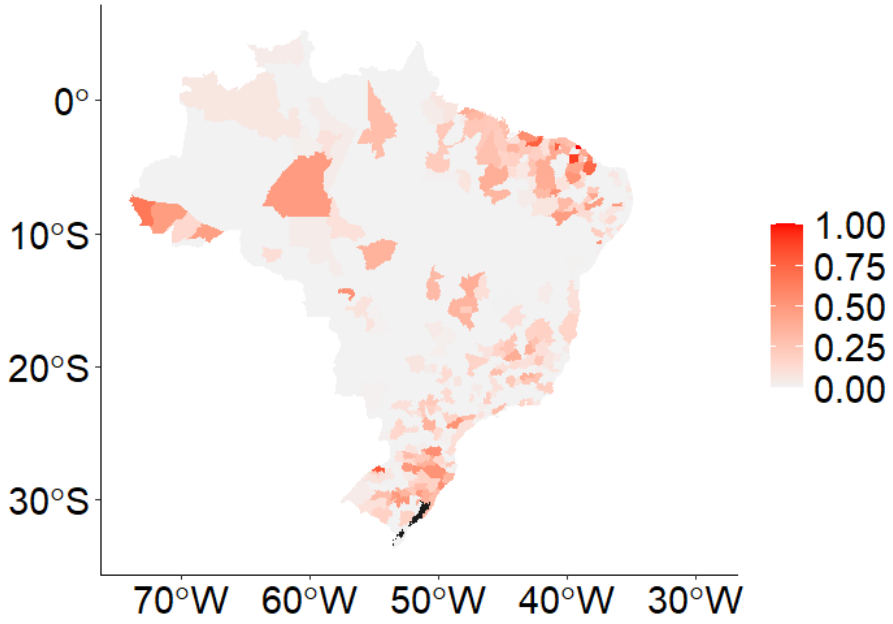


Figura 27: Incrementos na nota de matemática padronizados das variáveis de microrregiões do Brasil.

A Figura 28 apresenta o efeito das microrregiões na nota de matemática do ENEM de 2019, diferenciando efeitos positivos e negativos. Observa-se que 64 microrregiões obtiveram efeito negativo na estimação dos coeficientes, ou seja, apenas o fato da escola pertencer à estas determinadas microrregiões já impacta negativamente na nota do ENEM.

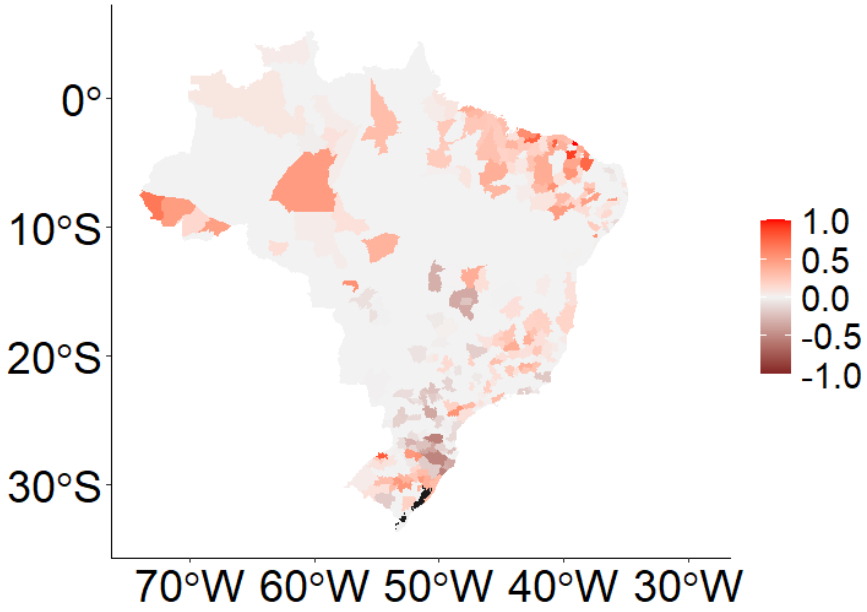


Figura 28: Efeitos das variáveis de microrregiões do Brasil.

Algumas análises foram realizadas a fim de entender o motivo desses efeitos negativos

nessas microrregiões, descritas na Tabela 5.

Observa-se pela Tabela 5 que em média a maioria das escolas das microrregiões que possuem efeito negativo possuem aguá potável (97%), são localizadas em áreas urbanas (91%) e possuem internet (98%). Essas variáveis demostraram ser influenciáveis positivamente na nota de matemática do ENEM de 2019, então como elas estão presentes na maioria das escolas das microrregiões que possuem efeito negativo não podemos dizer que elas são responsáveis por isto. Além disso, nota-se que em média apenas 24% das escolas são privadas, 4% possuem piscina, 33% possuem auditório, e ainda, apenas 10% são candidatos nos quais as mães possuem Pós-Graduação. Levando em consideração que essas variáveis possuem efeito positivo sob a nota de matemática no ENEM 2019, a proporção baixa delas não é um ponto positivo. Por fim, verificou-se que a em média existem 88 alunos nestas escolas.

Tabela 5: Proporção média das características dos candidatos e das escolas das microrregiões com efeitos negativos.

Variável	Proporção média das microrregiões
escola_privada	0,24
auditorio	0,33
agua_potavel	0,97
localizacao_Urbano	0,91
lab_ciencia	0,56
piscina	0,04
quadra	0,79
internet	0,98
prop_mulher	0,60
prop_solteiro	0,94
prop_branco	0,57
prop_maior_18	0,55
prop_pos_mae	0,10
prop_tem_carro	0,66
prop_tem_pc_casa	0,60
prop_possui_empregado	0,08
prop_qnt_pessoa_mais_10	0,62

As análises indicaram também que estas 64 microrregiões que possuem efeito negativo estão distribuídas pelas cinco grandes regiões do Brasil. Contudo, o Sul foi o que possuiu maior frequência de microrregiões diferentes, tendo destaque para duas Unidades Federativas: Paraná, com dezessete microrregiões distintas com efeitos negativos e Santa Catarina com quinze.

3.3.2 Modelo de português

A Figura 29 indica os coeficientes estimados do modelo de português. Após a aplicação do método de regularização LASSO, percebe-se que nenhum dos coeficientes de variáveis relacionados as características dos candidatos e/ou da escola foram zerados, ou seja, todos mostraram ser influenciáveis para a nota de português do ENEM de 2019.

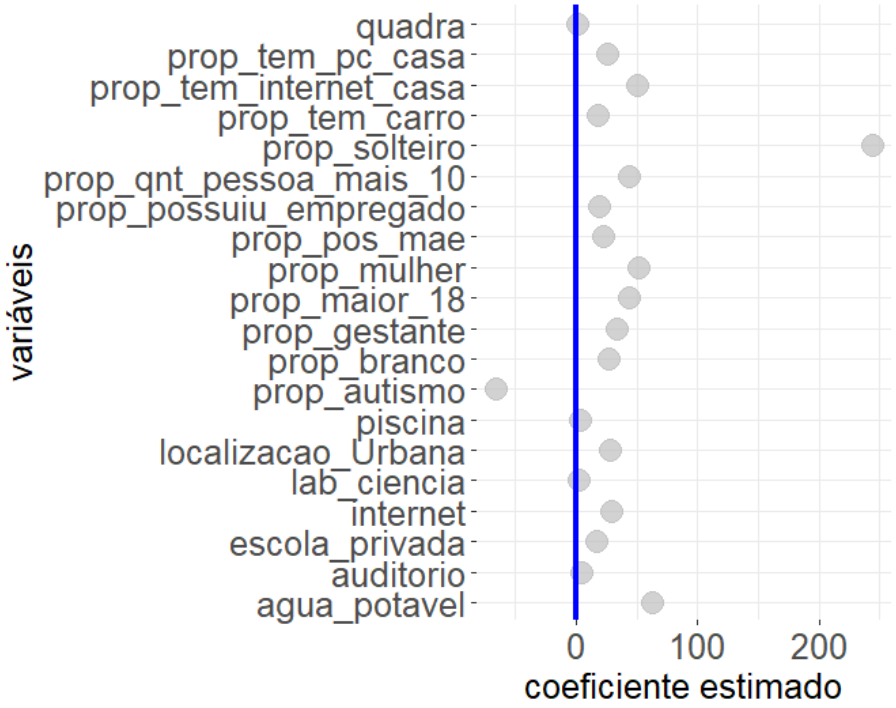


Figura 29: Coeficientes estimados das variáveis de características de alunos e infraestrutura para o Modelo da nota de Português.

Assim como no modelo anterior, a variável *prop_autismo* demostrou um efeito negativo, ou seja, a cada 1 ponto percentual de pessoas autistas na escola, espera-se que a nota média em português dessa escola reduza 66,3 pontos, sendo um impacto um pouco menor do que no modelo da nota de matemática.

As demais variáveis obtiveram a estimação de seus coeficientes positivo, com a seguinte interpretação:

- Apesar de não ter eliminado nenhuma variável do modelo, quatro demonstraram ser pouco relevante, tendo seus coeficientes estimados bem próximos de zero. São elas: *quadra* (1,39), *lab_ciencia* (2,91), *piscina* (3,52) e *auditorio*(4,54).

- Conforme esperado, a proporção de escolas com alunos brancos apresentou efeito positivo. A cada 1 ponto percentual de pessoa branca na escola, espera-se que a

nota média de português aumente 27,1 pontos, tendo um impacto um pouco menor do que na nota de matemática.

- Assim como no modelo de matemática, a variável que indica a proporção de solteiro na escola foi a que teve o maior efeito positivo. A cada 1 ponto percentual de aluno solteiro na escola, espera-se que a nota media de português aumente 244 pontos.

- Escolas com alunos que possuem computador, internet e carro em casa têm, em média, nota de matemática média maior do que candidatos que não possuem esses bens, sendo 26,2 e 17,5 pontos a mais na média, respectivamente.

- Candidatos que possuem empregado(a) e moram com mais de 10 pessoas em casa demostraram serem influenciáveis de forma positiva na nota do ENEM, sendo os seguintes coeficientes estimados : 18,9 e 43,7, respectivamente.

- A proporção de alunos nas escolas nos quais a mãe possui Pós-Graduação apresentou efeito positivo. Espera-se que a nota média de português aumente 22,2 pontos para cada 1 ponto percentual de alunos que se encaixam nesse requisito.

- Escolas com internet e aguá potável têm, em média, nota de português média maior do que escolas que não possuem essa infraestrutura, aumentando em média a nota em 28,7 e 62,7 pontos, respectivamente. Além disso, os alunos de escolas privadas e localizadas na área urbana também apresentam um melhor proficiência no ENEM.

A Figura 30 foi padronizada entre 0 e 1 para identificar o incremento na nota de português, sem considerar o sinal desse efeito. Após a aplicação do LASSO, 392 microrregiões tiveram seus coeficientes de estimação zerados, e consequentemente foram removidas do modelo. Ou seja, essas microrregiões não demostraram ser significantes para o comportamento da nota de português do ENEM de 2019.

Percebe-se que a maioria das microrregiões do Centro-Oeste e Sudeste não apresentaram nenhum efeito sob a nota de português. Em contrapartida, as regiões do Norte foram as que demonstraram melhor efeito, sendo a microrregião Cruzeiro do Sul (AC) a que possui melhor efeito espacial dentre todas.

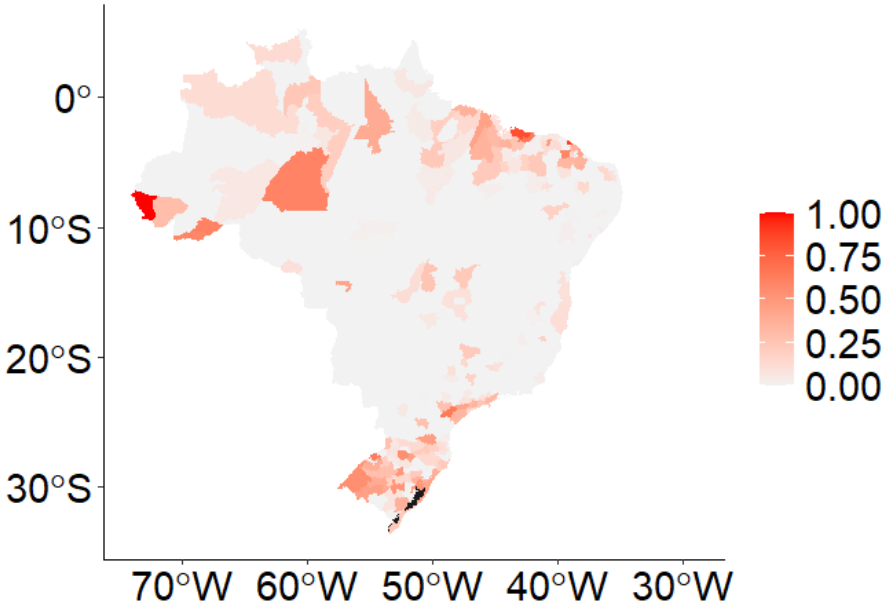


Figura 30: Incrementos na nota de português padronizados das variáveis de microrregiões do Brasil.

A Figura 31 apresenta o efeito das microrregiões na nota de matemática do ENEM de 2019, diferenciando efeitos positivos e negativos. Observa-se que 32 microrregiões obtiveram efeito negativo na estimação dos coeficientes, ou seja, apenas o fato da escola pertencer à estas determinadas microrregiões já impacta negativamente na nota do ENEM.

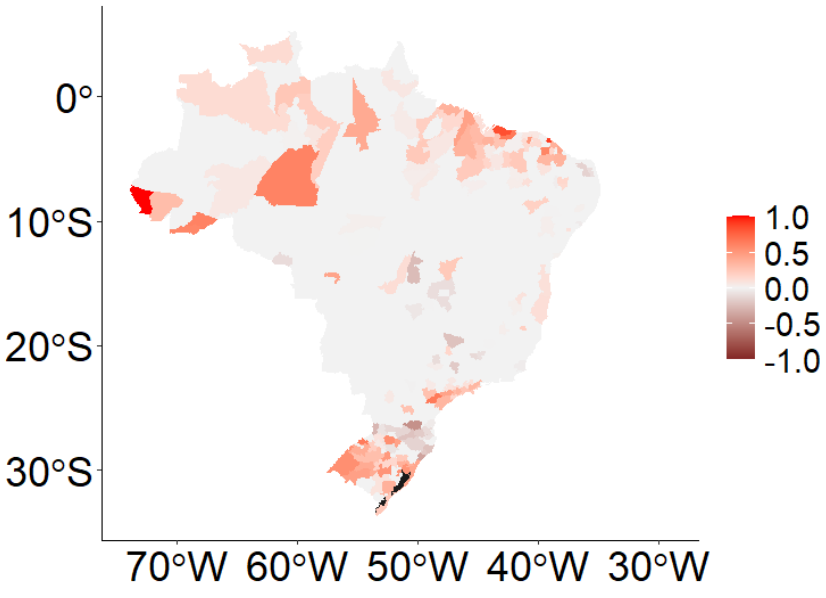


Figura 31: Efeitos das variáveis de microrregiões do Brasil.

Algumas análises foram realizadas a fim de entender o motivo desses efeitos negativos nessas microrregiões.

Observa-se pela Tabela 6 que o comportamento das variáveis foi bem similar ao do modelo de matemática. Em média a maioria das escolas das microrregiões que possuem efeito negativo possuem aguá potável (99%), são localizadas em áreas urbanas (91%) e possuem internet (98%), o que é um ponto positivo, visto que estas variáveis demostraram influenciar positivamente na nota de matemática do ENEM de 2019.

Por outro viés, nota-se que em média apenas 19% das escolas são privadas, 2% possuem piscina, 35% possuem auditório e somente 10% são candidatos nos quais as mães possuem Pós-Graduação. Sabendo que essas variáveis possuem efeito positivo sob a nota de matemática no ENEM 2019, a proporção baixa delas não é um ponto positivo. Por fim, verificou-se que a média geral de alunos dessas escolas é 45, uma quantidade não muito alta e inferior à verificada no modelo de matemática.

Tabela 6: Proporção média das características dos candidatos e das escolas das microrregiões com efeitos negativos.

Variável	Proporção média das microrregiões
escola_privada	0,19
auditorio	0,35
agua_potavel	0,99
localizacao_Urbana	0,91
lab_ciencia	0,47
piscina	0,02
quadra	0,72
internet	0,98
prop_mulher	0,60
prop_solteiro	0,95
prop_branco	0,61
prop_maior_18	0,64
prop_pos_mae	0,10
prop_tem_carro	0,65
prop_tem_pc_casa	0,60
prop_possui_empregado	0,07
prop_qnt_pessoa_mais_10	0,62

Por fim, as análises indicaram que essas microrregiões com efeito negativo estão distribuídas pelas cinco grandes regiões do Brasil, sendo o SUL o que possui maior quantidade de microrregiões diferentes. Além disso, tem como destaque principal a Unidade Federativa de Santa Catarina que obteve quinze microrregiões distintas com efeitos negativos, que é o mesmo valor encontrado no modelo de matemática.

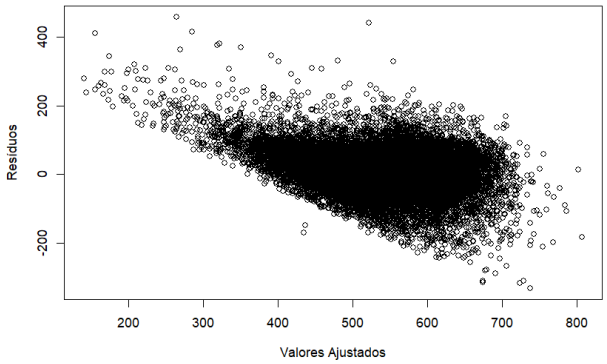
3.3.3 Análise dos Resíduos

Foi realizada a análise dos resíduos dos modelos para investigar se os pressupostos descritos na seção 2.2.1 são satisfeitos. A Figura 32 apresenta as análises para o modelo de matemática, enquanto a Figura 33 mostra as análises do modelo de português. A interpretação de ambos os modelos foi similar.

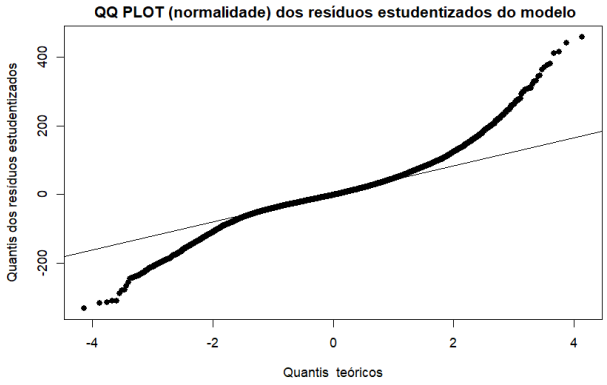
Observa-se pelas Figuras 32 e 33 que os resíduos dos modelos estimados possuem uma leve assimetria. Após realizar o Teste de Breusch–Pagan os pressupostos foram rejeitados, ou seja, os resíduos não possuem variância σ^2 constante. Além disso, pela análise gráfica há indicativos de que os modelos acabam subestimando algumas notas pra notas mais baixas, visto nos gráficos dos Resíduos versus valores ajustados, Figuras 32(a) e 33(a).

As Figuras 32(c) e 33(c) apresentam os histograma dos resíduos para o modelo de matemática e português, respectivamente. Apesar de aparentemente possuir um comportamento normal, após a realização do Teste de Lilliefors a hipótese dos dados seguirem a distribuição normal foi rejeitada para ambos os modelos (P-valor $< \alpha = 0,05$). A presença de caudas pesadas provavelmente faz com que a distribuição normal não consiga se adequar bem à esses modelos.

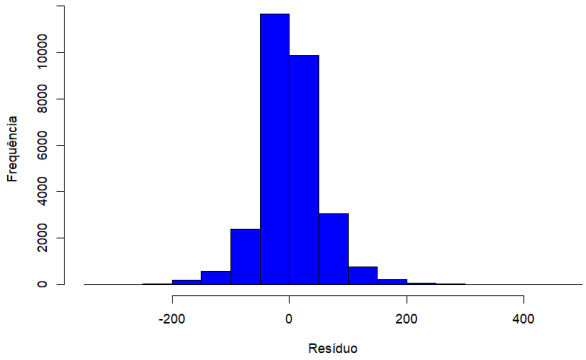
Logo, o ideal seria utilizar outros modelos que consigam acomodar melhor essas características dos dados. Por exemplo, modelos que não tenham pressupostos de variância constante e modelos com outras distribuições com caldas mais pesadas ou que consigam capturar algum tipo de assimetria nos dados.



(a) Ajuste do modelo

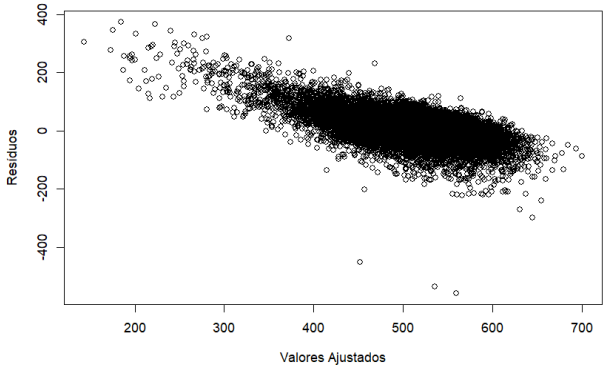


(b) Normalidade do Resíduo

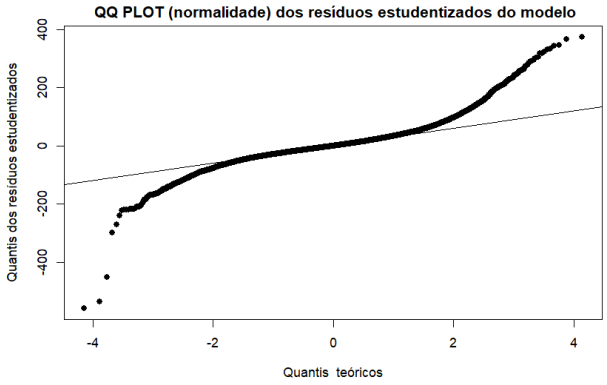


(c) Comportamento do Resíduo

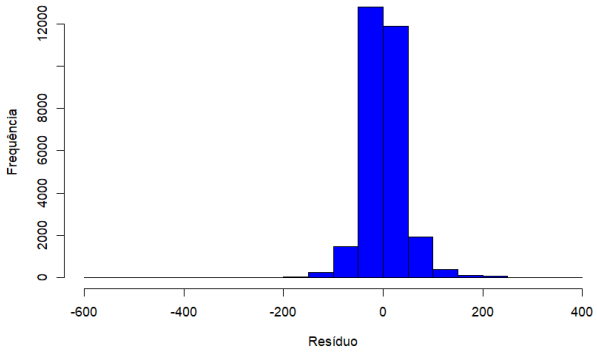
Figura 32: Análise dos Resíduos para o Modelo da nota de matemática.



(a) Ajuste do modelo



(b) Normalidade do Resíduo



(c) Comportamento do Resíduo

Figura 33: Análise dos Resíduos para o Modelo da nota de português.

4 Conclusão

Esse trabalho teve como objetivo principal avaliar a proficiência dass alunos a partir das notas médias de matemática e português no ENEM de 2019 do Brasil por escola. Através de métodos de estimação da inferência clássica e penalização por LASSO, foi viável selecionar as variáveis de maior relevância, bem como os efeitos fixos significativos relacionados às microrregiões.

A partir da análise descritiva ao nível Brasil, em média o rendimento dos candidatos na prova de português do ENEM de 2019 foi um pouco melhor do que na de matemática, sendo a maioria dos candidatos alunos entre 15 e 20 anos. Verificou-se que os estados que possuem mais candidatos no ENEM de 2019 são: São Paulo (SP), seguido de Minas Gerais (MG), e depois BA (Bahia). Na perspectiva da infraestrutura escolar, observou-se que a maioria das escolas são públicas (72%) e localizadas na área urbana (91%).

Já a análise descritiva no nível das grandes regiões do Brasil indicou que o Centro-Oeste e o Sudeste são as regiões que apresentam melhor rendimento dos candidatos nas provas, sendo a nota média de matemática um pouco superior. Ademais, reparou-se que o Sudeste é a região que mais possui candidatos de cor branca (72,5%) e que o Centro-Oeste é a região que menos possui candidatos maior de idade, com apenas 46,4%, diferentemente das demais regiões que possuem mais da metade. Além do mais, o Sul indicou ser a região que os candidatos possuem mais bens materiais de caráter socioeconômico como carro e computador, enquanto o Norte e Nordeste são as regiões mais prejudicadas neste quesito. Avaliando a infraestrutura escolar percebe-se que a localização da escola é influente. A região Sudeste apresentou melhor desenvolvimento escolar das seguintes características: presença de auditório, presença de laboratório de ciência, presença de internet na escola e presença de computador. Enquanto a região Centro-Oeste possui uma melhor infraestrutura relacionada à presença de água potável nas escolas e possui um bom desenvolvimento ligado a presença de internet na escola também.

As análises no nível das microrregiões do Brasil indicaram que a presença de empre-

gado(a) nas casas dos candidatos tem uma proporção bem baixa dentre todas as microrregiões do Brasil. Além disso, a análise da infraestrutura escolar indicou que as seguintes variáveis: água potável, internet, computador e cozinha na escola, não apresentam muita distinção entre as microrregiões do Brasil e positivamente estão presentes na grande maioria das escolas. Por outro lado, observa-se que a presença de auditório nas escolas é a mais baixa entre todas as variáveis analisadas por microrregiões. A proporção da presença de laboratório de ciência é um pouco maior em algumas microrregiões do Sul e do Nordeste.

A partir do modelo de regressão linear para descrever as notas de matemática do ENEM de 2019, após a aplicação do método de regularização LASSO, no modelo da nota de matemática duas variáveis demostraram não ser relevantes para o entendimento do comportamento da nota, são elas: *quadra* e *prop_gestante*; enquanto no modelo de português nenhum coeficiente estimado foi zerado. Além disso, a variável *prop_autismo* demostrou um efeito negativo em ambos os modelos, ou seja, escolas com mais pessoas que possuem autismo apresentam em média uma nota menor. Além disto, todas as seguintes variáveis indicaram efeito positivo na nota de matemática e português: *prop_branco, prop_tem_pc_casa, prop_tem_carro, prop_qnt_pessoa_mais_10, prop_possui_empregado, piscina, prop_maior_18, prop_pos_mae, localizacao_Urbana, lab_ciencia, internet, auditorio, escola_privada, agua_potavel* e com enfase em *prop_solteiro*, que foi a variável que apresentou o maior efeito positivo.

Ainda no modelo de regressão linear, a aplicação do LASSO nas variáveis de microrregiões que representam efeitos fixos indicou para o modelo de matemática que 295 microrregiões não são significantes para o entendimento do comportamento da nota, pois tiveram seus coeficientes estimados zerados. Por outro lado, sessenta e quatro microrregiões apresentaram efeito negativo, ou seja, apenas o fato da escola pertencer à estas determinadas microrregiões já impacta negativamente na nota do ENEM. O Sul foi a região que possui maior frequência de microrregiões diferentes com efeito negativo, tendo destaque para duas Unidades Federativas: Paraná, com dezessete microrregiões distintas com efeitos negativos e Santa Catarina com quinze. Já para o modelo de português, mais microrregiões tiveram seus coeficientes estimados zerados, totalizando em 360 e 32 microrregiões com efeito negativo. O Sul foi o que possuiu maior quantidade de microrregiões diferentes, tendo como destaque principal a UF de Santa Catarina que obteve quinze microrregiões distintas com efeitos negativos.

Logo, após todas as análises é possível perceber que a infraestrutura escolar e algumas características socieconômicas dos candidatos influenciam em seu desempenho na prova de

matemática e português do ENEM 2019, sendo coerente com os resultados encontrados em estudos anteriores. Além disto, as microrregiões que as escolas pertencem também demostraram relevantes em alguns casos. Com isto, o atual trabalho trouxe uma boa perspectiva de como é o comportamento das notas de português e matemática do ENEM 2019 ao longo das grandes regiões do Brasil e microrregiões, como também as variáveis que são relevantes no modelo.

Para trabalhos futuros pode ser interessante aprofundar mais nas microrregiões que possuem efeito negativo nas notas de matemática e português do ENEM, tentando entender quais variáveis podem influenciar nisto, realizando um estudo mais detalhista deste aspecto. Além disso, tentar utilizar outros modelos que consigam acomodar as características dos dados de forma melhor.

Referências

ALBERNAZ Ângela; FERREIRA, F. H. G.; FRANCO, C. Qualidade e equidade no ensino fundamental brasileiro. Instituto de Pesquisa Econômica Aplicada (Ipea), v. 6, p. 453–476, 2002.

BRANDÃO, C. R.; FAGUNDES, M. C. V. Cultura popular e educação popular: expressões da proposta freireana para um sistema de educação. 2016.

BRASIL. *Microdados do Enem 2019*. [S.l.]: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), 2020. ⟨https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem⟩.

BRASIL. *Malha Municipal*. [S.l.]: Instituto Brasileiro de Geografia e Estatística(IBGE), 2021. ⟨https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-malhas.html?edicao=27733&t=acesso-ao-produto⟩.

CONCEIçãO, P. Relatório do desenvolvimento humano 2019. Programa das Nações Unidas para o Desenvolvimento (PNUD), 2020.

FREITAS, E. de. *Educação, base do desenvolvimento*. 2022. ⟨https://brasilescola.uol.com.br/geografia/educacao-base-desenvolvimento.htm⟩.

GAUSS, C. F. 1870. 1-93 p. ⟨https://archive.org/stream/werkecarlf04gausrich#⟩.

GOMES, S.; MELO, F. Y. M. de. Por uma abordagem espacial na gestão de políticas educacionais: equidade para superar desigualdade. Educação Sociedade, 2021.

JALOTO, A.; PRIMI, R. Fatores socioeconômicos associados ao desempenho no enem. Em aberto, v. 34, p. 125–141, 2021.

KUTNER, M. H. et al. *Applied Linear Statistical Models*. [S.l.]: McGraw-Hill/Irwin, 2005. v. 5.

LILLIEFORS, H. W. On the kolmogorov-smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association, v. 32, p. 399–402, 1967.

MELO, R. O. et al. Impacto das variáveis socioeconômicas no desempenho do enem: uma análise espacial e sociológica. Revista de Administracao Publicar, 2021.

MORAES, C. P. de; PERES, R. T. Reflexões sobre diferenças de desempenho no enem: Uma análise socioeconômica e escolar do sudeste do brasil. Jornal de Políticas Educacionais, v. 16, p. 61–78, 2022.

OLIVEIRA, Y. C. de. *Modelando a relação entre desempenho escolar, infraestrutura e outros indicadores em Educação*. 1-126 f. Monografia (Graduação) — Universidade Federal Fluminense, Rio de Janeiro, 2022.

SOARES, D. J. M.; SOARES, T. E. A.; SANTOS, W. dos. Infraestrutura e desempenho escolar na prova brasil: aspectos e conexões. Olhar de Professor, v. 23, p. 1–18, 2020.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, v. 58, p. 267–288, 1996.

TORRES, R. et al. Determinantes do desempenho dos participantes da prova do enem: Evidências para o rio grande do sul. Desenvolvimento Em Questão, v. 18, p. 352–368, 2020.