Análise dos Microdados dos Desempenhos dos Estudantes das Escolas de São Paulo no ENEM 2015

Maria Nilza de Sousa Ramos
maria.nilza.s.ramos@gmail.com
Modelos de Regressão I

1 Introdução

Este trabalho analisa os fatores que influenciam a nota média de uma escola no ENEM 2015. Utilizamos os microdados por escola do exame e do Censo Escolar de 2015, aplicando um modelo de regressão linear múltipla com erros normais e variância constante. O objetivo é propor um modelo para explicar o desempenho escolar nas escolas de São Paulo, focando na variável "nota média do ENEM"(nota.media.enem).

2 Análise Descritiva

O ENEM é um exame realizado anualmente pelo MEC sob organização do INEP. Ele foi criado em 1998 com o objetivo central de avaliar o desempenho escolar dos estudantes ao término da educação básica. Porém, foi apenas em 2009 que o INEP passou a usar o exame como instrumento de avaliação para ingresso dos estudantes na educação superior (Inep, 2022).

Por este motivo, a variável que classifica a complexidade de gestão foi o filtro para prosseguir com o tratamento e elaboração da base para regressão. Essa variável pode ser classificada em cinco níveis diferentes, estas dizem respeito e são classificadas conforme o nível mais alto de ensino da escola. Deste modo, fez-se uma comparação das médias das notas entre inscritos que estudavam em escolas que não ofertavam ensino fundamental II, médio, e as que ofertavam ensino médio ou EJA.

Curiosamente, foi constatado que as notas médias eram maiores para os inscritos que sem dúvida não haviam cursado o ensino médio. Pode ser melhor visualizado nas figuras de 1 a 3 a seguir:

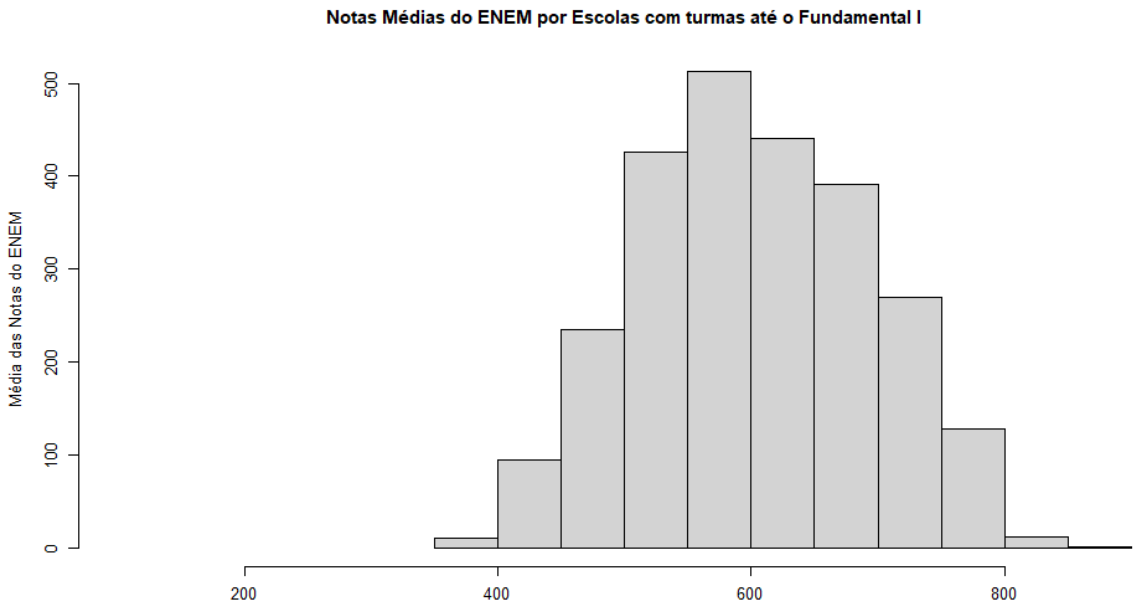

Figura 1: Notas Médias por Complexidade de Gestão até o Nível Fundamental I

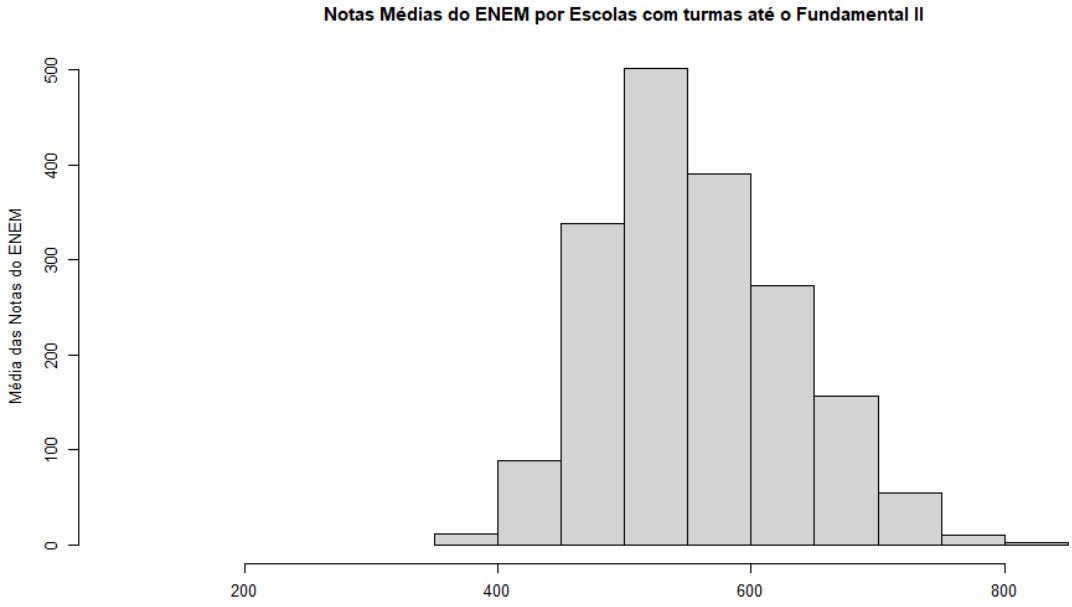

Figura 2: Notas Médias por Complexidade de Gestão até o Nível Fundamental II

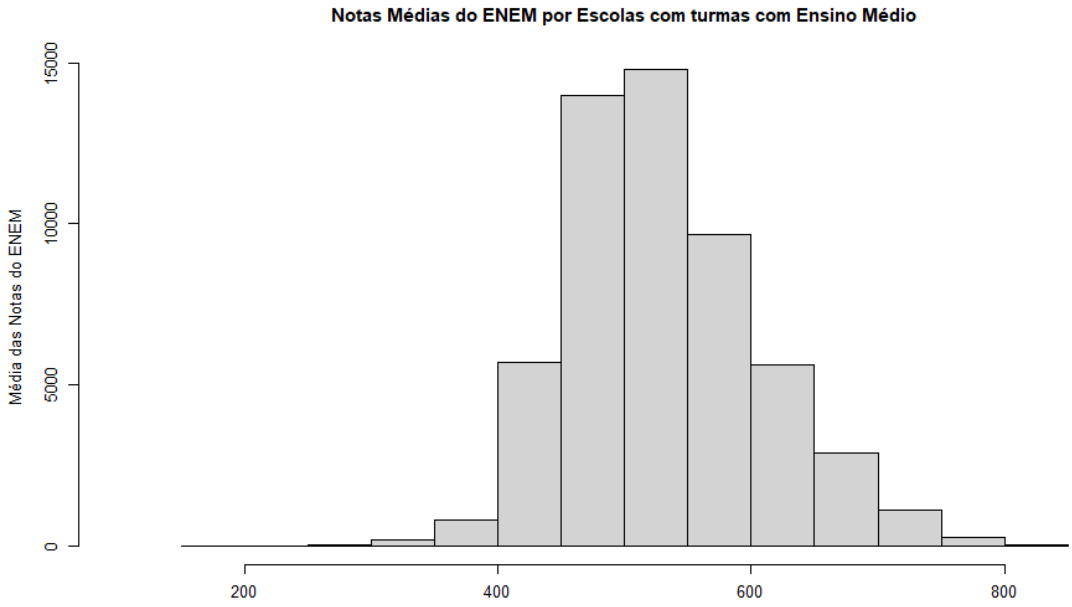

Figura 3: Notas Médias por Complexidade de Gestão até o Ensino Médio ou EJA

Conseguinte a isto, fora investigado outras variáveis por nível de complexidade de gestão educacional.

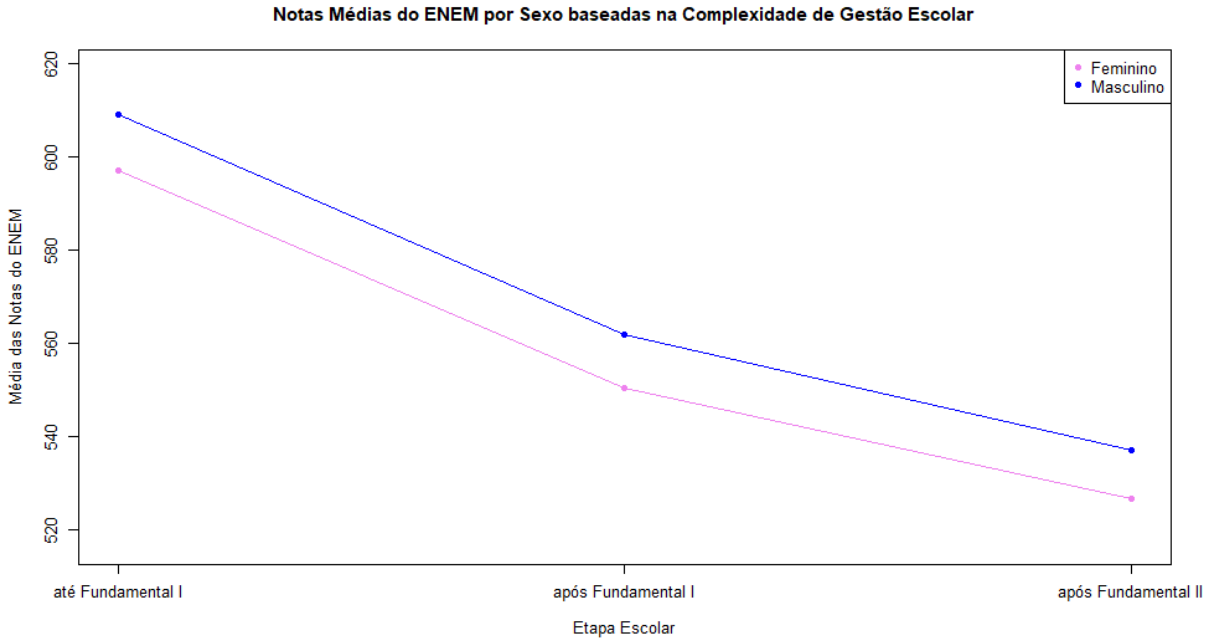

Figura 4: Notas Médias por Complexidade de Gestão até o Nível Fundamental I

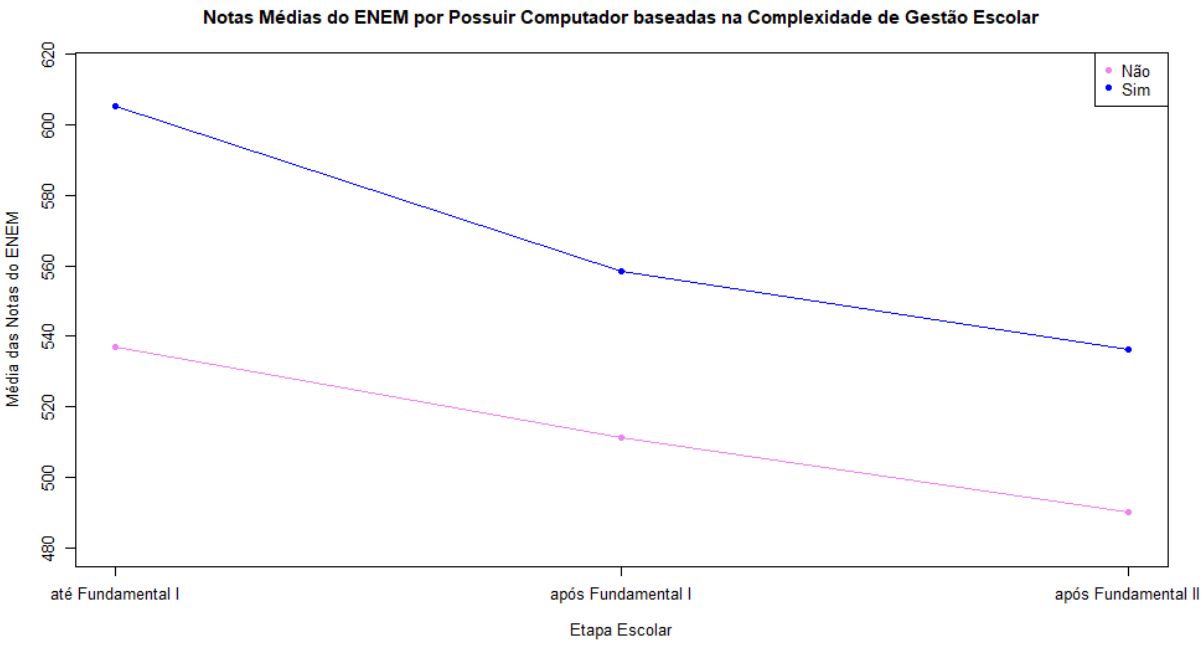

Figura 5: Notas Médias por Complexidade de Gestão até o Nível Fundamental II

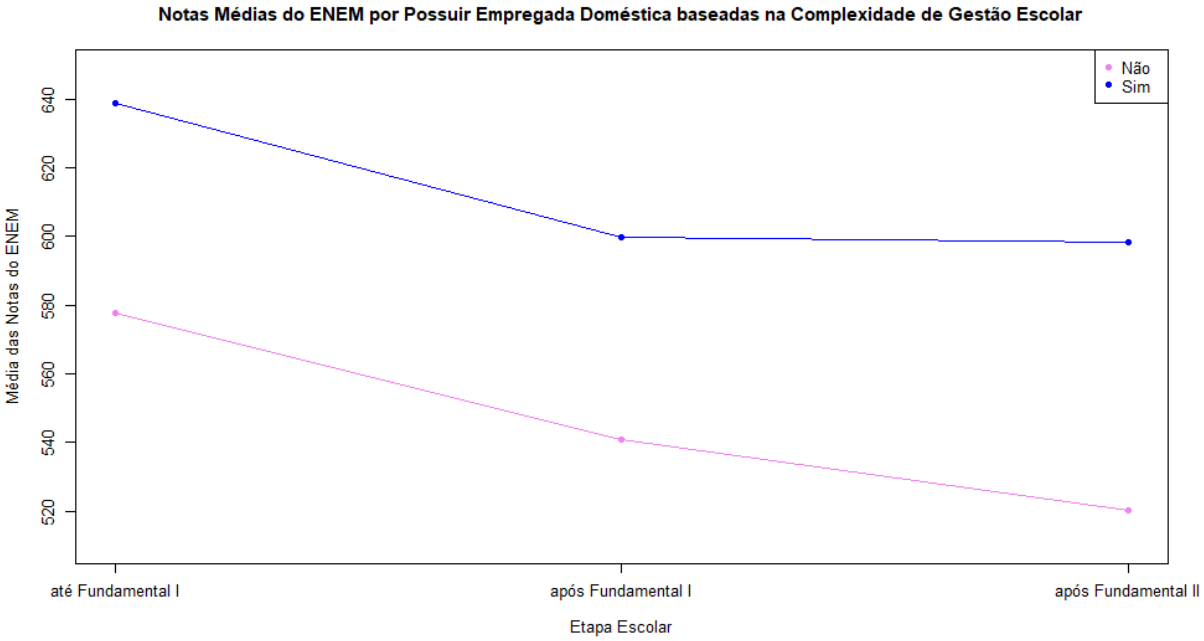

Figura 6: Notas Médias por Complexidade de Gestão até o Ensino Médio ou EJA

As notas médias do ENEM diminuem com o aumento da complexidade escolar. A diferença nas notas entre os grupos que possuem computador, ou emprega doméstica na residência e entre os sexos é vísivel e persiste ao longo das etapas escolares. (Ver figuras acima)

Cor/Raça	Nota Média ENEM	Docentes Qualificados (%)	Número de Alunos
Branca	546.37	71.31	136.16
Preta	506.04	69.87	142.66
Parda	508.47	69.64	138.29
Amarela	577.87	76.27	188.60
Indígena	492.01	70.66	133.45

Tabela 1: Dados Médios por Cor/Raça para Gestão de Complexidade de Escolas que possuam E.M. ou EJA

Os dados mostram que estudantes amarelos têm a maior nota média do ENEM e a maior proporção de adequação docente, enquanto estudantes brancos têm o segundo melhor desempenho geral.

Na construção da base de dados para a regressão linear múltipla, optou-se por apenas manter as variáveis de inscritos que estudem em colégios que ofereçam ensino médio ou EJA baseado no próprio objetivo de mensuração do exame, assim como, manter as seguintes variáveis independentes: "tp.sexo", "tp.cor.raca", "q007", "q024", "num.alunos.escola"e "prop.adeq.docente.alta". A variável dependente, ou variável resposta, escolhida foi "nota.media.enem".

3 Modelo de Regressão Linear Múltipla

O modelo de regressão linear múltipla ajustado pode ser representado pela seguinte equação:

$$\begin{aligned} y_i =& \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 s_{1i} + \beta_5 s_{2i} + \beta_6 rc_{1i} + \beta_7 rc_{2i} + \beta_8 rc_{3i} \\ &+ \beta_9 rc_{4i} + \beta_{10} q7_{1i} + \beta_{11} q7_{2i} + \beta_{12} q24_{1i} + \beta_{13} q24_{2i} + \epsilon_i \end{aligned} \tag{1}$$

3.1 Descrição das Variáveis

- y_i: Nota média do ENEM para o inscrito i.
- β_0: Intercepto do modelo.
- x_{1i}: Idade do inscrito i.
- x_{2i}: Número total de alunos na escola do inscrito i.
- x_{3i}: Proporção de docentes com alta qualificação na escola do inscrito i.
- s_{1i} e s_{2i}: Indicadores binários para o sexo dos alunos:
  - s_{1i}: Feminino (1 se feminino, 0 caso contrário).
  - s_{2i}: Masculino (1 se masculino, 0 caso contrário).
- rc_{1i}, rc_{2i}, rc_{3i} e rc_{4i}: Indicadores binários para cor/raça dos alunos:
  - rc_{1i}: Branca.
  - rc_{2i}: Parda.

  - rc_{3i}: Preta.
  - rc_{4i}: Indígena.

- $q7_{1i}$ e $q7_{2i}$: Indicadores binários indicando se o inscrito tem empregada doméstica na residência
  - $q7_{1i}$: Sim (1 se sim, 0 caso contrário).
  - $q7_{2i}$: Não (1 se não, 0 caso contrário).

- $q24_{1i}$ e $q24_{2i}$: Indicadores binários indicando se o inscrito tem computador na residência:
  - $q24_{1i}$: Sim (1 se sim, 0 caso contrário).
  - $q24_{2i}$: Não (1 se não, 0 caso contrário).

- ϵ_i: Termo de erro.

3.2 Modelo Preliminar

Obtemos então aproximadamente o modelo preliminar com os seguintes coeficientes:

$$\hat{y}_i = 647.70 - 11.08x_{1i} + 0.03x_{2i} + 0.71x_{3i} - 7.65s_{1i} -$$
$$26.64rc_{1i} - 25.07rc_{2i} + 22.96rc_{3i} - 41.53rc_{4i} + 62.61q7_{1i} + 31.31q24_{1i}$$

com $R^2 = 0.2235$, isto é, aproximadamente 22.35% da variabilidade de y pode ser explicada pela variabilidade dos regressores. À primeira vista, o modelo não parece ter uma boa adequação aos dados.

Na seleção dos modelos, empregamos o critério de seleção backward. Esse método começa ajustando o modelo com todas as variáveis disponíveis. A seguir, removemos as variáveis menos significativas uma a uma e ajustamos novamente o modelo sem essas variáveis em cada etapa. No nosso caso, o processo resultou no modelo preliminar com todas as variáveis.

3.3 Análise dos Pressupostos

A Regressão Linear Múltipla possui como pressuposto a homocedasticidade e normalidade dos resíduos. A investigação dos pressupostos mostrou sua falha, para um p-valor extremamente baixo foi rejeitada a hipótese de normalidade dos resíduos e através de um Q-Q foi mostrado desvio nos extremos, por possíveis outliers.

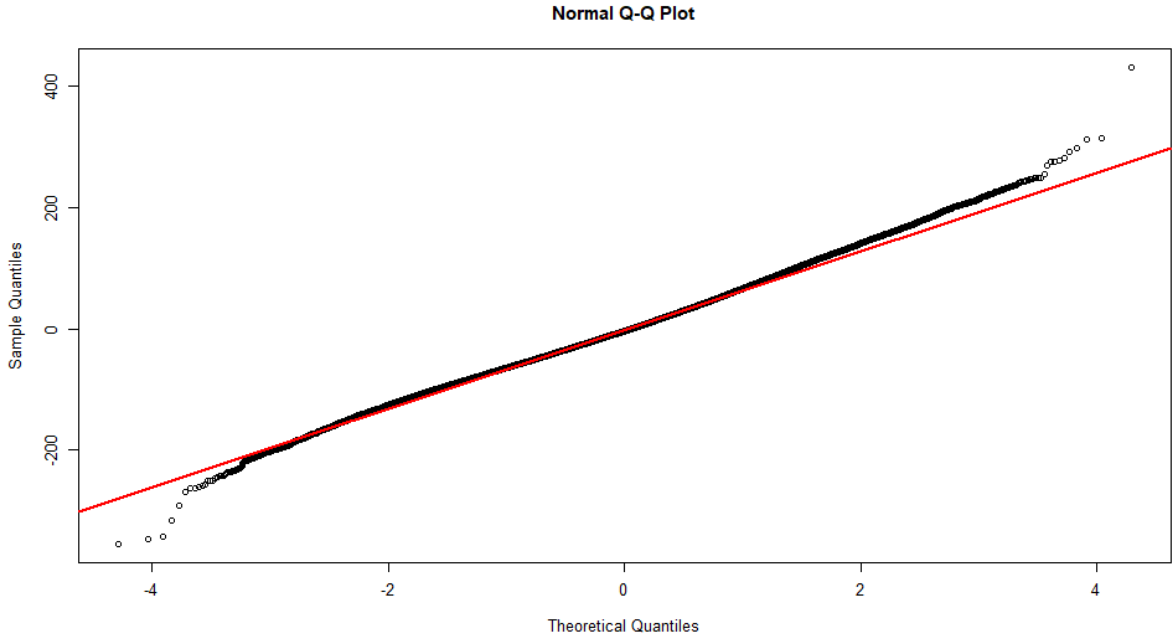

Figura 7: Gráfico Q-Q dos Resíduo do Modelo Preliminar

Foram usados quatro métodos diferentes para detecção de outliers: Alavancagem, Distância de Cook, Padronização e Studentização. Os pontos fora dos intervalos comuns da literatura ou com alta distância foram retirados do modelo, cerca de 0,6% dos dados. Podendo estes serem vistos na Figura 8.

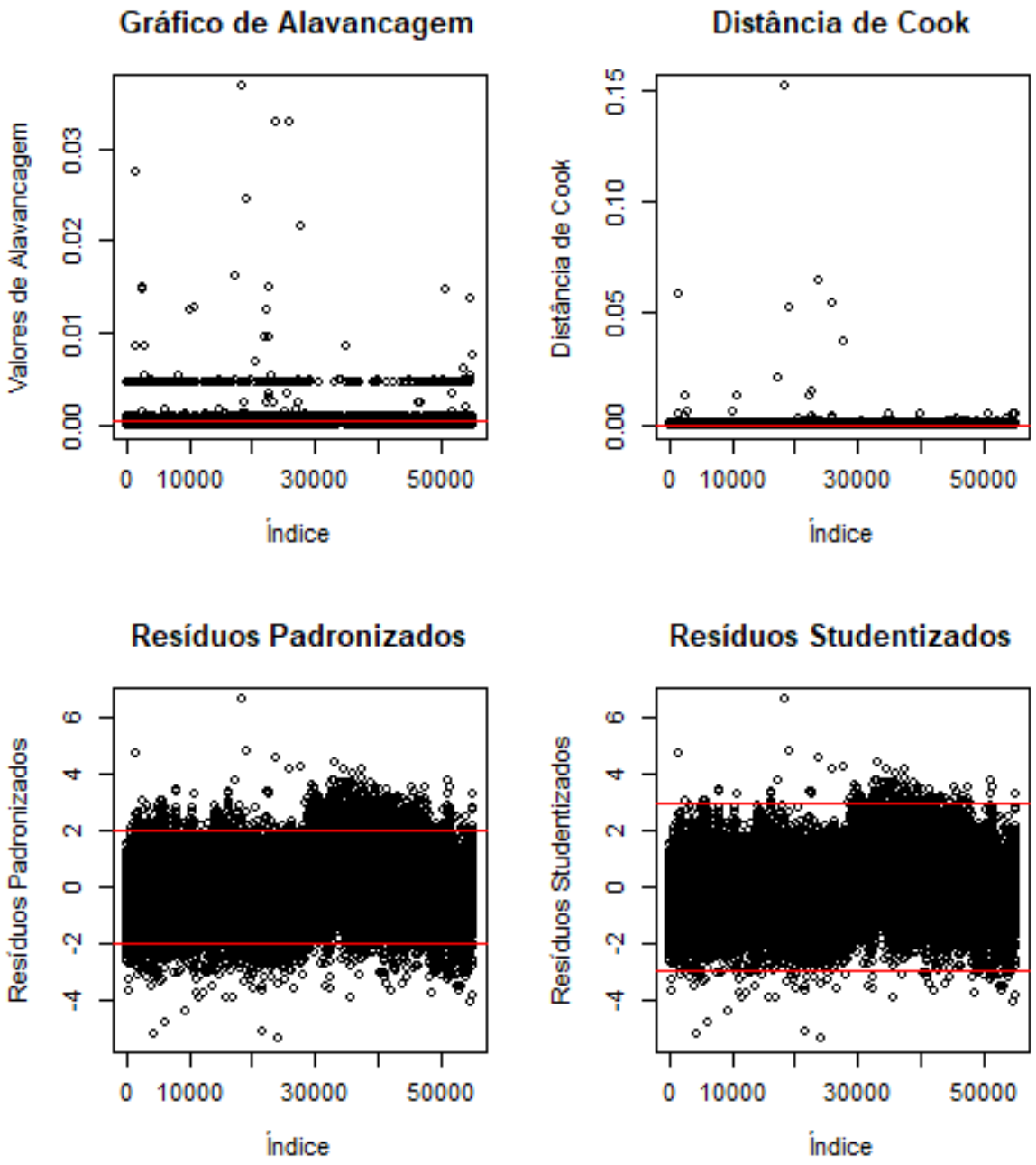

Figura 8: Gráficos para Visualização de Outliers

Após a retirada não houve mudanças significativas, então foi tentado o modelo de regressão linear múltipla pelo método dos mínimos quadrados ponderados e o R2 teve valor de aproximadamente 1. O teste de normalidade dos resíduos de Cramer-Von Mises teve a hipóstese nula rejeitada para o menor p-valor possível do software R estimar.

3.4 Modelo Escolhido

O ajuste do modelo pelo método dos mínimos quadrados ponderados foi o considerado melhor. Este foi realizado com pesos inversamente proporcionais aos quadrados dos resíduos após a remoção dos outliers. Os coeficientes estimados para as variáveis independentes

são apresentados a seguir:

- **Intercepto** (β_0): 647.704828
- $x1$: -11.083343
- $x2$: 0.029461
- $x3$: 0.705594
- **sexoF**: -7.666
- **cor_racaPreta**: -26.63
- **cor_racaParda**: -25.07
- **q007Sim**: 62.62
- **q024Sim**: 31.30

O R^2 ajustado é de 0.9999, indicando que aproximadamente 99.99% da variabilidade na variável dependente é explicada pelo modelo. Embora o valor de R^2 seja extremamente alto, deve-se tomar cuidado ao interpretar este resultado, pois a base de dados é muito grande. Os coeficientes não se alteraram, nem em valores nem em sinais devido a grande quantidade de números,

Podemos visualizar o ajuste do modelo WLS de valores ajustados versus os valores reais na 11:

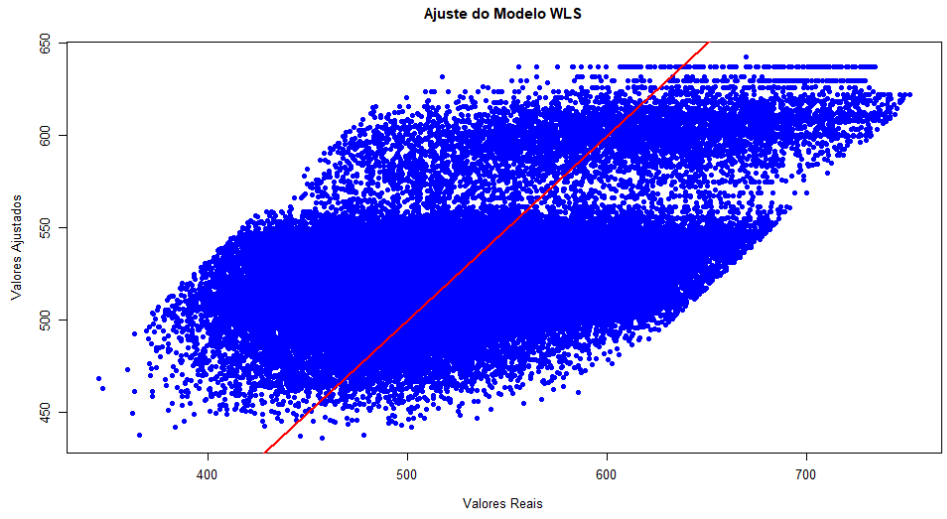

Figura 9: Ajuste do Modelo WLS

4 Discussão - Análise de Resíduos do Modelo Escolhido

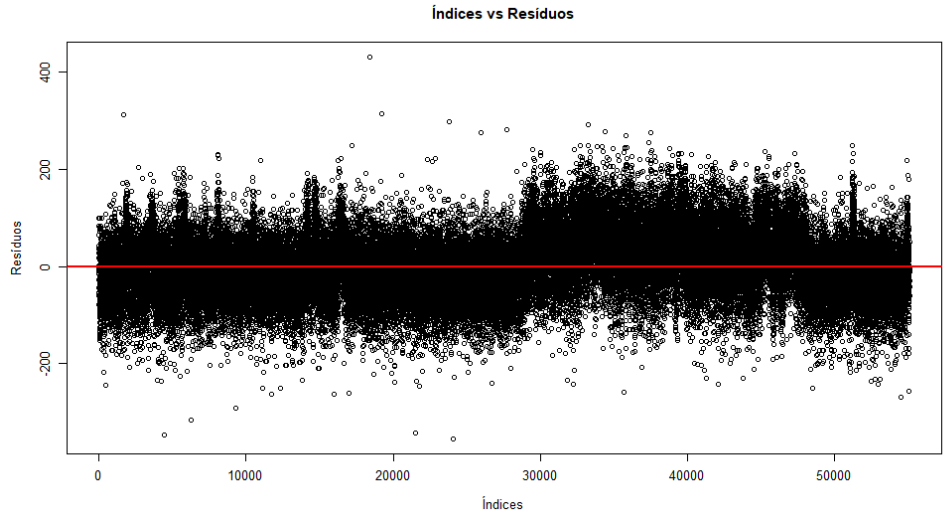

Figura 10: Ajuste do Modelo WLS

A heterocedasticidade persiste e na Figura 10 é possível ver uma certa periodicidade em determinados intervalos, dando indícios de que os resíduos são correlatados. Portanto, foi feito o teste de Durbin-Watson, e este mostrou evidência de dependência entre os resíduos (valor-p = 0), indicando autocorrelação.

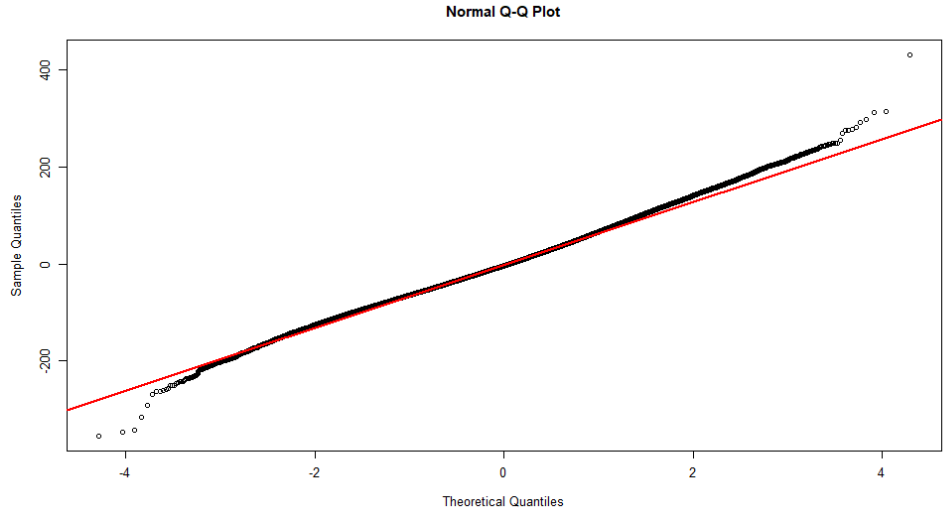

Figura 11: Ajuste do Modelo WLS

A retirada dos outliers também não se mostrou a melhor escolha pois, retirou o comparativo das outras raças como Amarela e Indígena e não trouxe alteração visível nos gráficos de resíduos.

5 Conclusão

Os coeficientes e os gráficos consoam entre si. As estimativas dos coeficientes corroboram para as premissas iniciais levantadas na análise descritiva dos dados, indivíduos do sexo feminino, negros, pardos e mais velhos são os que tem associação negativa com a nota média do ENEM.

No intercepto do modelo escolhido temos um valor positivo alto que contempla o indivíduo branco, nos demais coeficientes positivos temos por ordem de maior impacto para um bom desempenho no ENEM indivíduos que possuam computador na residência, possuam empregada doméstica, tenha uma proporção de docentes com alta qualificação na escola e número de alunos. Essas tendências indicam a influência de recursos e gênero no desempenho acadêmico.

Pelo grande volume de dados apesar da técnica de mínimos quadrados ter um R^2 satisfatório, não houve diferença nos coeficientes, nem atendeu as suposições dos resíduos. A quebra dos pressupostos para a regressão linear múltipla evidencia que ajustes no modelo são necessários para melhorar a precisão e tratar essas questões. Sugere-se investigar agrupamentos para entender melhor as influências no desempenho dos estudantes das escolas de São Paulo.

Referências

[1] Dutra, J. F., Firmino Júnior, J. B., & Fernandes, D. Y. S. *Fatores que podem interferir no desempenho de estudantes no ENEM: uma revisão sistemática da literatura*. Revista Brasileira de Informática na Educação, 31, 323-351. DOI: 10.5753/rbie.2023.3087.

[2] Souza, Rony Cardoso de. *Fatores que podem interferir no desempenho de estudantes no ENEM: uma revisão sistemática da literatura*. Universidade de São Paulo, Escola de Artes, Ciências e Humanidades, São Paulo, 2018.

[3] Conover, W. J. *Practical Nonparametric Statistics*. 3ª ed. John Wiley & Sons, 1998.

[4] McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. *Generalized Linear and Mixed Models*. 2ª ed. John Wiley & Sons, 2008.

[5] Ramos, Maria Nilza. *Mínimos Quadrados Ponderados*. Disponível em: https://www.kaggle.com/code/marianilzaramos/wls-m-nimos-quadrados-ponderados/edit.