UNIVERSIDADE FEDERAL DE OURO PRETO

TRABALHO FINAL – AJUSTE DE UM MODELO DE REGRESSÃO LINEAR MULTIPLA – Enem 2019

Grupo:

Diana Diniz 17.2.4348

Marco Antônio Baia Costa 18.2.4120

Paulo Vitor de Souza e Silva 18.2.4090

Raphaella Vitória Fernandes da Silva 17.2.4227

OURO PRETO - MG

2021

Sumário

1 Introdução

Neste trabalho iremos analisar uma base de dados composta por 1617 observações referentes aos candidatos que fizeram a prova do ENEM de 2019 na cidade de Ouro Preto. Esta base foi obtida ao realizarmos um filtro em base onde inicialmente haviam informação de todos os candidatos de Minas Gerais.

A base de dados contém as variáveis:

- NU_IDADE: Idade do inscrito em 31/12/2019.
- TP_SEXO: sexo do inscrito (‘Masculino’ ou ‘Feminino’)
- TP_ESTADO_CIVIL: Estado civil (‘Não informado’, ‘Solteiro(a)’, ’Casado(a)/Mora com companheiro(a)’, ‘Divorciado(a)/Desquitado(a)/Separado(a)’, ‘Viúvo(a)’)
- TP_COR_RACA: Cor/raça (Não declarado, Branca, Preta, Parda, Amarela, Indígena)
- TP_ESCOLA: Tipo de escola do Ensino Médio (Não Respondeu, Pública, Privada, Exterior)
- NO_MUNICIPIO_PROVA: Nome do município da aplicação da prova
- NU_NOTA_CN: Nota da prova de Ciências da Natureza
- NU_NOTA_CH: Nota da prova de Ciências Humanas
- NU_NOTA_LC: Nota da prova de Linguagens e Códigos
- NU_NOTA_MT: Nota da prova de Matemática
- NU_NOTA_REDACAO: Nota da prova de Redação
- Q001: Até que série seu pai, ou o homem responsável por você, estudou?
    - Nunca estudou.
    - Não completou a 4ª série/5º ano do Ensino Fundamental.
    - Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.
    - Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.
    - Completou o Ensino Médio, mas não completou a Faculdade.
    - Completou a Faculdade, mas não completou a Pós-graduação.
    - Completou a Pós-graduação.
    - Não sei.
- Q002: Até que série sua mãe, ou a mulher responsável por você, estudou? (mesmas respostas que a variável anterior)
- Q005: Incluindo você, quantas pessoas moram atualmente em sua residência?
- Q006: Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares).

- Q022: Na sua residência tem telefone celular? (Não. Sim, um. Sim, dois. Sim, três. Sim, quatro ou mais.)
- Q023: Na sua residência tem telefone fixo? (Sim. Não.)
- Q024: Na sua residência tem computador? (Não. Sim, um. Sim, dois. Sim, três. Sim, quatro ou mais.)
- Q025: Na sua residência tem acesso à Internet? (Sim. Não.)
- media: Média aritmética das variáveis referentes a notas.

1. A variável resposta é: media (Média aritmética das notas).
2. As variáveis explicativas são: "NU_IDADE", "TP_SEXO"," TP_ESTADO_CIVIL" "TP_COR_RACA", "TP_ESCOLA", "Q001", "Q002", "Q005", "Q006", "Q022", "Q023", "Q024" e "Q025".

A base de dados analisada foi fornecida ao grupo de trabalho por Carolina Silva Pena, Professora Adjunta do Departamento de Estatística da Universidade Federal de Ouro Preto, a qual agradecemos imensamente pelo apoio. Essa mesma base pode ser encontrada na internet no site do Inep.

2 Resultados

Nesta seção apresentaremos os resultados obtidos a partir do ajuste do modelo. Inicialmente, faremos uma análise exploratória da base de dados.

2.1 Análise exploratória

Podemos observar na Figura 1 a presença de quatro gráficos, sendo eles: (a) Gráfico de dispersão das notas no ENEM versus as idades, onde é possível observar uma leve relação linear decrescente, ou seja, quanto maior é a idade do candidato, menor sua nota. No gráfico (b) temos um Boxplot das notas no ENEM de acordo com o sexo, onde é possível observar distribuições bem parecidas, com notas similares e que possivelmente não sejam significativas no modelo de regressão, mas observamos que os candidatos do sexo masculino possuem uma média e mediana um pouco maior. O gráfico (c) apresenta um Boxplot das notas no ENEM em relação a Cor/raça Nota-se que pessoas brancas e não declaro, possuem notas maiores do que pessoas de outras raças. O gráfico (d) é um Boxplot das notas no ENEM de acordo com o tipo de escola do ensino médio, e através dele podemos concluir que pessoas de escola privada obtiveram notas maiores do que as pessoas de escolas públicas.

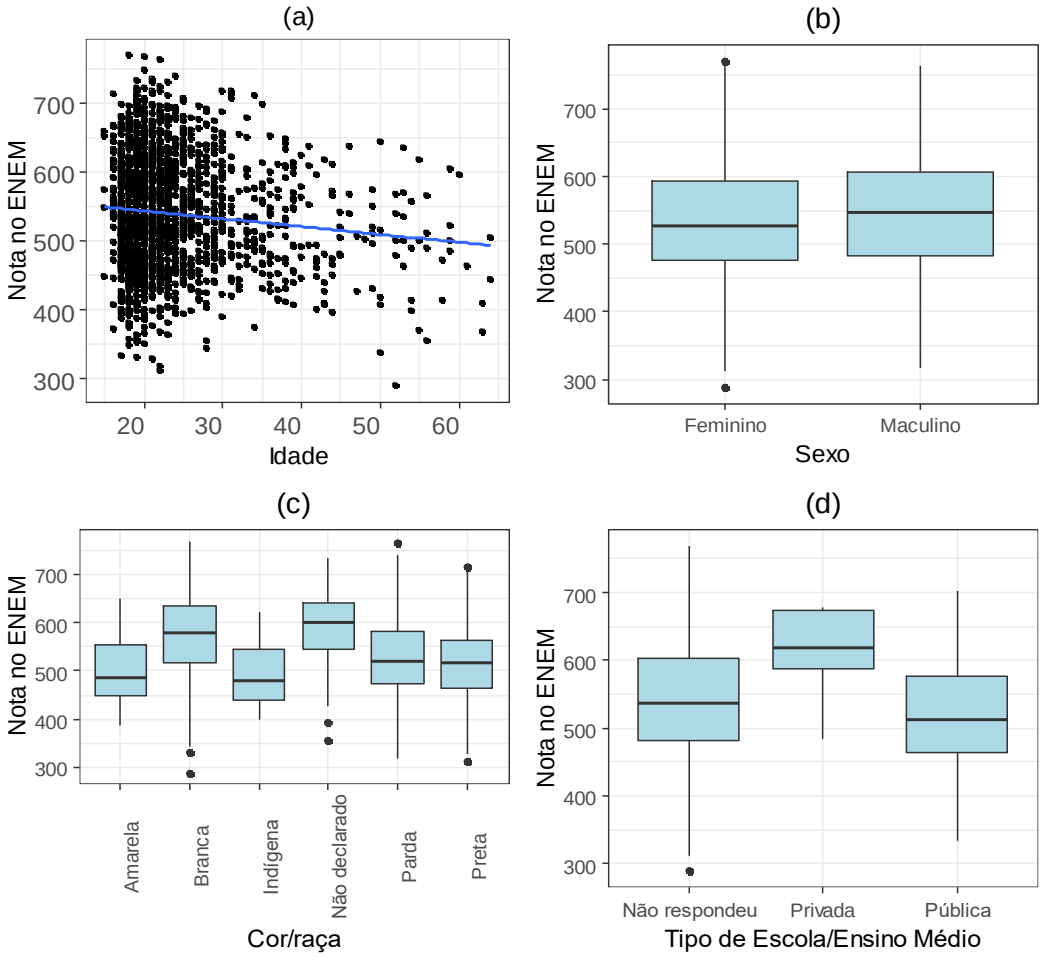

Figura 1: Análise exploratória das variáveis explicativas.

Podemos observar na Figura 2 a presença de três gráficos. O gráfico (a) é um *Boxplot* das notas no ENEM de acordo com o Estado civil, o qual torna-se possível dizer que viúvos(as) possuem as menores notas no ENEM. No gráfico (b) temos um *Boxplot* das notas no ENEM de acordo com a renda familiar. Nota-se que os candidatos que possuem as melhores notas estão com renda familiar acima de R$11.996,01. O gráfico (c) é um gráfico de dispersão das notas no ENEM versus número de residentes no qual observamos que não houve correlação entre as variáveis.

A Figura 3 é composta por dois gráficos, sendo eles: gráfico (a) - *Boxplot* das notas no ENEM de acordo com a escolaridade do pai, onde pode-se dizer que a maior média foram dos candidatos que possuem pai que completaram a Pós-Graduação e as menores médias foram de candidatos que não sabem sobre a escolaridade do pai, não estudou ou não conseguiu e pais que completaram a 4ª série/5º ano, mas não completaram a 8ª série/9º ano do ensino fundamental. O gráfico (b) é um *Boxplot* das notas no ENEM de acordo com a escolaridade da mãe. Ele mostra que os candidatos que tiveram menor média, a mãe não estudou ou o candidato não sabe o nível de escolaridade da mãe. É possível pressupor que a relação entre o nível de escolaridade dos pais e a nota do candidato pode impactar no resultado do candidato.

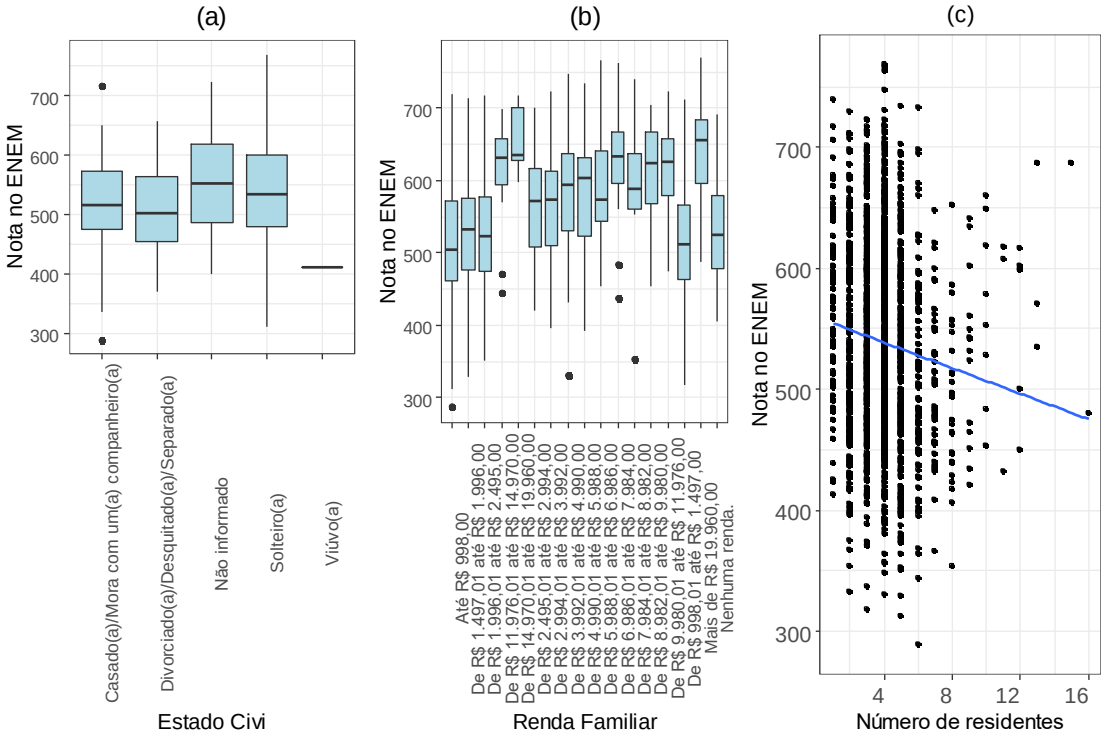

Figura 2: Análise exploratória das variáveis explicativas.

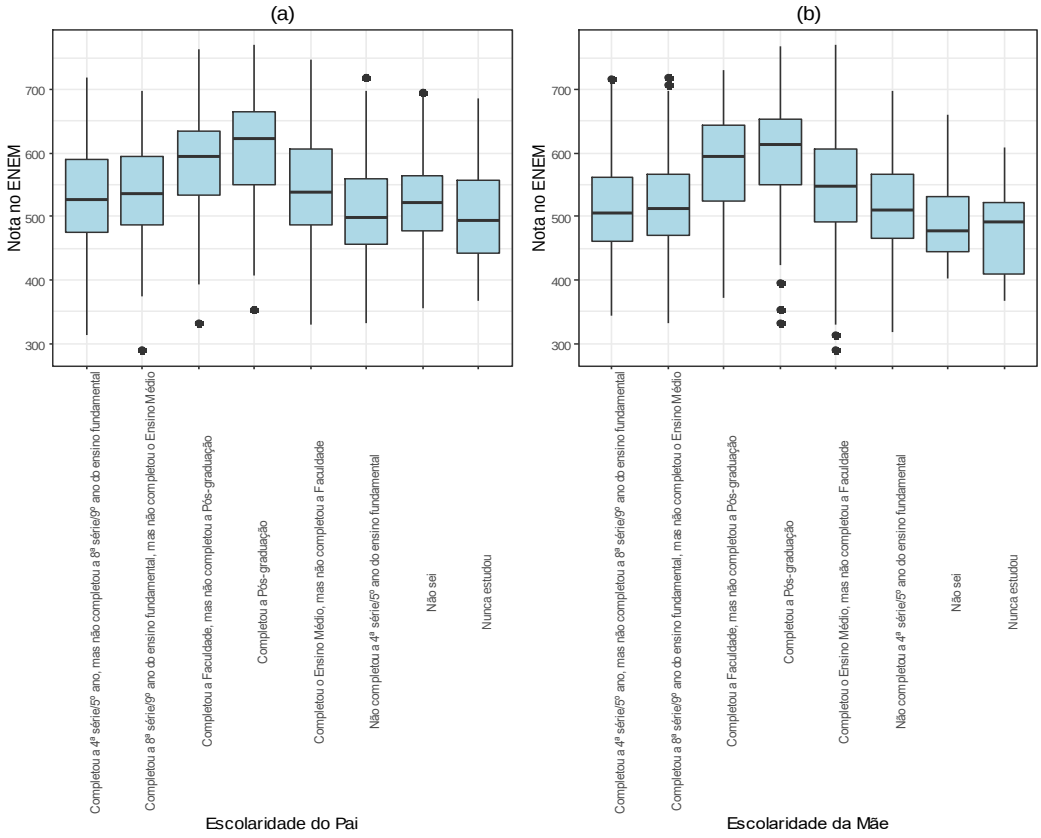

Figura 3: Análise exploratória das variáveis explicativas.

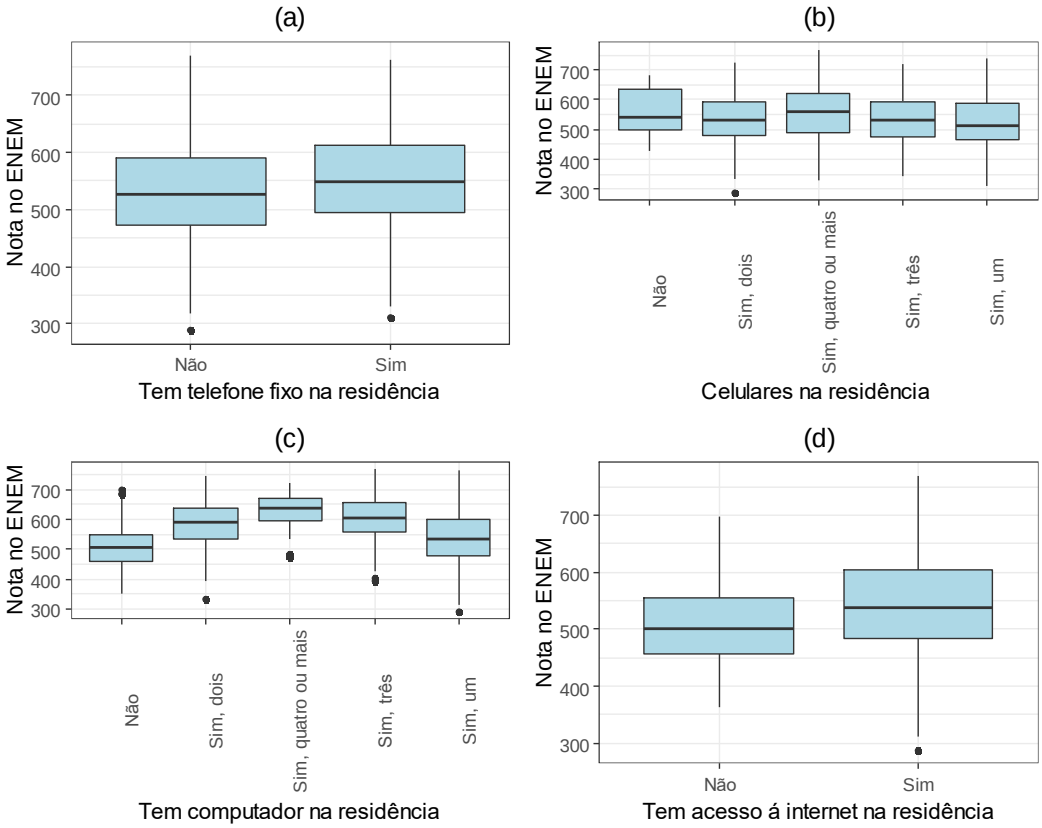

Figura 4: Análise exploratória das variáveis explicativas

Na Figura 4 temos a presença de quatro gráficos. O gráfico (a) é *Boxplot* das notas no ENEM de versus "Tem telefone fixo na residência", que mostra pouca diferença entre as variáveis, sendo a média um pouco maior para os candidatos que possuem telefone fixo. O gráfico (b) é um *Boxplot* das notas no ENEM de acordo com o número de celulares na residência. Podemos verificar uma variância relativamente baixa, mas sendo a maior média dos candidatos que possuem pelo menos 4 aparelhos celulares em casa. No gráfico (c) temos um *Boxplot* das notas no ENEM de acordo com o número de computadores na residência, nota-se uma variabilidade maior em relação ao gráfico (b), sendo a maior média também para candidatos que possuem pelo menos 4 celulares. O gráfico (b) é um *Boxplot* das notas no ENEM versus 'Tem acesso à internet na residência' mostra que candidatos com maiores médias possuem acesso à internet em casa.

2.2 Modelo Proposto

Inicialmente todas as variáveis explicativas foram inseridas no modelo. O ajuste do modelo foi realizado no software R. Verificamos que algumas variáveis apresentaram coeficiente não significativos, sendo assim, posteriormente faremos uma seleção de variáveis.

Para avaliar a significância do modelo iremos fazer uma Anova (análise de variâncias) para modelos aninhados. Isso significa comparar nosso modelo ajustado com um modelo derivado, ou seja, esse novo modelo possui a mesma variável resposta e se difere apenas no número de variáveis independentes. Esse novo modelo pode ser composto apenas pelo intercepto, então: modelo 1= (media ~1) e o modelo 2 será o modelo ajustado com todas as variáveis que foram prelecionadas e descritas na introdução. A anova irá comparar os modelos e fornecer um p-valor com base na estatística F. Essa análise de variância tem como hipótese nula (H0) que esses modelos são iguais, logo o desempenho deles na previsão das notas em estatística são iguais e a hipótese alternativa (Ha) é que eles são diferentes.

Tabela 1: Análise de variância para modelos aninhados.

	Res.Df	RSS	Df	SQ	F	Pr(>F)
1	199	794,19				
2	192	96,73	7	697,46	197,78	0

Para saber qual modelo é o melhor iremos olhar para o RSS (residual sum of squares). O melhor será o que possuir o menor valor. De acordo com a Tabela 1, o modelo 2 (modelo ajustado) possui o RSS menor e como o p-valor observado foi muito pequeno, podemos concluir ao nível de 5% de significância que existe pelo menos um coeficiente diferente de 0.

Tabela 2: Fator de inflação da variância (VIF).

	GVIF	Df	GVIF^(1/(2*Df))
NU_IDADE	2,018268	1	1,420658
TP_SEXO	1,051733	1	1,02554
TP_ESTADO_CIVIL	1,71872	4	1,070042
TP_COR_RACA	1,355022	5	1,030848
TP_ESCOLA	1,215804	2	1,050064
Q001	3,339943	7	1,089958
Q002	3,034286	7	1,082512
Q005	1,399585	1	1,18304
Q006	3,488983	16	1,039823
Q022	1,935423	4	1,086043
Q023	1,158573	1	1,07637
Q024	2,08181	4	1,095986
Q025	1,345289	1	1,159866

O VIF é utilizado para avaliar se existe multicolinearidade entre as variáveis explicativas do MRLM (modelo de regressão linear múltipla). Quando o VIF é maior que 10, há uma multicolinearidade significativa que precisa ser corrigida. Para isso é preciso remover uma (ou mais) das variáveis altamente correlacionadas. Como a informação fornecida pelas variáveis é redundante, o coeficiente de determinação não será muito prejudicado pela remoção. Caso seja necessário a excluir alguma variável do modelo, isso precisa ser feito com muita cautela e junto com o pesquisador pois essa variável com maior correlação pode ser a mais importante da pesquisa e uma forma de resolver esse problema é talvez excluindo uma outra variável de menor importância.

Na tabela 2 temos os valores do VIF para as variáveis do nosso modelo. Todas possuem valores menores que 10, então conclui-se que não existe uma multicolinearidade muito alta nessa base de dados.

Para realizar a seleção de variáveis no modelo usaremos o algoritmo "passo a passo" (Stepwise) que consiste na remoção das variáveis explicativas uma a uma, baseado em algum critério de seleção. Existem vários critérios, mas usaremos o AIC (Critério de Informação de Akaike) que é uma das medidas utilizadas para selecionar o melhor MRLM visando diminuir o número de parâmetros no modelo e aumentar a precisão para as inferências. Quanto maior for o número de variáveis incluídas no modelo, maior a possibilidade de colinearidade. O AIC pondera a complexidade do modelo (número de parâmetros) e a qualidade do ajuste. Quanto menor o valor dessa medida, melhor.

A seleção pode ser feita manualmente, porém no R, temos as funções 'step' do pacote stats e 'stepAIC' do pacote MASS que selecionam o modelo utilizando o método stepwise com opções backward, forward e both.

Na Tabela 3 são apresentadas as análises de variância da parte fixa do modelo. Podemos verificar que neste modelo final o intercepto estimado pelo modelo ($\hat{\beta}_0$) foi igual a 532,3416. Isso significa que quando as demais variáveis assumirem o valor zero, o valor esperado da média na nota do Enem é igual a 532,3. Essa interpretação é possível pois verificou-se, ao nível de 5% de significância, que o coeficiente de inclinação $\hat{\beta}_0$ é significativamente diferente de zero (p-valor = 0).

Tabela 3: Análise de variância da parte fixa do modelo reduzido a partir do modelo

	Estimativa	Erro padrão	Valor T	P-Valor
(Intercept)	532,3416	15,8644	33,5557	0

NU_IDADE	-0,9076	0,2593	-3,5001	0,0005
TP_COR_RACABranca	35,6483	12,6592	2,816	0,0049
TP_COR_RACAIndígena	-20,588	33,6634	-0,6116	0,5409
TP_COR_RACANão declarado	47,4948	16,4447	2,8882	0,0039
TP_COR_RACAParda	13,1933	12,3564	1,0677	0,2858
TP_COR_RACAPreta	9,5863	12,6405	0,7584	0,4483
TP_ESCOLAPrivada	48,6016	26,8944	1,8071	0,0709
TP_ESCOLAPública	-22,0183	4,9597	-4,4395	0
Q001Completou a 8ª série/9º ano do ensino fundamental, mas não completou o Ensino Médio	10,1679	6,6171	1,5366	0,1246
Q001Completou a Faculdade, mas não completou a Pós-graduação	11,1197	8,0491	1,3815	0,1673
Q001Completou a Pós-graduação	0,2417	9,8009	0,0247	0,9803
Q001Completou o Ensino Médio, mas não completou a Faculdade	1,4549	5,5163	0,2638	0,792
Q001Não completou a 4ª série/5º ano do ensino fundamental	-13,8047	5,9625	-2,3153	0,0207
Q001Não sei	5,5067	8,0389	0,685	0,4934
Q001Nunca estudou	-8,8186	12,8964	-0,6838	0,4942
Q002Completou a 8ª série/9º ano do ensino fundamental, mas não completou o Ensino Médio	-2,55	6,5953	-0,3866	0,6991
Q002Completou a Faculdade, mas não completou a Pós-graduação	26,1214	7,5555	3,4573	0,0006
Q002Completou a Pós-graduação	32,6254	8,1569	3,9997	0,0001
Q002Completou o Ensino Médio, mas não completou a Faculdade	16,311	5,4931	2,9693	0,003
Q002Não completou a 4ª série/5º ano do ensino fundamental	9,1925	6,3506	1,4475	0,148

Q002Não sei	-18,6567	13,9991	-1,3327	0,1828
Q002Nunca estudou	-15,2678	16,2148	-0,9416	0,3465
Q005	-5,4358	1,0399	-5,227	0
Q006De R$ 1.497,01 até R$ 1.996,00	5,7369	6,7802	0,8461	0,3976
Q006De R$ 1.996,01 até R$ 2.495,00	0,3409	6,9478	0,0491	0,9609
Q006De R$ 11.976,01 até R$ 14.970,00	55,0045	17,8013	3,0899	0,002
Q006De R$ 14.970,01 até R$ 19.960,00	66,038	25,0719	2,634	0,0085
Q006De R$ 2.495,01 até R$ 2.994,00	32,5648	8,8199	3,6922	0,0002
Q006De R$ 2.994,01 até R$ 3.992,00	24,4674	7,8249	3,1269	0,0018
Q006De R$ 3.992,01 até R$ 4.990,00	36,1905	9,5763	3,7792	0,0002
Q006De R$ 4.990,01 até R$ 5.988,00	25,3632	11,5313	2,1995	0,028
Q006De R$ 5.988,01 até R$ 6.986,00	39,6072	15,1456	2,6151	0,009
Q006De R$ 6.986,01 até R$ 7.984,00	46,81	17,2271	2,7172	0,0067
Q006De R$ 7.984,01 até R$ 8.982,00	35,7016	22,84	1,5631	0,1182
Q006De R$ 8.982,01 até R$ 9.980,00	51,2909	17,4972	2,9314	0,0034
Q006De R$ 9.980,01 até R$ 11.976,00	45,9891	15,9566	2,8821	0,004
Q006De R$ 998,01 até R$ 1.497,00	-5,606	5,4489	-1,0288	0,3037
Q006Mais de R$ 19.960,00	57,1268	21,1519	2,7008	0,007
Q006Nenhuma renda.	3,4951	15,3016	0,2284	0,8194
Q024Sim, dois	31,6174	6,9189	4,5697	0
Q024Sim, quatro ou mais	54,7171	14,4326	3,7912	0,0002

Q024Sim, três	36,4747	11,2205	3,2507	0,0012
Q024Sim, um	13,2012	4,274	3,0888	0,002

O coeficiente de inclinação associado a variável NU_IDADE (idade do inscrito) foi igual a -0,91. Isso significa que para cada aumento de um ano de vida do participante, espera-se uma redução de - 0,91 na nota do ENEM. Essa interpretação é possível pois verificou-se, ao nível de 5% de significância, que o coeficiente de inclinação é significativamente diferente de zero (p-valor = 0).

A variável TP_COR_RACA (Raça do participante) foi significativa para as categorias BRANCA e NÃO DECLARADO. Ao nível de 5% de significância, verificou-se que o coeficiente de inclinação associado a categoria 'BRANCA' em relação a categoria 'AMARELA' foi igual a 35,65. Isso significa que espera-se um aumento de 35,65 na nota do ENEM. O coeficiente de inclinação associado a categoria 'NÃO DECLARADO' foi igual a 47,5. Isso significa que espera-se um aumento de 47,5 na nota para aqueles que pertencem a categoria NÃO DECLARADO em relação a categoria AMARELA. As demais categorias não foram significativas.

A variável TP_ESCOLA (escola que o participante frequentou) foi significativa para a categoria 'Pública'. Ao nível de 5% de significância, verificou-se que o coeficiente de inclinação associado a categoria 'Pública' em relação a categoria 'Não respondeu' foi igual a -22,02. Isso significa que se espera uma redução de -22,02 pontos na nota do ENEM. As demais categorias não foram significativas.

A variável Q001 (escolaridade do pai) foi significativa para a categoria 'Não completou a 4ª série/5º ano do Ensino Fundamental'. Ao nível de 5% de significância, verificou-se que o coeficiente de inclinação associado a esta categoria em relação a categoria 'Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental' foi igual a -13,8. Isso significa que se espera uma redução de -13,8 pontos na nota do ENEM. As demais categorias não foram significativas.

A variável Q002 (escolaridade da mãe) foi significativa para quase todas as categorias. Com isso podemos afirmar que a escolaridade da mãe influencia no aumento da nota do ENEM. Por exemplo: espera-se um aumento de 32,06 pontos na nota daqueles que possuem mãe que 'Completou a Pós-graduação' em relação a aqueles que possuem mãe que 'Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.'

O coeficiente de inclinação associado a variável 'Q005' (número de pessoas que moram com o candidato) foi igual a -5,44. Isso significa que para cada aumento de uma pessoa na casa

do participante, espera-se uma redução de -5,44 na nota do ENEM. Essa interpretação é possível pois verificou-se, ao nível de 5% de significância, que o coeficiente de inclinação é significativamente diferente de zero (p-valor = 0).

A variável Q006 (renda mensal da família) foi significativa para praticamente todas as categorias. Com isso podemos afirmar que a renda familiar influencia no aumento da nota do ENEM. Por exemplo: espera-se um aumento de 57,13 pontos na nota daqueles que possuem renda acima de R$ 19.960,00 em relação a aqueles que possuem renda de até R$ 998,00.

Por fim temos a variável Q024(computadores na residência do participante), a qual foi significativa (ao nível de 5% de significância) para todas as categorias. Por exemplo: para aqueles que possuem Quatro ou mais computadores em sua residência se espera um aumento de 54,7 pontos na nota em relação a aqueles que não possuem computadores.

Ainda sobre o ajuste do modelo, o valor do coeficiente de determinação (R^2 ajustado) estimado pelo modelo foi igual a 0,2642736. Isso significa que 26,4% da variabilidade da variável resposta (média na nota do ENEM) pode ser explicada pelo modelo de regressão. O valor de R^2 varia entre 0 e 1 e quanto mais próximo estiver de 1, maior o poder de explicação. O valor estimado foi relativamente baixo

2.2.1 Análise do ajuste do modelo

A análise de resíduos é utilizada para verificar se as suposições a respeito dos erros ($\boldsymbol{\varepsilon}$) do modelo estão adequadas e se o modelo ajustado é adequado. Para verificar se a suposição de normalidade dos erros é verdadeira, testamos a normalidade dos resíduos utilizando o teste de shapiro wilk. O p-valor obtido foi igual a 0,1654, ou seja, ao nível de 5% de significância não rejeitamos a hipótese de normalidade dos resíduos.

No gráfico (a) da Figura 5 os pontos estão aderentes a linha, confirmando a suposição de normalidade dos erros. No gráfico *valores ajustados X resíduos padronizados* (gráfico (b)) os pontos estão distribuídos aleatoriamente em torno de 0, confirmando a suposição de homocedasticidade dos erros. No gráfico *ordem X resíduos padronizados* (gráfico (c)) os pontos também estão distribuídos aleatoriamente em torno de 0, confirmando a suposição de independência dos erros.

Na Figura 6 vemos que nenhum valor ultrapassou o critério $D_i > 1$ e poucos valores foram observados acima da linha que representa o critério $D_i > 4/n$. O valor máximo observado da distância de Cook foi igual a 0,0394. Os pontos que ultrapassaram o segundo critério devem

ser investigados, entretanto como nenhum deles ultrapassou o primeiro critério é possível que não estejam causando grande influência no ajuste. Destaca-se ainda que a remoção de qualquer ponto da base de dados só deve ser feita de acordo com a decisão da equipe, caso seja identificado algum erro de medição.

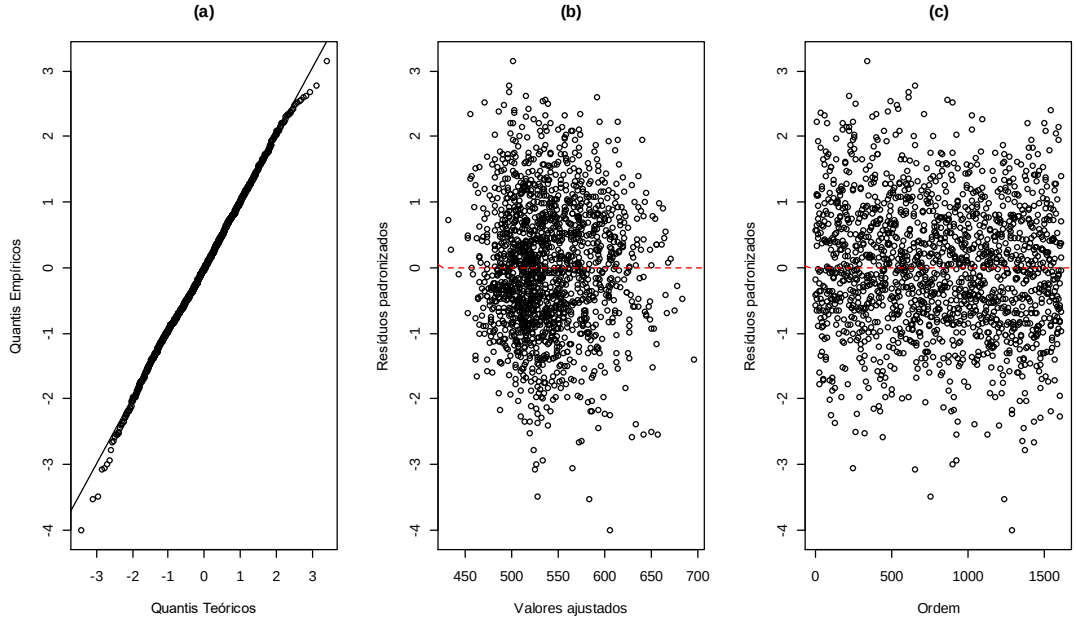

Figura 5: Análise dos resíduos padronizados

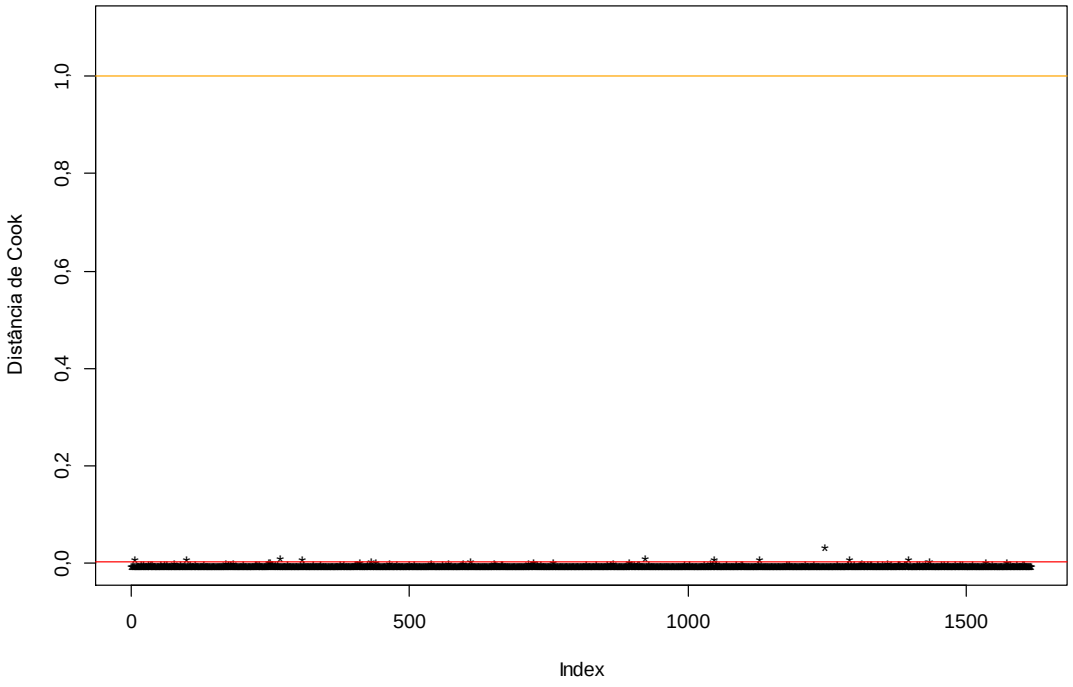

Figura 6: Distância de Cook para a identificação de pontos influentes.

Referências

INEP. **Microdados,** Microdados do Enem 2019. Disponível em:https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem/.Acesso em: 20 de ago. de 2021.

HOFFMANN, Rodolfo A. Análise de Regressão: Uma introdução à Econometria. USP, 2016.Disponivel em : https://www.esalq.usp.br/biblioteca/sites/default/files/Analise_Regress%C3%A3o.pdf .Acesso em 20 de ago. de 2021.