UNIVERSIDADE FEDERAL DE OURO PRETO
DEEST - Departamento de Estatística

ANÁLISE DE REGRESSÃO MÚLTIPLA

Diana Diniz

Marco Antonio

Paulo Vitor

Grupo 5

Trabalho Final da disciplina EST127- Análie de Regressão

Raphaella Vitória

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 havia informação de todos os candidatos de Minas Gerais.

Variável Resposta

Média (Média aritmética das notas).

Variáveis Explicativas

Idade do inscrito, Sexo, Estado Civil, Cor/Raça, Tipo de escola, Nível de escolaridade do pai/responsável do sexo masculino e da mãe/responsável do sexo feminino, Quantidade de moradores na mesma residência, Renda familiar, Na residência possui: telefone fixo? Celular? Computador? Tem acesso a internet?.

Dicionário das Variáveis

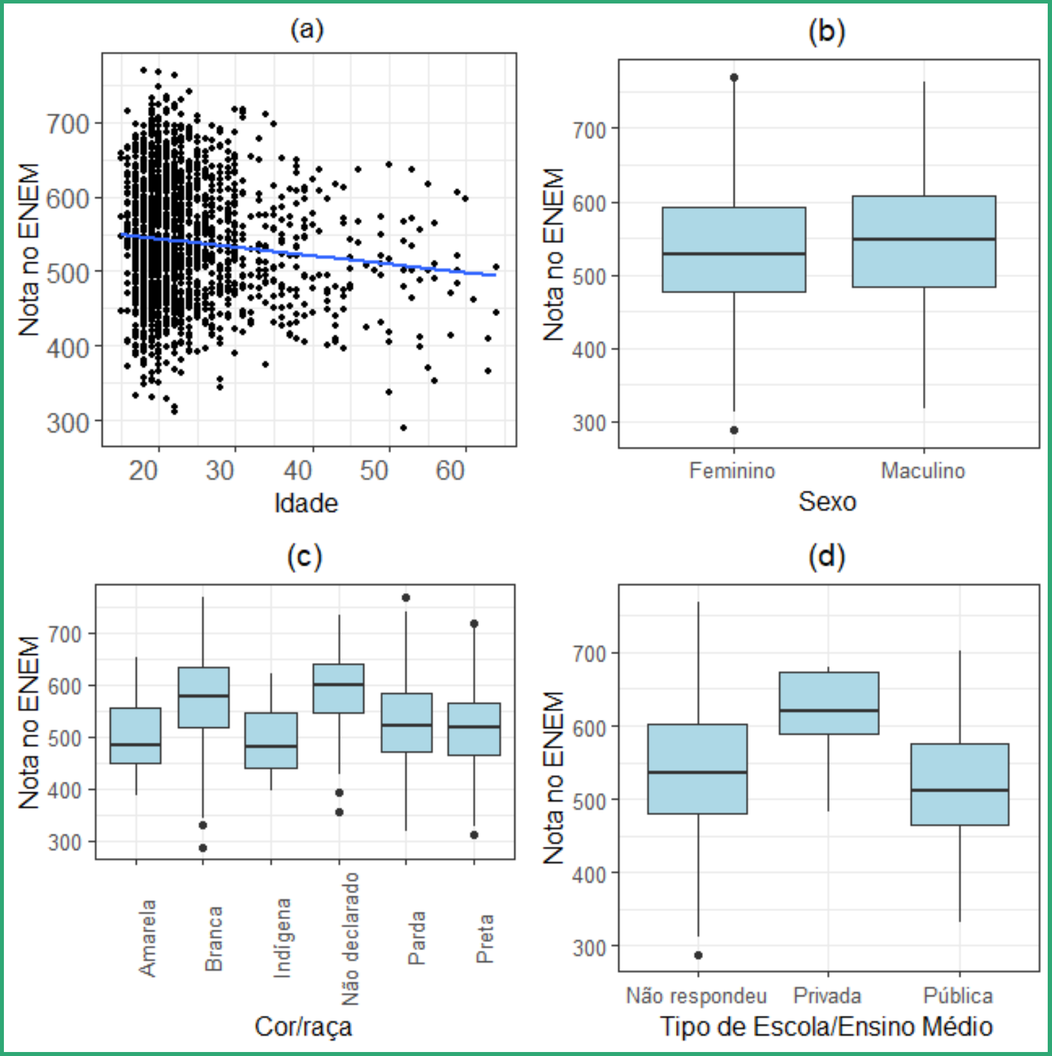
NU_IDADE	**Idade²**
TP_SEXO	Sexo
TP_ESTADO_CIVIL	Estado Civil
TP_COR_RACA	Cor/raça
TP_ESCOLA	Tipo de escola do Ensino Médio
Q001	Até que série seu pai, ou o homem responsável por você, estudou?
Q002	Até que série sua mãe, ou a mulher responsável por você, estudou?
Q005	Incluindo você, quantas pessoas 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?
Q023	Na sua residência tem telefone fixo?
Q024	Na sua residência tem computador?
Q025	Na sua residência tem acesso à Internet?

Ajuste do Modelo

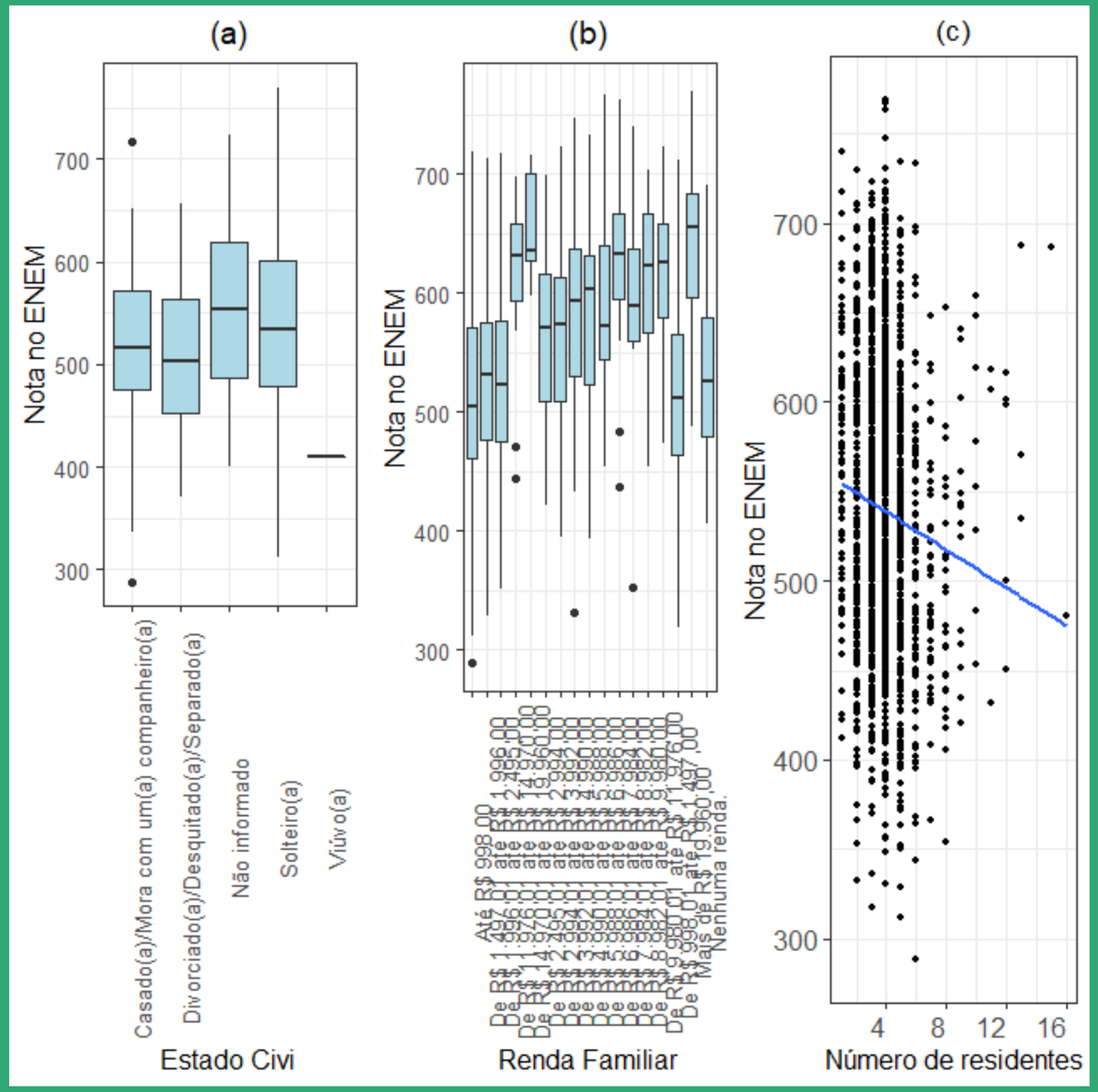
Inicialmente todas as variáveis explicativas da base de dados foram inseridas no modelo. O ajuste do modelo foi realizado no software R. Verificamos que algumas variáveis apresentaram coeficiente não significativos, sendo assim, posteriormente faremos uma seleção de variáveis.

1. Análise exploratória das variáveis explicativas
2. Análise de Variância - ANOVA
3. Fator de Inflação da Variável - VIF
4. Seleção de Variáveis
5. Coeficiente de Determinação
6. Análise de Resíduos
7. Distância de Cook

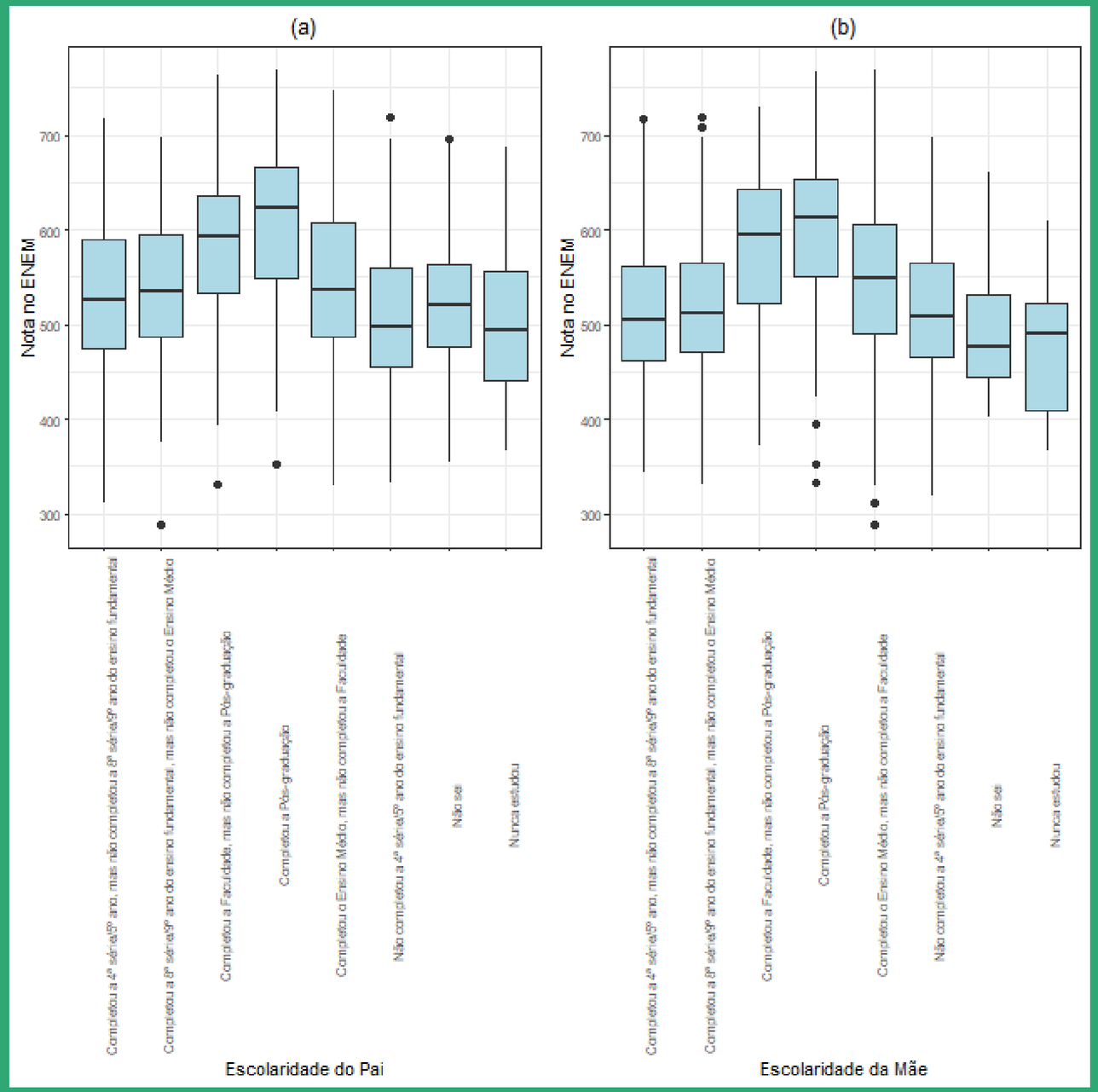
Análise exploratória das variáveis explicativas



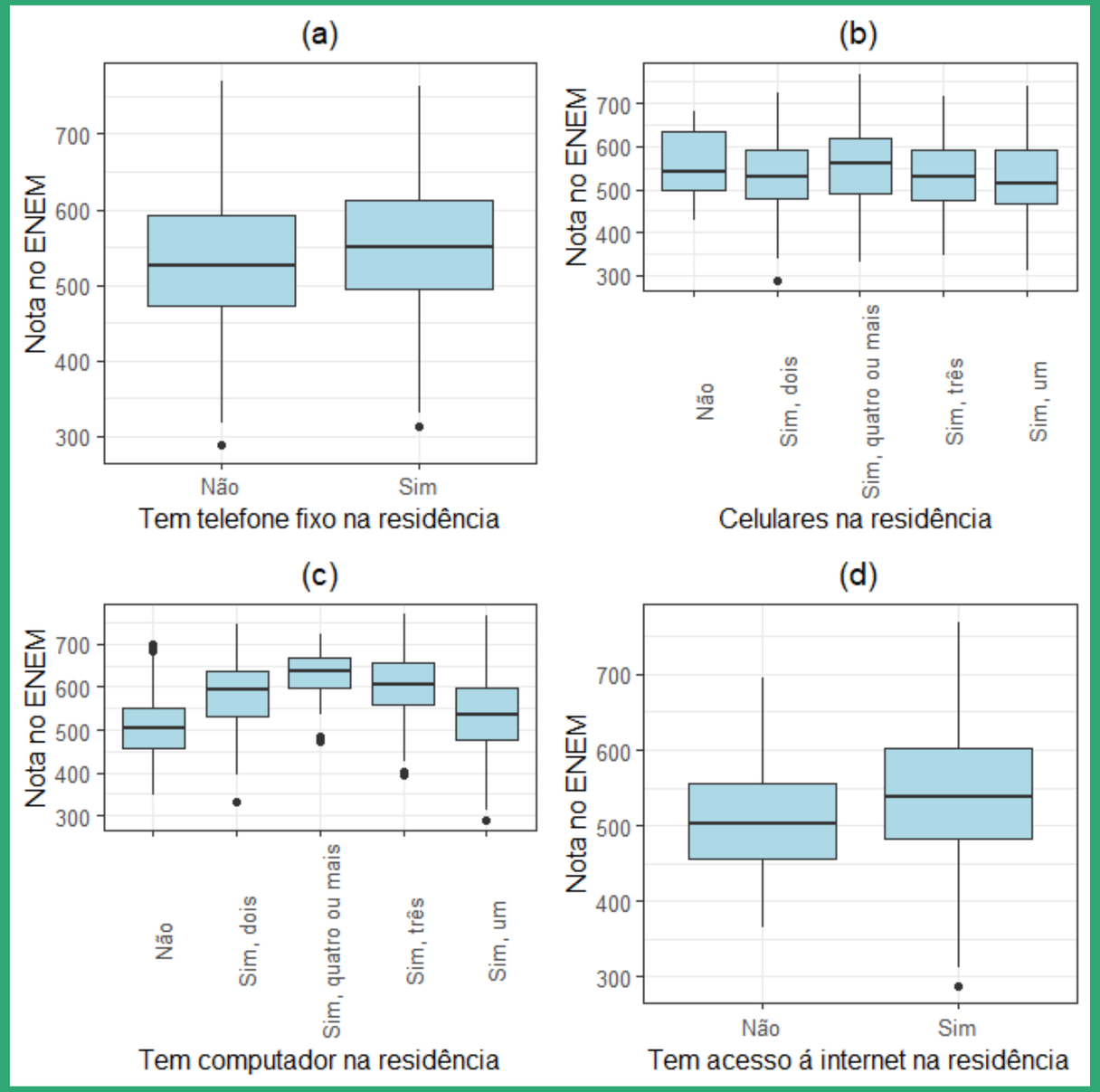
(a) Gráfico de dispersão das notas no ENEM versus as idades, onde é possível observar uma leve relação linear decrescente.
(b) Temos um Boxplot das notas no ENEM de acordo com o sexo, onde é possível observar distribuições bem parecidas, com notas similares e que possivelmente não sejam significativas no modelo de regressão, mas observamos que os candidatos do sexo masculino possuem uma média e mediana um pouco maior.
(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.
(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 tiraram notas maiores do que as pessoas de escolas públicas.



(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.
(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.
(c) é um gráfico de dispersão das notas no ENEM versus número de residentes no qual observamos que não houvecorrelação entre as variáveis.



(a) Boxplot das notas no ENEM de acordo com a escolaridade do pai, onde pode-se dizer que a maior média dos 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 estuou ou não conseguiu e pais que completaram a 4ª série/5º ano, mas não completou a 8ª série/9º ano do ensino fundamental.
(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 desempenho do mesmo.



(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.
(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 localizada nos candidatos que possuem pelo menos 4 aparelhos celulares em casa.
(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.

Análise de Variância - ANOVA

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 abaixo, mod2 (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.

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

Fator de Inflação da Variável - VIF

Temos , ao lado, 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.

	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

Modelo com todas as variáveis
X
Modelo ápos a seleção de variáveis

Antes

media = NU_IDADE+TP_SEXO+TP_ESTADO_CIVIL+TP_COR_RACA+TP_ESCOLA+
Q001+Q002+Q005+Q006+Q022+Q023+Q024+Q025

Depois

media = NU_IDADE + TP_COR_RACA + TP_ESCOLA + Q001+Q002 +Q005+ Q006 + Q024

Seleção de Variáveis

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.

	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_RACAIndigena	-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

	Estimativa	Erro padrão	Valor T	P-Valor
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

	Estimativa	Erro padrão	Valor T	P-Valor
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

	Estimativa	Erro padrão	Valor T	P-Valor
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

	Estimativa	Erro padrão	Valor T	P-Valor
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

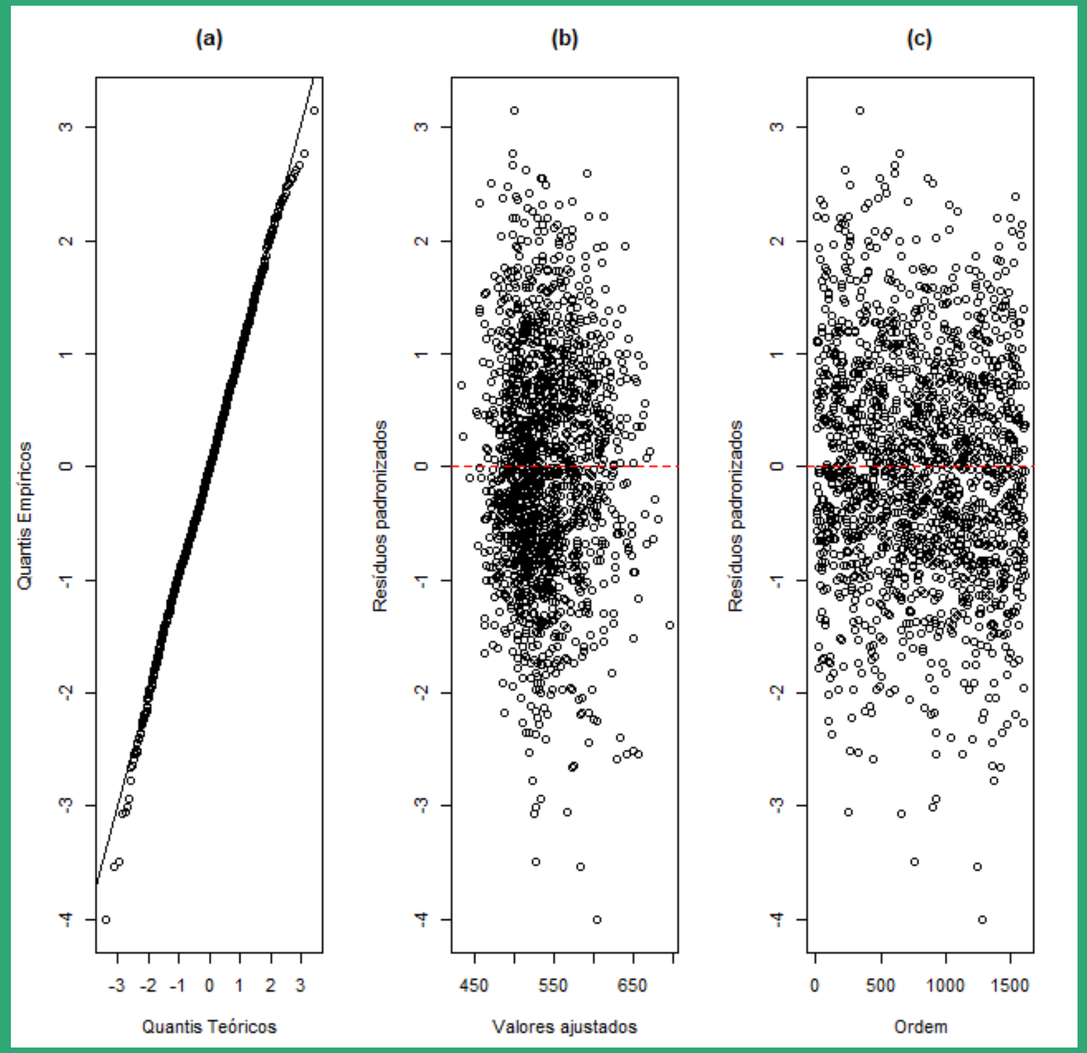
Coeficiente de Determinação

O valor do coeficiente de determinação 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.

Análise de Resíduos

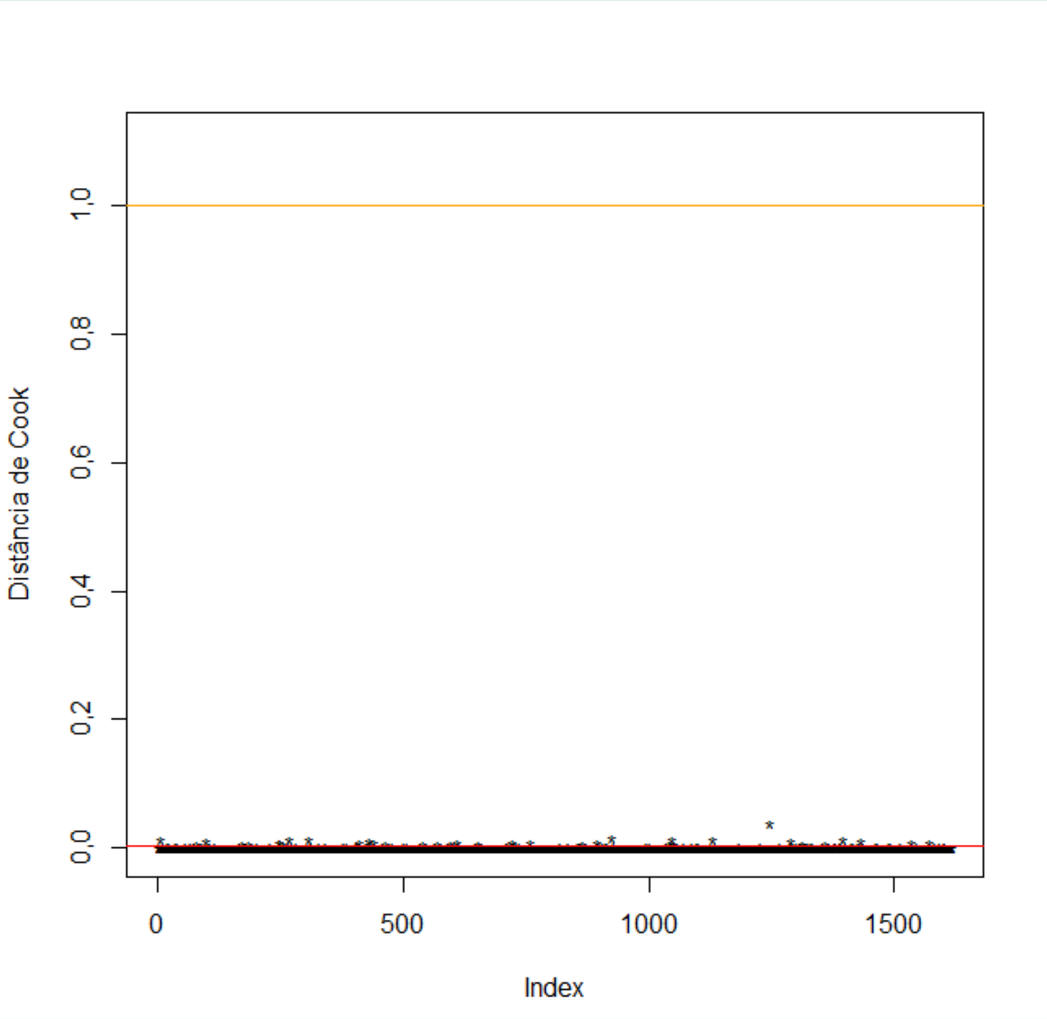
Para verificar se a suposição de normalidade dos erros é verdadeira, testamos a normalidade dos resíduos utilizando o teste de shapiro wilk. O p-valor obtido foi igual a 0,1654, ou seja, ao nível de 5% de significância não rejeitamos a hipótese de normalidade dos resíduos.

Análise dos resíduos padronizados



(a) Os pontos estão aderentes a linha, confirmando a suposição de normalidade dos erros.
(b) Os pontos estão distribuídos aleatoriamente em torno de 0, confirmando a suposição de homocedasticidade dos erros.
(c) Os pontos também estão distribuídos aleatoriamente em torno de 0, confirmando a suposição de independência dos erros.

Distância de Cook



Podemos observar que nenhum valor ultrapassou o critério 1 e poucos valores foram observados acima da linha que representa o critério 2. 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.

Obs: 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.

Agradecimentos

Agradecemos imensamente à Professora Carolina pelo apoio e atenção ao longo do perído e principalmente para realização deste trabalho.