Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Matemática Computacional - IMECC
Análise de Regressão linear - ME613

Análise de desempenho dos candidatos do ENEM-2019 (São Paulo)

Leonardo Mazzamboni Colussi

Campinas, 2021.

Conteúdo

1 Introdução

Este projeto consiste em realizar uma análise do desempenho médio dos candidatos (não-treineiros) do Exame Nacional do Ensino Médio (ENEM) de 2019, referente ao estado de São Paulo, identificando possíveis variáveis que podem estar associadas ao desempenho dos candidatos e, posteriormente, construir um modelo para uma análise quantitativa. Dessa forma, na primeira seção do projeto será feita uma análise descritiva dessas variáveis presentes no banco de dados utilizado. Por sua vez, na segunda seção, será feito uma modelagem estatística, assim como as etapas para a sua construção. Na última seção, por fim, será apresentado a interpretação dos resultados e a conclusão a respeito do trabalho realizado.

Desse modo, foi utilizado um banco de dados que apresenta 500 observações, exclusivamente, de candidatos do estado de São Paulo, as quais contém, além das notas médias dos candidatos, informações que podem estar associadas ao desempenho, como idade, sexo, estado civil, tipo de escola do Ensino Médio (pública ou privada), raça/cor, renda, número de pessoas que residem na casa onde mora, acesso à internet, renda familiar mensal e escolaridade dos pais.

2 Análise descritiva dos dados

Antes de iniciar a modelagem, foi realizado uma análise descritiva dos dados para entender o comportamento das variáveis que podem compor um estudo mais interessante conforme os dados estão disponíveis. Desse modo, há variáveis que são intuítivas e, de primeira vista, podem levar a acreditar que apresentam forte associação em relação ao desempenho dos candidatos no exame, no entanto, deve-se observar a disposição dos dados para ver se essa é representativa no conjunto utilizado.

Sendo assim, no que tange à variável `Internet`, tem-se a Tabela 1 que demonstra a quantidade e a porcentagem dos candidatos que possuíam acesso e os que não.

Tabela 1: Disposição dos dados referente ao acesso à internet.

Internet	Quantidade	Porcentagem (%)
Não	27	5,4
Sim	473	94,6

Note que, embora seja uma variável interessante de se considerar na modelagem e também intuitiva de que esteja associada ao desempenho dos candidatos, a quantidade de quem não tinha acesso $(5,4\%)$ é pouco representativa quando comparada a quem tinha $(94,6\%)$. Além disso, devido ao estado de SP ser um dos mais desenvolvidos do Brasil, o acesso à internet ser majoriatariamente positivo neste conjunto de dados pode ser reflexo desse fato, em contrapartida, pode não ser a realidade de outros estados brasileiros menos desenvolvidos no país, ou até de regiões mais desabastadas desse mesmo estado.

Tabela 2: Disposição dos dados referente ao estado civil.

Estado Civil	Quantidade	Porcentagem (%)
Solteiro(a)	496	99,2
Outros	4	0,8

A mesma situação ocorre com a variável `Estado Civil`, Tabela 2, que descreve a porcentagem de candidatos solteiros ou não. Notoriamente, essa variável se encontra muito mal distribuída, onde $99,2\%$ dos candidatos são solteiros e apenas $0,8\%$ não são, de modo que é pouco representativa para se usar na modelagem, podendo inflacionar o modelo com variáveis pouco informativas.

Para a variável idade, nota-se que, apesar de existir pessoas acima da faixa de 17-18 anos, ao observar para a mediana (Tabela 3), que é uma medida robusta, essa é igual a 17 anos, além disso, o terceiro quartil ainda está nesse intervalo, evidenciando que a maioria dos candidatos se encontra nessa faixa etária.

Tabela 3: Disposição dos dados referente à idade.

Mínimo	1º quartil	Mediana	3º quartil	Média	Máximo
16	17	17	17,45	18	46

Ao analisar a Figura 1, a qual compara os grupos de raças e o gênero em relação ao desempenho, é percetível que há diferença entre candidatos pertencentes ao grupo de raça branca/amarela com os de raça preta/parda/indígena, de forma que a mediana da nota média desse segundo grupo é menor que a do primeiro. Além disso, o grupo de raça branca/amarela apresenta maior variabilidade que o grupo de raça preta/parda/indígena, fato que pode estar associado a maior frequência relativa do segundo grupo na escola pública, como consta na Figura 2. Embora a categoria de candidatos pertencentes ao primeiro grupo com acesso a uma escolaridade mais precária também ocorra, para o segundo é ainda mais expressiva.

Por sua vez, o desempenho dos candidatos do sexo masculino é maior quando comparado com os candidatos do sexo feminino, tal fato não está associado à má distribuição entre os gêneros, dado que esses se encontram bem distribuídos ($57,4\%$ feminino e $42,6\%$ masculino). Ainda, é possível notar que as notas médias dos candidatos estão contidas no intervalo de 300 a 740 pontos.

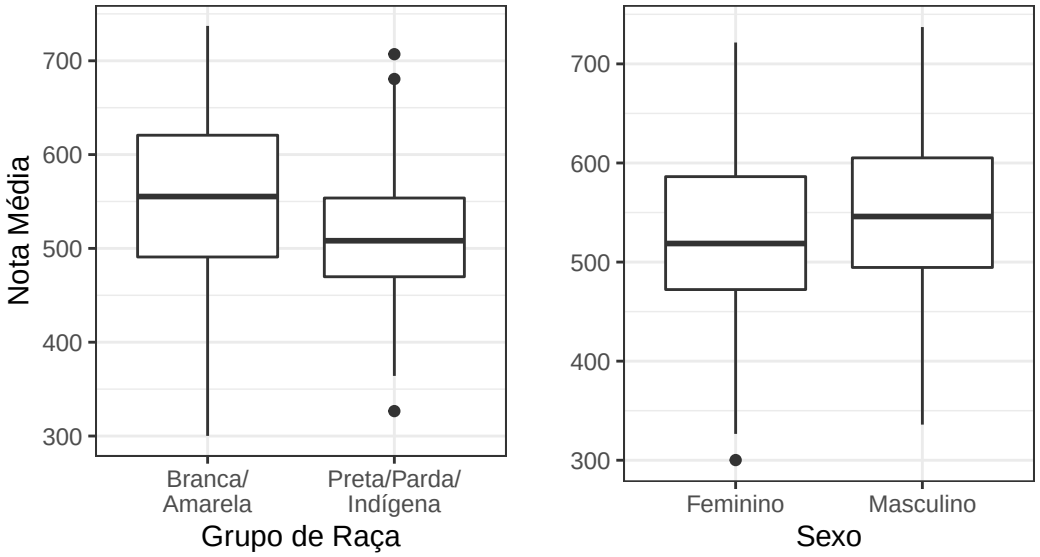

Figura 1: Variação do desempenho médio por grupo de raças.

Ainda, referente ao primeiro gráfico da Figura 2, observa-se que o desempenho médio no exame, tanto para os candidatos pertencentes às raças branca/amarela, quanto para às preto/pardo/indígena, são maiores quando alocados na classe D da renda familiar, no entanto, a representatividade deste segundo grupo nesta categoria é muito mais baixa ao comparar ao primeiro (Tabela 4), fato atrelado a contextos históricos de desigualdade social intrínseco ao contexto racial. Desse modo, as variáveis relacionadas à renda familiar e aos grupos raciais podem estar associadas. A variável renda está separada em quatro categorias: Até R$ 998,00 (A); De R$ 998,01 até R$ 1.996,00 (B); De R$ 1.996,01 até R$ 4.990,00 (C) e Mais de R$ 4.990,00 (D).

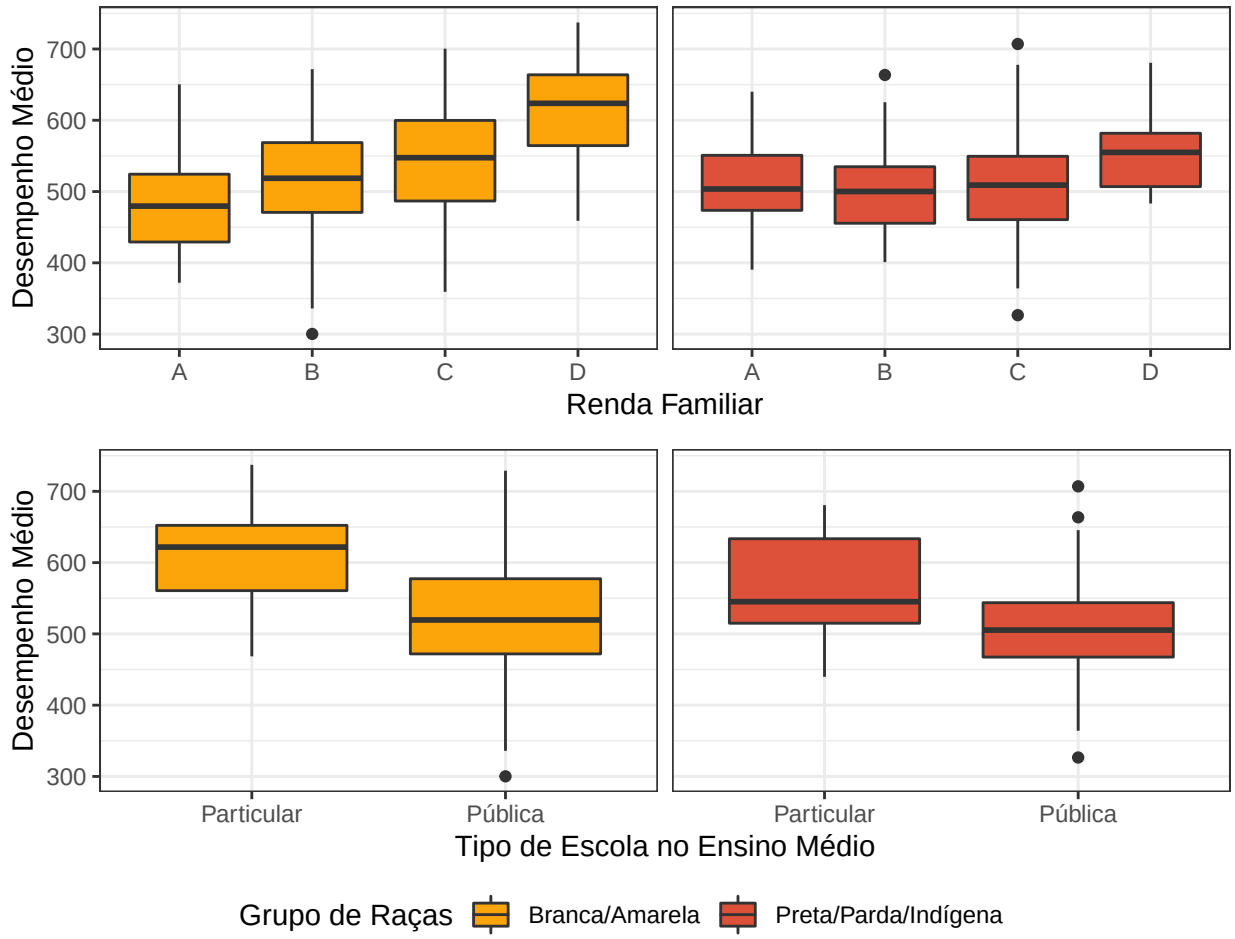

Figura 2: Boxplots entre grupos de raças e renda familiar em relação ao desempenho do candidato no exame.

Tabela 4: Disposição dos dados referente à renda familiar e ao grupo de raças

	Renda Familiar			
Grupo de Raças	A	B	C	D
Branca/Amarela	18	86	121	92
Preta/Parda/Indígena	18	76	69	20

Outra variável interessante de se analisar é a escolaridade dos pais (Figura 3) o que, a princípio, é intuitivo que há uma associação entra elas. Ademais, tanto a escolaridade do pai (ou homem responsável) quanto da mãe (ou mulher responsável), foram separadas em quatro categorias: fundamental, médio, superior e "não sabe".

Como se pode ver nos *boxplots* da Figura 3, há uma relação negativa entre o desconhecimento dos candidatos em relação à escolaridade dos pais, independente da renda familiar (exceto para a categoria D, de maior renda), com o desempenho médio do candidato. Isso pode estar associado ao abandono paterno e a falta de auxílio, como pensão, acarretando em uma renda familiar menor. Além disso, para ambos os casos, um maior nível de escolaridade é indício de melhor desempenho dos filhos no exame, dado que, para a categoria de ensino Superior, a mediana das notas é maior em relação as demais, inclusive nas maiores rendas.

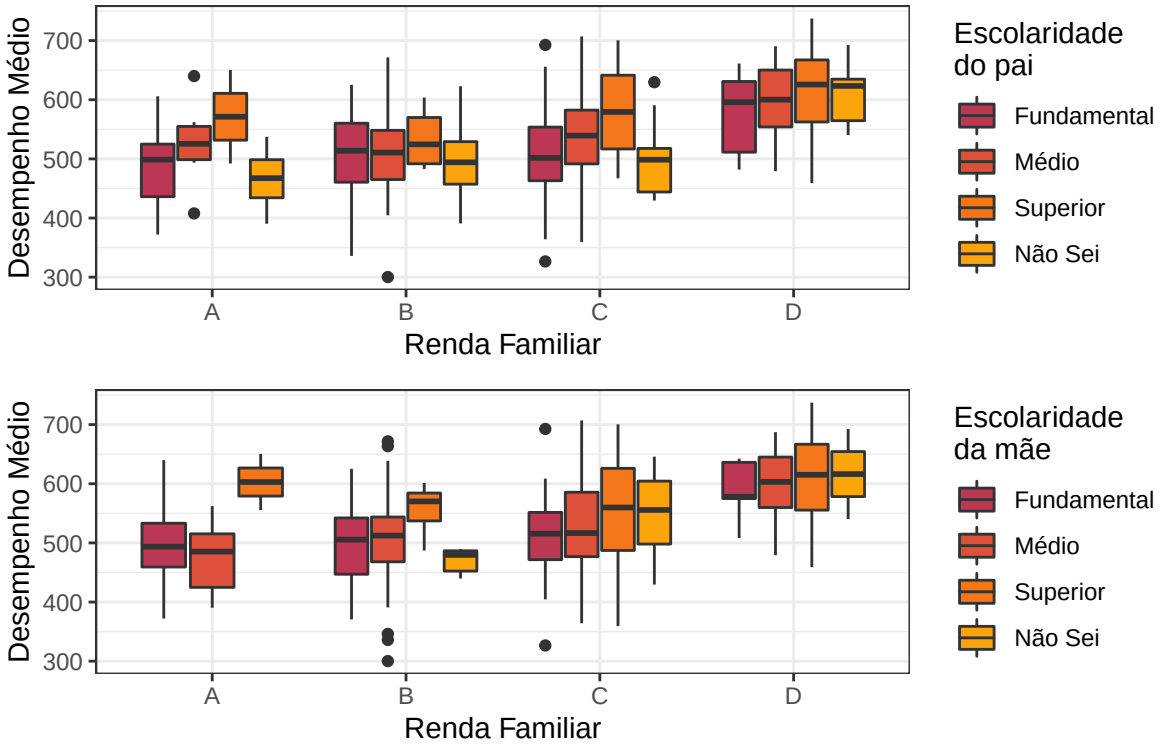

Figura 3: Disposição da escolaridade dos pais por renda familiar em relação ao desempenho dos candidatos no exame.

Tabela 5: Disposição dos dados referente ao Nº de pessoas que residem na mesma casa.

Mínimo	1º quartil	Mediana	3º quartil	Média	Máximo
2	3	4	3,93	4	10

A respeito da variável `Nº de pessoas que residem na casa onde mora` (Tabela 5), tem-se que a maioria dos candidatos moram com mais 2 ou 3 pessoas na casa, é de conhecimento geral que o mais comum nas famílias paulistas são casas que contém até 5 pessoas, característica de um estado urbanizado.

3 Modelagem Estatística

3.1 Seleção do Modelo

Para iniciar a modelagem, desconsiderou-se as variáveis pouco representativas (`Internet` e `Estado Civil`), vistas na seção anterior, dado que são pouco informativas e apenas inflacionariam o modelo com mais variáveis, também analisou-se a correlação entre as demais variáveis, de modo que nenhuma foi maior que $0,65$ e, consequentemente, descartou-se a possibilidade de estarem fortemente correlacionadas a ponto de prejudicar a construção do modelo. Dessa forma, ajustou-se inicialmente um modelo com o tipo de escola do Ensino Médio, renda, raça e sexo, obtendo a Tabela 6.

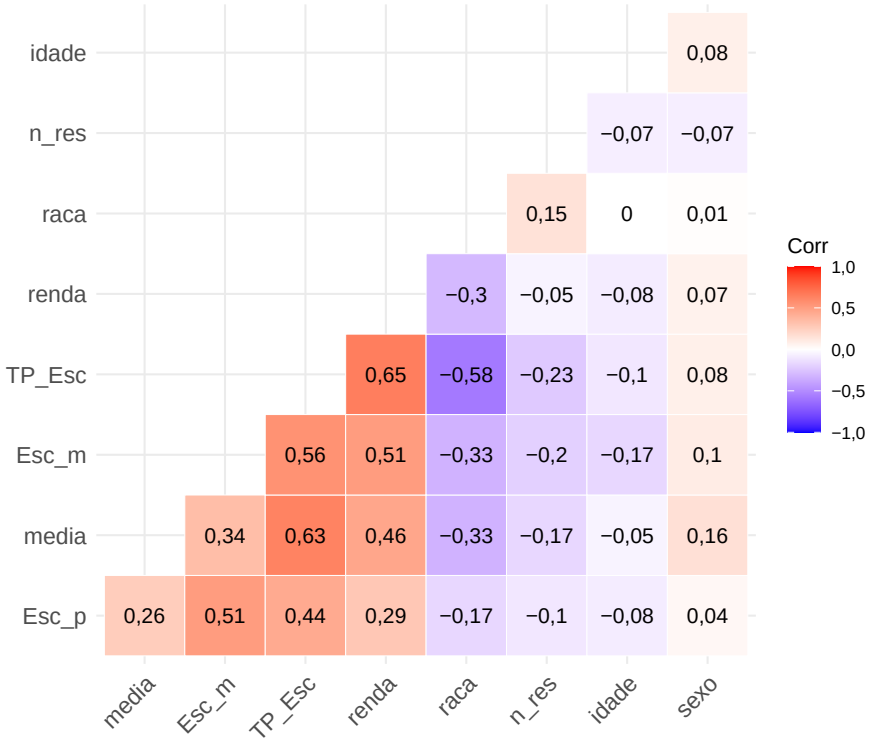

Figura 4: Gráfico de correlação entre as variáveis do banco de dados.

Tabela 6: Tabela dos coeficientes de efeitos fixos estimados no modelo.

Parâmetros	Estimativas	Erro Padrão	Estatística t	P-valor
Intercepto	496,10	11,70	42,42	<0,001
Tipo de Escola (particular)	57,02	7,95	7,18	<0,001
Renda B	9,50	12,19	0,78	0,4362
Renda C	18,45	12,17	1,52	0,1302
Renda D	67,81	13,50	5,02	<0,001
Raça (preto/pardo/indígena)	-15,50	6,51	-2,38	0,0176
Sexo (masculino)	17,26	6,00	2,88	0,0042

Nota-se que todas as variáveis escolhidas, a priori, para o modelo, segundo o $p-valor$ associados às estatísticas t, foram significativas a um nível de $\alpha = 5\%$ de significância, com destaque ao `Tipo de Escola` e a `Renda`, as quais foram muito significativas. No entanto, as categorias (B) e (C) da variável `Renda` não foram muito informativas. Dessa forma, decidiu-se juntá-las na casela de refência (categoria (A)), assim, a variável renda passou ter duas categorias: renda familiar menor que R$ $4990,00$ reais e maior que esse valor.

Ademais, da Tabela 7 para esse modelo inicial, dado que essas variáveis preditoras foram significativas e, também não se rejeitou a hipótese nula, associado à estatística F, de que os β's adicionados eram iguais a zero conforme as outras variáveis já estavam no modelo, decidiu-se mantê-las, pois eram informativas.

Tabela 7: Tabela ANOVA do primeiro modelo.

	GL	SQ	QM	Estatística F	P-valor
Tipo de Escola	1	749000	749000	171	<0,001
Renda	3	229000	76400	17,5	<0,001
Raça	1	23300	23300	5,32	0,0215
Sexo	1	36200	36200	8,27	0,0042
Resíduos	493	2160000	4380		

Agora, fazendo a alteração nas categorias da variável `Renda` e adicionando a variável `Escolaridade do pai` no modelo, obteve-se a Tabela 8.

Tabela 8: Coeficientes estimados no modelo com a variável "Escolaridade do pai".

Parâmetros	Estimativas	Erro Padrão	Estatística t	P-valor
Intercepto	502,39	6,27	80,16	<0,001
Tipo de Escola (particular)	53,30	8,10	6,58	<0,001
Renda > R$ 4999,00	47,02	8,34	5,64	<0,001
Raça (preto/pardo/indígena)	-14,63	6,48	-2,26	0,0244
Sexo (masculino)	15,62	6,01	2,60	0,0097
Escolaridade do pai (EM)	13,77	7,05	1,95	0,0514
Escolaridade do pai (Superior)	25,30	9,75	2,60	0,0097
Escolaridade do pai (Não sabe)	-5,67	12,07	-0,47	0,6384

Observe que, embora apenas uma das categorias relacionadas à variável `Escolaridade do pai` foi significativa, decidiu-se manter todas, pois torna o modelo mais condizente e interpretativo, dado que a categoria "Não sabe" pode estar associada a candidatos que não tiveram contato paterno e, de algum modo, refletir na nota média do candidato. Todavia, na Tabela 9, o $p-valor$ referente ao teste F deu menor que o nível de significância $\alpha = 5\%$, sendo assim, há evidências para rejeitar hipótese nula de que os β's associados à variável `Escolaridade do pai` é igual a zero. Logo, temos evidência que essa variável possui influência no desempenho do candidato, dado que as outras variáveis já estavam no modelo anterior.

Para finalizar a escolha do modelo, adicionou-se todas as outras variáveis, inclusive todos os possíveis pares de interação e, novamente, comparando o modelo anterior com este último, obteve-se o valor da estatística F de $0,887$, com o $p-valor$ de 0,72.

Tabela 9: Tabelas ANOVA entre modelos

Modelo inicial vs Modelo acrescentando a Escolaridade do pai					Modelo inicial com a Escolaridade do pai vs Modelo acrescentando as demais variáveis e interações				
GL Res.	SQ	QM	Estatística F	P-valor	GL Res.	SQ	QM	Estatística F	P-valor
490	2200000				490	2100000			
490	2100000	27000	6,2	0,013	430	1900000	250000	0,89	0,72

Desse modo, como não houve evidências para rejeitar a hipótese nula que os β's associados às variáveis `Escolaridade da mãe`, `Nº de pessoas na residência` e todas as possíveis interações são iguais a zero, optou-se pelo modelo anterior. Este fato ocorre quando as variáveis adicionadas não agregam ao ajuste ou quando a mesma informação está sendo explicada por outras variáveis já presentes no modelo. Ademais, foi utilizado o método automático de seleção via AIC (critério com valor de $4192,88$), o que, por fim, obteve-se o mesmo ajuste, exceto pela variável `Nº de pessoas na residência`, a qual decidiu-se não incluir em razão do teste F visto anteriormente. A Tabela 10 consta a ANOVA do modelo escolhido.

Tabela 10: Tabela ANOVA do modelo selecionado.

	GL	SQ	QM	Estatística F	P-valor
Tipo de Escola	1	749000	749000	173	<0,001
Renda	1	211000	211000	48,8	<0,001
Raça	1	25400	25400	5,85	0,0159
Sexo	1	38800	38800	8,96	0,0029
Escolaridade do pai	3	40000	13300	3,08	0,0273
Resíduos	492	2130000	4330		

Observe que todas as variáveis selecionadas para o modelo são significantes ao nível de $\alpha = 5\%$, com somas de quadrados (extra) relativamente altas, apesar deste ir diminuindo a cada variável acrescentada ao modelo, dado ao fato sequencial da tabela ANOVA, e parte das informações das variáveis selecionadas posteriormente já ter sido explicada pelas anteriores. Sendo assim, o modelo final foi o proposto pela Equação 1.

$$Y_i = 502,39 + 15,62X_{1i} + 53,30X_{2i} + 47,02X_{3i} - 14,63X_{4i} + 13,77X_{5i} + 25,30X_{6i} - 5,67X_{7i} + \varepsilon_i. \quad (1)$$

Onde $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$ é um erro aleatório. Ainda:

- Y: Nota média do candidato no ENEM 2019;
- X_1: Sexo Masculino;
- X_2: EM em Escola particular;
- X_3: Renda maior que R$ 4990,00;
- X_4: Raça (preto/pardo/indígena);
- X_5: Escolaridade do pai (EM);
- X_6: Escolaridade do pai (Superior); e
- X_7: Escolaridade do pai (Não sabe).

O modelo apresentou uma qualidade de ajuste relativamente baixa ($R^2 = 33,32\%$), no entanto, como o estudo é referente a uma área que envolve, dentre outros fatores, mas principalmente comportamento humano, é de se esperar que esse valor seja baixo, pois tais características são mais difíceis de se predizer com um simples modelo de regressão linear múltipla. Ademais, esse fato está relacionado à alta soma dos quadrados dos erros presente na Tabela 10 do modelo selecionado.

3.2 Diagnóstico do modelo

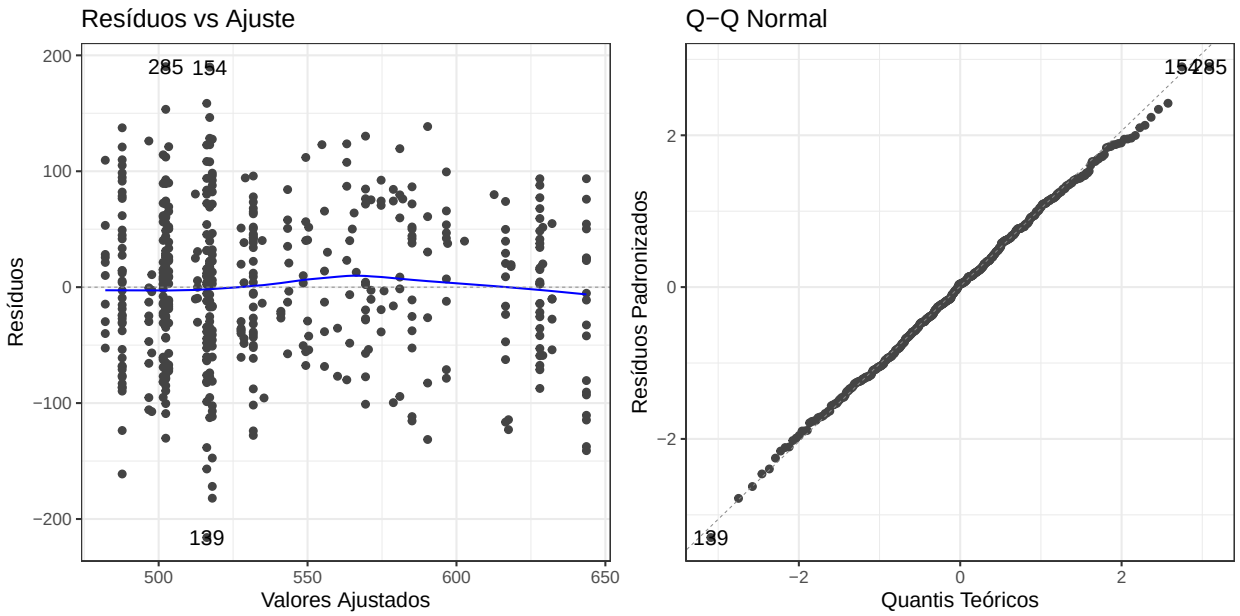

Figura 5: Gráficos referentes ao diagnóstico do modelo.

No que tange ao diagnóstico do modelo, para que se possa garantir a interpretação dos estimadores, deve-se conferir o comportamento dos resíduos, se esses seguem uma normalidade e se são homocesdásticos, i.e., a variância é constante, dada a suposição $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$. Logo, analisando o primeiro gráfico da Figura 5, nota-se que, de fato, os erros estão centrados em zero, no entanto, não é possível garantir a homocedasticidade apenas por esse gráfico, pois aparenta existir outliers próximo aos valores de 500 pontos na nota média ajustada. Sendo assim, foi analisado o maior resíduo semi-studentizado ($d_{máx}$), como $d_{máx} = 2{,}93 < 4$, não há outliers associado ao modelo selecionado em relação à variável resposta média.

Desse modo, recorrendo ao segundo gráfico da Figura 5, observa-se que os resíduos possuem uma normalidade bem condizente, então, para não se restringir apenas a aspectos visuais, foi realizado o teste de Shapiro-Wilk, obtendo o valor da estatística de 0,96, de modo que não há evidências para rejeitar a hipótese de normalidade dos erros. Para verificar a homocedasticidade, foi realizado o teste de Breusch-Pagan, cuja hipótese nula diz respeito se a variância do modelo é constante, sendo assim, o $p - valor$ associado à estatística desse teste foi de 0,71 e, portanto, não há evidências para rejeitar a hipótese de homocedasticidade.

4 Resultados

No modelo ajustado para a nota média no exame, conforme a Equação 1, tem-se que quando o candidato é de escola pública, com renda familiar de até R$ $998,00$ reais, pertencente ao grupo de raça branca/amarela, do sexo feminino e com pai cuja escolaridade é apenas o ensino fundamental, a nota média esperada é de $502,39$ pontos.

Agora, alterando-se uma variável e mantendo todas as outras inalteráveis, tem-se que, para o candidato que fez o Ensino Médio em um sistema de ensino particular, a nota tem um aumento de, em média, $53,30$ pontos. Ademais, tratando-se da variável renda, para candidatos que tem uma renda familiar maior do que R$ $4990,00$ reais, o desempenho aumenta, em média, $47,02$ pontos. Ainda, se o candidato pertencer ao grupo de raças preta/parda/indígena, a nota esperada diminui, em média, $14,63$ e, se o candidato for do sexo masculino, o desempenho médio esperado aumenta, em média, $15,62$ pontos.

Ainda, quando a escolaridade do pai for a nível de Ensino Médio, a nota do candidato aumenta, em média, $13,77$ e quando for de superior, o desempenho aumenta, em média, $25,30$ pontos. No entanto, quando o candidato desconhece a escolaridade do pai, seu desempenho reduz, em média, $5,67$ pontos.

5 Conclusão

Por fim, é evidente que o desempenho do aluno no ENEM de 2019 foi intrinsecamente associado a questões socioeconômicas que, infelizmente, permeiam até os dias atuais, principalmente quando se trata da desigualdade no contexto racial. Embora o coeficiente de determinação (R^2) tenha apresentado um baixo valor, as variáveis selecionadas para o modelo são pertinentes para a análise e estudo do caso. Obviamente que é muito difícil de se predizer exatamente a nota de um candidato, pois envolve inúmeras variáveis, entre elas variáveis latentes (como estresse e nervosismo, no momento do exame), e isso inflaciona a soma dos quadrados dos resíduos, diminuindo o valor da estatística R^2.

Referências

[1] https://blog.minitab.com/pt/analise-de-regressao-como-interpretar-o-r-quadrado-e-avaliar-a-qualidade-de-ajuste