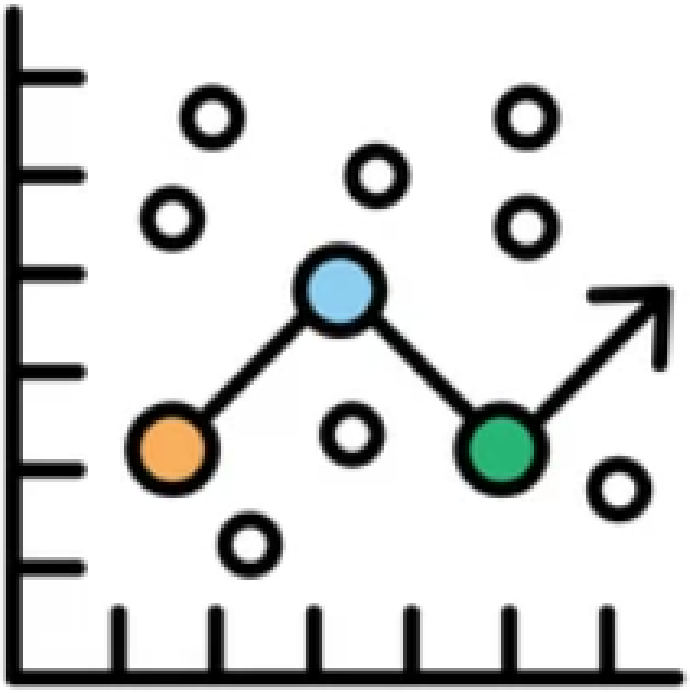
ME613 - Atividade 6

Andrei dos Santos Piscitelli RA:194181

30 junho, 2021



Descrição dos Dados

O conjunto de dados a ser utilizado é composto por uma amostra de participantes do ENEM com residência no estado do Ceará(CE). O conjunto é composto de 500 obervações com as seguintes informações a respeito de cada participante:“Sigla da Unidade da Federação de residência”,“Idade”(Idade do inscrito em 31/12/2019. Idades inferiores a 10 anos e superiores a 100 anos estão com o campo vazio na base.),“Sexo”, “Estado Civil”, “Cor/raça”,“Tipo de escola do Ensino Médio”,“Nota média nas provas”,“Escolaridade do pai ou homem responsável pelo participante”, “Escolaridade da mãe ou mulher responsável pelo participante”,“Número de moradores na residência atual”,“Renda mensal familiar” e “Acesso a internet na residência”. Após uma verificação foi constatado que não há presença de campos em vazio ou que foram preenchidos incorretamente, logo o conjunto de dados está completo. Segue abaixo uma descrição breve das estatísticas referentes as variáveis quantitativas.

Table 1: Medidas Sumárias das Variáveis Quantitativas

	Média	Desvio Padrão	Mediana	Valor Mínimo	Valor Máximo
Idade dos Participantes	18.2440	2.698709	18.00	16.00	56.00
Número de Moradores	4.2240	1.324134	4.00	1.00	10.00
Pontuação	496.1592	91.342994	488.41	257.78	779.48

A maior parte das variáveis do modelo são classificadas como categóricas sendo elas com respostas de 2 ou 4 níveis, ou seja, há 2 ou 4 tipos de resposta para cada variável. A seguir temos informações sumárias a respeito dessas variáveis.

Table 2: Medidas Sumárias das Variáveis Qualitativas com duas respostas

Gênero	Participantes
Feminino	263
Masculino	237

Estado Civil	Participantes
Solteiro(a)	493
Outros	7

Cor/Raça	Participantes
Branca/amarela	99
Preta/parda/indígena	401

Tipo de Escola do Ensino Médio	Participantes
Pública	451
Privada	49

Possuem Acesso a Internet	Participantes
Não	215
Sim	285

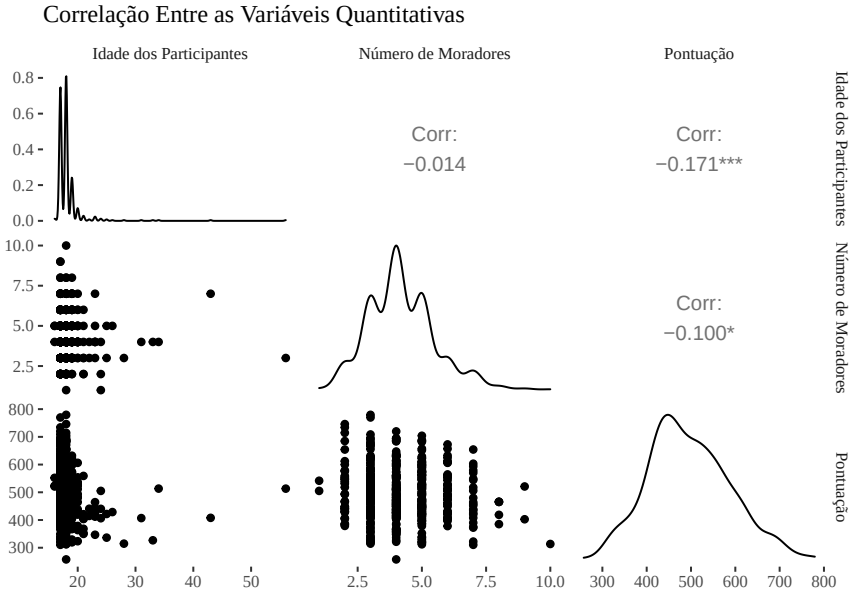
Table 3: Medidas Sumárias das Variáveis Qualitativas com quatro respostas

Escolaridade do Homem Responsável pelo(a) Participante	Participantes
Fundamental	282
Médio	88
Superior	27
Não sabe	103

Escolaridade da Mulher Responsável pelo(a) Participante	Participantes
Fundamental	286
Médio	115
Superior	39
Não sabe	60

Renda Mensal Familiar	Participantes
Até R$ 998,00	282
De R$ 998,01 até R$ 1.996,00	150
De R$ 1.996,01 até R$ R$ 4.990,00	44
Mais de R$ 4.990,00	24

Após analisar os dados acima podemos notar que quase todos os participantes são solteiros, tem por volta de 18 anos de idade e o número de homens e mulheres é parecido. O gráfico abaixo ilustra a correlação entre as variáveis quantitativas, ou seja, se há uma relação forte entre as variáveis.



Segundo o gráfico a menor correlação foi de 1,4% e a maior de 17,1% e todas as correlações tem caráter negativo. Uma vez que fizemos uma breve análise descritiva dos dados iremos, através de um

modelo de regressão, analisar de forma mais profunda a influência que as variáveis do conjunto de dados podem ter na nota dos participantes.

Modelagem Estatística

Para a análise será estimado, através do método de mínimos quadrados, um modelo de regressão linear múltipla. Em termos gerais o modelo é definido como:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1} + \epsilon_i \tag{1}$$

, onde:

- $\beta_0, \beta_1, ..., \beta_{p-1}$ são parâmetros.
- $X_{i1}, ..., X_{i,p-1}$ são constantes conhecidas.
- $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.
- $i = 1, 2, ..., n$.

Para a realização do modelo foi excluída a variável referente a União Federal uma vez que ela só tem um tipo de resposta. O modelo realizado com todas as variáveis produz o seguinte resultado:

Table 4: Coeficientes do Modelo

	x
Intercepto	592.0484067
Idade	-3.5265023
Sexo Masculino	-0.2883767
Estado Civil Outros	-16.8143174
Raça preta/parda/indígena	-31.8375912
Estudou em Escola Privada	72.8967313
O Pai cursou Ensino Médio	21.8932435
O pai cursou Ensino superior	51.4503751
Não sabe a escolaridade do pai	-21.4279281
A mãe cursou Ensino Médio	-17.2525322
A mãe cursou Ensino superior	10.5461319
Não sabe a escolaridade da mãe	-21.5044996
Número de Moradores na residência atual	-6.2176804
Renda mensal familiar de R$ 998,01 até R$ 1.996,00.	24.0500864
Renda mensal familiar de R$ 1.996,01 até R$ R$ 4.990,00.	38.0361129
Renda mensal familiar maior que R$ 4.990,00.	38.5097512
Residência possui acesso a internet	8.0205576

Para a avaliação do modelo serão realizados dois tipos de diagnósticos: O teste de Shapiro-Wilks, que busca analisar a normalidade dos resíduos e o teste de Breuch-Pagan que, supondo normalidade, avalia a heterocedasticidade. Para o primeiro modelo temos como resultado os seguintes dados:

Teste	p-valor	Resultado
Teste de Shapiro-Wilks	0.3850279	Sucesso
Teste de Breusch-Pagan	0.5751751	Sucesso

Segundo os testes temos que o modelo segue as condições de normalidade a também apresenta homocedasticidade, o que nos poupa do uso de ferramentas de correção como, por exemplo, transformação de variável. Serão analisados diferentes modelos com combinações distintas de variáveis preditoras para que seja selecionado um modelo mais adequado. Como o relatório está direcionado a pessoas sem profundo conhecimento na área de estatística iremos favoreçer modelos com menos variáveis pois eles são mais fáceis de serem interpretados, por tal motivo e pelo fato de o objetivo ser analisar a relação entre variáveis e não tentativas de predição será utilizado como critério de seleção o ***BIC***, abreviação para *"Bayesian information criterion"*. O critério ***BIC*** é mais exigente na sua avaliação, o que por consequência o faz favorecer modelos com menos variáveis. A seguinte tabela traz os 5 modelos com o menor número ***BIC***. Foi avaliado se todas as variáveis da cada modelo são significantes a nível de

5%.

BIC	Número de Variáveis Significantes a 5%
4365.091	Todas significantes
4366.920	Todas significantes
4367.307	Todas significantes
4369.249	1 não significante
4374.600	2 não significantes

Podemos notar que o primeiro modelo na tabela possui o menor valor ***BIC*** e todas as variáveis significantes a nível de 5%, o que o torna um forte candidato a ser o nosso modelo definitivo. Serão realizados os testes de Shapiro-Wilks e de Breuch-Pagan para verificar a sua aderência. Caso o modelo seja reprovado em um deles iremos analisar o modelo com o segundo menor valor ***BIC*** e que apresenta todas as variáveis significantes a nível de 5%, repetiremos o procedimento até achar o modelo mais adequado. Na tabela abaixo estão apresentados os resultados do modelo com o menor valor ***BIC***, tido como 4365.091, para os testes propostos.

Teste	p-valor	Resultado
Teste de Shapiro-Wilks	0.3150024	Sucesso
Teste de Breusch-Pagan	0.7688930	Sucesso

Como o modelo atendeu a todas as exigências propostas o usaremos como modelo definitivo para a nossa análise, o modelo possui R^2 no valor de 0,3104 e R^2 ajustado no valor de 0,302. A estatística F do modelo é de 36,98 com 6 e 493 graus de liberdade e p-valor menor que $2,2 \times 10^{-16}$. O modelo é definido da seguinte maneira:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 \tag{2}$$

as variáveis do modelo são:

- $X_1 = idade$
- $X_2 = \begin{cases} 0, & \text{caso participante seja de raça branca/amarela} \\ 1, & \text{caso participante seja de raça preta/parda/indígena} \end{cases}$
- $X_3 = \begin{cases} 0, & \text{caso participante tenha estudado em escola pública} \\ 1, & \text{caso participante tenha estudado em escola privada} \end{cases}$

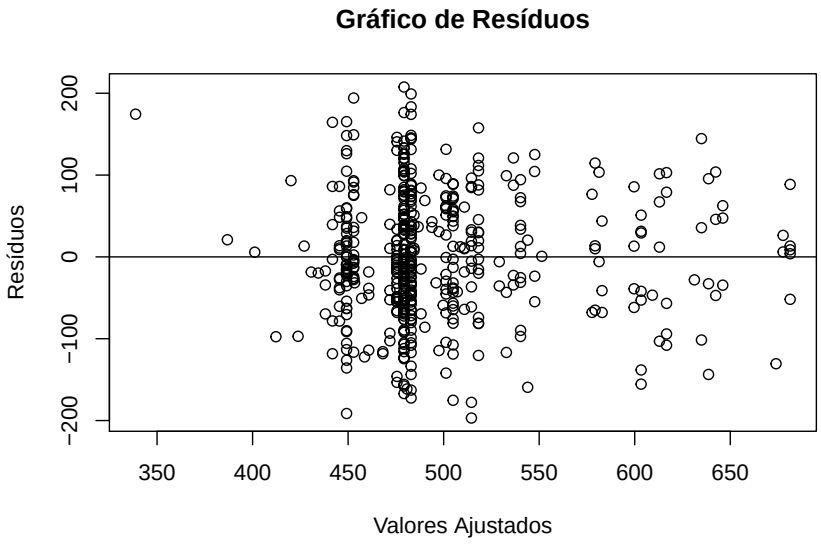
- $X_4 = \begin{cases} 0, & \text{caso contrário} \\ 1, & \text{caso escolaridade do pai ou homem responsável seja ensino médio} \end{cases}$
- $X_5 = \begin{cases} 0, & \text{caso contrário} \\ 1, & \text{caso escolaridade do pai ou homem responsável seja ensino superior} \end{cases}$
- $X_6 = \begin{cases} 0, & \text{caso contrário} \\ 1, & \text{caso não saiba inoformar a escolaridade do pai ou homem responsável} \end{cases}$

A seguir temos uma tabela com a informação de todos os coeficientes presentes no modelo selecionado:

Table 5: Coeficientes do Modelo Definitivo

	x
Intercepto	581.090334
Idade	-3.697061
Raça preta/parda/indígena	-35.281840
Estudou em Escola Privada	98.400158
O Pai cursou Ensino Médio	21.987352
O pai cursou Ensino superior	64.709937
Não sabe a escolaridade do pai	-30.064329

Uma vez analisados os elementos que compóem a equação, temos a seguir um gráfico de resíduos do modelo selecionado:



Interpetação e Conclusões

Ao interpetar o modelo percebemos que na amostra: pessoas de raça preta/parda/indígena costumam tirar menor pontuação na prova em comparação com pessoas de raça branca/amarela, pessoas oriundas de escolas da rede privada de ensino costumam ter pontuações maiores do que pessoas oriundas de escolas da rede pública, um maior nível de escolaridade do pai ou adulto responsável tem impacto positivo na nota do participante.Também nota-se que o avanço da idade tem um impacto negativo na nota do(a) estudante.

Supondo uma forte correlação entre etnia e raça pode-se concluir que ,entre os indivíduos da amostra, as minorias étnicas/raciais possuem um desempenho menor no ENEM quando comparado com o desempenho de outros povos. Participantes de camadas menos favorecidas socioeconomicamente também apresentam um desempenho menor quando comparado ao de concorrentes com melhores condições socioeconômicas.

Devido a sua abrangência e pelo fato de ser um processo seletivo para uma grande parte das instituições do ensino superior em território nacional o ENEM é tido como uma das maiores formas de avaliação educacional do país, fazendo com que os resultados do exame possibilitem, ainda, o desenvolvimento de estudos e indicadores educacionais. Através da analise dos dados aqui apresentados fica evidente na amostra uma relação entre condições socioeconômicas e a formação educacional. Assim como esperado, os membros da amostra considerados menos favorecidos pela sociedade, se comparados aos outros membros, se encontram em situações menos vantajosas quando se fala de educação.