**UNIVERSIDADE FEDERAL DE ALFENAS - UNIFAL-MG**

**THAIS DE PAULA ANDRADE**

# ANÁLISE SOCIOECONÔMICA DO DESEMPENHO DE PARTICIPANTES DO ENEM 2020 NO MUNICÍPIO DE VARGINHA-MG

**VARGINHA-MG**

**2023**

**THAIS DE PAULA ANDRADE**

# ANÁLISE SOCIOECONÔMICA DO DESEMPENHO DE PARTICIPANTES DO ENEM 2020 NO MUNICÍPIO DE VARGINHA-MG

Trabalho de conclusão de Piepex apresentado ao Instituto de Ciências Sociais Aplicadas da Universidade Federal de Alfenas como requisito parcial à obtenção do título de Bacharelado Interdisciplinar em Ciência e Economia.

Orientador: Profª Gislene Araujo Pereira

**VARGINHA-MG**

**2023**

## RESUMO

O Exame Nacional do Ensino Médio (Enem), aplicado anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), além de viabilizar o ingresso no Ensino Superior e atuar como base para programas de financiamento estudantil, tem entre seus objetivos avaliar o desempenho escolar geral dos alunos de escolas públicas e privadas no país, ao término do Ensino Médio. Neste contexto, estudos evidenciam a importância de analisar o impacto de diferentes fatores no desempenho dos estudantes, dado as diferentes condições da população brasileira. Logo, este trabalho visou identificar a relevância e mensurar os impactos de variáveis socioeconômicas no desempenho médio de participantes do Enem de escolas públicas e privadas da cidade de Varginha-MG, no ano de 2020. Para tanto, recorreu-se ao método de Regressão Linear Múltipla. Os resultados confirmaram a hipótese de que alunos de maiores rendas e estudantes de escolas privadas apresentam melhores desempenhos. Participantes do gênero masculino apresentam nota média maiores. Ademais, variável como raça não foram significativas na análise da nota média. A relevância deste estudo está na compreensão das variáveis de maior impacto sobre as notas do exame, embasando políticas e diretrizes para o contexto educacional do município.



**Palavras-chave:** Enem. Desempenho Educacional. Desigualdade Social. Regressão Linear.

**ABSTRACT**

The Exame Nacional do Ensino Médio (Enem), applied annually by the National Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), in addition to enabling entry into Higher Education and acting as a basis for student financing programs, has among its objectives to evaluate the general school performance of students from public and private schools in the country, at the end of high school. In this context, studies show the importance of analyzing the impact of different factors on student performance, given the different conditions of the Brazilian population. Therefore, we aim in this work to identify the relevance and measure the impacts of socioeconomic variables on the average performance of Enem participants from public and private schools in the city of Varginha-MG, in the year 2020. For this purpose, we used the Multiple Linear Regression method. The results confirmed the hypothesis that students with higher incomes and students from private school present better performances. Male participants have higher average scores. Furthermore, variables such as race were not significant in the analysis of the average grade. The relevance of this study lies in understanding the variables with the greatest impact on the exam scores, supporting policies and guidelines for the educational context of the municipality.

**Keywords**: Enem. Educational Performance. Social inequality. Linear Regression.

**SUMÁRIO**

# 1 INTRODUÇÃO

O Exame Nacional do Ensino Médio (Enem) foi criado para ser uma ferramenta de avaliação da qualidade do Ensino Médio no Brasil, em 1998. A princípio o Enem era utilizado pelo Ministério da Educação (MEC) para classificar as escolas, essa classificação tinha por objetivo embasar diretrizes governamentais para educação brasileira. (BRASIL, 2023a).

Com o desenvolvimento dessa ferramenta de avaliação, a partir do ano de 2009 o Enem passou a ter um papel mais relevante no âmbito educacional brasileiro de viabilizar e reduzir as barreiras de acesso ao Ensino Superior, ao ser utilizado como forma de ingresso às universidades públicas e privadas brasileiras, tornando-se o maior processo de seleção do país. Anteriormente, a forma de ingresso em uma universidade era restrita ao vestibular próprio das instituições de Ensino Superior e com o Enem, as condições formativas de ingresso e seleção passam a ser diversas.

Conforme informações do MEC, disponíveis no site do Gov.br, a partir da nota obtida no Enem há oportunidades tanto para o processo de seleção em cursos do Ensino Superior de universidades públicas quanto para universidades privadas. Para universidades públicas a seleção é feita por meio do Sistema de Seleção Unificada (Sisu). E para as universidades privadas a concessão de bolsas de estudo integral ou parcial em cursos de Graduação do Programa Universidade para Todos (Prouni), e para a introdução no Fundo de Financiamento Estudantil (Fies), que oferece financiamento do Ensino Superior (BRASIL, 2023b).

Embora a finalidade comum seja o ingresso no Ensino Superior, vários estudos anteriores, detalhados na seção 2, constataram que no Brasil ainda há a persistência de desigualdades sociais que excluem significativos contingentes populacionais.

A partir do atual modelo de avaliação para ingresso no Ensino Superior e o contexto das ocorrências das diferenças em âmbito socioeconômico, este trabalho tem como objetivo entender o impacto de determinados fatores no desempenho de participantes que realizaram o Enem no município de Varginha-MG. Para tanto, coletou-se uma amostra de alunos no ano de 2020, conforme disponibilizado pelos Microdados do Enem (2021). Nesse período, Varginha, uma cidade de 136 mil habitantes, contava com 13 escolas públicas, sendo uma federal e as demais estaduais, e 8 escolas privadas de ensino médio.

De acordo com informações de Censo Escolar nos municípios brasileiros, disponibilizados pelo portal QEdu, o percentual de alunos aprovados no Ensino Médio em Varginha, no ano de 2020, foi de 89,2% nas escolas públicas e 100% nas privadas. Com base nesses dados, e apesar da consolidação do Enem como critério de acesso ao Ensino Superior, pressupõem-se que o sistema público de ensino precisa de mais atenção, principalmente para estudantes de baixa renda. Mesmo com a presença de políticas públicas existentes desde 2012, que destina 50% das vagas para escolas públicas, com o intuito de incluírem no ensino superior estudantes de baixa renda terem aumentado, os dados mostram que políticas públicas voltadas para o ensino médio precisam ser aprimoradas e a qualidade do ensino melhorada.

Diante das considerações apresentadas, procurou-se analisar a significância e quantificar os impactos de alguns fatores condicionantes desse processo, recorrendo ao ajuste de um modelo de Regressão Linear Múltipla. A Regressão que por meio da especificação de um modelo matemático, estabelece as relações existentes entre uma variável dependente com uma ou mais covariáveis, a partir de $n$ observações destas variáveis.

A estrutura do trabalho contempla quatro seções. A primeira seção trata-se da introdução, seguida pelos trabalhos relacionados que faz uma recuperação de estudos anteriores que se aproximam da temático deste trabalho. Logo após, a terceira seção apresenta a metodologia usada para desenvolver o modelo de RLM, base do estudo, juntamente com os resultados encontrados pela estimação do modelo e a discussão acerca do assunto e, por fim, a última seção expõe as considerações finais.

# 2 TRABALHOS RELACIONADOS

Com relação às redes de Ensino Fundamental e Médio, Boneti e Oliveira (2017), ao analisarem o desempenho escolar nas edições de 2009 a 2013, auferiram que escolas que ocupam as últimas posições no ranking do exame são instituições públicas estaduais. Foi também constatado pelos autores, que em todas as edições analisadas não houve nenhuma instituição privada e federal entre as últimas posições.

Pode-se verificar também os impactos no desempenho de alunos no exame no que engloba a má distribuição de renda e acesso diferenciado aos recursos educacionais, nas diferentes regiões e classes sociais. O estudo de desempenho dos alunos de escola pública de 2012 a 2018, de Justiniano e Queiroz (2021) demonstraram que a renda familiar apresentou maior correlação com o as notas do exame. De modo geral, as menores proficiências do Brasil ocorreram nas regiões Norte e Nordeste do país, locais em que a renda *per capita* é frequentemente menor (até dois salários-mínimos).

Na análise realizada por Carvalho (2022), o impacto da pandemia do COVID-19 sobre o desempenho médio dos participantes do Enem no ano de 2020 mostra que os resultados dos estudantes do terceiro ano do Ensino Médio têm um impacto negativo nas notas do exame, se comparado aos resultados auferidos no ano de 2019, principalmente devido à falta de acesso de muitos estudantes de escola pública aos recursos necessários para assistirem as aulas online. Com relação ao tipo de escola, se pública ou privada, constatou-se que os alunos de escola pública, tem desempenho inferior aos de escola privada. Ademais, também foi observado que alunos autodeclarados de cor “Branca” obtiveram o melhor desempenho em relação aos demais. Em relação a situação socioeconômica, o estudo também mostra uma relação direta entre o nível de renda e o desempenho do participante. A respeito da Localização da Escola (Urbana ou Rural), constatou-se que os participantes advindos de escolas localizadas na zona urbana tiveram desempenho melhor.

Os resultados como os citados acima foram encontrados também nas análises de Lucena e dos Santos (2020), em que os autores constataram que os candidatos das escolas particulares têm melhor desempenho na nota geral que os demais, sendo possível observar que bolsistas das escolas particulares têm melhor desempenho que os não bolsistas. Além disso, os resultados evidenciaram que os candidatos que possuem

pai e mãe com maiores escolaridades também demonstram ter um melhor desempenho que os demais participantes.

# 3 FUNDAMENTAÇÃO TEÓRICA

## 3.1 ANÁLISE DE REGRESSÃO

A Análise de Regressão é um método estatístico baseado na construção de um modelo no qual se determina o relacionamento entre duas ou mais variáveis, onde uma variável pode ser predita a partir de uma ou mais.

Ao criar um Modelo de Regressão Linear (MRL), ajusta-se uma equação linear com base em uma amostra de dados. A partir dessa amostra, é possível ajustar um modelo de regressão linear simples, onde busca estabelecer uma relação linear direta entre uma variável independente (Y) e uma variável dependente ($X_1$). Esse modelo analisa como uma mudança na variável independente influencia a variável dependente.

Outra opção é ajustar um modelo de regressão linear múltipla, que de acordo com Tabachnick e Fidel (1996), considera múltiplos fatores simultaneamente ao prever a variável dependente. Cada coeficiente angular ($\beta_i, i = 1, \dots, p$) representa a mudança média na variável dependente (Y) quando a variável independente ($X_i, i = 1, \dots, p$) correspondente é aumentada em uma unidade, mantendo todas as outras variáveis constantes.

De acordo com Greene (2002), o relacionamento entre uma ou mais variáveis tem forma geral dada por:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{1}$$

em que,

- $\boldsymbol{y}_{n\times 1}$ é o vetor matriz que contém as observações da variável dependente (resposta);
- $\boldsymbol{X}_{n\times(p+1)}$ é a matriz contendo uma coluna com 1's e as p variáveis independentes;
- $\boldsymbol{\epsilon}_{n\times 1}$ é o vetor dos erros aleatórios;
- $\boldsymbol{\beta}_{(p+1)\times 1}$ é o vetor dos coeficientes desconhecidos do modelo que devem ser estimados.

Em notação matricial, tem-se que:

$$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{(n\times 1)} \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix}_{(n\times p)} \quad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{(p\times 1)} \quad \text{e } \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{(n\times 1)}$$

O termo aleatório é composto pelo vetor referente a variável $\boldsymbol{\epsilon}$ que contém os resíduos ($e_i$), que diz respeito a diferença de cada observação da variável dependente observada (**y**) e seu valor ajustado( $\widehat{\boldsymbol{y}}$) . Para Tabachnick e Fidel (1996), os resíduos são uma importante forma de medir os erros do modelo, e sua notação é dada da seguinte maneira:

$$\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

A variável $\boldsymbol{\epsilon}$ é aleatória independente e identicamente distribuída com uma distribuição de probabilidade normal em que:

$$E(\boldsymbol{\epsilon}) = \mathbf{0}$$

$V(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$ ; em que $\boldsymbol{I}$ é a matriz identidade

$$Cov_{(\epsilon i,\epsilon j)} = 0; i \neq j$$

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I}).$$

A parte determinística é composta por $\boldsymbol{X\beta}$. Deste modo tem-se que a variável **Y** é aleatória e normalmente distribuída em que, para dado um valor de **X**, o valor esperado de **Y** será dado por $E(\boldsymbol{Y/X}) = \boldsymbol{X\beta}$ e $Var(\boldsymbol{Y/X}) = \sigma^2\boldsymbol{I}$. Logo $\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$.

### 3.1.1 Ajuste de um MRL – O Método dos Mínimos Quadrados

O Método dos Mínimos Quadrados (MQM) tem como objetivo minimizar a soma de quadrados do erro da predição, ou seja, a quantidade total de erro de predição deve ser tão pequena quanto possível, dando a melhor linha matematicamente alcançável do conjunto de pontos dispostos num gráfico de dispersão. (Tabachnick e Fidel 1996, p.167).

O método é realizado minimizando a soma de quadrados do erro (SQE). Para isso,

$$SQE = \boldsymbol{\epsilon}'\boldsymbol{\epsilon}$$

$$SQE = (\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta})$$

$$SQE = \boldsymbol{Y'Y} - \boldsymbol{\beta' X'Y}$$

Então os estimadores de mínimos quadrados (EMQ) do modelo serão encontrados derivando a matriz que contém a SQE em relação a $\boldsymbol{\beta}$ e igualando a zero:

$$\frac{\partial \boldsymbol{\epsilon'\epsilon}}{\partial \boldsymbol{\beta}} = 0 \Rightarrow \widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$$

Logo, os valores estimados para Y em um MRLS ou MRLM será:

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$$

### 3.1.2 Propriedades dos Estimadores de Quadrados Mínimos (EQM)

Os coeficientes de regressão são:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$$

De acordo com Ye (2010), esses estimadores dos coeficientes de regressão são não viciados, logo:

$$E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

A variância de cada estimador é obtida através dos elementos da diagonal principal da matriz que é denominada de matriz de variâncias-covariâncias:

$$Var(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X'X})^{-1}$$

Deste modo, se os elementos da diagonal principal da matriz representam a variância dos coeficientes de regressão, os demais elementos da matriz representam a covariância que é dada por:

$$Cov(\hat{\beta}_i, \hat{\beta}_j) = C_{i,j}\sigma^2; i \neq j$$

Dessa forma $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X'X})^{-1})$. A normalidade decorre do fato de $\boldsymbol{\beta}$ ser uma função linear do vetor **Y**, cuja distribuição é normal.

### 3.1.3 O estimador de quadrados mínimos de $\sigma^2$

Segundo Ye (2010, p.292), para obter as estimativas da variância e, e consequentemente os erros-padrão desses estimadores, substituindo-se $\sigma^2$ pela estimativa apropriada obtida por meio dos dados experimentais.

Desta forma, para a equação de regressão linear, uma estimativa não viciada de $\sigma^2$é dada pelo quadrado médio do erro, ou seja:

$$\widehat{\sigma^2} = SQE/(n-p-1)$$

E, portanto, a estimativa do erro-padrão será:

$$\hat{\sigma} = \sqrt[2]{\widehat{\sigma^2}}$$

Sincich (2011), diz que o estimador do desvio total do modelo, erro-padrão, reflete o desvio em torno da reta de regressão.

## 3.2 QUALIDADE DO MODELO E ANÁLISE DE RESÍDUOS

Testes estatísticos de adequabilidade do modelo apresentados por Ye (2010) são feitos para diagnosticar a qualidade do modelo. Para Sincich (2011), essas técnicas são utilizadas para verificar a validade do modelo, ou seja, o nível de confiabilidade dos estimados de mínimos quadrados e a utilidade da reta de regressão.

### 3.2.1 Análise de Variância (ANOVA)

A análise de variância é utilizada para avaliar a qualidade total do modelo, se as variáveis independentes ($X_i, i = 1, \dots, p$) contribuem de forma expressiva para explicar as variações da variável dependente (Y). Ye (2010), define a análise de variância como um procedimento por meio do qual a variação total na variável dependente é subdividida em variação da regressão e variação do erro.

A decomposição da variação total (SQT) é dada pela soma de quadrados da regressão (SQR) que representa a variação nos valores y explicados pelo modelo e pela soma de quadrados do erro (SQE) que, como apresentado anteriormente, representa a soma dos quadrados das diferenças entre os valores de Y e $\hat{Y}$. Assim:

$$SQT = SQR + SQE$$

Esses valores obtidos são importantes para calcular a estatística do teste, baseada na distribuição F- Senedecor, e são dispostos numa tabela de ANOVA.

QUADRO1: Análise de Variância (ANOVA)

| Fonte de Variação (FV) | Graus de Liberdade (GL) | Soma de Quadrados (SQ) | Quadrado Médio (QM) | Estatística de teste -$F_o$ |
|---|---|---|---|---|
| Regressão (Reg) | p | $SQReg = \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \frac{(\sum y_i)^2}{n}$ | QMREG=SQReg/p | $F_o$=QMReg/QME |
| Erro (E) | n-p-1 | $SQE = \boldsymbol{y}'\boldsymbol{y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}$ | QME=SQE/ (n-p-1)= $\widehat{\sigma^2}$ | |
| Total (T) | n-1 | $SQT = \boldsymbol{y}'\boldsymbol{y} - n\acute{y}^2$ | | |

Fonte: Wooldridge (2014)

As hipóteses do teste que envolve essa análise, a um nível de significância α, serão:

$$H_0: \hat{\beta}_1 = \hat{\beta}_2 = \cdots = \hat{\beta}_j = 0$$

$$H_1: algum\ \hat{\beta}_j \neq 0$$

A hipótese inicial do teste será rejeitada se o valor da estatística do teste ($F_o$) for maior que o valor crítico $F_{(\alpha;p,n-p-1)}$, ou se o $p - valor$ for menor que o $\alpha$ estipulado. De acordo com Ye (2010,p.264), conclui-se com esse resultado que há uma quantidade significativa da variação na resposta que é explicada pelo modelo postulado.

### 3.2.2 Teste T

O teste de significância é utilizado para se determinar se há uma relação linear entre a função resposta Y e as variáveis independentes $x_j$, em que um teste de hipóteses baseado na distribuição t-student é realizado definindo assim, se a estrutura do modelo realmente representa um MRLS.

A hipótese inicial a ser testada, a um nível de significância α, será de que determinado coeficiente de regressão será igual a zero. Para Ye (2010), o teste t usado na regressão múltipla avalia assim a importância de cada coeficiente. Portanto se um coeficiente $\hat{\beta}$ não for significante a um nível de significância assumido, hipótese $H_0$ não é rejeitada, conclui-se também que a variável independente $x_j$ não é significativa para explicar as variações da variável dependente Y.

As hipóteses testadas serão que:

$$H_0: \hat{\beta}_j = 0$$

$$H_1: \hat{\beta}_j \neq 0$$

A estatística do teste é dada por:

$$t_j = \hat{\beta}_j / \sqrt{\widehat{\sigma^2} C_{j+1,j+1}}$$

em que, $C_{j+1,j+1}$ é o (j+1)-ésimo elemento da diagonal principal de$(X'X)^{-1}$, $j = 1,2, \dots, p$.

Para a hipótese inicial ser rejeitada a estatística do teste deve ser maior que o valor crítico calculado $t_{(\alpha/2),n-p-1}$, ou o $valor - p$ ser menor que o $\alpha$estipulado. Com isso, conclui-se que há uma relação linear significativa entre a variável dependente Y com a variável independente $x_j$.

### 3.2.3 O Coeficiente de Determinação (R²)

Uma medida de associação que mede, de forma quantitativa, a qualidade de um modelo de regressão segundo Navid (2012) é o coeficiente de determinação (R²) e o coeficiente de determinação ajustado ($R^2{}_{ajust}$). Eles determinam a proporção que a variável independente contribui para prever as variações na variável dependente, isto é, quanto da variabilidade de Y pode ser explicada pelas variáveis regressoras.

Navid (2012), ressalta que a diferença entre os dois é o fato de que o coeficiente de determinação ajustado dá uma melhor ideia da proporção de variação de Y explicada pelo modelo de regressão uma vez que tem em conta o número de variáveis independentes. Isso ocorre porque o coeficiente de determinação múltiplo aumenta

sempre, quando uma nova variável é adicionada ao modelo e o $R^2{}_{ajust}$ só aumenta se de alguma maneira houver vantagem na adição de uma nova variável.

Quando a diferença entre os dois é acentuada, há uma boa hipótese de que tenham sido incluídos no modelo termos estatisticamente não significativos.

O coeficiente de determinação múltiplo é dado por:

$$R^2 = SQReg/SQT$$

Já o ajustado por:

$R^2{}_{ajust.}$= 1- $\frac{(SQE/n-p)}{(SQT/n-1)}$

Quanto mais próximo o coeficiente de determinação múltiplo e o ajustado estiverem de 1, maior é a explicação da variável resposta pelo modelo.

## 3.3 ANÁLISE DE RESÍDUOS

Tanto na regressão linear simples quanto na regressão múltipla, as suposições do modelo ajustado precisam ser validadas para que os resultados sejam confiáveis. Assim para que o modelo de regressão seja apropriado, os resíduos devem apresentar as seguintes pressuposições:

$$\boldsymbol{Y = X\beta + \epsilon}$$

em que, $\boldsymbol{\epsilon} = (e_1, e_2, e_3, \ldots, e_n)$, e:

- $e_i, e_j$ são independentes $(i \neq j)$;
- $Var(e_i) = \sigma^2$ (constante);
- $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$ (normalidade);
- O modelo é linear;
- Não existir pontos discrepantes que possam ser influentes;
- Ausência de multicolinearidade entre as variáveis independentes.

A análise das pressuposições dos resíduos pode ser feita graficamente ou por meio de testes de diagnósticos.

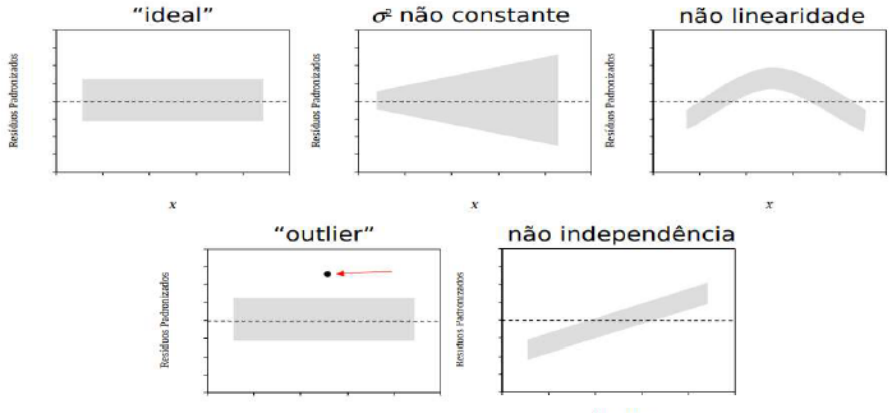
### 3.3.1 Análise Gráfica dos Resíduos

Segundo Hair et al (2009, p. 174), a representação gráfica dos resíduos versus as variáveis independentes, ou preditas, é um método básico para identificar violações das suposições para a relação geral. Assim, a partir da observação de padrões específicos dos resíduos, ou seja, a forma em que os pontos estão dispostos no gráfico as suposições podem ser identificadas.

De acordo com Charnet et al (2008), a análise gráfica dos resíduos pode ser realizada através de recursos computacionais, com auxílio de softwares.

A Figura 1 ilustras vários padrões e as possíveis implicações no gráfico dos Resíduos Studentizados versus valores ajustado.

Figura 1 – Diversos padrões e as possíveis implicações no gráfico dos Resíduos Studentizados versus valores ajustado




Fonte: Elaboração pelo autor.

Na Regressão Múltipla, além das suposições listadas acima, é necessário diagnosticar se há correlação e multicolinearidade entre as variáveis independentes para não afetar os resultados, causando inferências errôneas ou pouco confiáveis.

### 3.3.2 Testes de Diagnósticos

Os testes de diagnóstico são ferramentas importantes para verificar se as suposições dos resíduos em uma análise de regressão linear estão sendo atendidas. Segundo Wooldridge (2014), Os principais testes de diagnósticos utilizados para verificar as pressuposições dos resíduos em uma regressão linear são: Teste Shapiro-Wilk, Teste Breusch-Pagan e Teste de Durbin Watson**.**

1. **Teste Shapiro-WILK:** O Teste de Shapiro-Wilk tem como objetivo avaliar se uma distribuição é semelhante a uma distribuição normal, as hipóteses estatísticas testadas são:

$$\begin{cases} H_0: & Os\ resíduos\ seguem\ uma\ distribuição\ Normal \\ H_a: & Os\ resíduos\ não\ seguem\ uma\ distribuição\ Normal \end{cases}$$

2. **Teste Breusch-Pagan**: Avalia se a variância dos resíduos é constante. Um p-valor baixo sugere heterocedasticidade, as hipóteses estatísticas testadas são:

$$\begin{cases} H_0: & A\ variância\ dos\ resíduos\ é\ constante \\ H_a: & A\ variância\ dos\ \ resíduos\ não\ é\ constante \end{cases}$$

3. **Teste de Durbin Watson:** Testa a autocorrelação dos resíduos em diferentes defasagens, as hipóteses estatísticas testadas são:

$$\begin{cases} H_0: & Resíduos\ não\ são\ autocorrelacionados \\ H_a: & Resíduos\ \ são\ autocorrelacionados \end{cases}$$

## 3.4 MULTICOLINEARIDADE

Em um MRLM, pode ocorrer que as variáveis independentes não sejam apenas correlacionadas com a variável dependente em questão, mas correlacionadas entre si o que dificulta a precisão da influência das variáveis X sobre a variável Y. Se isso ocorre, pode-se falar que as variáveis independentes são multicolineares.

Conforme Hair et al (2009, pág. 190), cada variável independente se torna uma variável dependente e é regredida relativamente as demais variáveis independentes.

Assim, a multicolinearidade surge quando existe certo grau de dependência entre as variáveis regressoras (X), o que faz com que a variação total dos dados seja compartilhada entre todas as variáveis do modelo, afetando assim a capacidade de prever a medida dependente e, consequentemente, a análise dos papéis relativos que cada variável independente exerce no modelo.

Quando o objetivo da regressão é ajustar um modelo significativo que explique as variações de Y em relação a variação das variáveis X, essa inter-relação entre variáveis independentes se torna um problema, pois, segundo Hair et al (2009) à medida que a multicolinearidade aumenta, seus efeitos ficam mais visíveis, em termos de estimação e explicação, onde a variância total explicada pelo modelo tende a diminuir, bem como a variância especifica explicada pelos coeficientes de regressão, referente a cada variável independente, dificultando a avaliação dos efeitos isolados das variáveis independentes sobre a variável dependente.

Para Tabachnick e Fidel (1996), quando a correlação entre as variáveis independentes não são nulas, detecta-se a existência desse problema. Logo a identificação da existência de multicolinearidade pode ser realizada com uma simples análise da matriz de correlação entre as variáveis independentes, medidas de -1 a 1 em que quanto mais próximo de -1 ou 1, maior o grau do problema, e consequentemente, mais difícil identificar os efeitos únicos das variáveis independentes podendo chegar ao ponto de nenhum dos coeficientes de regressão ser estatisticamente significante.

Outra forma de identificar a multicolinearidade é através do uso de medidas, que expresse a intensidade que cada variável independente é explicada pelas outras variáveis independentes contidas no modelo. Duas medidas normalmente usadas para detecção de multicolinearidade são recomendadas por Hair et al (2009, p.190). A primeira delas é a medida de tolerância, definida como $1-R_i^2$, onde $R_i^2$, é o coeficiente de determinação que indica quanto uma variável independente "i" é explicada pelas outras variáveis independentes. Os valores de tolerância próximos de zero indicam que a variável é altamente predita pelas outras variáveis independentes e, por isso, há multicolinearidade. A outra medida é o VIF (Variance Inflation Factor) que é o inverso da tolerância. Valores altos de VIF indicam alto grau de multicolinearidade.

### 3.4.1 Métodos de Seleção de variáveis

Após a identificação da multicolinearidade, é possível utilizar técnicas de regressão que selecione, um subconjunto de variáveis independentes que melhor explique a variável dependente. Essas técnicas, segundo Costa (2011, pág. 61), têm como base um algoritmo que checa a importância das variáveis, incluindo ou excluindo-as do modelo se baseando em uma regra de decisão. A importância da variável é definida em termos de uma medida de significância estatística do coeficiente associado à variável para o modelo.

A seleção de variáveis independentes de acordo com Charnet et al (2008), é realizada normalmente com o uso de recursos computacionais, e tem como intuito selecionar o melhor modelo, sequencialmente, através da adição ou remoção de variáveis independentes em cada passo. Com essa finalidade o AIC e outros critérios, são usados para escolha do modelo e os procedimentos automáticos, como Backward e Stepwise são utilizados na seleção de variáveis que farão parte do modelo.

De acordo com Tabachnick e Fidel (1996) o procedimento Stepwise considera no processo de seleção apenas uma variável por vez, e decide se a elimina ou a adiciona no modelo, o método Forward começa sem nenhuma variável no modelo e adiciona variáveis a cada passo já no Backward se insere inicialmente no processo de seleção todas as variáveis e depois, por etapas, determina se remove ou não cada uma do modelo inicial determinado.

As etapas do procedimento Backward são apresentadas da seguinte maneira:

1. Inicialmente, o procedimento considera o modelo com todas as $p$ variáveis;
2. A decisão de retirada da variável é tomada baseando-se no valor-p que ela apresenta (valores acima são 0,15 excluídos do modelo);
3. Ajusta-se novamente o modelo, agora com as $p - 1$ variáveis;
4. Repetir o passo 2 até que nenhuma variável seja mais excluída. O algoritmo de eliminação termina quando não há nenhuma variável com valor p maior que 0,15.

## 3.5 MÉTODOLOGIA

Esta pesquisa foi realizada com informações públicas extraídas dos Microdados do Enem (2021), disponibilizas pelo (Inep), que reúnem um conjunto de informações detalhadas sobre os exames e avaliações da educação básica.

Os dados utilizados referem-se aos registros individuais dos participantes advindos de escolas públicas e privadas, que estavam cursando o Ensino Médio no ano de 2020, e que realizaram as provas no município de Varginha - MG. Foi obtida uma amostra final de 211 participantes, na qual foi considera apenas os participantes que tinham todas as informações completas. A título de exemplo não foi considerado participante que apresentavam pelo menos um item de nota em branco, dado que a falta de informações de notas implica que o participante não compareceu para realização da respectiva prova. Ademais, uma categorização de variáveis foi realizada sendo desconsiderados os registros em que não obtiveram respostas. A metodologia foi realizada por meio do *software* R (R CORE TEAM, 2021).

O método adotado neste estudo foi um Modelo de Regressão Linear Múltipla (MRLM). O ajuste dos coeficientes do modelo é realizado com o Método dos Quadrados Mínimos (MQM), que permite a obtenção de estimadores para determinar as relações e o impacto de relevantes fatores socioeconômicos associados ao desempenho no Enem auferido no período analisado no município de Varginha -MG.

A variável dependente ($Y$) utilizada no modelo é Nota Geral, obtida a partir da média auferida nas provas objetiva de Ciências da Natureza, de Ciências Humanas, Linguagens e Códigos, Matemática e na Redação. As covariáveis consideradas foram: Gênero, Faixa Etária, Tipo de Escola, Raça e Classes de Renda Familiar. Todas as covariáveis são categóricas, e assim foi criado variáveis Dummies para inclui-las no modelo.

Para as faixas de renda familiar a classificação por classes é dada com base no Instituto Brasileiro de Geografia e Estatística (IBGE) e nas respostas do questionário do Enem: classe A referente renda mensal domiciliar superior a R$ 22 mil, classe B referente a renda mensal domiciliar entre R$ 7,1 mil e R$ 22 mil, classe C referente a renda mensal domiciliar entre R$ 1,567 mil e R$ 7,1 mil e classes D/E renda mensal domiciliar de até R$ 1,045 mil. Para a variável raça temos as classificações de Branca, Preta, Parda e Amarela. Ademais, constituem variáveis dummies binárias. O Quadro 2

apresenta a descrição das variáveis. Antes do ajuste dos modelos, foi realizada também uma análise descritiva das variáveis estudadas.

**Quadro2** - Descrição das Variáveis Utilizadas.

| **Nome da Variável** | **Descrição da Variável** |
|---|---|
| Nota da Prova | Média aritmética das notas referentes as provas realizadas. |
| Gênero | Variável *dummie* para o gênero. 1 para o gênero masculino e 0 para o gênero feminino. |
| Faixa Etária | Variável *dummie* para faixa etária de idade. 1 para o participante menor de 18 anos e 0 caso contrário. |
| Tipo de Escola | Variável *dummie* para escola pública. 1 para o participante que cursou ensino médio em escola privada e 0 em escola pública. |
| Raça | Variável *dummie* para raça. Categorias definidas como Branca, Preta, Parda e Amarela, sendo o grupo base dado pela raça Branca. |
| Classe Renda Familiar | Variável *dummie* para renda familiar. Categorias definidas como A, B, C e D/E, sendo o grupo base dado pela classe A. |

Fonte: Elaboração pelo autor.

Para avaliar o desempenho no Enem dentre os participantes do município de Varginha -MG, o MRLM foi definido com variáveis relevantes para explicar o padrão de nota auferido pelos participantes a partir de uma equação linear é determinada da seguinte forma:

$$\boldsymbol{Y = X\beta + \epsilon}$$

Em que $\boldsymbol{\epsilon} \sim N(\mathbf{0}; \sigma^2 \boldsymbol{I})$ é o vetor de erros aleatórios, Y é o vetor que contém as observações das notas dos participantes, X é a matriz contendo as covariáveis. O vetor de $k$ coeficientes de regressão é dado por $\boldsymbol{\beta}$, sendo $\beta_0$ a média da nota quando todas as demais covariáveis forem zero e os demais $\beta$'s são os coeficientes associados as covariáveis do estudo, que correspondem à diferença na média das notas considerando fixos os valores das demais covariáveis.

O modelo inicial deste estudo será dado por:

$$\begin{aligned} nota = \beta_0 + \beta_1 masculino + \beta_2 menor18 + \beta_3 escolaprivada + \beta_4 preta \\ + \beta_5 parda + \beta_6 amarela + \beta_7 classeB + \beta_8 classeC \\ + \beta_9 classeD/E + \epsilon, \end{aligned}$$

Sendo as variáveis descritas conforme o Quadro 1. A estimação dos coeficientes de regressão, realizada via Mínimos Quadrados Ordinários, é dada por $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$.

Após o ajuste do modelo, foi testado individualmente a significância estatística do conjunto das variáveis independentes sobre a dependente. De acordo com Wooldridge (2014), para testar a significância individual de determinada covariável $X_j$, com $j = 1, \ldots, k$ as hipóteses são as seguintes:

$$H_0: \hat{\beta}_j = 0$$
$$H_a: \hat{\beta}_j \neq 0$$

Com estatística do teste é calculada como $t_{\beta_j} = \hat{\beta}_j / \sqrt{\hat{\sigma}^2 C_{j+1,j+1}}$ , sendo $C_{j+1,j+1}$ é o (j+1)-ésimo elemento da diagonal principal de$(\boldsymbol{X'X})^{-1}$, Se $t_{\beta_j} \geq t_{,(\alpha/2),n-k-1}$ ou se o $valor - p$ calculado for menor ou igual ao valor de $\alpha$ estabelecido, o teste é significativo.

Posteriormente, com o objetivo de obter uma métrica para a qualidade do ajuste do modelo, foi calculado o coeficiente de determinação ajustado, que representa a proporção da variabilidade das notas dos participantes explicada pelos fatores socioeconômicos considerados, e é dado por:

$$R^2{}_a = 1 - \frac{(SQE/n-k)}{(SQT/n-1)}$$

A multicolinearidade foi verificada com o cálculo do VIF (*Variance Inflation Factor*), em que se o VIF for menor que 10 não há multicolinearidade entre os fatores, mas se o VIF for maior que 10, as covariáveis podem estar altamente correlacionadas (WOOLDRIDGE, 2014).

Ainda, foi realizado um diagnóstico sobre os resíduos do modelo ajustado para verificar se as pressuposições iniciais referentes ao erro estão sendo respeitadas: independência, homocedasticidade, normalidade, linearidade e se há observações influentes que afetam o modelo, como apresentado em Wooldridge (2014).

# 4 RESULTADOS E DISCUSSÕES

## 4.1 ANÁLISE DESCRITIVA

Os resultados da análise descritiva referente a variável resposta, indica que a nota média dos participantes do Enem no município de Varginha-MG, no período analisado, foi de 583 pontos com um desvio-padrão de 86 pontos; já a nota mínima auferida foi 380 pontos e a nota máxima de 776. Na Tabela 1 são apresentadas essas estatísticas descritivas para a nota média dos participantes segundo as covariáveis consideradas.

Tabela 1 - Estatísticas Descritivas para a variável Nota Média segundo covariáveis.

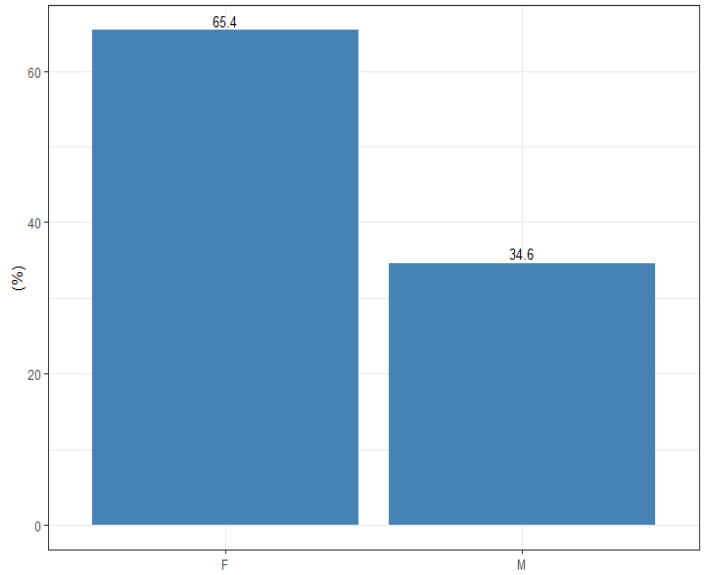
| **Covariáveis** | | **Média** | **Desvio Padrão** | **Mínimo** | **Máximo** |
|---|---|---|---|---|---|
| Gênero | Masculino | 605,4 | 84,8 | 446,1 | 776,0 |
| | Feminino | 571,7 | 84,9 | 379,7 | 750,1 |
| Faixa Etária | Menores que 18 anos | 584,6 | 89,7 | 425,4 | 755,8 |
| | 18 anos ou mais | 582,3 | 83,4 | 379,7 | 776,0 |
| Raça | Branca | 595,7 | 88,6 | 404,4 | 776,0 |
| | Preta | 549,6 | 90,5 | 401,4 | 735,8 |
| | Parda | 567,1 | 74,7 | 379,7 | 710,5 |
| | Amarela | 573,2 | 52,5 | 536,1 | 610,3 |
| Classe Renda Familiar | A | 711,0 | 67,0 | 583,8 | 776,0 |
| | B | 621,1 | 75,4 | 425,4 | 735,8 |
| | C | 572,1 | 81,5 | 379,7 | 755,8 |
| | D/E | 519,4 | 65,9 | 401,4 | 625,2 |
| Tipo Escola | Pública | 563,3 | 81,4 | 379,7 | 755,8 |
| | Privada | 649,6 | 66,4 | 498,9 | 776,0 |

Fonte: Elaboração pelo autor.

Conforme pode-se observar nas Figuras 2 a 6, a maior parte dos participantes são indivíduos do gênero feminino, que totalizam 65,4% dos participantes totais da amostra do ano de 2020 do município de Varginha-MG. Aproximadamente 45,5% dos participantes são menores que 18 anos de idade e 76,8% concluíram o Ensino Médio em escolas públicas. Em relação à raça, aufere-se na amostra que 62,6% dos participantes do Enem no município de Varginha-MG se declaram Brancos, 9,5% Pretos, 27,0% Pardos e 0,9% Amarelos. As faixas de renda dos participantes do Enem, indicam que a

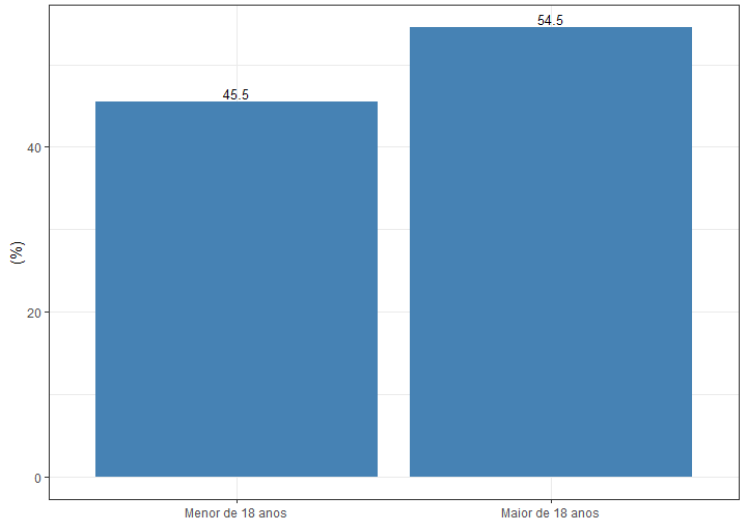
renda familiar mais usual auferida na amostra do município de Varginha-MG foi de R$ 1.045,01 até R$ 2.090,50. Analisando pela classificação estabelecida, a figura abaixo apresenta as faixas de distribuição de renda dos participantes do Enem no município de Varginha-MG, no período analisado.

Figura 2 - Distribuição de participantes do Enem no município de Varginha-MG no ano de 2020 por Gênero.



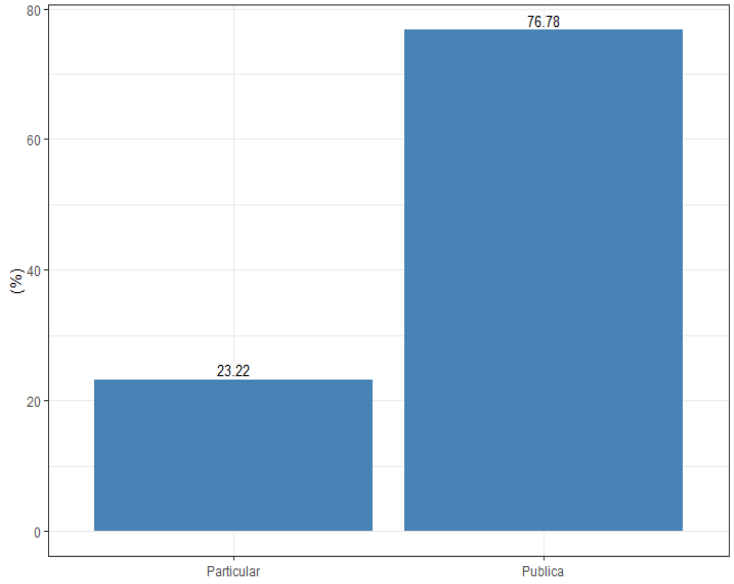

Fonte: Elaboração pelo autor.

Figura 3 - Distribuição de participantes do Enem no município de Varginha-MG no ano de 2020 por Faixa Etária.



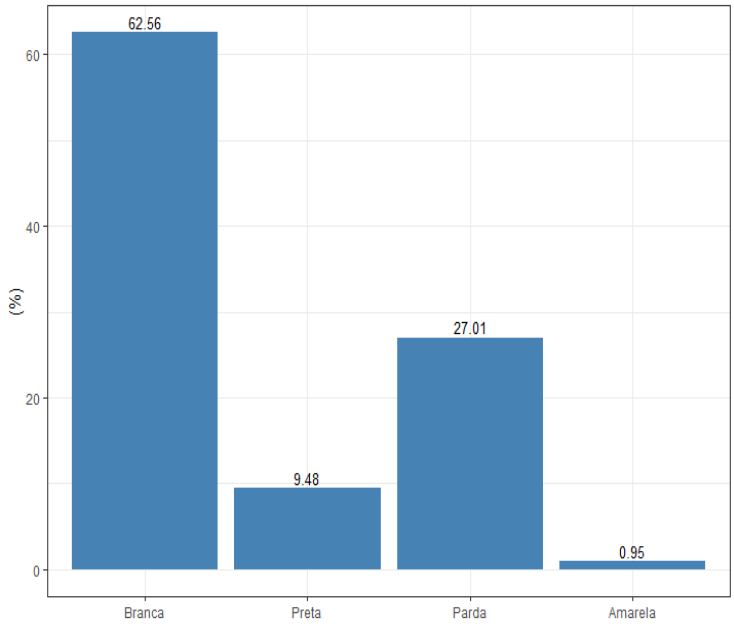

Fonte: Elaboração pelo autor.

Figura 4 - Distribuição de participantes do Enem no município de Varginha-MG no ano de 2020 por Tipo de Escola.



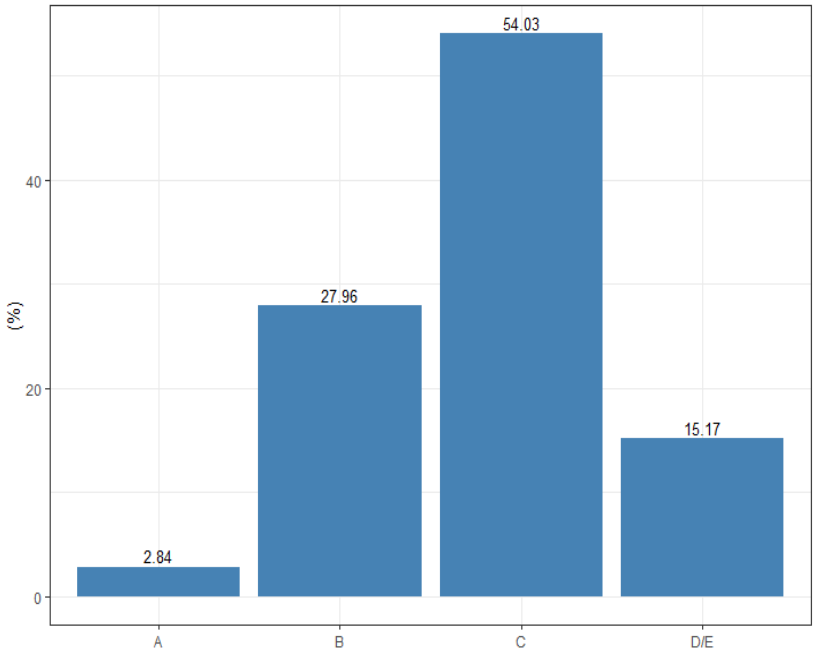

Fonte: Elaboração pelo autor.

Figura 5 - Distribuição de participantes do Enem no município de Varginha-MG no ano de 2020 por Raça.




Fonte: Elaboração pelo autor.

Figura 6 - Distribuição de participantes do Enem no município de Varginha-MG no ano de 2020 por Faixa de Renda Familiar.




Fonte: Elaboração pelo autor.

## 4.2 AJUSTE DO MODELO

O MRLM foi ajustado utilizando a função lm() do software R, e os resultados deste ajuste são apresentados na Tabela 2.

Tabela 2- Modelo Inicial ajustado.

| Coeficientes | Estimativas | Erro Padrão | Estatística t | Valor-p |
|---|---|---|---|---|
| Intercepto | 685,33 | 34,23 | 20,02 | < 2e-16 |
| Gênero(masculino) | 33,37 | 10,78 | 3,10 | 0,0022 |
| Faixa Etária (< 18 anos) | 2,35 | 7,29 | 0,32 | 0,7478 |
| Raça (Preta) | -0,06 | 36,08 | 0,00 | 0,9986 |
| Raça (Parda) | 17,53 | 28,32 | 0,62 | 0,5365 |
| Raça (Amarela) | -0,62 | 17,70 | -0,04 | 0,9722 |
| Escola (Publica) | -51,92 | 14,42 | -3,60 | 0,0004 |
| Classe Renda (B) | -52,78 | 32,41 | -1,63 | 0,1050 |
| Classe Renda (C) | -77,45 | 33,35 | -2,32 | 0,0212 |
| Classe Renda (D/E) | -130,22 | 35,77 | -3,64 | 0,0003 |

Fonte: Elaboração pelo autor.

Por meio dos resultados da Tabela 2, podemos concluir que a nível de significância de 5% as covariáveis Faixa Etária e Raça não são estatisticamente significativas para explicar a nota média dos participantes do Enem de Varginha-MG no período analisado. Além disso, a Classe de Renda Familiar B somente é significativa a nível de 10%, por considerar essa covariável extremamente relevante ela será mantida no modelo. Um novo modelo excluindo as covariáveis Faixa Etária e Raça foi ajustado, e os resultados desse ajuste são apresentados na Tabela3 .

Tabela 3 - Modelo Final ajustado.

| Coeficientes | Estimativas | Erro Padrão | Estatística t | Valor-p |
|---|---|---|---|---|
| Intercepto | 694,05 | 30,37 | 22,852 | < 2e-16 |
| Gênero(masculino) | 33,83 | 10,62 | 3,185 | 0,001675 |
| Escola (Publica) | -55,76 | 14,09 | -3,957 | 0,000104 |
| Classe Renda (B) | -55,53 | 32,25 | -1,722 | 0,086596 |
| Classe Renda (C) | -81,62 | 33,12 | -2,464 | 0,014557 |
| Classe Renda (D/E) | -131,59 | 35,53 | -3,704 | 0,000273 |

Fonte: Elaboração pelo autor.

Com base na Tabela 3, podemos escrever o modelo final ajustado da seguinte forma:

$$\widehat{Nota} = 694{,}05 + 33{,}83Maculino - 55{,}76EscolaPublica - 55{,}53ClasseRendaB - 81{,}62ClasseRendaC - 131{,}59ClasseRendaD/C$$

Observando os coeficientes estimados, chegamos as seguintes conclusões:

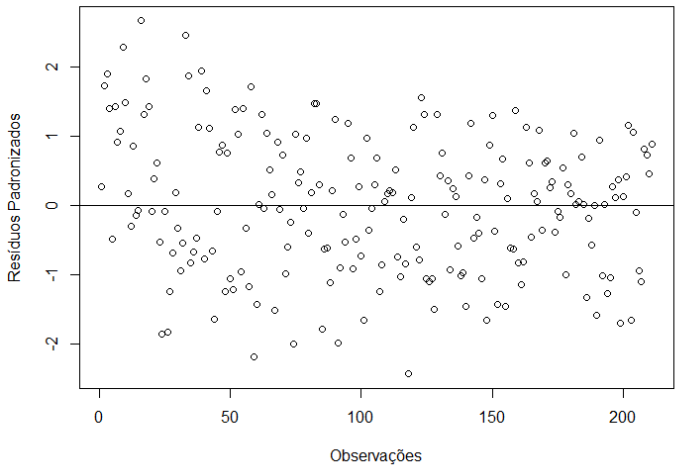
- Nota média dos participantes do Enem no município de Varginha-MG, considerando as demais covariáveis zero, foi de 694,05 pontos.
- Para o mesmo nível das demais covariáveis, participantes do gênero masculino apresentaram em média nota maiores que participantes do gênero feminino, com nota média final em torno de 34 pontos mais alta.
- Para o mesmo nível das demais covariáveis, participantes de escolas públicas apresentaram em média nota menores que participantes advindos de escolas privadas, com nota média final em torno de 56 pontos mais baixa.
- Para o mesmo nível das demais covariáveis, participantes da classe A apresentaram em média nota maiores que participantes advindos das demais classe, com a classe B apresentando nota média final em torno de 56 pontos menor que a classe A, mantendo-se as demais classes constantes; classe C apresentando nota média final em torno de 82 pontos menor que a classe A, mantendo-se as demais classes constantes; classes D/E apresentando nota média final em torno de 132 pontos menor que a classe A mantendo-se as demais classes constantes.

Esse modelo final apresentou um coeficiente de determinação ajustado $R_a^2 = 0{,}2769$ , indicando que o modelo tem um baixo poder de explicação. Nesse caso, pode se pensar que a inclusão de outros fatores além dos considerado, ou utilizar outra metodologia, poderá melhorar a explicação do modelo,

Para as covariáveis consideradas no modelo final foi calculado o VIF, e todos apresentaram valores menores que 10, indicando assim a ausência de multicolinearidade entre as covariáveis.

Ao realizar a análise de resíduos para o modelo final ajustado, conclui-se que nenhuma das pressuposições iniciais dos erros aleatórios foram violadas.

Figura 7. Resíduos padronizados do modelo final ajustado.




Fonte: Elaboração pelo autor.

A confirmação das pressuposições iniciais dos erros aleatórios, ao nível de significância de 5%, foram validades pelos testes: Shapiro-Wilk ($H_0$: Resíduos seguem distribuição Normal), Breusch-Pagan ($H_0$: Resíduos possuem variância constante (são homocedasticos)), e Durbin-Watson ($H_0$: Resíduos não são autocorrelacionados).

# 5 CONSIDERAÇÕES FINAIS

Este estudo analisou o desempenho dos participantes do Enem no município de Varginha-MG no ano de 2020, considerando os impactos das características socioeconômicas sobre as notas dos candidatos, a partir dos dados disponibilizados pelo Inep.

Os resultados obtidos no modelo final ajustado, indicam que no período analisado o desempenho médio dos participantes do Enem no município de Varginha-MG é explicado pelos fatores de gênero, tipo de escola e renda. Logo, demonstrou-se a associação existente entre desempenho dos alunos nas avaliações do Enem com o gênero, classe de renda familiar e os estudantes serem ou não de escola pública ou privada. Mesmo que as demais variáveis são relevantes em análises socioeconômicas educacionais, elas não apresentaram significância estatística neste estudo.

As notas obtidas pelos participantes homens foram em média melhores que as das participantes mulheres no período. Além disso, aspectos da desigualdade social são evidenciados a partir dos resultados de tipo de escola que se conclui o Ensino médio e renda, dada pelas classes sociais. A desigualdade social é um elemento presente da sociedade brasileira e, consequentemente, crianças de famílias mais pobres têm dificuldades de se dedicarem aos estudos, porque precisam se preocupar com a renda familiar, e até mesmo pelo acesso as escolas.

Em termos dos resultados associados as escolas públicas e privada, a educação pública brasileira é bastante questionada quanto a sua qualidade e eficiência. Segundo avaliações realizadas pelo Inep, apenas 5% dos alunos das escolas públicas apresentam desemprenho classificado como “Adequado”, e o sistema educacional está estagnado desde 2009. Além disso, há os impactos da pandemia do COVID-19 que afetou principalmente alunos de baixa renda, que dependiam de acesso à internet e aparelhos eletrônicos (tablets, computadores etc.) para terem acesso às aulas remotas.

A relevância deste estudo está na compreensão das variáveis de maior influência sobre as notas do exame, por meio de uma análise quantitativa de informações disponibilizadas pelo Inep, o que pode embasar políticas públicas e diretrizes para o contexto educacional do município, gerando impactos na redução das desigualdades e promoção de avanços educacionais.

**REFERÊNCIAS**


BONETI, L. W. ; DE OLIVEIRA, G. M. **Enem: análise do desempenho escolar nas edições de 2009 a 2013.** Revista Espaço Pedagógico, v. 24, n. 2, 2017. Disponível em: < http://seer.upf.br/index.php/rep/article/view/7420/4361>. Acesso em: 10 jan. 2023.

BRASIL. Ministério da Educação. **Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira** | Inep. Brasília, 2023a. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem.>. Acesso em: 05 jan. 2023.

BRASIL. **Oportunidades de Acesso ao Ensino Superior.** Brasília: MEC, 2023b. Disponível em: < https://acessounico.mec.gov.br/programas>. Acesso em: 05 jan. 2023.

CRUZ, R. C. **Uma avaliação empírica do Exame Nacional do Ensino Médio– ENEM: impacto da pandemia do Covid-19 no desempenho dos participantes do ENEM 2020.** Dissertação (Mestrado Profissional em Políticas Públicas). Faculdade Católica de brasília. Brasília, p. 36. 2022. Disponível em: < https://bdtd.ucb.br:8443/jspui/bitstream/tede/3002/2/RenatoCarvalhodaCruzDissertacao2022.pdf >. Acesso em: 19 jan. 2023.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA**. Rendimento, despesa e consumo.** Rio de Janeiro: IBGE. Disponível em: < https://www.ibge.gov.br/estatisticas/sociais/rendimento-despesa-e-consumo.html> . Acesso em: 02 jan. 2023.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Microdados do Enem 2020**. Brasília: Inep, 2021. Disponível em: < https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>. Acesso em: 03 jan. 2023.

JUSTINIANO, E. F.; QUEIROZ, A. P. **Renda, participação e desempenho no ENEM em São Paulo: uma abordagem espacial (2012-2018).** Confins. Revue franco-brésilienne de géographie/Revista franco-brasilera de geografia, n. 51, 2021. Disponível em: < https://journals.openedition.org/confins/38804>. Acesso em: 10 jan. 2023.

LUCENA, J. P. O.; DOS SANTOS, H. N. L. **A relação entre desempenho no Exame Nacional do Ensino Médio e o perfil socioeconômico: um estudo com os microdados de 2016**. Revista de Gestão e Secretariado, v. 11, n. 2, p. 1-23, 2020. Disponível em: < https://www.researchgate.net/publication/343467401_A_relacao_entre_desempenho_no_Exame_Nacional_do_Ensino_Medio_e_o_perfil_socioeconomico_um_estudo_com_os_microdados_de_2016>. Acesso em: 12 jan. 2023.

QEdu, 2023. **Censo Escolar Município de Varginha-MG**. Disponível em: < https://qedu.org.br/brasil/censo-escolar?7&brasil>. Acesso em: 12 jan. 2023.

R CORE TEAM . R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

WOOLDRIDGE, J. M. **Introductory econometrics: a modern approach**. Cengage learning, 2015.