

# List 05 - Regressão

Artur Damião \*

1 de dezembro de 2025

## 1 Ambiente

```
pacman::p_load(  
  rio,  
  tidyverse,  
  stargazer,  
  sjPlot,  
  lmtest,  
  sandwich,  
  performance  
)  
  
options(scipen = 999)
```

## 2 Importando e limpando dados do ENEM 2023

2. Lendo o documento usando a função `import` do pacote `rio`, que é mais rápido.

```
df <- rio::import("../dados/DADOS/MICRODADOS_ENEM_2023.csv")
```

3. Selecionando as variáveis de interesse.

```
df <- df %>%  
  select(NU_NOTA_CH, NU_NOTA_MT, NU_NOTA_CN, NU_NOTA_LC, TP_SEXO, TP_COR_RACA, TP_ESCOLA)
```

As variáveis selecionadas dizem respeito à nota da prova de um candidato em Ciências Humanas, Ciências da Natureza, Linguagem e Códigos, Matemática, bem como o Tipo de Escola cursado no Ensino Médio (Pública ou Privada), a Cor/Raça, de acordo com critérios do IBGE, e o Sexo (Masculino ou Feminino).

4. Limpando a base de dados.

---

\*Programa de Pós-Graduação em Sociologia da Universidade de São Paulo. N° USP 10701251. arturcar-doso@usp.br.

```

df_limpo <- df %>%
  filter(
    # Filtrando notas iguais a zero
    NU_NOTA_CH > 0,
    NU_NOTA_MT > 0,
    NU_NOTA_CN > 0,
    NU_NOTA_LC > 0,

    # Filtrando raça não declarada
    TP_COR_RACA != 0 & TP_COR_RACA != 6,

    # Filtrando escola não declarada
    TP_ESCOLA != 1
  ) %>%
  # Eliminando dados ausentes
  na.exclude() %>%
  janitor::clean_names()

```

### 3 Recortando os dados

O meu banco é composto por 1032361 linhas, após a filtragem de dados e a exclusão de NAs.

```

set.seed(666)

amostra <- df_limpo %>% slice_sample(n = 100000)

```

### 4 Analisando os dados: regressão linear simples

Objetivo: Utilizar as notas em ciências humanas para prever as notas em matemática por meio de uma regressão linear simples. Nossa equação é dada por:

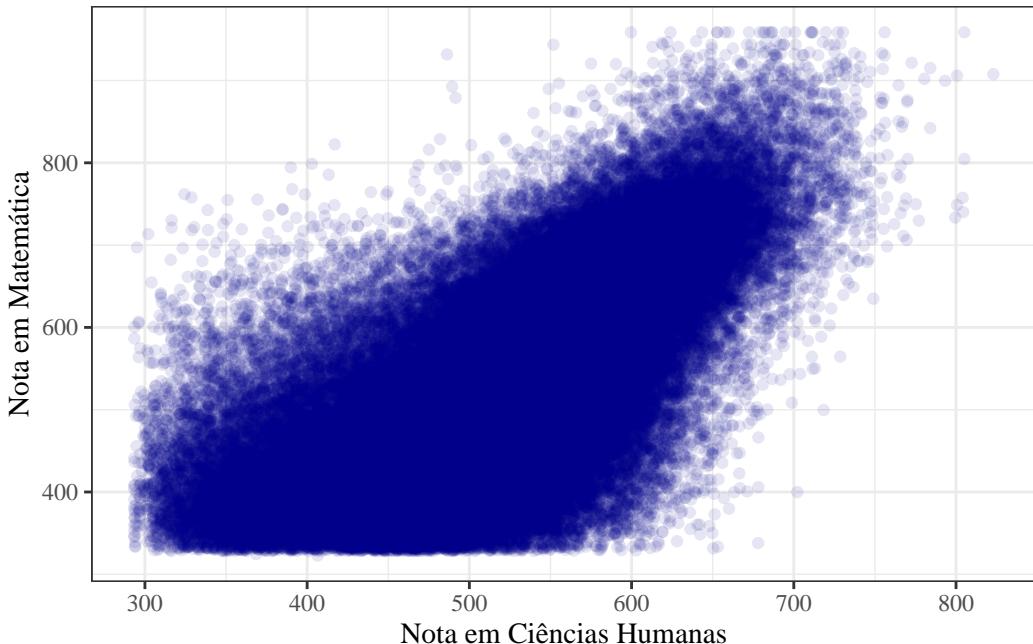
$$\text{NU\_NOTA\_MT} = \beta_0 + \beta_1 \times \text{NU\_NOTA\_CH} + \epsilon \quad (1)$$

1. Gráfico de dispersão

```

amostra %>%
  ggplot(aes(x = nu_nota_ch, y = nu_nota_mt)) +
  geom_point(
    alpha = 0.1,
    colour = "darkblue") +
  labs(
    x = "Nota em Ciências Humanas",
    y = "Nota em Matemática")
  +
  theme_bw(base_family = "serif")

```



## 2. Regressão linear simples<sup>1</sup>

```
reg_1 <- lm(nu_nota_mt ~ nu_nota_ch,
             data = amostra)

stargazer(reg_1, type = "text")
```

Dependent variable:	
	nu_nota_mt
nu_nota_ch	0.990*** (0.004)
Constant	18.921*** (1.851)
<hr/>	
Observations	100,000
R2	0.442
Adjusted R2	0.442
Residual Std. Error	92.116 (df = 99998)
F Statistic	79,065.360*** (df = 1; 99998)
<hr/>	
Note:	*p<0.1; **p<0.05; ***p<0.01

De acordo com o modelo, para cada aumento de 1 ponto na nota de Ciências Humanas, a nota de matemática é esperada aumentar em 0,990, pontos em média. Isso é constatado ao interpretar o valor do  $\hat{\beta}_1$ , que é altamente significativo, com  $p < 0.01$ . Além disso, o  $R^2$  é de 44,4%. Ou seja, a nota em Ciências Humanas explica 44% da variação total na nota de matemática.

<sup>1</sup>Prefiro utilizar a função `stargazer` porque o output é melhor.

## 5 Inferência estatística

1. Cálculo do intervalo de confiabilidade dos coeficientes ao nível de confiança de 95%.

$$IC = \hat{\beta} \pm 1.96 \times \text{Erro Padrão}$$

Para  $\hat{\beta}_0$  temos:

```
ic_inferior_b0 = 18.876 - (1.96 * 1.844)
ic_superior_b0 = 18.876 + (1.96 * 1.844)
```

Para  $\hat{\beta}_1$  temos:

```
ic_inferior_b1 = 0.990 - (1.96 * 0.004)
ic_superior_b1 = 0.990 + (1.96 * 0.004)
```

Usando a função `confint`:

```
confint(reg_1, level = 0.95)
```

```
2.5 %      97.5 %
(Intercept) 15.2923395 22.5491094
nu_nota_ch   0.9827309  0.9965272
```

2. Homocedasticidade e heterocedasticidade

*Homocedasticidade* e heterocedasticidade referem-se à variância dos erros em um modelo de regressão. A homocedasticidade significa que os erros variam de forma constante ao redor da linha de regressão. É a premissa ideal.

A *heterocedasticidade* significa que a dispersão dos resíduos varia à medida que a variável preditora muda. O modelo não é confiável quando a premissa é violada.

3. Recalculando o modelo de regressão

```
coeftest(reg_1, vcov = vcovHC(reg_1, type = "HC1"))
```

t test of coefficients:

	Estimate	Std. Error	t value	Pr(> t )			
(Intercept)	18.9207245	1.8540885	10.205 < 0.0000000000000022	***			
nu_nota_ch	0.9896290	0.0035632	277.732 < 0.0000000000000022	***			
---							
Signif. codes:	0	'***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

As estimativas do modelo permanecem as mesmas, ao passo que o erro padrão apresenta uma pequena variação. Agora, é calculado o erro padrão robusto. O valor de *p* continua significativo.

## 6 Checagem do modelo

1. Gere os resíduos do modelo executado anteriormente

```
residuos <- residuals(reg_1)
preditor <- amostra$nu_nota_ch
```

2. Crie uma nova tabela com as seguintes variáveis: os resíduos gerados no item anterior e a variável preditora do nosso modelo.

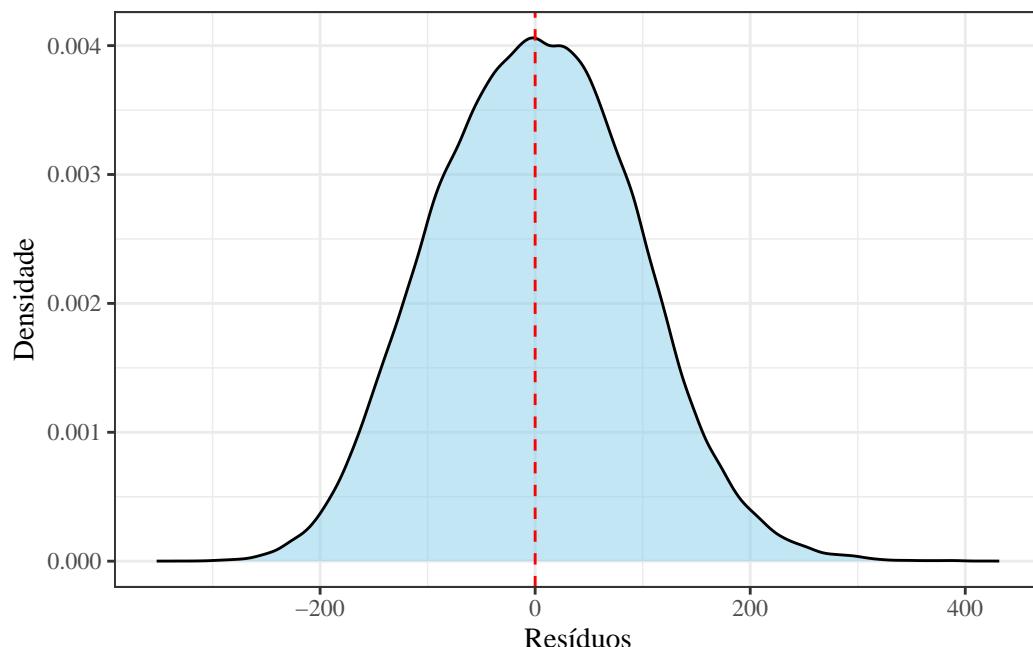
```
df_residuos <- data.frame(
  residuos = residuos,
  nu_nota_ch = preditor,
  id = row_number(amostra)
)
```

3. Análise dos resíduos.

A média dos resíduos é de 0. A Figura 1 apresenta a distribuição dos resíduos. A média dos resíduos é muito próxima de zero, e a distribuição segue uma distribuição normal.

```
df_residuos %>%
  ggplot(aes(x = residuos)) +
  geom_density(fill = "skyblue", alpha = 0.5) +
  geom_vline(xintercept = 0, color = "red", linetype = "dashed") +
  labs(
    x = "Resíduos",
    y = "Densidade"
  ) +
  theme_bw(base_family = "serif")
```

Figura 1: Distribuição de Densidade dos Resíduos do Modelo



#### 4. Gráfico quantil-quantil

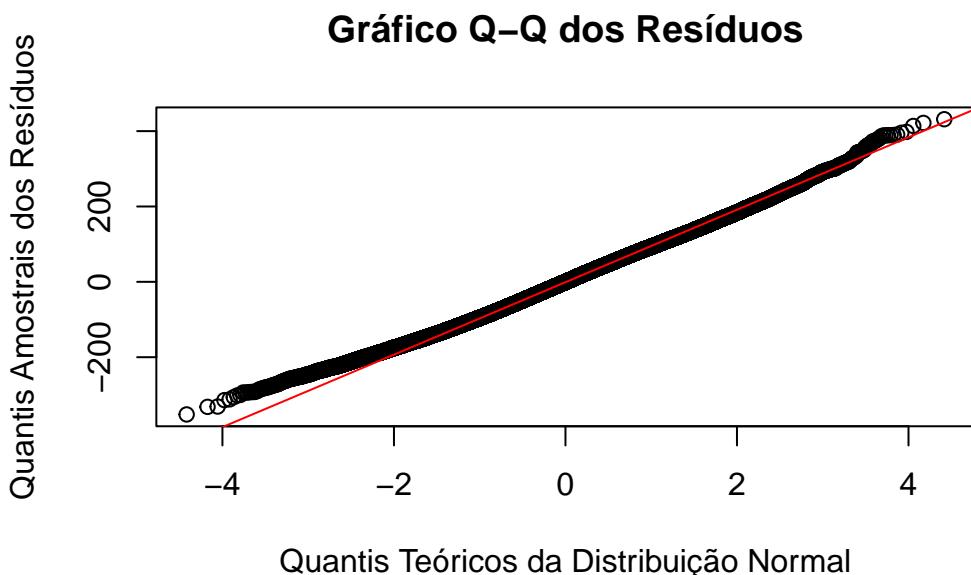
O qqplot checa a normalidade dos resíduos, ao comparar os quantis dos resíduos observados com quantis teóricos. Quando a distribuição dos resíduos é normal, os pontos no gráfico QQ se alinham com a reta.

#### 5. Funções `qqnorm()` e `qqline()`

```
# Configurando o layout para o gráfico Q-Q
par(mfrow = c(1, 1))

# Gráfico Q-Q dos resíduos
qqnorm(df_residuos$residuos,
       main = "Gráfico Q-Q dos Resíduos",
       xlab = "Quantis Teóricos da Distribuição Normal",
       ylab = "Quantis Amostrais dos Resíduos")

# Adicionar a linha de referência
qqline(df_residuos$residuos, col = "red")
```



Através do gráfico, há normalidade dos resíduos.

#### 6. Gráfico de dispersão

A Figura 2 apresenta os resíduos contra os valores ajustados do modelo. O pressuposto da linearidade avalia se a realação entre os preditores e a variável resposta é, em média, linear. Supõe-se que  $E[\epsilon|X] = 0$  (a média condicional dos erros é zero, logo, o preditor e o erro não estão correlacionados).

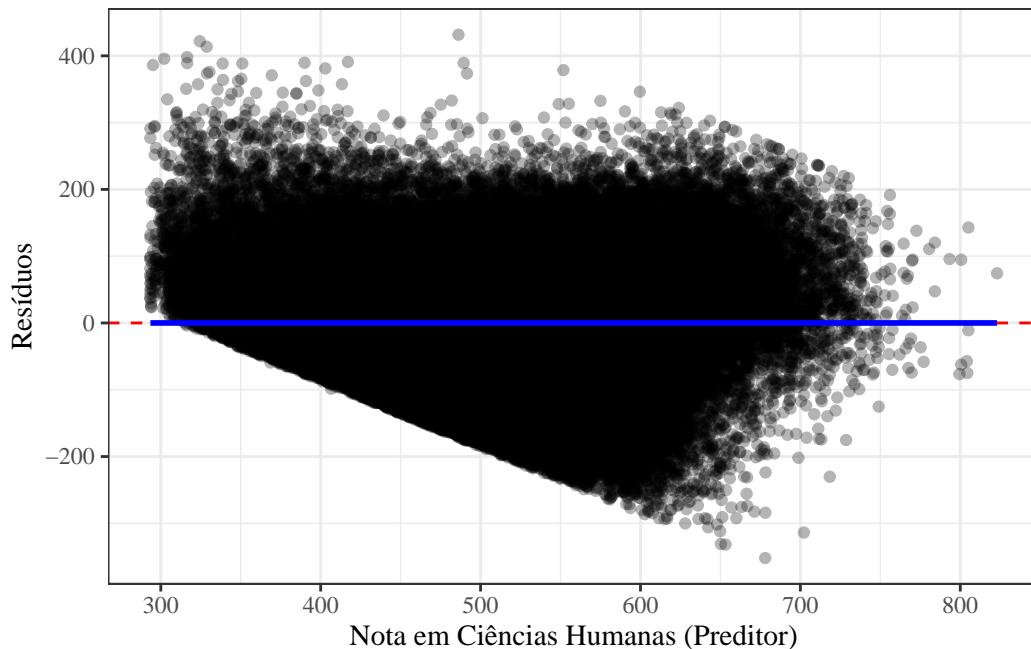
```
df_residuos %>%
  ggplot(aes(x = nu_nota_ch, y = residuos)) +
  geom_point(alpha = 0.3) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  geom_smooth(method = "lm", color = "blue", se = FALSE) + # Reta de Regressão
  labs(
    x = "Nota em Ciências Humanas (Preditor)",
```

```

y = "Resíduos") +
theme_bw(base_family = "serif")

```

Figura 2: Resíduos vs. Predictor (NU\_NOTA\_CH)



## 7. Quadrado dos resíduos

A Figura 3 confirma a presença de heterocedasticidade, ou seja, a variância nos erros não é constante.

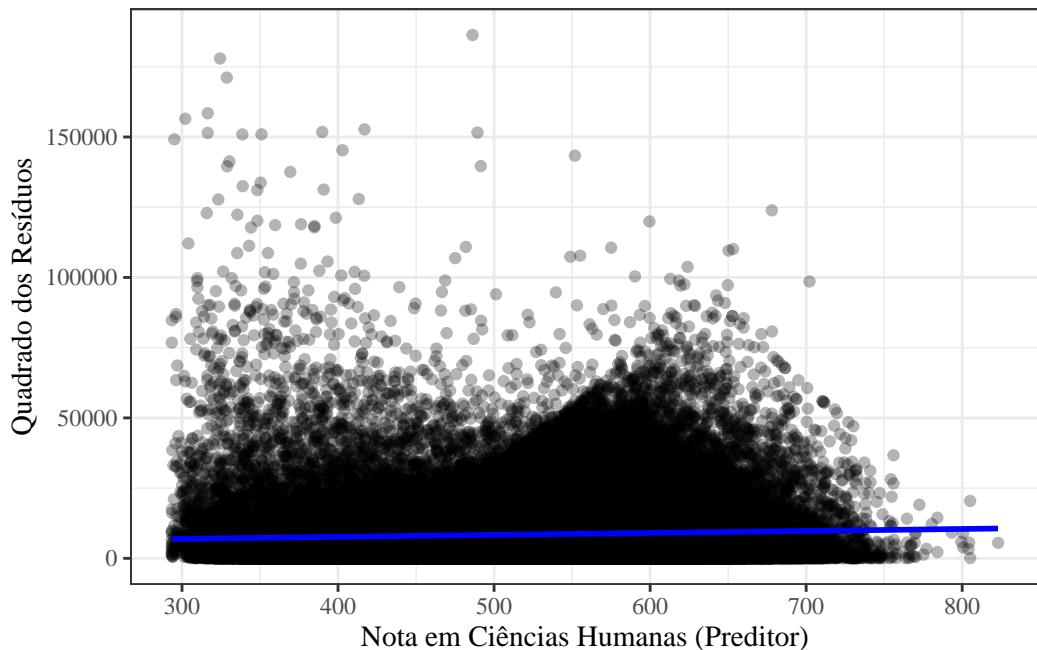
```

df_residuos <- df_residuos %>%
  mutate(residuos_quadrado = residuos^2)

df_residuos %>%
  ggplot(aes(x = predictor, y = residuos_quadrado)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", color = "blue") +
  labs(
    x = "Nota em Ciências Humanas (Predictor)",
    y = "Quadrado dos Resíduos") +
  theme_bw(base_family = "serif")

```

Figura 3: Quadrado dos Resíduos vs. Preditor (Checagem de Homoscedasticidade)



9. Geração da coluna “id”

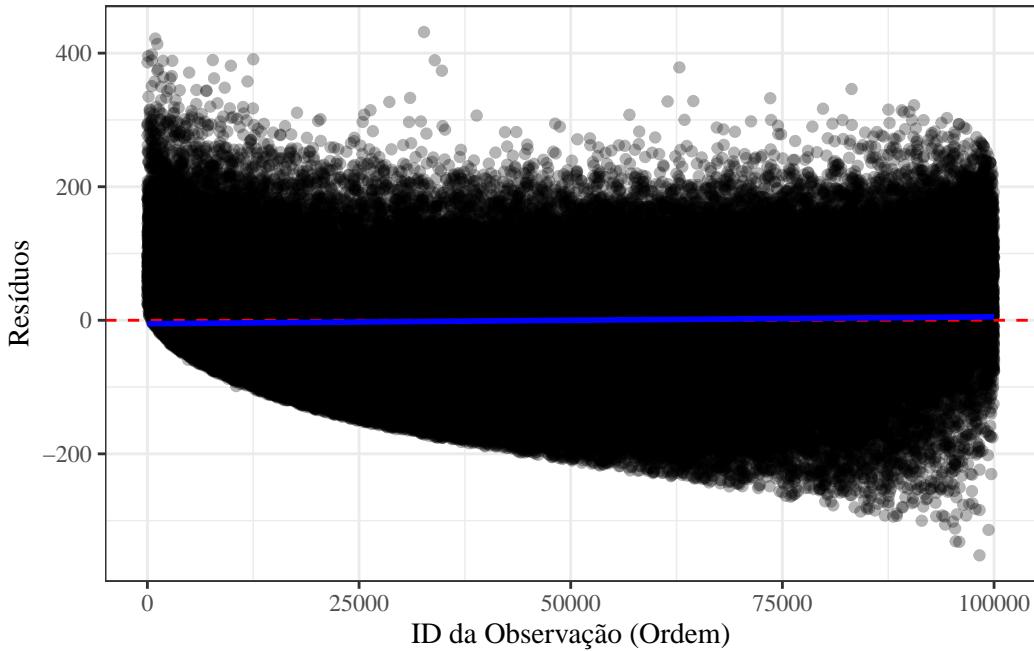
A coluna “id” foi feita no ponto 1.

10. Independência dos erros

A Figura 4 sugere que não há correlação entre a ordem da observação e o resíduo.

```
df_residuos %>%
  ggplot(aes(x = id, y = residuos)) +
  geom_point(alpha = 0.3) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  geom_smooth(method = "lm", color = "blue", se = FALSE) +
  labs(
    x = "ID da Observação (Ordem)",
    y = "Resíduos") +
  theme_bw(base_family = "serif")
```

Figura 4: Resíduos vs. Ordem da Observação (ID)



## 11. Conclusões gerais

A análise conduzida até aqui sugere que o modelo linear capta, de forma razoável, a relação entre as notas em ciências humanas e matemática. Entretanto, constatamos a violação de alguns pressupostos teóricos, como a presença de heterocedasticidade e leves desvios de normalidade dos resíduos. O pressuposto de independência dos erros foi satisfeito.

## 7 Regressão múltipla

### 1. Reclassificando variáveis

```
amostra <- amostra %>%
  mutate(
    tp_sexo = factor(tp_sexo),
    tp_cor_raca = factor(tp_cor_raca),
    tp_escola = factor(tp_escola),

    # Alterando os níveis
    tp_sexo = relevel(tp_sexo, ref = "M"),
    tp_cor_raca = relevel(tp_cor_raca, ref = "1"),
    tp_escola = relevel(tp_escola, ref = "2")
  )
```

### 2. Regressão linear múltipla

```
reg_2 <- lm(nu_nota_mt ~ nu_nota_ch + nu_nota_cn + nu_nota_lc + tp_sexo + tp_cor_raca +
  tp_escola, data = amostra)
```

### 3. Interpretando os coeficientes

```
stargazer::stargazer(reg_2, type = "text")
```

=====	
Dependent variable:	
-----	
	nu_nota_mt
-----	-----
nu_nota_ch	0.353*** (0.005)
nu_nota_cn	0.492*** (0.005)
nu_nota_lc	0.369*** (0.006)
tp_sexoF	-27.588*** (0.515)
tp_cor_raca2	-18.317*** (0.857)
tp_cor_raca3	-10.981*** (0.560)
tp_cor_raca4	-1.265 (2.091)
tp_cor_raca5	-15.938*** (3.331)
tp_escola3	37.811*** (0.678)
Constant	-67.401*** (2.177)
-----	-----
Observations	100,000
R2	0.589
Adjusted R2	0.589
Residual Std. Error	79.013 (df = 99990)
F Statistic	15,931.730*** (df = 9; 99990)
=====	=====
Note:	*p<0.1; **p<0.05; ***p<0.01

Este modelo, agora com mais covariáveis, explica aproximadamente 59% da variação nas notas de matemática. Um aumento de 1 ponto na nota de Ciências da Natureza está associado a um aumento de 0.493 pontos na nota de Matemática. Um aumento de 1 ponto na nota de Linguagens

e Códigos está associado a um aumento de 0,370 pontos na nota de Matemática. Um aumento de 1 ponto na nota de Ciências Humanas está associado a um aumento de 0,351 pontos na nota de Matemática. Anteriormente, vimos que o aumento de um ponto na em Ciências Humanas aumentava, em média, 0,990 na nota de matemática. Pessoas do sexo Feminino tendem a ter uma nota 27,813 pontos menor do que estudantes do sexo masculino.

Em relação à raça, cuja categoria de referência é a cor branca, pessoas de cor preta pontuam, em média, 18,3 pontos a menos que pessoas brancas; pessoas pardas pontuam, em média, 10,9 pontos a menos que pessoas brancas; pessoas amarela pontuam, em média, 1,2 pontos a menos que pessoas brancas; e pessoas indígenas pontuam, em média, 15,9 pontos a menos que pessoas brancas.

Pessoas que cursaram escola privada no ensino médio pontuam, em média, 37,8 pontos a mais que pessoas que estudaram em instituições públicas.

O valor do intercepto não faz sentido para essa análise, já que não é possível obter uma nota negativa no ENEM.

#### 4. Comparando os modelos

Como constatado pelo  $R^2$  do modelo de regressão múltipla, o poder de explicação deste modelo é maior, haja vista que ele considera outras variáveis e isola o efeito de cada preditor. Ao adicionar mais variáveis, mitigamos o viés de uma variável omitida.

#### 5. Erro padrão robusto

```
coeftest(reg_2, vcov = vcovHC(reg_2, type = "HC1"))
```

t test of coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-67.4009080	2.1063654	-31.9987	< 0.000000000000000022 ***
nu/nota_ch	0.3532758	0.0052369	67.4595	< 0.000000000000000022 ***
nu/nota_cn	0.4917453	0.0049282	99.7816	< 0.000000000000000022 ***
nu/nota_lc	0.3690748	0.0058418	63.1787	< 0.000000000000000022 ***
tp_sexoF	-27.5876571	0.5256101	-52.4869	< 0.000000000000000022 ***
tp_cor_raca2	-18.3171516	0.8572843	-21.3665	< 0.000000000000000022 ***
tp_cor_raca3	-10.9813105	0.5658572	-19.4065	< 0.000000000000000022 ***
tp_cor_raca4	-1.2652824	2.1296357	-0.5941	0.5524
tp_cor_raca5	-15.9376873	3.2498764	-4.9041	0.0000009401 ***
tp_escola3	37.8113995	0.7146308	52.9104	< 0.000000000000000022 ***
---				
Signif. codes:	0 ***	0.001 **	0.01 *	0.05 . ' 0.1 ' ' 1

Os coeficientes estimados permanecem os mesmos, considerando que o HC1 afeta apenas os erros-padrão. A significância estatística permanece. O efeito da cor amarela alterou ligeiramente.

#### 6. Checagem do modelo de regressão múltipla.

```
performance::check_normality(reg_2)
```

Warning: Non-normality of residuals detected (p < .001).

```
performance::check_collinearity(reg_2)
```

```
# Check for Multicollinearity
```

Low Correlation

Term	VIF	VIF 95% CI	adj. VIF	Tolerance	Tolerance 95% CI
nu_nota_ch	2.77	[2.74, 2.80]	1.66	0.36	[0.36, 0.36]
nu_nota_cn	2.02	[2.00, 2.03]	1.42	0.50	[0.49, 0.50]
nu_nota_lc	2.66	[2.63, 2.69]	1.63	0.38	[0.37, 0.38]
tp_sexo	1.03	[1.02, 1.03]	1.01	0.97	[0.97, 0.98]
tp_cor_raca	1.11	[1.10, 1.11]	1.01	0.90	[0.90, 0.91]
tp_escola	1.22	[1.22, 1.23]	1.11	0.82	[0.81, 0.82]

```
performance::check_autocorrelation(reg_2)
```

OK: Residuals appear to be independent and not autocorrelated ( $p = 0.752$ ).

```
performance::check_heteroskedasticity(reg_2)
```

Warning: Heteroscedasticity (non-constant error variance) detected ( $p < .001$ ).

Embora as premissas de independência dos resíduos e ausência de multicolinearidade do Modelo de Regressão Múltipla sejam satisfeitas, constatamos problemas de não normalidade dos resíduos e heterocedasticidade. A multicolinearidade é baixa, com todos os valores de VIF bem abaixo do limite de 5 ou 10, e o teste de Durbin-Watson indica que os resíduos são independentes, ou seja, não há autocorrelação. No entanto, a detecção de não normalidade dos resíduos e a heterocedasticidade nos chamam a atenção.

Ao analisarmos o HC1, que corrige a inconsistência da heterocedasticidade, validamos as inferências estatísticas. Ou seja, os valores são robustamente significativos. O HC1 não muda o impacto dos coeficientes, mas sim reforça a confiança que temos nos valores  $p$  e sua significância.