Trabalho 1 de Econometria

Pedro Henrique Ferreira de Souza e João Paulo de Souza Ferreira

18 de maio de 2025

1 Metodologia

Objetiva-se com essa seção apresentar a base de dados utilizada e explicitar a estratégia empírica utilizada.

1.1 Base de dados

A fonte dos dados utilizados no trabalho é a Relação Anual de Informações Sociais (RAIS). A plataforma possui diversos dados e níveis de agregação, contendo informações de 2.575.873 de firmas em 2021 (RAIS, 2021). Outrossim, a escolha do período selecionado (2021) se deve a disponibilidade de dados, já que o período em questão possui as informações necessárias para a construção das variáveis explicativas.

1.2 Estratégia empírica

Como forma de analisar a relação entre algumas variáveis e a remuneração média real dos trabalhadores das firmas no Brasil, o presente trabalho se utilizou do modelo de Mínimos Quadrados Ordinários (MQO). Segundo Gujarati e Porter (2009), o modelo de Mínimos Quadrados Ordinários (MQO) é o melhor estimador linear não viesado quando não há violação dos pressupostos, como amostragem aleatória, lianearidade nos parâmetros, exogeneidade estrita, não colineariedade perfeita e homocedasticidade.

Como forma de analisar a relação entre algumas variáveis e a remuneração média real dos trabalhadores das firmas no Brasil, o presente trabalho se utilizou do Modelo de Regressão Linear a partir do Método de Mínimos Quadrados Ordinários.

O Modelo de Regressão Linear é uma técnica estatística utilizada para tentar prever o comportamento de uma variável considerada dependente (Y) a partir da observação e análise de variáveis independentes ou explicativas $(X_1, X_2, ..., X_k)$. Intuitivamente, o modelo de Regressão se utiliza de uma amostra aleatória para estimar a equação de uma reta e com isso obter informações sobre os parâmetros que explicam o comportamento da variável dependente (Y).

Para estimar o modelo de regressão linear utilizaremos o Método de Mínimos Quadrados Ordinários (MQO) conforme descrito por Gujarati e Porter (2009).

O Método de MQO consiste em reduzir a diferença entre o valor real do Y e o valor estimado do Y para cada observação da amostra, através da minimização da soma dos quadrados dos erros. Com isso, é possível obter os melhores estimadores lineares não viesados, desde que as seguintes hipóteses sejam atendidas:

- Hipótese 1: O modelo de regressão é linear nos parâmetros, embora possa não ser linear nas variáveis;
- Hipótese 2: As variáveis X e o termo de erro são independentes, isto é, $\text{cov}(X_y, u) = 0$;

- Hipótese 3: A média condicional do termo de erro é zero em relação as variáveis explicativas é zero. Isso garante a exogeneidade do modelo, ou seja, que nenhuma variável omitida é correlacionada com as variáveis explicativas. A violação dessa hipótese torna os estimadores dos MQO tendenciosos;

- Hipótese 4: Homocedasticidade ou variância constante do termo de erro. A violação dessa hipótese causa Heterocedasticidade o que compromete a eficiência dos estimadores;

- Hipótese 5: Não há autocorrelação entre os termos de erro. A violação dessa hipótese torna os estimadores ineficientes, e testes de hipóteses tornam-se inválidos;

- Hipótese 6: O número de observações n deve ser maior que o número de parâmetros a serem estimados;

- Hipótese 7: Variabilidade dos valores das variáveis explicativas;

- Hipótese 8: Não Existência de Multicolinearidade Perfeita. Ou seja, nenhuma variável explicativa é uma combinação linear exata de outras. A violação dessa hipótese implica que a matriz X'X não é invertível, o que impossibilita a estimação dos coeficientes;

- Hipótese 9: Ausência de viés de especificação. O modelo precisa estar bem especificado, variáveis explicativas importantes não podem ser excluídas. A violação dessa hipótese causa viés por variável omitida;

- Hipótese 10: O termo de erro segue uma distribuição normal. Essa é uma hipótese fundamental para a parte de inferência estatística em MQO, pois sem ela não é possível realizar os testes T e F.

Para derivar os estimadores do MQO utilizaremos a sua forma matricial. Partindo da função de regressão amostral de k variáveis:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki} + \hat{u}_i \tag{1}$$

Em forma de matriz temos:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} + \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix} \tag{2}$$

Podemos resumir a equação acima, de modo que:

$$y = X\hat{\beta} + \hat{u} \tag{3}$$

Sendo:
y = vetor coluna n x 1 de observações da variável dependente Y;
X = matriz n x k com todas n observações das k - 1 variáveis explicativas;
$\hat{\beta}$ = vetor coluna k x 1 com os coeficientes de regressão do MQO $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$;
$\hat{u}$ = vetor coluna n X 1 de n resíduos.

Da equação 3, isolando o termo de erro, temos:

$$\hat{u} = y - X\hat{\beta} \tag{4}$$

Aplicando o conceito de soma dos quadrados dos resíduos (SQR) para obeter os estimadores de MQO, em forma matricial temos:

$$\mathbf{\hat{u}'\hat{u}} = \begin{bmatrix} \hat{u}_1 & \hat{u}_2 & \cdots & \hat{u}_n \end{bmatrix} \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix} = \hat{u}_1^2 + \hat{u}_2^2 + \cdots + \hat{u}_n^2 = \sum \hat{u}_i^2 \tag{5}$$

Portanto:

$$\begin{aligned} \hat{u}'\hat{u} &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \end{aligned} \tag{6}$$

Agora, para estimar os coeficientes $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ de modo que $\sum \hat{u}_i^2$ seja o menor possível, basta diferenciar parcialmente a equação 6 com relação aos coeficientes $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ e igualar o resultado a zero:

$$\frac{\partial(\mathbf{\hat{u}'\hat{u}})}{\partial\hat{\beta}} = -2\mathbf{X'y} + 2\mathbf{X'X}\hat{\beta} \tag{7}$$

Igualando a zero e fazendo as operações necessárias, temos:

$$(X'X)\hat{\beta} = X'y \tag{8}$$

Tendo em vista a validade da Hipótese 8, sabemos que a matriz (X'X) pode ser ivertida. Com isso, vamos multiplicar os dois lados da equação pela matriz inversa de (X'X) para obtermos a equação que estima os coeficientes do MQO:

$$\hat{\beta} = (X'X)^{-1}X'y \tag{9}$$

Após a estimação dos coeficientes do modelo de regressão linear, utiliza-se técnicas de inferência estatística, como o coeficiente de determinação (R^2), teste de hipóteses, nível de significância e intervalos de confiança, para verificar a significância e confiabilidade dos coeficientes estimados ($\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$) do modelo.

O coeficiente de determinação (R^2) mede a qualidade do ajustamento do modelo, ou seja, mede o quanto da variação na variável dependente (Y) é explicado pelas variáveis independentes ($X_1, X_2, ..., X_k$). O valor do (R^2) varia entre zero e um, em que sendo zero as variáveis independentes não explicam nada da variação de Y e sendo 1 indica que as variáveis independentes explicam toda a variação de Y.

Podemos obter o (R^2) da seguinte forma:

$$R^2 = \frac{\text{SQE}}{\text{STQ}} \tag{10}$$

Em que a Soma dos Quadrados Explicados (SQE) e a Soma Total dos Quadrados (STQ) é definida como:

$$\text{STQ:} \ \sum y_i^2 = \mathbf{y}'\mathbf{y} - n\bar{\mathbf{Y}}^2 \tag{11}$$

$$\text{SQE}: \ \hat{\beta}_2 \sum y_i x_{2i} + \cdots + \hat{\beta}_k \sum y_i x_{ki} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{Y}^2 \tag{12}$$

Logo:

$$R^2 = \frac{\hat{\beta}'\mathbf{X}'\mathbf{y} - n\,\bar{Y}^2}{\mathbf{y}'\mathbf{y} - n\,\bar{Y}^2} \tag{13}$$

A hipótese 10 de que o termo de erro segue uma distribuição normal é fundamental nessa etapa, pois ela garante que os coeficientes estimados ($\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$) também seguirão uma distribuição normal, permitindo a realização dos testes a seguir.

Para avalia a significância individual dos coeficientes estimados ($\hat{\beta}_i$), utilizamos o teste t, em que testamos a hipótese nula de que o coeficiente populacional (β_i) é igual a zero. A estatística t é calculada com a seguinte formula:

$$t = \frac{\hat{\beta}_i - \beta_i}{\text{ep}(\hat{\beta}_i)} \tag{14}$$

Se o valor da estatística t for maior que o t crítico, considerando o nível de significância (α) e os graus de liberdade (n - k), rejeita-se a hipótese nula e os coeficientes estimados (β_i) são estatisticamente significativos.

Ressalta-se que em caso de amostras muito grandes pode-se utilizar o teste z no lugar do teste t que obteremos o mesmo resultado, pois em amostras grandes, a distribuição t de Student se aproxima da distribuição normal padrão (Z).

Agora, para avaliar a significância do modelo como um todo, utilizamos o teste F, em que testamos a hipótese nula de que todos os coeficientes (exceto o intercepto) são simultaneamente iguais a zero. A estatística F é calculada com a seguinte formula:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \tag{15}$$

Se o valor da estatística f for maior que o F crítico, rejeita-se a hipótese nula e o modelo é estatisticamente significativo.

Ademais, o modelo estimado possui a seguinte característica em relação a sua forma funcional:

$$Y_i = \beta_0 + \beta X_i + \varepsilon_i \tag{16}$$

Onde Y_i representa a variável dependente (remuneração média da firma). Além disso, X_i representa o vetor de variáveis explicativas, já o β representa o vetor de parâmetros das variáveis explicativas e, por fim, ε_i o vetor do termo de erro. O subscrito i representa as firmas brasileiras da amostra.

Usou-se no trabalho diversas especificações de equações com intuito de testar as relações das variáveis explicativas com a variável dependente. Nesse sentido, segue as formulações usadas:

$$\text{remuneracao}_i = \beta_0 + \beta_1 \text{superior\_med}_i + \varepsilon_i \tag{17}$$

$$\text{remuneracao}_i = \beta_0 + \beta_1 \text{idade\_med}_i + \varepsilon_i \tag{18}$$

$$\text{remuneracao}_i = \beta_0 + \beta_1 \text{idade\_med}_i + \beta_2 \text{superior\_med}_i + \varepsilon_i \tag{19}$$

$$\text{remuneracao}_i = \beta_0 + \beta_1 \text{idade\_med}_i + \beta_2 \text{idade\_med}_i^2 + \beta_3 \text{superior\_med}_i + \varepsilon_i \tag{20}$$

$$\begin{split}\text{remuneracao}_i = {} & \beta_0 + \beta_1 \text{idade\_med}_i + \beta_2 \text{idade\_med}_i^2 + \beta_3 \text{superior\_med}_i \\ & + \beta_4 \text{sexo\_med}_i + \varepsilon_i\end{split} \tag{21}$$

$$\begin{split}\text{remuneracao}_i = {} & \beta_0 + \beta_1 \text{idade\_med}_i + \beta_2 \text{idade\_med}_i^2 + \beta_3 \text{superior\_med}_i \\ & + \beta_4 \text{sexo\_med}_i + \beta_5 \text{raca\_cor\_med}_i + \varepsilon_i\end{split} \tag{22}$$

$$\begin{split}\text{remuneracao}_i = {} & \beta_0 + \beta_1 \text{idade\_med}_i + \beta_2 \text{idade\_med}_i^2 + \beta_3 \text{superior\_med}_i + \\ & \beta_4 \text{sexo\_med}_i + \beta_5 \text{raca\_cor\_med}_i + \beta_6 \text{sexo\_raca\_interacao}_i + \varepsilon_i\end{split} \tag{23}$$

$$
\begin{aligned}
\ln(\text{remuneracao})_i = {} & \beta_0 + \beta_1 \text{idade\_med}_i + \beta_2 \text{idade\_med}_i^2 + \beta_3 \text{superior\_med}_i + \\
& \beta_4 \text{sexo\_med}_i + \beta_5 \text{raca\_cor\_med}_i + \beta_6 \text{sexo\_raca\_interacao}_i + \\
& \beta_7 \text{nivel\_tec\_1}_i + \beta_8 \text{nivel\_tec\_2}_i + \beta_9 \text{nivel\_tec\_3}_i + \\
& \beta_{10} \text{nivel\_tec\_4}_i + \beta_{11} \text{nivel\_tec\_5}_i + \varepsilon_i
\end{aligned}
\tag{24}
$$

Abaixo segue a Tabela 1 com a descrição das variáveis utilizadas nas estimações de Mínimos Quadrados Ordinários:

Tabela 1: Variáveis a serem usadas no modelo econométrico

Variável	**Descrição**
remuneracao_i	Média da remuneração real dos trabalhadores da firma
idade_med_i	Idade média dos trabalhadores da firma
$\text{idade\_med\_quadrado}_i$	Idade média dos trabalhadores da firma ao quadrado
superior_med_i	Proporção de trabalhadores da firma com grau de escolaridade igual ou superior a "superior completo"
sexo_med_i	Proporção de trabalhadores do sexo feminino na firma
raca_cor_med_i	Proporção de negros (pretos e pardos) e indígenas na firma
sexo_cor_raca_i	Interação entre proporção de trabalhadores do sexo feminino na firma e proporção de negros (pretos e pardos) e indígenas na firma
nivel_tec_1_i	Igual a 1 quando há intensidade tecnológica baixa da firma, 0 caso contrário
nivel_tec_2_i	Igual a 1 quando há intensidade tecnológica média-baixa da firma, 0 caso contrário
nivel_tec_3_i	Igual a 1 quando há intensidade tecnológica média da firma, 0 caso contrário
nivel_tec_4_i	Igual a 1 quando há intensidade tecnológica média-alta da firma, 0 caso contrário
nivel_tec_5_i	Igual a 1 quando há intensidade tecnológica alta da firma, 0 caso contrário

2 Resultados

2.1 Análise descritiva

A seguir, realizamos uma análise descritiva detalhada das variáveis incluídas no modelo. Para isso, apresentaremos uma tabela de estatísticas sumarizadas junto com a interpretação dos gráficos de densidade de Kernel, que permitem visualizar a forma das distribuições de cada uma das variáveis analisadas.

A análise da densidade kernel da remuneração média real das firmas (Figura 1) revela uma distribuição assimétrica à direita, com uma concentração massiva de valores próximos à média (1.694,87 reais), mas uma cauda extremamente alongada até 140.000 reais, o que evidência a presença de outliers extremos vinculados a setores ou cargos com alta remuneração, enquanto a maior parte das firma têm uma remuneração média muito baixa.

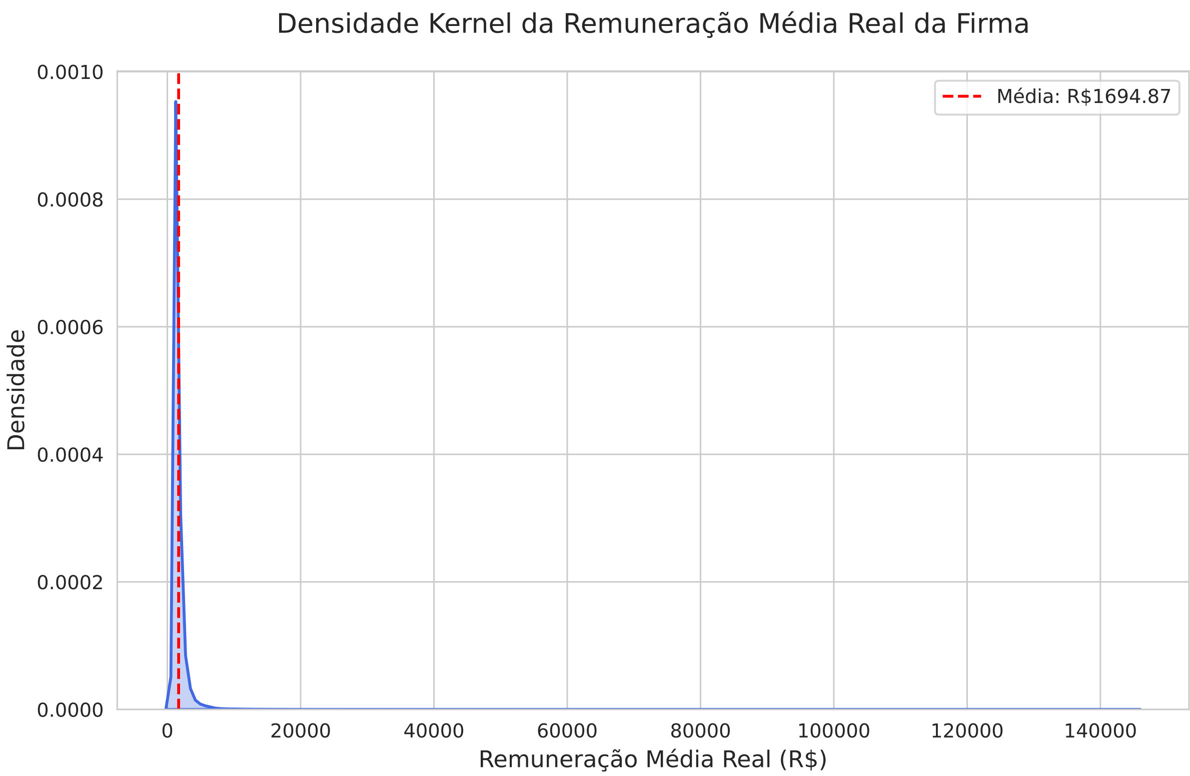

Figura 1: Densidade kernel da remuneração média real da firma

Tabela 2: Estatística Descritiva da Variável Remuneração Média

Estatística	**Valor**
Contagem (count)	$2.575\,873 \times 10^{6}$
Média (mean)	$1.694\,868 \times 10^{3}$
Desvio Padrão (std)	$1.305\,140 \times 10^{3}$
Mínimo (min)	0
Primeiro Quartil (25%)	$1.184\,857 \times 10^{3}$
Mediana (50%)	$1.455\,021 \times 10^{3}$
Terceiro Quartil (75%)	$1.863\,400 \times 10^{3}$
Máximo (max)	$1.457\,806 \times 10^{5}$

A densidade kernel da proporção de funcionários com ensino superior nas firmas (Figura 2) mostra uma distribuição multimodal e assimétrica. Há uma concentração expressiva de firmas com nenhum ou poucos funcionários com ensino superior e alguns picos menores, com firmas com uma proporção maior de funcionários graduados. A cauda alongada à direita revela organizações com proporções próximas a 1, comum em setores tecnológicos ou especializados. A ampla dispersão reflete a diversidade do mercado, desde firmas com quase nenhum profissional graduado até aquelas com predominância de formação superior, possivelmente associada a diferenças setoriais ou estratégias de contratação.

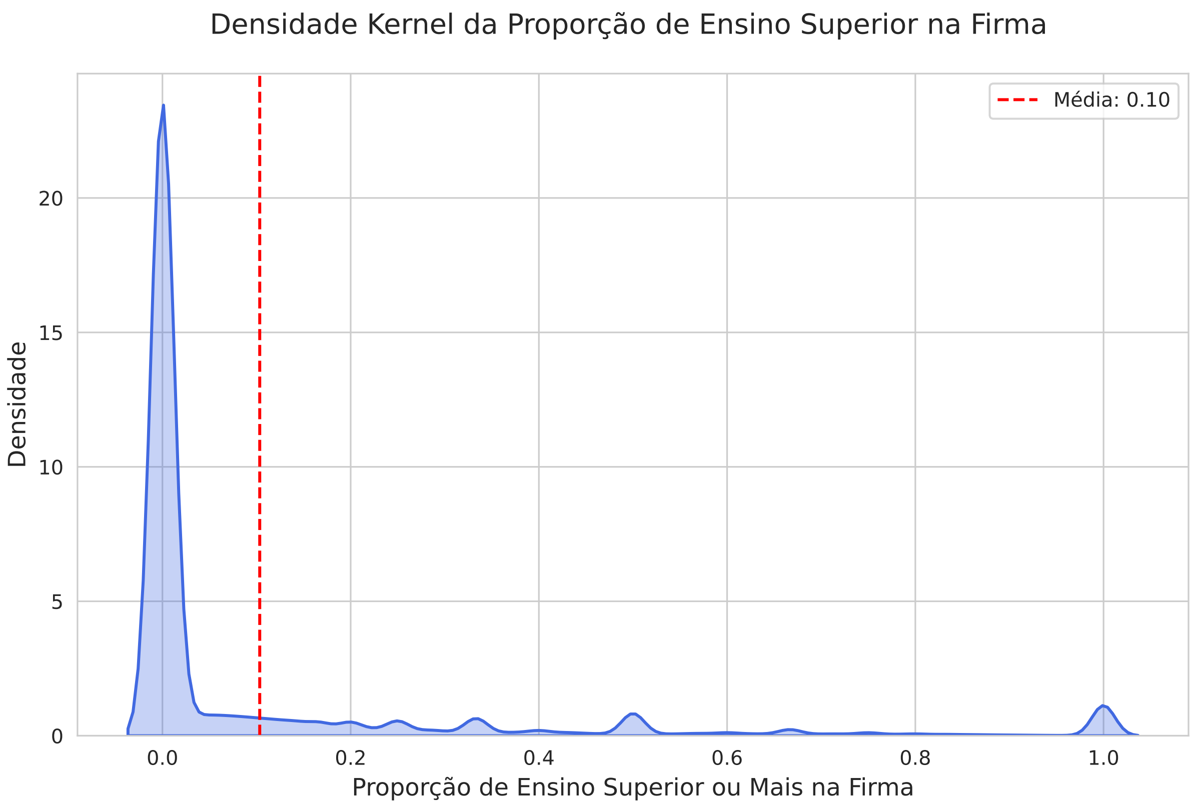

Figura 2: Densidade kernel da proporção de ensino superior ou mais na firma

Tabela 3: Estatística Descritiva da Variável Proporção de Ensino Superior

Estatística	**Valor**
Contagem (count)	$2.575\,873 \times 10^{6}$
Média (mean)	$1.033\,806 \times 10^{-1}$
Desvio Padrão (std)	$2.336\,733 \times 10^{-1}$
Mínimo (min)	0
Primeiro Quartil (25%)	0
Mediana (50%)	0
Terceiro Quartil (75%)	$5.882\,353 \times 10^{-2}$
Máximo (max)	1

A análise da densidade kernel da proporção de trabalhadoras do sexo feminino nas firmas (Figura 3) mostra uma distribuição assimétrica e multimodal, o que revela padrões distintos da participação feminina no mercado de trabalho. No gráfico, percebemos uma concentração de firmas com uma maior participação de trabalhadores do sexo masculino (pico à esquerda da média), enquanto um outro grupo significativo de firmas apresentam uma participação maior de mulheres (pico à direita da média) e no centro da distribuição, vemos alguns picos menores de firmas com uma participação mais equilibrada entre homens e mulheres.

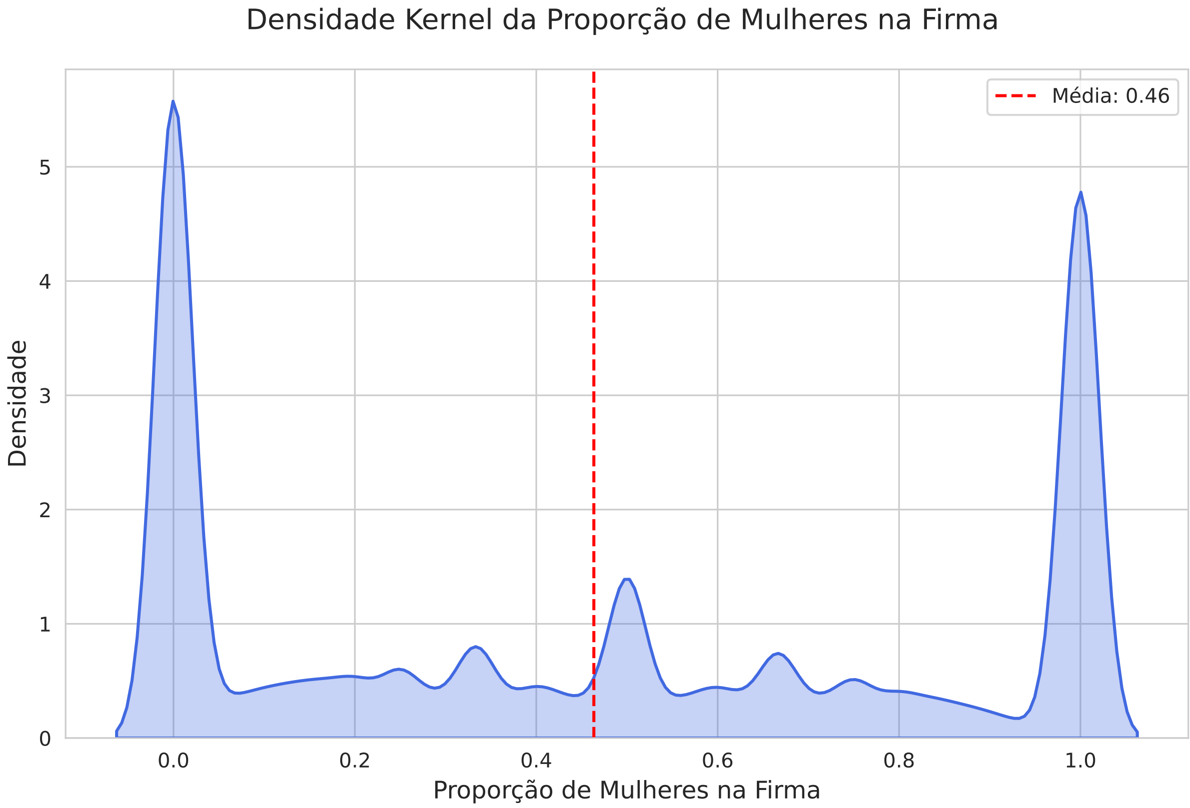

Figura 3: Densidade kernel da Proporção de Trabalhadoras do Sexo Feminino na Firma

Tabela 4: Estatística Descritiva da Variável Proporção de Mulheres

Estatística	**Valor**
Contagem (count)	$2.575\,873 \times 10^{6}$
Média (mean)	$4.634\,875 \times 10^{-1}$
Desvio Padrão (std)	$4.002\,703 \times 10^{-1}$
Mínimo (min)	0
Primeiro Quartil (25%)	0
Mediana (50%)	$4.461\,538 \times 10^{-1}$
Terceiro Quartil (75%)	$9.821\,429 \times 10^{-1}$
Máximo (max)	1

A análise da densidade kernel da proporção de trabalhadores negros nas firmas (Figura 4) revela uma distribuição multimodal e assimétrica, indicando a coexistência de padrões contrastantes de representação racial. O gráfico apresenta três picos principais: o primeiro (à esquerda da média), com um grade número de firmas com baixa participação de negros; um segundo (mais próximo da média), composto por firmas com uma participação mais equilibradas entre brancos e negros; o terceiro (à direita), com firmas formadas majoritariamente por negros.

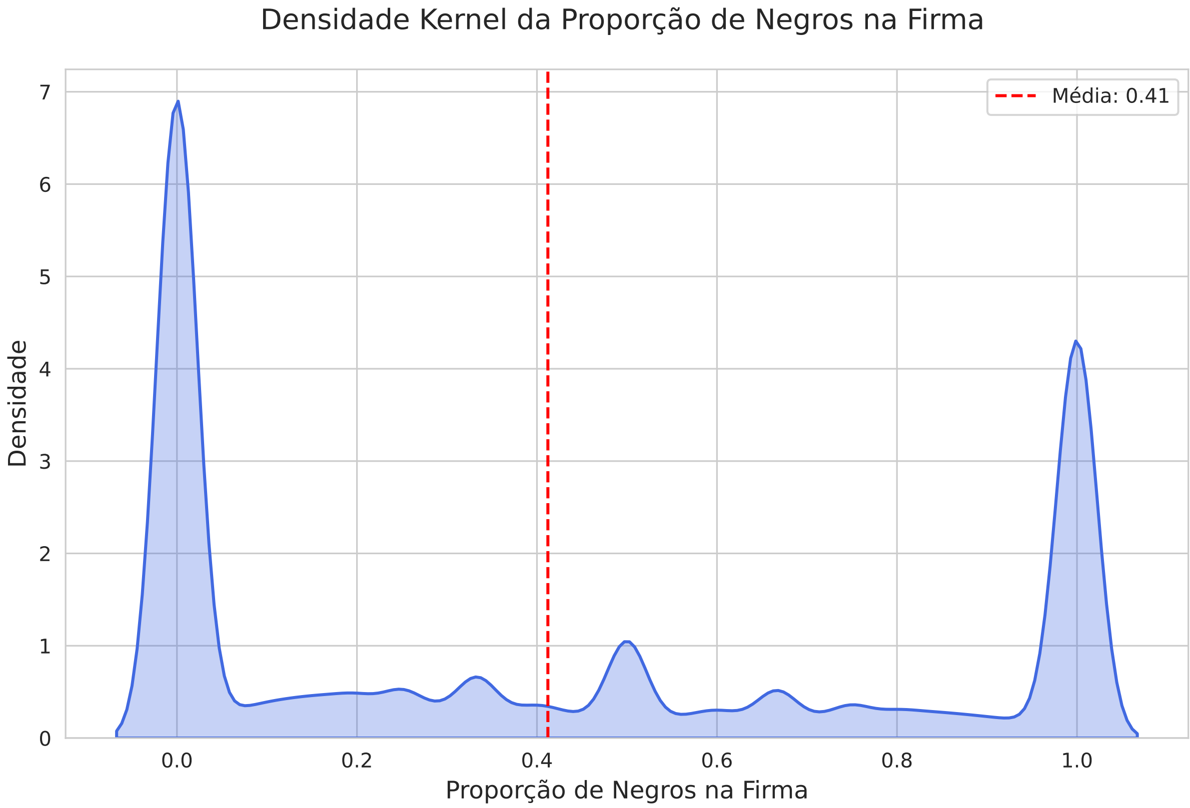

Figura 4: Densidade kernel da Proporção de Negros na Firma

Tabela 5: Estatística Descritiva da Variável Proporção de Negros

Estatística	**Valor**
Contagem (count)	$2.200\,281 \times 10^{6}$
Média (mean)	$4.121\,330 \times 10^{-1}$
Desvio Padrão (std)	$4.146\,033 \times 10^{-1}$
Mínimo (min)	0
Primeiro Quartil (25%)	0
Mediana (50%)	$3.000\,000 \times 10^{-1}$
Terceiro Quartil (75%)	$9.189\,189 \times 10^{-1}$
Máximo (max)	1

A análise da densidade kernel da idade média nas firmas (Figura 5) indica uma distribuição unimodal e aproximadamente simétrica, com o pico centralizado próximo à média de 35,75 anos. Isso sugere que a maioria das empresas possui uma idade média de seus trabalhadores em torno dos 35-36 anos, refletindo certa homogeneidade etária no mercado de trabalho. A ausência de picos secundários ou caudas pronunciadas aponta para uma baixa dispersão em torno da média, o que implica que poucas firmas se desviam significativamente desse valor central.

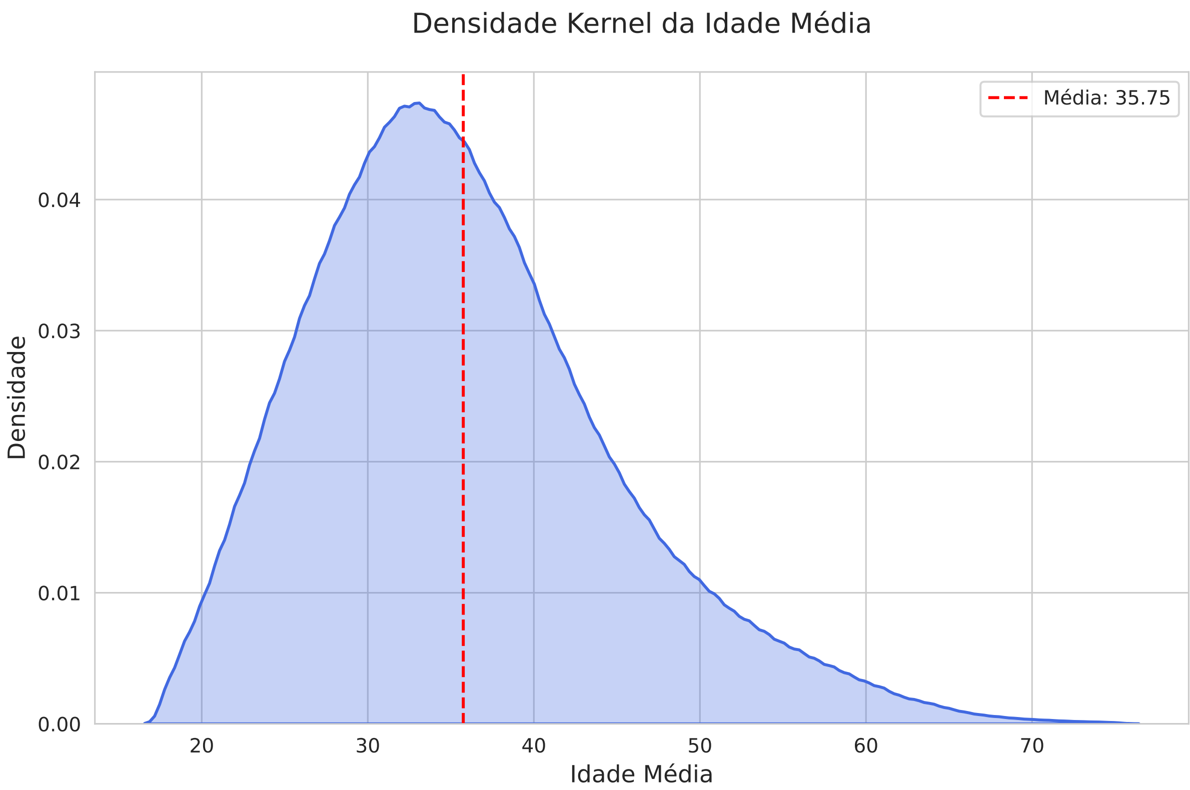

Figura 5: Densidade Kernel da Idade Média na Firma

Tabela 6: Estatística Descritiva da Variável Idade Média

Estatística	**Valor**
Contagem (count)	$2.575\,873 \times 10^{6}$
Média (mean)	$3.575\,006 \times 10^{1}$
Desvio Padrão (std)	$9.192\,819$
Mínimo (min)	$1.800\,000 \times 10^{1}$
Primeiro Quartil (25%)	$2.900\,000 \times 10^{1}$
Mediana (50%)	$3.466\,667 \times 10^{1}$
Terceiro Quartil (75%)	$4.100\,000 \times 10^{1}$
Máximo (max)	$7.500\,000 \times 10^{1}$

A Figura 6 apresenta a densidade kernel do nível técnico das firmas. Observa-se uma alta concentração em valores próximos a zero e um pico em torno de 1, sugerindo a presença de um grupo significativo de firmas com nível técnico próximo à unidade. A média da distribuição é de 0,79, revelando que, em média, as firmas possui um nível técnico inferior a 1. Notam-se ainda alguns picos menores em valores superiores a 1, indicando a existência de outliers ou de subgrupos de firmas com níveis técnicos relativamente mais altos. Essa configuração sugere uma distribuição assimétrica à direita, com a maior densidade concentrada em valores mais baixos.

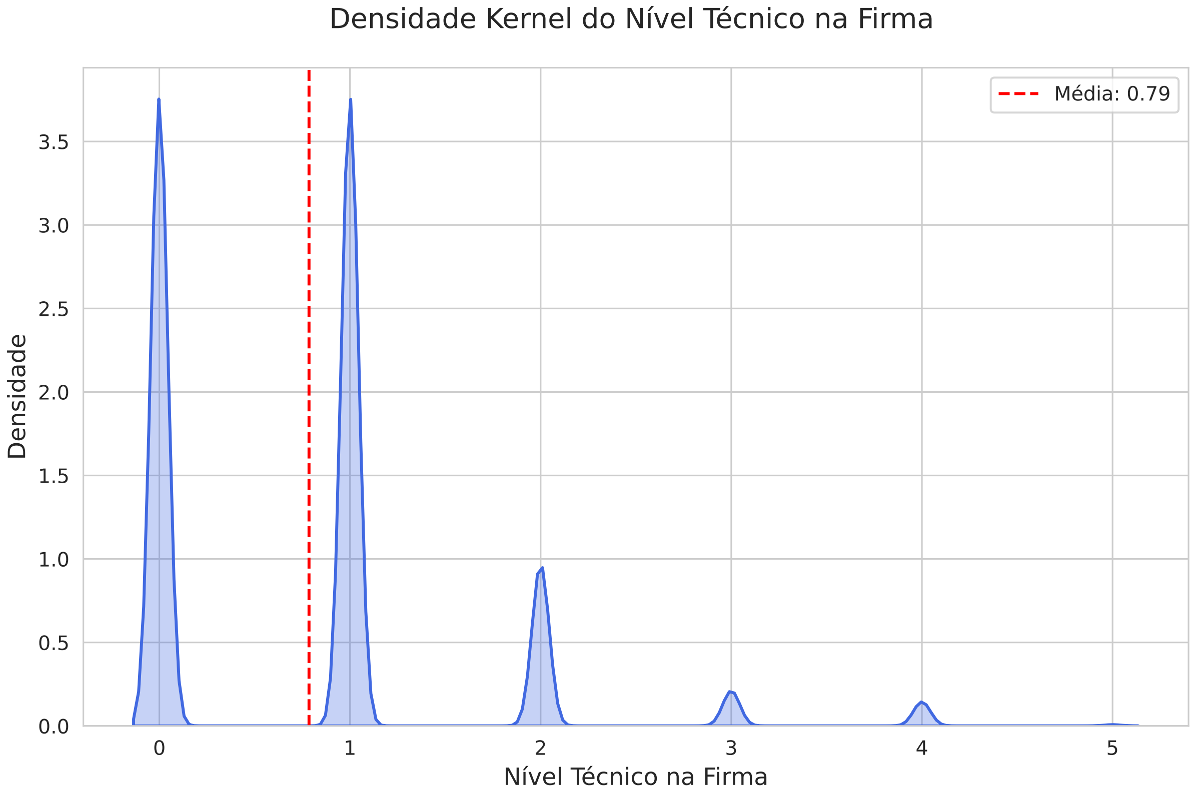

Figura 6: Densidade Kernel do Nível Técnico na Firma

Tabela 7: Estatística Descritiva da Variável Nível Técnico

Estatística	**Valor**
Contagem (count)	$2.575\,873 \times 10^{6}$
Média (mean)	$7.851\,090 \times 10^{-1}$
Desvio Padrão (std)	$8.622\,532 \times 10^{-1}$
Mínimo (min)	0
Primeiro Quartil (25%)	0
Mediana (50%)	1
Terceiro Quartil (75%)	1
Máximo (max)	5

2.2 Análise econométrica

2.2.1 Modelo 1

No primeiro modelo, foi feito uma regressão simples estimando o impacto da variável independente superior_med$_i$ sobre a variável dependente remuneracao$_i$, com a seguinte especificação:

$$\text{remuneracao}_i = \beta_0 + \beta_1 \text{superior\_med}_i + \varepsilon_i \tag{25}$$

Apesar da tabela 8 mostrar que a estimação normal por MQO apresenta heterocedasticidade, dado um p-valor muito baixo, os resultados apresentados são de uma estimação com erros-padrão robustos. Na tabela 9 temos os resultados dessa estimação. Vemos que a variável superior_med$_i$ exerce um impacto positivo na variável remuneracao$_i$. O coeficiente β_1 no valor de 1827.56 indinca que para cada aumento de 1 p.p. na proporção de funcionários com ensino superior na firma, a remuneração média da firma aumenta em R$18,27. Enquanto que o Coeficiente β_0 diz que quando a firma não possui nenhum funcionário com ensino superior, a remuneração média da firma vai ser de R$1505.93. Além disso, $\mathrm{R}^2 = 0,107$.

Tabela 8: Resultados do Teste de Breusch-Pagan para Heterocedasticidade

Estatística	**Valor**
Estatística LM (χ^2)	18913.8987
p-valor (LM)	0
F-estatística	19053.7906
p-valor (F)	0

Tabela 9: Resultado da Regressão Simples do Ensino Superior

Variável	**Coef.**	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Constante	1505.9328	0.592	2544.384	0.000	1504.773	1507.093
Superior Médio	1827.5639	9.273	197.082	0.000	1809.389	1845.739

Além disso, foram calculados os valores dos betas (β_0 e β_1) de forma manual. Para tal, segue abaixo os seguintes resultados:

$$\text{COV}(X, Y) = 99.7909 \tag{26}$$

$$\text{VAR}(X) = 0.0546 \tag{27}$$

$$\text{Média Amostral de } X = 0.1034 \tag{28}$$

$$\text{Média Amostral de } Y = 1694.8675 \tag{29}$$

$$\beta_1 = \frac{\text{COV}(X,Y)}{\text{VAR}(X)} = 1827.56464 \tag{30}$$

$$\beta_0 = \text{Média Amostral de } Y - (\beta_1 \times \text{Média Amostral de } X) = 1505.9328 \tag{31}$$

Todos os coeficientes estimados no modelo apresentam significância estatística ao nível de 5%, uma vez que os p-valores associados estão abaixo desse valor. Isso implica que a hipótese nula, que postula a inexistência de relação entre as variáveis (ou seja, coeficiente igual a zero), é rejeitada para todas as variáveis analisadas.

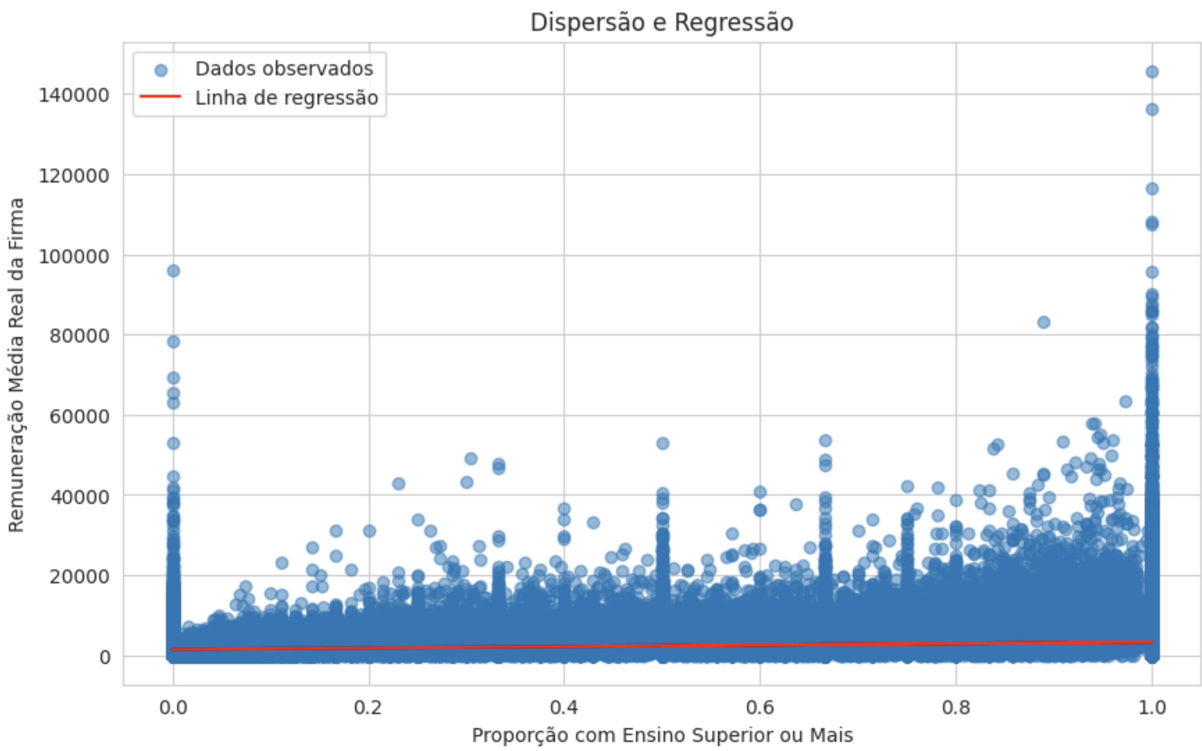

Figura 7: Enter Caption

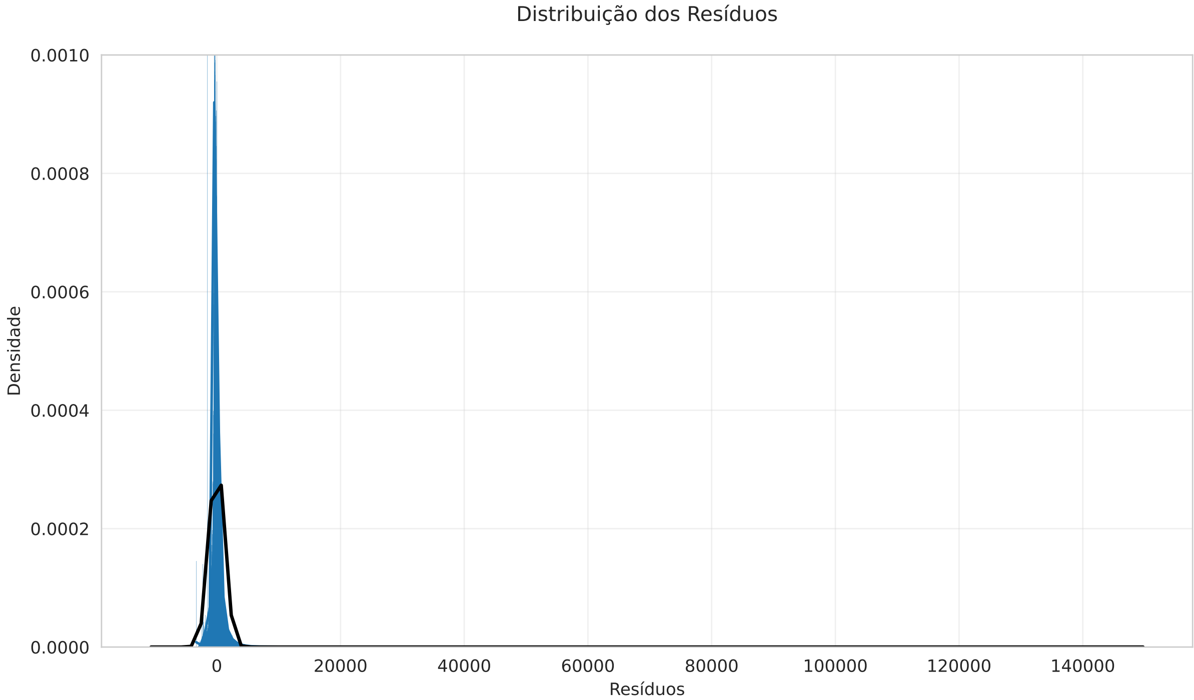

Figura 8: Enter Caption

2.2.2 Modelo 2

No segundo modelo, continuamos com uma regressão simples, mas agora estimando o impacto da variavel independente idade_med_i sobre a variável dependente remuneracao_i, com a seguinte especificação:

$$\text{remuneracao}_i = \beta_0 + \beta_1 \text{idade\_med}_i + \varepsilon_i \tag{32}$$

Na tabela 11 temos os resultados da estimação do segundo modelo. Vemos que a variável idade_med_i exerce um impacto positivo na variàvel remuneracao_i. O coeficiente β_1 no valor de 17.50 indica que para cada aumento de 1 ano na idade média dos funcionários, há um incremento médio de R$17,51 na remuneração média da firma.

Como não faz sentido que a idade média dos trabalhadores da firma seja igual a zero, podemos interpretar o Coeficiente β_0 como o valor da remuneração média da firma quando desconsideramos a idade dos funcinários, que nesse caso seria de R$1068,93. Além disso, o $\text{R}^2 = 0,015$.

Tabela 10: Resultados do Teste de Breusch-Pagan para Heterocedasticidade

Estatística	**Valor**
Estatística LM (χ^2)	1903.9312
p-valor (LM)	0
F-estatística	1905.3380
p-valor (F)	0

Tabela 11: Resultados da Regressão Simples de Idade

Variável	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Constante	1068.9382	3.321	321.832	0.000	1062.428	1075.448
Idade Média	17.5085	0.104	168.691	0.000	17.305	17.712

Todos os coeficientes estimados no modelo apresentam significância estatística ao nível de 5%, uma vez que os p-valores associados estão abaixo desse valor.

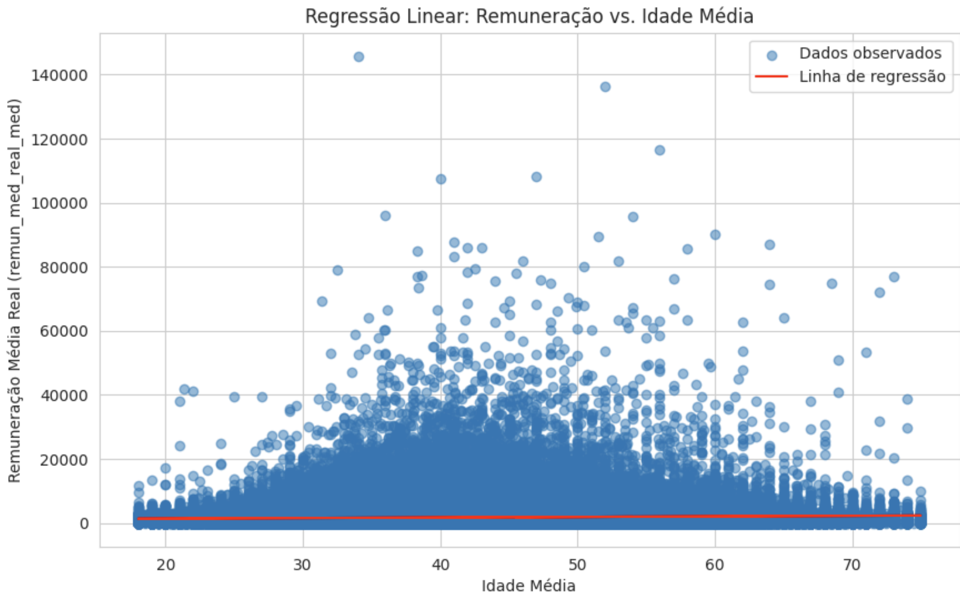

Figura 9: Enter Caption

Análise de Resíduos

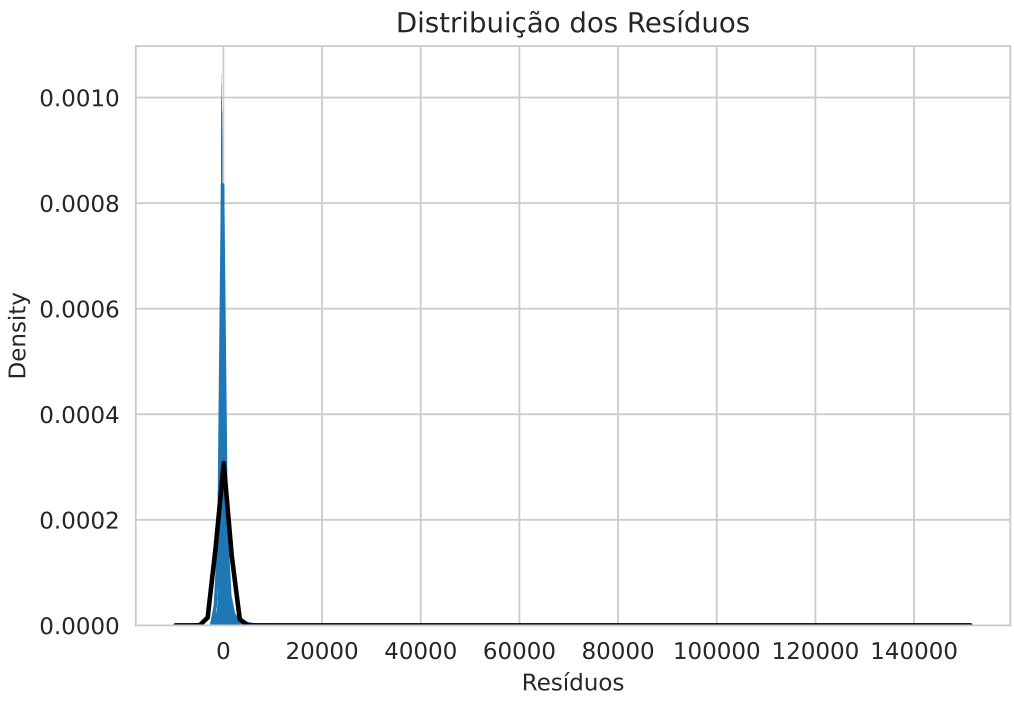

Figura 10: Enter Caption

2.2.3 Modelo 3

No terceiro modelo, introduzimos um modelo de regressão multipla, estimando o impacto das variáveis independentes idade_i e ensinosup_i sobre a variável dependente remuneracao_i, com a seguinte especificação:

$$\text{remuneracao}_i = \beta_0 + \beta_1 \text{idade}_i + \beta_2 \text{ensinosup}_i + \varepsilon_i \tag{33}$$

Na tabela 14 temos os resultados da estimação do terceiro modelo. Vemos que ambas as variáveis independentes, idade_i e ensinosup_i, exercem um impacto positivo na variàvel remuneracao_i.

O coeficiente β_0 estimado indica que na ausência de funcionários com ensino superior e desconsiderando a idade média dos funcionários, a remuneração média da firma é de R$923,64. Além disso, o $\text{R}^2 = 0,12$.

Tabela 12: Teste de Multicolinearidade (Fator de Inflação de Variância - VIF)

Variável	**VIF**
Constante	16.2416
Idade Média (idade_med)	1.0006
Educação Superior (superior_med)	1.0006

Tabela 13: Results of Breusch-Pagan Test for Heteroskedasticity

Test Statistic	**Value**
LM Statistic (χ^2)	20 196.864 36
LM p-value	0.000 00
F Statistic	10 178.225 69
F p-value	0.000 00

Tabela 14: Resultados da Regressão Múltipla Idade Média e Superior

Variável	**Coef.**	**Std.Err.**	**t**	**P>|t|**	**[0.025**	**0.975]**
Constante	923.6424	3.074	300.487	0.000	917.618	929.667
Idade Média	16.3352	0.083	196.819	0.000	16.173	16.498
Superior Médio	1811.1863	3.265	554.712	0.000	1804.787	1817.586

Já o coefiente β_1 mostra que, mantendo constante a proporção de funcionários com ensino superior, para cada ano adicional na idade média dos funcionários, a remuneração média da firma aumenta em R$ 16,34. Enquato que o coeficiente β_2 diz que, mantendo constante a idade média dos funcionários, para cada aumento de 1 p.p. na proporção de funcionários com ensino superior, a remuneração média da firma cresce em R$ 18,11.

Todos os coeficientes estimados no modelo apresentam significância estatística ao nível de 5%, uma vez que os p-valores associados estão abaixo desse valor.

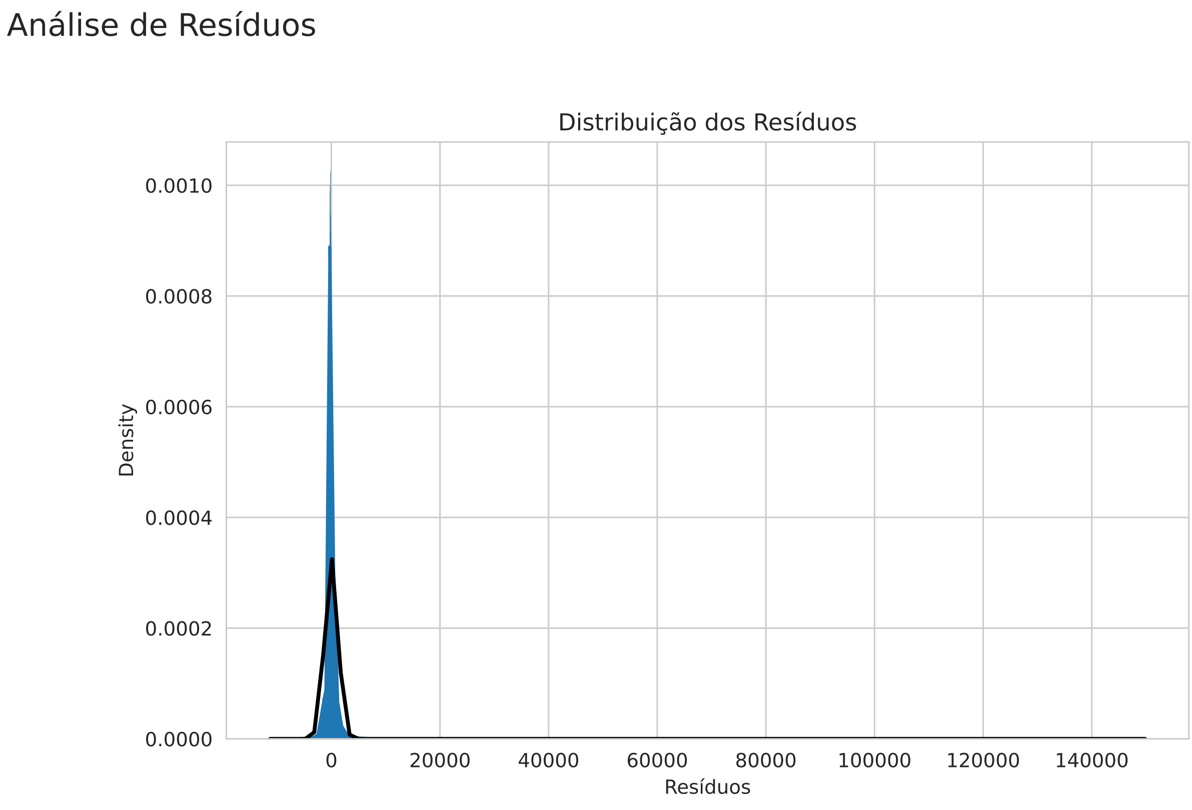

Figura 12: Enter Caption

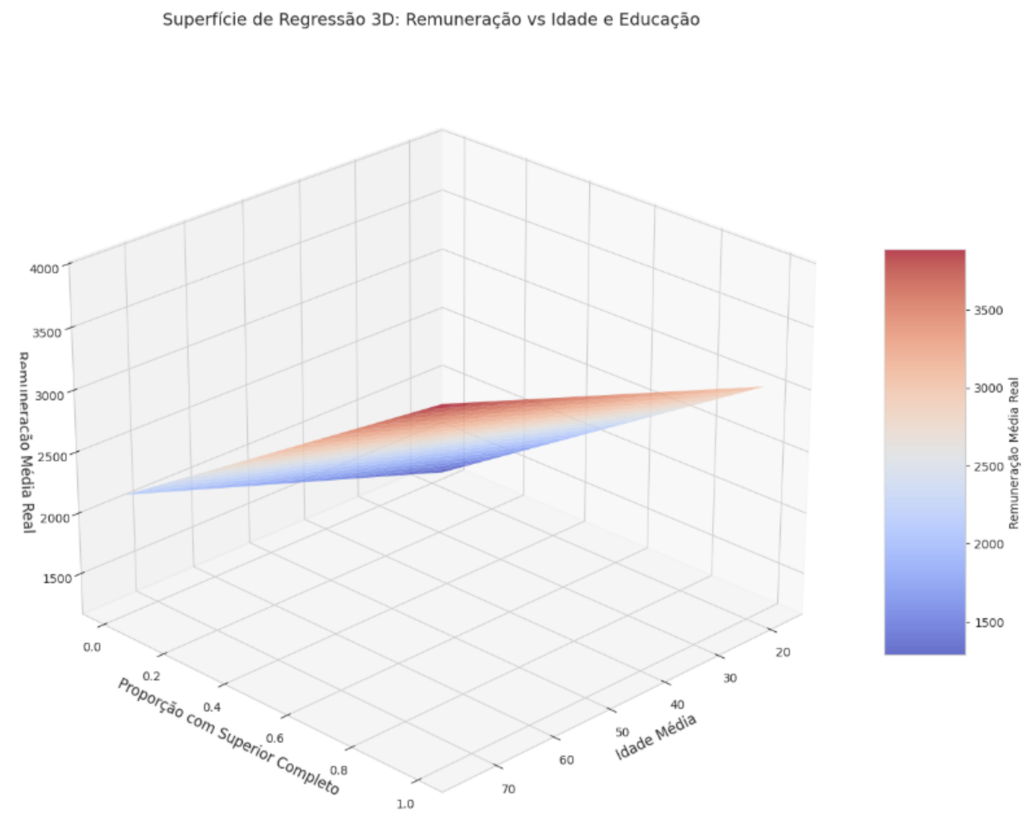

Figura 11: Enter Caption

2.2.4 Modelo 4

No quarto modelo, acrescentamos ao modelo 3 a variável quadrática idade_i^2 para tentar enteder melhor como o avanço da idade dos funcionários impacta na remuneração média da firma.

$$\text{remuneracao}_i = \beta_0 + \beta_1 \text{idade}_i + \beta_2 \text{idade}_i^2 + \beta_3 \text{ensinosup}_i + \varepsilon_i \tag{34}$$

O coefiente β_1 mostra que, mantendo constante as demais variáveis, para cada ano adicional na idade média dos funcionários, a remuneração média da firma aumenta em R$ 76,05. O coeficiente β_2 diz que, mantendo constante a idade média dos funcionários, para cada aumento de 1 p.p. na proporção de funcionários com ensino superior, a remuneração média da firma cresce em R$ 17,74. Além disso, o $\text{R}^2 = 0,125$.

Tabela 15: Variance Inflation Factors (VIF) Analysis

Variable	**VIF Value**
Constant	160.161 68
Average Age (idade_med)	37.695 28
Higher Education (superior_med)	1.009 42
Squared Age (idade_med^2)	37.674 48

Não há interpretação econômica para o coeficiente β_0 negativo. Já o sinal negativo do coeficiente β_3 representa que a medida que a idade média dos funcionários aumenta

Tabela 16: Resultados do Teste de Breusch-Pagan para Heterocedasticidade

Estatística	**Valor**
Estatística LM (χ^2)	20 060.904 33
Valor-p (LM)	0.000 00
Estatística F	6739.444 53
Valor-p (F)	0.000 00

Tabela 17: Resultados da Regressão Múltipla Idade Média, Idade Média ao Quadrado e Superior

Variável	**Coef.**	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Constante	−163.6543	10.352	−15.809	0.000	−183.943	−143.365
Idade Média	76.0579	0.611	124.397	0.000	74.860	77.256
Superior Médio	1774.8594	9.246	191.965	0.000	1756.738	1792.981
Idade Média Quadrado	−0.7662	0.009	−89.818	0.000	−0.783	−0.750

o impacto adicional de cada ano de idade diminui a remuneração média da firma em R$ 0,77.

Todos os coeficientes estimados no modelo apresentam significância estatística ao nível de 5%, uma vez que os p-valores associados estão abaixo desse valor.

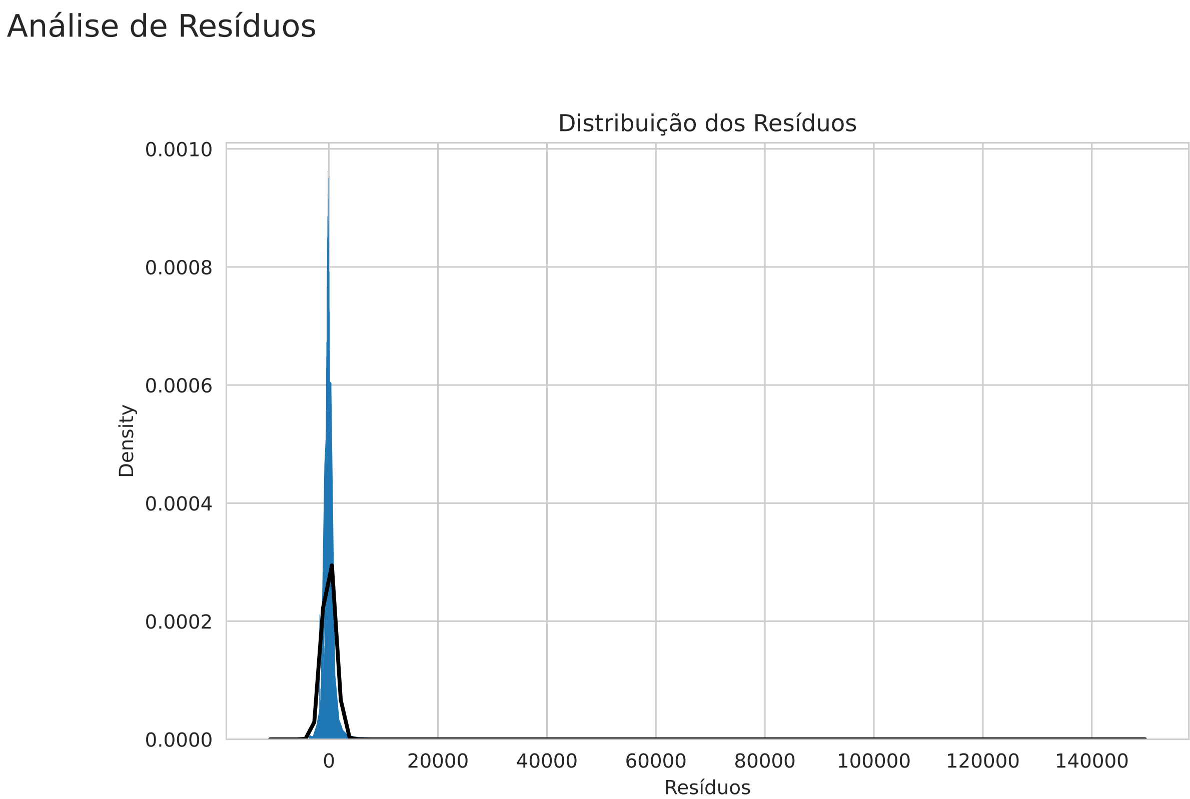

Figura 13: Enter Caption

2.2.5 Modelo 5

No quinto modelo, acrescentamos uma variável de gênero ao modelo 4. De acordo com o coeficiente β_0, quando não há funcionários com ensino superior, nem mulheres na firma e desconsiderando a idade média dos funcionários, a remuneração média da firma é de R$ 148,72. Além disso, o $R^2 = 0,138$.

$$\text{remuneracao}_i = \beta_0 + \beta_1\text{idade}_i + \beta_2\text{idade}_i^2 + \beta_3\text{ensinosup}_i + \beta_4\text{sexo}_i + \varepsilon_i \tag{35}$$

Tabela 18: Análise de Multicolinearidade (Fatores de Inflação de Variância - VIF)

Variável	**VIF**
Constante	164.59577
Idade Média (idade_med)	37.83672
Educação Superior (superior_med)	1.034 38
Idade Média Quadrado (idade_med_quadrado)	37.73680
Sexo (sexo_med)	1.041 23

Tabela 19: Resultados do Teste de Breusch-Pagan para Heterocedasticidade

Estatística	**Valor**
Estatística LM (χ^2)	21000.13542
Valor-p (LM)	0.00000
Estatística F	5293.17697
Valor-p (F)	0.00000

Tabela 20: Resultados da Regressão Múltipla Idade Média, Idade Média ao Quadrado, Superior e Sexo

Variável	**Coef.**	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Constante	148.7193	10.557	14.087	0.000	128.027	169.411
Idade Média	69.9890	0.614	113.902	0.000	68.785	71.193
Superior Médio	1875.1452	9.516	197.062	0.000	1856.495	1893.795
Idade Média Quadrado	−0.7152	0.009	−83.625	0.000	−0.732	−0.698
Sexo Médio	−378.1736	2.029	−186.349	0.000	−382.151	−374.196

O coefiente β_1 mostra que, mantendo constante as demais variáveis, para cada ano adicional na idade média dos funcionários, a remuneração média da firma aumenta em R$ 69,99. O coeficiente β_2 diz que, constante as demais variáveis, para cada aumento de 1 p.p na proporção de funcionários com ensino superior, a remuneração média da firma cresce em R$ 18,75.

O coeficiente β_3 indica que a medida que a idade média dos funcionários aumenta o impacto adicional de cada ano de idade diminui a remuneração média da firma em R$0,71. Já o coeficiente β_4 mostra que um aumento de 1 p.p. na proporção do sexo feminino na firma diminui em R$ 3,78 a renda média da firma.

Todos os coeficientes estimados no modelo apresentam significância estatística ao nível de 5%, uma vez que os p-valores associados estão abaixo desse valor.

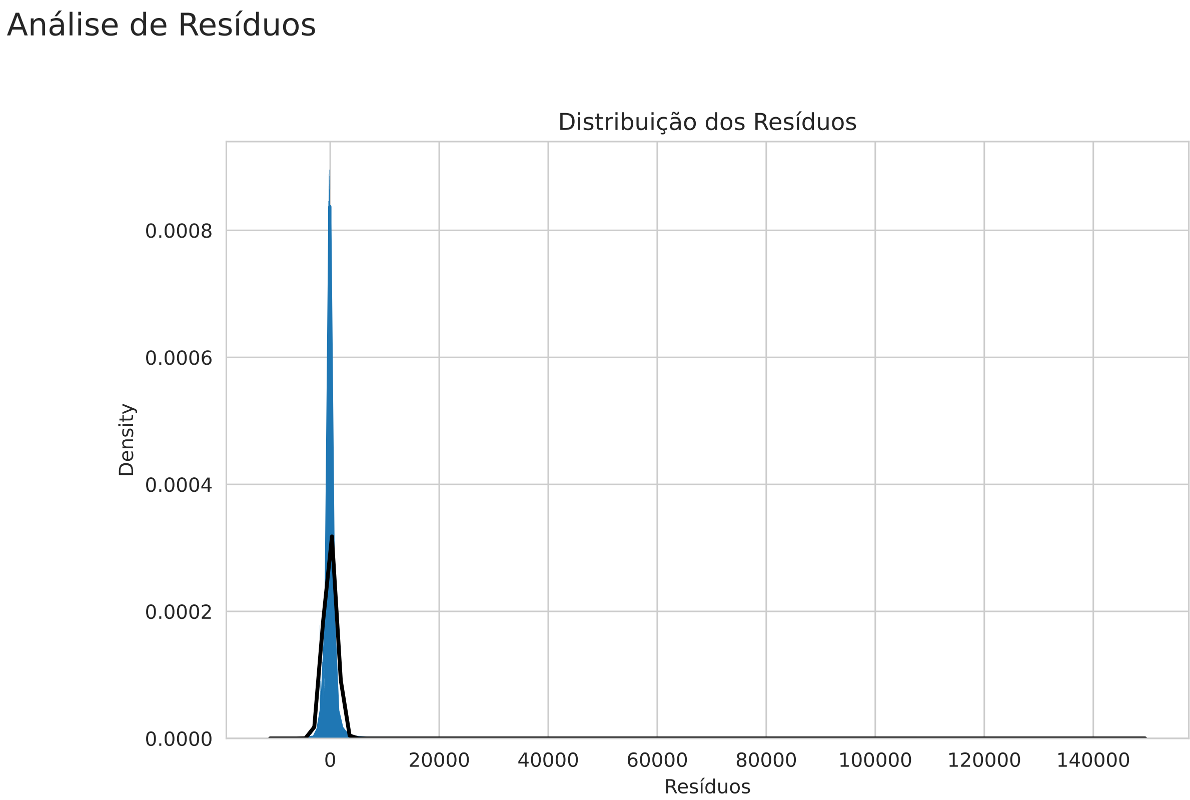

Figura 14: Enter Caption

2.2.6 Modelo 6

No sexto modelo, acrescentamos ao modelo 5 uma variável de raça e uma variável de interação sexo-raça. De acordo com o coeficiente β_0, quando não há funcionários com ensino superior, nem mulheres e nem negros na firma, e desconsiderando a idade média dos funcionários, a remuneração média da firma é de R$ 208,49. Além disso, o $\mathrm{R}^2 = 0,149$.

$$\begin{aligned} \text{remuneracao}_i &= \beta_0 + \beta_1 \text{idade}_i + \beta_2 \text{idade}_i^2 + \beta_3 \text{ensinosup}_i + \beta_4 \text{sexo}_i + \\ &\quad \beta_5 \text{raca}_i + \beta_6 \text{sexoraca}_i + \varepsilon_i \end{aligned} \tag{36}$$

Tabela 21: Análise de Multicolinearidade via Fatores de Inflação de Variância (VIF)

Variável	**VIFa**
Termo constante	175.931 76
Idade média (idade_med)	38.492 50
Educação superior (superior_med)	1.040 38
Idade média ao quadrado (idade_med_quadrado)	38.417 88
Sexo (sexo_med)	1.894 57
Raça/Cor (raca_cor_med)	2.269 38
Interação Sexo-Raça (sexo_raca_interacao)	2.979 75

Tabela 22: Resultados do Teste de Breusch-Pagan para Heterocedasticidade

Estatística	**Valor**
Estatística LM (χ^2)	18 850.709 81
Valor-p (LM)	0
Estatística F	3168.924 45
Valor-p (F)	0

Tabela 23: Resultados da Regressão com Interação

Variável	**Coef.**	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Constante	208.4909	11.975	17.411	0.000	185.020	231.961
Idade Média	76.0207	0.683	111.360	0.000	74.683	77.359
Superior Médio	1948.5084	10.521	185.194	0.000	1927.887	1969.130
Idade Média Quadrado	−0.7997	0.009	−84.736	0.000	−0.818	−0.781
Sexo Médio	−481.0298	3.425	−140.430	0.000	−487.744	−474.316
Raça/Cor Médio	−308.8916	2.936	−105.225	0.000	−314.645	−303.138
Interação Sexo-Raça	135.3052	4.230	31.985	0.000	127.014	143.596

O coefiente β_1 mostra que, mantendo constante as demais variáveis, para cada ano adicional na idade média dos funcionários, a remuneração média da firma aumenta em R$ 76,02. O coeficiente β_2 diz que, mantendo constante as demais variáveis, para cada aumento de 1 p.p. na proporção de funcionários com ensino superior, a remuneração média da firma cresce em R$ 19,48.

O coeficiente β_3 indica que a medida que a idade média dos funcionários aumenta o impacto adicional de cada ano de idade diminui em R$0,80 a remuneração média da firma. Já o coeficiente β_4 mostra que um aumento de 1 p.p. na proporção de mulheres na firma diminui em média R$ 4,81 a remuneração média da firma. Enquanto que o coeficiente β_5 diz que um aumento de 1 p.p. na proporção de negros e indígenas na firma diminui em média R$ 3,08 na remuneração média da firma. O coeficiente da interação entre sexo e raça β_6 mostra que quando o funcionário é simultaneamente do sexo feminino e negro ou indígena, um aumento de 1 p.p. na variável acarreta em diminuição de R$ 1,35 na remuneração média da firma.

Todos os coeficientes estimados no modelo apresentam significância estatística ao nível de 5%, uma vez que os p-valores associados estão abaixo desse valor.

Análise de Resíduos

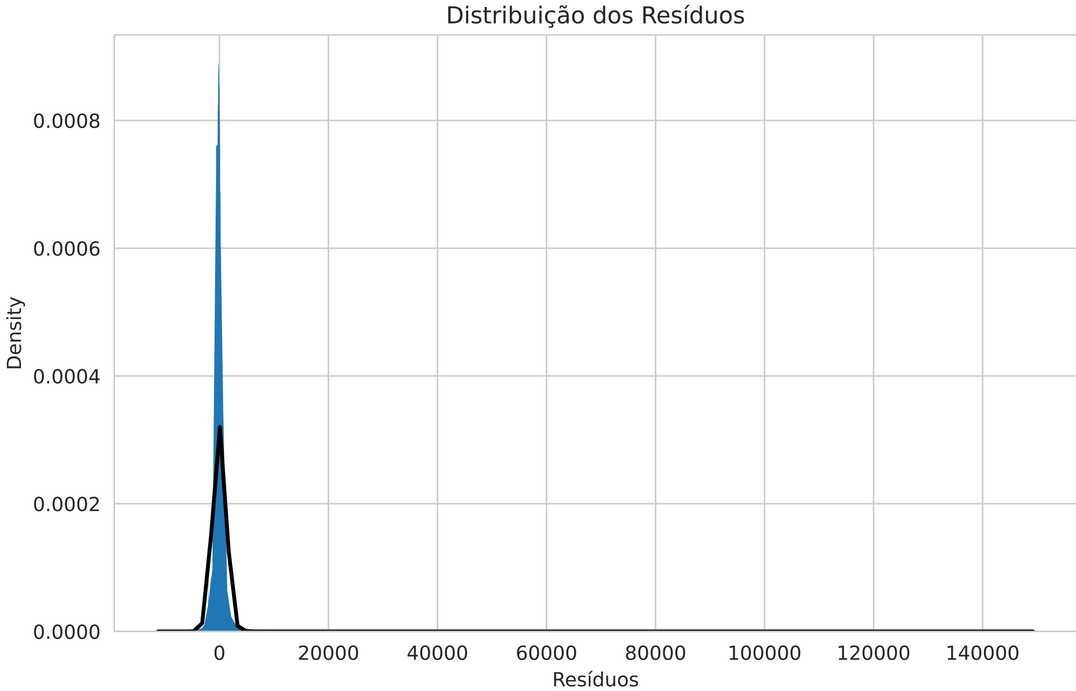

Figura 15: Enter Caption

2.2.7 Modelo 7

No modelo 7, acrescentamos ao modelo 6 algumas variáveis dummies sobre o nivél técnico da firma.

$$\begin{aligned}\ln(\text{remuneracao})_i = {} & \beta_0 + \beta_1 \text{idade}_i + \beta_2 \text{idade}_i^2 + \beta_3 \text{ensinosup}_i + \beta_4 \text{sexo}_i + \beta_5 \text{raca}_i + \\ & \beta_6 \text{sexoraca}_i + \beta_7 \text{tec1}_i + \beta_8 \text{tec2}_i + \beta_9 \text{tec3}_i + \beta_{10} \text{tec4}_i + \beta_{11} \text{tec5}_i + \varepsilon_i\end{aligned} \tag{37}$$

O coeficiente β_1 mostra que, mantendo constante as demais variáveis, para cada ano adicional na idade média dos funcionários, a remuneração média da firma aumenta em 12,44 %. O coeficiente β_2 diz que, mantendo constante as demais variáveis, para cada aumento de 1 p.p. na proporção de funcionários com ensino superior, a remuneração média da firma cresce em 57,96 %. Além disso, o $\mathrm{R}^2 = 0,154$.

O coeficiente β_3 indica que à medida que a idade média dos funcionários aumenta o impacto adicional de cada ano de idade diminui em 0,17 % na remuneração média da firma. Já o coeficiente β_4 mostra que um aumento de 1 p.p. na proporção de mulheres na firma diminui em 14,57 % a remuneração média da firma. Enquanto o coeficiente β_5 diz que um aumento de 1 p.p. na proporção de pessoas negras na firma diminui em 18,01 % na remuneração média da firma. O coeficiente da interação entre sexo e raça (β_6) mostra que quando o funcionário é simultaneamente do sexo feminino e negro ou indígena, um

Tabela 24: Análise de Multicolinearidade via Fatores de Inflação de Variância (VIF)

Variável	**VIF**
Termo constante	176.749 37
Idade média (idade_med)	38.539 15
Educação superior (superior_med)	1.054 23
Idade média ao quadrado (idade_med_quadrado)	38.445 80
Sexo (sexo_med)	1.970 35
Raça/Cor (raca_cor_med)	2.272 74
Interação Sexo-Raça (sexo_raca_interacao)	2.980 79
Nível técnico 1 (nivel_tec_1)	1.262 17
Nível técnico 2 (nivel_tec_2)	1.137 65
Nível técnico 3 (nivel_tec_3)	1.054 62
Nível técnico 4 (nivel_tec_4)	1.041 32
Nível técnico 5 (nivel_tec_5)	1.002 04

Tabela 25: Teste de Breusch-Pagan para Heterocedasticidade

Estatística	**Valora**
Estatística LM	40 818.179 92
p-valor (LM)	0.000 00
Estatística F	3780.863 51
p-valor (F)	0.000 00

Tabela 26: Resultados dos MQO para o modelo completo

Variável	**Coef.**	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Constante	5.1299	0.023	226.126	0.000	5.086	5.174
Idade Média	0.1244	0.001	96.500	0.000	0.122	0.127
Superior Médio	0.5796	0.006	94.132	0.000	0.568	0.592
Idade Média Quadrado	−0.0017	0.000	−97.589	0.000	−0.002	−0.002
Sexo Médio	−0.1457	0.005	−30.230	0.000	−0.155	−0.136
Raça/Cor Médio	−0.1801	0.005	−34.893	0.000	−0.190	−0.170
Interação Sexo-Raça	0.0457	0.008	5.643	0.000	0.030	0.062
Nível Técnico 1	0.0056	0.003	2.138	0.032	0.000	0.011
Nível Técnico 2	−0.0568	0.004	−14.310	0.000	−0.065	−0.049
Nível Técnico 3	−0.0003	0.008	−0.033	0.974	−0.016	0.015
Nível Técnico 4	0.1560	0.010	15.243	0.000	0.136	0.176
Nível Técnico 5	0.2944	0.040	7.392	0.000	0.216	0.373

aumento de 1 p.p. na variável acarreta numa diminuição de 4,57 % na remuneração média da firma.

Os coeficientes de β_7 a β_{11} são os coeficientes de cada uma das variáveis dummies do nível técnico da firma, eles representam o ganho a mais em rumuneração média da firma, em relação as firmas que não possem nenhum nível tecnológico, na medida que avança o nível técnico da firma.

Todos os coeficientes estimados no modelo apresentam significância estatística ao nível de 5%, exceto o nível técnico 1 e 3, uma vez que os p-valores associados estão abaixo desse valor. Além disso, cabe salientar que o modelo 7 possui o maior R^2.

Além disso, todos os modelos foram estimados com erro padrão robusto, ou seja, corrigindo algum tipo de heterocedasticidade presente em todos os modelos.

Análise de Resíduos

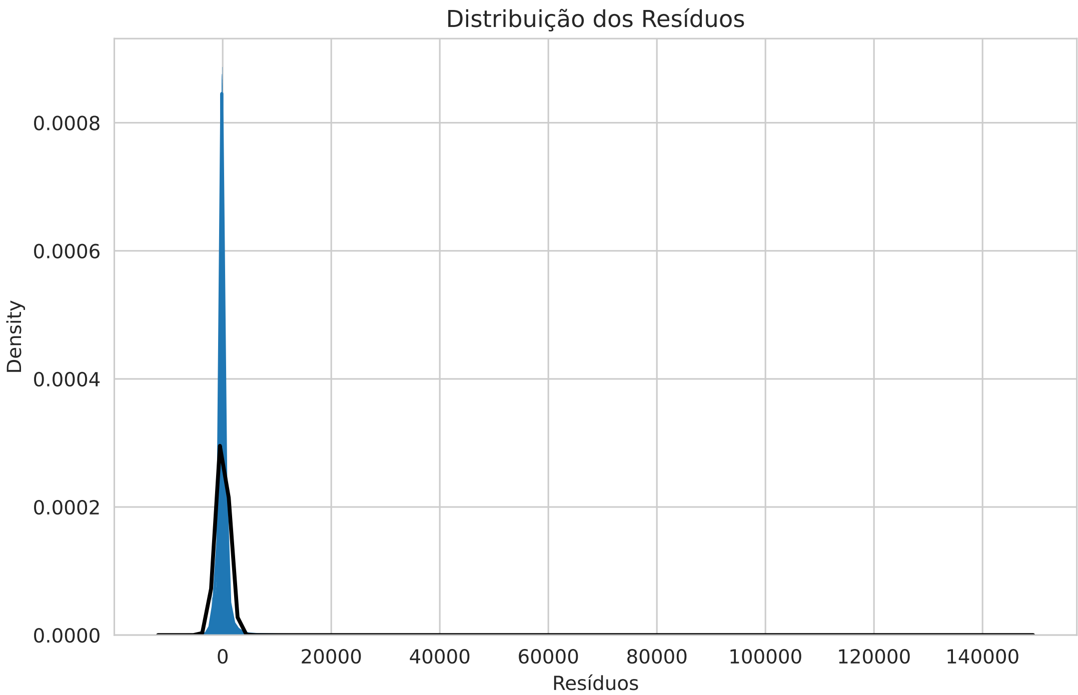

Figura 16: Enter Caption

Referências

[1] Gujarati, D. N., & Porter, D. C.*Basic Econometrics* (5ª ed.).Nova York: McGraw-Hill Education, 2009.

[2] BRASIL. Ministério do Trabalho e Emprego.*Relação Anual de Informações Sociais (RAIS)*. Disponível em: http://rais.gov.br. Acesso em: 10 abr. 2024.

[3] MALTA, Deborah C. et al. Fatores associados ao aumento do consumo de cigarros durante a pandemia da COVID-19 na população

brasileira. Cadernos de Saúde Pública, [s. l.], 2021. Disponível em: https://www.scielo.br/j/csp/a/Ldk3Ppq7Q4bSHt4TmthTyKh/?lang=pt. Acesso em: 8/6/2023.