Trabalho 2 de Econometria

Pedro Henrique Ferreira de Souza e João Paulo de Souza Ferreira

23 de Junho 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 65.601 firmas entre 2011 e 2021 (apenas os anos ímpares) (RAIS, 2021). Outrossim, a escolha do período selecionado (2011-2021) se deve à disponibilidade de dados, já que o período em questão possui as informações necessárias para a construção das variáveis explicativas.

1.2 Estratégia empírica

O objetivo deste trabalho é fazer uma aplicação prática dos seguintes tópicos abordados na disciplina de Econometria I:

- Endogeneidade: Estimadores de Variáveis Instrumentais e Mínimos Quadrados em 2 Estágios
- Modelos para Dados em Painel: Pooled Cross-Sections, Diferenças-em-diferenças, Estimação de Efeitos Fixos e Aleatórios;

Para isso, foi utilizado o Python para manipular e fazer as aplicações com os dados da RAIS, seguindo como base os roteiros do Heiss (2020). Na construção do referencial teórico sobre os tópicos listados acima, utilizamos o Wooldridge (2010).

Da base de dados da RAIS, serão selecionadas as variáveis descritas na Tabela 1 para construir os modelos econométricos.

Com isso, o presente trabalho está estruturado da seguinte forma: após esta descrição metodológica, tem-se uma síntese do referencial teórico utilizado e, em seguida, serão feitas as aplicações e construções dos modelos propostos.

Tabela 1: Variáveis a serem usadas no modelo econométrico

Variável	**Descrição**
y_log_i	Média da remuneração real dos trabalhadores da firma
$idade\_med_i$	Idade média dos trabalhadores da firma
$idade\_med\_quadrado_i$	Idade média dos trabalhadores da firma ao quadrado
$adultos\_med_i$	Proporção de trabalhadores com mais de 34 anos e menos de 55
$superior\_med_i$	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
$nivel\_tec_i$	Dummy que assume 0 caso a firma apresente CNAE sem nível de intensidade, 1 caso a firma apresente baixa tecnologia, 2 caso a firma apresente média-baixa tecnologia, 3 caso a firma apresente média tecnologia, 4 caso a firma apresente média-alta tecnologia e 5 caso a firma apresente alta tecnologia
$nivel\_tec\_classificado_i$	Dummy que assume 1 caso a empresa tenha o nível 5 de tecnologia (alta) e 0 caso contrário (níveis tecnológicos inferiores)
$ano\_classificado_i$	Dummy que assume 1 caso o ano da observação seja 2021 e 0 caso contrário (anos inferiores a 2021)

2 Referêncial Teórico

2.1 Endogeneidade

A endogeneidade em modelos econométricos surge quando uma ou mais variáveis explicativas são correlacionadas com o termo de erro do modelo. Essa correlação viola uma das premissas fundamentais do método dos Mínimos Quadrados Ordinários (MQO), resultando em estimadores inconsistentes e viesados. As principais causas de endogeneidade são:

- Variáveis Omitidas: Quando uma variável relevante que afeta tanto a variável dependente quanto uma variável explicativa é omitida do modelo;
- Erros de Medição: Quando as variáveis explicativas são medidas com erro. Se o erro de medição for correlacionado com a variável explicativa verdadeira, isso introduz endogeneidade;
- Simultaneidade: Ocorre quando há uma relação de causa e efeito bidirecional entre a variável dependente e uma das variáveis explicativas.

2.1.1 Estimadores de Variáveis Instrumentais

Para corrigir a endogeneidade dos modelos econométricos, utilizamos os estimadores de Variáveis Instrumentais (IV), em que buscamos incluir uma variável instrumental (Z) que satisfaça as seguintes condições:

1. Exogeneidade: A variável instrumental deve ser não correlacionada com o termo de erro (u) do modelo estrutural. Matematicamente, Cov(Z, u) = 0. Isso implica que a variável instrumental afeta a variável dependente (Y) apenas através da variável explicativa endógena (X), e não por outros canais diretos ou indiretos que estejam contidos no termo de erro.

2. Relevância: A variável instrumental deve ser correlacionada com a variável explicativa endógena (X). Matematicamente, Cov(Z, X) $\neq$ 0.

Com essas duas condições satisfeitas, os estimadores IV serão consistentes. A intuição por trás do IV é que ele utiliza a variação em X que é exógena, ou seja, que é explicada pelo instrumento Z, para estimar o efeito causal de X em Y, eliminando o viés da endogeneidade.

2.1.2 Mínimos Quadrados em 2 Estágios

Os Mínimos Quadrados em 2 Estágios (MQ2E) são uma extensão do método de Variáveis Instrumentais, aplicável quando há múltiplas variáveis endógenas ou múltiplos instrumentos. O nome dois estágios refere-se à sua implementação em duas etapas de regressão:

- Primeiro Estágio: Cada variável explicativa endógena é regredida em todas as variáveis exógenas do modelo (incluindo os instrumentos e as variáveis exógenas que já estavam no modelo original). O objetivo é obter a parte da variação da variável endógena que é explicada pelos instrumentos e pelas variáveis exógenas;

- Segundo Estágio: A variável dependente (Y) é regredida nas variáveis exógenas originais e nos valores preditos das variáveis endógenas obtidos no primeiro estágio. Este estágio é uma regressão de MQO padrão, mas com as variáveis endógenas substituídas por suas partes exógenas.

Os coeficientes estimados no segundo estágio são os estimadores MQ2E. Com isso, ao usar os valores preditos das variáveis endógenas (que são purgados de sua correlação com o termo de erro), o problema da endogeneidade é resolvido, e os estimadores resultantes são consistentes.

2.2 Modelos para Dados em Painel

Modelos para dados em painel são amplamente utilizados em econometria para analisar dados que combinam dimensões de corte transversal (indivíduos, empresas, países) e séries temporais (períodos de tempo). Essa estrutura de dados oferece vantagens significativas em relação a dados puramente de corte transversal ou séries temporais, principalmente na capacidade de controlar a heterogeneidade não observada e na análise de efeitos dinâmicos.

2.2.1 Pooled Cross-Sections

O método de Pooled Cross-Sections, ou Cortes Transversais Agrupados, é a abordagem mais simples para analisar dados em painel. Ele trata os dados de diferentes períodos como se fossem observações independentes de um único grande conjunto de dados de corte transversal. Essencialmente, todas as observações são agrupadas e uma regressão de Mínimos Quadrados Ordinários (MQO) é aplicada. Este método é apropriado quando a heterogeneidade não observada entre as unidades de corte transversal não é correlacionada com as variáveis explicativas, ou quando o objetivo é estimar um efeito médio sobre a população ao longo do tempo.

2.2.2 Diferenças-em-diferenças

O estimador de Diferenças-em-Diferenças (DiD) é uma técnica utilizada para avaliar o impacto causal de uma intervenção ou política. Ele é particularmente útil quando não é possível randomizar o tratamento e há um grupo de tratamento e um grupo de controle, observados antes e depois da intervenção. A ideia central do DiD é comparar a mudança na variável de resultado no grupo de tratamento com a mudança na variável de resultado no grupo de controle. Isso permite isolar o efeito da intervenção, controlando por tendências temporais comuns que afetariam ambos os grupos.

2.2.3 Estimação de Efeitos Fixos e Aleatórios

Os modelos de efeitos fixos (EF) e efeitos aleatórios (EA) são as abordagens mais comuns para lidar com a heterogeneidade não observada em dados em painel. A escolha entre EF e EA depende da natureza da correlação entre os efeitos não observados e as variáveis explicativas.

O modelo de Efeitos Fixos assume que a heterogeneidade não observada é correlacionada com as variáveis explicativas. Para controlar essa correlação, o EF remove os efeitos específicos da unidade. Isso pode ser feito através dos seguintes métodos:

- Within Transformation (Transformação Intra-grupos): Subtrai a média temporal de cada variável para cada unidade. Isso elimina a heterogeneidade não observada e permite estimar os coeficientes das variáveis explicativas. O es-

timador resultante é conhecido como estimador de Within ou Fixed Effects (FE);

- First Differencing (Primeiras Diferenças): Subtrai a observação do período anterior de cada variável. Isso também elimina os efeitos fixos. O estimador de primeiras diferenças é consistente sob as mesmas condições que o estimador de efeitos fixos, e em painéis com dois períodos, os dois estimadores são numericamente idênticos;
- Dummy Variables (Variáveis Dummy): Inclui uma variável dummy para cada unidade de corte transversal;

O modelo de Efeitos Aleatórios (EA), por outro lado, assume que a heterogeneidade não observada não é correlacionada com as variáveis explicativas. Em vez de remover os efeitos específicos da unidade, o modelo de EA os trata como um componente aleatório do termo de erro. Isso permite que o modelo estime o efeito de variáveis que não variam no tempo, o que é uma vantagem sobre o modelo de Efeitos Fixos. A estimação de Efeitos Aleatórios geralmente é feita usando Mínimos Quadrados Generalizados Factíveis (FGLS), que leva em conta a estrutura de covariância do termo de erro composto (erro idiossincrático mais o efeito aleatório). O estimador de Efeitos Aleatórios é mais eficiente que o de Efeitos Fixos se a suposição de não correlação entre a heterogeneidade não observada e as variáveis explicativas for válida. No entanto, se essa suposição for violada, o estimador de Efeitos Aleatórios será inconsistente.

Para comparar os estimadores de Efeitos Fixos e Efeitos Aleatórios é utilizado o Teste de Hausman. Esse teste verifica se há uma diferença sistemática entre os coeficientes estimados pelos dois métodos. Se a hipótese nula de não correlação entre os efeitos não observados e as variáveis explicativas for rejeitada, o modelo de Efeitos Fixos é geralmente preferido devido à sua robustez à endogeneidade dos efeitos não observados.

Ele também aborda o Teste de Hausman, que é comumente usado para comparar os estimadores de Efeitos Fixos e Efeitos Aleatórios. O teste de Hausman verifica se há uma diferença sistemática entre os coeficientes estimados pelos dois métodos. Se a hipótese nula de não correlação entre os efeitos não observados e as variáveis explicativas for rejeitada, o modelo de Efeitos Fixos é geralmente preferido devido à sua robustez à endogeneidade dos efeitos não observados.

3 Resultados

3.1 Análise descritiva

As tabelas apresentadas abaixo contêm todas as estatísticas descritivas das variáveis utilizadas nos modelos econométricos. Elas fornecem informações fundamentais sobre

a distribuição dos dados, como média, desvio padrão, valores mínimos e máximos, permitindo uma compreensão inicial das características das variáveis analisadas e auxiliando na interpretação dos resultados empíricos.

Tabela 2: Estatísticas Descritivas

Estatística	**y_log**	**adultos_med**	**idade_med**	**superior_med**
Contagem (count)	311862	311862	311862	311862
Média (mean)	7.3615	0.3873	35.2185	0.0871
Desvio Padrão (std)	0.4314	0.2899	7.7821	0.1960
Mínimo (min)	2.4869	0	18	0
Primeiro Quartil (25%)	7.0806	0.1667	29.8000	0
Mediana (50%)	7.2623	0.3636	34.3750	0
Terceiro Quartil (75%)	7.5406	0.5342	39.6667	0.0658
Máximo (max)	11.7546	1	75	1

Tabela 3: Estatísticas Descritivas

Estatística	**sexo_med**	**raca_cor_med**	**nivel_tec_1**	**nivel_tec_2**
Contagem (count)	311862	311862	311862	311862
Média (mean)	0.4632	0.3516	0	0
Desvio Padrão (std)	0.3643	0.3595	0	0
Mínimo (min)	0	0	0	0
Primeiro Quartil (25%)	0.1111	0.2500	0	0
Mediana (50%)	0.4444	0.6316	1	0
Terceiro Quartil (75%)	0.8000	1	1	1
Máximo (max)	1	1	1	1

Tabela 4: Estatísticas Descritivas

Estatística	**nivel_tec_3**	**nivel_tec_4**	**nivel_tec_5**
Contagem (count)	311862	311862	311862
Média (mean)	0.0315	0	0
Desvio Padrão (std)	0.1748	0	0
Mínimo (min)	0	0	0
Primeiro Quartil (25%)	0	0	0
Mediana (50%)	0	0	0
Terceiro Quartil (75%)	0	0	0
Máximo (max)	1	1	1

3.2 Análise Econométrica

Com base nos dados das firmas extraídos da Relação Anual de Informações Sociais (RAIS) para o período de 2011 a 2021, estimou-se um modelo econométrico do tipo Pooled OLS (Tabela 6). Esse modelo combina todas as observações do painel em uma única estrutura de dados, desconsiderando a heterogeneidade individual entre as firmas e ao longo do tempo. Essa abordagem permite obter uma visão geral dos efeitos médios

das variáveis explicativas sobre a variável dependente, ao tratar todas as unidades como homogêneas em relação às características não observadas que permanecem constantes no período analisado.

Tabela 5: Estimativas dos Parâmetros do Modelo (Pooled)

Parâmetro	**Coef.**	**Erro Padr.**	**T-stat**	**P-valor**	**LI (95%)**	**LS (95%)**
const	6.4625	0.0110	588.3100	0.0000	6.4410	6.4840
idade_med	0.0445	0.0006	72.0710	0.0000	0.0433	0.0457
superior_med	1.0839	0.0062	176.1400	0.0000	1.0719	1.0960
idade_med_quadrado	−0.0005	0.0000	−60.6890	0.0000	−0.0005	−0.0005
sexo_med	−0.2590	0.0024	−107.6900	0.0000	−0.2637	−0.2543
raca_cor_med	−0.0087	0.0028	−3.1101	0.0019	−0.0142	−0.0032
sexo_raca_interacao	0.0283	0.0044	6.3705	0.0000	0.0196	0.0370
nivel_tec_1	0.0555	0.0016	34.7310	0.0000	0.0523	0.0586
nivel_tec_2	0.0510	0.0019	26.7210	0.0000	0.0472	0.0547
nivel_tec_3	0.0495	0.0041	12.0630	0.0000	0.0414	0.0575
nivel_tec_4	0.0542	0.0034	15.8900	0.0000	0.0475	0.0608
nivel_tec_5	0.0556	0.0062	8.9282	0.0000	0.0434	0.0678

Interpretando os coeficientes da Tabela 6, cuja variável dependente é a remuneração média real da firma. A idade média dos trabalhadores exerce efeito positivo e significativo sobre a remuneração, sendo que um aumento de um ano na idade média está associado a um acréscimo de 4,45% na remuneração média. No entanto, o termo quadrático da idade média possui coeficiente negativo, indicando que esse efeito ocorre de forma decrescente — ou seja, a relação entre idade e salário é positiva, mas com retornos marginais decrescentes à medida que a idade aumenta.

A proporção de trabalhadores com escolaridade superior completa também está positivamente associada à remuneração, com coeficiente de 1.0839, reforçando a importância do capital humano mais qualificado na determinação dos salários médios das firmas. Por outro lado, a proporção de mulheres na firma está negativamente associada à remuneração, com um coeficiente de -0.2590, o que sugere a presença de desigualdades salariais de gênero. De maneira semelhante, a proporção de trabalhadores negros (pretos, pardos e indígenas) apresenta efeito negativo sobre a remuneração média, ainda que de menor magnitude (-0.0087).

A interação entre proporção de mulheres e proporção de negros na firma mostra um coeficiente positivo (0.0283), o que pode indicar que, em firmas com presença mais significativa conjunta desses dois grupos, os efeitos negativos individuais podem ser, ao menos parcialmente, compensados, resultando em níveis salariais um pouco mais elevados do que o previsto pela soma isolada dos efeitos de gênero e raça.

Por fim, observa-se que todas as categorias de intensidade tecnológica (nível-tec-1 a nível-tec-5) apresentam efeitos positivos e estatisticamente significativos sobre a remuneração média, em comparação com a categoria base (nível tecnológico mais baixo). Os coeficientes variam entre 0.0495 e 0.0556, sugerindo que firmas com maior intensidade

tecnológica tendem a pagar salários médios mais elevados.

Visando mitigar o viés proveniente de heterogeneidade temporal não observada nos resultados da Tabela 5, estimamos também os modelos de efeitos fixos e efeitos aleatórios exibidos, respectivamente, nas Tabelas 6 e 7.

Tabela 6: Estimativas dos Parâmetros do Modelo (Efeitos Fixos)

Parâmetro	**Coef.**	**Erro Padr.**	**T-stat**	**P-valor**	**LI (95%)**	**LS (95%)**
const	6.7884	0.0127	533.9000	0.0000	6.7635	6.8133
idade_med	0.0275	0.0007	38.6580	0.0000	0.0261	0.0289
superior_med	0.2781	0.0076	36.7000	0.0000	0.2632	0.2929
idade_med_quadrado	−0.0003	0.0000	−29.5810	0.0000	−0.0003	−0.0003
sexo_med	−0.0860	0.0043	−19.9560	0.0000	−0.0944	−0.0775
raca_cor_med	−0.0127	0.0038	−3.3292	0.0009	−0.0201	−0.0052
sexo_raca_interacao	0.0143	0.0056	2.5649	0.0103	0.0034	0.0252
nivel_tec_1	0.0009	0.0020	0.4268	0.6695	−0.0031	0.0048
nivel_tec_2	0.0014	0.0021	0.6695	0.5032	−0.0027	0.0055
nivel_tec_3	0.0018	0.0029	0.6243	0.5324	−0.0038	0.0074
nivel_tec_4	0.0024	0.0026	0.9382	0.3482	−0.0026	0.0075
nivel_tec_5	0.0048	0.0036	1.3193	0.1871	−0.0023	0.0119

Ao comparar os resultados da Tabela 6 (Efeitos Fixos) com a Tabela 5 (Pooled), observam-se alterações importantes nos coeficientes e na significância estatística:

- Idade: Os coeficientes de idade_med (0.275) e idade_med_quadrado (-0.0003) permanecem estatisticamente significativos (p-valor < 0.0001), mas suas magnitudes são menores em comparação com o modelo Pooled (0.0445 e - 0.0005 respectivamente). Isso sugere que parte do efeito da idade no modelo Pooled estava capturando a heterogeneidade não observada entre as firmas;

- Escolaridade: O coeficiente de superior_med (0.2781) é positivo e significativo, mas consideravelmente menor do que no modelo Pooled (1.0839). Esta redução na magnitude indica que uma parcela significativa do efeito positivo da escolaridade no modelo Pooled era atribuível a características não observadas das firmas que também estão correlacionadas com a proporção de trabalhadores com ensino superior;

- Gênero e Raça/Cor: Os coeficientes de sexo_med (-0.0860) e raca_cor_med (-0.0127) continuam negativos e estatisticamente significativos, embora com magnetude menor que no modelo de Pooled. A interação sexo_raca_interação permanece positiva (0.0143) e significativa. A persistência da significância, mesmo com magnitudes menores, sugere que, mesmo após controlar a heterogeneidade não observada constante no tempo, ainda existem disparidades salariais relacionadas a gênero e raça/cor que variam ao longo do tempo dentro das firmas.

- Intensidade Tecnológica: as variáveis de intensidade tecnológica no modelo de efeitos fixos não apresentam significanciância estatisticas pois são variáveis dummies que não variam significativamente ao longo do tempo para a mesma firma, ou seja, são quase constantes no tempo.

Tabela 7: Estimativas dos Parâmetros do Modelo (Efeitos Aleatórios)

Parâmetro	**Coef.**	**Erro Padr.**	**T-stat**	**P-valor**	**LI (95%)**	**LS (95%)**
const	6.7197	0.0120	560.2700	0.0000	6.6962	6.7432
idade_med	0.0310	0.0007	45.8760	0.0000	0.0296	0.0323
superior_med	0.4791	0.0076	63.1320	0.0000	0.4642	0.4940
idade_med_quadrado	−0.0003	0.0000	−35.8680	0.0000	−0.0004	−0.0003
sexo_med	−0.1406	0.0035	−40.2100	0.0000	−0.1474	−0.1337
raca_cor_med	−0.0106	0.0034	−3.0966	0.0020	−0.0173	−0.0039
sexo_raca_interacao	0.0118	0.0052	2.2779	0.0227	0.0016	0.0219
nivel_tec_1	0.0359	0.0017	20.8570	0.0000	0.0325	0.0393
nivel_tec_2	0.0359	0.0018	20.0200	0.0000	0.0324	0.0394
nivel_tec_3	0.0358	0.0027	13.3310	0.0000	0.0305	0.0410
nivel_tec_4	0.0372	0.0024	15.8020	0.0000	0.0326	0.0418
nivel_tec_5	0.0389	0.0035	11.0680	0.0000	0.0320	0.0458

Comparando os resultados do modelo de Efeitos Aleatórios (Tabela 7) com o modelo de Pooled (Tabela 5) e o modelo de Efeitos Fixos (Tabela 6), temos:

- Idade: Os coeficientes de idade_med (0.0310) e idade_med_quadrado (-0.0003) são estatisticamente significativos (p-valor < 0.0001). Suas magnitudes estão entre as do modelo Pooled e do modelo de Efeitos Fixos, o que é esperado, pois o EA é uma média ponderada entre os estimadores Between (que se assemelha ao Pooled) e Within (EF) ;

- O coeficiente de superior_med (0.4791) é positivo e significativo (p-valor $<$ 0.0001). Sua magnitude é menor que no Pooled, mas maior que no EF, novamente refletindo a natureza do estimador de Efeitos Aleatórios;

- Gênero e Raça/Cor: Os coeficientes de sexo_med (-0.1406) e raca_cor_med (-0.0106) são negativos e significativos. A interação sexo_raca_interacao é positiva (0.0118) e significativa. As magnitudes desses coeficientes também se situam entre as do Pooled e do EF.

- Intensidade Tecnológica: Diferentemente do modelo de Efeitos Fixos, as variáveis de intensidade tecnológica (nivel_tec_1 a nivel_tec_5) são estatisticamente significativas no modelo de Efeitos Aleatórios. Seus coeficientes variam de 0.0358 (nivel_tec_3) a 0.0389 (nivel_tec_3). Isso ocorre porque o modelo Efeitos Aleatorios não elimina completamente a variação entre as firmas, permitindo a estimação de variáveis que são constantes no tempo. Os coeficientes são menores do que no modelo Pooled, sugerindo que parte do efeito observado

no Pooled era devido à correlação entre o nível tecnológico e a heterogeneidade não observada

Tabela 8: Resultados do Teste (Efeitos Fixos vs Pooled)

Item	**Valor**
Nome do Teste	Teste F (Efeitos Fixos vs Pooled)
Hipótese Nula (H_0)	Efeitos são zero
Estatística F	15.1947
P-valor	< 0.0001
Distribuição	F(63238, 248612)

A Tabela 8 apresenta os resultados do Teste F para comparar o modelo de Efeitos Fixos com o modelo Pooled. A hipótese nula (H_0) deste teste é que os efeitos individuais são todos iguais a zero. Se a H_0 for rejeitada, isso indica que o modelo de Efeitos Fixos é preferível ao modelo Pooled, pois há evidências de heterogeneidade não observada que precisa ser controlada.

Os resultados da Tabela 8 mostram uma Estatística F de 15.1947 com um p-valor < 0.0001. Assim a hipótese nula de que os efeitos individuais são zero é rejeitada. Isso significa que há heterogeneidade não observada significativa entre as firmas que não é capturada pelo modelo Pooled. Consequentemente, o modelo de Efeitos Fixos é estatisticamente superior ao modelo Pooled.

Tabela 9: Comparação dos Modelos de Painel (Efeitos Fixos vs. Efeitos Aleatórios)

Item	**Efeitos Fixos**	**Efeitos Aleatórios**
Variável dependente	y_log	y_log
Estimador	PanelOLS	RandomEffects
Número de Observações	311862	311862
Cov._Est.	Clustered	Clustered
R-squared	0.0589	0.7305
R-Squared_(Within)	0.0589	0.0434
R-Squared_(Between)	0.1765	0.2657
R-Squared_(Overall)	0.1635	0.2403
F-statistic	1413.6000	76850.0000
P-value_(F-stat)	< 0.0001	< 0.0001

A Tabela 11 fornece uma comparação entre os modelos de Efeitos Fixos e Efeitos Aleatórios em termos de número de observações, tipo de estimador, estimativa de covariância, R-quadrado (Within, Between e Overall) e estatística F. Esta tabela é útil para uma visão geral do ajuste e das características de cada modelo antes de aplicar o Teste de Hausman.

As estatísticas F para ambos os modelos são altamente significativas (p-valor $<$ 0.0001). indicando que, em ambos os casos, pelo menos uma das variáveis explicativas é estatisticamente diferente de zero. No entanto, a escolha final entre EF e EA não se baseia apenas no R-quadrado ou na significância geral, mas sim na validade da suposição

de não correlação entre os efeitos não observados e as variáveis explicativas, que é testada pelo Teste de Hausman.

Tabela 10: Resultados do Teste de Relevância do Instrumento

Item	**Valor**
Estatística F do 1º Estágio	6 006 868.7177
Coeficiente do Instrumento	−0.9419
Valor-p do Instrumento	< 0.0001
Relevância do Instrumento	Instrumento forte ($F > 10$)

A Tabela 11 apresenta os resultados do Teste de Relevância do Instrumento do qual foi utilizada a variável Adultos_med como variável instrumental. Este teste é crucial para avaliar a qualidade dos instrumentos utilizados em estimações de Variáveis Instrumentais (IV) ou Mínimos Quadrados em 2 Estágios (MQ2E). A relevância do instrumento refere-se à sua capacidade de explicar a variação na variável explicativa endógena. Um instrumento fraco (pouco correlacionado com a variável endógena e/ou correlacionado com o termo de erro) pode levar a estimadores viesados e ineficientes, mesmo que o instrumento seja exógeno.

Os resultados da tabela 11 confirmam que o instrumento utilizado é forte e relevante, o que é uma condição necessária para a validade das estimativas do MQ2E.

Tabela 11: Resultados do Teste de Hausman

Item	**Valor**
Estatística de Teste	361.2270
P-valor	< 0.0001
Conclusão	Rejeita H_0: 'superior_med' é endógena

A Tabela 11 apresenta os resultados do Teste de Hausman, que é utilizado para comparar os estimadores de Efeitos Fixos (EF) e Efeitos Aleatórios (EA). A hipótese nula (H_0) do Teste de Hausman é que a diferença entre os coeficientes estimados por EF e EA não é sistemática, o que implica que os efeitos não observados não são correlacionados com as variáveis explicativas. Se a H_0 for rejeitada, isso sugere que o modelo de Efeitos Fixos é mais apropriado, pois o modelo de Efeitos Aleatórios seria inconsistente.

Dado que o p-valor é < 0.0001, a hipótese nula é rejeitada. Isso implica que o modelo de Efeitos Aleatórios é inconsistente, e, portanto, o modelo de Efeitos Fixos é o estimador estatisticamente mais relevante.

Tabela 12: Estimativas dos Parâmetros do Modelo (Modelo Estimado com o MQ2E)

Parâmetro	**Coef.**	**Erro Padr.**	**T-stat**	**P-valor**	**LI (95%)**	**LS (95%)**
Intercept	6.7588	0.0294	229.9500	0.0000	6.7012	6.8164
idade_med	0.0262	0.0017	15.0350	0.0000	0.0228	0.0296
idade_med_quadrado	−0.0003	0.0000	−12.9860	0.0000	−0.0003	−0.0002
sexo_med	−0.3440	0.0079	−43.5100	0.0000	−0.3595	−0.3285
raca_cor_med	0.0483	0.0042	11.6410	0.0000	0.0402	0.0565
nivel_tec_1	−0.0337	0.0073	−4.5962	0.0000	−0.0481	−0.0193
nivel_tec_2	−0.0391	0.0075	−5.1952	0.0000	−0.0539	−0.0244
nivel_tec_3	−0.0356	0.0084	−4.2385	0.0000	−0.0521	−0.0192
nivel_tec_4	−0.0316	0.0080	−3.9406	0.0001	−0.0473	−0.0159
nivel_tec_5	−0.0330	0.0105	−3.1332	0.0017	−0.0536	−0.0124
superior_med	2.4751	0.1103	22.4470	0.0000	2.2590	2.6912

A Tabela 12 apresenta as estimativas dos parâmetros do modelo após a aplicação dos Mínimos Quadrados em 2 Estágios (MQ2E). A necessidade de empregar o MQ2E foi reforçada pelos resultados do Teste de Hausman (Tabela 11), que indicou a endogeneidade da variável superior_med.

Ao comparar os resultados do MQE com os modelos anteriores, especialmente o de Efeitos Fixos (Tabela), observam-se mudanças importantes, particularmente no coeficiente da variável superior_med, que foi identificada como endógena:

- Idade: Os coeficientes de idade_med (0.0262) e idade_med_quadrado (-0.0003) permanecem estatisticamente significativos (p-valor < 0.0001), com magnitudes muito próximas às observadas no modelo de Efeitos Fixos. Isso reforça a relação quadrática entre idade e remuneração, indicando que a remuneração aumenta com a idade até um certo ponto e depois diminui.

- Escolaridade: O coeficiente de superior_med é 2.4751 e significativo (p-valor < 0.0001). Este é um aumento substancial em magnitude em comparação com o modelo Pooled (1.0839) e, especialmente, com o modelo de Efeitos Fixos (0.2781). A correção da endogeneidade via MQ2E resultou em um efeito muito mais forte e positivo da proporção de trabalhadores com ensino superior na remuneração. Isso sugere que o viés de endogeneidade estava subestimando o verdadeiro impacto da escolaridade na remuneração;

- Gênero e Raça/Cor: O coeficiente de sexo_med é negativo (-0.3440) e significativo (p-valor < 0.0001). Sua magnitude é maior (em valor absoluto) do que no modelo de Efeitos Fixos (-.0860), sugerindo que, uma vez corrigida a endogeneidade, o efeito negativo da proporção de mulheres na remuneração média da firma é mais pronunciado. O coeficiente de raca_cor_med é positivo (0.0483) e significativo (p-valor < 0.0001). Ressalta-se que nos modelos anteriores (Pooled, EF e EA), o coeficiente de raca_cor_med era negativo. A mudança de sinal e a significância positiva no modelo MQ2E sugerem que,

após o tratamento da endogeneidade, uma maior proporção de trabalhadores negros e indígenas está associada a uma remuneração maior, o que pode indicar que o viés de endogeneidade estava mascarando um efeito positivo ou que a variável instrumental capturou um aspecto não observado que inverteu o sinal.

- Intensidade Tecnológica: As variáveis de intensidade tecnológica (nivel_tec_1 a nivel_tec_5) apresentam coeficientes negativos e estatisticamente significativos (p-valores variando de 0.0000 a 0.0017). Isso contrasta fortemente com os modelos de Pooled (onde eram positivas e significativas) e de Efeitos Fixos (onde não eram significativas). A interpretação desses coeficientes negativos no modelo M2QE sugere que, após controlar a endogeneidade e a heterogeneidade não observada, firmas com maior intensidade tecnológica podem ter remunerações médias menores, o que é um resultado contrário ao esperado, podendo indicar que o intrumento não é tão relevante como o teste aponta.

As Tabelas 13, 14 e 15 foram construídas com o intuito de aproximar a estrutura do modelo de diferenças em diferenças (Dif-in-Dif), permitindo a comparação entre firmas tratadas e não tratadas ao longo do tempo. Foram classificados 2 grupos: (i) firmas classificadas com o nível tecnológico mais elevado (nível 5) (dummy igual a 1), enquanto as demais assumiram dummy igual a 0. A variável dependente adotada é o logaritmo da remuneração média real da firma. O ano de 2021 é considerado o período pós-tratamento, sendo contrastado com os anos anteriores (2011, 2013, 2015, 2017 e 2019), que compõem o período pré-tratamento. Essa abordagem permite avaliar se houve uma mudança diferencial na remuneração média associada à adoção de maior nível tecnológico.

Tabela 13: Resultados da Regressão Pooled OLS (2021)

Parâmetro	**Coef.**	**Erro Padr.**	**T-stat**	**P-valor**	**LI (95%)**	**LS (95%)**
const	7.3280	0.0022	3320.4000	< 0.0001	7.3236	7.3323
nivel_tec_classificado	0.0365	0.0198	1.8384	0.0660	-0.0024	0.0754

Na Tabela 13, estimamos o modelo apenas para o ano de 2021. Observa-se que, nesse período, as firmas com maior nível tecnológico apresentaram uma remuneração média real de 3,65% superior àquela observada nas firmas com menor nível tecnológico. Já na Tabela 14, o mesmo modelo foi estimado considerando os anos anteriores (2011, 2013, 2015, 2017 e 2019), e os resultados indicam uma diferença significativamente maior: nesse período, as firmas tecnologicamente mais avançadas apresentaram, em média, uma remuneração real 9% superior às demais.

Tabela 14: Resultados da Regressão Pooled OLS (2011 até 2019)

Parâmetro	**Coef.**	**Erro Padr.**	**T-stat**	**P-valor**	**LI (95%)**	**LS (95%)**
const	7.3655	0.0008	8878.0000	< 0.0001	7.3638	7.3671
nivel_tec_classificado	0.0900	0.0073	12.4000	< 0.0001	0.0758	0.1042

Na Tabela 15, o modelo foi estimado para todo o período de 2011 a 2021, com a inclusão de duas variáveis adicionais: a dummy ano_classificado, que assume valor 1 para o ano de 2021 e 0 para os demais anos, e a variável de interação tec_ano_interacao, construída a partir da multiplicação entre ano_classificado e nivel_tec_classificado. O coeficiente da variável de interação capta a variação no efeito do nível tecnológico especificamente em 2021, em comparação com os anos anteriores. Os resultados indicam que, em 2021, as firmas com maior nível tecnológico apresentaram uma remuneração média 5,23% inferior às demais em relação ao período pré_tratamento, podendo indicar um efeito suavizador entre a remuneração das firmas com diferentes tecnologias ao passar dos anos.

Tabela 15: Resultados da Regressão Pooled OLS (com interação)

Parâmetro	**Coef.**	**Erro Padr.**	**T-stat**	**P-valor**	**LI (95%)**	**LS (95%)**
const	7.3655	0.0008	8817.0000	< 0.0001	7.3638	7.3671
tec_ano_interacao	-0.0535	0.0204	-2.6272	0.0086	-0.0935	-0.0136
ano_classificado	-0.0375	0.0023	-16.4830	< 0.0001	-0.0420	-0.0330
nivel_tec_classificado	0.0900	0.0073	12.3140	< 0.0001	0.0757	0.1043

Referências

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

[2] HEISS, Florian. **Using R for Introductory Econometrics**.2. ed. [Sl]: Independently published, 2020. 337 p.

[3] PYTHON, Software Foundation. *Python: A programming language for general-purpose programming*. Disponível em: https://www.python.org. Acesso em: jun. 2025.

[4] WOOLDRIDGE, Jeffrey M. **Econometric Analysis of Cross Section and Panel Data**. 2. ed. Cambridge, MA: MIT Press, 2010. 1096 p.