Relatório de Pesquisa - EAE0324

Análise da influência racial sobre o Salário: PNAD COVID

Nome: Guilherme Dias Vianna
NUSP: 9301429
Data de Entrega: 27 de Novembro de 2023
Professora: Solange Ledi Gonçalves
Unidade: FEA - USP

1 Introdução

O presente trabalho tem como objetivo avaliar o efeitos da raça no salário dos brasileiros em um modelo no mesmo espírito de Mincer em (MINCER, 1958). É feita uma análise da influência da raça (e do gênero) sobre a remuneração, controlando por várias covariáveis já apontadas como significativas na literatura. Posteriormente é feito um teste de diagnóstico mais próprio da Estatística do que da Econometria para verificarmos se as hipóteses de modelo linear clássico são válidas.

2 Motivação/Revisão da Literatura

2.1 Diferenças no salário entre raças e gênero

A literatura especializada (tanto nacional quanto internacional) em economia do trabalho tem vários estudos de caso investigando tais relações entre raça/gênero e seus efeitos sobre a remuneração de pessoas comparáveis, citamos por exemplo (SALARDI, 2012), (MADALOZZO, 2010) e a meta-análise de (WEICHSELBAUMER; WINTER-EBMER, 2005)

3 Metodologia

3.1 Escolha do Modelo

Nosso modelo populacional consistirá de uma relação log-lin da forma:

$$\log(\mathbf{y}) = \boldsymbol{\beta}\mathbf{X} + \mathbf{u} \tag{1}$$

Aqui, $\mathbf{y}$ é nossa variável-resposta, $\boldsymbol{\beta}$ nosso vetor de parâmetros, $\mathbf{X}$ nossa matriz de observações amostrais e por fim $\mathbf{u}$ é nosso vetor de erros aleatórios, que está sujeito as hipóteses do **modelo linear clássico**:

$$\mathbf{u} \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{2}$$

$$\mathbb{C}\text{ov}[\mathbf{X}_i, \mathbf{u}] = 0 \tag{3}$$

Com $\boldsymbol{\Sigma}$ uma matriz de variância-covariância desconhecida.

3.2 Escolha de variáveis de controle

Escolhemos o seguinte subconjunto de dados da PNAD Covid para compôr $\mathbf{X}$:

Variável	Descrição	Tipo e valores assumidos
Educação	Nível de educação	Categórica (0 = sem ensino superior, 1 c.c)
Região	Subregião da família	Categórica (0 = Nordeste, 1 = Sudeste)
Situação	Distinção entre o domicílio urbano e rural	Categórica (0 = Rural, 1 = Urbano)
Raça	Cor da pele	Categórica (0 = não-branca, 1 = branca)
Condição	Condição entendida na estrutura do domicílio	Categórica (1 = Responsável pelo domicílio, 0 c.c)
Idade	Idade do entrevistado	Contínua ($\in (0, 130)$)
logsal	logaritmo em base 10 do salário	Contínua ($\in (0, +\infty)$)

Tabela 1: Descrição das variáveis explicativas

Como investigação preliminar, podemos ver como essas variáveis estão correlacionadas:

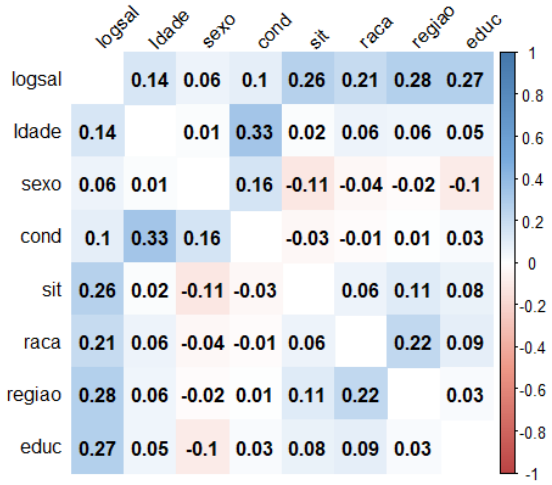

Figura 1: Matriz de correlação entre as variáveis de interesse.

4 Resultados e Teste de Robustez

4.1 Resultados

Nossa regressão gera as seguintes estimativas para cada uma das variáveis (e κ para o intercepto):

Tabela 2:

	Coeficiente	Erro Padrão	p-valor	t
Região	0.142^{***}	(0.002)	$\otimes$	59.322
Raça	0.097^{***}	(0.002)	$\otimes$	39.577
Condição	0.042^{***}	(0.002)	$\otimes$	17.109
Situação	0.210^{***}	(0.003)	$\otimes$	71.134
Educação	0.449^{***}	(0.005)	$\otimes$	87.219
Idade	0.002^{***}	(0.0001)	$\otimes$	24.186
Sexo	0.082^{***}	(0.002)	$\otimes$	34.254
κ	2.671^{***}	(0.005)	$\otimes$	554.150
Número de Observações		78,851		
R^2		0.241		
Estatística F		3570.356^{***}		

Legenda: O símbolo $\otimes$ daqui pra frente é utilizado para sinalizar que a estatística calculada é menor que 10^{-12}

Veja que, dada a hipótese (2), todos as nossas estimativas são estatisticamente significativas a 1% (com certeza o número alto de observações contribui para isso). O sinal dos regressores também mostra um gap salarial (*ceteris paribus*) entre brancos e não brancos (em linha com a literatura). Também é visto um prêmio salarial no ensino superior, também observado na literatura de forma bem recorrente. Raciocínios idênticos seguem para as variáveis de Região e Situação, também corroborando outros efeitos já documentados. Nosso modelo tem um R^2 de 0.241, ou seja este modelo reproduz 24% da variabilidade dos nossos dados.

4.2 Análise de Subgrupos

Separando nossa base por gênero temos para observações de homens:

Tabela 3:

	Coeficiente	Erro Padrão	p-valor	t
Região	0.158***	(0.003)	⊗	51.460
Raça	0.090***	(0.003)	⊗	28.367
Condição	0.077***	(0.003)	⊗	24.155
Situação	0.220***	(0.004)	⊗	62.011
Educação	0.510***	(0.008)	⊗	63.021
Idade	0.002***	(0.0001)	⊗	18.522
κ_h	2.671***	(0.006)	⊗	554.150
Número de Observações		46,008		
R^2		0.266		
Estatística F		2777.565***		

E para mulheres:

Tabela 4:

	Coeficiente	Erro Padrão	p-valor	t
Região	0.116***	(0.004)	⊗	30.819
Raça	0.103***	(0.004)	⊗	27.045
Condição	-0.007*	(0.004)	0.074	-1.787
Situação	0.196***	(0.005)	⊗	37.480
Educação	0.408***	(0.007)	⊗	60.862
Idade	0.002***	(0.0002)	⊗	13.922
κ_m	2.724***	(0.005)	⊗	351.977
Número de Observações		32,843		
R^2		0.211		
Estatística F		1466.090***		

Veja que, embora ambas as tabelas mostrem efeitos similares em termos de significância e sinal para a variável Raça, a variável Condição no caso feminino é menos estatisticamente significante e apresenta uma diminuição da remuneração.

4.3 Envelope Simulado

Por mais que nossos resultados até agora tenham sido estatisticamente significativos, uma boa parte disso se apoia somente em (2), faremos um envelope simulado (nos baseando em (ATKINSON, 1981) e (ATKINSON, 1987), com uma implementação de (EVERITT; HOTHORN, 2009)) para verificar se nosso modelo e hipótese de normalidade são adequados.
O procedimento segue em linhas gerais como:

1. Gerar n observações $\mathcal{N}(0,1)$
2. Rodar regressões como nosso modelo com cada uma dessas observações
3. Obter resíduos padronizados t_i (studentizados) para cada uma das regressões.
4. Repetir os passos (1)-(3) m vezes, gerando os resíduos t^*_{ij}, para $i = 1, \ldots, n$ e $j = 1, \ldots, m$.
5. Ordenar cada grupo de n resíduos para obter $t^*_{(1)j} \leq \ldots \leq t^*_{(n)j}$.

6. Obter os limites $t^*_{(i)l} = \min\{t^*_{(i)1}, \ldots, t^*_{(i)m}\}$ e $t^*_{(i)s} = \max\{t^*_{(i)1}, \ldots, t^*_{(i)m}\}$ e a mediana $t^*_{(i)M} = \text{mediana}\{t^*_{(i)1}, \ldots, t^*_{(i)m}\}$.

7. Juntar $(t^*_{(1)l}, \ldots, t^*_{(n)l}), (t^*_{(1)M}, \ldots, t^*_{(n)M}), (t^*_{(1)s}, \ldots, t^*_{(n)s})$ formando, respectivamente, o limite inferior, a mediana e o limite superior do envelope.

   Se os resíduos observados caírem fora do envelope, isso sugere que o modelo pode estar mal especificado, indicando possíveis violações das hipóteses (como homoscedasticidade, não-normalidade). Isso sugere que devemos ser muito mais céticos com nossas conclusões anteriores

Vejamos que, para n sendo nosso número de observações original e $m = 50$, temos a seguinte distribuição de resíduos e o envelope com 95% de confiança.

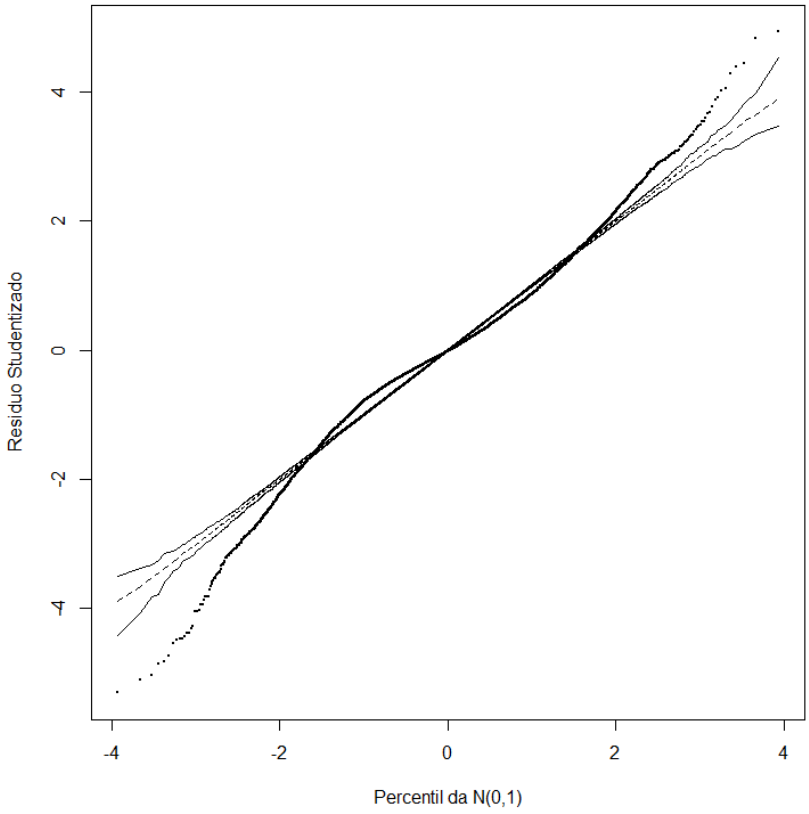

Figura 2: Resultados do envelope simulado.

Veja que, infelizmente nosso modelo parece estar mal especificado, com uma distribuição de resíduos bem pouco dentro do envelope esperado.
Seria interessante observamos outras famílias possíveis para modelar $\mathbf{u}$ e assim recalcularmos a significância estatística das estimativas, para termos uma noção mais acurada do efeito que está sendo observado.

5 Conclusão

Encerramos então com o diagnóstico de que, embora nosso modelo reproduza resultados apoiados pela literatura com uma alta significância, provavelmente persistem problemas de especificação no modelo que precisarão ser analisados com mais calma e que deixam as conclusões anteriores bem menos certeiras. São necessárias então novas análises com modelos mais robustos e outras hipóteses sobre os erros para que obtenhamos estimativas e p-valores robustos, caso os mesmos sobrevivam à novas rodadas de testes como os feitos aqui.

Referências

ATKINSON, A. *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. [S.l.]: Oxford University Press (Oxford Statistical Science Series), 1987.

ATKINSON, A. C. Two graphical displays for outlying and influential observations in regression. *Biometrika*, [Oxford University Press, Biometrika Trust], v. 68, n. 1, p. 13–20, 1981. ISSN 00063444. Disponível em: ⟨http://www.jstor.org/stable/2335801⟩.

EVERITT, B.; HOTHORN, T. *A Handbook of Statistical Analyses Using R*. [S.l.]: [Chapman and Hall], 2009.

MADALOZZO, R. Occupational segregation and the gender wage gap in brazil: an empirical analysis. *Economia Aplicada*, Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto da Universidade de São Paulo, v. 14, n. 2, p. 147–168, Apr 2010. ISSN 1413-8050. Disponível em: ⟨https://doi.org/10.1590/S1413-80502010000200002⟩.

MINCER, J. Investment in human capital and personal income distribution. *Journal of Political Economy*, v. 66, n. 4, p. 281–302, 1958. Disponível em: ⟨https://doi.org/10.1086/258055⟩.

SALARDI, P. Wage disparities and occupational intensity by gender and race in brazil: An empirical analysis using quantile decomposition techniques. In: . [s.n.], 2012. Disponível em: ⟨https://api.semanticscholar.org/CorpusID:16399428⟩.

WEICHSELBAUMER, D.; WINTER-EBMER, R. A meta-analysis of the international gender wage gap. *Journal of Economic Surveys*, v. 19, n. 3, p. 479–511, 2005. Disponível em: ⟨https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0950-0804.2005.00256.x⟩.