Economia Aplicada, v. 26, n. 1, 2022, p. 101-116

VIÉS DA LOG-LINEARIZAÇÃO: ESTIMANDO O RETORNO DA EDUCAÇÃO ATRAVÉS DE REGRESSÕES QUANTÍLICAS

Wallace Souza [*]
Erik Figueiredo [†]
Ana Cláudia Annegues [‡]
Marianne Zwilling Stampe [§] [¶]

Resumo

O presente estudo propõe estimar o retorno da educação sobre o salário dos trabalhadores no Brasil, através de regressões quantílicas. A estimação via MQO da equação de Mincer (1974) log-linearizada, tradicionalmente presente na literatura, pode gerar um viés de especificação resultado da Desigualdade de Jensen, a qual postula que a esperança do logarítmo de uma variável difere do logarítmo da sua esperança. As estimativas para a mediana bem como para a média dos quantis (regressão quantílica) apresentaram coeficientes menores que as estimativas por MQO, indicando uma possível superestimação do retorno da educação na média. Por fim, foi observado que a educação gera maiores ganhos salariais para os quantis superiores de renda.

Palavras-chave: retornos à educação, log-linearização, regressões quantílicas.
Códigos JEL: F1, C1.

Abstract

The present study proposes to estimate returns to education on Brazilian workers' wages through quantile regression. OLS estimation of log-linearized Mincer (1974) equation, which is traditionally present in literature, can generate a specification bias resulting from Jensen's Inequality, which postulates that the expected value of a variable's logarithm differs from the logarithm of its expected value. Median estimates, as well as the mean of quantile coeficients, presented lower coefficients than OLS estimates, indicating a possible superestimation on education returns in the mean. Ultimately, we observed that education generates bigger wage gains for upper income quantiles.

Keywords: returns to education, log-linearization, quantile regression.
JEL codes: F1, C1.

DOI: http://dx.doi.org/10.11606/1980-5330/ea147299

[*] Departamento de Economia, Universidade Federal da Paraíba. E-mail: wpsfarias@gmail.com
[†] Departamento de Economia, Universidade Federal da Paraíba. E-mail: eafigueiredo@gmail.com
[‡] Economista, Universidade Federal da Paraíba. E-mail: annegues.ana@gmail.com
[§] Departamento de Economia, Universidade do Estado de Santa Catarina. E-mail: maristampe@gmail.com
[¶] A autora agradece o apoio da FAPESC

Recebido em 18 de junho de 2018 . Aceito em 27 de julho de 2020.

1 Introdução

Um dos principais pressupostos dos métodos de estimação pela média condicional da variável dependente, como o método de mínimos quadrados - Ordinary Least Squares (OLS) - , é a linearidade do modelo com relação aos parâmetros. Em razão disso, uma forma recorrente de se estimar modelos não lineares é tomar o logaritmo das variáveis e assim linearizar a equação de regressão. Alguns exemplos são a estimação de funções de produção log-linearizadas, tradicional na literatura de crescimento econômico [ver Mankiw et al. (1992) e Duffy e Papageorgiou (2004)], e da equação de salários minceriana na literatura de retorno da educação [ver Sachsida et al. (2004), Maciel et al. (2001), Maciel, Campêlo e Raposo (2010), Coelho et al. (2010), Ramos & Reis (2009) e Reis e Ramos (2011)].

Contudo, Silva & Tenreyro (2006) e Arshad et al. (2016) verificam que a estimação de modelos log-linearizados via OLS compromete a interpretação dos parâmetros na presença de heterocedasticidade em decorrência da Desigualdade de Jensen, a qual postula que o valor esperado do logaritmo de uma variável aleatória é diferente do logaritmo do seu valor esperado ($E(\ln y) \neq \ln E(y)$). Em outras palavras, a estimação pela média condicional de uma transformação logarítmica de modelos multiplicativos exponenciais carregaria um viés de especificação às estimativas. Embora a Desigualdade de Jensen seja amplamente conhecida, tal fato tem sido negligenciado sistematicamente nessas aplicações econométricas.

Diante deste problema, Silva & Tenreyro (2006) propõem o estimador Pseudo Poisson Maximum Likelihood (PPML) como a melhor estratégia para estimação de modelos não lineares. Figueiredo, Lima e Schaur (2016) mostram, porém, que a principal hipótese de identificação do PPML, sobre o componente aleatório, não leva à identificação do modelo após a transformação logarítmica. Estes autores, então, sugerem a estimação de modelos log-lineares pelo método de regressões quantílicas devido à propriedade de equivariância dos quantis, a qual estabelece que para uma função crescente, o logaritmo da esperança é igual a esperança do logaritmo dessa função. Assim, a identificação do modelo linearizado conduziria à identificação do modelo não linear[1].

O uso de regressões quantílicas em comparação com estimações por OLS tem sido crescente na literatura. Por exemplo, Maiti (2019) mostra que decisões baseadas em OLS para dados financeiros, que possuem caudas longas, geralmente leva a decisões erradas em razão de o OLS ser baseado no valor médio das covariadas e, portanto ineficaz nas distribuições finais, sendo indicada a regressão quantílica ao permitir avaliar os valores extremos da distribuição.

Na área da saúde, Hossain & Majumder (2019) utilizaram a regressão quantílica para estimar os determinantes da idade da mãe no primeiro nascimento, indicando que esse modelo produz estimativas menos viesadas do que o modelo de regressão linear quando os dados não seguem a distribuição normal. Essa mesma ideia é utilizada como argumento de uso da regressão quantílica

[1] Não obstante, ainda um forte apelo das estimativas de mínimos quadrados, segundo Angrist et al. (2006), é a sua robustez à má especificação do modelo, ou seja, as estimativas OLS fornecem a melhor aproximação linear da esperança condicional. Essa aproximação não é garantida para o quantil condicional, embora a regressão quantílica seja uma boa alternativa para a estimação de modelos log-lineares na presença de heterocedasticidade, dada a sua invariância a transformações monótonas.

por Esteves et al. (2020) para estabelecer valores de referência apropriados para a condução do nervo da extremidade superior.

Ainda na área da saúde, Staffa et al. (2019) apresentam vantagens do uso da regressão quantílica em comparação com OLS, realizando aplicações ilustrativas para exemplos clínicos de anestesiologia. Um dos aspectos relevantes salientado pelos autores é que muitas vezes os pressupostos do modelo de OLS de que os resíduos são normais, homocedásticos e não correlacionados falham. Em alguns casos as variáveis podem ser transformadas, porém nem sempre isso é possível. Dessa forma, os autores sugerem o uso da regressão quantílica, tendo em vista que os quantis possuem propriedades da equivalência (Hao et al. 2007, Schultz 1961); Hao e Naiman, 2007; Hosseini, 2019) as transformações monotônicas (como linear ou logarítmicas), que mantém os dados na mesma ordem ascendente ou descendente podendo ser implementadas sem alterar os quantis estimados.

Na literatura sobre o retorno de investimentos educacionais[2], a regressão quantílica é geralmente utilizada para avaliar a desigualdade de renda (Cheruiyot 2019, Neves & Lima 2019, Rodrigues et al. 2014, Sampaio 2009), sendo pouca atenção dada aos problemas da estimação por OLS. Uma importante referência a esta discussão é o estudo de Arshad et al. (2016), que estimam retornos da educação para o Paquistão através da regressão quantílica. Os autores mostram que com a variável dependente y na forma de logaritmo natural não é possível aplicar OLS para estimar os efeitos marginais sobre a variável dependente em nível, dado que $\log[E(y|x) \neq E[\log(y)|x]$. Ao estimar os resultados por quantis é possível obter o efeito marginal sobre y pela propriedade de equivariância monotônica dos quantis apresentada por Hao et al. (2007).

No Brasil, problemas de identificação da equação minceriana já foram evidenciados por Moura (2008), embora muitas publicações subsequentes não considerem esse ponto em suas estimações. Tendo por base resultados encontrados para os EUA [Hungerford & Solon (1987), Heckman et al. (1996, 2006), Jaeger & Page (1996)], o estudo realizou testes de identificação (linearidade nos anos de estudo) rejeitando tal hipótese, e ainda constatou que as estimações mincerianas por OLS apresentam viés de superestimação.

Prieto-Rodriguez et al. (2008) também chama a atenção para o fato de que estimativas por OLS não são bons preditores de retorno da educação, uma vez que uma proporção importante das estimativas do quantil não está dentro dos intervalos de confiança das estimativas do OLS. Assim, a regressão quantílica tem vantagens importantes sobre o OLS para estimar os retornos da educação devido à grande disparidade na distribuição salarial. Okamoto (2016) mostra que estimações de modelos na qual a variável dependente está em logaritmo natural e as variáveis explicativas estão em nível (como é o caso da equação minceriana), a estimação por OLS pode ser inadequada mesmo corrigindo-se a heterocedasticidade. Como alternativa, o autor propõe um modelo duplo Pareto-Lognormal (dPLN), salientando a propriedade de equivalência e a heterogeneidade da população.

[2] A ideia de educação como sendo um investimento em capital humano surgiu nas décadas de 50 e 60, sendo Schulz (1961) e Becker (1962) os primeiros autores que associaram ganhos mais elevados com educação para os EUA e também indicaram a educação como sendo um limitador para o crescimento dos países pobres. Seguindo essa nova visão da educação, Castro (1970) e Langoni (1974) realizaram as primeiras estimativas de retorno da educação para o Brasil, indicando que o investimento em capital humano apresenta altas taxas de retorno da educação no país.

É sabido que os países em desenvolvimento apresentam elevados retornos da educação. No Brasil, o tema ganha importância pela alta desigualdade de renda e a heterogeneidade entre as regiões, o que pode causar variações entre os retornos dos investimentos em educação nas diferentes macrorregiões brasileiras. Lam & Levison (1992), por exemplo, estimam o retorno da educação entre homens do Brasil e Estados Unidos, chegando a um valor entre 15-16% para os homens no Brasil contra 10-11% nos Estados Unidos. Já para as regiões brasileiras, Dalcin (2015) estimam o retorno da educação com base nos dados da PNAD para 2003 e 2013 e encontram uma grande disparidade entre as regiões e quantis de rendimentos.Alguns trabalhos, como Sampaio (2009), Rodrigues et al. (2014) e Neves & Lima (2019), também fazem uso de regressões quantílicas para analisar o retorno da educação no resultado econômico individual.

Nesse sentido, o presente artigo procura mostrar o viés presente na estimação de modelos log-lineares por OLS utilizando como pano de fundo a estimação do retorno da educação através da equação de salários minceriana. Para isso, estimamos o modelo log-linear de retorno educacional por regressão quantílica, conforme o proposto por Figueiredo, Lima e Schaur (2016), e comparamos este resultado com o obtido pela metodologia tradicional de estimação pela média. Os dados utilizados para tal são provenientes da Pesquisa Nacional por amostra de Domicílios (PNAD) para o ano de 2015.

Como forma de conferir maior robustez à análise, procedemos à realização de uma simulação de Monte Carlo, construindo intervalos de confiança para os parâmetros estimados. De posse dos verdadeiros parâmetros dos dados gerados, será possível identificar com maior precisão se o viés da log-linearização de fato superestima o retorno à educação, comparando-se com o valor das estimativas de ambos os métodos.

Os resultados encontrados pelas estimações mostram que a estimação por regressão quantílica diminui o coeficiente da variável anos de estudo na equação minceriana, para todas as especificações consideradas. A diferença entre o coeficiente estimado via OLS em um modelo empírico menos completo e a média dos coeficientes gerados pela regressão quantílica no modelo com as variáveis de controle foi da ordem de 62%. A diferença entre os dois métodos foi corroborada na Simulação de Monte Carlo. Além disso, os resultados da regressão quantílica permitem também mostrar que o retorno da educação é menor para indivíduos mais pobres, evidenciando a existência de desigualdade de rendimentos.

O artigo conta com mais quatro seções, além desta introdução. A próxima seção descreve com maiores detalhes o viés de log-linearização e como a regressão quantílica conseguiria lidar com esse viés. A terceira seção é dedicada aos dados utilizados. A quarta seção apresenta os resultados encontrados das estimações com dados da PNAD. A quinta seção mostra os resultados da simulação de Monte Carlo e a sexta, e última seção, contém as considerações finais.

2 Procedimentos Metodológicos

Mincer (1974) propôs uma equação que não levasse em conta apenas a influência da educação no salário dos indivíduos, mas também o impacto do aprendizado adquirido pela experiência no trabalho. O modelo empírico esti-

mado no presente artigo inclui, adicionalmente às variáveis explicativas escolaridade e experiência no mercado de trabalho, as variáveis de controle raça, sexo, idade, unidade da federação, área urbana e metropolitana, entre outras:

$$\ln W_i = \beta_0 + \beta_1 Escol_i + \beta_2 Exper_i + \beta_3 Exper_i^2 + X_i + \epsilon_i \tag{1}$$

onde $\ln W_i$ é o log do salário por hora individual, $Escol_i$ a escolaridade, $Exper_i$ a experiência, $Exper_i^2$ a experiência ao quadrado para capturar a não linearidade do retorno desta variável e X_i o vetor das variáveis de controle.

O modelo acima se origina da transformação logarítmica de um modelo multiplicativo exponencial[3], com a seguinte especificação:

$$W_i = \exp(x_i\beta)\tau_i \tag{2}$$

A transformação objetiva tornar o modelo linear nos parâmetros, permitindo sua estimação via mínimos quadrados. Esta estratégia de identificação, no entanto, vem sendo questionada por algumas pesquisas recentes, com base nas implicações da Desigualdade de Jensen, um resultado estatístico segundo o qual o valor esperado do logaritmo de uma variável aleatória difere do logaritmo do seu valor esperado.

Segundo Figueiredo, Lima e Schaur (2016), sendo $\tau_i = \exp[(x_i\varphi)\epsilon_i]$, onde ϵ_i possui uma distribuição normal (μ, θ^2) i.i.d, τ_i possui distribuição log-normal com a variância sendo uma função de x_i. Reescrevendo a equação (2), tem-se:

$$\ln W_i = x_i\beta + (x_i\varphi)\epsilon_i. \tag{3}$$

Assim, $E(\ln \tau_i|x_i) = \frac{-1}{2}\theta_i^2$, e quando assumimos que a equação é identificada, temos que $E(\ln \tau_i|x_i) = 1$. Na equação (2), temos que $E(\ln \tau_i|x_i) = x_i(\beta - \frac{\theta_{\tau_i}^2}{2}) \neq x_i\beta$ de forma que a heterocedasticidade leva a inconsistência da regressão log-linear da equação (3) se $E(\tau_i|x_i) = 1$ e identifica o modelo multiplicativo.

Se supusermos que $E(\epsilon_i|x) = E(\epsilon_i) = 0$, temos que $E(\ln W_i) = x_i\beta + x_i\varphi E(\epsilon_i) = x_i\beta$. Pelas propriedades da distribuição log-normal, se $E(\epsilon_i) = 0$ teremos $E(\tau_i|x) \neq 1$. Portanto, se o modelo log-linear for identificado, o modelo multiplicativo não é identificado, pois $E(f_i|x) = \exp(x_i\beta)$. Em suma, a identificação do modelo exponencial não leva à identificação do modelo log-linear, e vice-versa, tendo em vista que a Desigualdade de Jensen nos mostra que $E(\ln \tau_i|x_i) \neq \ln[E(\tau_i|x_i)]$.

Uma das propriedades dos quantis é a equivariância, importante no âmbito de estudos aplicados pois permite que a escala da variável original possa ser alterada, sem perda de coerência nas conclusões baseadas nos resultados estimados da regressão (Figueiredo et al. (2014)), então propõem a estimação do modelo log-linearizado pelo método de regressões quantílicas. Os autores mostram que com isso é possível identificar simultaneamente os dois modelos.

Para qualquer variável aleatória Y, $Q_\theta(h(Y)) = h(Q_\theta(Y))$. Aplicando essa propriedade à (2), temos que:

[3]Os detalhes mostrados nesta seção se baseiam na descrição apresentada por Figueiredo, Lima e Schaur (2016)

$$
\begin{aligned}
Q_\theta(W_i|x_i) &= \exp(x_i\beta).Q_\theta(\tau_i|x_i), && (4)\\
&= \exp(x_i\beta).\exp[(x_i\theta)Q_\theta(\epsilon_i)], && (5)\\
&= \exp[x_i(\beta+\theta Q_\theta(\epsilon_i))], && (6)\\
&= \exp(x_i\beta(\theta)). && (7)
\end{aligned}
$$

onde $\beta(\theta) = \beta + \theta Q_\theta(\epsilon_i)$

Log-linearizando o modelo, teremos $\ln\tau_i = (x_i\theta)\epsilon_i$. Inserindo adicionalmente a propriedade da equivariância temos o modelo:

$$
\begin{aligned}
\ln W_i &= x_i\beta + \ln\tau_i && (8)\\
\ln W_i &= x_i\beta + (x_i\theta)\epsilon_i && (9)
\end{aligned}
$$

Para cada $\theta \in (0,1)$, a propriedade da equivariância nos diz que

$$Q_\theta(\ln(W_i)|x_i) = \ln[Q_\theta(W_i)] = \ln[\exp(x_i\beta(\theta))],$$

onde $\beta(\theta) = \beta + \theta Q_\theta(\epsilon_i)$. Assim, o $\beta(\theta)$ obtido será o mesmo obtido pela equação exponencial. Assim, a abordagem quantílica identifica tanto o modelo multiplicativo quanto o log-linear, de forma que a presente estimação terá como base o modelo das equações (8) e (9).

Para comparar as estimativas do retorno da educação obtidas por OLS e por regressões quantílicas, calculamos o efeito médio de x_i a partir dos coeficientes dos quantis e comparamos com o coeficiente de mínimos quadrados. Koenker (2005, p. 32) mostra que integrando a função quantílica em todo o domínio $(0,1)$ se chega a uma aproximação da função média:

$$E(y_i|x_i = x) \approx \int_0^1 Q_\tau(y_i|x_i = x)d\tau \qquad (10)$$

Em outras palavras, a média dos coeficientes estimados para cada quantil consistiria em uma boa aproximação do efeito na média condicional.

Figueiredo et al. (2014) citam ainda outras vantagens da regressão quantílica, como a facilidade de interpretar os parâmetros obtidos, a capacidade de lidar com variáveis censuradas e as estimativas para diferentes quantis, aumentando o espectro da análise com o impacto das variáveis explanatórias em diferentes pontos da distribuição. Ao fornecer um retrato geral do retorno da educação em toda a distribuição de salários, pode-se inferir resultados sobre desigualdade educacional [4] e de rendimentos[5].

3 Dados

Os dados utilizados no estudo são fornecidos pela Pesquisa Nacional por Amostra de Domicílios (PNAD) realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) para o ano de 2015. A PNAD é uma pesquisa por

[4] Figueiredo & Dantas (2011) mostram que a desigualdade educacional associada a uma desigualdade de retorno à escolaridade, faz da educação um componente importante da desigualdade de renda.

[5] Outra vantagem da regressão quantílica é a robustez a *outliers*(ver Silva & Porto Júnior (2006)).

amostragem probabilística de domicílios, representativa em todo o território nacional. Para dar maior consistência às estimativas, foram realizados alguns filtros na amostra total, mantendo apenas chefes de domicílios com idade economicamente ativa entre 18 e 65 anos. A variável dependente é o logaritmo da renda do trabalho por hora e a escolaridade é medida a partir dos anos de estudo dos indivíduos.

A variável experiência no mercado de trabalho foi criada a partir da idade do indivíduo menos os anos de estudo menos 6, que é idade que o indivíduo entra na escola (Experiência = Idade – Anos de estudo – 6). Já a variável sexo consiste em uma dummy que assume valor 1 se o indivíduo for do sexo masculino e 0 se o indivíduo é do sexo feminino, assim como a dummy de raça, que assume valor 1 se o indivíduo for branco e 0 se o indivíduo for não branco. Além dessas variáveis, foram incluídas uma dummy que indica se o indivíduo vive na área urbana, uma dummy que indica se o trabalhador é do setor formal e uma dummy para cada região da federação.

Variáveis relacionadas ao *background* familiar também são importantes na estimação da equação de salários, visto a alta persistência entre as gerações [Ver Ferreira & Veloso (2006)]. No entanto, a não inclusão de variáveis relacionadas a educação dos pais não inviabiliza o objetivo desse trabalho, cujo foco é no viés de log-linearização. Por outro lado, variáveis relacionadas às características do cônjuge ajudam a mitigar o viés de variáveis omitidas, onde indivíduos mais escolarizados tendem a ter pares também com maior grau de escolaridade ("casamentos seletivos"). Lam & Schoeni (1993), por exemplo, mostram que a correlação entre a escolaridade dos indivíduos e seus cônjuges é de 0,77 para dados da PNAD de 1982.

A Tabela 1 apresenta as estatísticas descritivas das variáveis relacionadas às características individuais, num total de 19.076 observações sobre indivíduos da PNAD de 2015. É observado uma média de anos de estudo de 10 anos com uma média de idade de pouco mais de 41 anos e 25,3 anos de experiência. Dentre os atributos pessoais, 72% da amostra é composta por homens chefes de família, aproximadamente 44% são de cor branca, 65% estão inseridos no mercado de trabalho formal e 53% moram num município diferente do que nasceram (migrante). Por fim, a quase totalidade reside na área urbana (96%), 59% em regiões metropolitanas e a maior parte nas regiões Sudeste (33%) e Nordeste (21%).

A Tabela 2 mostra as características do cônjuge. De fato, é possível observar uma média de anos de estudo praticamente idêntica a do chefe da família, dando indícios de que a amostra apresenta um *matching* parecido em termos de escolaridade. O percentual de cônjuges brancos na amostra é um pouco superior ao do chefe ao passo que possuem uma média de idade menor e menos experiência no mercado de trabalho. Por outro lado, um percentual maior de cônjuges está inserido no mercado de trabalho formal. Por fim, 27% dos cônjuges são do sexo masculino, complementando a informação da Tabela 1 com a maioria dos chefes de família homens (72%).

4 Resultados

Esta seção apresenta os resultados do retorno da educação sobre os salários, estimados por MQO, por regressão quantílica e para a média dos quantis (segundo Koenker (2005, p. 32)). Em todas as abordagens, as variáveis referentes

Tabela 1: Estatísticas Descritivas – PNAD 2015

Variáveis	Média	Desvio-Padrão	Mínimo	Máximo
Log da Renda/hora (Dependente)	7,753	2,476	3,218	27,631
Características Pessoais				
Anos de Estudo (Variável de interesse)	10,079	3,908	0	15
Sexo	0,720	0,448	0	1
Raça	0,442	0,496	0	1
Idade	41,453	10,044	18	65
Experiência	25,387	11,246	1	59
Formal	0,656	0,474	0	1
Migrante	0,533	0,498	0	1
Localização				
Urbana	0,964	0,184	0	1
Área Metropolitana	0,591	0,491	0	1
Norte	0,132	0,339	0	1
Nordeste	0,215	0,411	0	1
Sul	0,188	0,390	0	1
Sudeste	0,338	0,473	0	1
Centro-Oeste	0,125	0,331	0	1

Fonte: Elaboração dos autores com base nos dados da PNAD (2015).

Tabela 2: Características Pessoais do Cônjuge – PNAD 2015

Variáveis	Média	Desvio-Padrão	Mínimo	Máximo
Anos de Estudo	10,318	3,829	0	15
Sexo	0,278	0,448	0	1
Raça	0,475	0,499	0	1
Idade	39,684	9,972	18	65
Experiência	23,382	11,165	1	59
Formal	0,688	0,463	0	1

Fonte: Elaboração dos autores com base nos dados da PNAD (2015).

às características individuais e de localização são inseridas como regressores nas equações, com o objetivo de observar as mudanças no coeficiente referentes ao retorno da educação. Posteriormente são inseridas as características do cônjuge como regressores adicionais, buscando captar características não observáveis dos indivíduos (escolhas, ambiente familiar compartilhado) que podem afetar sua produtividade e consequentemente os salários, gerando assim estimativas mais consistentes.

Contudo, apesar do tratamento da omissão de variáveis, não é garantido que os modelos estimados na forma log-linear, tanto por OLS quanto por RQ, estejam corretamente especificados. É possível que exista viés decorrente da má-especificação do modelo. Para garantir que as estimações não tenham resultados viesados devido à sua forma funcional, foi realizada uma simulação de Monte Carlo. Esse processo simula um modelo no qual os verdadeiros parâmetros são conhecidos e permite calcular o mesmo modelo via OLS e RQ, podendo indicar de forma robusta o método de estimação que melhor se aproxima dos verdadeiros parâmetros.

A Tabela 3 apresenta as estimações por MQO, para o quantil 0,50 (mediana) e para a média dos quantis, respectivamente. Para uma comparação mais

direta com os resultados do MQO a estimação feita com a média dos quantis é a indicada, onde o retorno da educação é obtido pela estimação para diferentes quantis e posteriormente é calculada a média desses coeficientes[6]. O primeiro modelo estimado (coluna 1) é controlado apenas por características pessoais do próprio chefe da família (idade, idade ao quadrado, sexo, raça, experiência, experiência ao quadrado, trabalho formal, migração), que será mantida em todas as regressões, tal como descrita na seção dos dados. O modelo da Coluna 2 é estimado adicionando as variáveis de localização (região metropolitana, área urbana, dummies para região), mantendo-se as demais variáveis de controle. Por fim, na Coluna 3 são adicionadas nas estimações as características do cônjuge, buscando reduzir o viés de variável omitida.

Tabela 3: Retorno da Educação no Brasil

	Variável Dependente: Log do Salário por Hora		
	(1)	(2)	(3)
	OLS		
Escolaridade	0,161***	0,155***	0,099***
	(0,011)	(0,0012)	(0,014)
Observações	19.025	19.025	19.025
	Regressão Quantílica (Mediana)		
Escolaridade	0,125***	0,120***	0,087***
	(0,004)	(0,004)	(0,005)
Observações	19.025	19.025	19.025
	Média dos Quantis		
Escolaridade	0,114***	0,093***	0,061***
	(0,004)	(0,004)	(0,005)
Observações	19.025	19.025	19.025

Fonte: Elaboração própria a partir dos dados das PNAD de 2015.
Notas: (1) Controles: características Pessoais. (2) Controles: características Pessoais + Localização. (3) Controles: características Pessoais + Localização + Características do Cônjuge.
Notas: ***p-valor < 0,01. ** p-valor < 0,05. * p-valor < 0,10

Comparando as colunas para um mesmo método de estimação, é observada uma redução no coeficiente da escolaridade à medida que as demais variáveis são adicionadas, passando de 0,16 quando apenas as características pessoais são utilizadas (coluna 1) para 0,09, quando todas as variáveis são utilizadas (coluna 3) no caso da média estimada por OLS. Na estimação quantílica para a mediana a redução no coeficiente da escolaridade foi de 0,12 (coluna 1) para 0,08 (coluna 3) e para a média dos quantis de 0,11 (coluna 1) para 0,06 (coluna 3). Tais resultados evidenciam a possível superestimação dos coeficientes com a omissão de variáveis relevantes, independente do método de estimação.

Um ponto importante a destacar é que as estimativas por regressão quantílica apresentam coeficientes menores para cada uma das especificações utilizadas, tanto para o retorno da educação quanto para outras variáveis de

[6]Foram estimados os quantis de 0,01 ao 0,99 e retirado à média aritmética dos seus coeficientes.

controle.[7]. Isso sugere uma possível superestimação quando se usa a equação minceriana tradicional estimada pela média, devido a fatores como a presença de *outliers* ou erros de especificação. Quando se observa o resultado pela média dos quantis, são verificados coeficientes ligeiramente menores que os estimados para a mediana (quantil 0,50), o que sugere uma maior robustez dos resultados quantílicos. A diferença entre as abordagens fica ainda mais evidente quando comparamos o retorno da educação via OLS na especificação menos completa (1) e o retorno obtido com a média dos quantis no modelo com as características de localização e do cônjuge como controles (3). Há uma redução de 0,1, o que representa uma diferença da ordem de 62%.

Para melhor visualização dos resultados quantílicos, a Figura 1 mostra os coeficientes do retorno da educação para toda a distribuição de salários. Cada gráfico da Figura é construído a partir de uma das especificações da Tabela anterior, tal que o gráfico Modelo 1 corresponde a especificação da coluna 1 da Tabela, com a inclusão apenas das características pessoais, o gráfico Modelo 2 adiciona as variáveis de localização e o gráfico Modelo 3 adiciona as características do cônjuge, mantendo as demais. Em todas as especificações, vê-se que existe uma desigualdade de retorno educacional entre os quantis. Para os indivíduos com os menores salários um ano a mais de qualificação gera um acréscimo de renda menor do que para aqueles indivíduos localizados nos quantis mais altos de renda. À medida que são adicionadas as variáveis de controle, a curva dos coeficientes quantílicos fica menos inclinada e a diferença entre os retornos se torna sensivelmente menor.

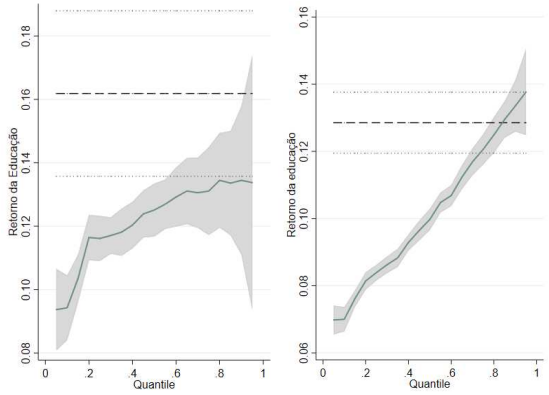
Este resultado corrobora a hipótese de que a educação é um fator preponderante para a desigualdade de rendimento entre os indivíduos. Os indivíduos mais pobres, em geral, possuem os menores níveis de escolaridade, além de ter acesso a um ensino de pior qualidade; a partir do momento que existe uma diferença de educação e de retorno educacional, há um reforço da desigualdade de rendimento já existente. Com a inclusão das demais variáveis, consegue-se controlar por mais fatores relacionados aos ganhos de renda dos indivíduos, de modo que tanto a magnitude quanto as variações entre os parâmetros estimados têm uma leve redução.

Em suma, a estimação por OLS da equação minceriana na sua forma log-linear tende a superestimar o efeito da educação sobre os salários. Embora a literatura tenha concentrado todos os esforços em favor da correção da endogeneidade (o que no nosso caso não pode ser completamente controlada por falta de bons instrumentos))[8], ao não levar em consideração que a função média não é invariante a transformação logarítmica negligencia-se um viés importante e que gera uma superestimação significativa do parâmetro do retorno educacional. No presente trabalho, o coeficiente do retorno da educação calculado pela média condicional possui um viés (viés da log-linearização) de 39,3% em relação à média dos coeficientes quantílicos. A possível superestimação das estimativas OLS será confirmada com maior precisão na simulação de Monte Carlo, cujos resultados são apresentados na seção seguinte.

[7] o objetivo do artigo não é discutir coeficientes e efeitos de variáveis relacionadas às demais cacacterísticas. Por isso, nos limitamos a apresentar o retorno da educação. No entanto, os resultados para as demais variáveis podem ser solicitados juntos aos autores

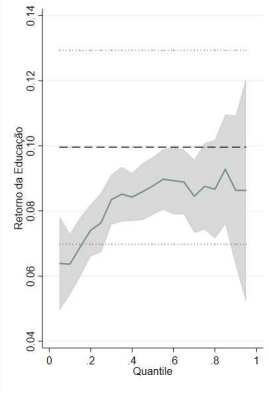
[8] A educação dos pais, por exemplo, consiste em um dos instrumentos tradicionalmente utilizados pela literatura. Entretanto, este procedimento recebe fortes críticas, dado que ela é significativa quando considerada como um regressor diretamente inserido na equação de salários, o que inviabiliza sua utilização como instrumento

Figura 1: Retorno Educacional para os Quantis de Salário



(a) Modelo 1

(b) Modelo 2



(c) Modelo 3

Fonte: Elaboração própria com base nas estimações.

5 Simulação de Monte Carlo

Angrist et al. (2006) mostra que o processo de regressão quantílica é limitado, uma vez que esta não é robusta a má especificação do modelo. Nesse caso, a função covariância não é proporcional à função covariância da função corretamente especificada. Caso exista problema de má especificação do modelo nos quantis condicionais não é possível recuperar a melhor aproximação linear deste quantil, diferente do que ocorre com a média condicional.

Não obstante a robustez do OLS a erros de especificação, o presente artigo procura mostrar que para a estimação de modelos log-lineares originados de modelos multiplicativos com heterocedasticidade, a regressão quantílica gera estimativas menos viesadas. Assim, foram realizadas simulações de Monte Carlo para comparar as estimativas sob o método de OLS, a abordagem de regressões quantílicas para a mediana e a média dos coeficientes quantílicos, este último tal como sugerido por Koenker (2005) para uma melhor comparação com a média condicional. Conhecendo os verdadeiros valores dos parâmetros β_1 e β_2, a Tabela 1 apresenta os resultados para o viés de especificação do modelo log-linear (%Viés) e o erro quadrático médio (RMSE - Root Mean Square Error) das simulações.

O Processo Gerador de Dados[9] é definido como $y = \exp(2+X_1+2X_2)\cdot\exp(1+(\gamma\cdot(X_1+X_2++X_2^2+X_2^2))\cdot u)$, com $(\beta_1,\beta_2) = (1,2)$, $u \sim iidN(0,1)$ e $X_1,i = 1,2$ obtidos de forma independente e normalmente distribuidos. O parâmetro γ controla a heterocedasticidade do modelo; atribuiu-se três diferentes valores nas estimações $\gamma = (0,1;0,5;1)$, tendo uma log normal heterocedastica. Caso considerássemos $\gamma = 0$ não teríamos heterocedasticidade. Observa-se ainda que o erro foi incluído na exponencial, não admitindo, portanto, erros negativos, dado que o modelo é estimado na forma log-linear. Por fim, foram considerados os tamanhos de amostra N = 500, N=1.000 e N=2.000, a partir do OLS, da regressão quantílica para a mediana e da média dos quantis usando 1.000 replicações.

As simulações buscaram observar o viés causado por transformações logarítmicas de modelos exponenciais com heterocedasticidade, como no caso da retorno da educação apresentado. Os resultados sugerem que o estimador de OLS apresenta um maior viés para a maior parte das simulações, com exceção de alguns coeficientes para amostras menores N=500 e um menor valor para gamma. Por outro lado, o desempenho da regressão quantílica para a mediana ($\tau = 0,50$) é superior em termos de redução do viés e possui menor RMSE que os demais, e essa diferença em relação ao OLS se eleva à medida que se aumenta o tamanho da amostra e a presença da heterocedasticidade. A estimativa usando a média dos quantis também apresenta um viés menor que as estimativas obtidas por OLS para amostras maiores com heterocedasticidade, porém o RMSE é maior que a mediana devido aos quantis extremos (caudas inferior e superior).

Em suma, as simulações mostram que a regressão quantílica consegue estimativas com menor viés e RMSE na presença de heterocedasticidade.A vantagem da estimação via regressão quantílica frente aos demais métodos aumenta quanto maior a amostra e o grau de heterocedasticidade. Nas mesmas condições que a mediana, a média dos quantis também consegue um desempenho melhor que os mínimos quadrados, porém apresenta um RMSE maior que a mediana, naturalmente devido aos quantis das caudas. Sem a heterocedasticidade, sem erros de esquecificação e em amostras pequenas, o OLS consegue uma estimação com menor viés.

6 Considerações Finais

O objetivo do presente trabalho consistiu em mostrar o viés produzido pela estimação de modelos log-lineares pelo método de mínimos quadrados. Para tanto, utilizamos como pano de fundo a estimação do retorno da educação com a equação minceriana de salários. Nesse sentido, procurou-se produzir estimativas mais robustas mostrando como solução do problema de identificação a aplicação de regressões quantílicas ao modelo log-linear, dada a propriedade de equivariância dos quantis. Para sustentar nossa hipótese, foi feita uma simulação de Monte Carlo, na qual de posse dos verdadeiros parâmetros do modelo simulado, conseguimos ver qual método produz as estimativas mais consistentes.

Os resultados mostraram que a estimação de modelos log-lineares pela média condicional tende a gerar coeficientes de retorno da educação tendenciosos. Quando comparamos as estimativas geradas por MQO com as gera-

[9] Em inglês *Data Generator Process (DGP)*

Tabela 4: Simulação de Monte Carlo

	$\gamma = 0,1$			$\gamma = 0,5$			$\gamma = 1$		
	OLS	**QR**	**MQR**	**OLS**	**QR**	**MQR**	**OLS**	**QR**	**MQR**
N = 500									
% Viés de β_1	0,001	0,007	0,006	0,006	0,037	0,029	0,012	0,075	0,06
% Viés de β_2	0,009	0,005	0,002	0,049	0,026	0,012	0,099	0,053	0,025
RMSE	1,749	0,459	1,65	8,746	2,295	4,255	17,492	4,59	8,511
N = 1.000									
% Viés de β_1	0,025	0,004	0,008	0,127	0,021	0,041	0,255	0,042	0,083
% Viés de β_2	0,003	0,003	0,003	0,019	0,017	0,019	0,038	0,034	0,038
RMSE	1,06	0,29	1,613	5,3008	1,45	4,069	10,601	2,9	7,139
N = 2.000									
% Viés de β_1	0,005	0,001	0,001	0,025	0,009	0,005	0,051	0,018	0,01
% Viés de β_2	0,025	0,005	0,006	0,126	0,028	0,034	0,253	0,057	0,068
RMSE	0,997	0,253	1,587	4,988	1,267	3,941	9,976	2,534	6,883

Fonte: Elaboração própria a partir das simulações.
Legenda: QR = Regressão quantílica para a mediana; MQR = Média dos Quantis.

das pelas abordagens de quantis (regressão quantílica tradicional e média dos quantis), verificou-se uma redução nos coeficientes em todos os modelos testados (com e sem controles).

Tais resultados são confirmados pela simulação de Monte Carlo, a qual mostra que a regressão quantílica produz as estimativas de retorno da educação mais próximas das verdadeiras, em comparação com as estimativas OLS. A diferença entre as estimativas dos dois métodos cresce à medida que se aumenta o tamanho da amostra, ratificando a vantagem da regressão quantílica na estimação de modelos log-lineares na presença de heterocedasticidade.

Referências Bibliográficas

Angrist, J., Chernozhukov, V. & Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the US wage structure. *Econometrica*, New Haven, v. 74, p. 539-563.

Arshad, I. A., Younas, U., Shaikh, A. W. & Chandio, M. S. (2016). Quantile regression analysis of monthly earnings in Pakistan. *Sindh University Research Journal*, Jamshoro, v. 48, p. 919–924.

Becker, G. S. (1962). Investment in human capital: a theoretical analysis. *Journal of Political Economy*, Chicago, v. 70, p. 9–49.

Castro, C. M. (1970). *Investment in education in Brazil: a study of two indus-trial communities*. 1970. Thesis (Ph.D. in Economics) - Vanderbilt University, Nashville.

Cheruiyot, K. (2019). Heterogeneous relationships between income levels and associated correlates in Gauteng province, South Africa: quantile regression approach. *Development Southern Africa*, Abingdon, v. 37, p. 871–887.

Coelho, D., Veszteg, R. & Soares, F. V. (2010). *Regressão quantílica com correção para a seletividade amostral: estimativa dos retornos educacionais e diferenciais raciais na distribuição de salários das mulheres no Brasil*. Brasília: Instituto de Pesquisa Econômica Aplicada. (Texto de Discussão do IPEA n. 1483).

Dalcin, Aline.; Annegues, A. C. A. R. (2015). Desigualdade de renda e retornos educacionais: uma abordagem quantílica. *In: VI Encontro de Economia do Espírito Santo*. Vitória: UFES.

Esteves, E. A., Guio, S. P., Carlos, A., Cantor, E., Habeych, M. E. & Malagón, A. L. (2020). Reference values of upper extremity nerve conduction studies in a Colombian population. *Clinical Neurophysiology Practice*, Amsterdam, v. 5, p. 73–78.

Ferreira, S. G. & Veloso, F. A. (2006). Intergenerational mobility of wages in Brazil. *Brazilian Review of Econometrics*, Rio de Janeiro, v. 26, p. 181–211.

Figueiredo, E. & Dantas, A. R. M. (2011). Retorno da educação nos estados nordestinos: Piauí, Rio Grande do Norte e Bahia. *Economia e Desenvolvimento*, Recife, v. 9, p. 79–99.

Figueiredo, E., Lima, L. R. & Schaur, G. (2014). Robust Estimation of gravity equations and the WTO impact on trade inequality. *In: CESifo Conference on Estimation of Gravity Model of Bilateral Trade*. Munich: CESifo.

Hao, L., Naiman, D. Q. & Naiman, D. Q. (2007). *Quantile Regression*. Califórnia: Sage.

Heckman, J. J., Lochner, L. J. & Todd, P. E. (2006). Earnings functions, rates of return and treatment effects: The Mincer equation and beyond. *Handbook of the Economics of Education*, Amsterdam, v. 1, p. 307–458.

Heckman, J., Layne-Farrar, A. & Todd, P. (1996). Human capital pricing equations with an application to estimating the effect of schooling quality on earnings. *Review of Economics and Statistics*, Cambridge, v. 78, p. 562–610.

Hossain, M. M. & Majumder, A. K. (2019). Determinants of the age of mother at first birth in Bangladesh: quantile regression approach. *Journal of Public Health*, Berlin, v. 27, p. 419–424.

Hungerford, T. & Solon, G. (1987). Sheepskin effects in the returns to education. *Review of Economics and Statistics*, Cambridge, v. 69, p. 175-177.

Jaeger, D. A. & Page, M. E. (1996). Degrees matter: New evidence on sheepskin effects in the returns to education. *Review of Economics and Statistics*, Cambridge, v. 78, p. 733-740.

Koenker, R. (2005). *Quantile Regression*. Cambridge: Cambridge University Press.

Lam, D. & Levison, D. (1992). Age, experience, and schooling: Decomposing earnings inequality in the United States and Brazil. *Sociological Inquiry*, Hoboken, v. 62, p. 220–245.

Lam, D. & Schoeni, R. F. (1993). Effects of family background on earnings and returns to schooling: evidence from Brazil. *Journal of Political Economy*, Chicago, v. 101, p. 710–740.

Langoni, C. G. (1974). *As Causas do Crescimento Econômico do Brasil*. Rio de Janeiro: APEC.

Maciel, M. C., Campêlo, A. C. & Raposo, M. C. F. (2001). A dinâmica das mudanças na distribuição salarial e no retorno em educação para mulheres: uma aplicação de regressão quantílica. *In: XXIX Encontro Nacional de Economia*. Salvador: ANPEC.

Maiti, M. (2019). OLS versus quantile regression in extreme distributions. *Contaduría y Administración*, Cidade do México, v. 64, p. 1–11.

Mankiw, N. G., Romer, D. & Weil, D. N. (1992). A contribution to the empirics of economic growth. *Quarterly Journal of Economics*, Cambridge, v. 107, p. 407–437.

Mincer, J. (1974). *Schooling*, *Experience*, *and Earnings*. New York: National Bureau of Economic Research.

Moura, R. L. (2008). Testando as Hipóteses do Modelo de Mincer para o Brasil. *Revista Brasileira de Economia*, Rio de Janeiro, v. 62, p. 407-449.

Neves, M. F. & Lima, A. C. C. (2019). Investimento em capital humano e retornos da educação nos mercados de trabalho brasileiros, 1991/2010. *Revista de Desenvolvimento Econômico*, Salvador, v. 1, p. 76–107.

Okamoto, M. (2016). *Mincer earnings regression in the form of the double Pareto-lognormal model*. Roma: Society for the Study of Economic Inequality. (Working Paper n. 407).

Prieto-Rodriguez, J., Barros, C. P. & Vieira, J. A. C. (2008). What a quantile approach can tell us about returns to education in Europe. *Education Economics*, Abingdon, v. 16, p. 391–410.

Ramos, L. & Reis, M. (2009). *A escolaridade dos pais*, *os retornos à educação no mercado de trabalho ea desigualdade de rendimentos*. Rio de Janeiro: Instituto de Pesquisa Econômica Aplicada. (Texto de Discussão do IPEA n. 1442).

Rodrigues, C. F. S., Oliveira, C. M. S. & Alves, J. S. (2014). Desigualdade de renda e retornos educacionais: uma abordagem quantílica. *In: Encontro Pernambucano de Economia*. Recife: ENPECON.

Sachsida, A., Loureiro, P. R. A. & Mendonça, M. J. C. (2004). Um estudo sobre retorno em escolaridade no Brasil. *Revista Brasileira de Economia*, Rio de Janeiro, v. 58, p. 249–265.

Sampaio, A. V. (2009). Estimação da equação de salário para o Brasil, o Paraná e o Rio Grande do Sul em 2007—uma abordagem quantílica. *Indicadores Econômicos FEE*, Porto Alegre, v. 37, p. 1–20.

Schultz, T. W. (1961). Investment in human capital. *American Economic Review*, Nashville, v. 51, p. 1–17.

Silva, E. N. & Porto Júnior, S. S. (2006). Sistema financeiro e crescimento econômico: uma aplicação de regressão quantílica. *Economia Aplicada*, Ribeirão Preto, v. 10, p. 425–442.

Silva, J. M. C. S. & Tenreyro, S. (2006). The log of gravity. *Review of Economics and Statistics*, Cambridge, v. 88, p. 641–658.

Staffa, S. J., Kohane, D. S. & Zurakowski, D. (2019). Quantile regression and its applications: a primer for anesthesiologists. *Anesthesia & Analgesia*, Philadelphia, v. 128, p. 820–830.