

# REGRESSÃO QUANTÍLICA COM CORREÇÃO PARA A SELETIVIDADE AMOSTRAL: ESTIMATIVA DOS RETORNOS EDUCACIONAIS E DIFERENCIAIS RACIAIS NA DISTRIBUIÇÃO DE SALÁRIOS DAS MULHERES NO BRASIL\*

Danilo Coelho\*\*

Róbert Veszteg\*\*\*

Fabio Veras Soares\*\*\*\*

Este texto estima os retornos educacionais e os diferenciais raciais na distribuição de salários das mulheres no Brasil, usando regressão quantílica com correção semiparamétrica para viés de seleção amostral. As estimativas mostram que os retornos educacionais são elevados e que não são constantes ao longo da distribuição salarial. Tanto os retornos educacionais quanto os diferenciais raciais são mais elevados nos pontos mais altos da distribuição de salário condicional, o que indica, no caso dos diferenciais raciais, que as mulheres negras enfrentam um teto de vidro (*glass ceiling*) nos níveis salariais mais altos. Para os diferenciais por anos de estudo, questões como a qualidade da educação podem ser um fator importante na explicação da desigualdade salarial entre as mulheres. O texto revela que o uso de uma especificação *probit* para a equação de participação, a fim de corrigir problemas de seleção, produz resultados muito semelhantes à correção semiparamétrica tanto para os retornos educacionais quanto para a discriminação racial.

## 1 INTRODUÇÃO

Embora tenha tido uma trajetória crescente,<sup>1</sup> a oferta de trabalho das mulheres ainda é bem inferior à dos homens nos países em desenvolvimento. Um aumento na participação feminina pode vir a ter um papel fundamental na redução da desigualdade de salários e de renda familiar *per capita*, assim como na diminuição dos diferenciais salariais entre homens e mulheres. Entretanto, há evidências de que o recente aumento da participação feminina no mercado de trabalho no Brasil teve um efeito pró-desigualdade de renda total (SCORZAFAVE, 2004), apesar de ter tido um efeito redutor da desigualdade salarial (SCORZAFAVE;

---

\* Os autores agradecem a Eduardo Pontual, Rafael Osório e Mauricio Reis pelos valiosos comentários. Veszteg agradece o apoio financeiro do Ministério da Educação e Ciência do governo espanhol. Este artigo foi finalizado durante a visita de Veszteg ao Institute for Social and Economic Research, Osaka University, Japão.

\*\* Técnico de Planejamento e Pesquisa da Coordenação de Estudos de Mercado e Regulação da Diretoria de Estudos e Políticas Setoriais, de Inovação, Regulação e Infraestrutura - DISET/Ipea.

\*\*\* Professor da Universidad Carlos III de Madrid.

\*\*\*\* Técnico de Planejamento e Pesquisa da Diretoria de Estudos e Relações Econômicas e Políticas Internacionais - DEINT/Ipea e pesquisador do Centro Internacional de Políticas para o Crescimento Inclusivo (CPI-CI).

1. Costa, Silva e Vaz (2009) mostram que no Brasil a taxa de participação das mulheres é de 54%. Para os homens era de 85% em 2006. O Chile apresenta a taxa de participação mais baixa para as mulheres na América Latina, cerca de 47%, segundo os dados da Encuesta de Caracterización Socioeconómica (Casen) de 2003.

MENEZES FILHO, 2005).<sup>2</sup> Desse modo, mesmo que o diferencial de salários entre homens e mulheres seja reduzido, a desigualdade de renda pode crescer, devido ao aumento da desigualdade entre as mulheres.

Neste caso, o retorno proveniente dos investimentos em educação – dado o maior estoque educacional das mulheres em relação aos homens – e o diferencial racial podem ser determinantes importantes da desigualdade salarial e, por meio desta, da desigualdade total dos rendimentos. Nesse sentido, é importante olhar para o efeito global dos determinantes dos salários das mulheres – tanto aqueles relacionados à remuneração de atributos produtivos, como a educação, quanto possíveis fontes de discriminação, como a cor – na distribuição salarial, em vez de se concentrar apenas no efeito condicional na média.

A regressão quantílica é uma poderosa ferramenta econométrica para examinar a distribuição dos salários das mulheres de uma forma útil e esclarecedora. Por exemplo, ela pode revelar se os retornos educacionais tendem a contribuir para uma desigualdade maior – quando os retornos são mais elevados no topo da distribuição de salários – e se os diferenciais raciais de salário tendem a ser mais prevalentes em quantis salariais mais elevados. Conforme apontam Martins e Pereira (2004), maiores retornos da educação para os quantis mais elevados podem indicar que a qualidade da educação ou externalidades e complementaridades relacionadas aos maiores níveis educacionais, entre outros fatores, podem ser importantes para entender os mecanismos que atuam no sentido de aumentar a desigualdade salarial. Neste caso, um acréscimo da participação feminina com um viés pró-mulheres escolarizadas pode ter o efeito de aumentar a desigualdade total, mesmo em um contexto de redução da desigualdade salarial, como discutido em Costa, Silva e Vaz (2009). Neste sentido, o aumento da participação de mulheres pobres e seu acesso a maiores salários são fatores extremamente importantes a serem levados em conta pelas políticas públicas, para diminuir tanto a desigualdade salarial quanto a desigualdade total.

Em termos metodológicos, o principal desafio na estimação dos determinantes da oferta de trabalho é que o salário de mercado não é observável para quem não está participando do mercado de trabalho. Este problema é particularmente importante no caso das mulheres, que enfrentam mais restrições para participar do mercado de trabalho. Ignorar o fato de que a distribuição salarial observada no mercado para as mulheres é truncada pode gerar um viés de seleção na estimativa dos parâmetros da equação de salários.

A econometria possui vários métodos de estimação que contornam o viés de seleção mencionado. O método convencional para estimar a oferta de trabalho sem

---

2. Hoffman e Leone (2004) também apontam que a renda do trabalho das mulheres contribuiu para o aumento da desigualdade na renda familiar *per capita* entre 1981 e 2002.

viés de seleção baseia-se na correção de Heckman, que consiste em duas equações: as equações de participação e de salário. A equação de participação explica a probabilidade de participar do mercado de trabalho, enquanto a equação de salário explica os salários de acordo com a abordagem minceriana de capital humano, incluindo o termo de correção derivada da distribuição do termo de erro da equação de participação. Em geral, este tipo de modelo pressupõe uma distribuição normal padrão para o termo de erro da equação de participação (modelo *probit*). No entanto, o uso de métodos semiparamétricos que não dependem de qualquer hipótese sobre a distribuição dos erros do modelo tornou-se comum na literatura.<sup>3</sup>

Buchinsky (2001) aplica o modelo de seleção amostral num arcabouço de regressão quantílica para estimar os retornos educacionais sobre os salários das mulheres nos Estados Unidos no período compreendido entre 1968 e 1990. Ele mostra que as estimativas semiparamétricas da equação de seleção (participação no mercado de trabalho) são consideravelmente diferentes daquelas obtidas a partir de um modelo paramétrico *probit*, e que existe um viés de seleção amostral importante para todas as faixas etárias e quantis.

A abordagem de regressão quantílica permite-lhe identificar que os retornos educacionais aumentaram muito para as coortes mais jovens, mas variaram pouco para as gerações mais velhas, e que, em geral, os retornos são mais elevados nos quantis mais baixos no início do período de observação e maiores nos quantis mais elevados no final do período de observação.

No que se refere à literatura brasileira, existem poucas aplicações de regressões quantílicas para investigar os determinantes dos salários das mulheres. Maciel, Campêlo e Raposo (2001), por exemplo, mostram um aumento nos retornos educacionais nos quantis superiores da distribuição salarial das mulheres entre 1992 e 1999. Entretanto, como a maioria das aplicações, os autores não controlam para um possível viés de seleção amostral.

Bartalotti e Leme (2007) investigam a discriminação salarial com base em cor e sexo, utilizando dados referentes ao Brasil fornecidos pela Pesquisa Nacional por Amostra de Domicílios (PNAD), do Instituto Brasileiro de Geografia e Estatística (IBGE) de 2004. Eles mostram que a discriminação baseada em cor e sexo é positivamente relacionada com o quantil da distribuição salarial. Também indicam que os retornos educacionais são maiores para os quantis mais elevados em todos os grupos analisados, e que isso aumenta as desigualdades entre as mulheres. Arcand e D'Hombres (2004) adotam o método de regressão quantílica, sem correção para seleção amostral, e constatam que a discriminação racial atinge mais os negros e pardos e aumenta à medida que se sobe ao longo da distribuição

---

3. Ver Martins (2001) para uma aplicação desta abordagem na oferta de trabalho das mulheres em Portugal.

condicional de salários. Santos e Ribeiro (2006) investigam diferencial salarial entre homens e mulheres aplicando regressão quantílica (também sem correção para seleção amostral), a fim de investigar a hipótese de teto de vidro (*glass ceiling*), segundo a qual existe uma barreira invisível para o crescimento do rendimento dos salários das mulheres nos níveis mais altos de rendimento, o que implicaria que o diferencial salarial entre homens e mulheres tenderia a aumentar ao longo da distribuição de salários. Suas estimativas mostram que o diferencial salarial de gênero tem a forma de  $U$  invertido, exceto para quantis mais elevados, nos quais há um enorme aumento do diferencial de salários. Este aumento para os decis superiores indica a existência de um teto de vidro.

O objetivo deste estudo é apresentar novas estimativas relativas ao salário das mulheres no Brasil, enfocando os retornos educacionais e diferenciais raciais de salário dentro deste grupo. Usa-se regressão quantílica com correção semiparamétrica para seleção amostral similar à de Buchinsky (2001), com o intuito de se estimar como estes determinantes variam ao longo da distribuição de salários, contribuindo, ou não, para uma desigualdade salarial maior. Será avaliado se a hipótese de normalidade da distribuição dos erros no modelo Heckit padrão não leva a estimativas ainda enviesadas dos parâmetros de interesse. Existe apenas uma aplicação de técnicas para a correção semiparamétrica de seleção amostral usando dados brasileiros. Tanuri-Pianto e Pianto (2002) aplicam esta metodologia para decompor o diferencial de salários entre trabalhadores formais e informais.

Os resultados confirmam que os retornos educacionais são, de fato, elevados na amostra de mulheres ocupadas na faixa etária entre 20 e 60 anos no Brasil. Além disso, os resultados da regressão quantílica mostram que os retornos educacionais e a discriminação racial tendem a ser maiores para as mulheres no topo da distribuição salarial.

O trabalho está organizado da seguinte forma. A seção 2 descreve brevemente o modelo de Heckman de seleção amostral e sua versão quantílica. A seção 3 expõe a abordagem semiparamétrica para corrigir vies de seleção. Os dados utilizados na estimativa são apresentados na seção 4. A seção 5 traz os principais resultados e a seção 6, as conclusões.

## 2 O MODELO DE SELEÇÃO AMOSTRAL DE HECKMAN E SUA VERSÃO QUANTÍLICA

### 2.1 Heckman: modelo de seleção amostral

No modelo de seleção amostral proposto por Heckman (1979), a variável dependente de interesse é o logaritmo do salário-hora observado apenas para os indivíduos que participam do mercado de trabalho. Uma das principais premissas do modelo estabelece que um indivíduo participa do mercado de trabalho sempre que seu

salário de mercado for maior do que seu salário de reserva. Este modelo pode ser descrito na seguinte forma:

$$Z_i^* = \underbrace{\left( 1, CR018, CR03, CR06, EDU, ID, ID2, BRANCA, CHEFE, \right.}_{X_i} \left. \begin{matrix} \\ TAM, NOSAL, OUTRAS, URBANA, R \end{matrix} \right)_i \cdot \beta + \varepsilon_i \quad (1)$$

A equação (1) é chamada de equação de participação e  $Z^*$  é a diferença entre o salário de mercado e o salário de reserva, que é linearmente dependente com relação a  $X$ . Por sua vez,  $X$  é um vetor de características individuais e do domicílio cuja especificação é baseada na teoria de capital humano minceriana e na hipótese de que cuidar dos afazeres domésticos e dos filhos são as principais alternativas femininas ao trabalho remunerado, particularmente ao exercido fora de casa. As variáveis  $CR018$ ,  $CR06$  e  $CR03$  são, respectivamente, o número de crianças menores de 18, 6 e 3 anos de idade na família;  $EDU$  é o número de anos de escolaridade;  $ID$  é a idade da mulher dividida por 10;  $ID2$  é o valor do quadrado de  $ID$ ; a variável  $BRANCA$  assume o valor de 1 se a cor for branca, e a  $CHEFE$  assume valor 1 se for chefe de família;  $TAM$  é o número de pessoas no domicílio;  $NOSAL$  corresponde ao total das rendas não salariais da pessoa;  $OUTRAS$  é a renda dos demais membros da família;  $URBANA$  é uma variável binária para domicílio em áreas urbanas; e  $R$  é um conjunto de variáveis binárias para regiões geográficas.<sup>4</sup>

A variável dependente  $Z^*$  não é observável, mas, uma vez que se pode observar se o agente trabalha ou não, pode-se verificar se  $Z^* \geq 0$  ou não. Logo,  $Z^*$  possui uma realização dicotômica e observável  $Z$  que está relacionada com  $Z^*$  como a seguir:

$$Z_i = \begin{cases} 1, \text{ se } \left( \underbrace{\left( 1, CR018, CR03, CR06, EDU, ID, ID2, BRANCA, \right.}_{X_i} \left. \begin{matrix} \\ CHEFE, TAM, NOSAL, OUTRAS, URBANA, R \end{matrix} \right)_i \cdot \beta + \varepsilon_i > 0 \\ 0, \text{ caso contrário} \end{cases} \quad (2)$$

A equação (3) é a equação de salário que depende linearmente de um conjunto de características do mercado de trabalho e individuais,  $H \subset X$ . Ou seja, todas as variáveis que afetam o salário individual e o salário de reserva. Formalmente:

$$SAL_i^* = \underbrace{\left( 1, EDU, BRANCA, ID, ID2, URBANA, R \right)_i}_{H_i} \cdot \gamma + u_i \quad (3)$$

4. Seguimos Buchinsky (1998) na escolha das variáveis que determinam a participação das mulheres na força de trabalho.

O viés de seleção amostral ocorre porque somente se observa a variável  $SAL$  descrita na equação (4) em vez de  $SAL^*$ :

$$SAL_i = \begin{cases} SAL_i^* & \text{se } Z_i = 1 \\ 0 & \text{se } Z_i = 0 \end{cases} \quad (4)$$

$$\begin{aligned} E(SAL_i | H_i, Z_i = 1) &= H_i \cdot \gamma_\tau + E(u_i | H_i, Z = 1) = \\ &= H_i \cdot \gamma_\tau + \underbrace{E(u_i | H_i, \varepsilon_i \geq -X_i' \beta)}_{\mu(X' \beta)} \end{aligned} \quad (5)$$

Ao se supor que o termo de erro  $(u, \varepsilon)_i$  tem uma distribuição normal bivariada, então  $\mu(\cdot)$  na equação (5) é igual ao produto entre a inversa da razão de Mill e a covariância entre  $\varepsilon$  e  $u$ . Este é o famoso procedimento de Heckit, introduzido por Heckman (1979), que pode ser utilizado para estimar os coeficientes  $\gamma$  de forma consistente regredindo  $SAL$  com relação a  $H$  e  $\frac{\phi(X' \beta_{probit})}{\Phi(X' \beta_{probit})}$  por mínimos quadrados ordinários (MQO).

## 2.2 Regressão quantílica para o modelo de seleção amostral de Heckman: Buchinsky (1998)

Por meio da regressão quantílica, introduzida por Koenker e Bassett (1978), estimam-se funções quantílicas condicionais, ou seja, estimam-se modelos cujos quantis da distribuição condicional da variável *resposta* são definidos como funções das covariadas observadas. O método de MQO permite que se determine a média condicional de uma variável aleatória,  $SAL^*$ , dadas algumas variáveis explicativas  $H$ . Por sua vez, a regressão quantílica vai além ao permitir esta determinação em qualquer quantil ( $\tau$ ) da função de distribuição condicional. Neste caso, a soma assimetricamente ponderada do valor absoluto dos resíduos é considerada para calcular os parâmetros:<sup>5</sup>

$$\min_{\gamma} \frac{1}{n} \sum_{i: SAL_i^* \geq \gamma H_i} w_i \tau |SAL_i^* - \gamma H_i| + \sum_{i: SAL_i^* < \gamma H_i} w_i (1 - \tau) |SAL_i^* - \gamma H_i| \quad (6)$$

5. Ver Koenker e Bassett (1978) e Koenker e Hallock (2001) para mais detalhes.

Assim, o método de regressão quantílica oferece uma visão mais profunda sobre as relações entre as variáveis na base de dados (KOENKER; HALLOCK, 2001). No caso particular que interessa a este trabalho, pode-se testar se os coeficientes associados aos retornos educacionais e os diferenciais raciais diferem ao longo da distribuição de salários.

A equação (5), formalmente, pode ser modificada para adaptar-se ao arcabouço do quantil condicional( $\tau$ ) como a seguir:

$$\begin{aligned} Quant_{\tau}(SAL_i | H_i, Z_i = 1) &= H_i \cdot \gamma_{\tau} + Quant_{\tau}(u_i | H_i, Z = 1) = \\ &= H_i \cdot \gamma_{\tau} + \underbrace{Quant_{\tau}(u_i | H_i, \varepsilon_i \geq -X_i' \beta)}_{\mu(X' \beta)} \end{aligned} \quad (7)$$

O termo quantil condicional na equação (7) não tem as mesmas propriedades do valor esperado truncado do termo de erro na equação (5). Observe-se que, diferentemente da equação (5), o termo de correção,  $\mu_{\tau}(X' \beta)$ , não é igual ao produto entre a inversa da razão de Mill e a covariância entre  $\varepsilon$  e  $\mu$ , ao se supor que  $(u, \varepsilon)_i$  possui uma distribuição normal bivariada. De fato, o quantil condicional possui uma forma desconhecida que é função de  $X' \beta$ .

### 3 ESTIMADOR SEMIPARAMÉTRICO COM DOIS ESTÁGIOS: NEWEY (1991) E BUCHINSKY (1998)<sup>6</sup>

Buchinsky (1998, 2001) aplica a técnica de regressão quantílica para investigar a oferta de trabalho feminina nos Estados Unidos. A natureza deste problema requer modelos de seleção amostral. Dado que o quantil condicional dos salários observados depende do termo de correção com forma desconhecida, o estimador de dois estágios semiparamétrico é usado conforme sugerido por Heckman (1979) e Newey (1991).

Neste artigo, segue-se um caminho semelhante, na medida em que se estima a equação de participação no primeiro estágio por meio de um método semiparamétrico e prossegue-se para a estimação da equação de salários – adicionando-se o termo de correção derivado do primeiro estágio – por regressão quantílica. O procedimento de dois estágios semiparamétricos utilizados neste trabalho é resumido a seguir.

- 1) Primeiro estágio – estimador semiparamétrico proposto por Gallant e Nychka (1987): estima-se o parâmetro  $\beta$  por quase máxima verossimilhança seguindo

6. Ver detalhes no anexo A.

Gerfin (1996) e Gabler, Laisney e Lechner (1993), que adotam a técnica de estimação proposta por Gallant e Nychka (1987). Nesta abordagem, a estimação é desenvolvida sem se supor uma pré-distribuição condicional especificada para o termo de erro  $(u, \varepsilon)_i$ . O modelo sugere a utilização de um polinômio interpolador de Hermite como aproximação de  $F_{\varepsilon|X}(\cdot)$ , a função de distribuição condicional desconhecida do termo de erro  $\varepsilon_i$ .<sup>7</sup> As estimativas do *probit* paramétrico e semiparamétrico não são diretamente comparáveis porque o último não tem variância unitária. No arcabouço semiparamétrico, um dos coeficientes é fixado em 1. No entanto, conforme demonstrado por De Luca (2008), pode-se usar a razão entre os coeficientes estimados para a comparação. Mais detalhes sobre a primeira etapa do processo de estimação podem ser encontrados no anexo A.

- 2) Segundo estágio – aproximação polinomial do termo de correção: estime-se o parâmetro  $\gamma_\tau$  por uma regressão quantílica do *SAL* em respeito a  $H_i = (1, \text{EDU}, \text{BRANCO}, \text{ID}, \text{ID2}, \text{URBANO}, R)_i$  e em respeito a uma aproximação do termo de seleção  $\mu(X'\beta_{sp})$ , como  $\mu(\vartheta) = \sum_{j=1}^3 \alpha_j \vartheta^{j-1}$ , onde  $\vartheta_i = X_i' \beta_{sp}$ .

#### 4 DADOS

As estimativas são baseadas em uma amostra da PNAD/IBGE de 2007. Ela contém 107.634 observações para mulheres com idade entre 20 e 60 anos. A tabela 1 apresenta estatísticas descritivas para a amostra total, usando os pesos amostrais. Cerca de 61% das mulheres da amostra trabalham em empregos remunerados. Portanto, a variável dependente *log* do salário-hora é censurada para quase 40% da amostra. Em termos de características demográficas, 52% das mulheres eram brancas, 23% eram chefes de família com, em média, oito anos de escolaridade, e 86% viviam em áreas urbanas.

7. Outra forma de relaxar a hipótese de distribuição gaussiana do termo de erro é seguir Klein e Spady (1993). Neste caso, a função de distribuição semiparamétrica *kernel* é utilizada como aproximação de  $F_{\varepsilon|X}(\cdot)$  a função de distribuição condicional desconhecida do termo de erro  $\varepsilon_i$ . Esta técnica é computacionalmente bastante custosa e, apesar dos esforços empreendidos, não foi possível executá-la, dado o grande tamanho da amostra.



TABELA 1  
Estatísticas descritivas da amostra

Variáveis	Observações	Médias	Desvio-padrão	Mínimo	Máximo
Dados ponderados					
CHEFE	107.634	0,23	9,17	0,00	1,00
CR018	107.634	1,22	27,98	0,00	13,00
CR06	107.634	0,36	14,17	0,00	8,00
CR03	107.634	0,17	9,19	0,00	6,00
TAM	107.634	3,95	37,92	1,00	22,00
EDU	107.031	8,00	96,96	0,00	15,00
ID	107.634	37,55	248,04	20,00	60,00
BRANCA	107.634	0,52	10,90	0,00	1,00
URBANA	107.634	0,86	7,49	0,00	1,00
OUTRAS	104.931	1.453,99	53.264,35	0,00	196.000,00
SAL	57.486	1,62	64,35	-5,47	25,47
Z	107.634	0,61	10,65	0,00	1,00
Região Norte	107.634	0,07	5,66	0,00	1,00
Região Nordeste	107.634	0,26	9,56	0,00	1,00
Região Sudeste	107.634	0,45	10,84	0,00	1,00
Região Sul	107.634	0,15	7,79	0,00	1,00
NOSAL	107.634	1,30	51,36	0,00	11,66

Fonte: PNAD/IBGE de 2007.

Elaboração dos autores.

## 5 RESULTADOS

A tabela 2 apresenta as estimativas dos coeficientes e dos respectivos erros-padrão para a equação de participação obtidas por dois métodos: o modelo paramétrico (*probit*) e o modelo semiparamétrico (o coeficiente de idade é fixado em uma unidade).<sup>8</sup>

Os dois métodos produziram resultados muito semelhantes. As variáveis positivamente relacionadas com o salário de reserva, como o número de filhos na família – *CH018*, *CH06*, *CH03* –, os rendimentos não laborais do indivíduo – *NOSAL* – e a renda total dos outros membros da família – *OUTRAS* – são significativas e afetam negativamente a decisão de participar no mercado de trabalho. O coeficiente da variável relacionada à educação (*EDU*) é positivo e é estatisticamente diferente de zero. O efeito da idade apresenta uma forma de *U* invertido, indicando que a participação no mercado de trabalho aumenta com a idade, mas em uma determinada idade começa a diminuir. Os resultados também indicam que as mulheres brancas e as que vivem em áreas urbanas são menos propensas a participar do mercado de trabalho, enquanto as mulheres chefes de família são mais propensas a participar do mercado de trabalho.

8. O estimador semiparamétrico identifica os coeficientes apenas em termos relativos e, portanto, fixou-se o coeficiente IDADE em 1.

TABELA 2

**Coefficientes e erros-padrão da equação de participação**

Variáveis	Probit		Gallant e Nychka	
	Coefficientes	Erro-padrão	Coefficientes	Erro-padrão
<i>Intercepto</i>	-1,4855	0,0026	0,1648327	0
<i>BRANCA</i>	-0,0291	0,0004	-0,22497	0,003811
<i>IDO</i>	0,1007	0,0001	1,00000	-
<i>ID2</i>	-0,0013	0,0000	-0,01297	2,48E-06
<i>EDU</i>	0,0611	0,0001	0,59963	0,000894
<i>TAM</i>	-0,00005	0,0002	-0,0229	0,001644
<i>CR018</i>	0,0253	0,0003	0,181029	0,002505
<i>CR06</i>	-0,0885	0,0005	-0,91336	0,004839
<i>CR03</i>	-0,1576	0,0006	-1,64769	0,006804
<i>Região Norte</i>	-0,1083	0,0010	-1,0087	0,009248
<i>Região Nordeste</i>	-0,0299	0,0008	-0,32931	0,007294
<i>Região Sudeste</i>	-0,0202	0,0008	-0,14173	0,006859
<i>Região Sul</i>	0,1691	0,0009	1,6216	0,007996
<i>URBANA</i>	-0,4784	0,0006	-4,17314	0,007573
<i>CHEFE</i>	0,3651	0,0005	3,458954	0,006372
<i>NOSAL</i>	-0,077	0,0001	-0,76418	0,001349
<i>OUTRAS</i>	-2,02E-05	1,91E-06	-0,00025	1,05E-06

Fonte: PNAD/IBGE de 2007.

Elaboração dos autores.

As tabelas 3 e 4 apresentam as estimativas de regressão quantílica para os retornos educacionais e os diferenciais raciais.<sup>9</sup> As estimativas de regressão quantílica sem termo de correção e com um termo de correção paramétrico<sup>10</sup> são apresentadas apenas como referência. Os resultados são ilustrados nos gráficos 1 e 2.

Nota-se que independentemente dos três métodos utilizados, após o quantil 0,30, os coeficientes associados à variável educação começam a aumentar, e no quantil 0,6 tornam-se maiores do que o retorno educacional médio estimado por MQO. Quanto aos coeficientes associados com o diferencial racial, estes mostram uma forma semelhante à descrita pelos retornos educacionais: caem até o quantil 0,30 e então passam a aumentar à medida que se sobe na distribuição de salários, sendo maior do que a estimativa de MQO a partir do sexto decil.

9. Os resultados apresentados nas tabelas 3 e 4 foram obtidos utilizando-se a rotina do SAS QUANTREG. Os intervalos de confiança (ICs) são baseados em erros-padrão estimados por *bootstrap*.

10.  $\mu(\vartheta) = \sum_{j=1}^3 \alpha_j \vartheta^{j-1}$ , onde  $\vartheta_i = X_i \beta_{probit}$ .

TABELA 3

**Resultados da regressão quantílica (coeficientes e ICs): retornos educacionais**

Quantis	EDU			EDU			EDU		
	(sem termo de seleção)			(termo de seleção paramétrico)			(termo de seleção semiparamétrico)		
	Coeficientes	IC 95%		Coeficientes	IC 95%		Coeficientes	IC 95%	
0,1	0,1084	0,1049	0,1119	0,0971	0,0920	0,1021	0,0976	0,0930	0,1022
0,2	0,1024	0,0997	0,1052	0,0964	0,0926	0,1002	0,0969	0,0931	0,1006
0,3	0,1004	0,0981	0,1027	0,0967	0,0932	0,1002	0,0975	0,0938	0,1012
0,4	0,1041	0,1022	0,1060	0,1008	0,0979	0,1037	0,1018	0,0988	0,1048
0,5	0,1085	0,1067	0,1103	0,1079	0,1054	0,1104	0,1090	0,1065	0,1115
0,6	0,1144	0,1125	0,1163	0,1146	0,1119	0,1173	0,1162	0,1133	0,1191
0,7	0,1197	0,1177	0,1216	0,1202	0,1170	0,1233	0,1215	0,1189	0,1241
0,8	0,1272	0,1251	0,1293	0,1297	0,1263	0,1331	0,1318	0,1285	0,1350
0,9	0,1371	0,1341	0,1402	0,1453	0,1420	0,1487	0,1456	0,1421	0,1491

Fonte: PNAD/IBGE de 2007.

Elaboração dos autores.

TABELA 4

**Resultados da regressão quantílica (coeficientes e ICs ): diferenciais raciais**

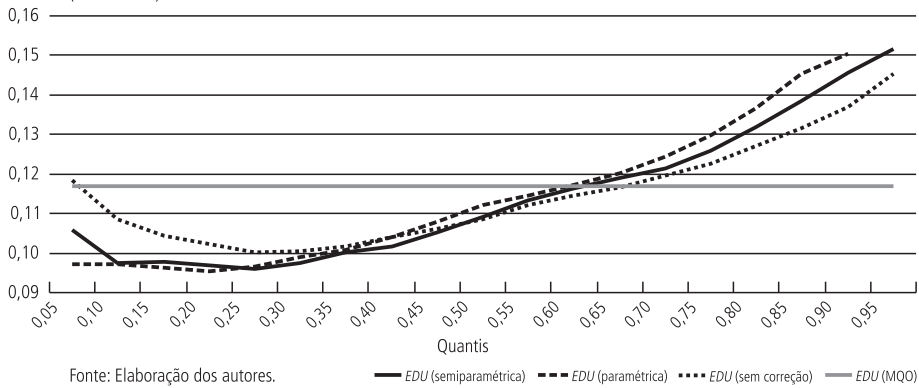
Quantis	BRANCA			BRANCA			BRANCA		
	(sem termo de seleção)			(termo de seleção paramétrico)			(termo de seleção semiparamétrico)		
	Coeficientes	IC 95%		Coeficientes	IC 95%		Coeficientes	IC 95%	
0,1	0,0994	0,0788	0,1200	0,1063	0,0856	0,1270	0,1053	0,0837	0,1269
0,2	0,0988	0,0831	0,1145	0,1019	0,0862	0,1175	0,1002	0,0853	0,1151
0,3	0,1018	0,0872	0,1164	0,1032	0,0895	0,1168	0,1026	0,0879	0,1173
0,4	0,1072	0,0930	0,1215	0,1078	0,0943	0,1213	0,1083	0,0956	0,1209
0,5	0,1221	0,1075	0,1366	0,1161	0,1026	0,1296	0,1160	0,1012	0,1308
0,6	0,1336	0,118	0,1492	0,1293	0,1128	0,1459	0,1276	0,1115	0,1436
0,7	0,1526	0,1332	0,1719	0,1430	0,1228	0,1632	0,1431	0,1243	0,1619
0,8	0,1598	0,1351	0,1845	0,1525	0,1341	0,1708	0,1509	0,1290	0,1727
0,9	0,1612	0,1217	0,2006	0,1543	0,1242	0,1844	0,1541	0,1229	0,1853

Fonte: PNAD/IBGE de 2007.

Elaboração dos autores.

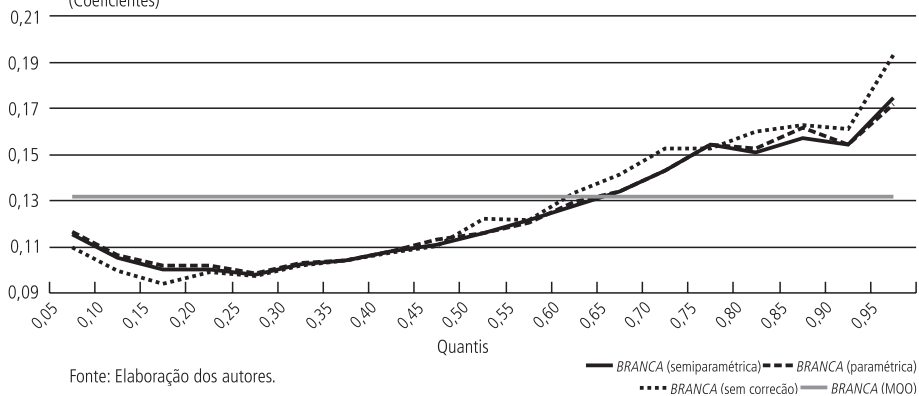
Percebe-se que as estimativas dos retornos educacionais semiparamétricas ao longo da distribuição de salários tendem a ser localizadas entre as estimativas dos dois modelos paramétricos com e sem correção. O gráfico 1 mostra que, enquanto para os quantis mais baixos a estimativa paramétrica sem correção tende a superestimar os retornos da educação, para os quantis mais elevados os retornos educacionais são subestimados. Os ICs de 95% para o modelo com correção semiparamétrica cruzam com a estimativa com correção paramétrica para todos os quantis relatados, mas não coincidem com as do modelo sem correção para o primeiro e último decil da distribuição salarial.

GRÁFICO 1  
Retornos educacionais: estimativas da regressão quantílica  
(Coeficientes)



Quanto ao coeficiente associado à cor, os resultados semiparamétricos no gráfico 2 mostram que as estimativas paramétricas sem correção levam a uma superestimação do impacto da discriminação no quantil superior (0,95). No entanto, os ICs de 95% para as três estimativas se sobrepõem uns aos outros, tornando as estimativas muito semelhantes.

GRÁFICO 2  
Diferenciais raciais: estimativas da regressão quantílica  
(Coeficientes)



Para o modelo semiparamétrico, fez-se um teste para ver se os coeficientes são diferentes entre quantis, de modo que se pode afirmar que ambos os retornos educacionais e diferenciais raciais tendem a ser maiores em quantis mais elevados. Os resultados são apresentados nas tabelas B.1 e B.2 no anexo B. Mostra-se claramente que para os retornos educacionais a partir do quantil 0,40 pode-se rejeitar qualquer igualdade dois a dois entre coeficientes de quantis diferentes. Resultados

semelhantes foram encontrados para os coeficientes associados à cor, salvo o coeficiente no quantil 0,10, que apenas se mostram inferiores aos coeficientes dos quantis superiores ao quantil 0,60. Entre os quantis mais elevados (entre 0,90 e 0,80 quantis), não há diferença na comparação de pares de coeficientes.

## 6 CONCLUSÕES

As estimativas para os retornos educacionais confirmam o fato estilizado sobre o elevado retorno oriundo dos investimentos em educação observado no Brasil, mesmo para o subgrupo populacional de mulheres entre 20 e 60 anos de idade. Elas também revelam que a análise do retorno condicional médio da educação esconde um importante fato: de que este retorno não é constante ao longo da distribuição salarial. Na verdade, os retornos educacionais têm uma forma de  $U$  assimétrico para a esquerda, revelando que para os quantis mais baixos os retornos educacionais diminuem antes de aumentarem nos quantis mais elevados. No modelo semiparamétrico, o retorno é bastante estável por volta de 10%<sup>11</sup> nos quantis de 0,1 até 0,3. No entanto, cresce de forma constante a partir do quantil 0,3, atingindo o máximo de 16% no quantil 0,9.

Este resultado está em consonância com a literatura, que apresenta retornos mais elevados no topo da distribuição (MACIEL; CAMPÊLO; RAPOSO, 2001; MARTINS; PEREIRA, 2004; BARTALOTTI; LEME, 2007). Ademais, revela que, mesmo que a desigualdade entre as mulheres seja menor que entre os homens, é possível que um aumento da participação da oferta de trabalho feminina leve a uma diminuição da desigualdade salarial concomitantemente a um aumento da desigualdade total. Fatores não facilmente observáveis, como a qualidade da educação, ou complementares a esta, como redes sociais, e outras externalidades relativas à educação podem explicar o aumento do retorno da educação ao longo da distribuição de salários. Tal fato requer um cuidado especial das políticas públicas no sentido de melhorar a qualidade da educação, diminuir sua heterogeneidade e, ao mesmo tempo, garantir às mulheres de famílias pobres maiores possibilidades de acesso a postos de trabalho melhor remunerados. Tais políticas incluem não somente treinamento e aquisição de habilidades gerais e específicas, mas também melhor oferta de equipamentos públicos como creches e escolas, que permitam às mães retornarem ao mercado de trabalho com menores privações em termos de tempo e custos domésticos.

Com relação às diferenças entre os resultados da correção semiparamétrica e o modelo sem correção para o caso do retorno da educação, fica evidente que a não correção do viés superestima os retornos da educação para os quantis mais baixos e subestima-os para os quantis mais elevados. No entanto, não há diferença entre

---

11. O diferencial salarial é calculado exponencializando-se o coeficiente e subtraindo-se uma unidade.

os resultados dos modelos paramétrico e semiparamétrico. As estimativas pontuais são muito parecidas e os ICs se interceptam em todos os quantis. Este resultado sugere que as correções paramétrica e semiparamétrica do viés de seleção no âmbito da análise de regressão quantílica apresentam resultados muito semelhantes. Portanto, ao contrário de Buchinsky (2001), pelo menos no caso brasileiro, a correção semiparamétrica não revela nenhum viés na correção paramétrica para o retorno da educação no contexto de regressões quantílicas.

Com relação aos coeficientes associados à cor, os resultados mostram comportamentos semelhantes ao observado para o caso da educação. No entanto, para isto a seleção amostral não parece afetar as estimativas dos diferenciais raciais de salários, dado que não há diferença substancial entre as três estimativas analisadas. No que diz respeito às variações ao longo da distribuição salarial, os diferenciais raciais de salários parecem ser muito maiores em quantis mais elevados. Isto leva à conclusão de que entre os trabalhadores do sexo feminino a discriminação racial de salários é mais flagrante. Ou seja, é possível que exista um teto de vidro para mulheres negras nas ocupações melhor remuneradas. O diferencial racial de salários de acordo com o modelo semiparamétrico é de 11% no quantil 0,1 e de 17% no quantil 0,9. Este resultado está alinhado, por exemplo, com Arcand e D’Hombres (2004) e Bartalotti e Leme (2007), que encontraram que a discriminação racial – aproximada pelo diferencial racial – seria mais forte nos quantis mais elevados. Além disso, Bartalotti e Leme (2007) mostram que a diferença salarial entre mulheres brancas e negras é maior no topo da distribuição salarial. Apesar de os autores não corrigirem para o viés de seleção amostral, os resultados do presente texto indicam que as três estimativas diferentes são muito semelhantes. Ou seja, neste caso, a seleção amostral não parece enviesar o coeficiente associado à cor.

Em resumo, este texto tem duas mensagens importantes. A primeira refere-se a uma questão técnica: trabalhar com uma distribuição paramétrica condicional errada dos erros do modelo para corrigir o viés de seleção pode levar a vários problemas na estimativa do retorno da educação nas equações salariais. Ou seja, pode levar à superestimação dos retornos educacionais em quantis mais baixos e à subestimação de quantis mais elevados. O mesmo não ocorre para estimativas de diferenciais raciais. A segunda mensagem refere-se à necessidade de se olhar para a distribuição condicional de salário por inteiro para se obter uma visão mais clara do impacto heterogêneo dos retornos educacionais e diferenciais raciais de salários. Os resultados da regressão quantílica mostram que os retornos educacionais tendem a ser maiores para as mulheres no topo da distribuição salarial. Da mesma forma, a discriminação racial tende a ser mais grave para as mulheres na parte superior da distribuição salarial. Este último resultado evidencia a existência de um teto de vidro para as negras na distribuição de salários das mulheres. Neste contexto, políticas que melhorem a qualidade da educação e assegurem um acesso maior das

mulheres negras mais educadas a postos mais bem remunerados são necessárias para garantir uma melhora na desigualdade salarial que se reverta também em melhora na desigualdade total dos rendimentos.

## ABSTRACT

We estimate the returns to education for women and the racial wage differential among women over the wage distribution in Brazil by using quantile regression with semiparametric correction for sample selection. Our estimates show that the returns to education are high and that they are not constant along the wage distribution. Both returns to education and the racial wage differentials are higher at higher points of (the conditional) wage distribution. Black women seem to be facing a glass-ceiling in the higher wage segment of the distribution. In addition, quality of education seems also to play some role in the inequality observed among higher paid women. The paper also reveals that using a probit specification to the participation equation in order to correct selection issues yields very similar results to the semiparametric correction for both returns to education and racial discrimination.

## REFERÊNCIAS

- ARCAND, J.; D'HOMBRES, B. Racial discrimination in the Brazilian labour market: wage, employment and segregation effects. *Journal of International Development*, v. 16, n. 8, p. 1.053-1.066, 2004.
- BARTALOTTI, O.; LEME, M. C. S. *Discriminação salarial por cor e gênero revisitada: uma abordagem de decomposição contrafactual utilizando regressões quantílicas*. 2007. Mimeografado. Disponível em: <<http://www.eesp.fgv.br/pessoas/alunos/mestrado-doutorado/A/1/36/>>
- BUCHINSKY, M. The dynamics of changes in the female wage distribution in the USA: a quantile regression approach. *Journal of Applied Econometrics*, n. 13, p. 1-30, 1998.
- \_\_\_\_\_. Quantile regression with sample selection: estimating women's return to education in the U.S. *Empirical Economics*, n. 26, p. 87-113, 2001.
- COSTA, J.; SILVA, E.; VAZ, F. *The role of gender inequalities in explaining income growth, poverty and inequality: evidence from Latin American countries*. International Policy Centre for Inclusive Growth, Apr. 2009 (Working Paper, n. 52).
- DE LUCA, G. SNP and SML estimation of univariate and bivariate binary-choice models. *The Stata Journal*, n. 8, p. 190-220, 2008.
- GABLER, S.; LAISNEY, F.; LECHNER, M. Semiparametric estimation of binary-choice models with an application to labour-force participation. *Journal of Business and Economic Statistics*, n. 11, p. 61-80, 1993.
- GALLANT, A. R.; NYCHKA, D. N. Semi-nonparametric maximum likelihood estimation. *Econometrica*, n. 55, p. 363-390, 1987.
- GERFIN, M. Parametric and semi-parametric estimation of binary response model of labour market participation. *Journal of Applied Econometrics*, n. 11, p. 321-339, 1996.
- HECKMAN, J. Sample selection as a specification error. *Econometrica*, n. 47, p. 153-161, 1979.
- HOFFMAN, R.; LEONE, T. Participação da mulher no mercado de trabalho e desigualdade da renda domiciliar *per capita* no Brasil: 1981-2002. *Nova Economia*, v. 14, n. 2, p. 35-58, 2004.

KLEIN, R.W.; SPADY, R. H. An efficient semiparametric estimator for binary response models. *Econometrica*, n. 61, p. 387-421, 1993.

KOENKER, R.; BASSETT, G. W. Regression quantiles. *Econometrica*, n. 46, p. 33-50, 1978.

\_\_\_\_\_.; HALLOCK, K. F. Quantile regression. *Journal of Economic Perspectives*, n. 15, p. 143-156, 2001.

MACIEL, M. C.; CAMPÊLO, A. K.; RAPOSO, M. C. F. A dinâmica das mudanças na distribuição salarial e no retorno em educação para mulheres: uma aplicação de regressão quantílica. In: ENCONTRO NACIONAL DE ECONOMIA, 21., 2001. *Anais...* Salvador: ANPEC, 2001.

MARTINS, M. F. O. Parametric and semiparametric estimation of sample selection models: an empirical application to the female labour force in Portugal. *Journal of Applied Econometrics*, n. 16, p. 23-39, 2001.

MARTINS, P. S.; PEREIRA, P. T. Does education reduce wage inequality? Quantile regressions evidence from 16 countries. *Labour Economics*, v. 11, n. 3, p. 355-371, 2004.

NEWKEY, W. *Two step series estimation of sample selection model*. MIT, 1991. Mimeografado.

SANTOS, R. V.; RIBEIRO, E. P. *Diferenciais de rendimentos entre homens e mulheres no Brasil revisitado: explorando o "Teto de Vidro"*. In: SEMINÁRIOS DE PESQUISA, 2006. Disponível em: <[www.eg.fjp.pg.gov.br/seminarioiv/download/vale.pdf](http://www.eg.fjp.pg.gov.br/seminarioiv/download/vale.pdf)>

SCORZAFAVE, L. *Caracterização da inserção feminina no mercado de trabalho e seus efeitos sobre a distribuição de renda*. 2004. Tese (Doutorado) – Instituto de Pesquisa Econômica da Faculdade de Ciências Econômicas da Universidade de São Paulo, São Paulo, 2004.

\_\_\_\_\_.; MENEZES FILHO, N. Impacto da participação das mulheres na evolução da distribuição da renda do trabalho no Brasil. *Pesquisa e Planejamento Econômico*, v. 35, n. 2, p. 245-266, 2005.

TANURI-PIANTO, M.; PIANTO, D. M. *Formal-informal earnings differentials in Brazil: a semiparametric approach*. 2002. Mimeografado. Disponível em: <<http://epge.fgv.br/portal/arquivo/1098.pdf>>

(Originais submetidos em março de 2008. Última versão recebida em abril de 2010. Aprovada em abril de 2010.)

## ANEXO A

A aproximação polinomial de Hermite é dada por  $h^*(u) = \sum_{i,j=0}^k \alpha_i \alpha_j u^{i+j} \exp[-(u/\delta)^2]$ .

Segue-se a normalização escolhida por Gabler, Laisney e Lechner (1993) e fixa-se  $\delta = \sqrt{2}$  e  $\alpha_0 = (2\pi)^{-1/4}$ . O parâmetro  $k$  fornece flexibilidade para a distribuição do termo de erro. À medida que  $k$  aumenta, aumenta-se o leque das distribuições consideradas, diferentemente do que ocorre na distribuição normal do modelo *probit*.



De Luca (2008) sugere a fixação de  $k$  pela aplicação de alguns critérios de seleção do modelo (por exemplo, os critérios de informação de Akaike e o bayesiano). No entanto, conforme apontado por Gerfin (1996) e Gabler, Laisney e Lechner (1993), o processo de estimação torna-se pesado e tende a não convergir para o limite máximo global à medida que  $k$  aumenta. Por um lado, especula-se que provavelmente (muito) mais observações são necessárias para que se possa escolher o valor do parâmetro correto. Por outro lado, argumenta-se que  $k = 3$  (o menor valor possível para o parâmetro) não é má escolha na prática.

Com esta especificação, a função de probabilidade pode ser definida como

$$F(X_n \gamma | X_n) = S^{-1} \int_{-X_n \gamma}^{\infty} h^*(u) du, \text{ sendo que:}$$

$$S = \int_{-\infty}^{\infty} h^*(u) du \text{ e } \underbrace{\left( \begin{array}{l} 1, CR018, CR03, CR06, EDU, ID, ID2, BRANCA, CHEFE, \\ TAM, NOSAL, OUTRAS, URBANA, R \end{array} \right)}_{\tilde{x}_i}$$

Para a estimação, usou-se o comando SNP do STATA desenvolvido por De Luca (2008).

## ANEXO B

TABELA B.1

### Teste de Wald para diferenças entre os coeficientes de *EDU* entre os quantis do modelo semiparamétrico

	$q = 0,2$	$q = 0,3$	$q = 0,4$	$q = 0,5$	$q = 0,6$	$q = 0,7$	$q = 0,8$	$q = 0,9$
$q = 0,1$	0,36502	0,04548	-1,78833	-4,76260	-7,34282	-9,03923	-11,8454	-16,54719
$q = 0,2$		-0,47078	-3,12444	-7,11413	-9,67195	-11,3164	-14,1423	-19,06323
$q = 0,3$			-4,34919	-9,33534	-11,8897	-13,3054	-16,3825	-21,12105
$q = 0,4$				-8,15305	-11,0789	-12,4040	-15,4860	-19,97940
$q = 0,5$					-7,85693	-9,51669	-13,2453	-17,74780
$q = 0,6$						-5,18583	-10,5122	-14,83515
$q = 0,7$							-8,41607	-12,93223
$q = 0,8$								-8,07499

Fonte: PNAD/IBGE de 2007.

Elaboração dos autores.

TABELA B.2

**Teste de Wald para diferenças entre os coeficientes de *BRANCA* entre os quantis do modelo semiparamétrico**

	$q = 0,2$	$q = 0,3$	$q = 0,4$	$q = 0,5$	$q = 0,6$	$q = 0,7$	$q = 0,8$	$q = 0,9$
$q = 0,1$	0,61354	0,28195	-0,29534	-0,98676	-1,89811	-3,13169	-3,41703	-2,87258
$q = 0,2$		-0,39884	-1,15214	-2,03920	-3,13955	-4,45697	-4,55794	-3,48114
$q = 0,3$			-1,18331	-2,13954	-3,22455	-4,40026	-4,53174	-3,42872
$q = 0,4$				-1,76977	-3,02369	-4,29718	-4,23918	-3,01860
$q = 0,5$					-2,30225	-3,80962	-3,81815	-2,53336
$q = 0,6$						-2,65237	-2,65735	-1,77844
$q = 0,7$							-1,15189	-0,78864
$q = 0,8$								-0,26600

Fonte: PNAD/IBGE de 2007.

Elaboração dos autores.