O preço da cor: diferenciais raciais na distribuição da renda no Brasil – réplica

NELSON DO VALLE SILVA *

O viés das variáveis omitidas e o viés da crítica

Moura Castro, em seu ensaio crítico sobre artigo de nossa autoria publicado anteriormente na *PPE,* [1] após enfatizar que não tem em relação a nosso trabalho "divergências doutrinárias ou de posturas teóricas" (o que nos deixa bastante aliviados), levanta o problema do possível efeito que variáveis não incluídas na análise poderiam ter nos resultados obtidos. De fato, o "viés de variáveis omitidas" é um problema básico de todo teste empírico de hipóteses e, dado um caso concreto, sempre se pode levantar uma lista de variáveis não incluídas que eventualmente produziriam um efeito importante na análise. No caso presente, Moura Castro lembra a importância de variáveis que caracterizam o *status* sócio-econômico da família de origem, argumentando que "é possível que, se incluídas na regressão, as variáveis de *status* captassem toda a variância de renda que o autor imputa à raça". Creio que podemos concordar com Moura Castro quanto à importância da omissão do "*status* familiar de origem" para a análise efetuada. No entanto, razões há que nos levam a acreditar que a omissão de variáveis desse tipo para a análise de diferenças raciais na distribuição de rendimentos são diversas, e mesmo opostas, às de Moura Castro, como tentamos mostrar a seguir.

* Do IBGE/SEPLAN.

1 Ver Nelson do Valle Silva, "O Preço da Cor: Diferenciais Raciais na Distribuição da Renda no Brasil", *in Pesquisa e Planejamento Econômico,* vol. 10, n.º 1 (abril de 1980), pp. 21-44.

Pesq. Plan. Econ.	Rio de Janeiro,	10 (3)	1007 a 1012	dez. 1980

Para o efeito de fixar a notação e simplificar a argumentação, vamos supor que tenhamos apenas duas variáveis, além de "rendimentos" (Y): "nível de escolaridade" (E) e a variável omitida "*status* familiar" (SF). Além disso, vamos admitir que são verdadeiras as seguintes hipóteses:

a) o *status* familiar (SF) produz não só um efeito direto, como um importante efeito indireto através do nível de escolaridade (E) sobre a variável rendimentos; e

b) para um dado nível de *status* familiar, o nível médio de escolaridade do grupo não-branco é menor ou igual ao do grupo branco.

A primeira hipótese corresponde ao argumento adiantado por Moura Castro; a segunda, que nos parece altamente plausível, implica dizer que os não-brancos não têm nenhuma vantagem relativa *vis-à-vis* os indivíduos do grupo branco no processo de escolarização, ou seja, o grupo não-branco está por hipótese em situação de igualdade ou de desvantagem competitiva no que diz respeito à educação.

Pela primeira hipótese, o modelo de rendimento "verdadeiro" é:

$$E = c_o + c_1\ SF + u \tag{1}$$

$$Y = b_o + b_1\ E + b_2\ SF + r \tag{2}$$

enquanto que, por não dispormos de informações relativas a SF em nossos dados, especificamos incorretamente a função rendimento, usando:

$$Y = a_o + a_1\ E + w \tag{3}$$

Seguindo as hipóteses acima, temos que $b_2 > 0$ e $a_1 > b_1 > 0$. Supondo que SF e E são fortes e positivamente correlacionadas, é plausível a hipótese 3 de que:[2]

$$a_1\ c_1 \cong b_1\ c_1 + b_2 \tag{4}$$

do que resulta:

$$c_1 \cong \frac{b_2}{a_1 - b_1} \tag{5}$$

[2] S. Masters, *Black-White Income Differentials* (Nova York: Seminar, 1975), p. 129.

Nosso objetivo será mostrar aqui que, dadas as hipóteses apresentadas acima, a nossa medida D de discriminação no mercado de trabalho será provavelmente maior quando usamos a especificação da equação 2 do que quando omitimos a variável SF, como na equação 3. Para isso, precisamos estabelecer mais alguns detalhes de notação. Escrevamos:

$$\overline{Y}_b = b^b_o + b^b_1 \overline{E}_b + b^b_2 \overline{SF}_b \tag{6}$$
$$\overline{Y}_n = b^n_o + b^n_1 \overline{E}_n + b^n_2 \overline{SF}_n$$

onde $\overline{Y}$, $\overline{E}$ e $\overline{SF}$ representam as médias das respectivas variáveis e os subscritos b e n representam os grupos branco e não-branco, respectivamente. A padronização da variável rendimentos pode ser escrita como:

$$f_b\,(\overline{N}) = b^b_o + b^b_1 \overline{E}_n + b^b_2 \overline{SF}_n \tag{8}$$

e a medida de discriminação D é:

$$D = (b^b_o - b^n_o) + \overline{E}_n\;(b^b_1 - b^n_1) + \overline{SF}_n\;(b^b_2 - b^n_2) = f_b\;(\overline{N}) - \overline{Y}_n \tag{9}$$

É razoável supor que $\overline{Y}_b > f_b\;(\overline{N}) > \overline{Y}_n$. Então provar que $D = f_b\;(\overline{N}) - Y_n$ é menor quando baseado na função de rendimentos "errada" – equação (3) – do que quando baseada na "verdadeira" – equação (2) – equivale a provar que $\varepsilon = \overline{Y}_b - f_b\;(\overline{N})$ é maior na mesma situação. Baseado na função de rendimentos "errada", temos (abandonando os superscritos dos coeficientes, uma vez que só usaremos as equações para o grupo branco):

$$\varepsilon^* = \overline{Y}_b - f_b\;(\overline{N}) = (a_o + a_1\;\overline{E}_b + w) - (a_o + a_1\;\overline{E}_n + w) = a_1\;(\overline{E}_b - \overline{E}_n) \tag{10}$$

enquanto que, baseados na função "verdadeira", teríamos:

$$\varepsilon = \overline{Y}_b - f_b\;(\overline{N}) = b_1\;(\overline{E}_b - \overline{E}_n) + b_2\;(\overline{SF}_b - \overline{SF}_n) \tag{11}$$

Donde:

$$\varepsilon^* - \varepsilon = (a_1 - b_1)\;(\overline{E}_b - \overline{E}_n) - b_2\;(\overline{SF}_b - \overline{SF}_n) \tag{12}$$

Agora:

$$\overline{E}_b = c_o + c_1 \overline{SF}_b \tag{13}$$

e

$$E_b (\overline{N}) = c_o + c_1 \overline{SF}_n \tag{14}$$

de tal forma que:

$$\overline{E}_b - E_b (\overline{N}) = c_1 (\overline{SF}_b - \overline{SF}_n) \tag{15}$$

Pela segunda hipótese, temos que E_b $(\overline{N}) > E_n$ e, portanto:

$$\overline{E}_b - \overline{E}_n > \overline{E}_b - E_b (\overline{N}) = c_1 (\overline{SF}_b - \overline{SF}_n) \tag{16}$$

Então:

$$\overline{E}_b - \overline{E}_n > c_1 (\overline{SF}_b - \overline{SF}_n) \tag{17}$$

Assim:

$$\varepsilon^* - \varepsilon > (a_1 - b_1) c_1 (\overline{SF}_b - \overline{SF}_n) - b_2 (\overline{SF}_b - \overline{SF}_n) \tag{18}$$

Usando a equação (5), temos então que aproximadamente:

$$\varepsilon^* - \varepsilon > (a_1 - b_1) \frac{b_2}{(a_1 - b_1)} (\overline{SF}_b - \overline{SF}_n) - b_2 (\overline{SF}_b - \overline{SF}_n) \tag{19}$$

do que resulta naturalmente que:

$$\varepsilon^* - \varepsilon > 0$$

e, portanto, nossa medida D é provavelmente uma subestimativa do "verdadeiro" nível de discriminação racial no mercado de trabalho.

Essas implicações derivadas a um nível puramente lógico parecem que são confirmadas também ao nível empírico. De fato, algumas medidas da situação nos Estados Unidos sugerem que a inclusão de variáveis de *status* familiar de origem pode aumentar as estimativas do efeito de discriminação sobre a distribuição da renda. Por exemplo, Blinder[3] comunica uma estimativa inicial de cerca de 40% para o efeito de discriminação nos diferenciais raciais de rendimentos. Quando são introduzidos *status* familiar e outras variáveis exógenas, a estimativa baseada no modelo em forma reduzida aumenta para cerca de 70%. Resultados semelhantes são obtidos por Masters e Duncan.[4]

Em conclusão, podemos concordar com Moura Castro que a omissão (inevitável, porque não se dispunha delas na base de dados utilizada, o Censo de 1960) de variáveis indicativas do *status* da família de origem pode alterar significativamente os resultados obtidos. Infelizmente, não podemos concordar que essa importância se deva a que possivelmente "as variáveis de *status* captassem toda variância de renda que o autor imputa à raça". Bem pelo contrário, tudo leva a crer que, de fato, a omissão dessas importantes variáveis na realidade nos conduziria a estimativas ainda mais elevadas do nível do efeito da discriminação na distribuição da renda.

3 A. Blinder, "Wage Discrimination: Reduced Form and Structural Estimates", *in Journal of Human Resources*, vol. 7 (1973), pp. 436-455.

4 S. Masters, *op. cit.*, e O. D. Duncan, "Inheritance of Poverty or Inheritance of Race?", *in* D. P. Moynihan (ed.), *On Understanding Poverty* (Nova York: Seminar, 1975).